MATEMATIČKI FAKULTET, UNIVERZITET U BEOGRADU

Primena metoda istraživanja podataka u razvoju modela zasnovanih na rejtingu-sklonost ka kupovini

Master rad

Mentor:
prof. dr. Nenad Mitić

Student:
Vladimir Marković

Predgovor

U radu je prikazana metodologija izrade matematičkih modela koji se koriste kao podrška prodaji u bankarskoj industriji. Objašnjeno je kako se definiše poslovni problem i priprema uzorak za razvoj modela. Nad pripremljenim podacima urađene su razne statističke analize i opisane metode redukovanja i izbora promenljivih. Nad izabranim promenljivama autor je razvio 10 matematičkih modela zasnovanih na logističkoj regresiji. Na kraju opisan je izbor najboljeg modela kao i njegova primena u sistemu „sledeća najbolja ponuda za klijenta“.

Analitiku predstavlja 6 datoteka (tabela) iste strukture na nivou klijenta sa 2043 promenljive izračunate u 7 vremenskih trenutaka/perioda sa 2,5 miliona opservacija. Za pripremu podataka korišćena je trogodišnja istorija poslovanja banke počev od transakcija klijenata, agregacija na nivou računa i klijenta pa do eksternih izvora kao što je kreditni biro. Za pripremu podataka autor je utošio 5 nedelja. Za preliminarne statistike, izbor i redukovanje promenljivih autor je potrošio nedelju dana, a za razvoj i ocenu modela 2 nedelje.

Zahvaljujem se mom poslodavcu Banka Intesa Beograd koja mi je omogućila da koristim hardverske i softverske resurse prilikom izrade ovog rada i koja mi je omogućila da primenom najsavremenijih tehnologija unapredim svoja znanja i veštine. Posebno se zahvaljujem Tamari Stanojević, direktorki CRM odeljenja, koja je pomogla da dobijem neophodne dozvole da rad objavim.

U skladu sa bezbednosnim principima kompanije u kojoj radim u ovom radu se ne prikazuju podaci niti je jasno opisana koja je ciljna grupa (uzorak) nad kojom je matematički model razvijan. Regresiona funkcija koja je rezultat modelovanja takođe nije prikazana. Statistički grafikoni prikazani u radu napravljeni su iz uzorka nad pomenutom ciljnom grupom i ne predstavljaju reprezentativni uzorak na nivou banke. Programski kod za *ETL* i *SAS Enterprise Miner* projekat nisu sastavni deo ovog rada i predstavljaju poslovnu tajnu i vlasništvo Banke Intese Beograd.

Svom mentoru prof.dr. Nenadu Mitiću se zahvaljujem na svesrdnoj pomoći, savetima i razumevanju koje mi je pružio tokom izrade ovog rada.

Zahvaljujem se svojim prijateljimi i kolegama, koji su mi na bilo koji način pružili pomoć i podršku u izradi ovog rada.

Posebno se zahvaljujem mojoj porodici na neizmernoj podršci, razumevanju i strpljenju koju su imali u toku izrade ovog rada.

Posvećeno ćerci Jani i supruzi Jeleni.

Beograd, jun 2014. godine Vladimir Marković

Sadržaj

Spisak slika

Spisak tabela

1 Uvod

Osnovni strateški cilj svake banke koja se bavi prodajom proizvoda ili servisa je uvećanje tržišnog udela tj. povećanje prodaje, a sami tim i profita. Tri osnovna pristupa u realizaciji ovog cilja su:

- uvećanje prodaje korišćenjem postojeće baze klijenata,
- proboj u okviru postojećeg tržišta akvizicijom novih klijenata,
- osvajanje potpuno novog tržišta.

Često vlada uverenje da je mnogo lakše prodati proizvod klijentu koji koristi ili je koristio vaše proizvode. Ovo je posebno izraženo u bankarskom sektoru, gde je bitno prvo uspostaviti poverenje između klijenta i banke. S druge strane bankama je mnogo jeftinije da prodaju proizvod svom klijentu nego da troše novac na akviziciju novih klijenata. Osim postojeće baze klijenata, banke imaju i informacije o ponašanju klijenata tj. kako klijenti koriste proizvode, kao i podatke dobijene iz eksternih izvora (kreditni biro, APR i sl.). Kvalitet i kvantitet ovih podataka je bolji od podataka koje banka može dobiti kupovinom od specijalizovanih agencija.

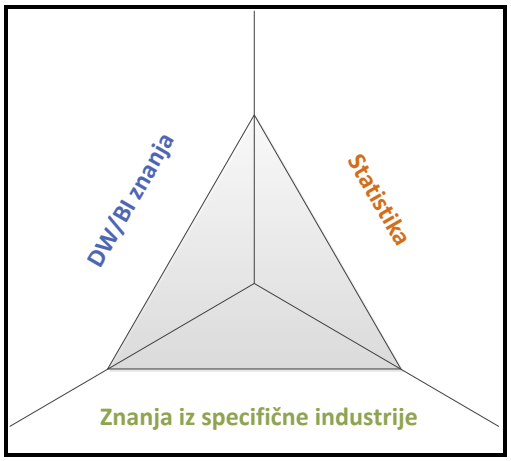
Zbog toga se banke u najvećoj meri oslanjaju na sopstvene klijente u cilju povećanja prodaje istih/sličnih proizvoda (eng. *up-sell*) ili različitih proizvoda (eng. *cross-sell*). Naravno, klijent koji je zadovoljan servisima banke ne mora nužno i da bude zainteresovan da kupi neki proizvod. Uspešnost prodaje istih ili različitih proizvoda svojim klijentima zahteva više od postojanja atraktivne ponude. Uspeh je jedino zagarantovan ako znamo kada, kome i šta treba da ponudimo.

Za podršku prodaji razvijaju se različiti matematički modeli koji ocenjuju sklonost klijenata da kupi neki proizvod (eng. *propensity to buy* model). Ovi matematički modeli određuju skorove za svakog klijenta koji predstavljaju verovatnoću da će klijent kupiti neki od proizvoda. Najbolje kotirani proizvodi za svakog klijenta predstavljaju najbolju ponudu za njega.

Ovaj rad opisuje izradu matematičkih modela zasnovanih na skoru/verovatnoći koristeći matematički metod logističke regresije. U radu je opisan proces izrade modela počev od definisanja i razumevanja poslovnog problema, specifikacije i pripreme podataka pre razvoja modela, razvoj modela, eksploatacije i nadgledanja modela.

Prilikom istraživanja podataka analitičar mora posedovati:

- znanja iz industrije gde se model primenjuje – u ovom slučaju bankarstvo (eng. *retail banking*)
- znanja iz DW/BI – priprema podataka, izrada *ad hoc* izveštaja i analiza,…
- znanja iz statistike – metode deskriptivne i prediktivne statistike



Slika 1. Balansirani odnos znanja je ključ uspeha

Ova znanja moraju da budu balansirana (Slika 1). Nedostatak znanja iz jedne od navedenih oblasti može značajno uticati na kvalitet modela i brzinu razvoja. Posedovanje znanja iz sve tri oblasti ne garantuje da će model biti uspešno razvijen. Osim navedenih znanja, potrebno je posedovati i veštinu da se poslovni problem uoči, izdvoji i opiše matematičkim modelom, a rezultate matematičkog modela neophodno je vratiti u poslovni kontekst razumljiv poslovnom korisniku.

Za potrebe rada biće napravljen model nad realnim podacima – gotovinski krediti. Takođe, u radu će biti opisana njegova integracija u sistem „sledeća najbolja ponuda“.

U praksi se prave i modeli koji ocenjuju sklonost ka kupovini (od strane klijenta banke) dozvoljenog prekoračenja, stambenog kredita, kredita za automobile, oročenog depozita, osiguranja, lizinga,… Svi ovi modeli mogu biti integrisani u sistem nazvan „sledeća najbolja ponuda za klijenta“.

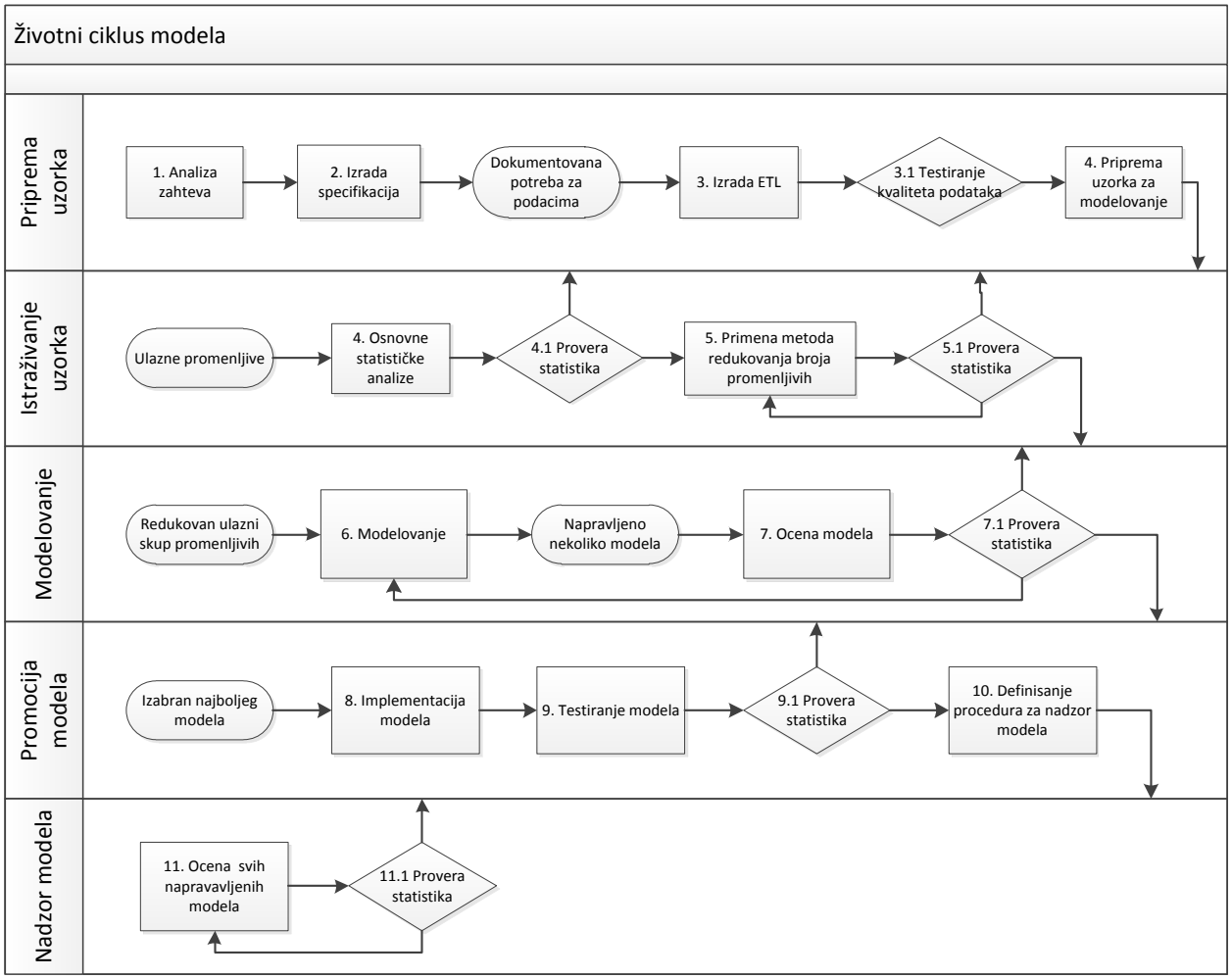
Rad se sastoji iz tri dela - glavnog dela i dva dodatka. U glavnom delu je opisana izrada modela na konkretnom primeru bez ulaženja u detalje koji se odnose na primenu SAS alata i matematičkih osnova na kojima se rad temelji. U slučaju da je neophodno dodatno objašnjenje postoje reference ka odgovorajućim pojmovima u dodacima i dalje ka odgovarajućoj literaturi.

Dodatak *A. „SAS Enterprise Miner“* predstavlja opis alata koji je korišćen u radu kroz SEMMA (***S**ample, **E**xplore, **M**odify, **M**odel, **A**ssesment*) pristup u razvoju modela. Detaljno su opisane samo one komponente koje su korišćene u radu. Ostale komponente su samo navedene.

Dodatak *B. „Matematičke osnove“* predstavlja ukratko opisane matematičke pojmove korišćene u samom radu, pri čemu je naglašena njihova poslovna primena.

1.1 Životni ciklus modela

Proces izrade modela je iterativan (Slika 2). Izrada modela počinje pripremom uzorka za modelovanje.



Slika 2. Životni ciklus modela

Analitičar koristi različite izvore i sve dostupne podatke kako bi pripremio uzorak. Uzorak se nalazi u tzv. ABT (***A**nalytic **B**ase **T**able*). Često se ovaj korak radi u IT po dostavljenoj specifikaciji analitičara. U slučaju da to radi analitičar, prilikom pripreme podataka moguće je odmah analizirati kvalitet podataka i upoznati se s poslovnim procesima banke kroz prizmu podataka[1].

[1] Ponekad se analitičari iznenade načinom na koji su prikupljeni podaci o nekom poslovnom procesu. Nemogućnost da se na osnovu ovih podataka izvedu kvalitetni atributi klijenta često inicira unapređenje poslovnih procesa i načina prikupljanja podataka, što je jedan od ciljeva svakog DW/BI rešenja.

Priprema uzorka je detaljno opisana u poglavljima *Priprema podataka za modelovanje* i *Specifičnosti modela koji „računaju“ sklonost ka kupovini*.

U slučaju da je uzorak pripremljen, istraživanje podataka predstavlja prvi kontakt analitičara sa podacima. Ovo je faza gde se rade osnovne statistike nad podacima i inicira eventualna izmena podataka. Zavisno od statistika moguće je vratiti se na prethodni korak ili krenuti u proces modelovanja. Istraživanje podataka je opisano u poglavljima *Preliminarni koraci u razvoju modela* i *Redukovanje i izbor promenljivih*.

Proces modelovanja predstavlja izbor algoritma i primenu algoritma nad različitim skupovima odabranih promenljivih. Nad izrađenim modelima primenjuju se razne tehnike za ocenu modela. U slučaju da nije moguće napraviti kvalitetan model moramo se vratiti jedan ili dva koraka unazad. Ovaj proces je opisan u poglavlju *Razvoj modela*.

Neposredno pre promocije modela radi se testiranje modela sa podacima koji nisu korišćeni u razvoju. Obično, ali ne i nužno, to su podaci koji imaju drugu vremensku dimenziju u odnosu na pripremljen uzorak. U ovoj fazi se definišu procedure za nadzor modela. Ova faza je opisana u poglavlju *Model u produkciji*.

Iako je proces izrade modela izrazito iterativan u produkciji se retko dešava da se model može „popraviti“. U slučaju da se prilikom nadzora modela utvrdi da model nije dobar („ne pogađa“), tada se pristupa ponovnoj izradi modela i prolazi se kroz sve faze u izradi modela. Svaka izmena modela predstavlja novi model, dok se stari model penzioniše.

2 Definisanje poslovnog problema

Banke se u najvećoj meri oslanjaju na sopstvene klijenata u cilju povećanja prodaje istih/sličnih proizvoda ili različitih proizvoda. Ovo je najjeftiniji i najsigurniji način povećanja prodaje i uspostavljanja čvrste veze između klijenta i banke.

Klijent koji je zadovoljan servisima banke ne mora nužno i da bude zainteresovan da kupi neki proizvod. Uspešnost prodaje istih ili različitih proizvoda svojim klijentima zahteva više od postajanja atraktivne ponude. Uspeh je jedino zagarantovan ako znamo kada, kome i šta treba da ponudimo.

2.1 Razumevanje poslovnog problema

Poslovni problem: Razviti matematički model koji ocenjuje sklonost klijenta da će kupiti gotovinski kredit. Analogno mogu se razviti i modeli za kreditnu karticu, potrošački kredit, kredit za automobile, stambeni kredit, oročeni depozit, tekući račun i dozvoljeno prekoračenje. Za svaku grupu proizvoda neophodno je izračunati skor. Najbolje kotirani proizvod se prvi nudi klijentu.

U sledećenih nekoliko pasusa[2] biće opisani osnovi bankarski pojmovi.

Tekući račun predstavlja osnovni bankarski proizvod. Na ovom računu klijent prima zaradu i ostale prihode. Raspolaže sa novcem u iznosu uplata. Isplata novca je gotovinska. Plaćanje roba i usluga sa računa klijent može uraditi čekom, debitnom karticom ili nalogom za plaćanje u ekspozituri ili elektronskim kanalom.

Avista račun predstavlja račun opšte namene gde klijent može štedeti po viđenju. Plaćanje roba i usluga može se izvršiti na isti način kao i kod tekućeg računa.

Oročeni depozit predstavlja mogućnost da klijent kratkoročno (do 12 meseci) ili dugoročno (preko 12 meseci) određenu svotu novca da na raspolaganje banci. Za uzvrat, po isteku roka, banka je obavezna da klijentu vrati depozit i isplati odgovarajuću kamatu.

Događaj prodaje računa predstavlja otvaranje računa.

Dozvoljeno prekoračenje predstavlja mogućnost da klijent bez posebne procedure koristi dodatna novčana sredstva. Prekoračenje obično ima granicu u visini plate. Kao proizvod ne postoji samostalno, već je jedan od servisa tekućeg računa.

Događaj prodaje dozvoljenog prekoračenje predstavalja dan prvog odobravanja granice. Ponekad se pod događajem prodaje može tretirati i povećanje granice.

[2] Uzeto iz „Leksikon bankarstva – Dobrivoje Milojević, , ISBN 86-904813-0-3, MeGraf 2003"

Kreditna kartica služi za bezgotovinsko plaćanje robe i usluga. Ona svom vlasniku omogućuje plaćanje raznih usluga, kupovinu proizvoda i podizanje gotovog novca uz obavezu da će potrošeni novac vratiti banci na ugovoren način. Kartica osim funkcije plaćanja ima i funkciju kreditiranja. Zavisno od kreditne sposobnosti klijentu se odobrava odgovarajuća granica i to najčešće u visini njegove plate.

Događaj prodaje kreditne kartice predstavlja datum prvog aktiviranja kartice. To je trenutak od kada klijent može da koristi odobrena sredstva.

Gotovinski kredit predstavlja novčana sredstva koja se klijentu odobravaju na tekućem ili nekom drugom računu. Ovi krediti nemaju posebnu namenu i klijent može koristiti isplaćen novac po svom nahođenju. Klijent novac vraća u dogovoru sa bankom.

Događaj prodaje gotovinskog kredita predstavlja transfer novca sa računa banke na račun klijenta.

Potrošački kredit predstavlja novčana sredstva koja banka u ime klijenta isplaćuje trećem licu za kupljenu robu ili usluge od strane klijenta. Ovim novcem klijent banke (dužnik) ne raspolaže slobodno već ih namenski koristi.

Događaj prodaje potrošačkog kredita predstavlja transfer novca sa računa banke na račun trećeg lica.

Stambeni krediti predstavljaju novčana sredstva koja banka u ime klijenta isplaćuje trećem licu za kupovinu stambene jedinice ili poslovnog prostora. Specifičnost ovih kredita je u sredstvima obezbeđenja koja obično uključuje i hipoteku na kupljenu nekretninu. Procedura odobravanja ovog kredita je specifična i može trajati nekoliko meseci.

Događaj prodaje stambenog kredita predstavlja prvi transfer novca sa računa banke na račun trećeg lica.

2.2 Scenario korišćenja modela

Postoje dva scenarija upotrebe modela. To su:

- Izračunavanje skora za pojedinačnog klijenta a na zahtev savetnika za prodaju u ekspozituri (u daljem tekstu prodavac).
- Izračunavanje skora za sve klijente banke i organizovanje kampanje.

Na zahtev prodavca moguće je za određenog klijenta izračunati rejting za sve važnije proizvode banke. Ovo se obično dešava u trenucima kada je klijent nekim drugim poslom došao u banku. Dok službenik radi sa klijentom, na prodajnom ekranu službenika pojavljuje se ekran sa listom proizvoda koje može ponuditi klijentu. Klijentu se uvek prvo nudi proizvod sa najvećim skorom/verovatnoćom. Ovakav vid organizovanja prodaje zove se *inbound CRM*.

Ponekad banka sa ciljem povećanja prodaje sama inicira kontakt sa klijentom i nudi odgovarajući proizvod. U ovom slučaju se računa skor za sve klijente. U ciljnu populaciju

ulaze samo oni klijenti sa najvećom verovatnoćom da će kupiti proizvod. Veličina populacije zavisi i od troškova kampanje. Ovakav vid kampanje obično koristi različite kanale kao što su: poštansko pismo, *e-mail*, SMS, MMS, kontakt centar. Ovakva organizovana prodaja kod koje banka inicira kontakt zove se *outbound CRM*.

3 Priprema podataka za modelovanje

U ovom poglavlju biće opisane strukture podataka koje se koriste u modelovanju ABT, specifikacija promenljivih za izradu modela i transformacija podataka do finalnih ABT-ova.

3.1 Strukture podataka pogodne za modelovanje

Za potrebe istraživanja podataka neophodno je napraviti odgovarajuće strukturu ABT. U RSUBP kontekstu, ABT je relacija (tabela). Pojedinačnu relaciju (red) nazivamo opservacija, a atribute relacije (kolone) zovemo promenljive. Za modele navedene u ovom radu podaci se pripremaju na nivou klijenta tj. u jednom vremenskom trenutku posmatra se stanje i istorija od 24 meseca unazad.

Promenljive u ABT mogu biti:

- Kategoričke (GENDER_CD, STD_OCUPATION_CD,…)
- Kontinualne
    - mera stanja (BALANCE_AMT, ACTIVE_ACCOUNT_CNT,…)
    - mere prometa (REPAY_AMT_M6, CA_LMTU50_CNT_M6 …)
    - mere proseka (BALANCE_AV_AMT_M3)
    - razne izvedene promenljive

Prema kardinalnosti domena promenljive možemo podeliti na:

- Binarne – promenljiva može imati dva stanja
- Intervalne – promenljiva može imati „beskonačno mnogo“ stanja pri čemu je rastojanje između susednih članova jednako.
- Nominalne – promenljiva ima „konačno mnogo“ stanja pri čemu ne postoji uređenost između članova niti je poznato rastojanje između susednih članova.
- Ordinarne – promenljiva ima „konačno mnogo“ stanja pri čemu znamo uređenost članova kao i njihovo rastojanje. Npr. nivo obrazovanja se može tretirati kao nominalna i kao ordinarna promenljiva pri čemu „rastojanje između“ nivou ne mora biti jednako (1-osnovna škola 2-srednja škola, 4-viša škola, 5-visoka škola, 7-master, 10-doktorat). Specifičnost ordinarnih promenljivih je što se za njih mogu raditi statistike i za intervalne i za nominalne promenljive.
- Unarna promenljiva može imati samo jednu vrednost/stanje i nije od interesa u procesu modelovanja.

3.2 Specifikacija promenljivih za izradu modela

Prvi korak u specifikaciji promenljivih je analiza izvora podataka za ABT. Po završenoj analizi identifikovane su grupe promenljivih i tada se za svaku grupu promenljivih definišu konkretne promenljive koje će biti korišćene u modelovanju.

U ovom radu, promenjive klijenata su podeljene u grupe. To su:

- sociodemografske i opšte promenljive klijenata (Tabela 1)
- promenljive ponašanja klijenata (Tabela 2) u smislu korišćenja bankarskih proizvoda
- zavisne (ciljne) promenljive i alternativne promenljive (Tabela 3)

Za svaku od ovih grupa napravljeno je od nekoliko do nekoliko desetina promenljivih. Ukupan broj promenljivih korišćen u ovom radu je 2043.

Customer Info	osnovi sociodemografski podaci o klijentu
Socio	pol, obrazovanje zaposlenje,…
Adress	opština i ogrug stanovanja
CRM	broj kontakata, da li zeli da bude kontaktiran,…
Risk & Default	kašnjenja i rejting
Default	kašnjenje klijenta
Credit Scoring	intretni kreditni rejting
CB	podaci iz kreditnog biroa
Q Report	poslednje dostupni kvartalni izveštaj
Contract Activity	aktivnosti vezano za dolazak klijenta u ekspoziturui i segmentacija
Bussines Location	najčešće korišćena ekspozitura, opština i okrug gde se nalazi ekspozitura
Business Segm.	interne segmentacije klijenata zasnovane na računu i aktivnosti, servisni model klijenta
Account Summary	aktivnosti vezane za otvaranje i zatvaranje računa

Tabela 1. Sociodemografske i opšte promenljive klijenta

CA	promenljive vezane za tekekuće račune
Base	osnovne premenljive, kao što su događaj prodaje, prosečna stanja, mesečni prometi,…
Profitability	profitabilnost
Salary Utilization	način korišćenja plate
Income Utilization	način korišćenja plate+ostalih prihoda
Overdraft Utilization	način korišćenja dozvoljenog prekoračenje
Standing Order	korišćenje trajnih naloga
Ebank	mogućnost korišćenja ebank
Debit Card	osnovni podaci o debitnim karticama
Debit Card Payment	način korišćenja debitne kartice (super market, putovanja, restorani,…)
Check Utilization	korišćenje čekova, broj realizovanih i izdatih
Credit Card	promenljive vezane za kreditne karticame
Base	osnovne promenljive kao što su broj izdatih kartica, broj korsnika, prosečna stanja,..
Profitability	profitabilnost
Limit Utilization	korišćenje limita
Payment	način korišćenja kreditne kartice (super market, putovanja, restorani,…)
Avista deposit	promenljive vezane za avista depozite
Base	osnovne promenljive, događaji otvaranja i zatvaranja
Balance	promenljive stanja i proseka
Profitability	profitabilnost
FX Payment	plaćanje i priliv iz inostranstva preko deviznih računa
Short Term Deposit	promenljive vezane za kratkoročne depozite
Base	osnovne promenljive, događaji otvaranja i zatvaranja
Balance	promenljive stanja i proseka
Profitability	profitabilnost
Long Term Deposit	promenljive vezane za dugoročne depozite
Base	osnovne promenljive, događaji otvaranja i zatvaranja
Balance	promenljive stanja i proseka
Profitability	profitabilnost
Consumer Loan	promenljive vezane za potrošačke kredite
Base	osnovne promenljive, događaji otvaranja i zatvaranja
Balance	promenljive stanja i proseka
Profitability	profitabilnost
Car Loan	promenljive vezane za kredite za automobile
Base	osnovne promenljive, događaji otvaranja i zatvaranja
Balance	promenljive stanja i proseka
Profitability	profitabilnost
Cash Loan	promenljive vezane za gotovinske kredite
Base	osnovne promenljive, događaji otvaranja i zatvaranja
Balance	promenljive stanja i proseka
Profitability	profitabilnost
Mortgage	promenljive vezane za stambene kredite
Base	osnovne promenljive, događaji otvaranja i zatvaranja
Balance	promenljive stanja i proseka
Profitability	profitabilnost
Leasing	promenljive vezane za lizing
Insurence	promenljive vezane za osiguranje

Tabela 2. Ponašanje klijenta

Target	ciljne promenljive i alternativne ciljne promenljive
Target Cash Loan	1 ako je klijent kupio proizvod 0 inače
Target Overdraft	1 ako je klijent kupio proizvod 0 inače
Target CC	1 ako je klijent kupio proizvod 0 inače
Target STD	1 ako je klijent kupio proizvod 0 inače
Target LTD	1 ako je klijent kupio proizvod 0 inače
Target Car Loan	1 ako je klijent kupio proizvod 0 inače
Target Mortgage	1 ako je klijent kupio proizvod 0 inače

Tabela 3. Ciljne promenljive

Detaljna specifikacija promenljivih se nalazi u dokumentu *MasterABT.xls* koji je prilog ovom radu.

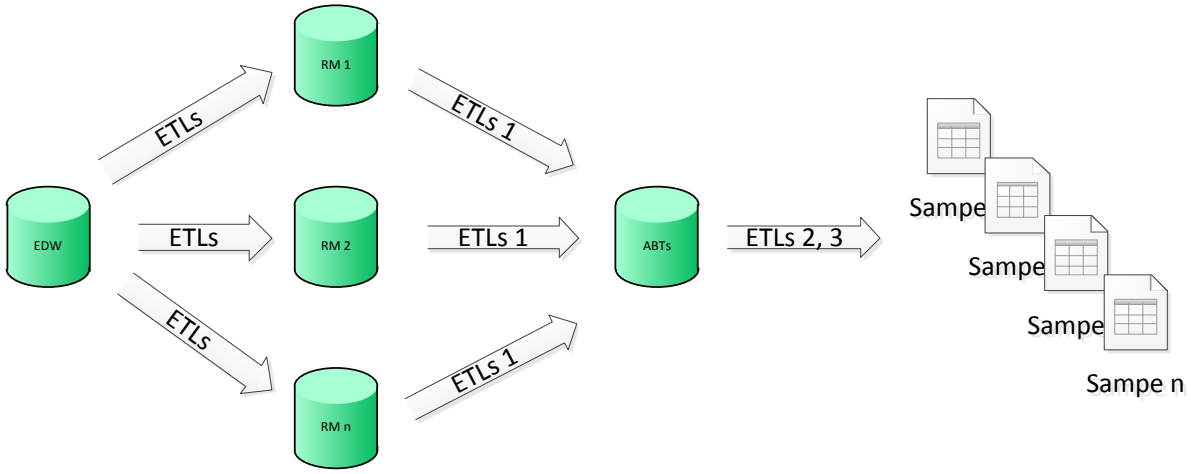
Imenovanje promenljivih je veoma važno u izradi specifikacije, a kasnije i u samom procesu modelovanja. Imenovanje primenjeno u ovom radu je opisano dodatku A poglavlje *Formiranje uzorka - Sample.* Važnost imenovanja se ogleda u proveri da li su promenljivima dodeljene ispravne uloge u procesu modelovanja[3].

Napomena: Imena promenljivih moraju da budu jasna i nedvosmislena tako da odmah asociraju na grupu promenljivih kao i na konkretno značenje. Prilikom imenovanja treba voditi računa na ograničenja koji sam alat ima. *SAS Enteprise Miner* (EM) podržava imena dužine najviše 32 znaka. Zbog toga, bilo bi dobro da u ABT-u nazivi promenljivih ne budu duži od 26 znakova. Na ovaj način ostavljamo mogućnost da u samom SAS EM izvedemo nove promenljive bez narušavanja imenovanja.

[3] Videti poglavlje *„Preliminarni koraci u razvoju modela/Povezivanje uzorka za modelovanje sa projektom“*

3.3 Transformacije podataka do finalnih ABT

Tok podataka u transformaciji od ulaznih do finalnih je prikazan na slici (Slika 3).



Slika 3. Tok podataka od EDW do ABT-ova

Na levoj strani slike nalazi se EDW[4] (eng. *Enterprise Data Warehouse*) baza. U slučaju da ona ne postoji onda je izvor transakciona baza. Strelice obeležene sa *„ETLs“* predstavljaju ETL[5] transformacije do izveštajnih baza (eng. Reporting/Data Mart[6]) koje su objavljene poslovnim korisnicima za ad hoc izveštavanje i analize.

Transformacije podataka za finalni ABT je urađeno na MS SQL 2012 RSUBP iz RM (Slika 3)

ETL za pripremu ABT je organizovan na sledeći način:

[4] EDW (eng. *Enterprise Data Warehouse*) centralizovano skladište podataka u kojem se nalaze konsolidovani, očišćeni, provereni i dobro struktuirani podaci jedne kompanije. U prilogu su dve *data warehouse* paradigme date od strane dva autoriteta iz DW/BI:

“Data warehouse is one part of the overall business intelligence system. An enterprise has one data warehouse, and data marts source their information from the data warehouse. In the data warehouse, information is stored in 3rd normal form” Bill Imnon

“Data warehouse is the conglomerate of all data marts within the enterprise. Information is always stored in the dimensional model.” Ralph Kimball

U slučaju da nemamo EDW bazu već samo RM kažemo da je primenjen Ralph Kimball pristup. Ako EDW baza postoji tada kažemo da je primenjen Bill Imnon pristup u dizajnu DW/BI sistema.

[5] ETL (eng. *Extract Transform Load*) su procesi (programi) koji prikupljaju, transformišu i učitavaju podatke u posebno dizajnirane strukture podataka. Pod transformacijom se podrazumevaju i procesi čišćenja i provere podataka (eng. *data quality, data validation, data cleansing*)

[6] Data Mart (DM) predstavlja strukture organizovane tako da obezbede zahtevane analize iz neke specifične poslovne oblasti. Ove strukture su poznate poslovnom korisniku i može ih koristiti za ad hoc analize i izveštavanje koristeći razne izveštajne alate bez podrške ICT.

- *ETLs1* – računa osnovne promenljive. Srodne promenljive se računaju u jednom prolazu. Rezultati se čuvaju u odgovarajućim tabelama (eng. *Base ABTs*). Ovo predstavlja osnovnu trasformaciju i uglavnom je u nadležnosti ICT.
- *ETLs2, ETLs3* – predstavlja dva fleksibilna sloja (obično implementirana kao SQL pogled ili stored procedura) nad baznim ABT tabelama u RSUBP:
    - Prvi sloj nam omogućava da izvedemo nove promenljive iz postojećih osnovnih promenljivih[7]. Takođe, u ovom sloju moguće je umanjiti rasipanje (standardnu devijaciju) mera stanja (npr. logaritmovanjem).
    - Drugi sloj predstavlja finalnu transformaciju do konkretnog uzorka interesantnog za dalje istraživanje podataka (npr. uzimamo samo aktivne klijente sa tekućim računom u poslednjih 6 meseci ili samo one koji imaju depozit veći od 10 EUR ili imaju uručenu kreditnu karticu).

  ETL2 i ETL3 su u nadležnosti osobe koja razvija model.

Tehnička napomena: Mnogi RSUBP imaju ograničenja koja se ogledaju u broju kolona u tabeli i broju kolona u rezultajućem skupu *select* komande (kod MS SQL je 1024 kolona, odnosno 4096 kolona u *select* komandi). Zbog toga se promenljive grupišu (CA, CC, DEPOZIT, Customer_Info) i za svaku grupu se napravi po jedna tabela, a zatim se napravi stored procedura koja vraća rezultujući skup (jedna select komada koja spaja više manjih ABTs).

Poslovna napomena: Filter uzet u ETL3 obično nije slučajan. On je rezultat statističkih analiza i istraživanja. Često se taj filter dobije taktičkom segmentacijom klijenata napravljenom isključivo za razvoj nekog konkretnog modela.

[7] Izvođenje promenljivih iz postojećih je česta tehnika koja se primenjuje u istraživanju podataka. Npr. odnos mesečnog prosečnog stanja sa tromesečnim prosečnim stanjem daje vrednosti oko 1. U slučaju da je odnos manji od 1 imamo trend opadanja stanja dok ako je veći od jedan imamo trend porasta stanja na računu. Stanje 0 može da se označi specijalnim znakom U slučaju da imamo 0/0 možemo tretirati specijalnim znakom ili jednostavno postaviti vrednost promenljive na 1.

4 Specifičnosti modela koji „računaju“ sklonost ka kupovini

4.1 Definisanje ciljne promenljive za modele

Ciljna ili zavisna promenljiva (eng. *target variable* ili *response variable*) predstavlja događaj koji želimo da opišemo nezavisnim promenljivama (eng. *independent variables, explanatory variables*). Konkretno u ovom slučaju, želimo da izračunamo verovatnoću da će se događaj prodaje desiti na osnovu podataka o klijentu.

Ovo je veoma osetljiv momenat u procesu razvoja modela i predstavlja početni korak u transformaciji poslovnog modela u matematički. Loše definisana ciljna promenljiva imaće uticaj na model, koji, iako perfektan sa statističkog gledišta ne opisuje dobro poslovni problem.

Ciljna promenljiva korišćena u ovom radu je binarna i ima vrednost 1 ukoliko se desio događaj prodaje (eng. *event, good*) u odgovarajućem vremenskom okviru i 0 ako se nije desio događaj prodaje (eng. *nonevent, bad*). U daljem tekstu skup svih opservacija kod kojih je ciljna promenljiva jednaka 1 zvaćemo *event* populacija, a pojedinačnu opservaciju *event*. Skup svih opservacija kod kojih je ciljna promenljiva jednaka 0 zvaćemo *nonevent* populacijom, dok pojedinačnu ospervaciju *nonevent*.

Moguće je napraviti i ternarnu ciljnu promenljivu tako da:

- 1 predstavlja događaj apliciranja,
- 2 predstavlja događaj prodaje,
- 0 inače

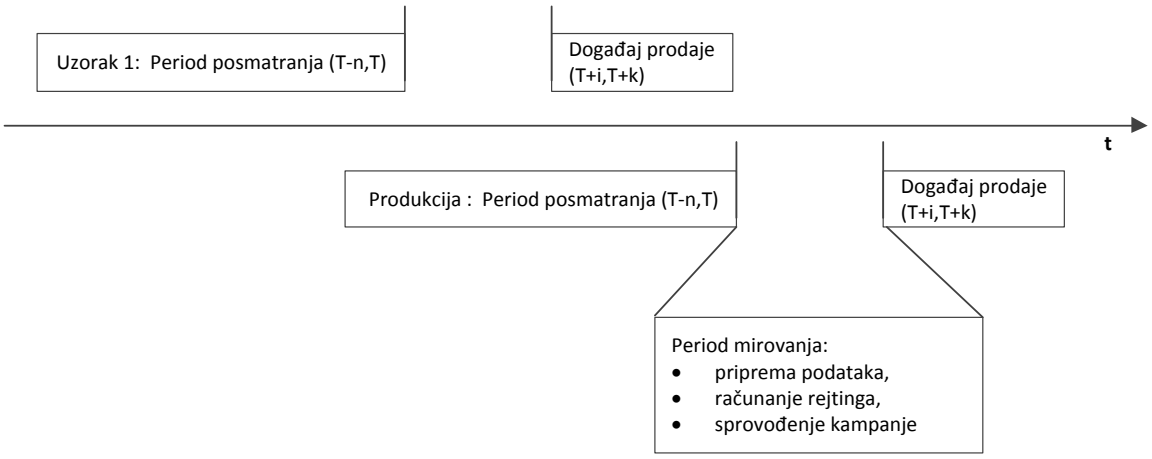
Ovakva promenljiva može dati precizniju sliku kod stambenih kredita kod kojih proces realizacije može trajati i nekoliko meseci. S druge strane komplikuje se izrada modela, a posebno interpretacija rezultata.

4.2 Metodologija pripreme uzorka u odnosu na ciljnu promenljivu

Događaj prodaje, osim što opisuje šta je klijent kupio, ima i vremensku dimenziju. Vremenski okvir u kojem posmatramo događaj je veoma bitan i mora se precizno definisati. On zavisi od scenarija upotrebe modela navedenog u poglavlju *2.2 Scenario korišćenja modela*.

Na slici (Slika 4) prikazan je proces pripreme uzorka za modelovanje. Uzorak na vremenskoj osi je podeljen u dve celine:

- nezavisne promenljive,
- događaj prodaje - zavisna promenljiva



Slika 4. Priprema ciljne promenljive u odnosu na uzorak

Nezavisne promenljive se posmatraju u vremenskom periodu *(T-n,T)*, pri čemu je n obično između 6 i 24 meseca. Događaj prodaje se posmatra u periodu *(T+i,T+k)*. Zavisno od scenarija upotrebe, *i* i *k* mogu da uzimaju sledeće vrednosti:

- *i= 0 meseci* i *k=1 mesec* ako se model primenjuje u *inbound* marketingu tj. ako se računa skor odmah po dolasku klijenta u ekspozituru
- *i=20/30/40 dana*, *k=2 meseca* ako se radi o *outbound* marketing tj. ostavlja se prostor potreban za računanje skora, pripremu i sprovođenje kampanja

Vremenski okvir za događaje prodaje je obično 1 mesec uz uslov da kijent nije aplicirao pre vremenskog trenutka T, odnosno T+i. Ovo je veoma važno ograničenje koje se mora postaviti radi finog podešavanja modela. U ovom slučaju podaci se obrađuju mesečno.

Napomena: Za kreditne kartice se često uzima datum odobrenja kartice bez obzira da li je klijent karticu aktivirao. Za stambene kredite čiji proces odobrenja može da traje mesecima vremenski okvir se može povećati na 6 meseci. U tom slučaju vrednovanje se radi nad podacima starim 7 meseci (1 mesec mirovanja i 6 meseci za događaj prodaje). Ovo značajno neće uticati na predviđanje jer se stambeni krediti ne kupuju često (obično jednom za života klijenta).

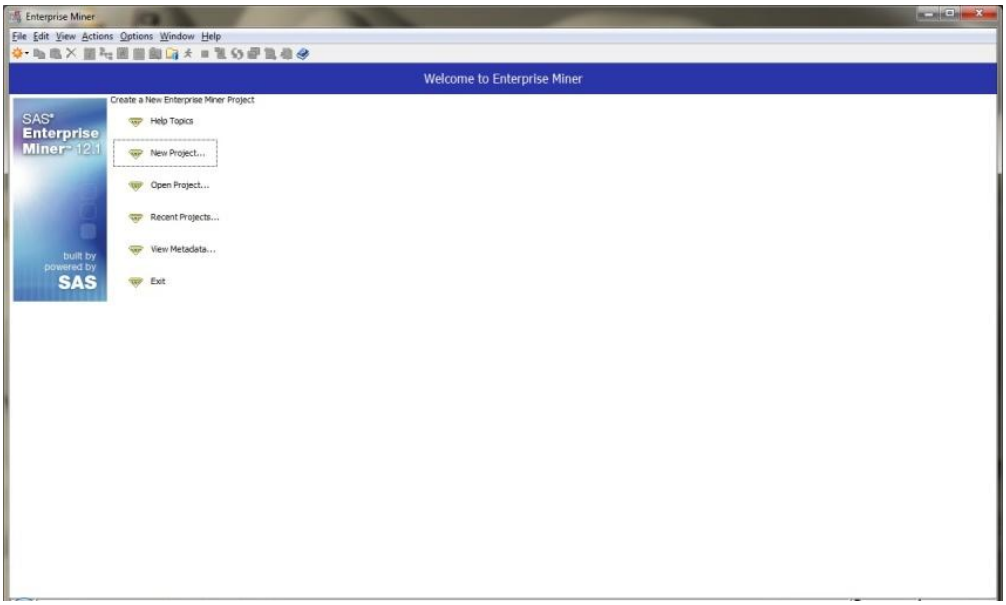
5 Podešavanje alata za razvoj modela

U ovom poglavlju su opisana podešavanje alata *SAS Enterprise Miner*[8] (SAS EM) koje se rade pre početka istraživanja podataka.

5.1 Izrada novog projekta

Novi SAS EM projekat u *workstation*[9] okruženju možemo napraviti na sledeći način:

1. Pokrenuti SAS EM

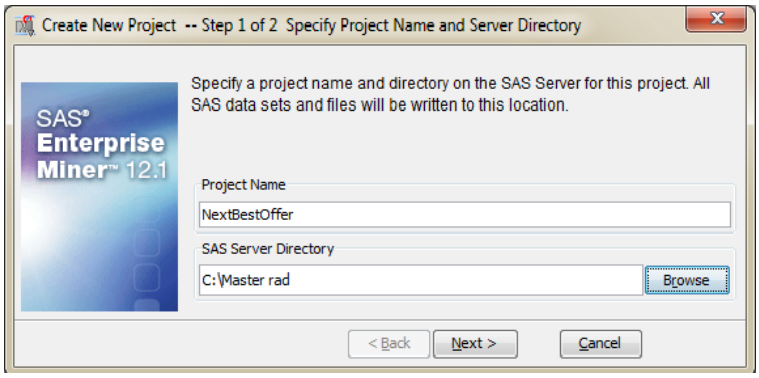


Slika 5. Pokrenut SAS EM

2. Odabrati link „New Project“

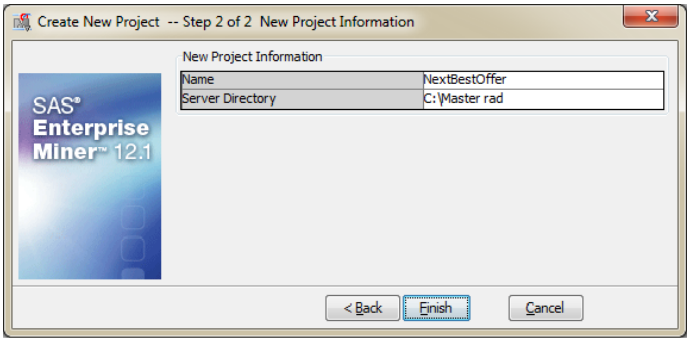
[8] http://www.sas.com/technologies/analytics/datamining/miner/

[9] SAS *workstation* instalacija predstavlja instaliranje serverski i klijentski komponenti softvera na jednu radnu stanicu.



Slika 6. Prvi korak – unos metapodataka

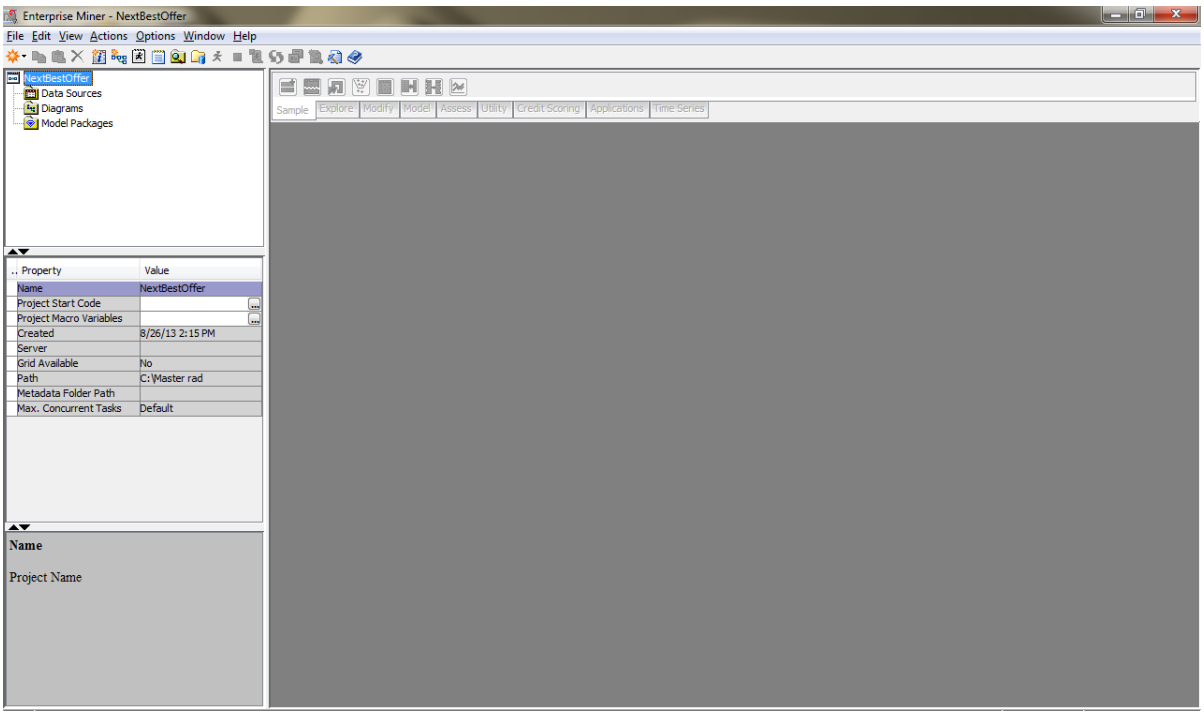
3. Uneti naziv projekta i lokaciju



Slika 7. Drugi korak - provera metapodataka

4. Izabrati opciju *Finish*

Otvoreni projekat je prikazan na narednoj slici



Slika 8. Otvoren SAS EM projekat - NextNestOffer

5.2 Struktura SAS EM

SAS EM je moguće koristiti na dva načina:

1. kao klijent – server,
2. u lokalnom okruženju (eng. *workstation/local)*

Kod klijent-server arhitekture na radnoj stanici je instaliran *SAS EM* klijent, dok su na serveru instalirani *SAS* metadata server[10] i *SAS/STAT*[11] komponenta neophodna za rad aplikacije. Klijentska aplikacija formira instancu na serveru gde se izvršavaju analize. Moguće je umesto klijentske aplikacije koristiti *Java aplet* i svaki put učitati aplikaciju sa odgovarajuće URL adrese na serveru (npr. http://mymachine:6098/EnterpriseMiner/). U oba slučaja komunikacija ne ide direktno, već se za to koristi SAS metadata server na kome se nalaze informacije o samom projektu (ime projekta, lokacija gde se nalazi, prava pristupa, okruženje koje se koristi za izvršavanje…). Za ovakav rad neophodno posedovati *SAS Enterprise BI*[12] okruženje.

U slučaju lokalne instalacije klijent i server se nalaze na istoj radnoj stanici. Za pristup projektu ne koristi se metadata server. Ovaj način rada je karakterističan za kompanije koje nemaju potrebu za *Enterprise BI* okruženjem[13].

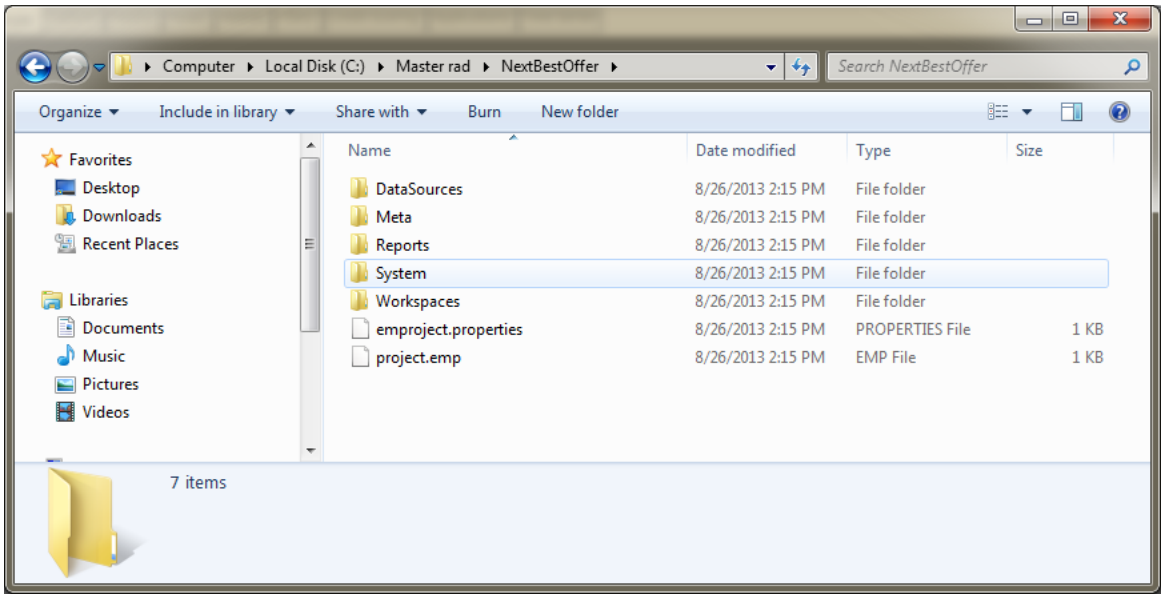
U oba slučaja SAS EM projekat fizički se nalazi u jednom direktorijumu

[10] http://www.sas.com/technologies/bi/appdev/base/metadatasrv.html

[11] http://support.sas.com/documentation/onlinedoc/stat/

[12] http://www.sas.com/technologies/bi/

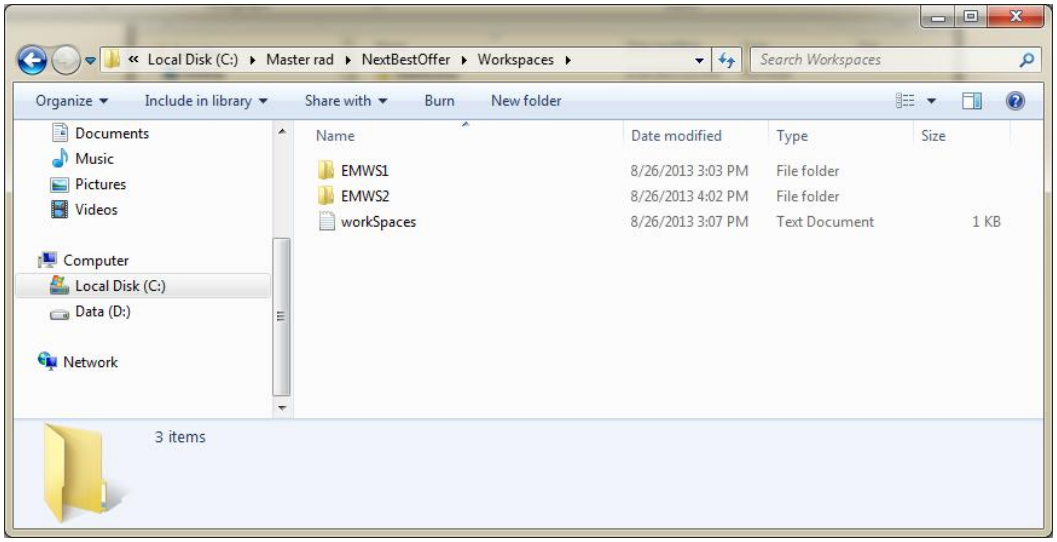
[13] Kompanije koje rade istraživanja za treća lica obično nemaju SAS Enterprisje BI okruženje. To su agencije kje se bave istraživanjem u farmaciji, marketinške agencije, statistički instituti,…



Slika 9. Struktura SAS EM projekta

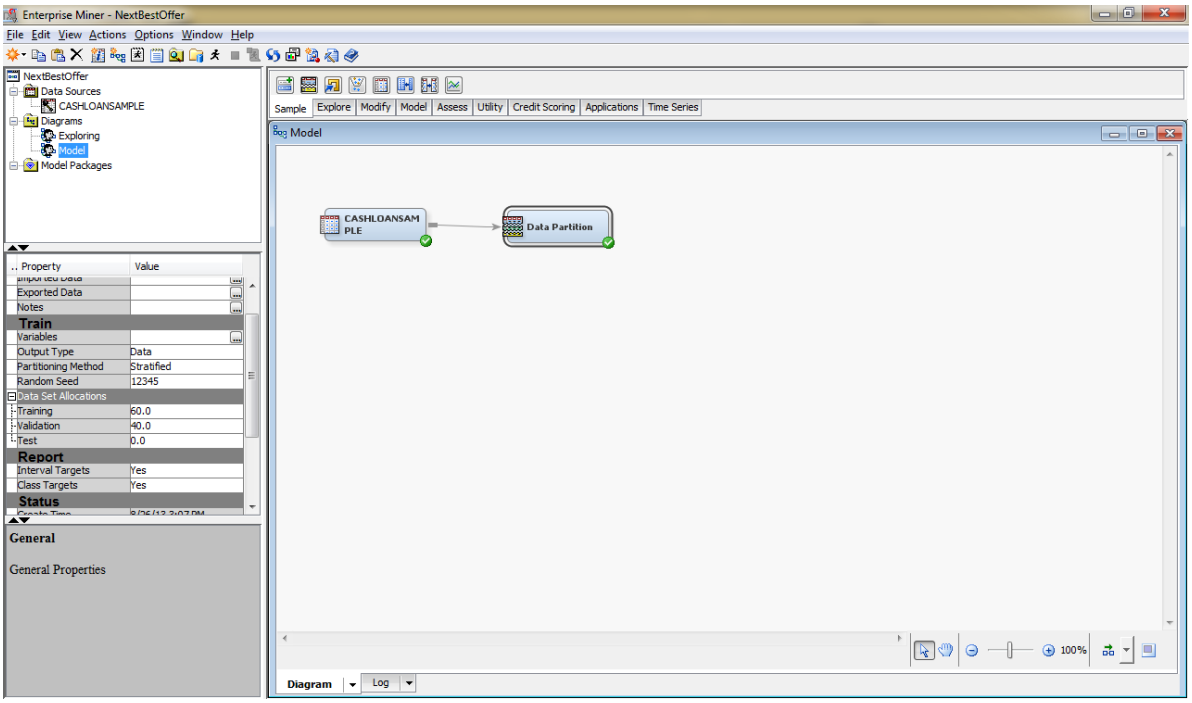
SAS EM projekat (Slika 9) je predstavljen jednom xml datotekom (*project.emp*). U direktorijumu *DataSources* nalaze se tabele sa podacima koji se koriste u procesu modelovanja.

U direktorijumu *Workspaces* nalazi se lista radnih površina i to za svaku radnu površinu po jedan poddirektorijum (Slika 10). Jedna radna površina predstavlja jedan tok podataka i koristi se za izradu jednog modela. Unutar svake radne površine za svaku komponentu u SAS EM kreira se poseban direktorijum u kome se nalaze metapodaci kao i rezultati istraživanja. Izlazni rezultati jedne komponente (eng. *Node*) predstavljaju ulazne podatke druge komponente.



Slika 10. Organizacija radnih površina u SAS EM

Na slici (Slika 11) nalazi se otvoren projekat *NextBestOffer*

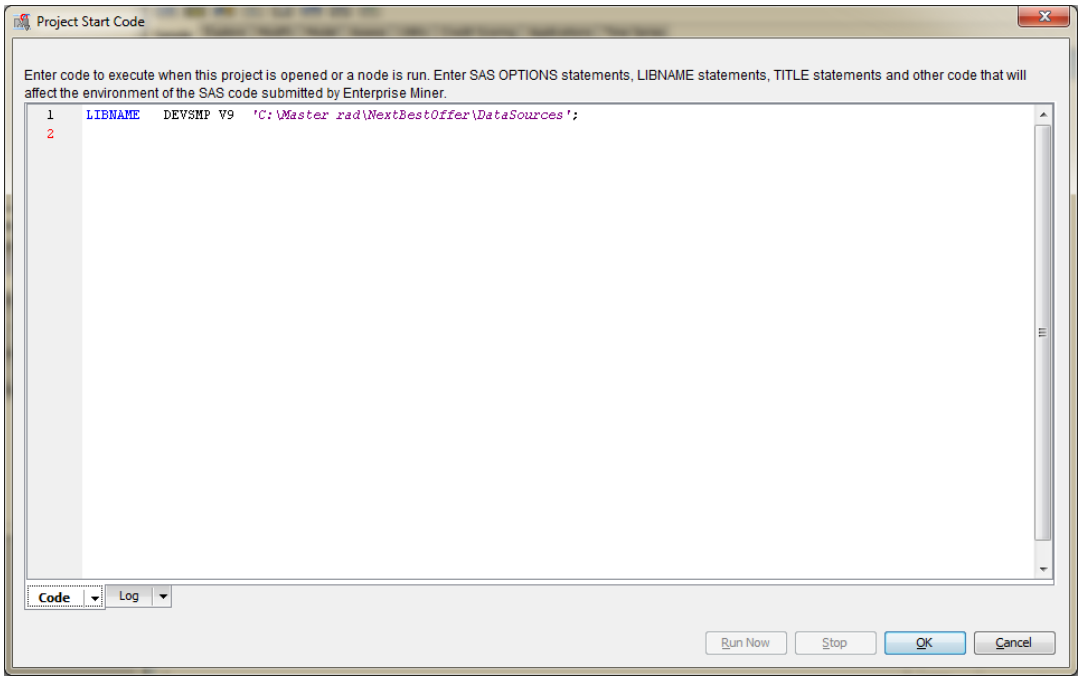


Slika 11. Početak rada u SAS EM

U gornjem levom uglu nalazi se *„Project Explorer”*. Ovde se mogu videti izvori podataka i dijagrami. Dijagram predstavlja radnu površinu gde se razvija model. Prilikom razvoja modele koriste se razne komponente koje su povezane strelicama. Strelice nam govore da rezultati jedne komponente predstavljaju ulazne podatke za drugu komponentu.

5.3 Povezivanje uzorka za modelovanje sa projektom

Po otvorenom projektu neophodno je povezati uzorak sa projektom. To se radi definisanjem tzv. *„startup“* koda. Ovaj SAS kod će se izvršavati uvek po pokretanju SAS EM projekta. Na slici (Slika 12) dat je primer takvog koda.

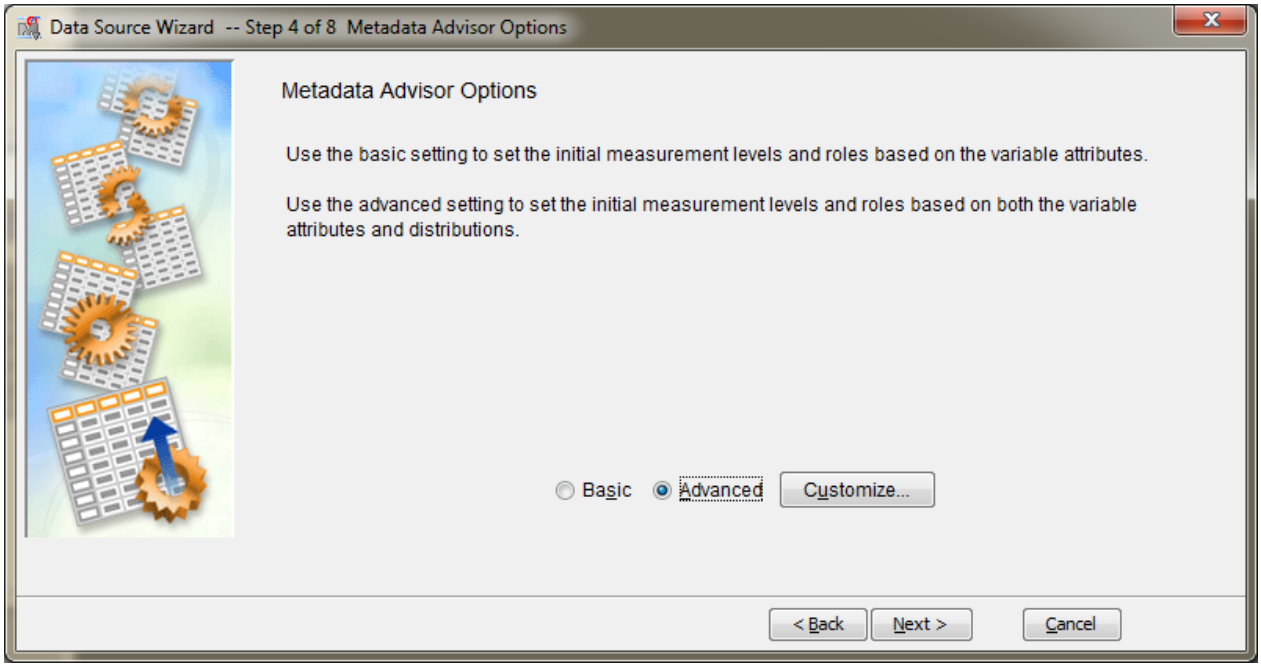


Slika 12. Podešavanje okruženja korišćenjem SAS startup koda

Na ovaj način registrovan je *DataSources* direktorijum kao mesto gde će se nalaziti ulazni podaci.

Neophodno je još uraditi i sledeće :

1. Premestiti uzorak u direktorijum *DataSources*
2. Pozicionirati se na *DataSources* i desnim klikom izabrati *Create Data Source*
3. Pratiti instrukcije do koraka 4 gde SAS EM nudi dva pristupa u formiranju metapodataka - osnovni i napredni.



Slika 13. Izbor pristupa pri formiranju metapodataka za skup ulaznih promenjivih

Kod osnovnog pristupa SAS će napraviti metapodatke za ABT bez ulaženja u same podatke već samo koristeći tipove podataka svake ulazne promenljive. Kod naprednog pristupa alat proverava i same podatke, tj. za svaku promenljivu proverava njenu kardinalnost. Tako, ako je kardinalnost jednaka 1, tj. u uzorku alat utvrdi da promenljiva ima samo jednu vrednost po automatizmu je odbacuje (postavlja rolu *rejected*). U slučaju da je kardinalost jednaka 2, alat postavlja promenljivu kao binarnu (postavlja rolu *binary*). U slučaju da promenljiva ima u nazivu „Target" postavlja je na Target. Ovo je veoma korisna opcija.

Prilikom izrade modela koristi se više hiljada promenljivih. U slučaju izbora opcije *basic* lako se može potkrasti greška, npr. da je jedna od promenljivih unarna (možda na čitavoj populaciji klijenta ona nije unarna ali je nad definisanim uzorkom jeste). Ovo može praviti probleme pri izračunavanju statistika, jer neke komponente koje rade izbor promenljivih ne očekuju unarne promenljive na izvoru. Pronalaženje unarne promenljive u skupu od preko 2000 promenljivih može biti naporno.

4. Pratiti dalje instukcije do završetka formiranja komponente.

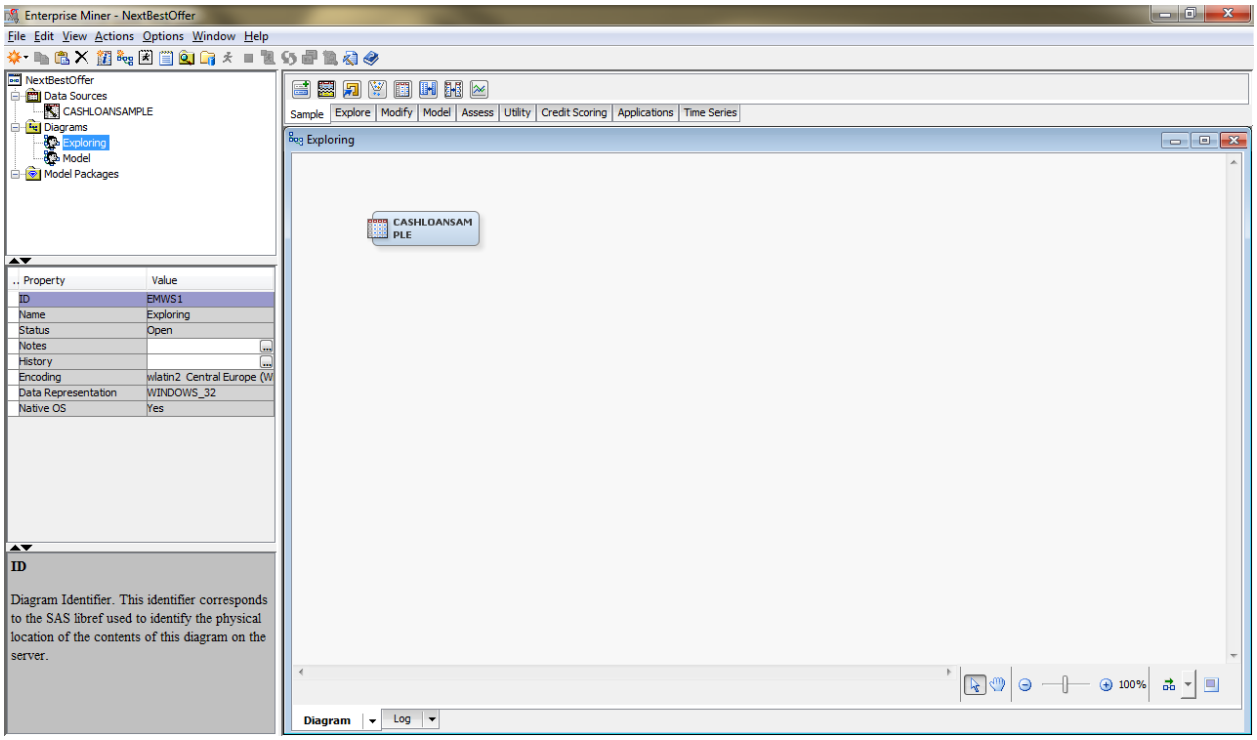
Posle formiranja ulazne koponente neophodno je proveriti sledeće:

- Obeležavanje promenljive identifikatora jedne opservacije (u radu to je CUSTOMER_RK) i promenljive koje predstavljaju vremensku dimenziju uzorka (INFORMATION_DT). Za ove promenljive treba postaviti na ulogu *ID* odnosno *rejected*[14].
- Od svih ciljnih promenljivih izabrati samo jednu, a ostalim ciljnim promenljivim postaviti ulogu na *rejected* jer u jednom trenutku razvijamo samo jedan model.

Više o samoj *Input* komponenti može se naći u dodatku A poglavlje Komponenta *Input Data*.

SAS EM projekat je spreman za istraživanje podataka

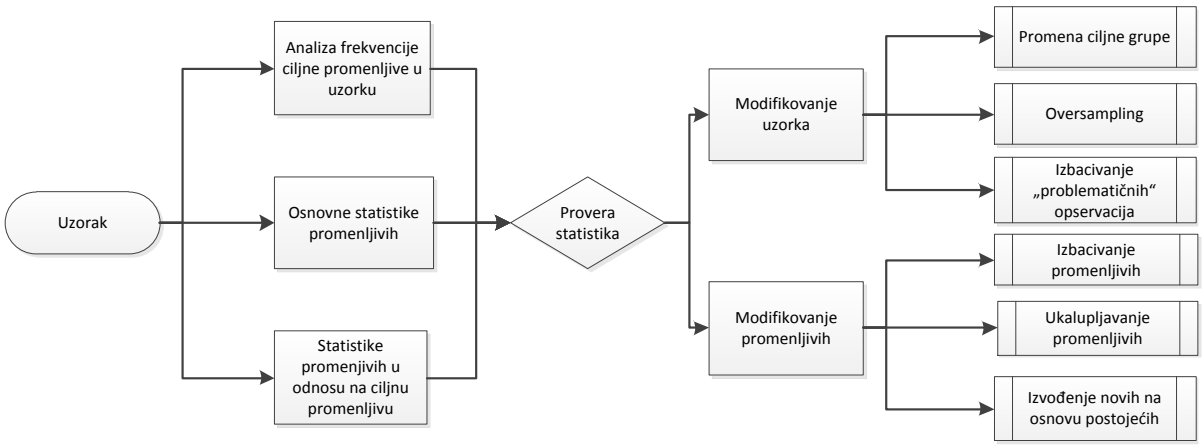
[14] Uloge koje promenljive mogu imati u procesu istraživanja opisane su u dodatku A - poglavlje „Formiranje uzorka (Sampe)"



Slika 14. Podešen SAS EM projekat pre početka istraživanja

6 Preliminarni koraci u razvoju modela

U ovom poglavlju su opisane osnovne tehnike za upoznavanje sa podacima. Upoznavanje je neophodno uraditi pre početka razvoja modela. Na slici (Slika 15) prikazan je dijagram koji opisuje proces preliminiranog istraživanja podataka.



Slika 15. Proces preliminernog istraživanja podataka

Prva tri podpoglavlja predstavljaju procese u preliminarnom istraživanju podataka. To su:

- *Analiza frekvencije ciljne promenljive u uzorku* – ovo poglavlje opisuje tzv. tehniku „*oversampling*” koja se primenjuje u nedostatku događaja prodaje.
- *Osnovne statistike promenljivih*
- *Statistike promenljivih u odnosu na ciljnu promenljivu*

Zavisno od dobijenih rezultata možemo ponoviti postupak pripreme podataka i to:

- Modifikovanjem uzorka – izbacivanje problematičnih opservacija, promenom ciljne grupe (tzv. taktička segmentacija klijenata)
- Modifikovanje promenljivih – izbacivanje problematičnih promenljivih, ukalupljivanje promenljivih, izvođenje novih promenljivih.

Za promenu ciljne grupe koristi se tzv. taktička segmentacija[15]. Ova segmentacija se može osloniti na ponašanje klijenata ali i na popunjenost nekih promenljivih (npr. klijenti koji primaju platu i/ili klijenti koji imaju depozit a nemaju tekući račun).

[15] Taktička segmentacije predstavlja *ad hoc* segmentaciju napravljenu za specifične potrebe. Ona može biti napravljena na osnovu iskustva ili koristeću SAS EM. Ako se koristi SAS EM najčešće se koristi *clustering* metoda gde se identifikuje grupe klijenata sa sličnim ponašanjem. Za svaku grupu se razvija model posebno. Prilikom primene modela prvo se klijent segmentira u grupu, a zatim se računa skor koristeći odgovarajući model.

Pre početka razvoja treba proveriti koliko su promenljive statistički značajne. Ovo je opisano u podpoglavlju *Preliminarni izbor značajnih promenljivih i njihovo istraživanje*.

Izvođenje novih promenljivih može biti dodavanjem novih promenljivih u samom ETL (npr. razni odnosi), grupisanjem vrednosti promenljive u binove[16] (eng. *interactive binning*) ili dimenzionom redukcijom prostora koristeći PCA[17] metodu (eng. *Principal Component Analysis*).

Po završenom preliminarnom istraživanju podataka neophodno je formirati uzorak za trening, proveru ispravnosti (eng. *validation*) i testiranje modela. Ovo je opisano u podpoglavlju *Formiranje uzorka za trening, proveru ispravnosti i testiranje*.

6.1 Analiza frekvencije ciljne promenljive u uzorku

Event populacija uzorka predstavlja podskup uzorka kod kojih je ciljna promenljiva jednaka 1. U ovom slučaju to su klijenti koji su kupili proizvod. *Nonevent* populacija predstavlja podskup uzorka kod kojih je ciljna promenljiva jednaka 0, u ovom slučaju to su klijenti koji nisu kupili proizvod.

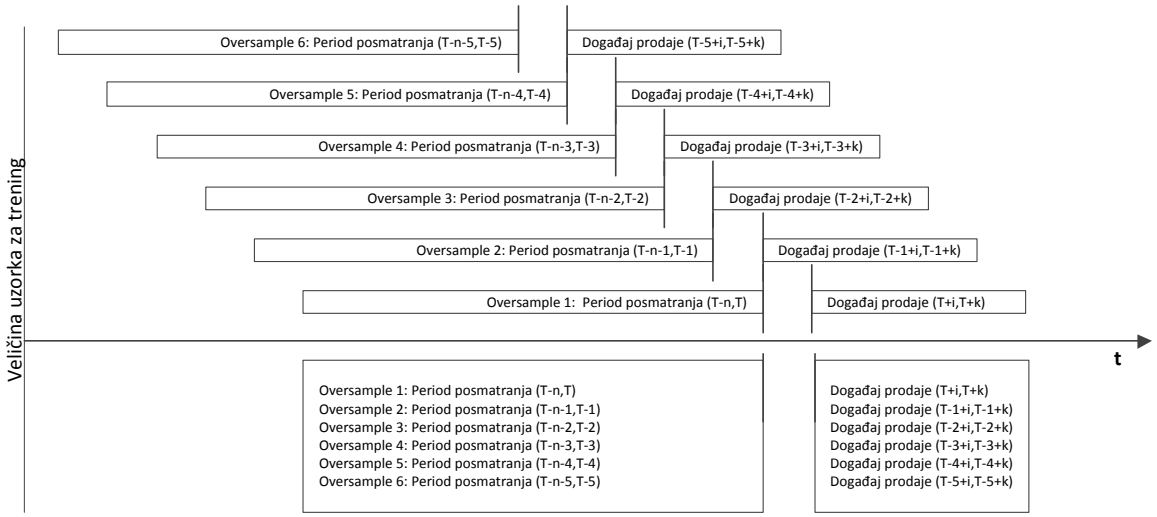
Često su događaji prodaje zastupljeni sa veoma malim procentom u ukupnoj populaciji. Ako je odnos

event/(event+nonevent)<7%

tada obično ne postoji dovoljno događaja na osnovu kojih možemo da razvijemo pouzdan model. Da bi povećali broj događaja prodaje često moramo da spajamo više od jednog uzorka. Na slici (Slika 16) prikazan je postupak pripreme uzorka tzv. *oversampling*.

[16] Binovi predstavljaju podelu domena promenljive na disjunkte intervale ako je promenljiva kontinulana, odnosno disjunkte grupe (podskupove domena) ako je promenljiva kategorička.

[17] PCA metoda je opisana u Dodatku A i B.



Slika 16. Priprema uzorka - oversampling

Za pripremu uzorka koristimo 6 vremenskih trenutaka. Algoritam je:

1. Za najsveži mesec iz populacije *nonevent* na slučajan način se bira 3 do 5 puta više opservacija u odnosu na broj *event* opservacija.
2. Za svaki mesec unazad ponovimo postupak ali vodeći računa da u celoj populaciji nemamo već uzete klijente
3. Spojimo generisani uzorak u jedan ABT.

Uzorak u ovako generisanom ABT koji se koristi u daljem radu ima odnos:

Event/(Event+Nonevent)≈15%

Napomena 1: Model napravljen iz uzorka koji je pripremljen na ovakav način ne daje realnu verovatnoću događaja. Ovako dobijenu verovatnoću neophodno je skalirati u odnosu na originalni uzorak. To se radi u slučaju da ovu verovatnoću treba porediti sa verovatnoćom ostalih modela. U slučaju da to nije potrebno iskrivljena verovatnoća čuva poredak tj. ako je $RV_1<RV_2$ tada je $OV_1<OV_2$ i obratno[18].

Napomena 2: Može se desiti da je *event* populacija mnogo veća od *nonevent* populacije. Zavisno od tehnika izrade modela koristi se sličan algoritam za umanjenje (eng. *undersampling*) *event* populacije u uzorku.

6.2 Osnovne statistike promenljivih

Ovo je prvi kontakt sa podacima osobe (u daljem tekstu *miner*) koja razvija model u slučaju da nije aktivno učestvovao/la u pripremi podataka. Bez obzira da li se radi o celoj

[18] RV – realna verovatnoća, OV-verovatnoća izračunata pomoću modela

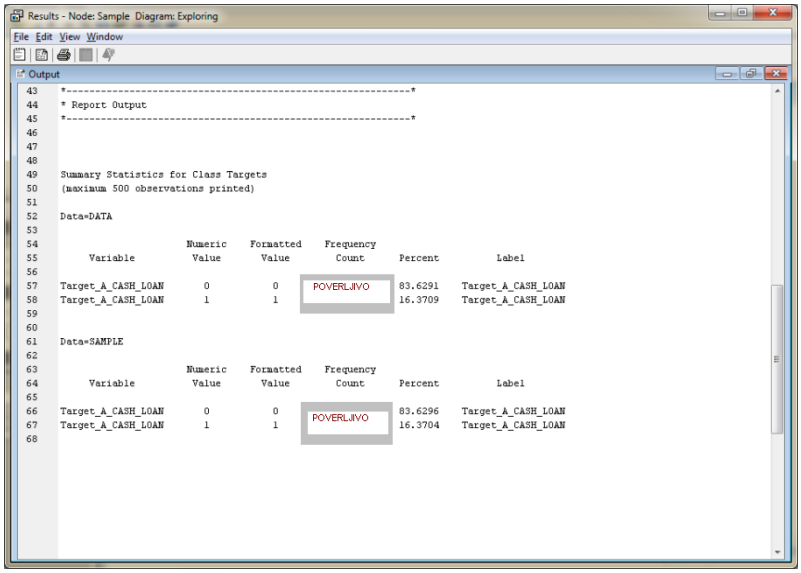
ili populaciji modifikovanoj tehnikom *oversampling*-a (ili *undersampling*-a), *miner* prolazi kroz sledeće faze:

- formiranje reprezentativnog uzorka - uzorak iz uzorka[19] kako bi se ubrzala izrada statistika,
- izrada osnovnih statistika,
- čitanje rezultata,
- preduzimanje akcija za korigovanjem uzorka.

Ovaj proces je iterativnan i može se ponoviti nekoliko puta.

6.2.1 Kreiranje reprezentativnog uzorka za preliminarno istraživanje podataka

Uzorak za modelovanje može biti veliki i zbog toga bi preliminarna istraživanja trebalo raditi nad manjim poduzorkom kako bi se ubrzala izrada statistika. U slučaju da izrada statistika ne traje dugo najbolje je iz uzorka uzeti populaciju koja je iste veličine kao i uzorak za trening modela. U ovom radu to je 60% od ukupnog uzorka. Prilikom izrade uzorka je primenjena metoda stratifikacije u odnosu na ciljnu promenljivu. Ovo znači da u 60% populacije proporcija *event/nonevent* je približna proporciji nad celim uzorkom.



Slika 17. Rezultati komponete *Samle*

Komponenta je opisana u dodatku A *Komponenta Sample*.

[19] Formiranje reprezentativnog uzorka iz uzorka se često primenjuje u preliminarnim istraživanjima. Sam uzorak može imati više stotina hiljada opservacija i formiranje reprezentativnog uzorka iz uzorka ubrzaće izradu statistika.

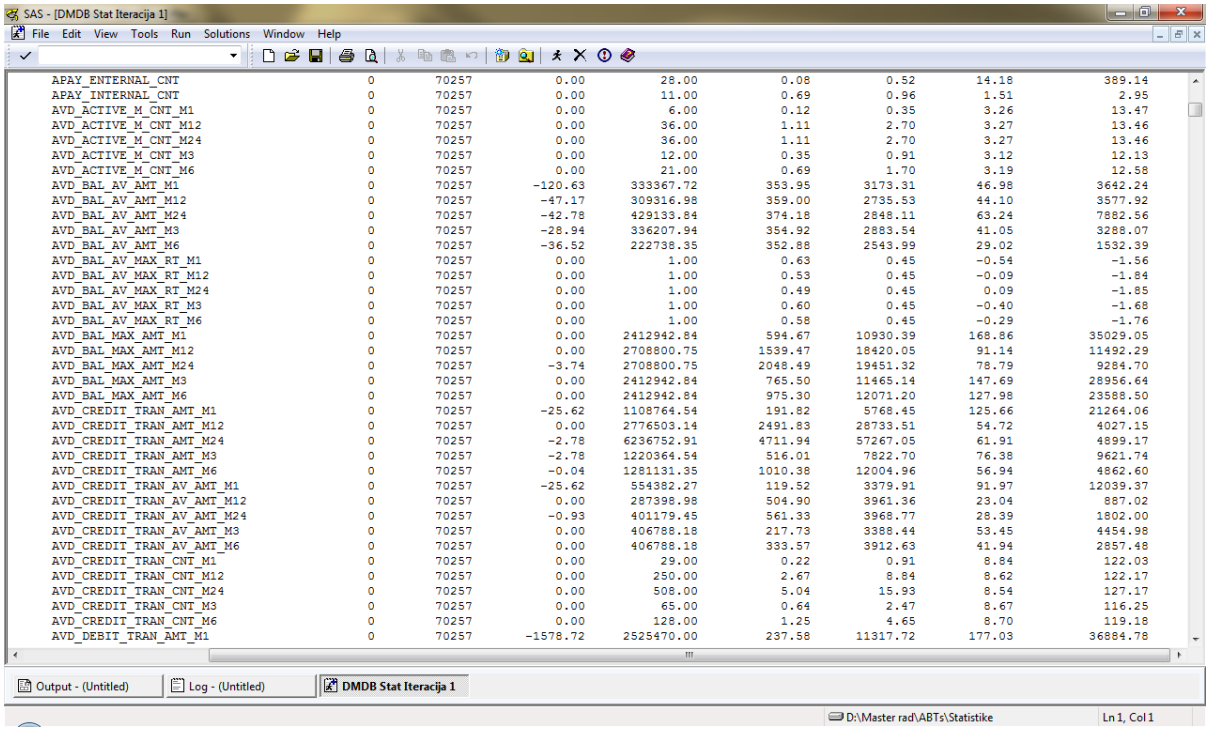
6.2.2 Izrada osnovnih statistika

Prve statističke mere za kontinualne promenljive koje se gledaju u procesu istraživanja su:

- Minimum
- Maksimum
- Prosečna vrednost (eng. *mean*)
- Standardna devijacija (eng. *standard deviation*)
- Izobličenje (eng. *skewenss*)
- Kurtosis

Ove statistike su opisane u dodatku B *Matematičke osnove*.

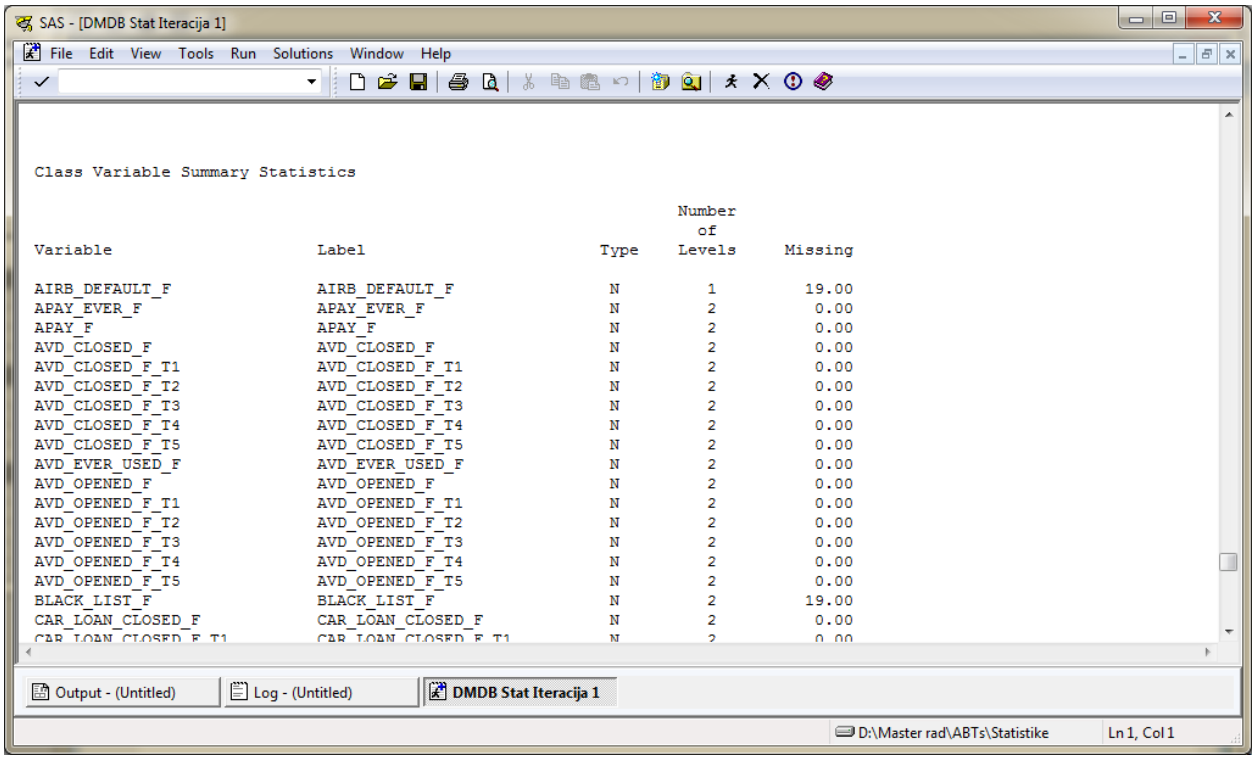
Na slici (Slika 18) su prikazane osnovne statistike kontinualnih promenljivih dobijenih korišćenjem DMDB komponente. DMDB komponenta je opisana u dodatku A poglavlje *Komponenta DMDB*.



APAY_ENTERNAL_CNT	0	70257	0.00	28.00	0.08	0.52	14.18	389.14
APAY_INTERNAL_CNT	0	70257	0.00	11.00	0.69	0.96	1.51	2.95
AVD_ACTIVE_M_CNT_M1	0	70257	0.00	6.00	0.12	0.35	3.26	13.47
AVD_ACTIVE_M_CNT_M12	0	70257	0.00	36.00	1.11	2.70	3.27	13.46
AVD_ACTIVE_M_CNT_M24	0	70257	0.00	36.00	1.11	2.70	3.27	13.46
AVD_ACTIVE_M_CNT_M3	0	70257	0.00	12.00	0.35	0.91	3.12	12.13
AVD_ACTIVE_M_CNT_M6	0	70257	0.00	21.00	0.69	1.70	3.19	12.58
AVD_BAL_AV_AMT_M1	0	70257	-120.63	333367.72	353.95	3173.31	46.98	3642.24
AVD_BAL_AV_AMT_M12	0	70257	-47.17	309316.98	359.00	2735.53	44.10	3577.92
AVD_BAL_AV_AMT_M24	0	70257	-42.78	429133.84	374.18	2848.11	63.24	7882.56
AVD_BAL_AV_AMT_M3	0	70257	-28.94	336207.94	354.92	2883.54	41.05	3288.07
AVD_BAL_AV_AMT_M6	0	70257	-36.52	222738.35	352.88	2543.99	29.02	1532.39
AVD_BAL_AV_MAX_RT_M1	0	70257	0.00	1.00	0.63	0.45	-0.54	-1.56
AVD_BAL_AV_MAX_RT_M12	0	70257	0.00	1.00	0.53	0.45	-0.09	-1.84
AVD_BAL_AV_MAX_RT_M24	0	70257	0.00	1.00	0.49	0.45	0.09	-1.85
AVD_BAL_AV_MAX_RT_M3	0	70257	0.00	1.00	0.60	0.45	-0.40	-1.68
AVD_BAL_AV_MAX_RT_M6	0	70257	0.00	1.00	0.58	0.45	-0.29	-1.76
AVD_BAL_MAX_AMT_M1	0	70257	0.00	2412942.84	594.67	10930.39	168.86	35029.05
AVD_BAL_MAX_AMT_M12	0	70257	0.00	2708800.75	1539.47	18420.05	91.14	11492.29
AVD_BAL_MAX_AMT_M24	0	70257	-3.74	2708800.75	2048.49	19451.32	78.79	9284.70
AVD_BAL_MAX_AMT_M3	0	70257	0.00	2412942.84	765.50	11465.14	147.69	28956.64
AVD_BAL_MAX_AMT_M6	0	70257	0.00	2412942.84	975.30	12071.20	127.98	23588.50
AVD_CREDIT_TRAN_AMT_M1	0	70257	-25.62	1108764.54	191.82	5768.45	125.66	21264.06
AVD_CREDIT_TRAN_AMT_M12	0	70257	0.00	2776503.14	2491.83	28733.51	54.72	4027.15
AVD_CREDIT_TRAN_AMT_M24	0	70257	-2.78	6236752.91	4711.94	57267.05	61.91	4899.17
AVD_CREDIT_TRAN_AMT_M3	0	70257	-2.78	1220364.54	516.01	7822.70	76.38	9621.74
AVD_CREDIT_TRAN_AMT_M6	0	70257	-0.04	1281131.35	1010.38	12004.96	56.94	4862.60
AVD_CREDIT_TRAN_AV_AMT_M1	0	70257	-25.62	554382.27	119.52	3379.91	91.97	12039.37
AVD_CREDIT_TRAN_AV_AMT_M12	0	70257	0.00	287398.98	504.90	3961.36	23.04	887.02
AVD_CREDIT_TRAN_AV_AMT_M24	0	70257	-0.93	401179.45	561.33	3968.77	28.39	1802.00
AVD_CREDIT_TRAN_AV_AMT_M3	0	70257	0.00	406788.18	217.73	3388.44	53.45	4454.98
AVD_CREDIT_TRAN_AV_AMT_M6	0	70257	0.00	406788.18	333.57	3912.63	41.94	2857.48
AVD_CREDIT_TRAN_CNT_M1	0	70257	0.00	29.00	0.22	0.91	8.84	122.03
AVD_CREDIT_TRAN_CNT_M12	0	70257	0.00	250.00	2.67	8.84	8.62	122.17
AVD_CREDIT_TRAN_CNT_M24	0	70257	0.00	508.00	5.04	15.93	8.54	127.17
AVD_CREDIT_TRAN_CNT_M3	0	70257	0.00	65.00	0.64	2.47	8.67	116.25
AVD_CREDIT_TRAN_CNT_M6	0	70257	0.00	128.00	1.25	4.65	8.70	119.18
AVD_DEBIT_TRAN_AMT_M1	0	70257	-1578.72	2525470.00	237.58	11317.72	177.03	36884.78

Slika 18. Rezultat primene DMDB komponente na uzorku

Za kategoričke promenljive (eng. *class variables*) proverava se kardinalnost (mera *Number of Levels* - Slika 19), kao i koliko ima nedostajućih vrednosti (mera *Missing* - Slika 19)



Slika 19. Statistike kategoričkih promenljivih

6.2.3 Rezultati istraživanja

Osnovne statistike kao što su:

- lokacija (minimum, maksimum, prosečna vrednost, modus),
- disperzija (standardna devijacija, percentili, interkvartilni ospeg i sl.)
- oblici (varijansa, izobličenje, kurtosis,…)

daju prvu sliku o podacima. Na osnovu ovih rezultata moguće je uočiti:

- nelogičnosti u samim podacima kao što su:
  - postojanje potrošnje po kartici koja je negativna
  - postojanje kredita u pretplati
  - datum izdavanja čeka je manji od datuma otvaranja tekućeg računa.
- kako su oblikovani podaci:
  - sve mere stanja (ima ih više od 200) imaju veliku standardnu devijaciju
  - sve mere imaju veliko pozitivno izobličenje
- kako su urađene transformacije nekih promenljivh
  - vrednosti koje menjaju nedostajuće vrednosti su prevelike (ovo je dobro za *interactive bining* ali nije dobro za direktno korišćenje regresije).

6.2.4 Preduzete akcije

Neophodno je uraditi čišćenje podataka da bi se ove promenljive ispravile. Dakle, imamo još jednu iteraciju pripreme podataka.

U ovom slučaju *miner* može lako prepraviti ETL2 (*view layer* –videti poglavlje „*Priprema podataka za modelovanje*“), tako da se ublaži efekat ovih anomalija. To može uraditi na sledeće načine:

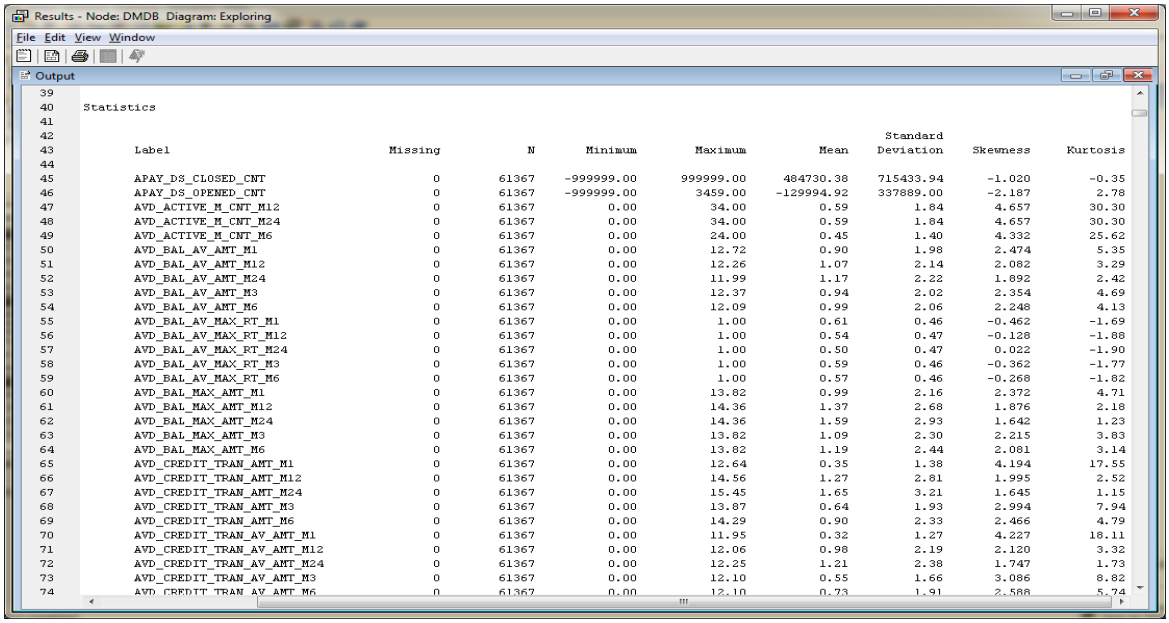
- Postavljanjem problematičnih vrednosti na NULL pa zatim na podrazumevanu vrednost za *NULL* (za većinu gore navedenih promenljivih je 0)
- Izostavljanjem problematičnih opservacija iz razvojnog uzorka.

Drugi metod treba sprovoditi samo u krajnjoj nuždi jer izbacivanjem problematičih opservacija se smanjuje populacija nad kojom se model može primeniti. Ako se pokaže da neka od sumnjivih promenljivih prediktivna tj. bude uzeta za modelovanje, model se ne može primeniti na opservacije koje imaju problematičnu vrednost. S druge strane ponekad je potrebno podesiti da ove promenljive budu postavljene na neke specijalne vrednosti, tako da se razlikuju od vrednosti u slučaju da klijent nema proizvod (nedostajuća vrednost).

Standardnu devijaciju moguće je popraviti logaritmovanjem. U radu je za sve pozitivne vrednosti stanja (%_AMT promenljive) primenjena funkcija LOG(X+1). Ovakav pristup ima svoje prednosti i mane. Prednost je sabijanje vrednosti u manji interval. Mana je otežano izvođenje novih promenljivih iz postojećih (logaritmovanih).

6.2.5 Rezultati ispravke

Posle ispravke kroz ETL2 i ponovnog procesiranja uzorka statističke mere imaju bolje vrednosti. Na slici (Slika 20) su prikazani rezultati ispravke.


Results - Node: DMDB Diagram: Exploring

File Edit View Window

Output

Statistics

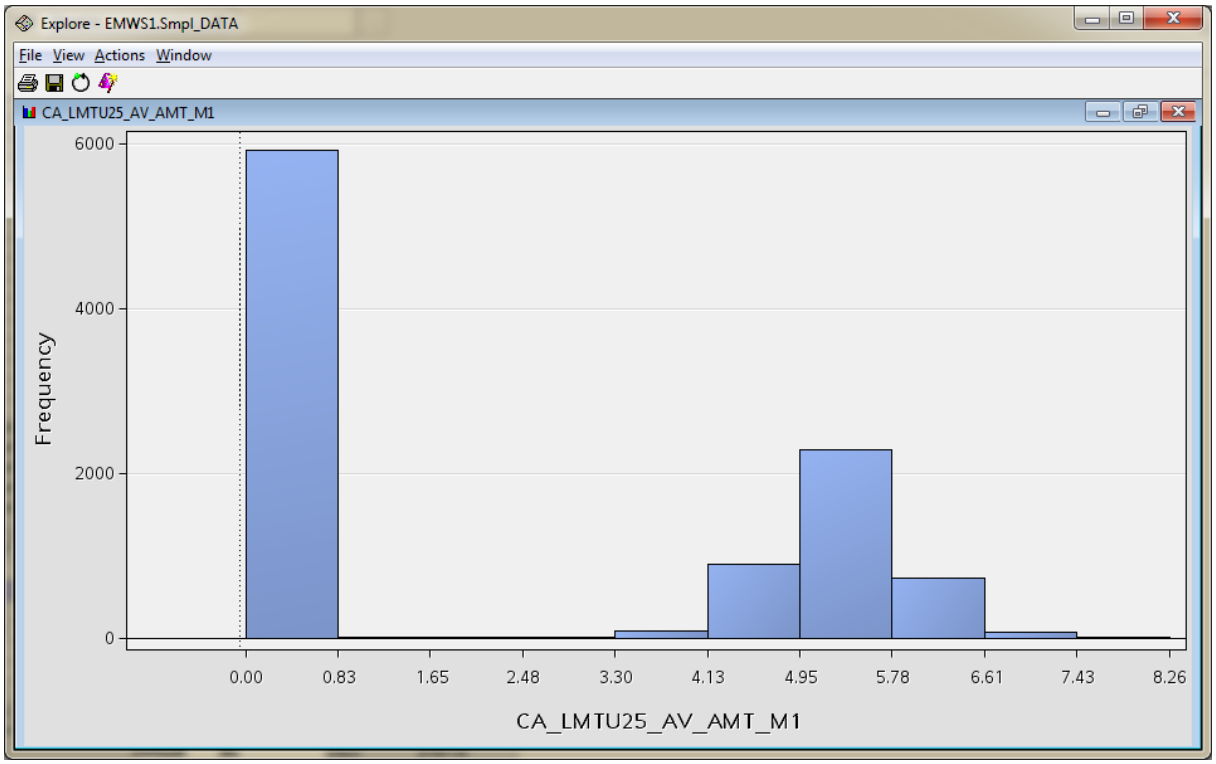
Label	Missing	N	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
APAY_DS_CLOSED_CNT	0	61367	-999999.00	999999.00	484730.38	715433.94	-1.020	-0.35
APAY_DS_OPENED_CNT	0	61367	-999999.00	3459.00	-129994.92	337889.00	-2.187	2.78
AVD_ACTIVE_M_CNT_M12	0	61367	0.00	34.00	0.59	1.84	4.657	30.30
AVD_ACTIVE_M_CNT_M24	0	61367	0.00	34.00	0.59	1.84	4.657	30.30
AVD_ACTIVE_M_CNT_M6	0	61367	0.00	24.00	0.45	1.40	4.332	25.62
AVD_BAL_AV_AMT_M1	0	61367	0.00	12.72	0.90	1.98	2.474	5.35
AVD_BAL_AV_AMT_M12	0	61367	0.00	12.26	1.07	2.14	2.082	3.29
AVD_BAL_AV_AMT_M24	0	61367	0.00	11.99	1.17	2.22	1.892	2.42
AVD_BAL_AV_AMT_M3	0	61367	0.00	12.37	0.94	2.02	2.354	4.69
AVD_BAL_AV_AMT_M6	0	61367	0.00	12.09	0.99	2.06	2.248	4.13
AVD_BAL_AV_MAX_RT_M1	0	61367	0.00	1.00	0.61	0.46	-0.462	-1.69
AVD_BAL_AV_MAX_RT_M12	0	61367	0.00	1.00	0.54	0.47	-0.128	-1.88
AVD_BAL_AV_MAX_RT_M24	0	61367	0.00	1.00	0.50	0.47	0.022	-1.90
AVD_BAL_AV_MAX_RT_M3	0	61367	0.00	1.00	0.59	0.46	-0.362	-1.77
AVD_BAL_AV_MAX_RT_M6	0	61367	0.00	1.00	0.57	0.46	-0.268	-1.82
AVD_BAL_MAX_AMT_M1	0	61367	0.00	13.82	0.99	2.16	2.372	4.71
AVD_BAL_MAX_AMT_M12	0	61367	0.00	14.36	1.37	2.68	1.876	2.18
AVD_BAL_MAX_AMT_M24	0	61367	0.00	14.36	1.59	2.93	1.642	1.23
AVD_BAL_MAX_AMT_M3	0	61367	0.00	13.82	1.09	2.30	2.215	3.83
AVD_BAL_MAX_AMT_M6	0	61367	0.00	13.82	1.19	2.44	2.081	3.14
AVD_CREDIT_TRAN_AMT_M1	0	61367	0.00	12.64	0.35	1.38	4.194	17.55
AVD_CREDIT_TRAN_AMT_M12	0	61367	0.00	14.56	1.27	2.81	1.995	2.52
AVD_CREDIT_TRAN_AMT_M24	0	61367	0.00	15.45	1.65	3.21	1.645	1.15
AVD_CREDIT_TRAN_AMT_M3	0	61367	0.00	13.87	0.64	1.93	2.994	7.94
AVD_CREDIT_TRAN_AMT_M6	0	61367	0.00	14.29	0.90	2.33	2.466	4.79
AVD_CREDIT_TRAN_AV_AMT_M1	0	61367	0.00	11.95	0.32	1.27	4.227	18.11
AVD_CREDIT_TRAN_AV_AMT_M12	0	61367	0.00	12.06	0.98	2.19	2.120	3.32
AVD_CREDIT_TRAN_AV_AMT_M24	0	61367	0.00	12.25	1.21	2.38	1.747	1.73
AVD_CREDIT_TRAN_AV_AMT_M3	0	61367	0.00	12.10	0.55	1.66	3.086	8.82
AVD_CREDIT_TRAN_AV_AMT_M6	0	61367	0.00	12.10	0.73	1.91	2.588	5.74

Slika 20. Ispravljene statistike mera stanja

Tako, mere stanja uglavnom imaju standardnu devijaciju manju od 2.

Skewenss i *Kurtosis* su takođe mali što nam govori da kriva distribucije ne naginje mnogo levo i desno kao i da ne postoj više od 2 „šiljka" (eng. *peak*) odnosno da „rame"[20] distibucije naglo pada. Jedan šiljak predstavlja vrednosti oko 0, a drugi predstavlja pravu distribuciju.

Na slici (Slika 21) prikazan je primer distribucije promenljive "*prosečno negativno stanje po tekućem računu u danima kada je klijent imao iskorišćenost limita veću od 25%*".



Slika 21. Distribucija ukalupljene promenljive CA_LMT25_AV_AMT_M1

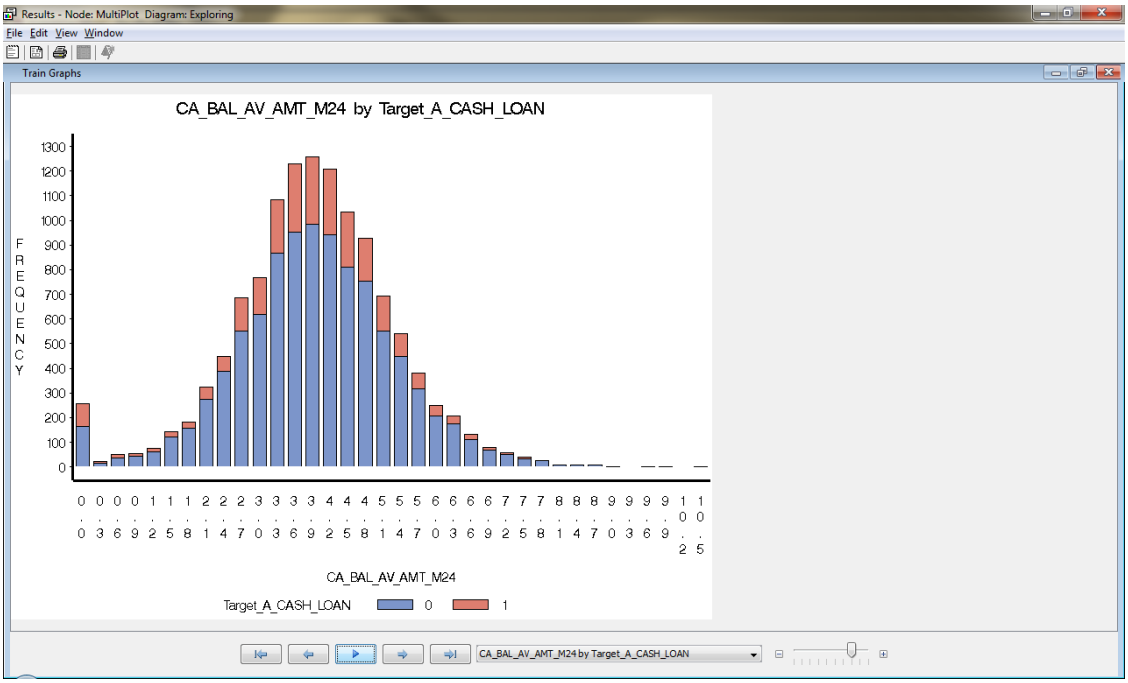
Naravno, i dalje postoje promenljive kod kojih i posle preduzetih akcija nemaju „dobre" statistike. Njih će alat najverovatnije sam odbaciti.

6.3 Statistike promenljivih u odnosu na ciljnu promenljivu

Veoma je važno napraviti analizu kako se neke promenljive ponašaju odvojeno za event i nonevent populaciju, a zatim uporediti statistike.

Ove analize je najbolje napraviti izradom različitih grafikona na kojima se mogu uočiti različita pravila. Za ovo svrhu koristi se *bar char* grafikon sa stubićima, stubićni dijagram, gde je na jednom stubiću prikazan i odnos *event/nonevent* u populaciji koje taj stubić prikazuje. Ovo je najjednostavnije uraditi koristeći komponentu *MultiPlot*.

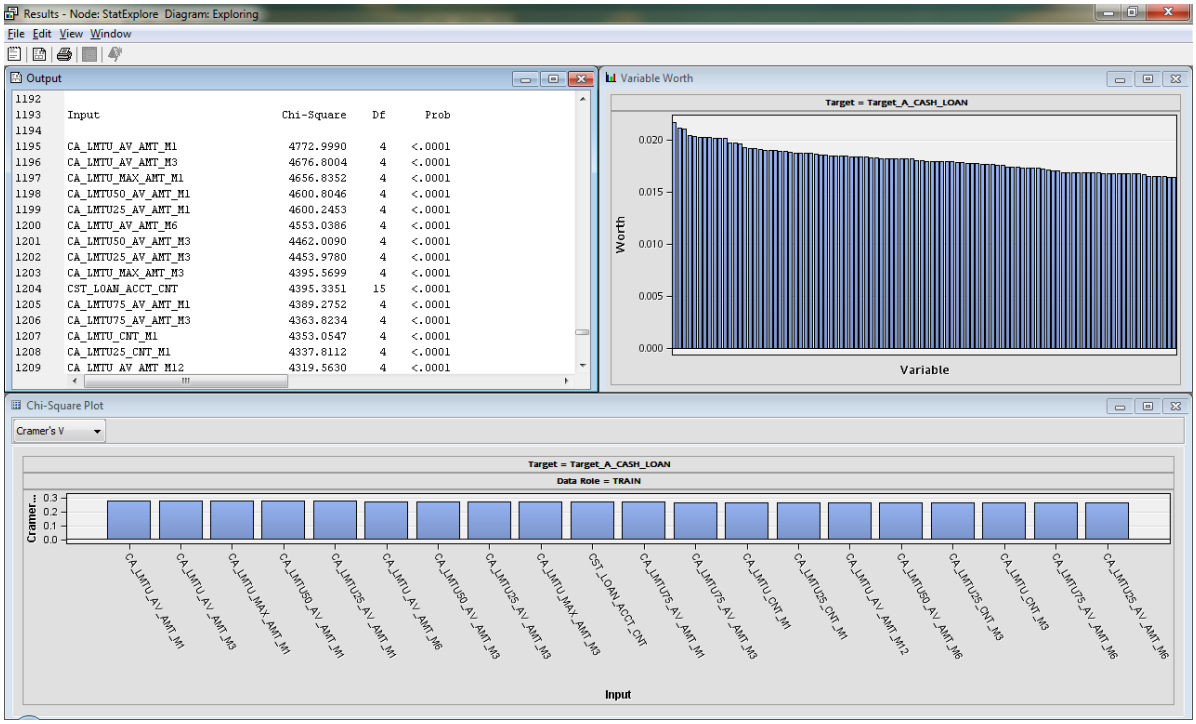
[20] Rame predstavlja središnju oblast između vrha i repa krive distribucije.



Slika 22. Distibucija promenljive CA_BAL_AV_AMT_M24 sa prikazanim odnosom event/nonevent

S obzirom da u radu postoji preko 2000 promenljivih pregled svih grafikona može biti izuzetno naporan.

Zbog toga je neophodno odvojiti statistički značajne promenljive. Za ovu *ad hoc* analizu korišćena je *StatExplore* komponenta koja pomoću *Chi-Square* i *Cramer's V* statistika određuje statistički znaćajne promenljive. Komponenta je opisana u dodatku A poglavlje Komponenta *Stat Explore*.



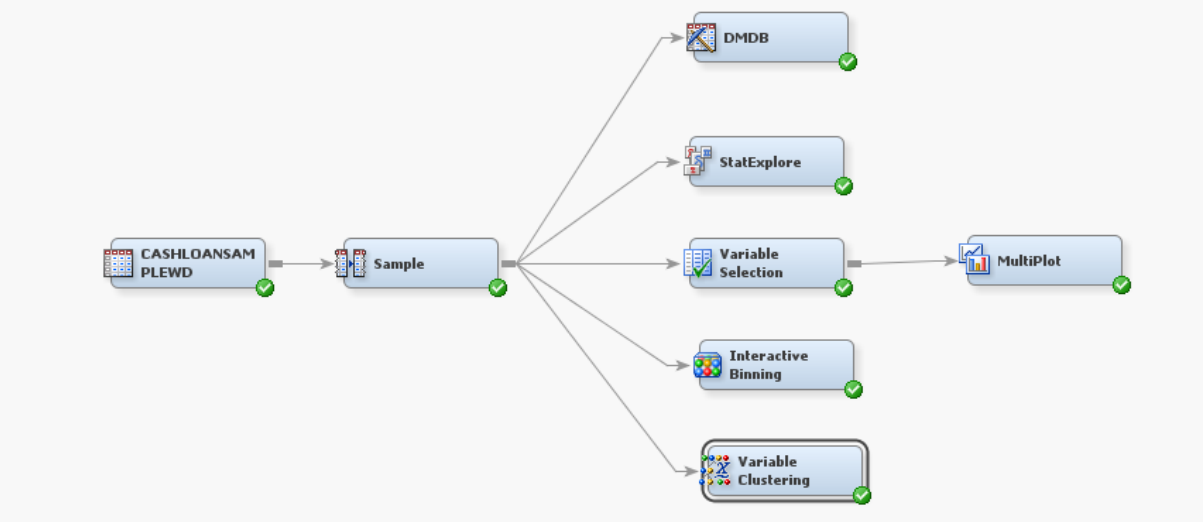
Slika 23. Statistički značajne promenljive dobijenje pomoću *StatExplore* komponente

Za svaku od statistički značajnih promenljivih (Slika 23) treba ispitati distribuciju i odnos *event* i *nonevent* populacije.

Napomena: Problem sa *StatExplore* komponentom je nemogućnost automatskog odbacivanja promenljivih koje nisu značajne. Zbor toga *MultiPlot* i *GraphExplore* komponente formiraju grafikone za sve promenljive što može biti poprilično sporo. Osnovne statistike izabranih promenljivih moguće je dobiti kroz skoro sve SAS komponente.

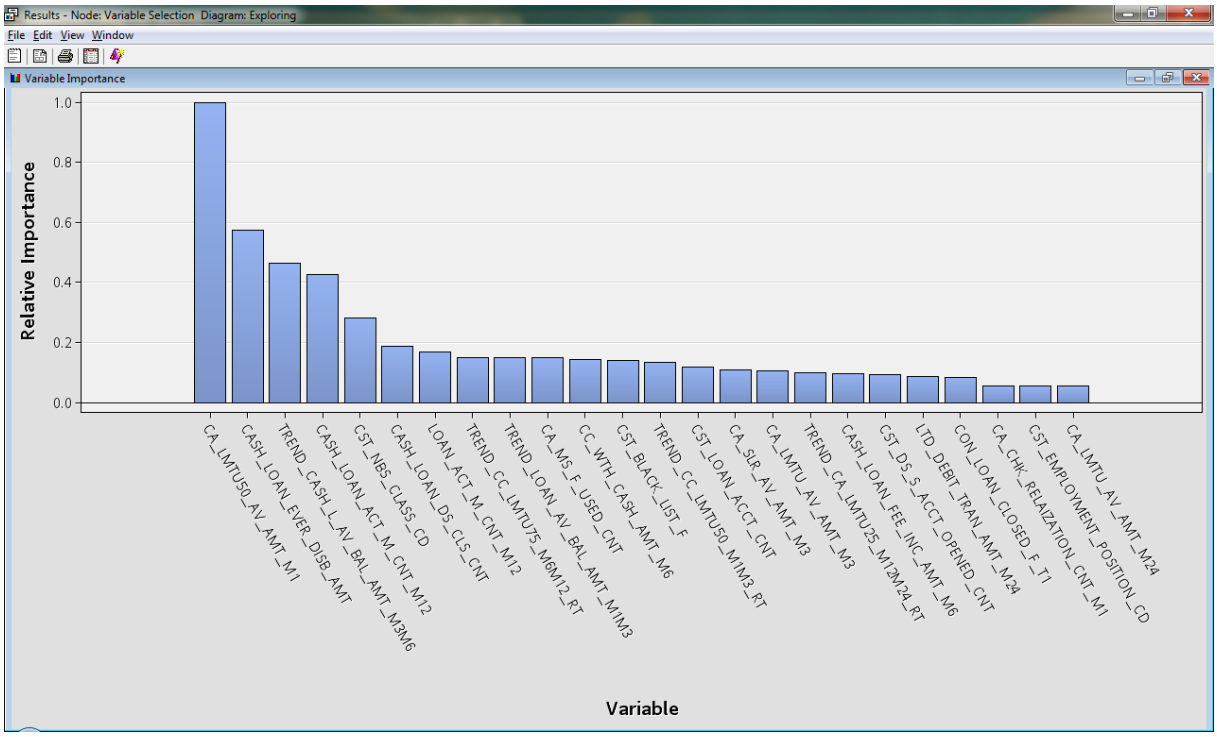
6.4 Preliminarni izbor značajnih promenljivih i njihovo istraživanje

Pre početka razvoja dobro je proveriti koliko su promenljive statistički značajne u odnosu na ciljnu promenljivu. *VariableSelection* komponenta bira promenljive na osnovu r-kvadrat (eng. *R-square*) i hi-kvadrat (eng. *Chi-square*) kriterijuma. Komponente su opisane u dodatku A *Upoznavanje sa podacima, istraživanje podataka - Explore.*



Slika 24. Radna površina SAS EM u fazi preliminarnog istraživanja

Rezultat izdvajanja je smanjen skup promenljivih. Nad ovim skupom se može primeniti komponeta *MultiPlot* radi analize svake prediktivne promenljive zasebno.

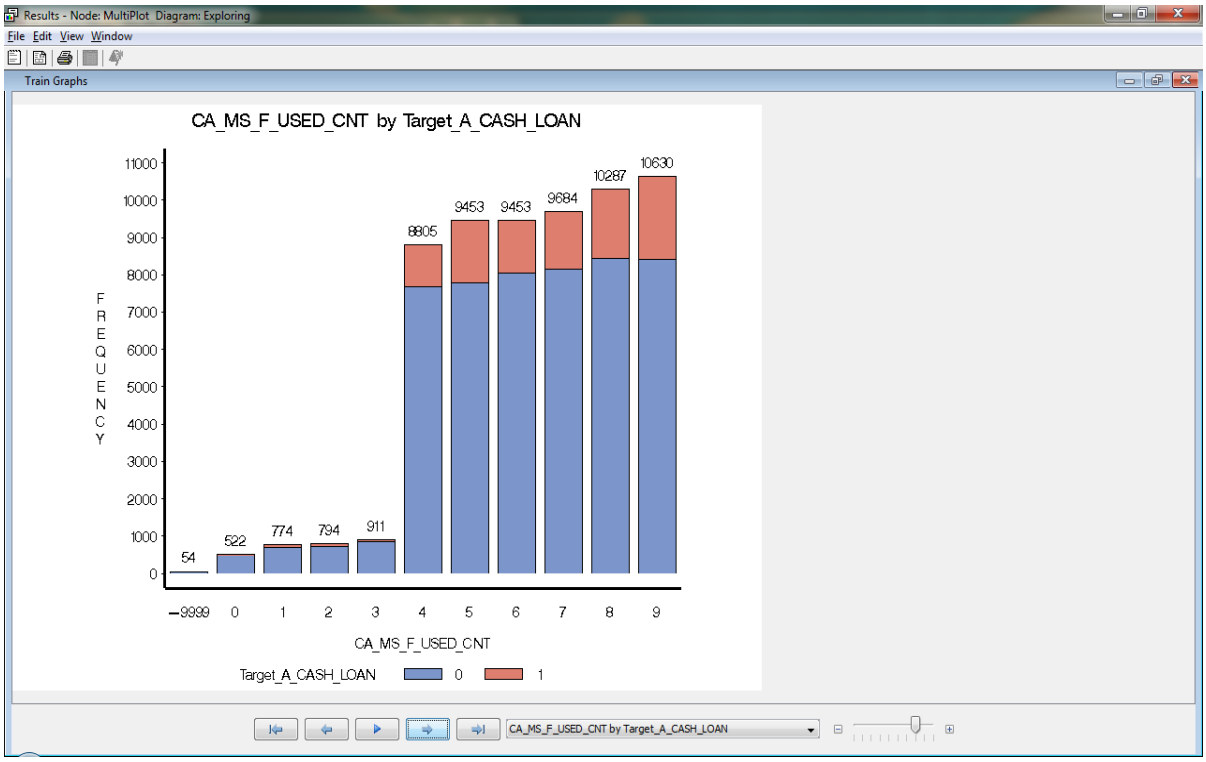


Slika 25. Rezultat selekcije promenljivih primenom VariableSelection komponente

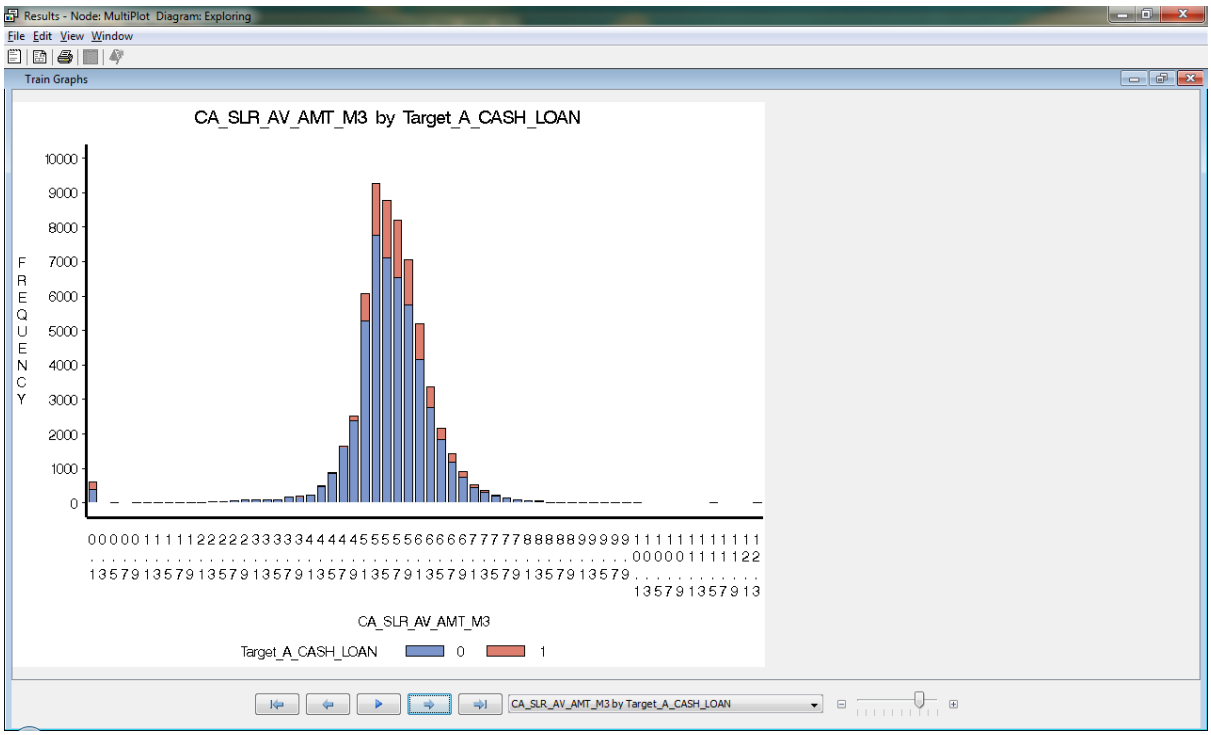
Napomena 1: U slučaju da uzorak nije dobar tj. greškom se neke opservacije ponavljaju lako se može desiti da *VariableSelection* odbaci sve promenljive sa upozorenjem da najverovatnije postoje duple opservacije. To se dogodilo u ovom radu gde je zbog manje od 1% duplih opservacija komponenta pokazala da nema prediktivnih promenljivih. Uzrok dupliranja koji je nastao u postupku *oversampling*-a je otklonjen u narednoj iteraciji.

Napomena 2: Kod dobro pripremljenih uzoraka može se desiti da su sve promenljive podjednako važne. U tom slučaju mera *Relative Importance* se ne može izračunati tj. sve promenljive će biti odbačene. Ovde imamo dve opcije: da malo „pokvarimo“ uzorak ili da primenimo druge tehnike u odabiru promenljivih.

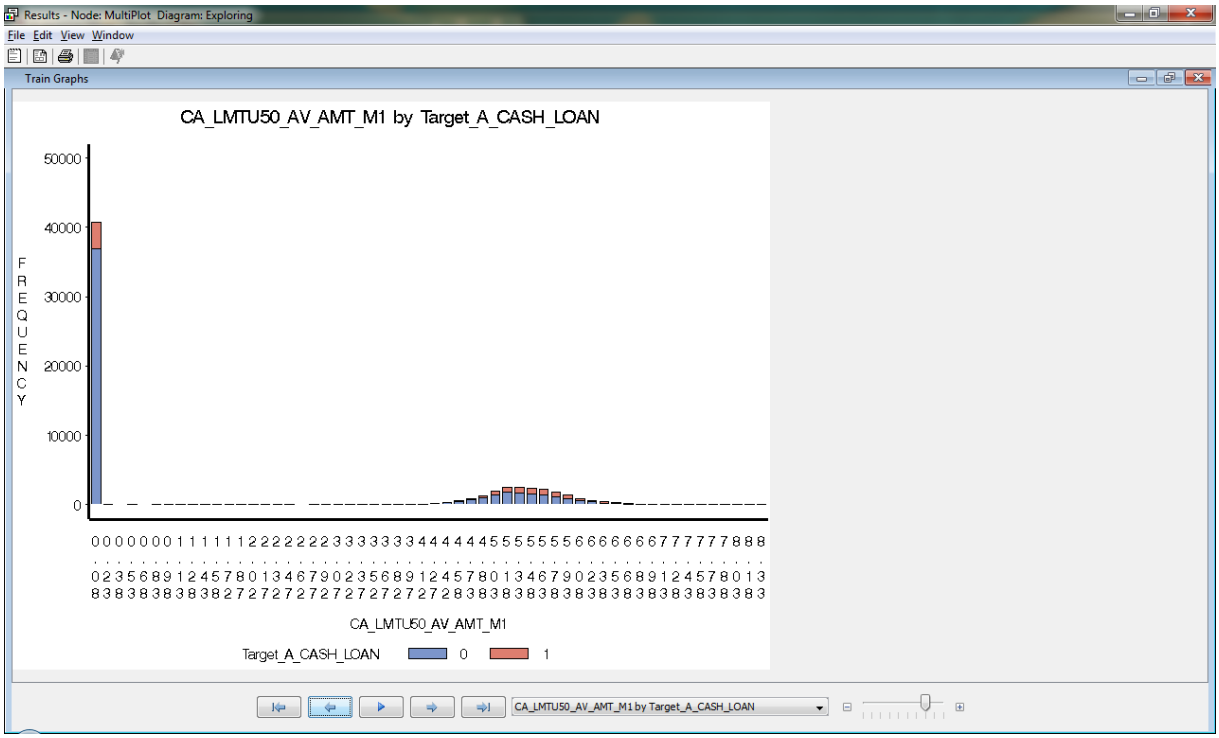
Nad ovim promenljivim možemo uraditi analizu udela *event* i *nonevent* populacije. Evo nekoliko slika:



Slika 26. Distribucija intervalne promenljive CA_MS_F_USED_CNT



Slika 27. Distirbucija promenljive CA_SLR_AV_AMT_M3 - prosečni tromesečni priliv po osnovu zarade



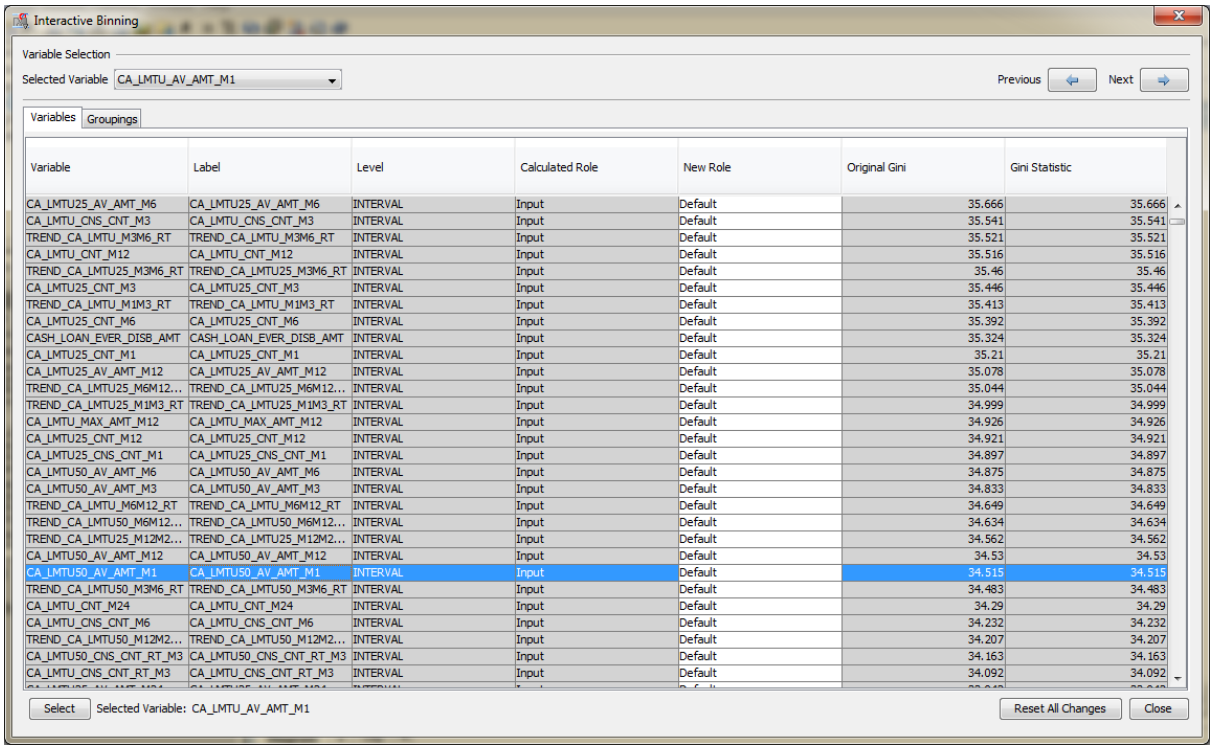
Slika 28. Distribucija promenljive CA_LMTU_AV_AMT_M1 – prosečno negativno stanje na tekućem računu u danima kada je klijent imao iskorišćenost granice veće od 50%.

Na slici (Slika 28) prikazana je promenljiva CA_LMTU50_AV_AMT_M1 koja predstavlja prosečno negativno stanje na tekućem računu u danima kada je klijent imao iskorišćenost granice veću od 50%. Distrubucija je ravnomerna osim u slučaju kada klijent nema proizvod (dozvoljeno prekoračenje) ili ima proizvod ali ga ne koristi. U ovom radu nedostajuće vrednosti uglavnom su menjane sa 0. Ponekad je dobro razdvojiti one klijente koji imaju proizvod ali ga ne koriste od onih koje nemaju proizvod. Ovo se obično radi u situacijama kada se za izbor značajnih promenljivih koristi Gini koeficijent[21] i komponenta *Interactive Binning*[22]. Zbog toga, preliminarno su napravljene grupe preko komponente *Interactive Binning* koristeći *Gini* koeficijent.

Na slici (Slika 29) je prikazana lista grupa koje imaju *Gini* koeficijent veći od 20.

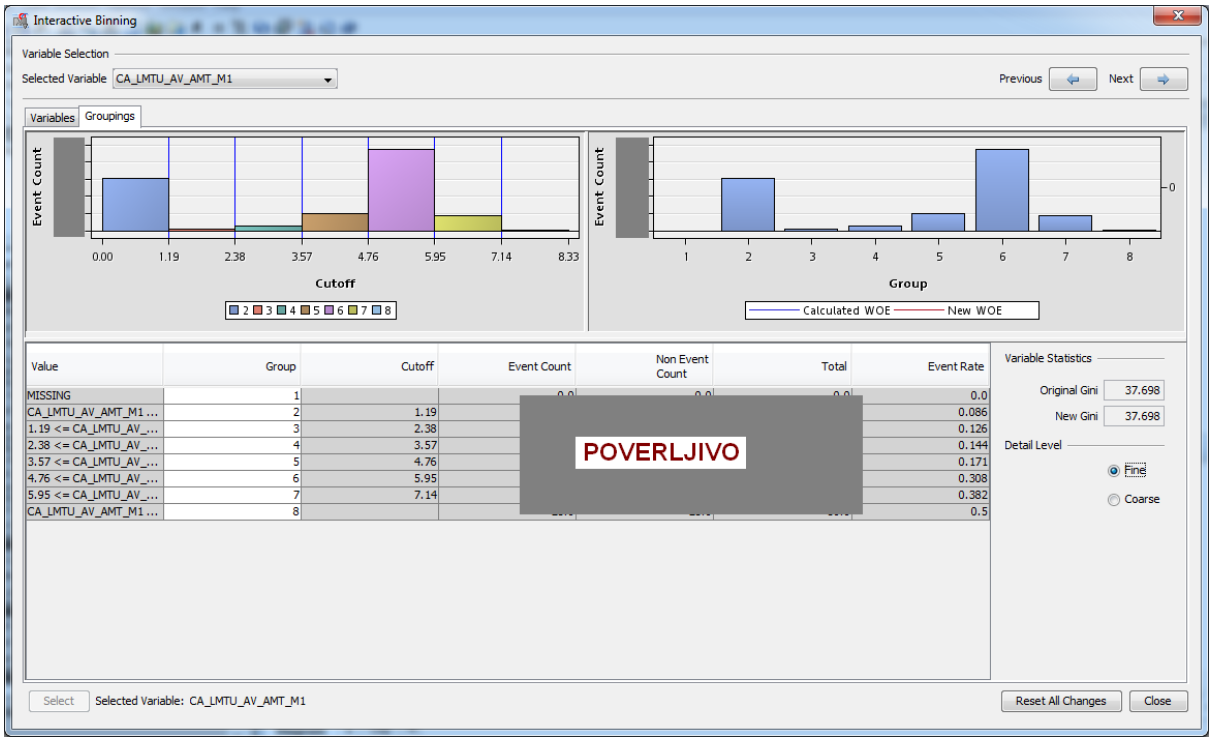
[21] Gini koeficijent je opisan u dodatku B „Matematičke osnove“

[22] Komponenta Interactive Binnig opisana je u dodatku A „SAS Enterprise Miner“



Slika 29. Rezultat izbora promenljivih pomoću *InteractiveBinning* komponente

Promenljivu CA_LMTU50_AV_AMT_M1 (distribucija se nalazi na slici 28) sa Gini koeficijentom 34,515 je visoko kotirana. Za nju je komponenta *InteractiveBinning* napravila grupe (Slika 30)



Slika 30. Intervali napravljeni pomoću Interactive Binning za promenljivu CA_LMTU50_AV_AMT_M1

Koristeći komponente *VariableSelection* i *InteractiveBinning* može se videti da je dosta „sličnih“ promenljivih označeno kao značajne. Tako su i promenljive CA_LMTU_AV_AMT_M3 i CA_LMTU_AV_AMT_M24 označene kao značajne iako je prva prosečno iskorišćeni limit po tekućem računu u prethodna 3 meseca, dok je druga u prethodna 24 meseca.

U ovom trenutku nećemo ispitivati kolinearnost ovih promenljivih, ali se pomoću komponete *Variable Clustering* mogu odrediti klasteri promenljivih tj. možemo proveriti da li će obe ove promenljive pripadati istom klasteru. Komponenta Variable Clustering opisana je u dodatku A Komponenta *Varijable Clustering*.

Zaključak: Imamo dovoljno statistički značajnih promenljivih da bi krenuli u proces izrade modela.

Ovo je trenutak kada je završena priprema podataka i prelazi se na razvoj modela.

6.5 Formiranje uzorka za trening, proveru ispravnosti i testiranje

Po završenoj preliminarnoj analizi neophodno je podeliti (*eng. data partition*) uzorak na dve do tri grupe. Jedan deo će se koristi za trening, drugi za proveru ispravnosti modela i treći za testiranje modela. Ovo se radi pomoću komponente *DataPartition*. Komponenta je opisana u dodatku A.

Za razvoj modela u ovom radu korišćen je odnos

trening:validation = 60:40

tj. uzorak je podeljen na populaciju za trening koja predstavlja 60% uzorka, a ostatak od 40% uzorka se koristi za proveru ispravnosti modela.

Testiranje modela biće sprovedeno pomoću celog uzorka obrađenog za T+1[23] u odnosu na uzorak za razvoj i proveru ispravnosti koji je obrađen u vremenskom trenutku T,T-1,…,T-5[24].

[23] T predstavlja vremenski trenutak u kome su pripremljeni podaci za modelovanje. T+1 predstavlja jedan mesec posle dok T-1,…T-5 predstavlaju vremenske trenutke koje označavaju jedan, dva,..,pet meseci pre vremenskog trenutka T.

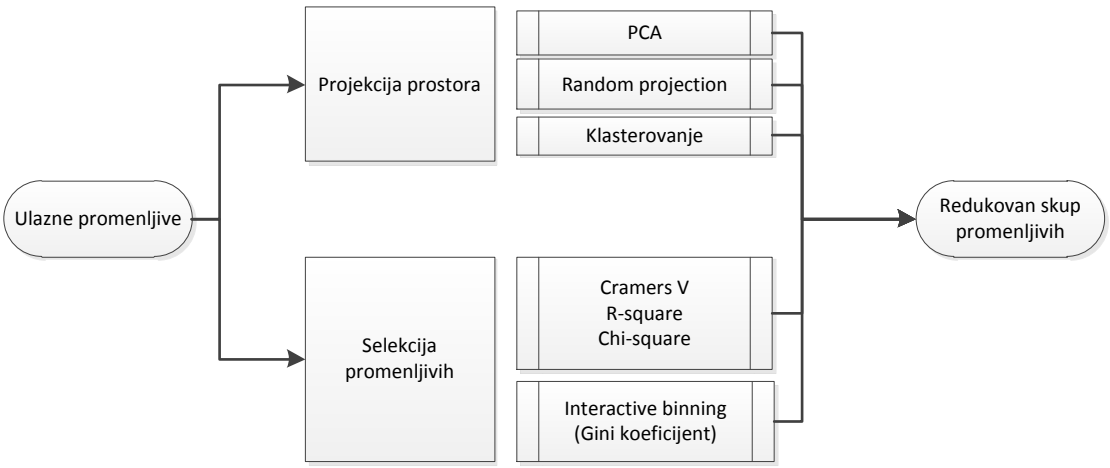
[24] Videti poglavlje 4.2 Metodologija pripremem uzorka i poglavlje 6.1 Analiza frekvencije ciljne promenljive u uzorku.

7 Redukovanje i izbor promenljivih

Redukovanje ulaznih promenljivih je jedan od ključnih zadataka u procesu razvoja modela. U ovom radu ABT sadrži preko 2000 promenljivih i pažljiv izbor ulaznih promenljivih je najteži zadatak.

Postoje dva metoda redukovanja broja ulaznih promenljivih. To su:

- izbor važnih (značajnih) promenljivih (eng. *variable selection*)
- dimenziona redukcija promenljivih



Slika 31. Redukovanje ulaznih promenljivih

Prilikom izbora važnih promenljivih obično se koristi *stepwise* regresija, korelacija i hi kvadrat test. Izborom promenjivih na ovaj način se omogućava jednostavan opis modela, jer se za opis modela koriste promenljive razumljive poslovnom korisniku.

Drugi način za redukovanje dimenzija ulaznog prostora je projektovanje ulaznog prostora koristeći metode dimenzione redukcije (*principal component analysis*, *singular value decomposition*, *random projection,…*). Ovaj pristup je mnogo jednostavniji ali s obzirom da je izlaz linearna kombinacija promenljivih veoma je teško opisati model poslovnom korisniku.

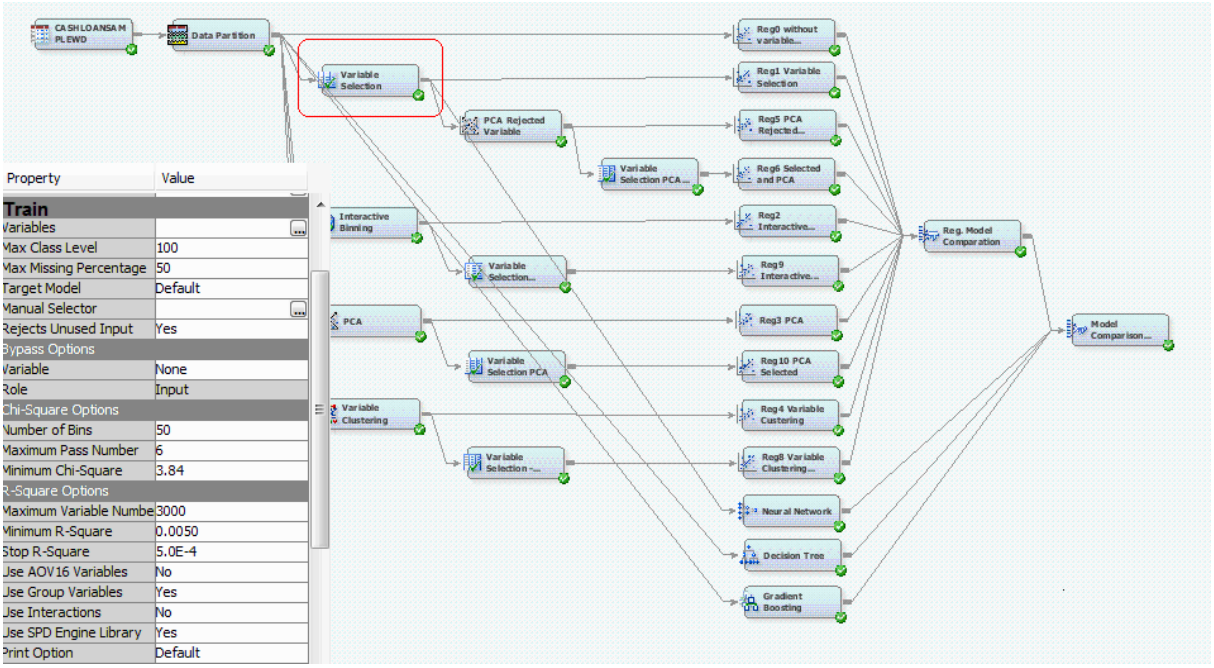
U ovom poglavlju biće prikazano redukovanje ulaznih promenljivih koristeći SAS komponente:

- *VariableSelection* (*stepwise* regresija, *chi-square* test)
- *InteractiveBinning* (Gini koeficijent)
- PCA
- VariableCustering

Takođe biće prikazano i kombinovanje ovih metoda sa ciljem dobijanja najboljeg skupa promenljivih za logističku regresiju.

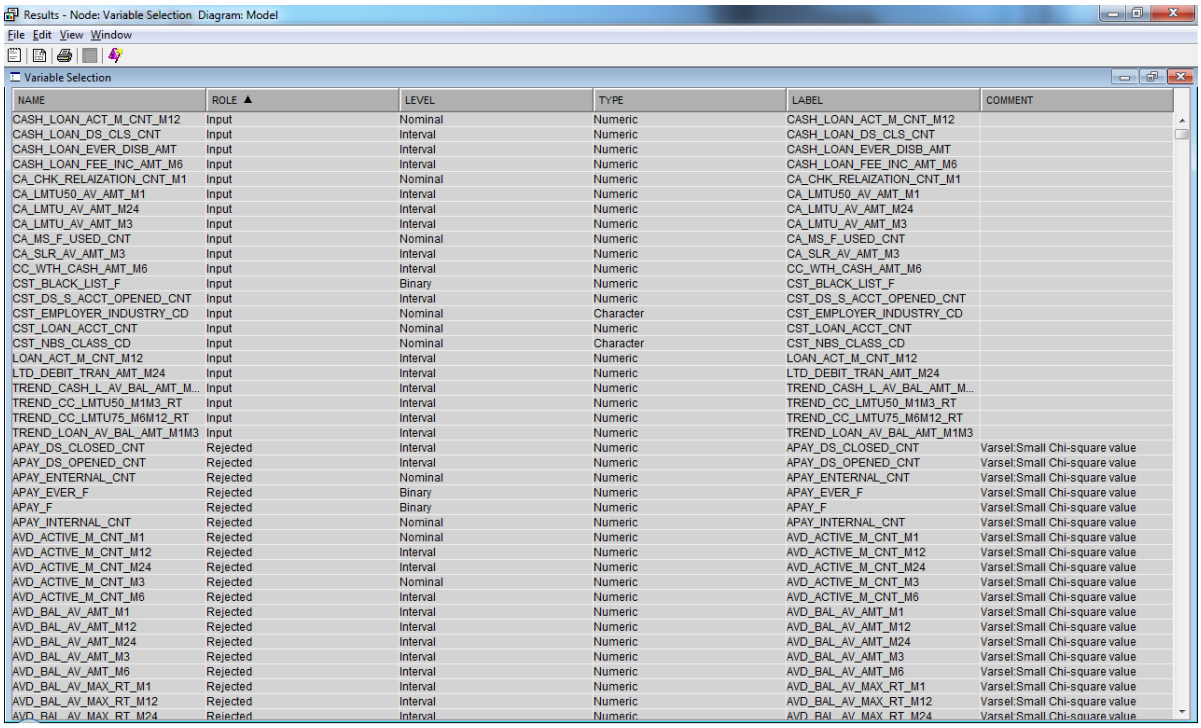
7.1 Izbor važnih promenljivih koristeći *VariableSelection* komponentu

Komponentom *Variable Selection* biraju se statistički značajne promenljive. Ova komponenta koristi metode hi-kvadrat i r-kvadrat prilikom odabira promenljih.



Slika 32. Korišćenje komponente Variable Selection.

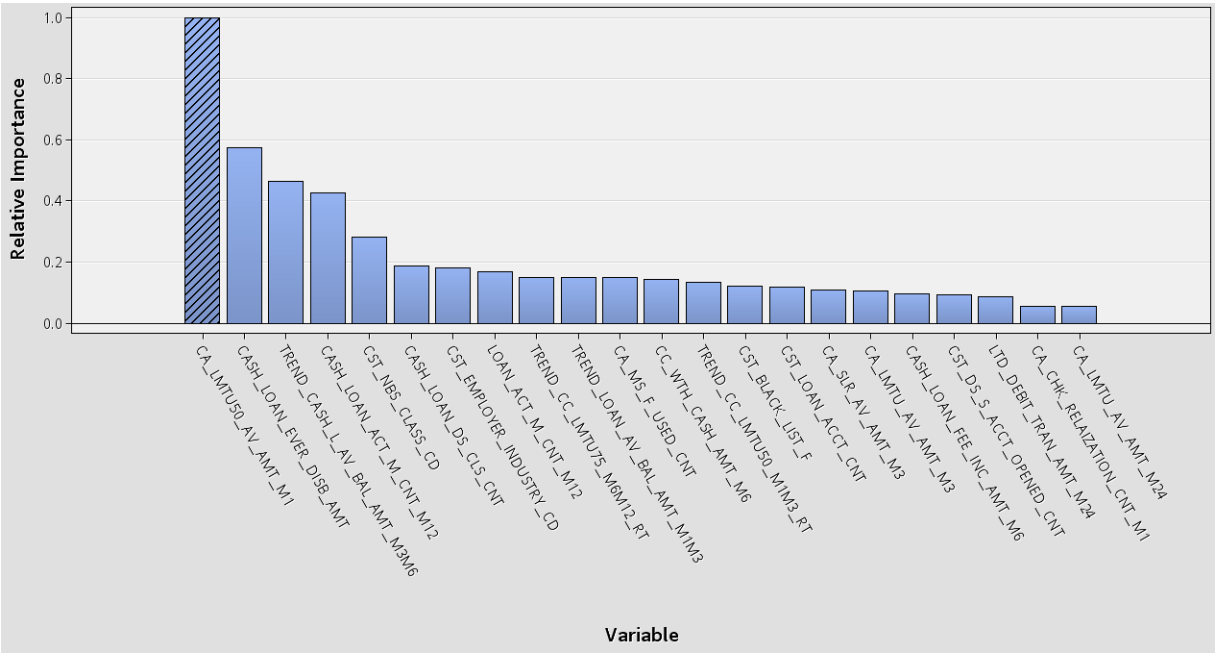
Komponenta je izabrala 22 promenljive (Slika 33) koje imaju hi-kvadrat veći od 3.84. Ostale promenljive su odbačene i neće biti korišćene u daljoj analizi.



NAME	ROLE ▲	LEVEL	TYPE	LABEL	COMMENT
CASH_LOAN_ACT_M_CNT_M12	Input	Nominal	Numeric	CASH_LOAN_ACT_M_CNT_M12	
CASH_LOAN_DS_CLS_CNT	Input	Interval	Numeric	CASH_LOAN_DS_CLS_CNT	
CASH_LOAN_EVER_DISB_AMT	Input	Interval	Numeric	CASH_LOAN_EVER_DISB_AMT	
CASH_LOAN_FEE_INC_AMT_M6	Input	Interval	Numeric	CASH_LOAN_FEE_INC_AMT_M6	
CA_CHK_RELAIZATION_CNT_M1	Input	Nominal	Numeric	CA_CHK_RELAIZATION_CNT_M1	
CA_LMTU50_AV_AMT_M1	Input	Interval	Numeric	CA_LMTU50_AV_AMT_M1	
CA_LMTU_AV_AMT_M24	Input	Interval	Numeric	CA_LMTU_AV_AMT_M24	
CA_LMTU_AV_AMT_M3	Input	Interval	Numeric	CA_LMTU_AV_AMT_M3	
CA_MS_F_USED_CNT	Input	Nominal	Numeric	CA_MS_F_USED_CNT	
CA_SLR_AV_AMT_M3	Input	Interval	Numeric	CA_SLR_AV_AMT_M3	
CC_WTH_CASH_AMT_M6	Input	Interval	Numeric	CC_WTH_CASH_AMT_M6	
CST_BLACK_LIST_F	Input	Binary	Numeric	CST_BLACK_LIST_F	
CST_DS_S_ACCT_OPENED_CNT	Input	Interval	Numeric	CST_DS_S_ACCT_OPENED_CNT	
CST_EMPLOYER_INDUSTRY_CD	Input	Nominal	Character	CST_EMPLOYER_INDUSTRY_CD	
CST_LOAN_ACCT_CNT	Input	Nominal	Numeric	CST_LOAN_ACCT_CNT	
CST_NBS_CLASS_CD	Input	Nominal	Character	CST_NBS_CLASS_CD	
LOAN_ACT_M_CNT_M12	Input	Interval	Numeric	LOAN_ACT_M_CNT_M12	
LTD_DEBIT_TRAN_AMT_M24	Input	Interval	Numeric	LTD_DEBIT_TRAN_AMT_M24	
TREND_CASH_L_AV_BAL_AMT_M...	Input	Interval	Numeric	TREND_CASH_L_AV_BAL_AMT_M...	
TREND_CC_LMTU50_M1M3_RT	Input	Interval	Numeric	TREND_CC_LMTU50_M1M3_RT	
TREND_CC_LMTU75_M6M12_RT	Input	Interval	Numeric	TREND_CC_LMTU75_M6M12_RT	
TREND_LOAN_AV_BAL_AMT_M1M3	Input	Interval	Numeric	TREND_LOAN_AV_BAL_AMT_M1M3	
APAY_DS_CLOSED_CNT	Rejected	Interval	Numeric	APAY_DS_CLOSED_CNT	Varsel:Small Chi-square value
APAY_DS_OPENED_CNT	Rejected	Interval	Numeric	APAY_DS_OPENED_CNT	Varsel:Small Chi-square value
APAY_ENTERNAL_CNT	Rejected	Nominal	Numeric	APAY_ENTERNAL_CNT	Varsel:Small Chi-square value
APAY_EVER_F	Rejected	Binary	Numeric	APAY_EVER_F	Varsel:Small Chi-square value
APAY_F	Rejected	Binary	Numeric	APAY_F	Varsel:Small Chi-square value
APAY_INTERNAL_CNT	Rejected	Nominal	Numeric	APAY_INTERNAL_CNT	Varsel:Small Chi-square value
AVD_ACTIVE_M_CNT_M1	Rejected	Nominal	Numeric	AVD_ACTIVE_M_CNT_M1	Varsel:Small Chi-square value
AVD_ACTIVE_M_CNT_M12	Rejected	Interval	Numeric	AVD_ACTIVE_M_CNT_M12	Varsel:Small Chi-square value
AVD_ACTIVE_M_CNT_M24	Rejected	Interval	Numeric	AVD_ACTIVE_M_CNT_M24	Varsel:Small Chi-square value
AVD_ACTIVE_M_CNT_M3	Rejected	Nominal	Numeric	AVD_ACTIVE_M_CNT_M3	Varsel:Small Chi-square value
AVD_ACTIVE_M_CNT_M6	Rejected	Interval	Numeric	AVD_ACTIVE_M_CNT_M6	Varsel:Small Chi-square value
AVD_BAL_AV_AMT_M1	Rejected	Interval	Numeric	AVD_BAL_AV_AMT_M1	Varsel:Small Chi-square value
AVD_BAL_AV_AMT_M12	Rejected	Interval	Numeric	AVD_BAL_AV_AMT_M12	Varsel:Small Chi-square value
AVD_BAL_AV_AMT_M24	Rejected	Interval	Numeric	AVD_BAL_AV_AMT_M24	Varsel:Small Chi-square value
AVD_BAL_AV_AMT_M3	Rejected	Interval	Numeric	AVD_BAL_AV_AMT_M3	Varsel:Small Chi-square value
AVD_BAL_AV_AMT_M6	Rejected	Interval	Numeric	AVD_BAL_AV_AMT_M6	Varsel:Small Chi-square value
AVD_BAL_AV_MAX_RT_M1	Rejected	Interval	Numeric	AVD_BAL_AV_MAX_RT_M1	Varsel:Small Chi-square value
AVD_BAL_AV_MAX_RT_M12	Rejected	Interval	Numeric	AVD_BAL_AV_MAX_RT_M12	Varsel:Small Chi-square value
AVD_BAL_AV_MAX_RT_M24	Rejected	Interval	Numeric	AVD_BAL_AV_MAX_RT_M24	Varsel:Small Chi-square value

Slika 33. Izabrane promenljive metodom chi-square

Promenljive poređane po „važnosti“ prikazane su na slici (Slika 34).



Slika 34. Promenljive poređane po važnosti

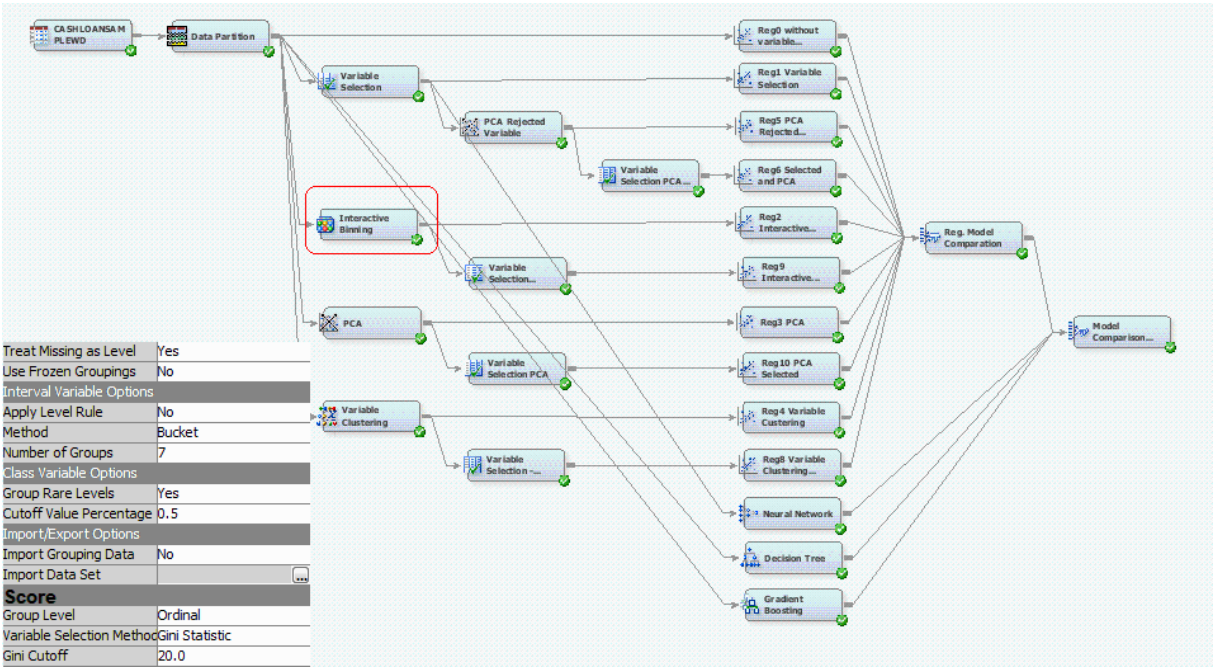
Ovom analizom redukovali smo prostor ulaznih promenljivih sa 2043 na 22. Ovo ne znači da ostale promenljive nisu dobre. Ako iz uzorka izbacimo ove 22 promenljive i dalje je moguće napraviti dobar model. Zbog toga se ne treba odreći manje važnih promenjivih. U poslednje vreme je sve više zastupljeno kombinovanje više metoda u izboru promenljivih. U poglavlju *Kombinovanje VariableSelection i PCA metoda* opisano je

kombinovanje Variable Selection i PCA. Ovo nužno ne mora da da bolji model ali vredi pokušati.

7.2 Izrada novih promenljivih komponentom *Interactive Binning*

Komponenta *Interactive Binning* grupiše vrednosti promenljive u unapred određen broj grupa. Za ovako grupisane vrednosti promenljive komponenta računa Gini koeficijent. Samo grupisane promenljive sa velikim Gini koeficijentom[25] dalje učestvuju u izradi modela.

Metod regresije je veoma osetljiv na kvalitet podataka[26]. Dovoljno je da neka prediktivna promenljiva ima pogrešene vrednosti na jednom delu uzorka i model koji bude napravljen neće opisati dobro poslovni problem. Transformacija kontinualnih promenljivih u grupe tj. u ordinarne promenljive može ublažiti problem loših podataka jer u slučaju greške manja je verovatnoća da će vrednost migrirati iz jedne u drugu grupu. Modeli napravljeni iz ovih promenljivih možda „slabije“ opisuju problem, ali dugoročno gledano model je stabilniji.



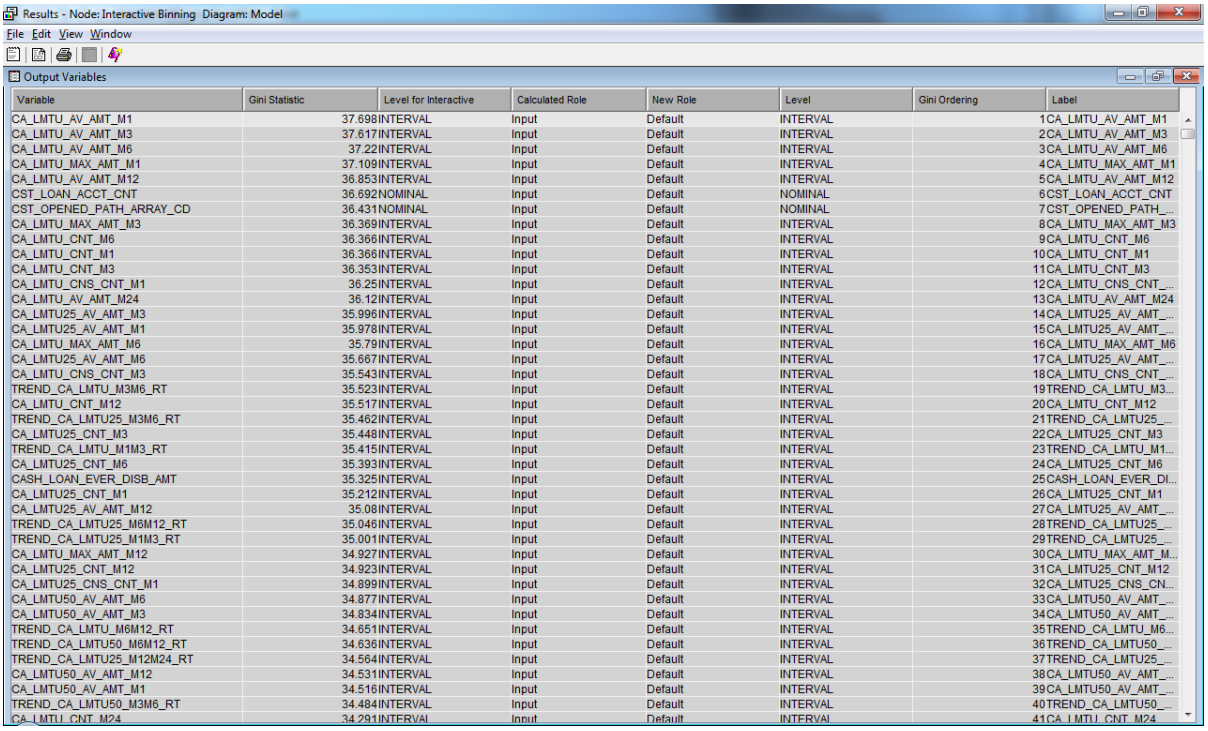
Slika 35. Komponenta Interactive binning

Vrednost Gini Cutoff=20 nam govori da će sve promenljive kod kojih je Gini koeficijent manji od 20% biti odbačene. Za promenljive kod kojih je Gini koeficijent veći od 20% izvešće se nove promenljive koje u imenu imaju prefix GRP. Na ovaj način će se svaka

[25] Ovo nam govori da broj pogodataka nije ravnomerno raspoređen u svim grupama. Gini koeficijent je opisan u Dodatku B, dok korišćenje komponente je opisano u Dodatku A.

[26] Videti poglavlje Linearna regresija u Dodatku B.

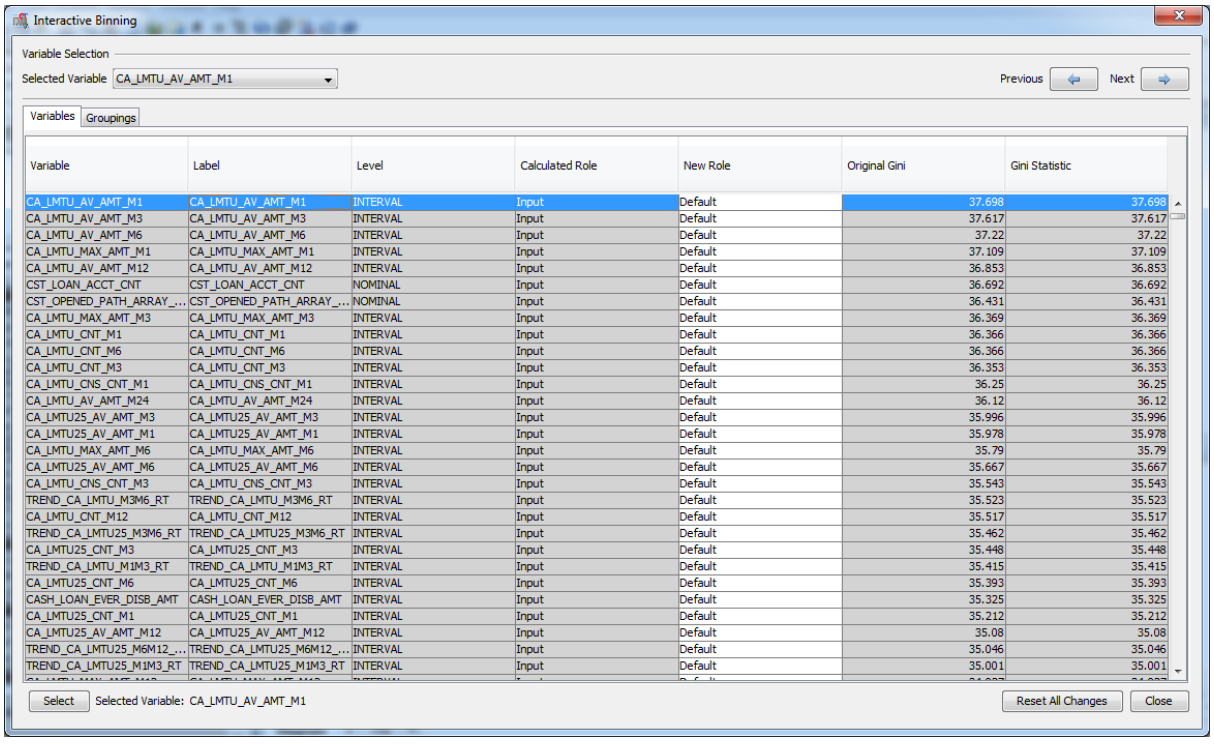
vrednost promenljive transformisati u redni broj grupe u kojoj vrednost pripada. Nad ovim vrednostima se dalje radi modelovanje.



Variable	Gini Statistic	Level for Interactive	Calculated Role	New Role	Level	Gini Ordering	Label
CA_LMTU_AV_AMT_M1	37.698	INTERVAL	Input	Default	INTERVAL	1	CA_LMTU_AV_AMT_M1
CA_LMTU_AV_AMT_M3	37.617	INTERVAL	Input	Default	INTERVAL	2	CA_LMTU_AV_AMT_M3
CA_LMTU_AV_AMT_M6	37.22	INTERVAL	Input	Default	INTERVAL	3	CA_LMTU_AV_AMT_M6
CA_LMTU_MAX_AMT_M1	37.109	INTERVAL	Input	Default	INTERVAL	4	CA_LMTU_MAX_AMT_M1
CA_LMTU_AV_AMT_M12	36.853	INTERVAL	Input	Default	INTERVAL	5	CA_LMTU_AV_AMT_M12
CST_LOAN_ACCT_CNT	36.692	NOMINAL	Input	Default	NOMINAL	6	CST_LOAN_ACCT_CNT
CST_OPENED_PATH_ARRAY_CD	36.431	NOMINAL	Input	Default	NOMINAL	7	CST_OPENED_PATH_...
CA_LMTU_MAX_AMT_M3	36.369	INTERVAL	Input	Default	INTERVAL	8	CA_LMTU_MAX_AMT_M3
CA_LMTU_CNT_M6	36.366	INTERVAL	Input	Default	INTERVAL	9	CA_LMTU_CNT_M6
CA_LMTU_CNT_M1	36.366	INTERVAL	Input	Default	INTERVAL	10	CA_LMTU_CNT_M1
CA_LMTU_CNT_M3	36.353	INTERVAL	Input	Default	INTERVAL	11	CA_LMTU_CNT_M3
CA_LMTU_CNS_CNT_M1	36.25	INTERVAL	Input	Default	INTERVAL	12	CA_LMTU_CNS_CNT_...
CA_LMTU_AV_AMT_M24	36.12	INTERVAL	Input	Default	INTERVAL	13	CA_LMTU_AV_AMT_M24
CA_LMTU25_AV_AMT_M3	35.996	INTERVAL	Input	Default	INTERVAL	14	CA_LMTU25_AV_AMT_...
CA_LMTU25_AV_AMT_M1	35.978	INTERVAL	Input	Default	INTERVAL	15	CA_LMTU25_AV_AMT_...
CA_LMTU_MAX_AMT_M6	35.79	INTERVAL	Input	Default	INTERVAL	16	CA_LMTU_MAX_AMT_M6
CA_LMTU25_AV_AMT_M6	35.667	INTERVAL	Input	Default	INTERVAL	17	CA_LMTU25_AV_AMT_...
CA_LMTU_CNS_CNT_M3	35.543	INTERVAL	Input	Default	INTERVAL	18	CA_LMTU_CNS_CNT_...
TREND_CA_LMTU_M3M6_RT	35.523	INTERVAL	Input	Default	INTERVAL	19	TREND_CA_LMTU_M3...
CA_LMTU_CNT_M12	35.517	INTERVAL	Input	Default	INTERVAL	20	CA_LMTU_CNT_M12
TREND_CA_LMTU25_M3M6_RT	35.462	INTERVAL	Input	Default	INTERVAL	21	TREND_CA_LMTU25_...
CA_LMTU25_CNT_M3	35.448	INTERVAL	Input	Default	INTERVAL	22	CA_LMTU25_CNT_M3
TREND_CA_LMTU_M1M3_RT	35.415	INTERVAL	Input	Default	INTERVAL	23	TREND_CA_LMTU_M1...
CA_LMTU25_CNT_M6	35.393	INTERVAL	Input	Default	INTERVAL	24	CA_LMTU25_CNT_M6
CASH_LOAN_EVER_DISB_AMT	35.325	INTERVAL	Input	Default	INTERVAL	25	CASH_LOAN_EVER_DI...
CA_LMTU25_CNT_M1	35.212	INTERVAL	Input	Default	INTERVAL	26	CA_LMTU25_CNT_M1
CA_LMTU25_AV_AMT_M12	35.08	INTERVAL	Input	Default	INTERVAL	27	CA_LMTU25_AV_AMT_...
TREND_CA_LMTU25_M6M12_RT	35.046	INTERVAL	Input	Default	INTERVAL	28	TREND_CA_LMTU25_...
TREND_CA_LMTU25_M1M3_RT	35.001	INTERVAL	Input	Default	INTERVAL	29	TREND_CA_LMTU25_...
CA_LMTU_MAX_AMT_M12	34.927	INTERVAL	Input	Default	INTERVAL	30	CA_LMTU_MAX_AMT_M...
CA_LMTU25_CNT_M12	34.923	INTERVAL	Input	Default	INTERVAL	31	CA_LMTU25_CNT_M12
CA_LMTU25_CNS_CNT_M1	34.899	INTERVAL	Input	Default	INTERVAL	32	CA_LMTU25_CNS_CN...
CA_LMTU50_AV_AMT_M6	34.877	INTERVAL	Input	Default	INTERVAL	33	CA_LMTU50_AV_AMT_...
CA_LMTU50_AV_AMT_M3	34.834	INTERVAL	Input	Default	INTERVAL	34	CA_LMTU50_AV_AMT_...
TREND_CA_LMTU_M6M12_RT	34.651	INTERVAL	Input	Default	INTERVAL	35	TREND_CA_LMTU_M6...
TREND_CA_LMTU50_M6M12_RT	34.636	INTERVAL	Input	Default	INTERVAL	36	TREND_CA_LMTU50_...
TREND_CA_LMTU25_M12M24_RT	34.564	INTERVAL	Input	Default	INTERVAL	37	TREND_CA_LMTU25_...
CA_LMTU50_AV_AMT_M12	34.531	INTERVAL	Input	Default	INTERVAL	38	CA_LMTU50_AV_AMT_...
CA_LMTU50_AV_AMT_M1	34.516	INTERVAL	Input	Default	INTERVAL	39	CA_LMTU50_AV_AMT_...
TREND_CA_LMTU50_M3M6_RT	34.484	INTERVAL	Input	Default	INTERVAL	40	TREND_CA_LMTU50_...
CA_LMTU_CNT_M24	34.291	INTERVAL	Input	Default	INTERVAL	41	CA_LMTU_CNT_M24

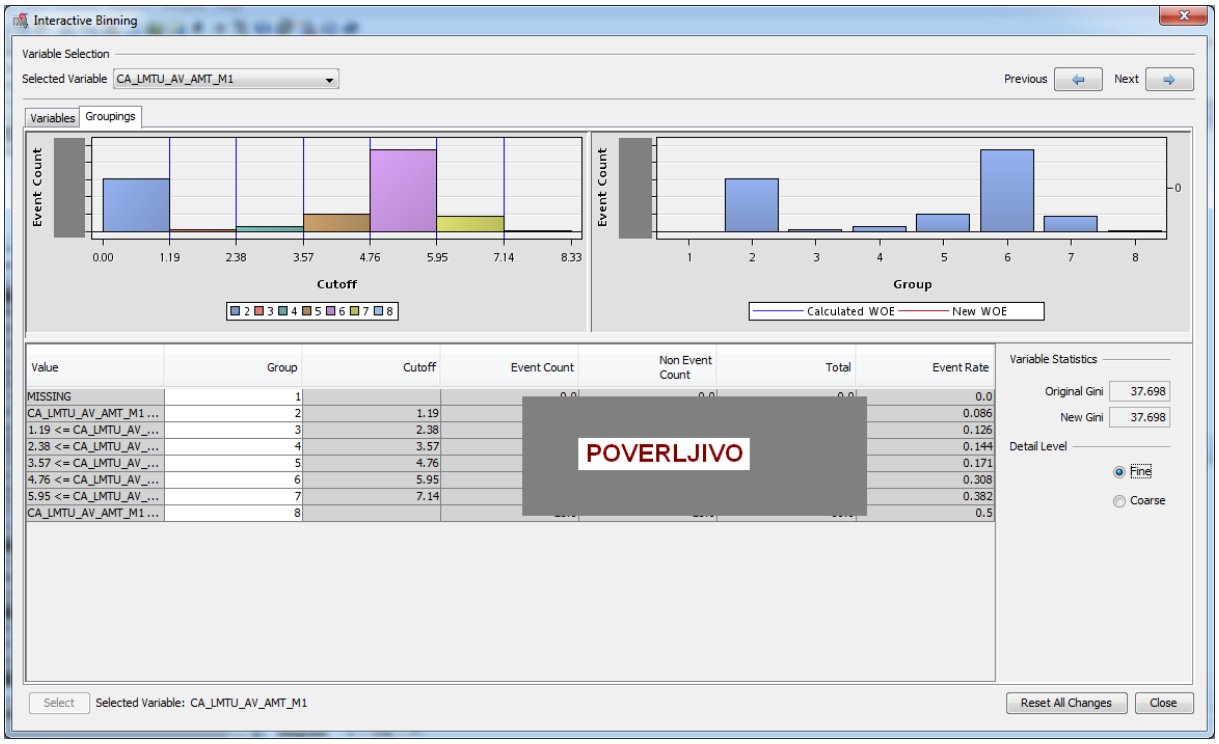
Slika 36. LIsta promenljivih koje su prošle „Gini Cuttoff" kriterijum

Komponenta sadrži aplikaciju *InteractiveBining* gde je moguće promentiti definiciju grupa za svaku promenljivu.



Variable	Label	Level	Calculated Role	New Role	Original Gini	Gini Statistic
CA_LMTU_AV_AMT_M1	CA_LMTU_AV_AMT_M1	INTERVAL	Input	Default	37.698	37.698
CA_LMTU_AV_AMT_M3	CA_LMTU_AV_AMT_M3	INTERVAL	Input	Default	37.617	37.617
CA_LMTU_AV_AMT_M6	CA_LMTU_AV_AMT_M6	INTERVAL	Input	Default	37.22	37.22
CA_LMTU_MAX_AMT_M1	CA_LMTU_MAX_AMT_M1	INTERVAL	Input	Default	37.109	37.109
CA_LMTU_AV_AMT_M12	CA_LMTU_AV_AMT_M12	INTERVAL	Input	Default	36.853	36.853
CST_LOAN_ACCT_CNT	CST_LOAN_ACCT_CNT	NOMINAL	Input	Default	36.692	36.692
CST_OPENED_PATH_ARRAY_...	CST_OPENED_PATH_ARRAY_...	NOMINAL	Input	Default	36.431	36.431
CA_LMTU_MAX_AMT_M3	CA_LMTU_MAX_AMT_M3	INTERVAL	Input	Default	36.369	36.369
CA_LMTU_CNT_M1	CA_LMTU_CNT_M1	INTERVAL	Input	Default	36.366	36.366
CA_LMTU_CNT_M6	CA_LMTU_CNT_M6	INTERVAL	Input	Default	36.366	36.366
CA_LMTU_CNT_M3	CA_LMTU_CNT_M3	INTERVAL	Input	Default	36.353	36.353
CA_LMTU_CNS_CNT_M1	CA_LMTU_CNS_CNT_M1	INTERVAL	Input	Default	36.25	36.25
CA_LMTU_AV_AMT_M24	CA_LMTU_AV_AMT_M24	INTERVAL	Input	Default	36.12	36.12
CA_LMTU25_AV_AMT_M3	CA_LMTU25_AV_AMT_M3	INTERVAL	Input	Default	35.996	35.996
CA_LMTU25_AV_AMT_M1	CA_LMTU25_AV_AMT_M1	INTERVAL	Input	Default	35.978	35.978
CA_LMTU_MAX_AMT_M6	CA_LMTU_MAX_AMT_M6	INTERVAL	Input	Default	35.79	35.79
CA_LMTU25_AV_AMT_M6	CA_LMTU25_AV_AMT_M6	INTERVAL	Input	Default	35.667	35.667
CA_LMTU_CNS_CNT_M3	CA_LMTU_CNS_CNT_M3	INTERVAL	Input	Default	35.543	35.543
TREND_CA_LMTU_M3M6_RT	TREND_CA_LMTU_M3M6_RT	INTERVAL	Input	Default	35.523	35.523
CA_LMTU_CNT_M12	CA_LMTU_CNT_M12	INTERVAL	Input	Default	35.517	35.517
TREND_CA_LMTU25_M3M6_RT	TREND_CA_LMTU25_M3M6_RT	INTERVAL	Input	Default	35.462	35.462
CA_LMTU25_CNT_M3	CA_LMTU25_CNT_M3	INTERVAL	Input	Default	35.448	35.448
TREND_CA_LMTU_M1M3_RT	TREND_CA_LMTU_M1M3_RT	INTERVAL	Input	Default	35.415	35.415
CA_LMTU25_CNT_M6	CA_LMTU25_CNT_M6	INTERVAL	Input	Default	35.393	35.393
CASH_LOAN_EVER_DISB_AMT	CASH_LOAN_EVER_DISB_AMT	INTERVAL	Input	Default	35.325	35.325
CA_LMTU25_CNT_M1	CA_LMTU25_CNT_M1	INTERVAL	Input	Default	35.212	35.212
CA_LMTU25_AV_AMT_M12	CA_LMTU25_AV_AMT_M12	INTERVAL	Input	Default	35.08	35.08
TREND_CA_LMTU25_M6M12_...	TREND_CA_LMTU25_M6M12_...	INTERVAL	Input	Default	35.046	35.046
TREND_CA_LMTU25_M1M3_RT	TREND_CA_LMTU25_M1M3_RT	INTERVAL	Input	Default	35.001	35.001

Slika 37. Aplikacija InteractiveBinning



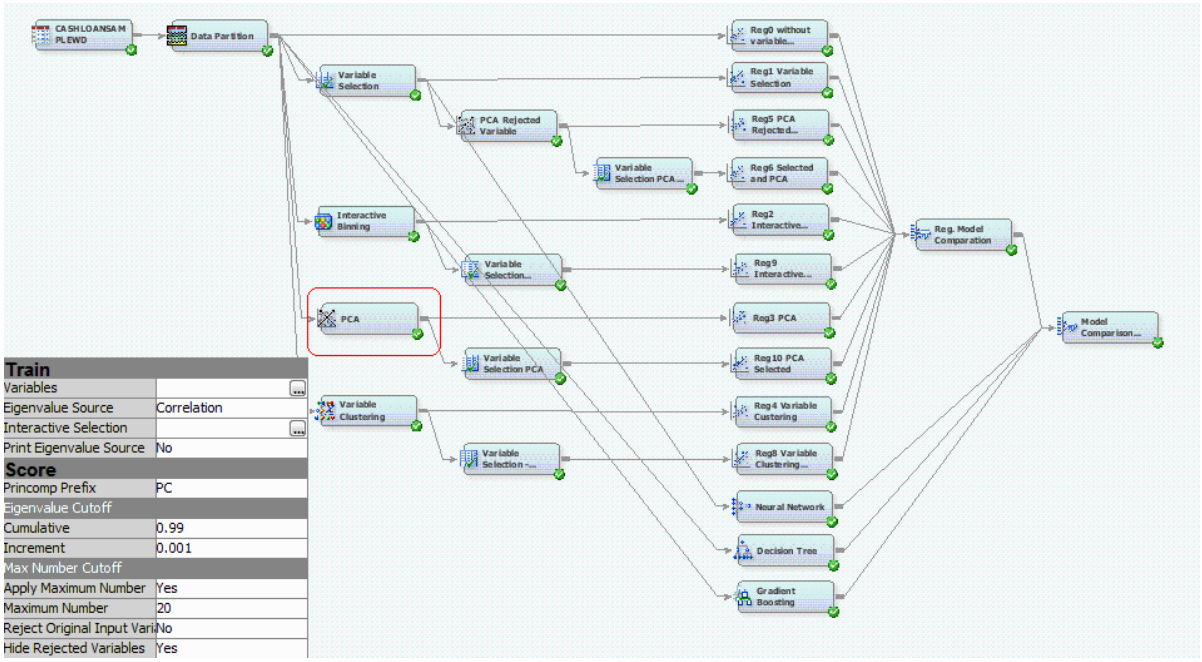
Slika 38. Vizuelizacija statistika za izabranu promenljivu u aplikaciji InteractiveBinning

Aplikacija (Slika 37 i Slika 38) nam omogućava da ručno promenimo intervale grupa. Na ovaj način možemo povećati ili umanjiti Gini koeficijent. Povećanjem Gini koeficijenta nova promenljiva postaje prediktivnija. Teoretski moguće je napraviti algoritam koji će podeliti domen promenljive na N grupa tako da Gini koeficijent bude najveći. Ovako definisane grupe se dalje mogu koristiti u procesu modelovanja. Postoji mogućnost učitavanja grupa (direktno menjanje metapodata SAS EM projekta) u komponentu i tako izbeći podelu na fiksirani broj grupa odnosno na kvantile (podrazumevana podela).

Komponenta *Interactive Binning* je opisana u dodatku A, dok je Gini koeficijent opisan u dodatku B.

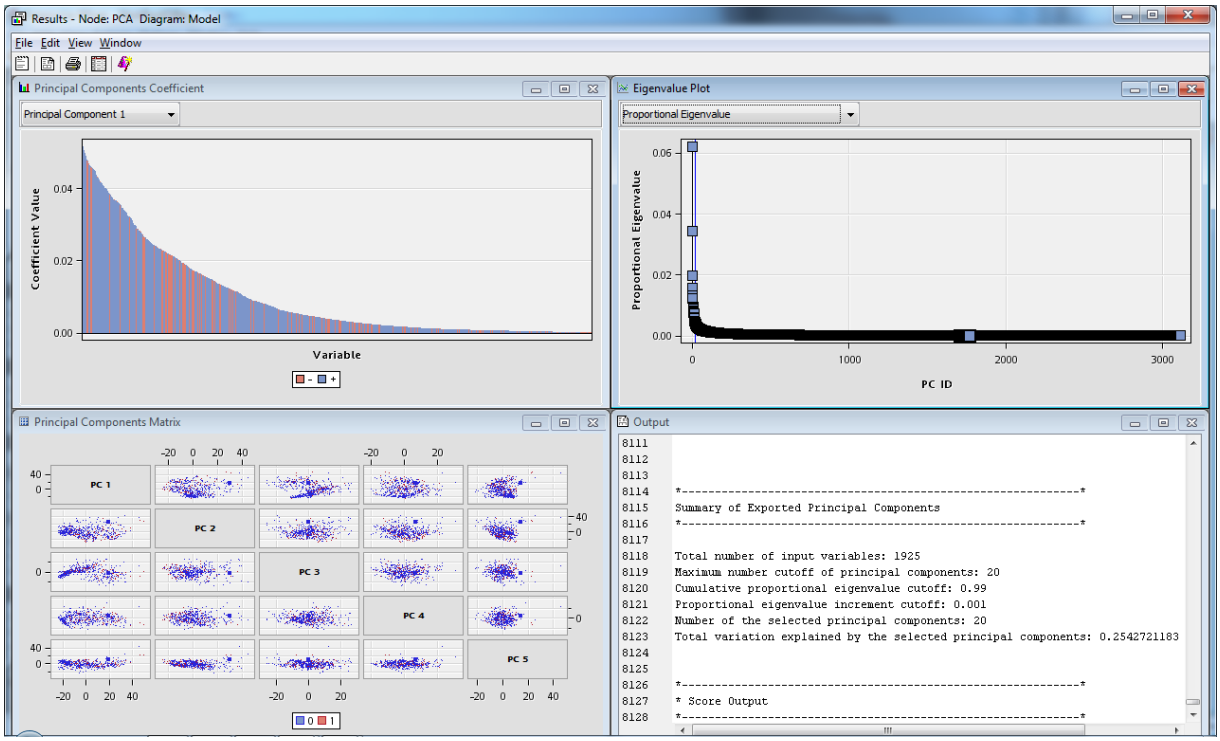
7.3 Projektovanje ulaznog prostora metodom *PCA*

Analiza glavnih komponenti (*Principal Component Analysis* – PCA) je metoda projektovanja ulaznog prostora zasnovanog nad promenljivima ($X_1, X_2,...,X_p$) (u ovom radu 2043 vektora) u podprostor koji obrazuju sopstveni vektori (Z_j), $j \leq p$, gde je p broj promenljivih.



Slika 39. Komponenta PCA u projektu

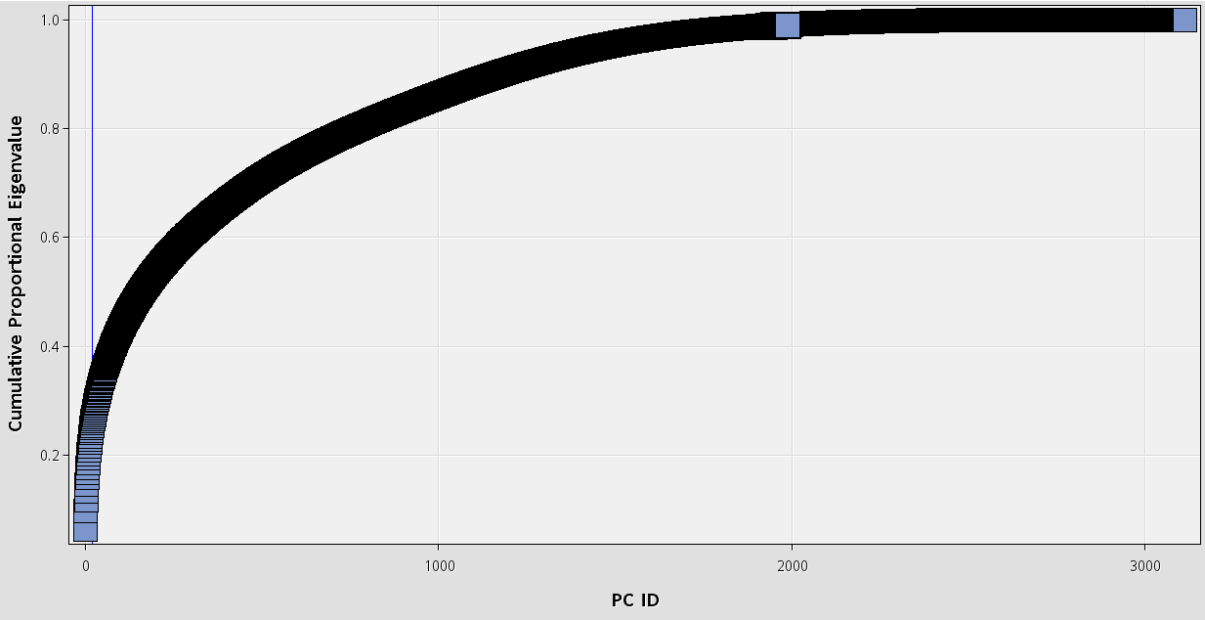
Kada se radi analiza glavnih komponenti želja je da varijanse većine novih promenljivih Z budu toliko male da su zanemarljive. U tom slučaju, veći deo varijacija originalnih podataka se može adekvatno opisati sa svega nekoliko glavnih komponenti.



Slika 40. Rezultat PCA analize

Na slici 40 opisan je rezultat PCA analize. Ideja je da se sa što manje sopstvenih vektora opiše ulazni prostor. U ovom slučaju prvih 20 sopstvenih vektora ima kumulativnu

sopstvenu vrednost od 0.25 (Slika 40) od ukupne (*Cumulative Proportional Eigenvalue=1*). Na slici (Slika 41) vidimo da sa 1000 sopstvenih vektora možemo dobiti *Cumulative Proportional Eigenvalue=0.85* dok sa 200 dobijamo *Cumulative Proportional Eigenvalue=0.60* što je možda najoptimalnije na ovom uzorku.



Slika 41. Cumulative Proportional Eigenvalue na uzorku za razvoj

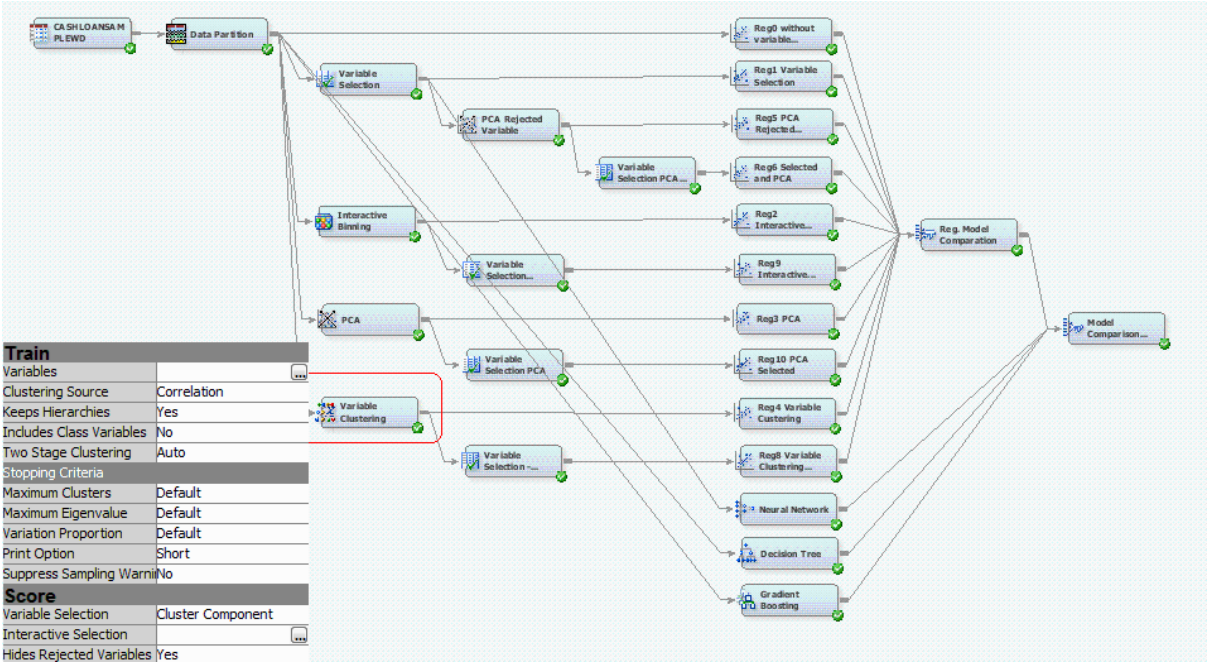
Nije nužno da sopstveni vektori sortirani po sopstvenim vrednostima budu ovim redosledom i najprediktivniji. Nad prvih 200 soptvenih vektora mogu se dalje primeniti druge tehnike redukcije i izbora promenljivih kako bi se došlo da najboljih promenljivih koje će biti ulaz za algoritam logističke regresije.

Komponenta PCA je opisana u dodatku A poglavlje *Komponenta Principal Component* matematičke osnove se nalaze u Dodatku B poglavlje *Analiza glavnih komponenti*.

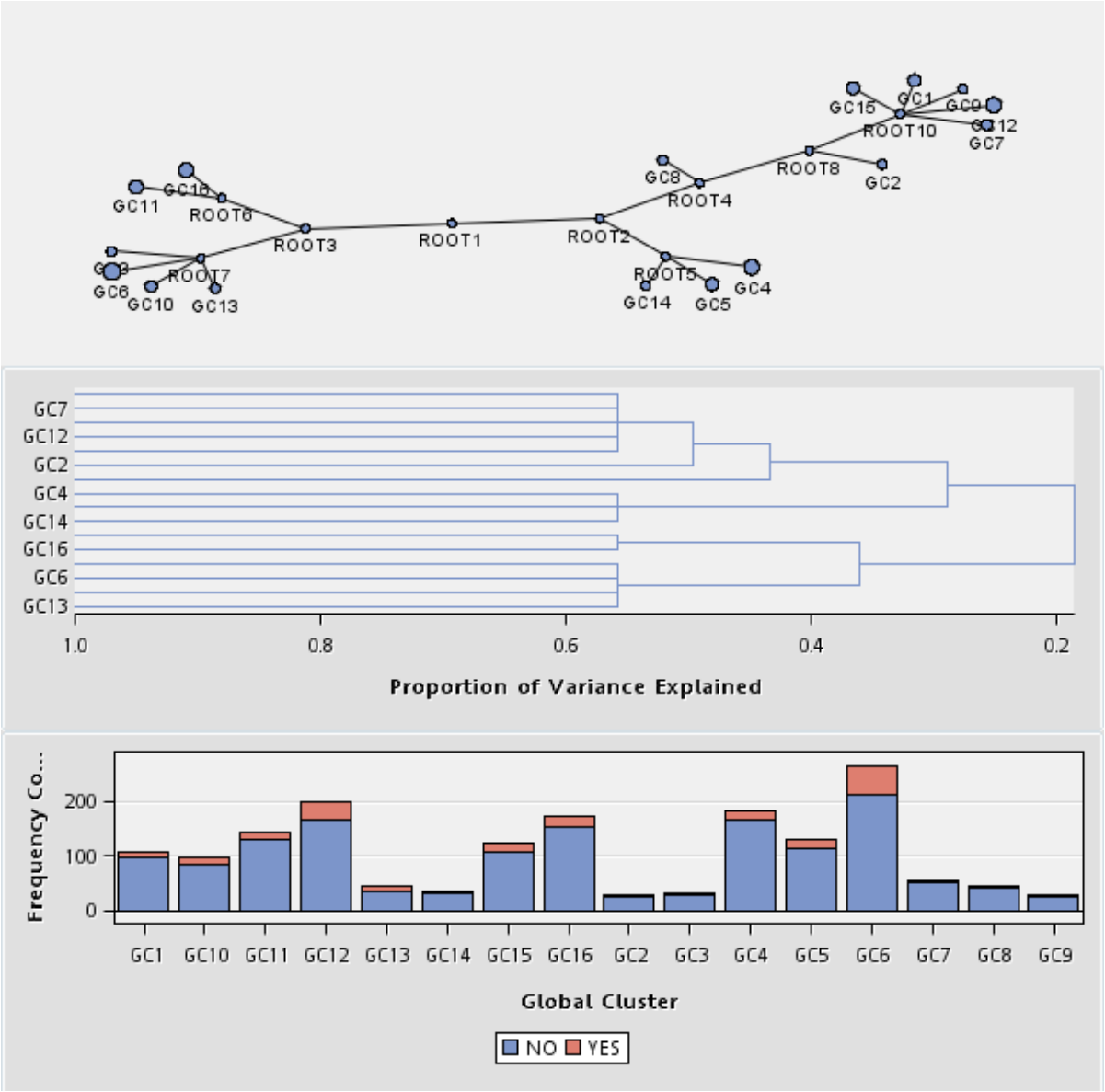
7.4 Grupisanje promenljivih pomoću komponente *Variable Clustering*

Klasterovanje promenljivih umesto nekoliko desetina promenljivih može značajno redukovati broj promenljivih za dalje analize. Klasteri nam obezbeđuju hetoregenost samih promenljivih što može biti značajno u daljim analizama.

Variable clustering (Slika 42) raspoređuje numeričke promenljive u nespojive i/ili hijerarhijske klastere. Rezultat klasteringa može se opisati kao linearna kombinacija promenljivih.



Slika 42. Primena Variable Clustering komponente u projektu.



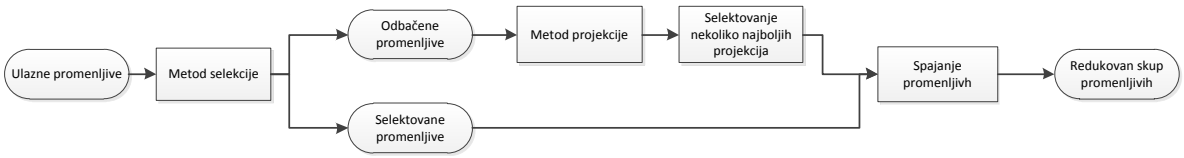
Slika 43. Klasteri koji opisuje grupe promenljivih

Algoritam klasterovanja je opisan u dodatku A poglavlje Upoznavanje sa podacima, istraživanje podataka - *Explore*.

7.5 Kombinovanje *VariableSelection* i *PCA* metoda

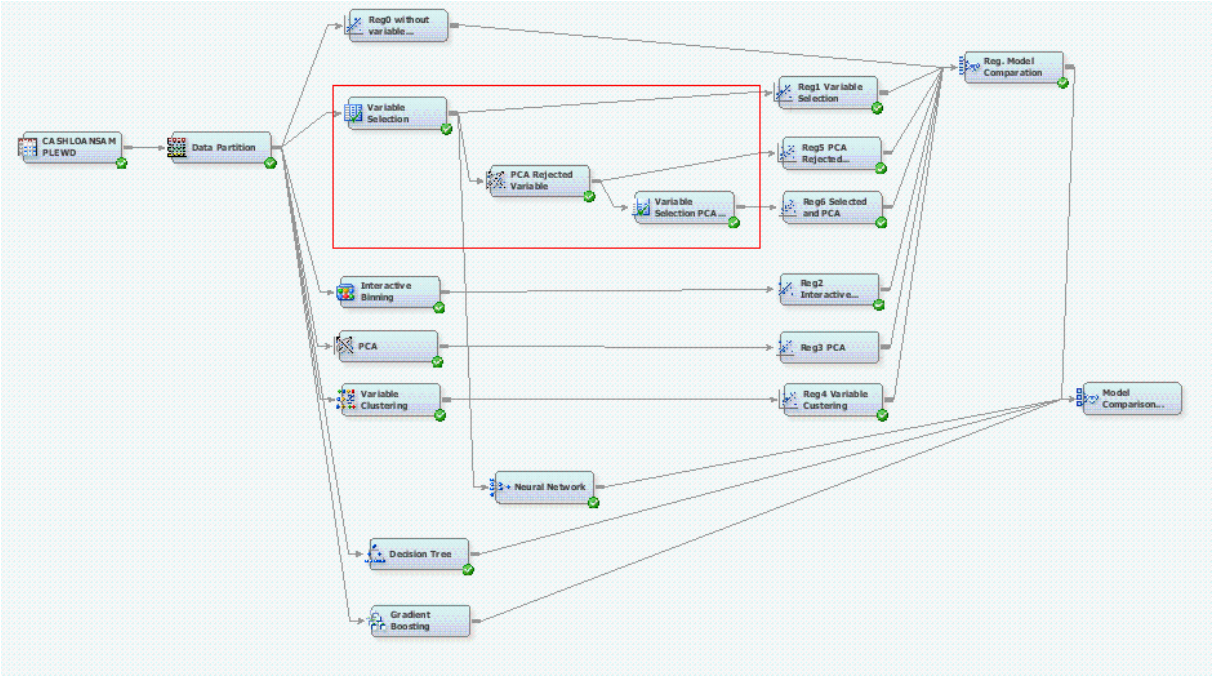
Metodom odabira promenljivih uzimamo samo nekoliko desetina promenljivih dok ostale promenljive dalje ne analiziramo jer predpostavljamo da one nisu statistički značajne. Praksa je pokazala da posle izbacivanja ovih „najprediktivnih“ promenljivih možemo napraviti isto tako dobar model kao što smo napravili od najprediktivnih promenljivih. Postavlja se pitanje da li možemo uvećati prediktivnost ako od odbačenih promenljivih izdvojimo „najprediktivnije“ i dodamo na već izabrane promenljive.

Da bi ovo ostvarili potrebno je kombinovati metode redukcije i metode izbora. Na sledećoj slici (Slika 44) prikazan algoritam kombinovanja.



Slika 44. Kombinovanje metoda selekciji i PCA

Prvo se uradi izbor pomoću r–kvardrat ili hi–kvadrat metode. Od odbačenih promenljivih napravimo redukciju prostora metodom PCA, a zatim izaberemo nekoliko najprediktivnijih glavnih komonenti koje dodamo na već izabrane promenljive (Slika 45).



Slika 45. Kombinovanje metoda selekcije i metoda redukcije promenljivih

U ovom radu je kombinovano *VariableSelection* i *PCA*. U praksi najbolji rezultati se dobijaju kombinovanjem *VariableSelection* i *RandomProjection*[27]. *RandomProjection* je SAS komponenta koja se zasebno kupuje i ovo je razlog zbog čega ona nije korišćena u radu.

Iako statistike pokazuju da ovaj pristup daje bolje rezultate nego korišećenje samo *VariableSelection* komponente ovo i dalje ne znači da je dobijeni skup promenljivih najbolji za ocenu (predikciju). Zavisno od uzorka može se desiti da neka treća metoda

[27] *Predictive Models Based on Reduced Input Space That Uses Rejected Variables - Taiyeong Lee, David Duling, and Dominique Latour , SAS Institute Inc., Cary, NC, 2009 (Paper 111-2009)*

izbora ili redukcije da bolje rezultate[28]. U svakom slučaju vredi pokušati sa kombinovanjem metoda jer utrošeno vreme na pripremu promenljivih na ovakav način nije veliko.

[28] U ovom radu model napraljen nad ovako izabranim promenljivama nije „pobedio“ (Slika 60. Izbor šampion modela)

8 Razvoj modela

U poglavlju 7 su prikazane tehnike redukovanja i izbora promenljivih. Ove tehnike izbora i redukcije promenljivih se mogu međusobno kombinovati tako da kao rezultat dobijemo nekoliko novih skupova promenljivih. Nad svakim skupom promenljivih treba naći najbolju matematičku funkciju koja opisuje podatke.

U ovom poglavlju biće opisana izrada modela na osnovu izabranih promenljivih metodom logističke regresije[29]. Osim metode logističke regresije kao kontrolne metode izabrane su i neuronske mreže, drveta odlučivanja i *gradient boosting*. Ovo je opisano u popoglavlju *Izrada modela.*

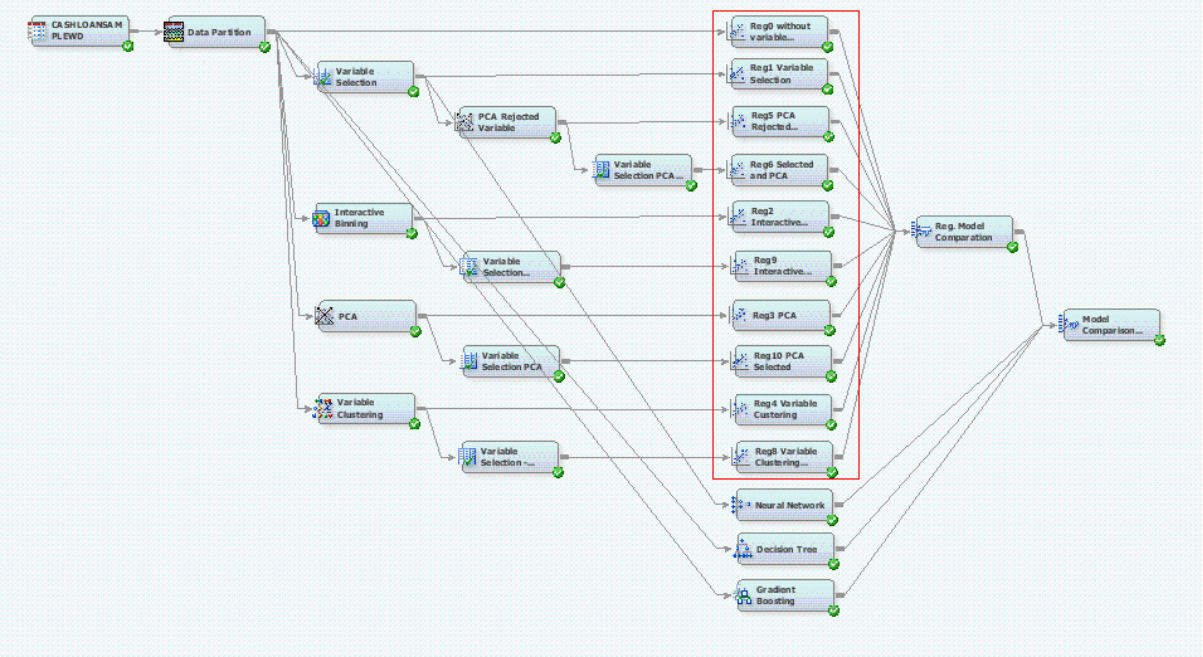
Nakon izrade svih matematičkih modela neophodno je izabrati najbolji metod, tj. metod kod kojeg je greška najmanja nad uzorkom odvojenim za proveru. Ovo je opisano u podpoglavlju *Ocena modela*.

8.1 Izrada modela

Napravljeno je 10 regresionih funkcija (Slika 46) nad sledećim skupovima izabranih promenljivih:

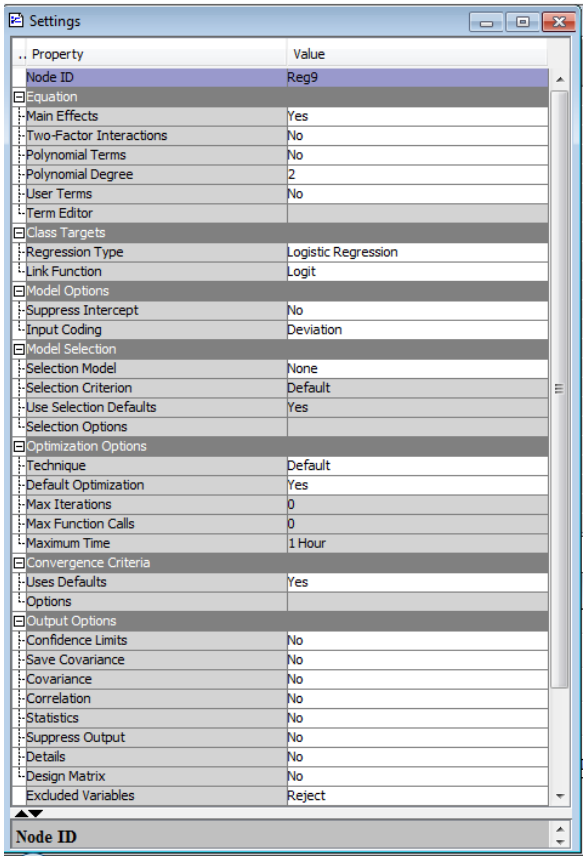
1. Bez izbora promenljivih (2043 promenljive).
2. Metodama hi-kvadrat i r-kvadrat.
3. Metodama hi-kvadrat i r-kvadrat, a potom nad odbačenim promenljivama primenjen je metod PCA. Sve PCA komponente su dodate predhodno izabranim promenljivima.
4. Metodama hi-kvadrat i r-kvadrat, a potom nad odbačenim promenljivama primenjen je metod PCA. Na PCA komponentama primenjen je metod hi-kvadrat i r-kvadrat. Samo najprediktivnije PCA komponente su dodate izabranim promenljivama.
5. Metodom *interactive grouping* sa Gini *cutoff* koeficijentom od 20%.
6. Metodom *interactive grouping* sa Gini *cutoff* koeficijentom od 20% pri čemu se biraju najbolje promenljive metodama hi-kvadrat i r-kvadrat.
7. Metodom PCA pri čemu se unapred bira maksimalan broj PCA komponenti.
8. Metodom PCA pri čemu se unapred bira maksimalan broj PCA komponenti, a zatim se biraju najbolje PCA komponente metodama hi-kvadrat i r-kvadrat
9. Metodom klasterovanja promenljviih
10. Metodom klasterovanja promenljivih pri čemu se biraju klasteri koji imaju najveći hi-kvadrat i r-kvadrat.

[29] Matematičke osnove linearne i logističke regresije opisane su u poglavljima u Dodatku B.



Slika 46. Izrada modela na osnovu izabranih promenljivih.

Komponente logističke regresije (Slika 46) imaju istu konfiguraciju kao na slici (Slika 47).



Slika 47. Podešavanja svih modela zasnovanih na regresiji

Moguće je finije podešavati svaki od ovih algoritama logističke regresije. Ovo se ne radi u ovom koraku jer je važno da prvo izaberemo jedan od 10 skupova promenljivih. Posle

izbora skupa promenljvih moguće je fino podešavati algoritam tako da dobijena regresiona funkcija bolje opisuje podatke. Moguće je napraviti algoritam koji će napraviti regresione funkcije od svih podskupova izabranog skupa promenljivih, a onda izabrati najbolju regresionu funkciju.

8.2 Ocena modela

8.2.1 Rezultat regresione analize

Za svaki od 10 modela moguće je videti mere koje ocenjuju model. Ove mere su izračunate na uzorku za trening, proveru ispravnosti i test. Mere možemo svrstati u dve kategorije. To su:

- Mere za procenu modela kroz prizmu poslovnog dobitka ili gubitka (eng. *lift measure)*
    - *Lift, Cumulative Lift, % Response, Cumulative % Response, Cumulative Captured % Response*
- Mere za procenu stabilnosti modela (eng. *model fit statistics*)
    - *Average Squared Error, Mean Squared Error, Root Average Sum of Squres,…*

Prilikom provere modela prvo koristimo *Model Fit* mere (Slika 48) kao što su *Average Squared Error, Mean Squared Erorr, Root Average Sum of Squared Error*. Regresiona funkcija je napravljena nad trening uzorkom. Ako *model fit* mere imaju „približno iste“ vrednosti nad trening i uzorkom za proveru možemo kazati da regresiona funkcija dobro opisuje uzorak.

Fit Statistics

Fit Statistics	Statistics Label	Train	Validation	Target	Target Label ▲	Test
AIC	Akaike's Information Criterion	45775.92		.Target_A_...	Target_A_CASH...	
ASE	Average Squared Error	0.114247	0.114107	Target_A_...	Target_A_CASH...	
AVERR	Average Error Function	0.370318	0.370972	Target_A_...	Target_A_CASH...	
DFE	Degrees of Freedom for Error	61203		.Target_A_...	Target_A_CASH...	
DFM	Model Degrees of Freedom	163		.Target_A_...	Target_A_CASH...	
DFT	Total Degrees of Freedom	61366		.Target_A_...	Target_A_CASH...	
DIV	Divisor for ASE	122732	81826	Target_A_...	Target_A_CASH...	
ERR	Error Function	45449.92	30355.19	Target_A_...	Target_A_CASH...	
FPE	Final Prediction Error	0.114855		.Target_A_...	Target_A_CASH...	
MAX	Maximum Absolute Error	0.982241	0.982086	Target_A_...	Target_A_CASH...	
MSE	Mean Square Error	0.114551	0.114107	Target_A_...	Target_A_CASH...	
NOBS	Sum of Frequencies	61366	40913	Target_A_...	Target_A_CASH...	
_NW_	Number of Estimate Weights	163		.Target_A_...	Target_A_CASH...	
RASE	Root Average Sum of Squares	0.338004	0.337797	Target_A_...	Target_A_CASH...	
RFPE	Root Final Prediction Error	0.338903		.Target_A_...	Target_A_CASH...	
RMSE	Root Mean Squared Error	0.338454	0.337797	Target_A_...	Target_A_CASH...	
SBC	Schwarz's Bayesian Criterion	47246.93		.Target_A_...	Target_A_CASH...	
SSE	Sum of Squared Errors	14021.72	9336.9	Target_A_...	Target_A_CASH...	
SUMW	Sum of Case Weights Times Freq	122732	81826	Target_A_...	Target_A_CASH...	
MISC	Misclassification Rate	0.153359	0.15313	Target_A_...	Target_A_CASH...	

Slika 48. Model fit statistike

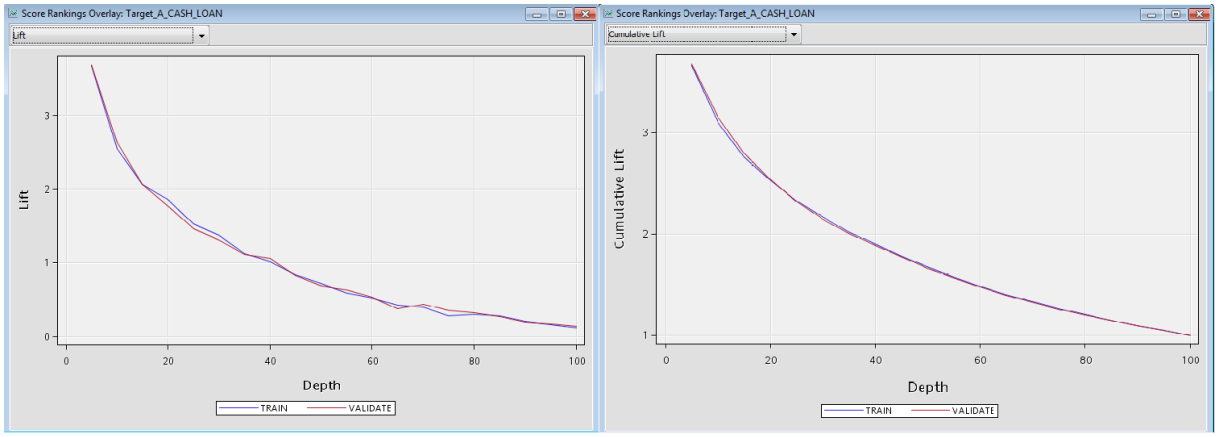
Da li model pogađa dovoljno dobro je pitanje na koje *Model Fit* mere ne mogu da daju odgovor.

Lift mere opisuju iznos profita primenom modela u odnosu na ***nepostojanje modela***[30]. Za sve *lift* mere naophodno je napraviti sledeću pripremu:

1. Izračunati verovatnoću za sve opservacije, a zatim sortirati rezultat po izračunatoj verovatnoći u opadajućem poretku.
2. Podeliti uzorak na *n* jednakih grupa (u ovom radu to je 20) tako da grupa koja ima veći redni broj ima manju verovatnoću.
3. Za svaku grupu izračunati broj *event* i *nonevent* opservacija.

Nad ovako kreiranim grupama se računaju mere.

Na slici (Slika 49) prikazane su *lift* i *cumulative lift* krive. X osa predstavlja veličinu sortiranog uzorka, dok je Y osa *lift* odnosno *cumulative lift*. Leva slika nam govori da grupa čija je verovatnoća između (P75, P80][31] (označeno na slici kao *Depth*=20) daje 1,7 puta bolje rezultate od *nepostojanja modela*, dok za prvih 5% klijenata model daje 3,6 puta bolje rezultate od *nepostojanja modela*.

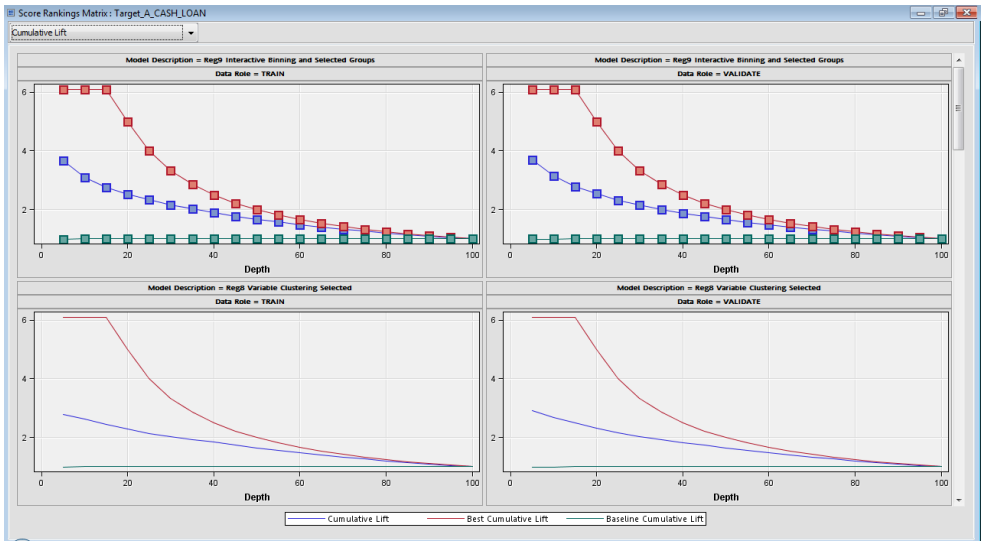


Slika 49. Lift i kumulativni lift modela

Na desnom delu slike (Slika 49) prikazan je kumulativni lift koji nam govori da prvih 40% uzorka sortiranih po verovatnoći u opadajućem poretku daje 2 puta bolje rezultate od *nepostojanja modela*.

[30] U matematičkom smislu, nepostojanje modela podrazumeva da se iz trening i uzorka za proveru uzme n disjunktnih slučajnih uzoraka i njima ponudi proizvod za koji se radi model. Svi uzorci bi trebalo da imaju istu verovatnoću kupovine i ona bi trebalo da bude približna event rate=*event/(event+nonevent)* u trening uzorku odnosno uzorku za proveru. U poslovnom smislu, nepostojanje modela označava odsustvo bilo koje smislene akcije izbora ciljne grupe klijenata kome će proizvod biti ponuđen. Proizvod se u ovom slučaju nudi svima ili slučajno izabranoj grupi.

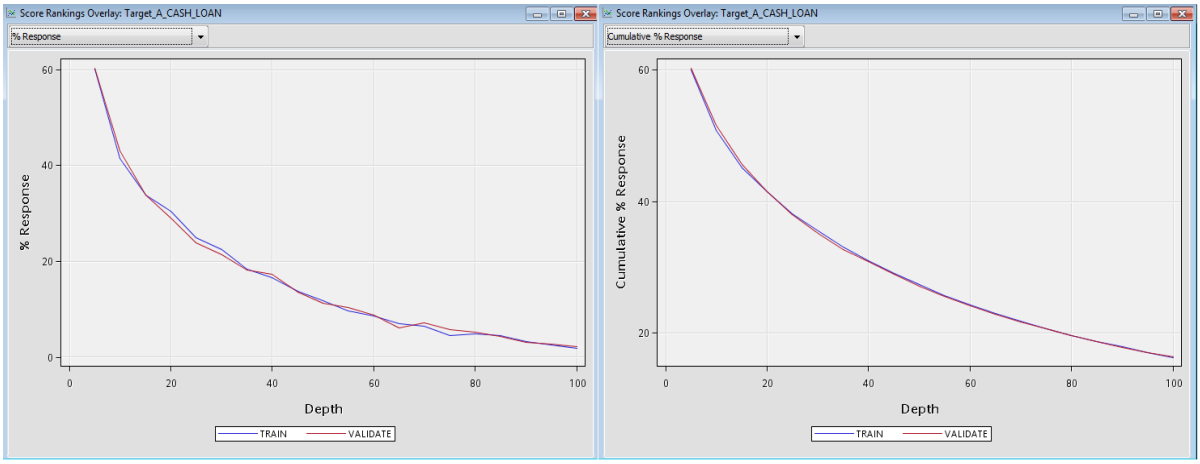
[31] (P75, P80] sve opservacije čija verovatnoća se nalazi između 75 i 80 percentila.



Slika 50. Score Ranking Matrix

Osim sa kumulativnim liftom *nepostojanja modela* (eng. *base line cumulative lift* tj. Y=1) se poredi i sa tzv. krivom kumulativnog lifta najboljeg modela (eng. *best cumulative lift*)[32]. Na slici (Slika 50) prikazano je poređenje ove tri krive.

Mera *lift* određuje koliko je model bolji od *nepostojanja modela* i lošiji od *najboljeg modela*, ali nam ne govori koliko „pogodaka" možemo da očekujemo u prvih n% sortirane populacije. Mere *% response* i *cumulative % response* nam to pokazuju. Na slici (Slika 51) levo grupa čija je verovatnoća između (P75, P80] ima 25% pogodaka (što je 1,7[33] puta više od „event rate"), dok na desnoj slici prvih 20% sortiranih opservacija ima 45% pogotaka (što je 2,7 puta više od „event rate").

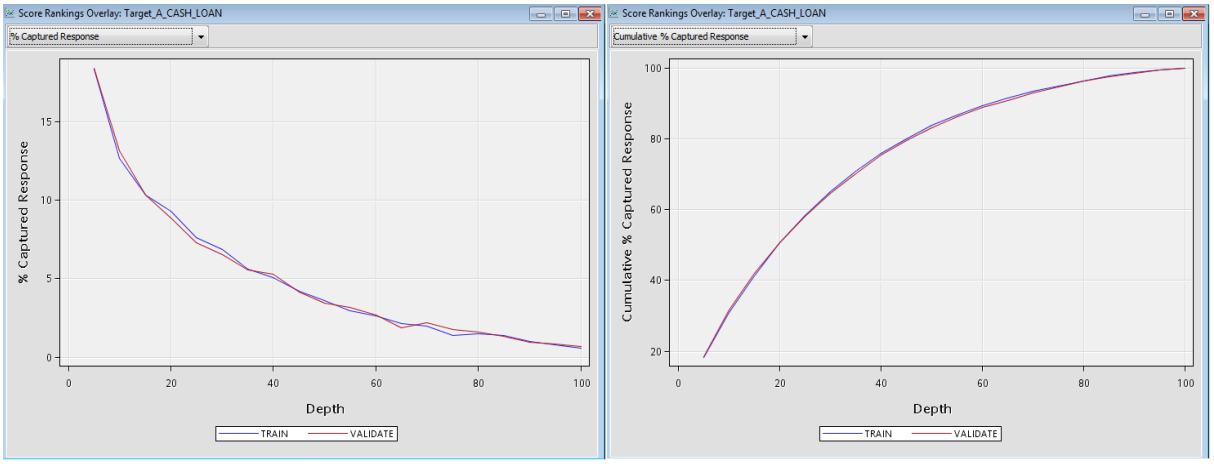


Slika 51. %reposne i cumulative % response kriva

[32] Najbolji model pogađa sa verovatnoćom od 100% tj. u ovako sortiranom uzoruku očekuje se da svi „pogoci" budu u prvih „*event rate*" procenata sortiranog uzorka. Kriva kumulativnog lifta najboljeg modela se dobije tako što se sve *event* opservacije stave u najbolje grupe (počev od 1,2,3,..). S obzirom da je *event rate* na testnom i uzorku za proveru 16% to će sve *event* opservacije biti raspoređene u tri najbolje grupe i jedan manji deo u četvrtoj (ukupno imamo 20 grupa).

[33] Videti lift i kumulativni lift

Mere *%Captured Response* i *Cumulative % Capture Response* računaju koliko ima pogodaka u ukupnom broju pogodaka. Na slici (Slika 52) među prvih 20% sortiranih opservacija ima oko 45% pogodaka od ukupnog broja pogodaka, dok na prvih 50% sortiranih opservacija ima 85% pogodaka od ukupnog broja pogodaka.

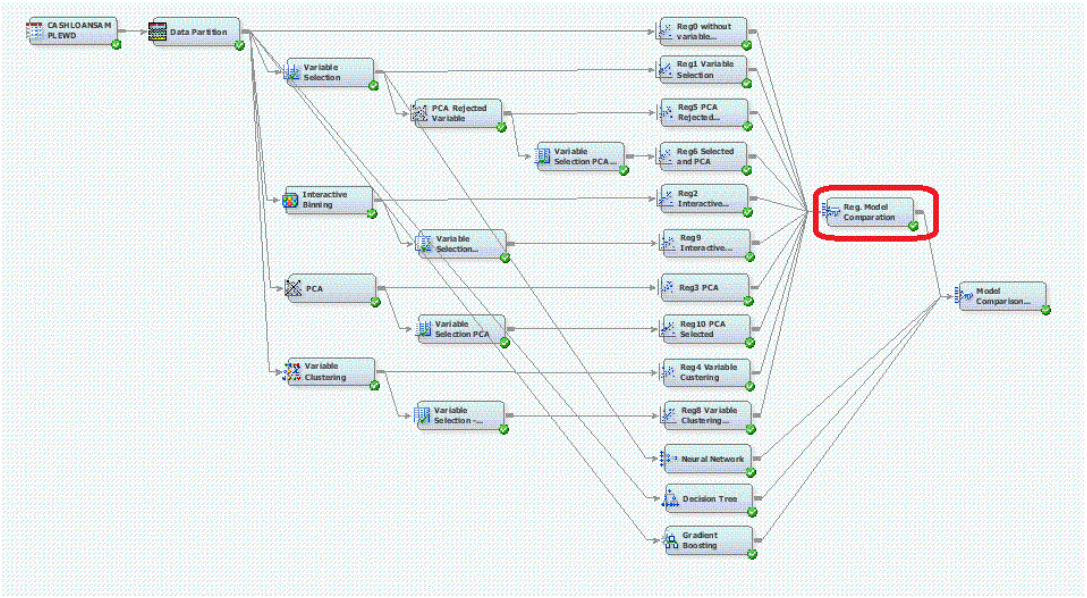


Slika 52. %captured response i cumulative % captured response

Regresioni model je napravljen nad trening uzorkom. Potrebno je proveriti funkciju modela nad uzorkom za proveru i testnim uzorkom. Sve gore navedene mere moguće je porediti nad sva tri uzorka. Svako odstupanje između grafikona nad ova tri uzorka može da bude znak da model neće biti stabilan u produkciji.

8.2.2 Izbor najboljeg modela

Prilikom izrade modela korišćeno je 10 istih regresionih algoritama nad različitim skupovima promenljivih. Sada je neophodno izabrati koji od ovih modela najbolje opisuje uzorak. Izbor modela se radi koristeći SAS komponentu „*Model Comparation*“ (Slika 53).

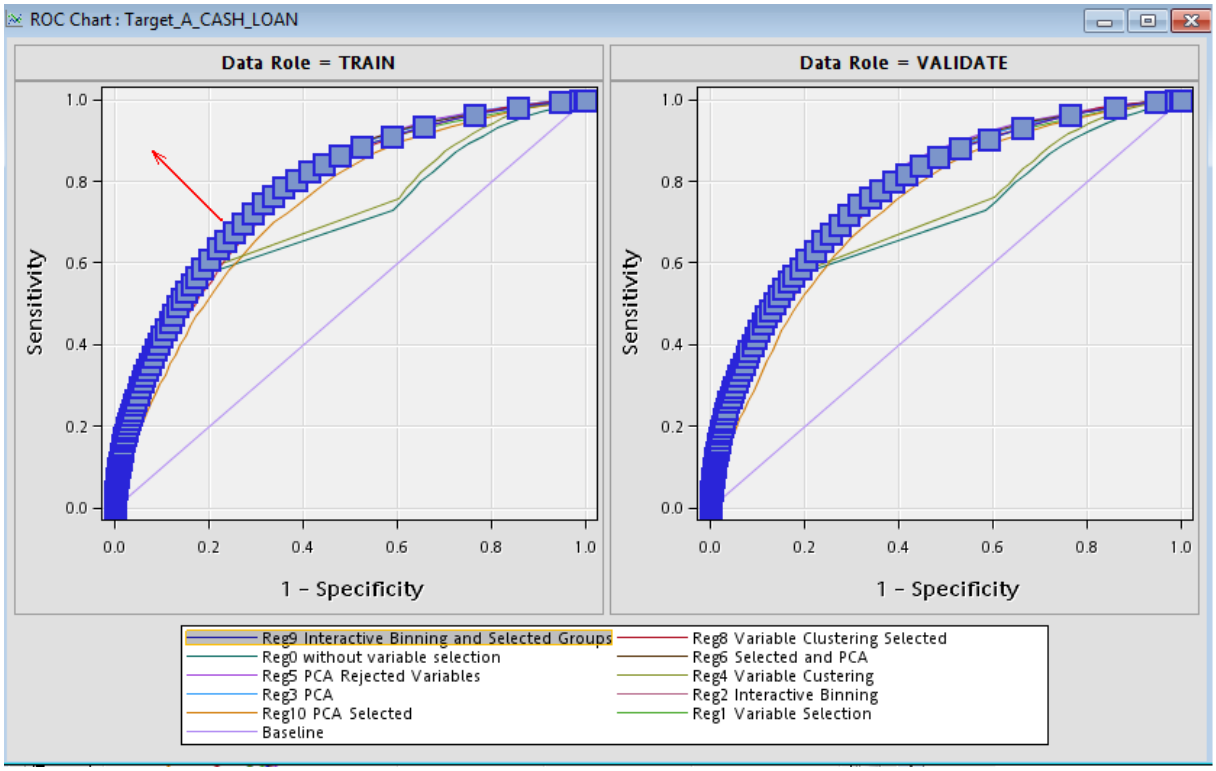


Slika 53. Izbor najboljeg modela

Komponenta *Model Comparation* poredi modele koristeći različite kriterijume i tehnike. Izbor kriterijuma zavisi od primene modela. Za binarne ciljne promenljive ti kriterijumi su grupisani po tipu analize i mogu biti:

- Klasifikacione mere kao što je ROC (***R**eceivier **O**perating **C**haracteristics*) grafikon i kriva, odnos klasifikacije (*classification rates*) i sl.
- Procena modela kroz prizmu profita i gubitka (eng. *lift measure*). Ove mere su opisane u poglavlju *8.2.1* Rezultat regresione analize
- Statističke mere kao što su BIC (eng. *Bayesian Information Criterion*), AIC (eng. *Akaike's Information Criterion*), Gini, Kolmogorov-Smirnov, *Bin-Best-Two-Way* Kolmogorov-Smirnov test

Na slici (Slika 54) prikazane su ROC krive za sve funkcije linearne regresije. *Baseline* (Y=X) predstavlja ROC krivu u slučaju *nepostojanja modela*.



Slika 54. ROC krive svih modela

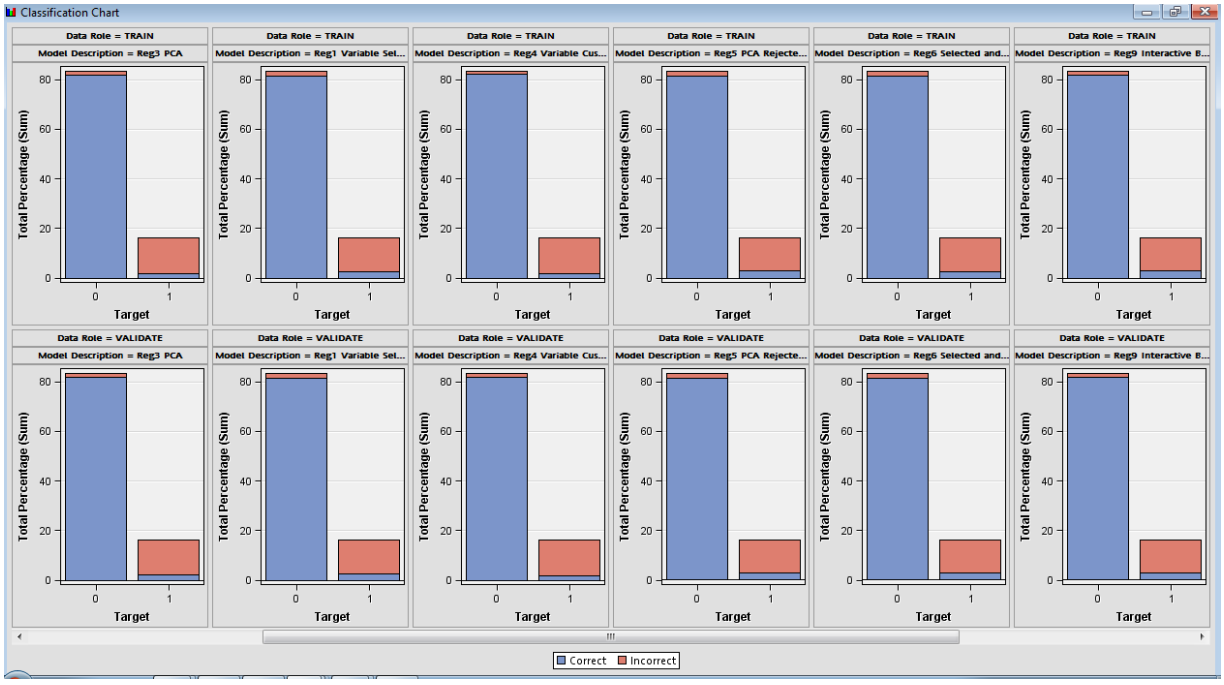
ROC kriva se računa tako što se podaci pripreme na sličan način kao i kod *lift* mera (videti 8.2.1 *Rezultat regresione analize*). Za ovako uzete grupe računaju se mere *Sensitivity* i *Specificity* (Slika 55).

		Condition	
		Condition positive	Condition negative
Test outcome	Test outcome positive	True positive	False positive (Type I error)	**Precision**= ∑ True positive/ ∑ Test outcome positive
	Test outcome negative	False negative (Type II error)	True Negative	**Negative predictive value**= ∑ True negative/ ∑ Test outcome negative
		Sensitivity= ∑ True positive/ ∑Conditional positive	**Specifity**= ∑ True negative/ ∑Conditional negative	

Slika 55. Formule za računanje *Sensitivity* i *Specificity* na uzorku.

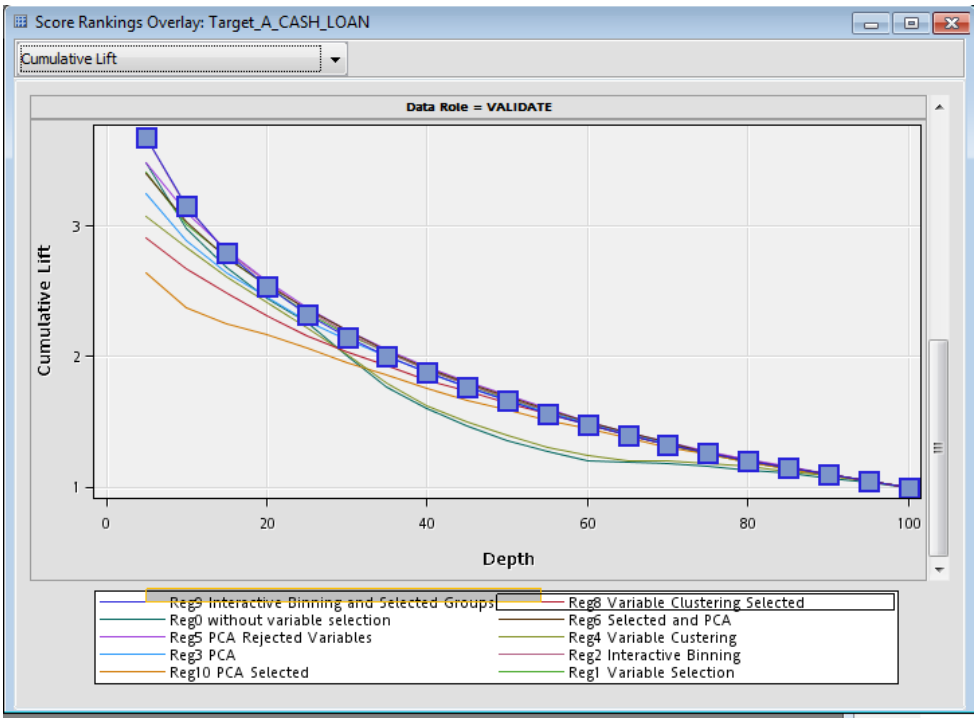
Najbolji model je u slučaju da je *Specifity*=1, *Sensitivity* =1 tj. funkcija modela pogađa sa verovatnoćom 100%.

Na slici (Slika 54) to je gornji levi ugao grafikona (1-*Specifity*=0, *Sensitivity*=1). Što je kriva bliže gornjem levom uglu (crvena strelica - Slika 54) to model bolje predviđa.

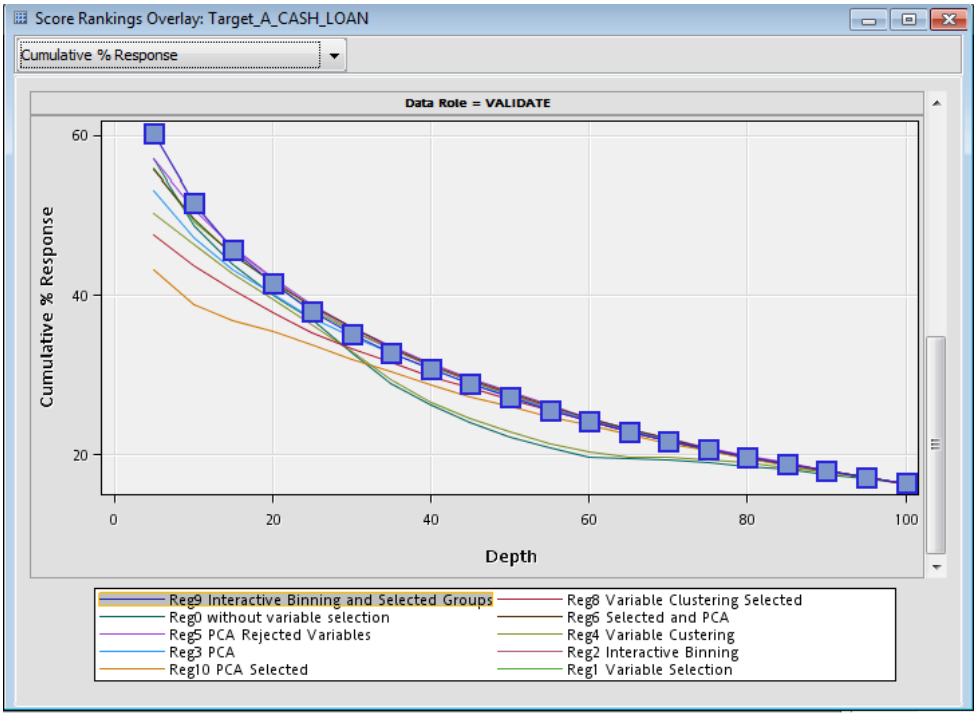


Slika 56. Matrica 2x2 iz slike 53 nad trening uzorkom i uzorkom za proveru prikazana grafički za sve modele

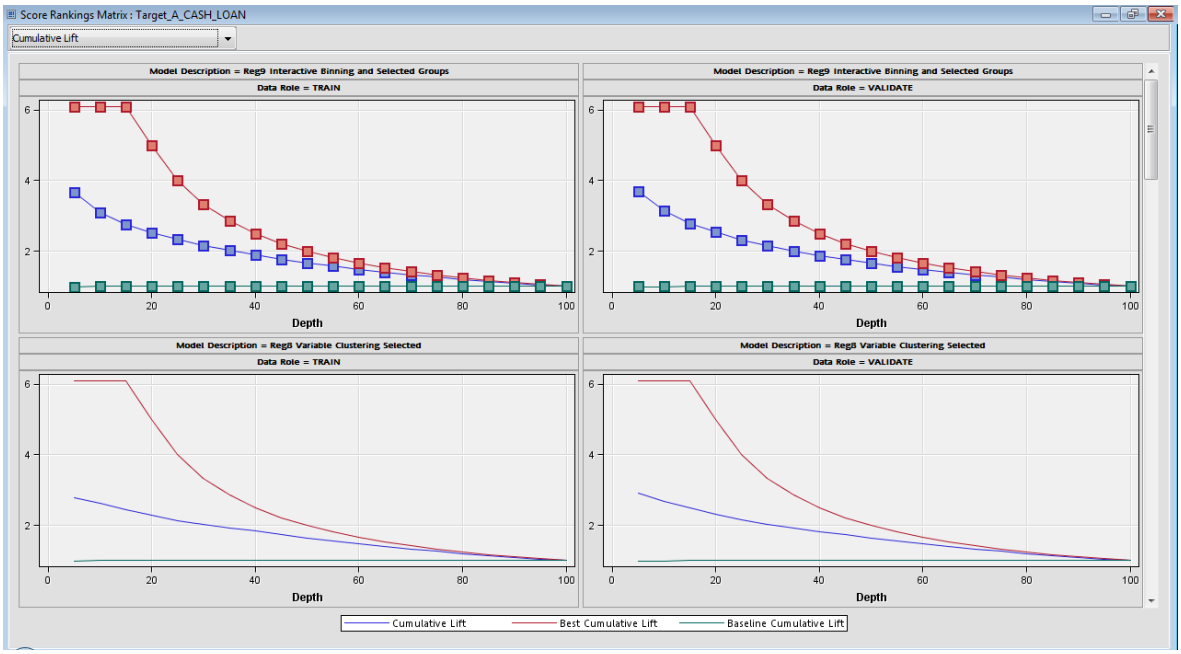
Lift mere su opisane u poglavlju 8.2.1. Narednih nekoliko slika pokazuju *lift* mere za sve modele.



Slika 57. Krive kumulativnog lifta modela nad uzorkom za proveru



Slika 58. Cumulative % response krive modela na uzorkom za proveru



Slika 59. *Score Ranking Matrix*

Na slici (Slika 60) prikazane su statističke mere koje se mogu koristiti u izboru najboljeg modela. Izbor modela je moguće napraviti po različitim merama izračunatim nad trening ili uzorkom za proveru. U ovom slučaju korišćena je mera *Misclassification Rate*. Model *Reg9* je izabran kao najbolji.

Fit Statistics

Selected Model	Predecessor Node	Selection Criterion: Valid: Misclassification Rate ▲	Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Average Error Function	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom	Train: Total Degrees of Freedom	Train: Divisor for ASE	Train: Error Function	Train: Final Prediction Error	Train: Maximum Absolute Error
Y	Reg9	0.15313	45775.92	0.114247	0.370318	61203	163	61366	122732	45449.92	0.114855	0.98224
	Reg2	0.153692	45481.24	0.113903	0.367966	61206	160	61366	122732	45161.24	0.114499	0.9898
	Reg5	0.156625	45314.56	0.114104	0.367423	61256	110	61366	122732	45094.56	0.114514	0.99563
	Reg7	0.157309	57831.64	0.124449	0.470225	61306	60	61366	122732	57711.64	0.124693	
	Reg	0.157676	46014.82	0.11567	0.373813	61298	68	61366	122732	45878.82	0.115927	0.99383
	Reg6	0.157994	45606.65	0.114841	0.37008	61273	93	61366	122732	45420.65	0.115189	0.99302
	Reg3	0.15924	46391.93	0.117296	0.37625	61259	107	61366	122732	46177.93	0.117706	0.9974
	Reg4	0.162369	50932.12	0.125517	0.411173	61132	234	61366	122732	50464.12	0.126478	0.99708
	Reg8	0.163029	47419.75	0.12042	0.386091	61349	17	61366	122732	47385.75	0.120487	0.99340
	Reg10	0.163615	48892.83	0.123754	0.398306	61362	4	61366	122732	48884.83	0.12377	0.96825

Slika 60. Izbor šampion modela

Cross Validate Misclassification Rate (False negative rate) se računa na sledeći način:

1. Pripreme se grupe iste veličine kao za lift mere nad uzorkom (poglavlje 8.2.1).
2. Nad tako pripremljenim grupama izračuna se greška svake grupe po formuli

   $Err_i=(FN_i)/(TP_i+FN_i)$,

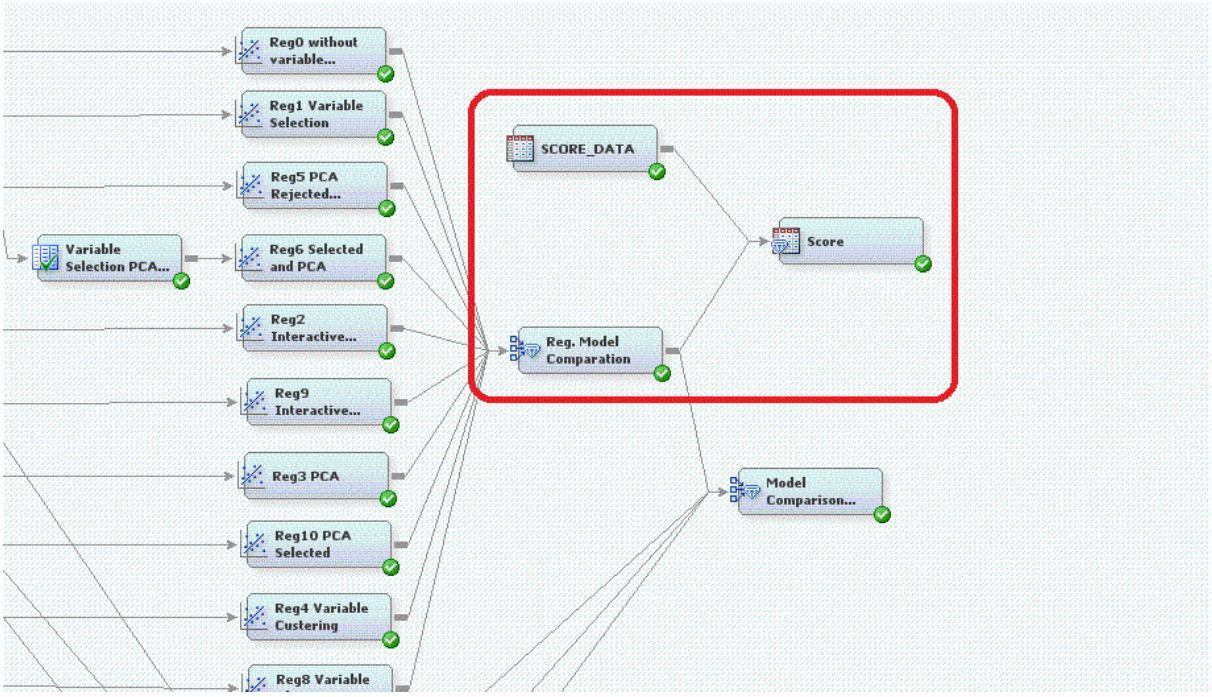
   gde je i=1,…n, a *FN - false* negative , T*P – true positive*, (Slika 55)
3. $Misclassification\ Rate=(Err_1+Err_2+…Err_n)/n$

9 Model u produkciji

9.1 Promocija modela

9.1.1 Računanje verovatnoće nad testnim uzorkom

SAS Komponenta *Score* računa skore/verovatnoću nad proizvoljim uzorkom (Slika 61).

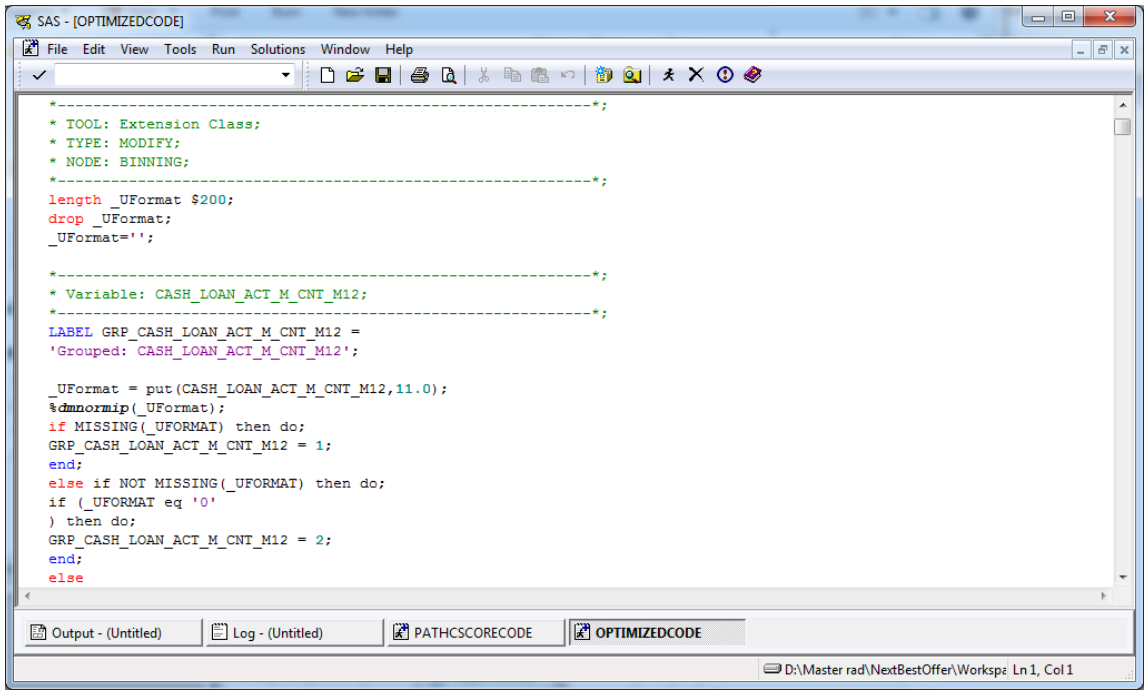


Slika 61. Računanje verovatnoće nad proizvoljnim uzorkom

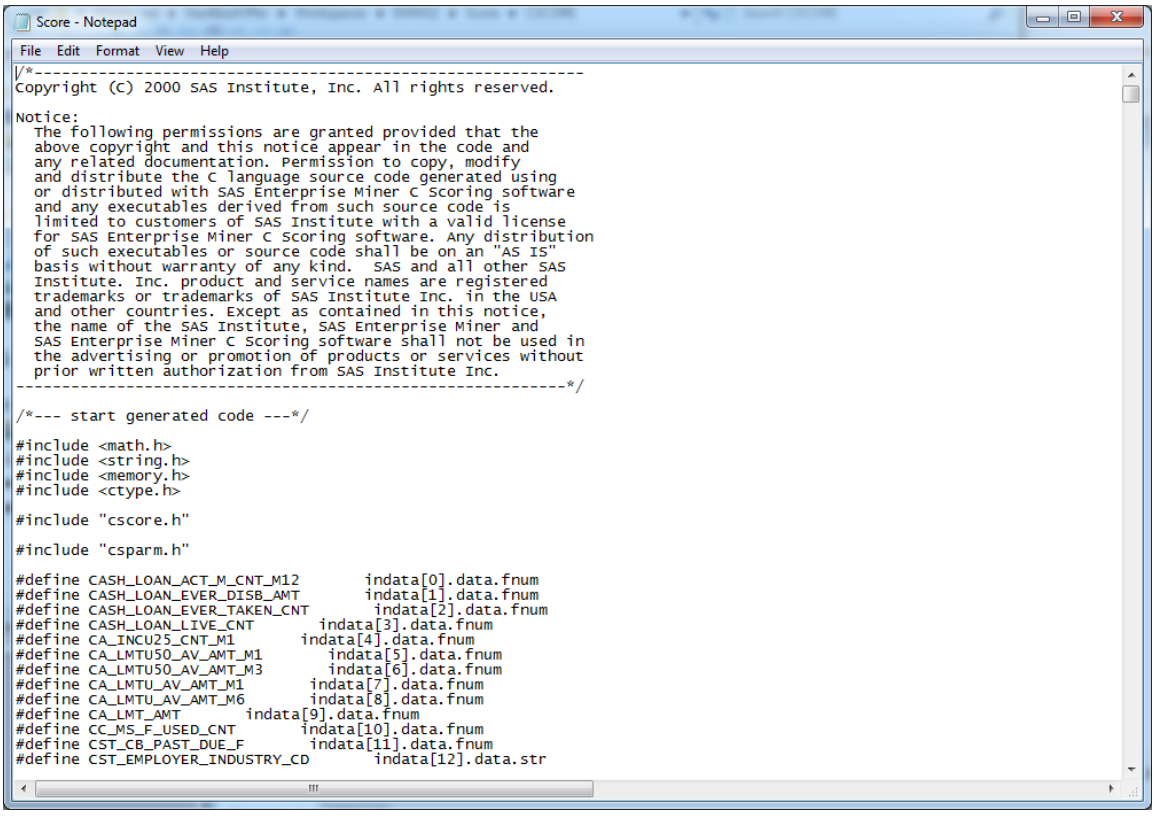
Verovatnoća se računa primenjujući funkciju izabranog modela u *Model Comparation* komponenti (videti. 8.2.2 Izbor *najboljeg modela*). Nad ovako izračunatim skorom moguće se primeniti sve statistike iz poglavlja *Ocena modela* kako bi proverili izabranu funkciju modela. Pomeranjem vremenske dimenzije testnog uzorka u odnosu na uzorak za razvoj obezbeđujemo bolju ocenu modela.

9.1.2 Priprema programskog koda za računanje verovatnoće

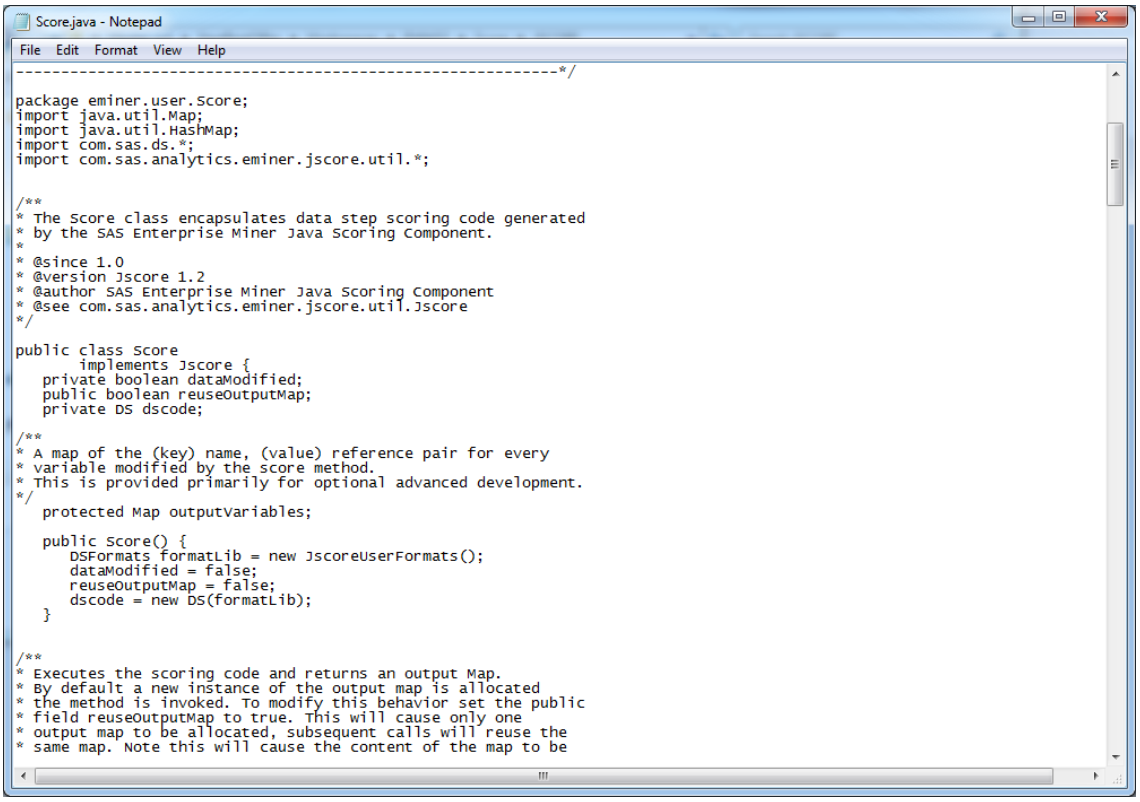
SAS komponenta *Score* osim što automatski računa skor/verovatnoću može generisati programski kod za računanje iste. Ovako generisan programski kod se može koristiti u drugim aplikacijama. Na narednim slikama su prikazani primeri generisanog SAS koda, DB2 skalarne funkcije, C koda i Java koda.



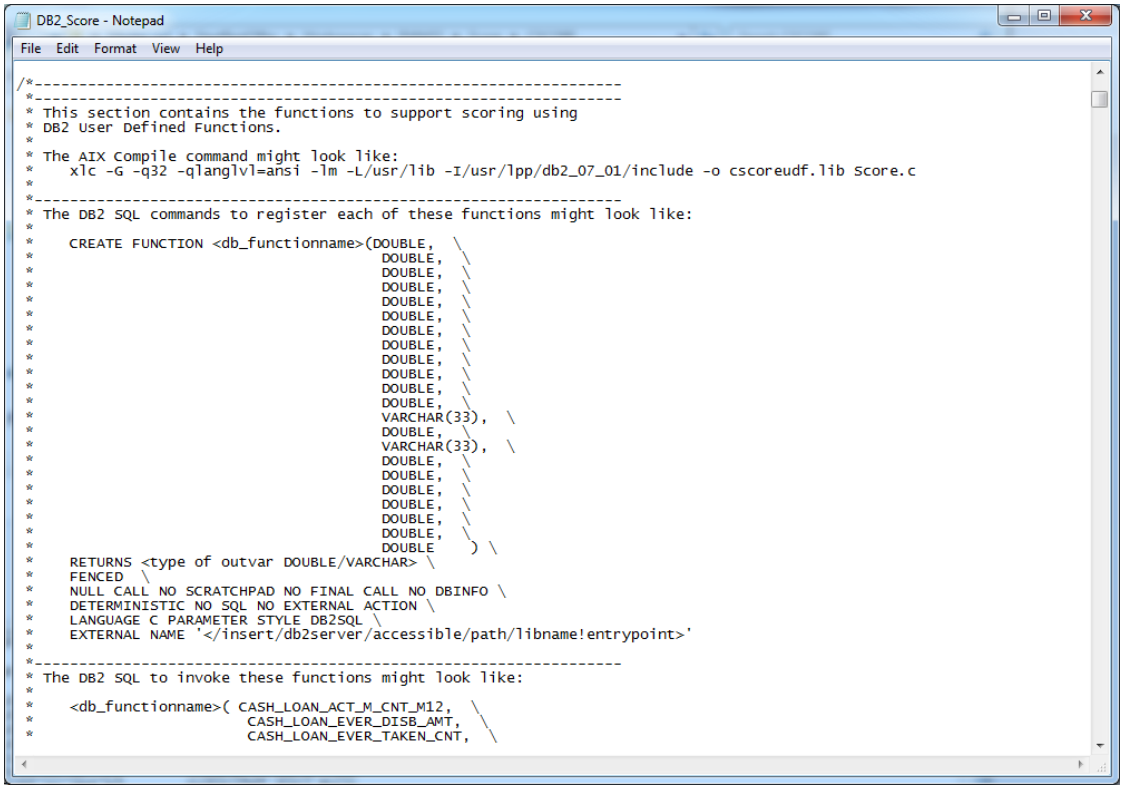
Slika 62. SAS kod za računanje verovatnoće



Slika 63. C kod za računanje verovatnoće

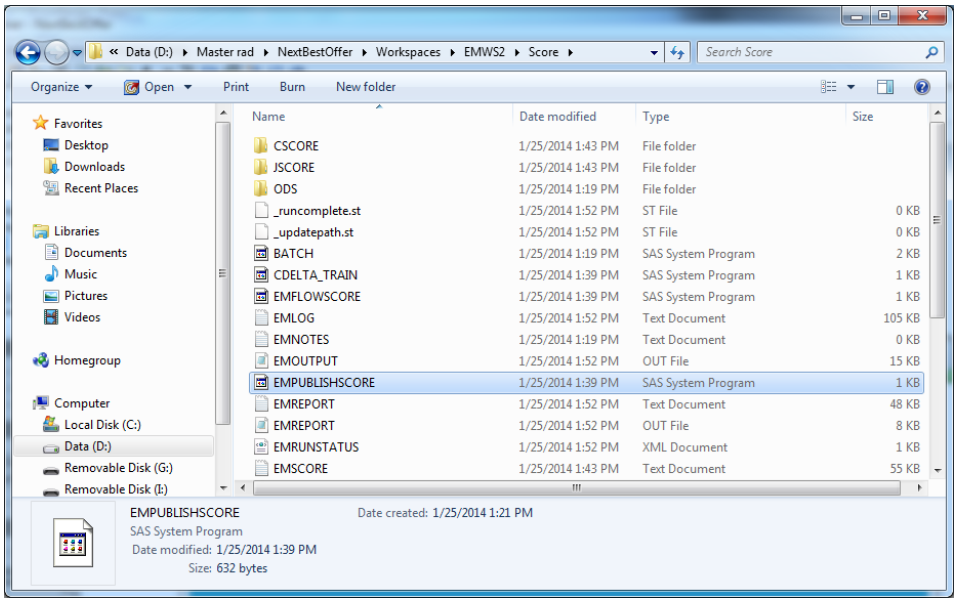


Slika 64. Java kod za računanje verovantoće



Slika 65. DB2 skalarna funkcija za računanje verovatnoće

Generisani kod se nalazi u datotekama unutar SAS EM projekta (Slika 66).



Slika 66. Generisani fajlovi sa kodom za računanje verovatnoće

U slučaju da kompanija ima *SAS Enterprise BI* rešenje moguće je objaviti projekat kao SAS paket (*SAS EM package*) u SAS metadata repozitorijumu. Ovako objavljen paket ima interfejs za ulazne podatke kao i mogućnost zakazivanja izvršavanja paketa. Izvršavanjem paketa računa se verovatnoća nad čitavim skupom podataka. Ovako izračunata verovatnoća se koristiti za nadgledanje modela (eng. m*odel monitoring*).

9.1.3 Korigovanje verovatnoće

Posle izračunatog skora neophodno je korigovati skor s obzirom da smo primenili tehniku *oversampling* (videti poglavlje 6.1 *Analiza frekvencije ciljne promenljive u uzorku*). Ovaj skor se koriguje pomoću sledeće formule:

correct probabilities = 1/(1+(1/***original fraction***-1)/(1/***oversampled fraction***-1)*(1/***scoring result***-1))

gde je

original fraction- procentualni udeo *event* populacije u originalnom uzorku

oversampled fraction – procentualni udeo *event* populacije u *oversample* uzorku

scoring result – rezultat skora dobijen primenom formule modela

SAS kod u ovom slučaju je:

```
FIXED_p_scoring1 = 1/(1+(1/0.0513-1)/(1/0.16-1)*(1/p_scoring1-1));
```

gde je p_scoring1 skor dobijen SAS SCORE komponentom.

9.2 Primena modela

Računanje skora/verovatnoće moguće je uraditi na dva načina:

- računanje skora mesečno masovnom obradom,
- računanjem skora na zahtev.

U slučaju da se promenljive koje se koriste u modelu računaju na mesečnom nivou računanje skora na zahtev ne daje drugačije rezultate u odnosu na mesečnu obradu. Računanje skora na zahtev ima smisla u slučaju da se u modelu koriste promenljive koje se mogu osvežiti u realnom vremenu.

Primer. Ulazna promenljivu CA_LMT25_AV_AMT_M1 (prosečna iskorišćenost limita po tekućem računa na mesečnom nivou u slučaju da je iskorišćenost bila veća od 25% limita) nije lako izračunati i malo je verovatno da će se ovo računati na zahtev prodavca u trenutku kada komunicira sa klijentom.

Ulazne promenljive CASH_LOAN_EVER_DISB_AMT (iznos do sada odobrenih kredita) ili CST_OPENED_PATH_ARRAY_CD (redosled kupovine proizvoda u banci) je izvesno da se mogu izračunati u realnom vremenu.

U ovom radu je prikazan razvoj jednog modela. Banka obično ima više različitih modela za različite proizvode. Za svakog klijenta se računaju verovatnoće po svim modelima za sve proizvode. Ovako izračunate verovatnoće mogu se porediti samo u slučaju da je verovatnoća korigovana (videti poglavlje 9.1.3 *Korigovanje verovatnoće*).

U slučaju *inbound CRM*, a zavisno od CRM aplikacije koju banka poseduje, moguće je primeniti izračunate verovatnoće na različite načine. Evo nekoliko primera:

- na prodajnom ekranu prodavca se implementira komandna tabla (eng. *dashboard*) na kojoj su prikazane verovatnoće za različite proizvode; prodavac prvo nudi proizvode sa najvećom verovatnoćom
- prilikom korišćenja platne kartice na ATM moguće je prikazati ponudu klijentu u obliku personalizovane poruke; ponuda obuhvata proizvod sa najvećom verovatnoćom
- prilikom korišćenja e-banking usluga moguće je prikazati korisniku personalizovanu reklamu na sajtu banke tj. ponudu za proizvod sa najvećom verovatnoćom
- prilikom uzimanja tiketa za čekenje u redu u ekspozituri moguće je na samom tiketu napisati personalizovanu poruku za klijenta u slučaju da se klijent autentifikovao (npr. karticom ili čipovanom ličnom kartom).

Svi napravljeni modeli se mogu koristiti i za *outbound* CRM tj. za sprovođenje CRM kampanja. U ovom slučaju se ponude šalju poštom, elektronskom poštom, SMS-om, MMS-om,…

9.3 Nadgledanje modela

Nadgledanje modela (eng. *model monitoring*) je veoma važan aspekt održavanja modela. Za svaki model se na mesečnom nivou računaju statistike iz poglavlja 8.2 *Ocena modela* i porede se sa izračunatim statistikama u ranijim mesecima. Svako odstupanje se mora dodatno analizirati i objasniti.

Osim ovih statistika veoma je važno pratiti i kvalitet ulaznih promenljivih funkcije modela. Može se desiti da model slabije pogađa ne zbog toga što je model loš, već zbog samog kvaliteta ulaznih podataka. Zbog toga, neophodno je proveravati i statistike ovih ulaznih promenljivih u odnosu na ciljnu promenljivu.

U ovom radu najbolji model je *Reg9* (Slika 53 u poglavlju 8.2.2 *Izbor najboljeg modela*) koji koristi *InteractiveBining* pa *VariableSelection*. Za svaku promenljivu neophodno je pratiti:

- Gini koeficijent grupe istorijski; nagla promena koeficijenta u nekum mesecima (eng. *peak*) ukazuje da nešto nije u redu sa podacima.
- Stabilnost grupa u nekom vremenskom periodu (T1,T3,T6) . Stabilnost grupe se računa pomoću matice migracije (GR_i,GR_j) i,j≤n, gde svaki element predstavlja procentulani udeo migriranih klijenata iz grupe GR_i u grupu GR_j. Indeks stabilnosti se nalazi na dijagonali matrice (GR_i, GR_i). Ako indeks stabilnosti odstupa od uobičajnog (npr. gleda se mesec za mesec) potrebno je uraditi dodatne analize i ispitati uzroke.

10 Zaključak

Istraživanjem podataka zadovoljavajaju se analitičke potrebe kompanije i ono se nalazi na vrhu piramide DWH/BI. Mnoge kompanije imaju DWH/BI, ali malo njih se bavi istraživanjem podataka opisanom u radu. Izrada DWH/BI sistema, koji za cilj ima istraživanje podataka i razvoj prediktivnih i deskriptivnih analitičkih modela je kompleksan, izrazito iterativan i skup. Potrebno je nabaviti adekvatan hardver i softver, zatim anagažovati inžinjere/konsultante (interno ili eksterno) da razviju DWH tj. da dobro struktuiraju, konsoliduju, očiste i pripreme podatke za proces istraživanja i na kraju treba angažovati analitičare za razvoj analitičkih modela.

U radu je prikazana metodologija izrade matematičkih modela koji se koriste kao podrška prodaji u bankarskoj industriji. Objašnjeno je kako se definiše poslovni problem i priprema uzorak za razvoj modela. Nad pripremljenim podacima urađene su razne statističke analize i opisane metode redukovanja i izbora promenljivih. Nad izabranim promenljivama autor je razvio 10 matematičkih modela zasnovanih na logističkoj regresiji. Na kraju opisan je izbor najboljeg modela kao i njegova primena u sistemu „sledeća najbolja ponuda za klijenta“.

Analitiku predstavlja 6 datoteka (tabela) iste strukture na nivou klijenta sa 2043 promenljive izračunate u 7 vremenskih trenutaka/perioda sa 2,5 miliona opservacija. Za pripremu podataka korišćena je trogodišnja istorija poslovanja banke počev od transakcija klijenata, agregacija na nivou računa i klijenta pa do eksternih izvora kao što je kreditni biro. Za pripremu podataka autor je utošio 5 nedelja (25 FTE[34]).

Istraživanje podataka je iterativan proces. Autor je napravio nekoliko iteracija koje se uglavnom odnose na ispravku i čišćenje podataka kao i pripremu uzorka za proces istraživanja. Preliminarno istraživanje podataka je uzelo 5 FTE, a izbor promenljivih za modelovanje, modelovanje i ocena modela uzelo je oko 10 FTE.

Kao što je opisano i prethodnim pasusima najviše vremena se utroši na pripremu podataka za modelovanje. Priprema podataka je kompleksna, izrazito iterativna i skupa. Na sreću, poslovna korist za kompaniju od pripremljenih podataka je višestruka. Jednom pripremljeni podaci mogu se koristiti za razvoj različitih prediktivnih i deskriptivnih analitičkih modela iz različitih poslovnih oblasti.

Životni vek modela „sklonost ka kupovini“ nije dug tj. nakon nekoliko meseci model ne daje dobre rezultate i neophodno je razviti novi model. Razlozi za kratak vek su različiti: promena tržišta, promena ponašanja klijenata, kvaliteta prikupljenih podataka, kvaliteta samog modela,... S druge strane razvoj modela je brz (2-3 nedelje bez pripreme podataka) i relativno jeftin što kompanijama daje mogućnost da brzo i bez nekih većih

[34] FTE (full time equivalent) predstavlja jednodnevno angažovanje (8h) jednog zapolenog (u ovom slučaju inžinjera/analitičara/konsultanta)

ulaganja reaguju i odgovarajućom akcijom uvećaju svoju prodaju. U ovom radu je prikazano da se prodaja može uvećati 2,5 do 3,5 puta u odnosu na odsustvo bilo koje akcije. U praksi, zbog raznih drugih okolnosti, poslovna korist obično je malo manja ali i dalje dovoljno velika da opravda uložena sredstva i inicira nova ulaganja na unapređenju izrade modela.

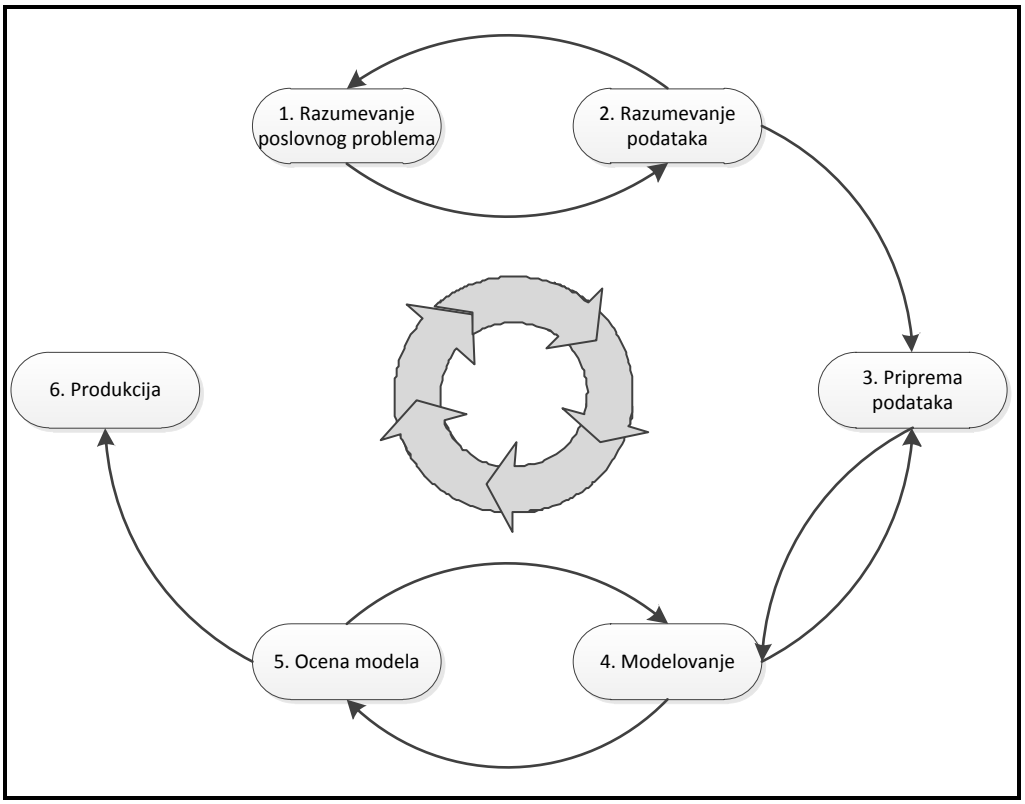
Metodologija i tehnike koje su opisane u radu mogu pomoći i dati ideje kako istraživati podatke i organizovati proces modelovanja, ali one ne mogu biti opšte pravilo ili šablon. Istraživanje podataka je specifična i kompleksna obast. Izbor metodoligije i tehnike istraživanja često zavisi od poslovnog problema kojeg treba opisati kao i od kvaliteta podataka sa kojim raspolažamo.

Autor i mentor ne mogu biti odgovorni za eventulane gubitke kompanije ili pojedinca koji mogu nastati primenom tehnika opisanog u ovom radu.

A. SAS Enterprise Miner

SEMMA eng. *Sample, Explore, Modify, Model, Asses* predstavlja pristup u izradi modela. Ovo nije metodologija i okvir u procesu modelovanja kao što su KDD i CRISP[35], ali je dobra praksa u izradi modela. SEMMA pristup obuhvata najvažniji deo u procesu izrade modela.

Na slici (Slika 67) prikazani su koraci u izradi modela. Ovaj proces je iterativan i u svakom trenutku moguće je vratiti se jedan ili više koraka unazad, a nekada čak i na početak.



Slika 67. Proces modelovanja

U poglavljima koji slede biće opisani SAS *Enterprise Miner* komponente/alati koristeći SEMMA pristup.

[35] Uporedna analiza je opisana u "*KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW – Ana Avezado, Manuel Filepe Santos, ISBN: 978-972-8924-63-8 © 2008 IADIS*"

A.1 Formiranje uzorka - *Sample*

U ovom poglavlju biće detaljno opisane SAS EM komponente/alati. To su: ***Input Data***, ***Sampe***, ***Data Partition***. Ove komponente se koriste za pripremu uzorka za modelovanje.

A.1.i Komponenta *Input Data*

Prvi korak u procesu razvoja modela je povezivanje ulaznih podataka (posmatrana populacija) sa metapodacima samog modela.

Populacija je u obliku SAS tabele (eng. *Analytic Base Table* – skraćeno ABT) u kojoj jedan red odgovara pojedinačnom uzorku (opservaciji) dok jedna kolona predstavlja promenljivu. Ovo je prvi kontakt sa populacijom unutar SAS EM alata.

Celoj populaciji neophodno je dodeliti odgovarajuću ulogu. Uloga može biti:

- ***Raw*** – sirovi podaci (imaju opštu namenu)
- ***Train*** – populacija se koristi za razvoj modela
- ***Validate*** – populacija se koristi za proveru modela
- ***Test*** – populacija se koristi za test modela
- ***Score*** – populacija je učitana radi računanja skora (koristi se prilikom promocije modela na produkciju)
- ***Transaction*** – radi se o transakcijskoj datoteci i mora imati bar jednu vremensku dimenziju

Za svaku promenljivu iz populacije neophodno je definisati njenu ulogu (eng. *role*) koju će imati u procesu istraživanja podataka. Uloga govori alatu o nameni same promenljive tj. da li u daljem procesu istraživanja treba praviti statistike ili ne. Najvažnije uloge su:

- ***Input*** – ulazna promenljiva; nezavisna promenljiva
- ***Target*** - ciljna ili zavisna promenljiva; obavezno je postaviti tačno jednu u slučaju da se radi o izradi modela zasnovanih na verovatnoći/skoru
- ***Rejected*** – napuštena promenljiva tj. neće biti propagirana dalje u procesu istraživanja
- ***Id*** – Id se ne koristi u procesu istraživanja i predstvlja jedinstevni identifikator pojedinačne opservacije.

Uloga se može promeniti u procesu modelovanja. Najčešće se neke promenljive napuštaju tj. postavlja im se uloga „*Rejected*“, a izvode se nove promenljive koje dobijaju ulogu „*Input*“

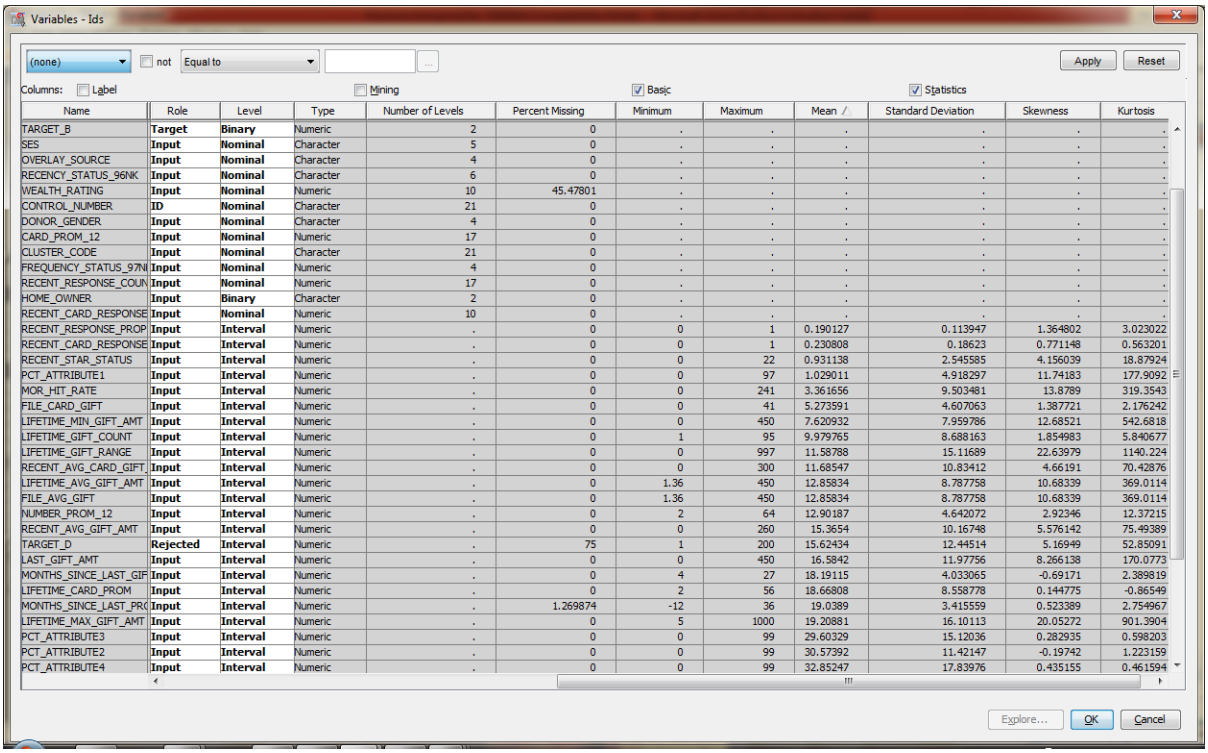
Za svaku promenljivu potrebno je proveriti kardinalnost domena (u SAS EM označena kao „*level*“). Promenljive po ovoj podeli delimo na:

- Binarne – promenljiva može imati dva stanja
- Intervalne – promenljiva može imati beskonačno stanja pri čemu je rastojanje između susednih članova jednako.
- Nominalne – promenljiva ima konačno stanja pri čemu ne postoji uređenost

između članova niti je poznato rastojanje

- Ordinarne – promenljiva ima konačno mnogo stanja pri čemu znamo uređenost članova kao i njihovo rastojanje. Npr. nivo obrazovanja se može tretirati kao nominalna i ako ordinarna promenljiva pri čemu „rastojanje između“ nivou ne mora biti jednako (1 – osnovna škola 2- srednja škola, 4- viša škola, 5- visoka škola, 7- master, 10- doktorat). Specifičnost ordinarnih varijabli je što se za nju mogu raditi statistike i za intervalne i za nominalne promenljive.
- Unarne promenljiva može imati samo jednu vrednost/stanje

SAS EM automatski generiše atribut „***level***“ na osnovu metapodataka i napravljenih statistika. Tako se sve promenljive znakovnog tipa obeležavaju kao nominalne, a sve promenljive numeričkog tipa obeležavaju se kao unarne, binarne ili intervalne. Zbog toga prilikom definisanja specifikacije za pripremu ABT neophodno je definisati tipove kolona u skladu sa predhodno navedenim pravilima.



Name	Role	Level	Type	Number of Levels	Percent Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
TARGET_B	Target	Binary	Numeric	2	0	.	.	.	.	.	.
SES	Input	Nominal	Character	5	0	.	.	.	.	.	.
OVERLAY_SOURCE	Input	Nominal	Character	4	0	.	.	.	.	.	.
RECENCY_STATUS_96NK	Input	Nominal	Character	6	0	.	.	.	.	.	.
WEALTH_RATING	Input	Nominal	Numeric	10	45.47801	.	.	.	.	.	.
CONTROL_NUMBER	ID	Nominal	Character	21	0	.	.	.	.	.	.
DONOR_GENDER	Input	Nominal	Character	4	0	.	.	.	.	.	.
CARD_PROM_12	Input	Nominal	Numeric	17	0	.	.	.	.	.	.
CLUSTER_CODE	Input	Nominal	Character	21	0	.	.	.	.	.	.
FREQUENCY_STATUS_97N	Input	Nominal	Numeric	4	0	.	.	.	.	.	.
RECENT_RESPONSE_COUN	Input	Nominal	Numeric	17	0	.	.	.	.	.	.
HOME_OWNER	Input	Binary	Character	2	0	.	.	.	.	.	.
RECENT_CARD_RESPONSE	Input	Nominal	Numeric	10	0	.	.	.	.	.	.
RECENT_RESPONSE_PROP	Input	Interval	Numeric	.	0	0	1	0.190127	0.113947	1.364802	3.023022
RECENT_CARD_RESPONSE	Input	Interval	Numeric	.	0	0	1	0.230808	0.18623	0.771148	0.563201
RECENT_STAR_STATUS	Input	Interval	Numeric	.	0	0	22	0.931138	2.545585	4.156039	18.87924
PCT_ATTRIBUTE1	Input	Interval	Numeric	.	0	0	97	1.029011	4.918297	11.74183	177.9092
MOR_HIT_RATE	Input	Interval	Numeric	.	0	0	241	3.361656	9.503481	13.8789	319.3543
FILE_CARD_GIFT	Input	Interval	Numeric	.	0	0	41	5.273591	4.607063	1.387721	2.176242
LIFETIME_MIN_GIFT_AMT	Input	Interval	Numeric	.	0	0	450	7.620932	7.959786	12.68521	542.6818
LIFETIME_GIFT_COUNT	Input	Interval	Numeric	.	0	1	95	9.979765	8.688163	1.854983	5.840677
LIFETIME_GIFT_RANGE	Input	Interval	Numeric	.	0	0	997	11.58788	15.11689	22.63979	1140.224
RECENT_AVG_CARD_GIFT	Input	Interval	Numeric	.	0	0	300	11.68547	10.83412	4.66191	70.42876
LIFETIME_AVG_GIFT_AMT	Input	Interval	Numeric	.	0	1.36	450	12.85834	8.787758	10.68339	369.0114
FILE_AVG_GIFT	Input	Interval	Numeric	.	0	1.36	450	12.85834	8.787758	10.68339	369.0114
NUMBER_PROM_12	Input	Interval	Numeric	.	0	2	64	12.90187	4.642072	2.92346	12.37215
RECENT_AVG_GIFT_AMT	Input	Interval	Numeric	.	0	0	260	15.3654	10.16748	5.576142	75.49389
TARGET_D	Rejected	Interval	Numeric	.	75	1	200	15.62434	12.44514	5.16949	52.85091
LAST_GIFT_AMT	Input	Interval	Numeric	.	0	0	450	16.5842	11.97756	8.266138	170.0773
MONTHS_SINCE_LAST_GIF	Input	Interval	Numeric	.	0	4	27	18.19115	4.033065	-0.69171	2.389819
LIFETIME_CARD_PROM	Input	Interval	Numeric	.	0	2	56	18.66808	8.558778	0.144775	-0.86549
MONTHS_SINCE_LAST_PRO	Input	Interval	Numeric	.	1.269874	-12	36	19.0389	3.415559	0.523389	2.754967
LIFETIME_MAX_GIFT_AMT	Input	Interval	Numeric	.	0	5	1000	19.20881	16.10113	20.05272	901.3904
PCT_ATTRIBUTE3	Input	Interval	Numeric	.	0	0	99	29.60329	15.12036	0.282935	0.598203
PCT_ATTRIBUTE2	Input	Interval	Numeric	.	0	0	99	30.57392	11.42147	-0.19742	1.223159
PCT_ATTRIBUTE4	Input	Interval	Numeric	.	0	0	99	32.85247	17.83976	0.435155	0.461594

Slika 68. Lista promenljivih sa dodeljnim ulogama i određenim tipovima promenljivih.

Veoma je važno proveriti da li je „*level*“ ispravno unet za svaku promenljivu ili grupu promenljivih, jer pogrešno uneta opcija „*level*“ alatu će dati pogrešne metapodatke o promenljivoj i samim tim neće biti tretirana ispravno u procesu modelovanja. U procesu modelovanja često se koristi nekoliko stotina, a nekada i nekoliko hiljada promenljivih, pa sama provera „odozgo na dole“ može biti naporna tj. praktično neizvodljiva. Zbog toga postoji filter u gornjem levom uglu (Slika 68), na osnovu kojeg je moguće izabrati neke promenljive. Imenovanje kolona u samom ABT može olakšati proveru istraživaču. Tako se prilikom imenovanja koristi sufiks koji bliže opisuje tip i svojstvo promenljive. Na primer, sufiksi mogu biti:

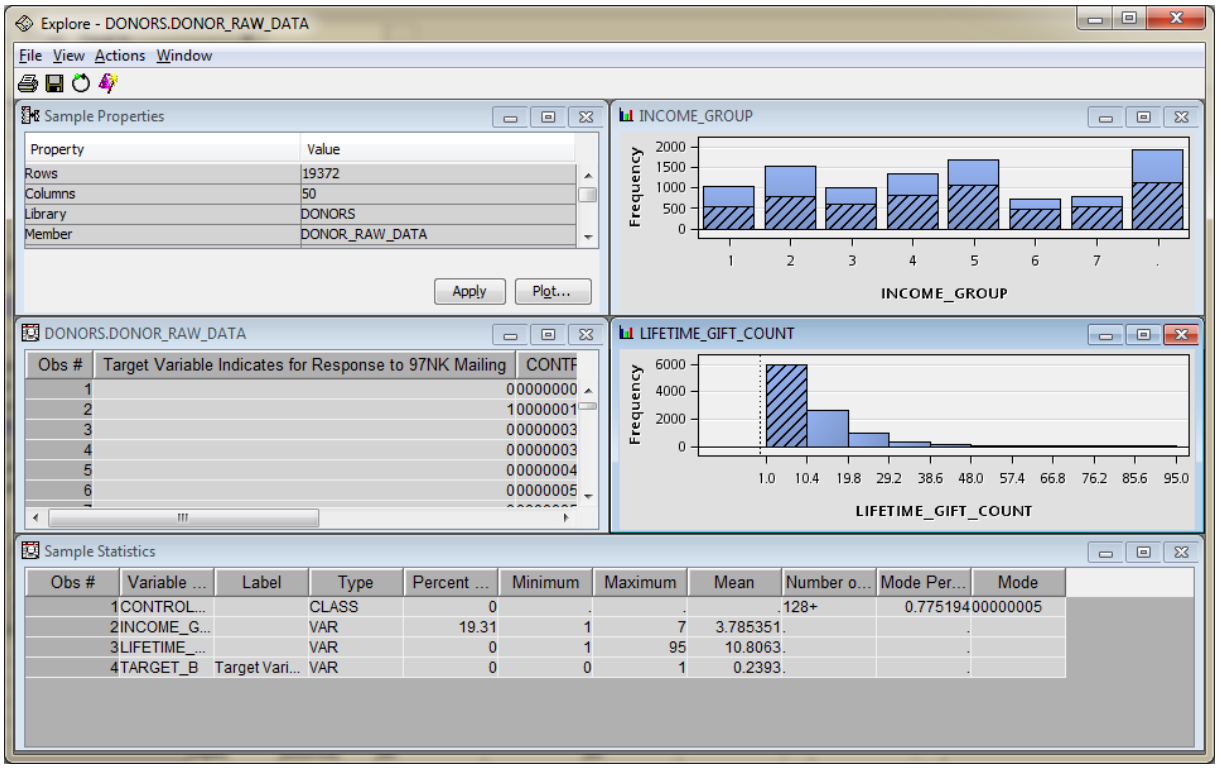
- *CD* – skraćeno od eng. *code* i označava nominalne (kategoričke) promenljive, obično je znakovnog tipa od 3 znaka

- *RK* – surogat ključ – ne koristi se u procesu modelovanja
- *ID* – poslovni ključ ne koristi se u procesu modelovanja
- *AMT* – skraćeno od eng. *amount* i označava uvek intervalnu promenljivu i predstavlja neki novčani iznos
- *CNT* – skraćeno od eng. *count* predstavlja neko brojanje za period ili na dan; obično predstavlja intervalnu promenljivu ali se ponekad može transformisati i u nominalnu
- *RT* – predstavlja procenat ili odnos (eng. *rate*) i uvek je intervalna promenljiva
- *DSC* – opisna promenljiva ne koristi se u procesu modelovanja
- *FLG* – obično je binarna promenljiva ima vrednosti 1,0 odnosno 'Y', 'N'.

Primer. Promenljiva WEALTH_RATING (Slika 68) je numerička, a predstavlja klasifikaciju klijenta. Inicijalno ona je bila intervalna, ali istraživač je morao ručno da joj promeni rolu u nominalnu.

Opcija „***Variable***" pruža nam mogućnost da napravimo neke korisne statistike. Unutar samog prozora moguće je uključiti opciju „***Statistic***" koja nam daje osnovne statistike promenljivih kao što su: broj članova nominalne promenljive (***Number of Levels***), minimalnu, maksimalnu, prosečnu vrednost, standardnu devijaciju, skewness, kurtosis.

Takođe, moguće je detaljnije istražiti promenljive obeležavanjem jedne ili više njih i izborom opcije „***Explore***". Na slici (Slika 69) izabrane su promenljive INCOME_GROUP i LIFETIME_GIFT_COUNT.



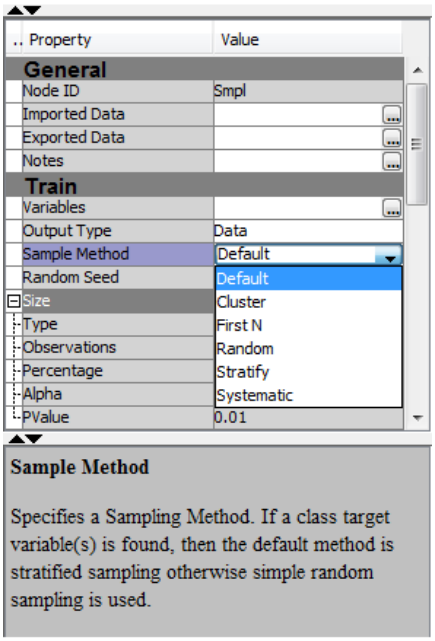
Slika 69. Rezultat istraživanja promenljivih INCOME_GROUP i LIFETIME_GIFT_COUNT

A.1.ii Komponenta *Sample*

Komponenta se koristi za izdvajanje reprezentativnog uzorka s ciljem bržeg i lakšeg istraživanja podataka. Na ovaj način iz cele populacije, koja može imati više miliona opservacija izdvaja se reprezentativni uzorak. Nad ovim uzorkom kreiraju se razne statistike na osnovu kojih možemo upoznati podatke. Komponenta dozvoljava izbor metode za izradu uzorka, kao što su:

- Prvih N (***Sample Method='First N'***). Metoda uzima prvih N opservacija.
- Slučajno izabrani (***Sample Method='Random'***) . Uzima određeni broj opservacija slučajnim uzorkom.
- Stratifikacija (***Sample Method= 'Stratify'***). Stratifikacija kontroliše distribuciju ciljne promenljive u napravljenom uzorku. Zavisno od potreba moguće je definisati različite kriterijume stratifikacije vodeći računa da se ta distribucija ne naruši u odnosu na celu populaciju. To su:
    - ***Proportional***. Proporcija unutar slojeva (eng. *strata*) je ista kao i u celoj populaciji tj. uzima se određeni broj opservacija slučajnim uzorkom pri čemu se vodi računa da u napravljenom uzorku proporcija broja opservacija po ciljnoj promenljivoj bude približna proporociji nad celom populacijom.
    - ***Equal***. Uzima isti broj opservacija za svaki sloj (eng. *stratum*)
    - ***Optimal***. Proporcija unutar slojeva kao i relativna standardna devijacija je ista kao i u celoj populaciji.

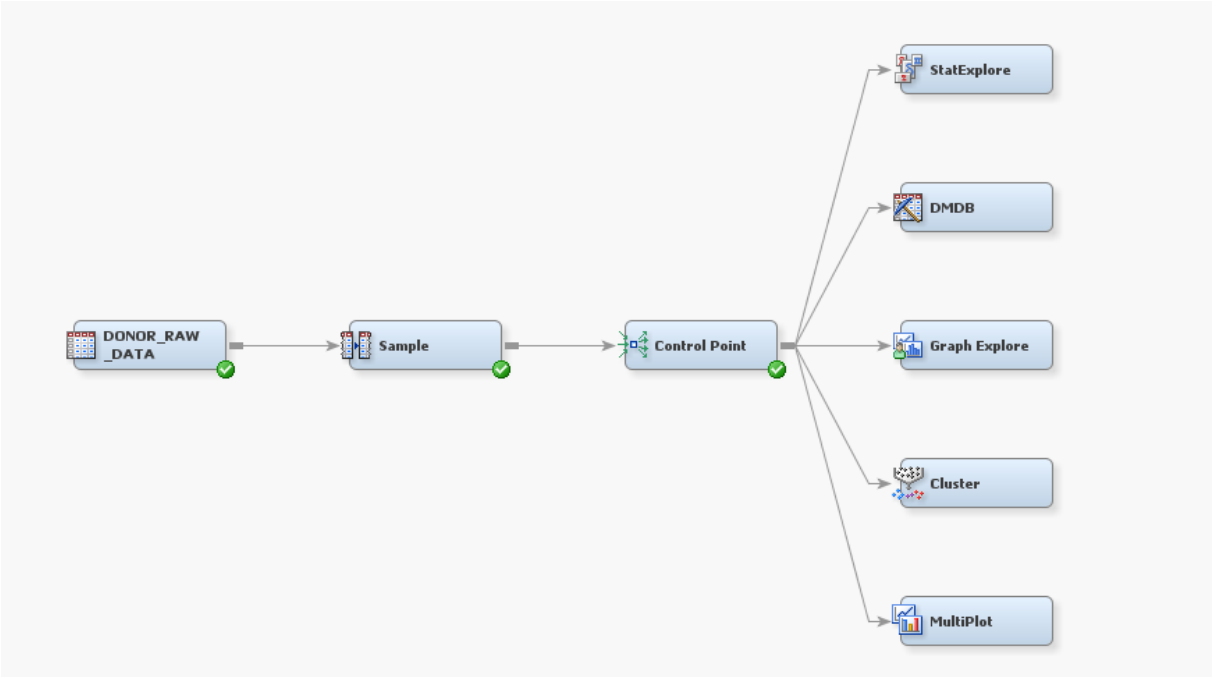
  Stratify(Proportional) je podrazumevani metod u slučaju da je SAS Enterprise Miner našao ciljnu promenljivu.
- Klaster (***Sample Method='Cluster'***) metodom mogu se napraviti klasteri nad populacijom, pa se zavisno od metode klastera (*FirstN, Random, Systematic*) selektuje uzorak. U nekim oblastima istraživanja ova metoda može da obezbedi odgovarajuću efikasnost. Međutim, ovo može dovesti i do gubitka preciznosti procene u poređenju sa neklasterovanim uzorkom iste veličine. Da bi se smanjio ovaj efekat jedinice unutar klastera sami klasteri treba da budu što je moguće više heterogenije prirode.
- *Systematic random sampling* (***Sample Method='Systematic'***) . Metoda bira fragmente kao fiksne intervale kroz populaciju ili sloj (*stratum*) ako se radi o stratifikaciji ali posle slučajnog starta. Frakcioni interval se obezbeđuje specificiranom veličinom uzorka. Interval je jednak N/n odnosno N_k/n_k za stratifikaciju. Verovatnoća izbora je jednaka n/N odnonsno n_k/N_k ako se radi o stratifikaciji. Ova metoda predstavlja implicitnu stratifikaciju. U slučaju da postoji ciljna promenljiva tada je metoda koristi kao sloj.



Slika 70. Podešavanje komponente Sample

Podešavanje veličina uzorka radi se preko opcija ***Type, Observations, Percentage, Alpha*** i ***PValue*** u odeljku ***Size*** (Slika 70)

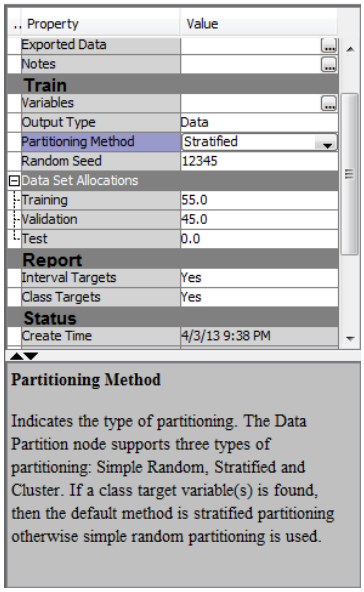
Veličinu uzorka moguće je odrediti procentualno u odnosu na veličinu ABT (***Size.Type='Percentage'***) ili apsolutno po broju opservacija (***Size.Type='Number of Obesrvations***). Moguće je dozvoliti da sama komponenta izračuna veličinu uzorka (***Size.Type='Computed'***) pri čemu se definiše ***Alpha*** i ***P-Value***.



Slika 71. Komponente *Input data* i *Sampe u procesu* prelimenarnog istraživanja podataka

A.1.iii Komponenta *Data Partition*

Ova komponenta je jedna od najvažnijih u modelovanju. Koristi se za deljenje populacije na dva odnosno tri dela. To su: uzorak za razvoj modela (eng. *traning*), uzorak za proveru modela (eng. *model validation*) i uzorak za testiranje modela (eng. *test*). Ovaj uzorak može biti proizvoljne veličine, ali je obično u odnosu 40:30:30, odnosno 60:40:0 ako ne želimo da imamo testni uzorak. Izbor opservacija se radi slučajnim uzorkom na sličan način kao i kod komponente *Sample*.



Slika 72. Osobine komponente *Data Partition*

Izbor metode je sličan kao i kod komponente *Sampe* samo što se ovde kreiraju disjunktni skupovi uzoraka u zadatom odnosu.

.

A.1.iv Ostale komponete koje se ređe koriste

File Import koristi se za učitavanje datoteka (obično tekstualnih) dobijenih iz eksternih izvora. Npr. ako je ABT na nekom RDBMS, a kompanija nije kupila SAS Enterprise BI rešenje tj. nema metadata server i mogućnost povezivanja sa RDBMS već ima samo SAS EM kao „*single instalation*“, tada je moguće tabelu učitati u datoteku iz RDBMS, a zatim datoteku koristiti kao izvor za istraživanje podataka.

Filter - izdvaja slogove iz već pripremljenog ABT

Append - spajanje dva uzorka. Podaci su obično pripremljeni i ova komponenta se retko koristi.

Time Series – omogućava čišćenje i agregaciju transakcionih datoteka po zadatom intervalu koristeći klasične analize vremenske serije. Koristi se za izradu uzorka iz transakcionih datoteka.

A.2 Upoznavanje sa podacima, istraživanje podataka - *Explore*

U ovom poglavlju biće opisane komponente koje se koriste u procesu upoznavanja sa podacima (eng. *Explore* u S**E**MMA pristupu). Ovde će biti opisane samo komponente koje su korišćene prilikom razvoja modela u ovom radu.

Prve tri komponente (***DMDB***, ***Graph Explore*** i ***Multi Plot***) se koriste za upoznavanje podataka, dok se preostala tri alata (***Stat Explore***, ***Variable Clustering*** i ***Varijable Selection***) koriste za izbor statistički značajnih (korisnih) promenljivih.

Ovi alati nam pomažu da bolje upoznamo podatke i uočimo:

- nedostajuće vrednosti za neke promenljive,
- rasipanje, repove, i sl.
- korelaciju sa ciljnom promenljivom,
- statistički značajne (korisne) promenljive.

Na osnovu rezultata istraživanja možemo da iniciramo:

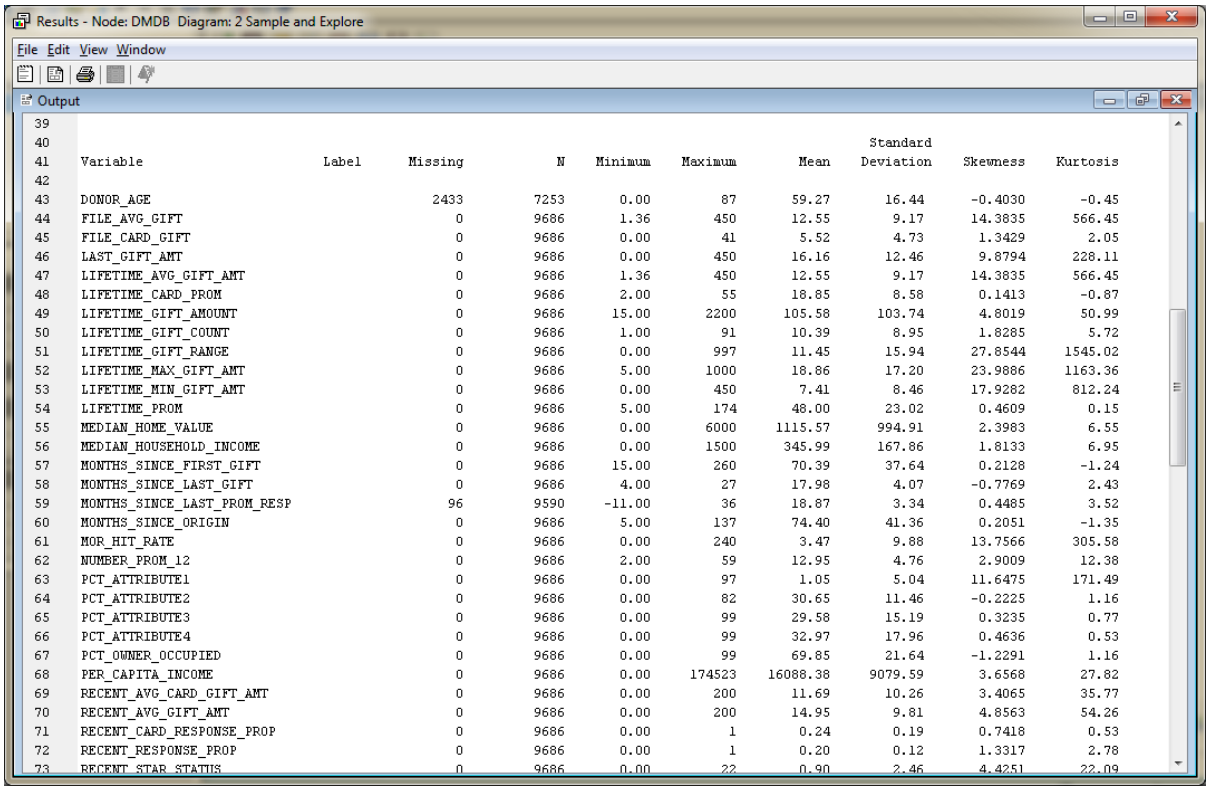
- zamenu nedostajućih vrednosti po izabranom algoritmu
- umanjenje efekta rasipanja i sabijanje repova – transformacija promenljivih (npr. logaritmovanje)
- redukovanje broja promenljivih na samo statistički značajne/korisne promenljive

Neki od alata zahtevaju pročišćene podatke, pa je zbog toga neophodno prvo koristiti neke alate iz poglavlja „Modifikovanje podataka – *Modify*“ ovog dodatka. Inače, koraci *Explore* i *Modify* se prepliću u S**EM**MA pristupu.

U poslednjem poglavlju navedeni su ostali alati koji se ređe koriste ili se uopšte ne koriste u razvoju modela zasnovanih na verovatnoći (skoru).

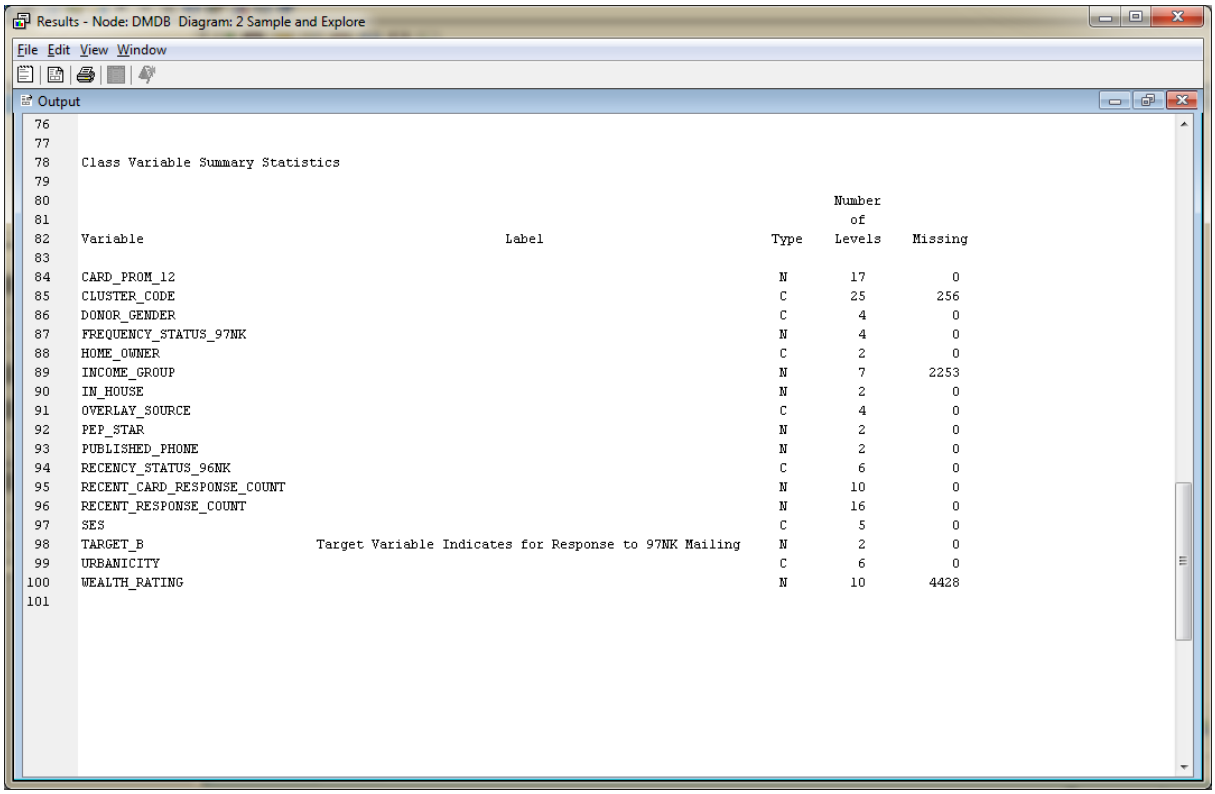
A.2.i Komponenta *DMDB*

Komponenta računa osnovne statistike za izabrane promenljive i rezultat smešta u tekstualnu datoteku u vidu izveštaja spremnog za štampu. Za računanje koristi se SAS DMDB procedura koja sve statistike računa u jednom prolazu. Statistike se prikazuju u dve tabele, posebno za intervalne i posebno za klasifikacione promenljive.



Variable	Label	Missing	N	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
DONOR_AGE		2433	7253	0.00	87	59.27	16.44	-0.4030	-0.45
FILE_AVG_GIFT		0	9686	1.36	450	12.55	9.17	14.3835	566.45
FILE_CARD_GIFT		0	9686	0.00	41	5.52	4.73	1.3429	2.05
LAST_GIFT_AMT		0	9686	0.00	450	16.16	12.46	9.8794	228.11
LIFETIME_AVG_GIFT_AMT		0	9686	1.36	450	12.55	9.17	14.3835	566.45
LIFETIME_CARD_PROM		0	9686	2.00	55	18.85	8.58	0.1413	-0.87
LIFETIME_GIFT_AMOUNT		0	9686	15.00	2200	105.58	103.74	4.8019	50.99
LIFETIME_GIFT_COUNT		0	9686	1.00	91	10.39	8.95	1.8285	5.72
LIFETIME_GIFT_RANGE		0	9686	0.00	997	11.45	15.94	27.8544	1545.02
LIFETIME_MAX_GIFT_AMT		0	9686	5.00	1000	18.86	17.20	23.9886	1163.36
LIFETIME_MIN_GIFT_AMT		0	9686	0.00	450	7.41	8.46	17.9282	812.24
LIFETIME_PROM		0	9686	5.00	174	48.00	23.02	0.4609	0.15
MEDIAN_HOME_VALUE		0	9686	0.00	6000	1115.57	994.91	2.3983	6.55
MEDIAN_HOUSEHOLD_INCOME		0	9686	0.00	1500	345.99	167.86	1.8133	6.95
MONTHS_SINCE_FIRST_GIFT		0	9686	15.00	260	70.39	37.64	0.2128	-1.24
MONTHS_SINCE_LAST_GIFT		0	9686	4.00	27	17.98	4.07	-0.7769	2.43
MONTHS_SINCE_LAST_PROM_RESP		96	9590	-11.00	36	18.87	3.34	0.4485	3.52
MONTHS_SINCE_ORIGIN		0	9686	5.00	137	74.40	41.36	0.2051	-1.35
MOR_HIT_RATE		0	9686	0.00	240	3.47	9.88	13.7566	305.58
NUMBER_PROM_12		0	9686	2.00	59	12.95	4.76	2.9009	12.38
PCT_ATTRIBUTE1		0	9686	0.00	97	1.05	5.04	11.6475	171.49
PCT_ATTRIBUTE2		0	9686	0.00	82	30.65	11.46	-0.2225	1.16
PCT_ATTRIBUTE3		0	9686	0.00	99	29.58	15.19	0.3235	0.77
PCT_ATTRIBUTE4		0	9686	0.00	99	32.97	17.96	0.4636	0.53
PCT_OWNER_OCCUPIED		0	9686	0.00	99	69.85	21.64	-1.2291	1.16
PER_CAPITA_INCOME		0	9686	0.00	174523	16088.38	9079.59	3.6568	27.82
RECENT_AVG_CARD_GIFT_AMT		0	9686	0.00	200	11.69	10.26	3.4065	35.77
RECENT_AVG_GIFT_AMT		0	9686	0.00	200	14.95	9.81	4.8563	54.26
RECENT_CARD_RESPONSE_PROP		0	9686	0.00	1	0.24	0.19	0.7418	0.53
RECENT_RESPONSE_PROP		0	9686	0.00	1	0.20	0.12	1.3317	2.78
RECENT_STAR_STATUS		0	9686	0.00	22	0.90	2.46	4.4251	22.09

Slika 73. Statistike kontinualnih promenljivih dobijene komponentom DMDB



Class Variable Summary Statistics

Variable	Label	Type	Number of Levels	Missing
CARD_PROM_12		N	17	0
CLUSTER_CODE		C	25	256
DONOR_GENDER		C	4	0
FREQUENCY_STATUS_97NK		N	4	0
HOME_OWNER		C	2	0
INCOME_GROUP		N	7	2253
IN_HOUSE		N	2	0
OVERLAY_SOURCE		C	4	0
PEP_STAR		N	2	0
PUBLISHED_PHONE		N	2	0
RECENCY_STATUS_96NK		C	6	0
RECENT_CARD_RESPONSE_COUNT		N	10	0
RECENT_RESPONSE_COUNT		N	16	0
SES		C	5	0
TARGET_B	Target Variable Indicates for Response to 97NK Mailing	N	2	0
URBANICITY		C	6	0
WEALTH_RATING		N	10	4428

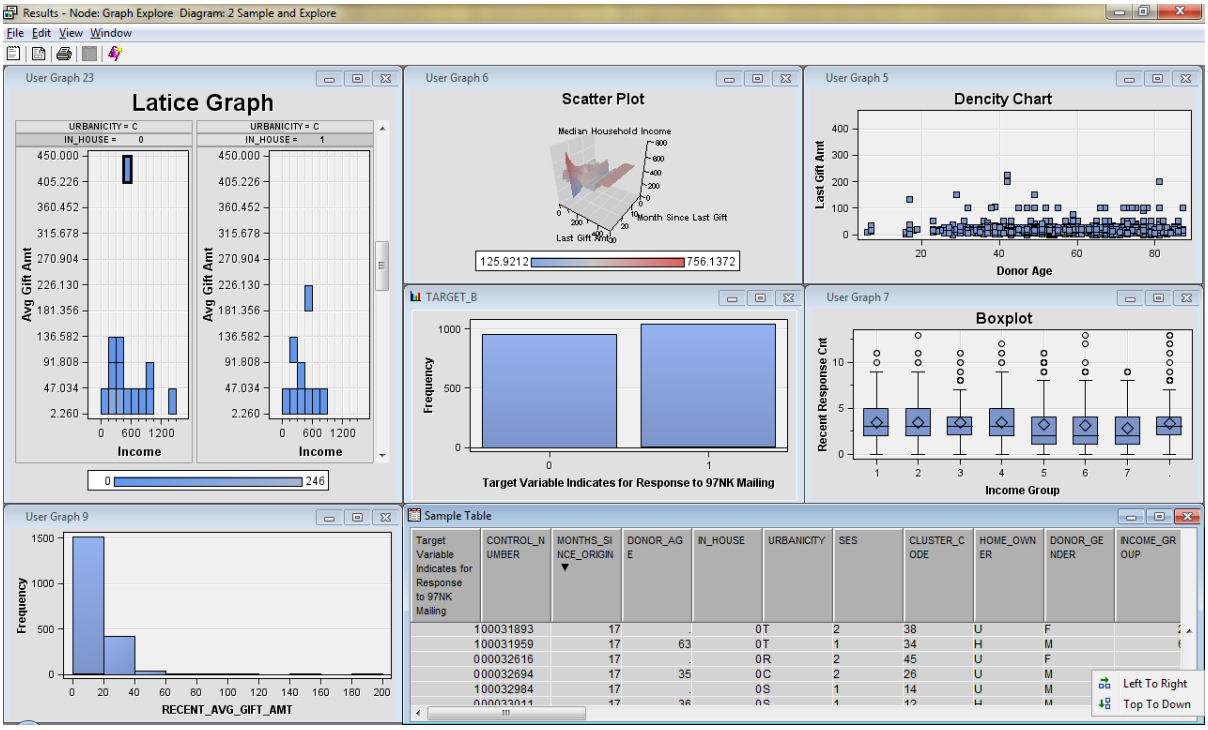
Slika 74. Statistike nominalnih promenljivih dobijene komponentom DMDB

A.2.ii Komponenta *Graph Explore*

Graph Explore je alat gde se rezultati istraživanja mogu umotati u intuitivnu grafičku vizuelizaciju. Ovde se mogu analizirati distribucije jedne ili više promenljivih, napraviti *scatter* i *box* grafikon, *constellation* i 3D grafikoni.

Rezultati analiza se mogu prikazati preko sledećih grafikona: *Scatter, Line, Histogram, Density, Box, Tables, Matrix, Lattice, Parallel Axis, Constellation, 3D Charts, Contour, Bar, Pie, Needle, Vector, Band.*

Na slici (Slika 75) prikazani su sledeći grafikoni *Latice Graph, Scater Plot, Density Chart, Frequency Bar, Boxplot* kao i detaljna tabela sa uzorcima.



Slika 75. Primer korišćenja Graph Explore komponente

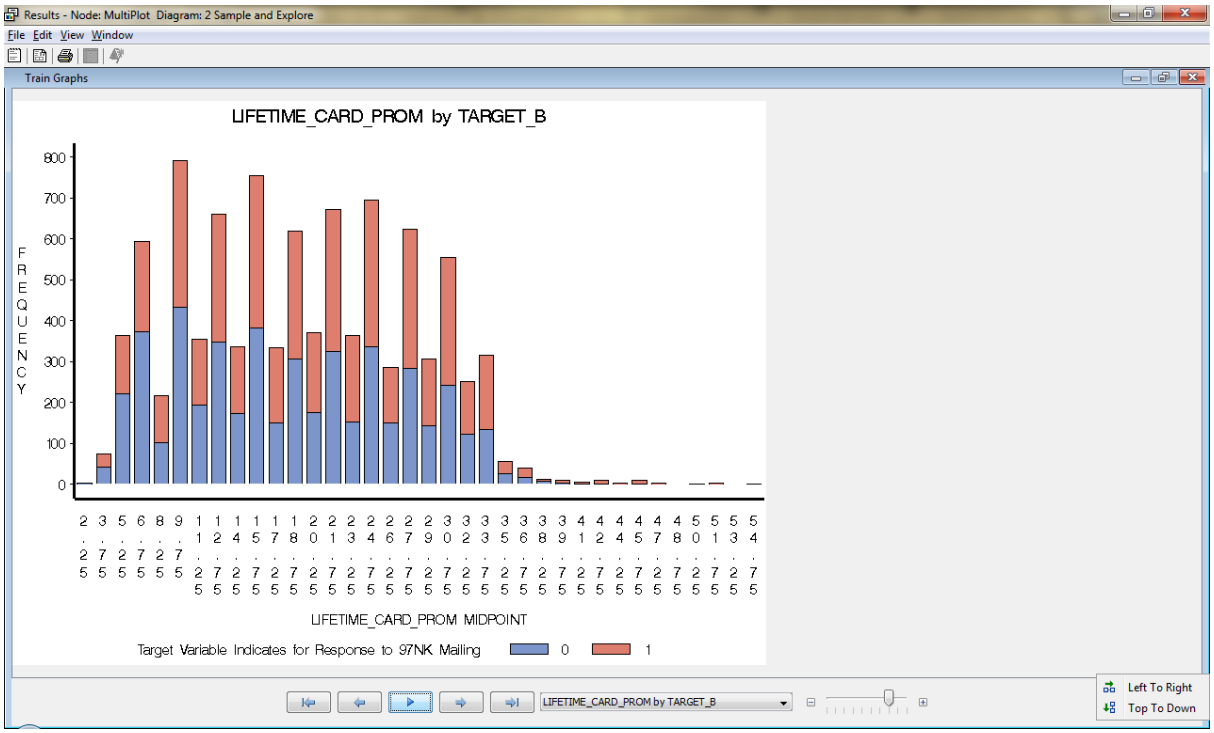
A.2.iii Komponenta *Multi Plot*

MultiPlot komponenta se koristi isključivo u procesu upoznavanja sa podacima. Komponenta ima mogućnost da prikaže distribuciju i relaciju sa ciljnom promenljivom za veliki broj promenljivih i to kroz niz slajdova.

Komponeta može kreirati sledeće grafikone:

- Bar Charts:
  - Histogram za svaku nezavisnu i ciljnu pormenljivu
  - *Bar chart* nezavisne promenljive u odnosu na ciljnu promenljivu
  - *Bar chart* nezavisne promenljiva grupisana po ciljnoj promenljivoj
- Scatter Plots:
  - *Plot* intervalna nezavisna promenljiva u odnosu na ciljnu promenljivu

- *Plot* klasifikaciona nezavisna promenljiva u odnosu na ciljnu promenljivu



Slika 76. Distribucija promenljive LIFETIME_CARD_PROM prikazane odvojeno za *event* i *nonevent* populaciju

A.2.iv Komponenta *Stat Explore*

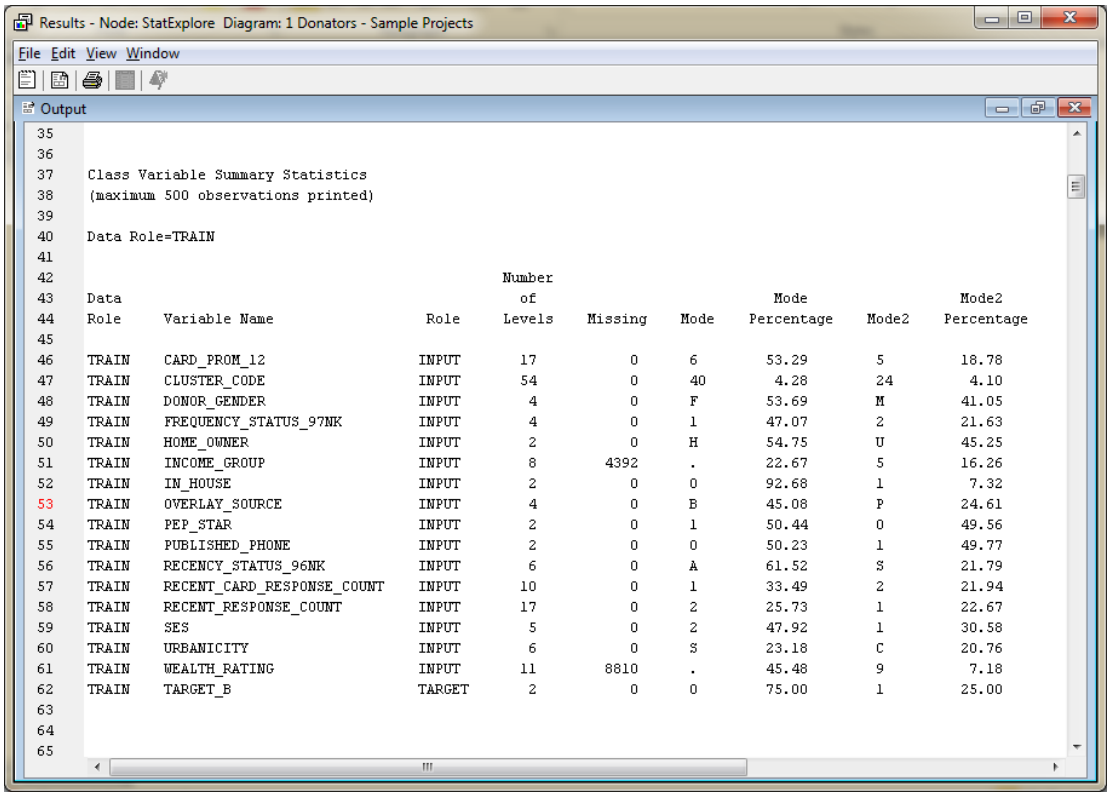
Komponenta *StatExplore* pravi sumarne statistike kao i statistike korelacije. StatExplore komponenta se koristi za:

- Iznor promenljivih za analize, profilisanje klastera i prediktivne modele.
- Izračunavanje standardnih statistika pojedinačanih promenjivih.
- Izračunavanje standardnih statistika u odnosu na ciljnu promenjivu.
- Izračunavanje korelacione statistike intervalnih promenljivih u odnosu na ciljnu promenljivu.

Rezultat ovih istraživanja je sledeći:

- Smanjenje skupa promenljivih tako što bi se manje značajne promenjive odbacile (promena uloge u metapodacima na "*rejected*").
- Na osnovu statistika može se sugerisati transformacija postojećih promenljivih.

Sumarne statistike klasifikacionih promenljivih su prikazane na slici (Slika 77).



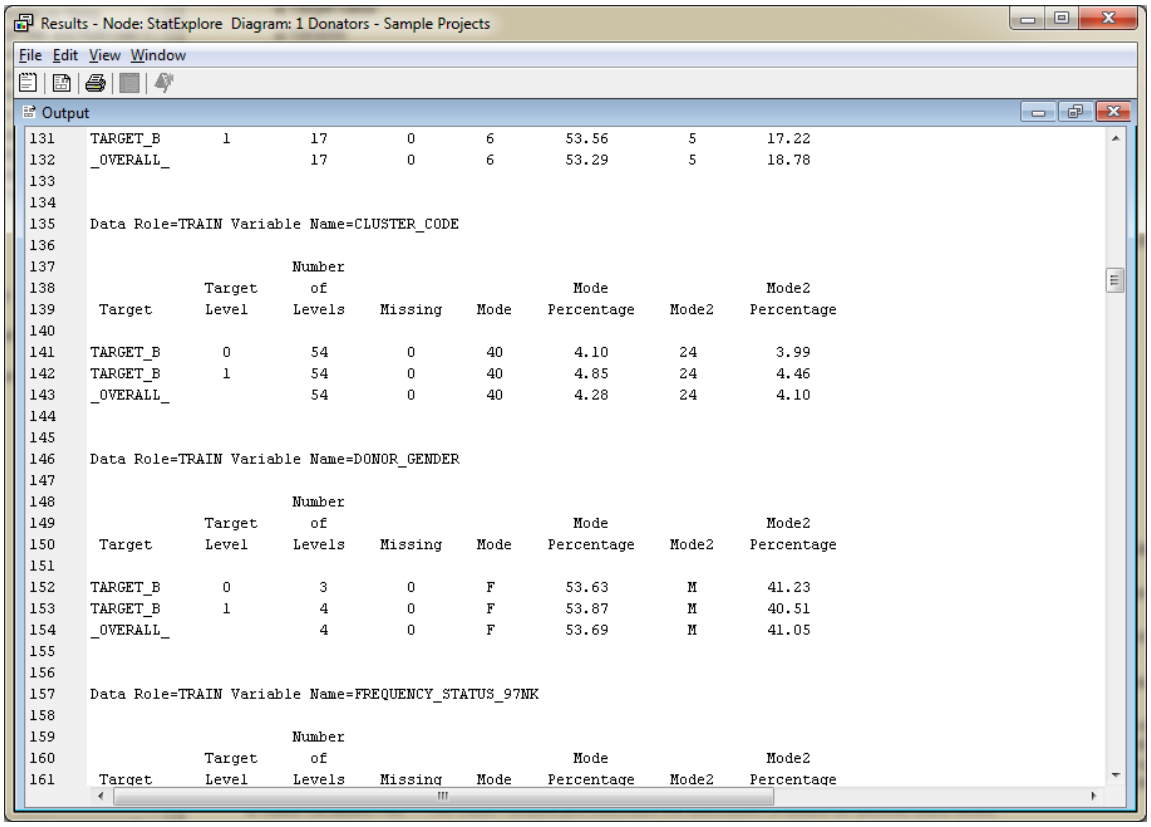
Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	CARD_PROM_12	INPUT	17	0	6	53.29	5	18.78
TRAIN	CLUSTER_CODE	INPUT	54	0	40	4.28	24	4.10
TRAIN	DONOR_GENDER	INPUT	4	0	F	53.69	M	41.05
TRAIN	FREQUENCY_STATUS_97NK	INPUT	4	0	1	47.07	2	21.63
TRAIN	HOME_OWNER	INPUT	2	0	H	54.75	U	45.25
TRAIN	INCOME_GROUP	INPUT	8	4392	.	22.67	5	16.26
TRAIN	IN_HOUSE	INPUT	2	0	0	92.68	1	7.32
TRAIN	OVERLAY_SOURCE	INPUT	4	0	B	45.08	P	24.61
TRAIN	PEP_STAR	INPUT	2	0	1	50.44	0	49.56
TRAIN	PUBLISHED_PHONE	INPUT	2	0	0	50.23	1	49.77
TRAIN	RECENCY_STATUS_96NK	INPUT	6	0	A	61.52	S	21.79
TRAIN	RECENT_CARD_RESPONSE_COUNT	INPUT	10	0	1	33.49	2	21.94
TRAIN	RECENT_RESPONSE_COUNT	INPUT	17	0	2	25.73	1	22.67
TRAIN	SES	INPUT	5	0	2	47.92	1	30.58
TRAIN	URBANICITY	INPUT	6	0	S	23.18	C	20.76
TRAIN	WEALTH_RATING	INPUT	11	8810	.	45.48	9	7.18
TRAIN	TARGET_B	TARGET	2	0	0	75.00	1	25.00

Slika 77. Statistike nominalnih promenljivih

Statistike nominalnih promenljivih u odnosu na ciljnu promenljivu su prikazane na slici (Slika 78).



TARGET_B	1	17	0	6	53.56	5	17.22
OVERALL		17	0	6	53.29	5	18.78

Data Role=TRAIN Variable Name=CLUSTER_CODE

Target	Target Level	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TARGET_B	0	54	0	40	4.10	24	3.99
TARGET_B	1	54	0	40	4.85	24	4.46
OVERALL		54	0	40	4.28	24	4.10

Data Role=TRAIN Variable Name=DONOR_GENDER

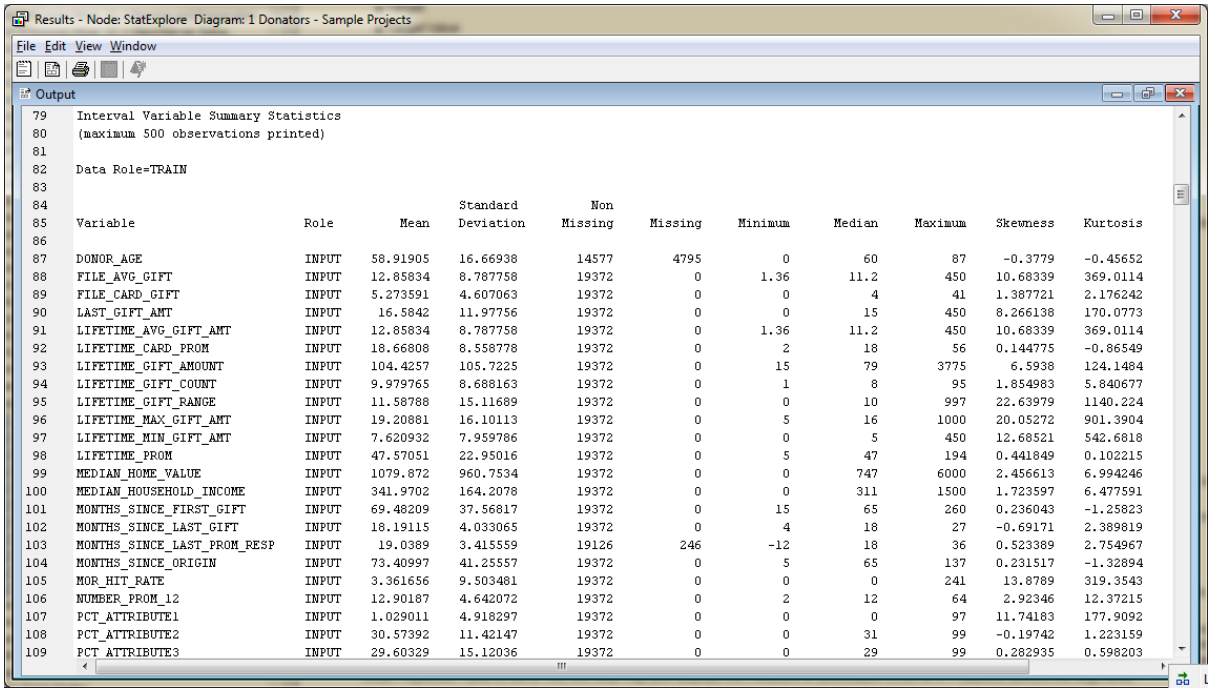
Target	Target Level	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TARGET_B	0	3	0	F	53.63	M	41.23
TARGET_B	1	4	0	F	53.87	M	40.51
OVERALL		4	0	F	53.69	M	41.05

Data Role=TRAIN Variable Name=FREQUENCY_STATUS_97NK

Target	Target Level	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage

Slika 78. Statistike nominalnih promenljivih u odnosu na ciljnu promenljivu

Sumarne statistike intervalnih promenljivih su prikazane na slici (Slika 79).



Results - Node: StatExplore Diagram: 1 Donators - Sample Projects

File Edit View Window

Output

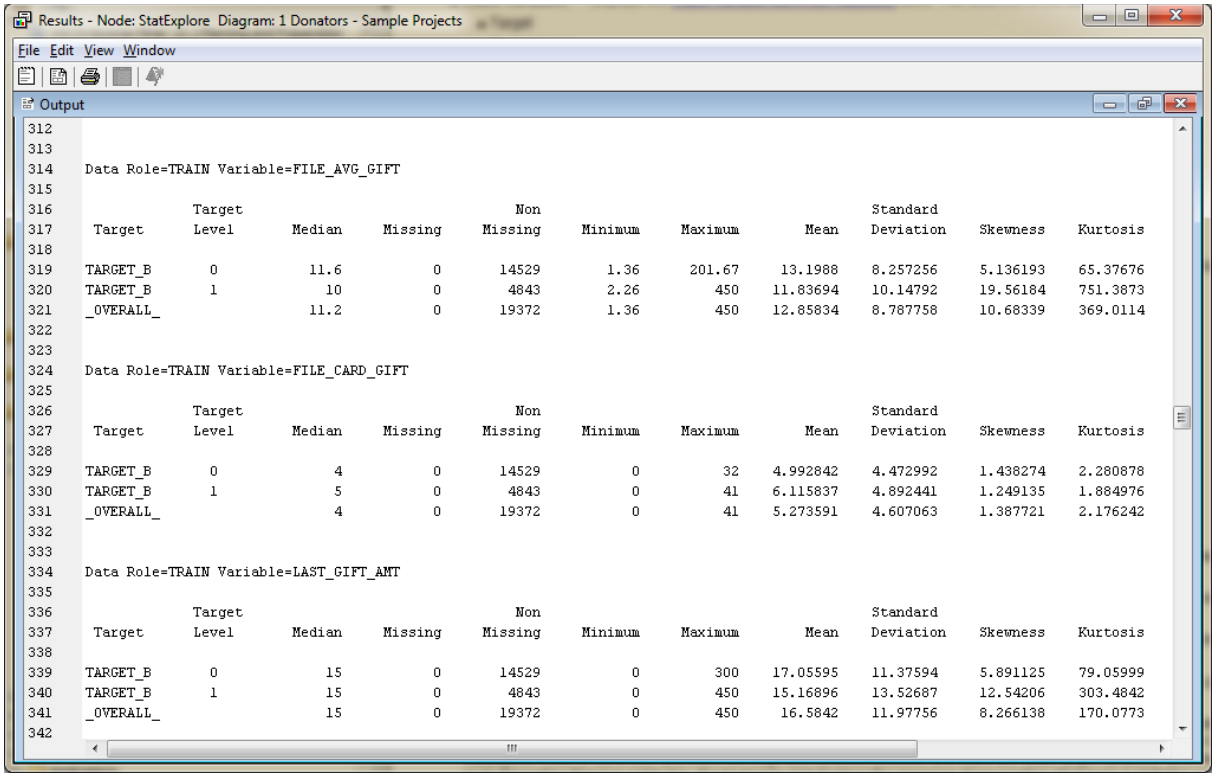
Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
DONOR_AGE	INPUT	58.91905	16.66938	14577	4795	0	60	87	-0.3779	-0.45652
FILE_AVG_GIFT	INPUT	12.85834	8.787758	19372	0	1.36	11.2	450	10.68339	369.0114
FILE_CARD_GIFT	INPUT	5.273591	4.607063	19372	0	0	4	41	1.387721	2.176242
LAST_GIFT_AMT	INPUT	16.5842	11.97756	19372	0	0	15	450	8.266138	170.0773
LIFETIME_AVG_GIFT_AMT	INPUT	12.85834	8.787758	19372	0	1.36	11.2	450	10.68339	369.0114
LIFETIME_CARD_PROM	INPUT	18.66808	8.558778	19372	0	2	18	56	0.144775	-0.86549
LIFETIME_GIFT_AMOUNT	INPUT	104.4257	105.7225	19372	0	15	79	3775	6.5938	124.1484
LIFETIME_GIFT_COUNT	INPUT	9.979765	8.688163	19372	0	1	8	95	1.854983	5.840677
LIFETIME_GIFT_RANGE	INPUT	11.58788	15.11689	19372	0	0	10	997	22.63979	1140.224
LIFETIME_MAX_GIFT_AMT	INPUT	19.20881	16.10113	19372	0	5	16	1000	20.05272	901.3904
LIFETIME_MIN_GIFT_AMT	INPUT	7.620932	7.959786	19372	0	0	5	450	12.68521	542.6818
LIFETIME_PROM	INPUT	47.57051	22.95016	19372	0	5	47	194	0.441849	0.102215
MEDIAN_HOME_VALUE	INPUT	1079.872	960.7534	19372	0	0	747	6000	2.456613	6.994246
MEDIAN_HOUSEHOLD_INCOME	INPUT	341.9702	164.2078	19372	0	0	311	1500	1.723597	6.477591
MONTHS_SINCE_FIRST_GIFT	INPUT	69.48209	37.56817	19372	0	15	65	260	0.236043	-1.25823
MONTHS_SINCE_LAST_GIFT	INPUT	18.19115	4.033065	19372	0	4	18	27	-0.69171	2.389819
MONTHS_SINCE_LAST_PROM_RESP	INPUT	19.0389	3.415559	19126	246	-12	18	36	0.523389	2.754967
MONTHS_SINCE_ORIGIN	INPUT	73.40997	41.25557	19372	0	5	65	137	0.231517	-1.32894
MOR_HIT_RATE	INPUT	3.361656	9.503481	19372	0	0	0	241	13.8789	319.3543
NUMBER_PROM_12	INPUT	12.90187	4.642072	19372	0	2	12	64	2.92346	12.37215
PCT_ATTRIBUTE1	INPUT	1.029011	4.918297	19372	0	0	0	97	11.74183	177.9092
PCT_ATTRIBUTE2	INPUT	30.57392	11.42147	19372	0	0	31	99	-0.19742	1.223159
PCT_ATTRIBUTE3	INPUT	29.60329	15.12036	19372	0	0	29	99	0.282935	0.598203

Slika 79. Statistike intervalnih promenljivih

Statistike intervalnih promenljivih u odnosu na ciljnu promenljivu su prikazane na slici (Slika 80).



Results - Node: StatExplore Diagram: 1 Donators - Sample Projects

File Edit View Window

Output

Data Role=TRAIN Variable=FILE_AVG_GIFT

Target	Target Level	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
TARGET_B	0	11.6	0	14529	1.36	201.67	13.1988	8.257256	5.136193	65.37676
TARGET_B	1	10	0	4843	2.26	450	11.83694	10.14792	19.56184	751.3873
OVERALL		11.2	0	19372	1.36	450	12.85834	8.787758	10.68339	369.0114

Data Role=TRAIN Variable=FILE_CARD_GIFT

Target	Target Level	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
TARGET_B	0	4	0	14529	0	32	4.992842	4.472992	1.438274	2.280878
TARGET_B	1	5	0	4843	0	41	6.115837	4.892441	1.249135	1.884976
OVERALL		4	0	19372	0	41	5.273591	4.607063	1.387721	2.176242

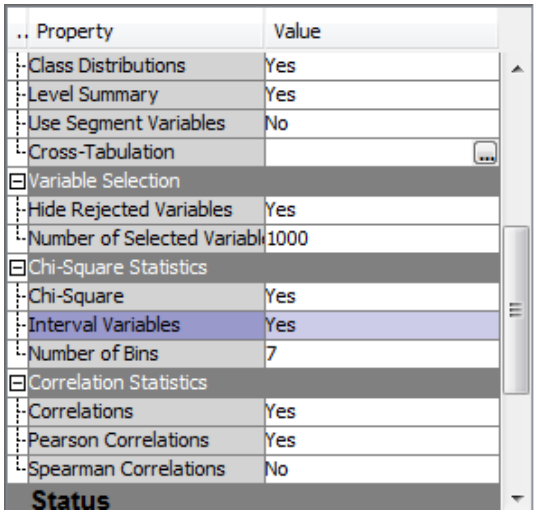
Data Role=TRAIN Variable=LAST_GIFT_AMT

Target	Target Level	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
TARGET_B	0	15	0	14529	0	300	17.05595	11.37594	5.891125	79.05999
TARGET_B	1	15	0	4843	0	450	15.16896	13.52687	12.54206	303.4842
OVERALL		15	0	19372	0	450	16.5842	11.97756	8.266138	170.0773

Slika 80. Statistike intervalnih promenljivih u odnosu na ciljnu promenljivu

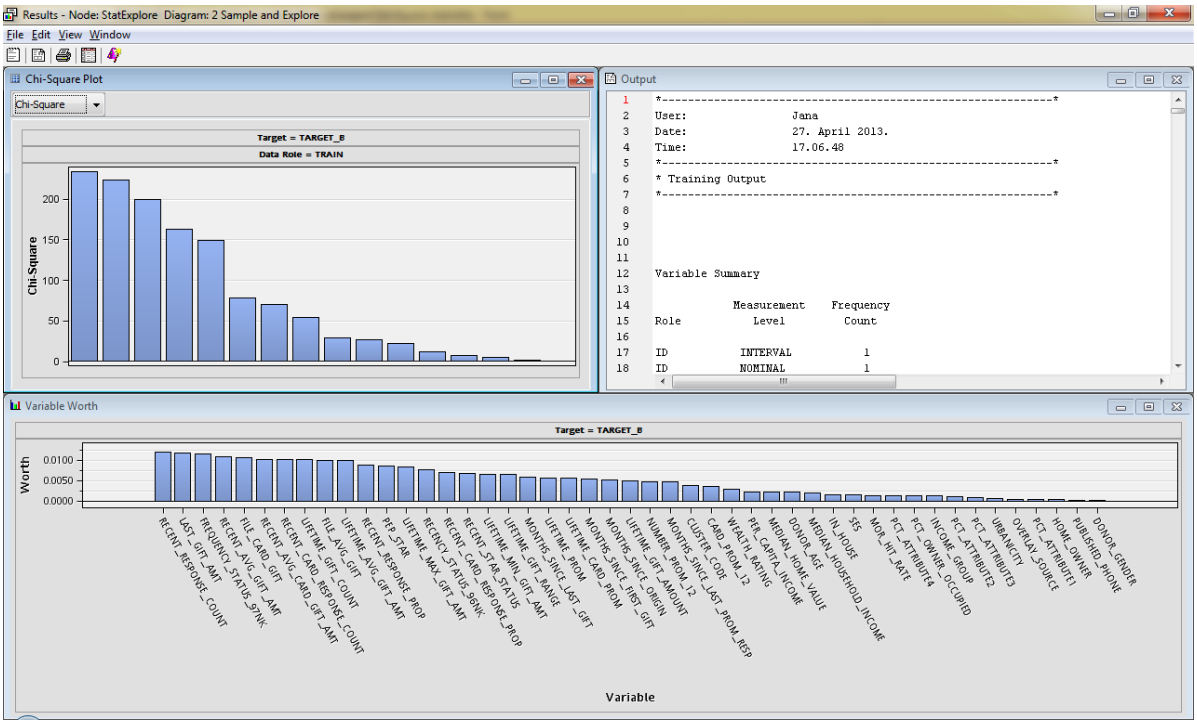
Pored osnovnih sumarnih statistika kreiraju se i sledeće korelacione statistike:

- ***Variable Worth*** pri čemu se važnost promenljive računa pomoći „*Gini split worth*“ statistike generisane iz drveta odlučivanja dubine 1.
- ***Cramer's V***
- ***Chi-Square*** – hi kvadrat statistika koristi se samo za nominalne promenljive. Za intervalne varijable neophodno je postaviti opciju „*Chi-Square Statistic.Interval Variables*“ na *'Yes'*. U tom slučaju SAS EM će automatski uraditi grupisanje (eng. *binning*) pri čemu se broj grupa definiše kroz opciju „*Bins*“ (Slika 81).

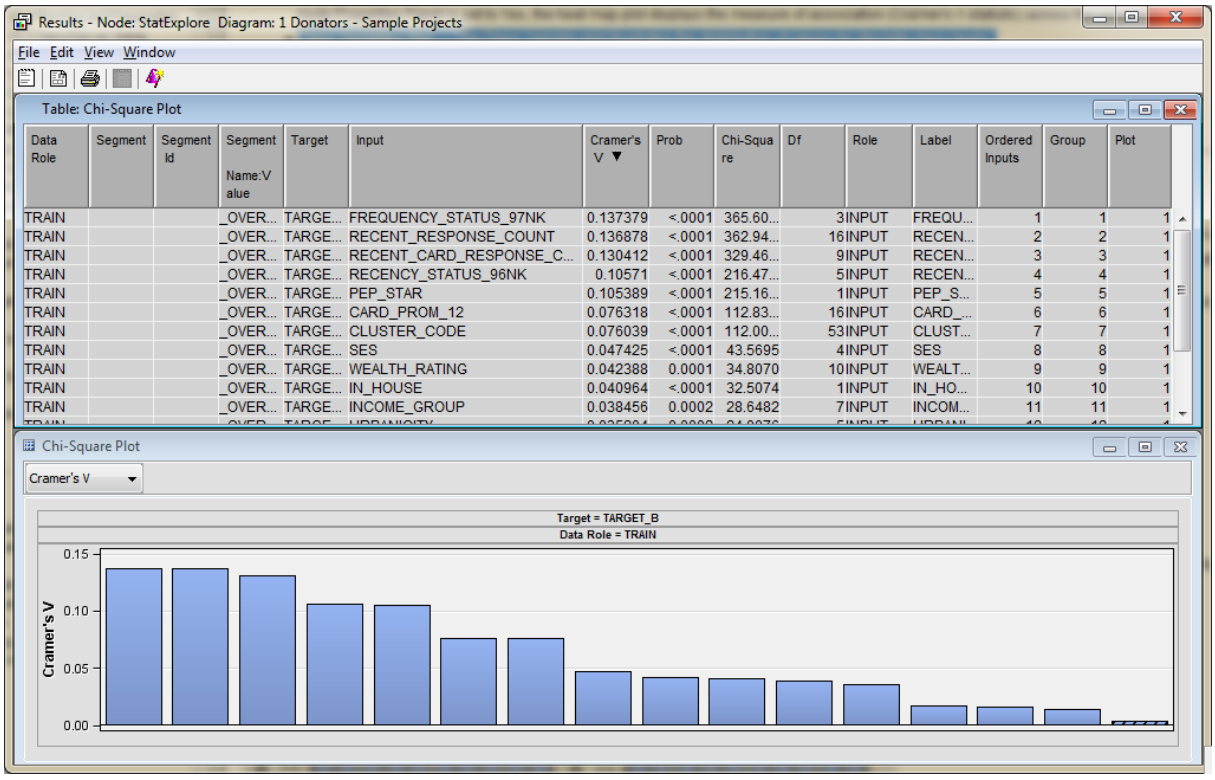


Slika 81. Osobine *Stat Explore* komponente

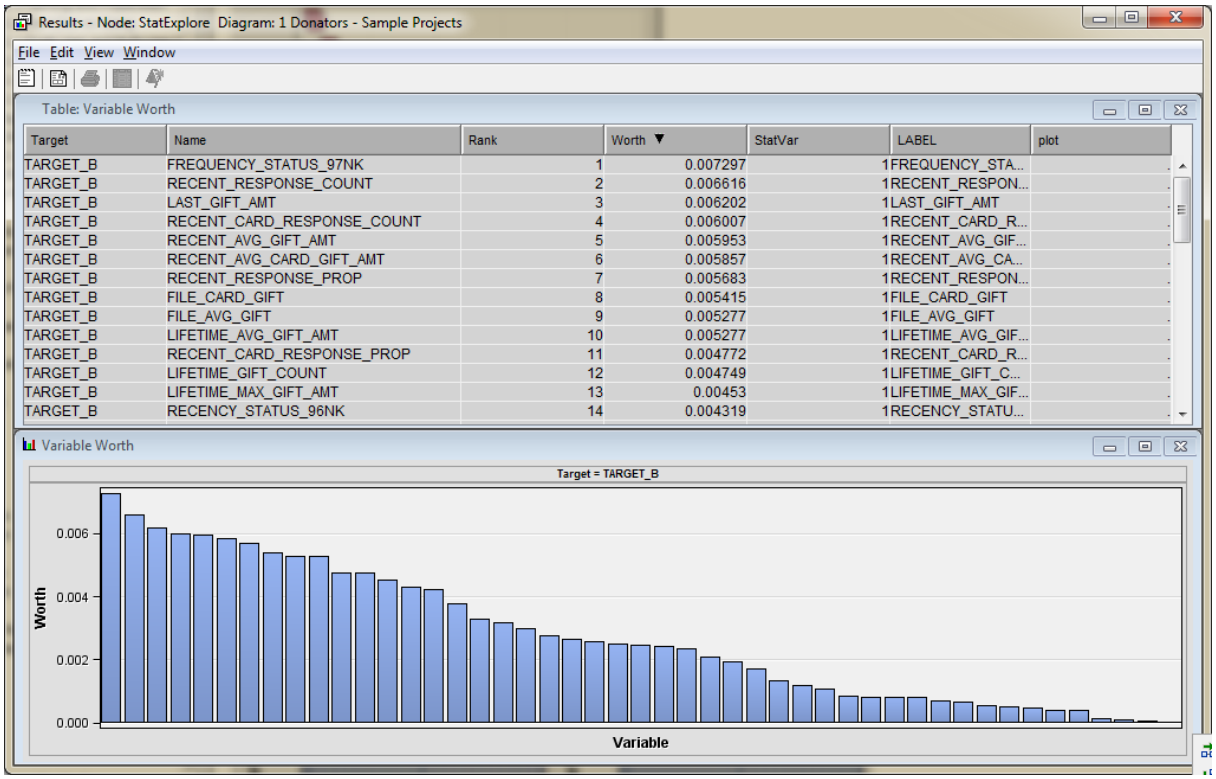
Korelacione statistike u odnosu na ciljnu promeljivu prikazane su tabelarno i kroz grafikone (Slika 82, Slika 83, Slika 84).



Slika 82. Hi-kvardrat statistike u komponenti *Stat Explore*



Slika 83. *Cramer's V* statistike korelacije promenljivih u odnosu na ciljnu promenljivu



Slika 84. Grafički i tabelarni prikaz „*Variable worth*“

A.2.v Komponenta *Varijable Clustering*

Komponenta predstavlja veoma koristan alat za izbor najboljih promenljivih ili klastera za dalje analize. Izbor klastera umesto nekoliko desetina promenljivih može značajno redukovati broj promenljivih. Klasteri nam obezbeđuju hetoregenost samih promenljivih što može biti veoma značajno.

Variable clustering raspoređuje numeričke promenljive u nespojive i/ili hijerarhijske klastere. Rezultat klasterovanja može se opisati kao linearna kombinacija promenljivih. Linearna kombinacija promenljivih je prva komponenta klastera (eng. *the first principal component of the cluster*) zvana klaster komponenta (eng. *cluster component*). Klaster komponenta obezbeđuje skor za svaki klaster. Skor se računa kao *weighted average of the variables* opisana kao varijanca.

$$\operatorname{Var}\left(\sum_{i}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \operatorname{Var}(X_i) + 2\sum_{1\le i} \sum_{<j\le n} a_i a_j \operatorname{Cov}(X_i, X_j)$$

gde je a_i=1/n za svako i=1,2,…n,

Postoje dva načina za analizu klaster komponente. Jedan koristi korelaciju između promenljivih, a drugi kovarijansu. Ako se koristi korelacija, sve promenljive se tretiraju kao podjednako važne. Ako se koristi kovarijansa, promenljive sa najvećom varijansom imaju veću važnost.

Variable Clustering algoritam

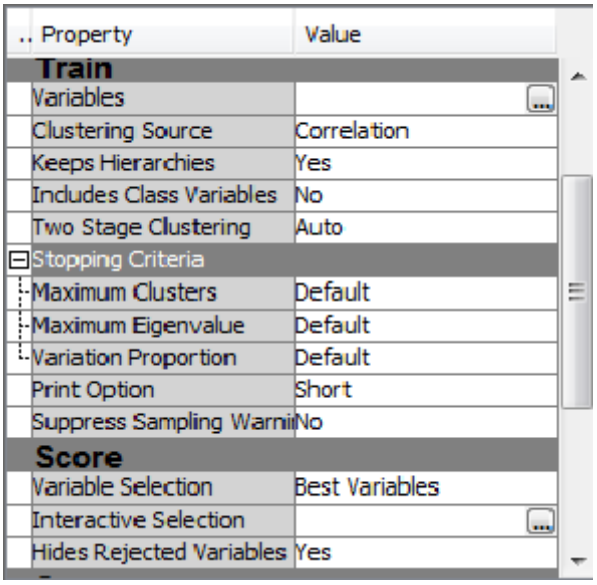
Algoritam je iterativan. U prvom koraku sve promenljive se pridružuju jednom klasteru.

Dalje se primenljuju sledeći koraci:

1. Zavisno od opcija izabrani klaster se deli u slučaju da ima najmanji procenat varijacije (opcija ***Variation Proportion***) ili najveću sopstvenu vrednost (eng. *eigenvalue*) pridružen drugoj klaster komponenti (opcija ***Maximum Eigenvalue***).
2. Izabrani klaster se dalje cepa nalaženjem prve dve klaster komponente, korišćenjem *„orthoblique“* rotacije (eng. *raw quartimax rotation on the eigenvectors; Harris and Kaiser, 1964*)
3. Iterativna reinicijalizacija promenljivih u klaster se odvija u dva koraka.
    a. Prvi korak je faza sortiranja (*nearest centroid sorting algorithms, Anderberg (1973)*). U svakoj iteraciji se svakoj promenljivoj dodelju odgovarajući klaster vodeći računa da se promenljivoj dodeli klaster sa najvećom kvadratnom korelacijom.
    b. U drugom koraku uključuje se algoritam provere koji svaku promenljivu proverava da li pridruživanjem drugom klasteru uvećavamo varijancu. Ako je promenljiva reinicijalizovana u ovoj fazi, statistike komponenti ova dva klastera će biti ponovo izračunate. Prva korak je mnogo brži od drugog ali postoji mogućnost da se u prvoj fazi neke promenljive dodele pogrešnom klasteru.

4. Kada je reinicijalizacija promenljivih urađena za svaki novi klaster se primeni korak 1.

Cepanje klastera se stopira kada je dosegnut maksimalni broj klastera (podrazumevana vrednost je broj promenljivih) ili kada svaki klaster zadovoljava kriterijum zaustavljanja specificiran u opcijama ***Variation Proportion*** i/ili ***Maximum Eigenvalue***.



Slika 85. Osobine Variable Clustering komponente

Ovaj algoritam je alternativa metodu najmanjeg kvadrata (eng. *least-squares*) i konvergira veoma brzo. Problem je korak 3b u slučaju velikog broja promenljivih. U slučaju da se koristi podrazumevana metoda inicijalizacije korak 3b veoma retko unapređuje rezultate dobijene pod 3a i obično se završi u nekoliko iteracija.

Korišćenjem hijerarhijskog klasterovanja (opcija ***Keep Hierarchies***) dodatno se uvodi restrikcija da u koraku 3 reinicijalizacija varijabli može da se desi unutar roditelj klastera.

Opcija ***Two Stage Clustering*** dopušta cepanje klastera na dva ili više. Ako se izabere „*Yes*“, klaster se uvek cepa na dva. U slučaju da je izabrana opcija *„No“* klaster se cepa na proizvoljan broj klastera sve dok ne bude zadovoljen kriterijum stopiranja.

U slučaju da je ***Two Stage Clustering***=*Auto* i da u ABT više od 200 promenljivih tada se klaster cepa na najviše *INT(broj promenljivih/100+2)* nova klastera.

NULL vrednosti

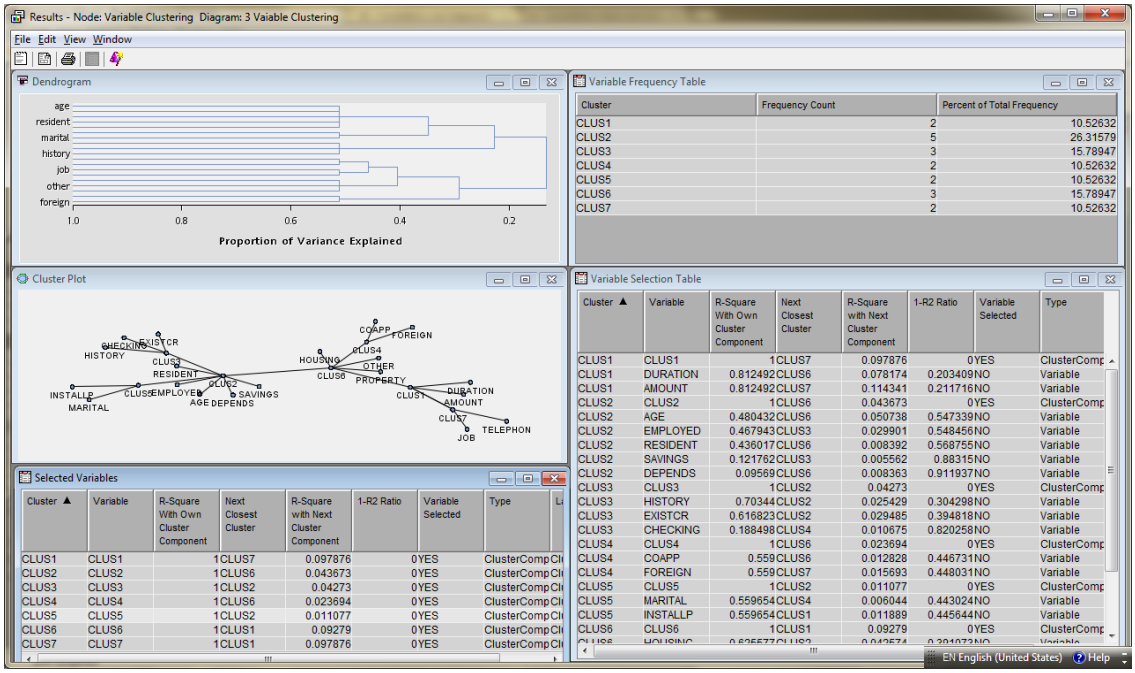
Ako opservacije sadrže nedostajuće vrednosti komponenta isključuje ove opservacije iz dalje analize, jer ne može da izračuna varijansu. U slučaju da imamo značajan broj nedostajućih vrednosti korisno je pre upotrebe komponente zameniti *null* vrednost odgovarajućim vrednostostima pomoću komponente *Variable Replacement* (videti poglavlje „Komponenta *Replacement*“) neposredno pre primene komponente *Variable Clustering*.

Ograničenja

Variable Clustering komponetna je veoma zahtevna kada se koristi nad velikim brojem opservacija i promenljivih. Ona je korisna kada ABT ima manje od 100 promenljivih i 100,000 opservacija. Pokretanje komponente nad velikim skupom podataka može značajno usporiti komponentu. U ovom radu je komponenta radila više od 7 sati (2000 promenljivih sa 80000 opservacija).

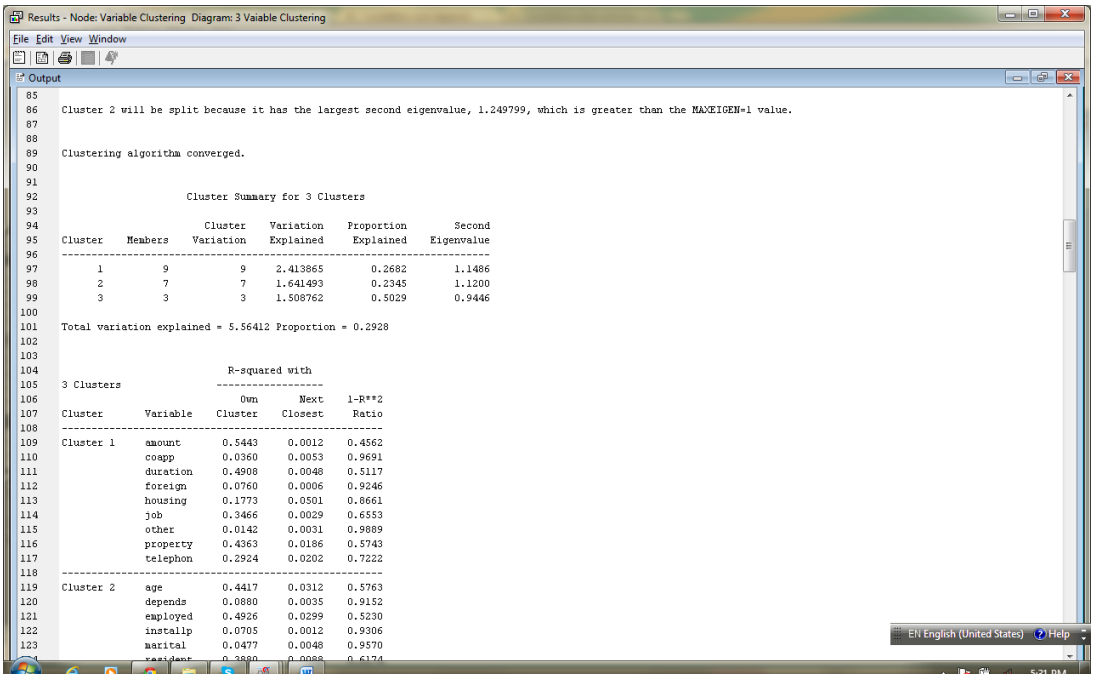
U slučaju da imamo veliki ABT tada (preko 100000 opservacija) moguće je koristiti *Sampe* komponentu radi dobijanje reprezentativnog uzorka nad kojim će se dalje raditi klasterovanje.

Rezultat komponente ***Variable Clustering*** je na slici (Slika 86).



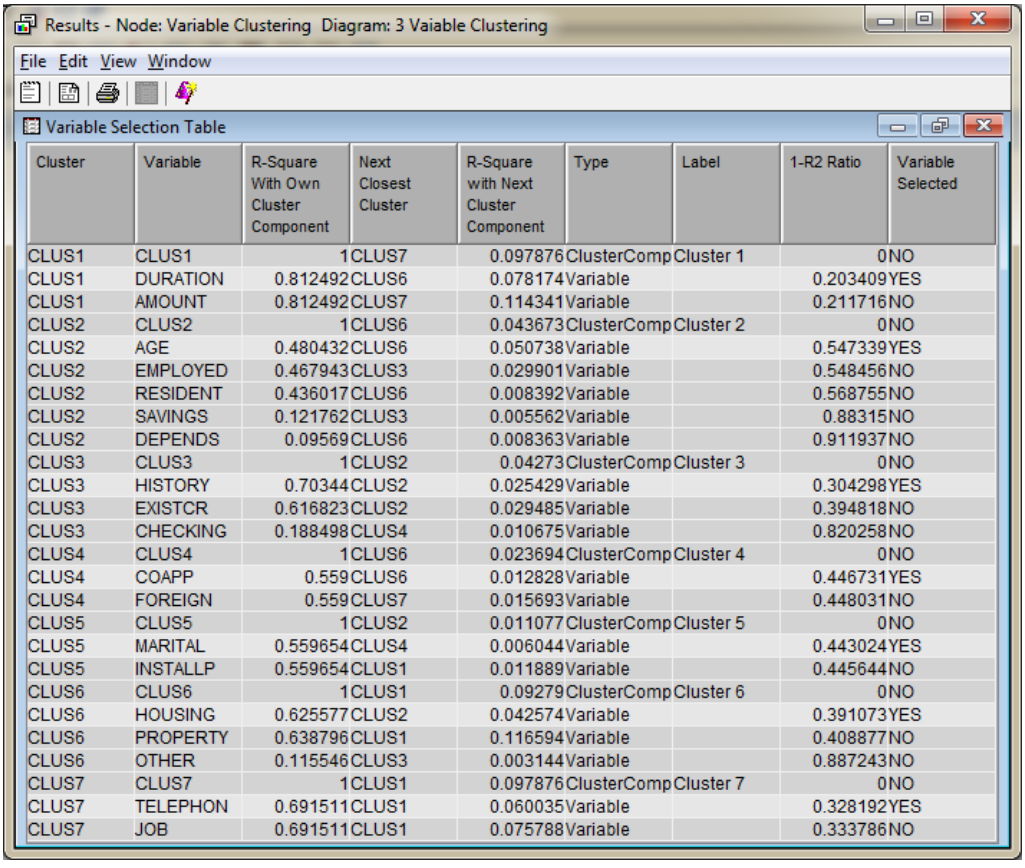
Slika 86. Rezultat komponente *Variable Clustering*

Svaki iterativni korak opsian je detaljno u izlaznoj datoteci (Slika 87).



Slika 87. Detaljne informacije o svim iteracijama klasterovanja

U slučaju da je potrebno izabrati najbolje promenljive dovoljno je promeniti svojstvo ***Variable Selection***=*Best Variables* i pokrenite analizu ponovo. Komponenta će sama izabrati najbolje promenljive tj. promenljive koje su najbliže klasteru (Slika 88). U koloni *Variable Selected* sa *YES* su obeležene najbolje promenljive.



Cluster	Variable	R-Square With Own Cluster Component	Next Closest Cluster	R-Square with Next Cluster Component	Type	Label	1-R2 Ratio	Variable Selected
CLUS1	CLUS1	1	CLUS7	0.097876	ClusterComp	Cluster 1	0	NO
CLUS1	DURATION	0.812492	CLUS6	0.078174	Variable		0.203409	YES
CLUS1	AMOUNT	0.812492	CLUS7	0.114341	Variable		0.211716	NO
CLUS2	CLUS2	1	CLUS6	0.043673	ClusterComp	Cluster 2	0	NO
CLUS2	AGE	0.480432	CLUS6	0.050738	Variable		0.547339	YES
CLUS2	EMPLOYED	0.467943	CLUS3	0.029901	Variable		0.548456	NO
CLUS2	RESIDENT	0.436017	CLUS6	0.008392	Variable		0.568755	NO
CLUS2	SAVINGS	0.121762	CLUS3	0.005562	Variable		0.88315	NO
CLUS2	DEPENDS	0.09569	CLUS6	0.008363	Variable		0.911937	NO
CLUS3	CLUS3	1	CLUS2	0.04273	ClusterComp	Cluster 3	0	NO
CLUS3	HISTORY	0.70344	CLUS2	0.025429	Variable		0.304298	YES
CLUS3	EXISTCR	0.616823	CLUS2	0.029485	Variable		0.394818	NO
CLUS3	CHECKING	0.188498	CLUS4	0.010675	Variable		0.820258	NO
CLUS4	CLUS4	1	CLUS6	0.023694	ClusterComp	Cluster 4	0	NO
CLUS4	COAPP	0.559	CLUS6	0.012828	Variable		0.446731	YES
CLUS4	FOREIGN	0.559	CLUS7	0.015693	Variable		0.448031	NO
CLUS5	CLUS5	1	CLUS2	0.011077	ClusterComp	Cluster 5	0	NO
CLUS5	MARITAL	0.559654	CLUS4	0.006044	Variable		0.443024	YES
CLUS5	INSTALLP	0.559654	CLUS1	0.011889	Variable		0.445644	NO
CLUS6	CLUS6	1	CLUS1	0.09279	ClusterComp	Cluster 6	0	NO
CLUS6	HOUSING	0.625577	CLUS2	0.042574	Variable		0.391073	YES
CLUS6	PROPERTY	0.638796	CLUS1	0.116594	Variable		0.408877	NO
CLUS6	OTHER	0.115546	CLUS3	0.003144	Variable		0.887243	NO
CLUS7	CLUS7	1	CLUS1	0.097876	ClusterComp	Cluster 7	0	NO
CLUS7	TELEPHON	0.691511	CLUS1	0.060035	Variable		0.328192	YES
CLUS7	JOB	0.691511	CLUS1	0.075788	Variable		0.333786	NO

Slika 88. Izbor promenljivih najbližih klasteru

A.2.vi Komponenta *Varijable Selection*

Često ABT-ovi imaju po nekoliko stotina, a ponekad i nekoliko hiljada promenljivih. Sve ove promenljive su mogu koristiti u procesu modelovanja. *Variable Selection* komponenta pomaže nam da redukujemo broj ulaznih promenljivih tako što odbacimo one promenljive za koje komponenta ustanovi da nisu u vezi sa ciljnom promenljivom. U daljem procesu modelovanja biće korišćene samo one promenljive koje nisu odbačene.

Ova komponenta brzo identifikuje promenljive koje su korisnije u procesu modelovanja za predikciju ciljne promenljive. Naravno, uvek je moguće odbačenu promenljivu ručno uključiti u proces modelovanja. Ova komponenta se obično koristi za redukciju ulaznih promenljivih za modele zasnovane na neuronskim mrežama, ali se može koristii i u ostalim metodama.

Komponenta koristi R-kvadrat (eng. *R-square*) i hi-kvadrat (eng. *Chi-square*) kriterijum izbora promenljivih.

R-kvadrat kriterijum izbora promenljivih

R-kvadrat izbor koristi *forward stepwise least square* regresiju da maksimizira R-kvadrat vrednost. Alogiritam obezbeđuju brzu preliminarnu ocenu promenljivih i brz razvoj prediktivnih modela sa velikim brojem promenljivih i opservacija. Na ovaj način se brzo identifikuju promenljive koje su korisne za predikciju ciljne promenljive na osnovu linearnih modela. R-kvadrat izbor promenljivih se izvršava u sledećim koracima:

- Izračunavanje kvadratne korelacije (eng. *Square Correlation*). Koeficijent kvadratne korelacije (R2) za svaku ulaznu promenljivu se računa i poredi sa podrazumevanim do tada izračunatim minimalnim R kvadratom (opcija komponente, podrazumevana vrednost je 0.005). Ako je koeficijent kvadratne korelacije manji od ***Minimum R-Square***, uloga ulazne promenljive se postavlja na odbačena (*Rejected*). Zavisno od potreba istraživanja podataka, moguće je promeniti *Minimum R-Square*. Specifična znanja iz oblasti za koje se radi istraživanje podataka mogu sugerisati da se ovaj kriterijum promeni. Uvećavanjem *Minimum R-Square* kriterijuma smanjujemo skup prediktivnih promenljivih i obratno. *Variable Selection* komponenta koristi jednostavnu linerarnu regresiju da obezbedi koeficijent kvadratne korelacije za intervalne promenljive odnosno „*one way frequency*“ analizu varijance da izračuna kvadratnu korelaciju kategoričkih promenljivih.
- *Forward Stepwise Regresion*. Nakon izračunavanja koeficijenta kvadratne korelacije za svaku promenljivu, preostale statistički značajne promenljive se ocenjuju koristeći R kvadrat regresiju. Proces sekvencijalnog pravolinijskog izbora počinje ulaznim promenljivama koje imaju najveću varijaciju u ciljnoj promenljivoj tj. promenljive sa najvećim koeficijentom kvadratne korelacije. U svakoj sledećoj iteraciji regresije dodaju se ulazne promenljive koje obezbeđuju najveći inkrementalni rast u R2 modelu. Iteracija se prekida kada od preostalih ulaznih promenljivih nije moguće obezbediti rast R2 u modelu (opcija ***Stop R-square criterion***; podrazumevana vrednost je 0.0005)

- Logistička regresija za binarne ciljne promenljive. Ako je ciljna promenljiva binarna tada se na kraju izvršava logistička regresija koristeći prediktivne vrednosti dobijene kao rezultat Forward Stepwise selekcije (nezavisne ulazne promenljive).

Hi- kvadrat kriterijum selekcije

Ovaj kriterijum selekcije je dostupan samo u slučaju da je ciljna promenljiva binarna. Ovaj kriterijum obezbeđuje brzu preliminarnu ocenu promenljivih. Selekcija koristi cepanje po binarnoj promenljivoj da maksimizira hi-kvadrat vrednost od 2x2 matrice frekvencije.

NULL vrednosti

Komponenta tretira nedostajuće vrednosti na sledeći način:

- Opservacije koje imaju nedostajaću vrednost u ciljnoj promenljivoj biće isključene iz dalje analize
- Nedostajuće vrednosti u kategoričkim ulaznim promenljivama biće tretirana kao nova kategorija.
- Nedostajuće vrednosti u intervalnim promenljivima zamenjuju se ponderisanim prosekom (eng. *weighted mean*)

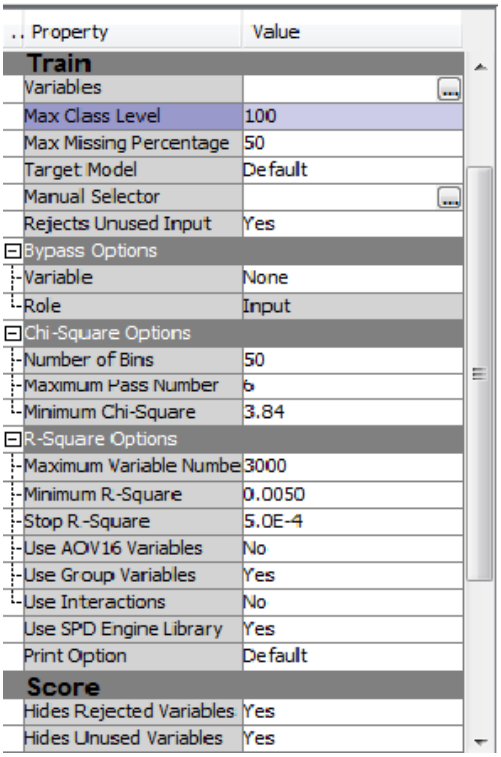
Osobine komponente

Max Missing Percetage – koristi se da se odbace promenljive kod koji je broj opservacija sa *NULL* vrednostima veći od navedenog

Target Model – kriterijum selekcije. Može biti *R–Square*, *Chi-Square* ili oba. Podrazumevana vrednost: Ako je ciljna promenljiva binarna i ako je stepen slobode (end. *degree freedom*) veći od 400 koristi se *Chi-Square* inače se koristi *R-Square*.

Manual Selector – mogućnost da se neke promenljive ručno selektuju. Ove promenljive ne ulaze dalje u selekciju i ne mogu biti dobiti ulogu *Rejected*.

Reject Unused Input – ako je postavljeno na *YES* komponenta automatski odbacuje promenljive koje nisu „prediktive“.



Slika 89. Osobine komponente Variable Selection

Specifičnosti hi-kvadrat algoritma

Number of Bins – Broj slojeva (strata) koji se koriste u transformaciji intervalne promenljive u nominalnu. Koristi se za određivanje hi-kvadrata za intervalne promenljive.

Maximum Pass Number – maksimalan broj prolaza prilikom određivanja optimalnog broja binarnih podela

Minimum Chi-Square – predstavlja donju granicu hi-kvadrat vrednosti u kojoj promenljiva ostaje dostupna za selekciju. Vrednost mora biti veća od 0. Hi-kvadrat vrednost predstavlja x osu odgovarajuće verovatnoće *chi- square* distribucije.

$$P(\text{ chi-square statistic} > 3.84) = 0.05$$

Verovatnoća	Hi-kvadrat vrednost
0.01	6.635
0.05	3.841
0.10	2.706

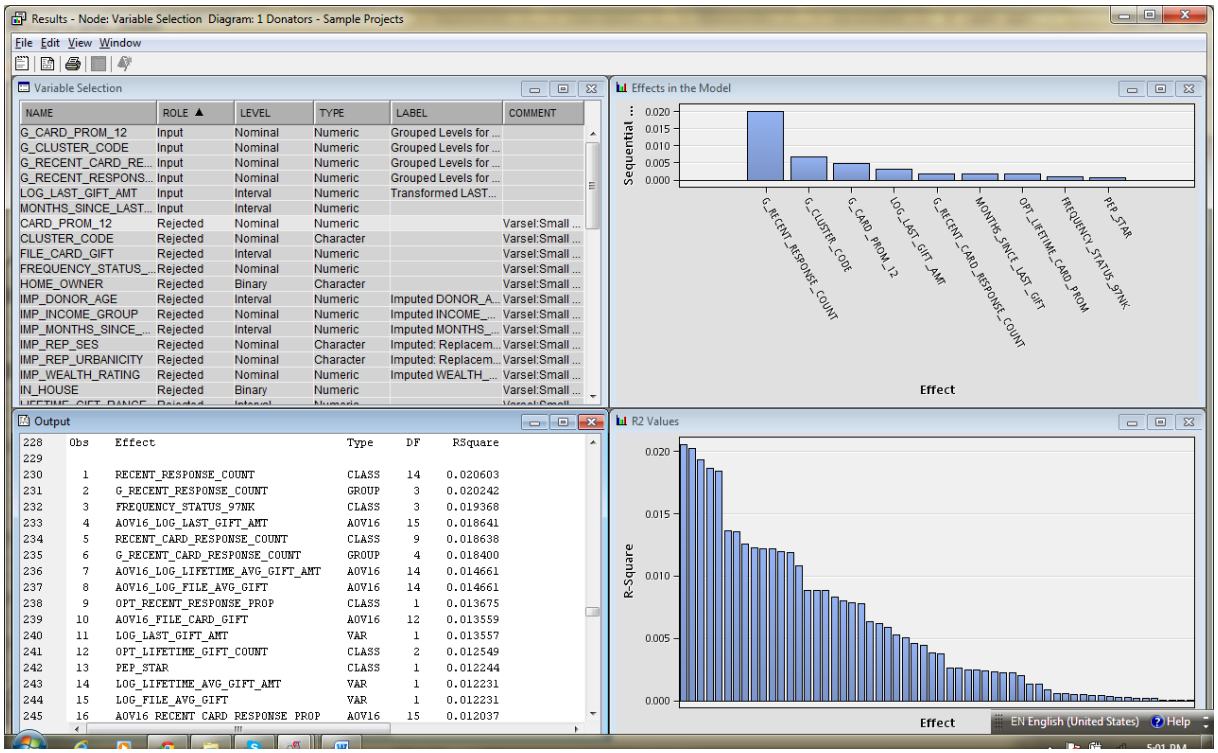
Tabela 4. P vrednosti za neke hi-kvadrat vrednosti

Specifičnosti R-kvadrat algoritma

Maximum Variable Number – maksimalan broj promenljivih koje možemo izabrati u model.

Minimum R-Square – predstavlja donju granicu u kojoj promenljiva „ostaje u igri“ (videti prvi korak R-kvadrat algoritma)

Stop R-Square – kriterijum zaustavljanja R-kvadrat algoritma.



Slika 90. Rezultat *Variable Selection* komponente

Na slici (Slika 90) je prikazan rezultat *Variable Selection* komponente. U gornjem levom uglu nalazi se spisak značajnih promenljivih (*Role=Input*). Ostale promenljive su odbačene. U gornjem desnom uglu nalazi se lista promenljivih koji imaju najveći efekat na model. Među njima su i neke odbačene promenljive (zbog *Stop R–Square* kriterijuma). U donjem desnom uglu prikazan je R-kvadrat za sve ulazne pomenljive. Pozicioniranjem miša na odgovarajući stubić može se dobiti ime promenljive i odgovarajuća R-kvadrat vrednost.

U donjem levom uglu nalazi se detaljna izlazna datoteka. Ovde su opisani koraci R–kvadrat algoritma kao i međurezultati svake iteracije.

Slovo *G* ispred promenljive znači da ove nominalne promenljive ima veliku kardinalost (broj članova) i zbog toga su one grupisane u manji broj grupa.

LOG u prefiksu znači da ovo nije originalna promenljiva već da je vrednost promenljive logaritmovana kako bi se smanjila standardna devijacija. Ovo kao i rešavanje problema nedostajućih vrednosti treba uraditi pre pokretanja komponente.

A.2.vii Ostale komponente koje se ređe koriste u izradi modela zasnovanih na skoru

Komponenta Cluster

Klasterovanje opservacija. Na ovaj način moguće je segmentirati bazu. Na osnovu samih podataka komponenta svrstava opservacije u klastere tako da se opservacije u jednom

klasteru mogu opisati. Klastering analiza se radi na osnovu izračunatog rastojanja (Euklidskog) između dve ili više promenljivih.

Rezultat klasterovanja se grafički može prikazati i opisati . Takođe moguće je generisati SAS kod, C kod ili SQL kod koji se može izvršiti nad validacionim i testnim podacima kao i u produkcionom okruženju.

U bankarskoj industriji koristi se za stratešku i taktičku segmentaciju klijenata (ponašanje klijenata).

Komponente Association i Market Basket

Asocijacija je identifikovanje objekata koji se pojavljuju zajedno u nekom događaju ili zapisu. Ova tehnika je poznata kao *Market Basket* analiza. Izvor za ovu komponentu je transakciona baza. Pravila asocijacije su zasnovana na brojanju učestalosti u vremenskom periodu u transakcionoj datoteci. Jedno od pravila asocijacije može da bude:

„Ako se objekat A pojavljuje u događaju/zapisu O_i tada se i objekat B pojavljuje u X% slučajeva“

Da bi se izvršila analiza neophodno je da ulazni skup podataka ima ulogu „*Transaction*“ sa jasno naznačenom vremenskom dimenzijom u ulaznom skupu podataka (*Role=Time ID*).

Koponenta Path Analysis

Koristi se u analizi WEB loga. Komponenta omogućava da se analizira putanja kojom se klijent kretao u odnosu na ciljnu promenljivu (u ovom slučaju putanja).

Takođe, komponenta može analizirati niz podataka tako da otkrije uzastopnu učestalost nekog podniza.

Komponenta SOM/Kohonen

Komponenta se koristi za nenadgledano učenje (eng. *unsupervised learnig*) koristeći Kohonenova kvantizacioni vektor (eng. *Kohonen vector quantization VQ, Kohonen self – organizing maps (SOMs)*).

A.3 Modifikovanje podataka – *Modify*

A.3.i Komponenta *Drop*

Drop komponenta se koristi za uklanjanje promenljivih iz skupa podataka odnosno njihovo skrivanje u metapodacima SAS EM projekta.

A.3.ii Komponenta *Replacement*

Replacment komponenta koristi se za zamenu vrednosti promenljive sa unapred definisanom vrednošću. Npr. ako imamo bimodalnu distribuciju i želimo da uklonimo manju „grbu“ možemo sve vrednosti manje grbe zameniti sa prosečnom vrednošću promenljive.

A.3.iii Komponenta *Impute*

Komponta *Impute* koristi se za zamenu nedostajućih vrednosti. Impute komponenta obezbeđuje sledeću zamenu nedostajućih vrednosti za intervalne promenljive:

- *Andrew's Wave*
- *Default Constant*
- *Distribution*
- *Huber*
- *Mean*
- *Median*
- *Mid-Minimum Spacing*
- *Midrange*
- *None*
- *Tree*
- *Tree Surrogate*
- *Tukey's Biweigh*t

Nedostajuće vrednosti za kategoričke promenljive se mogu zameniti sa:

- *Count*
- *Default Constant*
- *Distribution*
- *None*
- *Tree*
- *Tree Surrogate*

A.3.iv Komponenta *Transform Varijables*

Transform komponenta pravi novu promenljivu na osnovu postojeće promenljive. Npr. ako je standardna devijacija intervalne promenljive X (gde je X>0) velika tada se može kreirati nova promenljiva $LOG_{10}(X)$ kod koje je standardna devijacija mnogo manja.

A.3.v Komponenta *Interactive Binning*

U slučaju da je ciljna promenljiva binarna možemo koristiti *Interactive Binning* komponentu koja vrednosti svake intervalne i kategorične promenljive grupiše u unapred zadati broj grupa. Cilj grupisanja je da se poveća prediktivna snaga svake promenljive posebno.

Ova komponeta predstavlja alat za grupisanje koja koristi *Gini* statistiku koja je opisan u dodatku B poglavlje *Gini koeficijent*. Kreirane grupe unutar promenljive u nekim slučajevima mogu biti prediktivnije od samih promenljivih.

Grupisanje nam daje mnoge prednosti:

- Ovo je jednostavan način da se prevaziđe retke vrednosti nominalnih promenljivih kao i ekstremne vrednosti (eng. o*utliers*) kod kontinualnih promenljivih.
- Nelinerne zavisnosti mogu biti modelovane sa linearnim modelom.
- Omogućava punu kontrolu u procesu razvoja modela kao što je modifikovanje i izrada novih grupa od strane samog korisnika (videti aplikaciju *Interactive Selection*)
- Proces grupisanja omogućava korisniku da uđe unutar svake promenljive i sazna više o samoj promenljivoj.

Osnovne osobine komponete

Treat Missing as Level – ako je postavljeno na *YES* nedostajuće vrednosti se posebno grupišu.

Use Frozen Group – ako je postavljeno na *YES* predhodno napravljene grupe će se primeniti i nad novim ulaznim podacima tj. neće se računati nove grupe. Ovo može biti korisno u slučaju dodavanja novih promenljivih u trening populaciju što je čest slučaj. Tada će interactive binning biti primenljen samo nad novim promenljivama.

Method – može biti *Quantile* ili *Bucket*. Ovo se primenjuje samo za grupisanje intervalnih promenljivih.

Number of Groups – podešava se broj nonmissing grupa; obično je to od 4 do 7.

Apply Level Rule – ako je postavljeno na *YES* broj različitih vrednosti će biti poređen sa brojem grupa. U slučaju da je broj različitih vrednosti manji od broja grupa promenljiva će u procesu grupisana biti tretirana kao kategorička.

Group Rare Level – ako je postavljen na *YES* sve vrednosti kategoričke promenljive koje se pojavljuju manje od ***Cutooff Value Percentagle*** biće smeštene u istu grupu.

Variable Selection Method – može biti *Gini Statistic* ili *None*. U slučaju da je *None* ne postoji selekcija promenljivih već se ceo skup grupisanih promenljivih prosleđuje dalje. Izbor promenljivih se može uraditi na neki drugi način u kasnijoj fazi razvoja modela (npr.

koristeći *Variable Selection* komponentu ili umesto Gini koeficijenta koristiti *Information Value*).

Gini Cutoff – vrednost Gini koeficijenta koji se koristi za izbor promenljivih (podrazumevana vrednosti je 20). Sve promenljive koje imaju manji koeficijent biće odbačene u daljem procesu modelovanja.

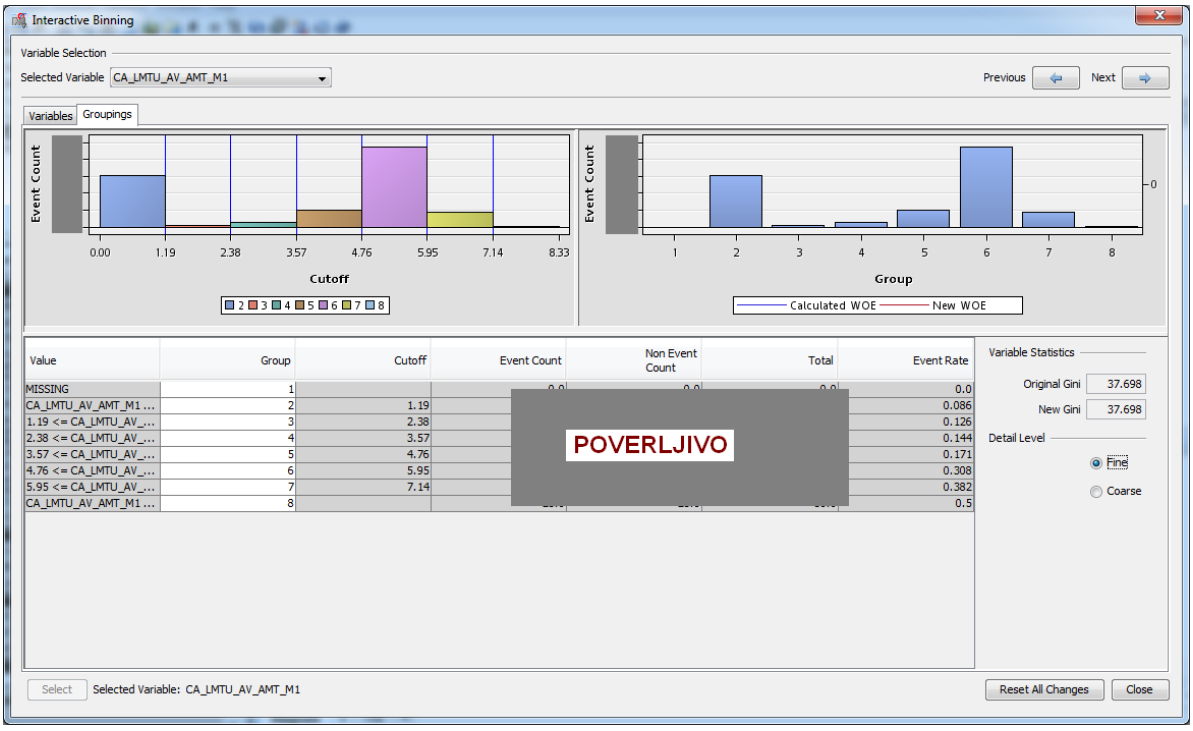
Import Grouping Data – postavljamo na *YES* u slučaju da želimo da definiciju grupa učitamo iz eksternog izvora.

Import Data Set – tabela sa definicijom grupa. Određivanje najboljeg grupisanja za svaku promenljivu se može uraditi nekim drugim alatom. Cilj određivanja grupisanja je maksimizacija Gini koeficijenta za svaku promenljivu. Učitavanje ovih metapodataka možemo SAS EG narediti da grupiše po našim pravilima i da dalje ovako kreirane promenljive koristi u procesu modelovanja.

Interactive Selection – videti sledeće poglavlje

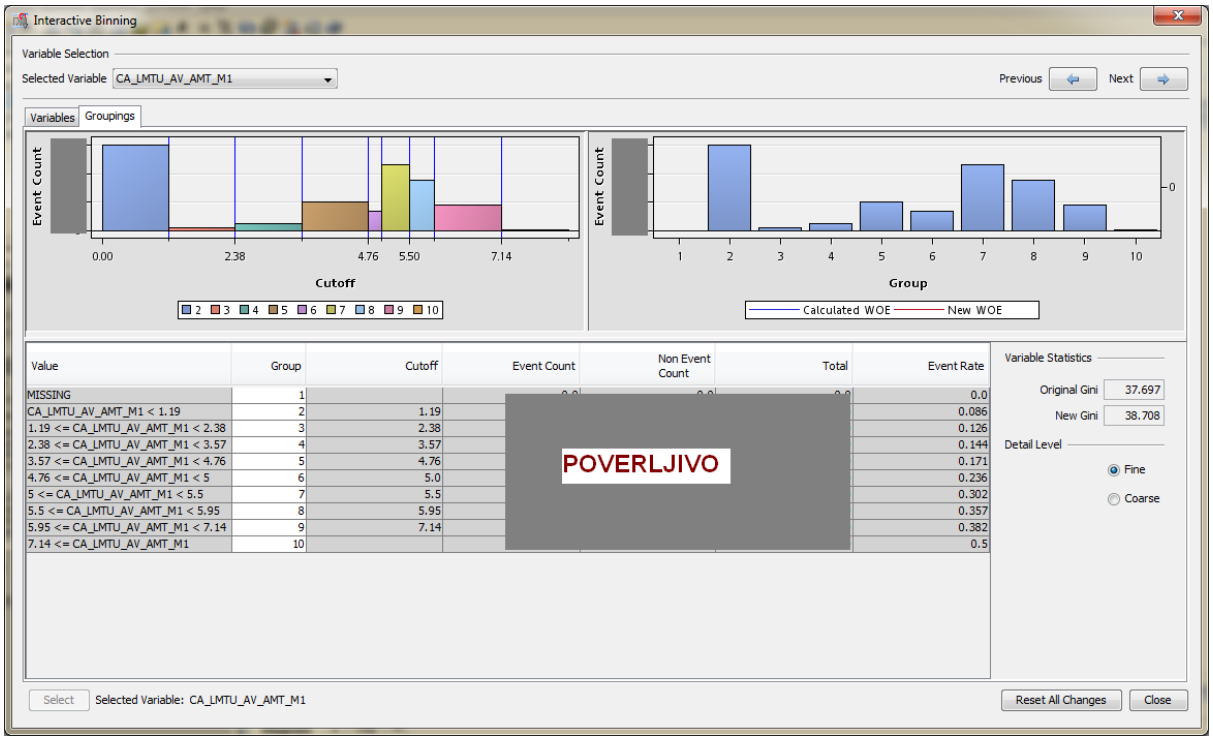
Aplikacija Interactive Selection

Ovo je veoma korisna aplikacija koja ostavlja mogućnost korisniku da za neke promenljive sam izračuna grupe. Prilikom ponovnog grupisanja automatski se računa Gini koeficijent. Ako je novi Gini koeficijenta veći od originalnog novo grupisanje je bolje od postojećeg i treba ga zadržati.



Slika 91. Aplikacija *Interactive Selection*

Na slici (Slika 91) nalazi se originalno grupisanje promenljive CA_LMTU_AV_AMT_M1. Koristeći ovu aplikaciju grupisanje je moguće podeliti tako da novi Gini koeficijent bude veći.



Slika 92. Promenjene grupe promenljive CA_LMTU_AV_AMT_M1

Na slici (Slika 92) stara grupa broj 6 je podeljena na tri nove grupe. Originalni Gini koeficijent je *37.697* dok je novi Gini koeficijent *38.708*. Ovo nam govori da je novo grupisanje bolje od starog i da ga treba zadržati.

A.3.vi Komponenta *Principal Component*

Ova komponenta se koristi za redukovanje broja ulaznih promenljivih. Komponenta koristi metod redukcije promenljivih projekcijom vektorskog prostora kreiranog od matrice korelacije ili matrice kovarijansi ulaznih promenljivih nad uzorkom za trening. Metoda je opisana u dodatku B poglavlje *Analiza glavnih komponenti*.

Komponenta računa sopstvene vrednosti i kreira sopstvene vektore od matrice $COV(X_{i,X_j})$ ili matrice $COR(X_i,X_j)$, gde je i,j<=n.

Rezultat rada komponente je novi skup promenljivih (sopstveni vektori), dok se ulazne promenljive odbacuju.

Ovom metodom se efikasno eliminiše linearna zavisnost promenljivih (kolinearnost), jer su sopstveni vektori ortogonalni, a zadržava se nelinearna zavisnost.

Interpretacija rezultata je često problematična ili nemoguća poslovnim korisnicima.

Osobine komponente

Eigenvalue Source – izvor za računanje sopstvenih vektora može biti:

- *Covariance* – kovarijansa
- *Corelation (default)* – korelacija

- *Uncorrected* – koristi nekorigovanu matricu ulaznih promenljivih

Interactive Selection – kriteriju za selektovanje sopstvenih vrednosti. Može biti:

- *Eigenvalue* – sopstvene vrednosti
- *Proportional Eigenvalue*
- *Cumulative Proportional Eigenvalue*
- *Log Eigenvalue*
- *Eigenvalue Table*

Cumulative – vrednost pomoću kojih se selektuju sopstveni vektori; predstavlja kumulativnu proporciju varijanse svakog principala u odnosu na ukupnu varijansu; u slučaju da je veća od ove vrednosti principal se ne prosleđuje dalje tj. dobija ulogu „rejected“. Podrazumevana vrednosti je 0.99

Increment – u slučaju da kumulativna proporcija varijanse dosegne 0.9 principal mora imati inkrement veći od zadatog. Podrazumevana vrednost je 0.

Aply Maximum Number – ako se postavi na *YES* onda se u osobini *Maximum Number* postavlja broj sopstvenih vektora.

A.4 Razvoj modela – *Model* (regresiona analiza)

U ovom poglavlju biće opisana komponenta *Regresion* SAS EM koja je korišćena u ovom radu. Matematičke osnove su date u dodatku B poglavlja *Linearna regresija Logistička regresija*.

A.4.i Tipovi regresione analize

Regression Type – tip regresione analize može biti „*Linear Regresion*“ ili „*Logistic Regresion*“

Link Function – za linearnu regresiju funkcija uvek je u obliku $g(M) = X\beta$; u slučaju da je izabrana logistička regresija moguće je izabrati sledeće funkcije: ***Cloglog***, ***Logit*** (podrazumevana vrednost), ***Probit***

U ovom radu je koričen ***Logit*** tj.

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \,.$$

A.4.ii Kodiranje kategoričkih promenljivih u regresionoj analizi

Suppress Intercept – u slučaju da je „*Yes*“ izbegnuto je kreiranje onoliko promenljivih koliko ima članova u nekoj klasifikacionij promenljivoj; u slučaju da je „*No*“ (podrazumevana vrednost) kreira se uvek n-1 promenljiva za n članova neke kategoričke promenljive

Input Coding – označava metod koji može biti primenjen to su ***GLM*** (*eng. non full rank General Linerar Model*) ili ***Deviation***

Primer. *Deviation Coding* i *GLM* kodiranja

Level	Job Clerical	Job Lawyer
Clerical	1	0
Lawyer	0	1
Paralegal	–1	–1

Tabela 5. *Deviation Coding*

Level	Job Clerical	Job Lawyer	Job Paralegal
Clerical	1	0	0
Lawyer	0	1	0
Paralegal	0	0	1

Tabela 6. *GLM*

Izbor kodiranja može uticati na modelovanje tj. može dati različite rezultate regresione analize.

A.4.iii Izbor metoda regresione analize

Selection Model – moguće je izabrati tri metoda selekcije promenljivih. To su:

- ***Backward*** – inicijlano sve promenljive su kandidati za regresionu funkciju, a zatim se jedna po jedna eliminiše ako ne zadovolje zadate uslove za značajnost promenljive (*Stay Significance Level*) ili dok nije zadovoljeni uslov za prestanak analize (*Stop Criteria*)
- ***Forward*** – nijedna promenljiva nije kandidat za regresionu funkciju već se dodaju jedna po jedna tako da se dosegne uslov značajnosti (*Entry Significance Level*) ili dok nije zadovoljen uslov za prestanak analize
- ***Stepwise*** - počinje kao forward regresiona analiza ali uz mogućnosti izbacivanja promenljivih sve dok nisu zadodovljeni odgovarajući uslovi
- ***None (default)*** – sve promenljive učestvuju u analizi

Selection Criterion - u slučaju da je izabran neki od algoritama izbora promenljivih, kruterijum izbora finalnog modela može biti:

- ***Default*** – koristi „profit/gubitak“ kriterijum nad uzorkom
- ***None*** – koristi standardni kriterijum izbora promenljivih pomoću p-vrednosti
- ***Akaike's Information Criterion*** (AIC) – model sa najmanjom AIC vrednoću biće izabran
- ***Schwarz's Bayesian Criterion*** (SBC) – model sa najmanjom SBC vrednoću biće izabran
- ***Validation Error*** – minimalna suma kvadrata grešaka za regresiju nad uzorkom za proveru
- ***Validation Misclassification*** – model sa najmanjim ROC (videti poglavlje 8.2 Ocena modela) biće izabran

Use Selection Defaults – *No* u slučaju da želimo da definišemo sopstvene kriterijume izbora promenljivih; *Yes* (default) inače.

Selection Options – kriterijumi izbora mogu biti:

- ***Sequential Order*** - u slučaju da je postavljen na *Yes* komponenta dodaje i izbacuje promenljive koristeći poredak značajnosti. *No* je podrazumevana vrednost.
- ***Entry Significance Level*** – u slučaju da je postavljen na *Yes* neophodno je uneti vrednost između 0 i 1. Podrazumevana vrednost je 0.05.
- ***Stay Signifince Level*** - u slučaju da je postavljen na *Yes* neophodno je uneti vrednost između 0 i 1. Podrazumevana vrednost je 0.05.
- ***Start Variable Number*** – u slučaju da je postavljen na *Yes* neohodno je definisati broj promenljivih od kojih se započinje regresiona analiza. Za *Forward* i *Stepwise*

regresiju podrazumevana vrednost je 0, dok je za *Backward* regresiju podrazumevana vrednost broj ulaznih promenljivih.

- ***Stop Variable Number*** – broj promenljivih posle kojeg se zaustavlja izbor novih promenljivih; podrazumevana vrednost je 0 za *Backward*, a ukupan broj promenljvih za *Forward* regresiju.
- ***Force Candidate Effects*** – predstavlja minimalan broj promenljvih koje moraju da učestvuju u regresiji;
- ***Maximum Number of Step*** – predstavlja maksimalan broj koraka u *stepwise* regresionoj analizi.
- ***Hierarchy Effect, Moving Effect Rule*** – videti sledeće poglavlje Efekat hijerarhije

A.4.iv Efekat hirerarhije

Osobina ***Effect Hierarchy*** dopušta da se kontroliše proces izbora promenljivih. Na primer, pretpostavimo da imamo tri promenljive A, B i C=A*B. Možemo zahtevati da modelovanje bude hijerarhijsko tako da uključi A*B samo ako su uključeni A i B. Slično važi i u slučaju da je neohodno neku od ove tri promenljive izbaciti. U tom slučaju se sve tri izbacuju.

Efekat hijerarhije postoji samo ako su promenljive klasifikacione ili ako u hierarhiji učestvuju i intervalne i klasifikacione promenljive.

Hierarchy Effects – postavlja se na *Class* ako želimo da samo klasifikacione promenljive razmatramo u hijerarhiji; *All* ako želimo da uključimo i klasifikacione i intervalne promenljive.

Movinig Effect Rule – moguće je izabrati jedan od sledećih efekata:

- ***None (default)*** – efekat hijerarhije se ne koristi
- ***Single*** – samo jedan uslov može biti uzet ili napušten u modelovanju u jednom trenutku (iteraciji). Npr. Ako su A i B uzeti u model u prvom koraku, u drugom koraku i A*B mora biti uzeta. Takođe, u slučaju da A i B treba da budu izbačene iz modelovanja prvo se mora izbaciti A*B
- ***Multiple*** – više uslova može biti uzeto ili napušetno u modelovanju u jednom trenutku.

A.4.v Optimizacija algoritma

Technique – predstavlja tehnike optimizacije. Mogu biti: ***Congra*** *(conjugate gradient optimization technique),* ***Dbldog*** *(Double Dogleg optimization technique),* ***Newrap*** *(Newton-Raphson with Line Search optimization technique),* ***Nrridg*** *(Newton-Raphson with Ridging optimization technique),* ***Quanew*** *(Quasi-Newton optimization technique)* ***Trureg*** *(Trust-Region optimization technique)*

Default Optimization – postavlja se na *No* ako želimo da sami promenimo ocobine optimizacije ispod

Max Iterations – maksimalan broj iteracija. Zavisno izabrane tehnike optimizacije podrazumevane vrednosti se nalaze u tabeli 7.

Optimization Technique	Default Max Iterations
Default	0
Congra	400
Dbldog	200
Newrap	50
Nrridg	50
Quanew	200
Trureg	50

Tabela 7. Podrazumevane vrednosti broja iteracija za različite tehnike optimizacije

Max Function Calls predstavlja maksimalan broj poziva funkcija

Optimization Technique	Default Max Function Calls
Default	0
Congra	1000
Dbldog	500
Newrap	125
Nrridg	125
Quanew	500
Trureg	125

Tabela 8. Podrazumevane vrednosti poziva funkcija modela zavisno od tehnike optimizacije

Maximum Time – predstavlja maksimalno vreme zauzeća CPU (npr. 5 minuta, 1 dan, 7 dana).

A.4.vi Kriterijumi konvergencije

Uses Defaults – *No* ako sami želimo da podesimo kriterijum konvergencije; *Yes* koristi podrazumevana podešavanja

Options – moguće je izabrati sledeće kriterijume konvergencije: *Absolute, Absolute Function, Absolute Function Times, Absolute Gradient, Absolute Gradient Times, Absolute Parameter, Absolute Parameter Times, Relative Function, Relative Function Times, Relative Gradient, Relative Gradient Times.*

A.4.vii Opcije izlaza

Confidence Limits – postavlja se na *Yes* ako želimo da generišemo granicu poverenja za parametre procene. Podrazumevana vrednost je *No*.

Save Covariance – postavlja se na *Yes* ako želimo da snimimo matricu kovarijanse parametara procene. Podrazumevana vrednost je *No*.

Covariance – postavlja se na *Yes* ako želimo da prikažemo matricu kovarijanse za parametre procene . Podrazumevana vrednost je *No*.

Correlation – postavlja se na *Yes* ako želimo da prikažemo matricu korelacije za parametre procene. Podrazumevana vrednost je *No*.

Statistics – postavlja se na *Yes* ako želimo da prikažemo jednostavnu deskriptivnu statistiku za sve ulazne promenljive. Podrazumevana vrednost je *No*.

Details – postavlja se na *Yes* ako želimo da prikažemo detalje svake iteracije u procesu.

Design Matrix – postavlja se na *Yes* ako želimo da prikažemo kodirane ulazne klasifikacione promenljive.

Excluded Variables – *None* nema efekta na promenljive, *Hide* – uklanja promenljive iz metapodataka, *Reject* – isključuje promenljive iz dalje analize, ali ih zadržava u metapodacima.

A.5 Ocena modela – Assess

A.5.i Komponenta *Model Comparation*

Komponenta *Model Comparation* poredi modele koristeći različite kriterijume i tehnike. Izbor kriterijuma zavisi od primene modela. Za binarne ciljne promenljive ti kriterijumi su grupisani po tipu analize i mogu biti:

- Klasifikacione mere kao što je ROC (eng. *Recivier Operating Characteristics*) grafikon i kriva, odnos klasifikacije (classification rates) i sl.
- *Data mining* mere tj. merenje modela kroz prizmu profita i gubitka (eng. *lift measure*)
- Statističke mere kao što su BIC (eng. *Bayesian Information Criterion*), AIC (eng. *Akaike's Information Criterion*) , Gini , Kolmogorov-Smirnov, *Bin-Best-Two-Way* Kolmogorov-Smirnov test

Ova komponenta je opisana u poglavlju *8.2 Ocena modela*.

A.5.ii Komponenta *Score*

Score komponenta se koristi za generisanje programskog koda regresione funkcije. Programski kod može biti SAS kod, C kode Java kod ili DB2 SQL funkcija. Takođe, ova komponenta se koristi i prilikom testiranja modela u slučaju da želimo da testiramo model koristeći uzorak koji ima drugu vremensku dimenziju u odnosu na uzorak za trening i proveru modela.

B. Matematičke osnove

B.1 Prosečna vrednost, medijana i najfrekventnija vrednost

Tip	Opis	Primer	Rezultat primera
Aritmetička sredina (*Arithmetic mean*)	Suma vrednosti podeljena sa brojem vrednosti	(1+2+2+3+4+7+9) / 7	4
Medijana (*Median*)	Vrednost u sredini koja razdvaja sortirani niz na dva jednaka dela	1, 2, 2, **3**, 4, 7, 9	2
Mode	Najfrekventniju vrednostu u skupu	1, ***2***, ***2***, 3, 4, 7, 9	2

B.2 Percentili

Kažemo da je u uzorku za promenljivu *X* vrednost *y n-ti percentil (centil)* ako u *n%* opservacija promenljiva *X* ima manju ili jednaku vrednost od *y*.

Primer. Percentili za promenljivu starost.

	P90	P10	P5	P1
Age	65	27	25	21

Tabela 9. Percentili za promenljivu starost (Age)

Iz tabele 9 vidimo da su klijenti u 90% opservacija mlađi od 65 godina, 10% opservacija mlađi od 27, 5% opservacija mlađi od 25 i u 1% opservacija mlađi od 21 godinu.

Dvadeset peti percentil zovemo prvi kvantil (Q1), pedeseti percentil je medijana ili drugi kvantil (Q2), sedamdesetpeti percentil predstalja treći kvantil (Q3). Dakle,

P25=Q1, P50=Q2=Median, P75=Q3

B.2.i Odsečeni prosek (eng. *truncated mean*)

Često u uzorku neka promenljiva ima ekstremne vrednosti (eng. *outliers*). To su vrednosti promenljive koje su mnogo veće ili mnogo manje od ostalih. U tom slučaju možemo računati prosek bez tih ekstrenimh vrednosti tj. samo za one vrednosti koje npr. pripadaju intervalu (P1,P99) . Ovako izračunatu prosečnu vrednost zovemo odsečena sredina. Ona mnogo bolje opisuje uzorak nego izračunata aritmetička sredina.

B.2.ii Interkvartalni prosek

Specifičan primer odsečenog proseka je prosečno stanje vrednosti u 2 i 3 kvartilu (eng. *interquartile mean*).

$$\bar{x} = \frac{2}{n} \sum_{i=(n/4)+1}^{3n/4} x_i$$

B.2.iii Interkvartalni opseg

Interkvartalni opseg (eng. *interquartile range, midspread, middle fifty,* IQR) predstavlja meru statističke disperzije:

IQR = $Q_3 - Q_1$

B.3 Standardna devijacija

Standardna devijacija (eng. *standard deviation*) pokazuje koliko mnogo varijacija i dispersija postoji u odnosu na prosečnu vrednost odnosno na očekivanu vrednost. Mala standardna devijacija ukazuje da vrednosti promenljive imaju tendenciju da budu blizu proseka. Visoka standardna devijacija ukazuje da postoji velika disperzija u odnosu na prosek.

Standardna devijacija od slučajne promenljive predstavlja kvadratni koren od varijance.

Neka je X slučajna promenljiva sa prosečnom vrednošću μ

$$\mathrm{E}[X] = \mu.$$

i neka E označava prosek ili očekivanu vrednost od X.

Standardna devijacija od X predstavlja

$$\begin{aligned}\sigma &= \sqrt{\mathrm{E}[(X-\mu)^2]} \\ &= \sqrt{\mathrm{E}[X^2] + \mathrm{E}[(-2\mu X)] + \mathrm{E}[\mu^2]} = \sqrt{\mathrm{E}[X^2] - 2\mu\,\mathrm{E}[X] + \mu^2} = \sqrt{\mathrm{E}[X^2] - 2\mu^2 + \mu^2} = \sqrt{\mathrm{E}[X^2] - \mu^2} \\ &= \sqrt{\mathrm{E}[X^2] - (\mathrm{E}[X])^2}.\end{aligned}$$

Standardna devijacija σ je kvadratni koren od varijance tj. kvadratni koren od proseka $(X - \mu)^2$.

U slučaju da X predstavlja konačan skup vrednosti $x_1, x_2, ..., x_N$, kod kojih svaka vrednost ima istu verovatnoću standardna devijacija je

$$\sigma = \sqrt{\frac{1}{N}\left[(x_1-\mu)^2 + (x_2-\mu)^2 + \cdots + (x_N-\mu)^2\right]}, \quad \text{where} \quad \mu = \frac{1}{N}(x_1 + \cdots + x_N),$$

odnosno

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}, \quad \text{where} \quad \mu = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

U slučaju da nemamo istu verovatnoću za sve vrednosti promenljive i neka x_1 ima verovatnoću p_1, x_2 p_2, ..., x_N p_N, standardnu devijaciju računamo kao

$$\sigma = \sqrt{\sum_{i=1}^{N} p_i(x_i - \mu)^2}, \quad \text{where} \quad \mu = \sum_{i=1}^{N} p_i x_i.$$

B.4 Kovarijansa

Kovarijansa (eng. *covariance*) je mera koja nam pokazuje koliko mnogo se dve slučajne promenljive menjaju zajedno. Ako velike vrednosti jedne promenljive odgovaraju velikim vrednostima druge promenljive, i obratno, ako male vrednosti jedne promenljive odgovaraju malim vrednostima druge promenljive tada je kovarijansa pozitivna. U suprotnom, kada male vrednosti jedne promenljive odgovaraju velikim vrednostima druge promenljive i obratno kovarijansa je negativna. Znak kovarijanse nam pokazuje kakva je linearna zavisnost između promenljivih. Magnitudu kovarijanse nije lako interpretirati. Normalizovana verzija kovarijanse, koeficijent korelacije nam pokazuje stepen linearne zavisnosti.

Kovarijansa između dve slučajne promenljive X i Y sa konačnim drugim momentom je definisan kao

$$\sigma(x, y) = \mathrm{E}\left[(x - \mathrm{E}[x])(y - \mathrm{E}[y])\right],$$

gde je E(x) očekivana vrednost odnosno prosečna vrednost.

Koristeći linearnost očekivane vrednosti dobijamo formulu

$$\begin{aligned}\sigma(x, y) &= \mathrm{E}\left[(x - \mathrm{E}\left[x\right])\left(y - \mathrm{E}\left[y\right]\right)\right] \\ &= \mathrm{E}\left[xy - x\,\mathrm{E}\left[y\right] - \mathrm{E}\left[x\right] y + \mathrm{E}\left[x\right]\mathrm{E}\left[y\right]\right] \\ &= \mathrm{E}\left[xy\right] - \mathrm{E}\left[x\right]\mathrm{E}\left[y\right] - \mathrm{E}\left[x\right]\mathrm{E}\left[y\right] + \mathrm{E}\left[x\right]\mathrm{E}\left[y\right] \\ &= \mathrm{E}\left[xy\right] - \mathrm{E}\left[x\right]\mathrm{E}\left[y\right].\end{aligned}$$

Za dva slučajna vektora ***x*** i ***y*** dimenzija **m** i **n** respektivno matrica korelacije je

$$\begin{aligned}\sigma(\mathbf{x}, \mathbf{y}) &= \mathrm{E}\left[(\mathbf{x} - \mathrm{E}[\mathbf{x}])(\mathbf{y} - \mathrm{E}[\mathbf{y}])^{\mathrm{T}}\right] \\ &= \mathrm{E}\left[\mathbf{x}\mathbf{y}^{\mathrm{T}}\right] - \mathrm{E}[\mathbf{x}]\,\mathrm{E}[\mathbf{y}]^{\mathrm{T}},\end{aligned}$$

Za vektor

$$\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \dots & x_m \end{bmatrix}^{\mathrm{T}}$$

od m slučajnih promenljivih sa konačnim drugim momentom, matrica kovarijanse predstavlja

$$\Sigma(\mathbf{x}) = \sigma(\mathbf{x}, \mathbf{x}).$$

Slučajne promenljive kod kojih je kovarijansa jednaka 0 kažemo da su nekorelisane.

Kovarijansa nad konačnim skupom od N opservacija i K promenljivih predstavlja matricu KxK

$$\bar{\bar{q}} = [[q_{jk}]]$$

gde je svaki element matrice računa kao kovarijansa između promenljive j i promenljive k

$$q_{jk} = \frac{1}{N}\sum_{i=1}^{N}(x_{ij} - E(x_j))\,(x_{ik} - E(x_k))$$

j,k=1,2,…K.

B.5 Korelacija i zavisnost

Koeficijent korelacije (eng. *correlation*) između dve slučajne promenljive X i Y sa očekivanim vrednostima μ_X i μ_Y i standardnom devijacijom σ_X i σ_Y je definisan kao

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y},$$

gde je E očekivana vrednost. Ova korelacija je poznata i pod nazivom Personova korelacija. Personova korelacija je definisana samo u slučaju da je standardna devijacija konačna i različita od 0.

Za razliku od kovarijanse Personova korelacija je u intervalu [-1,1]. Ona ima vrednost 1 ili -1, i u slučaju da postoji idealna linearna veza između promenljivih, pri čemu znak određuje da li se radi o pozitivnoj ili negativnoj linearnosti. Vrednosti između (-1,1) određuju stepen linearne zavisnosti, dok vrednosti blizu 0 nam govore o nedostatku linearne zavisnosti (promenljive su nekorelisane). U slučaju da je Personova korelacija 0 to nam ne govori da su promenljive nezavisne.

Primer. Pretpostavimo da je promenljiva X simetrično distribuirana oko 0 i da je promenljiva $Y = X^2$. Personova korelacije je jednaka 0 iako je jasno da su promenljive zavisne.

Ako u uzorku od n opservacija vrednosti promenljivih X i Y obeležimo sa x_i i y_i gde je i = 1, 2, ..., n, tada koeficijent korelacije između promenljivih X i Y se može izračunati kao

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

gde su x i y prosečne vrednosti od X i Y respektivno, a s_x i s_y su standardne devijacije od X i Y.

Ovo može biti zapisano kao

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\,\sqrt{n\sum y_i^2 - (\sum y_i)^2}}.$$

U realnosti granice koeficijenata korelacije ne mogu biti 1 ili -1, već vrednosti koeficijenta korelacije pripadaja manjem opsegu.

Matrica korelacije od n slučajnih promenljivih $X_1, \ldots, X_n$ je matrica dimenzije $n \times n$, pri čemu su elementi matrice koeficijenti korelacije svake dve promenljive $\text{corr}(X_i, X_j)$. Matrica koeficijenata korelacije je simetrična je važi: $\text{corr}(X_i, X_j) = \text{corr}(X_j, X_i)$.

B.6 Varijansa

Varijansa (eng. *variance*) nam govori koliko su vrednosti jedne promenljive „raširene“. To je jedan od načina da se opiše koliko brojevi „beže“ od prosečne vrednosti. Varijansa predstavlja prvi moment distribucije i veoma lako se računa.

Ako slučajna promenljiva X ima očekivanu vrednosti $\mu = \text{E}[X]$ tada varijanca od X je kovarijaca promenljive X sa samom sobom.

$$\begin{aligned}\operatorname{Var}(X) &= \operatorname{Cov}(X, X) \\ &= \operatorname{E}\left[(X - \mu)(X - \mu)\right] \\ &= \operatorname{E}\left[(X - \mu)^2\right].\end{aligned}$$

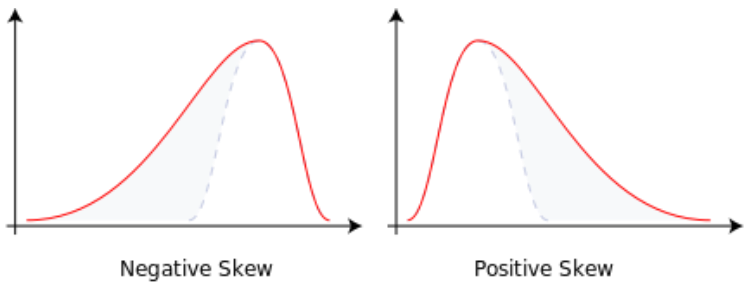
Ako je promenljiva X kontinulana i diskretna tada je varijansa može napisati i kao

$$\begin{aligned}\operatorname{Var}(X) &= \operatorname{Cov}(X, X) \\ &= \operatorname{E}[XX] - \operatorname{E}[X]\operatorname{E}[X] \\ &= \operatorname{E}\left[X^2\right] - (\operatorname{E}[X])^2.\end{aligned}$$

B.7 Skju

Skju (eng. Skewness) predstavlja statističku meru koja nam opisuje položaj krive distribucije u odnosu na prosečnu vrednost.

Kvalitetna interpretacija skju je komplikovana. Za unimodalnu distribuciju negativan skju nam govori da je „rep” na levoj strani krive distribucije „duži“ ili „deblji” u odnosu na desni.



Slika 93. Geometrijska interpretacija skju-a

Pozitivan skju nam govori obratno. Na slici (Slika 93) su dati primeri pozitivnog i negativnog skju-a.

U slučaju da je na jednoj strani rep „duži”, a na drugoj „deblji” skju nije lako interpretirati. U slučaju da sklju ima vrednost nula to ukazuje da su repovi krive distribucije dobro

izbalansirani ili da je jedan rep „mršav" i "dugačak", a drugi „debeo" i „kratak". U slučaju da imamo multimodalnu distribuciju interpretacija je još komplikovanija. Skju ne određuje odnos između srednje vrednosti i medijane.

Skju od slučajne promenljive *X* označen kao

$$\gamma_1 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} = \frac{\mathrm{E}\left[(X-\mu)^3\right]}{(\mathrm{E}\left[(X-\mu)^2\right])^{3/2}} = \frac{\kappa_3}{\kappa_2^{3/2}} \ ,$$

gde μ_3 treći momenat od proseka μ, σ je standardna devijacija a E očekivana vrednost.

Skju se često obeležava i kao *Skew[X]*.

Za uzorak od n opservacija *Skew[X]* je

$$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\right)^{3/2}} \ ,$$

gde je $\overline{x}$ prosečna vrednost promenljive x u uzorku, m_3 treći centralni moment na uzorku, m_2 predstavlja varijancu na uzorku.

B.8 Kurtosis

Kurtosis (grčki κυρτός, *kyrtos* or *kurtos*) je statistička mera koja opisuje oblik krive distribucije.

Tačno tumačenje Personove mere kurtosis se osporava. Klasična interpretacija se odnosi samo na simetrične distribucije gde je Skew[*X*]=0. U tom slučaju ona meri učestalost (stepen) „šiljaka" u distribuciji kao i odsustvo repova. Po nekim statističarima kurtosis meri odustvo „ramena" u distiribuciji pri čemu rame predstvlja deo između šiljka i repa (ovo je slobodna interpretacija).

Četvrti standardni moment je definisan kao

$$\beta_2 = \frac{\mathrm{E}[(X-\mu)^4]}{(\mathrm{E}[(X-\mu)^2])^2} = \frac{\mu_4}{\sigma^4}$$

gde je μ_4 četvrti momenat od proseka a σ standardna devijacija

Kurtosis definisan kao *excess kurtosis* je

$$\gamma_2 = \frac{\kappa_4}{\kappa_2^2} = \frac{\mu_4}{\sigma^4} - 3$$

Broj 3 na kraju formule se često objašnjava korekciom da bi se kurtosis od normalne distribucije izjednačio sa 0. Četvrti standardizovani moment može biti najmanje 1 pa kurtosis uzima vrednosti od -2 pa naviše.

Za uzorak od n vrednosti kurtosis se računa kao

$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\right)^2} - 3$$

B.9 Distribucija frekvencije (eng. *frequency distribution*)

Distribucija frekvencije predstavlja frekvenciju pojavljivanje vrednosti jedne ili više promenljivih unutar jedne grupe ili intervala. Najčeće se koristi distribucija frekvencije jedne (eng. *univatiate frequency distribution*) odnosno dve promenljive (eng. *bivariate joint frequency distribution*).

B.10 Gini koeficijent

Gini koeficijent meri nejednakost raspodele frekvencije jedne ili više promenljivih (npr. nivoa prihoda i događaja da je klijent kupio proizvod). Gini koeficijent 0 označava apsolutnu jednakost npr. da svi klijenti koji kupuju proizvod pripadaju podjednako istim grupama prihoda, dok 1 (odnosno 100%) označava apsolutnu nejednakost.

Definiciju Gini koeficijenta (eng. *bivariate join frequency distribution*) gde je jedna promenljiva kategorička (nominalna) a druga ciljna (binarna *target* promenljiva) data je pomoću algoritma:

1. Prvo se sortiraju podaci (grupe) u opadajućem poretku po učestalosti događaja u svakoj grupi. Prepostavimo da imamo m grupa od 1,2,…m. Prva grupa ima najveću učestalost događaja (pogodaka).
2. Za svaku od ovih grupa računa se broj događaja (n_i^{event}) broj nedogađaja ($n_i^{nonevent}$) u grupi i. Gini koeficijent se računa pomoću formule

$$Gini = \left(1 - \frac{2\sum_{i=2}^{m}(n_i^{event} * \sum_{1}^{i-1} n_i^{nonevent}) + \sum_{i}^{m}(n_i^{event} * n_i^{nonevent})}{N^{event} * N^{nonevent}}\right) * 100$$

   gde je N^{event} and $N^{nonevent}$ ukupan broj događaja i „nedogađaja“ u populaciji.

B.11 Hi-kvadrat selekcija

Jedan od najčećih problema u istraživanju podataka je proučavanje asocijacija između binarne ciljne promenljive Y i nominalne (kategoričke) promenljive X koja ima k grupa ($S = \{1, \ldots, K\}$) . Problem se ogleda u nalaženju pravila binarne podele nominalne promenljive X koja predviđa binarnu promenljivu Y. Promenljiva X se može transformisati u binarne promenljive na sledeći način:

$$X^{(S_j)} = \begin{cases} 1 & \text{if } X \in S_j \\ 0 & \text{otherwise,} \end{cases}$$

gde je S_j , $j = 1, \ldots, 2^K - 2$, j-ti podsukup skupa S (j-ti element particionog skupa P(S)).

Hi-kvadrat test (eng. *chi-squared test*) se često koristi kao test asoscijacije između dve binarne promenljive koristeći N nezavisnih opservacija. U ovom slučaju p-vrednost (eng. *p-value*) hi-kvadrat testa može se koristiti kao mera asocijacije. X dobro predviđa Y ako je p vrednost jako mala. Problem se svodi na računanje maksimalnog hi-kvadrata za promenljive Y i $X^{(Sj)}$.

$$\chi^2_{max} = \max_{S_j \in S} \chi^2_{S_j},$$

Ako je P(hi-kvadrat>3.84)=0.05 tada p-vrednost<0.05. Sledeća tabela daje p-vrednosti i odgovarajući hi-kvadrat.

p-vrednost	Minimalni hi-kvadrat
0.01	6.635
0.05	3.84
0.10	2.706

Tabela 10. P-vrednosti koji odgovaraju minimalnom hi-kvadratu

Izbor podskupa promenljivih (X_1,X_2,…X_n) koje dobro predviđaju Y možemo uraditi na sledeći način:

Izračunati maksimalni hi-kvadrata za svaki par promenjivih (X_i,Y), i=1,2,…n.

Promenljive koje imaju hi-kvadrat veći od unapred određene granice (Tabela 10) biće korišćene u daljoj analizi.

Više o hi-kvadrat statistici:

Maximally selected chi-square statistics and binary splits of nominal variables, Anne-Laure Boulesteix, Department of Statistics, University of Munich, October 18, 2005

B.12 Analiza glavnih komponenti

Analiza glavnih komponenti (eng. *Principal Component Analysis-PCA*) predstavlja jednu od najjednostavnih multivarijantih tehnika. Ona se primenjuje kada je veliki broj promenljivih u skupu redundantan, odnosno kada se više promenljivih odnosi na istu dimenziju i kada ne pružaju dodatnu informaciju koja već nije obuhvaćena nekom drugom promenljivom. Geometrijski gledano, to znači da na prostoru od *k* dimenzija imamo *p* promeljivih pri čemu je *k*<*p*. Očekuje se da će *k* najvećih glavnih komponenti biti dovoljno da opiše podatke u skupu. Na ovaj način polazni prostor od p dimenzija redukujemo projektovanim prostorom od k dimenzija.

Cilj analize je da se uzme *p* promenljivih (X_1, X_2,...,X_p) i da se pronađe kombinacija istih da bi se izračunale nove promenljive (Z_1,Z_2,...,Z_k) koje međusobno nisu u korelaciji i koje će opisivati varijacije podataka na sličan način kao i polazne. Nepostojanje korelacije znači da nove promenljive mere međusobno različite „dimenzije“ podataka i njihove varijanse su poređane u opadajući niz (Var(Z_1)≥ Var(Z_2)≥...≥ Var(Z_k)).

Promenljive Z predstavljaju zapravo glavne komponente. Kada se radi analiza glavnih komponenti, želja je da varijanse većine promenljivih Z budu toliko male da su zanemarljive. U tom slučaju, veći deo varijacija originalnih podataka se može adekvatno opisati sa svega nekoliko glavnih komponenti.

Analiza glavnih komponenti ne uspeva uvek u tome da veliki broj originalnih varijabli X smanji na mali broj izvedenih varijabli Z. Ako originalne varijable nisu u korelaciji, analiza neće postići željeni rezultat tj. nećemo imati projekciju prostora ulaznih promenljivih. Najbolji rezultati se postižu kada su originalne varijable u visokoj korelaciji, bilo pozitivnoj ili negativnoj.

B.13 Linearna regresija

Linearna regresija (eng. *Linear least squares regression*) je najrasprostranjeniji metod modelovanja. Osim svoje velike raspostranjenosti ona igra osnovu ulogu u mnogim drugim metodama nelinearne regresije.

Linearna regresija se koristi za estimaciju bilo kojeg skupa podataka sa funkcijom:

$$f(\vec{x};\vec{\beta})=\beta_0+\beta_1 x_1+\beta_2 x_2+\dots$$

U statističkom smislu svaka funkcija koja zadovoljava formulu iznad nazivamo lineranom funkcijom ako važi:

- svaka promenljiva u funkcije se množi sa nepoznatim parametrom
- postoji najviše jedan parametar koji „pripada“ jednoj promenljivoj
- svaki pojedinačni umnožak se sabira

Primeri lineranih funkcija u statističkom smislu:

$$f(x;\vec{\beta})=\beta_0+\beta_1 x+\beta_1 x_2,$$

$$f(x;\vec{\beta})=\beta_0+\beta_1 ln(x),$$

$$f(x;\vec{\beta})=\beta_0+\beta_1 sin(x)+\beta_2 sin(2x)+\beta_3 sin(3x).$$

Primer nelinerarnih modela:

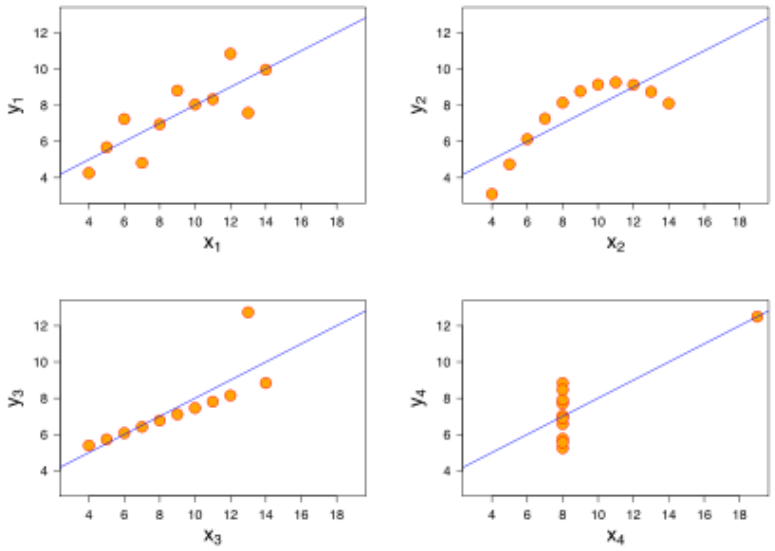
$$f(x;\vec{\beta})=\beta_0+ \beta_0\beta_1 x_1,$$

Linear least squares regression je dobila ime i po načinu na osnovu kojeg se procenjuju (izračunavaju) nepoznati parametri β. To je metod najmanjih kvadrata. U ovoj metodi nepoznati parametri su procenjeni smanjenjem zbira kvadrata odstojanja od podataka i modela. Proces minimizacije redukuje sistem od N jednačina sa p nepoznatih (N je broj opservacija u uzorku, a p-1 broj nezavisnih promenljivih), dobijenih iz podataka, na određen sistem od p jednačina sa p nepoznatih koji ima rešenje.

Ovaj metod je našao svoje mesto kao primarni alat u procesu modelovanja zbog svoje efikasnosti i potpunosti. Mnogi poslovni procesi su inherentno linearni (preko malih opsega) pa se mogu dobro aproksimirati pomoću lineranog modela. Dobri rezultati se mogu dobiti čak i sa malim skupom podataka. Teorija asocijacija sa linearnom regresijom je dobro poznata i omogućava izradu različitih tipova lako interpretabilnih statističkih intervala za predikciju, kalibraciju i optimizaciju koji mogu biti korišćeni da daju odgovore na različita naučna, inžinjerska i poslovna pitanja.

Osnovni nedostatak linerarne regresije su ograničenja u oblicima koje linearni modeli mogu opisati koristeći veće intervale promenljivih. Ovi modeli su veoma osetljivi na vrednosti koje značajno odstupaju od proseka (eng. *outliars*). Jedan ili dva outlier-a može

iskriviti rezultat (Slika 94). Linerarni modeli koji u nezavisnim varijablama imaju nelinearnost (Slika 94) loše opisuju problem pogotovo ako u nezavisnim promenljivima imamo ekstremne vrednosti. Ovo znači da linearni modeli ne mogu opisati problem kod kojih podaci ne mogu biti prikupljeni u „regionu interesa“.



Slika 94. Linearna regresija nad različitim skupovima podataka

B.13.i Kada linearna regresija nije dobra?

Kod linearne regresije želimo da zavisnu promenljivu Y opišemo linearnom funkciojom od nezavisnih promenljivih X. Na ovaj način dobijemo *i* jednačina:

$$Y_i=\alpha+\beta_1 X_{i1}+\beta_2 X_{i2},+\ldots+\beta_p X_{ip}+\varepsilon_i$$

gde su $\beta_1,\beta_2,\ldots,\beta_p$ regresioni koeficijenti,

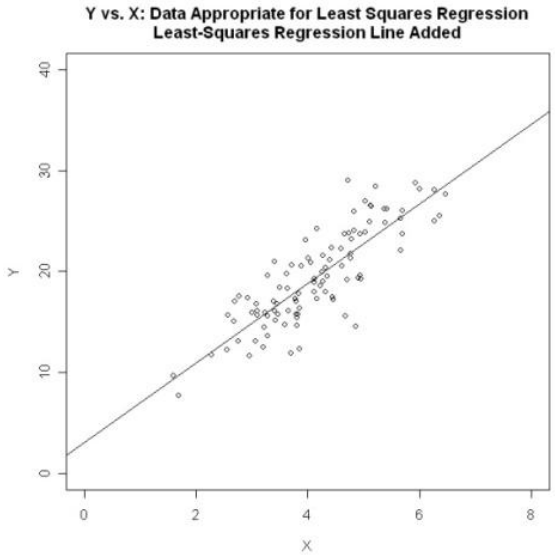
$X_{i1},X_{i2},\ldots,X_{ip}$ odgovarajuće vrednosti promenljivih X u i-toj opservaciji,

α presek (slobodni koeficijent eng. *intercept*),

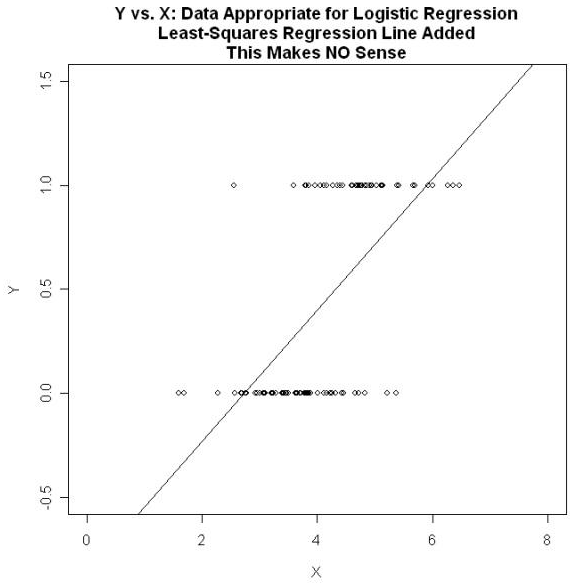
ε_i slučajna greška.

Problem je u tome što slučajna greška ima normalnu distribuciju oko 0. Ovo znači da **Y mora biti kontinualna** promenljiva, a ne binarna. Pretpostavimo da imamo samo dve promenljive X i Y, pri čemu je X kontinualna nezavisna promenljiva, a Y zavisna promenljiva.

Ako je Y kontinualna promenljiva tada možemo napraviti dobar model pomoću funkcije $Y=\alpha+\beta X+\varepsilon$ (Slika 95). U slučaju da je Y binarna promenljiva tada linearna regresija nema smisla (Slika 96).



Slika 95. Linearna regresija Y=f(X) gde je Y kontinualna promenljiva



Slika 96. Linearna regrecija Y=f(X) gde je Y binarna promenljiva

B.13.ii Ograničenja i pretpostavke

Ograničenja koja postoje kod modelovanja linerarnom regresijom su:

- U slučaju da imamo više promenljivih p od opservacija N logistička regresija ne može biti primenjena jer imamo manje jednačina od nepoznatih.
- U slučaju p=N moguće je napraviti model samo ako postoji linearna zavisnost Y od X
- U najvećem broju slučajeva p<N i obično imamo dovoljno informacija da izračunamo regresione koeficijente

Pretpostavke koje moraju biti ispunjene prilikom razvoja modela su:

- Prediktorske promenljive X se tretiraju pre kao fiksne vrednosti nego slučajne promenljive. Ovo znači da se pretpostavlja da su prediktorske promenljive perfektno tačne. U slučaju da prediktorske promenljive imaju na jednom delu uzorka grešku ona će značajno uticati na model.
- Zavisna promenljiva je linearna kombinacija regresionih paramatara. Ova pretpostavka je manje restriktivna nego što izgleda pošto se nezavisne promenljive X tretiraju kao fiksne vrednosti. Linearnost je ograničenje samo nad regresionim parametrima. Same prediktorske promenljive se mogu proizvoljno menjati, ustvari može da se doda više izvedenih prediktorski promenljivh iz baznih promenljivih ne vodeći mnogo računa o njihovoj linerarnosti. Ovo čini regresiju izuzetno moćnom.
- Odsustvo multikolinearnosti kod prediktor promenljivih. Za ovaj metod matrica X mora imati maksimalni rank p. U suprotom imamo stanje poznato kao kolinerarnost prediktor promenljive. U ovom slučaju nemamo jedinstveno rešenje za parametre regresije. U najboljem slučaju mi možemo da identifikujemo neke parametre regresije i suzimo prostor definisan matricom X.

B.14 Logistička regresija

Pomoću linearne regresije želimo da modelujemo očekivanu vrednost E(Y) zavisne promenljive Y. Tada imamo jednačine:

$$E(Y_i)=\alpha+\beta_1 X_{i1}+\beta_2 X_{i2},+\ldots+\beta_p X_{ip}+\varepsilon_i$$

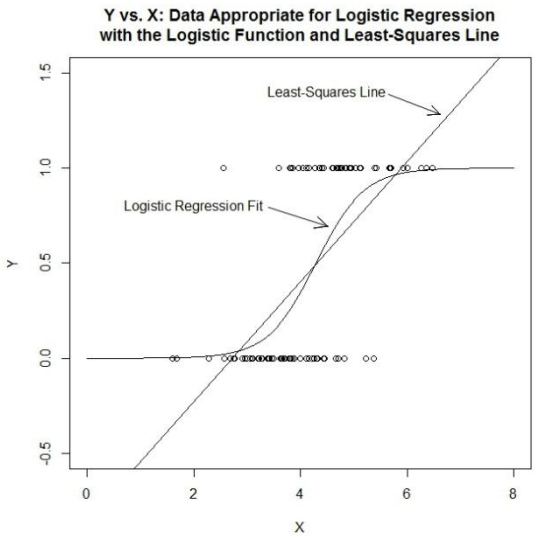
Ako je Y binarna promenljiva sa vrednostima 0 i 1, tada E(Y)=p , gde je p verovarnoća da je Y=1 i ako levu stranu jednačine zamenimo sa funkcijom koja vraća vrednosti od (0,1) dobijemo

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \,.$$

Na ovaj način možemo koristiti isti algoritam za računanje regresionih koeficijanata kao i kod linearne regresije pri čemu verovatnoću P(Y=1) računamo na osnovu formule:

$$E(Y_i) = p_i = \frac{e^{\alpha+\beta_1 X_{i1}+\beta_2 X_{i2}+\cdots+\beta_p X_{ip}}}{1+e^{\alpha+\beta_1 X_{i1}+\beta_2 X_{i2}+\cdots+\beta_p X_{ip}}}$$

Na slici (Slika 97) prikazane su funkcije logističke i linearne regresije. Funkcija logističke regresije ima vrednost između 0 i 1 i nema nelogičnih vrednosti kao kod linearne funkcije (npr. za X=8 dobije se očekivana vrednost od Y=1,5).



Slika 97. Funkcija logističke regresije

Ograničenja koja postoje kod modelovanja linearnom regresijom prisutna su i kod modelovanja logističkom regresijom.

Literatura

1. ***Leksikon bankarstva*** – Dobrivoje Milojević, ISBN 86-904813-0-3, MeGraf 2003
2. ***The Data Warehouse Lifecycle Toolkit. Second Edition*** – Ralph Kimball, Margy Ross, Waren Thornthwaite, Joy Mundy Bob Backer, ISBN 978-0-470-14977-5, Wiley 2007
3. ***Building the Data Warehouse*****. 3st Edition** – Bill Imnon, ISBN: 0-471-08130-2 Wiley 2002
4. ***Data Preparation for Analytics Using SAS*** - Gerhard Svolba, ISBN-13: 978-1599940472, SAS Press
5. ***Data Preparation for Data Mining*** - Dorian Pyle, ISBN 1-55860-529-0, Morgan Kaufmann Publishers
6. ***KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW*** – Ana Avezado, Manuel Filepe Santos, ISBN: 978-972-8924-63-8 © 2008 IADIS
7. ***Data Mining: Concepts and Techniques, Second Edition*****,** Jiawei Han**,** Micheline Kamber, ISBN 13: 978-1-55860-901-3, Morgan Kaufmann Publishers
8. ***Principles of Data Mining*** - David Hand , Heikki Mannila, Padhraic Smyth, ISBN-13: 978-0262082907, The MIT Press © 2001
9. ***Logistic Regression Using SAS: Theory and Application, Second Edition*** - Paul D. Allison, ISBN-13: 978-1599946412, SAS

15. ***Missing Data*** - Paul D. Allison, ISBN-13: 978-0761916727, SAGE University Paper, 2002
16. ***Predictive Models Based on Reduced Input Space That Uses Rejected Variables*** **-** Taiyeong Lee, David Duling, and Dominique Latour , SAS Institute Inc., Cary, NC, 2009 (Paper 111-2009)
17. ***Logistic Regression in Rare Events Data*** - Gary King, Langche Zeng, http://gking.harvard.edu/files/abs/0s-abs.shtml , 2001
18. ***Maximally selected chi-square statistics and binary splits of nominal variables***, Anne-Laure Boulesteix, Department of Statistics, University of Munich, October 18, 2005
19. ***Getting Started with SAS(R) Enterprise Miner(TM) 12.1***, Cary, NC: SAS Institute Inc., SAS Institute Inc 2012
20. ***SAS Enterprise Miner 12.3: Administration and Configuration, Second Edition***, Cary, NC: SAS Institute Inc., SAS Institute Inc 2012
21. ***SAS 9.3 Product Documentation,*** (http://support.sas.com/documentation/93/index.html)
22. ***Overrepresentation - "SAS"-Oversampling***, Guido Deutsch, 2010 (http://www.data-mining-blog.com/tips-and-tutorials/overrepresentation-oversampling/)
23. ***LogisticRegressionAnalysis.com*** (http://logisticregressionanalysis.com/)
24. ***Data Miners Blog*** (http://blog.data-miners.com/ , http://www.data-mining-blog.com)
25. ***Business intelligence (BI):How to build successful BI strategy***, Prashant Pant, Delloite, 2009