Analiza i predviđanje PM2.5 čestica u gradu Guangzhou

Aleksandra Borisavljević, IN17/2018, aleksandra.borisavljevic99@gmail.com

I. UVOD

Tema izveštaja je analiza i predviđanje PM2.5 čestica u gradu Guangz hou u Kini. Spadaju u grube i teške čestice i stoga imaju tendenciju taloženja. PM2.5 čestice imaju prečnik 2.5 mikrometara. One mogu doći iz više izvora kao što su strujna postrojenja, toplane, individualna ložišta, motorna vozila, avioni, gorenje drveta, šumski požari i mnogi drugi. Bitno je napomenuti da su vrlo štetne za ljudsko zdravlje i da preveliko izlaganje može dovesti do trajnih respiratornih problema kao što su astma, hronični bronhitis i srčana oboljenja. Stoga je analiza PM2.5 čestica od velikog značaja. Pravljenjem modela može da se na osnovu drugih parametara predvidi njihova količina i time se poboljšaju svakodnevni uslovi kako bise smanjila količina ovih štetnih materija.

II. BAZA PODATAKA

Skup podataka koji se obrađuje je GuangzhouPM20100101_20151231.csv. Sadrži 52584 uzoraka i 17 obeležja. Jedan uzorak predstavlja izmerene različite vremenske parametre i količina PM2.5 čestica u jednom satu. U skupu se nalaze podaci od 2010. godine do 2015. godine.

Kategorička obeležja koja se javljaju su: *year, month, day, hour, season, cbdw* (pravac vetra). Dok su numerička obeležja: *No, PM_City Station, PM_5th Middle School, PM_US Post, DEWP* (temperatura rose/kondenzacije u Celzijusima), *HUMI* (vlažnost vazduha u %), *PRES* (vazdušni pritisak u hPa), *TEMP* (temperatura u Celzijusima), *Iws* (kumulativna brzina vetra u m/s), *precipitation* (padavine na sat u mm) i *Iprec* (ukupne padavine u mm). Obeležja koja započinju sa PM predstavljaju koncentracija PM2.5 čestica na nekoliko lokacija merene u ug/$m^{3.}$

III. ANALIZA PODATAKA

Analiza podataka je započeta najpre odbacivanjem obeležja *No, PM_City Station i PM_5th Middle School.* Obeležje *No* je odbačeno zato što vrednosti ne daju značajniju informaciju o uzorku i jednistvene su za svaki uzorak. Dok su druga dva obeležja odbačena zbog samog zadatka.

A. *Analiza null vrednosti*

Drugi korak je bila analiza null vrednosti u skupu. Uočeno je da za 9 obeležja postoji samo jedna null vrednost i daljom analizom je zaključeno da se radi o tačno jednom uzorku. Kao najbolje rešenje odabrano je da se taj uzorak obriše iz skupa, zato što se time neće uneti greška u skup podataka, a takođe 1 uzorak ne predstavlja značajniji udeo u celom skupu.

Takođe uočeno je 40 % null vrednosti kod obeležja PM_US Post. Tom problemu se moralo ozbiljno pristupiti zato što se radi o obeležju koje će biti potrebno predviđati. Ne bi bilo odgovorno obrisati skoro polovinu uzoraka iz baze. Kao najbolje rešenje, odlučeno je da se skup podataka grupiše po godinama i tako je utvrđeno da za 2010. godinu ne postoji ni jedan podatak za PM2.5 čestice, dok za 2011. nedostaje vise od 70%. Zaključeno je da nije moguće na adekvatan način popuniti te podatke. Ako bi se popunili srednom vrednošću unela bi se velika greška u podatke, a samim tim i u model za predviđanje. Kao najbolje rešenje odlučeno je da se 2010. i 2011. godina obrišu iz baze, a da se ostali nedostajući podaci za obeležje PM2.5 čestica zamene sa susednom vrednošću. Time se postiglo da se ne izbace svi uzorci sa nedostajućim podacima, a da se pritom ne ugrozi tačnost podataka.

Nisu svi null podaci vidljivi odmah. Za neke je potrebno izanalizirati statističke parametre da bi se mogle uočiti nepravilnosti. Prvo takvo obeležje je temperatura rose/kondenzacije u celzijusima (DEWP). Primećeno je da postoji minimalna vrednost koja iznosi -9999. Takođe je primećeno da je medijana veća od sredje vrednosti, što takođe ukazuje da postoje neke male vrednosti koje su odvukle srednju vrednost. Daljom analizom je utvrđeno da postoji samo 4 uzorka koja imaju tu vrednost za DEWP obeležje. Stoga je odlučeno da se ta 4 uzorka obrišu iz baze. Nakon brisanja utvrđeno je da je minimum -11, što je potvrdilo mišljenje da su ove vrednosti zabeležene samo kao zamena za null vrednost.

Što se tiče pritiska (PRES), normalni pritisak je između 1009 i 1022 hPa. Barometrijsko očitavanje iznad 1022 hPa se generalno smatra visokim, visok pritisak je povezan sa vedrim nebom i mirsnim vremenom. Srednja vrednost pritisak koji se javlja je 1004 hPa, minimalna vrednsot je 975 hPa dok je maksimalna 1023 hPa. Sva ova merenja potvrđuju da su vrednosti za pritisak, koje se nalaze u skupu podataka, ispravne.

Ukoliko je vlažnost vazduha 100%, to je maksimalna vrednost koja se javlja, govori da je velika verovatnoća da će pasti kiša. Što naravno ne znači da vlažnost vazduha mora biti 100% da bi padala kiša. Velika vlažnost vazduha ne odgovara ljudima. Ljudi se oslobađaju toplote u svom telu u procesu znojenja. Ukoliko je kišno vreme, ova vlaga nema gde da ode. Minimalna vrednost koja se javlja je 13%, a takođe se može videti da 50% uzoraka ima vrednost manju ili jednaku sa 83% za ovo obležje. Sve ove informacije ukazuju da su vrednosti za ovo obeležje uredu.

Utvrđeno je, nakon analize, da se podaci za ostala obeležja nalaze unutar dozvoljenih opsega.

B. *Kategorička obeležja*

Što se tiče kategoričkih obeležja, jedino su vrednosti za obeležje pravac vetra dati slovnim vrednostima. Što znači da je neophodno, pre početka obuke modela prevesti ih u numeričke vrednosti. To je učinjeno na dva načina. Prvi način je da se iskoristi algoritam za kodovanje dummy varijabli. Na ovaj način se na osnovni skup podataka dodaje još obeležja. Broj obeležja koji se dodaje odgovara broju mogućih vrednosti za obeležje pravac verta umanjeno za jedan. Ovaj postupak je odgovarajući u ovoj situaciji zato što je broj vrednosti za obeležje pravac vetra 5 i time se dimenionalnost neće značajno povećati. Ako bi se radilo o većem broju mogućnosti onda bi se trebao potražiti negu drugi način. Drugi način je da se oznake za strane sveta zamene za stepene uglova. Prilikom obučavanja biće testirana oba načina, u cilju da se dobije što je moguće manja greška.

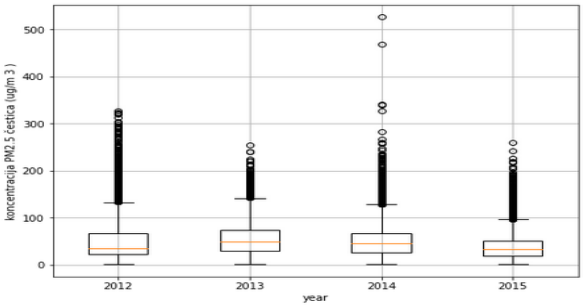
IV. OPIS ZAVISNOSTI OBELEŽJA

A. *PM2.5 i ostala obeležja*

Veoma je bitno dobro analizirati zavisnosti između PM2.5 obeležja, koje predstavlja predmet predviđanja, i ostalih obeležja.

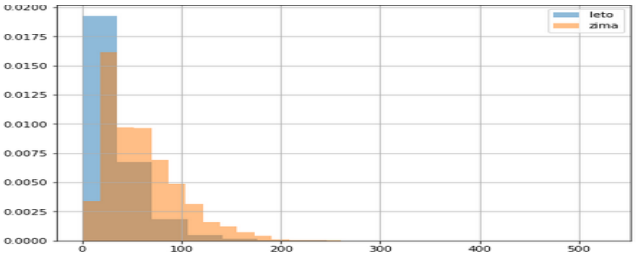
Na grafikonu (Slika 1) može se uočiti da se u 2014. godini javljaju najveće vrednosti za koncentraciju PM2.5 čestica. Najveća vrednost koja se pojavljuje je 526 ug/m^3. Po standardima vrednsoti koje su oko 500 predstavljaju ozbiljan rizik od nastanka respiratornih problema u opštoj populaciji. U ovakvim situacijama preporučuje se da osobe sa respiratornim i srčanim problemima, stariji i deca treba da ostanu u zatvorenom prostoru.

Može se uočiti da je u 2015. godini naglo opalo pojavljivanje velikih koncentracija zbog čega može da se pretpostavi da su se grad više posvećuje očuvanju zdravlja ljudi i sredina. Takođe vidi se da je se u 2013. godini nisu javljale ekstremne vrednosti u toku jednog sata, ali takođe se može uočiti da su se češće javljale veće koncentracije nego što je to slučaj u narednim godinama.



Slika 1: Boxplot - koncentracija PM2.5 čestica kroz godine

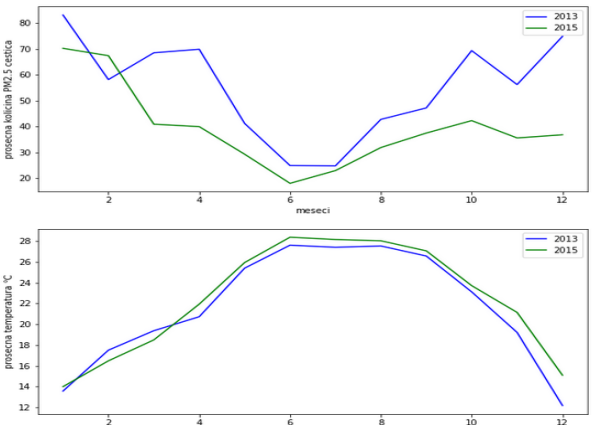
Na histogramu (Slika 2) može se videti da je koncetracije PM2.5 čestica znatno veća tokom zimskog perioda, nego letnjeg. To je potpuno u skaldu sa činjenicom da je tokom hladnijih meseci stanovništvu neophodno grejanje, pri čemu nastaju velike količine PM2.5 čestica. Ugalj koji se koristi je veoma loš, lignit sadrži dosta sumpora i puno teških metala. Što se privatnih ložišta tiče, ona su veliki problem zato što ljudi lože i ono što ne bi trebalo i tako emituju veliki broj štetnih čestica.



Slika 2: Histogram-prikaz koncentracije PM2.5 čestica u toku zime i leta

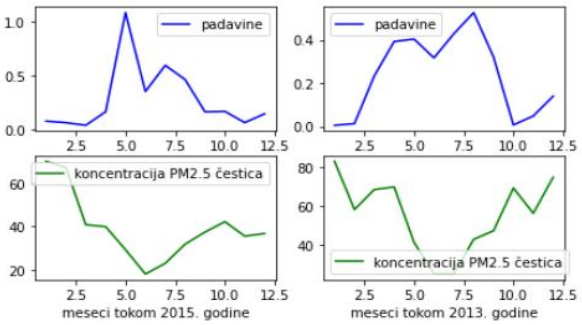
Na sledeća dva grafikona (Slika 3) može se uporedo videti promena prosečne mesečne količine PM2.5 čestica i prosečne mesečne temperature u 2013. i u 2015. godinu.

Može se uočiti je promena ova dva obeležja tokom 2013. i 2015. godine veoma slična. Vidi se da su najveće zabeležene temperature tekom juna, jula i avgusta, i tada koncentracija PM2.5 čestica bude najmanja. Dok prilikom pada temperature koncentracija raste, što je u skladu sa prethodnim grafikonom. Može se uočiti da su najhladniji meseci decembar,januar i februar.



Slika 3: Plot – uporedni prikaz promene koncentracije PM2.5 čestica i promene temperature kroz mesece u dve godine

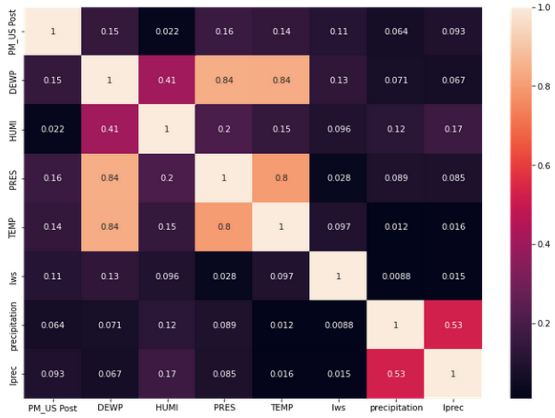
Na sledećem linijskom dijagramu (Slika 4) može se uočiti zavisnost količine padavina i koncentracije PM2.5 čestica posmatrano kroz dve godine. U prvom redu prikazana je promena količine padavina kroz mesece, gde se vidi da je u 2015. godini najviše padavina bilo tokom meseca maja i jula, dok se u junu vidi nagli pad. U 2013. godini padavine se u velikoj količini javljaju tokom prolećnih i letnjih meseci, dok je mesec kada je zabeleženo najviše padavina bio avgust. Što se tiče koncentracije PM2.5 čestica tokom obe godine najviše koncentracije se javljaju u zimskom periodu. Kada bi se ove dve analize uporedile moglo bi se zaključiti da uglavnom kada ima više padavina tada koncentracija PM2.5 čestica bude manja. Kiša utiče da se PM2.5 čestice uklone iz vazduha. Odnosno slaba kiša zimi povećava koncentraciju zagađujućih materija.



Slika 4: Plot – zavisnost padavina i koncentracije PM2.5 čestica kroz dve godine

B. Međukorelacija obeležja

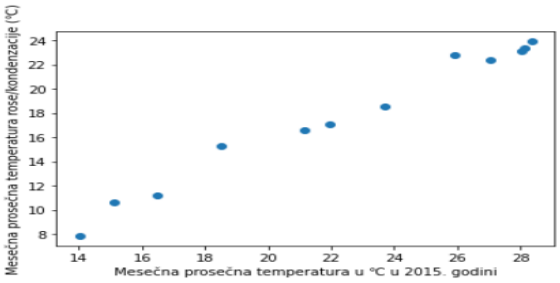
Na matrici korelacija (Slika 5) može se uočiti da je mali broj obeležja u korelaciji. Korelacija govori da li su dva obeležja povezana i na koji način. U korelaciji su obeležja temperatura rose i pritisak, temperatura rose i temperatura. Dok su u malo manjoj korelaciji pritisak i temperatura.



Slika 5: Heapmap – matrica korelacije

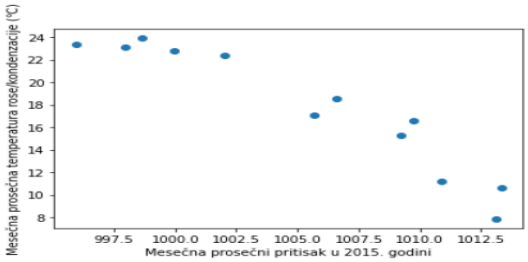
U nastavku sledi ispitivanje da li je u pitanju negativna ili pozitivna korelacija.

Na prvom scatter grafikonu (Slika 6) uočava se da je između temperature rose i temperature pozitivna korelacija, odnosno kada raste temperatura rose raste i temeratura, kada jedna opada, opada i druga.



Slika 6: Scatter – pozitivna korelacija temperature rose i temperature

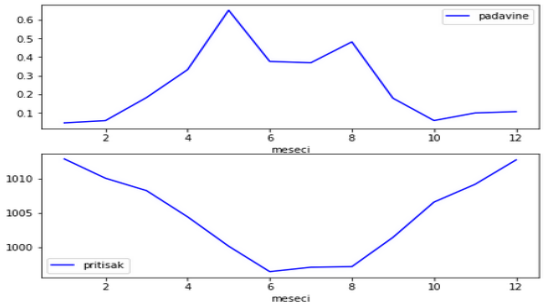
Na sledećem grafikonu (Slika 7) vidi se zavisnost pritiska i temperature rose, tu se uočava negativna korelacija, ondosno kada temperatura rose opada, pritisak raste. Takođe ista situacija je sa teperaturom i pritiskom. Što se slaže sa činjenicom da je topliji vazduh lakši i ređi, pa mu je i pritisak manji.



Slika 7: Scatter – negativna korelacija temperature rose i pritiska

C. Odnosi između preostalih obeležja

Na sledećem linijskom grafikonu (Slika 8) može se uočiti prosečna količina padavina tokom meseci uporedo sa prosečnim pritiskom u toku meseci. Veće količine padavina su zabeležene tokom prolećnih i letnjih meseci. Najveće količine padavina su zabeležene u julu i avgustu. Dok je nagli pad pritiska zabeležen upravo u tim periodima kada su padavine bile najobilnije. Slaže se sa činjenicom da je nagli pad pritiska povezan sa pojavom oluje. Vidi se da je period od juna do jula nepromenljiv što se tiče padavina, dok veoma slična situacija može da se uoči i sa pritiskom koji je u tom periodu takođe nepromenljiv.



Slika 8: Plot - Promena količine padavina i promena pritiska kroz mesece

V. Regresija

Na samom početku neophodno je iz skupa uzoraka izbaciti obeležje koje treba da se predviđa. Zatim je neophodno podeliti skup podataka na skup za obuku i skup za testiranje modela. Skup za bouku dobio je 90% uzoraka.

Nakon što je urađen osnovni oblik linearne regresije sa hipotezom y=b0+b1x1+b2x2+...+bnxn, dobijena je greška koja ukazuje da model koji je obučen pravi apsolutnu grešku 24.8, dok je R2 score iznosio 0,064. R2 score je najbolje da bude što bliži jedinici. Ova mera govori koliko obučen model dobro predviđa vrednosti u odnosu na srednju vrednost. Što znači da ovako kreiran model nema zadovoljavajuće karakteristike. Stoga je neophodne preduzeti mere koje će da dovedu do bolje obučenih modela.

Neki od načina su sledeći: da se razmisli koja obeležja utiču na obeležje koje se predviđa, promena početne hipoteze prilikom obuke modela, normalizacija obeležja, upotreba regularizatora. Regularizatori kažnjavaju pojave dominantnih obeležja.

A. *Selekcija obeležja*

Prvi način da se poboljša model linearne regresije jeste da se izbace obeležja koja ne utiču u velikoj meri na obuku modela. Postoji selekcija obeležja unapred i unazad. U ovom primeru analizirana je selekcija unazad. Za granicu pristanka postavljena je vrednsot od 1%. Stoga sva obeležja sa vrednošću verovatnoće P većom od zadate ne treba da se zadrži u skupu obeležja. Kada se izbace pojedina obeležja, dobra je praksa ponoviti ovaj postupak, kako bi se uvidela mogućnost da se izbaci još neko obeležje.

	coef	std err	t	P>|t|	[0.025	0.975]
const	8020.3825	362.559	22.122	0.000	7309.752	8731.013
year	-4.3331	0.182	-23.791	0.000	-4.690	-3.976
month	-0.6125	0.065	-9.416	0.000	-0.740	-0.485
day	0.0374	0.023	1.661	0.097	-0.007	0.082
hour	0.2260	0.030	7.563	0.000	0.167	0.285
season	0.6572	0.234	2.813	0.005	0.199	1.115
DEWP	-1.9697	0.338	-5.823	0.000	-2.633	-1.307
HUMI	0.4392	0.082	5.334	0.000	0.278	0.601
PRES	0.7120	0.060	11.912	0.000	0.595	0.829
TEMP	1.9418	0.338	5.737	0.000	1.278	2.605
Iws	-0.3188	0.017	-18.491	0.000	-0.353	-0.285
precipitation	-0.2857	0.125	-2.277	0.023	-0.532	-0.040
Iprec	-0.3994	0.040	-10.080	0.000	-0.477	-0.322
cbwd_NE	-5.2968	1.460	-3.628	0.000	-8.158	-2.435
cbwd_NW	4.0996	1.464	2.801	0.005	1.231	6.968
cbwd_SE	-5.4184	1.507	-3.595	0.000	-8.372	-2.464
cbwd_SW	-8.2044	1.558	-5.268	0.000	-11.257	-5.152

Slika 9: Prikaz verovatnoća P za sva obeležja

Može se uočiti (Slika 9) da su to obležja precipitation i day. Nakon što su ova obeležja izbačena, ponovo je primenjen osnovni oblik lenearne regresije, međutim podaci nisu poboljšani.

B. *Normalizacija obeležja*

Normalizacija obeležja je neophodna kako bi se kasnije mogle primeniti mere regularizacije. Normalizacija podrazumeva da se obeležja svedu pod isti opseg ili ista statistička svojstva. Normalizacija podrazumeva svođenje obeležja na nultu srednju vrednost i jediničnu varijansu. Time dobijamo mogućnost kasnijeg upoređivanja i kombinacije obeležja. Normalizacija može dovesti brže do rešenja, ali ne mora dovesti do boljeg.

Nakon normalizacije obeležja izvršena je obuka modela sa početnom hipotezom y=b0+b1x1+b2x2+...+bnxn,međutim rezultat nije poboljšan.

Zatim je primenjena obuka uz početnu hipotrezu koja uključuje interakciju između obeležja koja su međusobno korelisana. Ovde može da se primeti da su rešenja poboljšana. Srednja apsolutna greska iznosi 22.8, dok je R2 score porasla sa 0.06 na 0.18. Što predstavlja bolje karakteristike modela. Međutim i dalje ove vrednosti ne ukazuju da treba stati sa poboljšanjem kvaliteta modela.

Sledeća hipoteza koja je upotrebljena jeste hipoteza koja pored interakcija između obeležja uključuje kvadrate pojedinačnih obeležja. Nakon obuke ovaj model daje bolje karakteristike od prethodnog (Slika 10). Može se videtii da je vrednosti prilagođenog R2 score-a povećana sa 0.186 na 0.219.

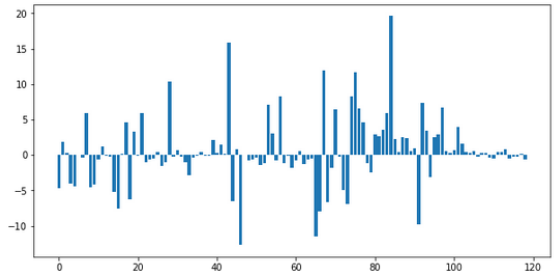
```
Mean squared error:  1022.0583813152846
Mean absolute error:  22.23273631841097
Root mean squared error:  31.969647813438367
R2 score:  0.22231425116798198
R2 adjusted score:  0.2193700649906839
```

Slika 10: Vrednosti mera uspešnosti regresora pri upotrebi klase PolynomialFeatures

Međutim stalno povećavanje stepena i broja međusobnih interakcija ipak nije ideja kojom se treba rukovoditi. Na ovaj nači dimenzija sistema vrlo brzo raste, pa se lako može doći do preobučenog modela. Što svakao nije cilj. Problem koji je ovde nastao je postojanje prevelikih težina. Ako su težine prevelike znači da je došlo do nadprilagođavanja. Može se uočiti da su neke vrednosti koeficijenata toliko male da se na slici ne mogu uočiti, pored koeficijenata koji su mnogo veći od njih.

C. Regularizacija

Neophodno je velike vrednosti težinskih koeficijenata smanjiti, odnosno sprečiti postojanje dominantnih obeležja. Regularizatori se bore protiv natprilagođavanja. Postoje dve mere regularizacije, Lasso i Ridge. U ovom primeru je korišćena Ridge regresije sa stepenom kažnjavanja 5. Nakon primene ove mere regularizacije mere uspešnosti regresora su ostale nepromenjene, međutim desila se velika promena u vrednostima težinskih koeficijenata (Slika 11). Ovaj model je svakako bolji od prethodnog iako su vrednsoti za mere uspešnosti regresora iste.



Slika 11: Vrednosti težinskih koenficijenata nakon primene Ridge regresije

Nakon Ridge regresije isprobana je i Lasso regresija. Međutim rezulati koje je ona dala za nijansu su lošiji od prethodnih, dok su koeficijenti takođe prihvatljivo raspoređeni. Lasso regresija pored što ima ulogu da regularizacije, radi i indirektnu selekciju obležja. Odnosno ako zaključi da je procenjena težina za neko od obeležja jako bližu nuli, tada donosi odluku da to obeležje izbaci iz skupa.

Celokupan postupak određivanja najboljeg modela isproban je i za drugačije formiran skup podataka, gde je kategoričko obeležje za pravac vetra prevedeno u stemene uglova. Međutim svaki korak određivanja davao je nešto veće vrednosti za mere uspešnosti regresora.

Kada se pogleda celokupna analiza dobijenih regresionih modela, može se doneti zaključak da je najbolji model dobijen kada je rađena Ridge regresija, čija početna hipoteza je bila, hipotezu sa kombinovanim obeležjima i sa kvadratima pojedinačnih obeležja. Stepen kažnjavanja Ridge regresije je 5. Ovim modelom dobijaju se najbolje vrednosti za mere uspešnosti regresora. Takođe i najbolje raspoređene vrednosti težinskih koeficijenata.