UNIVERZITET U BEOGRADU, FAKULTET ORGANIZACIONIH NAUKA – MAŠINSKO UČENJE

|  |
| --- |
| Carvana – Sell No Evil |

Miloš Nikić

July 1, 2019

# 1 Opis i razumevanje problema

## Opis problema

Kako je Carvana kompanija koja se bavi prodajom automobila, jedan od najvećih problema sa kojima se suočavaju jeste kupovina automobila koji predstavljaju „kiks“. Auto se smatra „kiksom“ kada poseduje neke mane koje sprečavaju dalju prodaju kupcu. Najčešći razlozi zbog kojih se pojavljuju ovakvi automobili jeste takozvano „vraćanje kilometraže“, neki od mehaničkih problema koje prodavac nije u mogućnosti da identifikuje.

## Izazovi

Da bi kompanija što bolje poslovala potrebno je smanjiti troškove. Pokušaćemo da predvidimo da li je određeni auto dobra kupovina ili ne na osnovu nekih njegovih karakteristika. Time ćemo smanjiti različite troškove kompanije i postići određenu dozu poverenja kod kupaca.

## Važnost

Svakako da je od velike važnosti izvršiti dobru klasifikaciju automobila. Sa aspekta kompanije potrebno je videti kolika je cena greške kada se određeni auto klasifikuje kao loša kupovina, a on je dobra (trošak koliko je moglo da se zaradi na tom automobilu) i kolika je cena greške kada se izvrši loša klasifikacija – auto se klasifikuje kao da je dobra kupovina, a ispostavi se da ima grešku (trošak vraćanja automobila, popravke, gubljenje poverenja korisnika, itd.).

# Opis i razumevanje podataka

Skup podataka se sastoji od 34 različita atributa i 6798 opservacija koje su označene sa nulama, ukoliko je kupovina bila dobra, i jedinicama, kada se kupovina ispostavila lošom.

## Opis atributa

Naš skup podataka sadrži podatke koji su numerički, kategorički i datumskog tipa. Svaka opservacija ima jedinstveni RefId kao identifikator. Atribut „PurchDate“ opisuje stvarni datum, mesec i godinu kada je vozilo kupljeno. „VehYear“ atribut nam govori kada je vozilo proizvedeno. Iz prethodna dva atributa moguće je izvući atribut „VehicleAge“ koji predstavlja razliku prethodna dva. „Auction“ je kategorički atribut koji nam govori na kojoj aukciji je kupljeno vozilo.

Takođe, imamo atribute koji nam bliže određuju samo vozilo poput „Make“, „Model“, „Trim“ i „SubModel“. Još neki od atributa su „Color“, „Transmission“, „Nationality“, „Size“, „WheelType“, „WheelTypeID“.

Kao što se može videti, neki od prethodnih atributa su u korelaciji i prilikom pretprocesiranja podataka potrebno je obratiti pažnu na te atribute.

Zatim imamo atribute koji su vezani za troškove.

„MMRAcquisitionAuctionAveragePrice“, „MMRAcquisitionAuctionCleanPrice“, „MMRAcquisitionRetailAveragePrice“,  
„MMRAcquisitonRetailCleanPrice“, „MMRCurrentAuctionAveragePrice“, „MMRCurrentAuctionCleanPrice“, „MMRCurrentRetailAveragePrice“, „MMRCurrentRetailCleanPrice“,  
„Acquisition“ se odnosi na cenu vozila po kojoj je prodato na aukciji, dok „Clean“ se odnosi na vozilo u dobrom stanju i samim tim su ta vozila češće i skuplja od onih „Average“.

„Auction“ se odnosi na očekivanu cenu vozila na aukciji.

„Retail“ se odnosi na cenu koju je kupac spreman da izdvoji za vozilo.

„VNZIP1“ nam govori o poštanskom broju gde je vozilo kupljeno, a „VNST“ o državi. Ova dva atributa su takođe visoko korelisana.

„PRIMEUNIT“ atribut nam govori o nivou tražnje vozila, dok „AUCGUART“ govori o riziku koji to vozilo nosi sa sobom, tj. garanciju koju prodavac može da ponudi.

Na kraju, imamo atribut „BYRNO“ koji predstavlja jedinstveni broj dodeljen kupcu, „VehBCost“ koje predstavlja cenu vozila. „IsOnlineSale“ da li je prodaja obavljena putem interneta i „WarrantyCost“ trošak garancije.

Atributi koji su od najvećeg značaja su pređeni kilometri, cena vozila kao i starost vozila.

## Disbalans klasa

Ono što je bilo očigledno jeste da postoji disbalans klasa. Čak 87,3% svih opservacija predstavlja dobru kupovinu. Način na koji smo rešili problem jeste uzorkovanje. Uvećali smo manjinsku klasu tako da se broj opservacija izjednačio sa dominantom.

## Vizuelizacija podataka

Kao što se može videti na sledećim graficima atributi VehYear i VehOdo imaju raspodelu nalik normalnoj, kao i VehAge. Sa druge strane, neki atributi imaju pozitivno asimetričnu raspodelu (asimetrične udesno), kao npr. VehBCost i WarrantyCost.

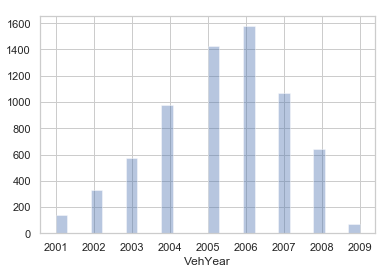


Figure 1 VehYear raspodela

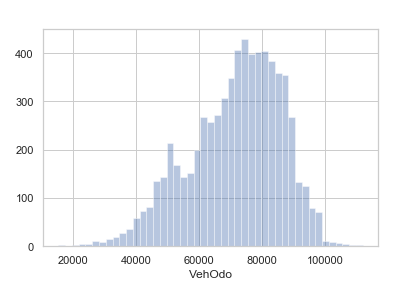


Figure 2 VehOdo raspodela

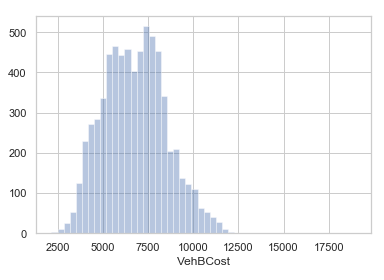


Figure 3 VehBCost pozitivno asimetrična raspodela

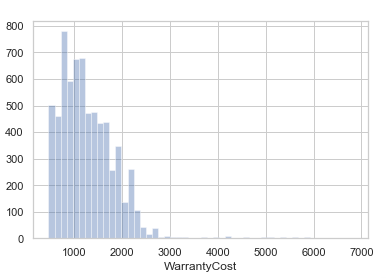


Figure 4 WarrantyCost pozitivno asimetrična raspodela

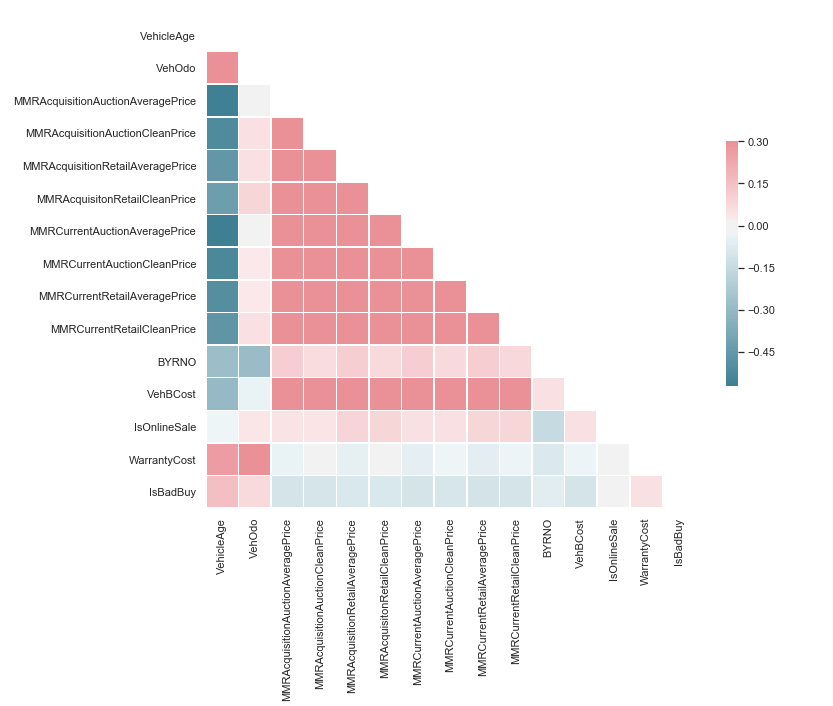


Figure 5 Korelaciona matrica

Iz matrice korelacije vidimo kako atributi vezani za cenu imaju linearnu zavisnost, kao i da je atribut godina automobila u negativnoj korelaciji sa cenom.

## Nedostajuće vrednosti

Kada se radi o nedostajućim vrednostima u celom skupu podataka imamo ukupno 33828.

Dva najkritičnija atributa za imputaciju nedostajućih vrednosti su PRIMEUNIT i AUCGUART jer imamo samo oko 5% podataka za oba atributa. Za numeričke atribute koristili smo metod imputacije srednjom vrednošću, a za kategoričke smo koristili metod najveće frekvencije.



Figure 6 PRIMEUNIT i AUCGUART nedostajuće vrednosti

# Priprema podataka

Atribute PRIMEUNIT i AUCGUART smo izbacili iz dalje analize zbog velikog broja nedostajućih vrednosti. Takođe smo izbacili „RefId“ koji je jedinstveni identifikator svake opservacije. Kada se pogleda pažljivo može se videti da je atribut „WheelTypeID“ samo identifikacioni broj za „WheelType“, pa smo zbog toga ostavili samo „WheelType“. Analogno tome smo izbacili „VNZIP1“ koji predstavlja poštanski broj države u kojoj je vozilo kupljeno. Atribut „VNST“ predstavlja države. Atribut „VehYear“ je izbačen zbog toga što imamo atribut „VehicleAge“.

## Transformacija tipova podataka

Kategoričke atribute smo prebacili u numeričke korišćenjem „dummy“ vrednosti.

Z transformacijom je izvršena normalizacija numeričkih atributa.

## Imputacija nedostajućih vrednosti

Nedostajuće vrednosti za „WheelType“ smo popunili vrednošći Alloy, koja je najčešća.

Za atribute MMR cene smo koristili srednju vrednost za imputaciju nedostajućih vrednosti.

Za sve kategoričke atribute korišćena je tehnika najveće frekvencije.

Atribut „VNST“ koji predstavlja državu u kojoj je vozilo kupljeno je imao veliki broj nedostajućih vrednosti. Na osnovu vrednosti poštanskog broja i skupa podataka pribavljenog sa interneta (taj skup podataka sadrži poštanske brojeve i države), vrednosti koje su falile su dopunjene.

## Izvođenje novih atributa

Iz atributa „Model“ je korišćenjem regularnih izraza izvučen broj ventila, kao i pogon automobila (2WD, 4WD). Dok je iz „SubModel“ atributa izvučen broj vrata svakog automobila.

# Algoritmi

Modeli koji su izabrani za dati problem klasifikacije su: Logistička regresija, K najbližih suseda, Naivni Bajes i Random forest. Za validaciju dobijenih rešenja korišćen je cross-validation pristup. Performanse modela ocenjivane su na sledećim metrikama: precision, recall, accuracy i AUC. Metrika od najvećeg značaja u ovom kontekstu je recall, budući da je poželjniji što manji broj „loših“ kupovina klasifikovanih kao dobre. Osim toga, AUC je dobar indikator klasifikacione sposobnosti modela i može da posluži u njihovom međusobnom poređenju. Rezultati svakog od modela na izabranim metrikama prikazani su u tabeli ispod.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Logistička regresija | KNN | Naivni Bajes | Random forest |
| Precision | 0.222 | 0.173 | 0.122 | 0.32 |
| Recall | 0.031 | 0.167 | 0.717 | 0.01 |
| Accuracy | 0.864 | 0.79 | 0.306 | 0.87 |
| AUC | 0.606 | 0.524 | 0.481 | 0.62 |

## Trening i test set

Za način validacije rešenja odabrana je cross-validation opcija. Razlog ovakvog izbora jeste mogućnost da se prilikom treniranja/testiranja obuhvati čitav skup podataka, bez rizika od loše slučajne podele na trening i test skup.

# Selekcija atributa

Zbog velikog broja atributa, nastalog usred predprocesiranja podataka neophodna je selekcija atributa. Ona je izvršena preko filter metode za izbacivanje neinformativnih atributa, pri čemu se njihov broj smanjio na 19. Pomoću Analize glavnih komponenti broj atributa sveden je na njih 500 , koji objašnjavaju čak 99% varijabiliteta.

# Optimizacija parametara

U svakom od korišćenih algoritama izvršena je optimizacija parametara , pri čemu je optimalno stanje istih mereno pomoću accuracy metrike. Za logističku regresiju podešen je parametar C na optimalnu vrednost 0.001, dok je kod KNN modela izabrana euklidska mera rastojanja i broj najbližih suseda podešen na k=25.

# Balansiranje klasa

Primenom SMOTE tehnike za rešavanje problema disbalansa klasa, početni skup podataka je proširen na 8288 opservacija. Rezultati primene modela na tako formiran dataset su se u svim posmatranim metrikama pokazali daleko boljim, što se i može videti u tabeli ispod.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Logistička regresija | KNN | Naivni Bajes | Random forest |
| Precision | 0.716 | 0.874 | 0.565 | 0.98 |
| Recall | 0.764 | 0.993 | 0.998 | 0.99 |
| Accuracy | 0.73 | 0.925 | 0.614 | 0.98 |
| AUC | 0.804 | 0.804 | 0.614 | 1.00 |

# Zaključak

Nakon postignutih rezultata zaključujemo da je potrebno više se posvetiti predprocesiranju podataka, najviše kategoričkih. U manjku znanja, neke od atributa nismo formirali na željeni način, što bi nas dovelo do manjeg broja atributa kasnije i lakše optimizacije modela. Svakako, za bolju interpretaciju rada modela, za rešavanje ovakvog problema klasifikacije, smatramo da bi trebalo odrediti težine grešaka iste.