Univerzitet u Beogradu, Fakultet organizacionih nauka – Mašinsko učenje

Carvana – Don't get kicked

Janko Gašić, Filip Branović, Ivan Prelić

1. Uvod i opis problema

Jedan od glavnih problema pri nabavci polovnih automobila na aukciji je nedostatak informacija o njihovom stvarnom stanju. Vraćanje kilometraže, prikrivanje mehaničkih problema i oštećenja su neke od strategija kojima trgovci automobila podižu vrednost vozilima na aukciji, što otežava kupovinu, jer ona dolazi sa većim rizikom, i time nastaje nesigurnost kod kupaca. Potencijalni troškovi se odnose na popravku, transport i gubitke pri ponovnoj prodaji automobila. Cilj analize jeste predvideti da li je vozilo oštećeno, uspešnim predviđanjem bi se smanjio broj prevara, iznos neplaniranih troškova i samim tim nezadovoljstvo kupaca.

2. Opis i priprema podataka

Skup podataka pod nazivom "*caravana.csv*" sadrži 6798 instanci (redova), koje su opisane pomoću 34 atributa (kolona), od toga 33 atributa predstavlja nezavisne promenljive, dok je atribut "*IsBadBuy*" izlazni atribut koji nam govori da li je određeni auto neispravan (vrednost atributa je 1) ili je ispravan (vrednost atributa je 0).

Skup podataka se sastoji od 15 kategoričkih i 19 numeričkih promenljivih. Takođe, postoji 18 kolona koje sadrže nedostajuće vrednosti. Neke od kolona smo popunili odgovarajućim vrednostima (medijanom kolone), u zavisnosti od tipa podataka, dok smo druge potpuno izostavili iz dalje analize.

Prvi atribut "*RefID*" predstavlja jedinstveni identifikacioni broj svakog vozila. Ova kolona postavljena kao indeks, jer je jedinstvena za svaki automobil i neće nam služiti u predikciji.

Atribut "*PurchDate*" sadrži datum kupovine vozila. Pošto su sve kupovine obavljene u januaru, iz ove kolone ćemo izostaviti dan, mesec i vreme i ostaviti samo godinu kupovine.

Naziv aukcije na kojoj je kupljen auto je dat pomoću atributa "*Auction*", ali postoji izuzetno veliki broj nedostajućih vrednosti (više od 50%), pa je ovaj atribut isključen iz dalje analize.

Atribut "*VehAge*", koji ukazuje na starost vozila, ćemo isključiti iz predikcije, jer postoji atribut "*VehYear*" koji nam daje iste informacije, odnosno starost automobila. Vidimo da je "*VehYear*" dosta korelisan sa izlaznim atributom.

IsBadBuy	0	1	% kick
VehYear			
2001	104	34	0.246377
2002	247	79	0.242331
2003	455	120	0.208696
2004	817	159	0.162910
2005	1247	182	0.127362
2006	1412	165	0.104629
2007	983	82	0.076995
2008	597	42	0.065728
2009	70	3	0.041096

Atribut "*Make*" predstavlja naziv proizvođača automobila.

Od kolone "*Model*", koja sadrži veliki broj informacija, ćemo izvesti nekoliko atributa. Prvo, izvodimo kolonu "*ModelNew*" koja će sadržati samo model automobila, bez oznake za zapreminu motora, pogon i broj cilindara. Nazivi modela se većinski sastoje od jedne reči, dok i one koji se sastoje od 2 možemo identifikovati pomoću prve, pa iz tog razloga uzimamo samo prvu reč iz naziva. Time smo smanjili broj jedinstvenih vrednosti ove kolone sa 632 na 187. Zatim, iz kolone "*Model*" izdvajamo atribut koji označava zapreminu motora, ali kako on ima previše nedostajućih vrednosti (oko 60%) njega zanemarujemo. Narednu kolonu koju smo izdvojili sadrži tip pogona automobila, i primećuje se da je najmanji broj neispravnih

vozila bio na prednju vuču, ali i ovu kolonu zanemarujemo zbog nedostatka podataka. Atribut "*Model*" smo napunili vrednostima iz "*ModelNew*", a "*ModelNew*" smo obrisali.

Kolona "*Trim*" ima dosta nedostajućih vrednosti, popunjavamo ih na osnovu najučestalije verzije modela za svaku od marki automobila.

"*SubModel*" je atribut koji u sebi sadrži više različitih informacija, pa ćemo od njega pokušati da napravimo nove atribute. Podataka o pogonu i o zapremini motora ima malo, pa ćemo ih zanemariti. Izvodimo kolonu "*CarBodyStyle*" koja predstavlja oblik karoserije automobila, nedostajuće vrednosti, kojih je samo nekoliko, popunjavamo uz pomoć pretrage modela na internetu. Još jedan novi atribut je "*Doors*", odnosno broj vrata na automobilu. Kako smo izvukli sve moguće podatke, atribut "*SubModel*" brišemo.

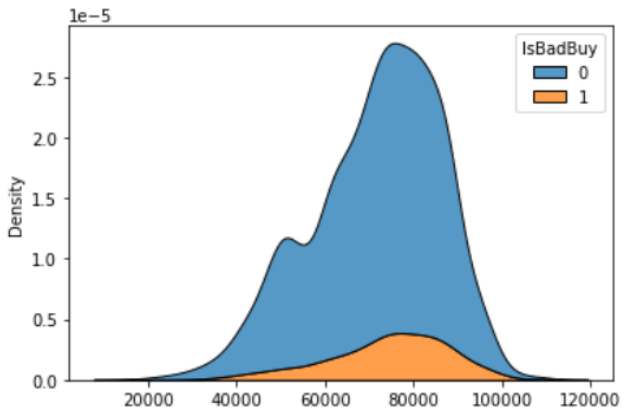
"*Color*" predstavlja boju automobila, dok "*Transmission*" određuje da li je menjač automatski ili manuelni.

"*WheelType*" i "*WheelTypeID*" su u visokoj korelaciji, pa "*WheelType*" izbacujemo iz analize, a nedostajuće vrednosti u koloni "*WheelTypeID*" popunjavamo najvećom vrednošću.

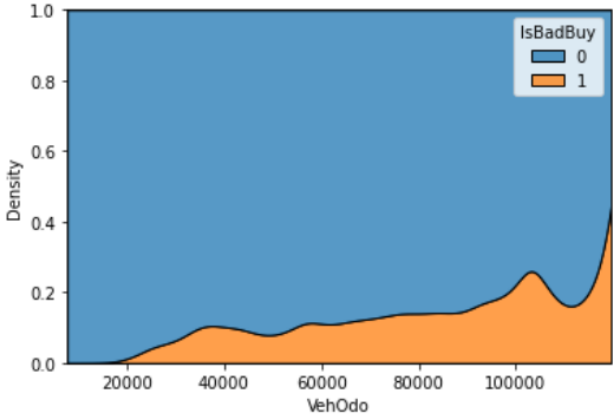
"*Nationality*" predstavlja zemlju porekla vozila, a najčešći su američki proizvođači.

"*TopThreeAmericanName*" sadrži slične informacije kao kolona "*Make*", pa ćemo je obrisati.

Atribut "*VehOdo*" sadrži numeričke podatke o pređenoj kilometraži vozila.



Graf iznad nam govori relativan odnos između zastupljenosti *kick*-ova i onih koji to nisu. Vidimo da oko vrednosti od 50000 graf gubi svoju konzistensnost. Naša pretpostavka je da je to "magična brojka" na koju prevaranti vraćaju kilometražu.



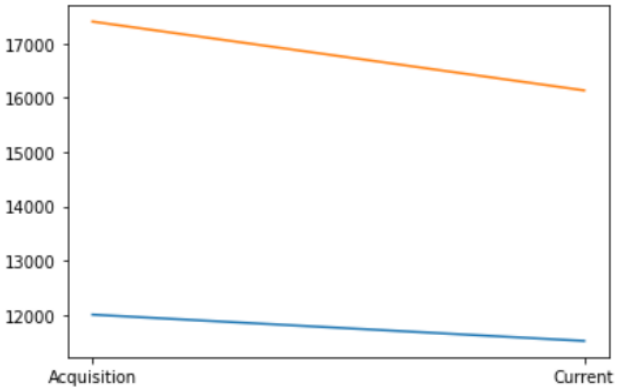
Atributi:
"*MMRAcquisitionAuctionAveragePrice*",
"*MMRAcquisitionAuctionCleanPrice*",
"*MMRAcquisitionRetailAveragePrice*",
"*MMRAcquisitonRetailCleanPrice*",
"*MMRCurrentAuctionAveragePrice*",
"*MMRCurrentAuctionCleanPrice*",
"*MMRCurrentRetailAveragePrice*",
"*MMRCurrentRetailCleanPrice*"
označavaju cene vozila u zavisnosti od stanja vozila i trenutka kupovine, kao i naznaku da li je cena tržišna ili aukcijska.

Atribute kojima se opisuju cene smo iskoristili da bi napravili nove atribute preko kojih bi objasnili kretanje relativne vrednosti automobila u trenutku kupovine do poslednjeg trenutka u skupu podataka ("*Perc_Diff_RetAVG*",
"*Perc_Diff_AuctAVG*",
"*Perc_Diff_RetAuctAVG*",
"*Perc_Diff_RetCLEAN*",
"*Perc_Diff_AuctCLEAN*",
"*Perc_Diff_RetAuctCLEAN*").

Jos jedan set atributa smo definisali na osnovu cena, čime smo želeli da vidimo da li je relativna razlika između prodajne i aukcijske cene na dan kupovine, kao i u poslednjem datom trenutku u skupu podataka, je povezana sa tim da

li je auto *kick* ("*Perc_Diff_PAST_RetAuctAVG*", "*Perc_Diff_PRESENT_RetAuctAVG*", "*Perc_Diff_PAST_RetAuct_CLEAN*", "*Perc_Diff_PRESENT_RetAuct_CLEAN*").

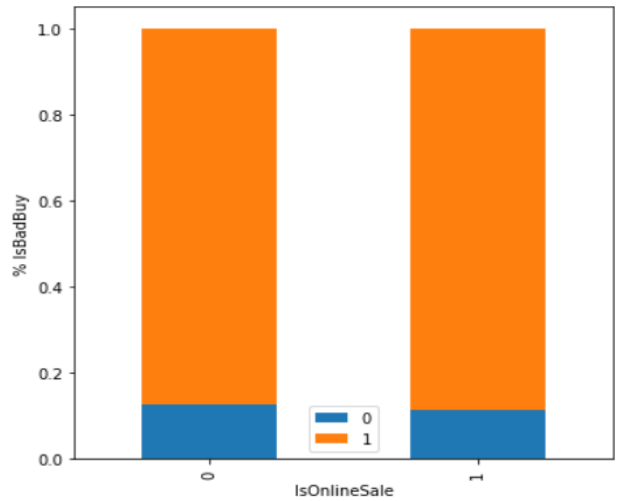
Sledeći graf nam daje prikaz kretanja cena za prosečno stanje automobila kroz vreme automobila rednog broja 8. Crveno je tržišna cena, dok je plava aukcijska.



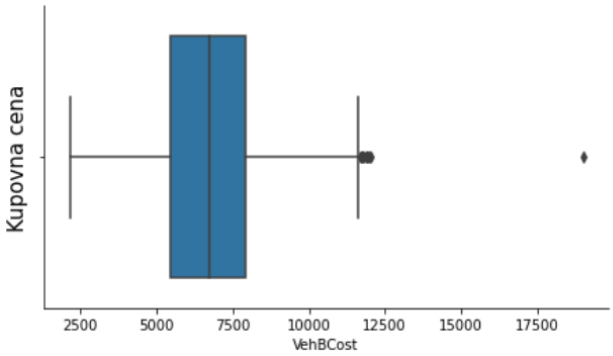
"*BYRNO*" i "*VNZIP1*" predstavljaju šifru kupca automobile, šifru gde je automobil kupljen respektivno, pa ih nećemo uključiti u predikciju. Kolona "*VNST*" predstavlja naziv države kupovine auta, ali sadrži 80% nedostajućih vrednosti, pa je isključujemo iz dalje analize.

"*PRIMEUNIT*" i "*AUCGUART*" imaju previše nedostajućih vrednosti pa ih izbacujemo.

Atribut "*IsOnlineSale*" nam govori da li je auto kupljen na *online* aukciji. Većina automobila je kupljena uživo, ali smo primetili da to svakako ne utiče na to da li je *kick*. Na narednom grafu vidimo da *"IsOnlineSale"* ne utiče na vrednost izlazne kolone.



Kolona "*VehBCost*" govori o troškovima nabavke vozila pri kupovini, dok "*WarranyCost*" nam daje informacije o cenama garancije.



Zanimljivo je da je jedan automobil kupljen za čak 19000 i da se ispostavio kao *kick.*

Nakon popunjavanja nedostajućih vrednosti i odabira kolona za predikciju, kategoričke podatke kvantifikujemo pomoću funkcija *replace* i *dummies.*

3. Treniranje i ocena modela

Dati problem se svodi na problem algoritama nadgledanog učenja, preciznije, algoritama klasifikacije. Cilj je napraviti model koji će uspešno predviđati koji automobili su potencijalno neispravni ili im je vraćena kilometraža (štelovan brojač).

Kako svi atributi ne doprinose podjednako pri predikciji, koristili smo određene metode selekcije atributa kako bismo smanjili broj atributa sa kojima se radi. Nakon toga smo upoređivali ocene modela na novim skupovima podataka kako bi pronašli najbolji algoritam za naš problem. Metode koje smo koristili su filter metode i jedna metoda obavijanja. Filter metoda koju smo prvu koristili računa varijansu na osnovu zadatog praga (0.05 u našem modelu) i odbacuje attribute sa varijansom manjom od praga. Druga filter metoda koristi računanje entropije između izlaznog i ostalih atributa i zadržava one sa najvećom vrednošću. Kao treću, koristili smo metodu obavijanja koja prolazi kroz skup podataka iterativno i u svakoj iteraciji izostavlja određeni broj atributa koje smo zadali (10 u našem modelu) i vrednuje modele na osnovu zadate roc_auc metrike. Na kraju je inicijalni skup atributa smanjen sa 388 na 42, 40 i 40 atributa respektivno.

Pre same klasifikacije kao metod evaluacije odabrali smo *unakrsnu validaciju* koja je kompleksniji ali i verodostojniji vid ocenjivanja kvaliteta modela od trening-test podele. *Unakrsna validacija* prolazi kroz delove skupa podataka i svaki od njih jedanput uzima za test dok ostale koristi za trening. Za finalnu ocenu se uzima prosek ocena svih delova. Kod trening-test podele zbog nasumičnosti podele podataka možemo dobiti dobre ocene modela iako to možda nije slučaj.

Za evaluacionu metriku ćemo uzeti *ROC* krivu, tj površinom ispod nje (AUC). *AUC* je korisna evaluciona metrika jer uzima u obzir sve granice odlučivanja i daje nam najbolju. Od dve vrste greške skuplje će nas koštati onda kada auto koji jeste *kick* klasifikujemo kao da nije. Ali pošto ne znamo cenu grešaka odnosno dobit pogodaka, ne možemo da koristimo *total_cost* meru. U slučaju da smo imali informacije iz prethodne rečenice, i sa pretpostavkom da je lažno negativna opservacija najskuplja, onda bi nam odziv bio izuzetno bitan tj. nastojali bi da ga povećamo korišćenjem modela koji daju verovatnoću (logistička regresija ili naivni bajes) tako što bi manipulisali pragom tolerancije.

Algoritmi klasifikacije koje ćemo koristiti:

- Stablo odlučivanja
- Naivni Bajes
- K najbližih suseda
- Logistička regresija
- Ansambl algoritmi (Glasanje, Stacking i Slučajne šume)

Nakon klasifikacije sa predefinisanim hiperparametrima, dobili smo sledeće rezultate:

Algoritam	X_filter_variance_thresh	X_filter_mutual_info	X_rfecv
Stablo odlučivanja	53.72	54.31	54.02
Naivni Bajes	62.06	62.98	63.15
KNN	52.17	55.39	54.79
Log regresija	61.39	60.81	62.20
Glasanje meko	60.86	60.78	62.86
Stacking lr	57.02	55.06	57.58
Random forest	64.62	62.83	65.11

Može se videti da najbolje rezultate daje skup atributa dobijen na osnovu obavijajuće metode i iz toga razloga u daljem radu ćemo se fokusirati samo na njega. Dobijeni rezultati za *obavijajuću metodu* nam govore da *model slučajne šume* daje vrednost od 65,11 *AUC*. Možemo zaključiti da je ovaj rezultat sigurno bolji od nasumičnog pogađanja, ali da definitivno ima mesta za poboljšanje modela.

4. Optimizacija parametara

Nakon optimizacije hiperparametara modela stabla odlučivanja, KNN, logističke regresije i slučajnih šuma dobili smo bolje rezultate od klasifikacije sa predefinisanim parametrima. Kod modela najbližih suseda testirali smo različite vrednosti broja suseda i dobili smo da je 475 vrednost blizu optimalne. U logističkoj regresiji parametar "C" bi trebalo da nosi vrednost 1,56. Za stablo odlučivanja idealna maksimalna dubina bi bila 6 listova. Kod modela slučajnih šuma gde je minimalan broj uzoraka potrebnih za razdvajanje čvora postavljen na 1, maksimalna dubina na 25 i maksimalan broj atributa na 10, dobijaju se bolji rezultati AUC metrike.

Algoritam	pre optimizacije	posle optimizacije
Stablo odlučivanja	54.02	62.89
KNN	54.79	60.96
Log regresija	62.20	62.28
Random forest	65.11	66.26

Vidimo da smo unapredili sve modele, neke više od drugih. Najbolji rezultat daje model nasumičnih šuma sa vrednošću 66,26 AUC metrike.

5. Zaključak

Najveći deo procesa kreiranja rada je bio vezan za samu pripremu podataka pre modelovanja. Shvatanje, opisivanje, dodatno uređivanje (problemi nedostajućih vrednosti, kvantifikacija kategoričkih atributa,...) i

konstrukcija novih atributa su zahtevali najviše pažnje.

Kao najbolja metoda za selekciju se pokazala metoda obavijanja, te smo sa tim novoformiranim skupom podataka primenili 7 različitih algoritama gde se kao najbolji, nakon optimizacije hiperparametara na 4 modela, pokazao algoritam Random Forest.

Pošto nemamo domensko znanje, smatramo da bi nam značili kompanijski podaci vezani za cenu grešaka i način na koji su svaku od njih kvantifikovali, dodatno pojašnjenje i opis termina “kick”. Određeni atributi su pokazali zanimljive raspodele i kretanja pa bi bilo korisno ispitati i njihovu dalju korelaciju sa modelom ali bi za to bila potrebna saradnja sa Carvanom.