Predviđanje isplativosti kupovine polovnih automobila

Danica Trajić 1043/19, Nađa Bulajić 372/17, Elena Dubić 186/17

1. Opis i razumevanje problema

Cilj analize je da rešimo probleme koji nastaju prilikom kupovine polovnih automobila. Kupcima je teško da procene na osnovu dostupnih informacija da li je automobil koji kupuju vredan kupovine ili ne, odnosno, da li je cena koja je ponuđena realna. Potrebno je napraviti model koji će moći na osnovu svih dostupnih atributa vozila da predvidi da li je automobil vredan kupovine. Ovo predviđanje je bitno jer bi moglo kupcima da olakša donošenje odluke o kupovini polovnog automobila, a dalo mogućnost samostalne kupovine onima koji nisu dovoljno upućeni u automobilsku industriju. Takođe, prodavci bi smanjili svoje troškove koji bi nastali usled žalbi nezadovoljnih korisnika koji su kupili defektan auto.

2. Opis, razumevanje i priprema podataka

Skup podataka sadrži 6798 opservacija i 33 atributa od kojih je jedan izlazni atribut i to je IsBadBuy. Proverom za nedostajuće vrednosti smo uvidele da varijable Auction, PRIMEUNIT, AUCGUART i VNST imaju previše nedostajućih vrednosti, tj. dostupan broj podataka je 2983, 339, 339 i 1131 respektivno. Zbog prevelikog broja nedostajućih vrednosti, ovi atributi su isključeni iz dalje analize.

Što se tiče ostalih atributa, oni imaju manji broj nedostajućih vrednost i pored ovih 3 ima ih još 14. Za svaki od njih smo posebnom analizom određivali da li ćemo ih izbaciti iz analize ili popuniti vrednosti.

Varijable *WheelTypeID* i *BYRNO* – ove varijable predstavljaju jedinstveni broj, pa i njih odmah izbacujemo iz dalje analize jer nam nisu značajni.

Varijabla *VNZIP1* – predstavlja poštanski broj mesta gde je automobil kupljen, nema nikakvog značaja, pa i nju izbacujemo.

Varijabla *VehYear* – predstavlja godinu proizvodnje automobila, s obzirom da pored nje imamo i varijablu *VehicleAge* koja nam govori o starosti vozila, one su u visokoj korelaciji i iz tog razloga smo je izbacile.

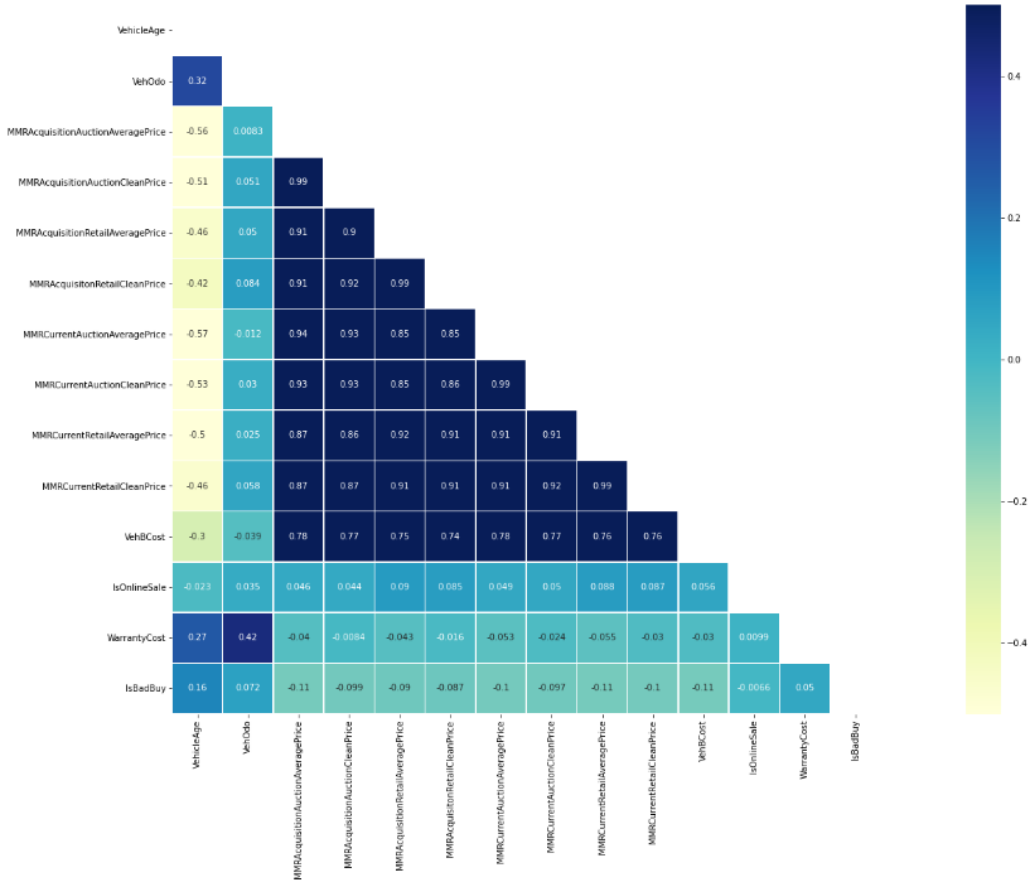
Varijabla *Trim* – uvidele smo da ova varijabla ima 104 jedinstvenih vrednosti, a veliki broj njih ima frekvenciju 1. Kako se radi o verziji modela, nije moguće spojiti određene grupe, a moguće je da bi ovakva dovela do pretreniranosti modela, odlučile smo da je izbacimo.

Varijabla *WheelType* – odnosi se na tip felni, imamo dve kategorije u kojima su opservacije približno raspoređene. Dominantna kategorija je Alloy, pa 339 nedostajućih vrednosti dopunjujemo njom.

Varijabla *Nationality* – odnosi se na zemlju proizvodnje vozila i visoko je korelisana sa varijablom *Make* koja se odnosi na proizvođača vozila, a i najveći broj opservacija pripada samo jednoj kategoriji, odlučile smo da je izbacimo.

Varijabla *Size* – odnosi se na pripadnost kategoriji na osnovu veličine vozila. Ima samo jednu nedostajuću vrednost i popunile smo je dominantnom kategorijom MEDIUM.

Varijabla *TopThreeAmericanName* – identifikuje da li određeni proizvođač pripada grupi top tri američka proizvođača i upravo zbog toga je u visokoj korelaciji sa gorepomenutom varijablom *Make*, pa smo je izbacile iz analize.



Možemo primetiti da su varijable:

- *MMRAcquisitionAuctionAveragePrice,*
- *MMRAcquisitionAuctionCleanPrice,*
- *MMRAcquisitionRetailAveragePrice,*
- *MMRAcquisitonRetailCleanPrice ,*
- *MMRCurrentAuctionAveragePrice ,*
- *MMRCurrentAuctionCleanPrice,*
- *MMRCurrentRetailAveragePrice ,*
- *MMRCurrentRetailCleanPrice*

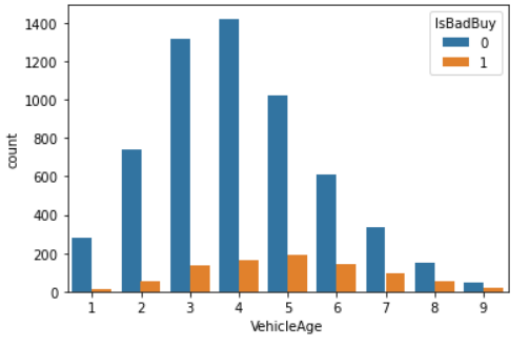
međusobno visoko korelisane, ali pored toga i veoma značajne za našu analizu.

Ovi atributi se odnose na cenu automobila uzimajući u obzir stanje vozila, veličinu tržišta, kao i trenutak procene njegove vrednosti. Kako bismo utvrdile pojedinačan doprinos, napravile smo odvojene dataset-ove, jedan koji sadrži sve ove atribute, a drugi koji sadrži samo one koji se odnose na automobile koji su u prosečnom stanju.

Opservacija sa indeksom 530 je imala nedostajuću vrednost za sve ove atribute, iz tog razloga smo je izbacile. Pored toga, 4 atributa koja se odnose na sadašnjost su imali jos 20 nedostajućih vrednosti i utvrđivanjem da oni ne podležu normalnoj raspodeli, vrednosti koje nedostaju smo zamenile medijanom.

Varijabla *PurchDate* – predstavlja datum kupovine vozila, a iz datuma nam je značajna samo godina, tako da smo odlučile da napravimo novu varijablu *PurchYear*, a zatim izbacile ovaj atribut.

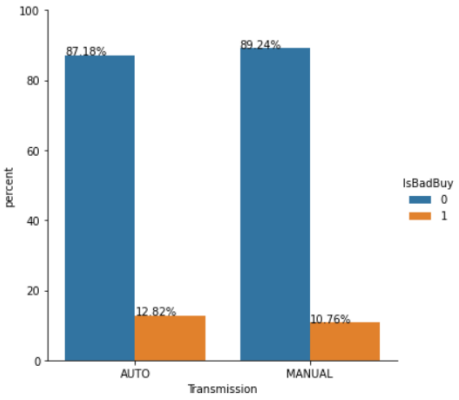
Varijabla *VehicleAge* – vidimo da različita starost vozila ima drugačiji uticaj na kupovinu, pa možemo da zaključimo da nam je ovaj atribut značajan za dalju analizu



Varijable *Model* i *Submodel* – odnose se na model i podmodel automobila. Uočile smo da imaju previše različitih, ali i nesrazmerno raspodeljenih kategorija, pa smo obe isključile iz dalje analize.

Varijabla *Color* – smatramo da je značajna za dalju analizu jer boja može znatno uticati na cenu istog modela automobila. Ima 5 nedostajućih vrednosti označenih kao ,,NOT AVAIL'' koje smo popunile kategorijom ,,OTHER'' iako nije dominantna, jer se u njoj već nalaze opservacije koje ne pripadaju nijednoj određenoj kategoriji boja.

Varijabla *Transmission* – kategorička varijabla koja razdvaja automatik i manuel vozila. Dominantna kategorija je automatik i iz toga razloga smo posmatrale proporcionalnu razliku između dve kategorije u odnosu na izlaznu promenljivu. Vidimo da ne postoji značajna razlika kupovine u zavisnosti od kategorije.



Varijabla *IsOnlineSale* – pokazuje nam da li je vozilo kupljeno online ili ne, pošto većina opservacija pripada klasi koja sadrži informacije o automobilima koji nisu kupljeni online, izbacile smo je.

Varijabla *IsBadBuy* – predstavlja izlaznu varijablu i govori nam o tome da li je kupovina koja je već izvršena bila loša ili dobra. Uočavamo da je nebalansirana klasa u kojoj je dominantnija kategorija ,,False''.

Analizom mogućih scenarija dolazimo do zaključka da nam je najgora greška koju možemo da napravimo da predvidimo da neka kupovina treba da se obavi, a zapravo nije trebalo, pa kupac nakon kratkog perioda otkrije da je vozilo defektno i da nije bilo vredno tog ulaganja.

3. Selekcija atributa

Prvi korak nam je bio da sagledamo novi dataset i varijable koje smo ostavile i da zdravorazumski zaključimo koje su značajne za našu analizu. Preko različitih grafikona smo donosile zaključke koje varijable i u kojoj meri utiču na izlaznu promenljivu, a nakon toga smo koristile filter metodu, preciznije VarianceThreshold funkciju, kako bismo izostavile varijable koje nam nisu dovoljno informativne (u odnosu na ciljanu varijablu ili samu varijablu koju posmatramo). Ovom metodom smo uspele da smanjimo broj varijabli sa 70 na 20.

4. Klasifikacija

Koristili smo sledeće algoritme klasifikacije:

- Naivni Bajes
- Stablo odlucivanja
- K najblizih suseda (KNN)
- Logističku regresiju
- Ansambl algoritme (Bagging, Voting, Random Forest)

Ovim algoritmima analiziramo podatke nad originalnim skupom podataka i nad podacima dobijenim korišćenjem različitih metoda za selekciju atributa, a sve u svrhu odabira najboljeg modela za naš konkretan problem. Pored njih smo koristile i Stacking kako bismo našle najbolju kombinaciju modela.

Algoritam	Inicijalni skup podataka	Potpuni skup podataka	Potpuni skup podataka sa ACC	Potpuni skup podataka filter	Potpuni skup podataka filter ACC
Naivni Bajes	50.34	59.91	14.36	61.44	81.46
KNN	55.23	55.11	85.68	56.20	85.31
Stablo odlucivanja	51.51	52.55	78.05	53.12	77.63
Logisticka regresija	67.14	67.20	87.16	65.53	87.18
Voting	61.01	61.73	87.18	63.25	87.18
Bagging	66.96	67.03	87.16	65.45	87.18
Random forest	62.77	63.36	87.11	59.78	83.18

Prvo smo primenili sve algoritme na inicijalni skup podataka – bez MMR atributa koji odnose na nadprosečna vozila. Najbolji rezultat smo dobile za logističku regresiju i Bagging koji se zasniva na logističkoj regresiji.

Kako bismo proverile da li je pravljenje dataseta bez navedenih atributa dalo poboljšanje primenile smo sve algoritme i na potpuni skup podataka. S obzirom da je došlo do malog poboljšanja odlučile smo da nastavimo rad nad potpunim skupom podataka.
Nakon izvršenog filtriranja podataka i ponovne primene algoritama nad tako dobijenim datasetom vidimo da je došlo do minimalnog poboljšanja svega osim Logističke regresije, Bagging-a i Random Forest-a.

Do sada smo kao meru evaluacije koristile ROC krivu ali smo došle na ideju da proverimo kakvi se rezultati dobijaju korišćenjem mere accuracy. Za ovu meru ćemo takođe analizirati potpuni i filtrirani skup podataka. Kod potpunog skupa uočavamo da svi modeli osim Naivnog Bajesa i Stabla odlučivanja imaju tačnost preko 85%.

S obzirom da nam Stablo odlučivanja i KNN daju najlošije rezultate pokušale smo da ih poboljšamo optimizacijom parametara. Dobile smo da je optimalna dubina stabla 2, a broj suseda 100. Kada smo to ubacile u model videle smo da je došlo do poboljšanja za približno 10% kod stabla i 5% kod KNN-a.

Algoritam	Pre optimizacije	Nakon optimizacije
Naivni Bajes	61.44	61.44
KNN	56.20	61.19
Stablo odlucivanja	53.12	63.08
Logisticka regresija	65.53	65.53
Voting	63.25	65.12
Bagging	65.45	65.38
Random forest	59.78	59.01

5. Zaključak

Na osnovu dobijenih rezultata zaključile smo da nam Logistička regresija daje najbolje rezultate i zbog toga ćemo taj model koristiti. Tačnost predviđanja ovim modelom iznosi 87.18%, preciznost iznosi 92%, a AUC 65.53%.
S obzirom da vidimo da i dalje ima prostora za poboljšanje predlog je da se izvrši regularizacija i na osnovu ridge odredi optimalna vrednost koeficijenata ili primenom lasso metode izvrši selekcija atributa.

Još jedan predlog za poboljšanje je da se obrati pažnja na varijablu *Model* jer sadrži dosta skrivenih informacija koje znatno mogu uticati na konačnu vrednost automobila (npr. jačina motora, prednja/zadnja vuča...).

Model koji smo napravile je tek polazna osnova razvoja cele ideje i kao ovakav može da nastavi da se koristi u analizama i za dalji rad na njegovom poboljšanju ali nije spreman da bude pušten u upotrebu kao gotov proizvod.