Kreiranje klasifikatora za predviđanje kupovine lošeg automobila na aukciji

Sanja Rančić 0133/2017 Milica Perišić 0052/2016

Fakultet Organizacionih nauka
Beograd, 2020

Rezime: *Prilikom kreiranja modela mašinskog učenja postoji više segmenata na koje treba obratiti pažnju. U radu je prikazano na koji način je izvršena i priprema podataka, kao i koji modeli su kreirani, na koji način su isti optimizovani i kako su evaluirani.*
Ključne reči: Mašinsko Učenje, Tačnost Algoritma, Balansiranje Klasa, Optimizacija Parametara, Analiza podataka

1. Uvod

Modeli mašinskog učenja se koriste svakodnevno za različite komercijalne potrebe. Jedna od potreba koja se javlja jeste upravo i predviđanje ispravnosti automobila. Kupovina polovnih automobila sa sobom nosi određene rizike. Bilo da je kupac fizičko ili pravno lice, bilo da ima iskustva u proceni automobila ili ne, rizik da će automobil u kratkom vremenskom periodu postati neispravan postoji. Do toga često dolazi usled prikrivanja mehaničkih kvarova koje nije lako uočiti, sređivanjem automobila samo za potrebe prodaje ili vraćanjem kilometraže. Ovo predstavlja veliki problem i za fizička lica, ali još veći problem za dilere automobila. Cilj je kupiti automobil na aukciji po jednoj ceni i prodati ga po višoj nakon toga. Kada je automobil koji je kupljen neispravan to postaje nemoguće. Često cela investicija propada i diler automobila trpi trošak.

Fokus ovog rada je na kreiranju modela koji bi rešio ovaj problem, tj na kreiranju klasifikatora za predviđanje kupovine lošeg automobila na aukciji. Postojanje ovakvog modela bi znatno smanjilo troškove tih preduzeća i unapredilo poslovanje.

Rad je struktuiran na sledeći način. U odeljku 2 definisana je metodologija pristupa kao i rezultati, dok je u odeljku 3 zaključen rad.

2. Metodologija

Glavni fokus ovog rada jeste kreiranje modela mašinskog učenja koji će na adekvatan način klasifikovati automobile u 2 grupe, ispravne i neispravne. Prilikom kreiranja ovakvog modela postoji više segmenata na koje treba obratiti pažnju.

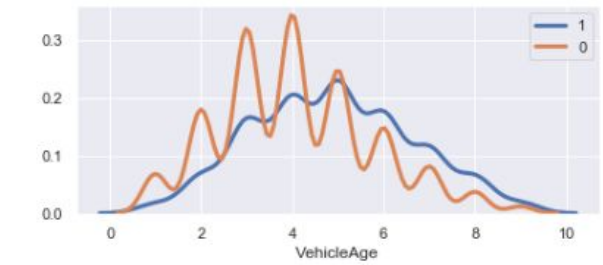
Metodologija se sastoji iz 5 delova koji predstavljaju 5 koraka u kreiranju modela mašinskog učenja.

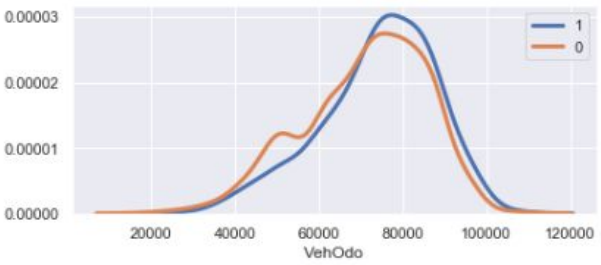
2.1. Analiza podataka

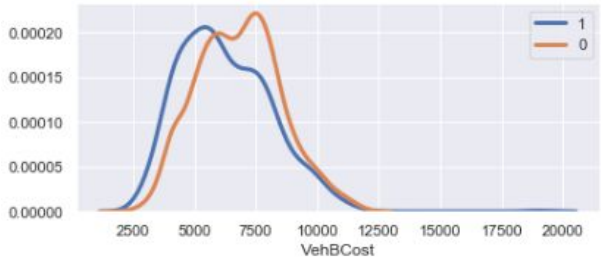
Set podataka nad kojim je model kreiran, carvana.csv, sadži 34 atributa i 6798 opservacija. Među atributima razlikujemo 19 numeričkih i 15 kategoričkih. Među kategoričkim atributima postoji i nekoliko binarnih varijabli. Varijabla koju je potrebno predvideti jeste “IsBadBuy”, binarna varijabla kod koje nula predstavlja automobile koji su bili dobrog kvaliteta, dok je jedan oznaka za lošu kupovinu. Ova varijabla je nebalansirana. Znatno je više onih kupovina koje su se pokazale kao dobre. Procentulano, samo nešto više od 12% kupovina su označene kao loše, odnosno 87% seta predstavljaju dobre kupovine. Ovo značajno utiče na sposobnost modela da nauči više o odlikama kupovina koje su klasifikovane kao loše.

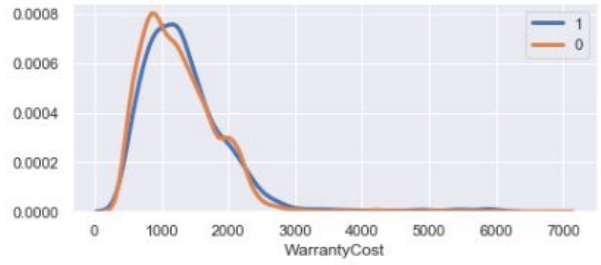
Dileri automobila bi bili na znatno većem gubitku ukoliko bi model predvideo da neki automobil neće biti loša kupovina, a on zapravo bude, što znači da je u ovom slučaju važnija greška prvog tipa (fn), te se kao cilj postavlja optimizacija odziva.

Raspon godina automobila se kreće između 1 i 9, i može se reći da godine podležu normalnoj raspodeli, kao i većina numeričkih atributa. Postoje neki automobili koji su prodati i za 0 dolara, što je upečatljivo, dok se najviše cene kreću između 20 i 30 hiljada. Trošak akvizicije je značajno veći od troška garancije. Najćešći brend je Chevrolet, boja srebrna, dok su skoro svi automobili automatik. Uglavnom se radi o automobilima američkog porekla, srednje veličine.

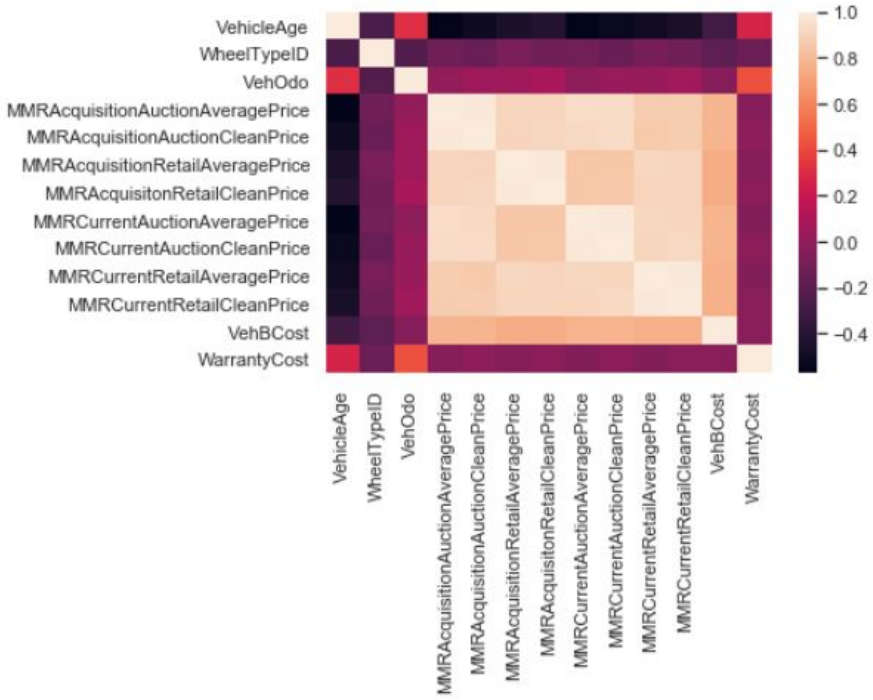








Iz grafika se uočava da postoji određena razlika između automobila koji su bili loša kupovina i onih koji nisu u cenama i trošku. Prosek cena je niži kod automobila koji su se kasnije pokazali kao kick, dok je trošak veći. Takođe, automobili koji su se pokazali kao kick su nešto stariji od onih koji nisu.



Na grafiku iznad je prikazana korelacija atributa. Korelacija između cena je očekivano visoka. Takođe, korelacija između cene i godišta automobila je negativna, dok je pozitivna između godišta i pređene kilometraže. Za ostale varijable uglavnom nema značajne korelacije, osim povezanosti troška sa kilometražom i godištem, što je očekivano.

Daljom analizom utvrđeno je da set podataka ima ukupno 23345 nedostajućih vrednosti. Nekoliko atributa - Auction, PRIMEUNIT, AUCGUART i VNST, ima veliki procenat nedostajućih vrednosti, dok je kod ostalih atributa taj procenat relativno zanemarljiv.

2.2. Priprema podataka

Priprema podataka je započeta neutralisanjem nedostajućih vrednosti. Primenjivanje neke od tehnika popunjavanja nedostajućih vrednosti nad gore pomenutim bi se loše odrazilo na model, te je bolji pristup neuključivanje ovih atributa u dalju analizu.
Nad ostalim atributima je korektno izvršiti popunjavanje. Kategorički atributi će biti popunjeni modusom. Numerički u zavisnosti od toga da li imaju ili nemaju normalnu raspodelu popunjeni su srednjom vrednošću ili medijanom.

Nakon sređivanja kolona sa nedostajućim vrednostima, atributi su analizirani i iz postojećih su izvedeni novi atributi. Atributi Model i SubModel imaju izuzetno veliki broj jedinstvenih vrednosti i kao takvi ne bi bili upotrebljivi u modelima, tj. ne bi donosili vrednost modelu. Međutim,

primećeno je da ovi atributi imaju u sebi podatke koji mogu biti značajni. Tako je iz Modela izveden atribut vrsta motora, dok je iz SubModel-a izveden atribut broj vrata. Ono što je još pokušano jeste dobijanje informacije o gradu, kao i o zapremini motora, međutim zbog velikog broja jedinstvenih vrednosti i nedostajućih vrednosti kad je u pitanju drugi atribut, oni nisu uključeni u dalju analizu.

Sledeći korak u pripremi podataka jeste transformacija kategoričkih atributa u numeričke. Pristup je bio takav da su one varijable kojima je moguće utvrditi poredak (ordinalne) mapirane putem rečnika, dok je su nominalne varijable postale dummy varijable.

Tokom analize je utvrđeno da neke kolone nemaju uticaj na odlučivanje modela, te su izbačene. Pre svega se to odnosi na id kupca za kojeg nisu poznate dodatne informacija, kao i na zip kod iz gore pomenutih razloga. Takođe, tokom eksploratorne analize je uočeno da boja nema velikog značaja na krajnji ishod i da procentualno nema velike razlike između loših i dobrih kupovina. Zbog toga je i boja izbačena iz dalje analize. Datum kada je vozilo kupljeno na aukciji takođe nije od značaja za predikciju, a s obzirom da već postoji podatak o starosti vozila uklonjena je i godina proizvodnje. Isključen je i Trim jer se odnosi na dodatne karakteristike koji neki automobil ima, a koje su specifične za svakog proizvođača.

Kako naš model ne bi učio na ekstremnim slučajevima potrebno je ukloniti ekstremne vrednosti. Nakon uklanjanja ekstremnih vrednosti moguće je izvršiti skaliranje podataka kako bi se uspostavio ispravan odnos izmedju promene vrednosti atributa, što je učinjeno MinMax Scalerom koji skalira vrednosti na raspon od 0 do 1.

Poslednji korak u pripremi podataka jeste balansiranje podataka, zbog nebalansirane klase koja se predviđa. Postoje tri pristupa. Prvi jeste da se klasa ostavi takva kakva je, što nije preporučljivo u ovako ekstremnim slučajevima. Drugi, undersample, odnosno smanjivanje broja opservacija. I treći, oversample, povećavanje broja opservacija. I jedan i drugi pristup dovode do izjednačavanja broja opservacija kojima je target klasa 0 i kojima je target klasa 1.

```
Bez balansiranja: (4626, 60)
Undersample: (1186, 60)
Oversample: (8066, 60)
```

Podaci koji su prikazani se odnose na train set podataka.
Eksperimentalnim pristupom, gde su primenjene sve tri opcije i testirane, odlučeno je da se radi undersample podataka.

2.3. Treniranje algoritama i interpretacija rezultata

Problem koji je potrebno rešiti jeste problem klasifikacije, te su primenjeni sledeći algoritmi : KNN, Logisticka regresija, Random forest, i Gradient Boosting. Na ovaj način je pristupljeno problemu sa više različitih grupa algoritama. Pre skaliranja podataka i balansiranja, skup podataka je podeljen na test skup i train skup u odnosu 1:3, kako bi bilo moguće evaluirati modele.

Prvo su modeli učili sa podrazumevanim postavkama parametara. Kao najbolji se pokazala Logistička regresija kada su uzeti u obzir tačnost i odziv. Dobijeni rezultati su sledeći:

```
LogisticRegression
Accuracy:  0.6101865859808371
Precision:  0.18518518518518517
Recall:  0.6300813008130082
Roc-auc:  0.6563999232393015
```

Rezultati svakako nisu idealni. Tačnost modela je 61%, dok je mera koju je potrebno optimizovati 63%. Takav odziv znači da bi ovaj model za 37% opservacija predvideo da će biti dobra kupovina iako to nije slučaj.

2.4. Optimizacija parametara i interpretacija dobijenih rezultata

Svakako modeli se ne treniraju sa podrazumevanim parametrima, nego je iste potrebno optimizovati. To se može postići traženjem optimalnih parametara. Tehnike koje su korišćene u ovom projektu jesu GridSearch i RandomizedSearch.
Svi modeli su sa optimalnim parametrima popravili rezultate u određenoj meri. Sledi prikaz rezultata optimizacije:

Logistička regresija
Optimizacijom parametara za logističku regresiju nisu mnogo popravljeni rezultati modela. Slični su gore naznačenim rezultatima. Optimalni parametri su sledeći:

```
{'C': 1.3403641664256707, 'penalty': 'l2'}
```

KNN
KNN je poboljšao tačnost u odnosu na prethodnu iteraciju, dok je zbog toga žrtvovan odziv. Optimalni parametri su sledeći:

```
{'metric': 'euclidean', 'n_neighbors': 9}
```

Random Forest
Radnom Forest je optimizacijom parametara poboljšao svoje rezultate za do 2%. Optimalni parametri su sledeći:

```
{'max_depth': 8, 'min_samples_split': 32, 'n_estimators': 80}
```

Gradient Boosting
Gradient Boosting je slično kao KNN poboljšao tačnost da 1,5% međutim to je značilo smanjivanje vrednosti odziva. Optimalni parametri su sledeći:

```
{'max_depth': 6, 'min_samples_split': 8, 'n_estimators': 40}
```

Zaključak nakon optimizacije jeste da je i dalje najbolja Logistička regresija.

2.5. Selekcija atributa i iterpretacija dobijenih rezultata

Trenutno modeli uče na osnovu informacija iz 60 atributa. Mnogi od tih atributa su neinformativni i dovode do šuma u modelu, zbog čega je potrebno izvršiti njihovu selekciju i ostaviti one koji su značajni za model. Od pristupa biće predstavljeni sledeći: Selekcija iz modela, Variance Threshold, Select K Best i Metoda obavijanja.

Selekcija iz modela
Selekcija iz modela je očekivano najviše uticala na Random Forest i popravila njegove performanse, dok kod ostalih nije napravila značajnu razliku. Neki su dali i lošije rezultate u odnosu na slučaj kada su sve varijable uključene. Broj atributa je sveden na 27.

Variance Threshold
Variance Threshold je sa pragom 0.02 izbacio nešto manje atributa, te je ostalo njih 39, i ponovo je to najviše uticalo na Random Forest koji do sada ima najbolji rezultat - popravljen za 2,5%.

SelectKBest
SelectKBest je primenjen nad već smanjenim brojem atributa iz prethodnog koraka i za K je prosleđen 20. Ovih 20 atributa daju do sada najbolje rezultate za sve modele, dok je trenutno najbolja Logistička regresija - poboljšana za više od 4% u odnosu na korišćenje svih atributa.

Metoda obavijanja
Metodu obavijanja je primenjena na svakom algoritmu posebno i ova metoda je najviše uticala na Gradient Boosting - poboljšan za 4,5%.
Ovim bi zaključili da u zavisnosti od modela, zavisi i koji bismo način selekcije primenili.

3. Zaključak
Pošto je cilj optimizovati odziv, odlučeno je koristiti Logističku regresiju nakon izbora optimalnih parametara i Select K Best selekcije atributa. Trenutni rezultati su sledeći:

```
LogisticRegression
Accuracy:  0.6167423096318709
Precision:  0.19693396226415094
Recall:  0.6788617886178862
Roc-auc:  0.6697136919555724
```

Postoji prostor za značajno poboljšanje ovog modela. Postoji mogućnost da su neke od kolona koje su izbačene informativnije u nekom drugom obliku koji nije uočen tokom rada na projektu. Obogaćivanje modela novim atributima i dublja analiza kada je u pitanju izbacivanje ekstremnih vrednosti ili balansiranje podataka bi možda uticali na poboljšanje modela.