|  |  |  |
| --- | --- | --- |
| Carvana | July 5  2022 | |
| Studenti:  Danica Zdravković 0076/2018  Željana Grujić 0045/2018 | | Predviđanje da li će kupovina automobile biti loša ili ne nad skupom podataka Carvana. |

Sadržaj

[1. Uvod 3](#_Toc107924412)

[2. Opis I razumevanje podatak 3](#_Toc107924413)

[3. Eksperiment 5](#_Toc107924414)

[4. Rezultati 6](#_Toc107924415)

[5. Zaključak 6](#_Toc107924416)

# Uvod

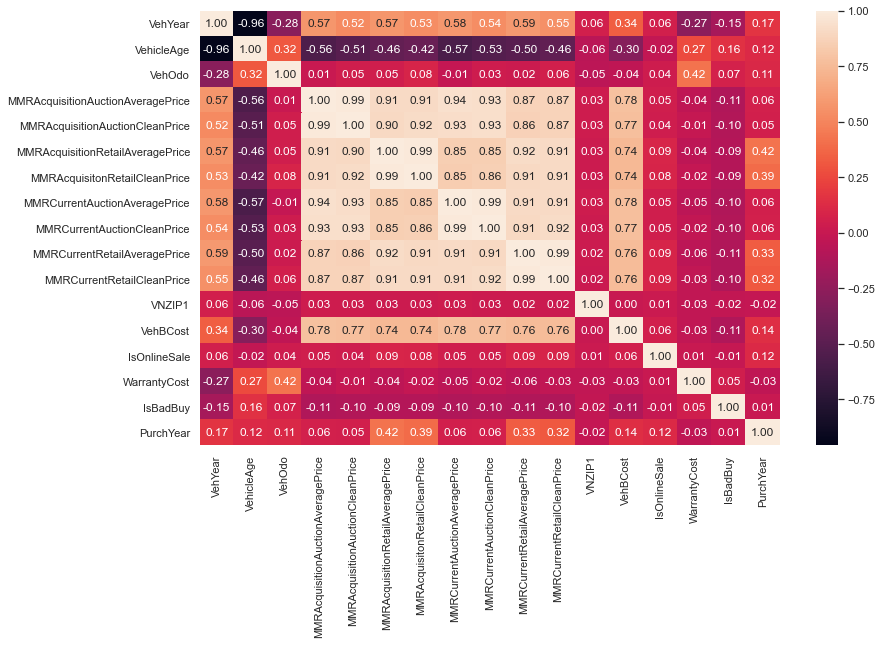
Carvana je Američki start-up, koji se bavi kupovinom polovnih automobile, sa ciljem njihove restauracije I promene fizičke infrastrukture novom, modernom tehnologijom. Kako bi se postarali da je kupljeni automobil u dobrom stanju, moraju ga kategorizovati klasu „dobra“ ili „loša“ kupovina. Kategorizacija se vrši na osnovu prikupljenih podataka vezanih za osnovne karakteristike automobila (proizvođač, marka, vrsta prenosa, tip automobila, model, boja, tip felni..), raznih cena koje se vezuju za automobil, trošak prodaje tog automobila, da li je kupljen online ili ne, troškovi garancije automobila...

Cilj ove analize jeste da se pomoću različitih klasifikacionih algoritama, naprave optimalni modeli koji će moći da na bazi ulaznih podataka kategorišu automobil kao dobru ili lošu kupovinu, kako bi omogućili bolje i uspešnije poslovanje. Korišćenjem optimalno napravljenih modela Carvana može ostvariti niže troškove, bolje poslovanje, manje napravljenih grešaka prilikom kupovine polovnih automobila, kao I generalno bolje finansijsko poslovanje.

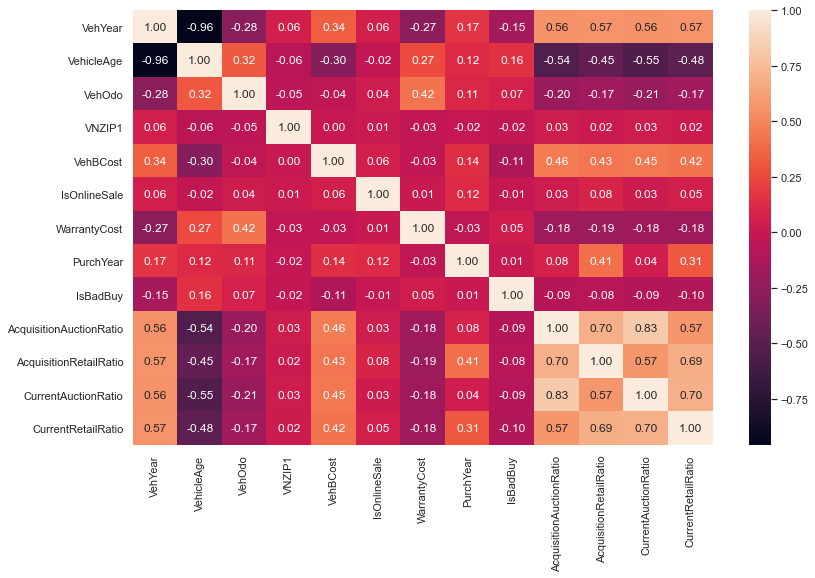
# Opis I razumevanje podatak

Početna baza podataka na osnovu koje kreiramo modele ima 6798 opservacija posmatranih na 30 varijabli, od kojih su 12 numeričke, a 18 kategoričke. Proverom nedostajućih vrednosti ustanovljeno je da varijable PRIMEUNIT, AUCGUART , VNST, Auction imaju veki broj NA vrednosti, pa sui z tih razloga izbačene iz razmatranja. Varijable RefId, WheelTypeID, BYRNO su izbačene iz razmatranja jer za svaku observaciju imaju jedinstvenu vrednost. Varijabla PurchDate u izvornom format ne nosi nikakve informacije, pa je ona iskorišćena za pravljenje nove varijable, tako što je iz celog datuma izvučena samo godina. Kategoričke varijable koje sui male NA vrednosti, su popunjene vrednošću koja se najviše puta pojavila u skupu podataka, dok je nad numeričkim varijablama prvo primenjen Shapiro test, koji ispituje da li varijabla ima Normalnu raspodelu. Za slučaj da varijabla nema normalnu raspodelu, NA vrednosti su popunjavane medijanom, dok za slučaj da imaju Normalnu raspodelu, popunjavane su prosečnom vrednošću.

Kreiranjem korelacione matrice, ispitana je zavisnost medju svim atributima. Visoka korelacija se posebno ističe kod varijabli koje nose informacije o ceni automobile (MMRAcquisitionAuctionAveragePrice, MMRAcquisitionAuctionCleanPrice, MMRAcquisitionRetailAveragePrice, MMRAcquisitonRetailCleanPrice, MMRCurrentAuctionAveragePrice, MMRCurrentAuctionCleanPrice, MMRCurrentRetailAveragePrice, MMRCurrentRetailCleanPrice). Kako bise smanjila korelacija između ovih varijabli, kreirani su novi atributi. Novi atributi su napravljeni kao racio koji treba da pokaže koliko je average cena veća/manja od clean cene, tako da preko tog odnosa uvek može da izračuna jedna ili druga cena.



Nova matrica korelacije pokazuje značajno manje multikolinearnosti.



Kako bi podaci bili pogodni za ulaz u algoritam, izvršeno je dummy kodiranje nad object varijablama kao I standardizacija numeričkih varijabli (zbog postojanja autlajera kod nekih varijabli). Ovako sređen dataset je sačuvan pod nazivom “carvana\_final.csv”.

# Eksperiment

Izabrani modeli koji će biti korišćeni za pravljenje predikcija jesu stablo odlučivanja I KNN klasifikator, a kako bismo poboljšali rezultate, na osnovu ovih osnovnih kreirani su I kompleksniji modeli, Voting Bagging.

U cilju boljeg predviđanja performansi naših modela na novim podacima, podelili smo celokupan skup podataka na train i test deo u odnosu 70:30, gde su pomoću kros validacije određene optimalne vrednosti parametara i hiperparametara naših modela na osnovu metrike roc\_auc. Na test delu su merene performanse tako kreiranih modela na novim podacima. Modeli na test setu evaluirani su pomoću standardnih evaluacionih metrika : accuracy, precision, recall, F1.

# Rezultati

Modeli su učeni na train delu skupa podataka, a zatim testirani na test datasetu. U tabeli1 u prikazani rezultati dobijeni nakon testiranja modela na test setu, pri čemu su ostavljeni default hiperparametri za svaki model, a korišćeni su svi atributi bez obzira na njihov značaj.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **ROC\_AUC** | **Accuracy** | **Precision** | **Recall** |
| **DT default** | 0.53127953563501 | 0.759313725490196 | 0.13897280966767372 | 0.18253968253968253 |
| **KNN default** | 0.551016590138326 | 0.8676470588235294 | 0.23529411764705882 | 0.031746031746031744 |
| **Voting default** | 0.5560882027793113 | 0.8083333333333333 | 0.15422885572139303 | 0.12301587301587301 |
| **Bagging default** | 0.5642404835737926 | 0.8715686274509804 | 0.3684210526315789 | 0.05555555555555555 |

Tabela

Pomoću metode GridSearchCV pronalazimo optimalne hiperparametre za modele, a zatim kreiramo nove modele, I testiramo ih na test setu. U tabeli2 su prikazani rezultati dobijeni nakon testiranja na test setu, nad modelima koji su podešeni sa optimalnim hiperparametrima, a gde su korišćeni svi atributi bez obzira na njihov značaj.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **ROC\_AUC** | **Accuracy** | **Precision** | **Recall** | **F1** |
| **DT1** | 0.5987646193261769 | 0.8764705882352941 | 0.0 | 0.0 | 0.0 |
| **KNN1** | 0.6072556955802424 | 0.8745098039215686 | 0.3 | 0.011904761904761904 | 0.022900763358778622 |
| **Voting1** | 0.6227119387350981 | 0.8764705882352941 | 0.0 | 0.0 | 0.0 |
| **Bagging1** | 0.6077317382713819 | 0.8759803921568627 | 0.4444444444444444 | 0.015873015873015872 | 0.03065134099616858 |

Tabela

Primenom metode SelectKBest izvlačimo 23 najznačajnijih atributa, u zavisnosti od njihove korelacije sa izlaznom varijablom, I pravimo finalne modele koji će biti trenirani nad train setom koji se sastoji upravo od podataka za tih 23 atributa, I podešavamo optimalne hiperparametre pomoću GridSearchCv metode. Rezultati su prikazani u tabeli3.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **ROC\_AUC** | **Accuracy** | **Precision** | **Recall** | **F1** |
| **DT final** | 0.5969727457104224 | 0.8735294117647059 | 0.3125 | 0.01984126984126984 | 0.03731343283582089 |
| **KNN final** | 0.6165521901060814 | 0.8759803921568627 | 0.3333333333333333 | 0.003968253968253968 | 0.00784313725490196 |
| **Voting final** | 0.6414620805126714 | 0.8764705882352941 | 0.0 | 0.0 | 0.0 |
| **Bagging final** | 0.6210331179022146 | 0.8774509803921569 | 1.0 | 0.007936507936507936 | 0.015748031496062992 |

Tabela

# Zaključak

Potencijalne mogućnosti za poboljšanje modela jesu da se prikupe dodatne informacije I na taj način doda broj atributa koji će dodatno opisati automobile npr. Potrošnja goriva, snaga, klasa, broj kvarova I slično. Ove dodatne informacije bi mogle da poboljšaju prediktivnu moć modela. Neka od mogućih poboljšanja su takođe detaljnije tumačenje I sređivanje podataka, s obzirom da je radi potrebe sprovođenja analize, u jednom trenutku sređivanja podataka, namerno unet šum, tako što su određene vrednosti koje su “pravile problem” zamanjene neutralnom nulom.