1. Opis i razumevanje problema

U ovom projektnom radu analiziran je jedan od najpoznatijih nemačkih drogerijskih lanaca prodavnica Rossman, koji poseduje preko 3000 prodavnica u 7 evropskih država. Menadžment kompanije Rossman je postavio zadatak da se za 6 nedelja unapred predvidi dnevna prodaja. Na prodaju utiču mnogi faktori, kao na primer konkurencija, školski i državni praznici, promocije, sezona i drugi.

2. Opis i razumevanje podataka

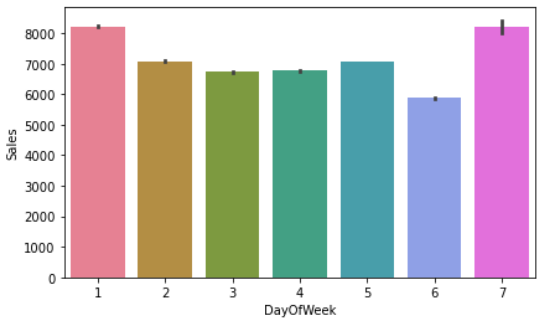
Za potrebe ovog projektnog rada korišćena su 3 skupa podataka, *Train*, *Test* i *Store.* Podaci o prodaju su prikupljeni u 1115 Rossman prodavnica.

Na *Train* skupu podataka je vršeno učenje modela i predviđanje prodaje, dok je na *Test* skupu podataka vršeno testiranje istih modela. Dodatne informacije o prodavnicama čine poseban skup podataka – *Store*.

U Train i Test skupu podataka su uključeni sledeći podaci: *Store* – jedinstveni ID za svaku prodavnicu, *DayOfWeek* – dan u nedelji na koji se podaci odnose, *Date* – datum kada je izvršeno prikupljanje podataka, *Open* – pokazatalj da li je prodavnica bila otvorena ili ne za određeni datum, *Promo* – pokazatelj da li je bila aktivna promocija tog dana, *StateHoliday* – parameter koji pokazuje da li je tog dana bio državni praznik i koji (državni, Uskrs, Božić), *SchoolHoliday* – parameter koji pokazuje da li su škole radile tog dana.

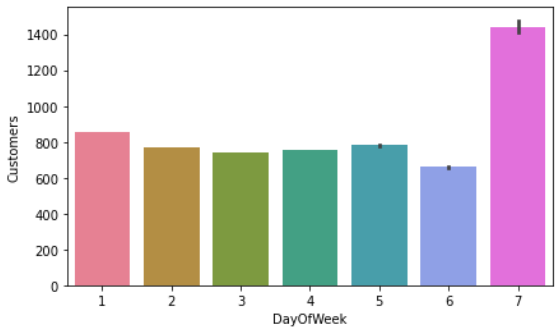
Postoje podaci koji su sadržani u Train skupu, ali nisu u Test skupu podataka, a to su: *Sales* – predstavlja ostvarenu prodaju i *Customers* – označava broj kupaca. *Store* skup podataka sadrži sledeće podatke: *Store, StoreType* – predstavlja tip prodavnice (a, b, c i d), *Assortment* – označava tip asortimana (osnovni, ekstra, prošireni), *CompetitionDistance* – pokazuje udaljenost najbliže konkurencije, *CompetitionOpenSinceMonth* – označava mesec kada je konkurencija otvorila svoju prodavnicu, *CompetitionOpenSinceYear* – označava godinu kada je konkurencija otvorila svoju prodavnicu, *Promo2* – parameter koji pokazuje da li prodavnica kontinuirno učestvuje u promociji, *Promo2SinceWeek* – označava kalendarsku nedelju od kada prodavnica učestvuje u *Promo2*, *Promo2SinceYear* – označava kalendarsku godinu od kada prodavnica učestvuje u *Promo2*, *PromoInterval* – predstavlja vremenske intervale za svaku prodavnicu u kojima *Promo2* počinje. U narednim vizualizacijama biće prikazani neki od podataka.

Na slici 1 je prikazana prosečna prodaja po danima u nedelji. Može se uočiti da je nedeljom i ponedeljkom najveća prodaja.



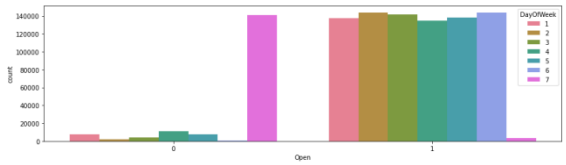
Slika 1 Prosečna prodaja po danima u nedelji

Slika 2 prikazuje prosečan broj kupaca po danima u nedelji. Slično kao i u slučaju prodaje, najveći broj kupaca je nedeljom i ponedeljkom u radnjama.



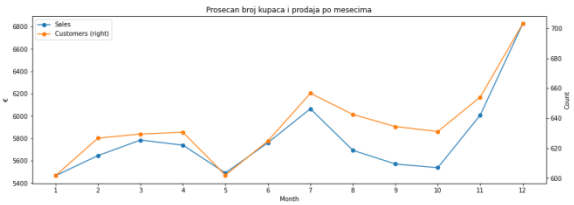
Slika 2 Prosečan broj kupaca po danima u nedelji

Na slici 3 dat je prikaz broja opservacija po danima u nedelji u zavisnosti od toga da li je prodavnica bila otvorena ili zatvorena.



Slika 3 Broj opservacija po danima

Na slici 4 je dat uporedni prikaz broja kupaca i prodaje po mesecima. Može se zaključiti da obe krive imaju sličnu putanju kretanja, tj. da su vrednosti ova 2 atributa najmanje u maju, a najveće u julu i decembru.



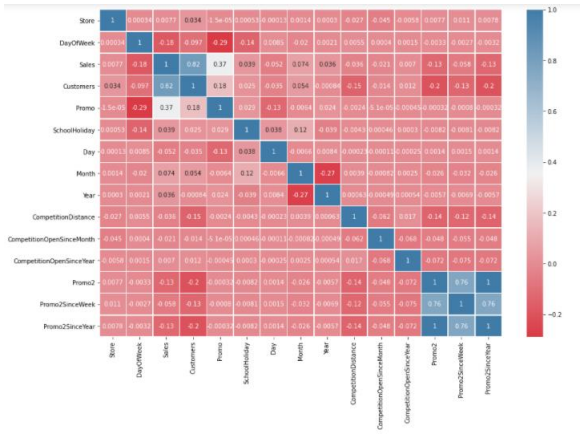
Slika 4 Prosečan broj kupaca i prodaja po mesecima

3. Priprema podataka

Nakon analaziranja i vizualizacije podataka bilo je neophodno popuniti nedostajuće vrednosti. *Train* skup podataka ne sadrži nedostajuće vrednosti, dok *Store* skup podataka sadrži nedostajuće vrednosti po sledećim atributima: *CompetitionDistance* – 3, *CompetitionOpenSinceMonth* – 354, *CompetitionOpenSinceYear* – 354, *Promo2SinceWeek* – 544, *Promo2SinceYear* – 544 i *PromoInterval* – 544. Vrednosti koje nedostaju u atributu *CompetitionDistance* su popunjenje srednjom vrednošću tog atributa budući da postoje samo 3 nedostajuće vrednosti, pa ovakav način popunjavanja istih ne utiče na rezultate. Nedostajuće vrednosti aributa *PromoInterval, Promo2SinceWeek i Promo2SinceYear* popunjene su 0 budući da opservacije koje ih sadrže imaju vrednost atributa *Promo2* koja je jednaka 0. Za atribute *CompetitionOpenSinceMonth i CompetitionOpenSinceYear* je korišćen KNN Imputer koji na osnovu broja najbližih suseda popunjava nedostajuće vrednosti. U *Test* skupu podataka nalazi se 11 nedostajućih vrednosti za atribut *Open*, koje su bile popunjene medijanom, jer je u većem broju slučajeva prodavnica bila otvorena.

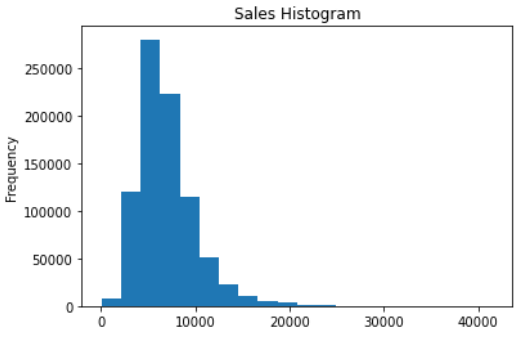
Kada su popunjene nedostajuće vredosti, bilo je potrebno spojiti podatke o prodavnicama sa *Train* i *Test* skupom. Iz skupa podataka se izbacuju oni slučajevi kada prodavnica nije radila i kada je prodaja bila jednaka 0. Pomenute opservacije mogu izazvati šum u podacima i zato je atribut *Open* izbačen. Matricom korelacije atributa koja je prikazana na slici 5 se može utvrditi koliko atributi utiču jedni na druge. Uočeno je da su atributi *Promo2SinceYear* i *Promo2SinceWeek* visoko

korelisani sa atributom *Promo2* i iz tog razloga se uklanjaju iz skupa podataka. Kategoričke promenljive iz *Train* i *Test* skupa podataka se pretvaraju u numeričke podatke, tako da će biti novih numeričkih, binarnih atributa onoliko koliko je bilo mogućih vrednosti kategorija za svaku kategoričku promenljivu. Atributi koji su pretvoreni u numeričke su *StateHoliday, StoreType, Assortment i PromoInterval.*



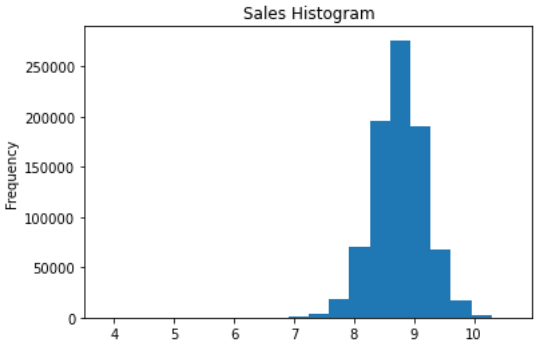
Slika 5 Matrica korelacije

Nakon sređivanja ulaznih podataka, bilo je potrebno pogledati raspodelu izlaznog atributa. Na slici 6 može se primetiti da je raspodela pomerena u levo i zato je bilo potrebno logaritmovati promenljivu *Sales*.



Slika 6 Histogram promenljive Sales

Nakon logaritmovanja raspodela promenljive Sales prikazana je na slici 7.

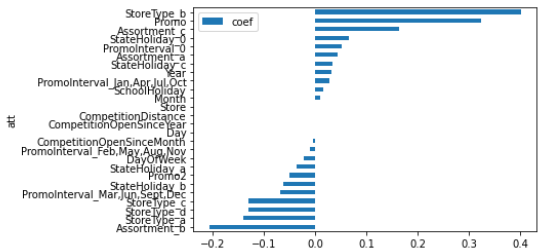


Slika 7 Historam promenljive Sales nakon logaritmovanja

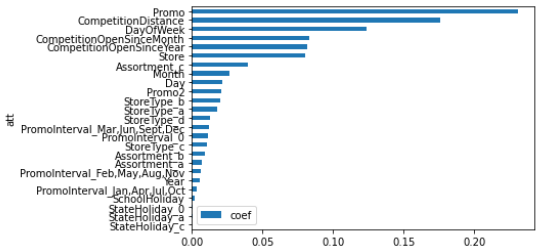
4. Podešavanje parametara algoritma i selekcija/otežavanje atibuta

U ovom projektnom radu korišćeni su sledeći algoritmi: *Linear regression, Decision tree, KNN, Random forest, XGBoost.* Vršena je optimizacija parametara algoritma i selekcija atributa u zavisnosti od njihove značajnosti. Izvršena je optimizacija po sledećim parametrima: *Decision tree - max_features, min_samples_leaf i max_depth, KNN – n_neighbors, Random forest – n_estimators, XGBoost - learning_rate, min_split_loss i max_depth.*

Prikaz atributa po značajnosti za određeni model da je na sledećim slikama (slika 8 i slika 9)



Slika 8 Značajnost atributa za Linearnu regresiju



Slika 9 Značajnost atributa za Drvo odlučivanja

5. Poređenje performansi algoritama

Model		RMSPE	
		Train	Test
LR	SEL	0.0436	0.0435
DT	OPT	0.0343	0.0348
	SEL	0.0297	0.0297
KNN	OPT	0.0	0.0203
RF	OPT	0.0057	0.0145
	SEL	0.0059	0.0151
XGB	OPT	0.0166	0.0169
	SEL	0.0162	0.0165

Tabela 1 Poređenje performansi

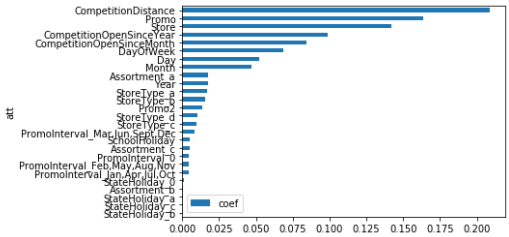
6. Evaluacija rešenja i primena modela

Nakon procesa učenja modela neophodno je evaluirati i oceniti koliko je on dobar. Mera evaluacije koja je korišćena je RMSPE – Root Mean Square Percentage Error (koren srednje kvadratne procentualne greške). Formula za RMSPE je:

$$\mathrm{RMSPE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2}$$

gde y_i predstavlja stvarnu vrednost prodaje, a yhat_i odgovarajuću predikciju za prodaju.

Na osnovu uporednog prikaza performansi algoritama koji je dat u tabeli 1, može se zaključiti da je najbolji model *Random forest* koji ima najmanju grešku – najmanju vrednost RMSPE - a. Izvršena je optimizacija parametra algoritma i dobijena je najbolja vrednost za n_estimators koja iznosi 30. Za ovaj model najznačajniji atributi su: *CompetitionDistance, Promo, Store, CompetitionOpenSinceYear, CompetitionOpenSinceMonth, DayOfWeek, Day, Month, Year, StoreType_a, Assortment_a, StoreType_b.* Ukoliko se uporedi greška pre i posle izvršene selekcije atributa uviđa se da promena nije značajna, tj. da je sa manjim skupom atributa dobijena slična greška.



Slika 10 Značajnost atributa za Random forest

7. Zaključak

U cilju uspešnijeg poslovanja kompanije Rossman neophodno je konstantno pratiti podatke i vršiti analizu i predviđanja u svrhu otkrivanja velikih promena u prodaji. Pre nego što se krene sa analizom i predviđanjem, od velikom značaja je pripremiti podatke na adekvatan i ispravan način. Prodaja koja je predviđena može se koristiti kao ulaz za neka druga predviđanja koja su korisna za poslovanje.