Primena algoritama mašinskog učenja

Rossman Store Sales Prediction

Student: Stevan Nešić 3159/2022

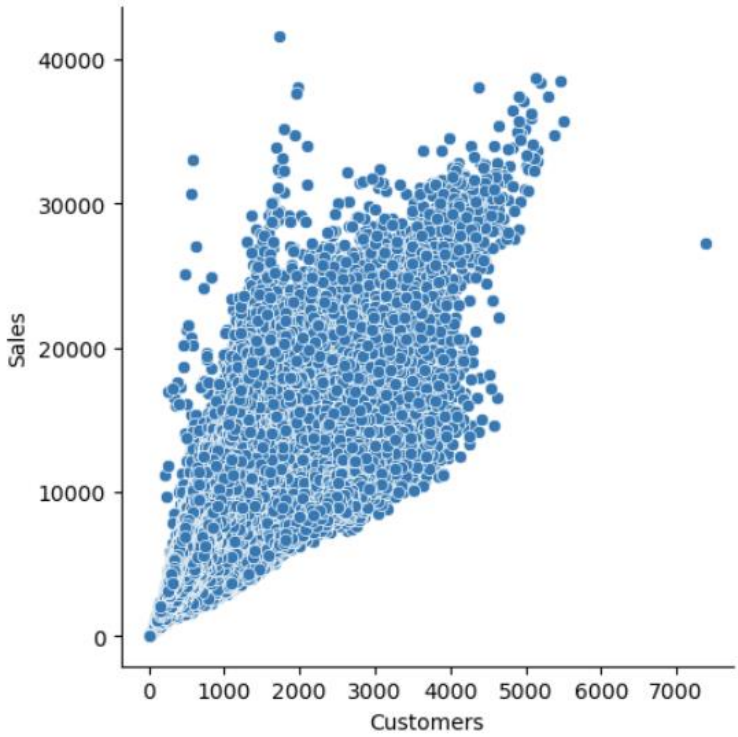
Uvod

Rossmann je kompanija koja se bavi prodajom u Evropi sa 1115 prodavnica. Tako velikoj firmi je da planiraju i predviđaju svoju prodaju. Cilj zadatka je da korišćenjem metoda mašinskog učenja, napravimo model koji najpreciznije predviđa prodaju po svakoj od prodavnica. Podaci su zadati u dnevnim prodajama I kretanjima kupaca za svaku prodavnicu, tip prodavnice, vremenski periodi (datum, dan u nedelji, neradni dani), kolika je blizina konkurencija i kada je otvorena konkurentska radnja u blizini, kao i da li prodavnica ima promocije.

Vizuelizacija i analiza podataka

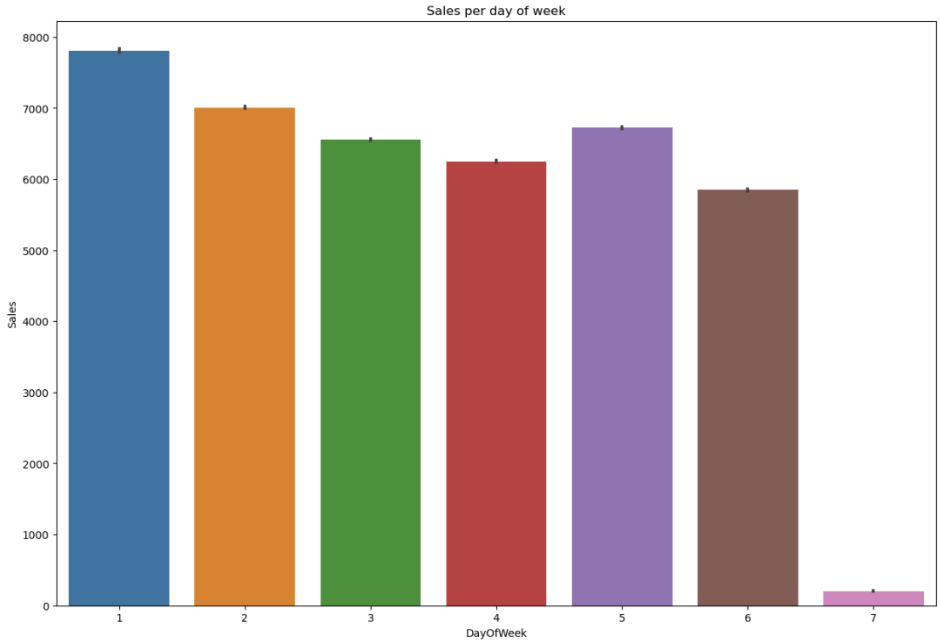
Skup podataka se sastoji iz 2 tabele, jedna u kojoj se nalaze podaci 1115 prodavnica i druge u kojoj se nalaze podaci o prodaji za 1017209 merenja. Prva tabela sadrži 10 kolona koje se odnose na same informacije o prodavnica, dok druga sadrži 9 kolona koje bliže opisuje stanje u kojoj je prodavnica bila u trenutku merenja.

Na Slici 1 se vidi kakav je odnos prodaje i broja kupaca i jednoj radnji, kao što se vidi na slici linearno raste sa brojem kupca.



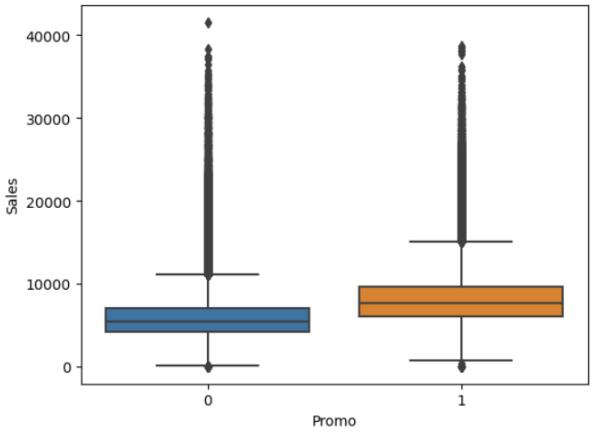
Slika 1. Odnos između broja kupaca i prodaje

Takođe vrlo važan podatak je kog dana se obično kupovina izvršava, obzirom na Sliku 2 možemo zaključiti da je najveći broj kupovina ponedeljkom, a razlog tome će najverovatnije biti to što prodavnice većinom ne rade nedeljom.

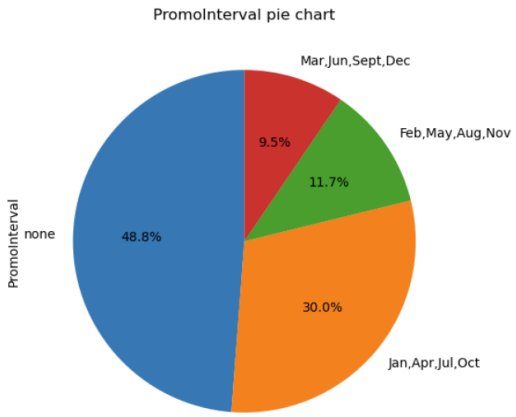


Slika 2. Kupovine po danima u nedelji

Najvažniji podaci u ovom skupu podatka su podaci vezani za promocije i to se može videti. Uz pomoć promocije se postiže veća prodaja (Slika 3) i promocija je korišćena strogo četiri meseca i to u samo tri različita termina (Slika 4).



Slika 3. Prodaja uz korišćenje promocije



Slika 4. Promotivni interval

Pripremanje podataka

Nakon analiz ei popunjavanja nedostajućih vrednosti u skupu podataka za prodavnice, Train skup i Store skup spojeni su u jedan skup podataka. Izvršeno je “inner join” spajanje po koloni Store.

Atribut Date je transformisan i na osnovu njega dobijene su 3 nove kolone: Day, Month i Year. Nakon toga je Date obrisan.

CompetitionOpenSinceYear i CompetitionOpenSinceMonth su integrisani u jednu kolonu koja je dobila naziv CompetitionOpenSince i vrednosti ove kolone predstavljaju ukupan broj meseci od otvaranja najbližeg konkurenta. Nakon toga su CompetitionOpenSinceYear i CompetitionOpenSinceMonth obrisane.

Takođe, vrednosti StateHoliday su transforimsane. Ukoliko je vrednost a, b ili c StateHoliday uzima vrednost 1, a ukoliko je 0 onda ostaje 0.

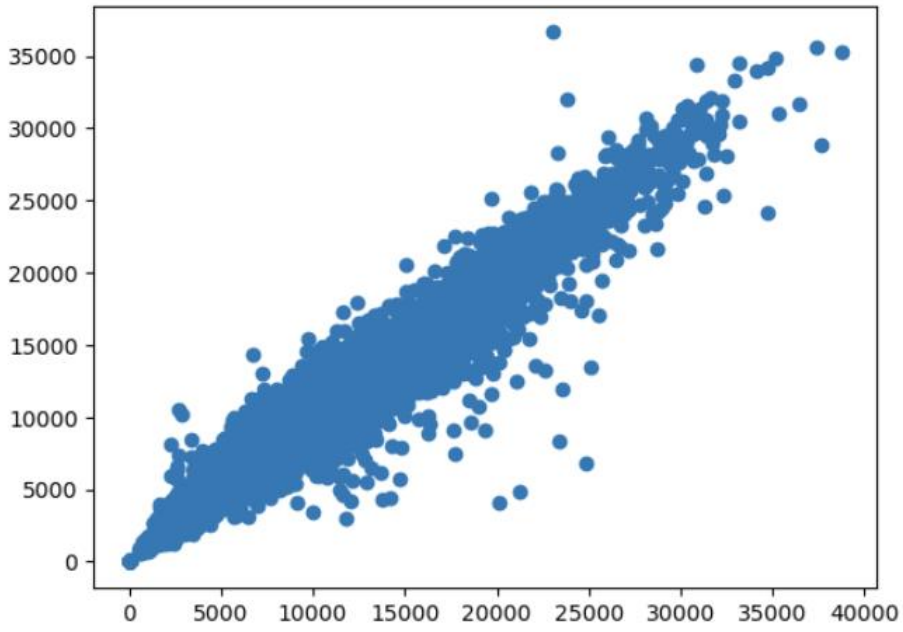
Sve kategoričke vrednosti su “dummy” kodovane i dobijamo konačan skup podataka od 27 kolona.

Modelovanje

Konačni skup podataka podeljen je u train i test skupove. Prvo je primenjen model Linearne regresije, Lasso, Ridge, Stabla odlučivanja i RandomForestRegressor, kao i Stacking ansambla. Na kraju se najbolje pokazao RandomForestRegressor koji je imao najmanje greške.

Kada je odrađena i dodatna selekcija atributa, korišćenjem metode VarianceThreshold koja otklanja atribute kojima je varijansa mala, zbog čega imaju slab uticaja na model predviđanja, broj atributa nam se smanjio na 24 Finalno, Random Forrest Regressor se opet pokazao kao najbolje rešenje, sa 98,4% objašnjenog varijabiliteta i greškom od 422,16.

Osim toga, možemo primetiti da je linearna regresija, čak i nakon regularizacija najgore rešenje za ovaj model sa objašnjenih 90% varijabiliteta i greškom od 1090,82.



Na osnovu velikog broja opservacija koje je kompanija Rossmann izmerila i u skladu sa pretpostavkama koje smo napravili i metodama koje smo koristili, računar je uspeo da dođe do modela, koji može sa velikom preciznošću da radi predviđanje prodaje po prodavnici za naredni period. Finalna preporuka bi bila da se koristi model dobijen metodom Random Forrest Regressor s obzirom na to da je u svakom testova pokazala najbolji rezultat, objasnivši najveći procenat varijabiliteta, sa najmanjim korenom srednje kvadratne greške. Konkretno smatramo da je najbolja opcija koristiti model dobijen,preciznijom selekcijom atributa i podrazumevanima parametrima.

Iako je su modeli dali vrlo dobre rezultate, možda bi najbolje bilo pokušati primeniti neki od modela za vremenske serije. Problem predviđanja prodaje se najčešće rešava primenom modela za vremenske serije i neki od mogućih pravaca je korišćenje neuronske mreže.