**Predikcija ocene drustvenih igara**

**1.Cilj**

Cilj ovog projekta je predikcija ocene drustvenih igara na osnovu karakteristika koje crpimo iz dataseta. Predikcija se vrsi pomocu:

* **linearne regresije** - Najcesce se linearna regresija odnosi na model u kojem je uslovna srednja vrednost od *Y*, uz datu vrednost *X*, funkcija od *X*. Slučaj sa jednom nezavisnom promenljivom se naziva jednostavna linearna regresija. Kad je obuhvaćeno više od jedne nezavisne promenljive, proces se zove visestruka linearna regresija.
* **random forest regresija** - Random forest regresija je metod za klasifikaciju, regresiju i ostale zadatke koji se mogu uraditi kreiranjem vise "drveca odluke" i koriscenjem mean predikcije(u slucaju regresije). Drveca u random forest regresiji su podlozna "overfittingu" i zato treba biti obazriv.

linearne regresije i pomocu random forest regresije i nakon toga se rezultati dve predikcije porede. Projekat je radjen u jupyter notebook-u u python programskom jeziku.

**2.Dataset**

Dataset koji je koriscen u izradi aplikacije se zove Board game geek data. Dataset je vlasnistvo sajta BoardGameGeek i otvoren je za koriscenje u nekomercijalne svrhe. Originalni dataset sadrzi 101 hiljadu igara sa 21 karakteristikom igara.

Karakteristike igrica su:

|  |  |
| --- | --- |
| *Id*(kljuc) | Integer |
| *Type* | String |
| *Name* | String |
| *Yearpublished* | Float |
| *Minplayers* | Float |
| *Maxplayers* | Float |
| *Playingtime* | Float |
| *Minplaytime* | Float |
| *Maxplaytime* | Float |
| *Minage* | Float |
| *Users\_rated* | Integer |
| *Average\_rating*(ovo se predvidja) | Float |
| *Bayes\_average\_rating* | Float |
| *Total\_owners* | Integer |
| *Total\_traders* | Integer |
| *Total\_wanters* | Integer |
| *Total\_wishers* | Integer |
| *Total\_comments* | Integer |
| *Total\_weights* | Integer |
| *Average\_weight* | Float |
| *Meta* | String |

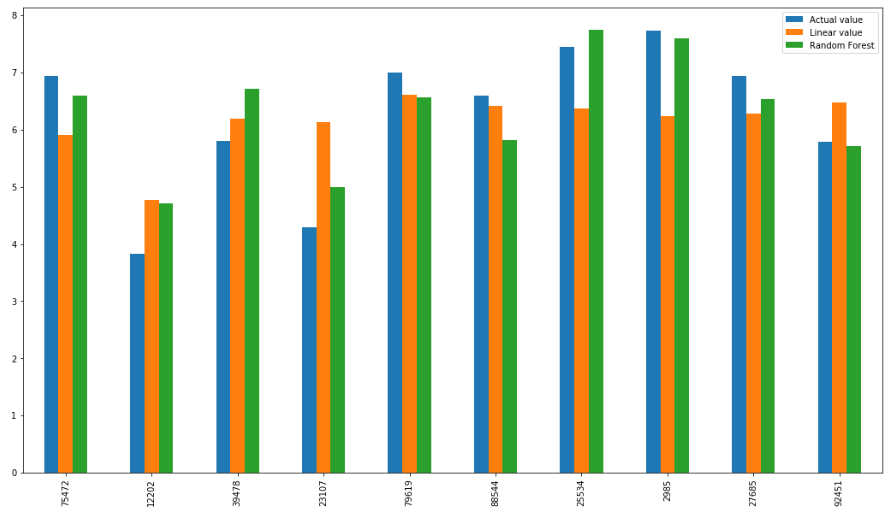
**3.Opis izrade**

Prva faza u izradi projekta je predprocesiranje podataka, pre svega na redu je bilo uklanjanje null vrednosti iz dataseta. Dobra stvar je sto null vrednosti nije bilo mnogo, na mesto int i float tipova podataka gde je umesto vrednosti stajao null je uradjena median funkcija koja trazi neku srednju vrednost iz dataseta i ubacuje je u dataset dok su na mesto gde umesto Stringa stoji null uklanjane cele igrice. Dalje su izbacene sve igrice koje imaju iste nazive, originalna igrica ostaje u datasetu dok se duplikati izbacuju. Zatim su izbacene sve igrice kojima je prosecna ocena koja se predvidja 0.

Posle toga je radjen Tukey InterQuartile range koji izbacuje ekstremne vrednosti nekog parametra, u ovom slucaju su izbacivane samo ekstremne vrednosti yearpublished karakteristike odnosno karakteristike koja govori koje godine je igrica nastala.

Dalje su prikazani grafici gde je radjena analiza.

Nakon toga podaci su razdvojeni na trenazne i test podatke na osnovu 75% - 25% podele. Zatim su uradjeni modeli za linearnu regresiju i random forest regresiju i poredjene vrednosti u graficima. Mean squared error je pre svega koriscen kao pokazatelj kvaliteta predikcije ali i mean absolute error, kao i root mean squared error.



**Slika 1.** Grafik predikcije za 10 igrica