Project - Football Player Price Prediction

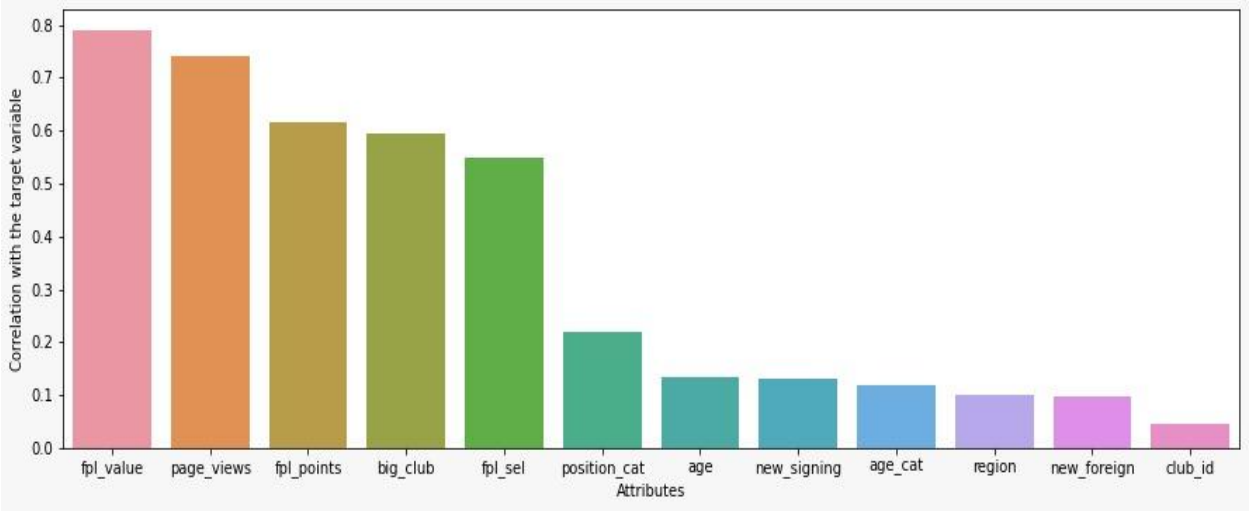
Team Members
Nisha Bharti
Atul Kumar
Lovepreet Singh

EXPLORATORY DATA ANALYSIS

- ➢ Importing various libraries for analysis (pandas, matplotlib, NumPy, seaborn)
- ➢ Reading the dataset
- ➢ Noting general observations
  - ❖ Rows, columns- (461, 17)
  - ❖ Continuous variable and Discrete Variables
  - ❖ Discrete Variables Count: 8
  - ❖ Discrete_feature-['position_cat','fpl_value','region','new_foreign','age_cat','club_id','big_club', 'new_signing']
  - ❖ Continuous feature Count 5
  - ❖ Categorical features 4
- ➢ Handling missing values, removing unwanted symbols, dropping unwanted columns/features

OBSERVATIONS USING GRAPHS AND PLOTS:

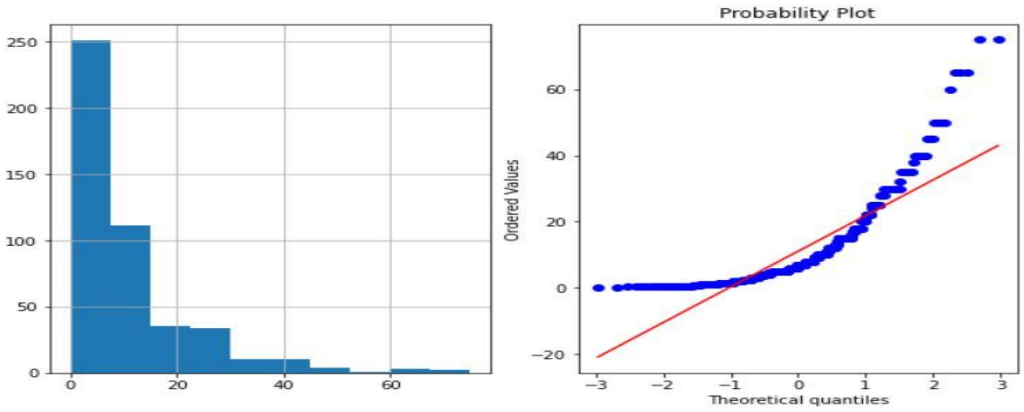
- ➔ Selection of features
- ➔ features = ['age', 'position_cat', 'page_views', 'fpl_value', 'fpl_sel','fpl_points', 'region', 'new_foreign', 'age_cat', 'club_id', 'big_club', 'new_signing']
- ➔ Find **correlation** for market values using a bar graph



- ➔ Implementation of various other plots such as heatmaps, scatterplot, line graph, boxplot, etc to analyze the data.

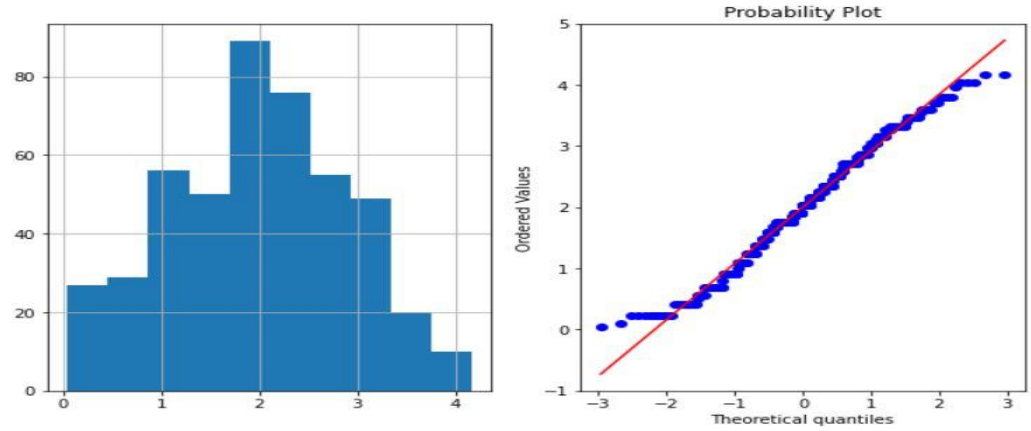
To check the skewness of the data:

- ➢ For example, market value



The second plot is a Q-Q plot that tells whether the feature is Gaussian distributed or not

- Applying boxcox transformation skewness would reduce



- For market_value skewness is 2.1625484427984265(before applying boxcox)
- For market_value skewness is 0.00392136767502309(after applying boxcox)

Implementation and Selection of model

Algorithm name	Accuracy	MSE Score
Linear regression	**0.6800122500857713**	**0.2306339052132473**
Lasso	**0.6870535253156608**	**0.4749303590211173**
Ridge	**0.68000640894826**	**0.4802479726797647**
Support Vector Machine	**0.705156474580491**	**0.46098914362286003**
Decision Tree	**0.5504786127846979**	**0.5692068000565597**
RandomForest	**0.7750325302250248**	**0.4026750877629798**
GradientBooster	**0.7895039278174102**	**0.3895084630974733**
KNN	**0.6613596017344051**	**0.494042468730294**

Based on the above table, GradientBooster is the best fit model as accuracy is highest in it.

Hyperparameter Tuning

Algorithm name	Old Accuracy	New Accuracy
Linear regression	0.6800122500857713	0.6800122500857706
Lasso	0.6870535253156608	0.6880855800872612
Ridge	0.68000640894826	0.6815343411071076
Support Vector Machine	0.705156474580491	0.7837269725235378
Decision Tree	0.5504786127846979	0.6571994877682152
RandomForest	0.7750325302250248	0.7912458769240002
GradientBooster	0.7895039278174102	0.8104958622978086
KNN	0.6613596017344051	0.7066582705660307

- **GradientBooster turned out to be the best model overall.**

**** We tried one more thing in the feature selection process where the club feature which is having 20 unique club names we divided them into three categories High, Medium, and Low based on the impact that they were having on the market value and then created dummy variables from those categories but the overall performance on the models remained same so we dropped that club column.**

Model Deployment

**** Link to Vist the web application: https://football-player-price.herokuapp.com/**

**** https://github.com/Lovepreet-Singh-ACET/Football-player-price-prediction**