# Predicting the Price of a Football Player

# USING PYTHON

## PROJECT REPORT


**GROUP MEMBERS:-**

**Neetu Singla**

**Vijay Garg**

**Sai Nikhilesh**

**Nikita Arora**

**Ajaypal Singh**

**Problem Statement:-**In the English Premier League, May - July represents a lull period due to the lack of club football. What makes up for it, is the intense transfer speculation that surrounds all major player transfers today. An important part of negotiations is predicting the fair market price for a player.

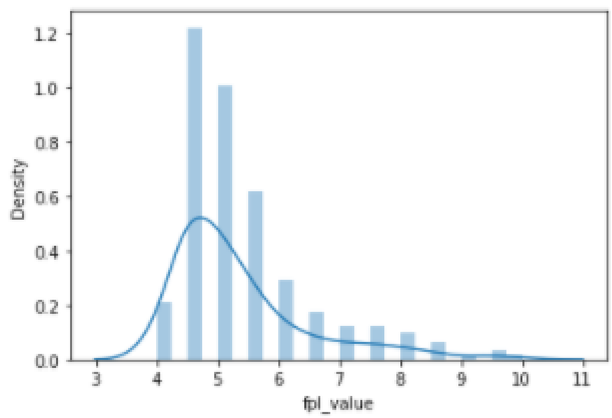
**Objective:-**To develop a model to calculate and predict the fair market price of EPL players using regression algorithms: Linear Regression, Lasso Regression, Ridge Regression, Nearest Neighbour Regression, Support Vector Regression, Tree Regression, Random Forest Regression and Gradient Boosted Regression.

## STUDY METHODOLOGY

### Data Preprocessing & EDA

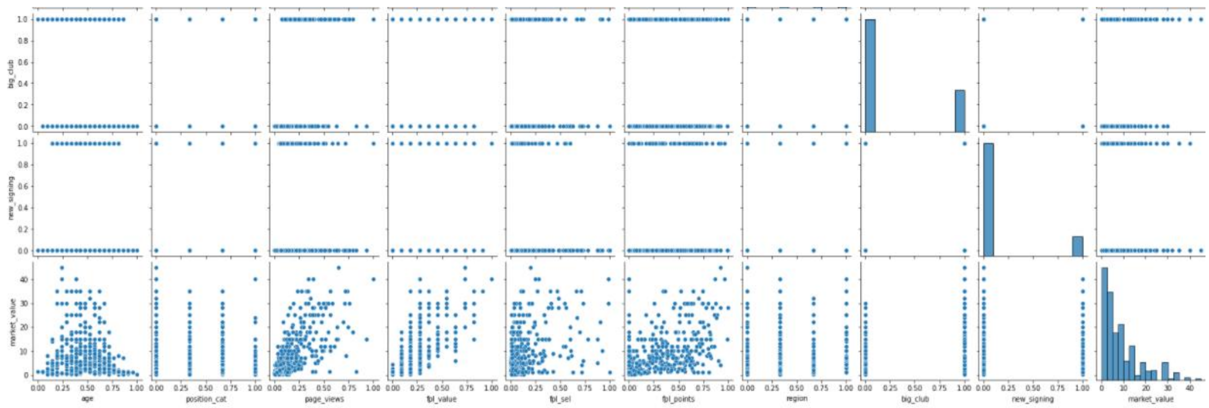
- **Defining variables**
  Regression model uses only quantitative variables hence all the required nominal variables were encoded to Numerical form using one hot encoding .
  We have a total 17 variables, 1 dep(numerical) and 16 indep(4 numerical,12 categorical).
  One missing value was there for the variable **region,** so we dropped that row.

- **Outlier Analysis**
  Outlier's analysis is done to prevent our statistical measures and data distributions from getting skewed because skewness could provide misleading interpretation of the underlying data and relationships.

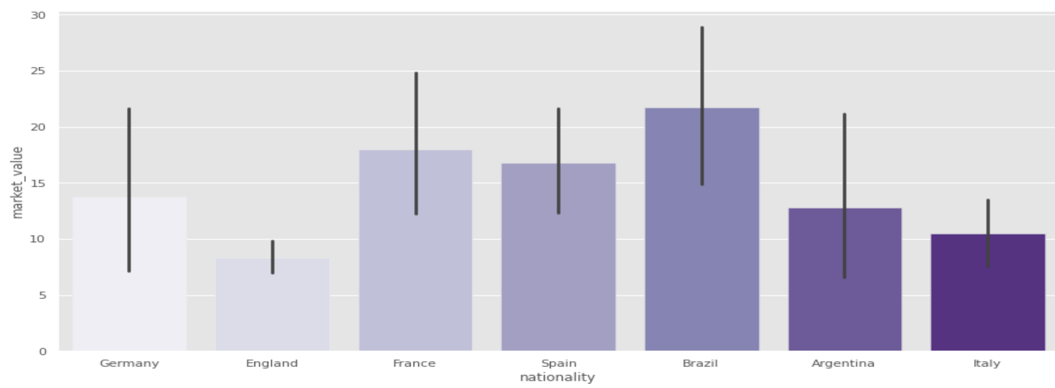



- **Exploratory Data Analysis (EDA)**
  EDA is the process of figuring out what the data can tell us and we use EDA to find patterns, relationships or anomalies to inform our subsequent analysis. The ultimate purpose of the EDA stage in a project is to find some valuable insights that can help with the problem. While there are an almost overwhelming number of methods to use in EDA,

one of the most effective starting tools is the pairs plot. A pair plot allows us to see both distribution of single variable & relationships between two variables.

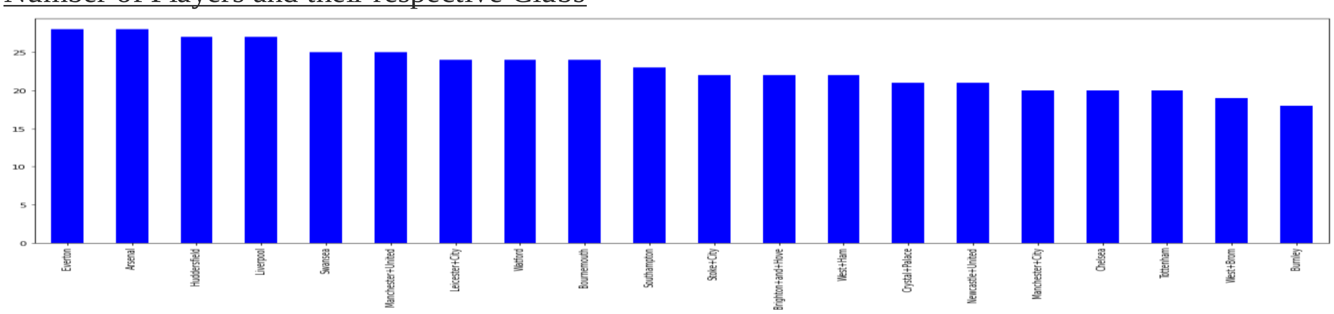



A box plot of nationality vs market_value




It is observed that, rather than England where most players are from, The Foriegn players especially players from France, Argentina and Germany have the highest market value. This is noteworthy since The clubs prefer choosing the best performers from the counties and pay them a good amount to keep them in their team.

Number of Players and their respective Clubs




Best Players for each position with their age, club and nationality based on their fpl_points

Best Players per each position with their age, club, and nationality based on their fpl_points

```
concat_train_data.loc[concat_train_data.groupby(concat_train_data['position'])['fpl_points'].idxmax()][['position', 'name',
                                                                              'age', 'club','nationality', 'page_views']].style.background_gradient('Reds')
```

| | position | name | age | club | nationality | page_views |
|---|---|---|---|---|---|---|
| 378 | AM | Christian Eriksen | 25 | Tottenham | Denmark | 1130 |
| 94 | CB | Gary Cahill | 31 | Chelsea | England | 1420 |
| 377 | CF | Harry Kane | 23 | Tottenham | England | 4161 |
| 376 | CM | Dele Alli | 21 | Tottenham | England | 4626 |
| 396 | DM | Ã‰tienne Capoue | 29 | Watford | France | 412 |
| 74 | GK | Tom Heaton | 31 | Burnley | England | 717 |
| 95 | LB | Marcos Alonso Mendoza | 26 | Chelsea | Spain | 3069 |
| 426 | LM | Chris Brunt | 32 | West+Brom | Northern Ireland | 242 |
| 0 | LW | Alexis Sanchez | 28 | Arsenal | Chile | 4329 |
| 96 | RB | Cesar Azpilicueta | 27 | Chelsea | Spain | 869 |
| 102 | RM | Victor Moses | 26 | Chelsea | Nigeria | 2537 |
| 97 | RW | Pedro | 29 | Chelsea | Spain | 1500 |
| 213 | SS | Roberto Firmino | 25 | Liverpool | Brazil | 2196 |

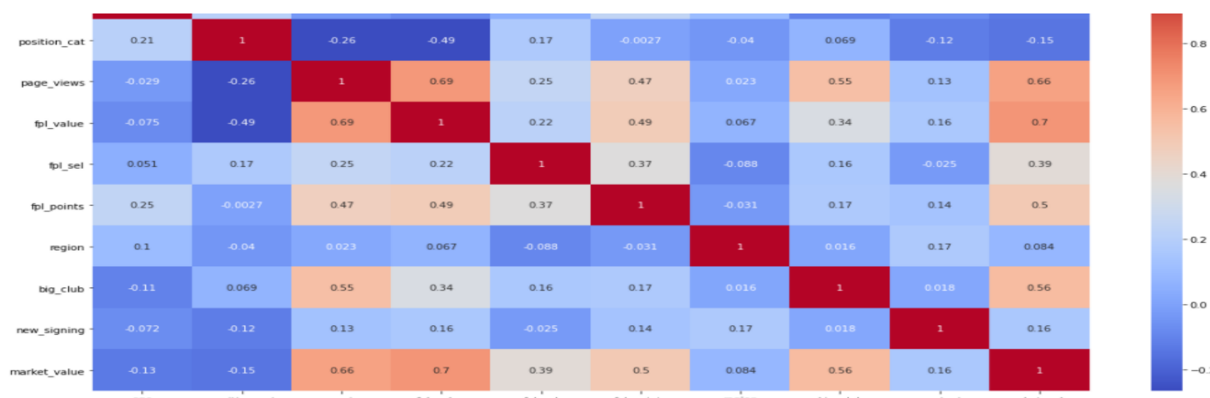
## Best Players for each club ranked by their page views

BEst players for each club ranked by their page views (popularity)

```
[ ] concat_train_data.loc[concat_train_data.groupby(concat_train_data['club'])['page_views'].idxmax()][['position', 'name',
                                                                      'age', 'club','nationality', 'page_views']].style.background_gradient('Reds')
```

| | position | name | age | club | nationality | page_views |
|---|---|---|---|---|---|---|
| 1 | AM | Mesut Ozil | 28 | Arsenal | Germany | 4395 |
| 29 | CF | Jermain Defoe | 34 | Bournemouth | England | 3213 |
| 62 | RW | Anthony Knockaert | 25 | Brighton+and+Hove | France | 726 |
| 74 | GK | Tom Heaton | 31 | Burnley | England | 717 |
| 93 | CF | Diego Costa | 28 | Chelsea | Spain | 4454 |
| 112 | RW | Wilfried Zaha | 24 | Crystal+Palace | Cote d'Ivoire | 1709 |
| 143 | SS | Wayne Rooney | 31 | Everton | England | 7664 |
| 172 | CM | Aaron Mooy | 26 | Huddersfield | Australia | 588 |
| 189 | CF | Jamie Vardy | 30 | Leicester+City | England | 2988 |
| 215 | LW | Sadio Mane | 25 | Liverpool | Senegal | 3219 |
| 251 | CF | Gabriel Jesus | 20 | Manchester+City | Brazil | 4254 |
| 263 | CM | Paul Pogba | 24 | Manchester+United | France | 7435 |
| 302 | CF | Dwight Gayle | 26 | Newcastle+United | England | 1351 |
| 322 | CF | Manolo Gabbiadini | 25 | Southampton | Italy | 2012 |

- **Correlation:**heatmap is used to check correlation between variables.




Plotting the heatmap of features and target (Market Value) reveals some interesting trends: Page_views, fpl_value, fpl_set, fpl_points and Big club seems to be correlated with the target variable - market_value of each player:

**Building models:** We have to implement Linear Regression, Lasso Regression, Ridge Regression, Nearest Neighbour Regression, Support Vector Regression, Tree Regression, Random Forest Regression and Gradient Boosted Regression.

As the range of variables vary so we had applied Min max scaling. One hot encoding was applied to 'nationality'. We dropped redundant and irrelevant variables such as name, age(as age_cat was correlated),club (club_id correlated), position(position_id correlated).We randomly split the entire data frame into 80% Training Set and 20% Test set, where we hold out the Test Set for final model evaluation. For Nearest neighbor algorithm, we have implemented a genetic algorithm to improve its performance for regression problem and came up with the optimum value of K (hyperparameter for KNN).

## Model Evaluation

There are 3 main metrics for model evaluation in regression:

*R Square/Adjusted R Square*,*Mean Square Error(MSE)/Root Mean Square Error(RMSE), Mean Absolute Error(MAE).*

| **Models** | **$R^2$ on train data** | **$R^2$ on test data** | **RMSE** |
|---|---|---|---|
| Linear | .79 | | |
| Lasso | .80 | .60 | |
| Ridge | .80 | .74 | |
| SVM | .77 | .79 | 6.03 |
| NN | .89 | .69 | 7.4 |
| **Decision Tree** | **.89** | **.75** | **7.3** |
| **Random Forest** | **.89** | **.90** | **4** |
| **Gradient Boost** | **.93** | **.75** | **6.6** |

We can clearly check that Random Forest(RF),Decision Tree and gradient Boost are performing better than others so we will perform hyperparametric tuning for these.

**Tune the hyperparameters and build the most accurate model:** *Hyperparameters*, are parameters that cannot be directly learned from the regular training process. They are usually fixed before the actual training process begins. Two best strategies for Hyperparameter tuning are:Grid SearchCv and RandomizedSearchCV.

After hyperparametric tuning, RF turns out to be a winner. So we deployed our model using RF.