Sabudh Foundation

Predicting the Price of a Football Player (Group Coursework)

Team H

Efforts By:

Ridhima Airi

Uday Mahajan

The Dataset contains the following information:

- Name: Name of the player
- Club: Club of the player
- Age: Age of the player
- Position: The usual position on the pitch
- Position category:
- 1 for attackers
- 2 for midfielders
- 3 for defenders
- 4 for goalkeepers

- Market value: As on transfermrkt.com on July 20th, 2017
- Page views: Average daily Wikipedia page views from September 1, 2016 to May 1, 2017
- fpl_value: Value in Fantasy Premier League as on July 20th, 2017
- fpl_sel: % of FPL players who have selected that player in their team
- fpl_points: FPL points accumulated over the previous season

- Region:
- 1 for England
- 2 for EU
- 3 for Americas
- 4 for Rest of World

- Nationality
- New foreign: Whether a new signing from a different league, for 2017/18 (till 20th July)
- Age category
- Cuboid
- Big club: Whether one of the Top 6 clubs
- New signing: Whether a new signing for 2017/18 (till 20th July)

Findings:

Size

(Original): 471 X 17, i.e., 471 players and their information on 17 features.

(After removing null values): 470 X 17.

The data includes the players from following clubs:

- Arsenal
- Bournemouth
- Brighton and Hove
- Burnley
- Chelsea
- Crystal Palace
- Everton
- Huddersfield
- Leicester City
- Liverpool
- Manchester City
- Manchester United
- Newcastle United
- Southampton
- Stoke City
- Swansea
- Tottenham
- Watford
- West Brom
- West Ham

We check the correlation between different attributes.

Results:

For the output variable ‘market_value’, correlation is maximum with the features ‘fpl_value’ and ‘page_views’ followed by ‘fpl_points’.

Using the describe() command, we can observe that:

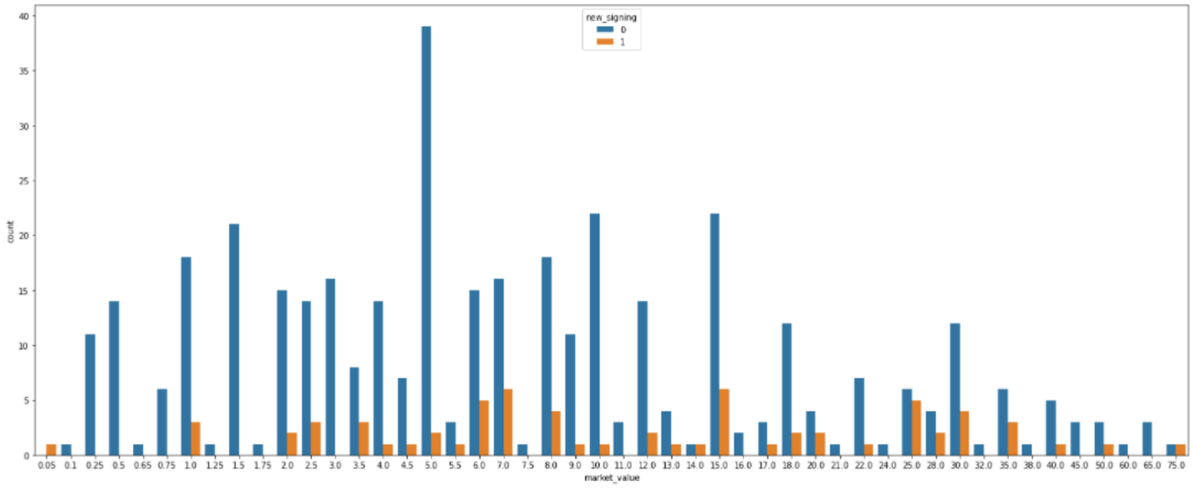
Mean (Age) = 26 years ranging between 17 years to 38 years.

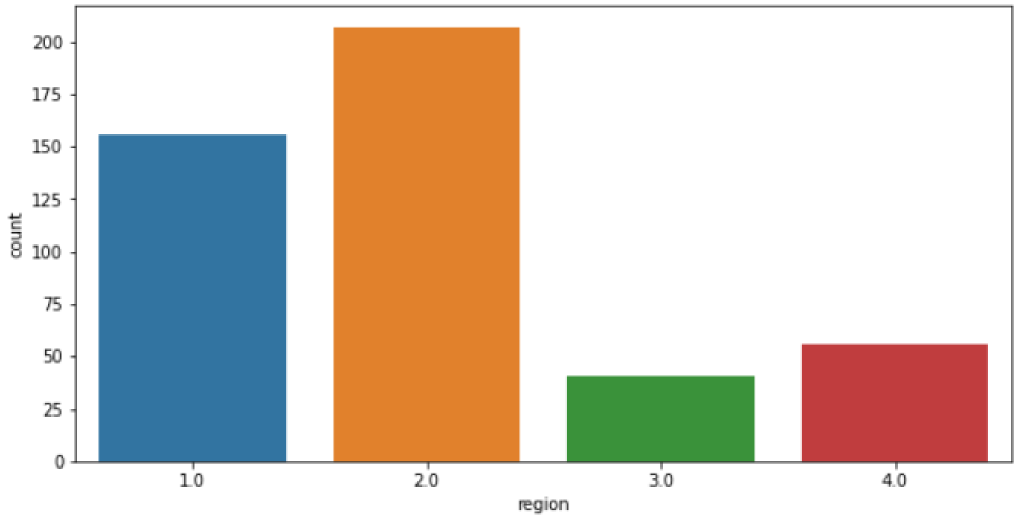
Mean (Market Value) = 11 ranging between 0.05 and 75.

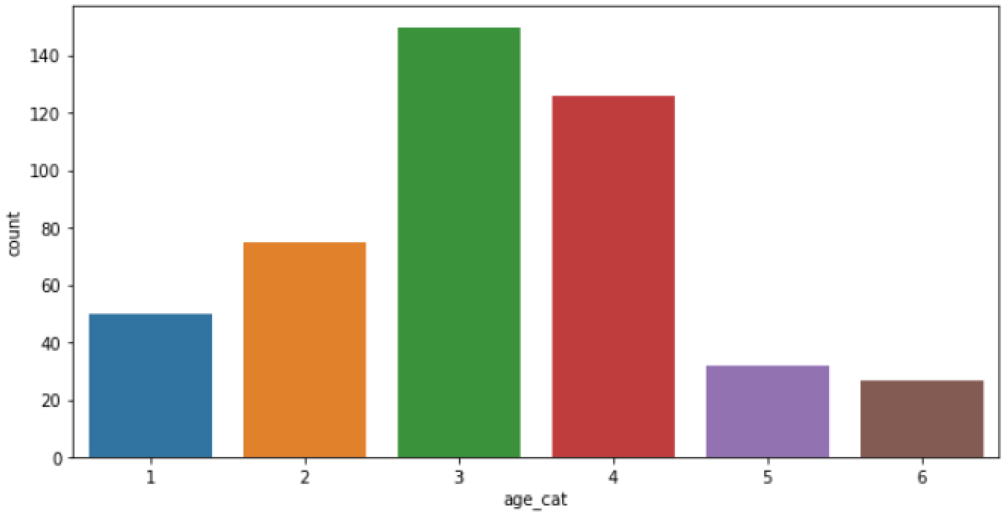
We define the input and output variables in order to train our model.

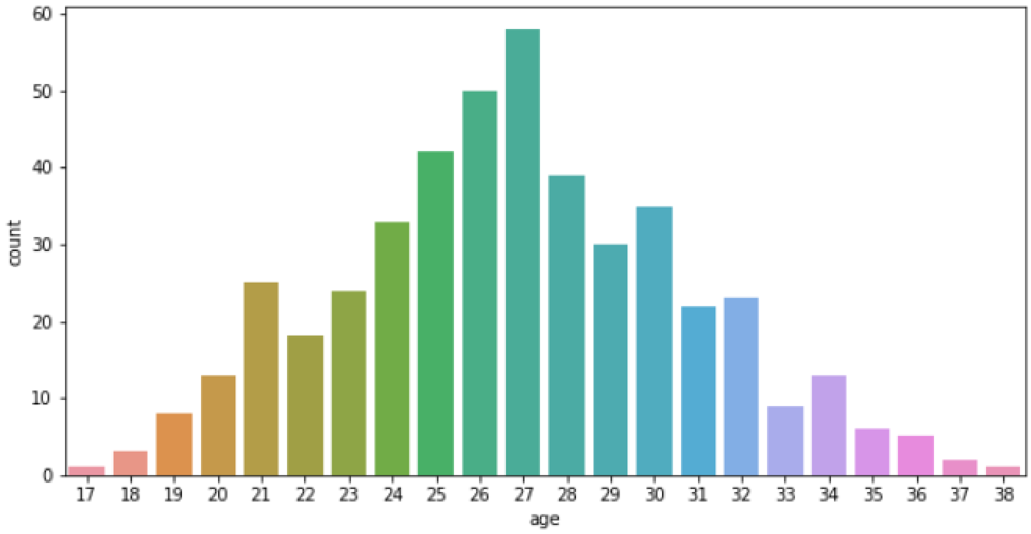
For input set, we drop the redundant features like “Name","Club","Age","Nationality","Position","Club_id,"fpl_sel” and "Market_value" as “Market_value" defines the output variable.

We use Seaborn to obtain countplot() to Show the counts of observations in each categorical bin using bars.



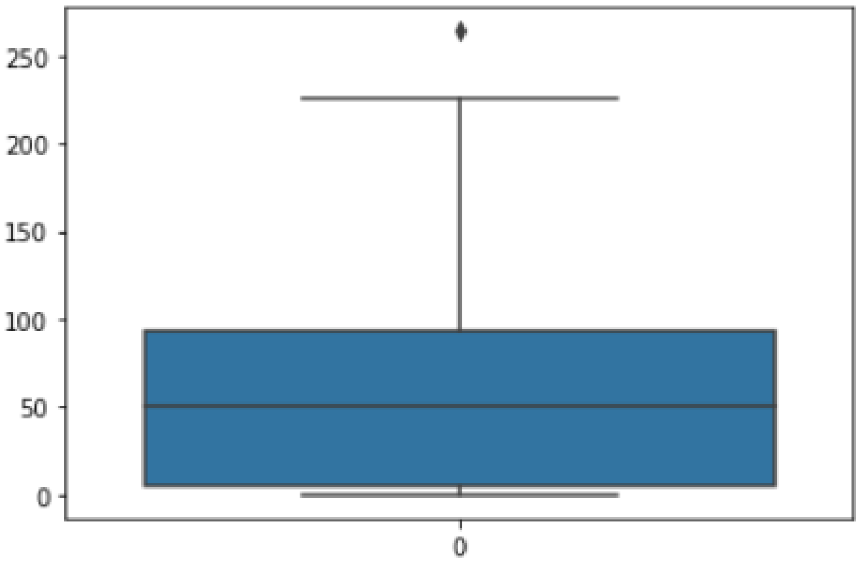




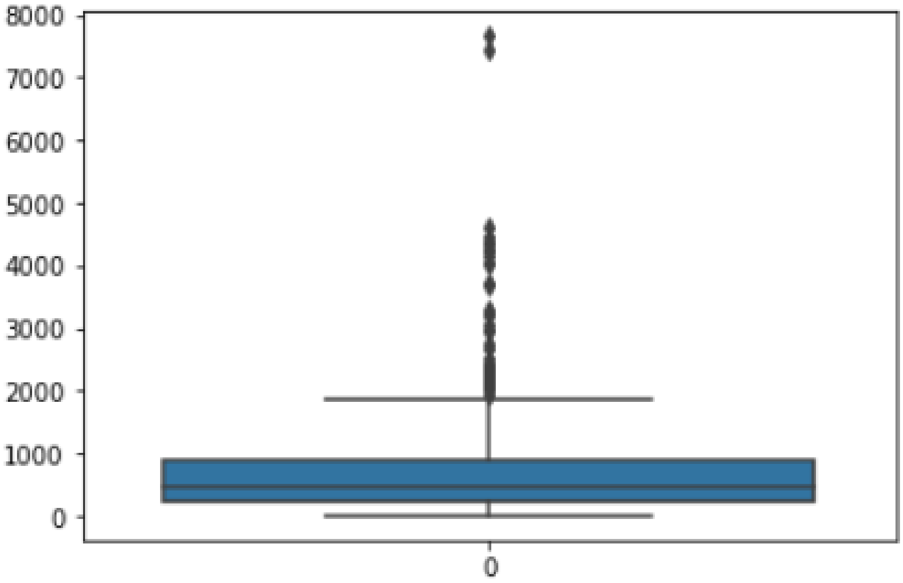


We plot Boxplots to obtain the 5-number summary: Minimum, Maximum, Median, first and third quartile.

- fpl_points



- Page views



Modelling:

We train the data using four algorithms:

Linear Regression

SVR

Random Forest

Extra Tree Regressor

We find that the Extra Tree Regressor gives the most accurate results.