Thanh Nguyen-Duong

DSC 550

Week 6

4/18/2020

Case Study Part 1 – Graph Analysis

Introduction

Soccer has always been my favorite sport. Been playing soccer when I was young made me realize that dreams can be pursuit if you really put your heart into it. However, the one thing I do not understand was many soccer athletes are paid enormous amounts compared to others. There are in fact many factors that go into decisions as to why certain athletes are given particular contracts and at what time in their careers they receive these opportunities. This dataset for the case study contains data about Premier League soccer players including statistics about their league history and their market value from 2017-2018 season. We will explore if there are any trends in player history, country of origin, and popularity in applying this data

Dataset

The data is for the 2017-2018 season of the Premier League. The dataset was sourced from Kaggle at the following link: https://www.kaggle.com/mauryashubham/english-premier-league-players-dataset

The variables in the dataset are as follows:

1) Name - Name of the player
2) Club - Club of the player
3) Age - Age of the player

4) Position - The usual position of the player
5) Position Category - Divided into four categories: Attackers, Midfielders, Defenders, Goalkeepers
6) Market Value - Value on transfermrkt.com on July 20th, 2017
7) Page Views - Average daily Wikipedia page views from September 1, 2016 to May 1, 2017
8) Fpl_value - Value in Fantasy Premier League as on July 20th, 2017
9) Fpl_sel - % of FPL players who have selected that player in their team
10) Fpl_points - FPL points accumulated over the previous season
11) Region - Categorized into four regions: England, EU, Americas, Rest of the World
12) Nationality - Nationality of the player
13) New_foreign - Binary. Whether a new signing from a different league, for 2017/18 (till 20th July)
14) Age_cat - ID number for age
15) Club_id - ID number for club
16) Big_club - Binary. Whether player is part of a Top 6 club.
17) New_signing - Binary. Whether a new signing for 2017/18 (till 20th July)

Here is a preview of the data:

In [36]:
```
#Step 3:  Look at the data
print(data.head(5))
```

```
                name     club  age position  position_cat  market_value  \
0     Alexis Sanchez  Arsenal   28       LW             1          65.0
1         Mesut Ozil  Arsenal   28       AM             1          50.0
2         Petr Cech  Arsenal   35       GK             4           7.0
3       Theo Walcott  Arsenal   28       RW             1          20.0
4  Laurent Koscielny  Arsenal   31       CB             3          22.0

   page_views  fpl_value fpl_sel  fpl_points  region     nationality  \
0        4329       12.0  17.10%         264     3.0           Chile
1        4395        9.5   5.60%         167     2.0         Germany
2        1529        5.5   5.90%         134     2.0  Czech Republic
3        2393        7.5   1.50%         122     1.0         England
4         912        6.0   0.70%         121     2.0          France

   new_foreign  age_cat  club_id  big_club  new_signing
0            0        4        1         1            0
1            0        4        1         1            0
2            0        6        1         1            0
3            0        4        1         1            0
4            0        4        1         1            0
```

Here are the types of variables in the data:

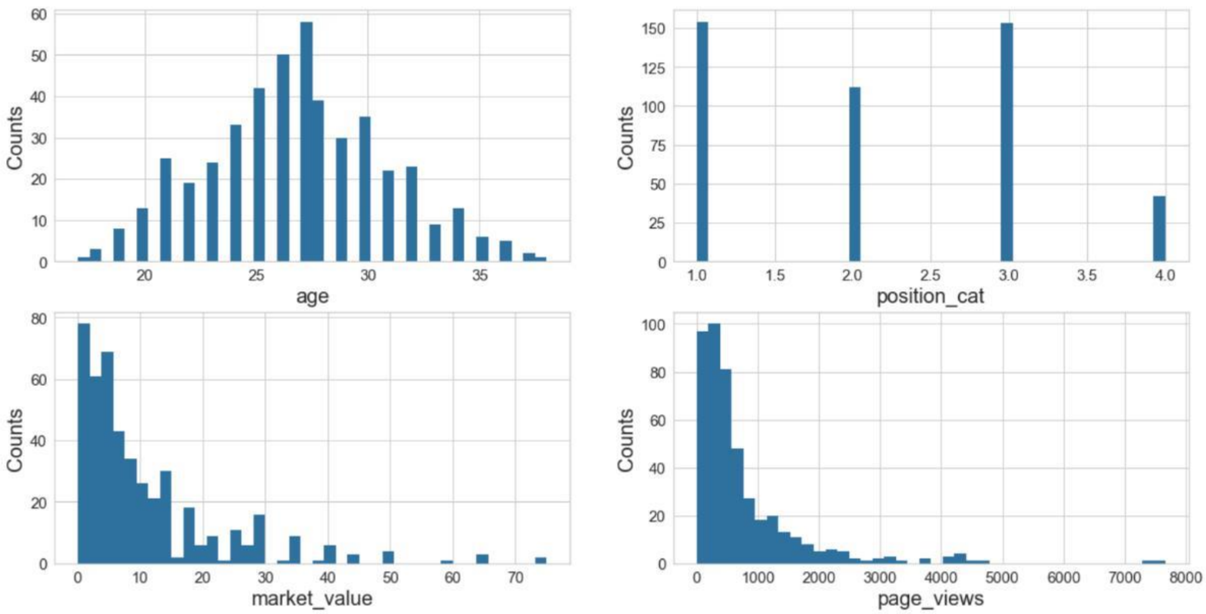
In [37]:
```
#Step 5:  what type of variables are in the table
print("Describe Data")
print(data.describe())
print("Summarized Data")
print(data.describe(include=['O']))
```

```
Describe Data
              age  position_cat  market_value   page_views   fpl_value  \
count  461.000000    461.000000    461.000000   461.000000  461.000000
mean    26.804772      2.180043     11.012039   763.776573    5.447939
std      3.961892      1.000061     12.257403   931.805757    1.346695
min     17.000000      1.000000      0.050000     3.000000    4.000000
25%     24.000000      1.000000      3.000000   220.000000    4.500000
50%     27.000000      2.000000      7.000000   460.000000    5.000000
75%     30.000000      3.000000     15.000000   896.000000    5.500000
max     38.000000      4.000000     75.000000  7664.000000   12.500000

       fpl_points      region  new_foreign     age_cat     club_id  \
count  461.000000  460.000000   461.000000  461.000000  461.000000
mean    57.314534    1.993478     0.034707    3.206074   10.334056
std     53.113811    0.957689     0.183236    1.279795    5.726475
min      0.000000    1.000000     0.000000    1.000000    1.000000
25%      5.000000    1.000000     0.000000    2.000000    6.000000
50%     51.000000    2.000000     0.000000    3.000000   10.000000
75%     94.000000    2.000000     0.000000    4.000000   15.000000
max    264.000000    4.000000     1.000000    6.000000   20.000000

         big_club  new_signing
count  461.000000   461.000000
mean     0.303688     0.145336
std      0.460349     0.352822
min      0.000000     0.000000
25%      0.000000     0.000000
50%      0.000000     0.000000
75%      1.000000     0.000000
max      1.000000     1.000000
Summarized Data
                 name     club position fpl_sel nationality
count             461      461      461     461         461
unique            461       20       13     113          61
top     Nemanja Matic  Arsenal       CB   0.10%     England
freq                1       28       85      64         156
```

Graph Analysis

First, I generated histograms of four variables to understand the spread of some of the variables. The histograms show the following initial insights:

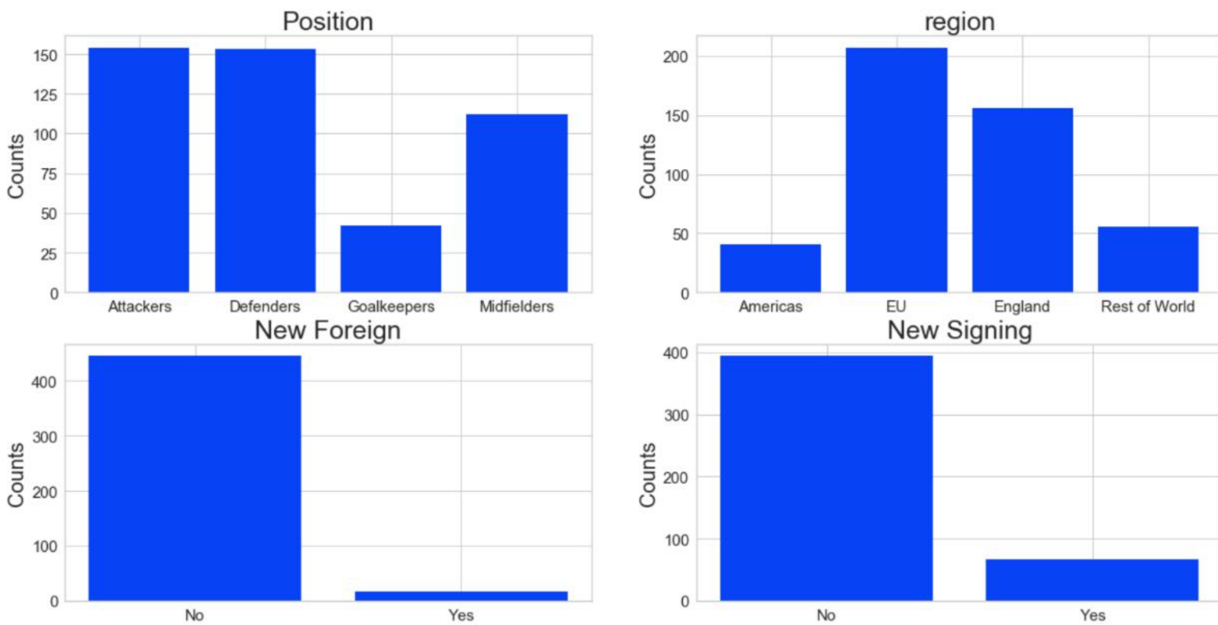
- Age - Normal distribution with an average age range of 26-28
- Position - Lowest count is for goalkeepers which makes sense since there is only one on the field per team per match
- Market Value - Most players are valued at 15 million or less
- Page Views - Most players receive 1,000 or less daily Wikipedia views



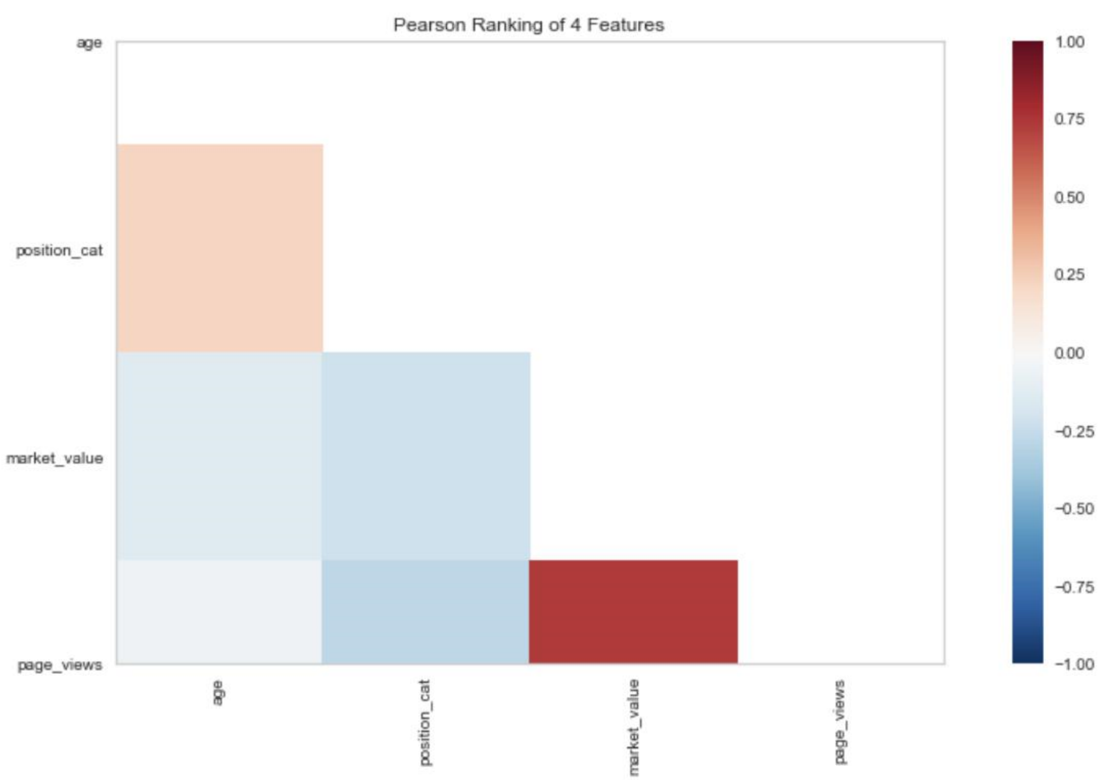
I explored four variables in bar charts to understand how the values compare. The following insights can be drawn from these bar charts:

- Position - Confirmed that goalkeepers are the least present in the dataset
- Region - Most players are from the EU

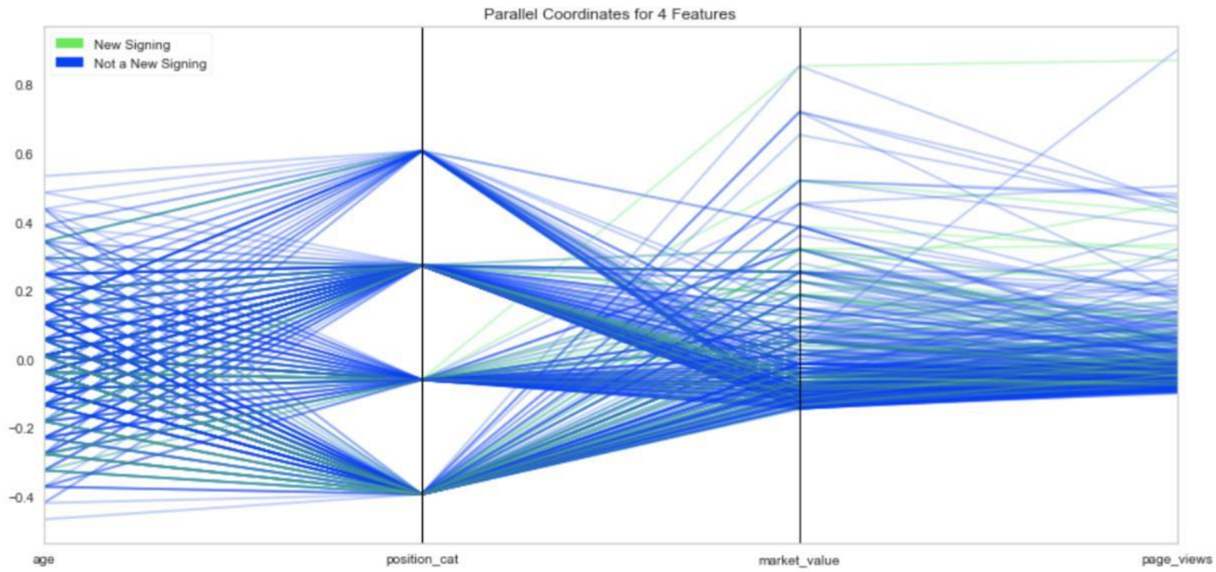
- New Foreign - Most players in the dataset are not new foreign players to the Premier League
- New Signing - Most players in the dataset are not new players to the Premier League



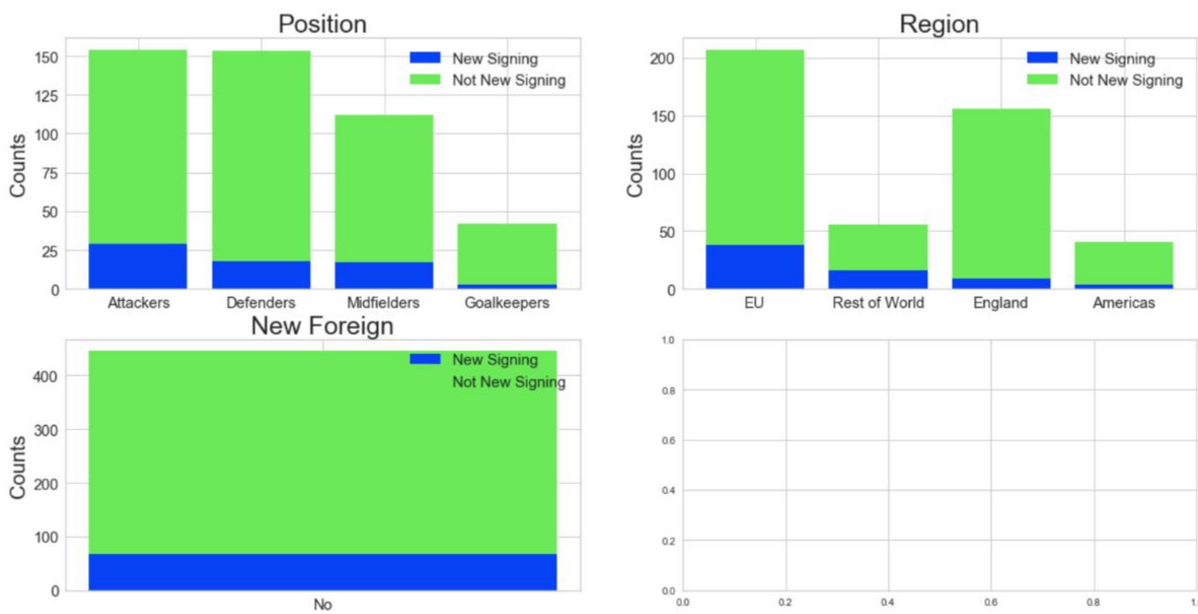
Pearson Ranking was done on the four variables I selected earlier. There appears to be a strong correlation between market value and page views signifying that popularity can be part of the value a player is seen as contributing to the team.



For the comparison part of this case study, I decided to perform analysis on the binary variable of whether the player was a new player to the league or not.



I then applied the New Signing variable to three additional variables for comparison. The most important insight is that there are no players in the dataset who are both new to the Premier League and a new Foreign player.



Case Study Part 2 – Dimensionality and Feature Reduction

Considering the dataset and my original question, the feature that made the most sense to predict was Market Value. Since the target vector is quantitative, I decided to use linear regression for my model.

The first step I took was to convert categorical data to numbers. I used One Hot Encoding on Position Category and Region. The resulting set of all features after this process are below.

	age	market_value	page_views	fpl_value	fpl_points	new_foreign	new_signing	position_cat_Attackers	position_cat_Defenders	position_cat_Goalkeepers	position_cat_Midfielders	region_Americas	region_EU	region_England	region_Rest of World
0	28	65.0	4329	12.0	264	0	0	1	0	0	0	1	0	0	0
1	28	50.0	4395	9.5	167	0	0	1	0	0	0	0	1	0	0
2	35	7.0	1529	5.5	134	0	0	0	0	1	0	0	1	0	0
3	28	20.0	2393	7.5	122	0	0	1	0	0	0	0	0	1	0
4	31	22.0	912	6.0	121	0	0	0	1	0	0	0	1	0	0
5	22	30.0	1675	6.0	119	0	0	0	1	0	0	0	1	0	0
6	30	22.0	2230	8.5	116	0	0	1	0	0	0	0	1	0	0
7	31	13.0	555	5.5	115	0	0	0	1	0	0	0	1	0	0

For my initial analysis, I wanted to include all Features available. I split the Features and Targets and then placed each row in its own array. The first five rows of each set are displayed below.

```
Features (First 5):
[[2.800e+01 4.329e+03 1.200e+01 2.640e+02 0.000e+00 0.000e+00 1.000e+00
  0.000e+00 0.000e+00 0.000e+00 1.000e+00 0.000e+00 0.000e+00 0.000e+00]
 [2.800e+01 4.395e+03 9.500e+00 1.670e+02 0.000e+00 0.000e+00 1.000e+00
  0.000e+00 0.000e+00 0.000e+00 0.000e+00 1.000e+00 0.000e+00 0.000e+00]
 [3.500e+01 1.529e+03 5.500e+00 1.340e+02 0.000e+00 0.000e+00 0.000e+00
  0.000e+00 1.000e+00 0.000e+00 0.000e+00 1.000e+00 0.000e+00 0.000e+00]
 [2.800e+01 2.393e+03 7.500e+00 1.220e+02 0.000e+00 0.000e+00 1.000e+00
  0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 1.000e+00 0.000e+00]
 [3.100e+01 9.120e+02 6.000e+00 1.210e+02 0.000e+00 0.000e+00 0.000e+00
  1.000e+00 0.000e+00 0.000e+00 0.000e+00 1.000e+00 0.000e+00 0.000e+00]]
Target (First 5):
[[65.]
 [50.]
 [ 7.]
 [20.]
 [22.]]
```

I then split each set into a test and training set with the test set being 30% of the data. Once that was complete, I created a scaler object that I fitted to the test and training set. Once complete, I ran both the L1 and L2 models with various strengths. I have included the results below.

```
Summary

L1
C: 10
Training accuracy: 0.34782608695652173
Test accuracy: 0.04316546762589928

C: 1
Training accuracy: 0.2453416149068323
Test accuracy: 0.02877697841726619

C: 0.1
Training accuracy: 0.13354037267080746
Test accuracy: 0.04316546762589928

C: 0.001
Training accuracy: 0.09316770186335403
Test accuracy: 0.02158273381294964
```

```
Summary

L2
C: 10
Training accuracy: 0.3198757763975155
Test accuracy: 0.04316546762589928

C: 1
Training accuracy: 0.2546583850931677
Test accuracy: 0.050359712230215826

C: 0.1
Training accuracy: 0.18944099378881987
Test accuracy: 0.04316546762589928

C: 0.001
Training accuracy: 0.11801242236024845
Test accuracy: 0.04316546762589928
```

The resulting Test scores in all cases are very close to zero so I made a couple of changes before running the model again. I increased the Training Set from 70% to 85% and reviewed individual variables.

After analyzing the statistical relevance of the individual features, it appeared that the features from all fantasy scores (variables Fpl_value, Fpl_sel, and Fpl_points) achieved the same results as each other. I decided to run the test again with these variables removed and compare the results.

```
Features (First 5):
[[   28 4329    0    0    1    0    0    0    1    0    0    0]
 [   28 4395    0    0    1    0    0    0    0    1    0    0]
 [   35 1529    0    0    0    0    1    0    0    1    0    0]
 [   28 2393    0    0    1    0    0    0    0    0    1    0]
 [   31  912    0    0    0    1    0    0    0    1    0    0]]
Target (First 5):
[[65.]
 [50.]
 [ 7.]
 [20.]
 [22.]]
```

Here are the results with the increased training dataset and the Fantasy League variables removed.

```
Summary

L1
C: 10
Training accuracy: 0.23529411764705882
Test accuracy: 0.05714285714285714

C: 1
Training accuracy: 0.17902813299232737
Test accuracy: 0.05714285714285714

C: 0.1
Training accuracy: 0.11508951406649616
Test accuracy: 0.07142857142857142

C: 0.001
Training accuracy: 0.10741687979539642
Test accuracy: 0.014285714285714285
```

```
Summary

L2
C: 10
Training accuracy: 0.22250639386189258
Test accuracy: 0.04285714285714286

C: 1
Training accuracy: 0.18925831202046037
Test accuracy: 0.02857142857142857

C: 0.1
Training accuracy: 0.16624040920716113
Test accuracy: 0.04285714285714286

C: 0.001
Training accuracy: 0.11764705882352941
Test accuracy: 0.05714285714285714
```

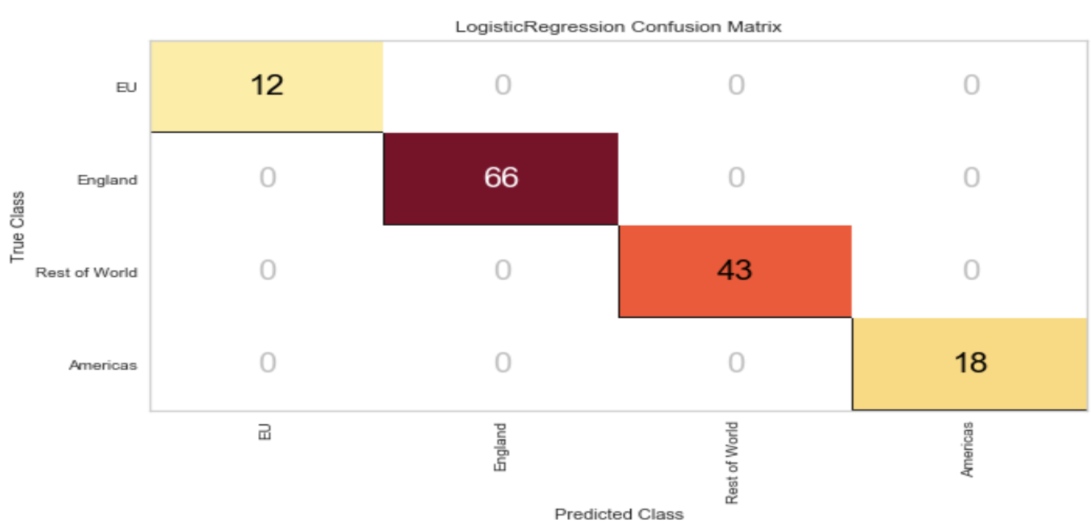
These changes did not improve the Test Accuracy of the model. Based on the analysis I have performed so far, it appears that this dataset does not include features that can accurately predict market value of a player.

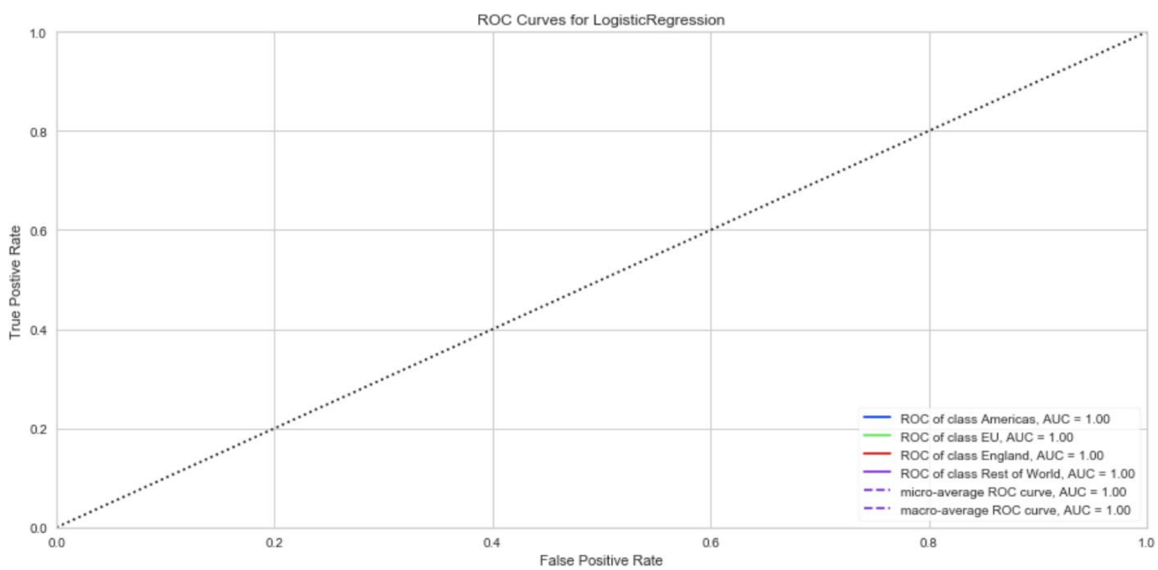
Model Evaluation and Selection - Part 3

I performed Model Evaluation to predict two features: position and region. For this case study, I am considering which features are the most aligned with all the features available to see if there are any trends with assessing market value of a player and these two each have four options which makes it most suitable for this week's task of selecting a supervised model.

For predicting the region feature, I started with the 70/30 split and the results presented a perfect accuracy.

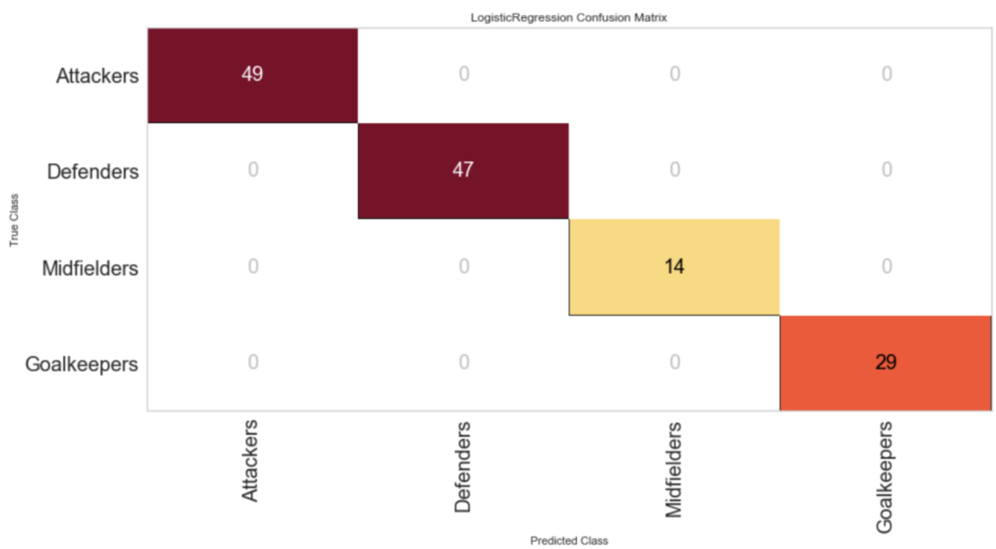
```
No. of samples in training set:  322
No. of samples in validation set: 139


No. of each region in the training set:
EU                 142
England            113
Rest of World       38
Americas            29
Name: region, dtype: int64


No. of each region in the validation set:
EU                 66
England            43
Rest of World      18
Americas           12
Name: region, dtype: int64
```

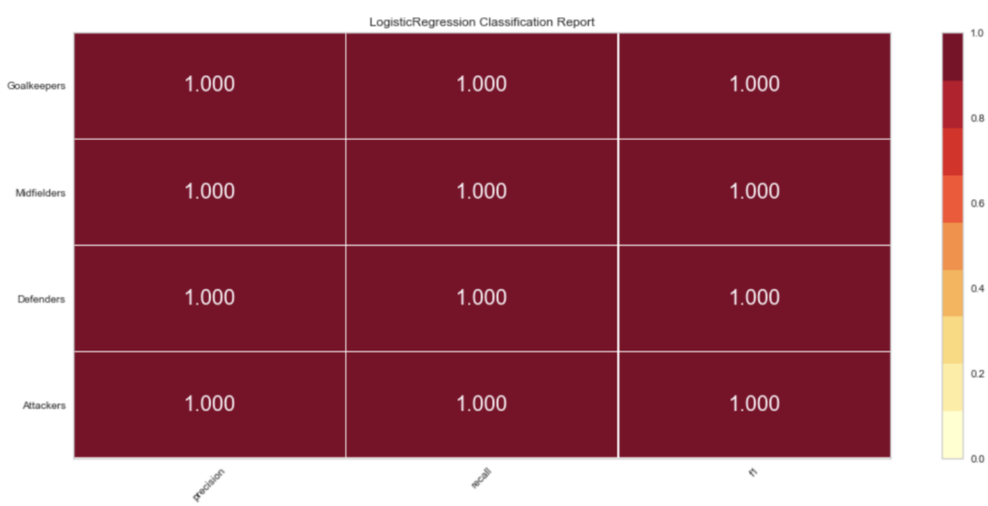


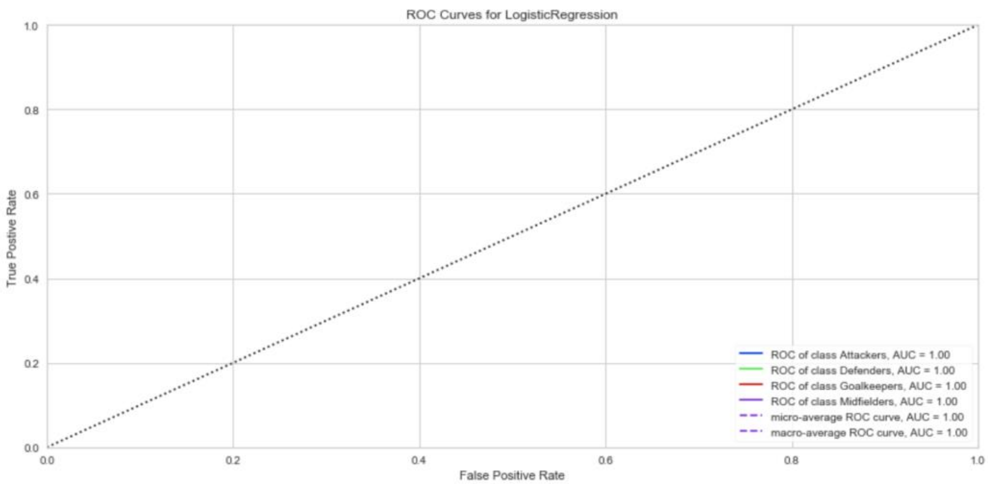


I tried the same ratios on the position features, and I got the same overall results.

```
No. of samples in training set:  322
No. of samples in validation set: 139


No. of each position in the training set:
Defenders       106
Attackers       105
Midfielders      83
Goalkeepers      28
Name: position_cat, dtype: int64


No. of each position in the validation set:
Attackers       49
Defenders       47
Midfielders     29
Goalkeepers     14
Name: position_cat, dtype: int64
```



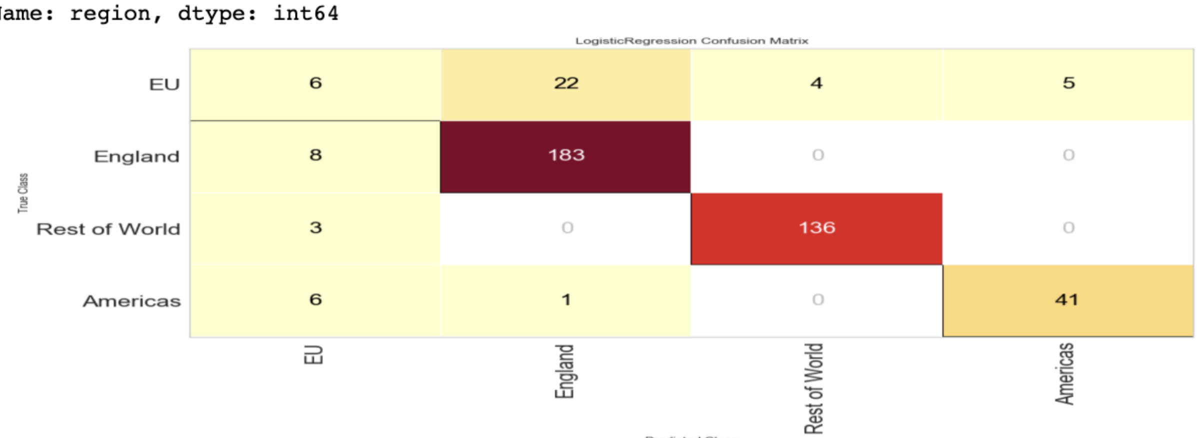


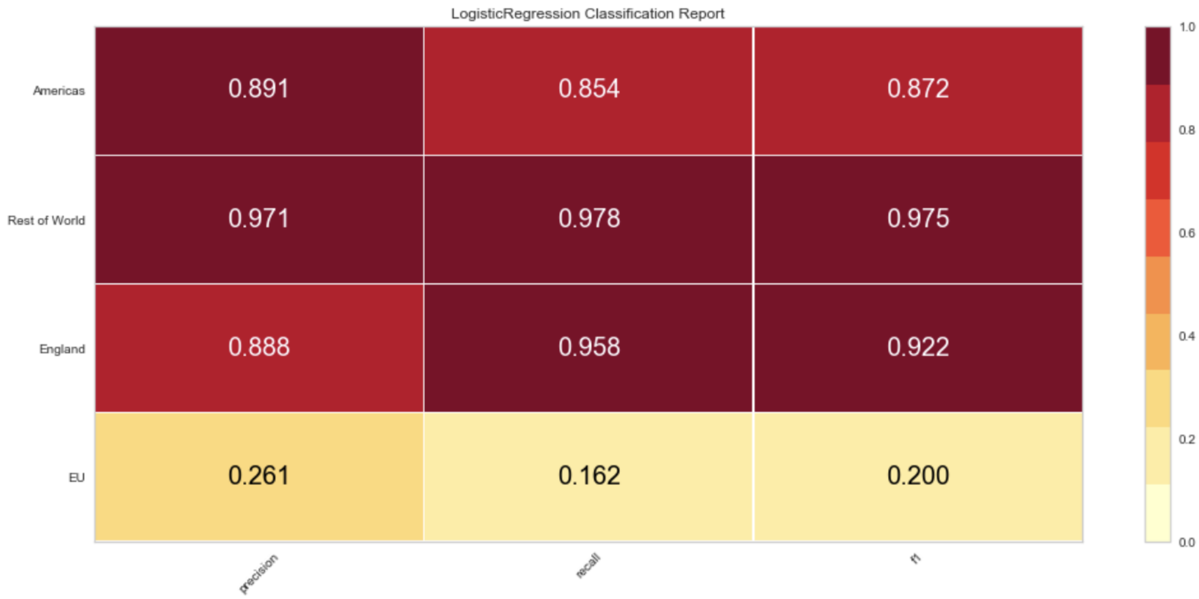


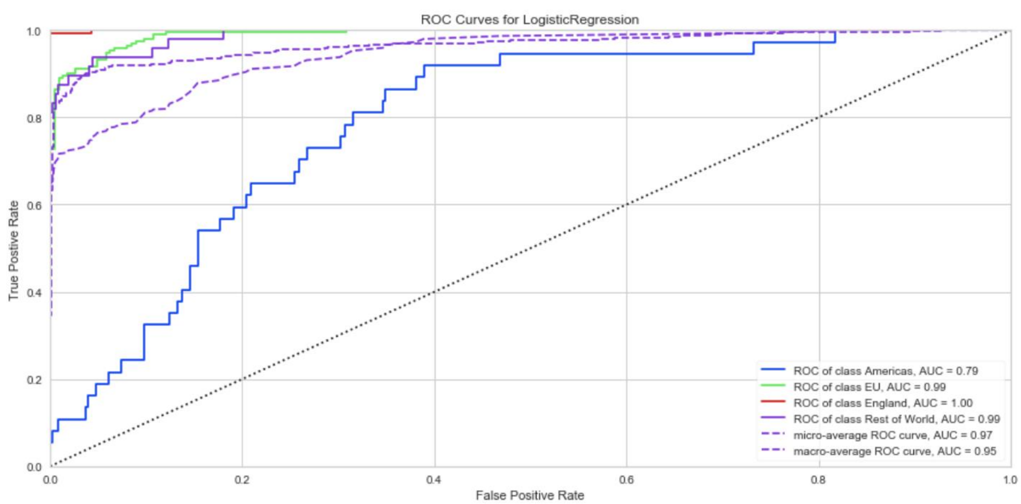
I wanted to test the validity of this scoring, so I dramatically reduced the training set to 10% with a 90% validation set and the scores did start to adjust but the accuracy was still significant in most categories.

Region adjustment to 90/10 for region.

```
No. of samples in training set:  46
No. of samples in validation set: 415


No. of each region in the training set:
England          17
EU               17
Rest of World     8
Americas          4
Name: region, dtype: int64


No. of each region in the validation set:
EU               191
England          139
Rest of World     48
Americas          37
Name: region, dtype: int64
```

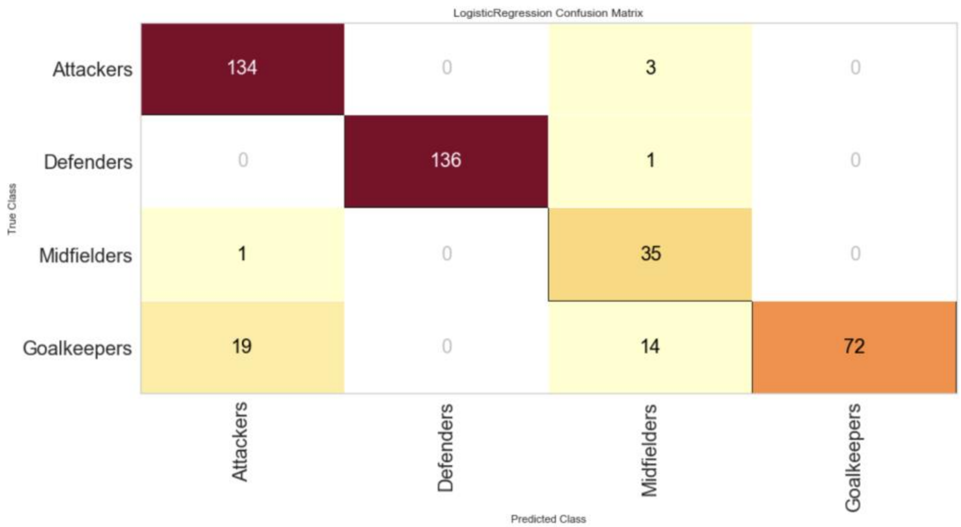


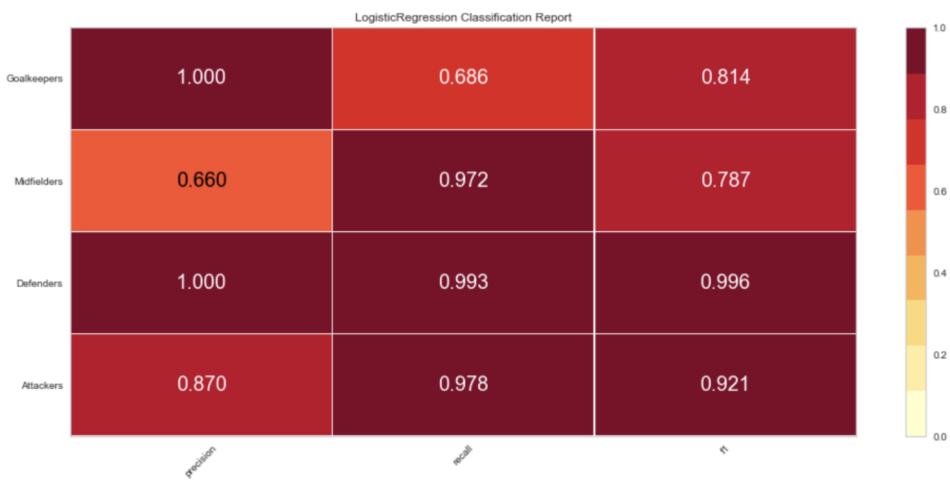


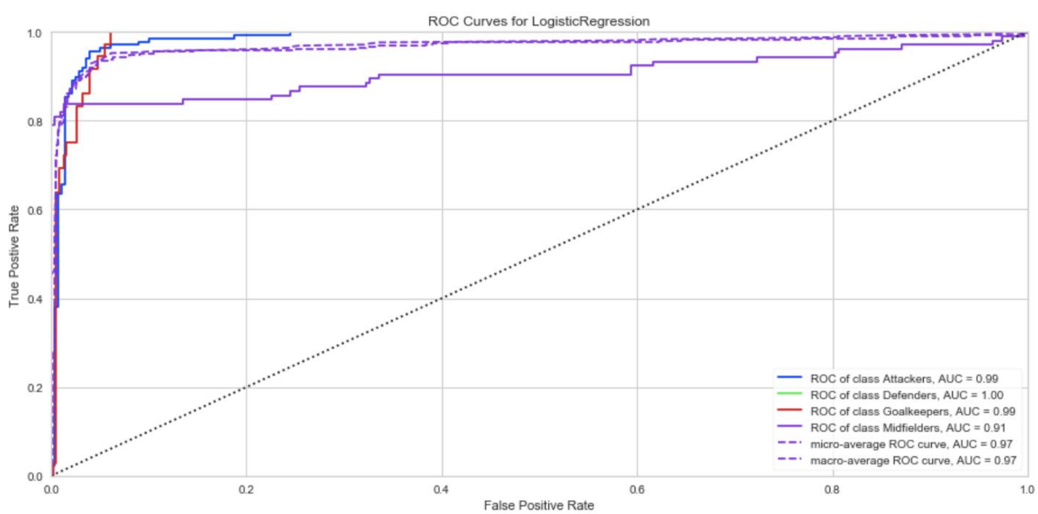


Position adjustment for 90/10 for position.

```
No. of samples in training set:  46
No. of samples in validation set: 415


No. of each position in the training set:
Attackers      17
Defenders      16
Midfielders     7
Goalkeepers     6
Name: position_cat, dtype: int64


No. of each position in the validation set:
Attackers      137
Defenders      137
Midfielders    105
Goalkeepers     36
Name: position_cat, dtype: int64
```


LogisticRegression Confusion Matrix
True Class
Predicted Class
Attackers
Defenders
Midfielders
Goalkeepers
134
0
3
0
0
136
1
0
1
0
35
0
19
0
14
72


LogisticRegression Classification Report
Goalkeepers
Midfielders
Defenders
Attackers
precision
recall
f1
1.000
0.686
0.814
0.660
0.972
0.787
1.000
0.993
0.996
0.870
0.978
0.921


ROC Curves for LogisticRegression
True Positive Rate
False Positive Rate
ROC of class Attackers, AUC = 0.99
ROC of class Defenders, AUC = 1.00
ROC of class Goalkeepers, AUC = 0.99
ROC of class Midfielders, AUC = 0.91
micro-average ROC curve, AUC = 0.97
macro-average ROC curve, AUC = 0.97

Conclusion

The original question for this project was to see if there were any trends for the market value of a player based on their experience in the Premier League, country of origin, position, and popularity. In Section 2, I was unable to show a collection of variables that could accurately predict the market value. However, when trying to predict position or region (where market value was an included variable), it was possible to create a model with a high accuracy. These two sections teach me that there may still be a way to predict the market value of a player with some different approaches. For example, if a heavier weight is placed on variables connected to popularity (wikipedia page views, presence in a big club), we may be able to improve the accuracy of predicting the market value of a player.