Fantasy Premier League Project Report Data Science Class

-Prathmesh Mengane

08

**Fall**

* **Problem statement and hypothesis**

I’m developing a Linear Regression Model to predict the total Fantasy Points scored by a player in the Fantasy Premier League Soccer (FPL) each week.  
 FPL is an online fantasy game for English Premier Soccer League. In this league a user selects a total of 15 players (11 starters + 4 substitutions) from the given pool of 624 players. Each Player has been assigned a value based on his current form and each user has been given a budget of £100m to select the 15 players in his team.

More Information: <http://fantasy.premierleague.com/rules/>

* **Description of your data set and how it was obtained**

I am extracting the data for each player from the official API of the FPL site (<http://fantasy.premierleague.com/web/api/elements/266/>).   
The API has data in the JSON format. I used the *requests* method learned in the class to grab data from the API and then feed it into a Dictionary.

* **Description of any pre-processing steps you took**

I created *for loop* to extract data from the API for all the 624 players and saved it an empty list. I used a random sleep parameter, to avoid being blocked by the API. Then I wrote the contents of the List into a CSV file (dict\_output1.csv) using *csv.DictWriter*. Then I read the data from the CSV file into a Panda’s Data Frame.

I decided to narrow down the scope of the project by only focusing for the players playing in Forward positions instead of goalkeepers, defenders and midfielders. Hence, I filtered the data from the original CSV containing all the players using condition *players\_df.type\_name==’Forward’* .   
 I then feed that data into a new CSV (players\_updated.csv) containing only the forward players, which constituted to 107 players as compared to previous 624 players and it still has same 58 columns.

I feed that CSV into a new Pandas Data Frame, containing only the Forwards.

* **What you learned from exploring the data, including visualizations**

I’ve narrowed down my data set to only contain data for Forwards. When I am running the model I am seeing some inaccuracy in predictions, which I’ve explained further below.   
I’ve explored and eye balled the data in the CSV and using pandas data frame. I observe that some of the players don’t have any significant data related to them because they haven’t been playing most of the weeks. (These are termed as “fringe players” in soccer, the one’s are used as backup if the star players get injured and hence they don’t get to play much throughout the season.)   
I am wondering if they are acting as anomalies and hence making the model weak. I’ll try and remove this players based on “*total\_minutes*” played condition and see if that improves my model.

Other than that most of my data seems to be clean in nature. It doesn’t have any missing data or Na data for the features I’m concentrating on.

* **How you chose which features to use in your analysis**

I have total of 58 columns/features in my data set currently. I initially started by eliminating the obvious ones, which don’t provide any insight to my model/prediction like “*photo*”, “*team\_name*”, “*current\_fixture*” etc.

I then eliminated the variables which I know have direct relationship with the Y variable. Meaning the variables, which are used to calculate the “*event\_total*” score for a particular player. These variables are “*goals\_scored*”, “assists”.  
(Reference: Appendix B explaining how the event score is determined)

I tried plotting some scatter plots to observe any relationship between remaining features and the Y.

Using my domain knowledge, I started to create a model with 6 features.  
(Reference: Appendix A explaining those 6 features)

Currently, I have not narrowed down all the features I want and I am still trying different methods to find relevant features. I am going to try some techniques Kevin suggested like scikit-learn feature selection module, lasso and ExtraTreesClassifier to compute feature importance.

* **Details of your modeling process, including how you selected your models and validated them**

Because I’m predicting a continuous variable as my output, I decided to use Linear Regression Model for my project.

I initially created a linear regression model from the Forward players data-frame using 6 variables

forwards\_model = smf.ols(formula='event\_total ~ selected\_by +   
value\_form + value\_season + form + ea\_index + bps', data=forwards\_df).fit()

I then got an error regarding multicollinearity . Thus, I created a Scatter Matrix and a Correlation Matrix plot for the 6 variables.

I found that there was a high correlation between “ea\_index” and “bps” and hence I decided to drop bps. The new model without bps still gives me multicollinearity error. So at this stage, I’m contemplating whether to drop “bps” all together from the model or use interaction terms technique and combine ea\_index and bps.  
(Reference: Appendix C showing Correlation Matrix Plot)

I then applied the “train\_test\_split” module to divide my data into train and test set. But then I got ValueError when I try to run my model on the train set. After exploring this error, I realized it was due to the “train\_test\_split” module. Hence I decided to use a different method using numpy’s “randn” function to create my test and train set.

As of now, I ran the Predict method on my Test set. But the results of the prediction are not accurate and thus I have to refine my model.

* **Your challenges and successes**

My Biggest Challenge so far has been feature selection. I have large number of X variables/features to select from ie 58 features. I’ve narrowed down some of them based on my domain knowledge but I’m not sure if the model created using them is giving me what I want. The step of Data Science workflow on which I am currently stuck, is the Feature selection step. And other challenge is once I’ve the features then how to combine them in a model to create the most accurate model possible.

There have been few successes along the way, I am proud of learning how to get a large data set from a public API without being recognized as a bot and being denied access.   
And how to store the large data set into a dictionary and data frames and then to perform functions on top of that.

* **Possible extensions or business applications of your project**

If I can get this model working accurately for the Forwards, then I plan on extending the project in the future for the other player types (defenders, midfielders and goal-keepers). If I can predict for all the types then I predict an entire team of players expected to do well in next weeks round of matches.

Overall the Online Fantasy Sports market is a growing industry. It’s a

huge market especially In Europe and Asia for English Premier League Soccer. . With more than 2.5 million registered users for FPL, the fantasy league is a million dollar (if not billion) industry. This can be a good guide for users to pick their teams with the highest total score potential.

* **Conclusions and key learnings**Coming Soon …..

**Appendix A:**

Variable Names Explanation.

|  |  |
| --- | --- |
| Variable | Explanation |
| event\_total (Y variable) | Total points scored by that player in that week (event) |
| selected\_by | The % of users who has that player in that team |
| value\_form | Custom index denoting value of player compared to current form of the player |
| value\_season | Custom index term denoting value of the player compared to the form of the player based on all the previous rounds so far |
| form | Custom index term denoting the current form of the player |
| ea\_index | Index created by the EA org. and used for ranking/comparing players, and its also changes based on the current performances of the player |
| bps | Index denoting the ranking of players based on the bonus points scored by the players |

**Appendix B:**

Scoring Rules



**Appendix C:**

Correlation Matrix denoting correlation between ea\_index and bps.

