Proposal.md

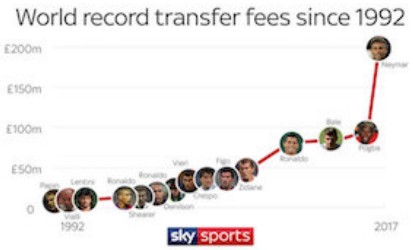
Machine Learning Engineer Nanodegree

Capstone Proposal

Sahil Goyal October 18th, 2017

Domain Background

Football (or *soccer*) is the biggest global sport and is a fast-growing multibillion dollar industry; with an estimate of 27 billion dollars in terms of annual revenue for the football clubs [1]. With more and more money pouring into the sport, the betting industry for predicting the outcome of matches is worth a billions pounds every year [2]. The *ridiculous* amount of money can be summed in the chart below, that shows the amount of money clubs have paid to secure a football player's services (*I being a mere mortal find myself counting the number of zeros in these huge sums of money*).



World record transfer fee over time

Prediction of soccer matches is a tough problem. Predicting the exact scoreline is a near impossible task. *On the field* often defies *on paper*. So much so, in the 2011-12 season of the *English Premier League (EPL)*, it was not known who the champions would be until the last five minutes of a nine month long season [3]. In the 2015-16 season of EPL, at the beginning of the season it was more likely that *Kim Kardashian* would become the US president, than the eventual team *Leicester* winning the title [4].

The game involves emotional factors such as the passion of the crowd, how pumped the players are that day, players' personal lives; factors which are beyond measure. In this project, I wish to tackle the problem using factors we can indeed measure.

Personal Motivation

I would divide my motivation to do this project into three reasons:

- ***My Interest and domain knowledge*** in this field. I follow European soccer with passion and interest. I would like to use my domain knowledge to come up with questions interesting to me, and help me solve those questions.
- ***Feature selection***, the biggest take away for me from this nanodegree course is the importance of feature selection, and finding correlations between features. The project on customer segments gave me a chance to apply some techniques such as *PCA* to reduce dimensionality; and I would like to apply them to this problem, in conjunction with my domain knoweldge.

- ***Visualization***, I am interested in the various ways we can visualize and create cool-looking visualizations, such as heat maps, layover over geo locations etc. I am hoping to use this project and come up with a few cool-looking visualizations [5]!

Problem Statement

I wish to find answers to a few of the following problems through this project:

- The *holy grail* is predicting the outcome of a football match, as mentioned in the kaggle link for the dataset I intend to use [6].
- Cluster matches based on the excitement level and interest in the game; possibly use this to model ticket prices. I would guess that a match between *Real Madrid* and *Barcelona* would be the most expensive. What factors make fans pay more for a match? If I follow a certain team, how much money can I expect to shell out?
- Model players and teams; and assign them a quantifiable number that serves as a rating of skill. I would like to find the answer to the *pull of star power*; many people tune in to football matches to watch *Messi*, *Neymar* or *Ronaldo*. Note that the capability of players also affects the excitement surrounding football matches.

Datasets and Inputs

The dataset I will be using for this task is the European soccer database over at Kaggle [6]. The database contains data from 25,000 matches; 10,000 players for the years 2008-2016, spread over 11 European countries. The matches are defined well in terms of data, with information about fouls, shots on target, possession etc. The dataset also contains betting odds from 11 different betting companies.; which will help me in evaluating my models.

The dataset contains attributes for both teams and matches, that I will be using to come up with a regression model to obtain a skill level rating.

The dataset is perfect for the questions I am trying to answer. Note that I intend to select one country's league first (probably the English Premier League because of familiarity), and then try to generalize it for different countries.

Solution Statement

Broadly speaking, I see my solution to be divided into four parts. Further details on my proposed soutions to each of the four parts is documented in the project design section.

- ***Data Preprocessing***: This dataset is spread across multiple files in the form of a SQLite database, with some anomalies mentioned in the description. After loading the data and cleaning it up for sure, I will apply techniques learnt in the course such as *feature scaling*, *one hot encoding* etc.
- ***Feature Selection***: Arguably the most important and exciting part for me. I would like to combine my domain knowledge and also use dimensionality reduction techniques to pick the best set of features.
- ***Model fitting and optimization***: After deciding on the features, fit various supervised learning models and compare performance; and tweak hyperparameters as needed to achieve the best performance.
- ***Prediction results and visualizations***: Visualize findings and predictions.

Benchmark Model

- ***Match predicition***: The kaggle page mentions that the best achieved thus far is 53%. I also have data from betting companies I can use to compare my predictions to.
- ***Player and team rating***: The dataset contains player and team ratings according to the models by the creators of *FIFA* video games. I will compare my regression models for team and player ratings to these ratings. Also, the top players in the world are widely known; when I cluster players based on the features I have selected, I would like to see that distinction between top players and average players.

- ***Cost and season tickets***: Although the dataset does not have data for this; this is obtainable through a variety of sources [7]. While it might be not be possible to compare the actual numbers, I certainly can compare a relative ranking in terms of prices.

Evaluation Metrics

- ***Match prediction***: The success rate of predicting matches correctly. Also I will be evaluating my performances vs the betting companies.
- ***Player and team ratings***: Comparison of my regression models with the ones by *FIFA*, with metrics such as *R2 score*.
- ***Cost of being a fan***: As stated in the section above, I would evaluate the difference between my financial modelling and the reality in a relative manner.

I have skimmed through a few existing kernels on the Kaggle page [8,9], that I might end up using to compare my performance. ***Note that if I do end up using any ideas/code from any kernels on the page, I will be sure to properly give credit where due.***

Project Design

I would like to be able to touch across all aspects of machine learning that I have learnt through this course.

Data Preprocessing

The dataset is in the form of a *sqlite* database. Analysis would require joins across multiple tables; wherein I will need to be careful to handle the cases of missing data/NaN values. As done in all of the Udacity projects, and I found it useful, I will be taking a few examples for each table and analysing the different types of variables at my disposal.

The dataset contains a mix of categorial variables (eg. a team's *buildUpPlaySpeedClass*), continuos variables (eg.: a player's *attacking_work_rate*) and boolean variables (eg. if a team is playing at home or not). This will be followed by preparing the data for machine learning, through techniques such as *one hot encoding*, *feature scaling* etc.

Feature Selection

This step is where I get to use my domain knowledge and dimensionality reduction techniques. I will be using the techniques I learnt in the customer segments project, to find correlated features; for eg. the *R2 score*. I will also analyse what to do with the outliers, if any. Note that sometimes a single player can win a match for their team, therefore I think that adding features such as *player-in-top_10*: the number of players in the team that are in the top 10 players of the world; *player-in-top-100*: number of players in top 100, would help in improving the accuracy. Often having a star player has more effect than just his skill, he can lift the players around him and bring the best in them. After selecting the features for players, I will select the features for matches that will help me best analyze the result.

I will measure the effectiveness of these new features by creating a pseudo-team of the best 11 players from the database, and see if this All-stars team beats everyone else most of the time.

To predict match excitement (from a neutral's perspective), I will select features such as *the number of shots on target*, *the number of red cards*, *the number of corners*, to name a few.

Model fitting and optimization

Once the features have been selected; I will be trying out various supervised learning models (*SVMs*, *Decision Trees*, *Ensemble methods* etc.) to fit the data. I also intend to give deep learning a try in this section. Once deciding on a good model, I will work towards hyperparameter tuning using *Grid Search Cross Validation*; and reiterating on the feature selection step above with any findings in this step.

Prediction results and visualizations

In this section, I plan to draw some interesting conclusions through visualizations:

- The best model(s) for the problems I have tried to solve.
- The evaluation of the accuracy of match predictions, and ratings regression model for the players.
- Visualizations to answer questions on the *most predictable league*, *most exciting matches*, *most expensive clubs to follow* etc.

I aim to use some of the techniques used in [5] to this end, teaching myself some python visualization techniques in the process.

References

[1] Call for Papers for a Special Issue in the Springer Journal

[2] Football betting - the global gambling industry worth billions

[3] City beat QPR in the last five minutes

[4] Things likelier than Leicester winning the league

[5] Seventeen Ways to Map Data in Kaggle kernels

[6] European Soccer Database

[7] Every club's cheapeat ticket

[8] Match outcome prediction in football

[9] EPL Weekly Predicting