**Introduction**

Hello fellow machine learning enthusiast,

This is a project that I undertook as a part of my internship.

Me and my supervisor were interested in a project pertaining to Machine Learning and we wanted it to be relevant in the current scenario.

We started this project around May and the Euro Cup 2016 was just around the corner.

So we decided to use machine learning to predict the goals scored in each match.

We had a little background with Machine Learning. But the result came out pretty good and I would like to share what we did, how we did it and why we did it.

**Data acquisition**

As any machine learning project goes, we needed data to work with. And a lot of it.

Machine learning algorithm basically tries to find a pattern in the provided data to come up with the predictions. Too less data, and you are at the risk of under fitting the model. Too much data, and you may generalize or over fit the model.

At this time of documentation, we haven’t debugged our model. So we are quite not sure whether we have an over fitted or under fitted model at our hand.

So coming back to data collection. We needed to collect historic data about the teams. Analyze the win, lose, tie ratio of each team. Calculate the attack strength, defense strength of each team based on their performance. Assign the FIFA points to them too. Finally once that was done, we had to clean up the data. Decide which teams were not important, decide on the features for the machine learning algorithm, normalize the features with a high variance.

So in order to perform all these tasks, we used our friendly neighborhood software, Microsoft Excel.

We copied every single football match ever played since January 2008 to June 2016. That was a lot of data. Precisely, 2502 rows in an excel sheet of data.

We used excel formulas to do some simple calculations and for some complex tasks, we resorted to visual basic coding.

So the final excel sheet had these following columns.

**Index**: An index number, just so that we didn’t get lost in the data.

**Date**: Date of the match

**Match Type**: The type of match. Can be a Friendly, World Cup, Euro Cup etc.

**Match Index**: A label starting from 1 was given to each type of match.

**Home Team**: Well, you know, the Home Team.

**Home Team Code**: The FIFA code for the team.

**Home Score**: The goals scored by the Home team.

**Away Score**: The goals scored by the Away team.

**Away Team Code**: The FIFA code for the team.

**Away Team**: Yes, exactly what you are thinking.

**Home Score Index, Away Score Index**: We gave labels for Win, Tie or Lose as 1, 0, -1 respectively.

**FIFA points home, FIFA points away**: The FIFA points for the home and the away team.

**FIFA points home normalized, FIFA points away normalized**: The FIFA points had to be normalized as they had a very high variance. For e.g. There were teams with FIFA points as low as 0 and as high as 1532. The algorithm may face trouble with such a high difference.

**Attack home, Attack away**: The attack strength of the Home and the Away team. Calculated based on the goals scored.

**Defense home, Defense away**: The Defense weakness of the Home and the Away team. Calculated based on the goals conceived.

**Recent Trend Home, Recent Trend Away**: The outcome of the last 5 matches played by the home and the away team.

**Encounter Trend**: The last 5 encounters of both the teams.

So this excel sheet acted as our database. Setting up the excel sheet was where we spent a lot of time.

**The Fun Part**

Now that we have the data to work with that is cleansed, normalized and baptized we searched for a good platform to start working with.

The first preference was using ipython notebook with scikit learn and pandas for data manipulation. We used a Raspberry pi at first, but the huge excel sheet was too much data for the poor chap. It was pretty slow and sometimes just gave up on me.

So I tried setting up the same environment on my mac and again ran into some trouble with python packages (matplotlib, yes I am talking about you).

So after a bit more research, we decided to use R software as it had a really good plotting and machine learning packages.

I watched a lot of tutorials in youtube and learned a bit about plotting the data, installing the required packages for machine learning.

I imported the excel sheet into R as a dataframe, removed the columns that are not relevant.

We used 2 algorithms to create 2 models.

The first one was a regression model and the second one was a classification model.

I used the built in lm() function to create a regression model.

For classification, I used the Random Forest model.

The features that we used to create the model were:

Normalized FIFA points of both the home and away team.

Attack strength and Defense strength of both the home and away team.

The 5 recent trends of the home and the away teams.

I created 2 models. One for predicting the Home score and the other one for predicting the away score.

Then I created a new excel sheet with the upcoming matches from the Euro cup, imported it into R as a dataframe.

I used the above model to predict the outcome of these matches.

The predictions were pretty good considering the fact that this was the first try and we never debugged the model.

For the Random Forest Model, out of the 45 matches played in the Euro Cup till date:

9 matches were predicted with the exact scores.

4 matches with the correct Difference.

7 matches with the correct Tendency.

For the Regression model, out of the 44 matches predicted in the Euro Cup till date:

8 matches were predicted with the exact scores.

3 matches with the correct Difference.

8 matches with the correct Tendency.

We have a bidding website inside our office where the colleagues compete against each other. So we thought it would be fun to see how good our machine learning algorithms would perform against Human intuition. So we enrolled the two algorithms into the website.

There were some days where the Classification algorithm was at the top ( out of the 27 Human players).

**Future Work**

We are planning to move onto different project that may make use of deep learning. But we would also like to improve our project. We believe that there is room for a lot of improvement and we are open to suggestions from fellow machine learning enthusiasts. We still have to plot and see how good our algorithm is doing and maybe try a whole new approach with a different set of features to predict the goals.