**DATA CLEANING**

As we analyze the data, we realize that for there are some examples like for a team, different spellings occur in the dataset. These differences may not be so important for the reader but it make difference in coding because we do not have team Ids.

For example, Manchester United is a very well-known team even among the people who are not fun of football. The name appears in the 4 different spelling in the data set as follows:

* manchester – united
* manchester united
* manchester utd

This leads us to lose the previous match analysis data that are important for our newly extracted features.

In order not to lose any information we firstly make the spelling correction in the team names.

**NEW FEATURE EXTRACTION**

In addition to the odd features extracted in the base code, match performance related features are extracted.

It is important to analyze the previous match results played between the home and away teams in predicting the future matches. Number of goals scored by each team in the previous matches, overall result summary of the previous matches, score that is assigned to each team for its performance in the recent matches are the features extracted. The detailed explanation is as follows.

1. Difference in GoalDifferences for each team matching

GoalDifferences is a measure that is considered as a performance measure in the most of the football leagues.The difference is calculated as GoalsForTheTeam-GoalsAgainstTheTeam.

As a feature we recommend to use the difference between the opponent teams’ GoalDifferences.

The features extracted for this topic are as follows:

* Total number of goals in the past 2 identical home-away matches
* Total number of goals in the past 5 identical home-away matches
* Total number of goals in the past 2 two-way home-away matches
* Total number of goals in the past 5 two-way home-away matches

In order to clarify the identical and two-way home-away matches we can give the following example.

For Arsenal – Wolves matching if Home = Arsenal and Away= Wolves identical matches are also Home = Arsenal and Away= Wolves matches. But the two-way home-away matches include matches for both Home = Arsenal and Away= Wolves or Home = Wolve and Away= Arsenal.

1. Match result summary of the previous matches

The most frequently winner team is extracted for the previous 2 and 5 matches for both two way and identical matchings. The features extracted for this topic are as follows:

* The most winning team in the last 2 identical matchings
* The most winning team in the last 5 identical matchings
* The most winning team in the last 2 two-way matchings
* The most winning team in the last 5 two-way matchings

1. Home/ Away Performance for the previous matches

Team performances may differ when the team is home or away. The mostly occurred match result for the previous matches that the team is in the same Home or Away position. The features extracted for this topic are as follows:

* The mostly occurred match result in the last 2 Home position matches
* The mostly occurred match result in the last 5 Home position matches
* The mostly occurred match result in the last 2 Away position matches
* The mostly occurred match result in the last 5 Away position matches

For each match the related feature (Home position feature for the Home team and Way position feature for the Away team) will be used.

1. FIFA Rating Differences

From the website … the FIFA Ratings at the beginning of the seasons are extracted. There are four different ratings for each team. You can see the rating from as follows: Attack, Defense, Midfield, Overall. We created 4 features by the following calculations:

* Attack Difference : Home Attack – Away Attack
* Midfield Difference : Home Midfield -Away Midfield
* Overall Difference : Home Overall -Away Overall
* Defense Difference : Home Defense -Away Defense

The detailed analysis can be seen from **makale makale**

1. ELO Ratings

From the website … the ELO Ratings are extracted for each team. Elo Ratings are used as an indicator of the performance for each team.

**DATA PREPERATION**

After feature extraction we end up with lots of features and most of them are correlated among each others. In order to prevent curse of dimensionality, PCA method is applied and we obtain 12 features that can explain almost %98 of the variance in the dataset.

The related graphs can be seen from below:

As you can see …..

Aa a result after feature extraction and dimension reduction, The selected features are as follows :

Colnames(bidid bidi)

**ALTERNATIVE MODELS**

After obtaining the feature set we tried lots of model that is applicable for the data set.

* Penalized Regression
* Random Forest
* Gradient Boosting method

are tried for the data set. Random forest gave better results. Although it has an important advantage that it can handle missing values, we preferred not use it because it was computationally inefficient. Our second choice was GLM that’s accuracy is very close to Random Forests’s.

The parameters for all the methods are extracted with 10fold cross validation.

**RESULTS**

After selecting our model, each round is simulated. The average RPS values can be seen from below