Exploration of the data

Finding sources of data

Kaggle

As the idea of this project came from the famous data science compentitions website Kaggle (kaggle.com) we started by looking the data provided there. The data there is collected and uploaded by the different teams. The most interesting for us files available there were:

- Teams.csv - contains all 365 teams in USA possible to play in the tournament. Each team has a 4 digit long ID number, which is used as a key throught all files.

- RegularSeasonCompactResults.csv - contains the results of previous year's tournaments (from 1985 until 2016). Which year, which team played with which team and their score and few more details.

- SampleSubmission.csv - the file containing the predictions for this year's tournament. This file contains the ID’s only of the 68 teams playing this season, with all possible matches between them (2 278) and the outcome prediction as a floating point number between 0 and 1 (showing the probablity that the first team wins over the second).

There are more files available containig detailed information about the teams and seasons for the past 13 years, but as it will be explained later on, we decided to not use them.

Kenpom

The data that we came across and decided to use to build our model upon comes from one of the best sites for advanced analysis of college basketball – Kenpom (kenpom.com). This website is build by Ken Pomeroy – an instructor of atmospheric sciences at the University of Utah. After working as a meteorologist for the U.S. government, he quit that job to focus full-time on his website (https://en.wikipedia.org/wiki/Ken\_Pomeroy#cite\_note-noaa-3). As a source of the ratings Ken uses The Sports Network, Inc. for the box score data and the play-by-play data is gathered from the NCAA (stats.ncaa.org). The accuracy of Pomeroy's rankings in predicting game outcomes has been noted by popular newspapers and blogs such as FiveThirtyEight (https://en.wikipedia.org/wiki/Ken\_Pomeroy#cite\_note-13), Mediaite (https://en.wikipedia.org/wiki/Ken\_Pomeroy#cite\_note-14), and The Wall Street Journal (https://en.wikipedia.org/wiki/Ken\_Pomeroy#cite\_note-15). Essentially the data provided by this website gives a snapshot of a team’s current level of play. The core of the system is the pythagorean calculation for expected winning percentage (http://kenpom.com/blog/ratings-explanation; https://en.wikipedia.org/wiki/Pythagorean\_expectation). The value of the exponent used since 2012 is 10.25.

Final choice of data

By using this already verified, calculated and proved rating data as a training data for the model we can prove that there is a relation between the rating of a team and it’s actual win or loss in each season.

The data from Kenpom shows the rating of a team for each season. The kaggle’s data give us information for each season which team won durring the tournament and which lost. Combining these two together and using this a a training data for the model will allow us to create a more precise prediction for the outcomes of the current season.

Exploration

Kaggle

The exploration started by getting to know the data from kaggle using Jupyter Notebook (Python 2.7 - Anaconda) and the libraries numpy and pandas we’ve loaded the CSV files into data frames which were used later on in the manipulations.

As simple preview of the data we printed the first and last few rows using the head and tail commands in almost every scipt. Another usefull command was shape, which showed us the number of rows and columns in the dataframe.

To see the teams that participated most in the tournament durring the years we used a group by and counted how many occurences a team had in the RegularSeasonCompactResults file. After that using the join command with the Teams file we joined the team names and visualised the top 20 in descending order of their occurances.

Using the SampleSubmission file and the Teams file we managed to create a list of the 68 teams playing this season (merge command – inner join) with their Ids and Names (CODE: DA - NCAA - Get Teams 2017.ipynb).

Kenpom

After the extraction of the data from the website and it’s merge with the kaggle data we started exploring the created dataset.

We’ve checked the distribution of the data (using histograms) and the relation of the features with the target. We used a free tool called Weka created for data analysis, exploration, visualisation and building of models. This exploration helped us to see if there are any outliers, missing data and anomalies in the data. The relation diagrams help us see if some of the features have more outstanding relation with the target.

(PUT THE GRAPHICS HERE from CODE: DA - NCAA - Feature Digrams A.ipynb)

As a result we didn’t notice any large outliers or any missing data. The distribution was in the expected normal distribution shape for most of the features.

The relation diagrams show some expected results. Mainly that the distribution is almost equal between the features and the target. It doesn’t matter if the offence or defence is high. Both have equal chance of winning or loosing the game. When comparing both offence and defence and plotting the win/loose with different color, we can see a small trend that the winning teams have higher offence and lower defence. The same with EM (marging) the higher the marging the higher the chance for win. On the other side the Luck also helps a bit, but the graphics which show the relation between Luck and AdjO and AdjD again confirm that higher offence and lower offence are more typical for a winning team.