Study on NCAA Basketball Tournaments

|  |  |  |
| --- | --- | --- |
| Chen Song SCI, University of Pittsburgh 135 N Bellefield Pittsburgh, PA chs222@pitt.edu | Junjia Guo SCI, University of Pittsburgh 135 N Bellefield Pittsburgh, PA jug44@pitt.edu | Yuxuan Li SCI, University of Pittsburgh 135 N Bellefield Pittsburgh, PA yul154@pitt.edu |

ABSTRACT

Every year, NACC Basketball tournament is one of the most popular game series and biggest sports gambling events. In this paper, we try to use the teams’ performance in regular seasons to predict their game results in the tournaments of the same season since we assume that the comprehensive team strengths should not change significantly in the same season. However, our work shows that this method only outperforms the baseline model (only using seed difference as predictor) to a small extent, which means that regular season performance may not be a good predictor of tournaments game results.

KEYWORDS

NCAA Basketball Tournaments, Machine Learning, Seed Difference

1 INTRODUCTION

Applying data driven method on sports area becomes more and more popular these recent years. All the technical statistics of games, teams and players forms a huge database and is waiting for people to explore. Basketball and football team managers now rely on machine learning algorithms to build their teams, prepare for future games and sell the tickets.

In all the excited games, NCAA Division I Men's Basketball Tournament, or so called “March Madness”, is one of the most concerned. This tournament was created in 1939 by the National Association of Basketball Coaches and currently featuring 68 college basketball teams from the Division I level of the National Collegiate Athletic Association. Among these 68 teams, 32 teams will be the champions from 32 Division I conferences (which receive automatic bids), and the other 36 teams will be awarded “at-large” berths which are chosen by an NCAA selection committee. The tournament will determine the national championship through a single-elimination competition system, which means that if a team loses once, the whole season ends for it. Therefore, every game will be very intensive and attractive. On the other side, this tournament is also the largest sports gambling event in the United States. “With over $3 billion wagered each year on the outcome of the tournament, bettors turn to expert rankings of teams for help with predictions.” (P. Kvam and J. S. Sokol, 2006)

For such a popular tournament with long history, as sports fans and data analysts, it is both possible and interesting for us to study on its data. In this project, we got the historical data of NCAA basketball games from Kaggle (including game results of both regular season and “March Madness” from 1985 to 2016), and aim to using machine learning algorithms to build a model and predict the result of a certain game in the tournaments.

2 RELATED WORKS

The 68 teams in NCAA Division I Men's Basketball Tournament will be "seeded", or ranked, within its region from 1 to 16 (totally 4 regions) by the selection committee. This rank could base on the teams’ performance in regular season and other information the committee may consider, and could represent the comprehensive strength of a team to a large extent. Therefore, some works used “seed difference” to predict the game results. This kind of models will predict that the team with stronger seed will always beat the team with weaker seed. For example, according to T. Smith and N. C. Schwertman (1999), only using simple linear models, seed difference could be a good predictor of margin of victory.

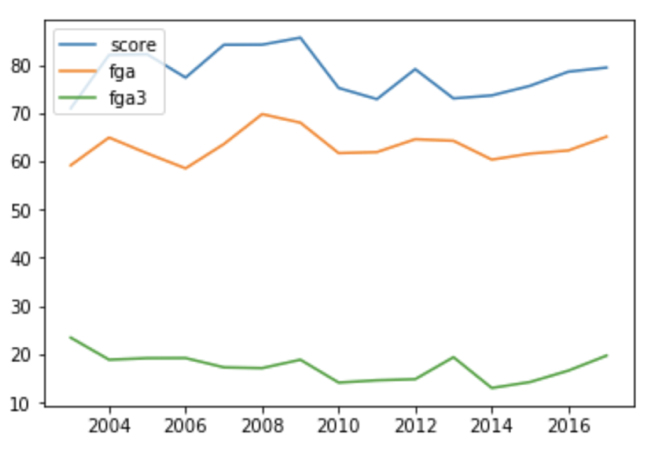
B. P. Carlin (1996) built an improved model to predict the probability of a given team could emerge as the four regional championship (“Final Four”) using external information. As we mentioned above, since this tournament is very popular in gaming industry, it already has some successful ranking systems to help for predicting game results. This paper used the point spreads available at the start of the tournament for the first-round games (Sagarin rating). The author indicated that the true point spreads are superior to the crude summaries provided by tournament seedings because it does not only concern the relative strengths of the teams, but also could include the consideration of new information like injury. However, the goal of our project is to build a better ranking system, so it could be weird to use existed ranking systems as features.

With the development of machine learning algorithm, more complex models were built to predict the NCAA basketball games. For example, P. Kvam and J. S. Sokol (2006) presented a combined logistic regression/Markov chain model for predicting the outcome of NCAA tournament games given only basic input data (daily online scoreboards) and the performance of the model is better than current used ranking systems. This paper gave good performance, but we also care about the evaluation of team strength as sports fans, which this model cannot provide.

3 DATA EXPLORATION AND PREPARATION

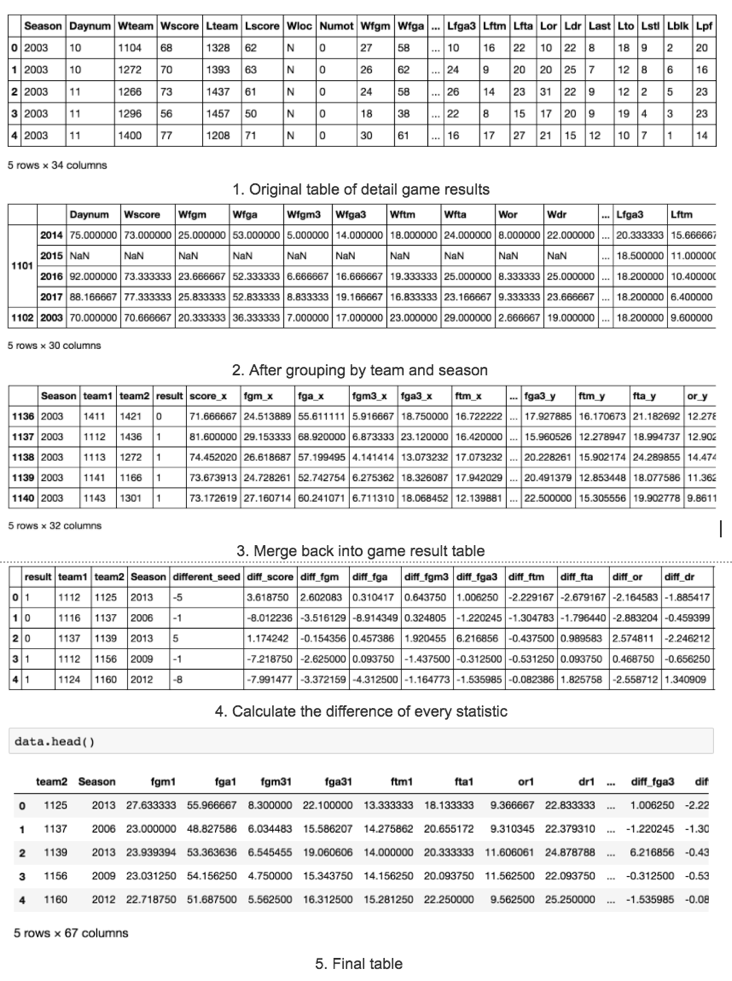
We got the historical data of regular season and tournaments games from the Kaggle competition "March Machine Learning Mania 2017". In the dataset, we have compact game results from 1985-2015, detail game results from 2003 to 2016 which have more technical details of each game, and seed information from 1985 to 2016.

We want to use the performance of each team in regular seasons and their differences as our features. Therefore, we need to aggregate the regular season game results to get the statistics like winning rate, average scores etc. for each team in every season. In these process, we can visualize a team’s performance change from a season to the next. For example, the figure below shows some regular season performance change of team 1314 (North Carolina, 2017 championship).



**Figure 1: Change of average score, fga and fga3 in regular season from 2003 to 2016, North Carolina.**

Then we merge the features of each team back to the tournament results table, get the difference of every statistic between the two teams in a game, and these will be the features we use in our model.



**Figure 2: process of data preparation.**

4 MODELS

Our basic idea to build the model is to use the performance of the teams in regular seasons to predict the game results of the tournament in the same season (Method 1). To compare with this method, we will also build a baseline model using only seed difference between the two teams in a game (Method 2), and using ELO ranking system to calculate the team strengths (Method 3).

Since we build a lot of features in Chapter 3 (67 columns), we will need to select the features carefully. With the three methods motioned above, we totally gave six rounds of tries:

* Round 1: Seed Data Only (baseline)
* Round 2: Seed Data, Time, Average Scores and Win-rate
* Round 3: All without Seed Data
* Round 4: All Features into the Pool
* Round 5: Delta values
* Round 6: ELO

4.1 Baseline Model: Seed Difference (Round 1)

Since this is a classification problem and has only one feature (seed difference) in the baseline model, we use several different classification algorithms like logistic regression, SVM and Adaboost. The logistic regression got the best performance: the accuracy is 0.716, the precision is 0.673, the recall is 0.731, and the log loss is 0.532.

4.2 Statistics of Regular Season Results (Round 2-5)

The basketball statistics can be divide into two classes, one is

basic statistics that include average score, rate of win and time of game each season, other one is advanced and detailed statistics data which more specific present the performance of each team. Firstly, we fit basic statistics into models, unfortunately, the performance of basic statistics wasn’t better than baseline.

At 3 and 4 rounds, we fit all features with or without seed into models, on one hand, we want to evaluate the impact of “seed” feature, on the other hand, just for thoroughness sake, lets add in all of the season-average basketball stats and see what comes out. Compare to without seed, all features in pool have a better performance, but that still worse than baseline.

Let’s try to cut the number of variables in half by using the ‘delta values’ for each game, representing the difference in average stats between the two teams. This will incorporate two variables’ worth of data in one variable.

Now we’re cooking. Practical accuracy is up to 72.7%, higher than our baseline, and we no longer receive the warning about overfitting. This seems like a good balance between quality and fit.

4.3 ELO Ranking System (Round 6)

ELO ranking system is a widely-accepted ranking system for the Go, Chess Game and some other sport games. It represents the performance of a team or player. The higher score the player or team has, the better the performance is. For each game between A and B, the probability of A winning the game can be calculated from the Formula 1. When the game ends, two ELO scores will be updated. The Formula 2 shows how to recalculate A’s ELO score. will be 1 if A wins, or 0.

**Formula 1: Calculation of A’s win rate from A’s and B’s ELO score.**

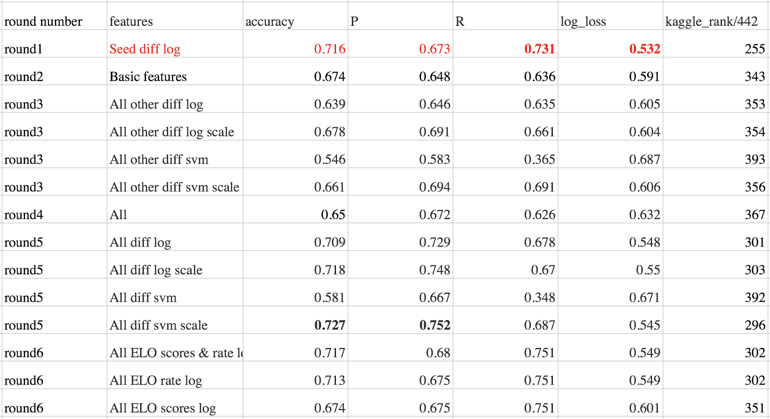
**Formula 2: Calculation of A’s new ELO score after the game between A and B.**

In this paper, we introduced ELO rating system into the prediction of NCAA Game results. We sorted all games, including regular games and tourney games, to update all teams’ ELO scores and store all scores each year into a matrix. Next, we combine the ELO scores and the win rate calculated from the scores into the existed features to become new features to establish the prediction model. It is expected to help improve the accuracy of prediction.

5 EVALUATION AND DISCUSSION

To evaluate these models, we fit the models with training data and predict with testing data. Then we calculate the prediction accuracy, precision, recall and log loss of every model. As the table below shows, the best model we got is using all differences with scaling and SVM classifier. If we focus on which team would win the game, this model outperformed the baseline model (only using seed difference and logistic regression) to a small extent. But if we consider about the probability prediction, which could be more important to determine the odds of each game in sports gambling, the baseline model has the best log loss.

**Table 1: Evaluations**



6 CONCLUSION AND FUTURE WORKS

After all the hard works, we sadly found that the best model we got is just almost as good as only using seed difference, which means that regular season performance is not a really good predictor for tournaments results. This could because that the tournaments are much more intensive than regular season games, or because each team will have different opponents in regular seasons and tournaments.

Therefore, for future work, we have two general plans: first, we can try to choose only intensive games in regular seasons to calculate all the features (i.e. score difference within a certain threshold). Second, we can try to use tournaments performance from the previous seasons to build the features. But if a team is going to the tournaments for the first time, it will have no data at all. We should try to solve this problem first.

ACKNOWLEDGMENTS

Yuxuan did the data preparation and main models. Chen finished the baseline model (seed difference) and drafted the report. Junjia finished the ELO part.

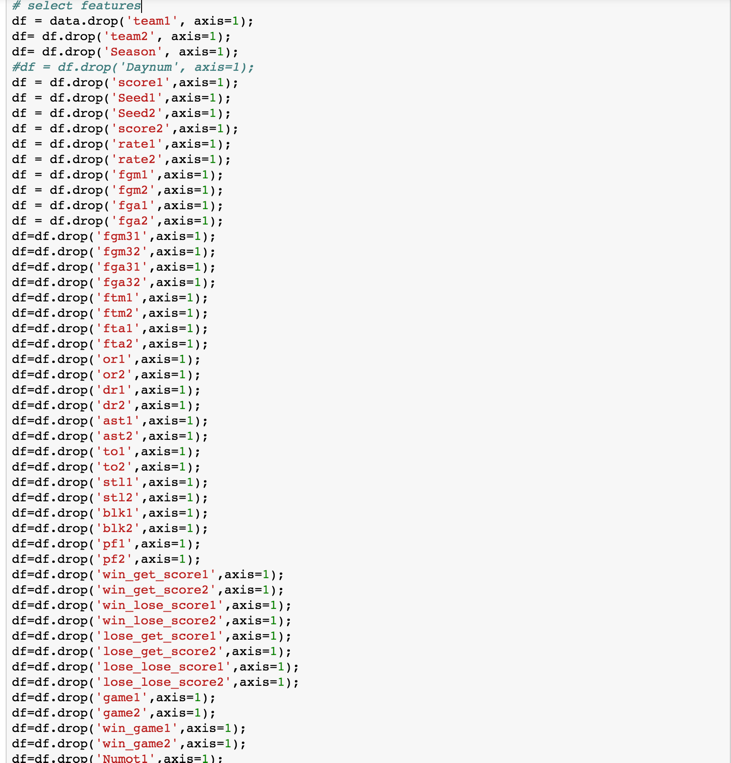
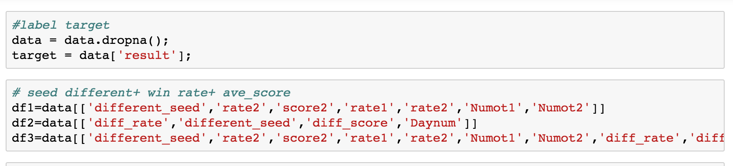
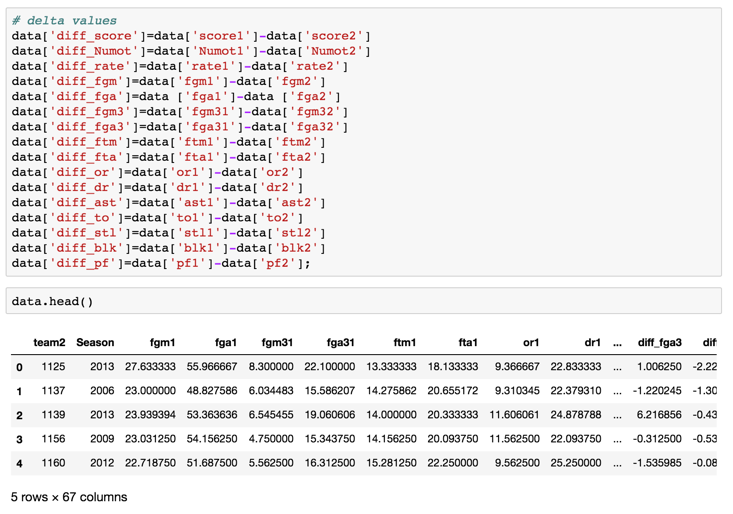
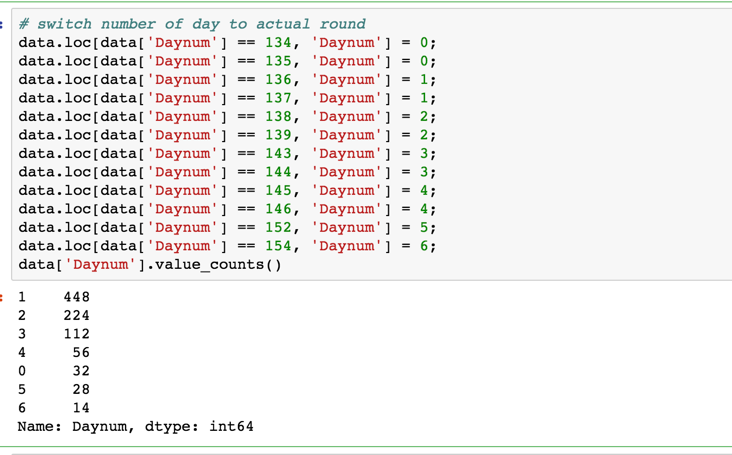
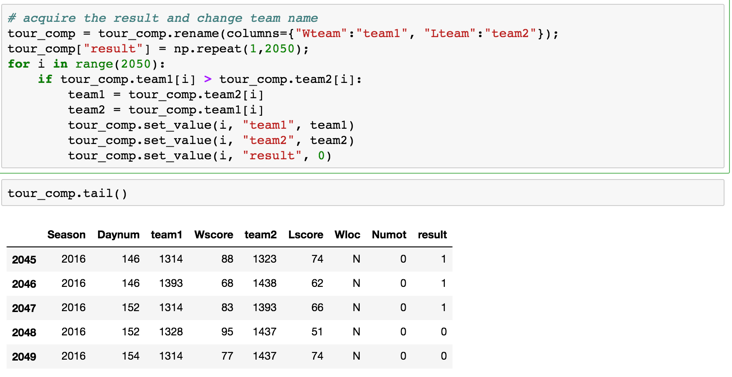
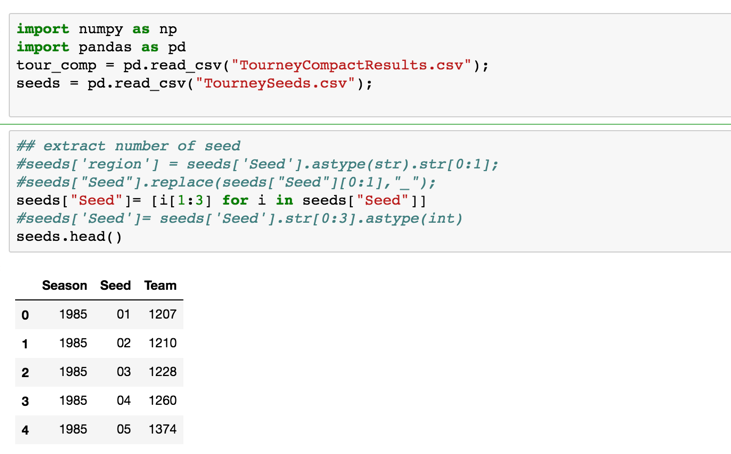
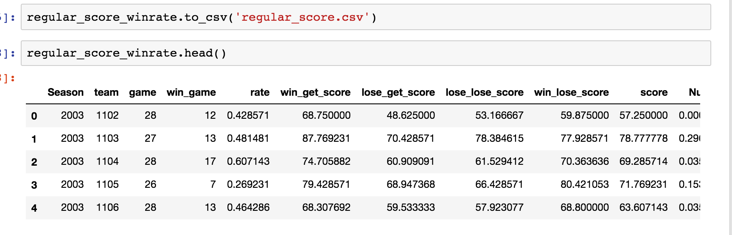
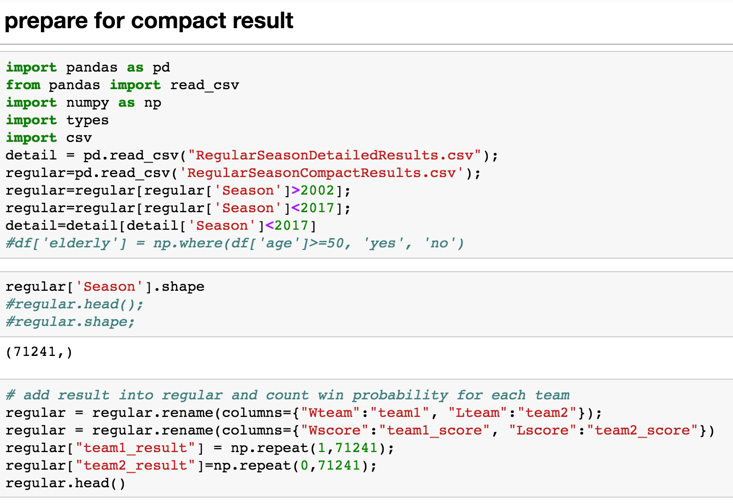
REFERENCES

|  |  |
| --- | --- |
| [1] | Wiki: NCAA Division I Men's Basketball Tournament. |
| [2] | T. Smith and N. C. Schwertman, "Can the NCAA Basketball Tournament Seeding be Used to Predict Margin of Victory?" *The American Statistician*, vol. 53, (2), pp. 94-98, 1999. |
| [3] | P. Kvam and J. S. Sokol, "A logistic regression/Markov chain model for NCAA basketball," *Naval Research Logistics (NRL),* vol. 53, (8), pp. 788-803, 2006. |
| [4] | B. P. Carlin, "Improved NCAA Basketball Tournament Modeling via Point Spread and Team Strength Information," *The American Statistician*, vol. 50, (1), pp. 39-43, 1996. |
| [5] | Wiki: ELO ranking system |

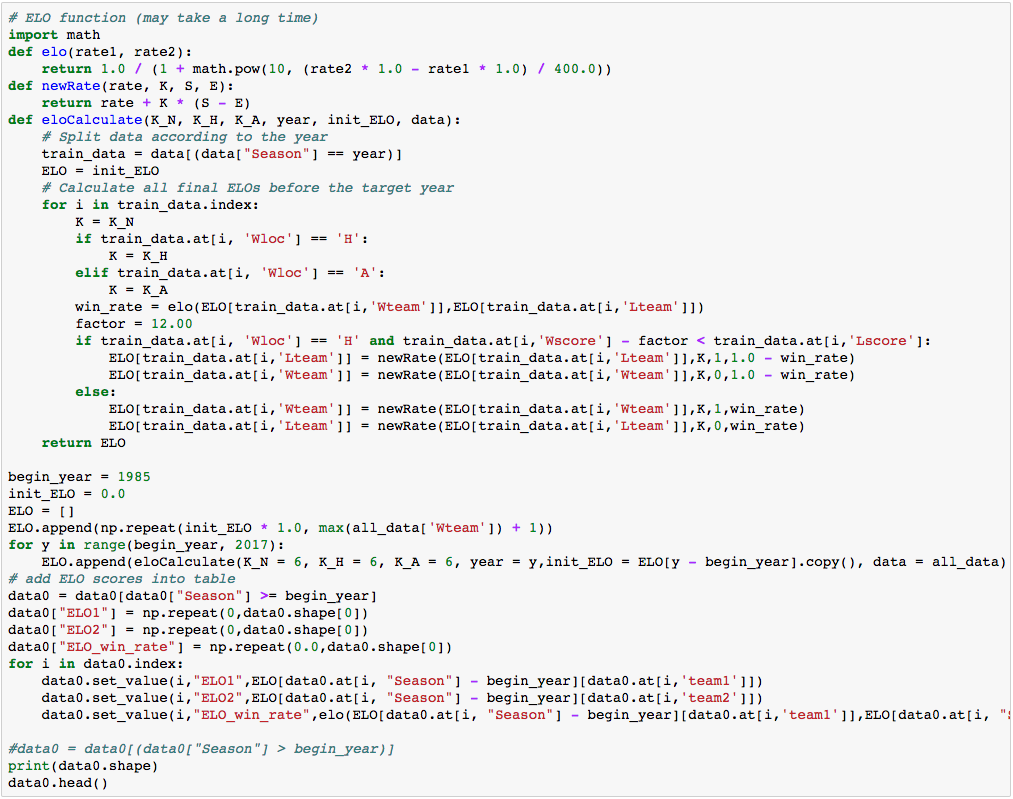
APPENDICES: Code of the Project

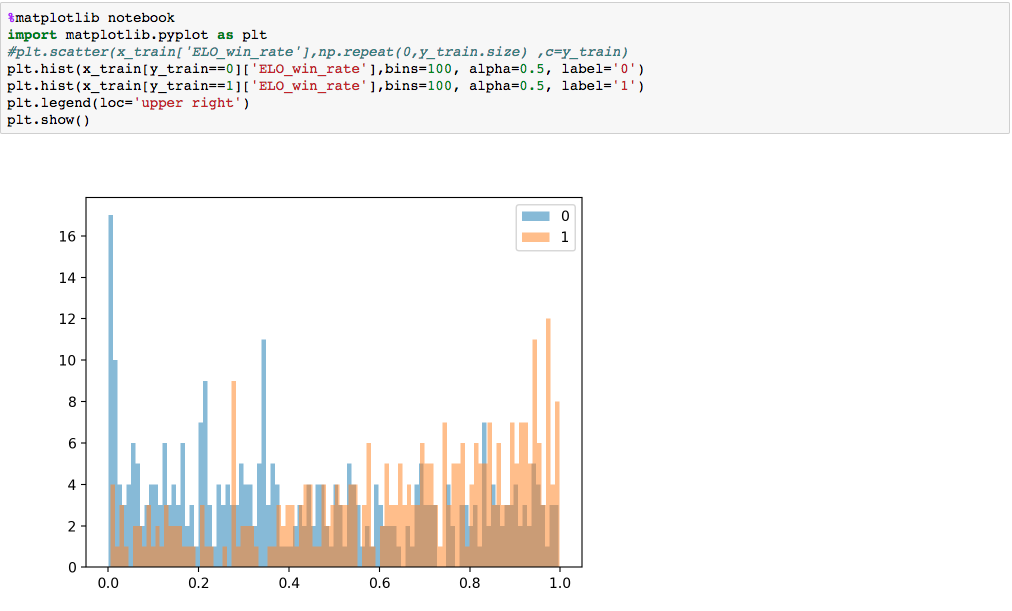
1.Data prepossessing

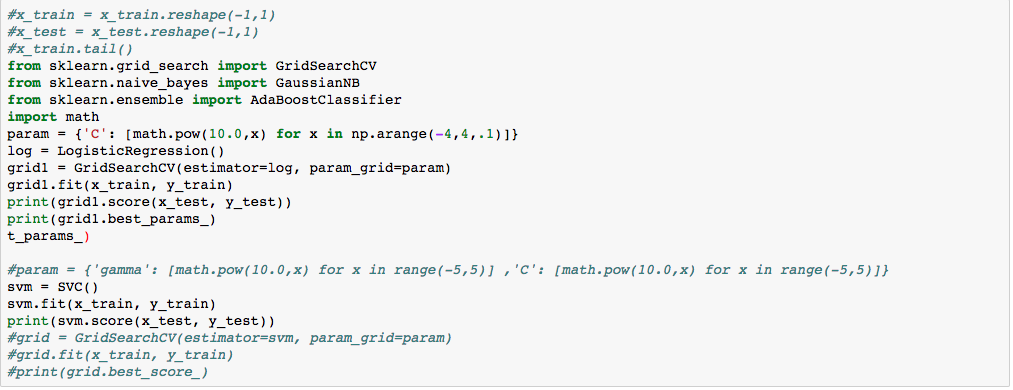




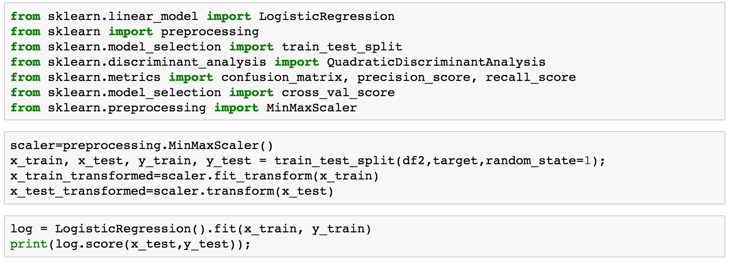








2. Fit data into model

../../../../Desktop/Screen%20Shot%202017-12-07%20at%208.37.42%20