EECS 349 Project Final Report

Student Names: Zihao Chen, Siqi Zhuo

Introduction

Our project is to predict the outcomes of an incoming soccer match in Spanish League (La Liga) based on statistics of the two soccer teams, along with other factors such as weather, and live sports betting odds from some soccer betting websites. Predicting the outcomes of sport events represents a natural application for machine learning. The precise game outcome can also be critical to soccer fans and bettors as a source of advice. As a consequence, using machine learning technology to make the prediction for soccer games is also a good way to apply what we have learnt in this course into the practice.

Dataset Description

We collected and processed the raw data we found an open source European Soccer Database. The data we collected includes 25k+ matches, players & teams attributes for European Professional Football, and FIFA data API with players and teams attributes together. After pre-processing the raw data, we generated the input dataset using the last 10 years match statistics of La Liga. We used 2500 examples in training set, and 500 examples for testing set. The features we employed will be elaborated in Analysis & Experiment Results part.

Analysis & Experiment Results

Initially, we extracted useful features in the dataset and created features that we need to generate based on the raw data. And then we performed experiments on the selected features, during which we removed features by checking whether their deletion led to improvement in accuracy.

The features we employed as follows:
Home Team Lineup: the overall ratings of lineup of home team.
Away Team Lineup: the overall ratings of lineup of away team.
HomeTeamRank: The league ranking of home team in last season.
AwayTeamRank: The league ranking of away team in last season.
WinRateHome: The win rate of home team in current season.

WinRateAway: The win rate of away team in current season.
LastFiveGameWinRateHome: The win rate of last 5 games of home team in current season.
LastFiveGameWinRateAway: The win rate of last 5 games of away team in current season.
Odds difference between Home Win and Away Win from betting websites.

Initially we tested our dataset in Weka using some methods.

	ZeroR	Decision Tree	KNN	Multi-layer Perceptron
Accuracy	48.83%	48.09%	50.86%(k=7)	48.53%

Then we wrote our own neural network implementation in R to train the final model.
The model has four layers, the first layer is input layer, with 30 feature values from dataset. the second hidden layer has 128 cells, with 0.2 dropout to avoid overfitting, the activation function is ReLu, which have better efficiency and performance than Sigmoid. The third hidden layer has 64 cells, with 0.1 dropout, activation function is ReLu as well. The last output layer we choose Softmax to be the output function, we think Softmax is suitable to deal with multi-classification problem, where we need predict among Win, Draw and Loss.
The NN model achieved around 85% accuracy in the training process(Figure 1), and the accuracy of testing set is around 55%(Figure 2).

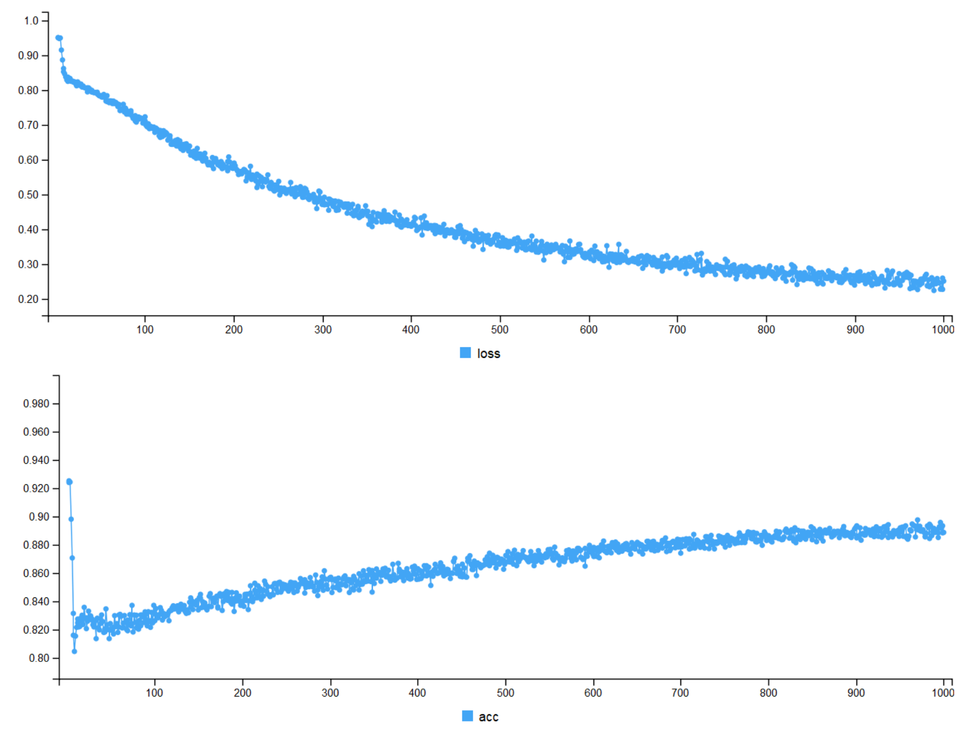


Figure 1

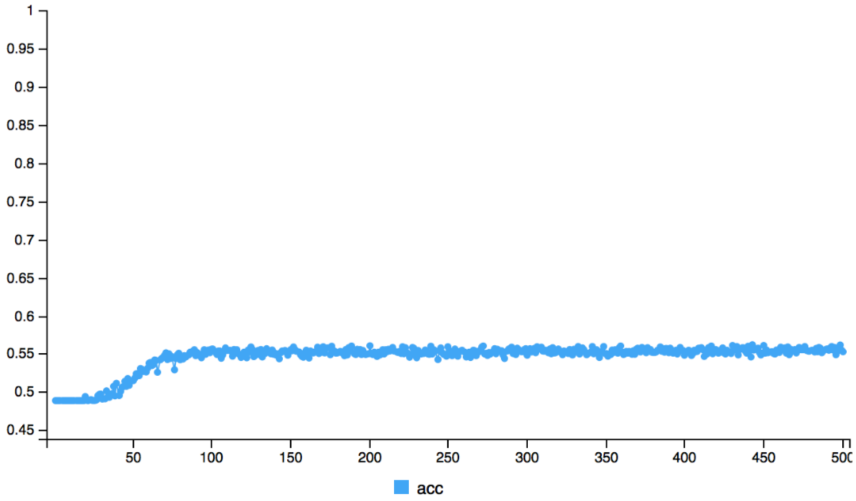
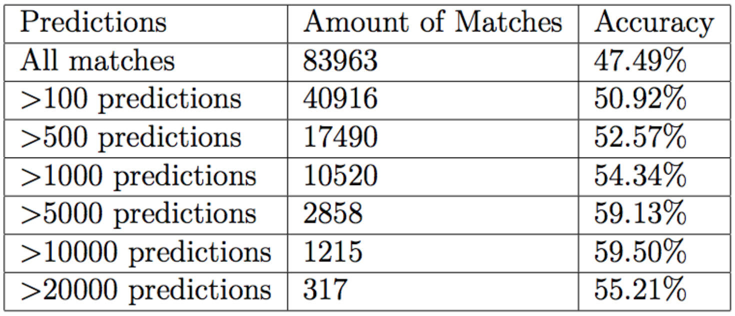


Figure 2

We searched the prediction accuracy of soccer game from many big betting websites and newspapers, and we found that their prediction accuracies before game are always between 50% to 60%. Therefore, we think the fact that our testing accuracy dropped to 55% is reasonable and acceptable.

Figure 3 shows the accuracies for Forza Football Users predictions.



Predictions	Amount of Matches	Accuracy
All matches	83963	47.49%
>100 predictions	40916	50.92%
>500 predictions	17490	52.57%
>1000 predictions	10520	54.34%
>5000 predictions	2858	59.13%
>10000 predictions	1215	59.50%
>20000 predictions	317	55.21%

Figure 3

After our analysis on the results of our Win-Draw-Lose model, we tried simplified binary prediction model (win or not), which bring a higher accuracy when predicting whether one team will win or not. We reduced the outcome labels from Win-Draw-Lose to Win-NotWin, and achieved a testing accuracy around 65% (Figure 4), which is better than ternary model.

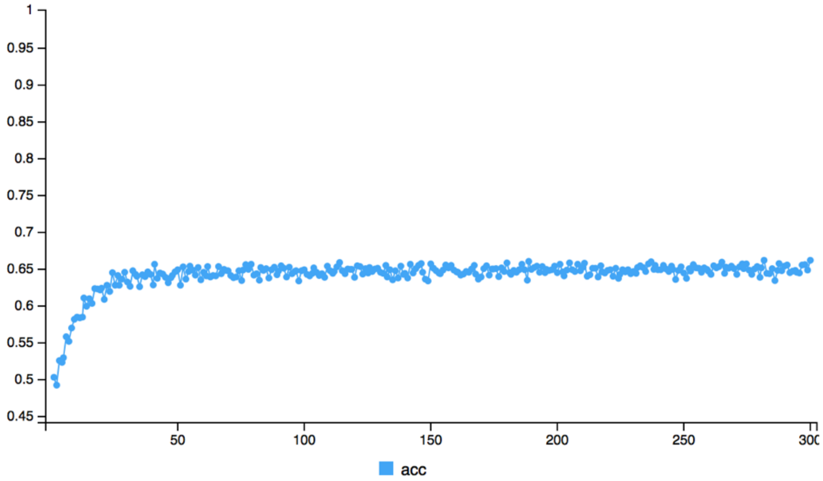


Figure 4

Conclusion

In our experiments, we built two models for both original ternary(Win-Draw-Lose) model and simplified binary(Win-NotWin) model.

	Training Accuracy	Testing Accuracy
Ternary Model	85%	55%
Binary Model	88%	65%

From the evaluation on our selected features and improvement on accuracy, we found overall ratings of team lineup and team win rate are most useful features in our model. This makes sense because higher overall ratings of lineup always means more advantages and opportunities to win in the game. And team win rate in the season is representative for the average state for one team, which is high related to team's performance.

Future Work

What’s we have planned for this project is to complete the dataset with more data about the team average match statistical, such as data as average shoot, average pass, pass success rate etc. In this way we can model one team more specifically and accurately, which may provide improvement on our prediction accuracy. We also can only predict before game start, we plan to improve our model from the first minute of game, with more and more data generate by game progress and input into our model, the model can make dynamic prediction base on game time.