MET INSTITUTE OF ENGINNERING, NASHIK

DATA MINING AND WAREHOUSING MINI-PROJECT REPORT
ON

"CLASSIFYING RESULTS OF GAME"

SUBMITTED BY

Kalyani Dawalkar-11
Rutuja Bhamare – 21
Rituje Desale - 23

Under the guidance of

Mr.Vishal Patil

DEPARTMENT OF COMPUTER ENGINEERING

Academic Year 2021-22

Contents

1 Problem Statement

Consider a labeled dataset belonging to an application domain. Apply suitable data preprocessing steps such as handling of null values, data reduction, discretization. For prediction of class labels of given data instances, build classifier models using different techniques (minimum 3), analyze the confusion matrix and compare these models. Also apply cross validation while preparing the training and testing datasets.

2 Abstract

Classification is a form of data analysis that extracts models describing important data classes. Such models, called classifiers, predict categorical (discrete, unordered) class labels. For example, we can build a classification model to categorize bank loan applications as either safe or risky. Such analysis can help provide us with a better un- derstanding of the data at large. In this project we use multiple classification models to analyse the outcome of hockey game played between various teams. Use apply suitable data preprocessing steps.We then compare performance of classification models to find which one is the best.

3 INTRODUCTION

We have been provided with the data regarding various aspects of the home team, the opposition team and their supporters for a number of hockey games.
The Data fields are

1. Id – Unique id given to each game.
2. game_seq – Sequence of the game in the history of FIH (International Hockey Federation).
3. season_end – Year in which the corresponding season ended.
4. date – Date on which the game was played.
5. season_game_seq – Sequence of the game in the corresponding season.
6. playoff – Whether the game is a playoffs game.
7. team_id – Unique id for the home team.
8. Elo – Elo rating for the home team before the game.
9. opp_team_id – Unique id for the opposition team.
10. opp_Elo – Elo rating for the opposition team before the game.
11. win_equivalent – Equivalent number of wins for the home team in a season.
12. bet_ratio – Fraction of bets placed on the home team.
13. home_crowd – Number of supporters for the home team.
14. opp_crowd – Number of supporters for the opposition team.
15. total_crowd – Total number of attendees for the game.
16. game_result – Win or loss for the home team (Win - 1, Loss - 0).

The train set contains 45000 records while the test set contains 13107 records.We drop the date column from our analysis. The null entries are as follows

Attribute	Null Count
Elo	9197
opp_Elo	7006
win_equivalent	12263

Table 1: Null Counts

We fill the null Elo and opp elo entries with the mean value of Elo and opp Elo attribute respectively i.e. 1501.184 1501.837

The boxplots of some attributes are as follows:

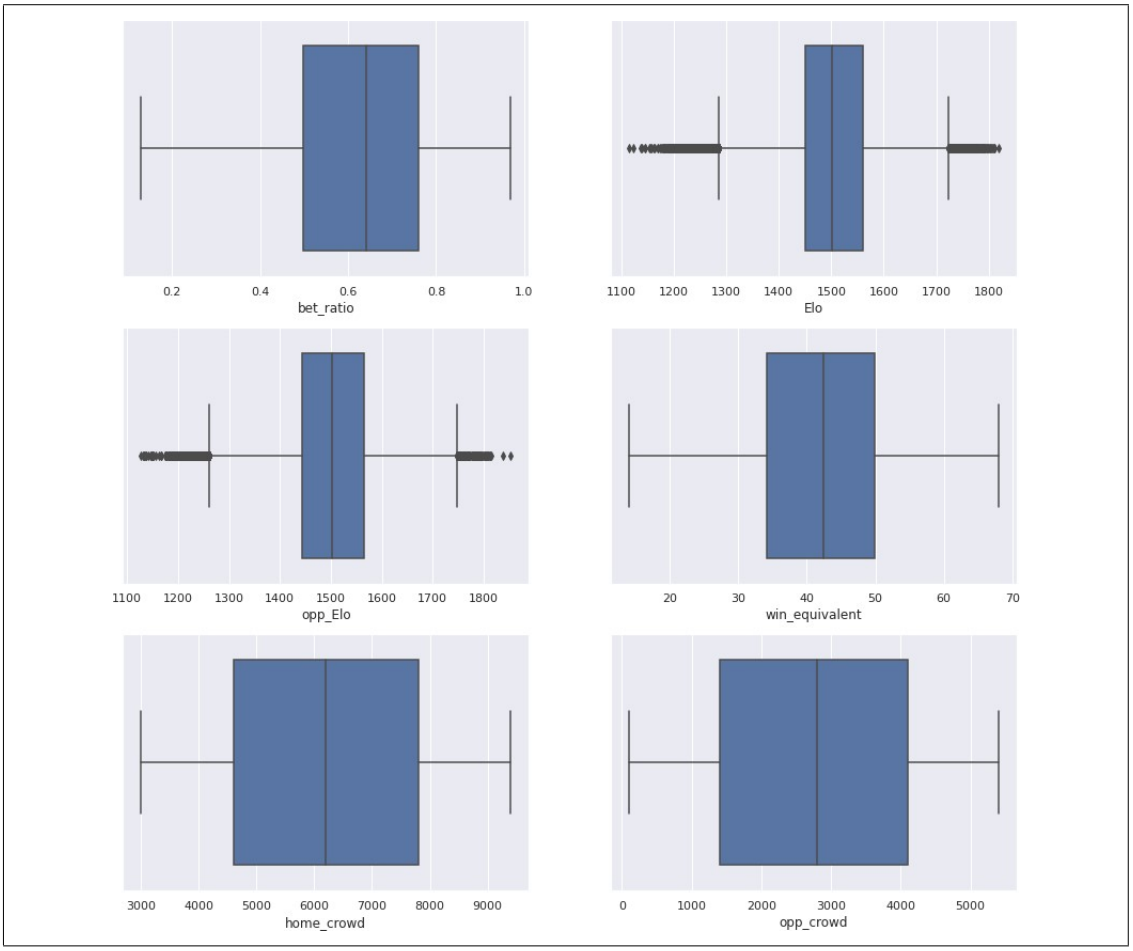

Figure 1: Boxplots

We drop rows with bet_ratio below 0.13 and above 0.97. Similarly we drop rows with opp_Elo below 1200 and above 1750, Elo below 1175 and above 1780 and win_eqivalent below 14 and above 68

We extract the id from string team id and opp id fields. We drop columns Id,game seq,team id and opp_id. We convert playoff to categorical data.

We have trained using two models Logistic Regression, KNN classifier , XGBoost Classifieer and RandomForest Classifier. We find that XGBoost Classifier performs better.

4 OBJECTIVE

- To understand data preprocessing
- To perform classification on dataset and predict labels for test dataset.

5 Test Cases

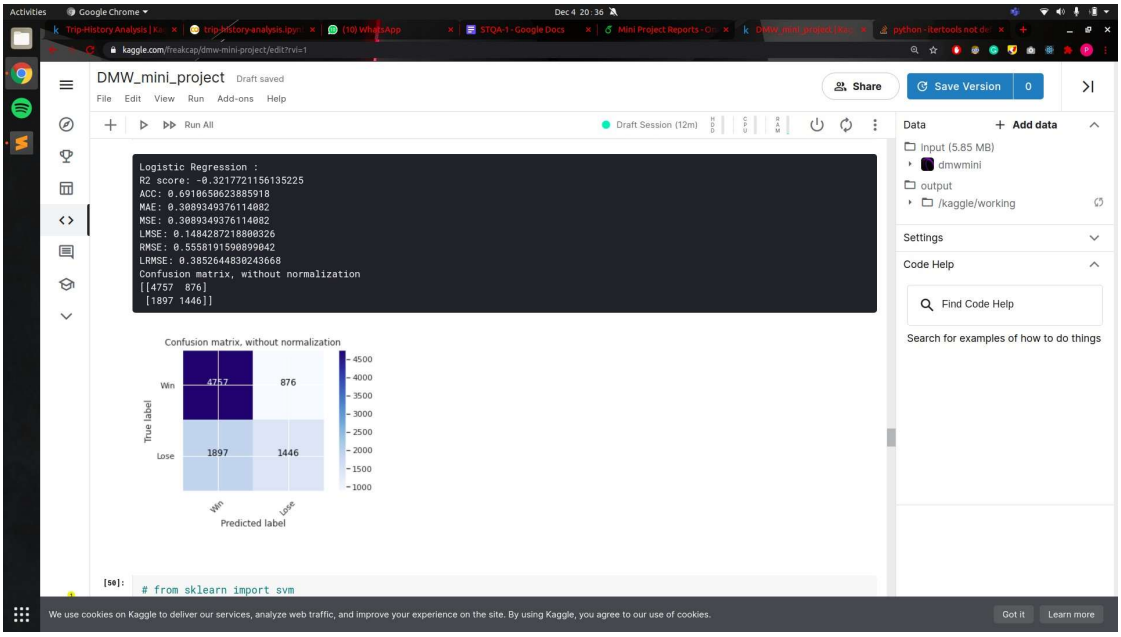

Figure 2: Output for logistic regression

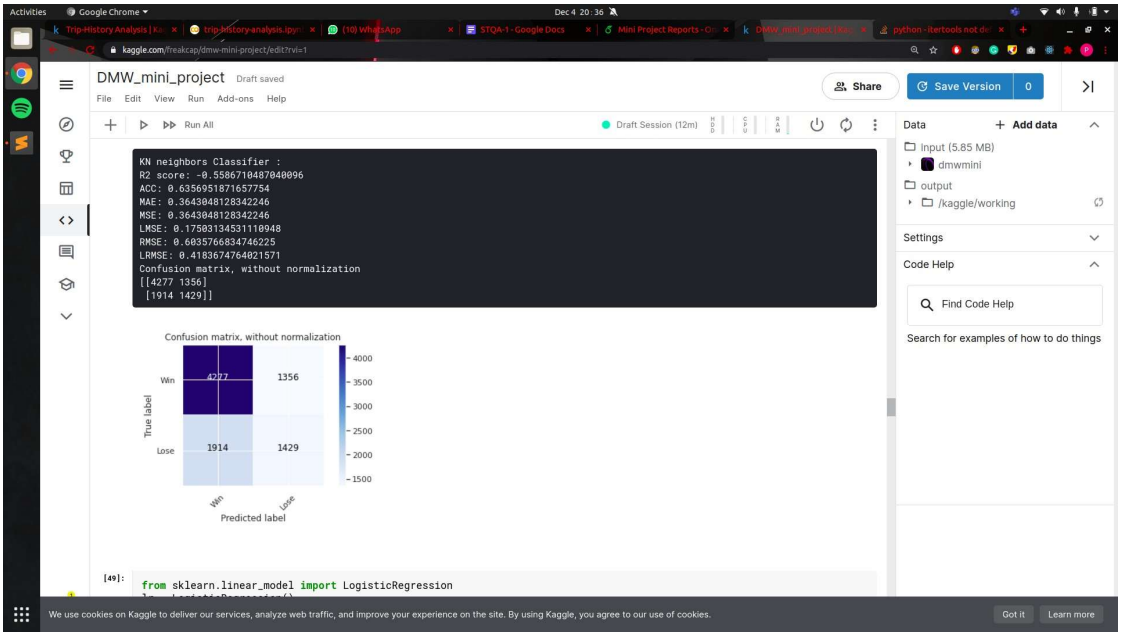

Figure 3: Output for K Neighbours classifier

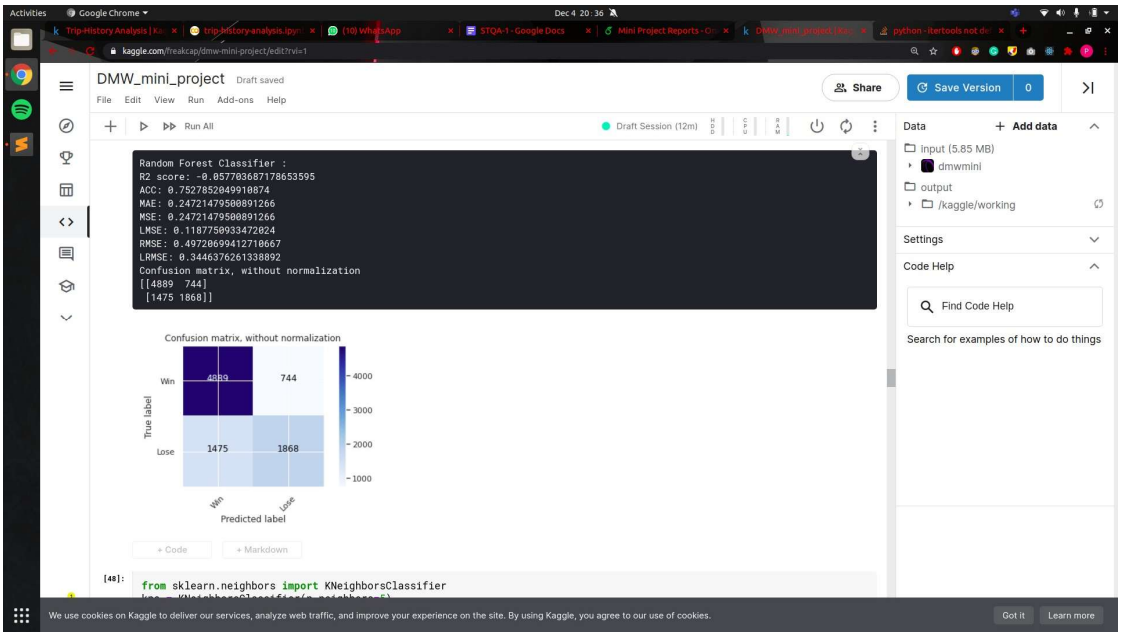

Figure 4: Output for random forest classifier

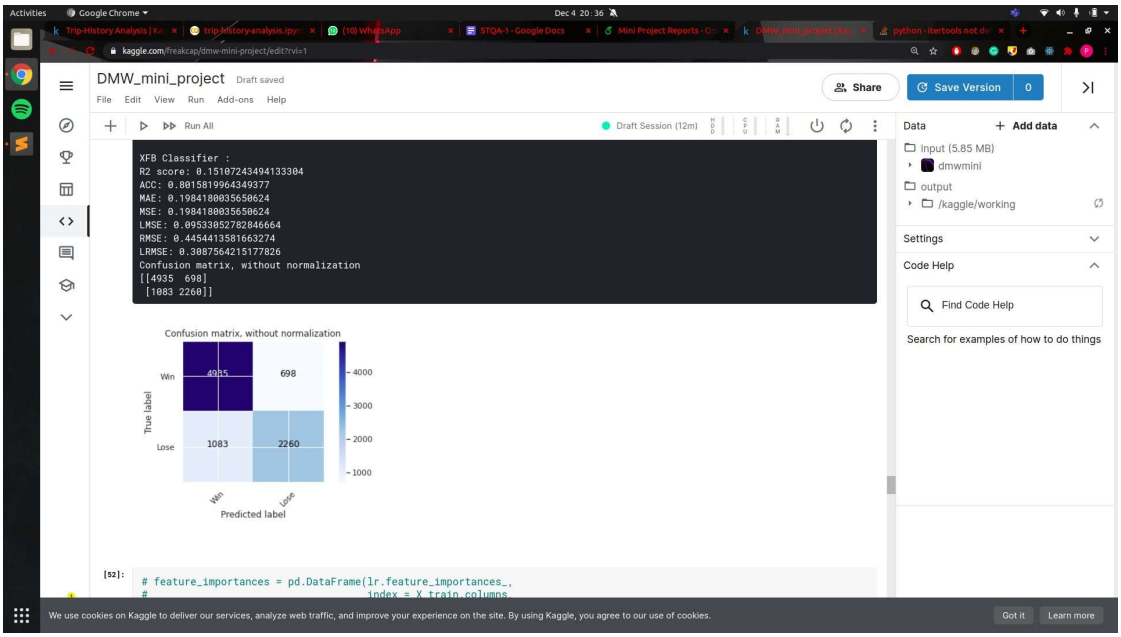

Figure 5: Output for XGBoost classifier

6 Result

The accuracy for XGBoost classifier is arround 80%. while that of other models is lesser. The following are confusion matrices of various model outputs.

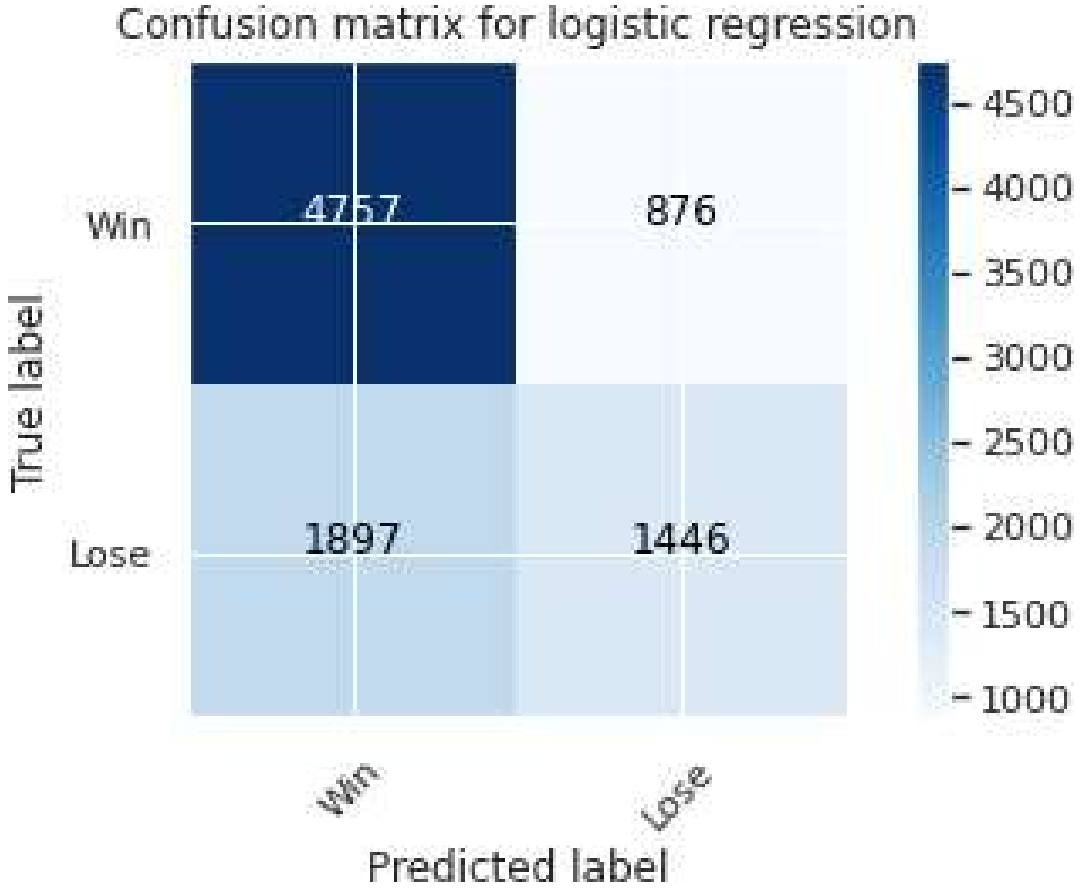

Figure 6: Logistic regression

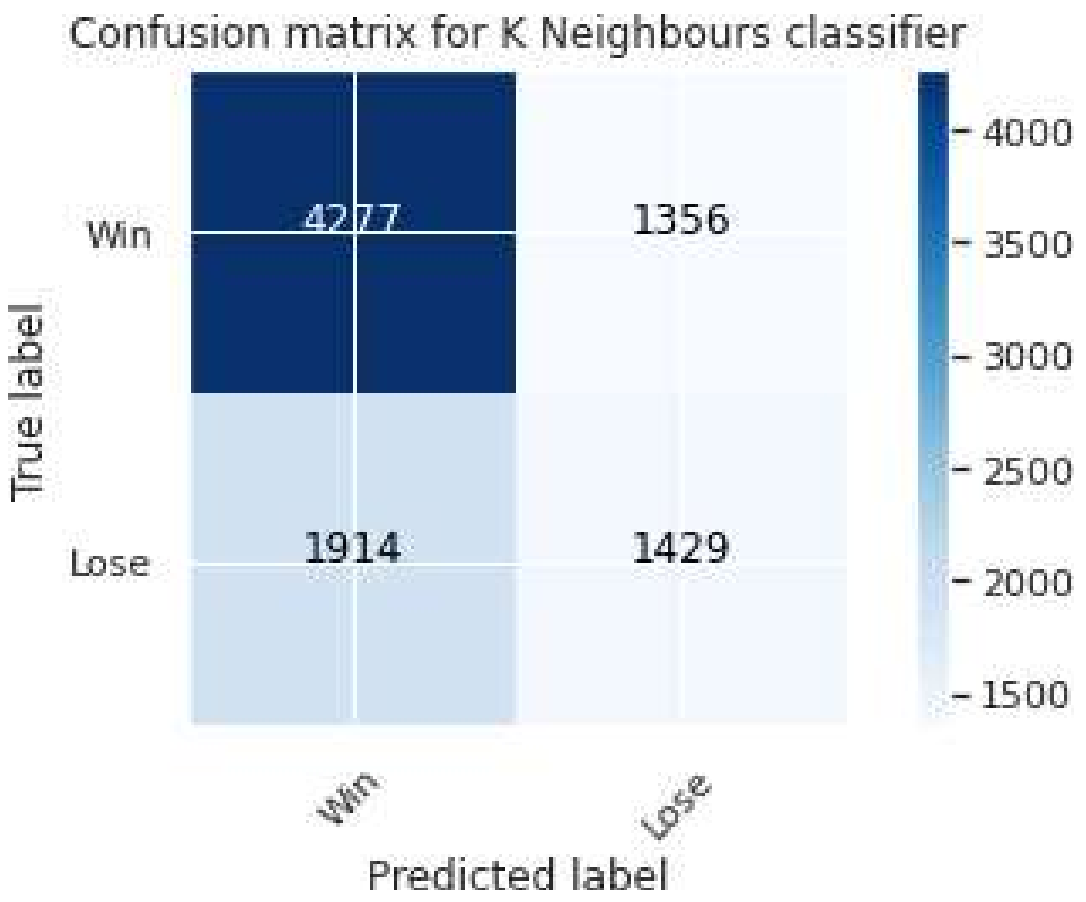

Figure 7: K Neighbours classifier

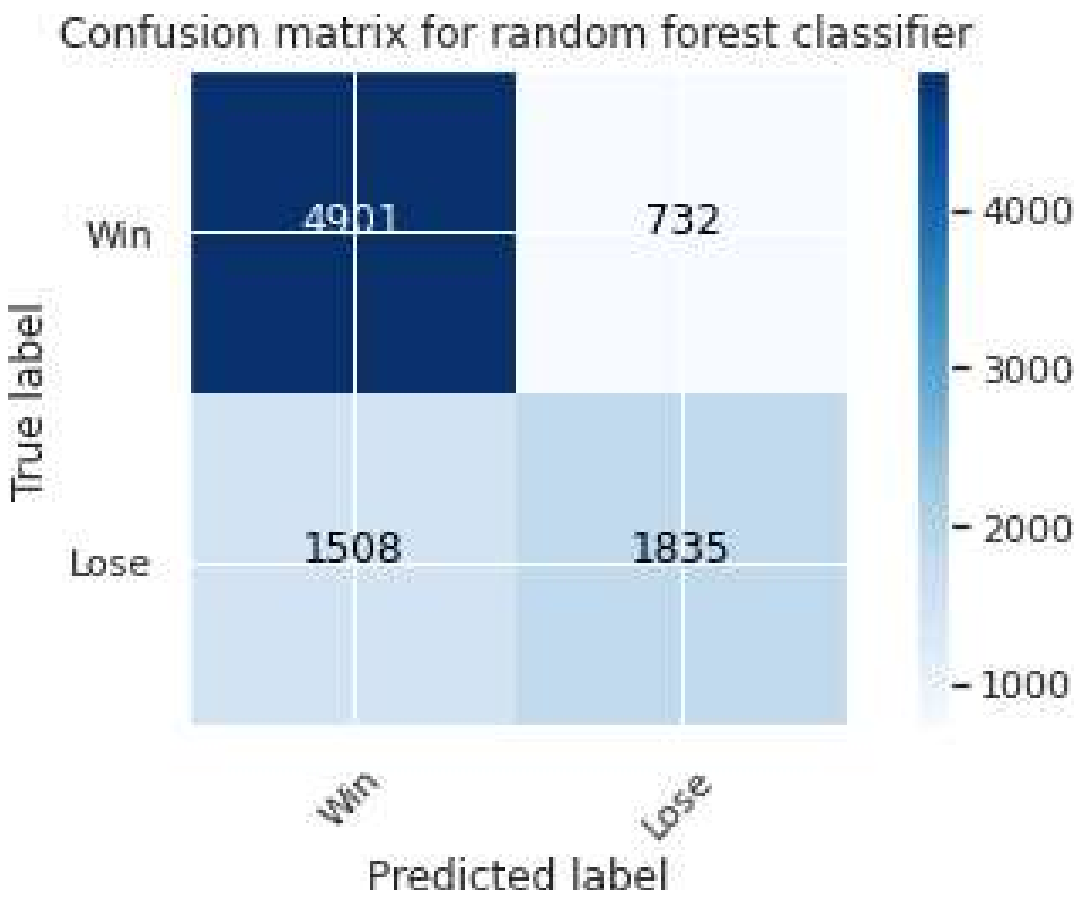

Figure 8: random forest classifier

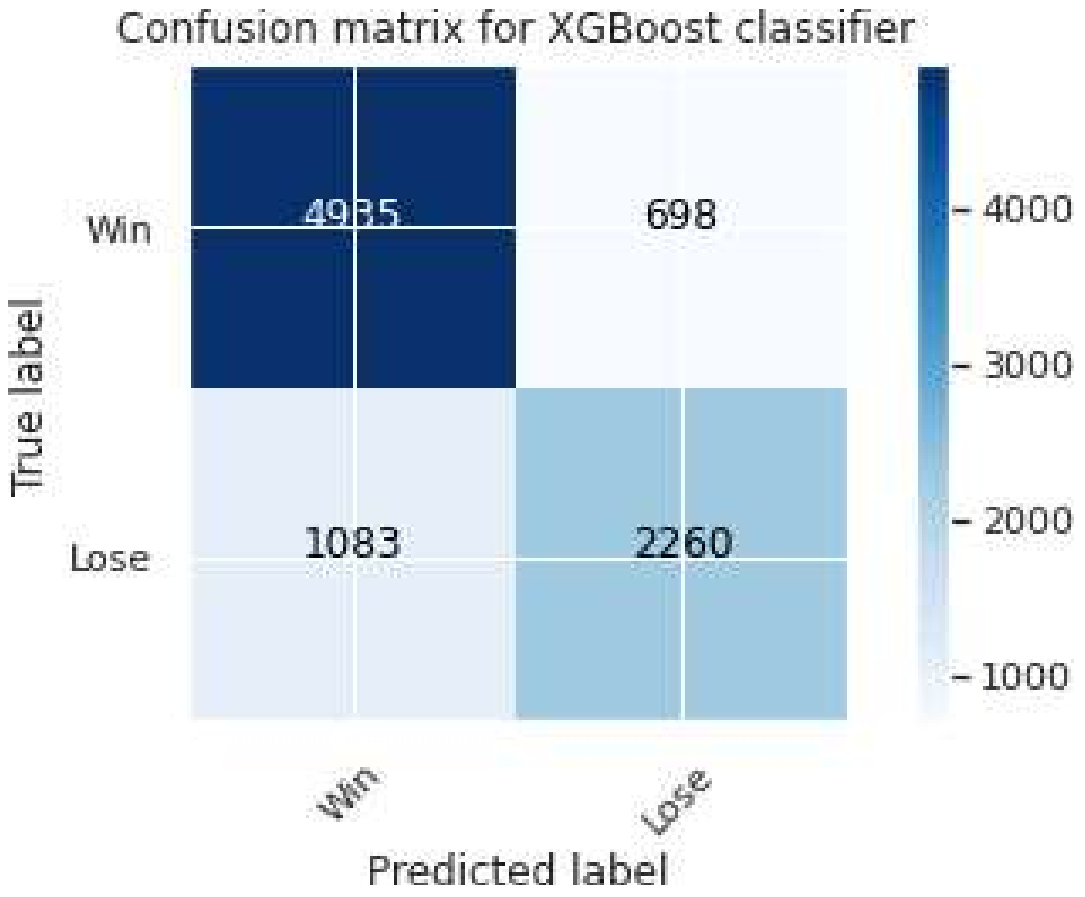

Figure 9: XGBoost classifier

7 Conclusion

We have analysed the hockey game dataset and performed data pre-processing steps. We have experimented multiple classification models and found out the best performer amongt them. We have then used this model to mae predictions on test dataset.

References

[1] https://www.kaggle.com/c/datawiz19round1/data

[2] https://seaborn.pydata.org/index.html

[3] Jiawei Han,Micheline Kamber, Jian Pei, *Data Mining Concepts and Techniques*

List of Figures

List of Tables