CSE 482 FINAL PROJECT

Project Title: NBA Game Predictor

Summary of Team Member Participation:

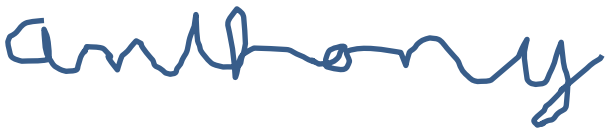
Fill out the following table for each team member of the group.

Name	**Participate in data collection**	**Participate in preprocessing**	**Participate in data analysis/ experiment**	**Participate in writing the final report**	**Completed Assigned Tasks**
Anthony Ficarra	**X**	**X**		**X**	**X**
Abhinav Thirupathi	**X**	**X**	**X**		**X**

Team Member Roles and Contributions:

Name	**Roles and Contributions**
Anthony Ficarra	Responsible for collecting and preprocessing the data; help writing the final report
Abhinav Thirupathi	Responsible for collecting and preprocessing the data; perform data analysis;

I approve the content of the final report (please add your signature below):

Anthony Ficarra: 

Abhinav Thirupathi: 

NBA Game Predictor

Abhinav Thirupathi
Anthony Ficarra
Project URL:
https://github.com/abhit20/NBA-Game-Predictor

ABSTRACT

The objective of this project was to create a model that can predict the winner of an NBA game. The data was collected from basket-reference and ESPN. To achieve this, we gathered data for each match from the 2017-2018 season which is a total of 1230 games for training the model and used 410 games played in 2018 from the 2018-2019 season for evaluating the model. All this data was preprocessed, and 52 feature vectors were created based on the averages of each teams last 7 games. The model was a classification problem solved through logistic regression, so the winner of each game was the predictor. 1 if the home team won and -1 if the away team won. First it was performed without sampling. The outcome of that model was 61.74% split between the home and away team. After that, sampling with replacement was performed to see if there could be a better outcome. The outcome of sampling with replacement was 64.98% split. The sampling with replacement model provided a much better outcome for the predictor than the model without sampling.

1. INTRODUCTION

NBA basketball games are played between two teams from different locations in the United States and Canada. The winner of each game is determined by each team's final score. Whichever, team scored the most points wins. One of the five players on the court for their team can score points by shooting the ball into a basket. The game of basketball has been loved by many since Dr. James Naismith invented it in 1891. Even though the NBA is primarily located in North America, it is a global brand. People from around the world enjoy following NBA games.

Some fans of the NBA even enjoy predicting games. Many like to predict games to prove that their favorite team is better than another. Others may want to predict a game so they can gamble on it. This interest in predicting games led to the idea of creating a model to predict the outcomes of NBA games based on statistics from games that were previously played.

The goal of this project was to determine if an NBA game could be predicted with a classification model. If an accurate prediction to an outcome of an NBA game based on team statistics from previous games was found, the classification model would be successful.

The goal to predict an NBA game through a classification problem would be achieved through a logistic regression model with and without sampling. In both these models, 52 featured vectors were used. These featured vectors contained the average statistics from the home team, away team, and the opponents of the home and away team from their last 7 games. The logistic regression model is evaluated using alternative measures such as the confusion matrix and F-measure.

In order to make this model, multiple statistics from NBA games were needed. Statistics such as assists, turnovers, and rebounds were collected in order to help predict games. These statistics help determine which team would have an advantage based on there averages of each one. There were also some statistics that proved to benefit the home and away team if they were more efficient.

Collecting and analyzing the data for the logistic regression model also had some obstacles. One major challenge occurred when the data was be collected. Each NBA game had its own CSV file. To properly train and test the model we needed to collect 1641 box scores. In order to collect and preprocess the data before the project deadline we ended up collecting the data manually from the box score of each game.

After collecting, preprocessing, and analyzing the data, the results that were achieved from the logistic regression model with and without sampling were gathered. The results from the test case without sampling proved to have a test accuracy of 63.39%. The accuracy was also gathered by using a confusion matrix to calculate the F score. The average F score achieved from the model without sampling was 61.74%. Next the results from the model with sampling was gathered. A testing accuracy of 65.03% was achieved along with an average F score of 64.98%. Therefore, the goal of being able to predict NBA games with a classification model was successful as the best case was able to predict games correctly over 65% of the time.

2. DATA

The data from the NBA games were collected from Basketball Reference and ESPN (URLs given in References section). The data from these sites were then collected manually into excel spreadsheets. One spreadsheet was for the 2017-2018 NBA regular season, which was the training data, and another was for the 2018-2019 NBA regular season, which was the testing data. Each statistic (table 1) that was used to predict the games in the model were copied from the box scores for the game provided by Basketball Reference and ESPN. Finally, an extra column was added to determine the winner of each game. If the value was a 1 the home team won, if -1 the away team won. This was the target class that was used in this classification problem. The format of the data was the websites HTML code.

Attribute name	Type	Description
Date of Game	Nominal	The day the teams playing, played the game on.
Team Name	Nominal	The Name of the team playing
Team Points	Addition	Points scored by the team playing
Team Field Goal Attempts	Addition	Attempted shots by the team playing
Team Field Goal Percentage	Multiplication	The percentage of attempted shots made
Team 3 Point Attempts	Addition	Attempted 3point field goal by team
Team 3 Point Percentage	Multiplication	The percentage of 3point field goals that were made
Team Free Throw Attempts	Addition	The number of free throws attempted by team
Team Free Throw Percentage	Multiplication	The percentage of free throws made
Team Total Rebounds	Addition	The total number of rebounds the team had
Team Assists	Addition	The total number of assists the team had
Team Steals	Addition	The total number of steals the team had
Team Blocks	Addition	The total number of blocks the team had
Team Turnovers	Addition	The total number of turnovers the team had
Team Personal Fouls	Addition	The total number of personal fouls the team had
Target Class	Addition	Used to determine the winner of the game. 1 if the home team won and -1 if the away team won.

Table 1: Attributes of the data acquired from Basketball Reference and ESPN. These attributes were for one team after one game. These stats were taken from every game in the 2017-2108 NBA season and 411 games in the 2018-2019 NBA season. These stats were then put into a script were the averages of each stat was calculated.

There were many characteristics to the raw data collected from both Basketball Reference and ESPN. The data was collected from the entire 2017-2018 (training data) and select games from the beginning of the 2018-2019 (testing data) NBA regular season. There were many problems when collecting the data from these websites. One issue was that each game had its own web page. Another issue was the fact that there were csv files available to download but only for one team at a time for one game. To overcome these problems, the data from each game was manually input into an excel document. There were also multiple data points that were included that were low factors in predicting the outcome of a game for a certain team. Some of these data points included a player's own individual statistics. Those data points were not placed into the excel spreadsheet where the manually entered data was input. Overall, there were some issues that occurred when collecting the data but were solved once the decision was made to enter the data manually.

Number of Games	411 games (Oct 2018-Dec 2019) 1230 games (Oct 2017 – Apr 2018) (187 KB)
Number of missing values	0%
Number of Attributes	30 columns: 1 date of game, 1 name of home team, 13 home team statistics, 1 away team name, and 13 away team statistics

Table 2: Summary statistics of the raw data from Basketball Reference and ESPN.

Number of Games	999 games (training dataset) 183 games (testing dataset)
Number of missing values	0%
Number of attributes	52 feature vectors were created based on the 4 sets of 13 averages calculated from the original two sets of 13 original statistics collected for each home and away team. 1 target class

Table 3: Summary of statistics in the processed datasets

There were also some preprocessing steps that were needed in order to get the proper data for the models. When collecting the data an extra column was needed in order to determine the winner of the game. A script was also written in order to calculate the average statistics for the home team, away team, home opponent, and away opponent's average statistics from the previous 7 games. That data was then placed into the training and testing files. The training file got the team averages from the 2017-2018 NBA season. The testing file, on the other hand, had the team averages from the 2018-2019 NBA season. The data from those files were then placed into 52 feature vectors. These vectors where then used in the logistic regression model. After performing the model there were statistics that were revealed to favor the away or home team. Finally, sampling was performed on the data where the away team was the winner in order to make the training data larger. After performing sampling on the data, a more favorable outcome occurred.

This project was a classification problem solved through a logistic regression model. This being a classification problem the main class that was used to predicate games was the winner class. The winner class had 1 and -1 in it to determine who won the game, the away or home team, respectively. This was the main target class used to help the model predict the outcome of each future game.

After collecting and preprocessing the data, each file was a different size. The data from the 2017-2018 NBA season was .140 Mbytes. Then after sending this data through the script the train-averages file had the size of .291 Mbytes. The data from the 2018-2019 NBA

season was .047 Mbytes. After sending this data through the script, the size of the test-averages file was .054 Mbytes. After collecting the data for the training and testing file, that data was placed in a feature vector with the size of 52. The training dataset initially had had 53 columns and 999 rows when there was no sampling. With sampling there were 1158 rows and 53 columns. The testing data, on the other hand, had 183 rows and 53 columns. All in all, there were many different data sets of many different sizes.

3. METHODOLOGY

This project was completed after performing many necessary steps needed to solve this problem. First the data needed to perform the analysis had to be collected. So, the data was then collected manually from Basketball Reference and ESPN and places in an excel document. After collecting the data, the files were converted into csv files in order to run the script that was created on the data. The script then read the data from the NBA 2017-2018.csv and NBA 2018-2019.csv files and placed-placed them in the train and test files. The data from the 2017-2018 regular season was placed in the train-averages.csv file and the data from the 2018-2019 season was placed in the test-averges.csv file. Finally, the data analysis was done on the classification problem using a logistic regression model.

With this problem being a prediction problem, training and testing data was needed in order to solve it. The training and testing data were created from the Python preprocess.py that was implemented on the data that was collected from the 2017-2018 and 2018-2019 NBA seasons. After that data went through the script, the average statistics of the home team, away team, and the home and away team's opponents were calculated and placed into files named training-averages.csv and testing-averages.csv with the target variable being the true outcome of that game. These files were then placed into 52 feature vectors and 1 target variable to be used in a logistic regression model to solve the classification problem.

The code for this project was separated into many different formats. The collection of the data was done manually. The data was copied from either Basketball Reference or ESPN and then placed into an excel document. The preprocessing step was then done in the preprocess.py python file. This is the python program used to calculate the average statistics from the previous 7 games of the home team, away team, and the home and away team's previous 7 opponents. The output was indeed the averages of each team's last 7 games statistical averages. The target class was the true outcome the game being predicted (-1 or 1). Finally, the modeling was done in a Jupyter notebook file named modeling.ipynb. This is were the classification problem was performed using a logistic regression model with cross validation for the selection of the best hyperparameter.

In summary:

- Manually: The statistics for each game was manually input into an excel document. All the statistics were taken from the box scores of Basketball Reference or ESPN. Another column was also added to the excel document to show who the winner was. That was the target class in this classification problem.
    - Result after this step: NBA 2017-2018 (1230 games) and NBA 2018-2019 (411 games) files. Both files have 30 columns: 1 date column, 1 home team name column, 13 columns for the home team stats, 1 column for away team, 13 away team stats, and 1 target class (-1 or 1).
- preprocess.py: this is the python script file to read the game files, calculate the averages of the home and away teams and the home and away team's opponents previously seven games. It then wrote to the separate files the averages, and target class that were used to perform the classification task.
    - Result from the preprocess.py: One train-averages file (999 games) and one test-average file (183 games). Both files have 53 columns: 13 columns of home team averages, 13 columns of home opponents' averages, 13 columns of away team averages, 13 columns of away opponents' averages, and 1 target class (the true outcome of the game).
- modeling.ipynb: this is the Jupyter notebook file to perform the classification task of the project.

4. EXPERIMENTAL EVALUATION

4.1 Experimental Setup

This project was completed using personal MacBook Pros for hardware and Jupyter Notebook and Pycharm for the software. The first step in creating the experiment was to collect the data. After that the data was preprocessed and using a script and placed in 52 featured vectors and the target class. The feature vectors were then analyzed using a logistic regression model. One of the models did not use sampling and just used the data provide in the 52 featured vectors. The other model used sampling on the games were the outcome had the away team winning (-1 in the winner class). Finally, the result of the best hyperparameter, test accuracy, root mean squared, R-square, and F score were then gathered from the model with and without sampling.

4.2 Experimental Results

	Logistic Regression without sampling	Logistic regression with sampling with replacement
Target class value counts	1: 579, -1:420	1:579, -1:579
Best hyperparameter	C = 0.05	C = 0.01
Test Accuracy	0.63387	0.6503
Root Mean Squared Error	1.2102	1.1828
R-square	-0.5201	-0.4520
Average F1-score	0.6174	0.6498

Table 4: Summary of the value counts, final performance and evaluation metrics for without and with sampling with replacement

After the experiment was setup, the results were gathered from the models with and without sampling with replacement. In the model without sampling, as summarized in Table 4, the results of a test accuracy of .63387, a root mean squared error of 1.2102, an R-square of -0.5201 and the best hyperparameter at .05. The results in this test case were not great. The test accuracy was only a little over 63% successful. When calculating the root mean square a low number is better and higher number is better for square. The result of the root mean square, however, had a result over 1 and the R-square was below 0. After that, a confusion matrix and F1 score was calculated for further evaluation, which also shows the accuracy of the test. The F1 score gave a result of .61732. However, the confusion matrix (Table 4) shows that the model was much more efficient in predicting home teams winning, but poorly predicted the away team's victory. Therefore, the model that did not use sampling was far better at predicting home team wins (1) than at predicting away team wins due to the slight skew in the data set (579 home team wins compared to only 420 away team wins). The model coefficients show that home opponents blocks' (-0.157), away blocks (-0.155), and away steals (-0.126) were weighted more in the prediction of the away team wins (-1). While, away opponents' turnovers (0.072), and away team assists (0.053) were some of the features that played a crucial role in the prediction of the home team wins (1).

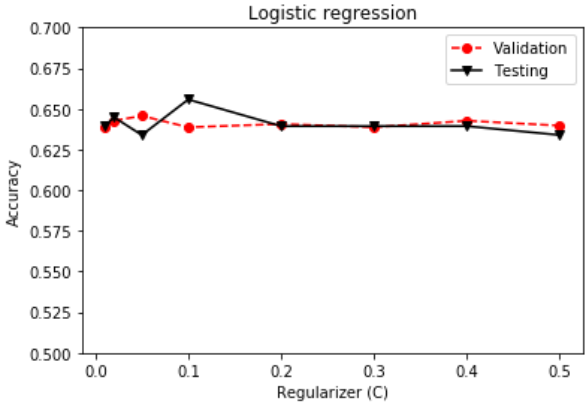


Figure 1: Graph of test set accuracy and validation set accuracy against the regularization parameter C for without sampling

	0	**1**
0	39	35
1	32	77

Table 5: Results of the confusion matrix. The 0 represents the games won by the away team and the 1 represents the games won by the home team.

To reduce the slight skew in the data towards home team wins (1), the decision was made to sample with replacement. The games were the away team won were sampled as they were poorly predicted as shown by confusion matrix in Table 4 (only 39 compared to 77 correctly predicted home team wins). The outcomes tested much better than the model without sampling. The results of the sampling data included a test accuracy of .65027, a root mean squared error of 1.1828, and R-square of -0.4520. The test accuracy increased by almost 2 percent. However, even though the root mean squared and R-square improved there were still not the best results. The confusion matrix also proved to be much better when predicting the away team's winning. The F score produced by the confusion matrix was .64976, which was much higher than the F score produced without sampling. Overall, the results produced by the model with sampling were much more successful the model without it. The model coefficients show that home opponents blocks' (-0.205), away team steals (-0.146), and away team blocks (-0.105) were weighted more in the prediction of the away team wins (-1). While, home team blocks (0.122), away team turnovers (0.116), and away personal fouls (0.058) were some of the features that played a crucial role in the prediction of the home team wins (1).

To further compare models trained without and with sampling, the confusion matrixes (Table 5 and Table 6) clearly show some tradeoff. The model without sampling was very good at predicting home team wins, but the average was 0.6174. After sampling with replacement, the model was better at predicting away team wins. This, however, made the model worse at predicting home team wins. There was a tradeoff made to make the model better at predicting away team wins through the use of sampling with replacement, nevertheless sampling with replacement made the model better at predicting away team wins from the initially model which had no sampling.

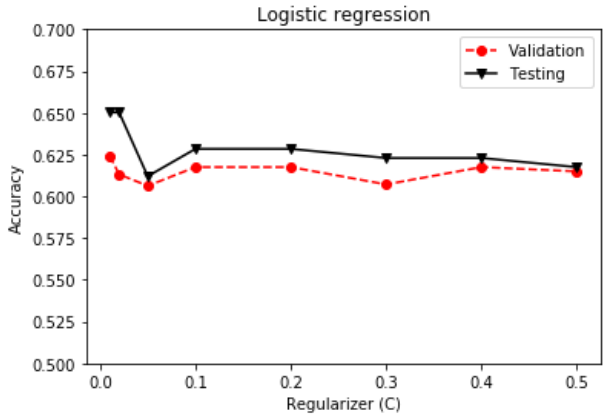


Figure 2: Graph of test set accuracy and validation set accuracy against the regularization parameter C for with sampling.

	0	**1**
0	56	18
1	46	63

Table 6: Results of the confusion matrix. The 0 represents the games won by the away team and the 1 represents the games won by the home team.

5. CONCLUSIONS

This project was a classification problem solved with a logistic regression method. The data collected from Basketball Reference and ESPN were then preprocessed and analyzed. After that the

results from the with and without sampling models were collected. The results from the test case without sampling proved to have a test accuracy of 63.39%. The accuracy was also gathered by using a confusion matrix to calculate the F1 score. The F1 score achieved from the model without sampling was 61.74%. Next the results from the model with sampling was gathered. A testing accuracy of 65.03% was achieved along with an F score of 64.98%. The model using sampling was better than the one without as it had a higher F1 score and test accuracy. Even though the best model successfully predicted games over 65% of the time accurately, it could have been improved if more statistics from NBA games were gathered than the 13 used as well as gathering those statistics from more games then the total used. Overall, the prediction model was a success though as the best model predicted 65.03% of the games accurately. In conclusion, logistic regression model trained with data without sampling was better at predicting home team wins than away team wins, but the model using sampling with replacement became better at predicting away team wins while sacrificing prediction accuracy of home team wins.

6. REFERENCES (at least 3 references)

[1] "2018-19 NBA Schedule and Results." *Basketball Reference*, www.basketball-reference.com/leagues/NBA_2019_games.html.

[2] "NBA Basketball Scores - NBA Scoreboard." *ESPN*, ESPN Internet Ventures, www.espn.com/nba/scoreboard..

[3] "Scores." *NBA.com*, www.nba.com/scores#/.

[4] The pdf notes from Lecture 12 and 13 available on the D2L course website were also referenced multiple times in this project.