Walter Meeker

Comp 4449

Midterm Project

**NBA Salary Prediction**

**Introduction**

The project presented herein is a series of regression models used to predict the salary of NBA players spanning the years 1984 to 2017. The features used in this model are listed below and included key statistics unique to each player, such as: round drafted, career games played, career field-goal percent, season year, team played for in given season, and more. The models tested included Linear Regression, Random Forest Regression, and XGBoost. After preparing the data, each model was trained and the results were compared and the best model was selected for hyperparameter tuning.

**Data**

The data were given in two separate CSV files. The first, called Players, included static information on 4,658 unique players, with a given shape of 4,658 rows by 7 columns. The values of features given in players do not change year-over-year. The second file, called Salaries, contained time-series data on 2,408 unique players’ salaries and team in each season played. Salaries was given with a shape of 14,163 rows by 6 columns, implying an average number of salaries / seasons of 5.9 per player. The features of each dataset are listed in table 1 below.



**Table 1:** Feature names of Salaries and Players datasets

**Merging the Data**

A full outer join was used to merge the two sets of data. Recall that the information given in the “Players” dataframe did not change over time, there was only a single instance for each player given. For instance, a player’s height, hometown, and career games played did not change season-over-season, but their team, their salary, and the year of the season did, which are examples of features given in the “Salaries” dataframe. Therefore, performing a full outer join meant that if a player appeared 5 times in the “Salaries” dataframe, then their statistics from the “Players” dataframe would be repeated year-over-year. This imperfect solution will be discussed more in the conclusion section.

**Feature Creation**

**Age**

Player age was not given; however using their birth year and the season year the player’s age in that season was calculated.

**Rookie**

A dummy variable for whether a player is a rookie was also created using the current season year and the respective player’s draft year.

**Data Cleaning**

**Height**

Player height was given as a string. For instance, six-foot ten-inches was given as, “6-10.” To address this, a function was written to convert height to a number in inches.

**Position**

Player position was given as a string and did not contain a consistent method of labeling. Therefore, to simplify a function was written to assign a player as “Offensive” if the word “guard” appeared in their position-name and “Defensive” otherwise.

**Other**

Other features that were cleaned included “shoots,” which indicated a player’s handedness and “round drafted,” which indicated the round a player was drafted.

**Feature Selection**

A relatively naïve approach was used for feature selection in this project, which in hindsight is one of the primary reasons for model performance, which will be discussed more later. The final features selected for modeling are listed in Table 2 and 3 below. The method used to select these features was based on a qualitative opinion of their significance to the target variable.

**Table 3:** Categorical Features Used for Modeling

**Table 2:** Numeric Features Used for Modeling



**NA Handling and Datatypes**

Of the features selected, all were converted to the proper datatype as shown above and any remaining rows with NAs were dropped. Table 4 below shows the percent of NAs that appeared and that were subsequently dropped.

**Table 4:** NA Percent



**Visualization**

The relationship between each of the features and the target variable were visualized using bar plots and scatter plots, shown below. A few relationships were apparent. For instance, players who can shoot with their left and right hand make more money, on average, than those who only shoot with one. First round draft picks, on average, receive higher salaries throughout their career. Players make more money as they age into their early twenties and then begin making less money as they approach their thirties. Player salaries have increased steadily since the nineteen eighties. Seeing these relationships is a positive sign and signals that there is relevant information in the features to predict the target.

Chart, bar chart, histogram

Description automatically generated

**Figure 1:** Barplots of Salary vs Selected Features

Graphical user interface

Description automatically generated with medium confidence

Chart, scatter chart

Description automatically generated

**Figure 2:** Scatterplots of Salary vs Selected Numeric Features

**Preprocessing**

The data were scaled using sklearn’s MinMaxScaler, which was chosen due to the high number of one-hot-encoded features. The categorical features were one-hot-encoded using pandas “get\_dummies” function. The data were split into seventy percent training and thirty percent testing using sklearn’s TrainTestSplit.

**Modeling**

Three models were tested: linear regression, random forest regression, and XGBoost regression. Of the three models, only the XGBoost was tuned. The results of these are shown in Table 5 below and the plots of the test predictions vs actuals are shown in the accompanying Jupyter notebook.

**Table 5 and 6:** Results of each model and mean, standard deviation of the target





Other than the linear regression model, each model tested suffered from overfitting. The results of the XGBoost model were the most favorable in terms of RMSE of the test predictions, with a value of $2,015,046. Comparing this value to the standard deviation of the target variable shows that we are well within one standard deviation, which is positive. However, given the overfitting there is likely a better solution to be found.

**Conclusion**

The results of this project suffered from overfitting and despite being within one standard deviation of the mean target value, a better solution may be found. The primary problem that this project has suffered from is feature selection, which will be a primary focus in next steps. One example of a potential problem variable is “team.” The relationship between team and the salary paid is likely to have changed over time as certain clubs have risen to be more successful while others have faltered. Additionally, performing a full outer join on the two initial datasets is likely to have provided too much repeated information and likely caused the model to bias certain feature-values. These are primary examples of issues that need to be addressed.

**Bonus Model**

To test the issue caused with the full outer join a second dataset was created, which only included the last instance for each player listed in the dataset evaluated here. The results of this did not provide better results and therefore I am not discussing it here. This can be seen in the jupyter notebook submitted with this report.