ORIE 4741 Midterm Report

Genghis Shyy (gs484) , Heesun Chang (hc483), Mohammad Kamil (mk848)

November 8, 2020

1 Data Preprocessing

We began with two separate datasets, with one providing statistical data and the other providing contractual data. In order to examine the potential relationships between an NBA player's statistical performance and his contract details, we therefore combined these two datasets, such that any player with missing statistical data or missing contractual data was filtered out. In addition, we aggregated all statistics for players who played for multiple teams during the 2017-18 NBA regular season, so as to accurately reflect their performance throughout the entire season. (This included aggregate computations for points, rebounds, assists, different field goal percentages, etc.) Similarly, for players who were listed as having received multiple contracts, we combined their contract salaries so as to obtain these players' total guaranteed and non-guaranteed salaries. Finally, proper type conversions were completed, ensuring that we would not encounter unexpected floating point errors, misplaced empty strings, etc. while performing data analysis.

2 Dataset Overview

Overall, our preprocessed dataset focuses on 415 different NBA players, providing 1) all major 2017-18 regular season statistics, and 2) all contractual data over the 2018-19 through 2023-24 seasons for each player. More specifically, our dataset contains 36 distinct features for each player, such as points per game, total rebounds per game, assists per game, field goal percentage, total guaranteed salary, etc. This ensures we can flexibly measure a player's statistical performance in a variety of ways, such as by shooting efficiency (by looking at a player's different field goal percentages); defensive impact (by looking at steals and blocks per game); ball-handling ability (by computing players' assist-to-turnover ratios), etc. Additionally, all missing or corrupted data was removed or fixed during preprocessing (as already discussed in Section 1 above.)

Having obtained this preprocessed dataset, we began initial data analysis by exploring the distribution of values in regards to players' points, turnovers, rebounds, steals, blocks, and assists per game (see Figure 1 in appendix). The red dotted line in each histogram indicates the average value, whereas the blue dotted line indicates the median value.

From there, we also grouped players by their position, taking into account all five traditional positions: Center (C), Power Forward(PF), Small Forward (SF), Shooting Guard (SG), and Point Guard(PG). This allowed us to analyze how age and position affects any given player's guaranteed salary by creating a scatter plot ((Figure 2).) For example, we can see here how younger players tend to get higher guaranteed salary than older players.

To more clearly delineate any potential correlations between position and salary, we also visualized the average guaranteed salary for each position ((Figure 3).) From the bar graph, we can see how power forwards tend to receive lower guaranteed salaries than players of other positions, whereas point guards and centers positioned tend to receive the highest guaranteed salaries.

Finally, we also computed the average values for some of the dataset features ((Figure 4)), such as points per game, age, field goal percentage, etc.

3 Preliminary Analysis

Before fitting a model, we split the training and test data in a 60 percent to 40 percent ratio. The reason behind this is because we need to have a higher test set size due to the small number of columns in our data set while also making sure that the model is not simply regurgitating predictions learned from our training set.

We then use the least squares regression model to predict Guaranteed salary for each player using their points (PTS), rebounds (TRDBS), and assists (AST) statistics. Points, rebounds, and assists were chose as features as its expected that players with a higher number within these features are more likely to have a higher salary. Upon fitting a regression line to the data, we find that both the training and test data error is incredibly high ((Figure 5)). One reason for this could be that the number of labels we are using to predict the guaranteed salary is low and we have to introduce more features into our feature space to possibly gain a better prediction.

4 Conclusion

Based on the mean squared errors of our train and test model, we find that our linear model is likely overfitting given that our train error is notably lower than the test error. To address this issue, we plan to fit our linear model using additional features beyond just points, rebounds, and assists per game—with such additional features potentially including blocks per game (to account for defensive impact); minutes per game (to account for per-minute production); effective field goal percentage (to account for shooting efficiency); and age. More importantly, we also plan to investigate how guaranteed salary varies depending on player position, as our initial analysis fails to account for how positional differences may lead to drastic statistical and contractual differences. For example, centers have traditionally been expected to score less points, but contribute more defensively (in terms of rebounds, blocks, and steals) compared to guards; such positional discrepancies would therefore require us to at least partition players by position, and perhaps fit multiple linear models based off a different set of features for each player.

In addition to developing more nuanced linear models, we intend to explore other models and strategies that will help us decrease the mean squared error on both our testing and training data sets. Specifically, we aim to utilize the perceptron algorithm to test whether players of a given position are predicted to reach a certain salary threshold or not; we also intend to investigate feature engineering to reduce overfitting and different means of clustering to account for positional differences.

5 Appendix

Figure 1

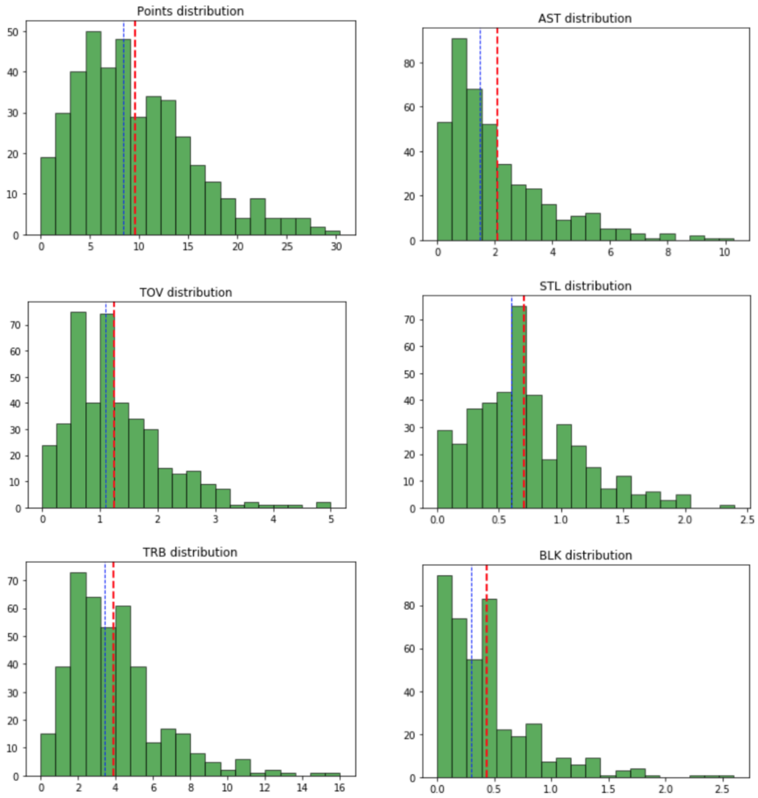


Figure 2

Scatter plot to show the relationship of players' age and position with their guarantees

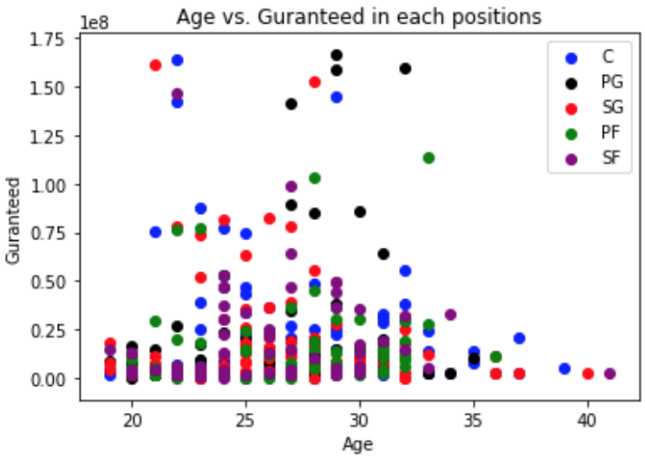
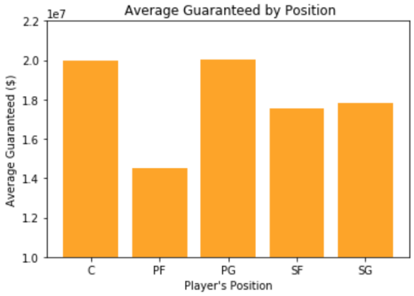
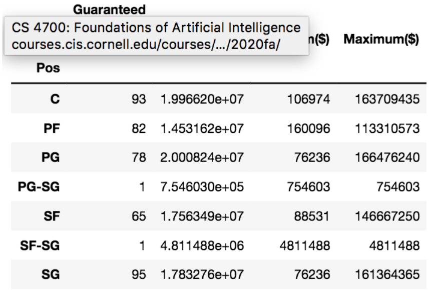


Figure 3

Bar graph and table to show the average, min and max of guarantees in each position groups





CS 4700: Foundations of Artificial Intelligence
courses.cis.cornell.edu/courses/.../2020fa/

	Guaranteed			
			n($)	Maximum($)
Pos				
C	93	1.996620e+07	106974	163709435
PF	82	1.453162e+07	160096	113310573
PG	78	2.000824e+07	76236	166476240
PG-SG	1	7.546030e+05	754603	754603
SF	65	1.756349e+07	88531	146667250
SF-SG	1	4.811488e+06	4811488	4811488
SG	95	1.783276e+07	76236	161364365

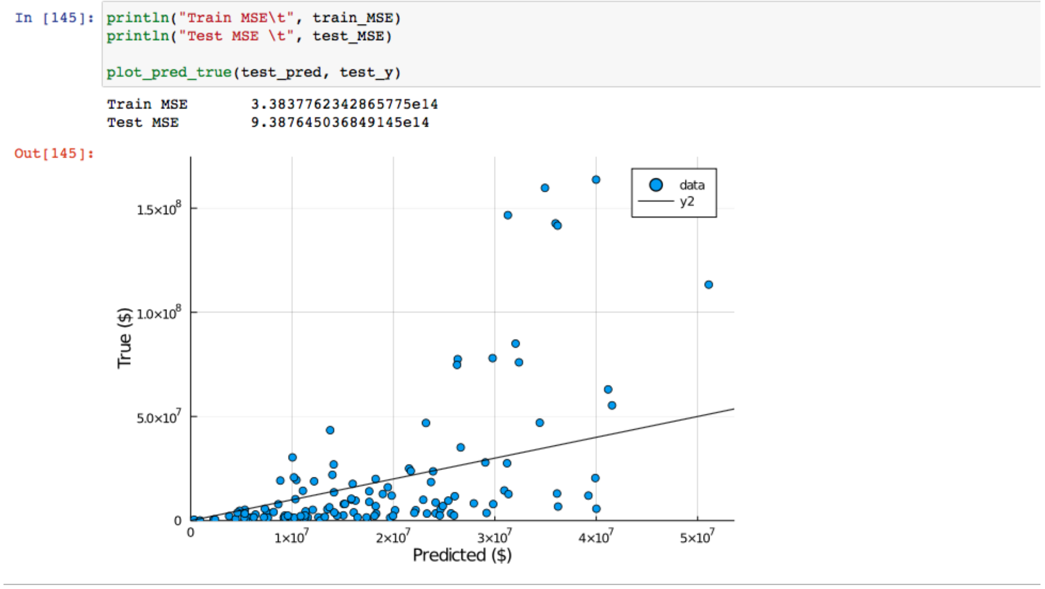
Figure 4

Computed averages for some dataset features

```
an average NBA players points per game 9.549277108433731
an average NBA player is 25.990361445783133 years old
an average NBA players FG% is 0.45539759036144545
an average NBA players effective field goal %: 0.512313253012048
an average NBA players total rebounds per game: 3.8899036144578325
an average NBA players assists per game: 2.0860481927710834
an average NBA players steals per game: 0.7014698795180727
an average NBA players blocks per game: 0.43337349397590347
an average NBA players turnovers per game: 1.2452530120481926
```

Figure 5

Regression line that is fit on the data as well as the MSE values for the training and testing data set.



```
In [145]: println("Train MSE\t", train_MSE)
          println("Test MSE \t", test_MSE)

          plot_pred_true(test_pred, test_y)

          Train MSE       3.3837762342865775e14
          Test MSE        9.387645036849145e14

Out[145]:
```