Mid-term Project Report

Jayesh Paunikar (jap452), Manas Koppar (mk2235), Trevor Xavier (tx28)

Data Description

The dataset fields used for training the models:

Game Attribute Fields	Physical Attribute Fields
Position Games Played Usage rate- the percentage of how long the player is handling the ball Turnover rate- The rate at which the player loses the ball FTA- Free throws attempted FT%- Success percentage of free throws 2PA- 2 pointers attempted 2P%- Success percentage of 2 pointers PPG- Points per game RPG- Rebounds per game APG- Assists per game SPG- Steals per game BPG- Blocks per game TOPG- Turnovers per game ORTG- Offensive rating DRTG- Defensive rating	Height Wingspan Age

Data Preprocessing

We began with a dataset from nbastats.com which had data related to Game Attribute Fields and scraped data from basketballreference.com to get the physical attributes. To get a full representation of all the features related to a player we combined these two datasets using vlook-up to form a final dataset with all the features required and used it for training the models. As of now, we are using just the previous season's data i.e 2019-2020.

From the dataset, we removed the entries with missing values in the feature TOPG. We also processed string fields like WingSpan to real numbers which are useful inputs for the models. Similarly, the position of each player was one hot encoded to enrich the dataset. Additionally, subtle changes were manually made to the values in the fields which were inconsistent with the rest of the values from the field. Our dataset has 626 rows and 34 features, out of which we explored the above features. We standardized the dataset, to transform the features to comparable scales. In total, we removed 31 entries that were missing important feature information as represented by the NaN in the dataset. On further analysis, by cross-referencing the dataset with (NBA.com), we found that there were some corrupted data points with respect to the height of the players that were corrected. However, this change was very minor.

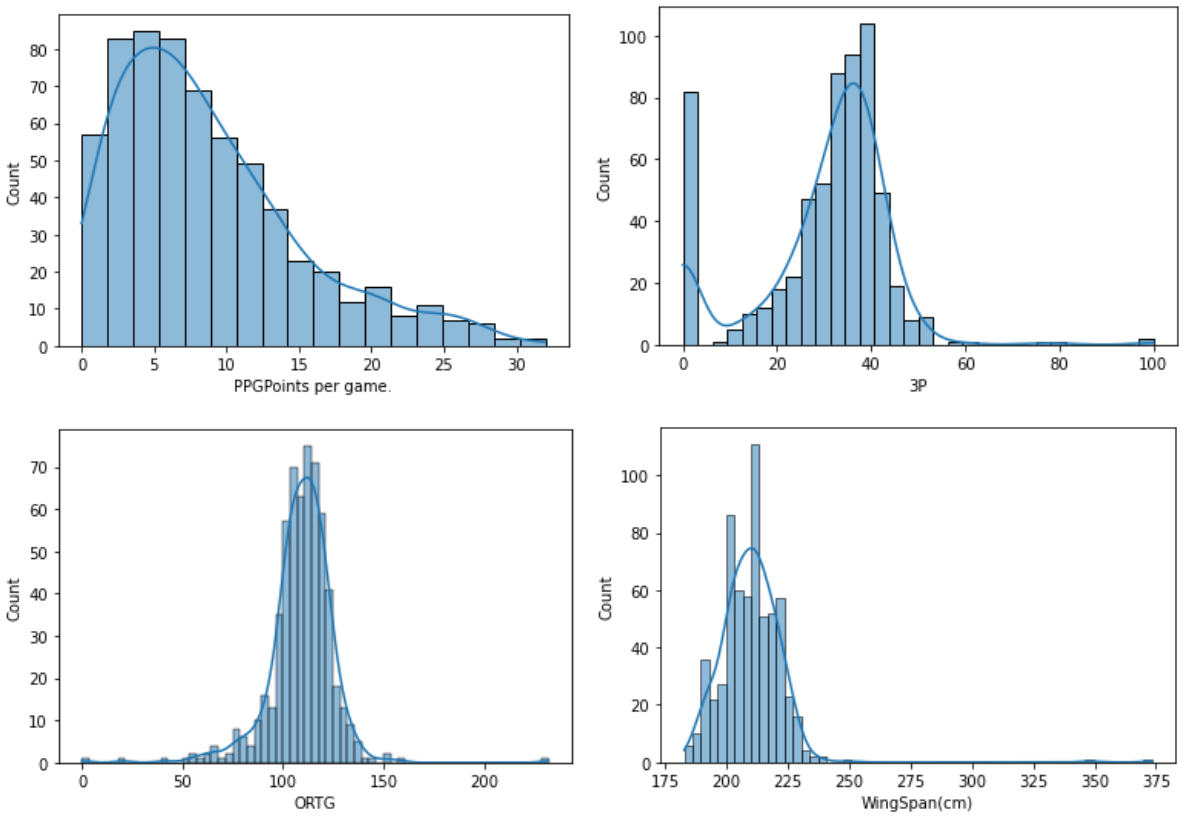
Avoiding Overfitting and Underfitting

To avoid overfitting, we trained our model by using fewer features to keep the model simple. Additionally, we have added L2 regularization and tested it on the validation set. We trained a cubic (order 3) linear model, to prevent further overfitting. We will be adding previous years' data as well to avoid underfitting.

Testing Effectiveness of the Model

To gauge the effectiveness of the model, we have split the data set into three parts, training, validation, and test. We have employed linear regularized, higher-order linear models and ControlBurn to predict the three-pointer efficiency of the validation set. The complete dataset was divided among these three sections in a 60,20,20 train, validate and test split and was shuffled prior to the split. The model with the least error on the validation set will be used for prediction on the test set.

Descriptive Dataset Statistics



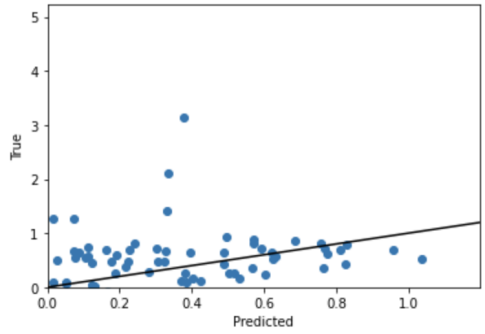
The average three-point percentage of the players during the NBA season was 29.68%. This was due to a massive outlier of 80 players who didn't play valuable minutes or shoot a three-pointer. Disregarding them, the average bumps up to 34.15% which is still lower

compared to the elite shooters who usually perform above 38%. The majority of NBA players were at the guard position as well. Finally, the ORTG(Total points achieved by the player per 100 individual possessions) was 109.08. The mean Wingspan of the players was 216cm. The mean

Preliminary Analysis

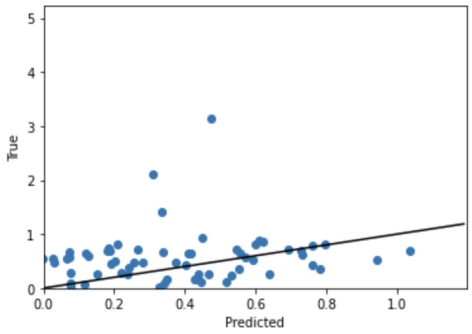
The team went about employing feature transformations. This included one-hot encoding of the Positions of each player. There were seven different positions encoded such as G, F, G-F, C, F-C, C-F. There were standard positions played in the NBA in the last decade, however, the game has become increasingly positionless giving rise to hybrid players who could play more than one position. Prior to this, we set a linear model with regression on the entire standardized training data that was then used to predict the 3P% on the test data. The same process was used when using the one-hot encoded position feature and used L2 regularization to reduce the Test MSE marginally. All these simple models are used to see the effectiveness of linear models on the dataset. The Train and test errors are small due to the standardized dataset, so they must be compared relative to scale.

```
Train MSE       0.5627285745359488
Test MSE        0.8328210020800751
```



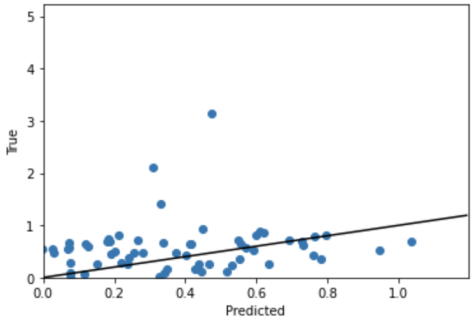
Regression w/o Positions one hot encoded

```
Train MSE       0.5547761818994067
Test MSE        0.8260719042854777
```

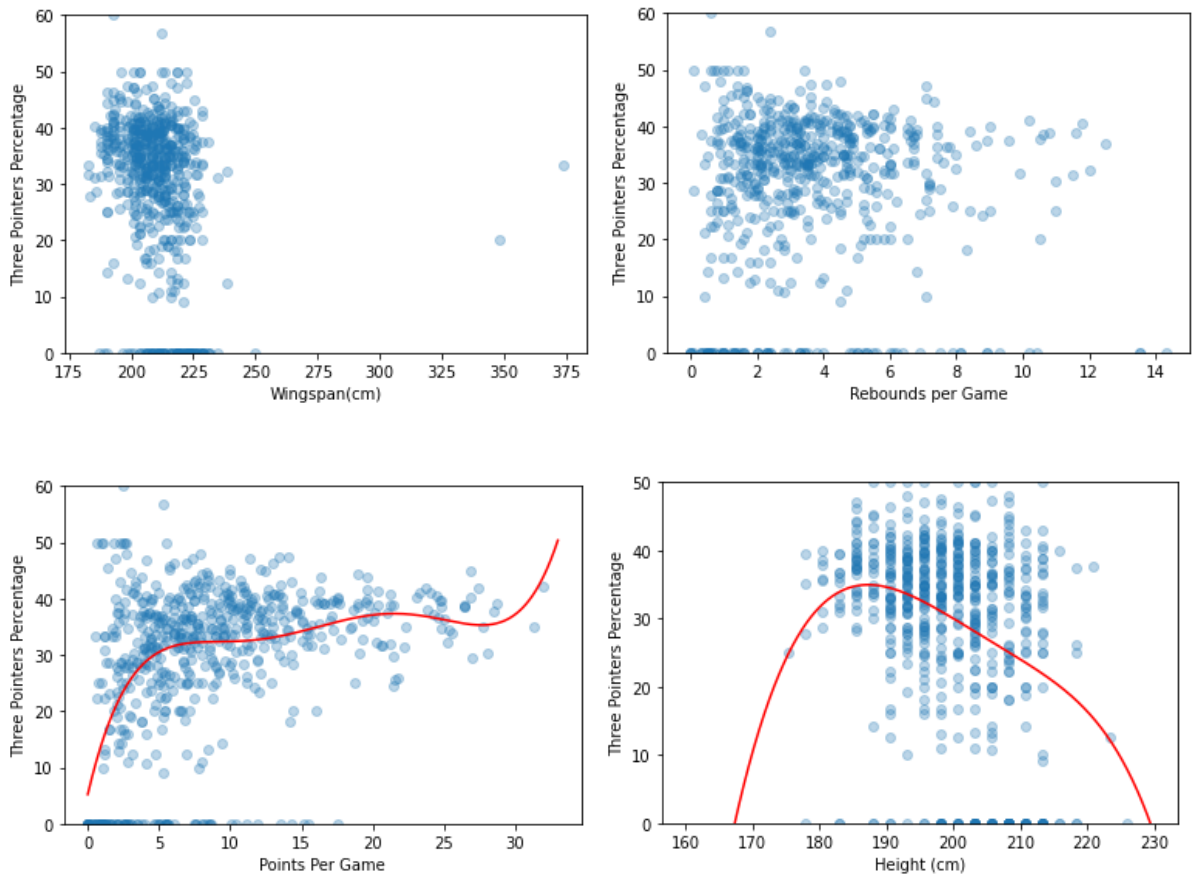


Regression with Positions one hot encoded

```
Test MSE is: 0.825900911071386
```



Ridge Regression



We looked at the relationship between the wingspan of an NBA player vs their 3P%, and the rebounds acquired per game vs 3P%. As seen above, players with a wingspan between 200 and 225 cm shoot the ball more efficiently nearing 40%(elite). This could be due to shorter players modifying their game to the perimeter(3P line) to score more points. This line of reasoning is given more footing as players who rebound the ball less tend to shoot better. Historically, shorter players (6 foot 5 or less) rebound the ball less. However, this is just a preliminary EDA.

We tried to establish which order of the feature value would help to get better results. For example, we tried to plot PPG vs 3P% and an order-5 curve seems to fit the data. So, using the order-5 feature value of PPG would help predict better. On further exploration, 3P% decreases with PPG which reconfirms that players who score more per game make a lot of 3-pointer attempts and miss because they are trying to grab every opportunity they get, whereas the players who score less per game attempt fewer 3-pointers and only attempt when they are very sure that the ball is going to go in, thus making their 3P% is a bit better.

Another example is Height(cm) vs 3P% and an order-5 curve seems to fit the data. And we can also conclude from the plot that shorter people have a better 3P%. The line of reasoning is similar to that of people with a shorter wingspan as these two are related, where the shorter

players have adjusted their playing style to score from far away so that their height is not a disadvantage.

Future Scope

We plan to include the data (from 2005-2019) and train ensemble tree models, ControlBurn, and neural networks on the dataset.
During the dataset analysis, we figured that with respect to each of the features, the distribution is over a region that we think would be better represented using trees.
Using neural nets, we can input all the features that we have into the neural networks, and the neural networks will tune the hyperparameters accordingly. Thus, making the step of formation of valuable features by feature transformation and feature selection redundant.