Determining Predictive Features of NBA Playoff Teams

Olivia Roy
University of Notre Dame
oroy@nd.edu

Tina Wu
University of Notre Dame
ywu6@nd.edu

Matthew Gregory
University of Notre Dame
mgregor4@nd.edu

Francis Schickel
University of Notre Dame
fschicke@nd.edu

1 INTRODUCTION

An NBA team's success can be measured by whether they are able to compete in the playoffs. The NBA playoffs consists of the top 8 teams of the East division and the top 8 teams for the West division totaling 16 teams each year. Franchises consistently seek to bring about the best possible team for any given year and team management will invest heavily in order to achieve the goal of reaching the playoffs. By examining the playoff teams for the past decade, we discovered that team statistics are helpful in predicting whether a team will successfully make the playoffs.

However, there is a multitude of data present in a single team's statistics. Due to the high dimensionality of data available, it can be very difficult to understand which data features are the most crucial factors that affects a team's chances of playoff competition. For example, a team holds a number of different dimensions of data including average points per game, average rebounds per game, salary cap, and so on.

The high dimensionality of the data causes an increase in the time and computing power needed to predict whether a team will make the playoffs. Additionally, various features of data may have very little to contribute to a team's playoff chances, and therefore should not be considered. In order to reduce the dimensionality of NBA data, we plan to use feature selection. Feature selection consists of testing data with different models in order to determine which features of data, or new classes of data, are useful for accurately predicting an output.

We sought to produce a smaller set of features by testing out different combinations of features with the following statistical prediction models: Gaussian Naive Bayes, CART Decision Trees, and Support Vector Machines (SVM). Our data consisted of all team data of all 30 NBA teams for every season from 2008-2009 to 2017-2018. The predictions are be binary, outputting either YES (make playoffs) or NO (do not make playoffs). Each model was created for a single season and used the actual NBA playoff lineup as the test case. F1 scores were used to compare each different model with each different combination.

Our work was split into two steps: 1) Hold-1-Out feature selection and 2) pruning and evaluation. Our results concluded that the Gaussian NaÃŕve Bayes model best determined a teamâĂŹs likelihood of making the playoffs with a reduced dimensionality of 10 features with an F1 score of 0.750. This is a 0.032 better F1 score than the best CART feature selection model with 6 features, and a 0.044 better F1 score than the SVC model with 7 features selected. The features we found most common between all models include 3-Point shots made, personal fouls, and steals. The success of Gaussian Naive Bayes is supported by such, as each of the three most common features also follow a Gaussian distribution.

2 RELATED WORK

Predicting whether the success within the NBA is a frequent study. Some studies include judging individual players to consider whether they will be named the MVP for the season using data science techniques such as Data Mining Discriminant[2] [5]. Other models go further than ours in order to predict the entire outcome of a seasons playoffs [4]. In order to do so, the model required feature selection in order more simply and clearly create a model of prediction. Additionally, the model utilizes prediction models such as Gaussian naive Bayes and SVM in order to predict outcome.

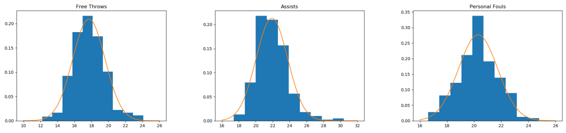
Additionally, much research has been conducted within the field of feature selection itself [3]. Research in this field is due to the high costs of high dimension data, and the cost and generalization benefit of feature selection.

ESPN itself has created a Basketball Power Index to predict a team's overall quality in a match up which can eventually be used to consider a team's number of wins and ability to play in the playoffs [1]. Such also only utilizes a few features in order to consider playoff contention, again providing the importance of feature selection.

With consideration for past models that utilized feature selection, we will also attempt feature selection considering team data and test using similar prediction models.

3 METHOD

We chose three classification models, Gaussian Naive Bayes, CART (decision tree), and SVC (Support Vector Classifier) to be used for our analysis. These models are deployed because most numerical features in the dataset have Gaussian distributions (as shown in graphs below), and the label for playoff prediction is categorical.



Because of the high dimensionality of features, we utilized the Hold-1-Out feature selection method. We first ran the three models on all features, calculating the F1 scores as base values for each. Then, we reran the models on all features excluding only one, and calculated the difference in F1 score, or the difference in performance of the models, after that particular feature is excluded. For example, the first iteration excluded the number of games played feature, the next iteration excluded the number of minutes played per season by each team, and the third iteration excluded the number of field goals; and the process continued. The goal was to see how removing a feature affects the F1 scores of the models. If the F1 score increased, then the model was more accurate when trained and tested without the particular feature. The more the F1 score decreased, the more insightful the feature was while predicting whether teams would make the playoffs or not in the next season. This method gives us insights on which features has the most (positive/negative) impact on the result, thus allowing us to explore more intensively on the better features.

We then used a pruning method to further analyze which combination of features yields the best results. After sort all average difference in F1 scores for the 3 models, we specified a K value that indicates how many features are included. For example, if k=1, we only use the best predicted feature as the training set; if k=2, we use the first two features as training sets; if k=15, we include all data as training set. We calculated new F1 scores for each k value (combination of features) and graphed them on scatter plots. Lastly, we analyzed the trends in the plots and reached conclusions of which combination of features is the most accurate in predicting whether a team makes the playoffs.

4 DATA AND EXPERIMENTS

The data we will use for our project is from Basketball Reference (https://www.basketball-reference.com/). The data consists of two datasets, which are game features of the 30 NBA teams (1) and their salary data (2). Starting with the 2008-2009 NBA season, these datasets give us 10 years worth of training data. The features captured in the data include common statistics such as points, assists, and defensive rebounds, as well as free throw attempts and personal fouls. These features are the average or per game statistics, so the total amount of that statistic divided by number of games played.

Our experimental methods mentioned in the previous section are considered reliable because of the data cleaning and preprocessing we performed. The data was able to be exported to Excel in csv file format directly through their website. Using GitHub, we were able to get the datasets directly onto the student machines where we then did an initial round of cleaning for the team statistics. Subsequently, we loaded around 20 datasets into Python and converted them all into a dataframe. There were two main preliminary preprocessing modifications we made. First, in the team statistics, each team that made the playoffs had a * after their name. Instead of having to parse this to determine which team made the playoffs, we simply added an additional column to the dataset that was a binary Y or N field to denote making the playoffs. This way, a column (Playoff) easily captures making the playoffs. Second, three teams during the 10 years worth of data we captured changed their name. The New Jersey Nets became the Brooklyn Nets, the Charlotte Bobcats became the Charlotte Hornets, and the New Orleans Hornets became the New Orleans Pelicans. The dataset did not account for these name changes, so we converted those former names to their current names to ensure consistency between years. Both of these modifications were made using Python scripts for all the datasets.

After the initial preprocessing, we had the most important preprocessing modification of removing redundant features. This was done so our model would not have an "incorrect" higher F1 score in the prediction. This is so because if statistic 1 contributes a lot to the prediction, another statistic 2 that includes statistic 1 should therefore also contribute similarly to the prediction. For example, in our team statistics we only selected 15 of the 23 features because most of the statistics were either similar to our team statistics or redundant. This is seen in including free throws made (FT) and free throw attempts (FTA), but not including FT percentage, which is just FT/FTA. Having redundant data would skew the model, so we had to remove these to get a true F1 score.

5 EVALUATION

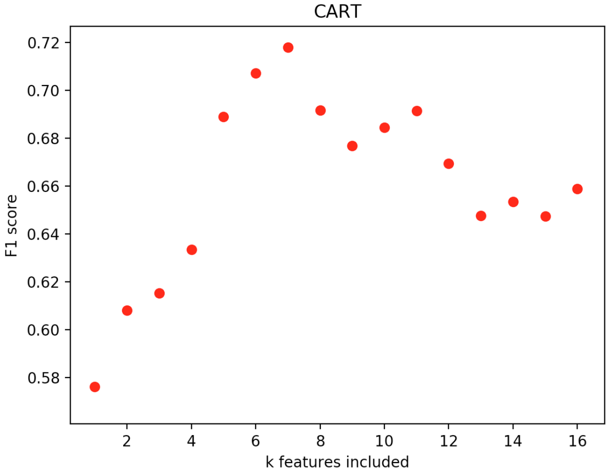
For each feature in team data, we evaluate the average F1 scores over nine years (09-10 season to 17-18 season). The F1 score for when no feature is excluded is subtracted from each F1 scores, giving us the result of how excluding each feature influences the performance of each model. An incomplete table displaying the data for five features is shown below as an example.

	None	FG	FGA	3P	3PA
9to10	0.82353	0.0098	-0.06677	-0.02353	-0.04575
10to11	0.73333	0.0	-0.04367	-0.02365	0.0
11to12	0.72727	-0.02139	-0.03977	0.0	0.0
12to13	0.8125	0.0	0.0	-0.02462	-0.02462
13to14	0.75676	-0.03454	-0.03454	-0.01992	-0.01992
14to15	0.6875	0.0	0.0625	-0.04044	-0.02083
15to16	0.7	-0.03333	0.03171	-0.03333	0.01795
16to17	0.78049	0.0	0.0	-0.03049	-0.01126
17to18	0.73333	-0.04367	0.0	0.0	0.0
Average Difference	0.0	-0.014	-0.01	-0.022	-0.012

The average F1 scores are ranked in ascending order, the lowest F1 score difference indicates the best feature for a certain model, and the highest F1 score difference indicates the worst feature for a certain model. If the F1 score difference is 0, it is an indication that the feature has little influence on the model's performance.

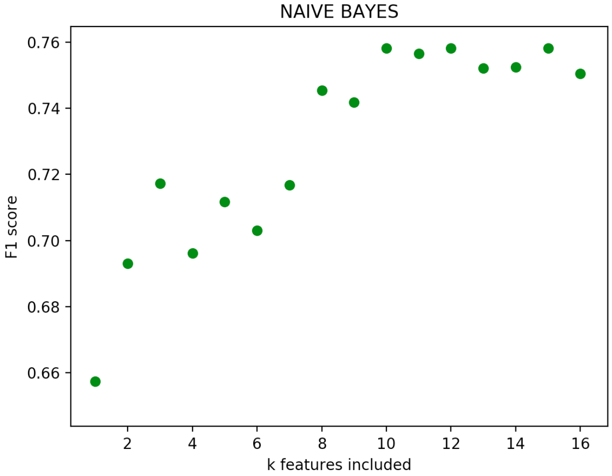
As previously mentioned, after sorting the F1 scores, as a result of Hold-1-Out, from best to worst, the models were used again to compute F1 scores for the top k features. The averages F1 scores were calculated, and plotted for each method below. To evaluate each model, their highest F1 score was compared to the original F1 score where all of the features are used.

CART Model



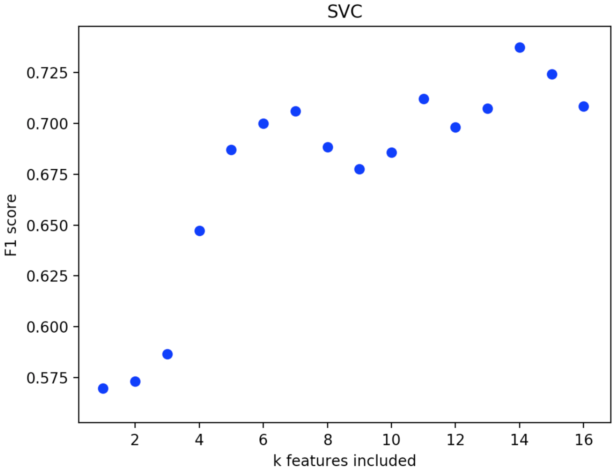
The F1 score when all 16 features are used is 0.673. The highest F1 score is 0.718, which is achieved when k=6. This is excellent because it is a 12% improvement, the most improvement of all three models. Additionally, the fact that the highest F1 score is when k is less than half, really narrows down which features are the most significant. The top six features were Defensive Rebound, Personal Fouls, 3-Point Attempts, Steals, 3-Point Field Goals, Blocks, and Turnovers.

Guassian Naive Bayes Model



The baseline F1 score, which includes all features, is 0.75. The highest F1 score is 0.7583 when k=12. This was only a 1% increase. While this increase was much smaller, the F1 score was still higher. The other difference is that k=12 is still over half of the features being used, which is less than ideal. However, the Guassian model is not close to a parabola. Instead, it seems to converge to about 0.75. If k needed to be reduced, will maintaining a high F1 score, an acceptable k value would be 10, having a F1 score of 0.7581. Again, this is only a 1% increase, but has $\frac{2}{3}$ of the original features. Even if we chose k to be 8 (half of the features), the F1 score would be 0.745. Although this is worse than the complete Gaussian model, it remains better than the F1 scores from CART. The top 10 features of this model were, in-order: Defensive Rebounds, Turnovers, Assists, Offensive Rebounds, 3-Point Field Goals, Personal Fouls, 2-Point Attempts, Steals, Free Throws, Field Goals, 3-Point Attempts, Field Goal Attempts, Points, Free Throw Attempts, 2-Point Field Goals, and Blocks.

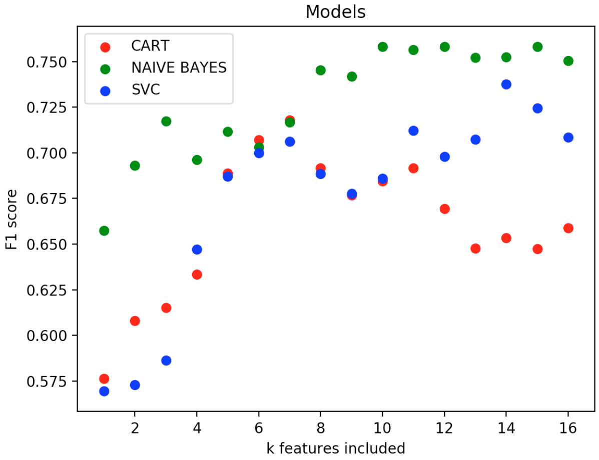
SVC Model



For the SVC model, the baseline F1 score is 0.708, while the highest achieved F1 score is 0.734 when k=14, having a 4% improvement. Immediately, it is seen from the graph that, unlike the previous two models, there is no clear pattern. Additionally if one goal is to lower k, only eliminating 2 features is not very effective. What we can do, is look at the first peak that happens at k=7 (about half of the features); it has a 0.706 F1 score, which is only a 0.3% decrease. This is not better, nor worse than including all of the features. The top 7 features were 3-Point Attempts, Points, 3-Point Field Goals, Offensive Rebounds, Field-Goal Attempts, Personal Fouls, and Steals.

Complete Comparison

While it is helpful to look at each graph individually, it is also helpful to overlay them to see their relationship to each other. In the graph below, it is clear that the Gaussian Naive Bayes consistently has the highest F1 score. The one exception is when k=6, where all the models actually preformed about the same; however they were using six different features based on the Hold-1-Out results.



6 CONCLUSION

Based on our evaluation of the three models, Gaussian Naive Bayes is the best model for several reasons. The first is that overall, regardless of k, the Gaussian Naive Bayes preforms better, as seen with the higher F1 scores. This could be attributed to the fact that the features, when graphed as a histogram, create a Gaussian curve. Additionally, because the model seems to converge at 0.75, the features 11-16 of that model add no value. This means it is acceptable to focus on just the first 10 features.

Asides from concluding which model worked better, we wished to determine the features that were most significant in determining if a team made it to playoffs. To do this, we looked at the similarity for the first k=8 features.

3 of 3 Models	2 of 3 Models	1 of 3 Models
3-Point Field Goals	3-Point Attempts	Field Goal Attempts
Personal Fouls	Points	2-Point Field Goals
Steals	Offensive Rebounds	Assists
	Defensive Rebounds	2-Point Attempts
	Turnovers	Blocks

Thus, if we were to recommend coaches, gamblers, or other NBA sports fans, what statistics are most predictive of a team making it to the NBA playoffs, based on the previous year, we suggest examining 3-Point Field Goals, Personal Fouls, and Steals. Of course, for the best accuracy using the Gaussian Naive Bayes model with the 10 features listed in the previous section, would yield the best results.

7 FUTURE WORK

An extension of this work, would be to include additional features, such as individual player data. Additionally, it would be interesting to see how each of these models preforms with other sports that have similar features, such as football, in attempt to make a more generalized model.

REFERENCES

[1] 2018. ESPN's Basketball Power Index. http://www.espn.com/nba/story/_/page/Basketball-Power-Index/espn-nba-basketball-power-index. Accessed: 2018-10-01.

[2] Mason Chen. 2017. Predict NBA Regular Season MVP Winner. (Oct. 2017). http://ieomsociety.org/bogota2017/papers/9.pdf

[3] Feiping; Li Xuelong; Yi Dongyun; Wu Yi Hou, Chenping; Nie. 2014. Joint Embedding Learning and Sparse Regression: A Framework for Unsupervised Feature Selection. In *IEEE Transactions on Cybernetics*. https://doi.org/10.1109/TCYB.2013.2272642

[4] Logan; Sundaresan Vishnu Lin, Jasper; Short. 2014. Predicting National Basketball Association Winners. (2014). http://cs229.stanford.edu/proj2014/Jasper%20Lin,%20Logan%20Short,%20Vishnu%20Sundaresan,%20Predicting%20National%20Basketball%20Association%20Game%20Winners.pdf

[5] Li K. TallÃşn-Ballesteros, A.J. (Ed.). 2017. *Fuzzy Systems and Data Mining III: Proceedins of FSDM 2017.* Vol. 299.