Predicting the NBA Rookie of the Year with Machine Learning (December 2019)

Joshua Mathew and Tucker Howard

Abstract - In this paper we will examine how machine learning can be used to predict the Rookie of the Year for the 2019-2020 NBA season. Three separate methods were used, logistic regression, k-nearest neighbors, and neural network algorithms. The ROY winners from the seasons between 2001-2019 were used as the training dataset for the system and the accuracy of each machine learning method were tested on the ROY winners from 1990-2000. All statistics used in this system were scraped from a well-known website, Basketball Reference [1], and a total of sixteen predictors were found to be significantly correlated with winning the ROY award. All three models were fairly accurate, and all predicted Ja Morant to win the ROY honor at this point in the season. The neural network algorithm performed the best of the three models. The model had the highest TPR of 75%, correctly predicted 91% of the ROY's from 1990 to 2000, and had the highest overall accuracy of 97% on the validation set.

I. INTRODUCTION

This paper will outline the methodology and results of using machine learning algorithms to predict the 2020 Rookie of the Year (ROY) for the National Basketball Association (NBA). It must be noted that a rookie is defined as a player in their first season of the NBA and we are using current data from a third of the way through the 2019-2020 season to predict the ROY for this season. We examined previous ROY winners using the detailed statistics provided from Basketball Reference [1], a well-known website documenting an extremely wide range of statistics for each player previously or currently in the NBA. We scraped all the data from Basketball Reference for each of the previous ROY winners from 2001-2019 to determine which statistics were highly correlated with winning and those which were not. This chart can be seen in Fig. 1.

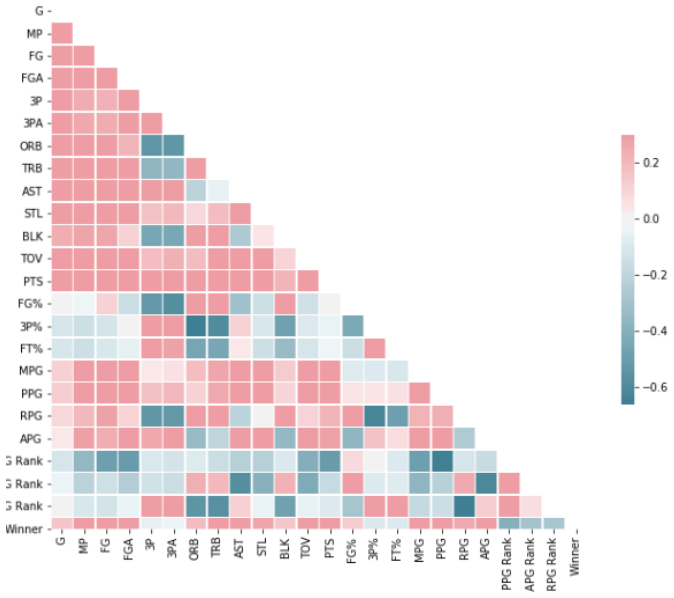

Fig. 1 - Correlation Matrix

Examining the chart, it is evident that certain statistics should be excluded from the data because they are either highly correlated with another statistic or are poorly correlated with winning the ROY. Variables least correlated with winning include three-pointers made (3P), three-pointers attempted (3PA), and three-point percentage (3P%). Total rebounds per game (TRB), rebounds per game (RPG), and offensive rebounds per game (ORB) were correlated with one another and therefore only RPG were included in the dataset for predicting ROY. Minutes played (MP) and games played (G) were also highly correlated with one another and therefore only MP was included in the dataset. We also created three variables: RPG rank, assist per game (APG) rank, and points per game (PPG) rank. These statistics were added as a way to gauge how a rookie compared against their individual class which is information that would otherwise be lost when all player data is combined for training and validation.

We used a recursive feature elimination algorithm (RFE) to eliminate predictors that were not useful to the outcome of the dataset. These variables were field goals attempted (FGA), ORB, and steals per game (STL). It is also worth noting that any rookie who averaged less than 10 PPG was removed from the dataset because no ROY has ever averaged less than 10 PPG. The final set of predictors for the models were determined to be: MP, field goals per game (FG), RPG, assists per game (AST), blocks per game (BLK), turnovers per game (TOV), points per game (PTS), field goal percentage (FG%), FT%, minutes per game (MPG), PPG, APG, PPG Rank, APG Rank, RPG Rank, and TRB.

Once we established the most effective predictors, the dataset of ROY winners from 2001-2019 was randomly split into separate training and validation datasets. We then used three different methods: logistic regression (LR), k-nearest neighbor (KNN) and neural networks (NN) to fit the dataset. Each model was trained on the training data and tested on each validation dataset.

The overall model accuracy on the test data is not a great indicator of how good the model is because there is an unequal distribution of the two classes, there are many more non-ROYs in the data than ROYs. Because of this we are more concerned with the true positive rate (TPR) which is the proportion of actual ROYs that were correctly predicted to be ROY.

Selecting the correct ROY's from a large group of players containing multiple ROY's from many different NBA seasons is difficult. A more realistic test of each model is feeding several seasons, that were not used for training, into each model and selecting the rookie with the highest assigned probability as the predicted ROY for each season. The data for each rookie class from 1990 to 2000 were fed into the models and the ROYs for those years were predicted using this method. The accuracy of these predictions is a better indicator of how well the models performed. All algorithms were implemented in Python using the NumPy, Pandas, Seaborn, Scikit-learn, Tensorflow, and Keras libraries.

II. Logistic Regression

The first model we created was a logistic regression (LR) model. This model had an overall 95% accuracy on the test data.

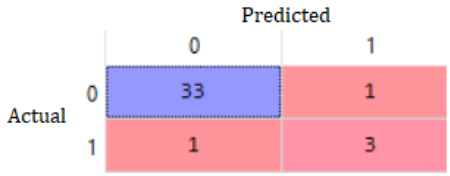

Fig. 2 - Confusion Matrix for LR

Based on Fig. 2, the model correctly predicted 33 players that were not ROY, 3 players correctly that were ROY, 1 player was predicted incorrectly to be ROY, and 1 player was not predicted to be ROY that actually was. The true positive rate (TPR) was ¾ or 75% and the false positive rate (FPR) was 1/34 or 2.9%. Fig. 3 below is a table displaying the ROY's that the LR model predicted for the 1990-2000 season compared to the actual ROY winners.

Index	Actual	Predicted
1990	David Robinson*	David Robinson*
1991	Derrick Coleman	Lionel Simmons
1992	Larry Johnson	Larry Johnson
1993	Shaquille O'Neal*	Shaquille O'Neal*
1994	Chris Webber	Chris Webber
1995	Grant Hill*	Grant Hill*
1996	Damon Stoudamire	Damon Stoudamire
1997	Allen Iverson*	Antoine Walker
1998	Tim Duncan	Tim Duncan
1999	Vince Carter	Vince Carter
2000	Elton Brand	Elton Brand

Fig. 3 - Actual vs Predicted ROY Winners with LR

The LR model correctly predicted the ROY winner during these 11 seasons all but two times for a total of 81.8% accuracy. Fig. 4 displays the weights the LR model assigned to each predictor. PPG Rank was the most influencing variable towards winning followed by APG Rank. Surprisingly, rebounding was negatively correlated with winning. This may be because rebounding and assists are negatively correlated.

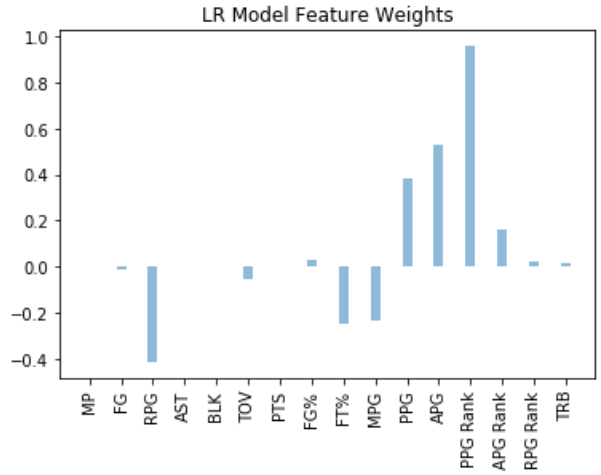

Fig. 4 - Predictor Weights for LR Model

The LR model was then fed the data for the current 2019-2020 rookies in the NBA and the results conclude that Ja Morant has the greatest chance to win ROY. The results can be seen below in Fig. 4.

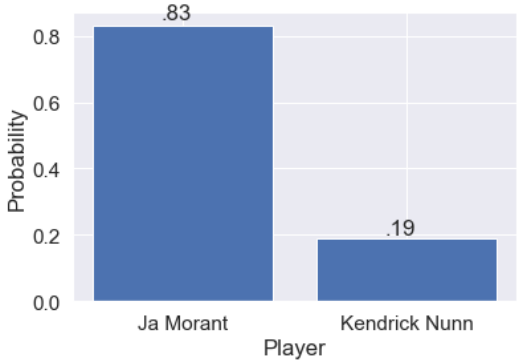

Fig. 5 - LR Prediction of 2020 ROY Winner

III. K-Nearest Neighbors

The second model used was the k-nearest neighbors algorithm. The data was first normalized, and the number of features was reduced from 16 to 9 using principal component analysis (PCA) to retain 95% variance before training and testing the model.

The model was tested using k values ranging from 1 to 40. A k value of 5 resulted in the model with the lowest mean error. This model had an overall accuracy of 89% on the test data.

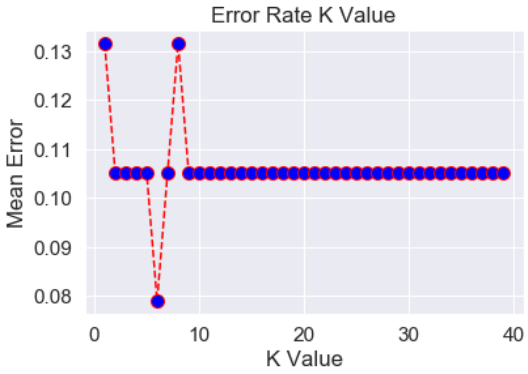

Fig. 6 - Model Error vs. K

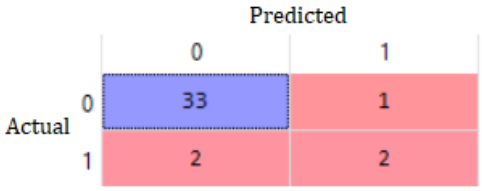

Fig. 7 - Confusion Matrix for KNN

Examining Fig. 7, the model correctly predicted 33 players that were not ROY, 2 players correctly that were ROY, 1 player was predicted incorrectly to be ROY, and 2 players not to be ROY that actually were. The true positive rate (TPR) was 50% and the false positive rate (FPR) was 2.9%.

The KNN model correctly predicted the ROY from 1990-2000 72.7% of the time only missing 3 out of 11. This is slightly worse than the LR method, but when given the current 2020 player data, the KNN model also predicted Ja Morant to win the ROY.

Index	Actual	Predicted
1990	David Robinson*	Tim Hardaway
1991	Derrick Coleman	Derrick Coleman
1992	Larry Johnson	Larry Johnson
1993	Shaquille O'Neal*	Christian Laettner
1994	Chris Webber	Anfernee Hardaway
1995	Grant Hill*	Grant Hill*
1996	Damon Stoudamire	Damon Stoudamire
1997	Allen Iverson*	Allen Iverson*
1998	Tim Duncan	Tim Duncan
1999	Vince Carter	Vince Carter
2000	Elton Brand	Elton Brand

Fig. 8 - Actual vs Predicted ROY Winners with KNN

IV. Neural Network Model

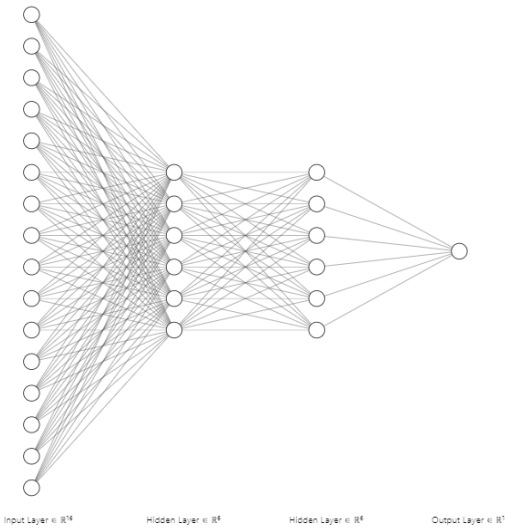

Fig. 9 - Neural Network Architecture

The final model used was a neural network (NN). The neural network was created with 2 hidden layers containing 6 nodes each. The ReLu activation function was used for the hidden layers. Because the output is binary, the sigmoid function was used for the output layer. The structure of our NN can be seen in Fig. 9.

The data was again normalized before training and testing. The neural network was trained using a batch size of 10 over 250 epochs. The cross-entropy loss function and Adam optimization algorithm were employed. The loss of the neural network over the training period can be seen in Fig. 10.

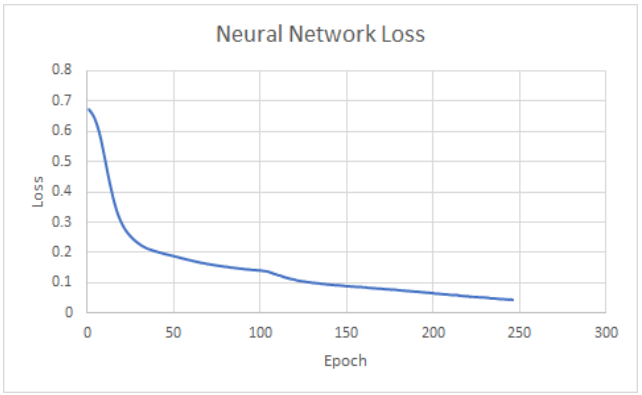

Fig. 10 - NN Loss

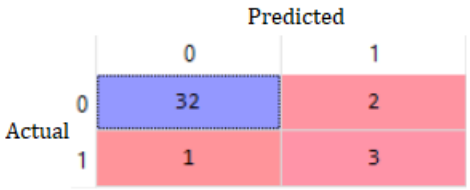

Fig. 11 - Confusion Matrix for NN

Based on Fig. 11, the model correctly predicted 32 players that were not ROY, 3 players correctly that were ROY, 2

players were predicted incorrectly to be ROY, and 1 player not to be ROY that actually was. The true positive rate (TPR) was 75% and the false positive rate (FPR) was 5.9%. The model achieved an overall accuracy of 97% on the test data.

Index	Actual	Predicted
1990	David Robinson*	David Robinson*
1991	Derrick Coleman	Derrick Coleman
1992	Larry Johnson	Larry Johnson
1993	Shaquille O'Neal*	Shaquille O'Neal*
1994	Chris Webber	Chris Webber
1995	Grant Hill*	Grant Hill*
1996	Damon Stoudamire	Damon Stoudamire
1997	Allen Iverson*	Shareef Abdur-Rahim
1998	Tim Duncan	Tim Duncan
1999	Vince Carter	Vince Carter
2000	Elton Brand	Elton Brand

Fig. 12 - Actual vs Predicted ROY Winners with NN

Examining Fig. 12, the neural network model correctly predicted the ROY winner 10/11 seasons for a total of 90.9% accuracy. Fig. 13 displays the predictor weights of the model. Stats associated with assists were the biggest predictors. Surprisingly turnovers were positively associated with winning even though turning the ball over is bad. This may be because turnovers are positively correlated with points and assists.

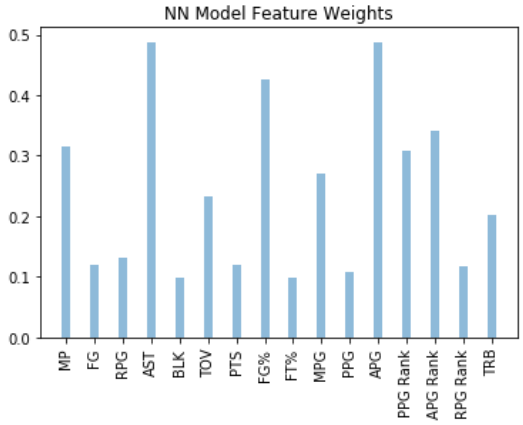

Fig. 13 - Predictors Weights for NN Model

After plugging in the 2019-2020 rookie data into the model Ja Morant was again predicted to be the 2020 ROY.

V. Conclusion

All the models predicted Ja Morant to be the 2020 NBA Rookie of the Year. This is a reasonable prediction as Ja Morant is currently the consensus best performing rookie in the NBA. All models show a wide gap in winning probability between Morant and any other rookie.

	NN	LR	KNN
Validation Set TPR	75%	75%	50%
Correct Prediction Rate 1990-2000	91%	82%	73%
Overall Validation Set Accuracy	97%	95%	89%

Fig. 14 - Model Comparison

Overall the neural network was the best performing model. It was tied with the linear regression model for highest TPR, correctly predicted the most ROY's from 1990 to 2000, and had the highest overall accuracy on the validation set.

The linear regression was the second best performing model followed by k-nearest neighbors. All models performed significantly better than guessing. For example, if only looking at players who have a reasonable chance of winning (>10ppg) there are generally between 3-10 potential winners. Guessing will result in a 10-33% chance of selecting the correct ROY. All the models were much more accurate than this.

This project is different from the cited work related to machine learning and the NBA because other models have only tried to predict the NBA MVP or rookie stats but not whether they will win the rookie of the year award. Rookie of the year is the highest honor that a first year NBA player can receive so it is interesting to try and predict who will win only a third of the way into the current NBA season.

VI. References

[1] Basketball Statistics and History," *Basketball Reference*. [Online]. Available: https://www.basketball-reference.com/. [Accessed: 27-Nov-2019].W.-K. Chen, *Linear Networks and Systems*. Belmont, CA, USA: Wadsworth, 1993, pp. 123–135.

[2] D. Bratulić "Predicting 2018–19 NBA's Most Valuable Player using Machine Learning," *Medium*, 15-May-2019. [Online]. Available: https://towardsdatascience.com/predicting-2018-19-nbas-most-valuable-player-using-machine-learning-512e577032e3. [Accessed: 17-Dec-2019].

[3] S. Kannan, "Predicting NBA Rookie Stats with Machine Learning," *Medium*, 30-Jun-2019. [Online]. Available: https://towardsdatascience.com/predicting-nba-rookie-stats-with-machine-learning-28621e49b8a4. [Accessed: 17-Dec-2019].

[1] "