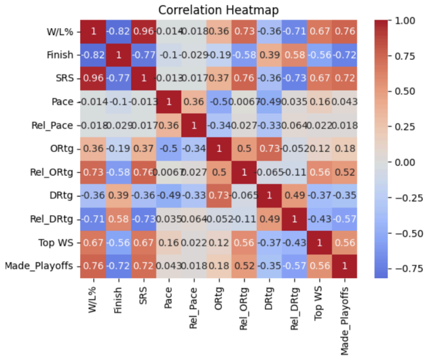
Jayden Nguyen, Andi Nguyen, Johnathan Pham
CSCI 5523
02 Dec 2025

Predicting NBA Playoff Qualification Using Machine Learning Models

The National Basketball Association, more popularly known as the NBA, is one of the most globally followed professional sports leagues with millions of fans, extensive media coverage, and substantial financial investments. Every year, organizations in the NBA pour money and resources into being the standalone winner; the champion of the NBA. Playoff qualification in the NBA carries significant implications for an organization's financial outlook. Players' contract, ticket pricing, and staffing decisions are all impacted by the playoff potential of a team. As analytics continue their exponential growth across professional sports, data-driven decision making has become essential and common across numerous leagues like the NBA. The aim of this project is to explore whether regular-season team statistics can be used to accurately predict qualification in the NBA. As explained, the playoff qualifications are impactful and knowing if a team is in position to make the playoffs or not can help teams make decisions earlier, saving them time and money. In addition to the teams themselves, these problems can help fans who invest their own time and money to support their teams. All in all, developing an accurate predictive system can support analytical decision making in the hands of teams, leagues, and fans. With a simple foundational model, it can help future modelers understand the effects and importance of every aspect of the model when computing a more advanced modeling system.

The project used the comprehensive dataset titled "NBA Season Records from Every Year" from Kaggle. The dataset created by Boon P contained team statistics across decades of NBA history with each record being an annual team. Features of the set includes wins, losses, offensive rating, defensive rating, pace, win shares, Simple Rating System (SRS) and more. These statistics are all potentially correlated to the playoff outlook of a team and studying the correlation between the two will be important in understanding the cause of a model's resulting metrics. Before fitting the model to the data, some cleaning and visualizations had to be done to the dataset to make it effective with the model and to help us understand the model's performance. The dataset contained categorical data, missing values, and unwanted records immediately after download. To fix this, the dataset was first stripped of non-NBA league teams since we want to study the effects in the NBA and not other leagues where statistics are different. Next, a playoff category was created and marked as binary where teams that made it to the playoffs were marked with 1 and non-playoff teams were marked with 0. This matches the goal of this project and removes the playoff stage of a team which is what was originally present. Finally, the irrelevant and redundant features were dropped including the old playoff category, team name, season, coach, and more. These features served no help in identifying the playoff position of a team and would be unnecessary complexity that is not wanted.

With the clean dataset, visualizations were created to help understand the possible feature correlation and possible issues in the models. First, a histogram comparing playoff and non-playoff teams in the dataset was created. As a result, we knew to balance the dataset when it came to fitting our model to it. Another visualization tool created was a correlation heatmap and a pairplot. The two can help us see the correlation between playoffs and features. The heatmap gives us a numerical and color value on the correlation index between every feature but the most important one was every feature against playoffs.



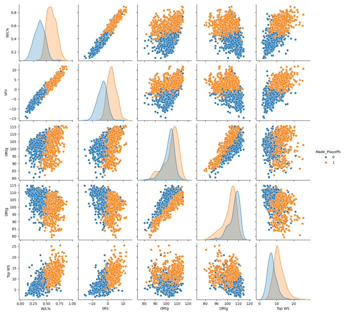


Figure 1.

Figure 2.

The diagram shown in figure 1 can point to most of the features having some degree of correlation to playoffs. Taking these correlation values, decisions on model changes were made later on. The other diagram created is shown in Figure 2 and shows the correlation between 2 feature interaction and playoffs. Together, these diagrams helped us understand and make future decisions concerning our models and the fitted features.

Three methods of classification were used in the project. Logistic regression, support vector machine (SVM), and k-nearest neighbor (KNN). These models were chosen because each classification method and their resulting metrics can be compared to understand patterns in the features that relate to the playoffs. Looking at these problems can help determine if the playoff qualification is a linear, nonlinear, or local similarity trend. KNN was included to serve as a simple baseline that relies on distance between data points. Because this is such a simple model, it can be used to see if the other two models actually improves classification above a fixed model structure. Logistic regression was selected due to its well known performance on classifying binary data as well as its interpretability property. The coefficients produced by logistic regression can be used to help understand the features importance and how each feature affects a team's playoff chances. Finally, SVM was chosen because of its performance in high dimensional datasets and its performance in nonlinear boundary conditions. As we saw in the scatterplots, the feature interaction compared with playoff outcomes shows clear boundaries between playoff and non-playoff teams. This follows the strengths of SVMs where it performs well on datasets where there are clear boundaries between classes. The radial basis function (RBF) kernel was used in the SVM model. These three models together provide good insight into the patterns of an NBA playoff team's performance data.

While ensuring proper evaluation and avoiding data leakage, the project used a 70-30 train-test split where 70 percent of the dataset was reserved for model training and the remaining 30 percent was used purely for the testing and evaluation. Stratification was applied to the split to preserve the proportion of playoff and non-playoff teams in both subsets. This makes sure that the test set plays no role in model fitting or tuning and provides an unbiased estimate of the generalization performance. Normalization was then applied to the numerical data and was computed using only the training data and then applied to the test data. This ensures that the test data does not influence the learned scaling parameters, preventing data leakage. Z-score normalization was the method used and is important as it prevents some features with larger numerical values from dominating the training as much of NBA statistics are on different scales. With the training portion of the data, hyperparameter tuning and model evaluation was then carried out using 5-fold cross validation on each model. The use of cross validation allowed each model to be trained and evaluated on multiple subsets of the training data which makes the estimate of generalization performance more robust. The logistic regression and SVM models had computed metrics of accuracy, precision, recall, AUC, and F1-score, averaging among the folds. Note that all models were fitted onto weighted versions of the dataset to fix the slight imbalance the dataset has between playoff and non-playoff teams. The KNN model had an elbow curve created and ran on multiple validation sets in order to find the appropriate k value to use in the model. After the optimal k value was found, it was used to test the model in the same manner as logistic regression and SVM. On top of the performance metrics mentioned, a ROC curve was also created. This process of evaluation ensured a consistent and true generalization of the models.

The KNN model, which was used as a baseline, outputted the elbow curve seen in Figure 3. The k was chosen by looking at the point of decreasing improvement as it will produce a good accuracy while avoiding underfitting. Based on this analysis, the optimal value of k was determined to be 11. The model produced an accuracy of 0.9108, a precision of 0.9091, a recall of 0.9383, an F1-score of 0.9229, and an AUC of 0.9649. The resulting ROC curve can also be seen in Figure 4. The results show that KNN is capable of identifying playoff teams but still got outscored by the other two models.

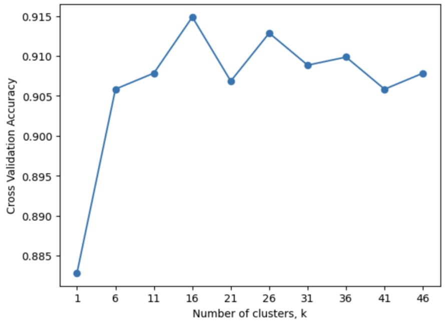
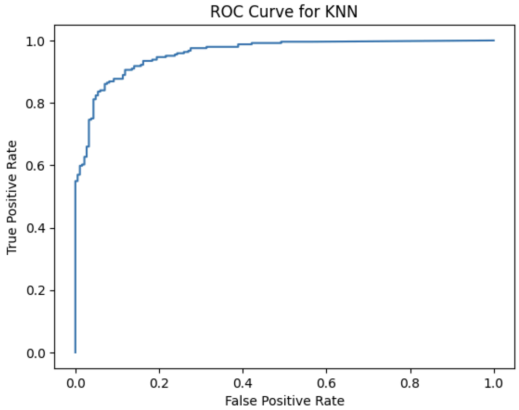


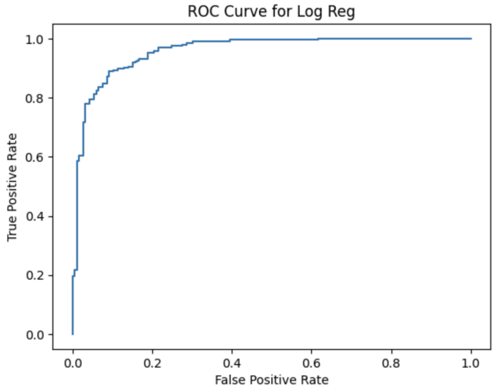
Figure 3.

Figure 4.



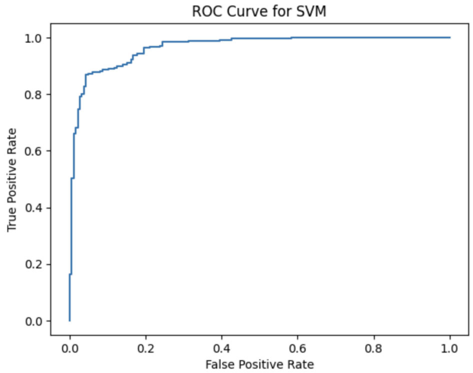
The logistic regression model also achieved strong results and slightly outperformed the KNN model. The logistic regression model produced an accuracy of 0.9139, a precision of 0.9305, a recall of 0.9171, an F1-score of 0.9236, and an AUC of 0.9689. Compared to KNN, almost all of the metrics slightly outperformed KNN with the exception of recall. As with the coefficients that the model produced, we saw similar results to the correlation heatmap where win percentage, finish ranking, and SRS were the features with the largest coefficient magnitudes at 2.5825, -1.1826, and -1.0301, respectively. The absolute values of the coefficients ranged from 0.0477 to 2.5825. The absolute value of the correlation is the value of interest as the positive or negative correlation in this case is not the study of interest. The resulting ROC curve for the logistic regression model can also be seen in Figure 5.

Figure 5.



The support vector machine with radial basis function kernel produced the highest achieving performance out of the three by a minimal margin. The SVM model produced an accuracy of 0.9239, a precision of 0.9365, a recall of 0.9295, an F1-score of 0.9328, and an AUC of 0.9721. The results are a strong balance between all measures and show that it is the best overall classifier between the three when all features are available. The ROC curve for the SVM model is shown in Figure 6.

Figure 6.



After analyzing the results of the first run of the models, the models seem to be heavily relying on the strongest associated features like win percentage, finish ranking, and SRS. This can be seen as true when studying the feature coefficients produced in the logistic regression model. While these variables are valid predictors, they are also extremely correlated with the playoff position. This makes sense as the highest win percentage teams are always going to be towards the lowest finishing positions and therefore into the playoffs. As a result, a second test of the models was orchestrated where the top three features mentioned before were to be removed. The goal of this second test was not to see improvements in the score but to see whether each model can still produce similar results given a group of less correlated features. This allows for a more meaningful comparison of the models and their effectiveness. The feature was reimported and the same data cleaning was applied with win percentage, finish ranking, and SRS added to the list of dropped features.

The KNN's elbow curve was reproduced and produced the curve seen in Figure 7. The k chosen in this run is the same as the first run although the graph is different. With k = 11, the 5-fold cross validation was run to compute the metrics. The KNN trial resulted in an accuracy of 0.8617, a precision of 0.8552, a recall of 0.9136, an F1-score of 0.8826, and an AUC of 0.9388. As we can see, the performance declined all metrics with the biggest dip in accuracy and precision. Recall and AUC was able to stay relatively high compared to the other metrics. The ROC curve for this run can also be seen in Figure 8.

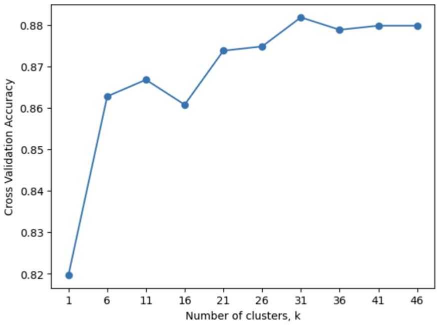


Figure 7.

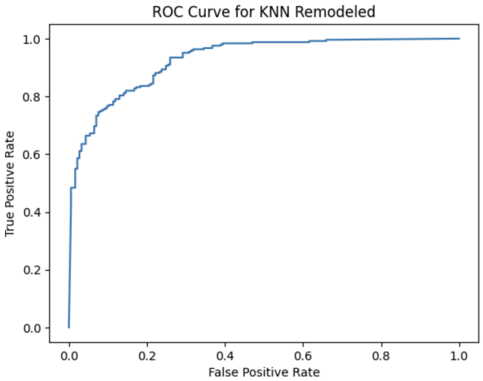
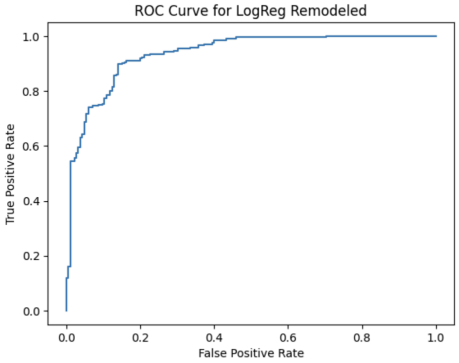


Figure 8.

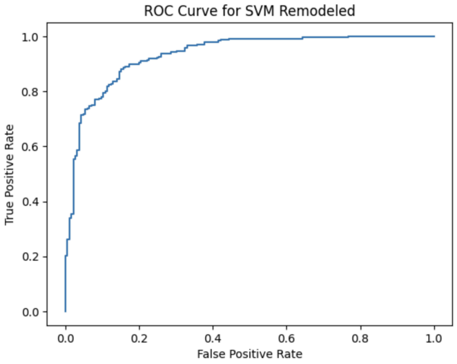
The logistic regression also saw a dip in performance with the new feature set. The new logistic regression run produced an accuracy of 0.8818, a precision of 0.9108, a recall of 0.8785, an F1-score of 0.8938, and an AUC of 0.9506. Once again, while the metrics were lower than the first run with the superset of features, the model remained competitive. The coefficients also shifted with most features gaining more correlation. Relative offensive rating, relative defensive rating, and defensive rating were the top correlated features this time around with an absolute value correlation value of 1.6700, 1.8972, and 1.0097, respectively. The shift indicates that in the absence of win percentage, SRS, and finish, the model relied on these features more. The ROC curve for this model can be seen in Figure 9.

Figure 9.



The last rerun was with SVM and the pattern of decreasing metrics continued. The subset of features fitted onto the SVM model produced an accuracy of 0.8868, a precision of 0.9013, a recall of 0.8996, an F1-score of 0.9001, and a AUC of 0.9529. The resulting ROC curve can also be seen in Figure 10. The SVM remained the highest performing model for the subset features despite also losing performance with the smaller feature size.

Figure 10.



The three different models on two different feature sets showed meaningful patterns about NBA playoff qualifications. With KNN, the features cause playoff teams to cluster tightly and become detectable with KNN's distance-based detection. However, when run on the second feature set where the highest correlated features are removed, the structure of the space becomes less clear, causing the steeper drop in performance compared to the other models. These results suggest that the baseline model can find distinguished clusters but can struggle more when separations decrease and correlation in the features decrease.

The logistic regression model consistently outperformed KNN and demonstrated greater robustness across both trials. As we saw in the first trial, the logistic regression model benefits from strong linear relationships as the model is based on the linear combination of the features.

The model is particularly effective when features with high correlation like win percentage, finish, and SRS exist. The coefficients from the first trial confirm that the model depended heavily on these features. Although those features were removed in the second trial, logistic regression remained strong with offensive and defensive relative ratings becoming the dominant predictors. This means that playoff teams are typically the teams that not only produce high scores offensively and defensively, but are teams that can separate their offensive and defensive ratings from other teams in the league. It also shows that the defensive rating of a team is slightly more meaningful than the offensive rating, backing up the old saying, “Defense wins championships”. All in all, the model’s performance suggests that logistic regression can capture meaningful linear patterns in the data even when the most obvious predictors are absent.

The SVM model outperformed both models with a balance between all the scores and a higher AUC in both runs. Unlike the models before, SVM with the RBF kernel was able to model nonlinear interactions between features. NBA playoff qualification is not determined by a single threshold or linear border and can be a combination of statistics that act in a nonlinear way. This explains why the SVM was able to achieve the highest score, because it was able to capture those patterns. The model’s ability to work in higher-dimensional space allows the study of these interactions to be properly taken into account. Overall, the SVM managed to make use of nonlinear feature interaction to outperform both logistic regression and KNN in model performance metrics.

The two different feature sets fitted onto the three models shows the playoff classification is indeed possible with feature characteristics of team average stats. The first run showed that win-based metrics push predictions to high accuracy as they align heavily with playoff position. The second run showed that even without win-based metrics, offensive and defensive ratings can still predict playoff qualification with reasonably high accuracy and AUC. With this, it can be determined that the regular season team statistics can in fact be used to determine a team’s playoff capabilities. This also suggests that playoff teams distinguish themselves not only from wins, but also with sustained efficiency offensively and defensively while pace of play from the team is less important.

In terms of the history and future of projects relating to this subject, many researchers have found similar results and future extensions to the analysis can be done to give further information. A recent study published in 2025 by Manho Yeung at the ITM Web of Conferences also found that models like KNN, logistic regression, and random forest can perform well at predicting NBA playoffs positions. Yeung's study focuses more on the individual game stats to predict but similarly to the second run conducted, found that efficiency level stats can help predict playoff potential. This combination of this project and the study done by Yeung shows that game level and season levels statistics can both be used to classify a team as playoff worthy or not.

The project conducted has some key limitations present that can be fixed in future updates with the project. The main limitation with the current approach is the ignorance of player availability, trades, and strength of schedules. Incorporating statistics that take these into account may significantly boost the classifying abilities, especially for borderline teams. The change of teams, injuries, and difference in team opponents are all factors constantly present and should not be ignored if a comprehensive model is desired. Additionally, ensemble methods like random forest could be tested to capture more feature importance and accuracy measures. These models have proven effective in past studies like the one conducted by Yeung. Hyperparameter tuning for models like KNN and SVM can also be adjusted to obtain additional performance. Finally, moving past the current binary classification to make the model into a percentage or seeding predictor can also serve useful to NBA front offices. All in all, certain limitations are clearly present in the current state of the project but further improvements can be made to boost the capabilities of the study.

Overall, the results suggest a clear hierarchy among the models and features. All three models present a unique approach to classification but produce a positive result. The project demonstrates a positive predictive performance while laying a foundation for understanding and extensions in the sport analytic realm.

GitHub Link: https://github.com/nguy4531/Data-Mining-Project

References

- https://www.itm-conferences.org/articles/itmconf/pdf/2025/01/itmconf_dai2024_04024.pdf
- https://www.kaggle.com/datasets/boonpalipatana/nba-season-records-from-every-year/data