NBA analytics: What makes a team win?

Final Project Report

Firas ABO MRAD
MSc DSBA student 2022-2023
firas.abomrad@student-cs.fr

Karim EL HAGE
MSc DSBA student 2022-2023
karim.elhage@centrale-supelec.fr

Konstantina TSILIFONI
MSc DSBA student 2022-2023
konstantina.tsilifoni@student-cs.com

ABSTRACT

This paper constitutes the final project report for the third assignment in the Foundations of Machine Learning module of the MSc in DSBA, offered by CentraleSupelec, with Fragkiskos Malliaros as supervising professor. This project is closely related to previous research on basketball statistics, which aims to understand what factors a team or a player should have to maximize performance and ultimately the winning chance of a basketball game. However, this study focuses on identifying important statistics for points scored and game outcome, in addition to the previous studies which have focused on short-term game performance or long-term team performance. Moreover, this paper aims to identify key statistics that contribute to a basketball team winning a game in the NBA using machine learning techniques. Four models were used in this study: Linear Regression, Logistic regression, Neural networks, and XGBoost. The results of our analysis showed that the 2-point shot percentage, defensive rebound percentage and the turnover expected ratio were the most important statistics for winning. The models performed well on the test set, with logistic regression achieving the highest accuracy. Additionally, it was observed that statistics such as steals were not statistically important in determining the outcome of the game. The findings of the research should generalize over any future games and will exclude statistics from games happening during the exceptional COVID-19 period. Furthermore, the findings cannot be immediately extrapolated to leagues other than the NBA without first conducting a similarity analysis.

Keywords

Machine Learning, Basketball, NBA, Linear Regression, Logistic Regression, XGBoost.

1. INTRODUCTION AND MOTIVATION

The world of sports generates a tremendous amount of data. Data that was previously unleveraged. However, in the past decades, starting with NFL, data science, and more specifically AI and machine learning, have left a distinct imprint on the sports industry, radically altering how sports are perceived [5]. The uniqueness and, at the same time, the complexity of sports analytics stem from the large number of interrelated aspects and factors that add a substantial amount of uncertainty to any prediction attempt. There is no model that can represent all those traits, although many attempts have been made to come as near as feasible.

Basketball analytics in specific, is a growing sector with a variety of possible applications of machine learning. In this industry the data is widely available, however the documentation is not exhaustive; there is still opportunity for contribution simply because there is a near endless pool of potential applications. The interested parties of basketball analytics span from team owners to coaches to even financial investors.

Sports analytics are extremely important to increase revenue and improve player performance and the team's quality of play. Nowadays sports teams are using these analytics to have a competitive advantage. It is also known that basketball itself as a game evolves with time. The way basketball was played 30 years ago is potentially quite different from the one played today. Different characteristics are important, new skills arise and others fall, changing the game fundamentally.

In this project, we aim to identify what makes a team win a game. Is it the assists? The 2-point shot percentage? To do that we will predict the points scored and whether a team will win or lose in an NBA matchup based on the statistics collected for a game. Our objective is to pinpoint the relevant metrics used to define whether the team is capable of winning or not.

These metrics are evidently relevant to the coaching staff, providing them with information on what traits they should be looking for in a player, and what they need to focus on for their game style. Coaches must understand which factors are the most important to prepare a team for basketball games, develop the best strategies, and make appropriate judgments during a game. The interested parties do not stop there however, administrators and advertisers and potential sponsors of each team should also take an interest in such research since applications of the results will theoretically result in more wins for a team and therefore increase profitability for the team.

2. PROBLEM DEFINITION

A little background shall be given first on how the scope was redefined before defining the problem. Initially, the goal of the project was to build a model able to accurately predict the outcomes of basketball games based on previous games. Upon greater analysis of the problem, it was evident that the exercise was too complex to run due to the forecasting component of the problem. Essentially the models that would be trained will be using features composed of game and player statistics. However, when evaluating the models, these features need to be forecasted or else the goal is not being achieved - predicting the outcomes of future NBA games. There would be an error in forecasting the value of the features as well as how an error in the actual game prediction. Hence, it was proposed to alter the scope of the work to instead determine the statistics that best contributed to the outcomes of NBA games due to the values such scope of work can add as discussed in Section 1.

Firstly, a few basketball notations shall be defined to better understand analysis in later sections:

- FGM - Field Goal Made is defined as any point scored during the game except in the context of a free throw.
- FGA - Field Goal Attempted is defined as any attempt to score a point in a game except in the context of a free throw
- FG_PCT - Field Goal Percent is a ratio of FGM/FGA

- FG3M - 3- Point Field Goal Made
- FG3A - 3-Point Field Goal Attempted
- FG3_PCT - 3 - Point Field Goal Percent is a ratio of FG3M/FG3A
- OREB - Offensive Rebound is defined as a ball picked up after hitting the rim by the offensive team after an attempted shot that did was not made by the team on offense
- DREB – Defensive Rebound is defined as a ball pick after hitting the rim by the defensive team after an attempted shot was not made by the team on offense
- AST - Assist is defined as a pass made that can be directly attributed as leading to a point scored
- BLK - Block is defined as a shot that was blocked
- STL - Steal is defined when a ball is taken from the offensive team in the cases of a pass or possession
- PF - Personal Foul is committed by a team member on an individual player of the opposing team
- TOV - Turnover is defined as whenever a team loses possession of the ball to the opposing team prior to a shot being taken
- PTS_OFF_TOV - Points off Turnover is defined as points earned because of a Turnover being awarded
- WL - Win/Loss Outcome
- PTS_SCORED - Points scored by a team

Since the goal is to find feature importance rather than game outcome prediction, optimization and scoring metrics are only required to ensure that the models used to extract feature importance are valid. In general, there are two different functions that need to be optimized:

1- Mean Square Error (MSE)

$$argmin\,[(PTS\_SCORED\ -\ PTS\_SCORED_{PRED})^2]$$

where PTS_SCORED is the number of points scored in a particular game and PTS_SCORED_{PRED} is the predicted points scored with the selected features.

2- Negative Log likelihood Function of n independent Bernoulli Trials

$$argmin[-Wlog(W_{PRED}) - (1-W)log(1-W_{PRED})]$$

where W is a binary classifier of whether a team has won or lost a game (Win = 1) and W_{PRED} is the model's predicated game outcome given a set of features.

Optimizing the MSE function will allow us to evaluate the validity of the model used to determine which game statistics are important predictors of the number of points scored. Optimizing the latter holds the same for predictors of winning the game.

Since the findings should be able to generalize over any future, statistics from games happening during the exceptional Covid-19 period shall be excluded from analysis. Furthermore, the findings cannot be immediately extrapolated to leagues other than the NBA without first conducting a similarity analysis.

3. RELATED WORK

There is a growing interest in the study of basketball statistics to understand what factors a team or a player should have to maximize performance and ultimately the winning chance of a basketball game. Different approaches have been taken aiming many times at different objectives.

Some studies for example follow a short-term focus, trying to pinpoint the game-related statistics that lead a team to perform better and win a game, others focus on long-term team performance studying several basketball seasons, not focusing on the immediate outcome of a game but of a season as a whole [1].

Plenty of previous research [6];[7];[13];[9] focused on immediate game performance (win or loss) has demonstrated that teams that win tend to have superior shooting percentages and more successful defensive rebounding. However, according to [8] in certain game situations, statistics such as fouls and free throws become more crucial. Other statistics such as offensive rebounds, turnovers, steals, and assists have not been consistently shown to be differentiating factors between winning and losing teams. The above findings suggest that a team's success is related to the quality of decision-making by players and their field-goal efficiency within a strategic and tactical team environment [2]. Additionally, the ability to recover the ball after an opponent's missed shots through defensive rebounding can lead to more opportunities to score and win games [2];[7]. High-level rebounding performance is linked to players' physical attributes, muscular fitness, and technical and tactical preparation [11].

In the paper [1], focused on teams' regular season final ranking, instead of the short-term win or loss outcome, it is concluded that the most important factors for a teams' season-long performance are the assists, the steals and the blocks.

What seems less documented so far is the identification of important factors in relation to points scored in a game, which will be attempted in this project. Additionally, the significant game-related statistics leading to win or loss will also be researched and the results of both of the above will be compared to see whether their results coincide or not.

4. METHODOLOGY

4.1 General Approach

The *Basketball Dataset* has collected 149 statistically different metrics from the years 1946 – 2021, when excluding the target column (Win or Loss in a specific game). The dataset contains 62.4k rows, each representing a game played in the NBA league, either during the regular season or during the playoffs. Each game is characterized by 149 columns that make up its features and provide various information and statistical measures about it. The first step would be to discard features that are outright irrelevant to the target objective followed by a diligent dive into the raw data to ensure no entries are missing. It could be appropriate at this stage to conduct an initial exploratory data analysis to locate visually identifiable relationships between variables and to attempt to reconcile any factors such as large variations of the data from their assumed distribution. The next stage shall be feature selection. Following that, the goal is to use multiple machine learning algorithms, create models that can accurately predict game outcomes and extract feature importance. Baseline algorithms shall be first developed to ensure that more complex algorithm implementations are modeled appropriately.

4.2 Preliminary Data Analysis, Processing and Description

The Basketball Dataset is a sqlite file composed of several different tables. Each table corresponds to tabulated data scraped from the NBA's official website. Due to the sheer volume of data, the first task was to identify the tables pertinent to the analysis. These were:

1- The "Game" table contains 149 different pieces of data including statistics on the play of each team such as the number of blocks, steals, etc. This will be the main table from which game statistics will be extracted to train a model and interpret feature importance.

2- The "Draft Combine" table contains information on how players performed on the yearly Draft Combine event. The Draft Combine is defined as a "multi-day showcase event" where college basketball players are measured and take medical tests, are interviewed, perform various athletic tests and shooting drills, and play in five-on-five drills for an audience of National Basketball Association (NBA) coaches, general managers, and scouts" [15]. These are secondary statistics that can be used for model prediction to see if a team with overperforming players in this event yields a successful team and whether there are specific performance metrics at this event that teams should be paying close attention to. For the problem of this report, it could be relevant to take the mean and standard deviation of each Draft Combine metric per team to see if any metric has any influence on game outcomes. To do this, players would have to be linked to teams per season (since players can be released, traded etc).

3- The "Draft" table contains information on when players were drafted and when. The purpose of this table is to link players to the teams they are drafted so that statistics such as their performance on the Draft Combine can be also linked to the team.

Due to changes in NBA regulations over the years, it was decided to limit analysis to games starting from the 2015 season. Hence, analysis of the Game table would only be done starting from this date. It was found that the oldest drafting occurring in this season was in 1995. Therefore, analysis of the Draft and Draft Combine tables is limited to a minimum back to this year.

Firstly, the Draft Combine table was analyzed. It was found that 13 out of 116 columns presented the key information metrics of the players. However, it was found that on average approximately 60% of the values per metric were missing with the most complete metric having 48% missing values. Whilst it is found that most of the metrics follow more or less the shape of a normal distribution (with a few exceptions), it would be very misleading to fill in the missing values with the mean as it would conclude that most players performed "average" and this could yield misleading results later on in analysis. Furthermore, it was found that on average only 80% of the players per analyzed NBA season starting 2015 actually performed in the Draft Combine. Considering these two findings, it was decided to exclude Draft Combine metrics from analysis. An interesting future analysis with the available analysis would be to assess the usefulness in the Draft Combine performance on the future performance of players as this can be done with only the available information.

The Game table with game outcomes and metrics is then analyzed. Whilst there are 149 variables, not all the columns are true game statistics but other information about each game. Therefore, all non-game statistic columns are removed from analysis. Furthermore, each game has statistics for the Home and Away Team. The table there needed to be processed in a manner such that each row corresponds to the game statistics of a unique team and an additional binary variable was introduced to indicate whether the game was a Home game for a specific team. The table is then filtered so that only games starting from the 2015 season are included for analysis. The remaining 12 game statistics (excluding Points Scored and Win/Loss) are analyzed. Most statistics follow a normal distribution, which is ideal for the assumptions that are made in the used algorithms. Some variables have a slight skewness that could not be resolved with a log transformation. These were filled with a value of 0 as these two metrics are in fact only awarded in special circumstances that could technically not occur in a game. The number of Points Scored also approximately follow a normal distribution with a slight skewness to the Left. The Win/Loss outcome is in perfect balance which makes sense for this problem since for each game, there is a winner, and a loser and NBA basketball games cannot end in a tie.

All features that cannot be used to predict Game Outcomes such as whether a game is played at home or not (which surely could help build a better model) shall be excluded from the features used in the model. This is because these are features that teams cannot use to refine or improve on (match schedules are fixed).

4.3 Exploratory Data Analysis

The pre-selected game statistic features are plotted against one another to explore any immediate visible correlation between features, points scored, and game outcome. This was achieved by first plotting a correlation heat map to observe any linear relationship and then through plots to observe any other type of relationships. In the case of all features and points scored the reasonable plot was a line plot whereas since Game Outcome is boolean, it makes more sense to plot histograms of W/L for each feature separately.

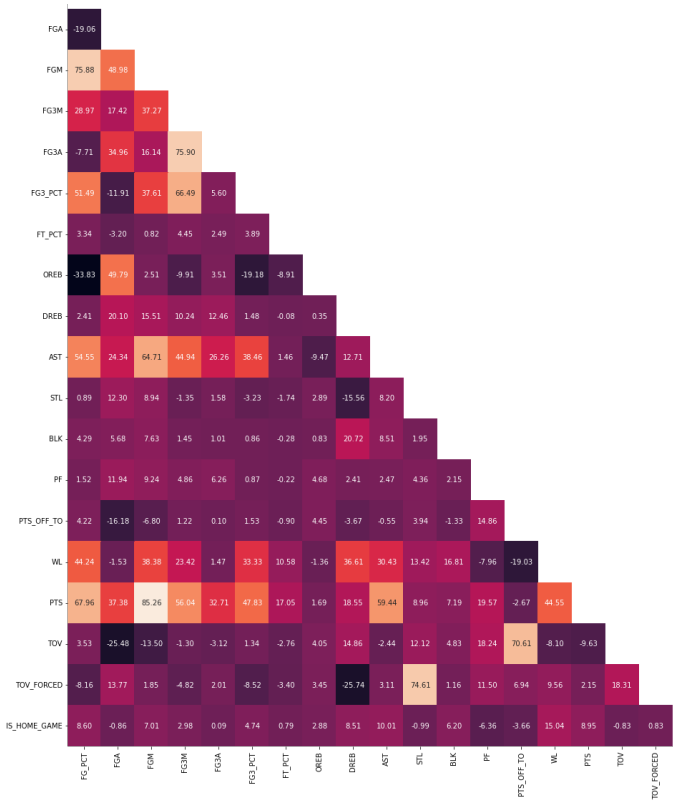

Figure 1: Correlation heatmap of features

Figure 1 above shows some strong linear relationships between PTS_scored and FGM, F3GM, AST. This is expected since all of these variables increase as a direct consequence of points increasing. Whilst it may seem that such variables, therefore, are not pertinent to analysis, it must be noted that the purpose is to see which metrics win games rather than score more points. For example, it would be interesting to see whether 3- point shots are more important than 2 - points when the game outcome is at stake.

Another important phenomenon is the collinear relationship between all Field Goal features. This makes sense considering that FG_PCT is a ratio of FGM to FGA and FG is a summation of FG3 another variable. Other collinear trends can be observed between TOV_FORCED and STL as well as TOV and PTS_OFF_TO. This also makes sense since steals automatically lead to a turnover and points from a turnover also originate from a turnover (if the point is scored). In section 4.4, solutions shall be proposed to resolve this whilst still retaining important features.

Since WL is a binary variable, it was more appropriate to plot Wins and Losses separately to observe any trends in the data rather than rely on the data in the heat map. The plot did not yield any obvious trends between the features and WL.

4.4 Feature Engineering

Proper feature selection and engineering is imperative to ensure that the coefficients of features in models are not severely impacted by collinearity issues observed in section 4.3. This is especially critical since in some of the models discussed in section 4.5, coefficients will be used to interpret feature importance, and consequently determine the most important predictors of game outcomes.

FG3_PCT, FG3A, and FG3M have a strong correlation to FG_PCT, FGA, FGM. This makes sense since each FG feature is in fact a sum of FG3 and FG2, the latter being the 2-point Field Goals. Therefore, the features FG2_PCT, FG2A, FG2M are extracted from the difference between FG_PCT, FGA, FGM and FG3_PCT, FG3A, and FG3M respectively. Having the FG3 and FG2 features allow now to drop all the FG features. Furthermore, since all the PCT (percentage) is the ratio of Made shots and Attempted shots, only FG3_PCT and FG_PCT need to be kept as features in the model.

The task to create new features proved to be difficult if the aim is to also distinguish what review papers have done. Furthermore, the features need to be easily interpretable for use by officials. In general, it is better to stick to generally known statistics. One statistic that was engineered however is the Turnover Exploitation Ratio which is the ratio of PTS_OF_TOV to TOV_FORCED (Points from Turnovers to forcing another team to do a turnover). This could give an idea of the usefulness of exploiting turnovers when forcing the other team to make them.

Whilst other statistics also had some linear relationship, they are critical to the overall analysis of the model and so must be kept. Unfortunately, it is not possible to make use of dimensionality reduction methods such as SVD, PCA, or LDA as this would contradict the objective of this study by losing the interpretability of the features.

For training all the models below, the data is split into 80% for training and 20% for testing to establish the ability of the model to generalize on new data.

4.5 Algorithms and models

Linear Regression:

Linear Regression is used to train and evaluate a model for predicting a continuous outcome, specifically, points scored (PTS) in a game. The LinearRegression object is instantiated and then fit to training data, task2XTrain and task2YTrain, to generate the model. For evaluation, the Root mean squared error (RMSE) is employed as a metric. The RMSE metric will be used as a baseline for other models that will be trained to predict scores of NBA games.

$$RMSE = \sqrt{\frac{\sum_i^n (PTS\_SCORED_i - PTS\_SCORED_{PRED,i})^2}{n}}$$

Specifically, the RMSE on the test set shall be used as the baseline.

The statsmodel package shall be used instead of sklearn to perform the regression due to the ability to output a detailed summary of the results of the model to interpret feature importance. Features with large positive coefficients will signify high positive feature importance whereas large negative coefficients will signify a detrimental impact to the model. In fact, it could give an indication of how much 1 additional unit of the feature could improve the number of points scored and vice versa if the coefficient is negative. The second important indicator of importance is the p-value which explains statistical significance. Any feature with a p-value less than 0.05 could be statistically significant (whether positively or negatively) to the number of points score. Whilst R^2 can also be used to understand how much the model explains the number of

points scored, it must be noted that some features were intentionally not introduced (like whether a game is played at Home) since they cannot be improved upon or lead to deeper strategy analysis. A 5-fold cross validation will be run to ensure that the linear regression model is not over fitting. If the RMSE of the average cross-validation set is comparable to the RMSE of the test set, it can be concluded that the model is not overfitting.

Logistic Regression:

Logistic Regression is used to train and evaluate a model that predicts a binary outcome, in this case, win or loss (WL) in a task. The model out will be given between 0 and 1. All values greater than 0.5 shall be deemed as a win. To evaluate the model, the W/L shall be directly compared that of the actual outcome for both the train and test set to determine the model's accuracy. This accuracy shall be used as a baseline accuracy for the more complex models to be developed to predict game outcomes. For comparison, the accuracy of the test set shall be used as a baseline. As with Linear Regression, the statsmodel package shall be used instead of sklearn to perform the regression due to the ability to output a detailed summary of the results of the model to interpret feature importance. The coefficients cannot be interpreted in the same way as Linear Regression. This is because although the dependent variables are taken as linear combinations of the features, the relationship is defined as the sigmoid function define below, considering the dependent variable is a probability:

$$p = \frac{1}{1 + e^{-X}}$$

The magnitude should still give a strong indication of the importance of each feature. The second important indicator of importance again is the p-value, which explains statistical significance. Any feature with a p-value less than 0.05 could be statistically significant (whether positively or negatively) to the number of points score. The logistic regression model shall be trained with a BFGS optimizer and increase the default number of iterations to ensure convergence. As with the linear regression model, a 5-fold cross fold validation shall be conducted to ensure that the model is not overfitting. The average cross-validation score shall be compared to the test set score.

Neural Networks

Whilst it may seem counterintuitive to use a neural network due to interpretability concerns, a non-complex neural network shall be developed to still be able to retain interpretability if deemed to perform better than the baseline.

We are using a feedforward neural network with two layers, and 128 neurons in each layer. The architecture shall not be made deeper for interpretability purposes. The architecture of the network is defined using the Sequential module of PyTorch, which is a container for holding a linear sequence of PyTorch modules. The network is trained firstly on the WL task. If deemed better than the baseline model, it will also be trained on the (PTS) task. In order to introduce a non-linearity component on each iteration, an activation is placed between each hidden layer. Due to ease of computation, the activation function of choice shall be ReLU. Whilst the ReLU function is linear when the output of the layer is positive, it then always zero whenever the output of the layer is negative and hence being overall a nonlinear function. Since WL is a classification problem, activation function to output shall be a sigmoid function as in the logistic regression model.
The input data is first converted into PyTorch tensors. Adam is defined as the model optimizer with initially an arbitrary learning rate as baseline. Adam is a popular optimization algorithm that uses moving averages of the parameters to provide a running estimate of the second raw moments of the gradients; the gradient is divided by the square root of the moving average. The Adam optimizer will work to minimize the Binary Cross Entropy Loss. This is in fact an optimization similar to the log likelihood function of the logit function defined in Section 2. An epoch limit of 20000 is defined, and the model is trained using the training data. With the data split it batches, an epoch can be defined as the number of times the model is trained on all training batches. As with the baseline logistic model, accuracy shall be used as a metric to evaluate model performance. Model accuracy on the test set shall be stored every 100 epoch. To test for overfitting, the Binary Cross Entropy Loss of the training and test set are plotted per epoch to observe when the test loss increases as the training loss decreases. The optimal number of training epochs can therefore be defined after observing this. The test set accuracy at the optimal epoch shall be compared to the baseline Logistic Regression model and see if performance is similar or better otherwise shall not be used for feature importance analysis.

The neural network tasked to train for PTS score shall follow the same methodology as the for WL. The main differences arise in that ReLU shall use the final activation function before output. Furthermore, the optimization function that needs to be solved is the MSE, as in linear regression. The RMSE on the test set at the optimal epoch would then be compared to the baseline linear regression model and see if performance is similar or better otherwise shall not be used for feature importance analysis.

XGBoost

The XGBoost (Extreme Gradient Boosting) is an ensemble learning algorithm, built on decision trees. It essentially combines individual decision trees, with the goal of performing a better classification. Each tree is trained to correct the mistakes of the previous tree. It uses a gradient boosting framework, which optimizes the cost function by minimizing the loss (e.g. mean squared error or log loss) using gradient descent.
Once the trees are trained, predictions are made by traversing the trees and combining the results from each tree. The final prediction is an average or a weighted average of the predictions from all the trees in the ensemble, depending on the configuration.
Two different types of XGBoost algorithms were used, the XGBClassifier and the XGBRegressor, with the purpose of creating a model for win/loss and points scored respectively. For each of the algorithms a baseline model was initially built in order to have a comparison standard. Following that the model's parameters were tuned using GridSearch for a certain group of parameters for each algorithm. For model hyperparameter tuning, a grid search with 5-fold cross-validation was conducted on each algorithm. The result of GridSearch is a model having the best combination of the possible parameters given based on an evaluation metric, root mean squared error for the regression and accuracy for the classification task. The hyperparameters tuned in both models were: (explanation taken from the official documentation of XGBoost [10];[14])

Table 1: XGBoost Parameters List

Hyperparameter	Definition	Best found (classifier)	Best found (regressor)
eta	learning rate	0,05	0,05
n_estimators	number of boosting rounds	70	160
max_depth	maximum tree depth for base learners.	7	5
min_child_weight	minimum sum of instance weight(hessian) needed in a child	5	4
eval_metric (always set)	defines the metric used for monitoring the training result and early stopping.	'error' (same as accuracy)	'rmse' (root mean squared error)
gamma	minimum loss reduction required to make a further partition on a leaf node of the tree	0,5	0,3
objective (always set)	the specific learning task and the corresponding learning objective or a custom objective function to be used	'binary:logistic' (logistic regression for binary classification, output is probability)	'reg:squarederror' (regression with squared loss)

For the classification task the additional parameter of **max_delta_step** was used, which represents the maximum delta step we allow each tree's weight estimation to be, the optimal was found to be **1**. **Early_stopping_rounds** were also used, which activates early stopping. Validation metric needs to improve at least once in every early_stopping_rounds round(s) to continue training. The best found was **6**. As for the regression task the **alpha** parameter was also used, which depicts the L1 regularization term on weights and who's optimal was found to be **0**, namely no L1 regularization is used.

More specific steps were taken in the classification task to tune the win probability threshold. In other words, the model can predict a probability that a team won, which is then compared to a threshold value to classify it finally actually as a win or loss. The default value for that threshold is 0,5 and it was checked whether that should be changed or not. The optimal threshold value was defined as the one maximizing the accuracy of the model. Unsurprisingly and since our data is balanced the optimal value was indeed very close to 0,5.

Apart from training and predicting, XGBoost also has built-in feature importance evaluation, which can be used to understand which features are most important in the model. This can be useful in the current project to understand the underlying relationships between input features and the output and is one of the main reasons that the algorithm was chosen.

For each of the tasks, two feature importance plots are produced. The one quantifies the importance as the weights [10], namely the number of times a feature appears in a tree and the other ranks the features on importance related to the average gain of splits which use the feature [10]. Those plots are used to reach the final objective of this project, to identify the most important factors that predict the points a team will score and whether it will win or not.

4.6 Limitations

It shall be recognised and mentioned that there are certain limitations in pursuing this project. Because of the nature of the problem, there are various factors that might possibly impact the results of the games, which are characterized by high uncertainty and would be exceedingly difficult to model. The emotional state of players, injuries to players, mid-season transfers, and the effect of covid are examples of such elements. These occurrences are considered abnormal throughout the season and, as such, will be assumed to be missing for the purposes of this project.

5. EVALUATION

According to prior papers on sports analytics [3];[4], the preferred evaluation metric for the performance of a predicting algorithm is the accuracy of the predictions, namely the percentage of the correctly predicted outcomes. The dataset to be used as mentioned in the general approach section is the *Basketball Dataset*, and in specific the *Game* tables, found in Kaggle. In our analysis, not all the 149 features available will be used but rather more specific, tangible ones. The evaluation of the produced predictions will be made in comparison to the actual outcomes of the games for the years 2015-2019. For every model run, there is an initial split of the data available, using 80% as train data and 20% as test data, while keeping the balance (50%) of the class win/loss in each split subset. However, what is important to note is that predicting the outcome of a game is not the purpose of this study but rather identifying the features that are used to reach a certain prediction. It is obvious that an accurate model is still beneficial since accurate predictions will

most likely give accurate estimations as to what factors and game-based statistics are important.

In every model, a 5-fold cross-validation was used in order to identify potential overfitting as it can provide an estimate of the performance of the model on unseen data, which can later be compared to the actual performance of the model. To identify potential overfitting, we shall compare the cross-validation accuracy with the accuracy on a separate test set (20% of data that was kept aside). Specifically, the stratified k-fold was used, which ensures that each fold contains roughly the same proportions of the different types of class labels. This is particularly useful for an imbalanced dataset, which is not the case here, however, it is still good practice to use it.

For a 5-fold cross-validation the dataset will be split into 5 equal folds. The model is trained on 4 of the folds and then tested on the remaining one. This process is repeated 5 times, with a different fold being used as the test set in each iteration. The final performance metric is the average of the performance metric obtained in each iteration.

The tables below contain the results of the validation process.

Table 2: RMSE Score Model Comparison – PTS SCORED

Model	Test Set Score (RMSE)	Cross-validation Score (RMSE)
Linear Regression (Baseline)	6,48	6,65
Neural Network	NA	NA
XGBoost	6,43	6,51

Table 3: Accuracy Score Model Comparison - WL

Model	Test Set Score (Accuracy)	Cross-validation score (Accuracy)
Logistic Regression (Baseline)	82,48 %	83,94 %
Neural Network	80.88%	NA
XGBoost	82,1 %	82,80 %

Linear Regression

The RMSE of the Linear Regression model was found to be 6,48 in its average deviation from the true number of points scored. This shall be used as a baseline for the more complex models.

After comparing the results of the 5-fold cross-validation average RMSE to that of the test score found after training, it can be concluded that the Linear Regression model is not prone to overfitting due to the proximity between both RMSEs.

Logistic Regression

The Test Set accuracy of the Logistic regression model was found to be 82.48% which can be interpreted as successfully classifying 82 out of 100 games on the test as Win/Loss. This accuracy score shall be used as the baseline for the performance of the more complex algorithms.

After comparing the results of 5-fold cross-validation average accuracy to that of the test score found after training, it can be concluded that the Logistic Regression model is not prone to overfitting due to the proximity between both accuracies.

Neural Network

The Neural Network built for WL classification was evaluated for overfitting by comparing the training BCE loss to the test BCE loss every 100 epochs.

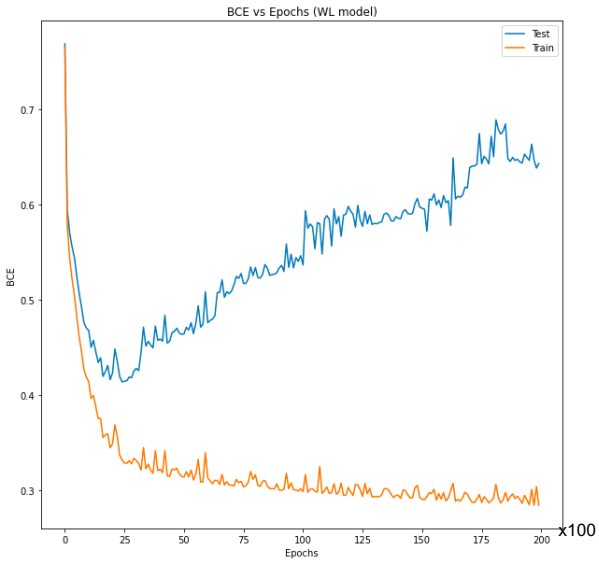

Figure 2: Binary Cross Entropy Loss Per Epoch

Figure 2 shows the Neural Network Binary Cross Entropy Loss per 100 epochs. The model overfits at around 25 epochs and so the parameters with the best cross validation score are stored and used for further evaluation. Due to this methodology for testing overfitting, cross-validation accuracy in the Table 3 is seen as NA.

The test set Accuracy score in the Neural Network performs worse than the baseline Logistic Regression model. Hence, the Neural network shall neither be used for feature importance analysis nor for pursuing a PTS_scored model. This also explains the NA that can be seen in Table 3.

XGBoost

Comparing the cross-validation score and the test score for each of the regression and classification models indicates that there is no overfitting from the models used. As for the performance of the models on the test set the results are satisfying, with the regression model being off on average by 6,43 points in its estimation and the classification model predicting on average 82 of 100 game outcomes correctly.

6. CONCLUSIONS

In conclusion, this study aimed to identify the key statistics that contribute to a basketball team winning a game in the National Basketball Association (NBA). By using machine learning techniques to predict the number of points scored and the outcome of a game based on statistics collected during the game, we were able to pinpoint the relevant metrics that are pertinent to a game outcome.

In the Appendix, Figure 4 - Figure 8 are used to interpret feature importance from the XGBoost model. Figure 7 shows the regression summary of the linear regression model whereas Figure 8 that of the logistic regression. As discussed previously, the coefficients and p-value of each feature are used for the interpretability of the Linear and Logistic Regression model. Here, it is paramount to emphasize that to interpret the importance of the coefficient of FG3_PCT and FG2_PCT, the coefficients need to be divided by 100 since these features are in percentage form. This is because, in the linear regression model, the coefficients explain the impact 1 full unit of a feature can have on the change of a dependent variable. However, this interpretability would mean the value of the feature is greater than 100%, which is impossible. Therefore, we instead divide the value of the coefficients of these percentages by 100 to interpret the realistic impacts that their variation could have.

The results of the XGBoost classifier showed that the 2-point shot percentage, the 3-point shot percentage as well as the defensive rebounds were the most important statistics for winning (Figure 4 - Figure 3. This is consistent with previous research that has shown that teams that win tend to have superior shooting percentages and more successful defensive rebounding. Additionally, the turnover exploitation ratio was found to be an important statistic in determining the outcome and points scored in the game. This is a metric that measures the extent to which a team took advantage of the opponents' turnovers.

Moreover, XGBoost and Logistic Regression gave interesting results to help identify the most important statistics for predicting points scored and game outcome. Our analysis revealed that the 3-point shot percentage shows slightly more important than the 2-point shot percentage according to the XGBoost model, while the Logistic Regression showed significantly more important for the 2-point percentage. Logistic Regression had a higher accuracy for win prediction. Despite this discrepancy in feature importance between the models, both models agree that the defensive rebound percentage and turnover expected ratio were important statistics in point and win prediction. Our results suggest that teams with a higher 3-point shot percentage, defensive rebound percentage and lower turnover expected ratio are more likely to score more points and win games.

It is important to note that the coefficients of the variables in the logistic regression model are interpreted as the change in the log-odds of winning for a one unit increase in the variable, holding all other variables constant. On the other hand, steals were found to not be statistically important in determining the outcome of the game. Interesting is to note that there were some results in the analysis opposite to what was expected, regarding the assists and the turnover exploitation ratio. The coefficient for assists was found to be negative, which suggests that teams that have more assists are less likely to win. This may be due to teams that have more assists typically playing at a faster pace and taking more risks, which can lead to more turnovers and ultimately fewer wins. The negative coefficient for turnover expected ratio also seems counterintuitive since teams which take greater advantage of an opponent's turnovers should logically be more likely to win.

In addition to identifying key statistics that contribute to winning in the NBA, there are several areas for future research that could build on the findings of this study. One potential area for future research would be to analyze the relationship between combined statistics and performance in the NBA. The NBA combine is a pre-draft event where draft prospects perform various physical and skill-based tests and drills. Measuring the correlation between Draft Combine statistics and in-game performance could provide valuable insights for teams when drafting players. Another area for future research would be to analyze the impact of coaching and strategy on game outcomes. While this study focused on team statistics resulting from player performance, it would be interesting to investigate the impact of different coaching styles and strategies on winning. Additionally, it would be interesting to investigate the relationship between team chemistry and winning. Team chemistry can be a difficult variable to quantify, but it is widely believed to be important for a team's success. Measuring the impact of team chemistry on winning could provide valuable insights for teams and coaches. Finally, it would be beneficial to replicate this research using other leagues such as the EuroLeague, the National Basketball League, and other professional leagues around the world to validate the findings and to compare the results with the NBA. In summary, this research opens a door for further exploration in basketball analytics by identifying key statistics that contribute to winning in the NBA. Future research can build on these findings by analyzing the relationship between combine statistics and performance in the NBA, the impact of coaching and strategy on game outcomes, the relationship between team chemistry and winning, and replicating the research in other leagues around the world.

7. REFERENCES

[1] S. J. Ibáñez, J. Sampaio, S. Feu, A. Lorenzo, M. A. Gómez, and E. Ortega, 'Basketball game-related statistics that discriminate between teams' season-long success', European Journal of Sport Science, vol. 8, no. 6, pp. 369–372, Nov. 2008, doi: 10.1080/17461390802261470.

[2] S. Trnini, D. Dizdar, and E. Luk, 'Differences Between Winning and Defeated Top Quality Basketball Teams in Final Tournaments of European Club Championship', Coll. Antropol., 2002.

[3] V. Sarlis and C. Tjortjis, "Sports analytics - evaluation of basketball players and Team Performance," *Information Systems*, 23-May-2020. [Online]. Available at: https://www.sciencedirect.com/science/article/pii/S0306437920300557?casa_token=jO_mfAnI7MwAAAAA%3AqZ61JuLQNY5F0SVx1I_nhYLC1wc8V9FwE9Gg5ZWe2eB-4eDiAdjXu7HEvlKE_7yxxtmQ4FMb7M2NMg. (Accessed: 30-Oct-2022).

[4] C. Cao, "Sports data mining technology used in basketball outcome prediction," *ARROW@TU Dublin*. [Online].

Available at: https://arrow.tudublin.ie/scschcomdis/39/. (Accessed: 30-Oct-2022).

[5] N. Chmait and H. Westerbeek, “Artificial Intelligence and machine learning in sport research: An introduction for non-data scientists,” *Frontiers*, 01-Jan-1AD. [Online]. Available at: https://www.frontiersin.org/articles/10.3389/fspor.2021.682287/full. (accessed Oct. 30, 2023).

[6] M. Akers, S. Wolff, and T. Buttross, ‘An Empirical Examination of the Factors Affecting the Success of NCAA Division I College Basketball Teams’, Journal of Business and Economic Studies, Jan. 1992, [Online]. Available at: https://epublications.marquette.edu/account_fac/72 (accessed Jan. 09, 2023)

[7] Ibáñez, S. J.; et al. "Game statistics discriminating the final outcome of junior world basketball championship matches (Portugal 1999)." Journal of Human Movement Studies, vol. 45, no. 1, pp. 1-20, 2003.[Online] Available at: https://scholar.google.com/scholar_lookup?hl=en&volume=45&publication_year=2003&pages=1-19&author=S.+J.+Ib%C3%A1%C3%B1ez&author=J.+Sampaio&author=P.+S%C3%A1enz-L%C3%B3pez&author=J.+Gim%C3%A9nez&author=M.+A.+Janeira&title=Game+statistics+discriminating+the+final+outcome+of+Junior+World+Basketball+Championship+matches+%28Portugal+1999%29%20 (accessed Jan. 09, 2023).

[8] B. Kozar, R. E. Vaughn, K. E. Whitfield, R. H. Lord, and B. Dye, ‘Importance of Free-Throws at Various Stages of Basketball Games’, Percept Mot Skills, vol. 78, no. 1, pp. 243–248, Feb. 1994, doi: 10.2466/pms.1994.78.1.243.

[9] Karipidis, A.; Fotinakis, P.; Taxildaris, K.; Fatouros, J. "Factors characterizing a successful performance in basketball." Journal of Human Movement Studies, vol. 41, no. 5, pp. 385-397, 2001 [Online] Available at: https://scholar.google.com/scholar_lookup?hl=en&volume=41&publication_year=2001&pages=385-397&author=A.+Karipidis&author=P.+Fotinakis&author=K.+Taxildaris&author=J.+Fatouros&title=Factors+characterising+a+successful+performance+in+basketball (accessed Jan. 08, 2023).

[10] ‘Python API Reference — xgboost 1.7.3 documentation’. [Online] Available at: https://xgboost.readthedocs.io/en/stable/python/python_api.html (accessed Jan. 13, 2023).

[11] J. Carter, T. Ackland, D. Kerr, and A. Stapff, ‘Somatotype and size of elite female basketball players’, Journal of Sports Sciences, vol. 23, no. 10, pp. 1057–1063, Oct. 2005, doi: 10.1080/02640410400023233.

[12] J. Sampaio and M. Janeira, ‘Statistical analyses of basketball team performance: understanding teams’ wins and losses according to a different index of ball possessions’, International Journal of Performance Analysis in Sport, vol. 3, no. 1, pp. 40–49, Apr. 2003, doi: 10.1080/24748668.2003.11868273.

[13] R. F. Ittenbach and I. G. Esters, ‘Utility of team indices for predicting end of season ranking in two national polls’, Journal of Sport Behavior, vol. 18, no. 3, pp. 216–225, Sep. 1995.

[14] ‘XGBoost Parameters — xgboost 1.7.3 documentation’. [Online] Available at: https://xgboost.readthedocs.io/en/stable/parameter.html (accessed Jan. 13, 2023).

[15] "NBA Draft Combine," 28 10 2018. [Online]. Available: https://en.wikipedia.org/wiki/NBA_Draft_Combine.

8. APPENDIX

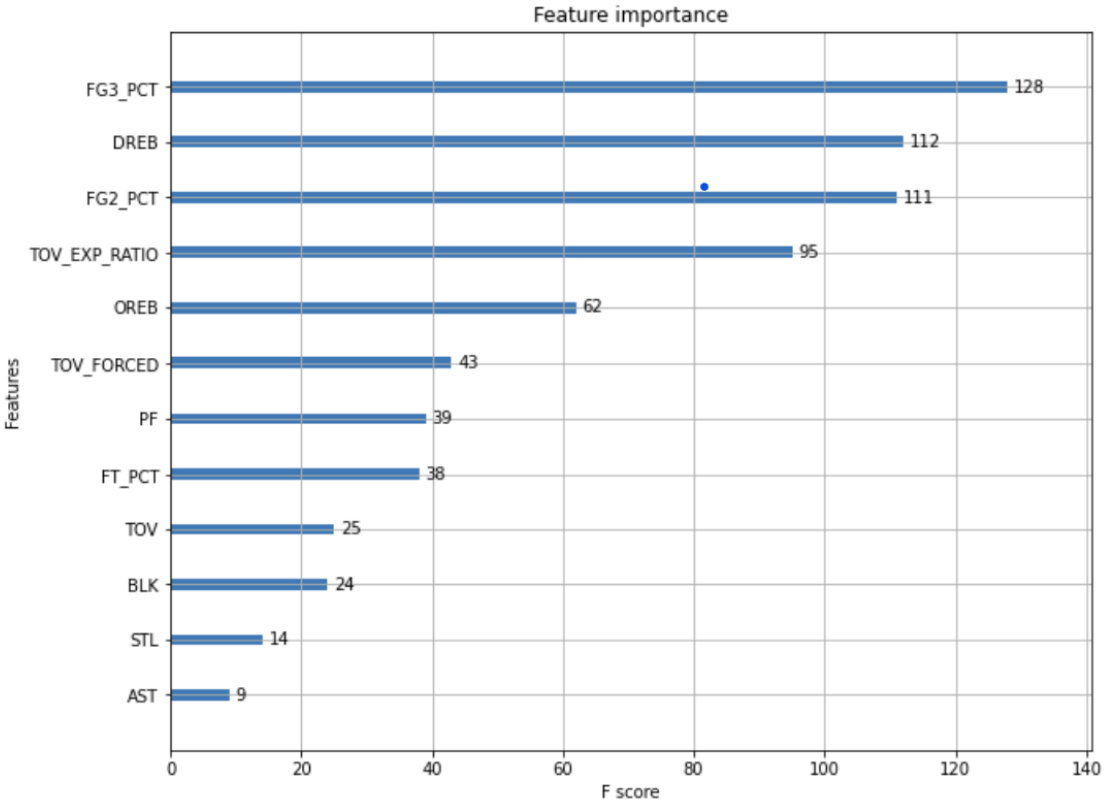

Figure 4: XGBoost WL Weights

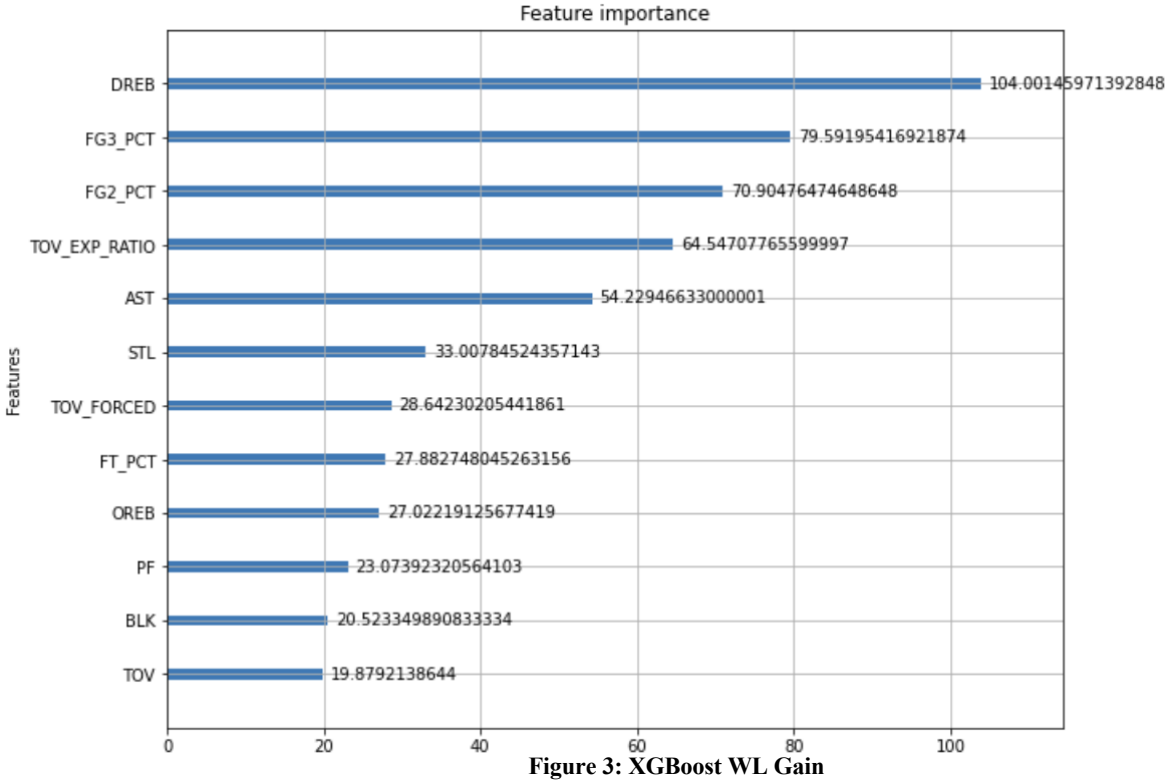

Figure 3: XGBoost WL Gain

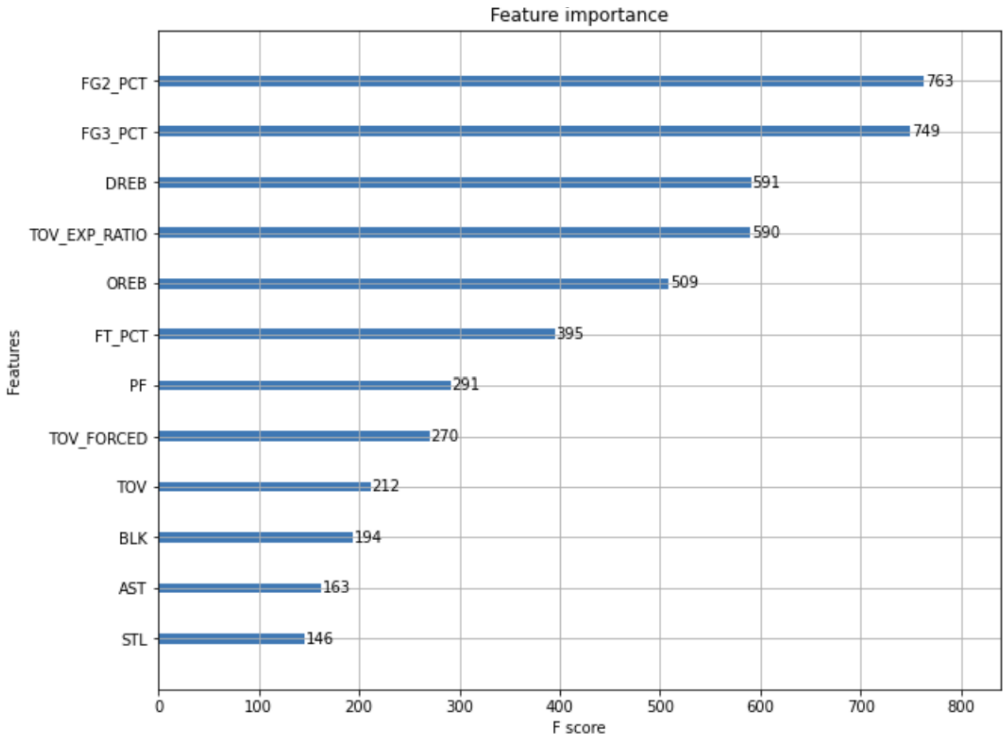

Figure 6: XGBoost PTS SCORED Weights

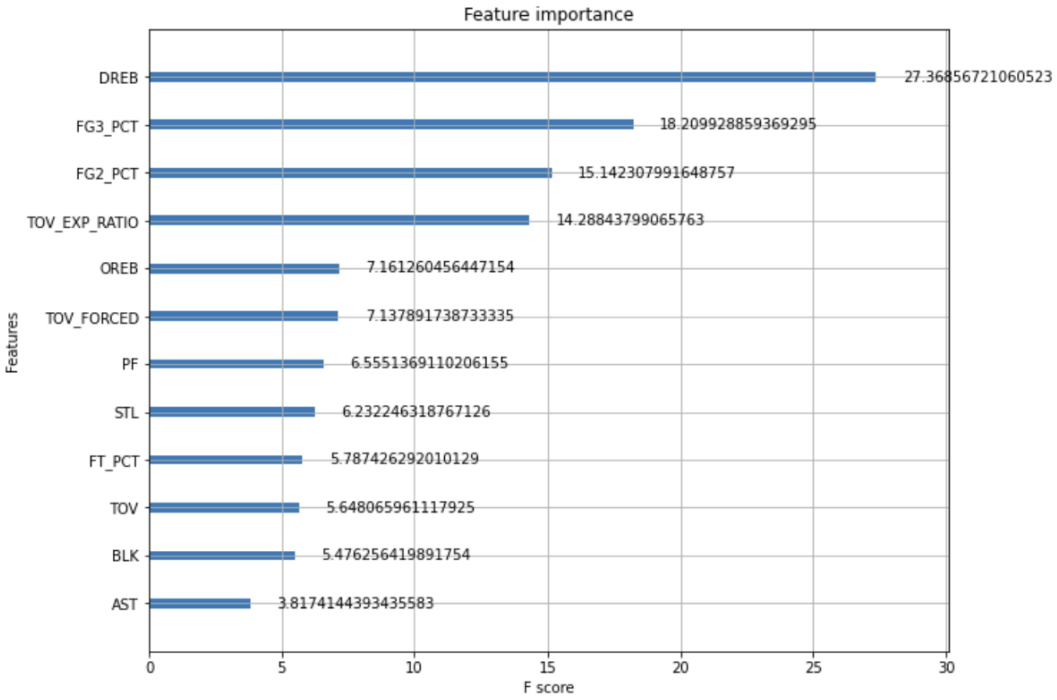

Figure 5: XGBoost PTS SCORED Gain

OLS Regression Results

Dep. Variable:	PTS	R-squared:	0.746
Model:	OLS	Adj. R-squared:	0.746
Method:	Least Squares	F-statistic:	3058.
Date:	Fri, 13 Jan 2023	Prob (F-statistic):	0.00
Time:	13:33:51	Log-Likelihood:	-41429.
No. Observations:	12508	AIC:	8.288e+04
Df Residuals:	12495	BIC:	8.298e+04
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
FG3_PCT	66.7759	0.742	90.032	0.000	65.322	68.230
FT_PCT	20.8543	0.579	36.046	0.000	19.720	21.988
OREB	1.0380	0.017	59.614	0.000	1.004	1.072
DREB	0.5048	0.012	41.380	0.000	0.481	0.529
AST	0.4048	0.014	27.925	0.000	0.376	0.433
STL	0.1035	0.031	3.372	0.001	0.043	0.164
BLK	-0.0268	0.024	-1.100	0.271	-0.074	0.021
PF	0.5417	0.014	38.631	0.000	0.514	0.569
TOV	-0.8858	0.020	-43.906	0.000	-0.925	-0.846
TOV_FORCED	0.6160	0.029	21.506	0.000	0.560	0.672
FG2_PCT	100.4782	1.007	99.780	0.000	98.504	102.452
TOV_EXP_RATIO	1.6748	0.137	12.185	0.000	1.405	1.944
const	-31.0742	0.971	-31.991	0.000	-32.978	-29.170

Omnibus:	527.469	Durbin-Watson:	1.998
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1595.058
Skew:	-0.131	Prob(JB):	0.00
Kurtosis:	4.730	Cond. No.	1.11e+03

Figure 7: Linear Regression PTS SCORED Model Summary - Baseline

Logit Regression Results

Dep. Variable:	WL	No. Observations:	12508
Model:	Logit	Df Residuals:	12495
Method:	MLE	Df Model:	12
Date:	Fri, 13 Jan 2023	Pseudo R-squ.:	0.4921
Time:	13:34:05	Log-Likelihood:	-4403.0
converged:	True	LL-Null:	-8669.7
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
FG3_PCT	19.6755	0.456	43.124	0.000	18.781	20.570
FT_PCT	4.2290	0.269	15.731	0.000	3.702	4.756
OREB	0.2310	0.009	26.979	0.000	0.214	0.248
DREB	0.3604	0.008	46.039	0.000	0.345	0.376
AST	-0.0650	0.007	-9.830	0.000	-0.078	-0.052
STL	0.0199	0.014	1.438	0.150	-0.007	0.047
BLK	0.1107	0.011	9.969	0.000	0.089	0.132
PF	-0.0949	0.006	-14.687	0.000	-0.108	-0.082
TOV	-0.1759	0.010	-17.728	0.000	-0.195	-0.156
TOV_FORCED	0.2558	0.014	18.483	0.000	0.229	0.283
FG2_PCT	24.1667	0.589	41.041	0.000	23.013	25.321
TOV_EXP_RATIO	-0.7821	0.071	-11.009	0.000	-0.921	-0.643
const	-34.6090	0.695	-49.794	0.000	-35.971	-33.247

Figure 8: Logistic Regression WL Model Summary - Baseline