

Kobe Bryant Shot Analysis

By: Anujan Thanabalasinkam

Contents

[Abstract 4](#_Toc57995272)

[Introduction 4](#_Toc57995273)

[Literature Review 4](#_Toc57995274)

[Dataset Explanation 7](#_Toc57995275)

[Data Statistics Summary 7](#_Toc57995276)

[Lists of Features and Descriptions 8](#_Toc57995277)

[Approach 9](#_Toc57995278)

[Approach Diagram 9](#_Toc57995279)

[Approach Details 10](#_Toc57995280)

[*Step 1: Attribute Study* 10](#_Toc57995281)

[*Step 2: Data Cleaning and Initial Insights* 10](#_Toc57995282)

[*Step 3: Insights and Data Visualization* 10](#_Toc57995283)

[*Step 4: Predictive Analysis* 10](#_Toc57995284)

[Data Preparation 10](#_Toc57995285)

[Issues with Dataset 10](#_Toc57995286)

[Major Feature Changes 11](#_Toc57995287)

[*Time Features* 11](#_Toc57995288)

[*Shot Distance Features* 12](#_Toc57995289)

[*Season Feature* 12](#_Toc57995290)

[*Date Feature* 12](#_Toc57995291)

[*Matchup Feature* 12](#_Toc57995292)

[*Opponent Feature* 12](#_Toc57995293)

[*Features to be eliminated* 13](#_Toc57995294)

[Missing Value Investigation 13](#_Toc57995295)

[Boxplots & Outlier Data 13](#_Toc57995296)

[Cartesian Distance Boxplots 13](#_Toc57995297)

[Global Positioning Boxplot 14](#_Toc57995298)

[Correlation Matrix 15](#_Toc57995299)

[Exploratory Analysis 16](#_Toc57995300)

[Action type 16](#_Toc57995301)

[Shot Location Visualization 18](#_Toc57995302)

[Shot location 18](#_Toc57995303)

[Shot Location “Hexbin” Plot 19](#_Toc57995304)

[Shot Location Percentage Plot 20](#_Toc57995305)

[Shot Type 21](#_Toc57995306)

[Regular Season vs Playoffs 22](#_Toc57995307)

[Opponents 23](#_Toc57995308)

[Game Date 24](#_Toc57995309)

[Game Year 24](#_Toc57995310)

[Dimension Reduction 25](#_Toc57995311)

[Shot Description Features 25](#_Toc57995312)

[Shot Location Features 25](#_Toc57995313)

[Shot Location Description Features 25](#_Toc57995314)

[Time Related Features 26](#_Toc57995315)

[Unrelated Features 26](#_Toc57995316)

[Results 26](#_Toc57995317)

[Predictive Analysis 26](#_Toc57995318)

[Features to be used for Predictive Analysis 26](#_Toc57995319)

[Training Set 27](#_Toc57995320)

[Test Set 28](#_Toc57995321)

[Receiver Operating Characteristic Curve (ROC) & Area Under Curve (AUC) 29](#_Toc57995322)

[Post Predictive Analysis 29](#_Toc57995323)

[Significance of this Study 30](#_Toc57995324)

[Conclusion 30](#_Toc57995325)

[References 31](#_Toc57995326)

[Github link 31](#_Toc57995327)

# Abstract

Kobe Bryant is considered one of the greatest Lakers and one of the greatest to ever play in the National Basketball Association. Over the course of his 20-year career with the Los Angeles Lakers, Kobe has scored many points and has made even more shot attempts. A dataset obtained from Kaggle competition called “Kobe Bryant Shot Selection” details every single shot Kobe took in his career with features that explain the shot location, shot type, opponent amongst others [1]. With the objective of creating a predictive model using supervised learning through Random Forest, the dataset was cleaned and analyzed to investigate the features that played the most significant roles in predicting the outcome of a shot attempt. Through exploratory analysis, details regarding his preferences in shot location and shot type as well as his output over his tenure was also studied. While Random Forest classification used on a cleaned and normalized dataset was able to create a model that could predict shot outcomes with relative accuracy, other machine learning algorithms must be investigated to create a model that has better accuracy overall in future work.

# Introduction

Kobe Bryant is a well-known name in the NBA and among basketball fans. Considered to be one of the greatest players to ever play the game, Kobe Bryant has shown his abilities on the court through his play style, athleticism, work ethic and championship mentality. In his 20-year career with the Los Angeles Lakers, Bryant has brought the franchise 5 Championships , broken many statistical records and has left a lasting impact on both the franchise and the league. He currently stands in 4th place for most regular season points behind fellow Laker great Kareem Abdul-Jabbar, Karl Malone and LeBron James. Finishing his career with an outstanding 33,643 regular season points, Bryant has taken many shot attempts at crucial times to win games.

This dataset found through Kaggle.com contains details about every single shot attempt Kobe Bryant has in his career and their outcomes [1]. The objective of this project will be to clean the dataset, analyze it and build a predictive model that can successfully predict the outcome of a given shot attempt using the important features. As the objective of this project will be to classify which observations produced “made” shots, Random Forest will be used to build the predictive model. R programming packages such as Dplyr, Tidyverse and GGPlot2 will be used for both data cleaning and exploratory analysis.

# Literature Review

Presently, NBA analytics are mostly consisting of individual player statistics such as points per game, assists, blocks, rebounds etc. While this information is useful for players and coaches, it is still arbitrary in the sense that other key parts of the game are not considered such as defense formations, player to player interaction and ball trajectory. Players and coaching staff use a combination of statistics, observation and game footage to study opponents and build their game plans. The NBA currently uses the STATS SportsVU camera system which can capture 30 frames per second of any game. This increase in available data has allowed the NBA to analyze and visualize other nuances of the game. These types of insights can be used for variety of purposes both within and outside of the NBA. While the scope of this project is within the available statistics presented by the NBA, Data Scientists have begun to study other nuances of the game which will all play key roles in building robust predictive models that are better in analyzing the whole game.

While the act of shooting may be a matter of mechanics and an individual player’s skill on the surface, many maneuvers must be made within a basketball game to create an opportunity to make a shot. A key player in the right position at the right time can decide the outcome of the entire game by a single shot. An example of this can be Kawhi Leonard’s final buzzer beater shot on May 12th,2019 against the Philadelphia 76ers. A paper published by the MIT Sloan Sports Analytics Conference titled “How to Get an Open Shot: Analyzing Team Movement in Basketball using Tracking Data” goes into the in-depth analysis of the 3 seconds before the shooter releases a shot [2]. The paper delves into the changes in roles and team formation that may be required to create an effective shot opportunity. The paper used data procured from NBA STATS SportsVU from the 2012-2013 season for 3-point shots taken by 13 teams. The paper delves into the differences between an “open” shot and a “pressured” shot and the differences in probability of that shot going in. Predictably, the chances of making an open 3-point shot is higher than a pressured one (an instance when there is a defender guarding the shooter between 0 to 8 feet away). The paper also investigates the intricacies of role swaps between players to increase the chances of obtaining an open shot and which players are most likely to switch roles to create an open shot [2]. It is concluded that the player who occupies the Centre position on the team are the likely candidates for a role swap during an offensive play and coaches position their players accordingly to make these changes efficiently. The key take-away from this paper is the series of obstacles faced by the shooter with respect to the movement and distance of defenders before the player makes the decision to make the shot. In terms of the dataset used for this project, information regarding the defender when Kobe made a shot attempt and their distance from Kobe could provide the basis to create a model that is more congruent with the real-world.

Another important facet of basketball is defensive play. In other words, the techniques and strategies used by teams to defend the rim in a defensive play. While it is far easier to quantify the efficiency of the offensive players through tangible statistics such as points and assists, quantifying defensive metrics is more complicated due to the purpose of defense. While the ability to assess individual players and team defensive metrics have improved due to the use of NBA STATS SportsVU, there is a greater amount of statistical information available with respective to offensive plays. A paper published by the MIT Sloan Sports Analytics Conference titled, “ Counterpoints: Advanced Defensive Metrics for NBA Basketball” uses player tracking and spatial regression modeling to create metrics that could be used to effectively quantify defensive metrics [3]. The authors created new metrics such as Volume Score, which measures the total number of shots attempts a defender faces over the course of a game and Counterpoints, which is a weighted average of points scored against a defender per 100 possessions. The further study of these types of metrics will greatly assist with the development of offensive and defensive plays during crucial points in games. These metrics can also improve the NBA’s ability to assess the value of defensive players with better statistics when it comes to player usage and trade value. With respect to Kobe’s shot effectiveness, studying defender information and metrics could further assist in developing a model that is able to predict Kobe’s shot attempt more accurately.

A crucial part of any jump shot is the arc in which the ball travels upon release. This arc is unique to each player with different levels of efficiency based on the game scenario. “Applying Deep Learning to Basketball Trajectory” investigates how to predict the success of a 3-point shot using Recurrent Neural Networks in the form of sequence modeling [4]. The authors of this paper used 20,000 data points from 631 games. The authors focus on strictly the player and not the ball movement as this adds further uncertainty to the model. The authors conclude that RNN had performed better with respect to other statistically approaches. These types of models can be used for many applications in the training of players and study of game footage. An example of where this model could be used is when players and coaches study three-point shots of opponents. This information could be crucial with respect to line up formation as it will give insights on the strengths and weaknesses of opposing players and the best defenders to place against them in during games.

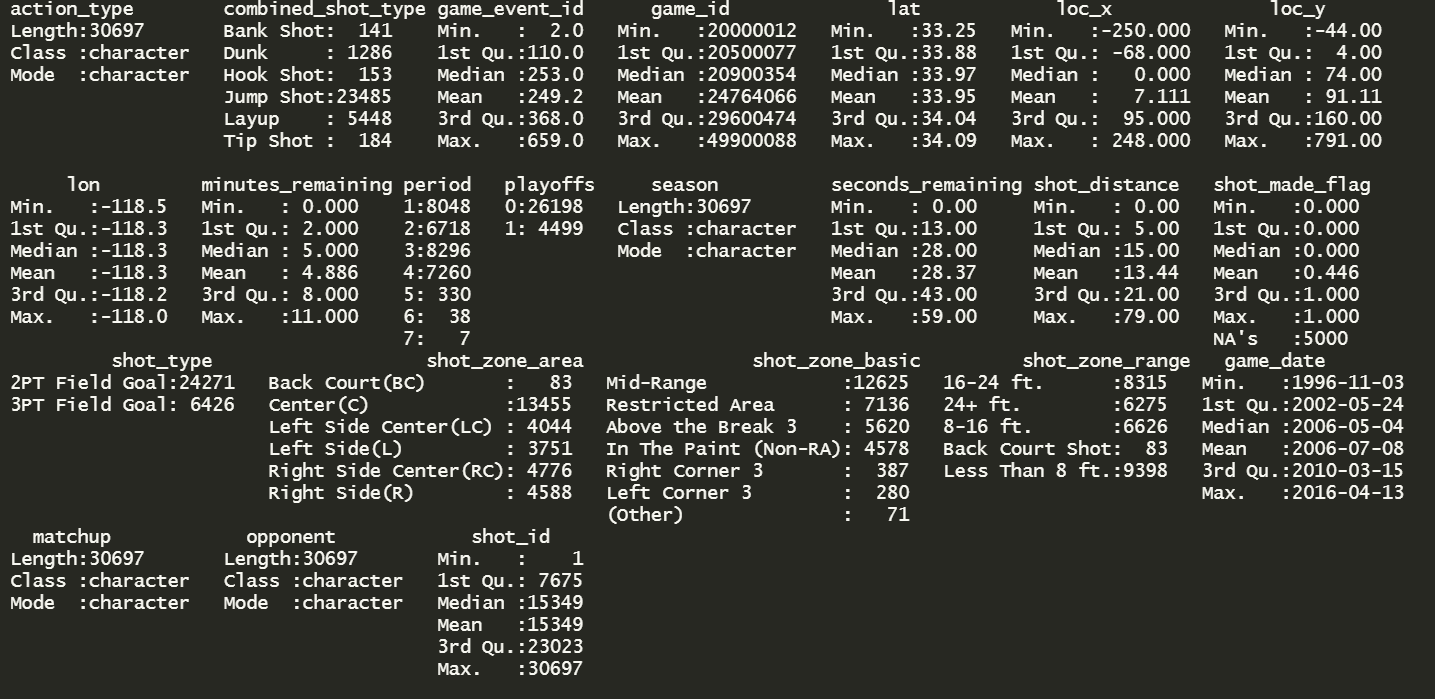
A research article published in the 2017 Optik Journal from Elsevier titled “ Applying deep bidirectional LSTM and mixture density network for basketball trajectory prediction” investigates another way in which shot trajectory can be studied and predicted [5].In terms of practice, the mechanics can be studied to determine the optimal posture for shot making with respect to the ball as well as where a player should aim. When shooting, the physics can be studied with respect to the force and direction when a shot is taken for the best chances to make the shot. This study looks into other aspects of shot trajectory aside from the 2-D location of the shot release such as initial height of the ball, the time it takes for the ball to make it to the basket as well as the course the ball takes. The study goes on to explain that the BLTSM + MDN model has better accuracy overall when predicting 3-point shots compared to other models. This is because while other models may be able to predict the ball's trajectory with comparable accuracy, this model has the capability of creating its own sample trajectories by creating several probability density functions . This type of modelling can one day allow coaching staff to make informed play formations and lineup decisions based on the opposing teams’ strengths and weaknesses. It can also be used to help players visually understand the changes to their posture and shot arc that need to take place for increase accuracy during games.

Beyond shot making ability and defensive maneuvers, individual talent with respect to a team’s lineups and a player’s overall contribution to a team’s success and failure is a crucial metric for coaching staff. An article titled, “ Evaluating Basketball Player Performance via Statistical Network Modeling” attempts to quantify individual players contributions to their teams and their 5-man line ups [6]. Like the other studies noted above, this study also investigates another hardly explored facet in the game of basketball. The authors use network-based algorithms to study how significant individual players are to their 5-man units and how well they perform in their role. The study goes on to build three weighted networks for offensive, defensive and total efficiency. This is quantified using centrality scores that provide a rank for each player studied. The study uses Kobe Bryant as an example to show how certain players possess significantly higher centrality scores in either offence or defense and significantly lower in another while maintaining a relatively high total efficiency. This type of model can be used to effectively choose players are not draft lottery players but have the potential of becoming great role players. These types of models can also help reduce the load on star players as well by providing effective co-stars and role players.

Currently, the NBA relies mostly on individual metrics and certain play by play level statistics to analyze a team's ability , their strengths, weaknesses and likelihood of winning the championship each season. With the introduction of SportsVU cameras in all NBA, there is an unprecedented amount of data that is available for analysis by teams and their coaching staff. Quantifying aspects of defensive players and understanding player contribution at a statistical level can significantly increase the abilities of NBA franchises to make lucrative draft and free agent selections to improve depth and line-up quality. Possessing and understanding how to use predictive models to study shot trajectory and offensive plays to increase opportunities to make open shots can significantly assist coaching staff with assisting in the improvement of a player’s skills. While each of these studies investigates single aspects of the game, with further research, models can be combined to create more robust predictive modelling products. While the scope of this dataset investigates individual offensive statistics, insights into Kobe’s opponents, the team’s he has faced, team line ups, his defensive prowess as well as his maneuverability on the court during every shot attempt can be beneficial in building better a model that can make predictions that resemble reality.

# Dataset Explanation

## Data Statistics Summary



##### Figure 1: Summary of initial dataset

## Lists of Features and Descriptions

The data set used in this project was obtained from Kaggle.com [1]. This dataset contains a record of every shot that Kobe Bryant took during his 20-year long career in the NBA. It contains a total of 23 features and 30697 observations. The following table lists and describes each individual feature found in the data set:

|  |  |
| --- | --- |
| **Feature name** | **Description** |
| Action\_type | Describes the shot type as well as the movement associated when the shot was made |
| Combined\_shot\_type | Summarized description of shot type. 6 types of shots |
| Game\_event\_id | Event occurrence with respect to all events recorded by the NBA |
| Game\_id | Game occurrence with respect to all games recorded by the NBA |
| Lat | Latitude of shot location (geographical court location and location on court) |
| Loc\_x | X-Coordinate on a 2-dimensional Cartesian plane (rim considered as origin) |
| Loc\_y | Y-Coordinate on a 2-dimensional Cartesian plane (rim considered as origin) |
| lon | Longitude of shot location (geographical court location and location on court) |
| Minutes\_remaining | Minutes remaining in the Quarter (regulation=12-minute, overtime=5 minute) |
| Period | The quarter in which the shot was taken |
| Playoffs | Denotes if a game was a playoff game or not (0 means regular season, 1 means playoffs) |
| Season | In which season was the shot was taken |
| Seconds\_remaining | Seconds remaining in the quarter when shot was taken |
| Shot\_distance | Straight line distance between Kobe’s shot location and the rim (origin) |
| **Shot\_made\_flag** | **Attribute to be predicted – (0 = Missed, 1 = Made)** |
| Shot\_type | Points scored in shot (2 points or 3-points) |
| Shot\_zone\_area | General location in which shot was taken (length and width of court used) |
| Shot\_zone\_basic | General location in which shot was taken (width distance of court only) |
| Shot\_zone\_range | General location in which shot was taken (length distance of court only) |
| Team\_id | Team ID number for Los Angeles Lakers (internal NBA codes) |
| Team\_name | Los Angeles Lakers |
| Game\_date | Date on which game was played in (YYYY-MM-DD) format |
| Matchup | Specifies the Opponent as well as game location (home or away) |
| Opponent | Opposing Team during shot attempt |
| Shot\_id | Designation of shot with respect to dataset |

##### Table 1: Features and descriptions.

# Approach

## Approach Diagram

**Predictive Analysis**

**Initial insights and data visualization**

**DATA CLEANING**

**ATTRIBUTE STUDY**

##### Diagram 1: Summarized Approach Diagram

## Approach Details

### *Step 1: Attribute Study*

Features studied individually to assess the initial need and elimination of obvious redundant features. This phase will also include listing obvious issues that should be rectified in the data cleaning phase. Two features that need to be eliminated before data cleaning are Team\_ID and Team\_name as Kobe Bryant only played for a single team in his career.

### *Step 2: Data Cleaning and Initial Insights*

1. Initial assessment of dataset via MS Excel.
2. Import dataset into R Studio
3. Conduct initial descriptive statistics to find potential missing values and outlier data
4. Check the classes of each and assess changes that need to be made
5. Eliminate features that are deemed unnecessary for the study
6. Assess correlation between features
7. Normalize data
8. Split dataset into Test and Training sets

### *Step 3: Insights and Data Visualization*

1. Visualize the data for shot location with respect to shot\_made\_flag feature
2. Determine the areas on the court with highest shot concentration
3. Visualize the occurrences of each shot type
4. Visualize other aspects of shot making in the dataset such as opponents, location of the game, points made during each quarter, times in which points are made with respect to the clock etc.

### *Step 4: Predictive Analysis*

1. Run Random Forest algorithm
2. Assess OOB error rate (out-of-bag)
3. Tune hyperparameters (number of trees (mtrees) and splits of variables (mtry))
4. Create a confusion matrix to assess accuracy precision and recall

# Data Preparation

## Issues with Dataset

|  |  |
| --- | --- |
| **Attribute name** | **Data Cleansing action taken** |
| Action\_type | Character class – **changed to Factor, Formalize the word structure** |
| Combined\_shot\_type | Character class – **changed to Factor, Formalize the word structure** |
| Game\_event\_id | Internal NBA code, insignificant to the scope **(eliminated)** |
| Game\_id | Internal NBA code, insignificant to the scope **(eliminated)** |
| Lat | Latitude is between 33.2533 and 34.0883  Latitudes will be calculated in decimals **(no action taken)** |
| Loc\_x | X coordinate is between -250 and 250, dataset does not have a uniform measure of distance. Based on regulation NBA court sizes (50 ft by 90 ft), it is evident loc\_x has been increased by a factor of 10. **Dividing each observation by 10 will indicate the values in feet (ft).** |
| Loc\_y | Y coordinate is between -44 and 791, dataset does not have a uniform measure of distance. Based on regulation NBA court sizes (50 ft by 90 ft), it is evident loc\_y has been increased by a factor of 10. **Dividing each observation by 10 will indicate the values in feet (ft).** |
| lon | Longitude is between -118.5198 and -118.0218   * Latitudes will be calculated in decimals **(no action taken)** |
| Minutes\_remaining | Numbers are in decimals, there is no uniform measure of time in dataset. - **Convert to seconds** |
| Period | Character class – **changed to Factor** |
| Playoffs | Character class – **changed to Factor** |
| Season | Split in various data types such as character, dates and numbers. (ex: 1996-97, Jan-01 etc.),  **-replace with Uniform designation for seasons** |
| Seconds\_remaining | **No action taken** |
| Shot\_distance | Distance measured in feet; **no action taken** |
| **Shot\_made\_flag** | 5000 missing values |
| Shot\_type | Character class – **changed to Factor** |
| Shot\_zone\_area | Character class – **changed to Factor** |
| Shot\_zone\_basic | Character class – **changed to Factor** |
| Shot\_zone\_range | Character class – **changed to Factor** |
| Team\_id | Only one team id for Kobe Bryant **(eliminated)** |
| Team\_name | Internal NBA code, insignificant to the scope **(eliminated)** |
| Game\_date | Includes month, day and year in one feature. Could potentially be more useful as separate features as factors |
| Matchup | Specifies 2 things:   * Home or away * Team being faced |
| Opponent | Character class – **changed to Factor** |
| Shot\_id | Does not influence the outcome of Kobe’s shot attempt **(eliminated)** |

##### Table 2: Problems that need to be addressed for data cleaning

## Major Feature Changes

### *Time Features*

The dataset specifies two separate features to represent the time in the game when a shot attempt was made. These features are minutes\_remaining and seconds\_remaining. For the ease of use during predictive analysis, minutes\_remaining was converted into seconds by multiplying each observation by a factor of 60. A new feature called “time remaining” was created by adding both time-oriented features. As such, the original features were eliminated.

### *Shot Distance Features*

Loc\_x, Loc\_y and shot distance all describe the locations on the court from where shots were taken but do not follow the same unit of measure. According to the NBA, the measurements of a regulation court is 50 ft by 90 ft. When the maximum and minimum values of loc\_x and loc\_y were studied they were -250 to 250 and -44 to 791 respectively. When the loc\_x and loc\_y values are divided by a factor of 10, the maximum and minimum values than become -25 to 25 and -4.4 and 79.1 respectively. When these points are plotted on a cartesian plane, the relationship between these three features can be described by the Pythagorean theorem. In other words, sqrt((loc\_x^2) + (loc\_y)^2) = shot\_distance. As such, loc\_x and loc\_y was divided by a factor of 10 to have all three features follow a single unit of measurement.

### *Season Feature*

This feature contained observations that denote season in two separate ways. Observations were either in a YYYY-YY format (ex: 1996-97) or in an MMM-YY format (Jan-01). In order to standardize the observations, these two formats were replaced with a numerical designation between 1 and 20. The class of this feature was also changed to a factor. This was done to prevent redundancy in the dataset because the game\_date feature already contains the year in which the shot attempt was made.

### *Date Feature*

The game\_date feature contains 3 separate pieces of information – Year, Month and Day. Kobe’s performance overall changed from year to year. Kobe’s output, with respect to the number of plays that featured Kobe and subsequent shot attempts, was higher during seeding games (regular season games in March and April which factor into playoff contention ) and during the playoff runs. The game\_date feature was split into two separate features called “game\_month” and “game\_year” and the original game\_date feature was eliminated. These new features were also made into factors.

### *Matchup Feature*

The matchup feature contained two pieces of information – the name of the opponent and if the game was played at the Lakers home arena or away from it. It should also be noted that the Lakers played in two home arenas during Kobe’s tenure: The Forum and Staples Centre. As the opponent variable exists in the dataset, it became redundant to have the opponents name in this feature. In order to minimize the need to specify “The Forum” or “Staples Centre”, observations containing “LAL @”, which denote the Lakers playing in another city, were replaced with “Away”. Observations containing “LAL vs.”, were replaced with “Home”. This change was made so the relationship between Kobe’s shot effectiveness with respect to game location (home or away) could be studied. This feature was turned into a factor.

### *Opponent Feature*

During Kobe’s career, there were a few team relocation and name changes. These major changes were:

* New Jersey Nets relocate to Brooklyn to become the Brooklyn Nets in 2012
* Vancouver Grizzlies relocate to Memphis become the Memphis Grizzlies in 2001
* Seattle Supersonics relocate to Oklahoma City to become the Oklahoma City Thunder in 2008
* New Orleans Hornets become the New Orleans Pelicans in 2013

For the sake of uniformity, observations containing information regarding the New Jersey Nets, Vancouver Grizzlies, Seattle Supersonics, and the New Orleans Hornets were replaced with the names of their relocations.

### *Features to be eliminated*

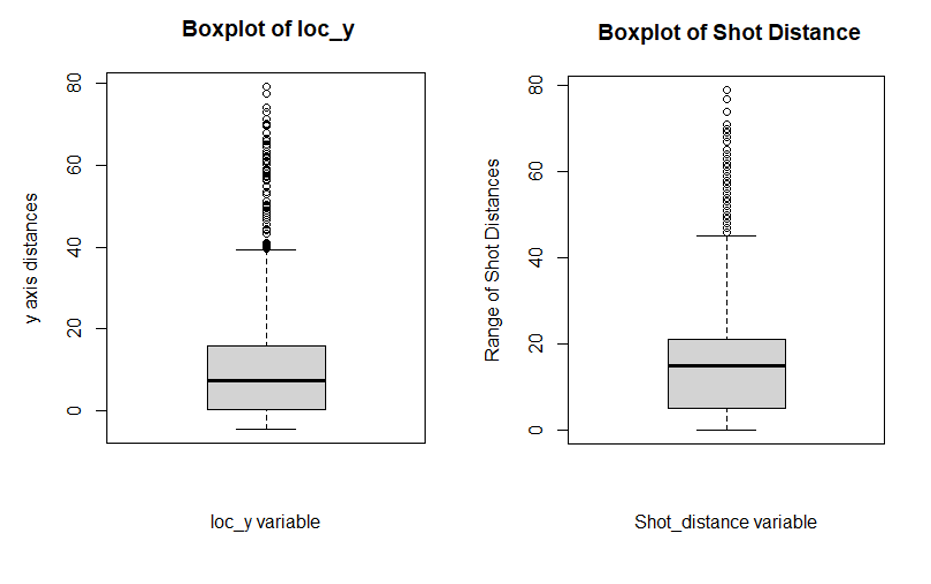
The game\_event\_id, game\_id, team\_id and team\_name features were eliminated from the dataset because they have no influence in Kobe’s shot attempt. These features are all related to internal NBA codes that span all the teams in the league and all the games played since the league started recording statistical information. Shot\_id only serves as a number designation in the dataset and does not factor into the outcome of Kobe’s shot.

## Missing Value Investigation

As per Figure 1, there are 5000 missing values that are only in the “shot\_made\_flag” feature. This dataset is from a Kaggle competition where a predictive model will need to be used to predict these missing values. Contestants would predict these missing values in the shot\_made\_flag feature and submit these. Since these missing values only account for roughly 16% of the dataset, these values were removed from the dataset leaving a total of 25,697 observations for analysis.

## Boxplots & Outlier Data

### Cartesian Distance Boxplots



##### Figure 2 & 3: Boxplots of loc\_y and shot\_distance showing significant outlier data

As per the boxplots of Cartesian Distance Features, it is evident that loc\_y has many outliers. Through analysis it was determined that there was a total of 76 outlier points in the data for this feature. The range of loc\_y is between -4.4 and 79.1 feet. The size of a regulation NBA court is 50 feet by 90 feet. In this case half court will be located 45 feet from either rim. Observations that are more than 45 feet suggest shots being taken from beyond half court.



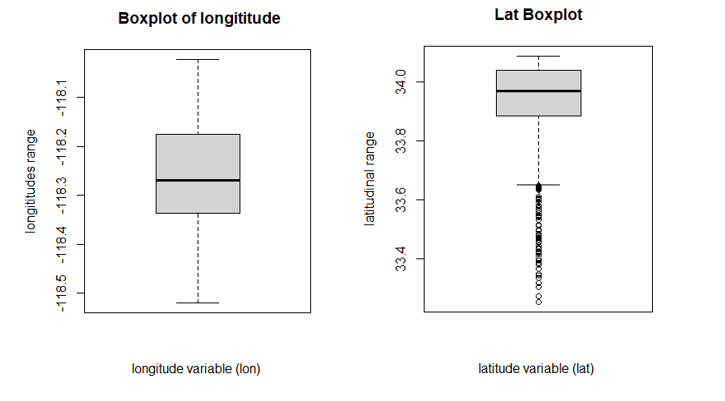
##### Figure 4: Distribution of loc\_y feature

Figure: 4 suggests that most of Kobe’s shot attempts take place between 0 feet and 16 feet from the rim. This can be validated by Kobe’s aggressive style of play which encompassed mainly of jump shots near the rim. In the NBA, the three-point line is situated 23.75 feet from the rim.

A percentile test was used to determine when 98% of Kobe’s shot attempts were made. The data suggested 98% of Kobe’s shots were made between -1.8 feet and 27.1 feet from the rim. This percentile test considers Kobe’s three-point shot attempts as well. This percentile test however does not consider the 76 outlier observations. These outliers were found to be actual observations and not mistakes in the data collection. When the time\_remaining feature and loc\_y feature was compared, it became evident that these shots were made when the time\_remaining was between 0 and 5 seconds. These shots were essentially unsuccessful shot attempts made as the game clock approached 0.

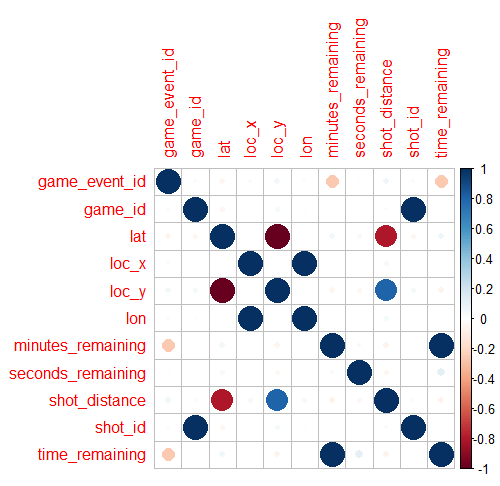
As these outliers were not caused by input error and were correctly recorded and were kept as valid data for the predictive analysis. This information was also found in the Lat and shot\_distance features.

### Global Positioning Boxplot



##### Figure 5 & 6 : Boxplots for Latitude and Longitude during shot attempt

## Correlation Matrix



##### Figure 7: Correlation Matrix before dimension reduction

Strong Positive Correlation:

* **Shot\_distance and loc\_y:** The reason behind the strong positive correlation between these features is due to them measuring distance from the rim in the same direction. Loc\_y represents distance up and down the court while shot\_distance represents the distance in a diagonal (as per the relationship noted between loc\_x, loc\_y, shot\_distance)
* **Time\_remaining and minutes\_remaining:** the time\_remaining feature was created using the minutes\_remaining feature
* **Lon and loc\_x:** both units of measure are measured left to right (east and west)
* **Shot\_id and game\_id:** both are numerical designation of a game event

Strong Negative Correlation:

* **shot\_distance and lat:** both are measured north/south or lengthwise. Lat however has negative coordinates while shot\_distance has only positive coordinates.
* **loc\_y and lat:** both are measured north/south or lengthwise. Lat however has negative coordinates while shot\_distance has mostly positive coordinates.

Due to there being features that represent the same or similar aspects of Kobe’s shot outcome as well as there being features that play minimal role in the outcome of Kobe’s shot such as shot\_id and game\_id, it is expected the correlation between features will change significantly with their removal.

# Exploratory Analysis

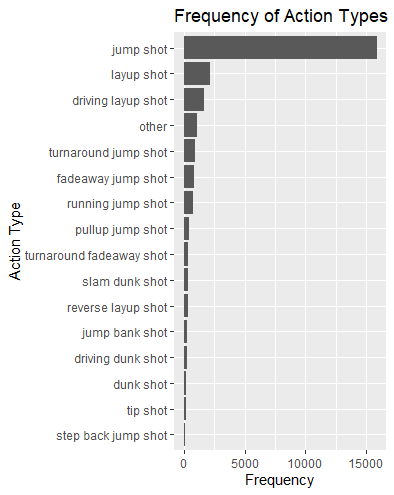
## Action type

As noted in Table 1, Action\_type provides greater detail of the type of shot Kobe took in each attempt. This feature has 57 distinct type of shots that Kobe has taken in a shot attempt. The descriptions provided the type of shot taken (out of the 6 types) plus a description of the movement taken with the shot. For example, there are 83 observations in the dataset that indicate a “driving reverse layup shot” was made. Layup shot would be the actual shot type whereas the “driving reverse-“ would be the movement Kobe made to position himself to make the layup shot.



##### Figure 8: Descriptors and frequency of shots attempted by Kobe

As Figure: 8 shows, there are many action\_type observations that have thousands of instances such as “jump shot” and “layup shot”. There are also many instances such as “turnaround hook shots” or “running tip shot” that have less than 10 observations. In order the reduce redundancy ,it was determined that any action\_type with less than 100 instances will be combined into a new action type called “other”.



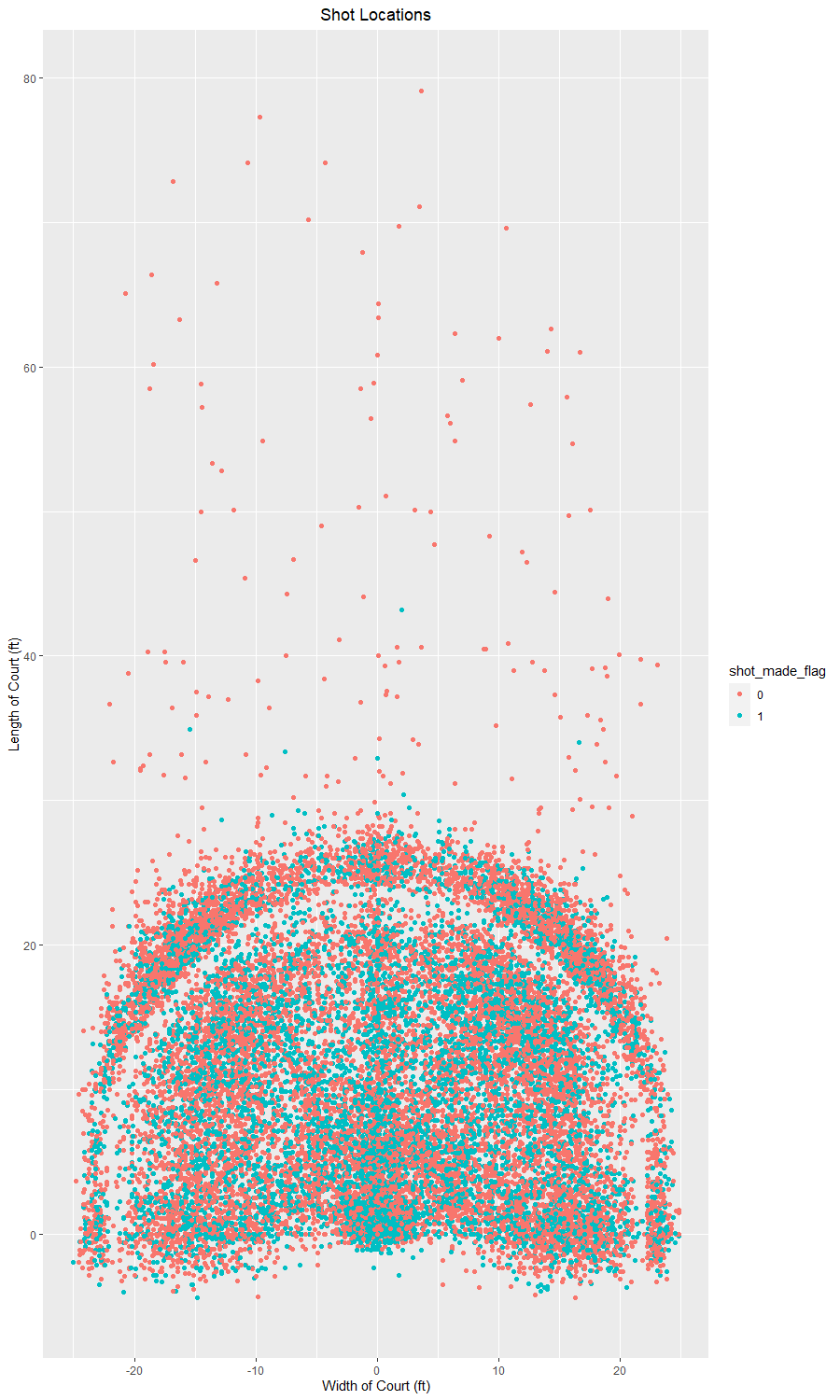
##### Figure 9: Action type Bar Graph

As Figure: 9 shows, combining all action\_type instances that have less than 100 observations reduces the total action\_types to 17 types.

* Kobe’s preferred choice of shot are Jump Shots, Layups, and driving layups.
* While the step back jump shot is used mainly for 3-point shot attempts, Kobe was known to take more of these shot attempts later in his career as the trend became more prevalent in the league. Kobe also preferred to take a step back before attempting a 3-point shot usually avoiding the 3-point line all together. This is done by players to both create space between them and the defender as well as insuring they are behind the line to guarantee 3 points if the shot was successful.

## Shot Location Visualization

### Shot location

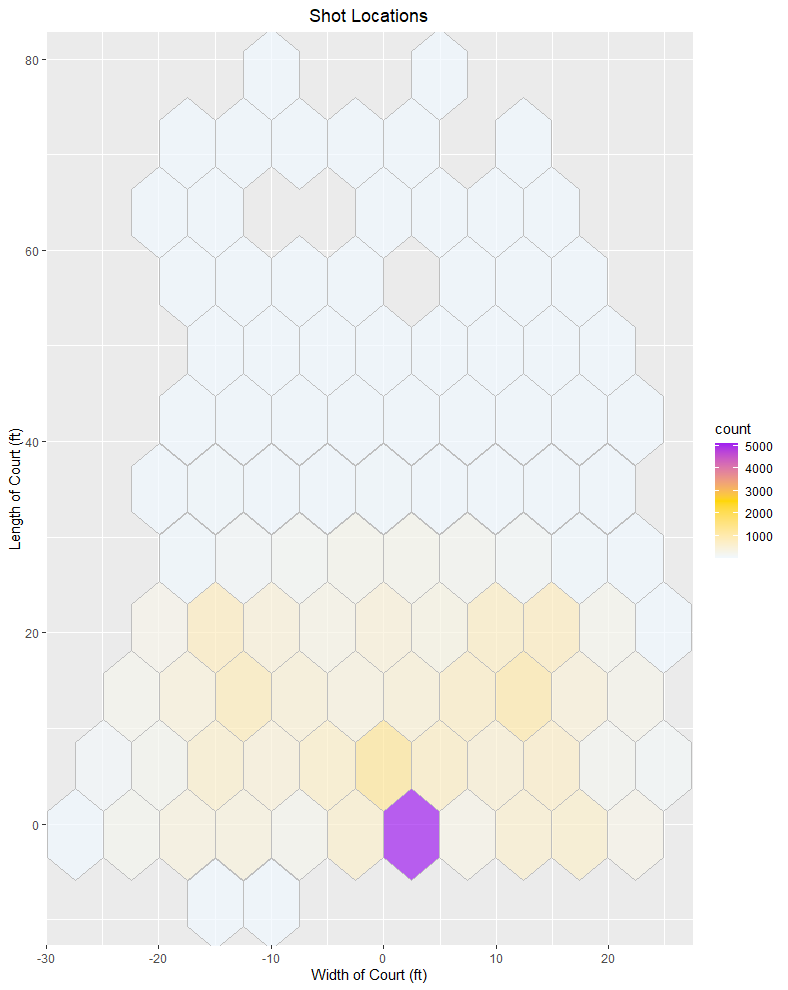


##### Figure 10: Shot location vs shot flag

This plot visualizes all loc\_x and loc\_y locations from which Kobe took shots. The red shapes represent the shots made while the blue shapes represent shots missed. Due to the sheer number of shot observations, the plot above provides little information with respect to the places where Kobe shot with higher accuracy. The plot does however convey two pieces of information.

* Most of the backcourt shots (beyond 47 feet away from the rim) were missed and were attributed to last second shots made before the expiration of the game clock.
* This plot also shows Kobe’s preference to take 3-point shots a couple of feet behind the 3-point line

### Shot Location “Hexbin” Plot

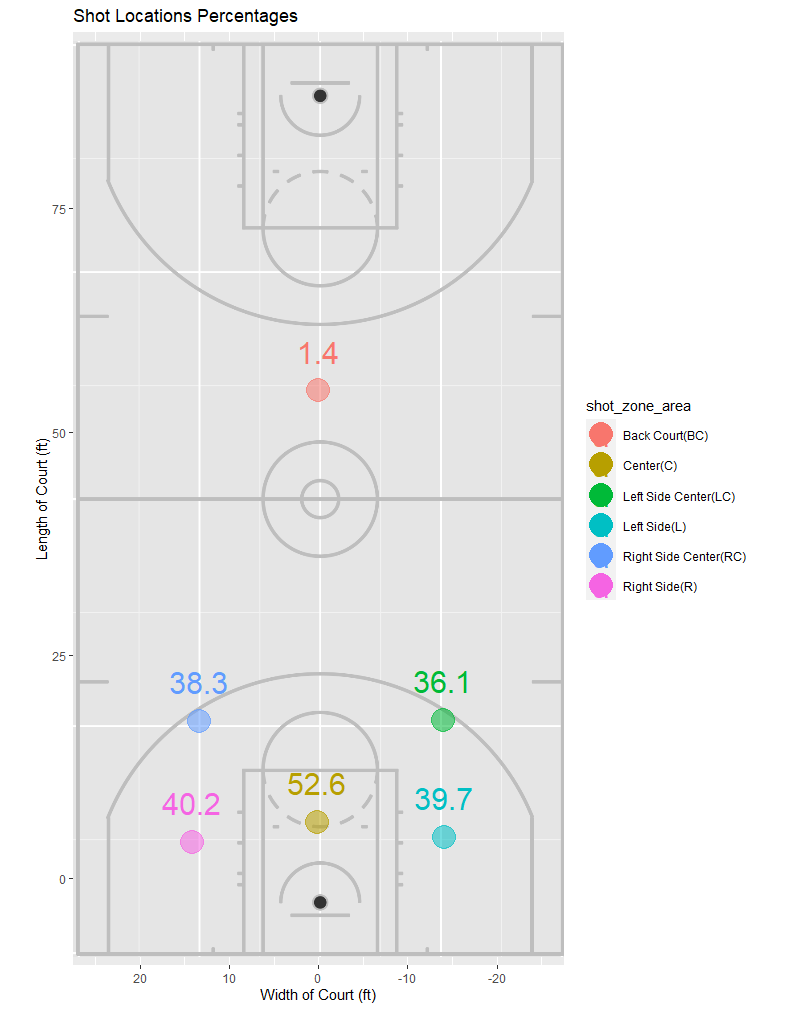


##### Figure 11:Hexbin plot to show shot area frequency visually

Figure: 11 provides more information about Kobe’s frequency in shot location. It must be noted that Kobe played in an era of the NBA where more physical play or “bully ball” was common and the three-point shot was uncommon when compared to the current era. Figure 11 provides a visual representation of the frequency of Kobe’s shot location.

* Most of Kobe’s shots were taken very close to the rim which mostly comprised of jump shots, layups and dunks
* Most of the three-point attempts made also appear to be around the Left Side Centre and Right-Side Centre. There were less corner-3 attempts (shots made from the left or right corners)
* Most of Kobe’s non three-point attempts were made in the center area

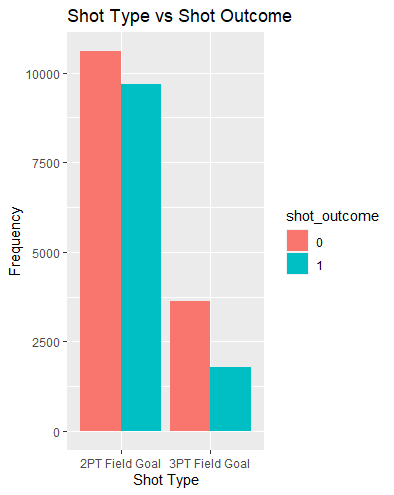
### Shot Location Percentage Plot



##### Figure 12: Shot Accuracy in Each Zone Area

To further illustrate the information provided by Figure:11, This plot further elaborates on Kobe’s accuracy in each part of the court. With respect to “Back Court” shots, Kobe has a 1.4% accuracy. This makes sense as these shots (76 outliers) were all shots made at the last moment when the game clock was approaching zero. This graph proves that most of Kobe’s shots and his most accurate shot location is in the “Centre” zone. Based on the frequency of shot types, these shots would be “Jump Shots” or “layups”. Kobe also has a preference in taking shots from the right side of the court with respect to 3-point shots at 38.3% of total successful shots. Kobe’s accuracy on the left and right side while in the paint is close to the same.

## Shot Type

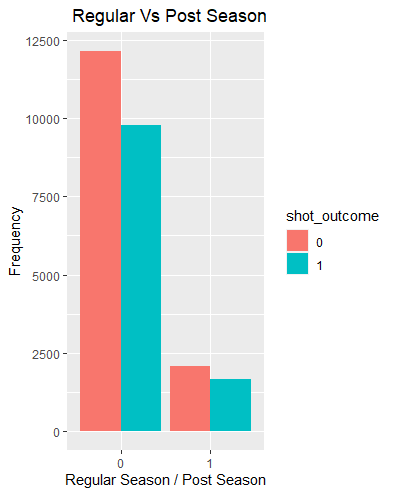


##### Figure 13: Frequency of 2 PT and 3 PT shots

When the shot type frequency and shot outcome are compared, there are three things that should be noted:

* Kobe took significantly more 2 PT Field Goal attempts when compared to 3 PT Field Goals. This also verifies Figure 11 where it is noted that Kobe makes a higher concentration of shots that are closer to the rim using moves such as Layups or Jump Shots.
* It should be noted that the frequency of 2PT Field Goal shots made is slightly less than the ones not made. The accuracy of this plot can be verified through the official ESPN site which states his overall field goal percentage (FG%) is about 44.7 percent. [7]
* This plot also shows that Kobe’s overall 3-point efficiency was less as well. While he primarily took jump shots over his career, they were closer to the rim. A possible reasoning for this type of play can be attributed to the more physical nature of basketball during the prime of Kobe’s career (2000 to 2010) and the prominence of “Big Men” who typically filled the role of Centre (C) with roles in defensive and rim protection. Kobe is known to be quick and physical despite his stature compared to “Big Men” like Shaquille O’Neal or Dwight Howard. Kobe preferred to maneuver himself closer to the net and take jump shots in order to increase space between him and the defender.

## Regular Season vs Playoffs



##### Figure 14: Regular season shot attempts vs Post Season

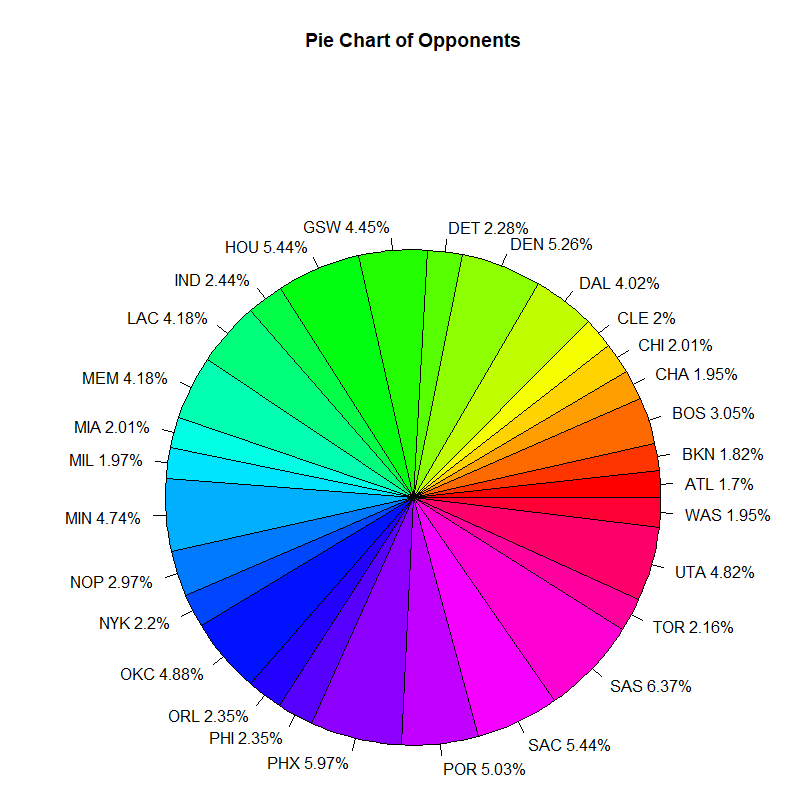
In his career, Kobe has taken the Lakers to the playoffs for 17 years, took them to the finals 7 times (10 total times to the playoffs without making a finals appearance) and was not in playoff contention 4 times total.

* The considerably lower frequency of shots taken during the post season can be attributed to:

1. Kobe being considered the second option of the team during the first half of his career as he played those years with the more prominent (at the time) Shaquille O’Neal who had more ball usage for those seasons.
2. Kobe had less appearances in the post season in the 2nd half of his career (after 2011) due to injury and age especially his final 3 seasons.

* While less overall shot attempts were made in the post season, his overall efficiency as visually higher in the post season when compared to regular.

## Opponents



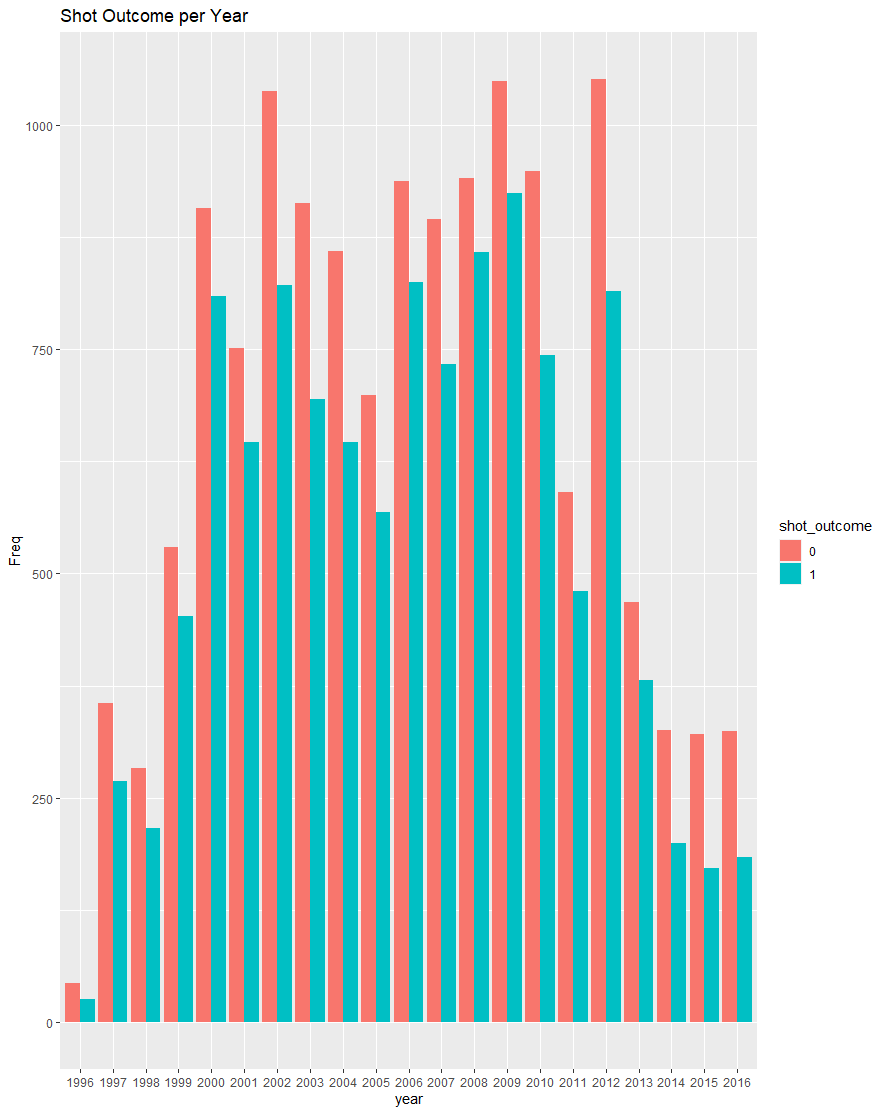
##### Figure 15: Number of shots made against each team

Insights into Kobe’s opponents:

* When not considering shorten seasons due to lockouts or other reasons, a regular NBA season is 82 games total. The 30 teams of the NBA are split between two conferences (15 teams per conference)
* Each team will play 4 games (2 home/2 away) with teams in their conference and 2 games with the other conference (1 home/1 away)
* Kobe played 67% of his games in the western conference frequently facing the following teams: (GSW,HOU,LAC,MEM,MIN,NOP,OKC,PHX,POR,SAC,SAS,UTA,DAL,DEN)

## Game Date

### Game Year



##### Figure 16:Shot attempts/results per year

Insights into Shot outcome by year:

* Kobe had significantly less ball usage and shot attempts earlier in his career and at the end of his career due to 1996 to 1999 being his initial years in the league and 2013 to 2016 due to injury.
* The drop in production in the year 2005 was due the team did not make the playoffs during this time

## Dimension Reduction

### Shot Description Features

The dataset contains the following features which describe Kobe’s shot:

* action\_type
* **combined\_shot\_type.**

While combined\_shot\_type distilled shot description to 6 types, action\_type provides the shot type and more information about the movements before the shot was taken. As such combined\_shot\_type was removed from the dataset used for predictive analysis

### Shot Location Features

The dataset contained the following features to describe shot location on the basketball court:

* loc\_x
* loc\_y
* **lon**
* **lat**
* **shot\_distance**

While loc\_x and loc\_y described the court in the form of a cartesian plane, lon and lat describe the court with respect to world. The lon and lat features were removed from the dataset as they described shot locations in decimal places which would make understanding results cumbersome. As loc\_x , loc\_y and shot\_distance is described in feet, it made more sense to use these features as they better describe the court.

Shot\_distance while related to loc\_x and loc\_y only provided a straight distance with the rim as a starting point. This feature is also related to loc\_x and loc\_y through the Pythagorean theorem. As loc\_x and loc\_y describes the shot location adequately, lon, lat and shot\_distance was removed from the dataset.

### Shot Location Description Features

The dataset contained the following three features to describe the basketball court in layman terms and basketball jargon:

* **Shot\_zone\_basic**
* **Shot\_zone\_range**
* Shot\_zone\_area

While shot\_zone\_area describes the entire length and width of the court, shot\_zone\_range describes only the length. Similarly shot\_zone\_basic describes only the width. As the shot\_zone\_area feature describes the entire court as a single feature, shot\_zone\_basic and shot\_zone\_area was removed from the dataset for predictive analysis.

### Time Related Features

The dataset contains the following features to describe the passage of time on the court. These features are:

* **Minutes\_remaining**
* **Seconds\_remaining**
* Time\_remaining

Minutes\_remaining and seconds\_remaining are features that contain two different units of measure. As time\_remaining is a combination of these features in a single unit of measure, the need for these features become redundant for use in the predictive model. Minutes\_remaining and seconds\_remaining were removed from the dataset.

### Unrelated Features

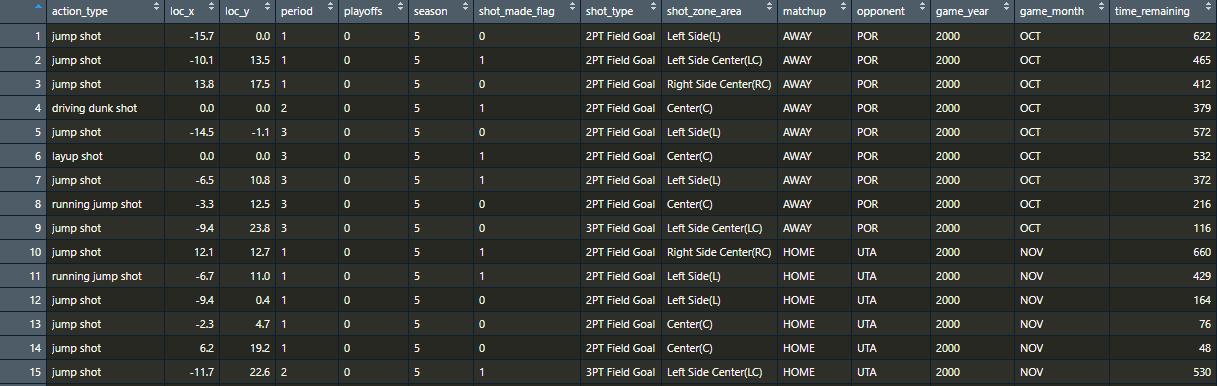
The following features do not have a significant role in the outcome of Kobe’s shot attempt. These features are row designations and features related to the Los Angeles team and the NBA. The following features were removed from the dataset as they have no effect on Kobe’s ability or shot attempt probability.

* **Game\_id**
* **Game\_event\_id**
* **Shot\_id**
* **Game\_date**
* **Team\_id**

# Results

## Predictive Analysis

### Features to be used for Predictive Analysis



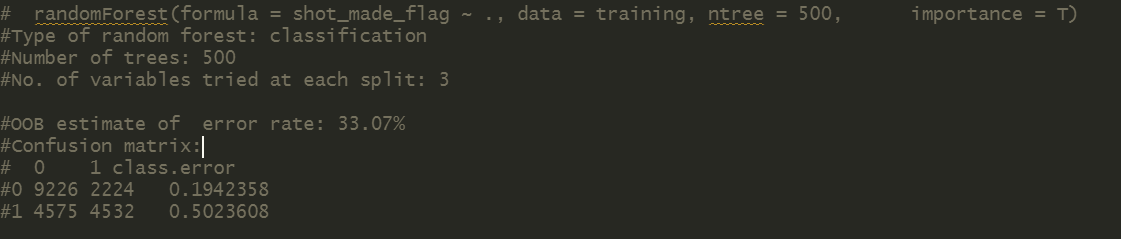
##### Figure 17: Features to be used in Predictive Analysis

### Training Set

Once the redundant features were removed from the dataset, the dataset was normalized. Uniform normalization was used to further minimize data redundancy and improve integrity. The various scales of measure numeric features were also the reason uniform normalization was used.

Once normalized, the data set was split into training and test sets.(80 percent training and 20 Percent) The feature to be predicted is the shot\_made\_flag which consists of two classes: shot made=1 and shot missed=0. N-Fold cross validation was not used in this situation as out-of-bag performance (OOB) cross validation performed the same function when training the model with the training set.

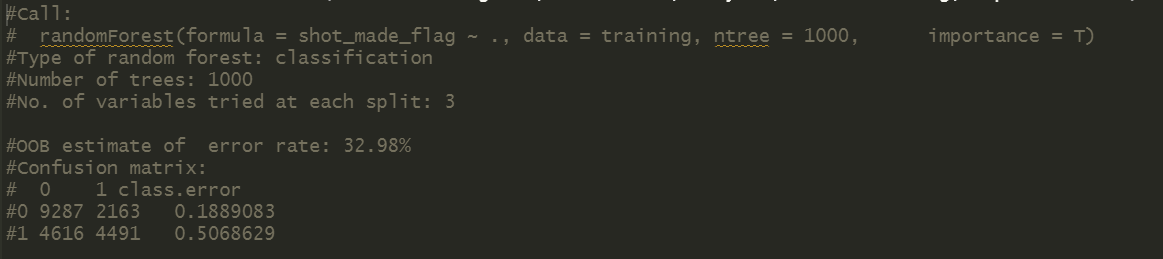
#### Training set – Hyperparameters (ntrees=500, mtry=3)



##### Figure 18: Initial results from predictive model with default hyperparameters

The first step in the predictive modeling process was to run the Random Forest classifier on the training set with the RStudio default hyperparameters of 500 trees with 3 variables tried at each split. When the model was run multiple times under these hyperparameters, the average OOB error rate was averaged out to be 33%. In order to obtain a lower error rate, the “ntrees” hyperparameter was increased to 1000 trees total.

#### Training set – Hyperparameters (ntrees=1000, mtry=3)

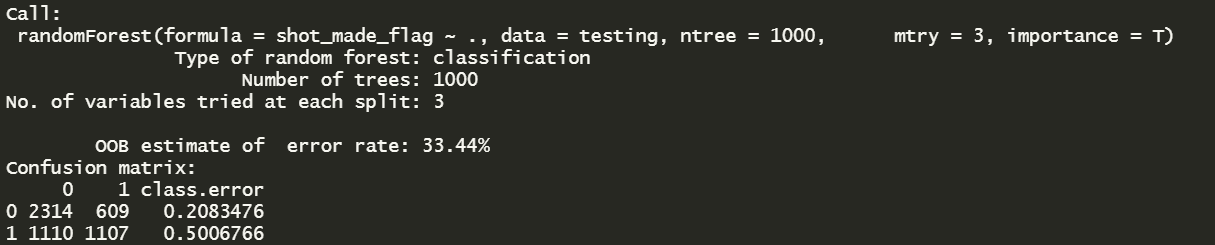


##### Figure 19: Results from predictive model with tuned hyperparameters

The model has a slightly less error rate when the hyperparameters were tuned. When the error rates between trees is visualized, the OOB error rate tapers off at 33% when more than 700 trees are used.

### Test Set

#### Hyperparameters(ntrees=1000, mtry=3)



##### Figure 20: Results from predictive model with tuned hyperparameters on Test set

When the tuned training set hyperparameters are used for the test set of data, the error is slightly higher at 33.44%. The OOB error rate being similar between test and training set suggests there is no overfitting or underfitting. To further test the error rate, the model was conducted on the test set again using a 70/30 split between training and test sets. The error rate has a slight increase approaching 34% but does not change by a huge factor overall.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Shot Missed (predicted) | Shot Made (Predicted) | Class Error |
| Shot Missed (Actual) | 2314 (TN) | 609 (FP) | 0.1986029 |
| Shot Made (Actual) | 1110 (FN) | 1107 (TP) | 0.4963235 |

##### Table 3:Confusion Matrix of Test set results

* **True Negative:** (Scenarios when the shot was missed, and the predictive model also predicted this correctly) 🡪 2314 instances
* **False Negative:** ( Scenarios when the shot was made but the predictive model stated it was missed) 🡪1110 instances
* **False Positive:** (Scenarios when the shot was missed but the predictive model stated it was made) 🡪 609 instances
* **True Positive:** (Scenarios when the shot was made, and the predictive model also predicted this correctly) 🡪 1107 instances

#### Precision/Recall

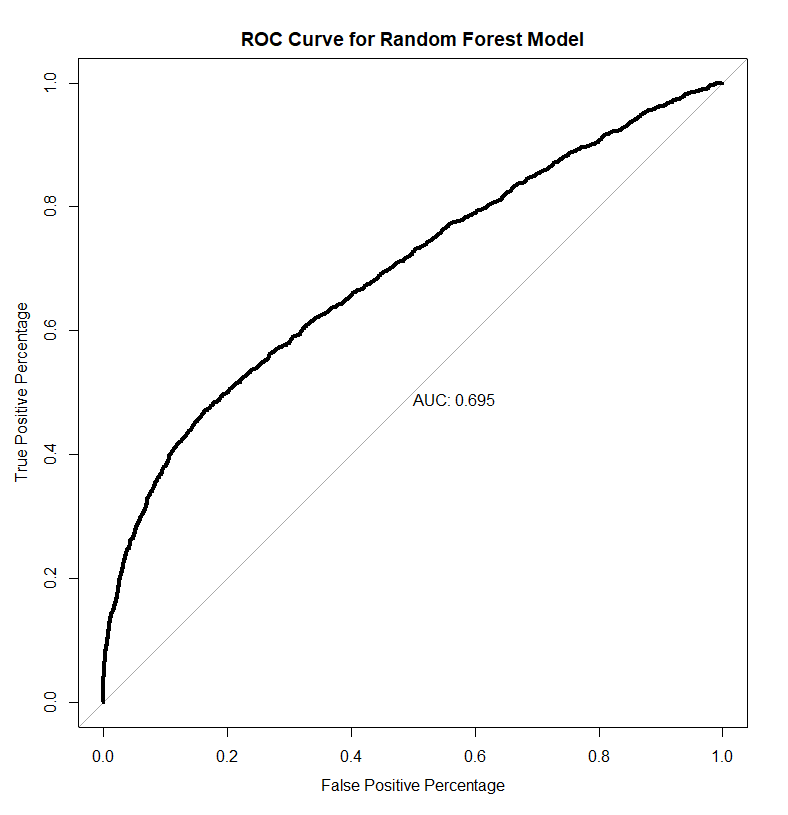
* **Precision**= TP/(TP+FP) = 1107/(1107+609) = 0.645 🡪 **65%**
* **Recall**= TP/(TP+FN) =1107/(1107+1110) = 0.499 🡪 **50%**
* **F1** = 2\*((precision\*recall)/(precision+recall))

= 2\*((0.3219)/(1.144))

= 2\*(0.2814)

= **0.5728**

### Receiver Operating Characteristic Curve (ROC) & Area Under Curve (AUC)



##### Figure 21: ROC & AUC

The ROC curve shows an area under curve of 0.695 ~ 69.5%. This indicates that there is a 69.5% chance that the model will be able to distinguish between the positive (shot is made) and negative (shot is missed) class.

# Post Predictive Analysis

The dataset found in a Kaggle Competition had a total of 23 features and 30697 observations which contained 5000 blanks in the shot\_made\_flag feature. Contestants were supposed to build predictive models to predict these blanks as submissions to win the competition. This project removed these missing values and built a supervised machine learning model instead with the remaining 25697 observations.

While the model has some accuracy, there is still room for improvement with respect to precision and recall. Actions to improve the overall accuracy could be taken via experimenting with other supervised learning models such as K-Nearest Neighbor or Naives Bayes. Unsupervised learning in the form of clustering allows one to make suggestions on the best parameters that need to be in place for Kobe to make a successful shot. Similar to the literature reviewed, studying aspects of the game that cannot be readily quantified such as defense formations and opponents could open new facets of the dataset that would be worth exploring to build a model that is closer to predicting real life scenarios for Kobe Bryant and by extension other players in the NBA.

# Significance of this Study

One of the main objectives of the video game industry and game developers are to bring the mechanics and visuals of their games closer to real life. Such examples are EA’s NBA 2k series who is working towards building player models that are closer visually to real life as well as play style. Sports games have come a long way and now show off significant differences between the play style and power levels between players to differentiate teams and individual players. Building robust and more encompassing predictive models can help significantly with game developers’ abilities of building games that mimic individual basketball player’s tendencies and play styles in the games. This will ultimately make the gaming experience closer to real life for consumers.

Further study into sports analytics can open new avenues in the entertainment industry as well. While farfetched currently, an example predictive models using sports analytics of could come in the form of VR. A current topic amongst NBA fans is: “Who is the greatest player of all time?”. With enough data and predictive power/accuracy, fabricated games could be created pitting superstars of different eras against each other as if it were a real-life sporting event. The opportunity to watch matchups such as Michael Jordan vs. Larry Bird, or Lebron James vs Michael Jordan or even, Kobe Bryant vs Michael Jordan in their primes can become a reality. Predictive models coupled with increasingly realistic graphics can be used to create almost simulations of these matchups in an arena like environment through VR. While this type of technology is far away, predictive modeling and the study of sports statistics can open new avenues for business and entertainment outside of traditional sports entertainment.

# Conclusion

The scope of this project was to take a dataset that contained information on every shot attempt Kobe Bryant has made in his career and build a predictive model that is successfully able to predict the shot outcome. After cleaning the data and understanding the features. Insights into shot locations, shot type, points scored amongst others were studied to determine which features were important and which were redundant to the model. The model was built using the random forest classifier algorithm and was able to achieve an overall accuracy close to 65%. Further work upon this project can be done by building models using other supervised learning algorithms such as KNN (K-nearest neighbor) or Naives Bayes with the objective of improving the overall accuracy.

While this project is limited to the scope of the data procured, this project could be viewed as a small part or steppingstone to building models that could make more complicated predictions that are closer to reality. While the model built for this project as well as the models mentioned in the literature reviewed study specific parts of the game and a single players stat, further research and study of data can assist in building models that consider more than one aspect of the game. Furthermore, models can be used to build models for other sports, leagues and industries in general.

# References

|  |  |
| --- | --- |
| [1] | National Basketball Association, "kaggle.com," National Basketball Association, 2016. [Online]. Available: https://www.kaggle.com/c/kobe-bryant-shot-selection. [Accessed September 2020]. |
| [2] | P. Lucey, A. Bialkowski, P. Carr, Y. Yue and I. Matthews, ""How to Get an Open Shot": Analyzing Team Movement in Basketball using Tracking Data," Disney Research, Pittsburgh, 2014. |
| [3] | A. Franks, A. Miller, L. Bornn and K. Goldsberry, "Counterpoints: Advanced Defensive Metrics for NBA Basketball," MIT Sloan Sports Analytics Conference, Boston, 2015. |
| [4] | R. C. Shah and R. Romjinders, "Applying Deep Learning to Basketball Trajectories," ArXiv, Chicago, 2016. |
| [5] | Y. Zhao, R. Yang, G. Chevalier and R. C. Shah, "Applying deep bidirectional LSTM and mixture density network for basketball trajectory prediction," Elsevier GmbH, 2017. |
| [6] | J. Piette, L. Pham and S. Anand, "Evaluating Basketball Player Performance via Statistical Network Modeling," MIT Sloan Sports, Analytics Conference , Boston, 2011. |
| [7] | ESPN, "ESPN.com," ESPN Enterprises, 2020. [Online]. Available: https://www.espn.com/nba/player/\_/id/110/kobe-bryant. [Accessed 20 11 2020]. |

## Github link

https://github.com/Anujan429/Kobe-Shot-Analysis