MSDS 6372 - Project 3

Kobe Bryant Shot Selection

Author:
Evangelos Giakoumakis

Introduction

Kobe Bryant is widely regarded as one of the greatest basketball players of all time. He was born August 23, 1978 in Philadelphia Pennsylvania. Was drafted right out of high school at the age of 17 (1996) by the Charlotte Hornets and then traded to the Los Angeles Lakers. He spent his entire 20-year-old career playing for the Lakers, leading them to 5 NBA Championships while amassing various accolades such as leading the NBA in scoring during two seasons, ranking third on the league's all-time regular season scoring and fourth on the all-time postseason scoring list. He also holds the NBA record for the most seasons playing with one franchise for an entire career. Kobe retired at the end of the 2016 season. Over the course of his long career various statistics have been collected ranging from shot type to game time and x-y coordinates on the court. We will attempt to leverage all that information to initially answer different questions such as is Kobe's shooting percentage subject to home field advantage, or do the odds of him making a shot decrease (linearly or not) with respect to distance from the hoop and finally predict whether one of his shots will go in or miss the target. This predictive model will aim to provide further insight to the greatness of one of the world's best basketball players.

Data Description

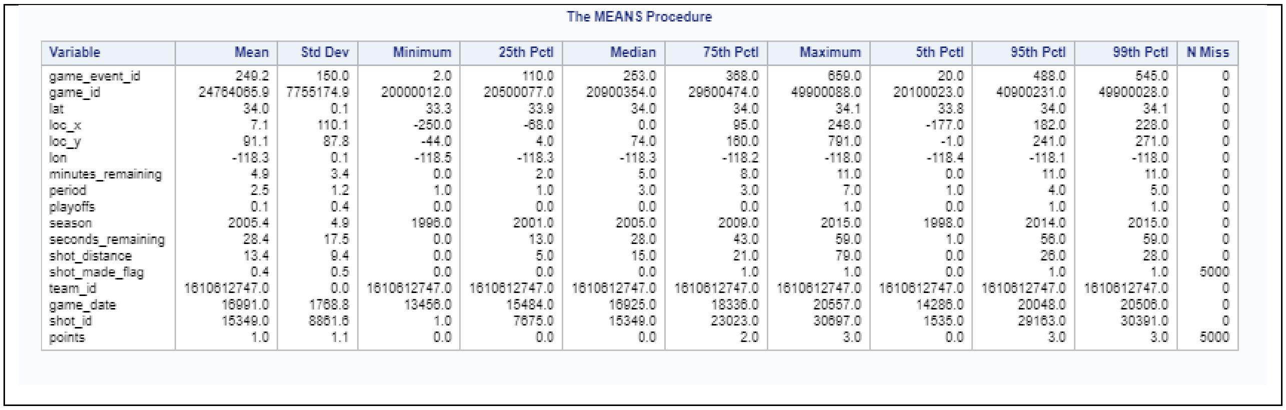
This data set contains historical information describing various attributes of Kobe Bryant's shot attempts spanning his 20-year-old career with the Los Angeles Lakers. It consists of 25 variables. The response variable is '*shot_made_flag'* which identifies whether the shot was made or missed (Binary). The 24 explanatory variables provide various details about Kobe's shot attempts. They consist of categorical and quantitative variables such as: '*action_type'* categorizes the type of shot attempted, '*period*' informs us on which period was the shot taken, '*opponent*' reveals against which team was the shot attempted, and '*shot_distance*' confirms how far from the hoop was the shot taken (feet), *'loc_x' & 'loc_y'* location on court using an x-y grid that shot was attempted. This data set contains 30,697 entries, from which we will use 25,697 on our training model and 5,000 on our testing model. A table of all independent variables follows below. More details surrounding all variables can be found on the appendix.

Kobe Shot Selection Variables		
action_type	minutes_remaining	shot_zone_basic
combined_shot_type	period	shot_zone_range
game_event_id	playoffs	team_id
game_id	season	team_name
lat	seconds_remaining	game_date
loc_x	shot_distance	matchup
loc_y	shot_type	opponent
lon	shot_zone_area	shot_id

Table 1: Data Set Variables

Exploratory Data Analysis

We begin our exploratory data analysis by importing the data set in SAS. We are immediately welcomed by an import error concerning '*season*' variable which SAS cannot categorize. To solve issue, we decide to drop the dual year representation (1997-98) and instead use the start year to identify seasons (1997). Additionally, we created a '*points*' variable to help us identify how many points has Kobe scored, combining '*shot_made_flag'* and '*shot_type' to generate the new variable.* Upon completing the above, we move on to examine whether missing variables are present as well as display the 5 number summaries.



The MEANS Procedure

Variable	Mean	Std Dev	Minimum	25th Pctl	Median	75th Pctl	Maximum	5th Pctl	95th Pctl	99th Pctl	N Miss
game_event_id	249.2	150.0	2.0	110.0	253.0	368.0	659.0	20.0	488.0	545.0	0
game_id	24764065.9	7755174.9	20000012.0	20500077.0	20900354.0	29600474.0	49900088.0	20100023.0	40900231.0	49900028.0	0
lat	34.0	0.1	33.3	33.9	34.0	34.0	34.1	33.8	34.0	34.1	0
loc_x	7.1	110.1	-250.0	-68.0	0.0	95.0	248.0	-177.0	182.0	228.0	0
loc_y	91.1	87.8	-44.0	4.0	74.0	160.0	791.0	-1.0	241.0	271.0	0
lon	-118.3	0.1	-118.5	-118.3	-118.3	-118.2	-118.0	-118.4	-118.1	-118.0	0
minutes_remaining	4.9	3.4	0.0	2.0	5.0	8.0	11.0	0.0	11.0	11.0	0
period	2.5	1.2	1.0	1.0	3.0	3.0	7.0	1.0	4.0	5.0	0
playoffs	0.1	0.4	0.0	0.0	0.0	0.0	1.0	0.0	1.0	1.0	0
season	2005.4	4.9	1996.0	2001.0	2005.0	2009.0	2015.0	1998.0	2014.0	2015.0	0
seconds_remaining	28.4	17.5	0.0	13.0	28.0	43.0	59.0	1.0	56.0	59.0	0
shot_distance	13.4	9.4	0.0	5.0	15.0	21.0	79.0	0.0	26.0	28.0	0
shot_made_flag	0.4	0.5	0.0	0.0	0.0	1.0	1.0	0.0	1.0	1.0	5000
team_id	1610612747.0	0.0	1610612747.0	1610612747.0	1610612747.0	1610612747.0	1610612747.0	1610612747.0	1610612747.0	1610612747.0	0
game_date	16991.0	1768.8	13456.0	15484.0	16925.0	18336.0	20557.0	14286.0	20048.0	20506.0	0
shot_id	15349.0	8861.6	1.0	7675.0	15349.0	23023.0	30697.0	1535.0	29163.0	30391.0	0
points	1.0	1.1	0.0	0.0	0.0	2.0	3.0	0.0	3.0	3.0	5000

Table 2: Data Set Means Procedure

As we can see from *Table 2* there are only 5000 missing variables which is what we will use to test our final model on. Nothing else appears to be out of the ordinary so we proceed with histogram inspection of the data set.

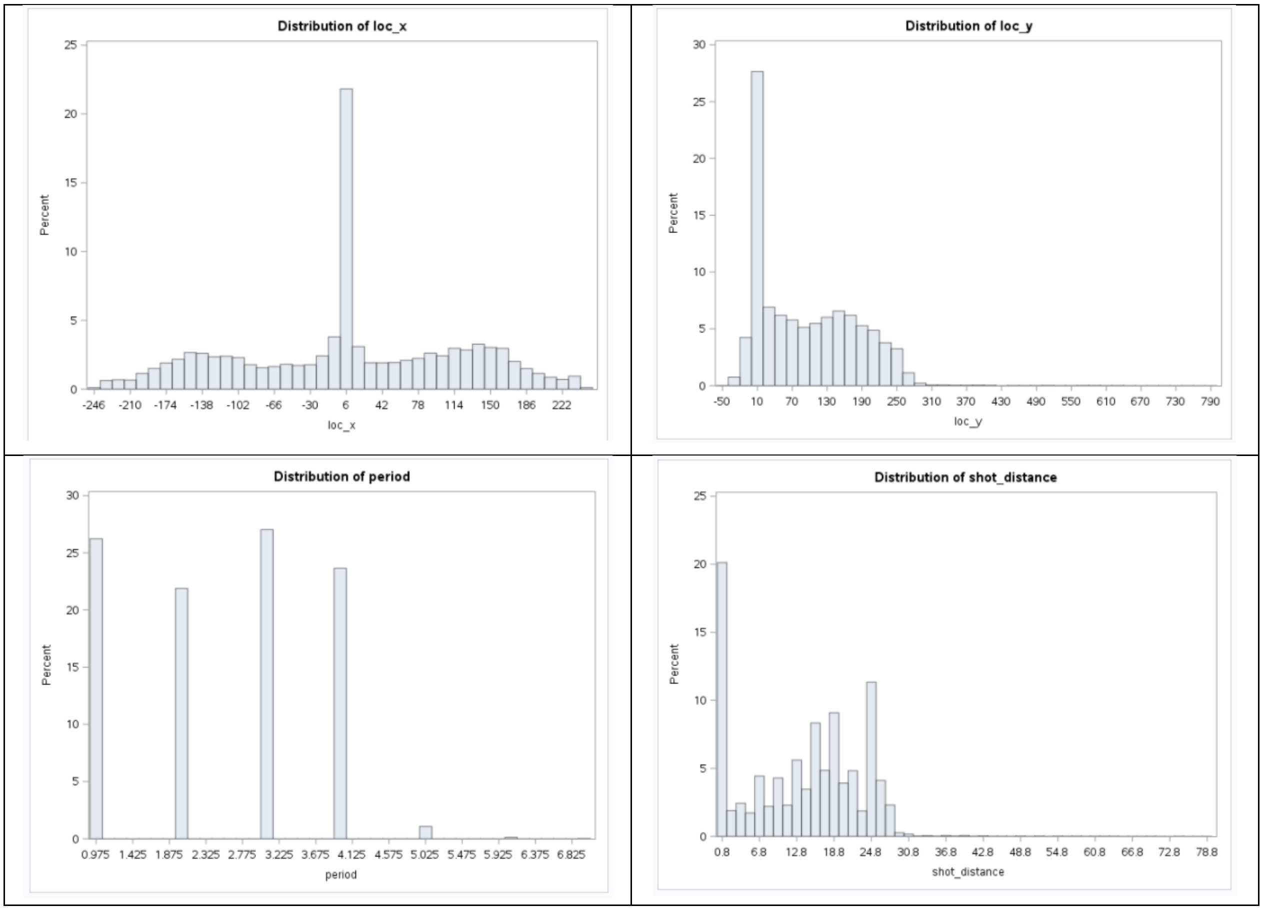


Table 3: Histograms of Data Set

From the top 2 graphs we can visualize the distribution of Kobe's shot attempts. It is evident that he took most of his shots facing the hoop head on from around 10 inches away. On the 3rd histogram we can see the distribution of shots taken per quarter. It is clear that Kobe took most of his shots on the third period and least of his shots on periods five and six. That is to be expected since a regular basketball game consists of 4 periods and only if the game is tied will there be an extra 5th period and if tied again a 6th one. Last histogram depicts the distribution of his shot attempt distance from the hoop. Again, we can see that Kobe took most of his shots under the basket (1 foot), followed by around the 3-point line (23 feet).

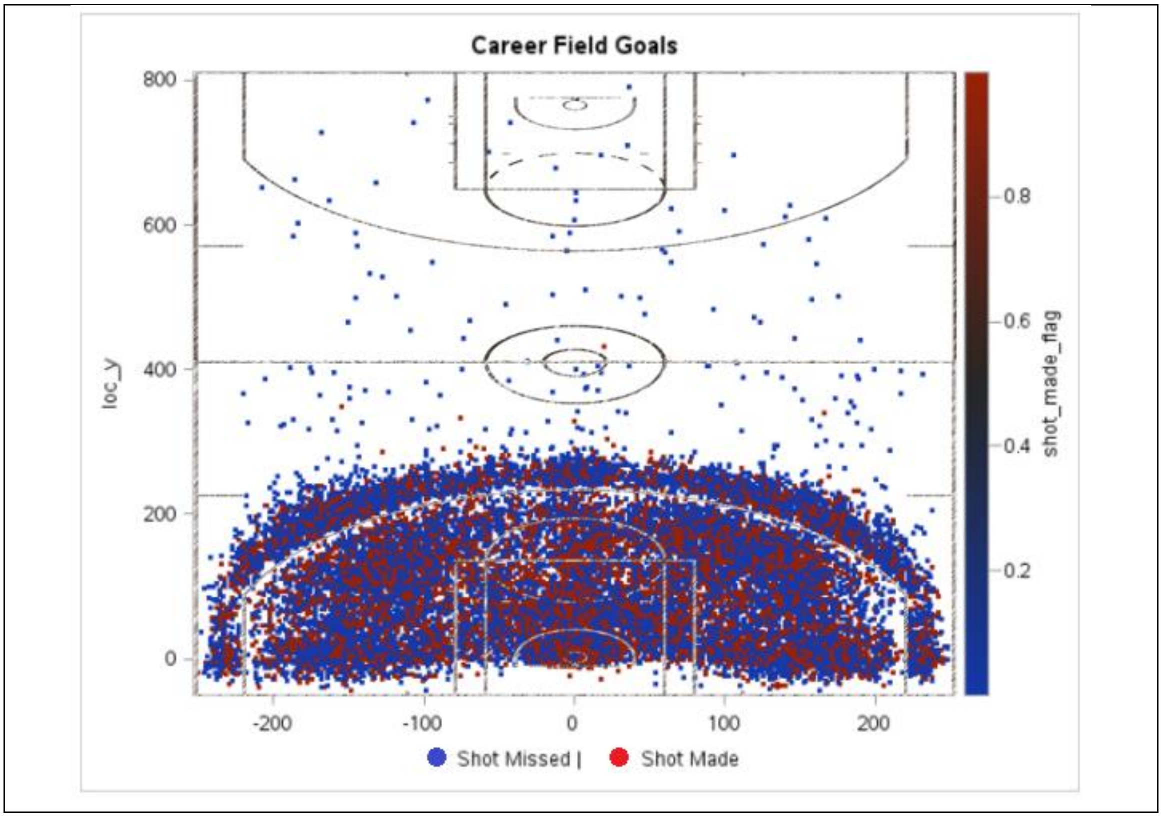


Table 4: Shot Attempts Superimposed on Court

On the picture above, we can get a better visual of Bryant's shot attempts on a basketball court throughout his career. To generate this image, we first plotted all the coordinates (x, y) of his shot attempts and then superimposed it to an image of a basketball court. More information about this can be found on the appendix.

Interpretation Models

1. We will attempt to prove if Kobe's shooting percentage is subject to a home field advantage. In other words, is Kobe's shooting percentage better or worse at home than when he is away? We begin by plotting points scored each season in home and away games. However, it's hard to identify a difference on the visualization below since both lines seem very close and often cross each other.

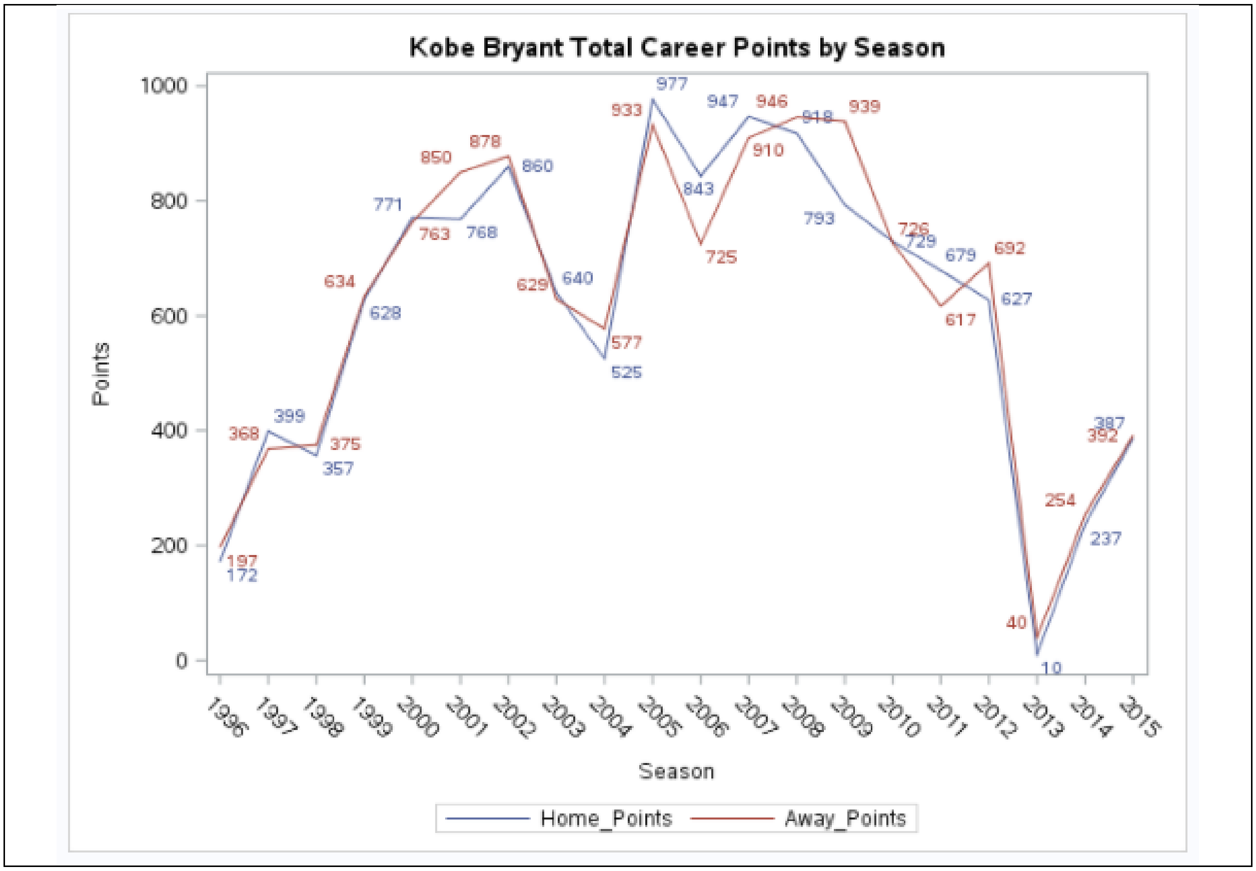
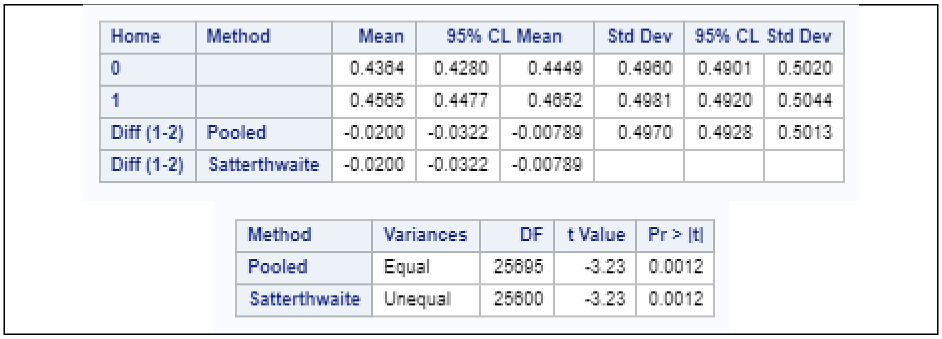


Table 5: Career Points by Season

To better grasp whether Kobe's shooting percentage is subject to a home field advantage we performed a paired t-test. Our null hypothesis is that there is no mean difference between shots made home and away and our alternative hypothesis is that there is a difference. Below are the results.



Home	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		0.4364	0.4280	0.4449	0.4960	0.4901	0.5020
1		0.4565	0.4477	0.4652	0.4981	0.4920	0.5044
Diff (1-2)	Pooled	-0.0200	-0.0322	-0.00789	0.4970	0.4928	0.5013
Diff (1-2)	Satterthwaite	-0.0200	-0.0322	-0.00789			

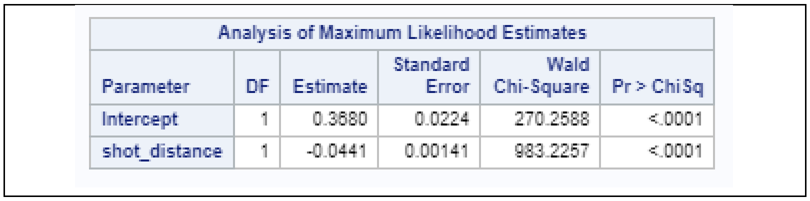
Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	25695	-3.23	0.0012
Satterthwaite	Unequal	25600	-3.23	0.0012

Table 6: T-test Procedure Results

As we can see from *Table 6* with a p-value of 0.0012 we can reject the null hypothesis of no difference for the alternative, that there is a difference in shooting percentage between home and away games (45.65%) and (43.64%) respectively. This informs us that Kobe has a 2.1% advantage at home games compared to away games.

2. Do the odds of Kobe making a shot decrease with respect to the distance he is from the hoop? If there is evidence of this, quantify this relationship. To answer this question, we will use regression with the following model:

$$logit(shot\_made\_flag) = \beta 0 + \beta 1 * shot\_distance$$



Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.3680	0.0224	270.2588	<.0001
shot_distance	1	-0.0441	0.00141	983.2257	<.0001

Table 7: Logistic Procedure Results

As we can see from table 7 above and model used we can deduce that Kobe shoots with a probability of 58% from 1 foot from the hoop and 35% from 22 feet from the hoop (3-pointer).

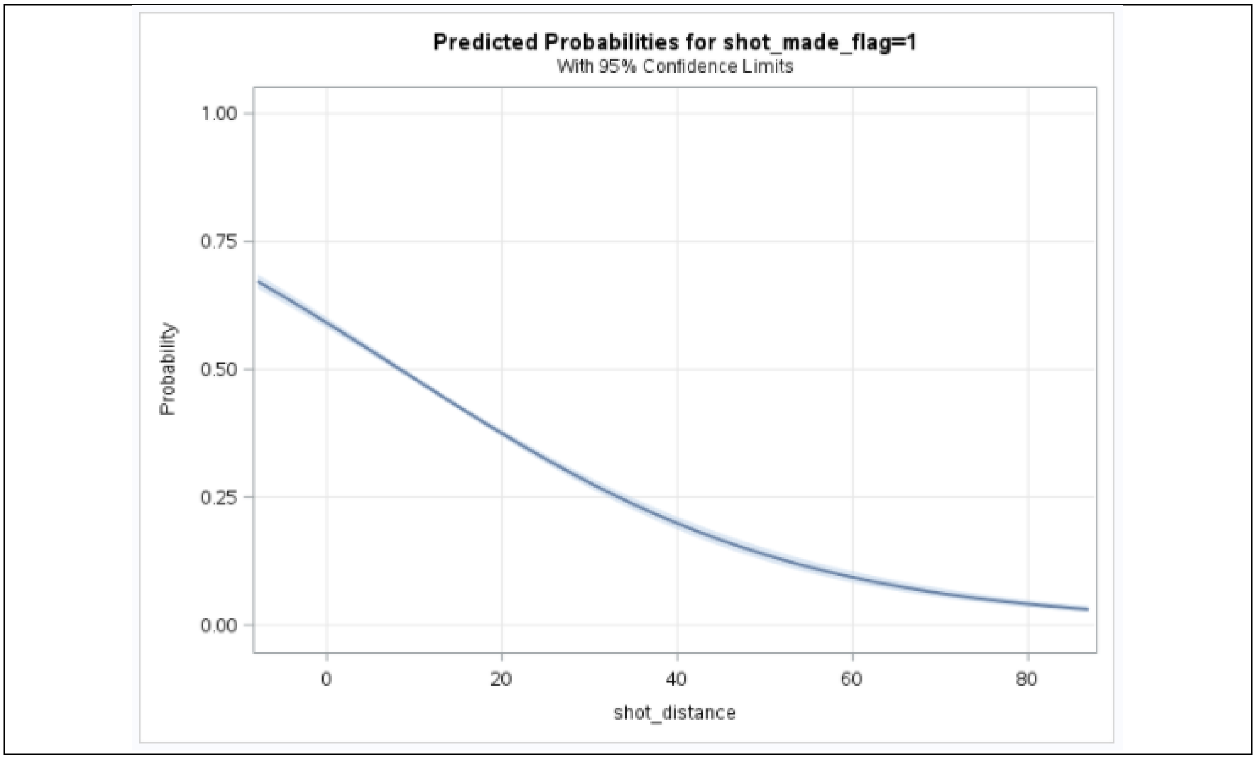


Table 8: Shot Probability with Confidence Limits

It is evident from all the above that the odds of Kobe making a shot decrease the further he gets from the basket.

3. Does the probability of Kobe making a shot decrease linearly with respect to the distance he is from the hoop? If there is evidence of this, quantify this relationship. Again, to answer this question we will default to regression with the same model as before.

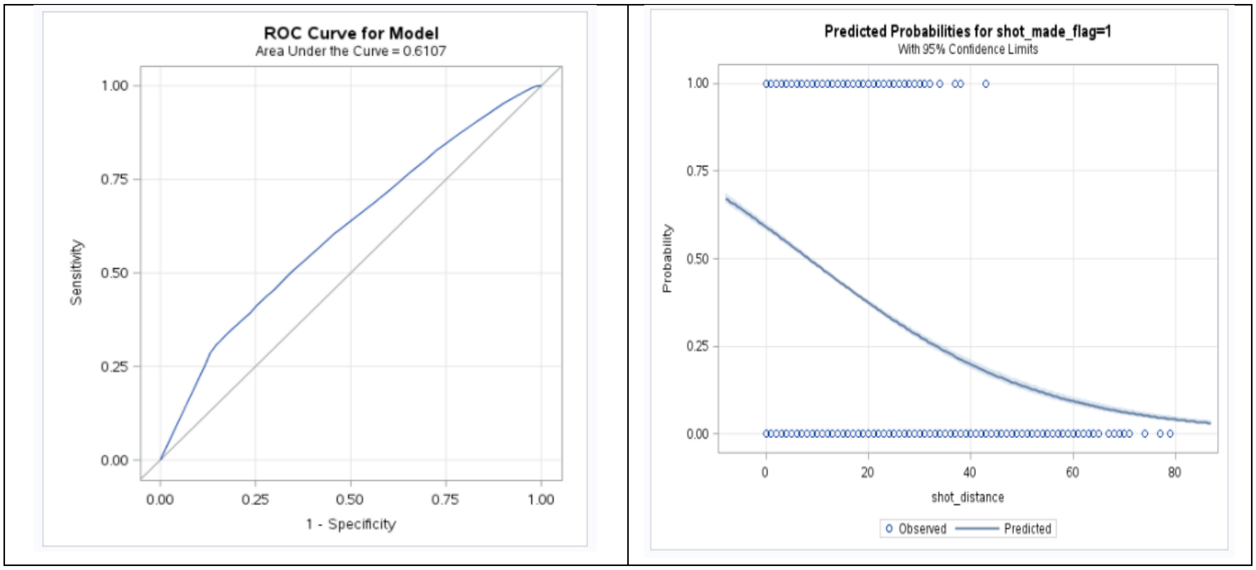
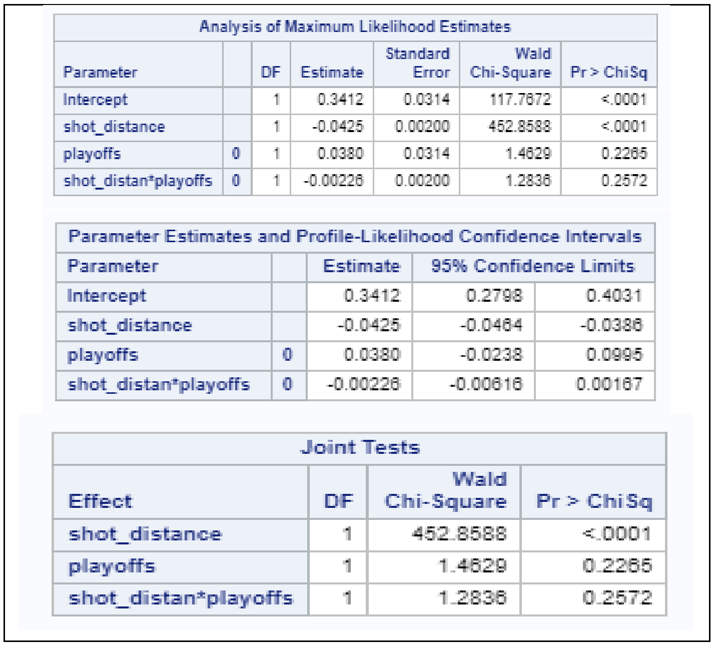


Table 9: Logistic Procedure Plots

As we can see from *Table 9* above there is a linear negative trend for shot distance that is up to 40 feet away from the basket, with an *R2* = 0.9982. After that linear relationship continues but gradient changes (tends to decrease less) the further away we get from the basket.

4. Does the relationship between the distance Kobe is from the hoop and the odds of him making the shot, any different if they are in the playoffs? If there is evidence of this, quantify this relationship. To answer this question, we will use regression analysis with the following model:

$$logit(shot\_made\_flag) = \beta 0 + \beta 1 shot\_distance + \beta 2 playoffs + \beta 3 shot\_distance*playoffs$$



Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.3412	0.0314	117.7672	<.0001
shot_distance		1	-0.0425	0.00200	452.8588	<.0001
playoffs	0	1	0.0380	0.0314	1.4629	0.2265
shot_distan*playoffs	0	1	-0.00226	0.00200	1.2836	0.2572

Parameter Estimates and Profile-Likelihood Confidence Intervals				
Parameter		Estimate	95% Confidence Limits	
Intercept		0.3412	0.2798	0.4031
shot_distance		-0.0425	-0.0464	-0.0386
playoffs	0	0.0380	-0.0238	0.0995
shot_distan*playoffs	0	-0.00226	-0.00616	0.00167

Joint Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
shot_distance	1	452.8588	<.0001
playoffs	1	1.4629	0.2265
shot_distan*playoffs	1	1.2836	0.2572

Table 10: Regression Results for Regular Season VS Playoffs

As we can see from *Table 10* above there is weak evidence that the probability of Kobe making a shot in the regular season is different compared to the probability of him making a shot in the playoffs. With a p-value = 0.2265 we can see that '*playoff'* variable and its interaction with '*shot_distance'* are not significant. Additionally, Table 11 below verifies what we already know ('playoff' is not statistically significant).

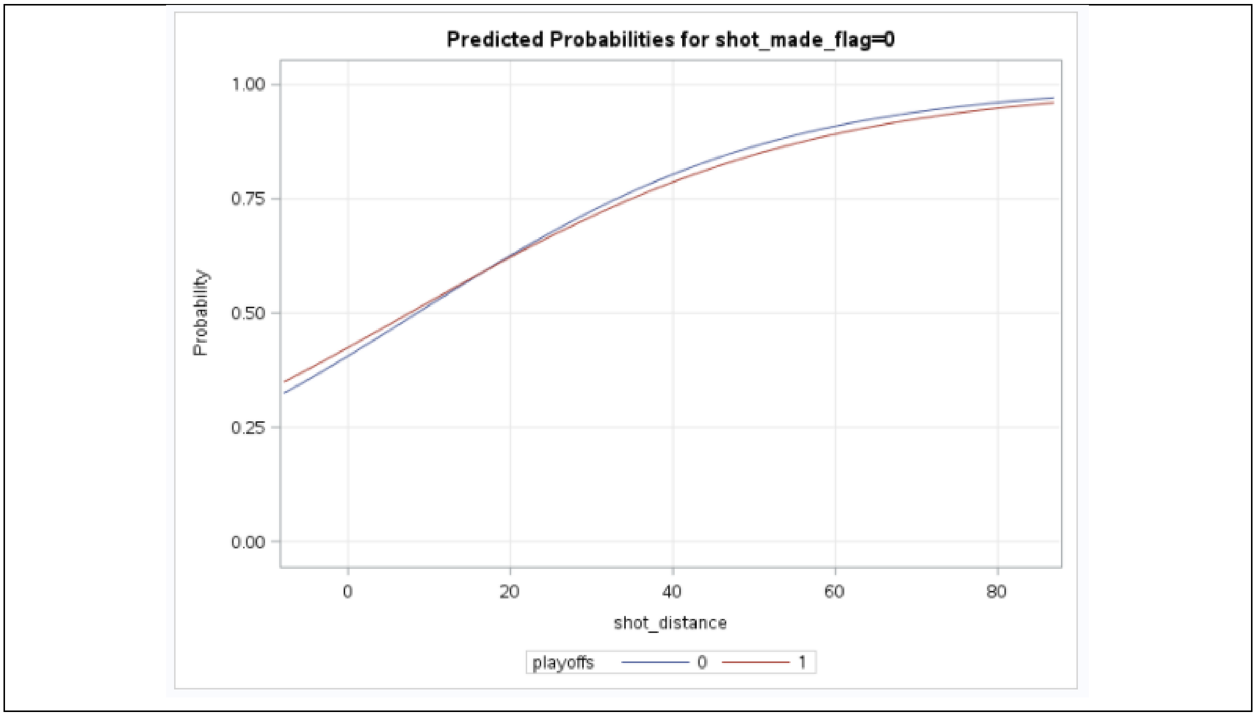


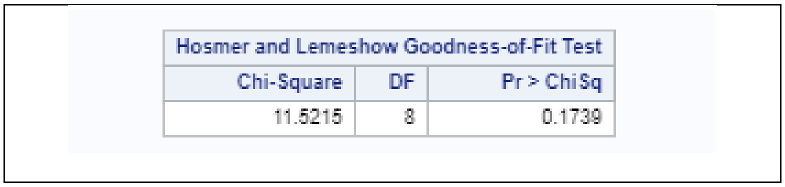
Table 11: Probability shot misses target given Distance and Playoff

Predictive Model

In this section we will examine the steps followed to generate our predictive model and testing it on the data set provided before uploading to Kaggle for scoring. Initially we must create new variables that depict additional features from the raw dataset, to increase our models predictive power. For instance, a ‘*home’* variable was created from ‘*matchup’* to help us classify whether Kobe was playing home or away. Another variable created was ‘*time_remaining*’ by combining ‘*minutes_remaining’* and ‘*seconds_remaining’*. This helped us determine the overall time that a shot was taken helping us account for fatigue and stress. Another binary variable generated was ‘*clutch*’. This one descends from ‘*seconds_remaining’* and depicts the increased difficulty of that shot given that time was running out and defense was most likely very intense. Additionally, the binary variable ‘tough’ was created from ‘loc_y’ helping us identify the difficulty of a shot if very far away or behind the backboard. Finally, the binary variable ‘*fire’* was generated by utilizing the previous value of ‘*shot_made_flag’* with its current value. This helps us identify a unique basketball condition that great players find themselves in called “the zone”. Once a player enters that condition he gets “in sync” with the hoop making him able to score at will and sink very difficult shots. After creating all those variables, we decided to start dropping various duplicates such as ‘*lat’*, ‘*lot’*, ‘*team_name’*, ‘*team_id’*, ‘*matchup*’, ‘*minutes_remaining’*, ‘*seconds_remainng’*, and ‘*game_event_id’ since they are not needed in our model*. Various models were tested (appendix) but in the end the one that performed best (lowest Kaggle Score) was the following:

$$\text{logit(shot\_made\_flag)} = \beta_0 + \beta_1\text{action\_type} + \beta_2\text{shot\_zone\_area} + \beta_3\text{shot\_zone\_basic} + \beta_4\text{shot\_zone\_range} + \beta_5\text{shot\_distance} + \beta_6\text{clutch} + \beta_7\text{tough} + \beta_8\text{fire} + \beta_9\text{home}$$

The assumptions of logistic regression are that our model must share a binary response, linearity of the odds log with the explanatory variables, and independence of observations. The binary response ‘*shot_made_flag’* (0 or 1) cannot help us determine independence of shots, however it is a reasonable assumption to make since there is no evidence against it. With a p-value = 0.1739 running a Hosmer and Lemeshow Goodness of fit test helps us verify that our assumptions are met (Table 12).



Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
11.5215	8	0.1739

Table 12: Model Goodness of fit Test

Below (Table 13) we can see the ROC Curve of our predictive model, which helps us visualize the performance of our binary classifier. Goal of this is to have an AUC (area under the curve) with a value close to 1. Here the average AUC is 0.7036 which is not bad, and higher than a random guessing model which would have an AUC of 0.50.

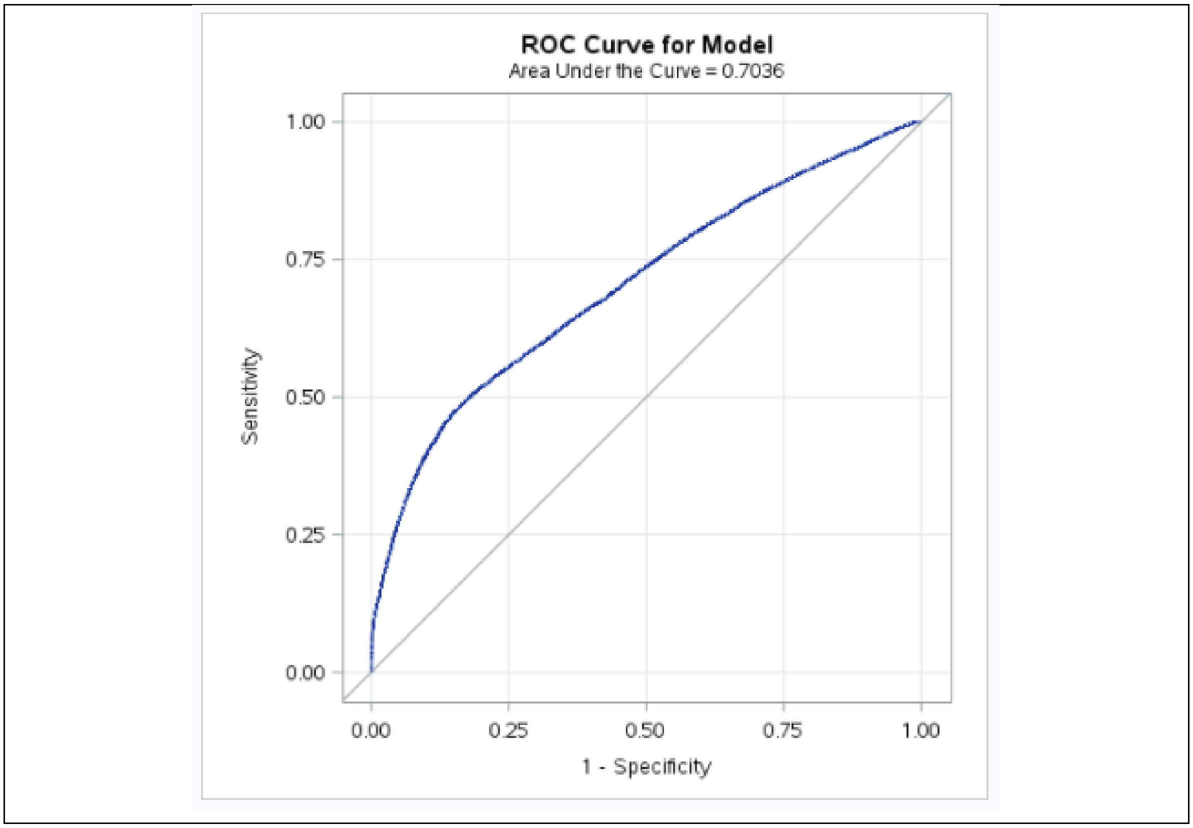
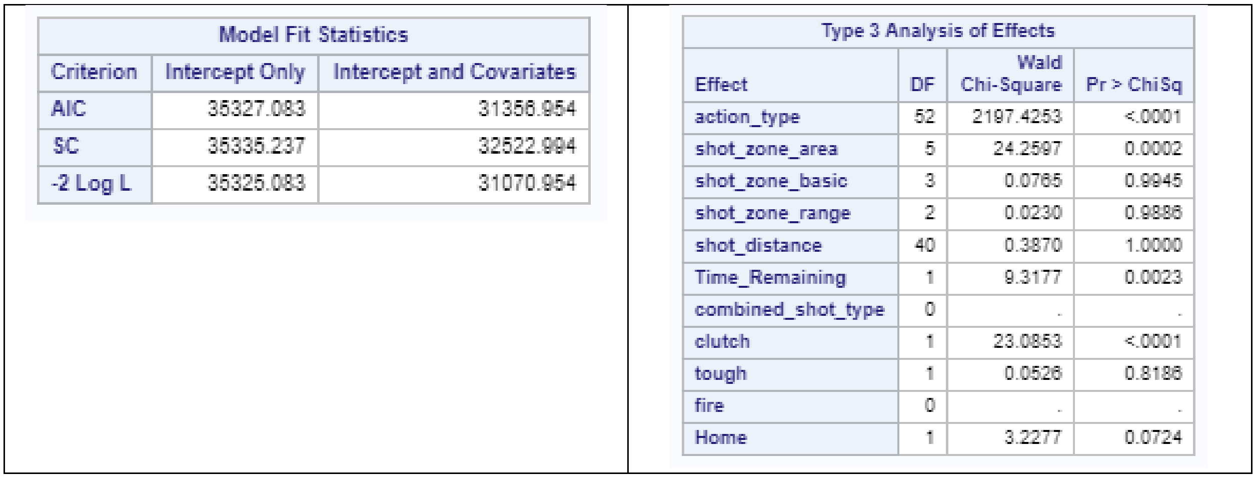


Table 13: ROC Curve

The Model fit statistics can be seen below (Table 14). Also, the effects analyzed help us get a better picture on the final model and the importance of the variables used in it. With an AIC score of 31356 (lowest we got from all models) we are certain that this is the best possible one. An interesting phenomenon discovered examining the analysis of effects table was that even though ‘*shot_distance’*, ‘*shot_zone_basic’* and ‘*shot_zone_range’* seem almost useless to our predictive model (high p-values) they improve our Kaggle score, so a decision was made to keep them in the model.

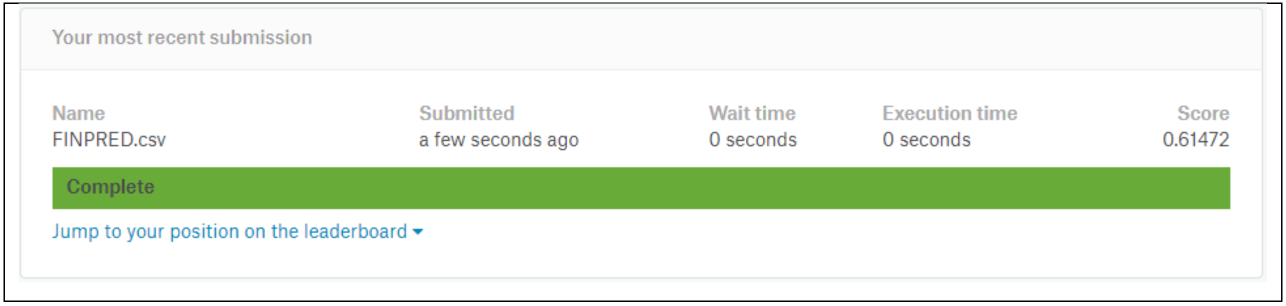


Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	35327.083	31356.954
SC	35335.237	32522.994
-2 Log L	35325.083	31070.954

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
action_type	52	2197.4253	<.0001
shot_zone_area	5	24.2597	0.0002
shot_zone_basic	3	0.0765	0.9945
shot_zone_range	2	0.0230	0.9886
shot_distance	40	0.3870	1.0000
Time_Remaining	1	9.3177	0.0023
combined_shot_type	0	.	.
clutch	1	23.0853	<.0001
tough	1	0.0526	0.8186
fire	0	.	.
Home	1	3.2277	0.0724

Table 9: Model Fit Statistics and Effects

Finally, below is a screenshot of our best (lowest) Kaggle attempt (score = 0.61472).



Your most recent submission

Name	Submitted	Wait time	Execution time	Score
FINPRED.csv	a few seconds ago	0 seconds	0 seconds	0.61472

Table 10: Kaggle Score

Conclusion

In this project we attempted to find the best predictors for whether Kobe's shot goes in or misses the target. Most important predictors detected were '*time remaining'*, '*shot_zone_area'*, '*action_type'*, '*clutch*', and '*home*'.

What we discovered in the process of answering all the questions of interest was that Kobe's probability of making a shot is 2.1% higher during Home games as opposed to Away games. Also, the distance from the hoop is a strong indicator of whether Kobe's shot will find the target or not. We found that Kobe shoots with a probability of 58% from 1 foot from the hoop and 35% from 22 feet from the hoop (3-point line). That relationship is not affected by whether he is in the playoffs or regular season games.

For further analysis we would be interested in analyzing defenders guarding him when a shot was attempted as well as health information (injuries, sickness) and psychological factors (marriage, kids, problems with teammates, sexual accusations). All the above would give us invaluable information helping us delve deeper in the mind of the black mamba. Unfortunately, this data was not available at the time of this study, but we remain positive that with the advent of A.I and Big Data we will one day be able to gather and use information similar to that and create even better predictive models.

Appendix

Variables:

Variables in the Kobe shot data explained, some of these variables have also been transformed into ordinal variables by assigning a number to each level.

- action_type – type of shot taken – 57 different levels
- combined_shot_type – combined action types into 6 levels
- game_event_id – NBA code for a particular event in a game
- game_id – NBA code for each game, 1559 different games
- lat – like loc x with lon it creates a position on the court
- loc_x - must be inches from basket in x direction on a grid of the court
- loc_y - must be inches from basket in y direction on a grid of the court
- lon - like loc y with lat it creates a position on the court
- minutes_remaining – minutes shown remaining on the clock in the period
- period – 4 quarters in a game but overtime means more, ordinal from 1-7
- playoffs – 1 means game in playoffs, 0 means not in playoffs
- season – 2000-01 means the 2000 through 2001 season, 1996-97 means the 1996through 1997 season
- seconds_remaining – seconds shown remaining on the clock in the period
- shot_distance – distance from basket in feet
- shot_made_flag - this is what you are predicting, 0 means shot missed, 1 means shotmade
- shot_type – 2 levels, either a 2 point or 3 point shot, free throws not included
- shot_zone_area – 6 levels different shot areas on court, 6 levels
- shot_zone_basic – 7 levels normal shot zones
- shot_zone_range – 5 levels, shot distances in groups
- team_id – just 1 team id, 161061247, Los Angeles Lakers (LAL)
- team_name – team that Kobe played for, only one team: Los Angeles Lakers(LAL)
- game_date – date of game, 1559 different dates
- matchup – example LAL @ ATL or LAL vs. ATL, 74 levels
- opponent - abbreviation for a team's city, ATL – Atlanta, 33 opponents
- shot_id – each shot were given a number, there are 30697 total shots taken

SAS Code:

```
/* Stat 2 - Project 3 SAS Code */

/* Import data set */
FILENAME REFFILE '/home/egiakoumakis0/sasuser.v94/Stat 2/Project3/data-2.csv';

PROC IMPORT DATAFILE=REFFILE
```

```
            DBMS=CSV
            OUT=stat_kobe;
            GETNAMES=YES;
RUN;

/* EDA */
data kobe;
set stat_kobe;
if shot_type EQ '2PT Field Goal' THEN points = 2 * shot_made_flag;
if shot_type EQ '3PT Field Goal' THEN points = 3 * shot_made_flag;
run;

data test_kobe;
set kobe;
if shot_made_flag = . ;
run;

data train_kobe;
set kobe;
if shot_made_flag = . then delete;
run;

proc print data=kobe;
   var season points;
   where season = 2016;
   sum points;
run;

proc means data=kobe mean std min p25 median p75 max p5 p95 p99 maxdec=1 nmiss; run;

proc univariate data=kobe noprint;
  var loc_x loc_y period shot_distance ;
 histogram;
run;

Data Train;
set train_kobe;

Home = 1;
if find(matchup,'@') then Home=0;
Time_Remaining = minutes_remaining*60+seconds_remaining;

drop lat;
drop lon;
drop team_id;
drop team_name;
*drop Season;
drop matchup;
drop minutes_remaining;
drop seconds_remaining;
drop game_event_id;

run;

/*updating test data set to match the train*/
Data Test;
set test_kobe;

Home = 1;
if find(matchup,'@') then Home=0;
Time_Remaining = minutes_remaining*60+seconds_remaining;

drop lat;
drop lon;
drop team_id;
drop team_name;
drop Season;
drop matchup;
drop minutes_remaining;
drop seconds_remaining;
shot_made_flagN = input(shot_made_flag, BEST12.);
drop shot_made_flag;
rename shot_made_flagN=shot_made_flag;
drop game_event_id;

run;

data home_train;
set Train;
if Home = 1;
run;

data away_train;
set Train;
if Home = 0;
run;

/* Question 1 */

proc sql;
create table hseasonpoints as
 select *,case
   when Home=1 then select sum(points) from home_train where season=a.season
   else .
   end as hseasonpoints
  from home_train as a;
```

```
quit;

proc sql;
create table aseasonpoints as
 select *,case
   when Home=0 then select sum(points) from away_train where season=a.season
   else .
   end as aseasonpoints
  from away_train as a;
quit;

data seasonpoints;
infile '/home/egiakoumakis0/sasuser.v94/Stat 2/Project3/kobess.csv' firstobs=2 dlm=',';
input season $ Home_Points Away_Points;
run;

PROC SGPLOT DATA = seasonpoints;
 xaxis label='Season';
 yaxis label='Points';
 SERIES X = season Y = Home_Points / datalabel;
 SERIES X = season Y = Away_Points / datalabel;
 TITLE "Kobe Bryant Total Career Points by Season" ;
RUN;

proc ttest data=Train;
var shot_made_flag;
class home;
run;

/* Question 2 */

PROC SGPLOT DATA = Train;
 xaxis label='Red: Shot Missed | Blue: Shot Made';
 scatter X = loc_x Y = loc_y / colorresponse=shot_made_flag markerattrs = (size = 3 symbol =
circlefilled);
 TITLE "Career Field Goals" ;
RUN;

proc logistic data=Train plots=all outest=estimates1;
model shot_made_flag(event='1') = shot_distance / clparm=both;
run; quit;

/* Question 3 */

proc logistic data=Train plots=all outest=estimates1;
class playoffs(ref="1");
model shot_made_flag(event='1') = shot_distance|playoffs/ clparm=both;
run; quit;
```

```
proc logistic data=Train plots=all outest=estimates1 PLOTS(MAXPOINTS=NONE);
model shot_made_flag(event='1') = shot_distance / clparm=both;
run; quit;

/* Question 4 */

proc logistic data=Train plots=all outest=estimates1;
class playoffs(ref="0");
model shot_made_flag(event='0') = shot_distance|playoffs/ clparm=both;
run; quit;

/* PDA */

/*create final dataset for model*/
Data Final;
set kobe;

by shot_id;
prevshot=lag(shot_made_flag);
if first.shot_id then prevshot = . ;
if shot_made_flag = prevshot then fire = 1;
else fire = 0;
drop prevshot;

Home = 1;
if find(matchup,'@') then Home=0;
Time_Remaining = minutes_remaining*60+seconds_remaining;
clutch = 0;
if seconds_remaining < 5 then clutch = 1;
tough = 0;
if loc_y < 0 OR loc_y > 200 then tough = 1;

drop lat;
drop lon;
drop team_id;
drop team_name;
drop Season;
drop matchup;
drop minutes_remaining;
drop seconds_remaining;
shot_made_flagN = input(shot_made_flag, BEST12.);
drop shot_made_flag;
rename shot_made_flagN=shot_made_flag;
drop game_event_id;

run;
```

```
proc logistic data=kobe2 plots=all;
class combined_shot_type_num action_type game_event_id;
model shot_made_flag(event='1') =  combined_shot_type_num dist angle action_type
game_event_id;
output out = predictions predicted = I;
run; quit;

proc logistic data=Final plots=all;
class  action_type combined_shot_type  shot_zone_area  shot_zone_basic shot_zone_range
shot_distance;
model shot_made_flag(event='1') = action_type shot_zone_area  shot_zone_basic
  shot_zone_range  Time_Remaining combined_shot_type clutch tough fire Home; */  lackfit;
output out = SS_PRED predicted = I;
run; quit;

data PredOut;
set SS_PRED;
where shot_made_flag=.;
keep shot_id shot_made_flag I;
run;

data finPred;
set PredOut;
shot_made_flag=I;
drop I;
if shot_made_flag = . then shot_made_flag = 0.451416;
run;

/* Kaggle score 0.61472*/
```