MSDS 6372 - Project 3

Predicting Kobe Bryant Shot Selection Success

Authors: Rajni Goyal, Damon Resnick, Ruby Vazquez Pena
4/29/2017

Introduction

This report details the analysis of the many basketball shots of Kobe Bryant and predicts the probability of specific shots being successful. In this analysis, we explore many of the different aspects that affect his shots, and focus on how the last few seconds of a basketball period is different than the rest of the game and what that means for a shooter like Kobe.

The last few seconds of a close basketball game can often be laborious to watch, but sometime those last few seconds are breathtaking and the stuff of legends. Even casual fans know about Jordan's last shot where he broke Bryon Russell's ankles to clinch the 1998 title. Anecdotally speaking, the last few seconds of an NBA game seem to operate differently. For example, the flow of the game changes, as trailing teams look for quick possessions to chip away at leads and teams ahead look to manage the clock. Occasionally the final seconds produce odds-defying shots from the back court, under heavy pressure, that find the net to the cheers or gasps of the crowd. With all this myth-making and hype around the final seconds there is also the idea that some players perform better under this pressure than others.

This report is an analysis of some of the ways the final seconds are different for Kobe Bryant. We are admittedly not NBA analysts and are limited in our understanding of the NBA, but we will apply statistical models in an attempt to learn more about Kobe Bryant's participation in the NBA. This analysis will focus on the ways Kobe changes his behavior with the clock running down in an effort to find the greatness and weakness in Kobe's game.

Data Description

The data set we analyzed contains 25 different variables describing every shot Kobe Bryant attempted in his 20 yearlong NBA career (30,697 rows). The response variable is *shot_made_flag* identifying whether the shot attempt was made (1) or not made (0). The 24 explanatory variables consist of nominal, ordinal, and continuous variables. The *shot_distance* variable may be the easiest variable to interpret as it is simply just the distance in feet that the shot was attempted round up to the nearest foot. One notable variable is *season*, a nominal variable which we turned into a simple ordinal variable, *ssn_numb*, which goes from 1 to 20 categorizing each individual season for which Kobe was an active player. Another set of interesting variables are *loc_x* and *loc_y*, the location of the shots taken on a x-y grid of the court. One of our favorite variables *action_type*, is a 54 level categorical variable corresponding to the many different types of shots that were performed. One variable *game_event_id* is a numerical nominal variable with 620 levels used by the NBA to describe each type of event in a game.

A detailed list of the variables and an explanation of their meanings are outlined in the appendix, while the exploratory analysis below shows some visualizations of selected variables.

Exploratory Data Analysis

With 25 variables describing 30,697 shot attempts, we truly have a diverse data set with multiple relationships between these variables.

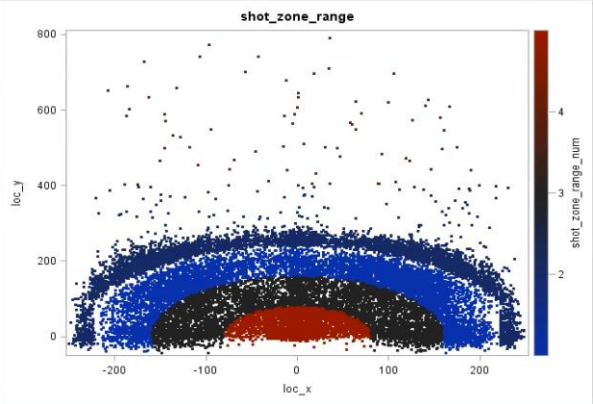


Figure 1. We have utilized the explanatory variables loc_x and loc_y to plot the location of every shot Kobe took and plot them on a x-y grid of the court. The color coding is due to the five different levels of *shot_zone_range*.

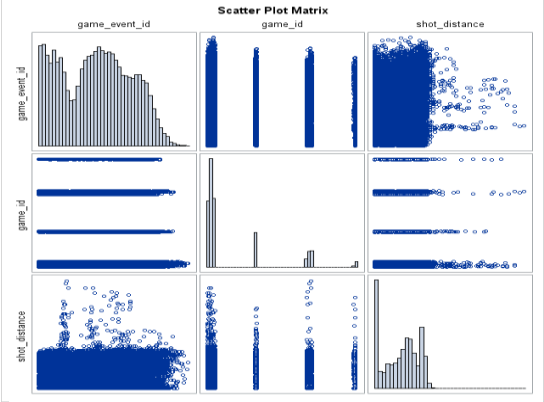
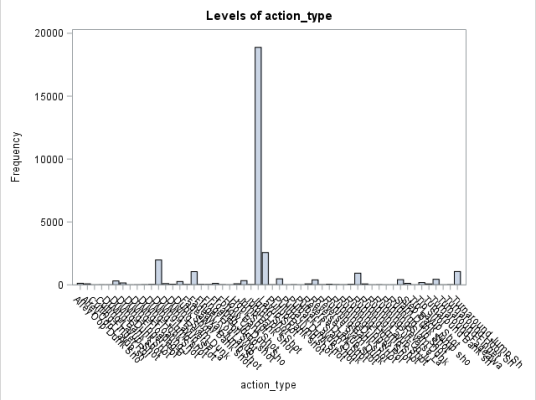


Figure 2. This is a matrix of figures of 3 of the numerical variables. It is interesting to note *game_id*, which is a code to identify each one of the 1559 games, seems to be made up of only 4 major groupings of these values.



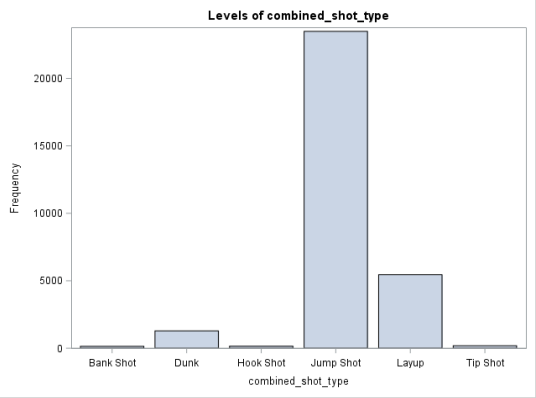


Figure 3 and 4. Below you can see bar charts of the categorical variables *action_type* and *combined_shot_type*. Clearly Kobe favored the jump shot but he could get to the hole pretty well for those layups.

Interpretation Models

1. Kobe's shooting percentage is subject to a home field advantage. That is, Kobe's shooting percentage is better or worse at home than when he is away.

There appears to be a significant difference in shooting percentage between home (45.6%) and away (43.6%) games. (p-value = 0.0012) . The percentages tell us that Kobe has a 2% advantage at a home game compared to an away game. We utilized the model below:

$$\texttt{logit(shot\_made\_flag)} = \beta_0 + \beta_{1j}\texttt{HomeField}$$

Table 1 below shows the maximum likelihood estimates and confidence intervals for Kobe's shooting percentage at home games.

$$Probability\ of\ making\ a\ basket\ at\ home\ games = \frac{e^{(intercept\ +\ HomeField\ 1)}}{1 + e^{(intercept\ +\ HomeField\ 1)}} = \frac{0.839}{1.839} = 0.456$$

Table 1

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.2150	0.0126	293.3598	<.0001
HomeField	1	1	0.0406	0.0126	10.4375	0.0012

Parameter Estimates and Profile-Likelihood Confidence Intervals				
Parameter		Estimate	95% Confidence Limits	
Intercept		-0.2150	-0.2398	-0.1905
HomeField	1	0.0406	0.0160	0.0652

Table 2 below shows the maximum likelihood estimates and their confidence intervals for Kobe's shooting percentage at away games. Using these estimates, we find that the

$$Probability\ of\ making\ a\ basket\ at\ away\ games = \frac{e^{(intercept\ +\ HomeField\ 0)}}{1 + e^{(intercept\ +\ HomeField\ 0)}} = \frac{0.774}{1.774} = 0.436$$

Table 2

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.2150	0.0126	293.3598	<.0001
HomeField	0	1	-0.0406	0.0126	10.4375	0.0012

Parameter Estimates and Profile-Likelihood Confidence Intervals				
Parameter		Estimate	95% Confidence Limits	
Intercept		-0.2150	-0.2398	-0.1905
HomeField	0	-0.0406	-0.0652	-0.0160

2. 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. (CIs, plots, etc.)

The distance a shot was attempted from has been seen to have significant effect on the odds of making a shot. (p-value <0.0001) Table 3 below shows the maximum likelihood estimates and the confidence intervals for Kobe's shooting percentage dependent on the distance of the shot using this model,

$$\texttt{logit(shot\_made\_flag)} = \beta_0 + \beta_1\texttt{shot\_distance}$$

Table 3

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

Parameter Estimates and Profile-Likelihood Confidence Intervals

Parameter	Estimate	95% Confidence Limits	
Intercept	0.3680	0.3242	0.4120
shot_distance	-0.0441	-0.0469	-0.0413

Using the estimates in Table 3 we constructed Table 4 below.

Table 4

Distance from Basket	Probability	Odds	95% Confidence Limits for Odds	
1 ft.	58%	1.38	1.319	1.448
22 ft.	35%	0.54	0.492	0.608

As you can see in Figure 5 below the odds of Kobe making a shot depends heavily on how far away the ball must travel to get into the hoop. When he is 1 foot away from the basket the odds of him making the shot are 1.38. Whereas for the start of three point range (22 ft) it looks to be about 0.54 and then tails off.

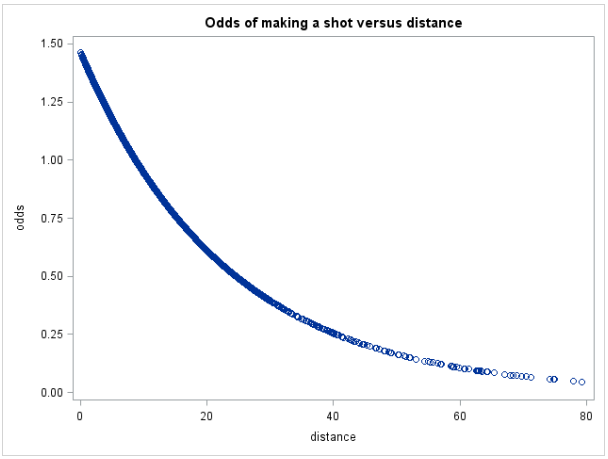


Figure 5. This is a graph of the predicted odds of Kobe making a shot versus distance in feet. We can see that the odds of Kobe making a shot get smaller as the shot comes from farther from the basket.

3. The probability of Kobe making a shot decreases linearly with respect to the distance he is from the hoop. If there is evidence of this, quantify this relationship. (CIs, plots, etc.)

Looking at Figure 6 below, we see that there is a linear trend for shot distances less than 40 feet. If we look more closely at that region, Figure 7, we see that it fits a linear function very well with an R^2 of 0.9982. The linear function has these features:

$$\text{Probability} = 0.59012 - 0.01072 \cdot \text{shot\_distance}$$

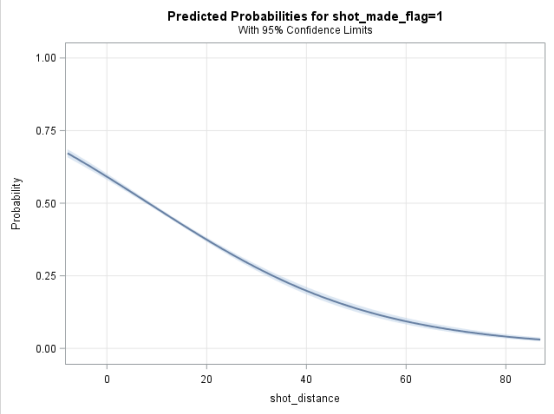


Figure 6

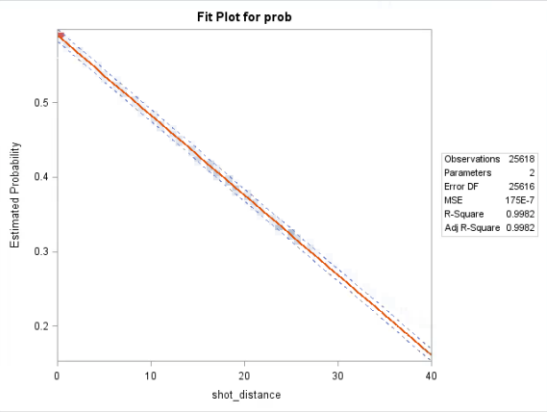


Figure 7

4. The relationship between the distance Kobe is from the hoop and the odds of him making the shot is different if they are in the playoffs. If there is evidence of this, quantify this relationship. (CIs, plots, etc.)

There is weak evidence that the probability of Kobe making a basket in the regular season compared to the probability of Kobe making a basket in the playoffs are different. (p-value = 0.8404) When Kobe is attempting a shot in the regular season the odds of him making a shot are 0.8064 (44.64%) and 0.8008 (44.47%) in the playoffs. Thus, there appears to be no significant difference when just looking at average shooting percentage. However, when shot distance is taken into account there does **appear** to be a difference between shooting percentage in the playoffs at the same distance. The model below was used:

$$\text{logit(shot\_made\_flag)} = \beta_0 + \beta_1\text{shot\_distance} + \beta_{2j}\text{playoffs} + \beta_{3j}\text{shot\_distance*playoffs}$$

Table 5: Regular Season

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

Table 6: 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	1	1	-0.0380	0.0314	1.4629	0.2265
shot_distan*playoffs	1	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	1	-0.0380	-0.0995	0.0238
shot_distan*playoffs	1	0.00226	-0.00167	0.00616

In Figure 10, you can see that the two probability/odds curves cross and their probabilities seem to differ at different distances. In Figure 11 you can also see the difference in odds. This shows that the probabilities appear to be different in the playoffs for short and long distances in opposite ways. It appears Kobe is **less** likely to make a short shot in the playoffs, conversely it appears he is **more** likely to make the longer 3 point shot in the playoffs. There is

no way to be certain why this would occur, however it could be argued that the NBA playoffs are much more physical than the regular season. Playoff teams may place a higher premium on rim defense in the playoffs, which could have forced Kobe and his team to set up better shots beyond the perimeter enhancing the probability of longer shots.

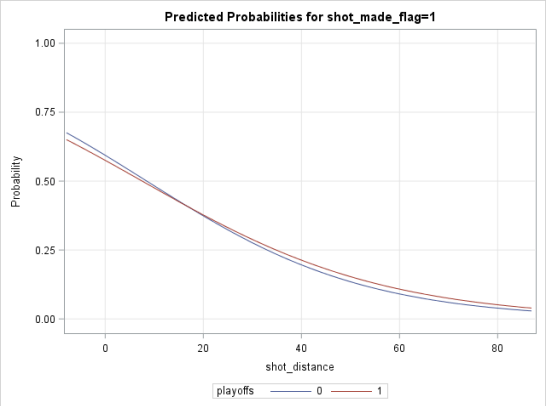


Figure 10. Shows the probability of Kobe making a shot during the regular season and during the playoffs versus *shot_distance*.

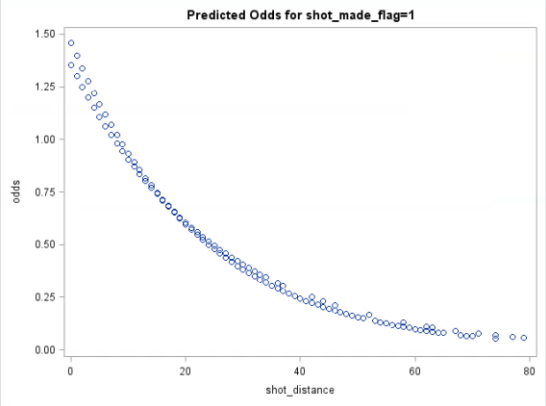


Figure 11. Shows the odds of Kobe making a shot during the regular season and during the playoffs versus *shot_distance*.

Now notice closely that in the last two paragraphs we used the word appeared several times. This is because Figure 10 may be misleading. There definitely appears to be a gap between the two curves at short and long distances indicating that there is a difference. However, we can't seem to determine if this difference is statistically significant! Looking closely at Table 7 below, the confidence intervals for predicted odds at specific distances clearly overlap by an extreme amount leading to a conclusion that this method for determining if there is a significant difference in shooting percentage in the playoffs compared to the regular season is flawed. However, the p-values for the playoff terms in the model are both greater than 0.05 so it seems likely that there is no significant difference.

Table 7					
Season	Distance from Basket	Probability	Odds	95% Confidence Limits for Odds	
Regular	1 ft.	57%	1.34	1.2628	1.4397
Playoff	1 ft.	56%	1.300	1.1413	1.4835
Regular	22 ft.	35.5%	0.552	0.4766	0.6401
Playoff	22 ft.	35.8%	0.558	0.4159	0.7506

5. With respect to question 4, is there evidence of a difference after accounting for a home field advantage? That is, does the answer to question 4 depend on a home field advantage? If there is evidence of this, quantify this relationship. (CIs, plots, etc.)

Even more interesting is accounting for home field advantage in the playoffs. As we have seen there is a 2% advantage in the probability of Kobe making a shot at home compared

to away. Now in the playoffs when we take distance into account, we have 4 probability curves as a function of distance (Figure 12). For short distances Kobe appears to have nearly the same probability of making a shot at home or away during the regular season and it appears about a 2% less chance of making those short shots in the playoffs either at home or away. For longer shots, something very interesting happens. Kobe's shooting percentage at long range in the playoffs seem to elevate to that of his regular season home field long range percentage whether he is at home or not. His long-range field goal percentage appears to be lowest during the regular season away while his long range percentage appears to be higher at home and about the same in the playoffs no matter if he plays at home or away. One might be eager to conclude that Kobe truly does live for the playoffs. He appears to show up when it matters more to his team's success, making the shots that are harder to defend in the playoffs.

$$\text{logit}(\text{shot\_made\_flag}) = \beta_0 + \beta_1\text{shot\_distance} + \beta_{2j}\text{playoffs} + \beta_{3j}\text{shot\_distance*playoffs} + \beta_{4j}\text{HomeField} + \beta_{5j}\text{shot\_distance*HomeField} + \beta_{6j}\text{playoffs*HomeField} + \beta_{7j}\text{shot\_distance*playoffs*HomeField}$$

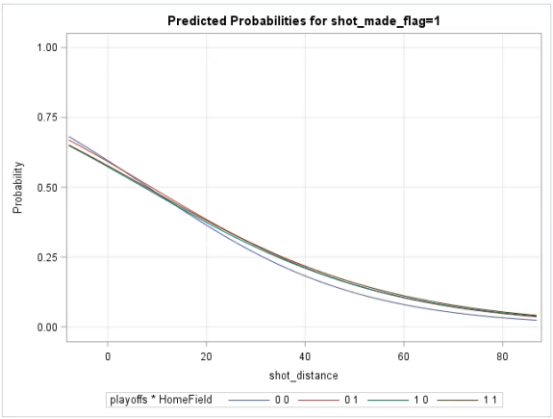


Figure 12. This graph displays 4 probability curves as a function of distance. Blue: Regular season away game, Green: Playoff away game, Red: Regular season home game, Brown: Playoff home game

Looking at Figure 12 there definitely appears to be separation between the probability curves at different distances. However if you look closely and calculate the confidence intervals for the predicted odds taking into account distance, what you find is that there doesn't appear to be a significant difference between shooting percentages at the same distance. Looking at the Table 8 below you can see that all the confidence intervals for a specific distance overlap considerably. While this does not prove there is not a significant difference between the shooting percentages from the same distance, it is highly suggestive that Kobe's shooting percentage is unchanged at away games when shot distance and playoffs are taken into account. In other words, Figure 12 appears to tell an incredibly interesting story, but may be misleading.

Table 8						
Season	Game	Distance from Basket	Probability	Odds	95% Confidence Limits	
Regular	Home	1 ft.	58.1%	1.3906	1.0697	1.8079
Regular	Away	1 ft.	58%	1.4043	1.0803	1.8257
Playoff	Home	1 ft.	56.7%	1.3102	1.0078	1.7031
Playoff	Away	1 ft.	56.3%	1.2908	0.9933	1.6786
Regular	Home	40 ft.	21.2%	0.2693	0.1125	0.6454
Regular	Away	40 ft.	18.2%	0.2227	0.0931	0.5338
Playoff	Home	40 ft.	21.7%	0.2778	0.1160	0.6651
Playoff	Away	40 ft.	20.9%	0.2657	0.1109	0.6359

6. Is Kobe clutch? After accounting for the distance of the shot, does Kobe's shooting percentage increase when he is taking a shot in the last 30 seconds of a period?

This is perhaps the most interesting part of the analysis. Kobe who many have called an assassin, or our favorite "The Black Mamba", in reference to one of the most most deadly of venomous African snakes, for his ability to cut the enemy down in a cold unfeeling way, would appear to be far from "clutch". Looking at Figure 13 below you can see his field goal percentage is lower across the entire spectrum. (p-values for all estimates are below 0.05) This probably should not be surprising because the last 30 seconds of a period is probably the hardest time to score a bucket because of the defensive intensity and the multitude of coaching time outs that allow coaches to sub in the best defenders with a specific plan to stop Kobe Bryant from scoring. Nevertheless, Kobe Bryant always wanted the ball in crunch time, and he was one of the best at making difficult shots. The model below was used:

$$\text{logit}(\text{shot\_made\_flag}) = \beta_0 + \beta_1 \text{shot\_distance} + \beta_{2j} \text{shot\_distance} * \text{clutch}$$

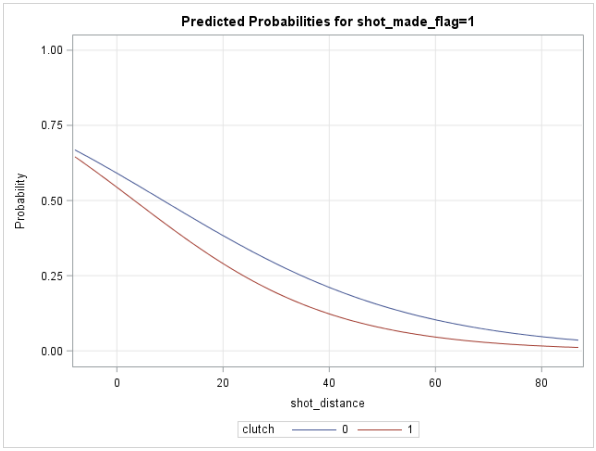


Figure 13. Displayed are the probability curves for clutch and no clutch versus *shot_distance*.

Like almost every high-volume shooter Kobe's field goal percentage was significantly lower in the last 30 seconds of periods and games. Not having data for the entire league during Kobe's 20 seasons we cannot say with certainty whether this difference in shooting percentage during crunch time made him more or less clutch than other similar players. However doing a simple google search and looking at percentages for similar players in similar situations we found that Kobe was above average in his ability to make a shot in the last 30 seconds. He didn't shoot better on average in the last 30 seconds than he normally would, but he did shoot better than most players in the last 30 seconds.

Table 9					
Clutch	Distance from Basket	Probability	Odds	95% Confidence Limits for Odds	
Yes (1)	1 ft.	53.1%	1.1307	0.9401	1.3599
No (0)	1 ft.	58.1%	1.3847	1.1511	1.6653
Yes (1)	22 ft.	26.8%	0.3677	0.2494	0.5393
No (0)	22 ft.	36.4%	0.5719	0.3889	0.8398

Predictive Model

For predicting individual shots made or missed we used a complex Logistic model. Variable selection was done by hand, using Kaggle, and using a cross-validation code that selected out randomly from the train data a set of data equal in size to the 5000 data observations in the test set. The cross-validation code then computed a log-loss score identical to the one that Kaggle uses for their leader board. Simple models with just a few variables known to have a strong correlation to making or missing a shot, such as *shot_distance* and *combined_shot_type,* were used initially and then other variables were added or subtracted depending on whether the CV log-loss score got better or not.

Several new variables were created based on the original variables in an attempt to model the data well. The variable *dist* was created based on loc_x and loc_y along with *angle*. Other variables like *clutch*, and total seconds remaining in a period or the game were also created using variables in the original data set.

Several interesting effects were noticed during the variable selection process. For instance, *shot_distance* gave a better prediction score if used as a categorical or class variable with 74 levels rather than a continuous numerical variable. Also, large leveled categorical variables such as *season*, *action_type*, and *game_event_id* which has 620 levels also helped seemed to help the prediction score significantly at first. Attempts were made to regroup some of these levels using cluster analysis but were not successful in reducing the Kaggle score.

After most of the variables were tried in different combinations a final best model was determined to consist of:

$$\text{logit}(\text{shot\_made\_flag}) = \beta_0 + \beta_1 \text{ttl\_sec\_remn\_gam} + \beta_{2j} \text{action\_type} + \beta_{3j} \text{ssn\_numb} + \beta_{4j} \text{shot\_zone\_area\_num} + \beta_{5j} \text{shot\_zone\_basic\_num} + \beta_{6j} \text{shot\_zone\_range\_num} + \beta_{7j} \text{shot\_distance} + \beta_{8j} \text{clutch}$$

It should be noted that all variables in this model except for *ttl_sec_remn_gam* are treated as categorical variables. Even *shot_distance* helped to predict shot made better as a categorical variable with 74 levels.

The assumptions of logistic regression are that you model a binary response, that the log odds be linearly related to the explanatory variables, and that the observations be independent of one another. This analysis models the binary response of shot_mdae_flag which is either 0 for missed and 1 for made. It cannot be determined that every shot is independent of one another but it is a reasonable assumption to make that the shots are mostly independent. A Hosmer and Lemeshow Goodness of Fit test was used to help check these assumptions. The p-value for this test was found to be 0.1855 implying the assumptions were met.

Table 10

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
11.2964	8	0.1855

Table 11

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	35327.083	31253.004
SC	35335.237	32557.665
-2 Log L	35325.083	30933.004

The model fit statistics are also reported for this model in Table 11. The cross-validated log-loss score computed was 0.60204 and it obtained a Kaggle score of 0.61228. Something interesting to note is that the variable `shot_zone_range_num` is almost totally useless to the fit as other variables can combine with it as linear combinations to zero out some of its levels. You can see that in the table below as the p-value = 0.9896. However, it helps the Kaggle score a bit. Without `shot_zone_range_num` the log-loss score was nearly the same at 0.60206 but the Kaggle score was 0.61258 so we left it in the model.

Table 12

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
ttl_sec_remn_gam	1	24.4647	<.0001
action_type	52	2243.8725	<.0001
ssn_numb	19	107.8694	<.0001
shot_zone_area_num	5	24.8123	0.0002
shot_zone_basic_num	6	16.3543	0.0120
shot_zone_range_num	2	0.0209	0.9896
shot_distance	72	99.1065	0.0188
clutch	1	40.7648	<.0001

Kaggle Scores

Below is an example of how to use the model. The parameter estimates for a specific combination of variables are selected in then added together to get the log odds of Kobe making a shot.

Model for 60 secs remaining in the game, Intercept = 4.5698, ttl_sec_remn_gam(60) = 0.000081, action type(Jump Shot)= -2.2935, ssn_numb(1) = 0.0260, shot_zone_area(Center(C)) = 1.4577 , shot_zone_basic(Mid-Range) = -27.8766, shot_zone_range(8-16 ft.) = 0, shot_distance(10 ft) = 22.8791, and clutch(0) = 0.1817.

$$\text{logit(shot\_made\_flag)} = 4.5698 + 0.000081 * 60 - 2.2935 + 0.0260 + 1.4577 - 27.8766 + 0 + 22.8791 + 0.1817$$

$$\text{logit(shot\_made\_flag)} = -1.04213$$

$$\text{Odds of making a shot} = e^{-1.04213} = 0.3527$$

$$\text{Probability of making a shot} = (0.3527/1+0.3527) = \mathbf{0.2607}$$

This combination of category types in this example gives Kobe only a 26% chance to make this particular shot. He probably doesn't choose to shoot it this way very often.

Conclusion

This report details the analysis of the many basketball shots of Kobe Bryant and provides a model which predicts the probability of specific shots being successful. In this analysis, we explored many of the different aspects that affect his shots, and focus on how the last few seconds of a basketball period is different than the rest of the game and what that means for a shooter like Kobe.

Kobe is a high-volume shooter and his shot patterns are as expected. He basically can shoot from anywhere on the court. Kobe's shooting percentage seems to be higher at home than at away games. Also like most shooters Kobe had a higher probability of making a shot at short distance than at longer distances, and his shots displayed a linear relationship for the probability of making a shot for shots less than 40 ft. In the playoffs shooting percentage does not change in a significant way. Even when looking taking in home-court into account Kobe appears to have a different shooting percentages but it is difficult to tell if the differences are significant. The last interesting piece is how Kobe performs in the "clutch" or last 30 seconds of periods. Kobe appears to have a lower shooting percentage in the clutch like most high-volume shooters.

Our final prediction model produced a Kaggle score of 0.61226. This model was obtained after weeks of variable selection and the crafting of new and interesting variables. No extra data was used and leakage was kept to a minimum. Only the original data was used and transforms of that data based on basketball knowledge was used to create new variables to help prediction. We also tried using cluster analysis to regroup several of the large categorical

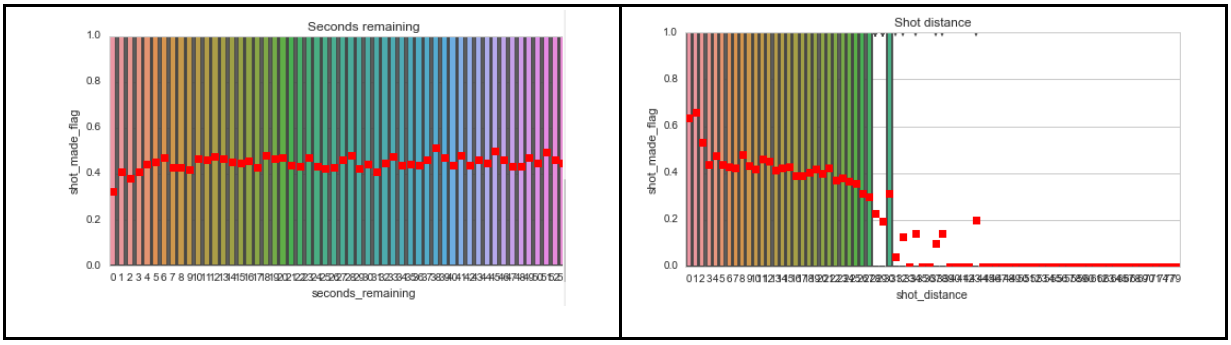
variables but found that for most of them regrouping was not helpful. A heatmap was even used to look at how many clusters could be used. Interaction terms were also tried for several terms and while the seemed to help the prediction at first it was later found that different combinations of non-interacting variables produced a better score. We think regrouping these large categorical variables could provide better prediction models but we were unable to provide a scheme here utilize that technique effectively.

Bonus Question 2

For this question, we have used Python and the Random Forest algorithm to solve the Kaggle Prediction for Kobe Bryant's Shot Selection. Below are the steps which we used:

1. We started with importing the data and any basic libraries we need to analyze the data.
2. Utilized dtypes() and describe() functions to get to know the data including each variable.
3. Next, we summarized data.
4. We obtained descriptive statistics and data visualizations by plotting the total shots_made, shot_made_flag against lat, lon, loc_x ,loc_y, shot_distance and secs_remn in each game.
5. When cleansing the data, we assumed independence between each shot, therefore we dropped columns not useful to the model like lat, lon, game_id etc.
6. Transformed variables to create new features and encoded categorical variables.
7. We computing indicator and transformed categorical variables into a "dummy" matrix. If a column in a data frame has k distinct values, a matrix containing k columns containing all 1's or 0's. Pandas has a get_dummies function.
8. We reduced the number of features Hyperparameter tuning.
9. We predicted the probability of shot_made_flag for the missing shot_ids using Random Forest algorithm and submitted to Kaggle.

<u>Analysis:</u> We used the Random Forest algorithm to test variables we identified during the initial exploratory analysis. Our data set and any models that we have developed for this project are not sophisticated enough to tell us whether or not we should have expected Kobe to get himself into a better scoring position as the clock winded down.



Kaggle Score Screenshot

sub.csv
an hour ago by Rajni Goyal
add submission details

0.74965

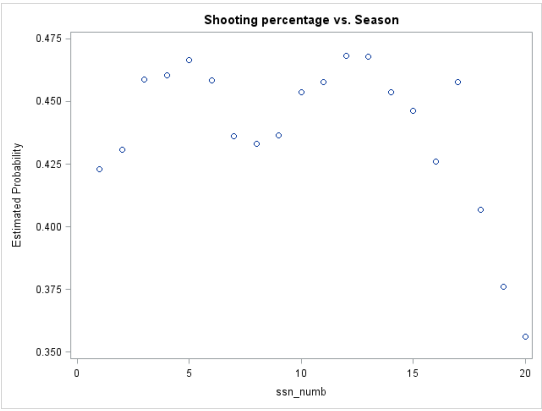
Bonus Question 3

Model Kobe's shooting percentage over time. Does he appear to get better over time? Use your knowledge of the methods we have studied so far to answer this question the best way possible. (Again, you may have 1 additional page to answer this question.)

In order to look at how Kobe's shooting percentage changed over time we look at how his shooting average was affected by season. It was found that season by itself was a significant factor in determining shooting percentage. (p-value < 0.0001, Wald Chi^2 of 63.176) In Figure 16 below we see that Kobe shot at his highest percentage (> 0.45) over a 3-4 season span twice in his career. Those seasons correspond to him first winning 3 championships with Shaquille O'Neal and later 2 championships with Pau Gasol. After those two peaks the decline in shooting percentage coincides with Gasol and head coach Phil Jackson leaving the Lakers. Further decline follows as Kobe gets older and suffers multiple season ending injuries in the last 3 seasons.

The seasons in which his percentage deviates in the most significant way are seasons 5, 12, 13, 19, and 20 (p-values = 0.0269, 0.0125, 0.0135, 0.0017, <0.0001 respectively) which correspond to winning championships in 2000-01 (season 5) and 2008-09 (season 13), also including his final season in 2015-16 where his percentage fell well below 40%

Figure 16



References

Jordan's last shot: https://www.youtube.com/watch?v=vdPQ3QxDZ1s

Code can also be found on the Githup repo for this project: https://github.com/daresnick/Stats-Project-3

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 1996 through 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 shot made
- 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

Some new variables made.

- ttl_sec_remn_per – total seconds remaining in a period – (minutes*60+seconds)
- ttl_sec_remn_half – total seconds remaining in the first half – (minutes*60+seconds if period=2)
- ttl_sec_remn_gam – total seconds remaining in a game – (36*60+minutes*60+seconds if period=1,24*60+ minutes*60+seconds if period=2, 12*60+minutes*60+seconds if period=3, minutes*60+seconds if period=4)
- home_field – 1 if home and 0 if away (use the matchup variable and if there is a vs. then home (1) and if there is a @ then away (0)
- angle - Convert location variables into polar coordinates

SAS Code:

Visit the Github Repo in the Reference to load_clean the data.

```
/* Question 1*/
proc logistic data=kobe_train plots=all outest=estimates1;
class homefield(ref="0");
model shot_made_flag(event='1') = homefield/ CLPARM=PL;
run; quit;
data probmade1;
set estimates1;
prob1=(EXP(Intercept+HomeField1))/(1+(EXP(Intercept+HomeField1
)));
keep prob1;
run;
proc print data=probmade1; run;
proc logistic data=kobe_train plots=all outest=estimates0;
class homefield(ref="1");
model shot_made_flag(event='1') = homefield/ CLPARM=PL;
run; quit;
data probmade0;
set estimates0;
prob0=(EXP(Intercept+HomeField0))/(1+(EXP(Intercept+HomeField0
)));
keep prob0;
run;
title 'Probability of Making a Shot at Home';
proc print data=probmade1; run;
title 'Probability of Making a Shot Away';
proc print data=probmade0; run;
/* Question 2 and 3*/
proc logistic data=kobe_train plots=all outest=estimates1;
model shot_made_flag(event='1') = dist/ clparm=both;
output out=probkobe prob=prob;
run; quit;
```

```
/* Question 4*/
proc logistic data=kobe_train plots=all outest=estimates1;
class playoffs(ref="0");
model shot_made_flag(event='1') = shot_distance|playoffs/
clparm=both;
run; quit;
proc logistic data=kobe_train plots=all outest=estimates1;
class playoffs(ref="0");
model shot_made_flag(event='1') = shot_distance|playoffs/
clparm=both;
output out=prob1k prob=prob;
run; quit;
data prob2k;
set prob1k;
odds=prob/(1-prob);
run;
title 'Predicted Odds for shot_made_flag=1';
proc sgplot data=prob2k;
scatter x=shot_distance
                    y=odds;
Run;
/* Question 5 */
proc logistic data=kobe_train plots=all outest=estimates1;
class playoffs(ref="1") homefield(ref="1");
model shot_made_flag(event='1') =
shot_distance|playoffs|homefield/ clparm=both;
run; quit;
/* Question 6 */
proc logistic data=kobe_train plots=all outest=estimates1;
class clutch(ref="1");
model shot_made_flag(event='1') = shot_distance|clutch/
clparm=both;
run; quit;
```

```
data kobe_sub;
set probkobe;
if shot_distance > 40 then DELETE;
run;
proc sgplot data=kobe_sub;
scatter x=shot_distance
                        y=prob;
run;
proc reg data= kobe_sub;
model prob=shot_distance;
Run;
```

```
/*Predictive Model*/
proc logistic data=kobe1 plots=all;
class  action_type  ssn_numb  combined_shot_type_num
shot_zone_area_num  shot_zone_basic_num
  shot_zone_range_num  shot_distance  clutch period;
model shot_made_flag(event='1') = ttl_sec_remn_gam  action_type
ssn_numb  shot_zone_area_num  shot_zone_basic_num
  shot_zone_range_num  shot_distance  clutch/clparm=both ;
output out = SS_PRED predicted = I;
run;
/* Bonus 3 */
proc logistic data=kobe_train plots=all outest=estimates1;
class ssn_numb;
model shot_made_flag(event='1') = ssn_numb;
output out=probdate prob=prob;
run; quit;

title 'Shooting percentage vs. Season';
proc sgplot data=probdate;
scatter x=ssn_numb
          y=prob;
run;
```