**Is KOBE Going to Score His Next Shot: Let Us Tell You**

Allen Ansari, Yongjun Chu and [Solange Garcia de Alford](https://2ds.datascience.smu.edu/my/index.php?id=1394)

# **INTRUDUCTION**

Kobe Bryant is one of the greatest basketball players in the history of National Basketball League. He scored 38,024 point over 20 years of his professional career and won 5 NBA championships. He took a lot of shots. Some went in the basket, while other didn’t. By identifying what factors affect his scoring potential, we may be able to predict or classify whether his next shot will get in the basket. The potential findings may be interesting and beneficiary to those current or future NBA players who want to improve their scoring potential of a shot. In this paper, our goal is to generate at least 2 models to best predict the shots made by Kobe Bryant on 5,000 occasions, based on the detained information about the rest of shots he took during his career. Specifically, we will build two different prediction models: Logistic Regression Model and Linear Discriminant Analysis (LDA) model with cross-validation using a training partition to derive the rule of the models’ algorithms and a test partition of the data to apply classification rules and predict the result of 5,000 shots made but with outcomes unknown.

## **MATERIAL AND METHODS**

## **Data Description**

## The original data set contains a total of 30,697 shot attempts by Kobe Bryant in his 20 years’ career with related data to his shots attempts. The data was partitioned as follows:

## 25,697 records which were used as the training set in the models.

## 5,000 records with the shot\_made\_flag values removed for use as test set in LDA predictions.

The data contains 29 variables which listed as below including brief description.

We found that transformations were not necessary for any of the variables. We found that we needed re-coding of several categorical variables into indicator variables for generating one of the LDA models when using SAS’ Proc DISCRIM’s Var statement. Proc LOGISTIC automatically re-codes the variables and presents these results in its resulting Design Variables matrix. Refer to Appendix D – Indicator Variables.

**Data Mining Approach and Evaluation**

**Data Mining Methods**

**Variable and Model Selection**

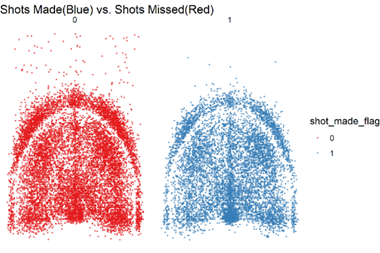
Schwarz Bayesian information criterion (SBC) was selected for each step of modeling evaluation for different techniques. Cross Validation PRESS (CVPRESS) was chosen as the stop criteria for further modeling regression.

|  |  |
| --- | --- |
| **action\_type:** The type of shot attempted, such as jump shot, dunk, etc.**combined\_shot\_type:** Classifies the shots under 6 larger categories: Bank Shot, Dunk, Hook Shot, Jump Shot, Layup, and Tip Shot.**Matchup:** The two teams in the specific match. Since Kobe was always on the Lakers and opponent contains all the information in matchup, we decided to reduced number of levels for this variable by dividing all games into Home and Away category.**Opponent:** Opponent in the specific match.**Season:** The basketball season (2000, 2001, etc.)**shot\_type:** includes categories 2pt or 3pt.**shot\_zone\_area:** Area from which shot was attempted (Right, Left, Center, Back Court, Right Center, Left Center)**shot\_zone\_basic:** Further area information (Mid-range, restricted area, in the paint, above the break 3, backcourt, left corner 3, right corner 3)**shot\_zone\_range:** Range (<8 ft, 8-16, 16-24, 24+, backcourt)**team\_name:** Name of Kobe’s team, the Lakers, so we decided to remove it from dataset.**arena\_temp:** average temperature of are*na***attendance**: Number of people who watched the game | **avgnoisedb:** Average noise level in the arena in decibels**game\_date:** Date of the specific match.**game\_event\_id:** **game\_id:** NBA game ID**lat:** The latitude of Kobe’s position during the shot attempt.**loc\_x:** The x-location on the court.**loc\_y:** The Y-location on the court.**Lon:** The longitude of Kobe’s position during the shot attempt.**minutes\_remaining:** The minutes remaining in the specific match**period**: The period in the specific match**playoffs:** binary, 1/0 values**recId****seconds\_remaining:** The seconds remaining in the specific match**shot\_distance: The** distance from which the shot was attempted, in ft.**shot\_id:** (from 1 to 30,697) of the attempted shot**shot\_made\_flag:** it is response variable and indicates if shot was successful (as 1) or not (as 0)**team\_id**: ID of Kobe’s team. Always the Lakers, so removed |

**RESULTS AND DISCUSSIONS**

## **Exploratory Data Analysis**

We explored the remaining predictors to find meaningful relationships with shot\_made\_flag. First, we looked for correlation between variables

**Figure 1. Title…. Figure 2**

**Figure 1** shows there is high correlation between loc\_y and lat and shot\_distance which make sense because loc\_y, lat shows distance from basket. Additionally, there is high correlation between attendance and average noise db.

## Next, we examined location data as shown in **Figure 2**. By plotting loc\_y vs. loc\_x a visualization of the shots Kobe made and missed by location is depicted. It is rather difficult at first glance to discern any differences. One point that becomes obvious, however, is the impact of range. There are many more misses than makes at longer ranges, meaning the 3-point line and beyond. Within the 3-point area the data is too noisy to analyze.

**Figure 3** provides a visualization of shot\_zone\_area, showing the on-court representation of each zone. The accuracy of shots in each zone shown in **Figure 4**. As expected, the accuracy for shots from the backcourt is extremely low.

A close up of a map

Description automatically generatedA screenshot of a cell phone

Description automatically generated

**Figure 3 ….. Figure 4 Title**

Now, we are looking at shot\_zone\_basic. Figure 5 provides a visualization of the on-court locations, and Figure 6 the accuracy and number of shots by location. Kobe’s accuracy by shot\_zone\_basic actually varies substantially and surprisingly, Kobe’s left corner accuracy is higher than right corner accuracy.

A close up of a piece of paper

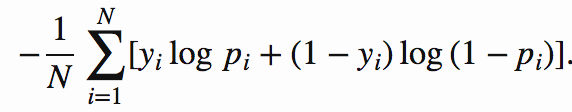
Description automatically generatedA screenshot of a cell phone

Description automatically generated

**Figure 5 Figure 6**

**Logistic Regression Model More Accurate at Predicting the fate of a shot**

To predict if Kobe’s next shot will go into the basket, two kinds of classification models were chosen for comparison: logistic regression model and linear discriminant analysis model. To compare how these two models perform we first generated a train data set (80%) and test dataset (20%) based on the whole dataset containing 25697 observations. For logistic regression, we first identified 11 statistically significant predictor variables based on the training dataset: action\_type, attendance, arena\_temp. These variables were found using Proc Logistic procedure with stepwise selection method. Forward selection method yielded the same set of significant variables. For LDA modeling, we only kept the quantitative variables in the modeling (SAS Proc Discrim) due to the requirement from the LDA model assumptions. These quantitative variables are: shot\_distance, … The regression parameter estimates were then derived for each model and further were applied to the test dataset. We obtained the mis-classification rate, sensitivity, specificity and for each case with test dataset. To obtain the AUC (Area Under the Receiver Operating Characteristic (ROC) Curve), we used the R package , “MLmetrics”, which calculates the AUC based on the ROC between true positive rate and false positive rate. To access the model fit, we also calculate the log loss function, which is defined as the following,



Where N is the total number classifications, yi is the shot\_made\_flag and pi is the probability from the model of each outcome (shot made). Table 2 shows the comparison results.

**Table 2.** The comparison of model fitting with test dataset for Logisitic regression model and LDA.

|  |  |  |
| --- | --- | --- |
| **Metric** | **Logistic Regression Model** | **Linear Discriminant Analysis** |
| AUC | 0.6877 | 0.5680 |
| Mis-Calculation Rate | 0.3257 | 0.4276 |
| Sensitivity | 0.4613 | 0.6245 |
| Specificity | 0.8485 | 0.5290 |
| Log LOSS function value | 0.6182 | 0.6972 |

The comparison results in Table 2 clearly suggested that logistic regression model out performed the LDA with the input data. The former has higher AUC value by more than 10% than the latter. Logistic modeling also has better specificity and lower log LOSS function value. All these results indicate that logistic regression model fits the data clearly better. Therefore, we used the logistic regression model to further study specific questions of interest within the dataset.

**Checking the Fitting of Logistic Model on Whole Input Data by Cross Validating (with one left out method)**

This time we applied logistic model to check how it performed on the whole input dataset (25697 observations). In the modeling, we first used stepwise or forward methods to identify significant variables. With no surprise, the 11 variables identified earlier with training dataset were identified again here. The Hosmer and Lemeshow Goodness-of-Fit test gave a p-value of 0.56, suggesting there is no evidence that the logistic modeling with 11 variables is a bad fit. To find out how well this model with 11 variables fit the data, we carried out the classification for each observation in the whole input dataset. To reduce the biases due to the predicting of itself, we used the cross validation by leaving one out method, which is easily carried out in SAS Proc Logistic procedure. In this method, all the rest of observations are used like a training dataset and prediction is carried out on that one observation. The fit statistics is shown in **Table 3**. The fit results suggest that the logistic model with 11 variables fit the data well, with a log LOSS function value of 0.605. The ROC curve is shown in supplementary (**Figure S1**). When plotting the diagnostic results for finding inferential data points, we found that observation 5850 may have relatively large influence on model fit and parameter estimates (**Figure S2**). Further investigation suggests the parameter estimates don’t change much without this observation.

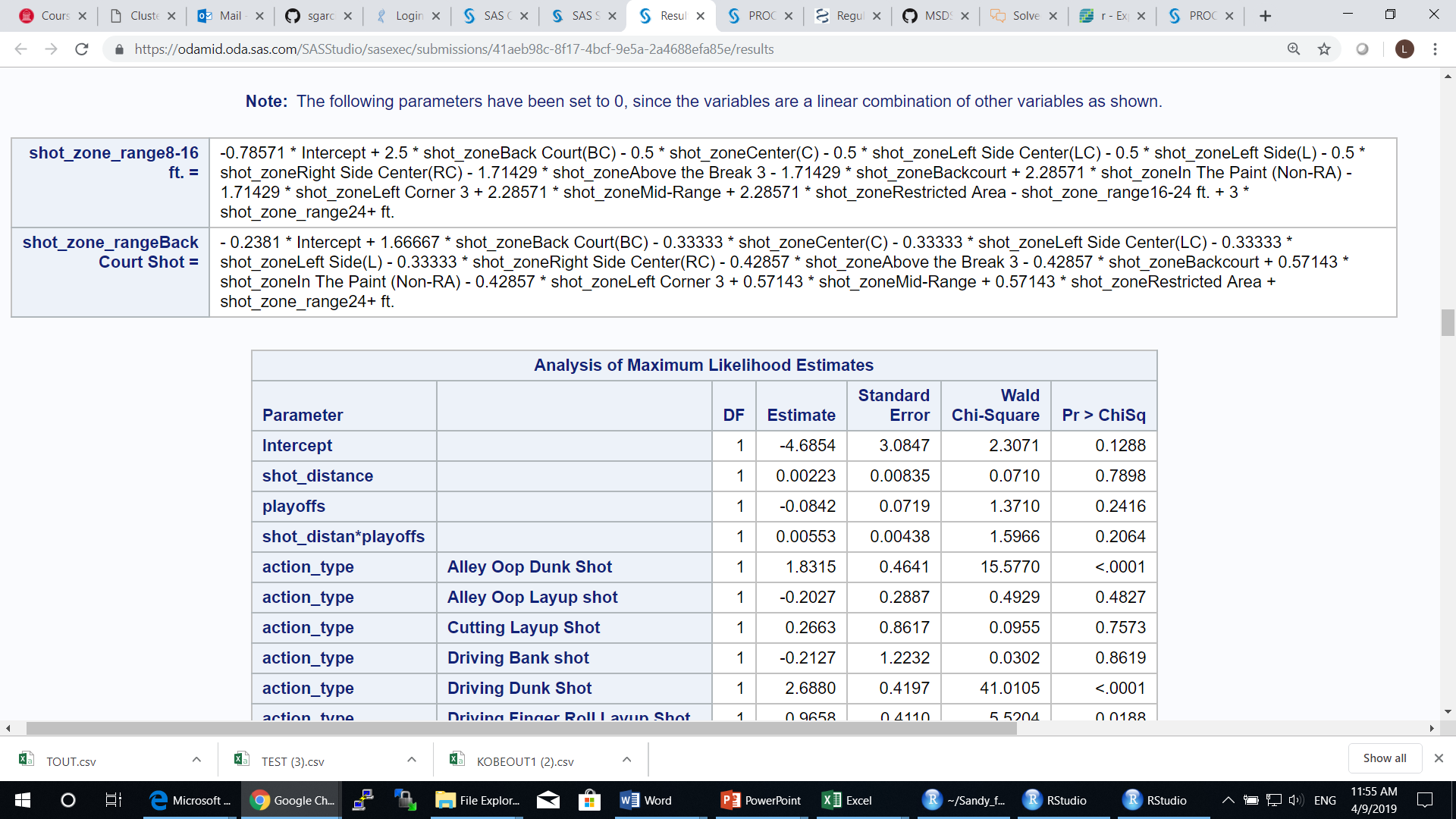
**Table 3**. The modeling fit results of logistic regression with whole input dataset.

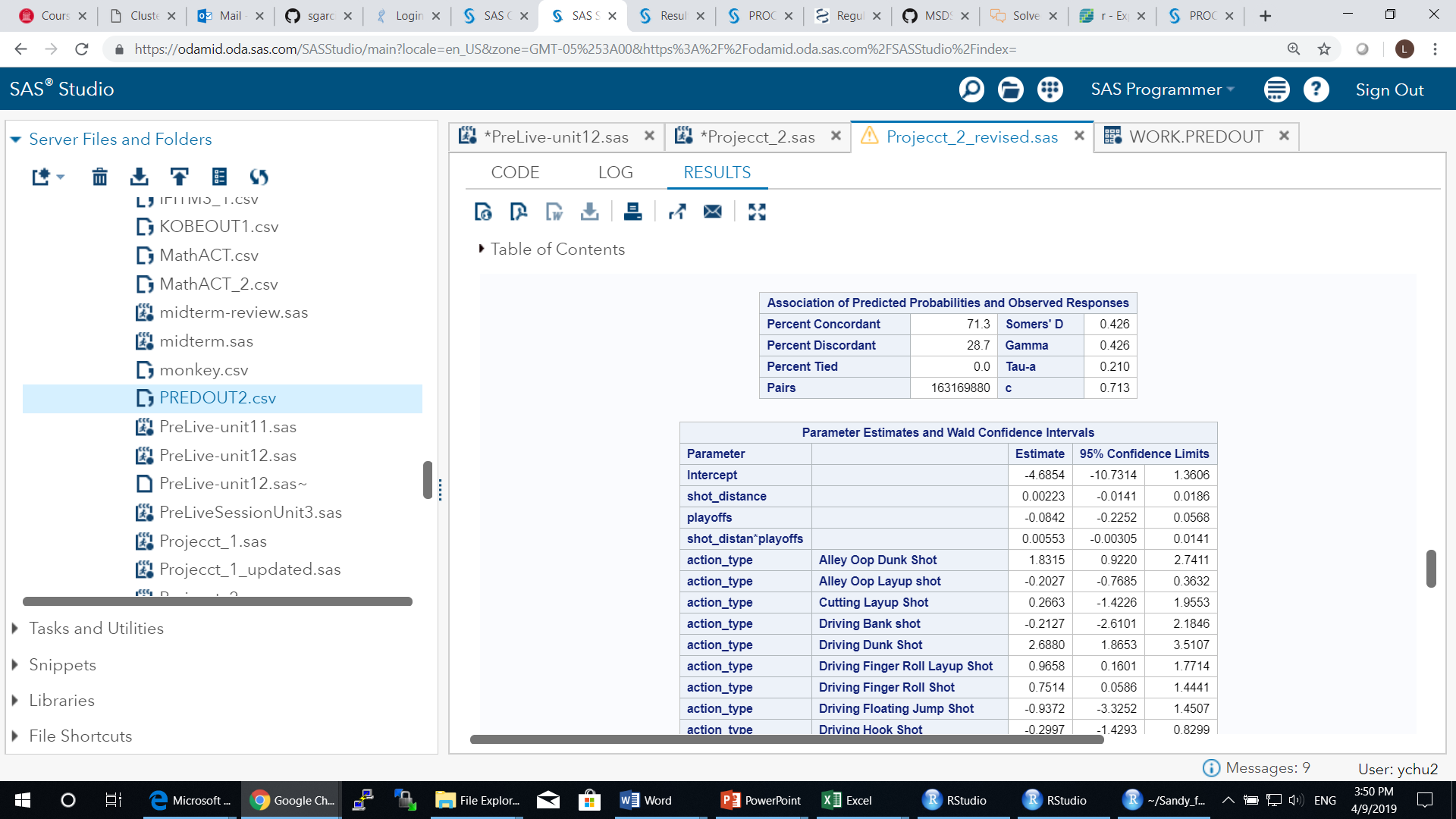
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **AUC** | **Mis-Calculation Rate** | **Sensitivity** | **Specificity** | **Log LOSS function value** |
| 0.6956 | 0.316 | 0.4633 | 0.8618 | 0.605 |

**Shot Distance and Playoffs not Significant Predictors on Making a Shot**

A common perception is when a basketball player is closer to the basket, it would be easier to make a shot. The reality is often more complicated than that. This is simply because the defensive players can do a better job of defending when an offensive is closer to the basket for scoring. Another interesting question is how Kobe performed in playoff games relative to the regular season games and whether the playoffs affect the relationship between shot distance and his scoring potential. To answer these questions we applied a logistic regression model to the whole input dataset (25697 observations). Although shot\_distance and playoffs were not among those 11 significant variables identified above, we included them and also the interaction term shot\_distance\*playoffs along with other 11 significant variables in the model fitting. The Hosmer and Lemeshow Goodness-of-Fit test gave a p-value of 0.3678, suggesting there is no evidence that the logistic modeling with 14 variables is a bad fit. The fitting statistics of shot\_distance, playoffs, and their interaction term are shown below.

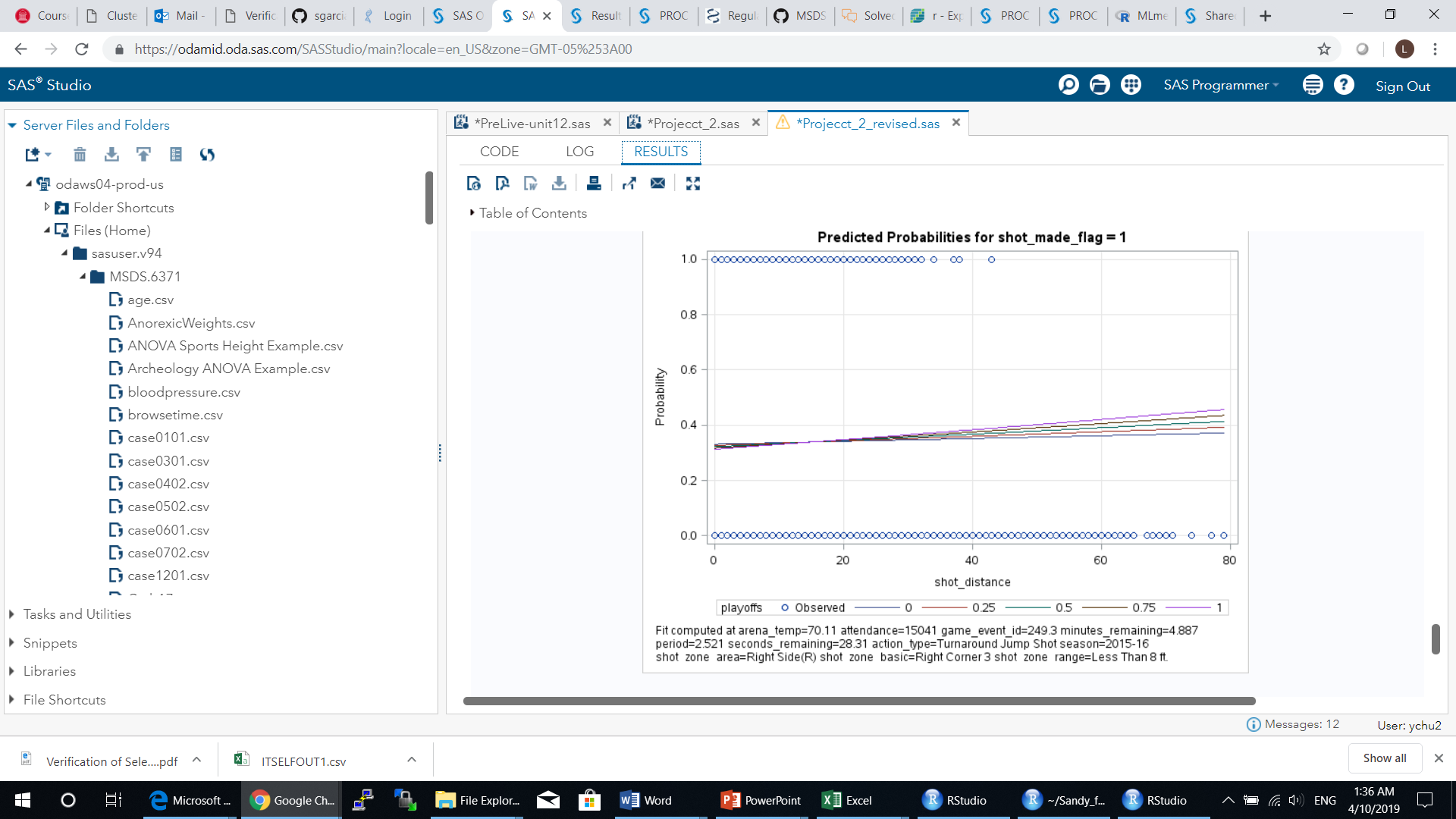
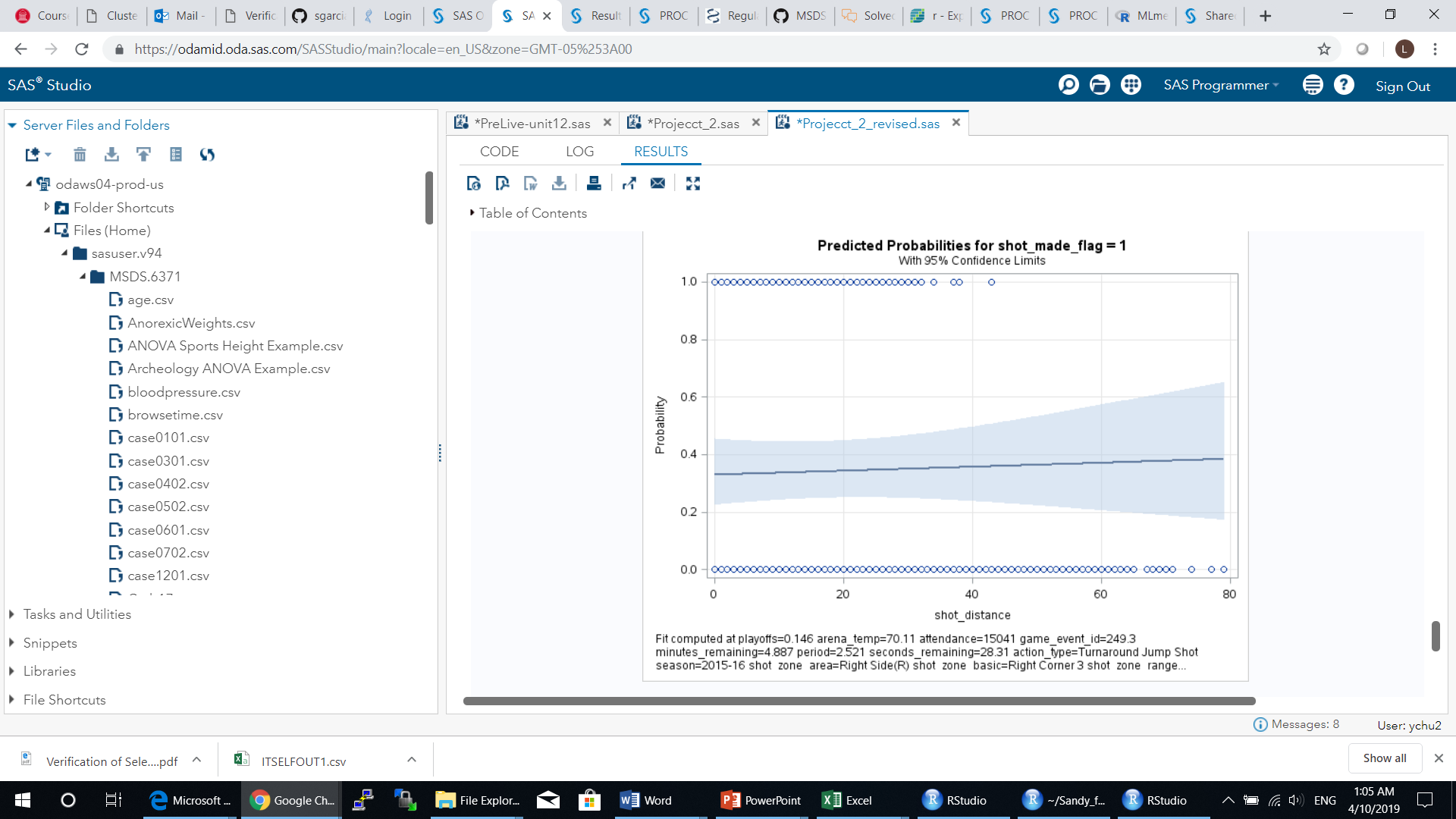
**Table 3**. The fitting statistics of shot\_distance, playoffs, and shot\_distance\*playoffs.





As shown in Table 3, log(odds of making a shot) increase 0.00223 for every unit increase of shot\_distance, or the odds of scoring a shot is 1.0022 when Kobe is one unit away from the basket. Although this seems odd, it can be understood in this way: when he is not close to basket, he has more space to operate by using favorable moves to score. When he is closer to basket, pressure goes up as more defensive players would crowd him, making him harder to score. The relationship between the probabilty of Kobe making a shot and the distance he is from the hoop is shown in **Figure 3** after accounting for other variables. It appears that there is a slightly positive linear relationship between distance and probalibity to score. The 95% CI band is also shown in the **Figure 3**.

On the other hand, the p-vlaue of 0.7898 and 95% CI of [-10.73, 1.36] which contains zero for the coefficient estimate mean that there is no significant relationship between shot\_distance and whether he will make shots after accounting for other variables. This indicates that other significant variables aleady explained well the chances that he will make a shot or not. Similar explainatin can be applied to playoffs, which has a p-value of 0.2416 and 95% CI [-0.2252, 0.0568]. The estimated coefficient for interaction term between shot\_distance and playoffs is not significant either (p-value=0.2064, 95% CI [-0.003, 0.014]) (**Figure 3**). Based on this analysis, we concluded that whether Kobe made a shot or not is not significantly related to the distance he is from the hoop. The relationship between the distance Kobe is from the basket and the odds of him making the shot is not different if they are in the playoffs.



**Figure 3**. The relationship between probability of Kobe scoring and the distance to the hoop.

**Predicting the Outcomes of 5000 Kobe’s Shots**

By using the best model identified, that is, the logistic regression model, we predicted the outcome of 5000 shot Kobe took during his career. In the prediction, we used all the input data points of (25697 observations) as the training dataset and 11 significant variables we identified earlier. The prediction results have been uploaded.

**CONCLUSIONS**

In summary, we have found that logistic regression model is a better fit than LDA for the input data presented here. The logistic model yield higher prediction accuracy, sensitivity and lower log LOSS function values. Eleven variables were found to be significantly affecting Kobe’s scoring potential.

It may seem obvious that the closer a basketball player to the basket, the odds of scoring is increasing. This is should be case if there is no defense around. Our investigation of this question with Kobe’s shooting dataset suggests that this is not the case. Distance is not that important. Other factors affect his scoring potential more significant. For example, when Kobe is close to the basket, the action type of a slam dunk has much higher chance to score than a bank shot. We also looked at if the playoffs changed his scoring potential with regard to his distance from basket. The results suggest no existing of such significant impact from playoff circumstances.

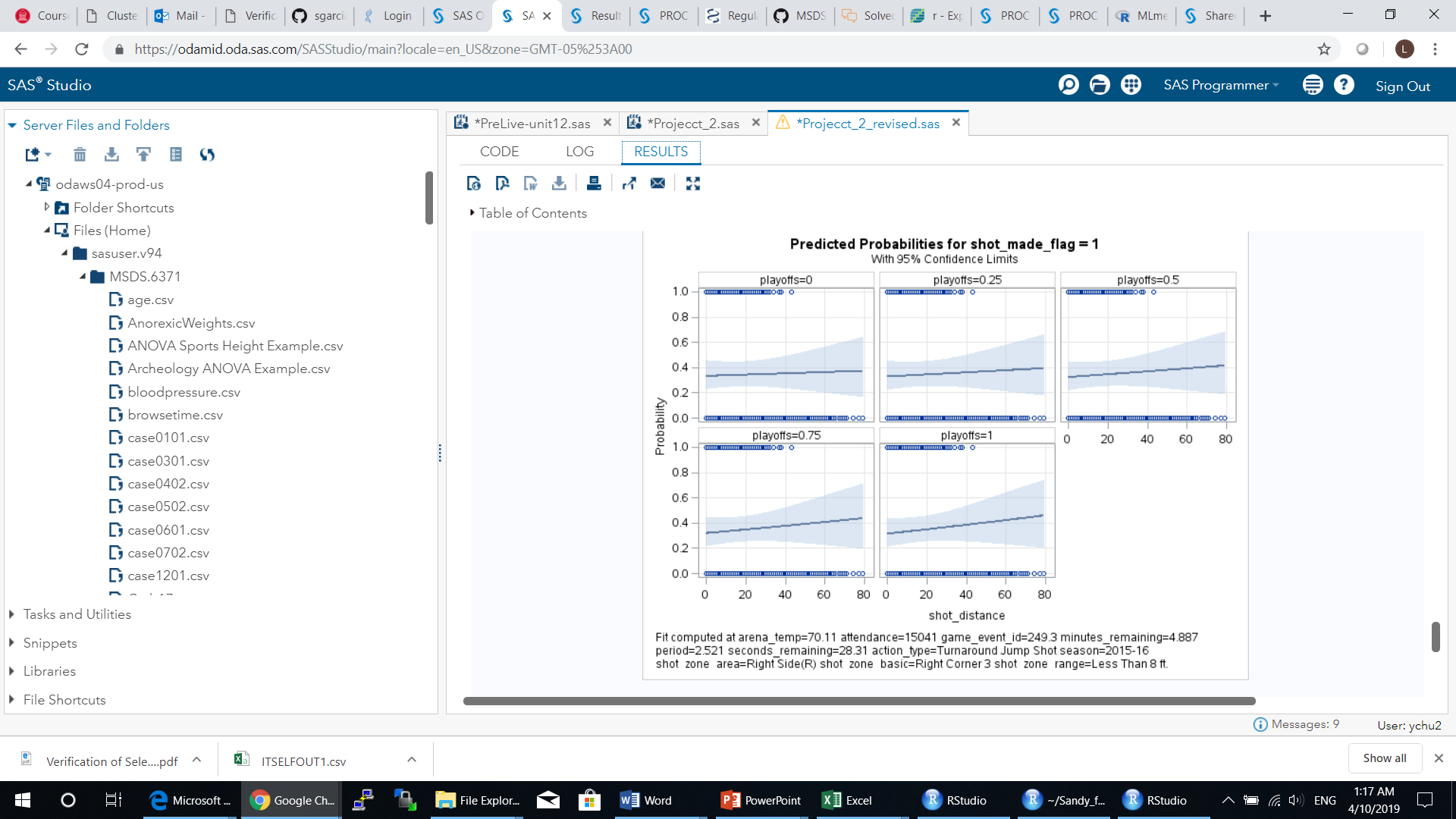
**REFERENCES**

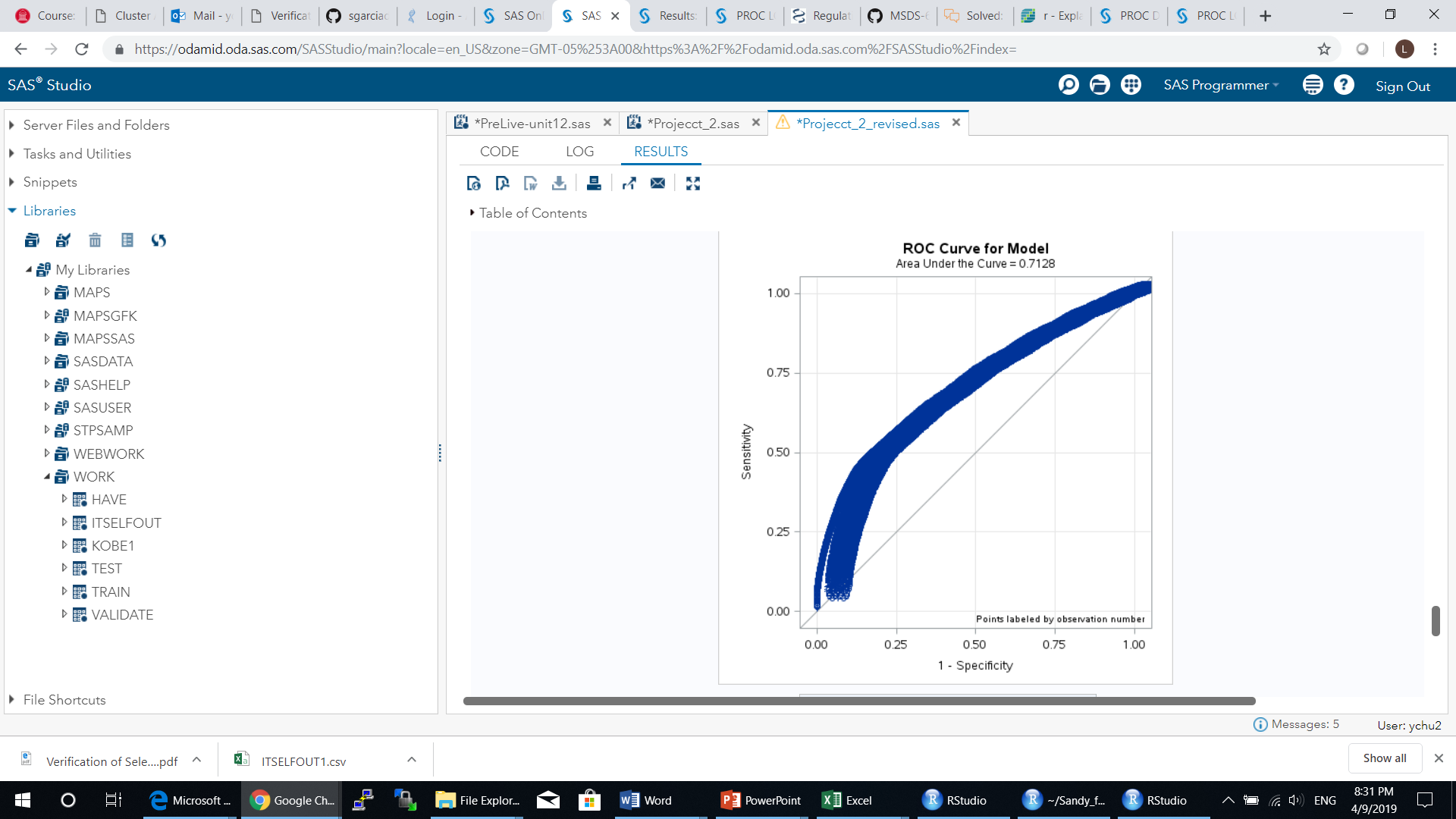
1.

2.

3.

**Supplementary figures:**





Plus other interesting facts on Kobe’s number of shots from different zone\_basics:

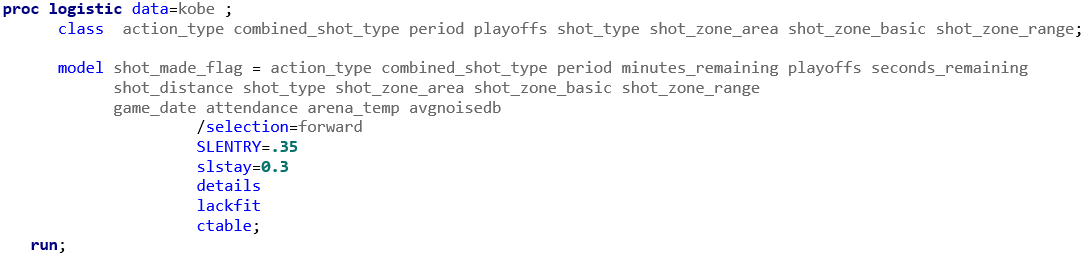
A close up of text on a black background

Description automatically generated

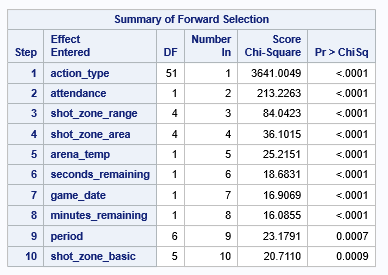
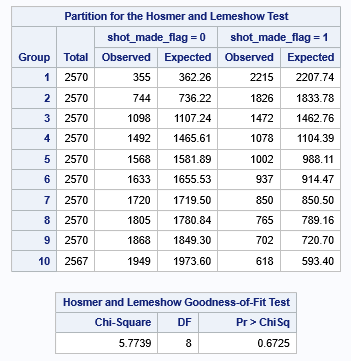
# Statistical Modeling

## Logistic Regression Model

We ran the model below with forward variable selection:



The model’s summary of forward selection is shown below.

A question of interest in the project was what are the odds that Kobe would make a shot decrease with respect to the distance he is from the hoop. We found no evidence of this, the forward selection model found that shot\_distance was not a significant, as you can see in the results above. Similarly, the playoffs variable was not significant. We concluded that Kobe was performed well and similarly in all games.

1. CONCLUSION

# We failed to reject the null-hypothesis that the model has a good fit at alpha 0.05 p-value (0.6725) with a Hosmer and Lemeshow Goodness of Fit test.

# At the 50% probability level, our model had a 68.2% accuracy, a sensitivity of 85.9% and a specificity of 46.%. Please refer to the Classification Table below.

