PROJECT FINAL REPORT

2019 - R STATISTICAL PROGRAMMING (ISGB-799V-001)

Group 6

*Gege Tao, Ila Srivastava, Lakshmi Sravya Katta, Junyan Zhao, Zhen Wu*

**Introduction**

For our project, the problem we are trying to solve involves predicting whether former NBA star Kobe Bryant makes or misses a shot based on several factors of the shot, such as distance from basket, type of shot, and opponent. This will help us analyzing the relationships between these variables and finally, coming up with an accurate model predicting shot accuracy.

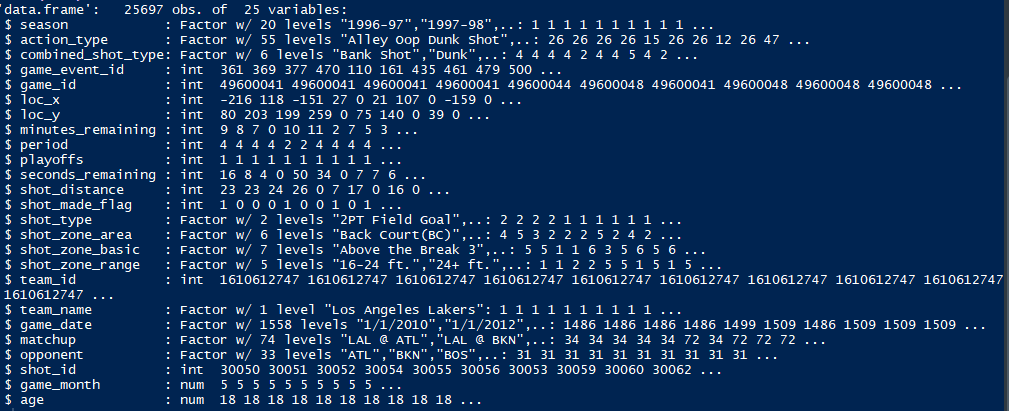
Initially, we intend to use exploratory data analysis (EDA) i.e., an approach to analyzing data sets to summarize their main characteristics, often with visual methods. EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. We will try to analyze various indicators of shot accuracy such as age, month, shot count, type of shot, shot distance to understand the correlation between these factors, while accounting for control variables. We will do extensive visual analysis to understand the importance of basketball court and competition condition. We will build regression models such as logistic regression, ridge and lasso to compare their performance in terms of accuracy, and coefficients of the lambda that produces the smallest mse (mean squared error) in order to identify the best one for predicting whether a shot is successful or not.

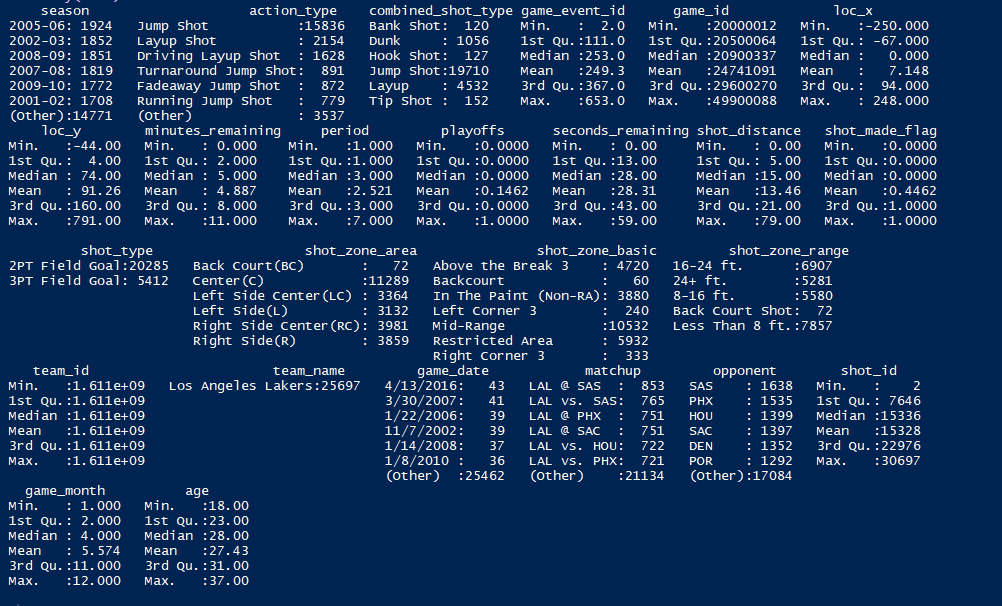
**A. Pre-analysis**

***1. Summary Statistics***

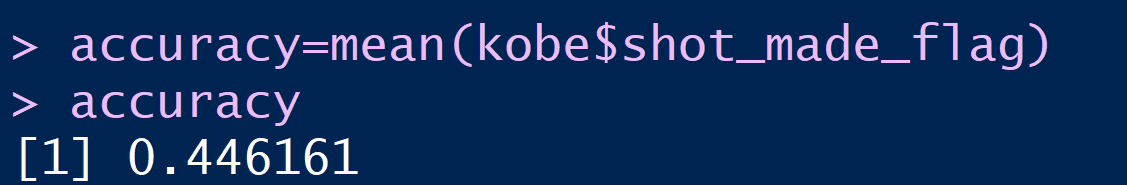
Firstly, we took a primary look at the dataset. As the following figure shows, there are 25697 records and 23 original variables and two new variables we developed ‘game\_month’ and ‘age’. The datatypes of each variable are mainly numeric and factor.

Also, the figure indicates that there is no missing value and We can see the mean, median, quartiles, etc. for all the columns.





And we calculated Kobe’s general shot accuracy shown as the following figure, which is 0.446161.



***2. Feature Development:***

The following changes have been made to our predictor variables up to make them have more impact on the model.

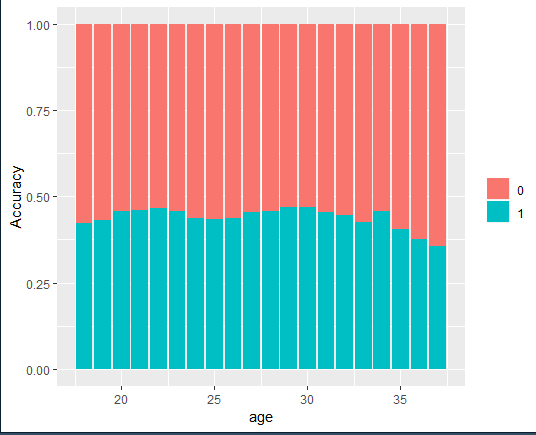
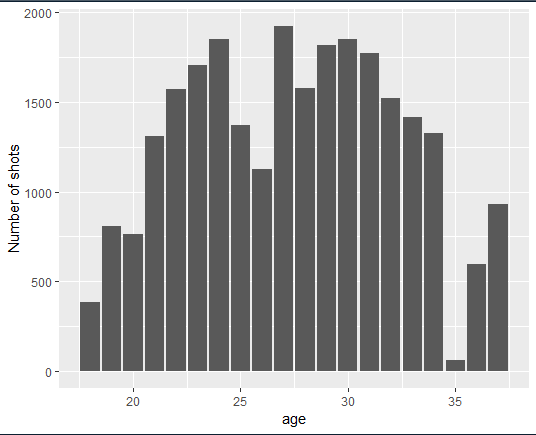
* Extracted month from column 'game\_date' generating a new column called 'game\_month'.
* Created a new column called ‘body\_age’ by adding Kobe's age data of each season. The reason why we added this new column was because we considered that if we wanted to analyze in which month Kobe performed better.

***3. Exploratory Data Analysis***

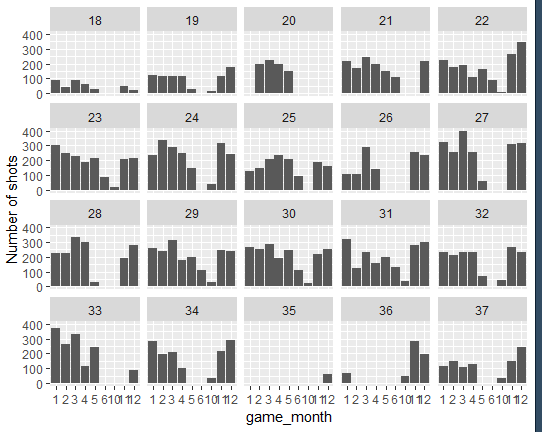
* *Relationship between Kobe’s demographic information and his shot accuracy*

For the EDA, we firstly analyzed the relationship between Kobe’s demographic data and his shot accuracy.

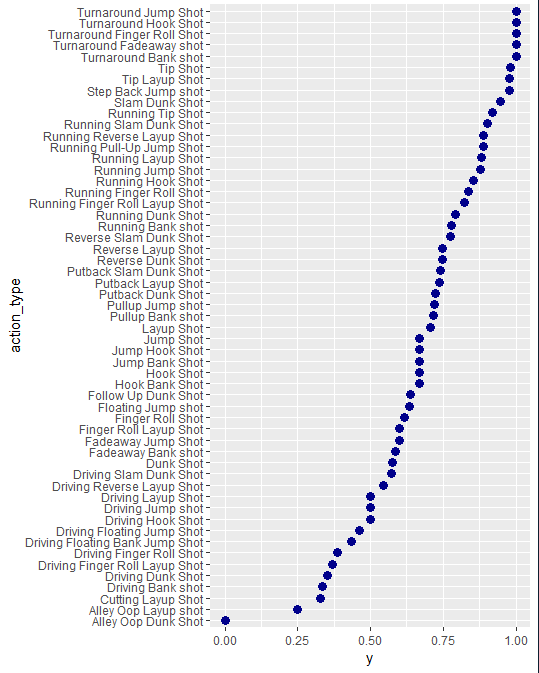
The following figure shows the relationship between Kobe’s age and shot accuracy. When Kobe was 21 to 34 years old, he made more shots and during Kobe’s NBA career, he always had a steady performance in accuracy although there is a decreasing trend after age 34, probably because of his higher body age.



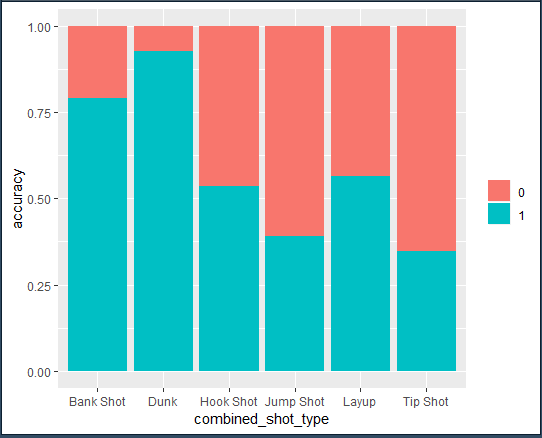
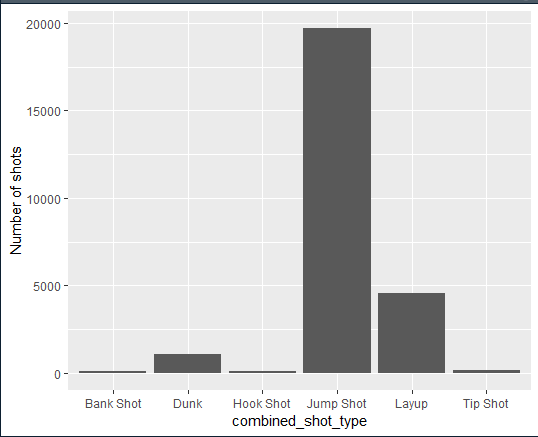
The following figure shows the Relationship between month and Kobe's shot count, accuracy with Kobe's age as control variable. This is to answer which month he played better when he was a certain (x) age. Again confirmed that when Kobe was 21 to 34 years old, he made more shots and during Kobe’s NBA career, he always had a steady performance.



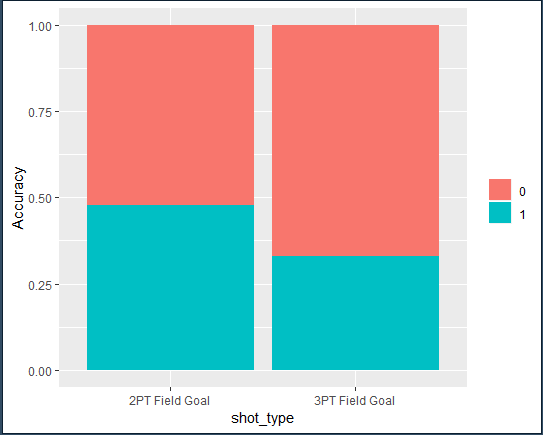
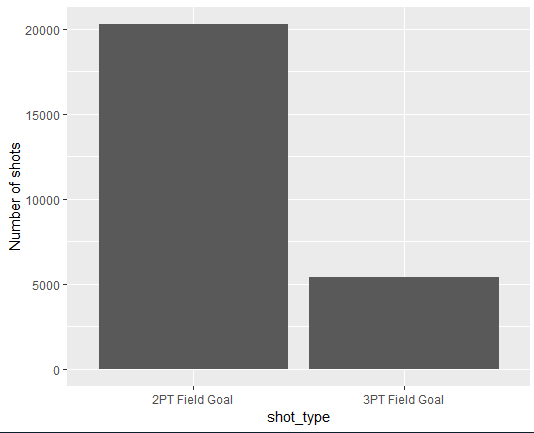
The following figure shows the relationship between Kobe's shot accuracy and his action type. There are 55 types of his action type. The shot accuracy varies substantially by his shot type.



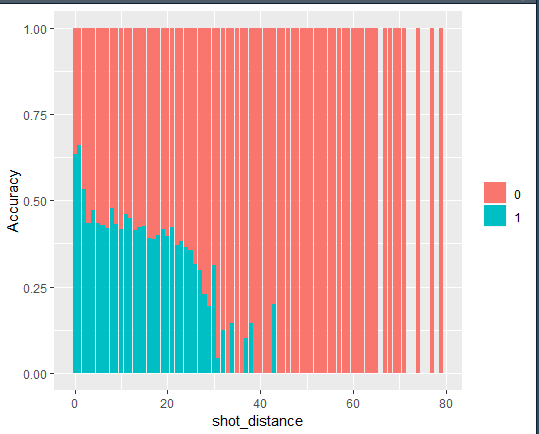
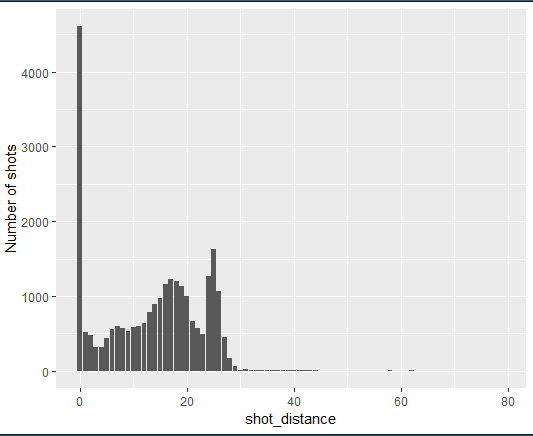
The following figure shows the analysis of Kobe’s combined shot type and its relationship with shot accuracy. Kobe made the maximum number of shots in jump shot and he seldom made shot in bank shot, hook shot and tip shot. But surprisingly, jump shot has the second lowest accuracy and bank shot, dunk have the highest accuracy.



The following figure shows the analysis of Kobe’s 2pt/3pt shots and its relationship with shot accuracy. Kobe made more 2pt field goal shots and the 2pt field goal shots have a higher accuracy.



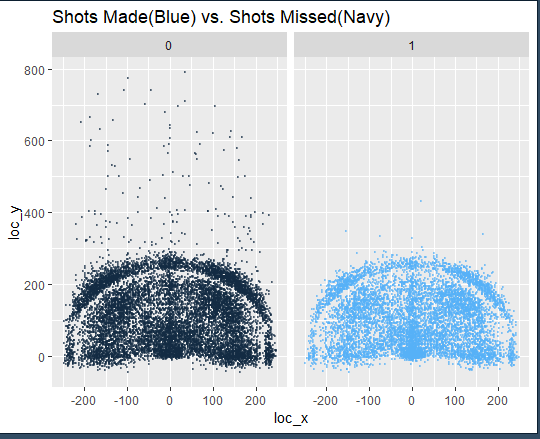
The following figure shows the analysis of Kobe’s shot distance and its relationship with shot accuracy. Instead of long-distance shot, he preferred very short-distance and middle-distance shot and he actually performed better in very short-distance shot, which corresponds to the above analysis that he made more 2pt field shots and performed better in this.



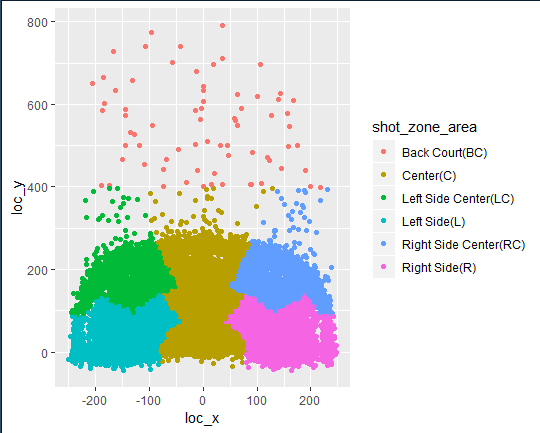
* *Relationship with Basketball Court and Kobe’s Shot Accuracy*

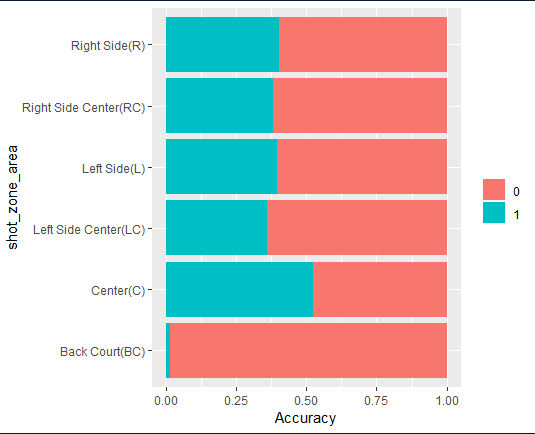
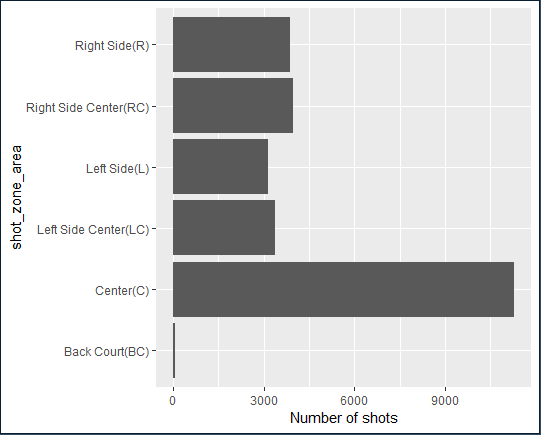
Then, we analyzed the basketball court influence on Kobe’s shot accuracy.

The following figure shows the analysis of whether shot has been made or not by position. 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.

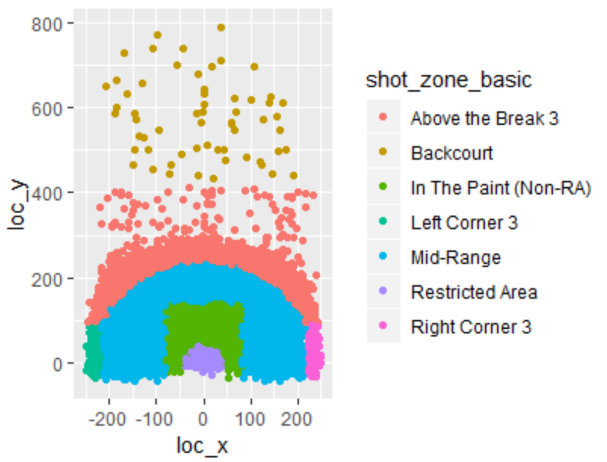


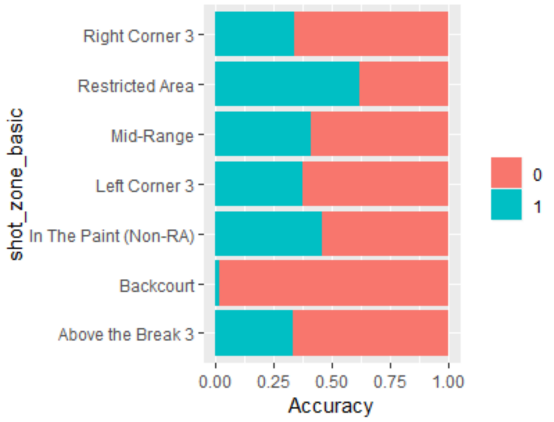
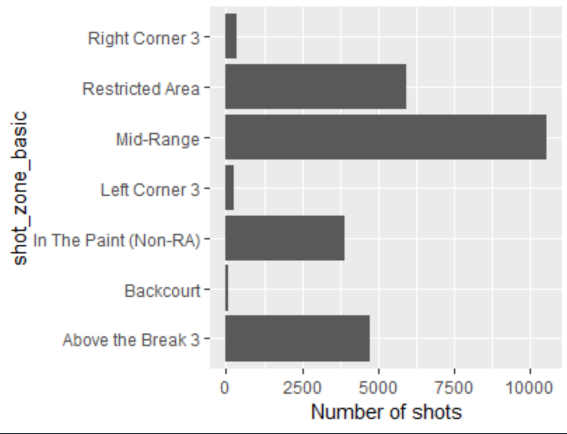
The following figure shows the analysis of the shot zone area in which Kobe performed better, also shows the on-court representation of each zone. Clearly, Kobe preferred shot in center area and he did perform the best in this area. In the back-court area, Kobe made the least shots and had the worst performance, but this is reasonable because shots in such area are hardly ever attempted by basketball players. Among other areas, Kobe slightly preferred the right side probably his right hand is dominant.



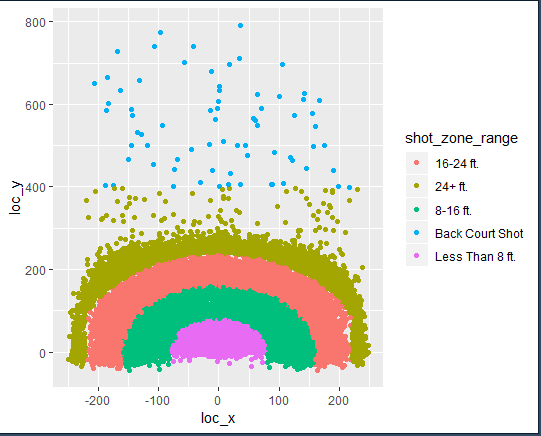


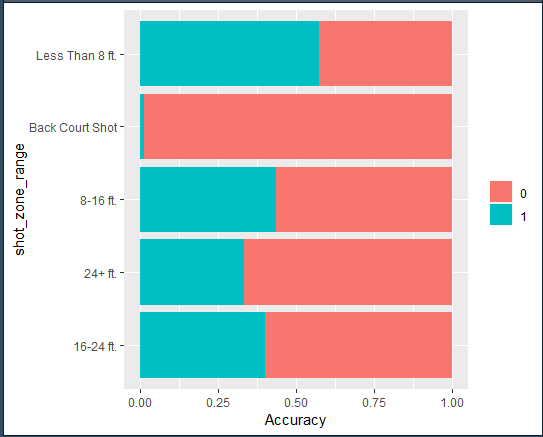
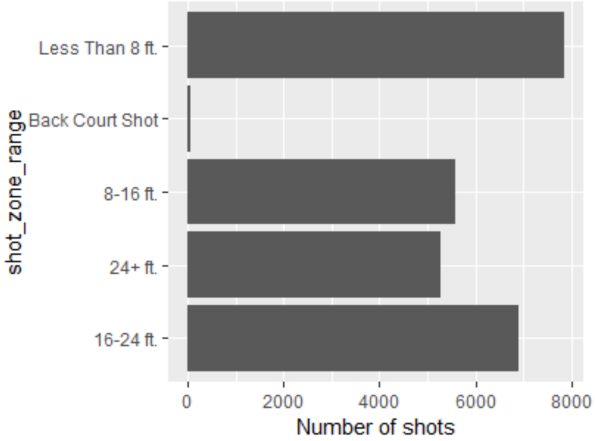
The following figure shows the analysis of the shot zone basic in which Kobe performed better. The number of shot and accuracy vary substantially by shot\_zone\_basic. Clearly, Kobe preferred shot in mid-range area and restricted area and he did perform the best in these two areas. Also, although Kobe seldom made shots in left corner, right corner and backcourt, he did well in the left corner and right corner. Surprisingly, Kobe performed better in left corner that he did in right corner, which indicates that he is an all-round basketball player.





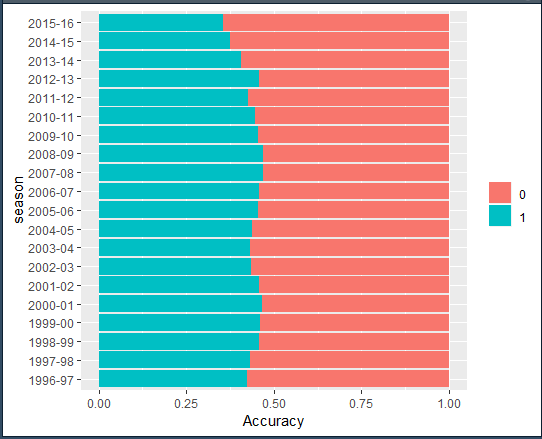
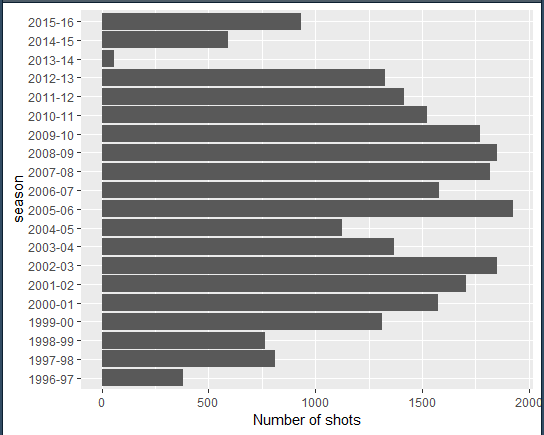
The following figure shows the analysis of shot zone range in which Kobe performed better. The outcome is corresponding to the above analysis of Kobe’s shot distance and its relationship with shot accuracy.



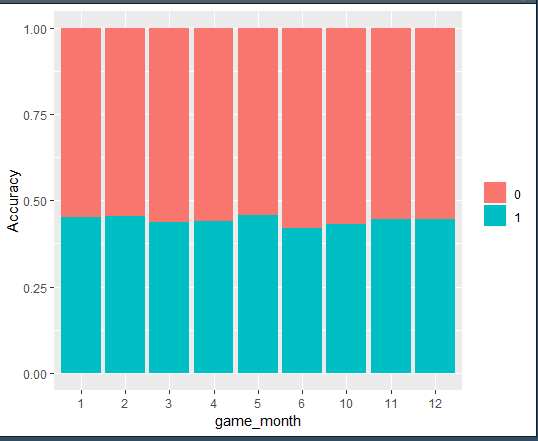
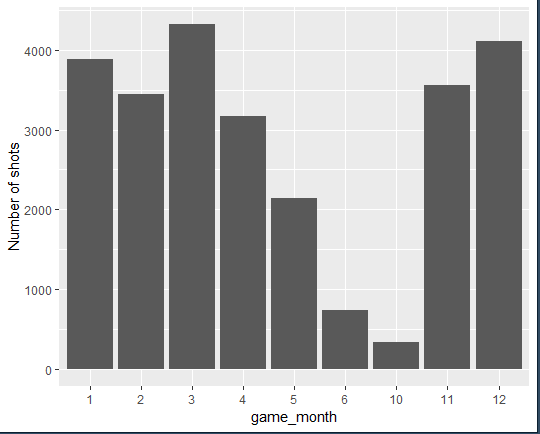


* *Relationship between Competition Condition and Shot Accuracy*

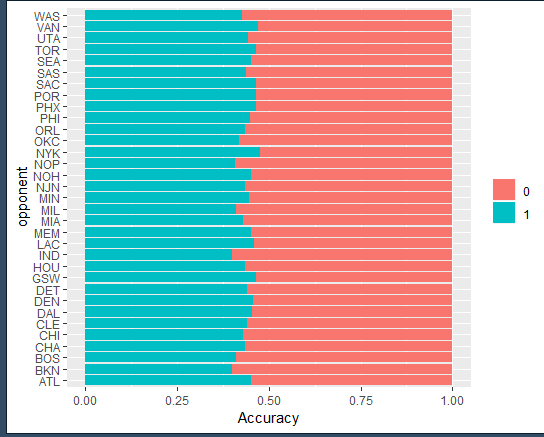
The following figure shows the analysis of in which season Kobe performed better. The period from season 1997-98 to season 2012-13 was Kobe’s career heyday. During this period, Kobe made more shots. He always had a steady performance in accuracy although there is a decreasing trend after season 2012-13, probably because his higher body age. The outcome is corresponding to the analysis of Kobe’s age influence on his shot accuracy.



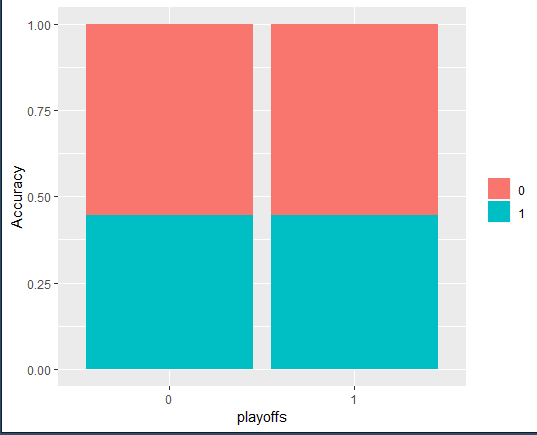
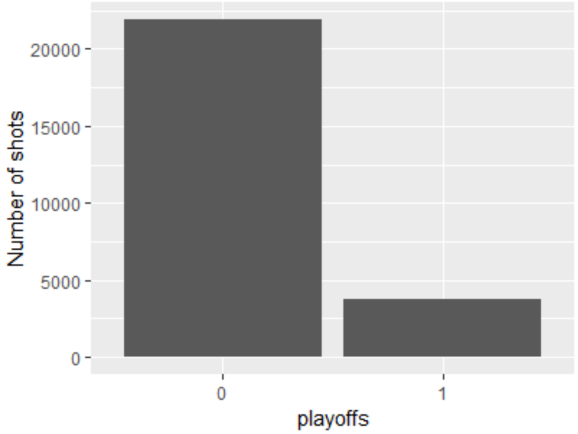
The following figure shows the analysis of month in which Kobe performed better. Although Kobe made more shots in autumn and winter, he always had a steady performance no matter in which month the competition was held.



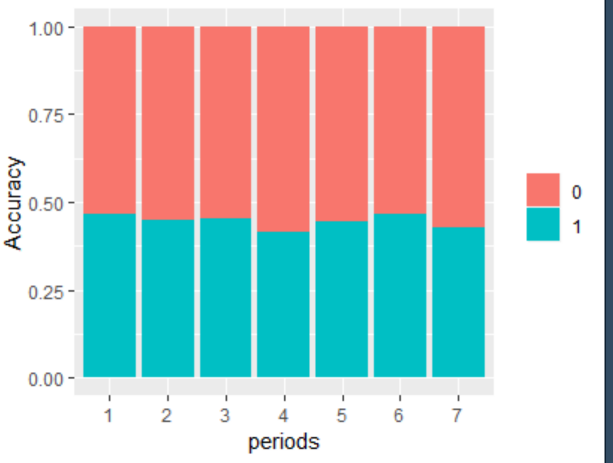
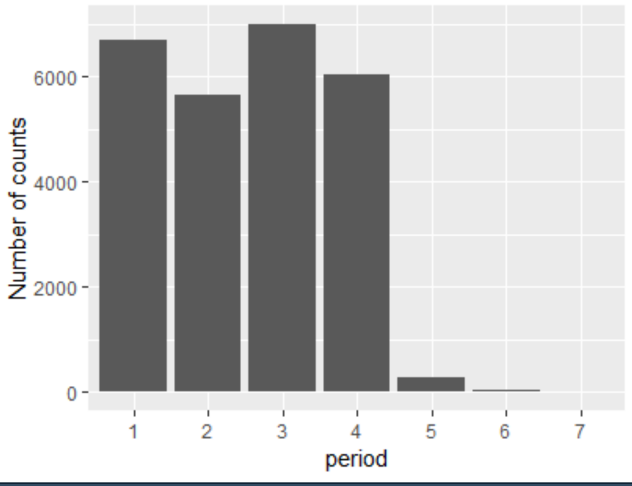
The following figure shows the analysis of Kobe’s performance when played certain opponents. There is no very substantial difference of Kobe’s performance when he played with different opponents. But he performed a little better when his opponent was team NYK



The following figure shows the analysis of Kobe’s performance in regular season and playoff. Clearly Kobe made more shots in regular season, this is normal since there are more games in regular season than in playoff. What surprising is that Kobe always maintained a steady accuracy level, this indicates that Kobe deserves his credit

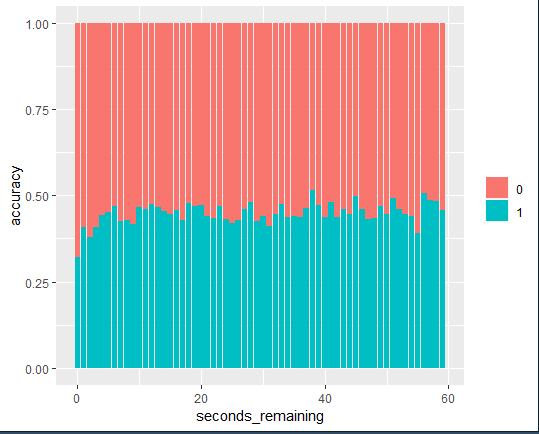
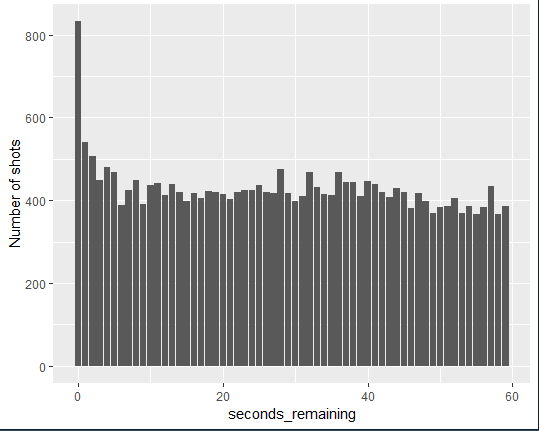


The following figure shows the analysis of Kobe’s performance in different periods. Kobe made the least number of shots in period 5, 6 and 7, this is normal since these three periods represents the overtime in NBA, which did not happen a lot only happened when there is no a winner of the two teams. But surprisingly, in these periods, Kobe always maintained a steady performance although it appears that in period 4, Kobe performed a little bit worst.



The following figure shows the relationship between seconds remaining and Kobe’s shot accuracy.

Note: minutes remaining and seconds remaining are really the same thing, so we only analyzed seconds\_remaining. As shown in the figure, Kobe always kept a steady performance with time, except for the final seconds. Also, in the final seconds, Kobe made the maximum number of shots. This is probably due to the pressure or riskier shots, which means that he wanted to make more points so he would take low probability shots that could be performed from longer range, not his specialty.



***4. Hypothesis***

Based on the plots developed above, we have come up with the following hypotheses

* Kobe’s age has no significant influence on predicting shot\_made\_flag.
* Kobe’s action type and combined action type could have a significant influence on predicting shot\_made\_flag.
* Shot distance has a significant influence on predicting shot\_made\_flag.
* Shot zone area, shot zone basic and shot range in a basketball court could have a significant influence on predicting shot\_made\_flag.
* Game season, month, opponent, playoffs or not and periods have no significant influence on predicting shot\_made\_flag.
* Although Kobe did not perform very well in the final seconds, it does not seem that Kobe’s performance varies enough for time to be a significant variable.

**B. Modeling:**

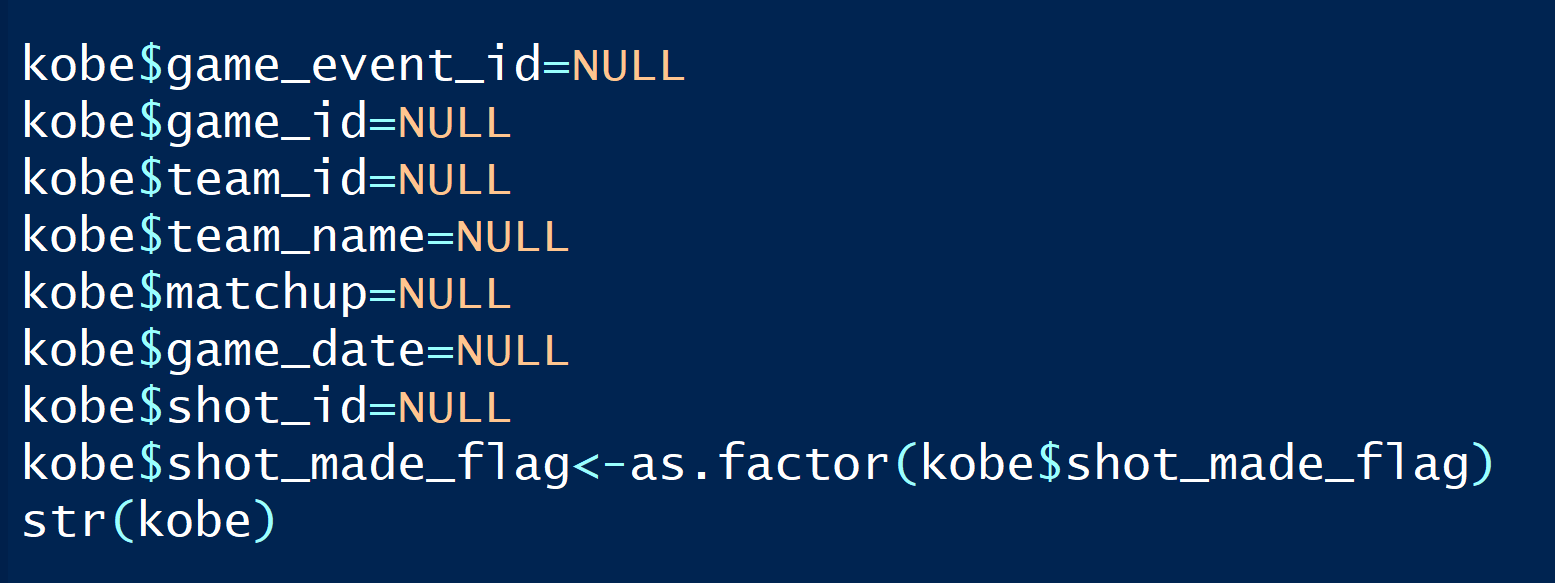
1. ***Models Used***

We chose to use Ridge logistic regression model and Lasso logistic regression because there are many variables in the dataset the for some categorical variables, there are many categories in them. Also, we could do the cross validation test in these two models.

1. ***Variables chosen***

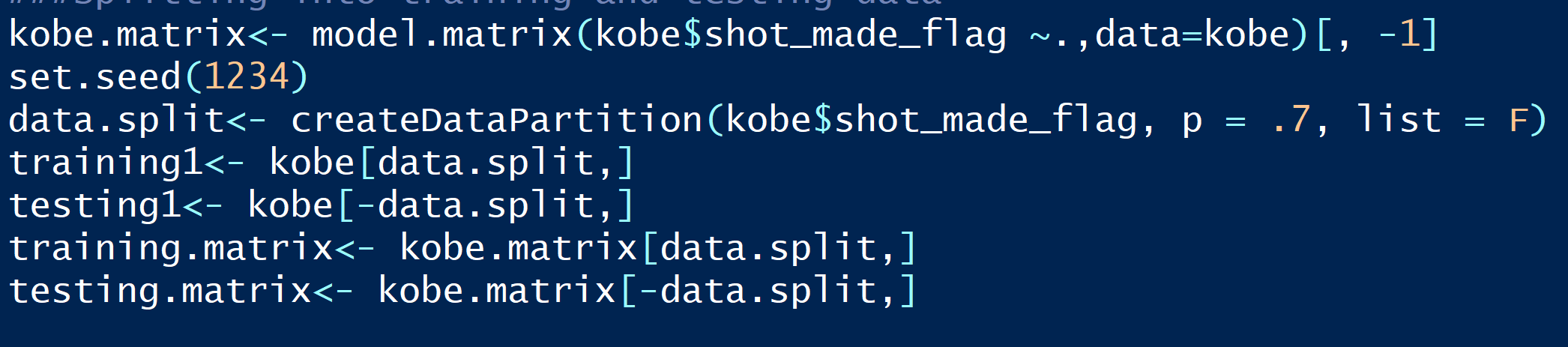
First, we extract month from column 'game\_month' generating a new column called 'game\_month', we also add Kobe's age data of each season.

Then, we changed game\_event\_id, game\_id, team\_id, shot\_id, team\_name, matchup, and game\_date column into NULL, because we do not need these columns. Change the data type of shot\_made\_flag into factor since we should do a classification problem. Just as the following figure shows.



For the rest variables, we did not select them by hand because we wanted the ridge and lasso models to help us select important model.

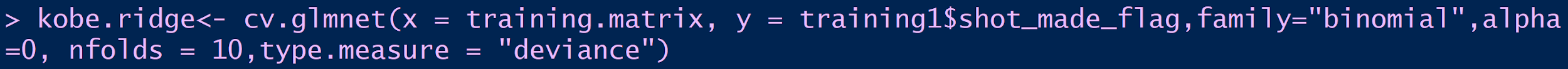
Also, we split the data into training and testing data by 70:30 just as the following figure shows.



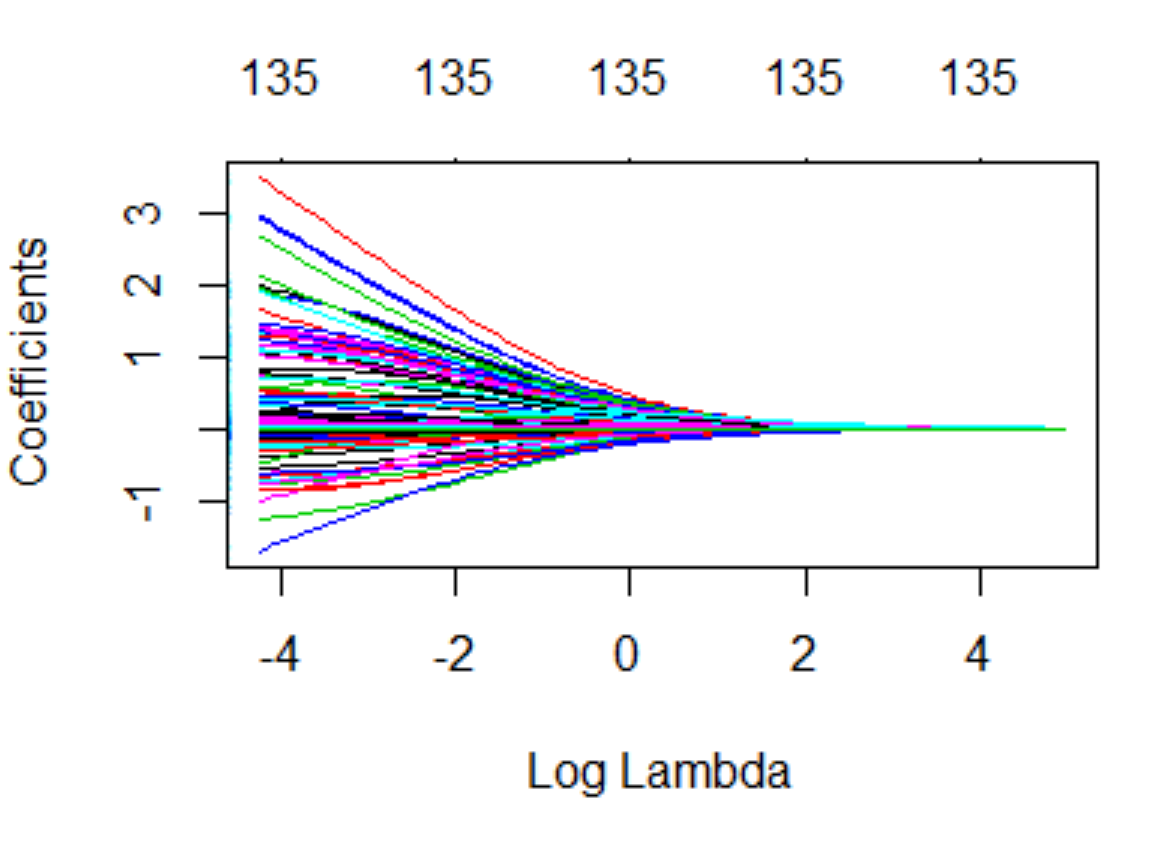
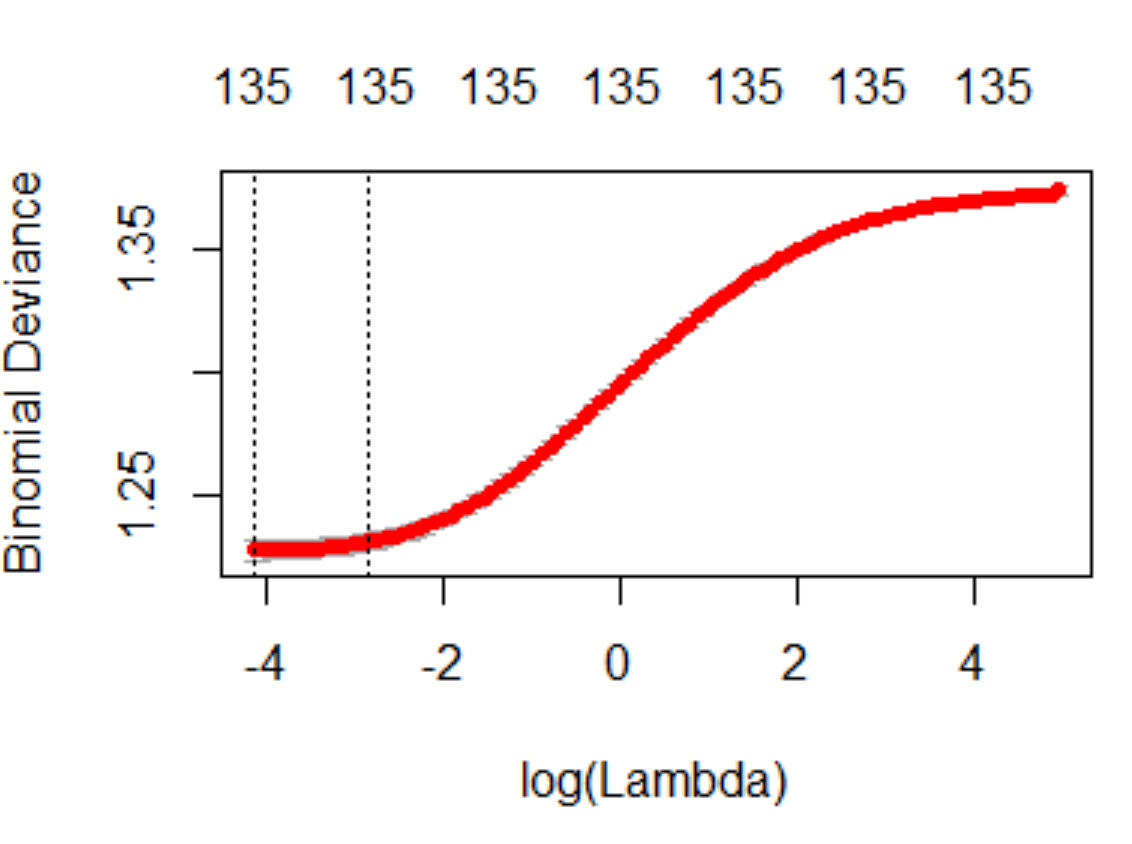
1. ***Model output***

*Ridge logistic regression model*

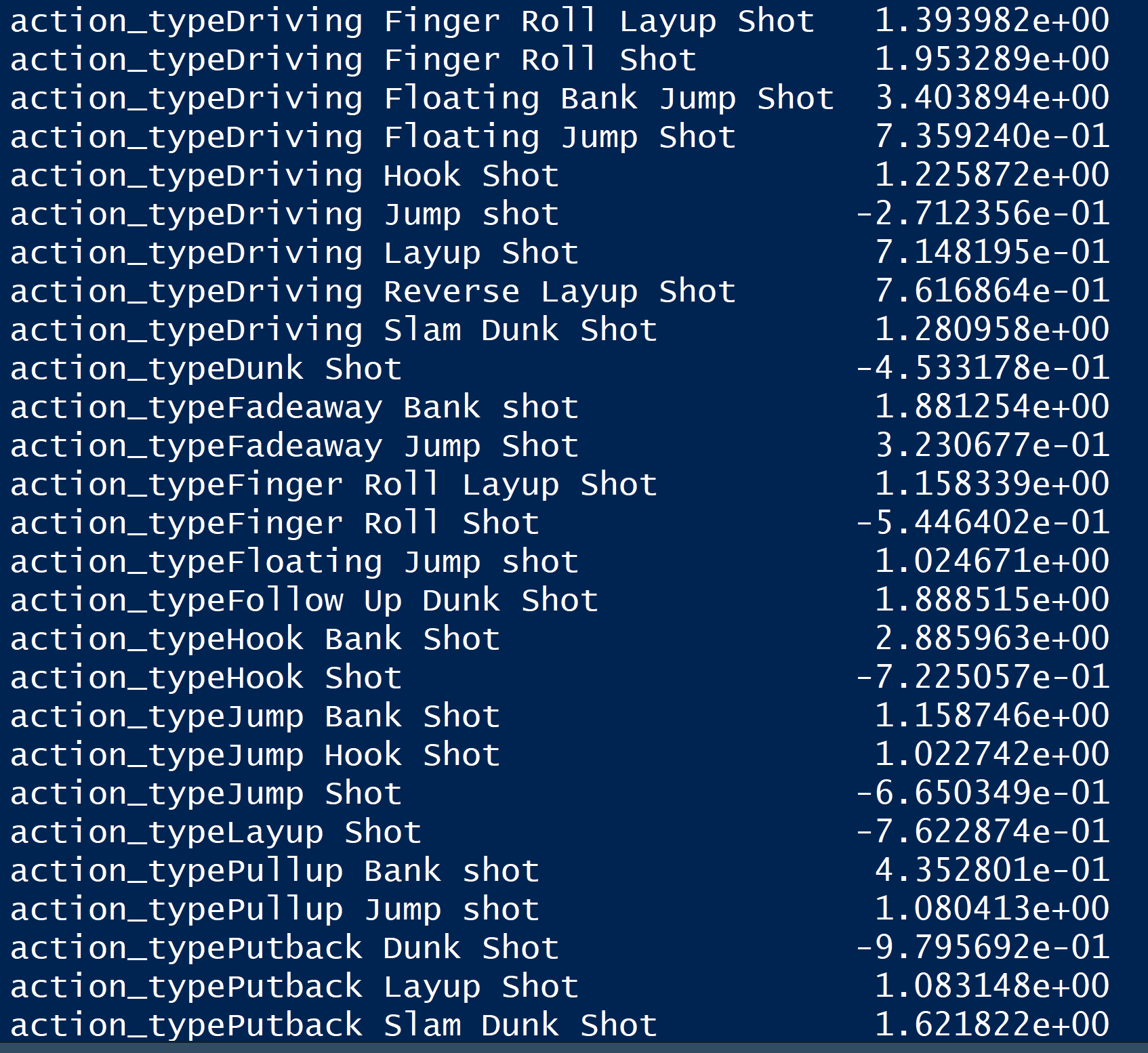
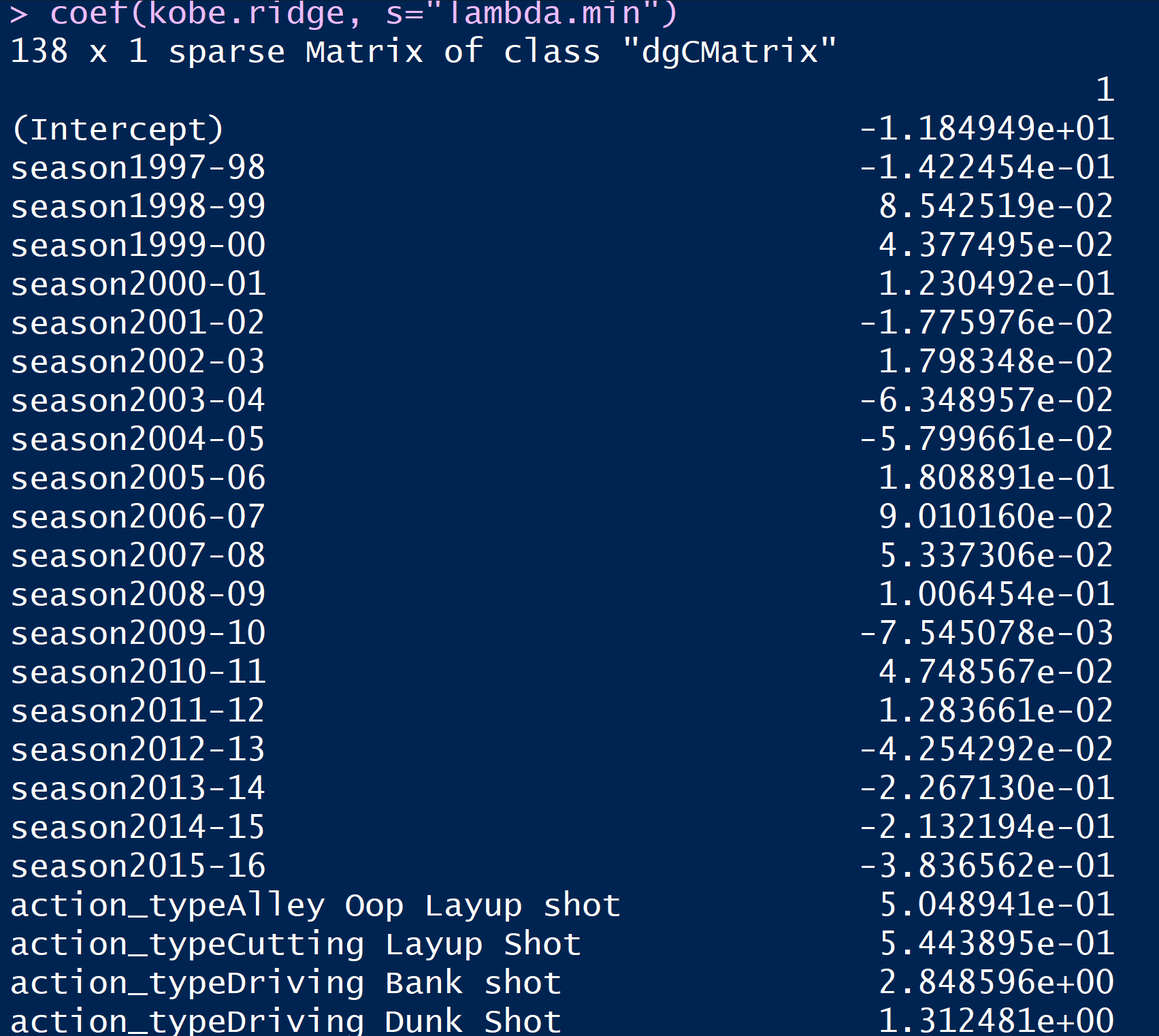
As the following figure shows, we built the ridge logistic regression model. We select 10 folds for the cross-validation test, also we set ‘deviance’ type measure for the logistic regression.

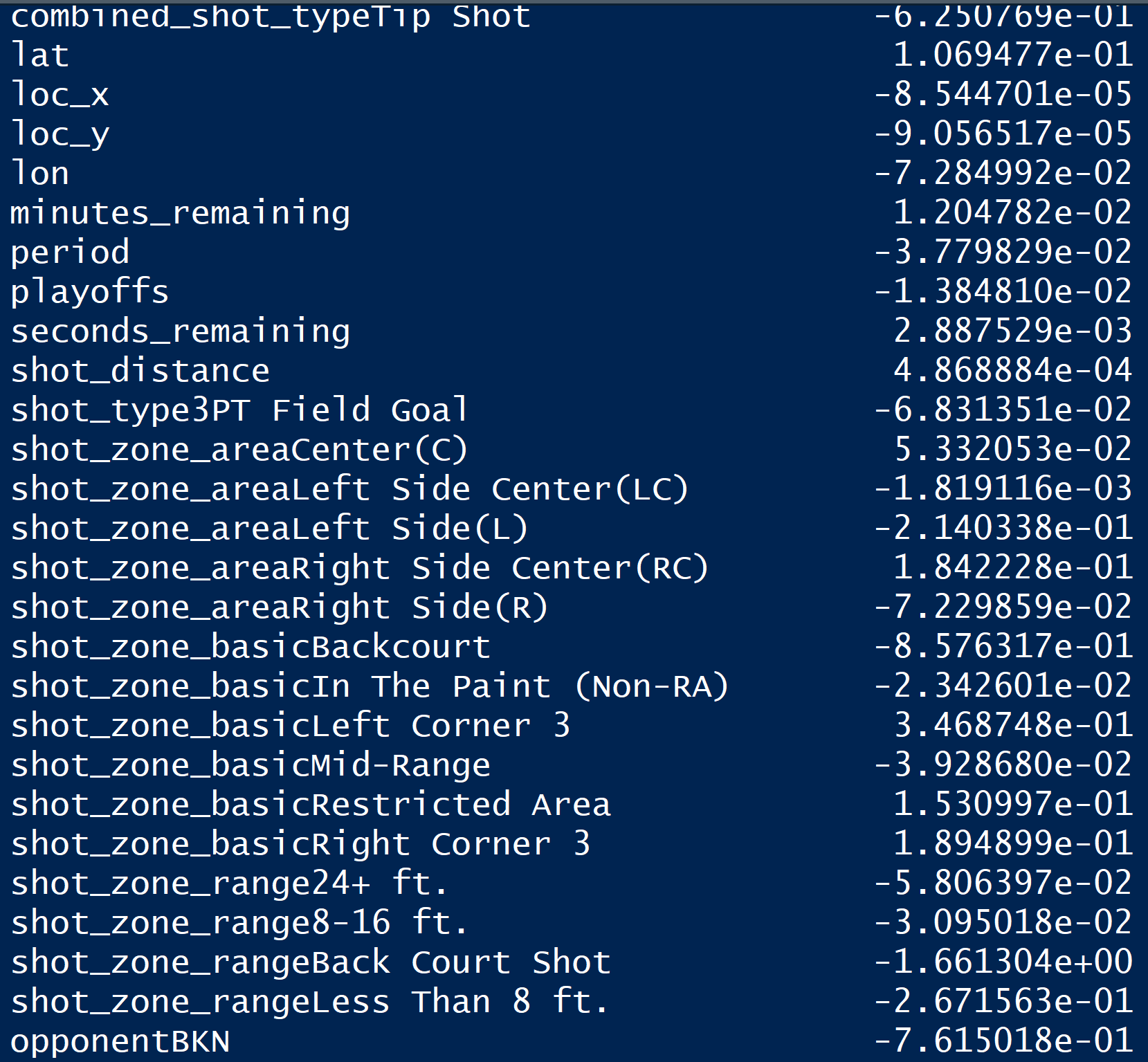
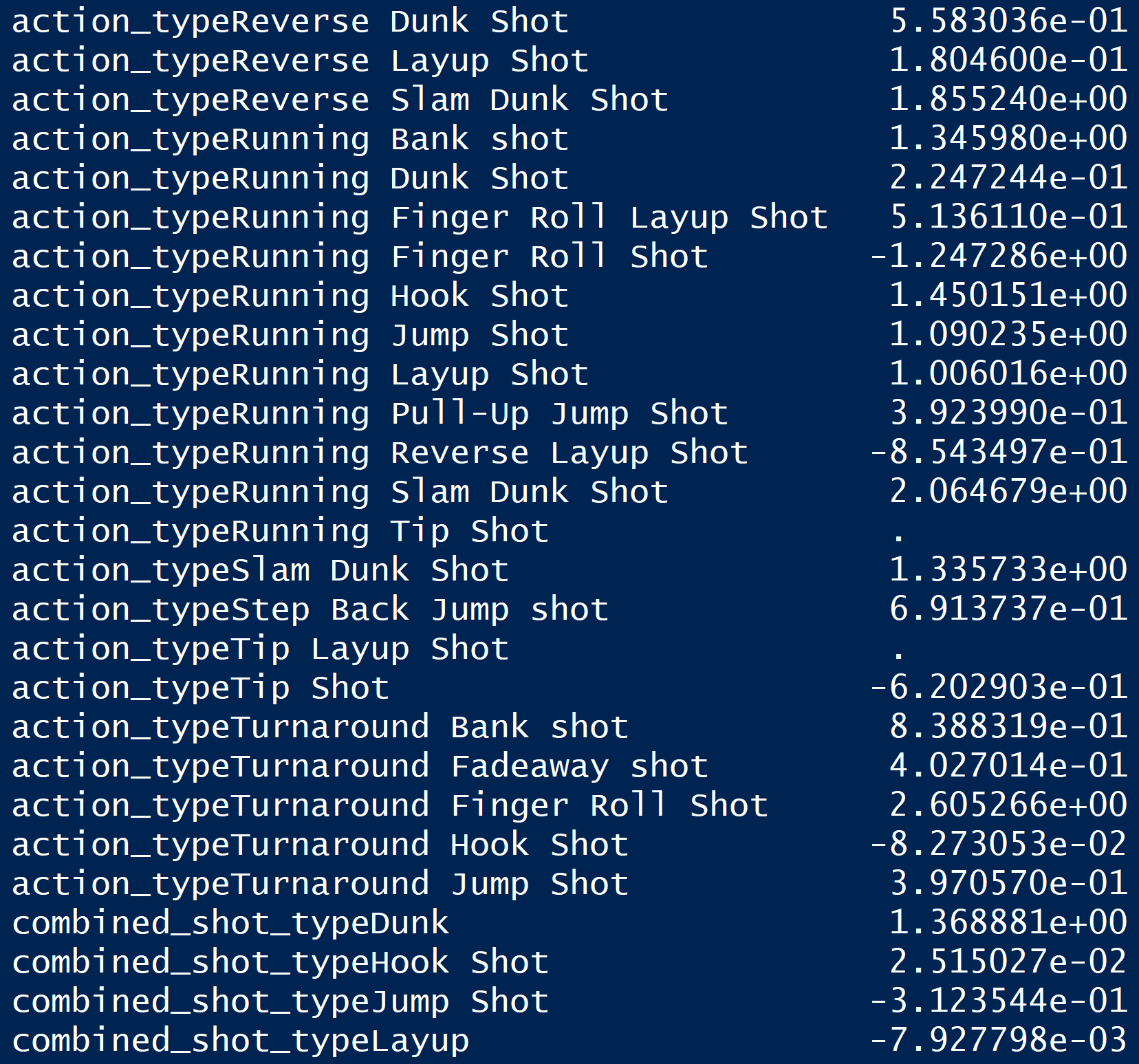


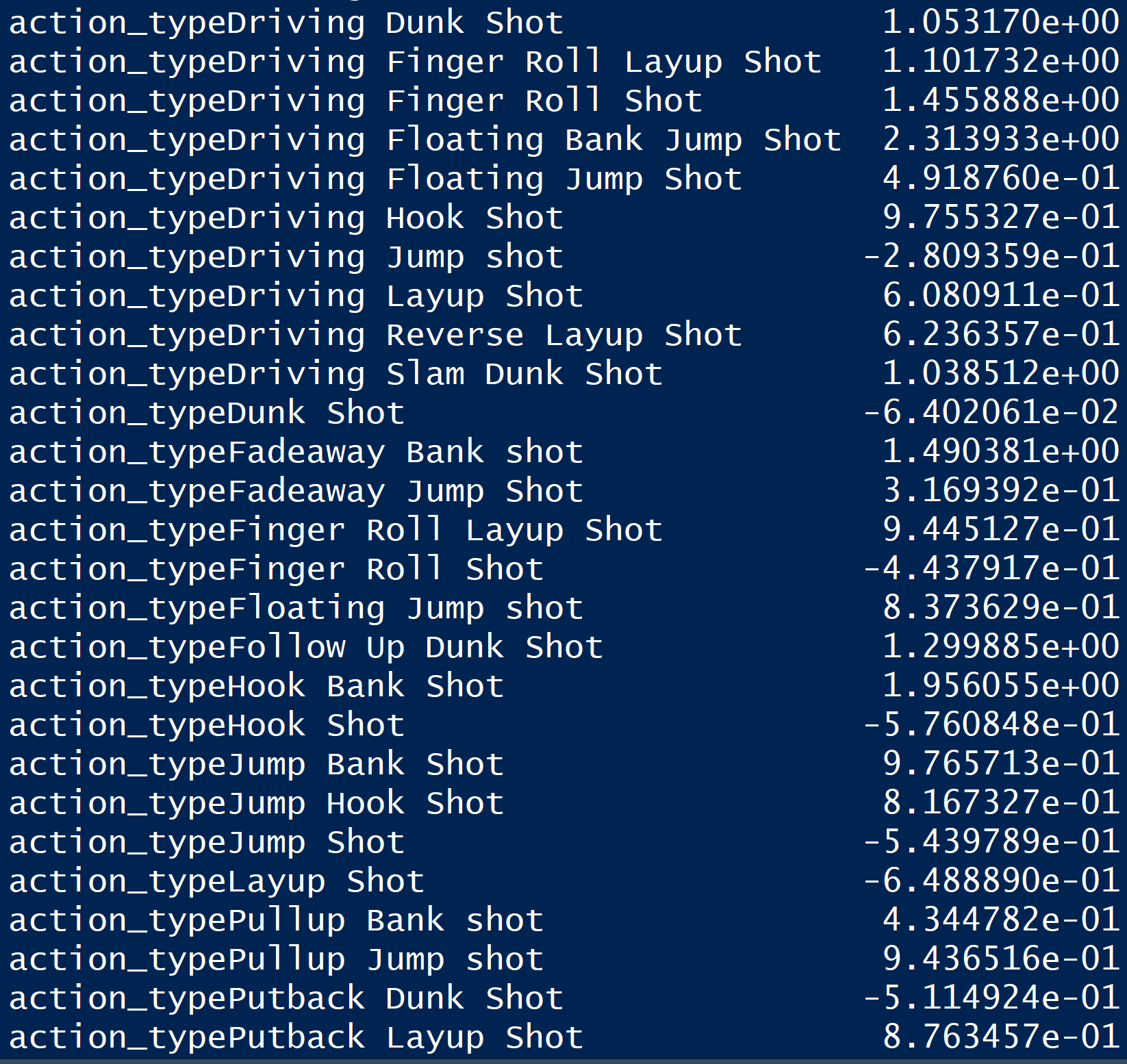
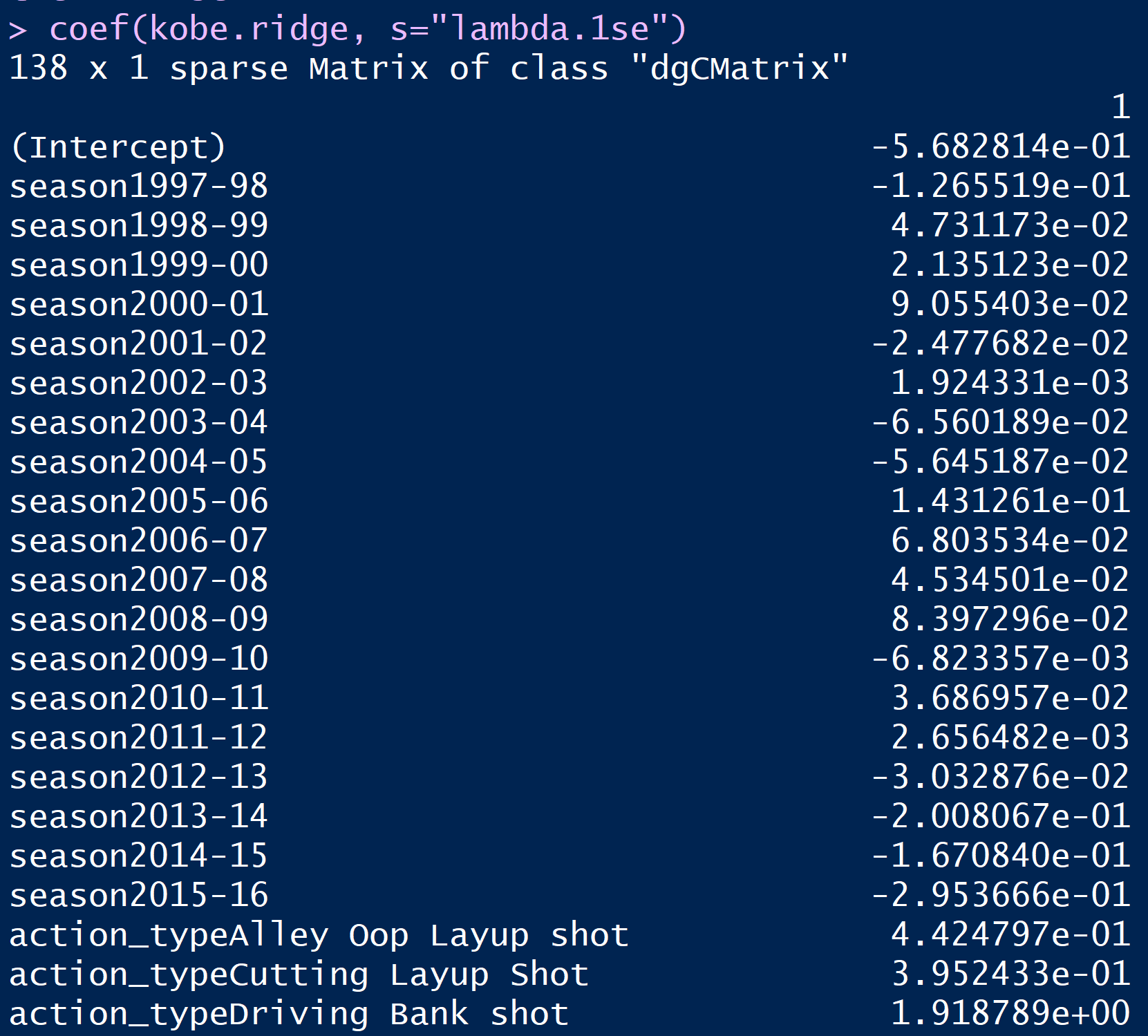
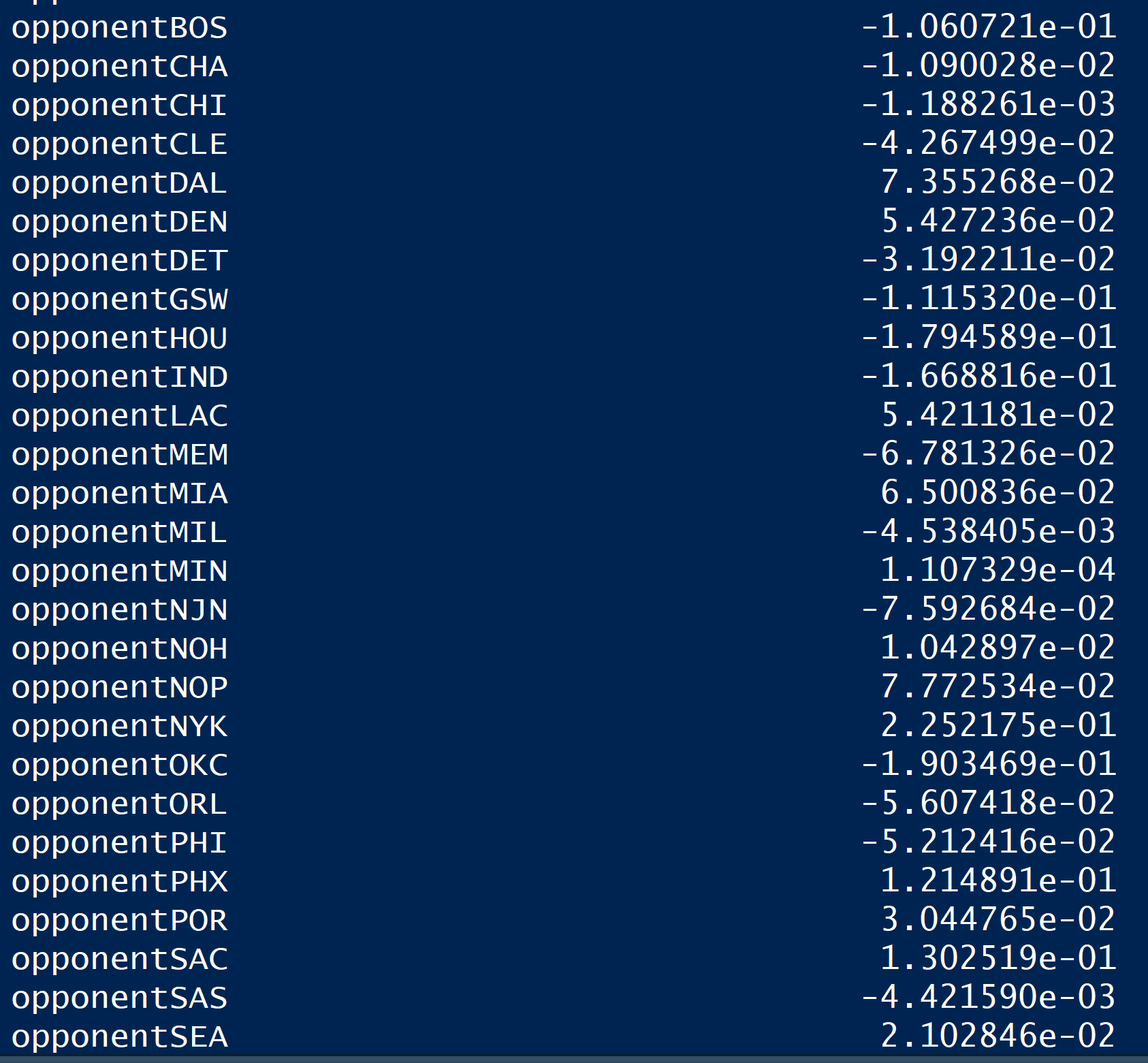
The following figure shows the value of lambda that could lead to smallest mse and the way how the coefficients shrink over different values of lambda.

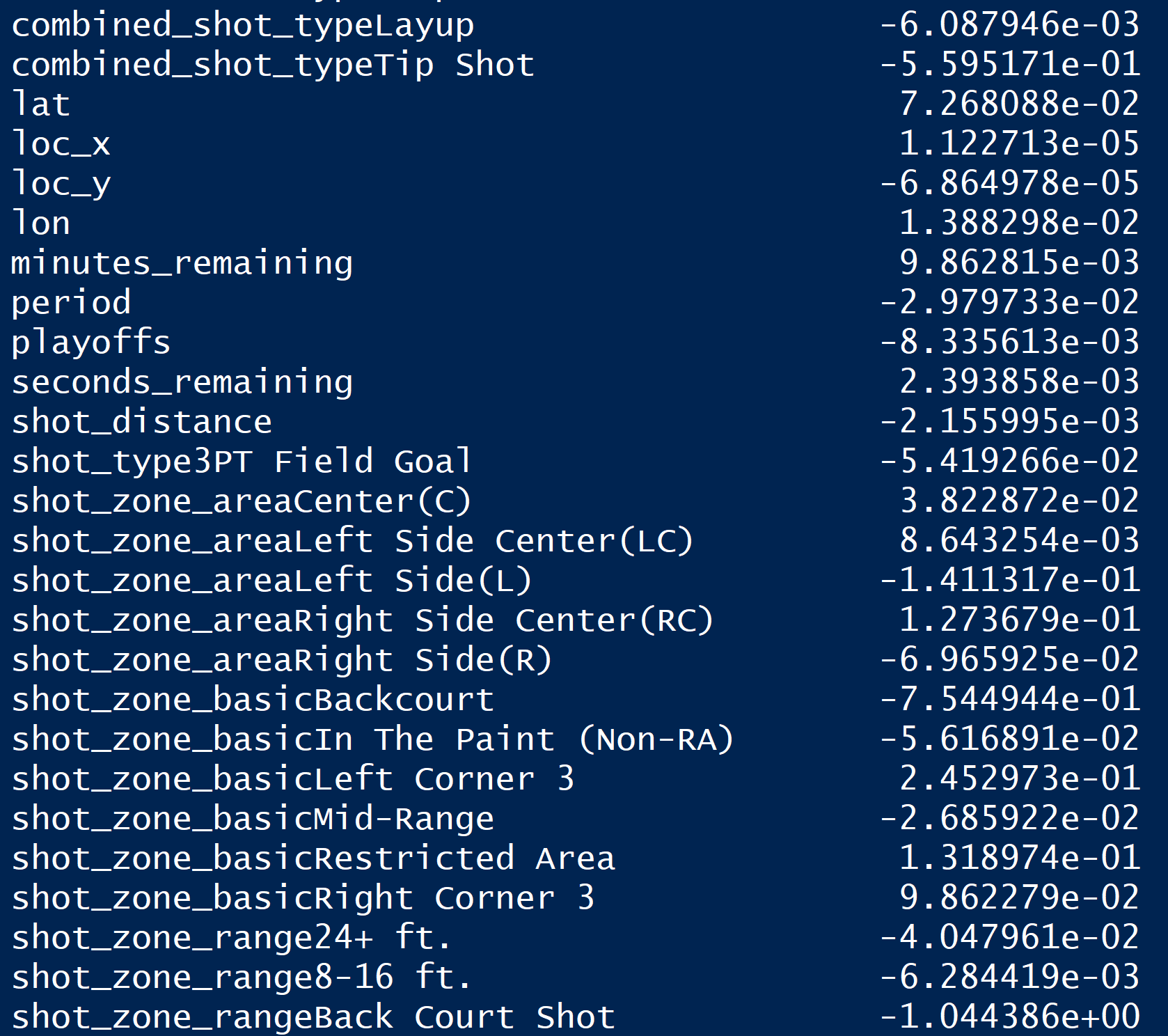
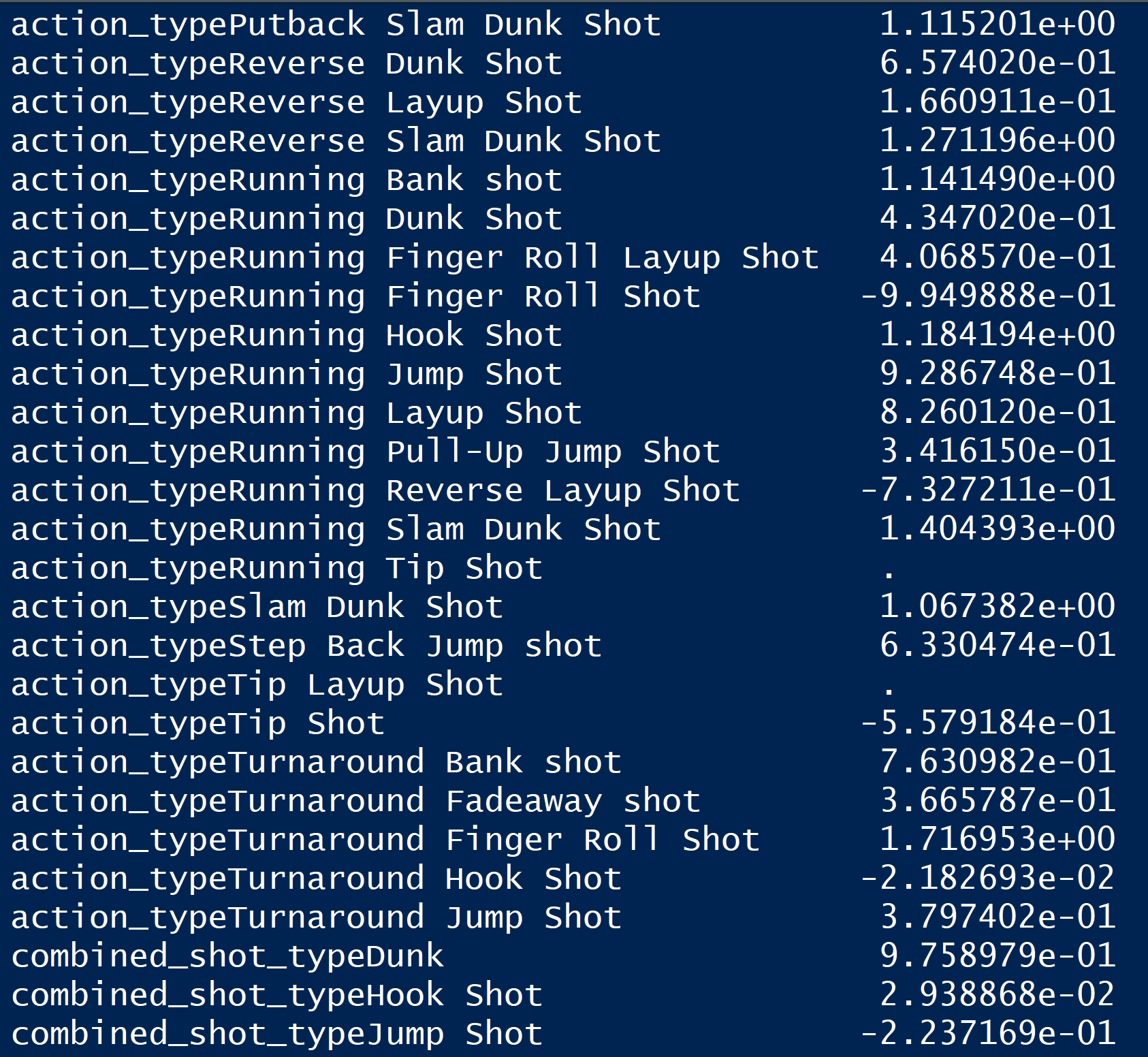


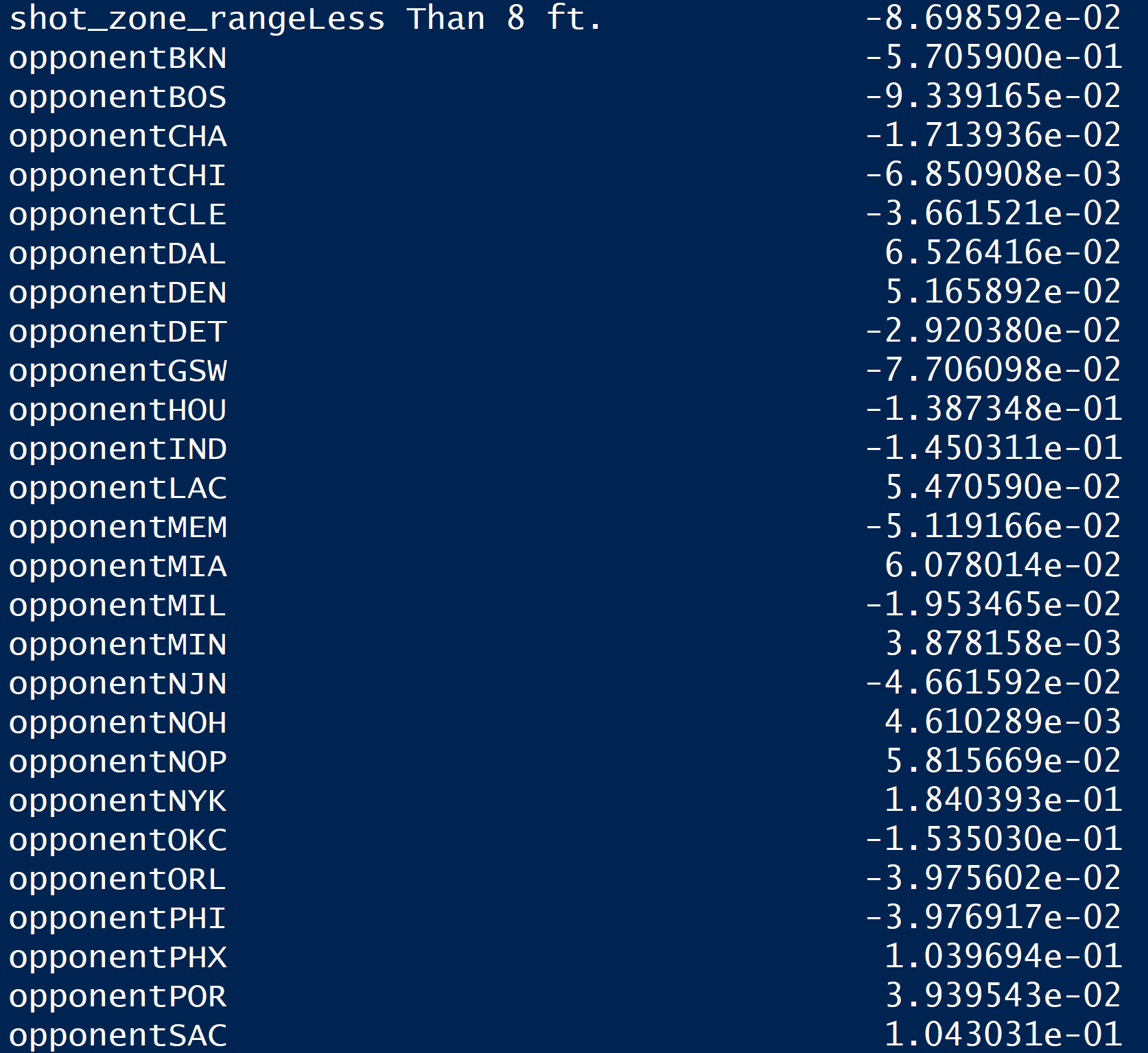
The following figure shows coefficients of the lambda that produces the smallest mse and coefficients of the one that is the most parsimonious. The coefficients mean that for each additional point of one independent variable, log odds increases or decreases by some number with all other independent variables held constant. And for both these two kinds of coefficients, ridge model deletes two features-action\_type Running Tip Shot and action\_type Tip Layup Shot and kept all the other variables as significant variable.





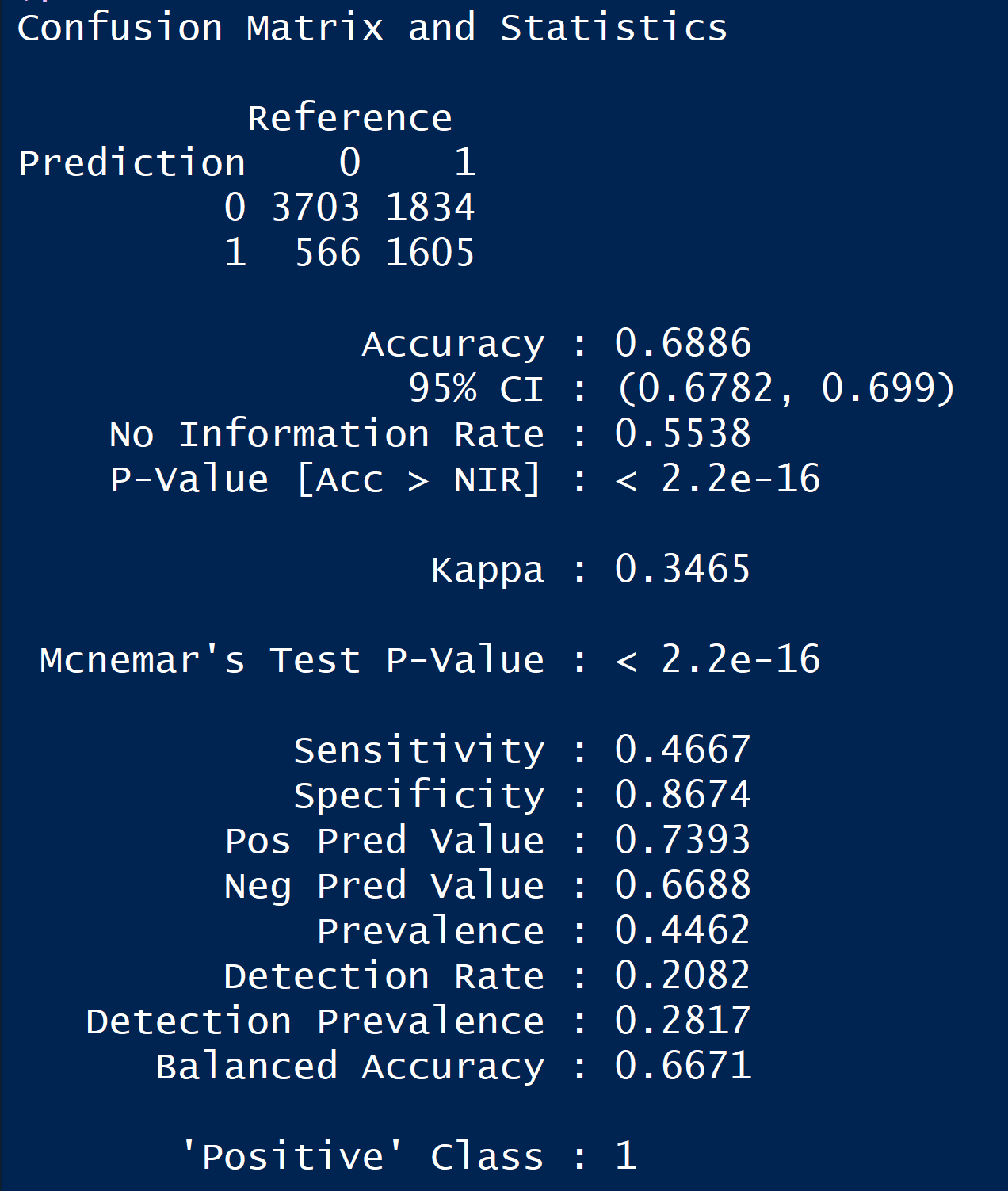




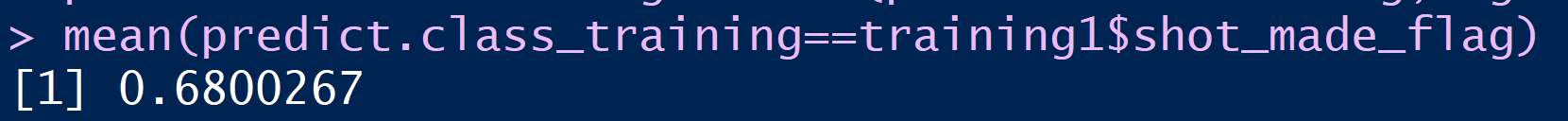


Then we checked the model’s performance on testing dataset.

The following figure shows the confusion matrix. As it shows, the accuracy of testing dataset is 0.6886, which means that our model could predict 68.86% of testing dataset correct, the sensitivity is 0.4667, which is not very good and the specificity is 0.8673.



We also checked the overfitting problem. As the following shows, the accuracy of training dataset is 0.68, which does not differ with the accuracy of testing dataset substantially. This indicates that there is no serious overfitting problem in the ridge logistic regression model.

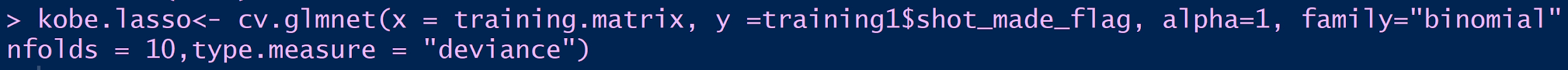


Also, we checked the assumptions of this ridge logistic regression model.

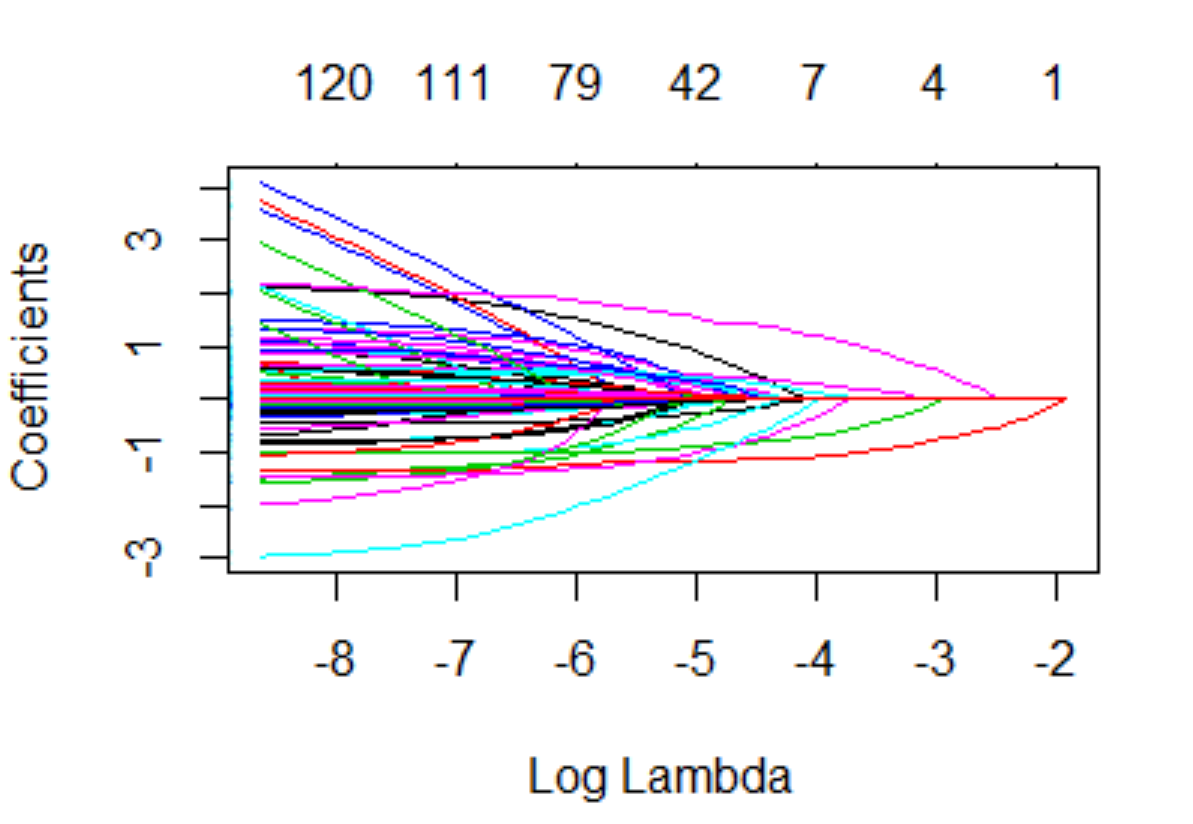
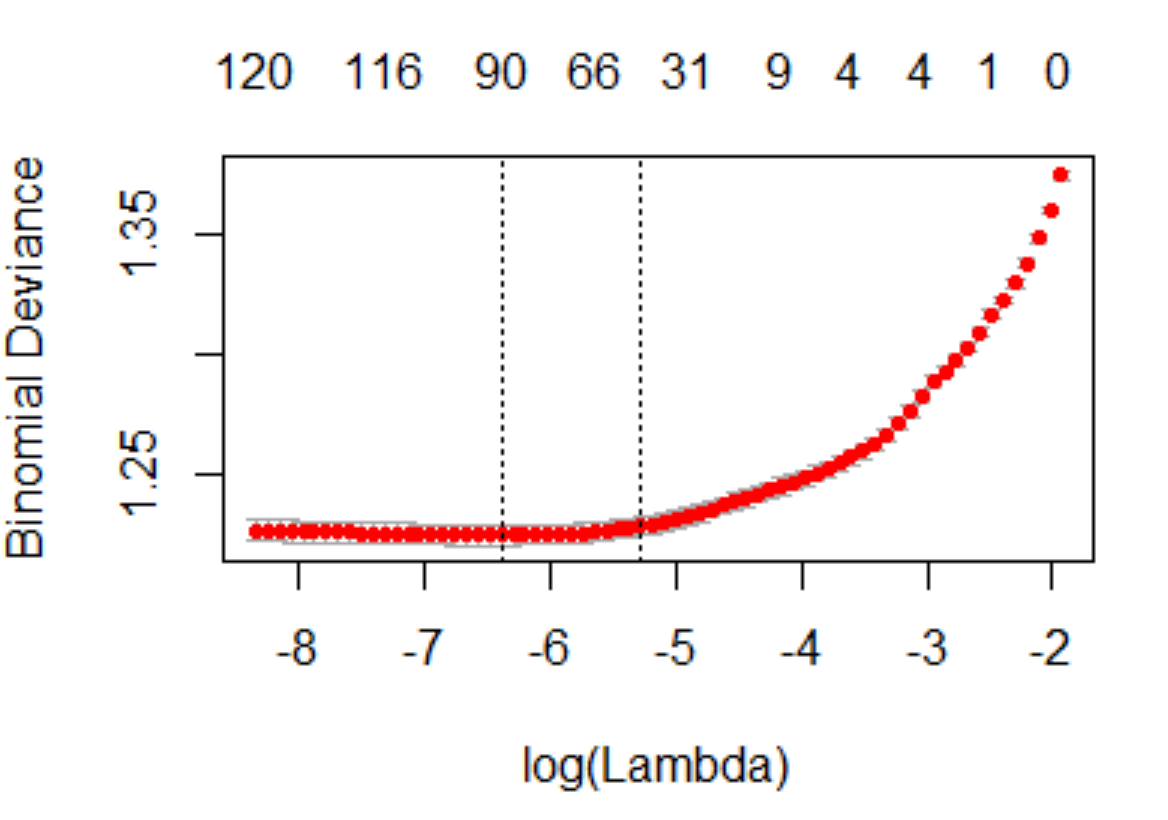
* Response in binary: for the dependent variable ‘shot\_made\_flag’, there are only two classes-‘0’ and ‘1’. So, the model fits this assumption.
* Independent observations: the records of Kobe dataset were gathered independently, so the model fits this assumption.
* Large sample size: since there are 25697 records in the dataset, which is greater than 30, the model fits this assumption.

*Lasso logistic regression model*

As the following figure shows, we built the lasso logistic regression model. We selected 10 folds for the cross-validation test, also we set ‘deviance’ type measure for the logistic regression.

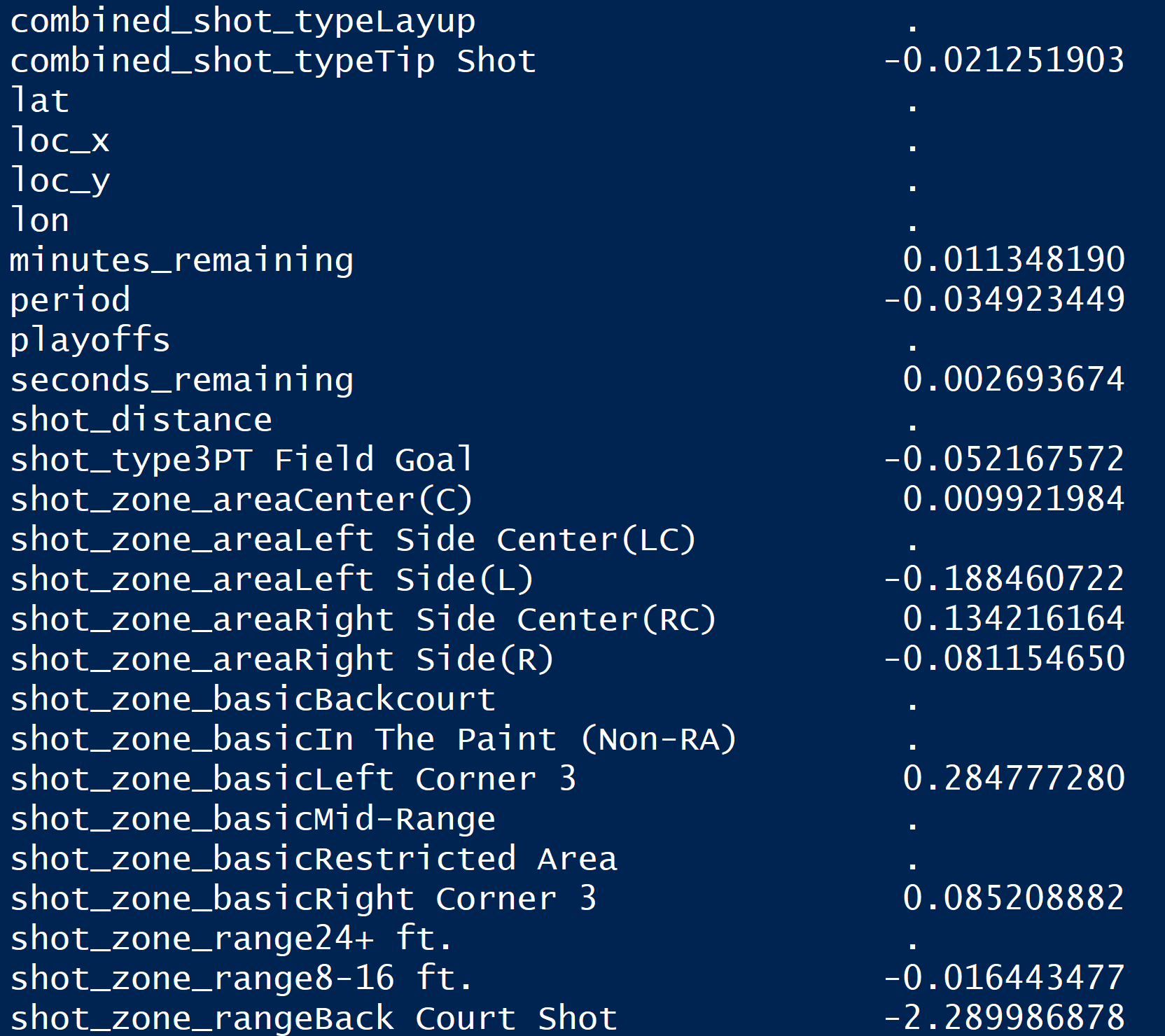
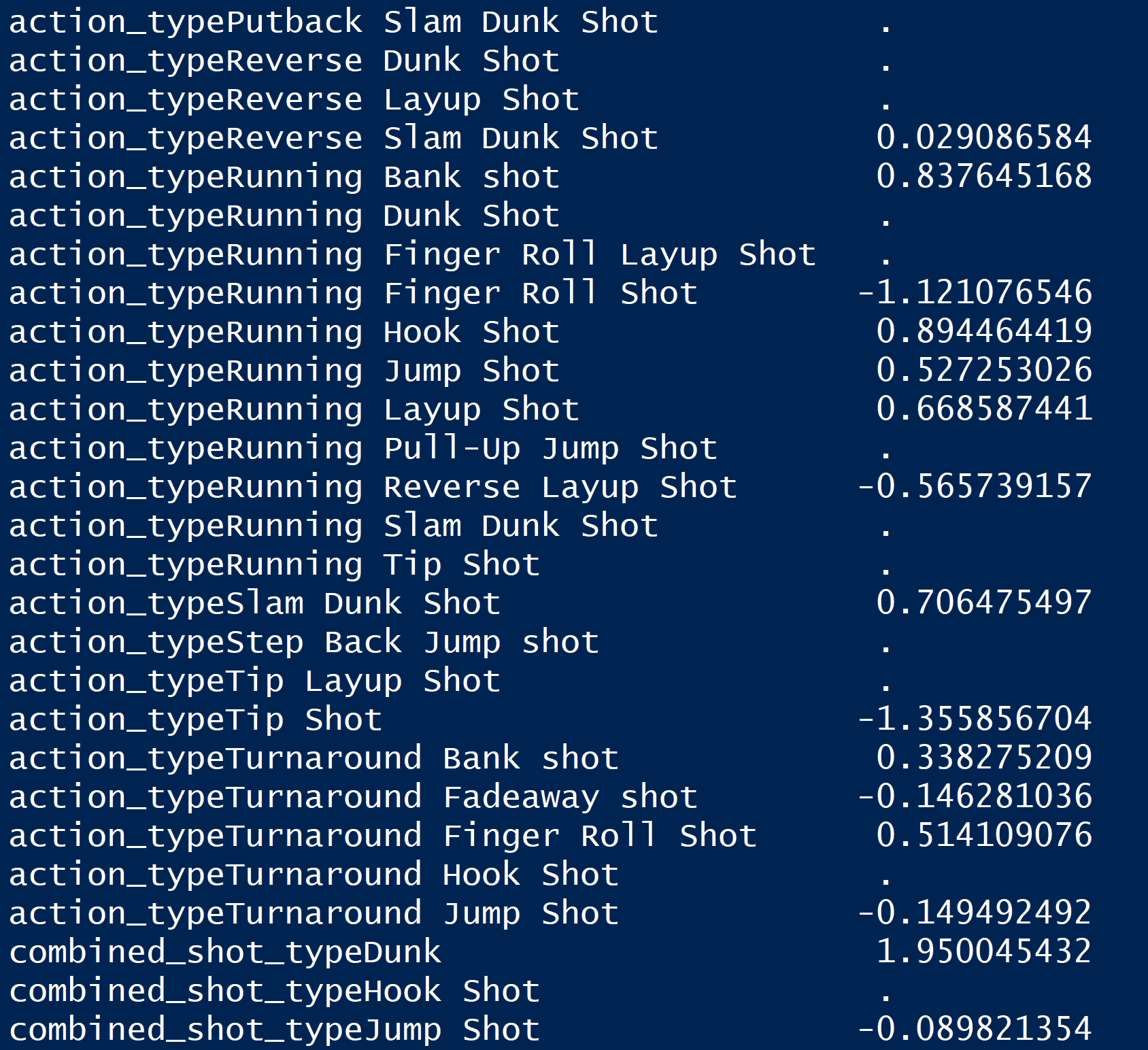
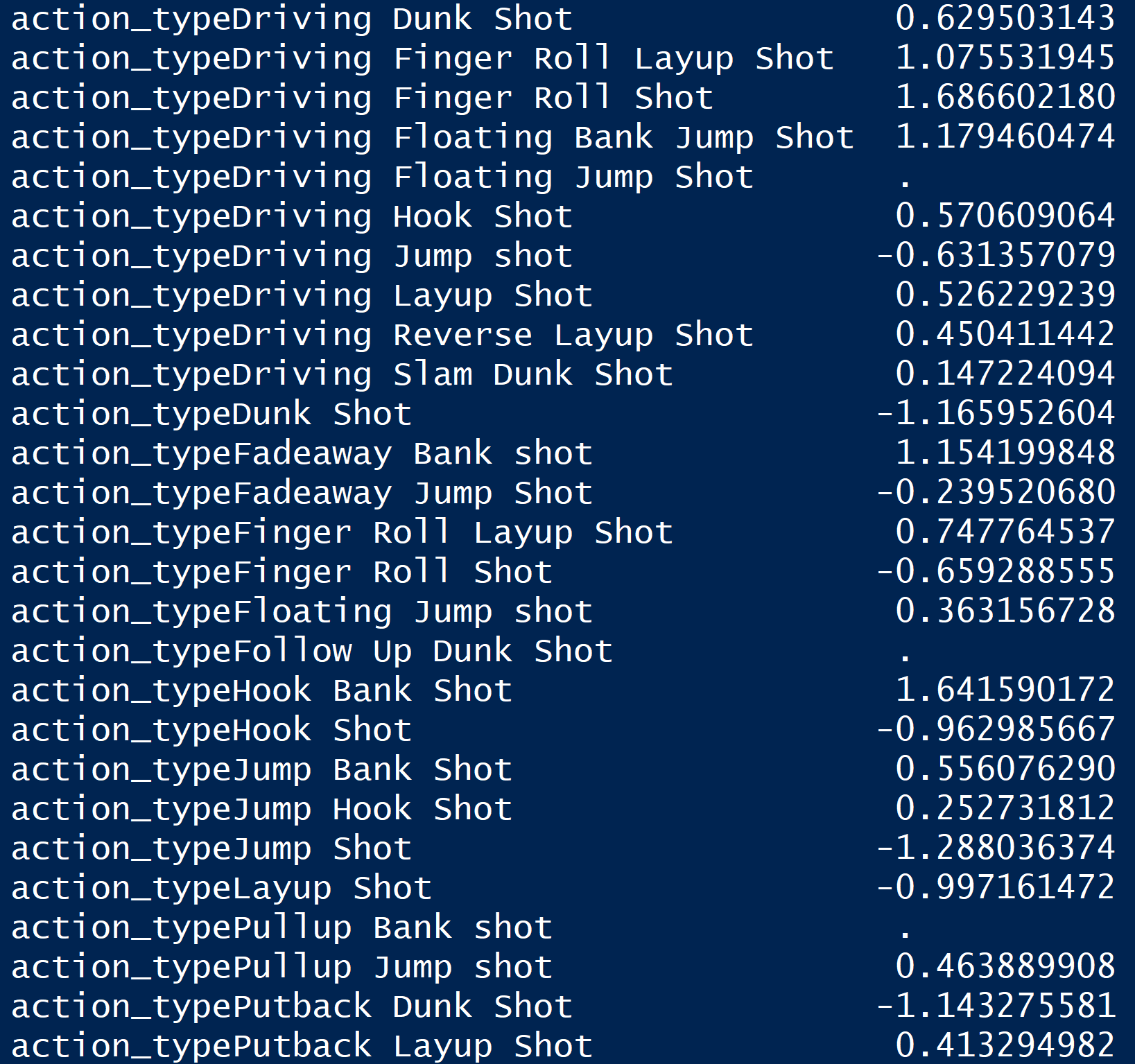
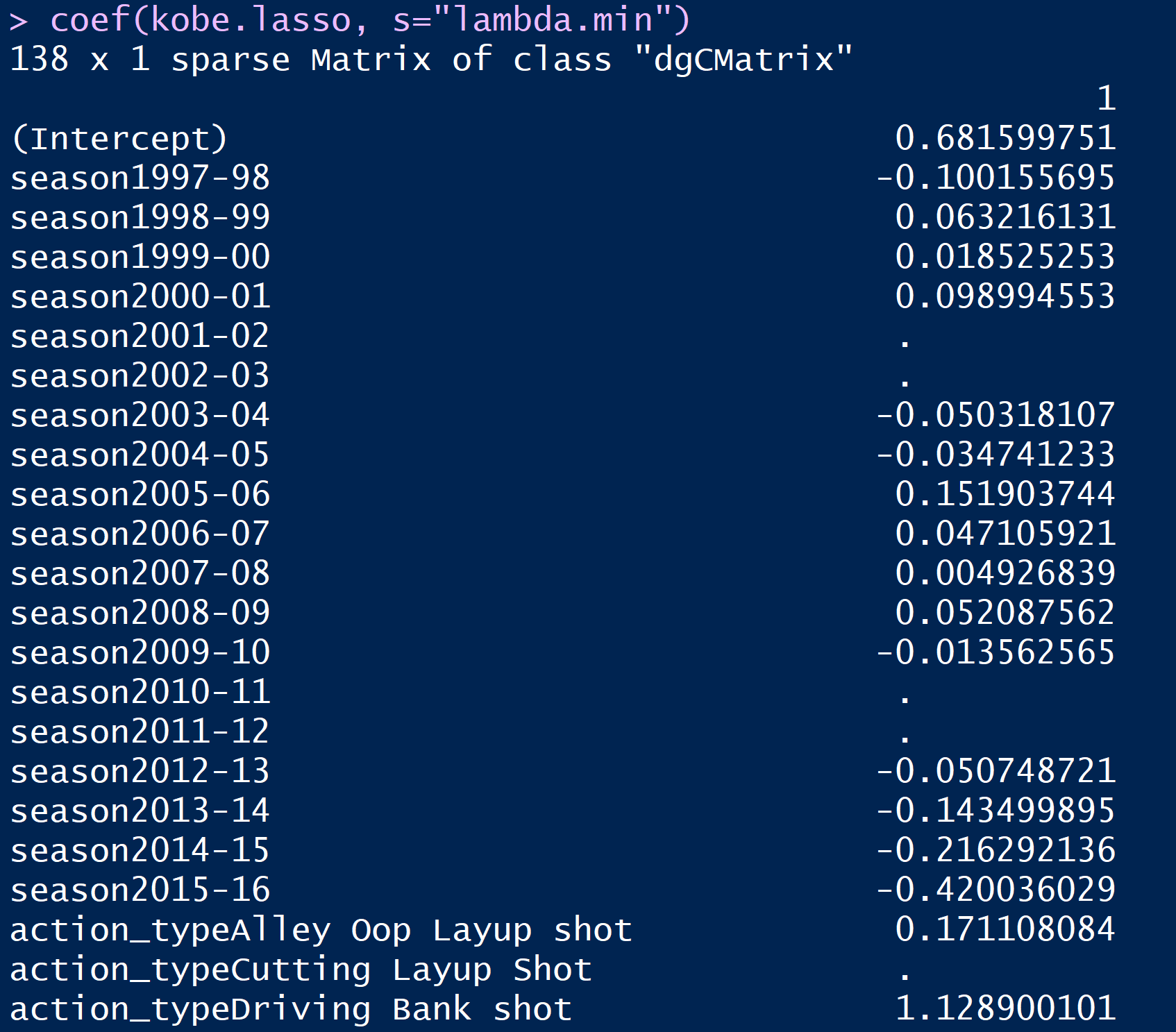


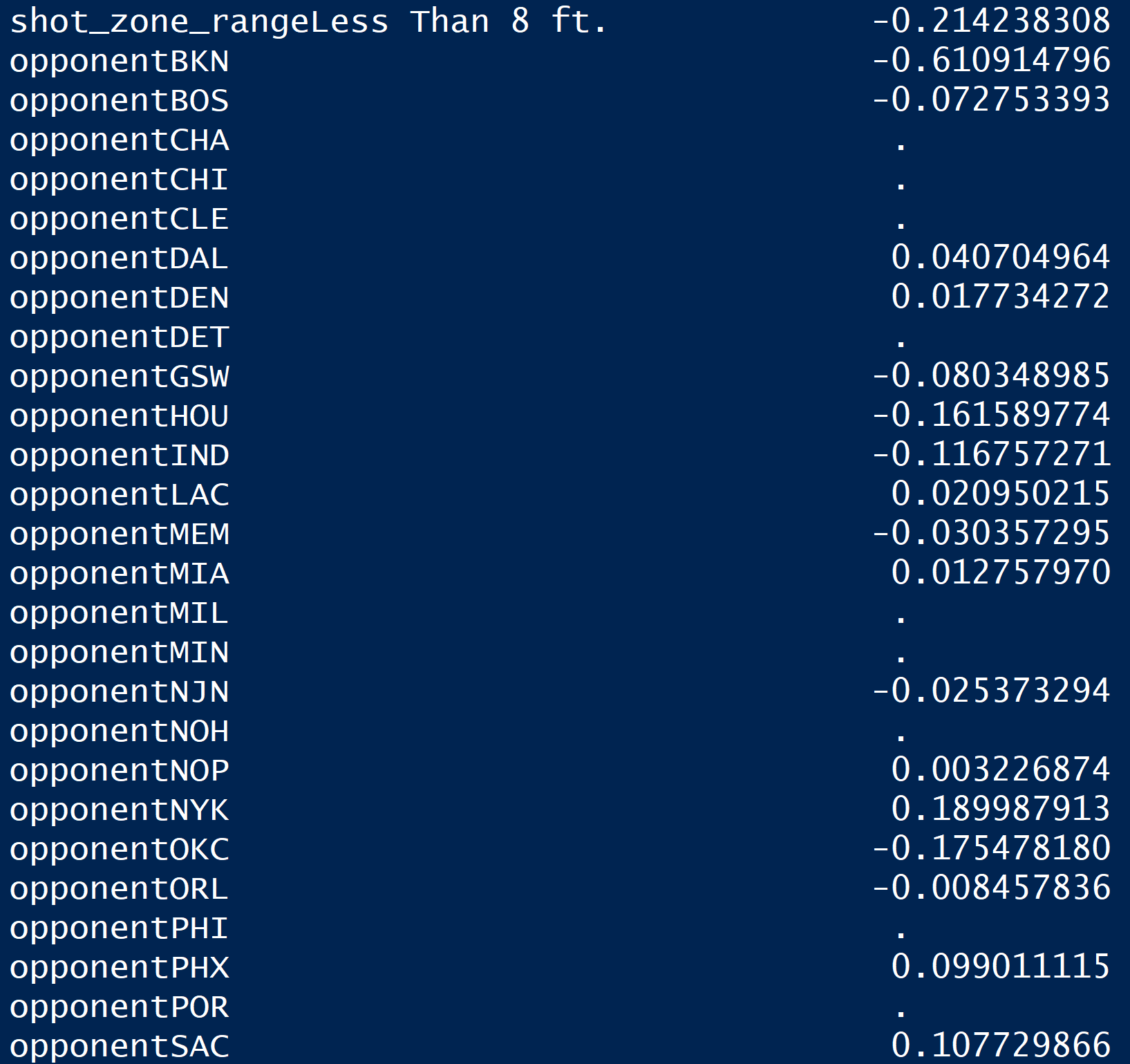
The following figure shows the value of lambda that could lead to smallest mse and the way how the coefficients shrink over different values of lambda.



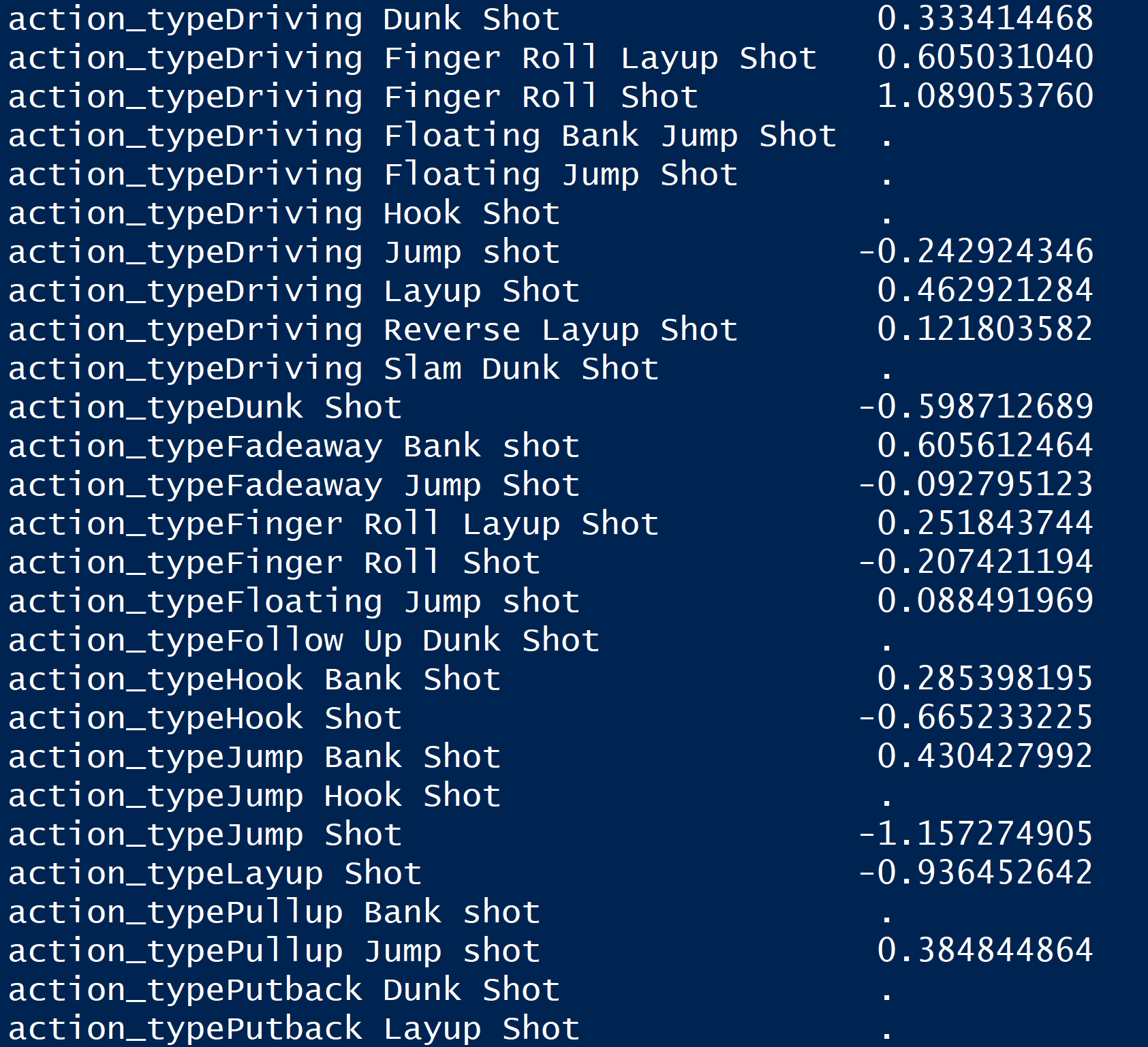
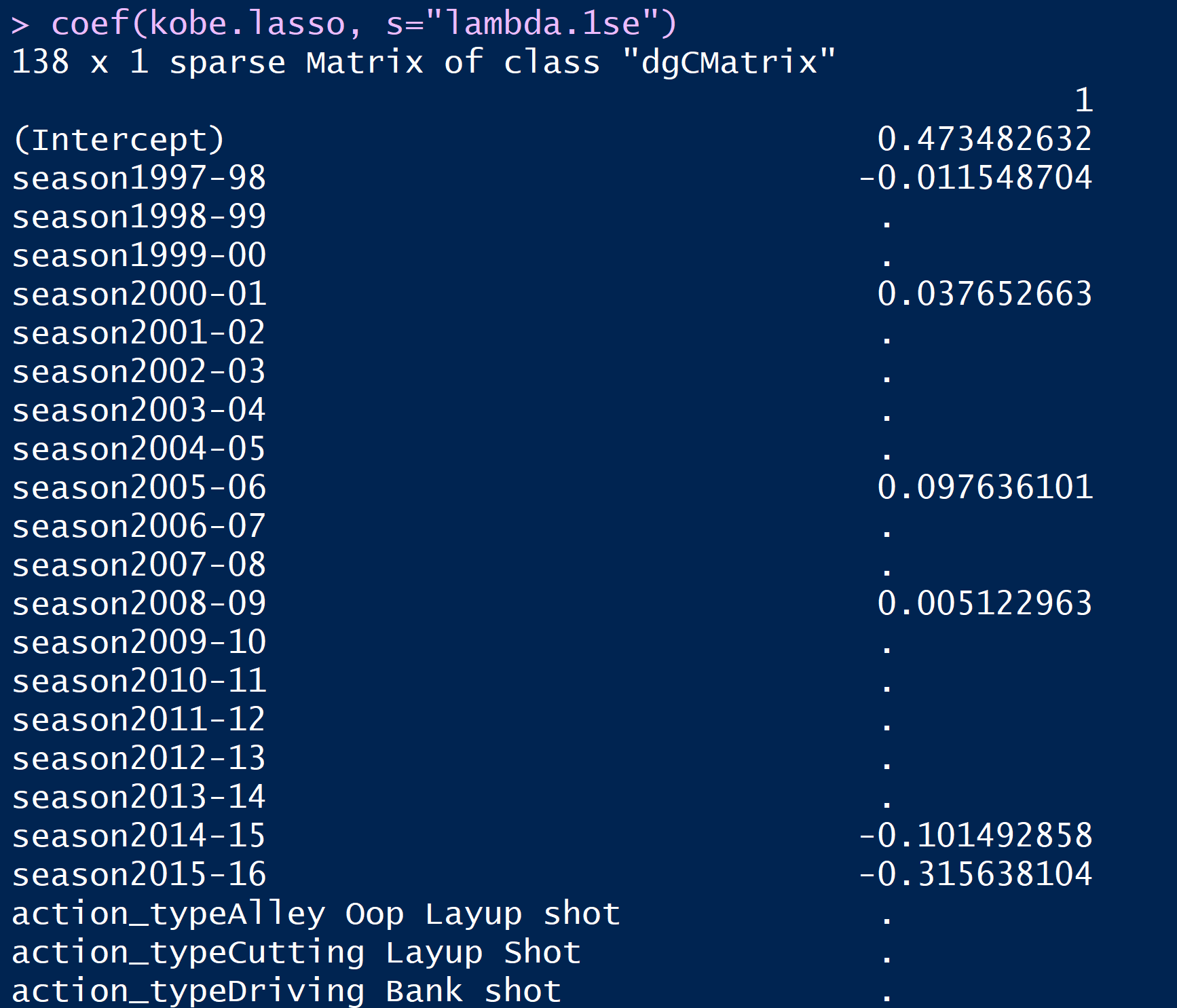
The following figure shows coefficients of the lambda that produces the smallest mse and coefficients of the one that is the most parsimonious. The coefficients mean that for each additional point of one independent variable, log odds increases or decreases by some number with all other independent variables held constant and.

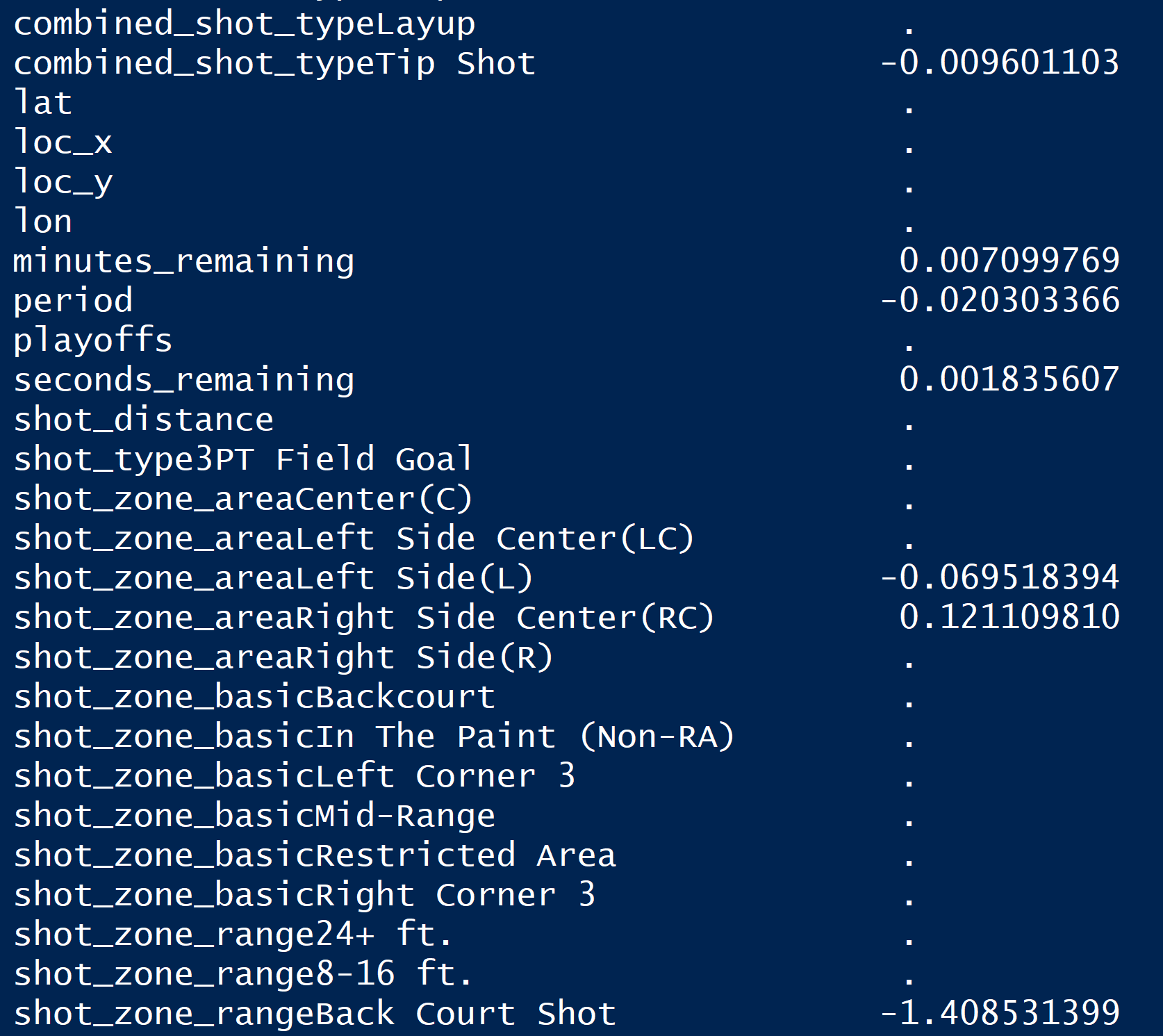
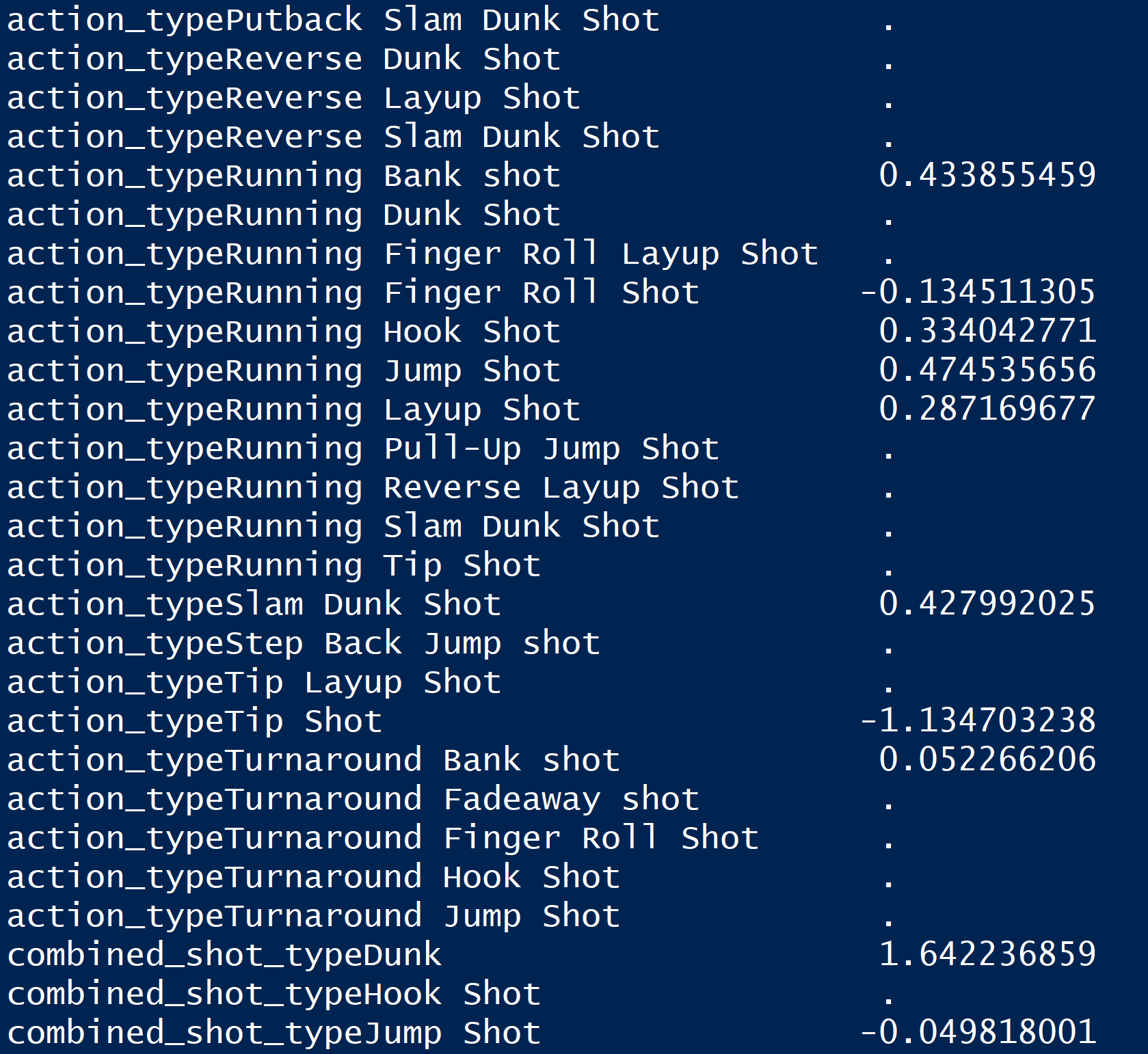
For coefficients of the lambda that produces the smallest mse, lasso model deleted 47 features and kept all the other features.

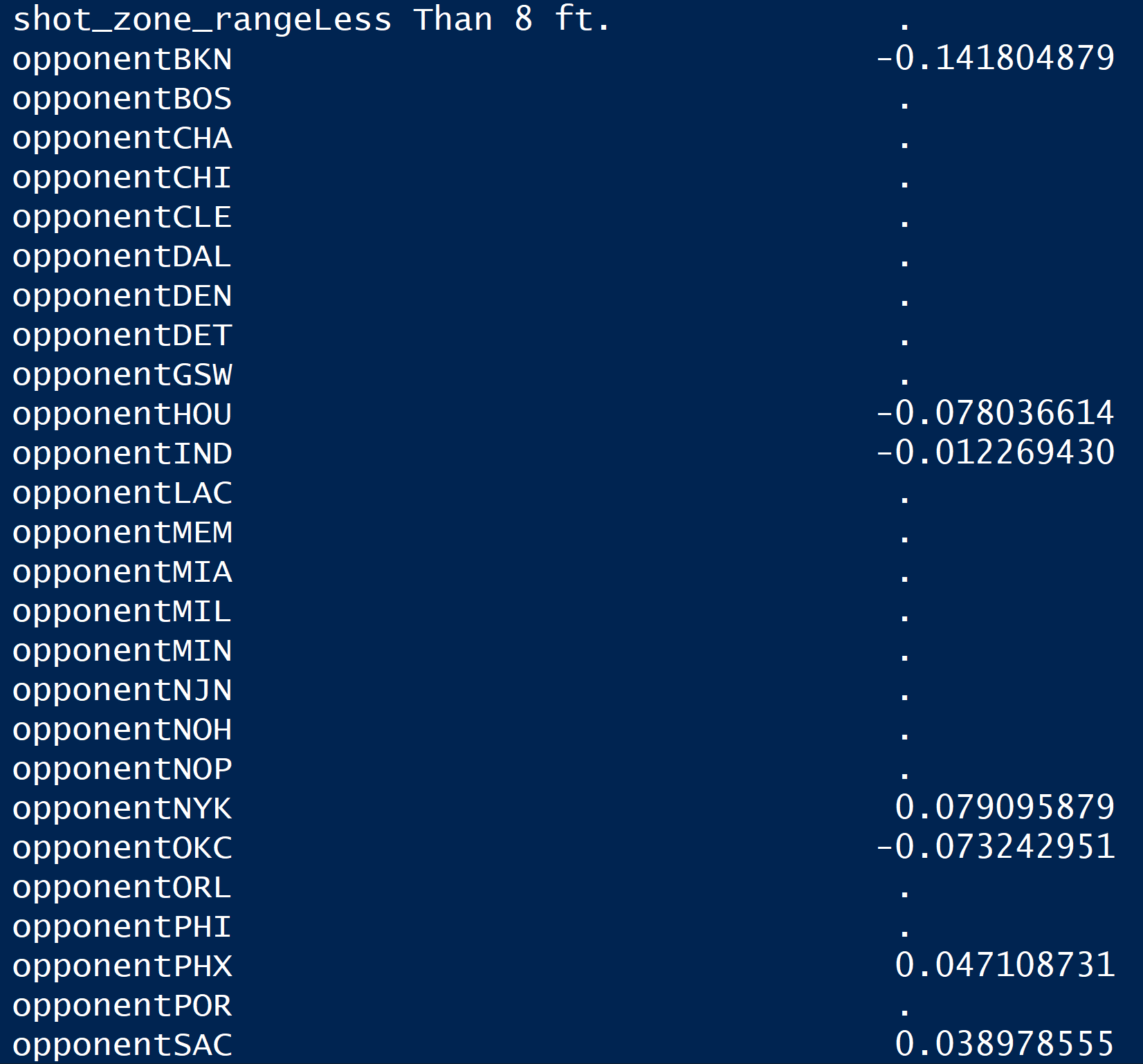




For coefficients of the one that is the most parsimonious, lasso model kept only 49 features.

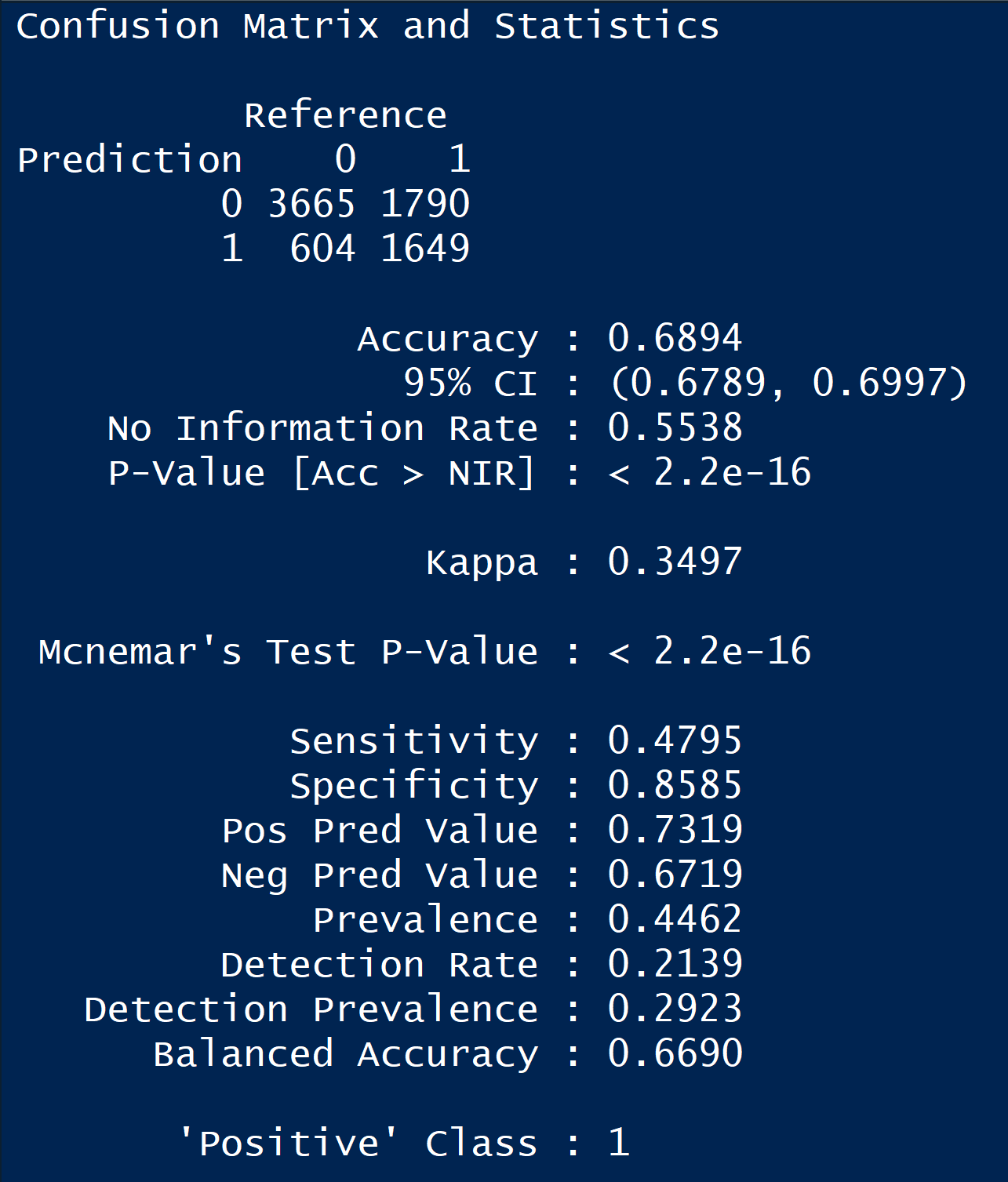




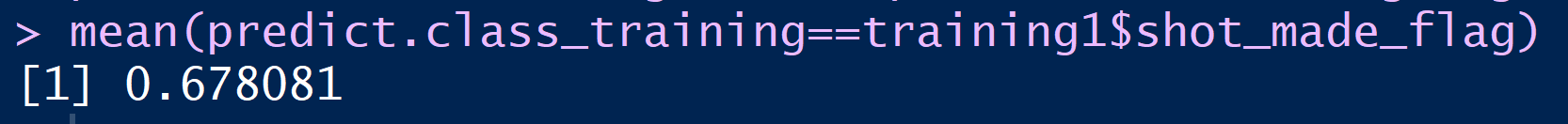


Then we checked the model’s performance on testing dataset.

The following figure shows the confusion matrix. As it shows, the accuracy of testing dataset is 0.6894, which means that our model could predict 68.94% of testing dataset correct, the sensitivity is 0.4795, which is not very good and the specificity is 0.8585.



We also checked the overfitting problem. As the following shows, the accuracy of training dataset is 0.678, which does not differ with the accuracy of testing dataset substantially. This indicates that there is no serious overfitting problem in the lasso logistic regression model.



Also, we checked the assumptions of this ridge logistic regression model.

* Response in binary: for the dependent variable ‘shot\_made\_flag’, there are only two classes-‘0’ and ‘1’. So, the model fits this assumption.
* Independent observations: the records of Kobe dataset were gathered independently, so the model fits this assumption.
* Large sample size: since there are 25697 records in the dataset, which is greater than 30, the model fits this assumption.

Overall, based on all the analysis above, the Lasso model performs the best, which indicates that we do not need all the features.

**C. Conclusion:**

***What can we take away from this project?***

For us:

* We learned how to extract different packages to research our project.
* We understand three different regression models and how to compare the output.
* We command of visualizing package to make people better understand the correlation between various features.

For project:

Based on our initial hypotheses, analysis of coefficients and accuracy, we learned that lasso logistic regression model is the most parsimonious and preferred out of the models developed.

Conclusions that agree with initial hypotheses:

* Kobe’s age, his shot distance and shot type, game’s month and playoffs or not do not have a significant influence on Kobe’s shot accuracy.
* Kobe’s action type, combined shot type and shot zone could have a significant influence on Kobe’s shot accuracy. For the shot zone, just as the outcome of pre-analysis, left side and back court do have a significant influence on Kobe’s shot accuracy since Kobe did not perform very well in both these two zones.

Conclusions that differ from initial hypotheses:

* Although in the pre-analysis, Kobe seems to perform better in short shot distance and worse in long shot distance, the lasso model shows that shot distance has no clearly significant influence on Kobe’s shot accuracy.
* Although in the pre-analysis, we thought that Kobe had a steady performance in each season, but the lasso model shows that season 1997-98, 2014-15, 2015-16 have a significant influence on Kobe’s shot accuracy probably because in these seasons, Kobe were young without any experience or old body age.
* Although in the first hypothesis, we thought that minutes\_remaining, seconds\_remaining and periods are not enough to be a significant variable, the lasso model shows that these three variables are significantly important. We could see the importance of psychological status in a basketball game.

***What would you like to improve on in the future?***

* We can try do to more pre-analysis methods to explore other aspects of the project.
* We can develop more models to recheck our results.
* We can do more visualization to understand the correlation between variables.
* We can extract some data to do cross-validation in order to recheck.
* If possible, we can even try to add some creative new variables to improve the accuracy of the predicted model.
* For future study, we need to keep the focus on the statistical learning and put what we learned in class into real-world practice.