

American University of Armenia

Manoogian Simone College of Business and Economics

Final Project

**Bank Direct Marketing Campaign**

Course: BUS 321- Data Mining for Business Decisions

Professor: Dr. David Rosenfeld

Student: Naira Kocharyan

,

Spring 2022

# Introduction

The topic of this project is the Bank Direct Marketing Campaign; the goal is to create a predictive model to forecast the success of telemarketing calls made by a Portuguese bank to its clients. Success, in this case, is a client who subscribed to a term deposit. Therefore, the research question of this project is how well this model predicts whether or not the client will subscribe. The potential model can help with marketing strategy by identifying clients’ characteristics (e.g., age, education) and campaign characteristics (e.g., type of contact, number of times the client was contacted) that affect the success of a marketing operation.

This question is important because analyzing clients' data will help the bank use the information more efficiently in future direct marketing campaigns. The task is to identify a set of clients who will potentially subscribe. The economic question here is whether or not direct marketing campaigns increase the number of subscriptions, and who the bank should target in terms of clients who are most likely to subscribe. So the motivation is the following: The Portuguese bank may be motivated by targeting potential subscribers more efficiently as a result of choosing the best available model.

To answer these questions, in my project I propose Machine Learning (ML) methods to forecast the success of the bank campaign. The main models are Lasso Model with Cross-Validation (CV), AIC, and BIC, as well as Random Forest (CAT) and Classification Tree (CART). I will present the data cleaning process as the first important step of my analysis as well as discuss the limitations and issues regarding the data and my models. My assumption is that future data will be similar to the past data. This reasoning lets me split my data into two parts: *training\_set* and *test\_set*. I regard the first part of the data(*training\_set*) as the past data and the other part(*test\_set*) as the future data. This is true because both parts are representatives of the same population. I then run my analysis on *training\_set* (CV lasso, random forest, regression tree, etc.). Finally, I will evaluate the error using *test\_set*.

A similar project conducting a data-driven approach to predict the success of bank telemarketing operations was originally done by Sérgio Moro, Paulo Cortez, and Paulo Rita in 2014(Sérgio Moro, 2014). The authors proposed a Data Mining (DM) approach for their analysis. They analyzed a large set of 150 client characteristics and explored a semi-automatic feature selection method which allowed them to reduce the characteristics to 21. My project uses a dataset with a reduced number of characteristics.

# Data

The data used for this project can be found on the UC Irvine Machine Learning Repository (http://archive.ics.uci.edu/ml)); it comes from a Portuguese banking institution’s series of 17 direct marketing campaigns, where an attractive long-term deposit application with reasonable interest rates was pitched and ranges from May 2008 to November 2010. As it was a telemarketing campaign, the direct marketing efforts were conducted through phone calls to clients.

The full dataset contains 41,188 observations with reduced 21 variables (20 input variables and one output variable). The input variables include bank client data: *age*, a numeric variable representing the age of each client; *job*, a categorical variable representing the type of job each client has; *marital*, a categorical variable representing each client’s marital status; *education*, a categorical variable representing each client’s education level; *default*, a categorical variable representing whether or not each client has credit in default; *housing,* a categorical variable representing whether or not each client has a housing loan; and *loan*, a categorical variable representing whether or not each client has a personal loan.

The dataset also contains input variables related to the last contact of the current campaign: *contact*, a categorical variable representing the contact communication type; *month*, a categorical variable representing the last contact month of the year; *day\_of\_week*, a categorical variable representing the last contact day of the week; and *duration*, a numeric variable representing the duration of the last contact.

Also included are input variables related to the campaign in general: *campaign*, a numeric variable representing the number of contacts performed for each client; *pdays*, a numeric variable representing the number of days since each client was last contacted from a previous campaign; *previous*, a numeric variable representing the number of contacts performed before this campaign for each client; and *poutcome*, a categorical variable representing the outcome of the previous marketing campaign for each client.

Lastly, the input variables include social and economic context attributes: *emp.var.rate*, a numeric variable representing the employment variation rate; *cons.pric.idx*, a numeric variable representing the consumer price index; *cons.conf.idx*, a numeric variable representing the consumer confidence index; *euribor3m*, a numeric variable representing the Euribor(Euro Interbank Offered Rate) 3 month rate; and *nr.employed*, a numeric variable representing the number of employees. The output variable *y* is a binary variable representing whether or not the client has subscribed to a term deposit.

# Methods

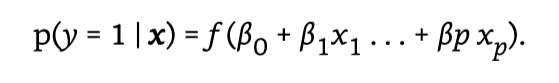
I cleaned the dataset by removing duplicates and ran some simple analysis (average) in order to see outliers, missing data, or structural issues. Before getting too far in my analysis, I made the decision to remove the variable *default* because including it in my vanilla logit regression resulted in an error. This was due to the large proportion of “no” to “yes” values(32,588 no; 3 yes). With the seed I chose, my *training\_set* was left with only values of “no” for *default*, which caused the logit error.

The second step in cleaning my data was to identify issues with missing values such as “NA”(Not Available)*.* I used the is.na() command along with a simple table() command to test if it returned any TRUE values, indicating missing data. I found out that my data doesn’t contain any “NA” values; however, multiple variables did contain a large number of “unknown” values. I addressed this by removing all observations containing a value of “unknown” for any variable. The next important step was to identify categorical variables and convert them into factors. I used str() and names() commands to figure out which variables were categorical, then proceeded to factor them. The final step in cleaning my data was converting my outcome variable *y* from “yes” and “no” to “1” and “0”, respectively, using the as.integer() command. Then I used the naref() command to ensure that R would not set a reference level for any of the categorical variables.

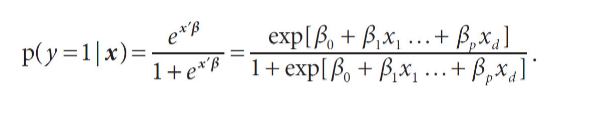
I created a model using k-fold cross-validation and logit with lasso penalty in order to use only the most impactful variables. There are only 20 input variables in this dataset, so I don’t expect many to be left out of the final model.

I will compare different binary classification models in order to pick the best prediction model. I will build these models by applying various machine learning algorithms to my dataset that find an underlying function that calculates the probability of the bank’s client subscribing to a deposit based on input variables explained in the data description section above. Those are the client’s characteristics, telemarketing features, and economic indicators. I will build each model using 90% of the observations and reserve 10% of the observations to evaluate the predictive power of each model. The Machine Learning (ML) methods I will build with this data are Logistic Regression, Logistic Regression with Lasso Regularization, Classification Tree (CART), and Random Forest.

Because this is a binary classification problem, the glm model that will be applied is logit instead of linear regression. I will apply my data to the logit model and build a model that predicts the probability of a client subscribing (y=1) based on the function:



The function will spit out a number between 0 and 1. Because logit is a regression, the function is easier to understand. The coefficients for each of my variables, therefore, will show the expected effect on the probability that the client subscribes. Following the same logic as linear regression, e^b\_k is then interpretable as the multiplicative effect for a unit increase in X\_k on the odds for event y=1:



I will build 2 different logit models using two distinct variable selection methods: AIC and CV. Each will apply different criteria to calculate and select the model with the lowest out-of-sample prediction error to select.

The classification tree, then, is a nonparametric regression model that is made of branching logic. Because trees are nonparametric, they make fewer assumptions about relationships making them powerful for large data sets with non-linear relationships. They are also prone to overfitting.

The final model I will create is a random forest.I evaluate each model using 2 metrics: Mean Squared Error or R2(proportional to each other) and ROC AUC score. The AUC function in my code is a modified version of Matt Taddy’s ROC function. There are many evaluation methods for predictive modeling. Mean Squared Error was chosen because the data didn’t contain many outliers.

ROC AUC score measures the area under the ROC curve from [0,1]. This metric is more applicable to my dataset than other standard classification model metrics such as accuracy, specificity, sensitivity, precision, and recall which rely on a defined threshold. Because the bank dataset class variable is skewed(i.e. 92% didn’t subscribe(y=0)), there is a high chance any model will predict correctly. ROC AUC measures the quality of prediction and ignores threshold and scale. R2 is mainly used to compare in-sample to out-of-sample fit for that model.

In my code, I calculate multiple other evaluation metrics such as accuracy, mean squared error, and precision. I found that this did not give a good measure of comparison to how well a sample actually worked.

# Main Results

My data is imbalanced when it comes to representing clients who subscribed versus refused to subscribe as illustrated in *Figure 1.* 92% of clients respond “no” to long-term deposit subscriptions. With such an unbalanced outcome variable, I would expect most models to score well in any metric of predictive fit. This is seen in my AUC Scores. These AUC scores describe the probability that each model will predict a client’s outcome correctly. The logistic regression models applying CV and AIC did the best.

|  |  |  |  |
| --- | --- | --- | --- |
| **ROC AUC Scores** | | | |
| Logit- CV | Logit-AIC | Classification Tree | Random Forest |
| 0.941 | 0.944 | 0.915 | 0.939 |

Looking at the top covariates by sorting covariates with the largest absolute value for β, it can be seen there was overlap in most effective outcomes.

This predictive model needs to make multiple assumptions to say it can be used by the bank to estimate. A major assumption for predictive models to be useful is that the future population will look similar to the sample population. Intuitively, this makes sense as the bank reaches out to its clients and I assume that clients are stable and have enough similarities to choose a certain bank. However, there are also external factors that make this assumption unlikely to hold true. For example, the economic context variables *emp.var.rate*, *cons.price.idx*, *cons.conf.idx*, *euribor3m*, and *nr.employed* are unrelated to client demographics or marketing campaign statistics; they are all liable to change over a period of time. Because of this, I cannot assume that only historical data is enough for predictions.

Because of these assumptions and possible omitted variables that could create bias, I do not feel confident in making causal inferences. Further, it may not be necessary for the use case presented. If the goal of the bank is to have a tool to predict the outcome of a marketing call, they can have a sense of if the call is worth whatever cost it incurs to implement such a campaign. Looking at the estimated correlation between covariates related to previous outcomes

# Appendix

## Figure 1: Distribution of Outcome

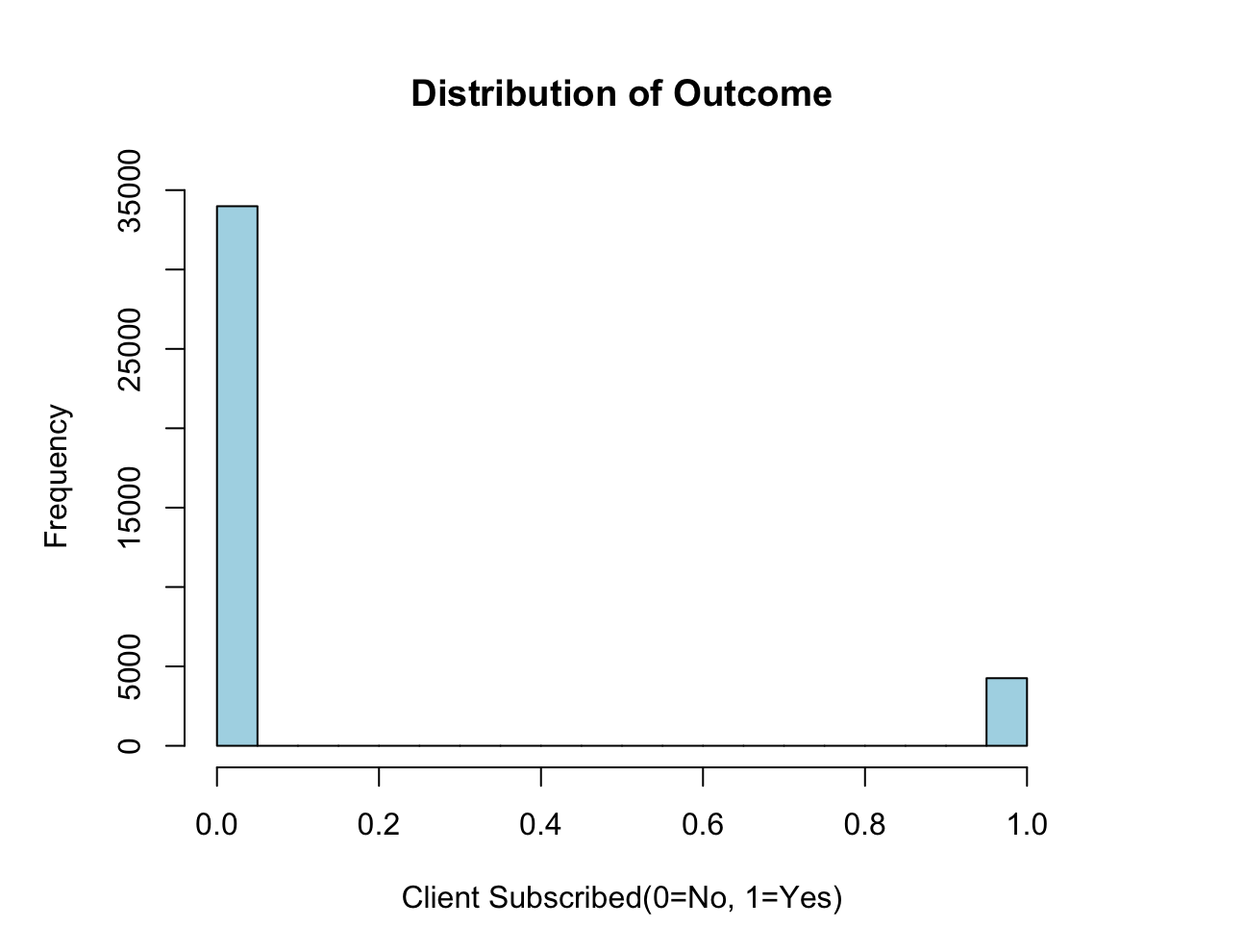


Figure 2: CV Top Coefficients

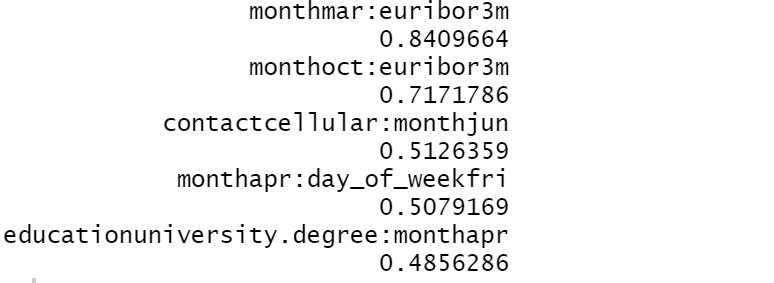


Figure 3: AIC Top Coefficients

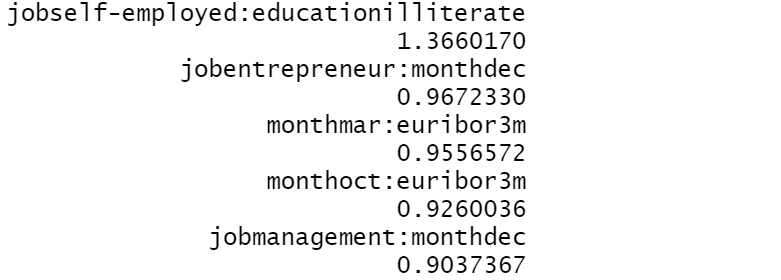


Figure 4: Age Density

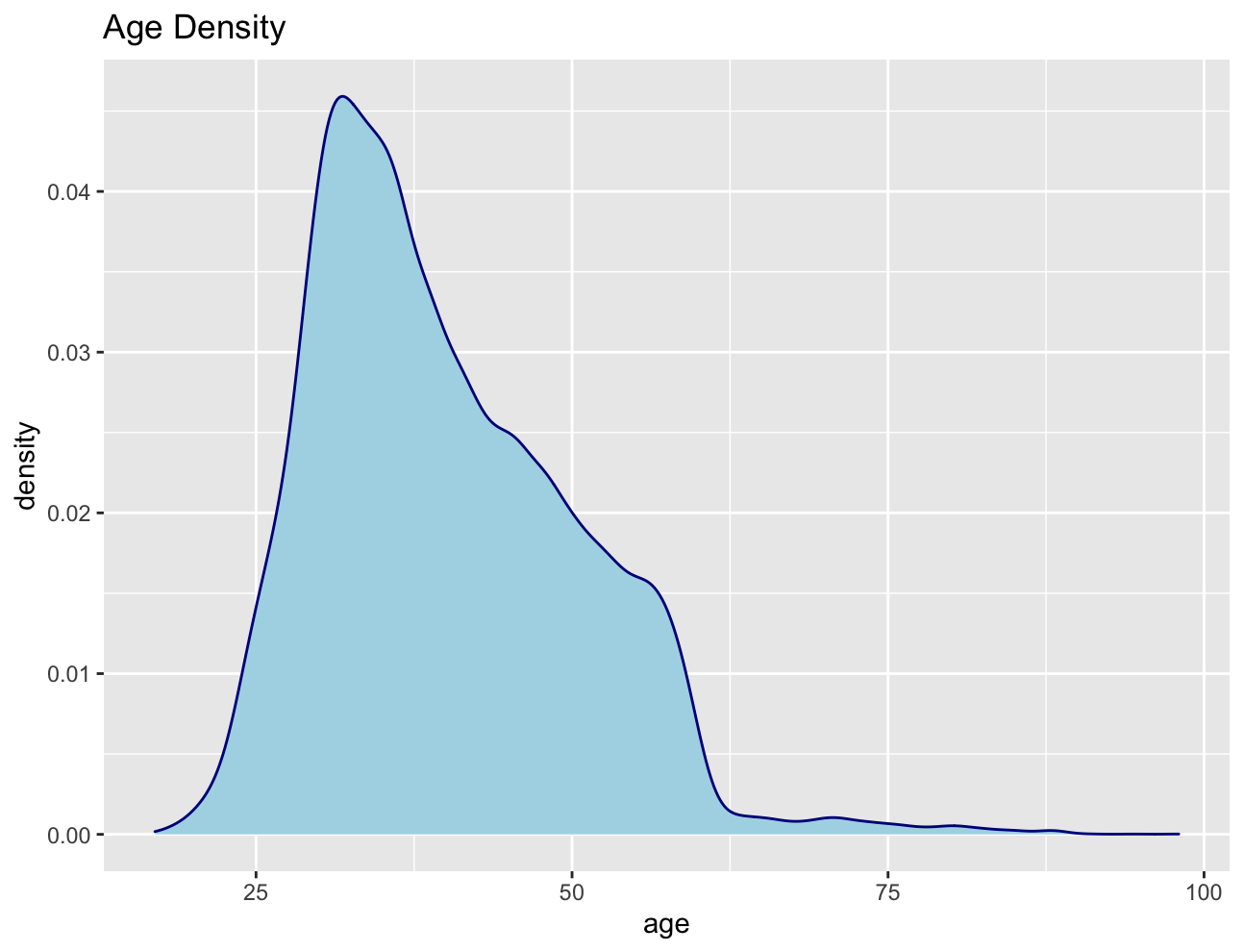


Figure 5:Out of Sample R2 for Logit and MSE for Nonparametric

# 

Figure 6: Plots of Regression Coefficients and Regularization Path

# 

Figure 7: ROC Curve IS

# 

Figure 8: ROC AUC Scores

# 

Figure 9: Plot of AICC Against Lambda

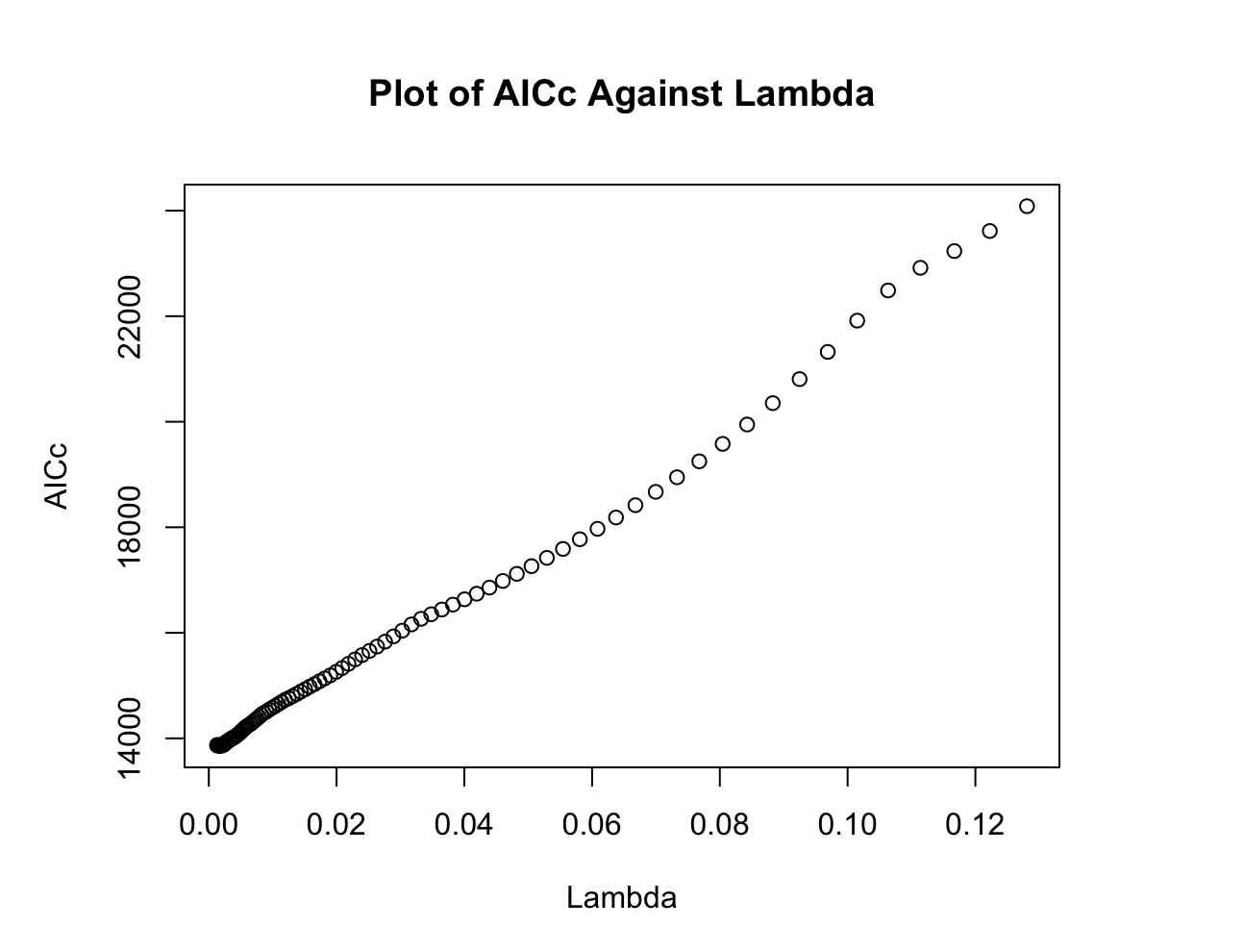


Figure 10: Duration Density

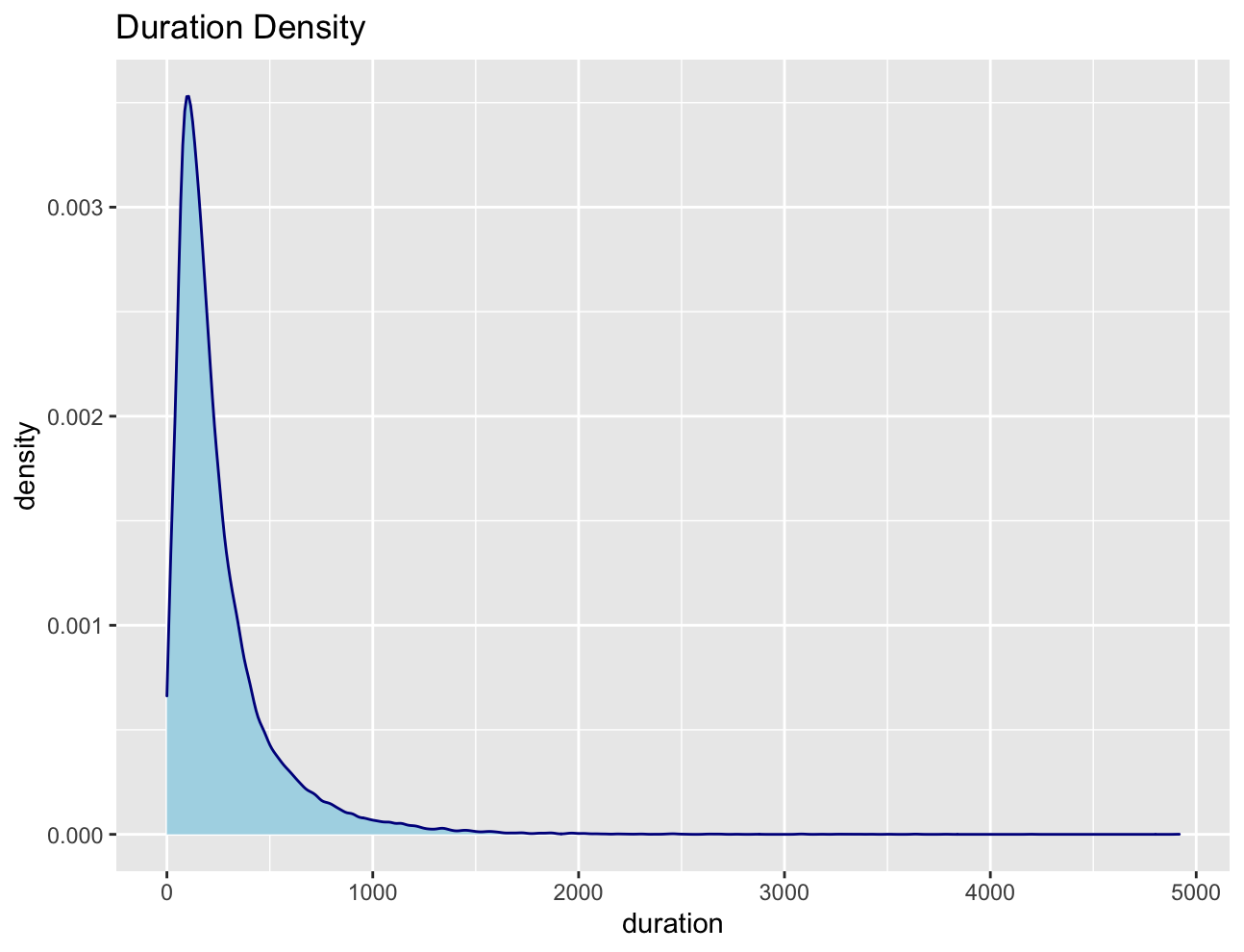


Figure 11: Model Sensitivity

# 

# 

# Works Cited

Sérgio Moro, P. C. (2014). A data-driven approach to predict the success of bank telemarketing, Decision Support Systems. *ScienceDirect, Volume 62*(0167-9236), 22-31. <https://www.sciencedirect.com/science/article/pii/S016792361400061X>

*UCI Machine Learning Repository*. Retrieved (2022, May 5) from https://archive.ics.uci.edu/ml/datasets/Bank+Marketing#