STAT 208 Final Project

Statistical Data Mining Methods

Professor Brandon Wales

Bank Telemarketing Project Report

Cindy Miao, Haiying Lin, Susie Liang, Eunyoung Kwak

June 4, 2024

1. Introduction

1.1 Background

Term deposits are an important financial product for banks and a funding source for income and lending activities. However, acquiring new term deposit customers is a challenging and resource-intensive task. In today's internet-driven world, fewer banks use traditional telemarketing campaigns to acquire loyal customers. By analyzing the records of this telemarketing campaign, we would like to know whether this Telephonic marketing campaign is effective in customer acquisition.

1.2 Objectives

- **Objective 1: Customer Segmentation**
  - Identify customer groups most receptive to term deposit offers.
- **Objective 2: Campaign Optimization**
  - Determine the most effective communication channels and contact frequency.
- **Objective 3: Predictive Modeling**
  - Develop models to predict customer subscription likelihood, allowing for targeted marketing efforts.

2. Data Description

2.1 Data Collection

We were interested in how direct marketing actually influences customer decision-making, so we found the "Banking Dataset - Marketing Targets" on Kaggle. The data is related to the direct marketing campaigns of a Portuguese banking institution.

2.2 Data Overview

This data is based on the results of a telephone direct marketing campaign for a Portuguese banking institution, where only cellular or landline phones are used to call customers to encourage them to sign up for a term deposit and to record the results. The data also includes the results of previous calls on the same clients, showing how the results of previous calls influenced the current call.

The dataset contains 45,211 rows and 17 columns. The target variable in this dataset is ***y***, which is a categorical variable that takes on the value of either "yes" or "no" for whether the customer has subscribed to a term deposit as a result of the current marketing. And there are 16 predictor variables related to the customer's demographic information and financial status, and the previous and current marketing campaign results for customers. In terms of data type, there are 7 numerical variables and 9 categorical variables.

bank *dataset (45,211 rows, 17 columns)*

Category	Name	Description	Type	Values
Customer	age	Age	Numerical	18 ~ 95

Information (Demographics, Financials)	job	Type of job	Categorical	services, admin,...
	marital	Marital status	Categorical	married, …
	education	Education level	Categorical	secondary,primary,...
	default	Has credit in default?	Categorical	Yes, No
	balance	Average yearly balance, in euros	Numerical	-8019 ~ 102127
	housing	Has a housing loan?	Categorical	Yes, No
	loan	Has a personal loan?	Categorical	Yes, No
Marketing Campaign Details	contact	Contact communication type	Categorical	telephone,cellular, ...
	day	Last contact day of the month	Numerical	1 ~ 31
	month	Last contact month of year	Categorical	jan, feb, …, dec
	duration	Last contact duration, in seconds	Numerical	0 ~ 4918
	campaign	Client outreach count	Numerical	1 ~ 63
	pdays	Previous campaign contact lag	Numerical	-1 ~ 871
	previous	Pre-campaign contacts	Numerical	0 ~ 275
	poutcome	Previous campaign results	Categorical	failure,success, …
Target Variable	**y**	**Term deposit subscribed?**	**Categorical**	**Yes, No**

2.3 Data Preprocessing

- **NA values**: The bank dataset contains no NAs.
- **Data type conversion**: For ease of use in EDA and graphing, we converted the categorical variables to factors and redefined the order of the levels of the 'month' and 'poutcome' (results from previous campaign) categorical variables.

```
bank[sapply(bank, is.character)] <- lapply(bank[sapply(bank, is.character)], as.factor)
bank$month <- factor(bank$month, levels = c("jan", "feb", "mar", "apr", "may", "jun", "jul", "aug", "sep", "oct", "nov", "dec"))
bank$poutcome <- factor(bank$poutcome, levels = c("success", "failure", "unknown", "other"))
```

2.4 Data Gap

- **Data disparity - "Unknown" values in variables: *education, contact, poutcome***
  - The "Unknown" values in the categorical variables education, contact, and poutcome, represent missing or unavailable information. These missing values could introduce bias or lead to incomplete or inaccurate insights if not handled properly.
- **Data disparity - Imbalanced Target Variable "y", whether the client subscribes**
  - An imbalanced target variable "y", where the class of non-subscribers is significantly larger than the subscribers, may cause the machine learning models to be biased towards the majority class, potentially resulting in poor performance for the minority class (subscribers). This means that the model may achieve high overall accuracy by predicting the majority class for most instances, but it will perform poorly in identifying the minority class (subscribers).
- **Missing Data - Long-term Outcomes on Deposit Subscription**
  - The dataset only captures the initial subscription decision but does not provide information about the long-term outcomes of the subscribed term deposits, such as "Subscription Duration" and "Churn Indicators", which could be relevant for understanding the overall effectiveness of the Telephonic marketing campaigns.

3. Exploratory Data Analysis (EDA)

3.1 Summary Statistics

- Predictor variables - Customer information

```
      age                job          marital          education     default       balance          housing      loan
 Min.   :18.00   blue-collar:9732   divorced: 5207   primary  : 6851   no :44396   Min.   : -8019   no :20081   no :37967
 1st Qu.:33.00   management :9458   married :27214   secondary:23202   yes:  815   1st Qu.:    72   yes:25130   yes: 7244
 Median :39.00   technician :7597   single  :12790   tertiary :13301               Median :   448
 Mean   :40.94   admin.     :5171                    unknown  : 1857               Mean   :  1362
 3rd Qu.:48.00   services   :4154                                                  3rd Qu.:  1428
 Max.   :95.00   retired    :2264                                                  Max.   :102127
                 (Other)    :6835
```

- Predictor variables - Marketing campaign details

```
    contact           day             month          duration          campaign          pdays            previous          poutcome
 cellular :29285   Min.   : 1.00   may    :13766   Min.   :   0.0   Min.   : 1.000   Min.   : -1.0   Min.   :  0.0000   success: 1511
 telephone: 2906   1st Qu.: 8.00   jul    : 6895   1st Qu.: 103.0   1st Qu.: 1.000   1st Qu.: -1.0   1st Qu.:  0.0000   failure: 4901
 unknown  :13020   Median :16.00   aug    : 6247   Median : 180.0   Median : 2.000   Median : -1.0   Median :  0.0000   unknown:36959
                   Mean   :15.81   jun    : 5341   Mean   : 258.2   Mean   : 2.764   Mean   : 40.2   Mean   :  0.5803   other  : 1840
                   3rd Qu.:21.00   nov    : 3970   3rd Qu.: 319.0   3rd Qu.: 3.000   3rd Qu.: -1.0   3rd Qu.:  0.0000
                   Max.   :31.00   apr    : 2932   Max.   :4918.0   Max.   :63.000   Max.   :871.0   Max.   :275.0000
                                   (Other): 6060
```

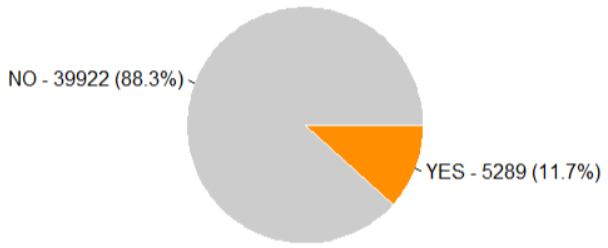
- Target variable - y (is term deposit subscripted?): “yes” - 5,289, “no” - 39,922

3.2 Visual Analysis

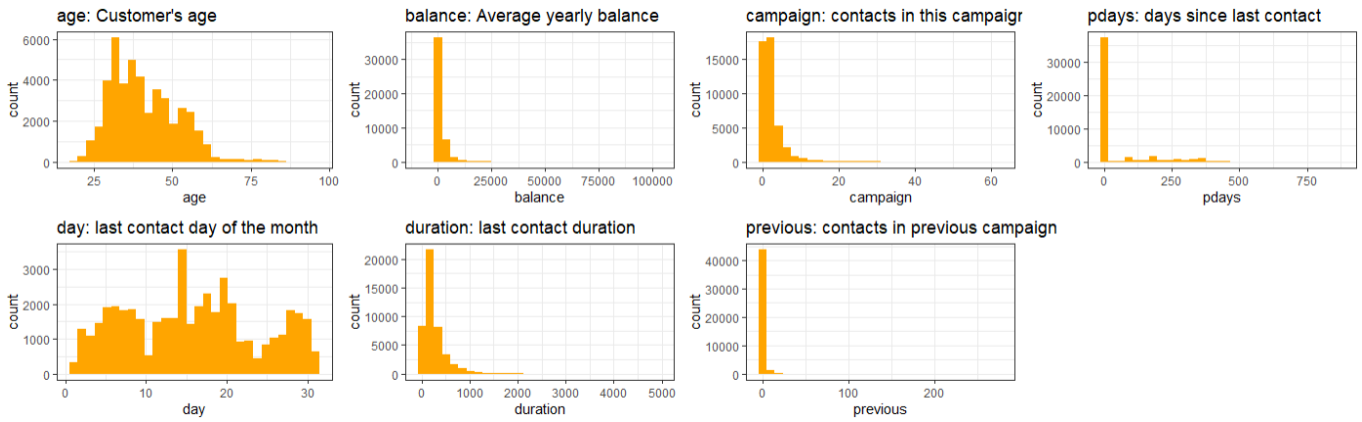
3.2.1 Variable Analysis

Overall, the variables in the bank data are not as evenly distributed as they are in the real world; they are mostly skewed and have many outliers. Therefore, our challenge is how to best utilize these skewed and outlier variables to answer our objectives.

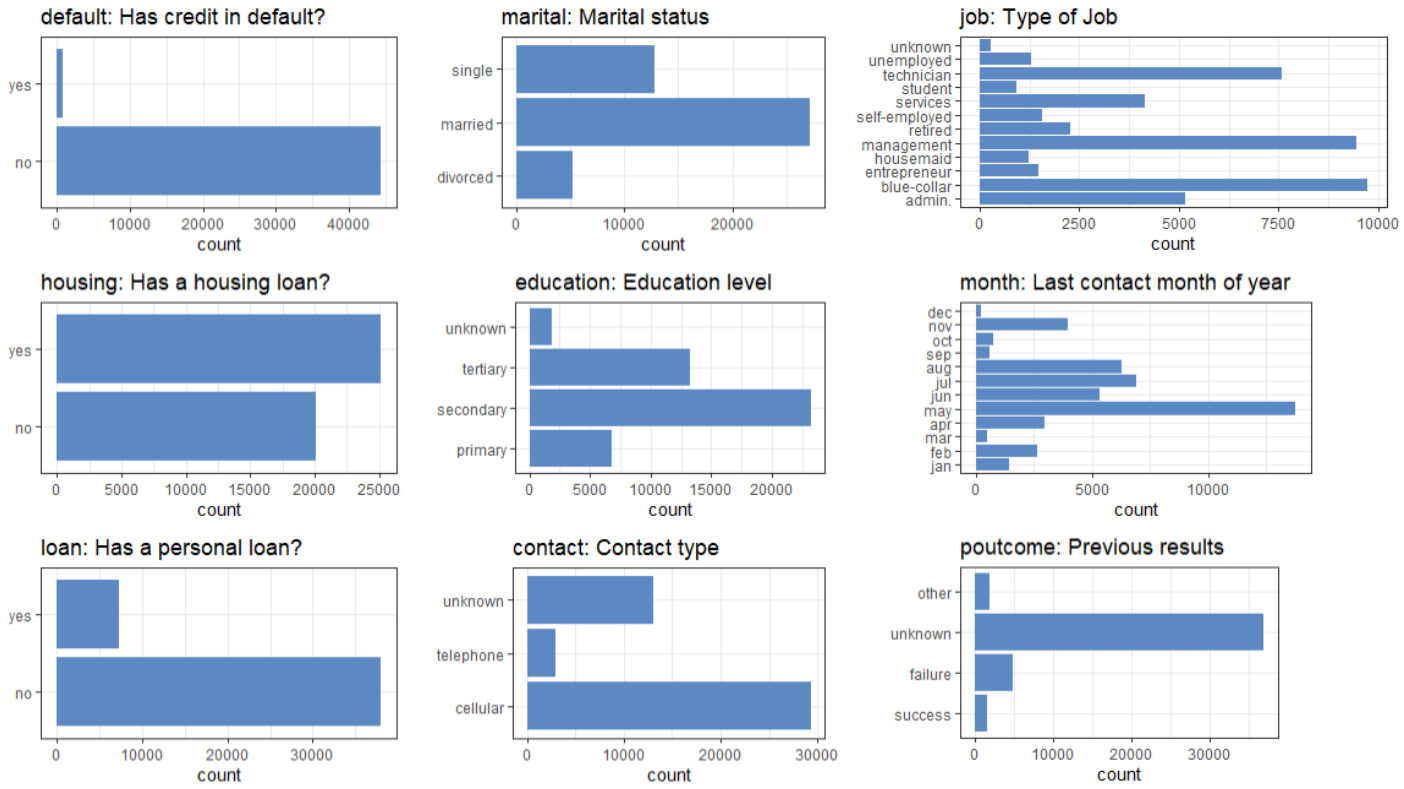
- **Target Variable y**: There is a large difference between the percentages of yes(11.7%) and no(88.3%).



- **Numerical variables** : Many of them are right-skewed and have outliers.

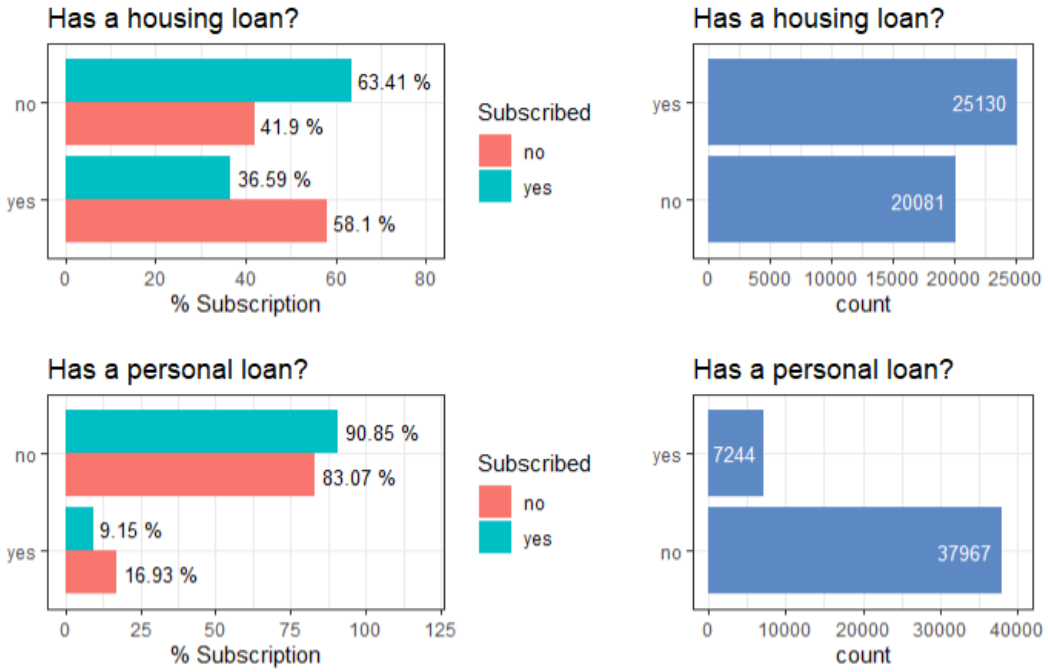


- **Categorical variables**: Some of the variables have unknown values and some have unbalanced values.

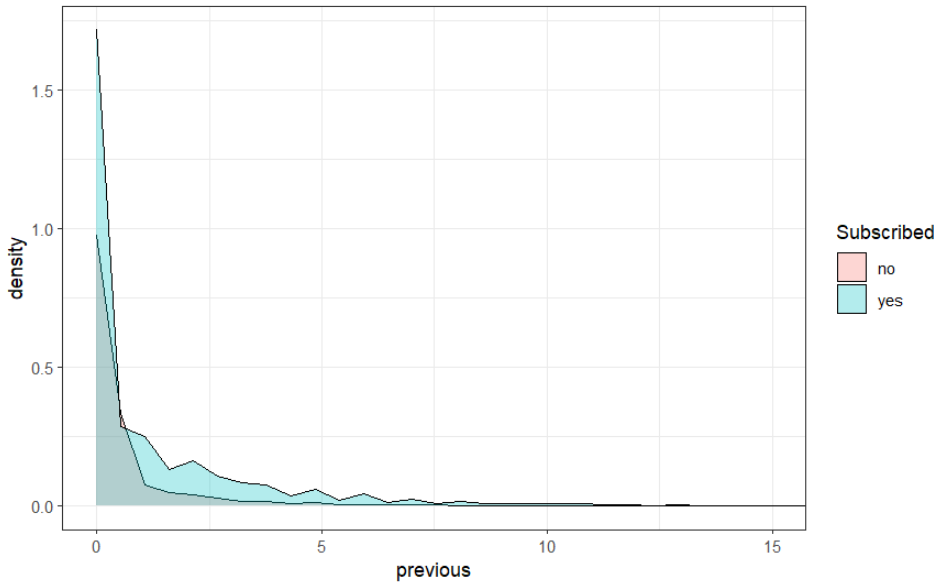


3.2.2 Relationship between variables

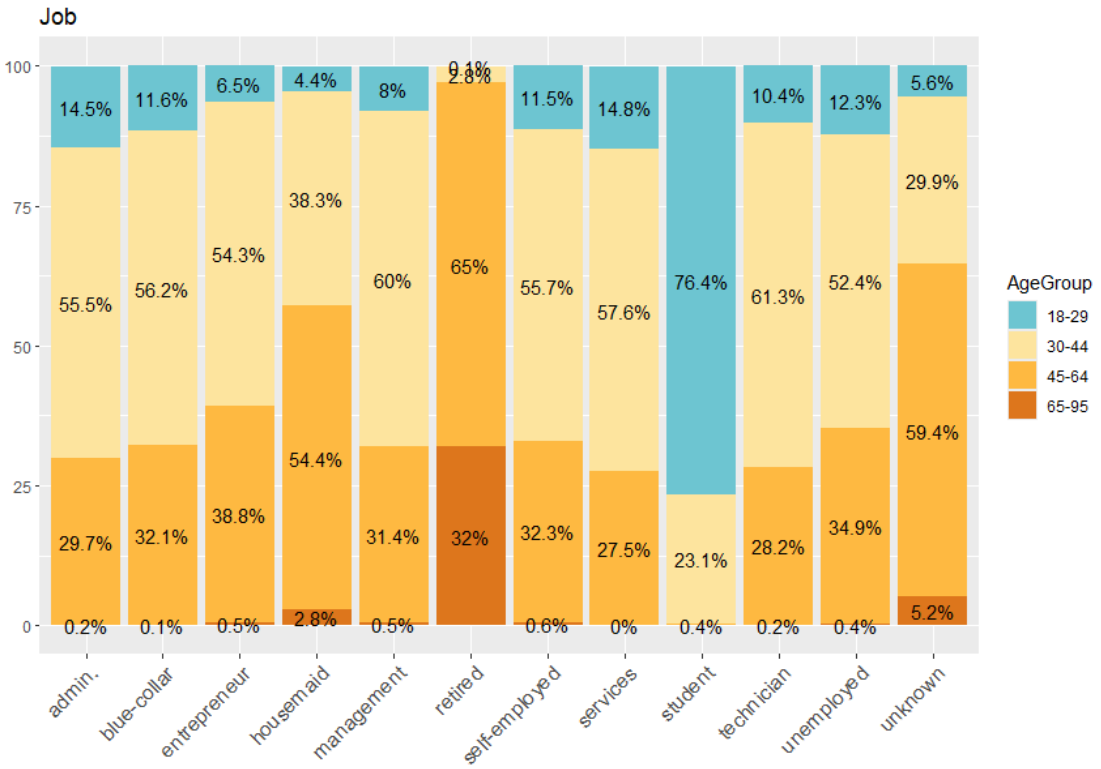
- **Relationship between *loan* status and *subscription rates***: People who don't have a housing loan or a personal loan have a higher subscription rate for term deposits. This may be because they are more financially stable without a loan. It might be worth considering targeting people without loans as potential customers.



- **Relationship between *the number of contacts performed before this campaign* and *subscription rates***: The density plot of the relationship between the number of contacts performed before this campaign (previous) and term deposit subscriptions shows that calling customers frequently before a new campaign definitely helps with subscription rates.



- **Fun fact**: There are 0.1% of our target audience between the ages of 18 and 29 who are already retired, and 2.8% between the ages of 30 and 44. We're jealous. However, it seems unlikely that these customers are truly retired. Perhaps they were just joking about their occupation in this telephone campaign.



3.3 Correlation Analysis

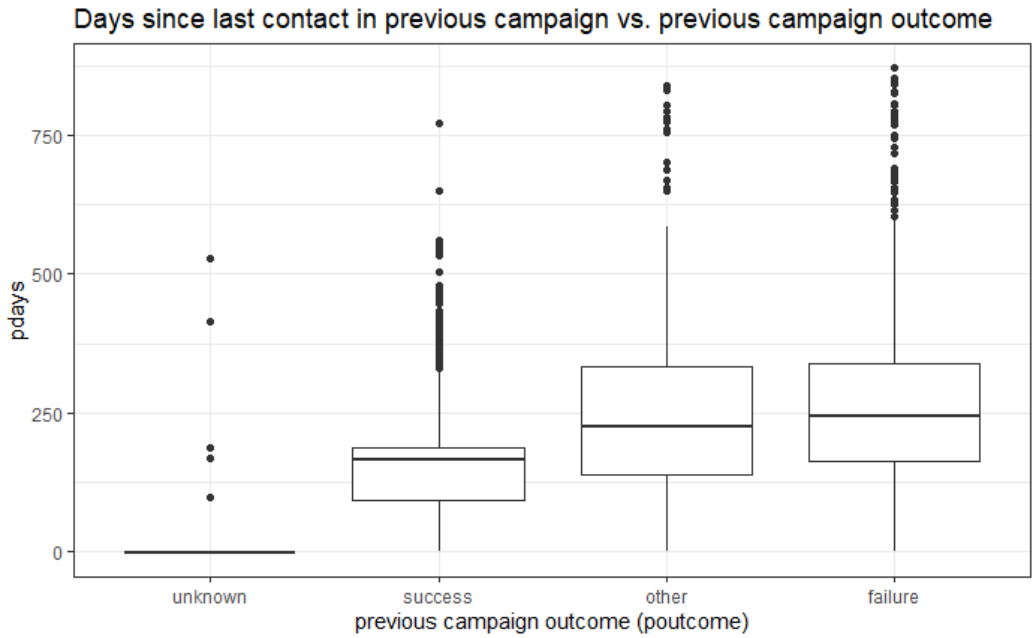
The following first four pairs of variables have an absolute correlation value greater than 0.5, excluding correlations between dummy variables derived from the same variable. We'll look at these four pairs and then see the duration variable that has the highest correlation with our target variable. Let's take a look at them one by one.

- pdays - poutcome(unknown) : -0.87
- pdays - poutcome(failure) : 0.70
- education(tertiary) - job(management) : 0.60
- poutcome(unknown) - previous : -0.53
- duration - y : 0.39

3.3.1 pdays - poutcome(unknown) : -0.87 and pdays - poutcome(failure) : 0.70

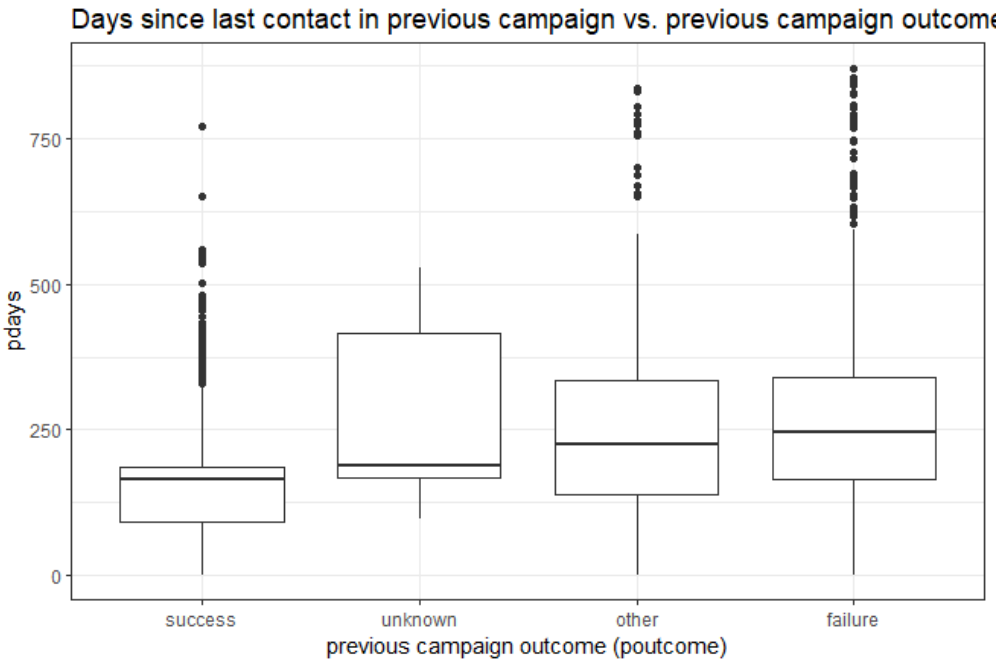
The correlation between *days since last contact in previous campaign*(pday) and *previous campaign outcome* (poutcome) has the highest correlation among all correlations. This indicates these:

- If the outcome of the last campaign was *unknown* and *pday is -1*, they are new target customers for the current campaign who were not targeted by the previous campaign. However, there are 5 customers whose previous campaign outcomes were really unknown. (correlation: -0.87)



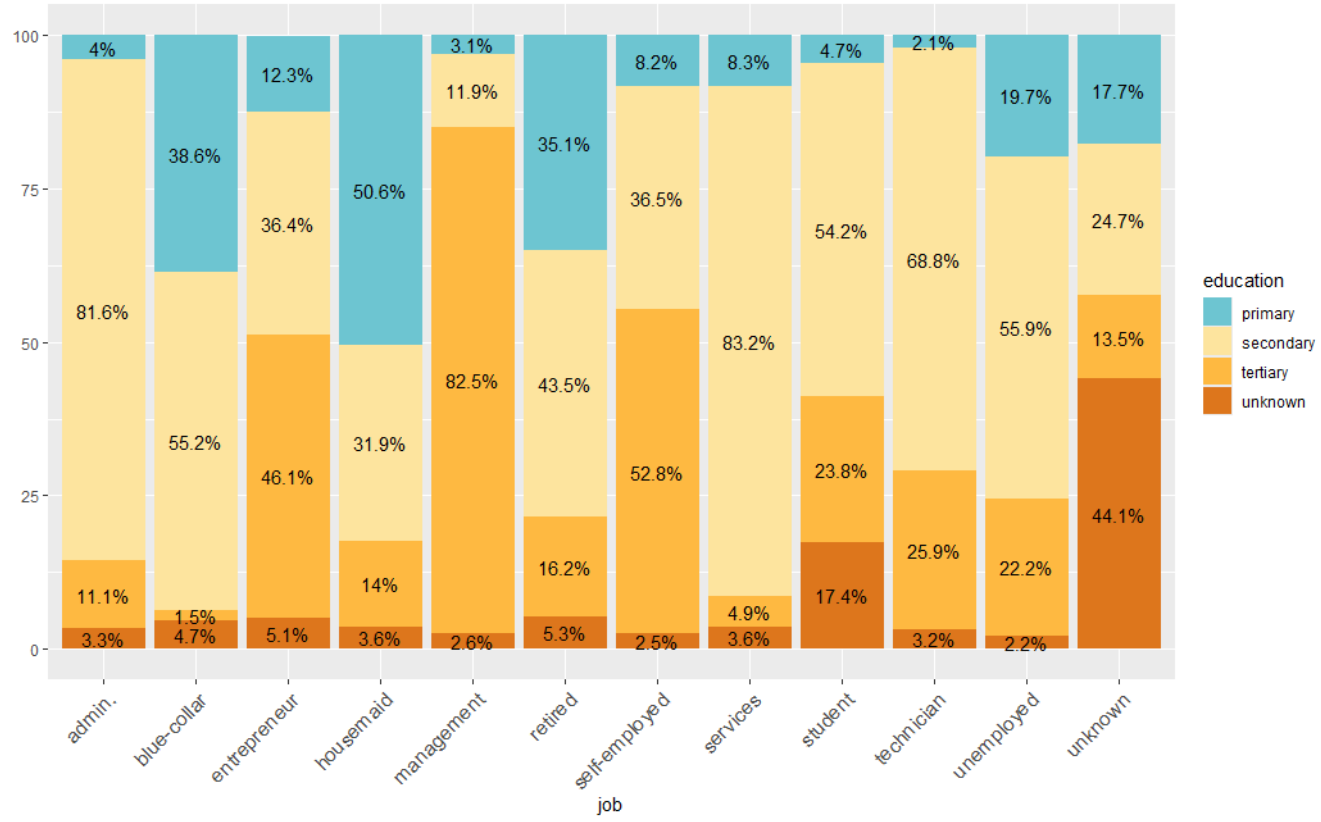
Therefore, to see how quickly the bank reaches out to customers who were already targets as a result of the last campaign, we excluded the new target customers and then looked at the relationship between the two variables again. Now we can see a clear picture.

- If the results are *success*ful, the bank will reach out to the customer again relatively quickly because they are likely to be successful with this campaign as well.
- If the campaign was unsuccessful (*failure* or other), the bank reaches out to the customer at a longer interval because this campaign is likely to be unsuccessful as well. (correlation: 0.70)



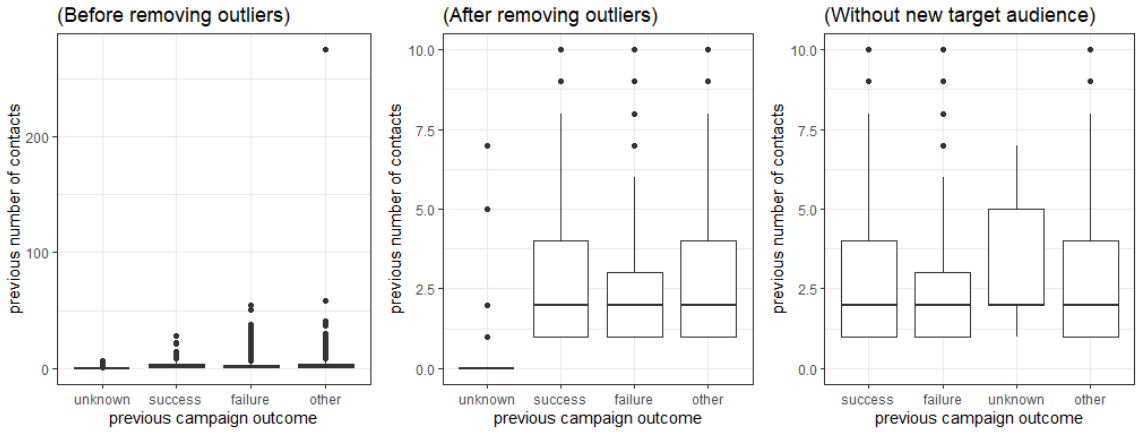
3.3.2 education(tertiary) - job(management) : 0.60

The correlation that people with *a college degree or higher*(education-tertiary) are more likely to *work in management*(job-management) is very real. As you can see in the bar chart below, tertiary-educated clients are overwhelmingly represented in management roles (82.5%) as opposed to other occupations.



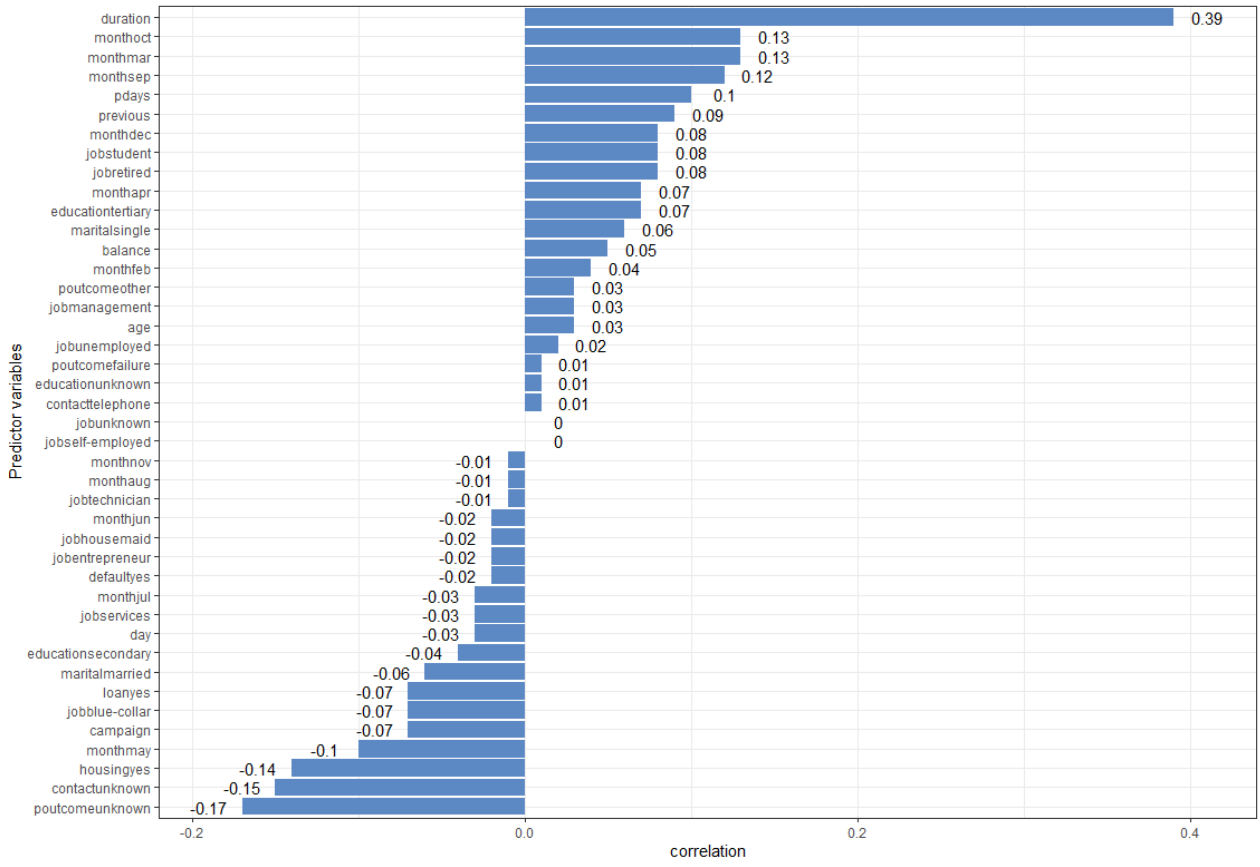
3.3.3 poutcome(unknown) - previous : -0.53

If the results of the last campaign are unknown, then these are new target customers who were targeted for the first time in this campaign. Therefore, if we exclude the 5 existing customers with unknown previous results, then the number of past calls cannot exist. Therefore, the correlation is relatively high. In future analyses, we recommend that you exclude new audiences when looking at the results of past and current campaigns. However, looking at this box plot, it seems like it would be a good idea to make at least 4 calls on average to avoid failure. We'll keep an eye on future analyses to see if this assumption is correct.



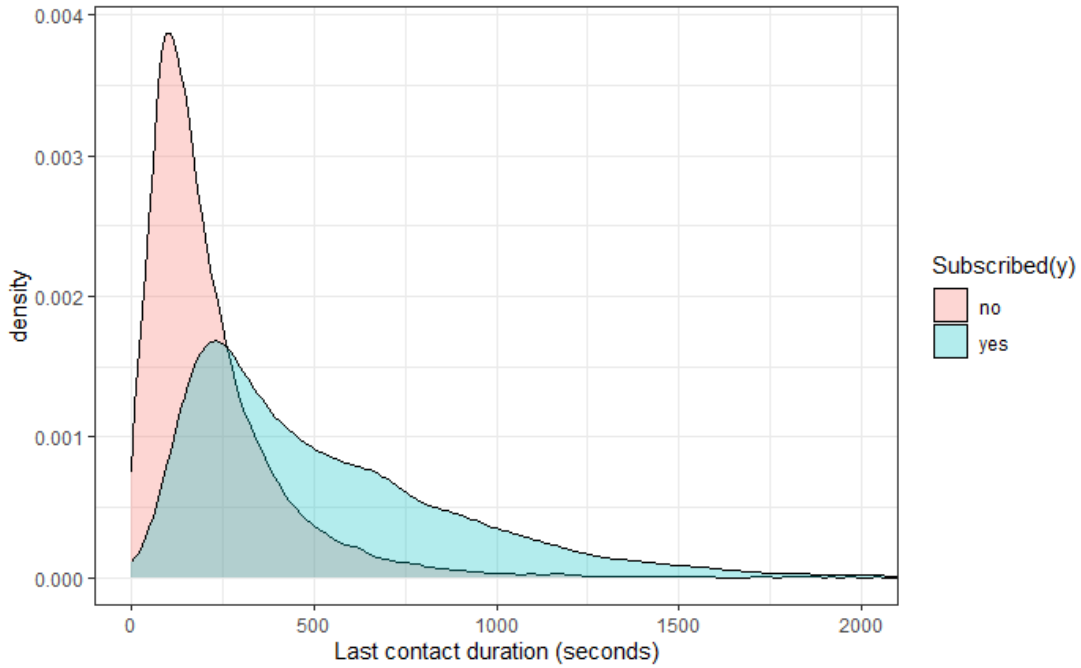
3.3.4. Correlation between target variable(y) and predictor variables

The target variable, y, is not highly correlated with most of the predictor variables overall, but its correlation with *last contact duration*(duration) stands out at 0.39, so let's take a look at that relationship.



3.3.5 duration - y : 0.39

From the density plot below, it appears that the longer you talk to a customer, the more likely you are to get them to sign up for a term deposit. To the naked eye, it looks like between 200 and 300 seconds is the most likely to result in a successful signup, but we'll see if this conjecture holds true in further analysis.



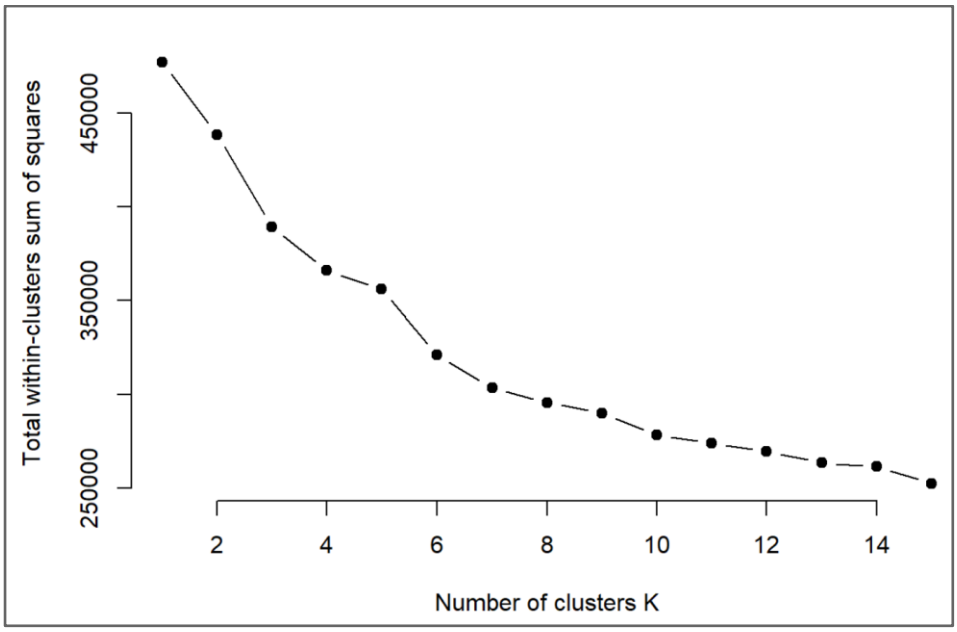
4. Objective 1: Customer Segmentation

4.1 Problem Definition

We will be using the K-mean Clustering method to identify whether the entire customer group can be separated into different sub-segments by analyzing the similar characteristics of their personal information and campaign records. Moreover, we want to find the target customer segment that generates the most term subscriptions. Understanding the specific needs and preferences of each segment, will help us to create future campaign strategies to optimize the telephone marketing process that resonates with more term subscriptions.

4.2 Segmentation Techniques

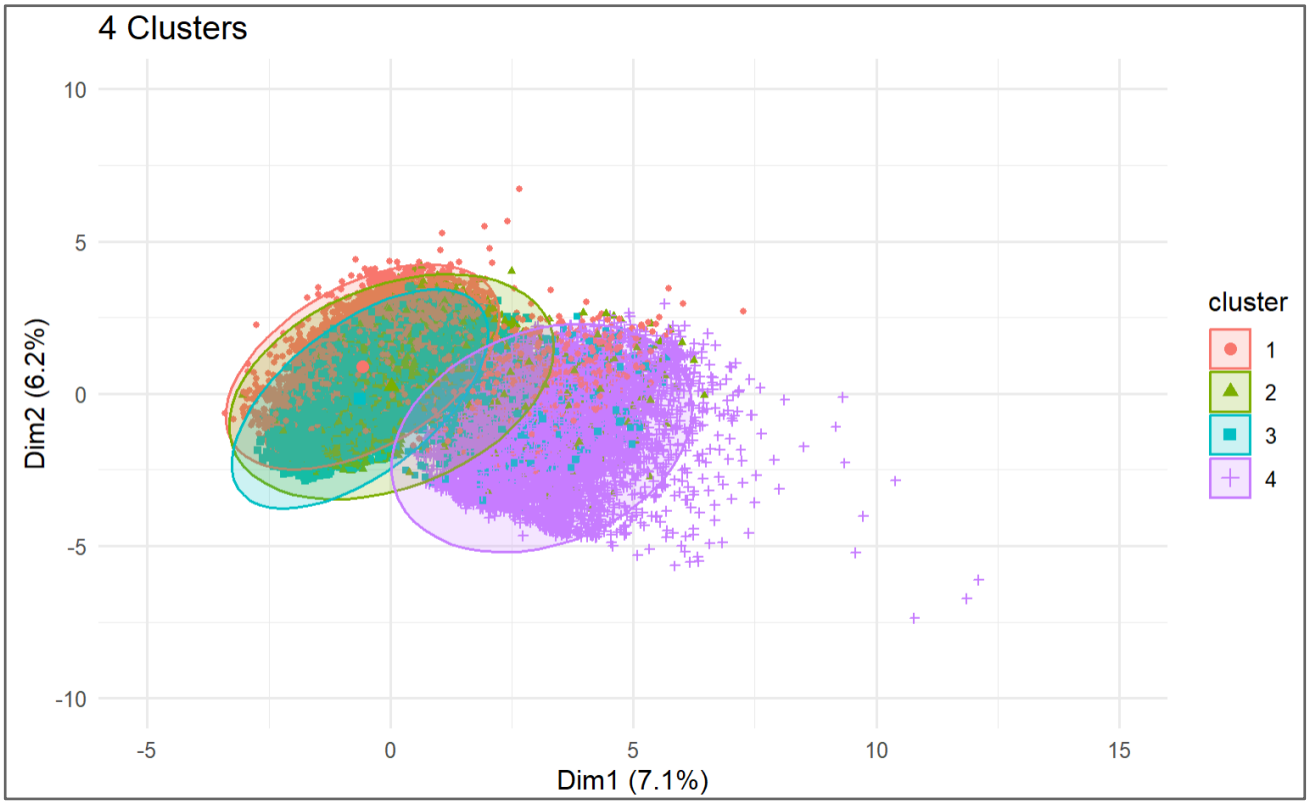
4.2.1 Elbow Test

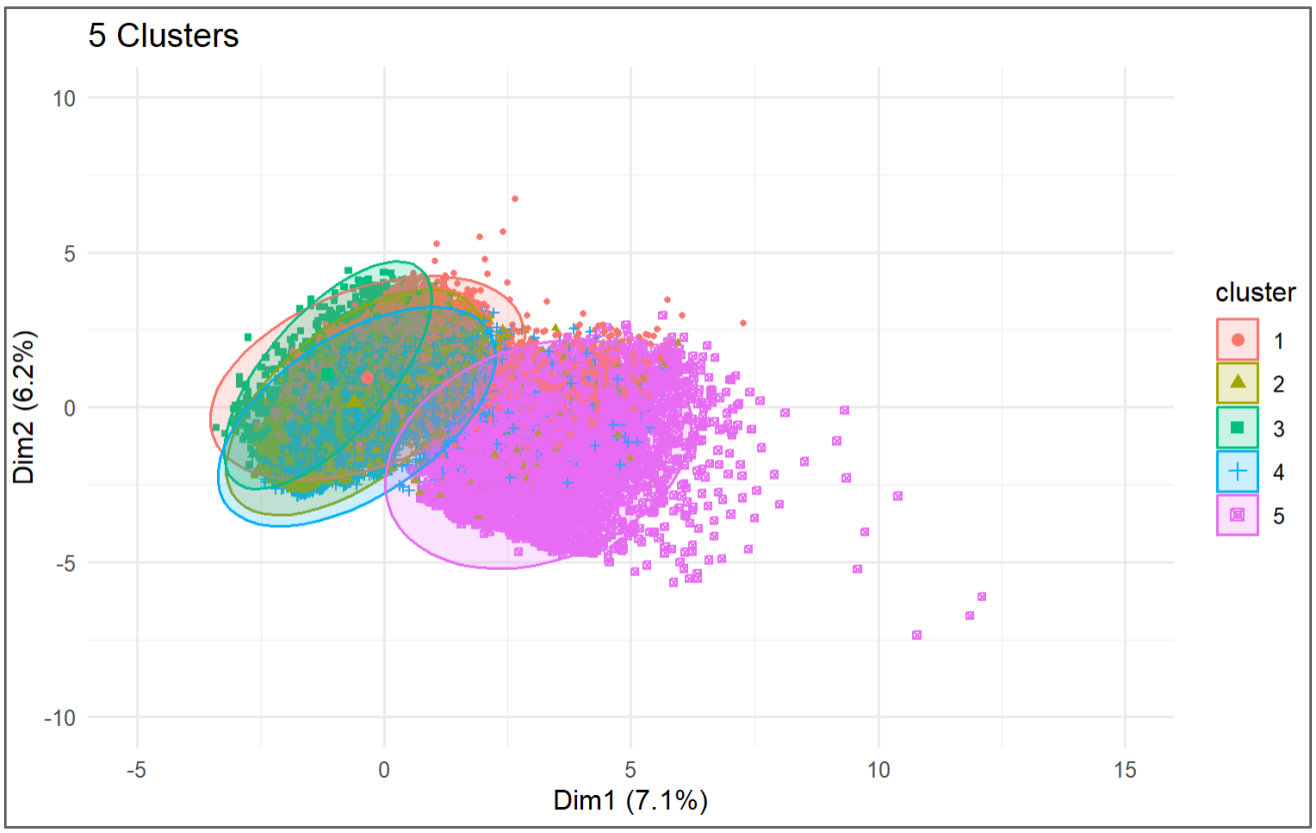


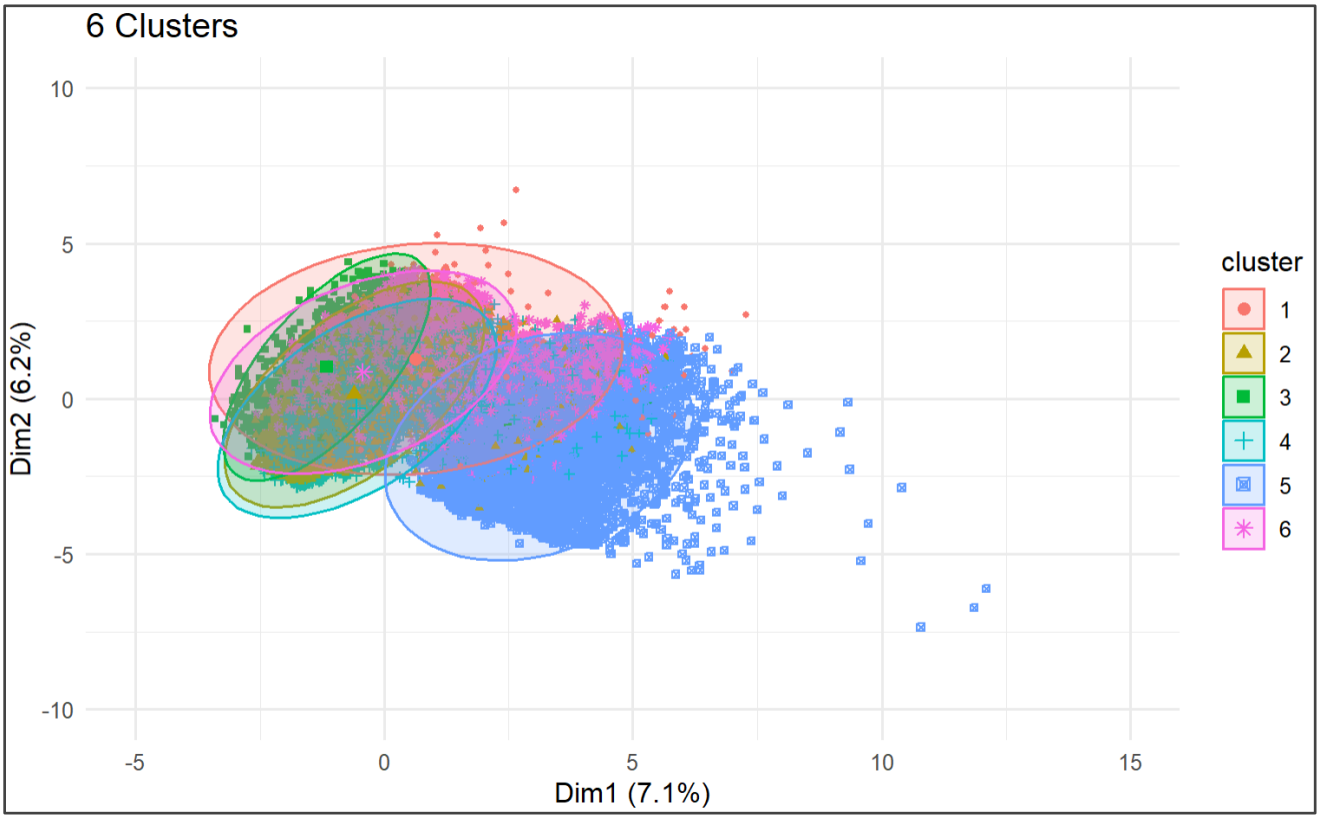
Based on the Elbow Test, we can identify the optimal number of clusters using K-means clustering is within the range of 4-6 as the diminished return starts to occur. We will test the optimal clusterings of 4, 5, and 6, and graph the cluster plots to pick the best customer profile to target.

4.2.2. K-Means Clustering

Cluster Plots with k = 4, 5, 6



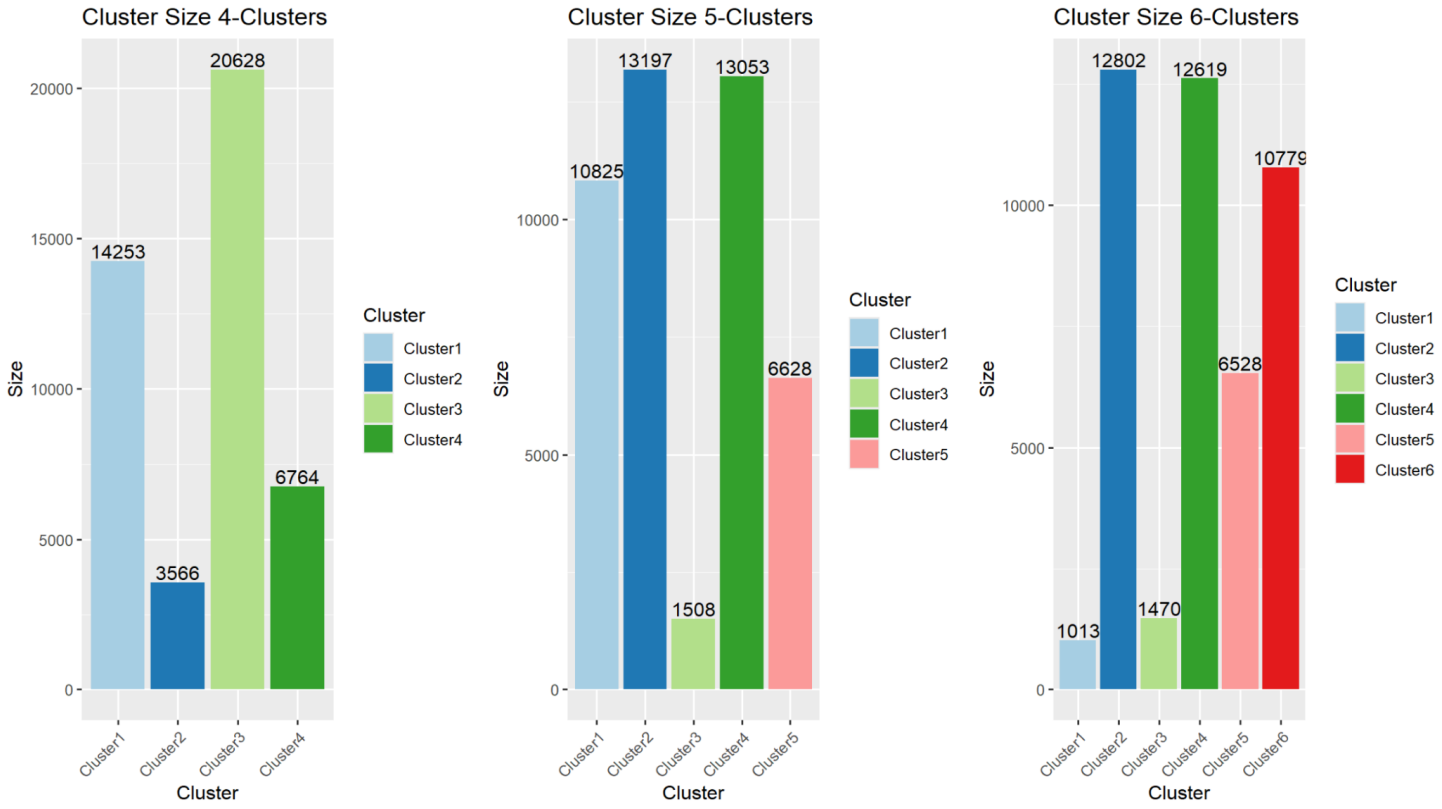




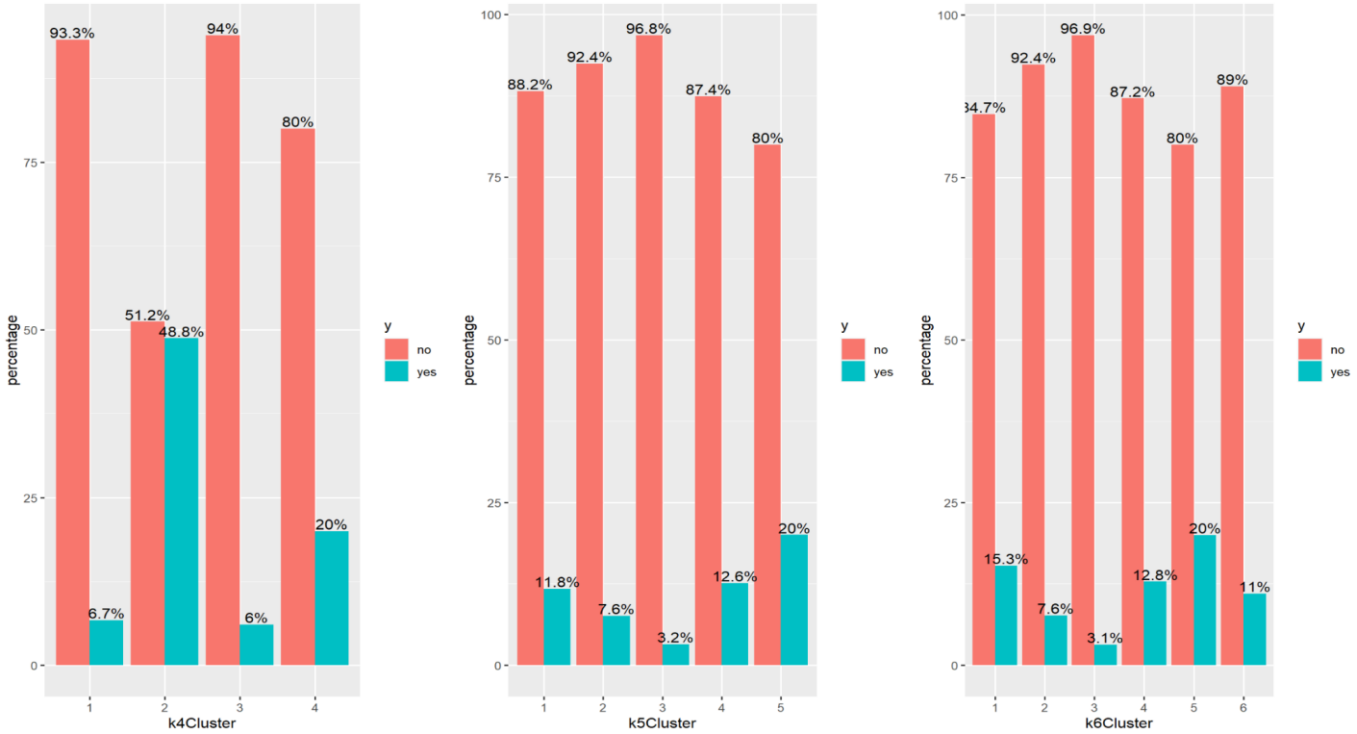
Based on the cluster plots, it's still hard to decide which cluster is the ideal one. Cluster 4 and 5 plot looks very extremely similar to each other, while Cluster 6 has one group that stands out a little differently. Since our objective is to target the

customers that have a higher rate on the term deposit subscription, we will look graph the subscription rate per cluster plot next to help us determine the right target cluster.

Customer Size of Cluster 4, 5, 6



Term Deposit Subscription Rate within Cluster 4, 5, 6



4.2.3. Target Cluster

From the group bar charts of the subscription rate of each cluster 4, 5, and 6, to generate the most term deposit subscriptions, we should divide the entire customers into **4 optimal clusters** and choose the **2nd sub-segment** to target, because it generated the **highest subscription rate of 48.8%** among all clusters.

In reality, if about half of the total potential customers subscribe to the term deposit, this telephone campaign is significantly successful overall.

4.2.4. Interesting Insights

From all the cluster plots, cluster size plots, and subscription rate per cluster plots for different clustering of 4, 5, and 6, we noticed that there is one sub-segment of the customer group that exists among all three clusterings with very little variation of the cluster size but the same customer subscription rate. This represents this segment has distinct and stable customers, and is likely to have high homogeneity within them, meaning the customers within these clusters are very similar to each other. This makes the clustering algorithm consistently group them together.

They are the **4th segment in 4 clustering, the 5th segment in 5 clustering, and the 5th segment in 6 clustering**, with cluster sizes of 6000+ and **subscription rates of 20% yes vs. 80% no**. Also, the 2nd, 3rd, and 4th clusters in both 4 and 5 clusterings have similar trends.

4.3 Segmentation Analysis

4.3.1 Cluster Means of all Variables

4.3.2 Customer Segment Profiles

Based on the cluster mean of every variable, we gain an overall picture of how each variable affects different customer segments and whether they have no effect or significant effect on the y subscription rate to summarize each customer segment profile.

Variables with no effect:
The factors of education secondary and tertiary, management job roles, and having personal loans do not have a significant impact on splitting the customer segments.

Cluster 1

- **Age: Mid-age to Old customers.**
- **Balance: High Balances.**
- **Campaign: High number of Contact Frequency.**
- Contact: High use of 'unknown' contact methods.
- Duration: Shortest calling duration.
- Education: Higher proportions with unknown education.
- Housing: Low likelihood of having a housing loan.
- **Job: Higher proportions in Entrepreneur and House-maid jobs, and Retired customers.**

- **Marital: Higher proportion of Married customers.**
- **Month: High contact rates in Summer Months: June, July, and August.**
- Poutcome: High proportion of Unknown previous campaign outcome, low proportion of success and others.
- Previous: Low number of previous contacts.
- y: Low Subscription Probability.

Cluster 2 (Target Custer)

- Age: Moderate aged customers.
- Balance: Moderate balances.
- Campaign: Moderate contact frequency.
- Contact: High use of 'unknown' contact methods.
- **Duration: Longer calling duration.**
- Housing: Moderate likelihood of having a housing loan.
- **Job: Predominantly in blue-collar, self-employed, and unemployed.**
- Marital: Moderate proportion of married customers.
- **Month: High contact rates in July, and Moderate in May, June, Nov.**
- Poutcome: High proportion of Unknown previous campaign outcomes, low proportion of success, and others.
- Previous: Low number of previous contacts.
- **y: High Subscription Probability.**

Cluster 3

- **Age: Younger customers.**
- Balance: Low and negative balances.
- Campaign: Moderate low contact frequency.
- Contact: Moderate high use of telephone contact.
- **Default: Higher proportion in Credit Default.**
- Duration: Short calling duration.
- Housing: Moderate high likelihood of having a housing loan.
- **Job: Higher proportion work as Services and Technicians, customers that are Students.**
- **Marital: Higher proportion of Single customers.**
- **Month: High contact rates in May, and Moderate in June, July, and August.**
- Poutcome: High proportion of Unknown previous campaign outcomes, low proportion of success, and others.
- Previous: Low number of previous contacts.
- y: Low Subscription Probability.

Cluster 4

- Age: Moderate aged customers.
- Balance: Low balances.

- Campaign: Low contact frequency.
- Contact: Balanced mix of contact methods with Low unknown contacts.
- Default: Lower proportion in credit default.
- Duration: Moderate Short calling duration.
- **Housing: Higher likelihood of having a housing loan.**
- **Job: Higher proportion of Admin Jobs.**
- Marital: Moderate proportions of single and marital customers.
- **Month: High contact rates in Winter and Early Spring Months: Dec, Jan, Feb, Mar and Sep, Oct**
- **Poutcome: High proportion of success and other outcomes**, low of unknown.
- **Pdays: Longer time since the last campaign**
- Previous: Low number of previous contacts.
- y: Moderate Subscription Probability.

5. Objective 2: Campaign Optimization

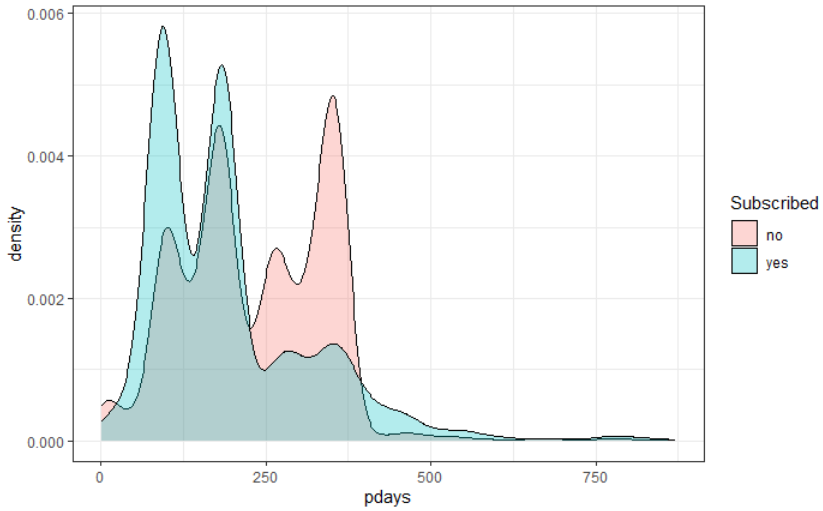
5.1 Problem Definition

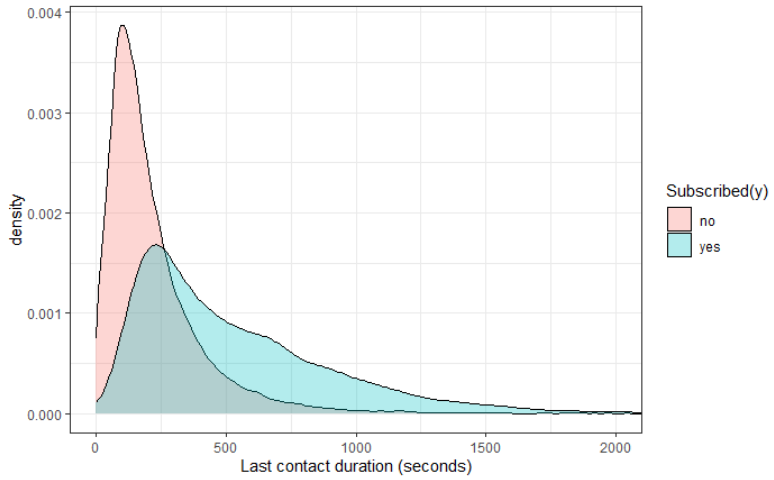
In this section, our objective is to optimize the communication channels and contact frequency for banking telephone advertising campaigns. This involves identifying the most effective channels for different customer segments and determining the optimal frequency of contact to maximize customer engagement and conversion rates.

5.2 Top Correlated Factors with Subscription

5.2.1 Contact Frequency and Duration

For new target audiences, there is no information about previous campaigns. Therefore, we excluded new target audiences and only used customer data which has previous campaign results. Based on the correlation between the target variable and the *days since the last contact in the previous campaign* (pdays), we know that 'pdays' is a critical factor affecting customer subscriptions. The higher the frequency of contact with the customer, the higher the probability of success. The sales department should contact the customer within 200 days.



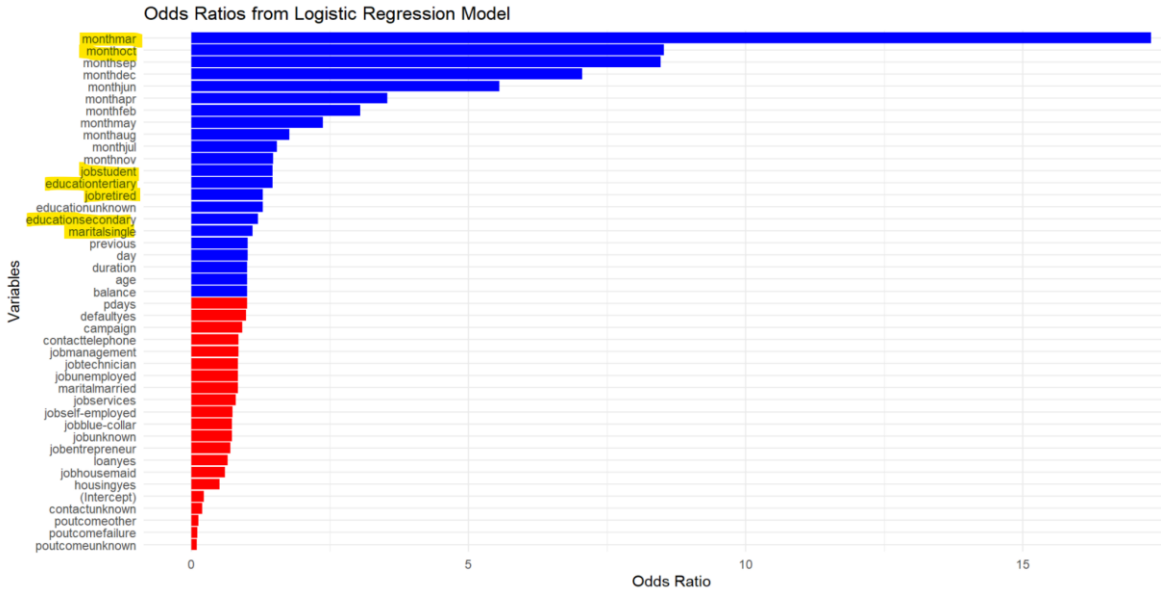


As our EDA part showed it before, the contact 'duration' is right-skewed. Most customers prefer contact durations between 200 and 300 seconds, approximately 5 minutes. If the contact duration exceeds 5 minutes, the likelihood of subscription may decrease.

duration	4.241e-03	7.268e-05	58.347	< 2e-16 ***

The logistic regression model and the clustering indicates that the 'duration' of contact is a crucial factor influencing customer subscription decisions. According to the decision tree analysis, customers prefer contact durations of either less than 522 seconds (8 minutes) or more than 828 seconds (13 minutes). This insight suggests that the sales team can gauge customer purchase intentions within the first 5 to 8 minutes of the conversation. If a customer shows interest in a more in-depth discussion, the sales representative can continue introducing the product. Otherwise, they can move on to the next customer, optimizing their time and efficiency.

5.2.2 Key Drivers of Subscription Success



After calculating the odds ratio based on our logistic regression model, we found that March, October, and September have the highest odds ratios among the months in which customers were contacted during the current campaign. It might be a good idea to analyze the seasonal trends these three months and focus telemarketing promotions on them.

In terms of jobs, students and retired customers seem to be more likely to subscribe to a savings account, so it would be good to create two different promotions for students and one for retired customers. In terms of education, it would be good to target customers with an education level above college.

5.3 Target Clustering (Cluster2) Analysis

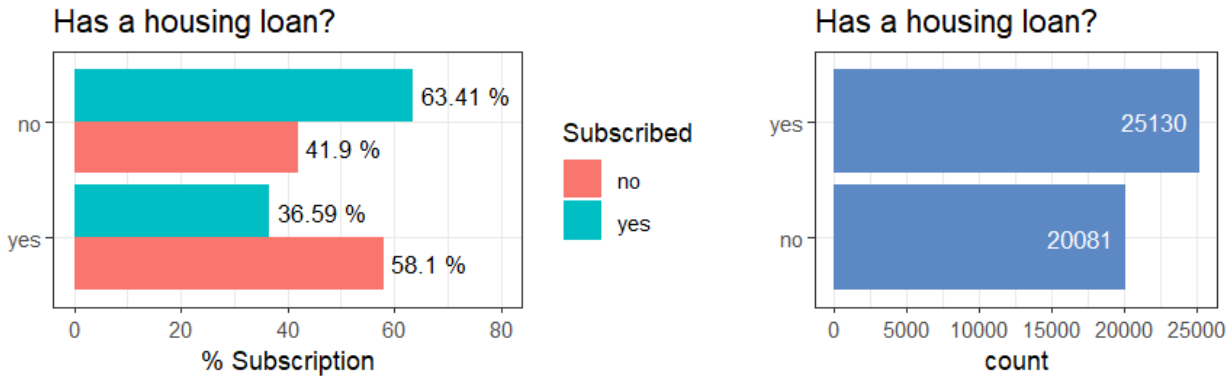
5.3.1 Logistic Regression

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3623  -1.0634  -0.7036   1.1213   1.9205

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.101e+00  5.662e-01   1.944 0.051931 .
age                -7.273e-03  4.509e-03  -1.613 0.106758
`jobblue-collar`   -3.227e-01  1.412e-01  -2.285 0.022306 *
jobentrepreneur    -2.356e-01  2.251e-01  -1.047 0.295192
jobhousemaid       -9.453e-01  2.547e-01  -3.711 0.000206 ***
jobmanagement      -3.580e-01  1.606e-01  -2.229 0.025782 *
jobretired          8.686e-02  2.273e-01   0.382 0.702322
`jobself-employed` -3.015e-01  2.121e-01  -1.421 0.155327
jobservices        -2.550e-01  1.636e-01  -1.558 0.119135
jobstudent          3.758e-01  3.735e-01   1.006 0.314449
jobtechnician      -2.533e-01  1.464e-01  -1.730 0.083659 .
jobunemployed      -2.995e-01  2.185e-01  -1.371 0.170456
jobunknown         -4.063e-01  5.747e-01  -0.707 0.479575
maritalmarried     -4.178e-01  1.109e-01  -3.769 0.000164 ***
maritalsingle      -3.910e-02  1.264e-01  -0.309 0.757042
educationsecondary -7.380e-02  1.131e-01  -0.653 0.513968
educationtertiary  -2.607e-03  1.421e-01  -0.018 0.985357
educationunknown   -3.335e-01  2.120e-01  -1.574 0.115602
defaultyes          1.489e-01  2.741e-01   0.543 0.586879
balance             4.242e-06  1.254e-05   0.338 0.735113
housingyes         -1.943e-01  8.582e-02  -2.264 0.023562 *
loanyes             1.197e-02  1.011e-01   0.118 0.905782
contacttelephone   -2.764e-01  1.707e-01  -1.619 0.105384
contactunknown     -1.213e+00  1.301e-01  -9.320  < 2e-16 ***
day                 1.265e-02  5.137e-03   2.462 0.013826 *
monthfeb            8.912e-01  2.843e-01   3.135 0.001718 **
monthmar            3.095e+00  8.000e-01   3.869 0.000109 ***
monthapr            7.951e-01  2.489e-01   3.195 0.001397 **
monthmay            1.400e+00  2.471e-01   5.664 1.48e-08 ***
monthjun            1.888e+00  2.792e-01   6.764 1.34e-11 ***
monthjul            7.936e-01  2.277e-01   3.485 0.000492 ***
monthaug            1.257e+00  2.457e-01   5.116 3.11e-07 ***
monthsep            1.753e+00  4.213e-01   4.160 3.18e-05 ***
monthoct            1.094e+00  3.767e-01   2.904 0.003685 **
monthnov            5.449e-01  2.396e-01   2.274 0.022936 *
monthdec            1.543e+00  6.831e-01   2.259 0.023876 *
duration            7.924e-04  1.070e-04   7.404 1.32e-13 ***
campaign           -1.375e-02  1.524e-02  -0.902 0.366834
pdays              -2.091e-03  1.439e-03  -1.453 0.146211
previous           -7.899e-02  8.154e-02  -0.969 0.332672
poutcomefailure    -1.363e+00  3.716e-01  -3.668 0.000245 ***
poutcomeunknown    -1.851e+00  4.358e-01  -4.247 2.17e-05 ***
poutcomeother      -1.563e+00  4.153e-01  -3.763 0.000168 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4941.4  on 3565  degrees of freedom
Residual deviance: 4596.9  on 3523  degrees of freedom
AIC: 4682.9

Number of Fisher Scoring iterations: 4
```

Blue-collar, HouseMaid, Duration, Month: March, May, June, August, Sep, Married, housing loan—yes, poutcome(outcome of the previous marketing campaign)

Based on the customer segmentation analysis, we found the target customer cluster:
-Demographic and Employment Factors:

- Marital Status: Married
- Employment: Student and retired
- Loan Status: Not having a housing loan.



-Behavioral Factors:

- Previous Campaign Outcome: Successful outcome in previous campaigns.
- Contact Duration: Longer duration of previous contact.

-Temporal Factors:

- Optimal Contact Months: March, May, June, August, September.

```
# Check for Multicollinearity

Low Correlation

      Term  VIF   VIF 95% CI Increased SE Tolerance Tolerance 95% CI
       age 1.58 [1.52, 1.66]         1.26      0.63      [0.60, 0.66]
       job 3.59 [3.40, 3.80]         1.90      0.28      [0.26, 0.29]
   marital 1.25 [1.21, 1.31]         1.12      0.80      [0.77, 0.83]
 education 2.39 [2.28, 2.52]         1.55      0.42      [0.40, 0.44]
   default 1.03 [1.01, 1.10]         1.01      0.97      [0.91, 0.99]
   balance 1.06 [1.03, 1.11]         1.03      0.94      [0.90, 0.97]
   housing 1.46 [1.41, 1.53]         1.21      0.68      [0.65, 0.71]
      loan 1.09 [1.06, 1.13]         1.04      0.92      [0.88, 0.95]
   contact 3.03 [2.87, 3.20]         1.74      0.33      [0.31, 0.35]
       day 1.41 [1.35, 1.47]         1.19      0.71      [0.68, 0.74]
  duration 1.07 [1.04, 1.12]         1.03      0.94      [0.89, 0.96]
  campaign 1.12 [1.08, 1.16]         1.06      0.90      [0.86, 0.92]
     pdays 3.99 [3.77, 4.23]         2.00      0.25      [0.24, 0.27]
  previous 2.85 [2.70, 3.01]         1.69      0.35      [0.33, 0.37]

Moderate Correlation

     Term  VIF   VIF 95% CI Increased SE Tolerance Tolerance 95% CI
    month 7.31 [6.88, 7.76]         2.70      0.14      [0.13, 0.15]
 poutcome 6.70 [6.31, 7.12]         2.59      0.15      [0.14, 0.16]
```

Checking the Multicollinearity, we see that `month` and `poutcome` are moderate correlation variables with VIF values greater than 5. As well as in the logistic regression summary above, all of the dummy variables of `month` and `poutcome` showed a high significance. Thus, we will remove these two variables.

After Removing High Correlated Variables: month, poutcome:

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2805  -1.1031  -0.7725   1.1346   1.7830

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         4.043e-01  3.010e-01   1.344 0.179100
age                -4.739e-03  4.398e-03  -1.077 0.281261
jobblue-collar     -3.148e-01  1.390e-01  -2.265 0.023531 *
jobentrepreneur    -2.923e-01  2.218e-01  -1.318 0.187495
jobhousemaid       -8.183e-01  2.469e-01  -3.315 0.000918 ***
jobmanagement      -3.469e-01  1.575e-01  -2.203 0.027616 *
jobretired          1.922e-01  2.213e-01   0.869 0.385115
jobself-employed   -3.291e-01  2.087e-01  -1.577 0.114802
jobservices        -2.314e-01  1.613e-01  -1.435 0.151390
jobstudent          4.748e-01  3.664e-01   1.296 0.195107
jobtechnician      -2.331e-01  1.433e-01  -1.627 0.103803
jobunemployed      -3.671e-01  2.127e-01  -1.726 0.084321 .
jobunknown         -6.158e-01  5.619e-01  -1.096 0.273093
maritalmarried     -4.025e-01  1.087e-01  -3.704 0.000212 ***
maritalsingle      -2.705e-02  1.239e-01  -0.218 0.827256
educationsecondary -6.196e-02  1.114e-01  -0.556 0.578123
educationtertiary   7.844e-02  1.395e-01   0.562 0.573806
educationunknown   -2.753e-01  2.087e-01  -1.319 0.187120
defaultyes          1.620e-01  2.712e-01   0.597 0.550394
balance             1.962e-06  1.217e-05   0.161 0.871926
housingyes         -2.598e-01  7.516e-02  -3.456 0.000549 ***
loanyes            -8.175e-02  9.821e-02  -0.832 0.405188
contacttelephone   -2.850e-01  1.659e-01  -1.718 0.085770 .
contactunknown     -6.007e-01  8.105e-02  -7.412 1.25e-13 ***
day                -4.668e-03  4.317e-03  -1.081 0.279540
duration            7.183e-04  1.035e-04   6.940 3.92e-12 ***
campaign            1.603e-03  1.423e-02   0.113 0.910315
pdays              -2.658e-04  9.437e-04  -0.282 0.778192
previous            8.173e-02  6.228e-02   1.312 0.189375
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4941.4  on 3565  degrees of freedom
Residual deviance: 4714.0  on 3537  degrees of freedom
AIC: 4772

Number of Fisher Scoring iterations: 4
```

We could clearly abstract out the important variables “housemaid”, “married”, “housing loan yes”, “contact unknown” and “duration”. Target customer groups can be found with fewer metrics

```
Waiting for profiling to be done...
                            OR     2.5 %    97.5 %
(Intercept)          1.4983138 0.8307636 2.7037794
age                  0.9952721 0.9867233 1.0038875
jobblue-collar       0.7299010 0.5555288 0.9582752
jobentrepreneur      0.7465103 0.4825323 1.1523882
jobhousemaid         0.4411852 0.2701519 0.7123237
jobmanagement        0.7068560 0.5187541 0.9620513
jobretired           1.2119147 0.7861955 1.8732248
jobself-employed     0.7195550 0.4774344 1.0828501
jobservices          0.7934151 0.5780155 1.0881003
jobstudent           1.6076185 0.8034335 3.4153539
jobtechnician        0.7920793 0.5978033 1.0485950
jobunemployed        0.6927205 0.4561530 1.0509510
jobunknown           0.5402119 0.1706457 1.6078857
maritalmarried       0.6686271 0.5400782 0.8270770
maritalsingle        0.9733146 0.7630840 1.2406638
educationsecondary   0.9399174 0.7555529 1.1695364
educationtertiary    1.0815954 0.8229824 1.4219211
educationunknown     0.7593348 0.5028958 1.1409635
defaultyes           1.1758132 0.6898886 2.0075420
balance              1.0000020 0.9999780 1.0000259
housingyes           0.7712400 0.6655383 0.8936198
loanyes              0.9214993 0.7598640 1.1168769
contacttelephone     0.7520233 0.5426657 1.0406115
contactunknown       0.5484219 0.4676459 0.6425688
day                  0.9953427 0.9869512 1.0037981
duration             1.0007185 1.0005181 1.0009243
campaign             1.0016044 0.9736345 1.0298259
pdays                0.9997342 0.9978894 1.0016017
previous             1.0851669 0.9627390 1.2305076
```

5.4 Recommendations for Campaign Strategy

5.4.1 Target Customer Segmentation Optimization

When we do the advertising campaign for these customers, we should develop personalized call scripts that address the specific needs and interests of the target segments. Use engagement techniques to ensure longer call durations. Create exclusive offers and incentives for married customers and those with housing loans, such as lower interest rates on additional loans or special savings plans.

For customers who do not have a housing loan, the focus should be on promoting the flexibility, security, and growth potential of deposit term products. Emphasize how deposit term products provide financial flexibility. Promote the higher interest rates of deposit term products compared to regular savings accounts.

Schedule intensive telemarketing campaigns during the months of May, June, July and Nov.
Analyze past call data to identify the best times of day to reach customers within the selected months.

5.4.2. Entire Campaign Optimization

Based on the analysis, we found that high potential customers: Married, Retired, Students, Entrepreneurs, Previous campaign success, high and moderate yearly balances,and avoided Blue-collar, Housemaids, those with housing loans, and the moderate potential customers: Single, without housing loans, longer duration contacts, previous campaign success.

There are following strategies could improve our success subscription based on the description of our customer:

Yearly Balance

For customers with high and moderate yearly balances, selling deposit terms can be approached with a focus on stability, growth, and tailored benefits. Highlight personalized investment plans based on their yearly balances, offering them higher interest rates or bonus rewards for committing to longer deposit terms. Offer a range of term options to suit different financial goals and lifestyles. For moderate balances, shorter terms might be more appealing, while high balance customers might prefer longer terms for maximum returns.

Timing and Scheduling

Focus campaigns during high-response months: March, April, May, June, July, August,and December.

Personalized Messaging

For married individuals, the introduction of Joint savings accounts, family insurance plans, and education savings plans is paramount. These offerings underscore themes of stability, security, and long-term planning, crucial pillars for safeguarding familial prosperity.

Retirees, on the other hand, benefit from retirement savings plans and annuities. Communicating the advantages of these instruments is essential in empowering them to secure their financial future and navigate retirement with confidence.

Students and entrepreneurs necessitate financial products characterized by flexibility and growth potential. Introducing savings accounts with lower minimum balances and business savings accounts with minimal transaction fees caters to their unique requirements. Emphasizing the opportunities for financial independence and expansion further augments the appeal of such offerings.

Recognizing the nuanced needs of each customer segment, a personalized approach to financial guidance ensures that individuals can effectively manage their finances in alignment with their respective life stages and aspirations.

Follow-Up and Feedback

Establish a follow-up schedule to re-engage customers who express interest but do not immediately commit. Use reminders and additional incentives to convert these leads. After each campaign, gather feedback from customers to refine strategies and scripts. Understand what worked well and what can be improved.

6. Objective 3: Predictive Modeling

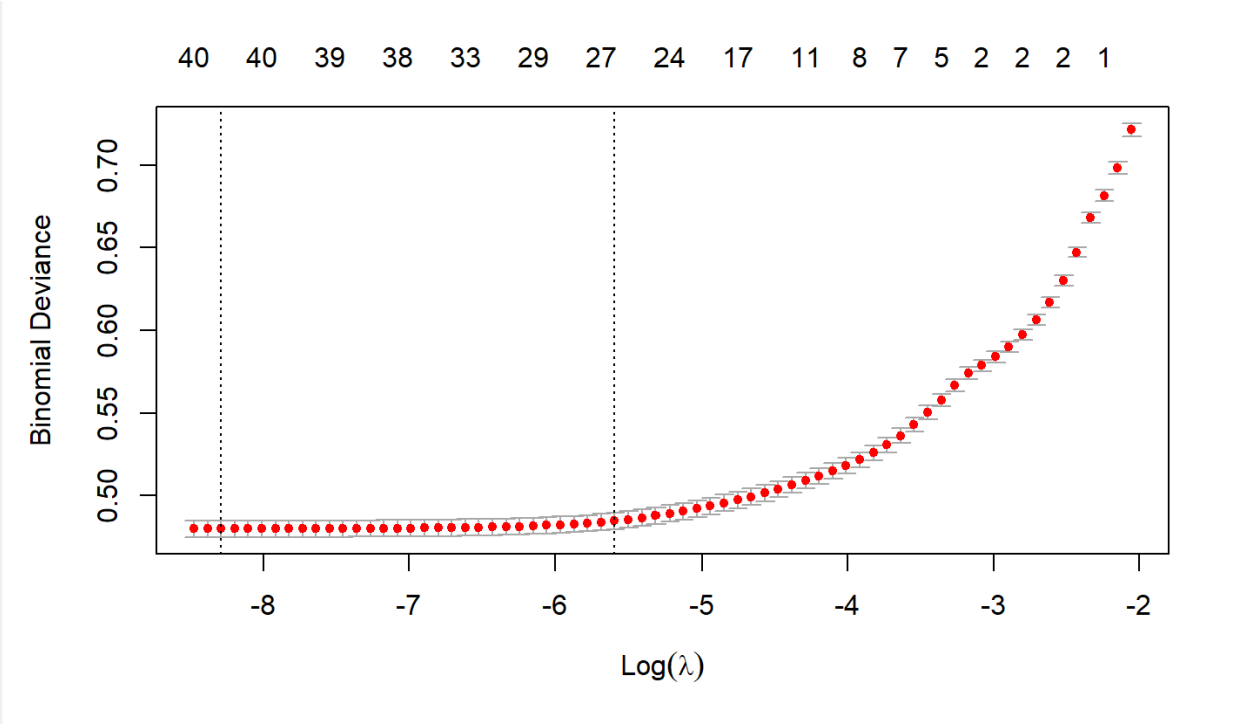
6.1 Problem Definition

The objective of this part is to develop and evaluate predictive models to accurately forecast whether a customer will subscribe to a term deposit based on their profile and past interactions. By leveraging machine learning techniques, we aim to identify the most significant factors influencing subscription decisions.

6.2 Model Selection and Model Predicting

6.2.1 Lasso Variable Selection

Lasso Variable Selection is employed for variable selection and regularization, aiming to improve the prediction accuracy and interpretability of the resulting statistical model. Given the imbalance in our data, we first segregate the data into 'yes subscription' and 'no subscription' categories. Subsequently, we allocate 80% of the data for training and the remaining 20% for testing.




```
Call:  glmnet(x = x, y = y, family = "binomial", alpha = 1, lambda = cv.lasso$lambda.min)

  Df  %Dev    Lambda
1 41 33.87 7.488e-05
```

We chose the lambda value (7.487831e-05) that minimizes the cross-validation error, indicating it is the best regularization strength for this model. The Lasso model selected 42 predictors including dummy variables, which are considered the most relevant for predicting the outcome while imposing penalties on the coefficients of less important variables.

```{r}
coef(lasso_model, lasso_model$lambda.min)
```

```
43 x 1 sparse Matrix of class "dgCMatrix"
                                s0
(Intercept)         -9.291168e-01
(Intercept)          .
age                 -2.033663e-03
jobblue-collar      -3.525898e-01
jobentrepreneur     -3.023273e-01
jobhousemaid        -5.576261e-01
jobmanagement       -1.659069e-02
jobretired           2.752205e-01
jobself-employed    -1.394618e-01
jobservices         -2.087772e-01
jobstudent           3.452913e-01
jobtechnician       -1.459844e-01
jobunemployed       -6.604177e-02
jobunknown          -3.695652e-01
maritalmarried      -1.816724e-01
maritalsingle        8.152557e-02
educationsecondary  -2.010157e-02
educationunknown     3.292698e-02
defaultyes          -1.217850e-01
balance              1.648555e-05
housingyes          -7.131975e-01
loanyes             -4.082641e-01
contacttelephone    -1.678085e-01
contactunknown      -1.615159e+00
day                  5.390038e-03
monthfeb             7.659520e-01
monthmar             2.590788e+00
monthapr             9.885731e-01
monthmay             5.764715e-01
monthjun             1.387814e+00
monthjul             1.704651e-01
monthaug             2.766837e-01
monthsep             1.871558e+00
monthoct             1.824289e+00
monthnov             1.691982e-01
monthdec             1.745887e+00
duration             4.223889e-03
campaign            -9.426637e-02
pdays               -1.281590e-04
previous             1.161780e-02
poutcomefailure     -2.238261e+00
poutcomeunknown     -2.308462e+00
poutcomeother       -2.042925e+00
```

Since we have only 16 independent variables, we will proceed by using all 16 independent variables in the subsequent prediction models.

6.2.2 Logistic Regression

Set the one-time logistic model and then check for Multicollinearity

```
Low Correlation
```

Term	VIF	VIF 95% CI	Increased SE	Tolerance	Tolerance 95% CI
age	2.13	[2.10, 2.17]	1.46	0.47	[0.46, 0.48]
job	3.87	[3.81, 3.94]	1.97	0.26	[0.25, 0.26]
marital	1.43	[1.41, 1.45]	1.19	0.70	[0.69, 0.71]
education	2.25	[2.22, 2.29]	1.50	0.44	[0.44, 0.45]
default	1.02	[1.01, 1.03]	1.01	0.98	[0.97, 0.99]
balance	1.04	[1.03, 1.05]	1.02	0.97	[0.95, 0.97]
housing	1.22	[1.21, 1.24]	1.11	0.82	[0.81, 0.83]
loan	1.04	[1.03, 1.05]	1.02	0.97	[0.95, 0.97]
contact	1.20	[1.19, 1.22]	1.10	0.83	[0.82, 0.84]
day	1.03	[1.02, 1.04]	1.01	0.98	[0.96, 0.98]
month	1.08	[1.06, 1.09]	1.04	0.93	[0.92, 0.94]
duration	1.10	[1.09, 1.12]	1.05	0.91	[0.90, 0.92]
campaign	1.04	[1.03, 1.06]	1.02	0.96	[0.95, 0.97]
pdays	3.68	[3.61, 3.74]	1.92	0.27	[0.27, 0.28]
previous	1.26	[1.24, 1.28]	1.12	0.79	[0.78, 0.80]
poutcome	4.12	[4.05, 4.19]	2.03	0.24	[0.24, 0.25]

The VIF values (below 5) and corresponding tolerance values (higher than 0.2) suggest that most predictor variables exhibit low to moderate multicollinearity. Variables such as “job”, “pdays”, and “poutcome” have higher VIF values but are still within an acceptable range (below 5).

We set up a 10-fold cross-validation and did the Logistic Regression Analysis.

```
# Set up 10-fold cross-validation
train_control <- trainControl(method = "cv", number = 10)

# Train the logistic regression model using 10-fold cross-validation
logistic_model_cv <- train(y ~ ., data = train_data, method = "glm", family = "binomial",
trControl = train_control)
```

```
##
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.7717  -0.3745  -0.2522  -0.1488   3.4855
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -1.334e+00  2.253e-01  -5.921 3.20e-09 ***
## age                -8.292e-04  2.457e-03  -0.338 0.735719
## `jobblue-collar`   -2.901e-01  8.191e-02  -3.541 0.000398 ***
## jobentrepreneur    -3.691e-01  1.418e-01  -2.603 0.009231 **
## jobhousemaid       -5.030e-01  1.527e-01  -3.295 0.000985 ***
## jobmanagement      -1.446e-01  8.256e-02  -1.752 0.079776 .
## jobretired          2.969e-01  1.088e-01   2.729 0.006361 **
## `jobself-employed` -2.287e-01  1.233e-01  -1.855 0.063629 .
## jobservices        -2.088e-01  9.398e-02  -2.222 0.026300 *
## jobstudent          3.433e-01  1.224e-01   2.804 0.005050 **
## jobtechnician      -1.933e-01  7.789e-02  -2.482 0.013061 *
## jobunemployed      -7.455e-02  1.232e-01  -0.605 0.544960
## jobunknown         -3.977e-01  2.608e-01  -1.525 0.127281
## maritalmarried     -1.775e-01  6.583e-02  -2.696 0.007024 **
## maritalsingle       7.183e-02  7.530e-02   0.954 0.340167
## educationsecondary  2.155e-01  7.243e-02   2.975 0.002929 **
## educationtertiary   3.848e-01  8.393e-02   4.585 4.54e-06 ***
## educationunknown    2.792e-01  1.163e-01   2.401 0.016335 *
## defaultyes         -1.206e-01  1.905e-01  -0.633 0.526767
## balance             1.585e-05  5.783e-06   2.740 0.006141 **
## housingyes         -7.162e-01  4.916e-02 -14.568  < 2e-16 ***
## loanyes            -4.129e-01  6.720e-02  -6.145 8.02e-10 ***
## contacttelephone   -1.590e-01  8.369e-02  -1.900 0.057467 .
## contactunknown     -1.621e+00  8.213e-02 -19.735  < 2e-16 ***
## day                 6.397e-03  2.789e-03   2.294 0.021819 *
## monthfeb            8.924e-01  1.444e-01   6.180 6.42e-10 ***
## monthmar            2.699e+00  1.664e-01  16.218  < 2e-16 ***
## monthapr            1.107e+00  1.323e-01   8.370  < 2e-16 ***
## monthmay            7.038e-01  1.313e-01   5.360 8.33e-08 ***
## monthjun            1.518e+00  1.467e-01  10.347  < 2e-16 ***
## monthjul            2.943e-01  1.302e-01   2.259 0.023862 *
## monthaug            3.882e-01  1.317e-01   2.947 0.003210 **
## monthsep            2.000e+00  1.659e-01  12.050  < 2e-16 ***
## monthoct            1.932e+00  1.549e-01  12.472  < 2e-16 ***
## monthnov            2.864e-01  1.351e-01   2.119 0.034050 *
## monthdec            1.860e+00  2.200e-01   8.454  < 2e-16 ***
## duration            4.241e-03  7.268e-05  58.347  < 2e-16 ***
## campaign           -9.587e-02  1.151e-02  -8.331  < 2e-16 ***
## pdays              -1.507e-04  3.404e-04  -0.443 0.658060
## previous            1.209e-02  7.315e-03   1.653 0.098349 .
## poutcomefailure    -2.243e+00  9.105e-02 -24.630  < 2e-16 ***
## poutcomeunknown    -2.313e+00  9.568e-02 -24.175  < 2e-16 ***
## poutcomeother      -2.041e+00  1.099e-01 -18.567  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 26108  on 36169  degrees of freedom
## Residual deviance: 17244  on 36127  degrees of freedom
## AIC: 17330
##
## Number of Fisher Scoring iterations: 6
```

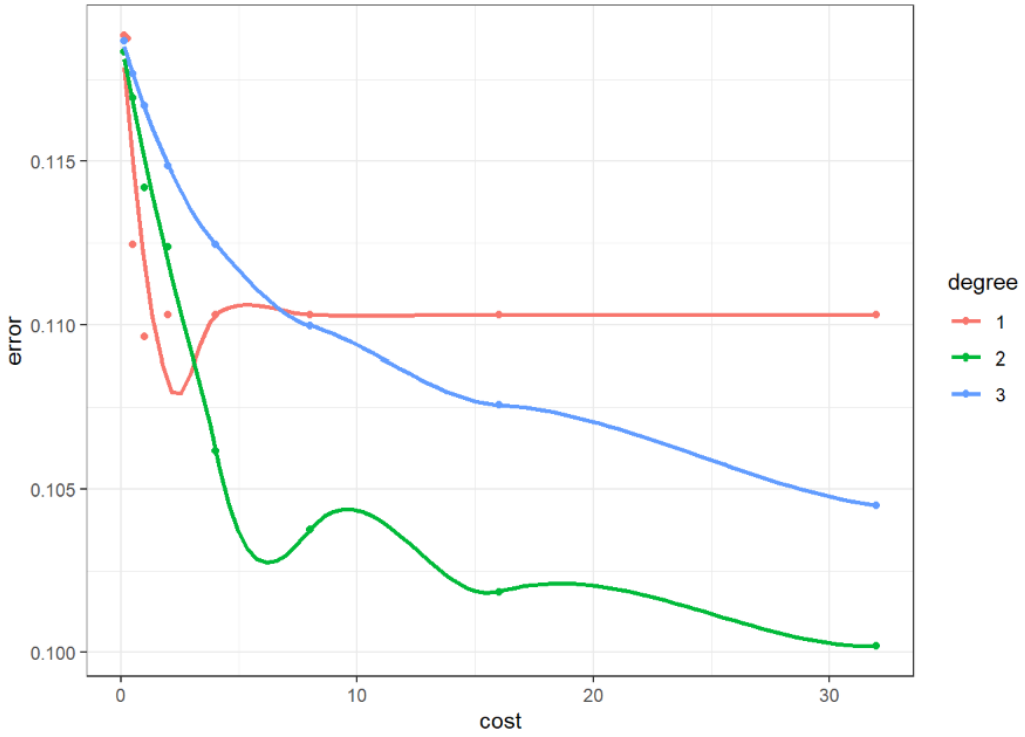
The selected predictors are important for predicting the outcome, with variables like " job blue-collar", "job entrepreneur", "balance", and "duration" "campaign" "poutcome" showing strong significance in their relationship with the dependent variable. But there are some different between each variable's categories, the estimated coefficients for each job category indicate the direction and strength of the association with the outcome variable, compared to the baseline category (typically one of the job categories not explicitly listed, such as 'jobadmin')

They are not at a level that typically requires corrective measures.

6.2.3 Support Vector Machine(SVM) —Polynomial

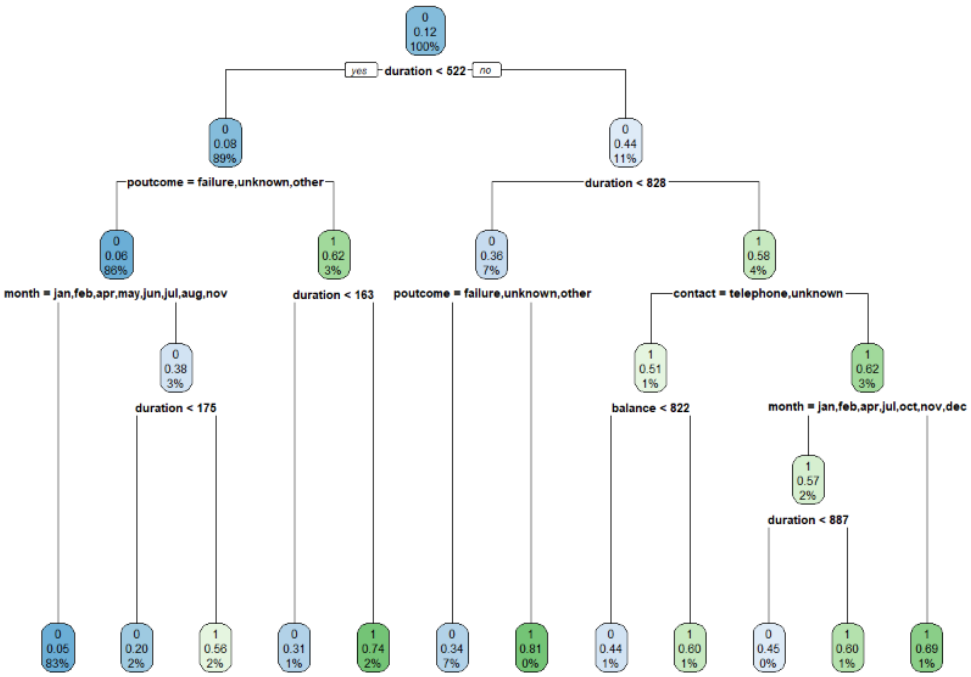
```
svm.polynomial <- tune(svm, y~., data=train_data, tunecontrol=tune.control(sampling="fix"),
                       kernel="polynomial", ranges=list(cost=2^seq(-3,5,1), degree=c(1,2,3)))

svm.polynomial$performances %>% arrange(error) %>% head(1)
```

```
##   cost degree     error dispersion
## 1   32      2 0.1001908         NA
```



The SVM model was tuned to find the best hyperparameters using a polynomial kernel. The best model had a cost parameter of 32 and a polynomial degree of 2, with an error rate of about 10.02%. This means that these hyperparameters were optimal in minimizing the error during the cross-validation process. The visualization confirms that this combination of hyperparameters achieves the optimal balance between model complexity and predictive accuracy.

6.2.4 Decision Trees



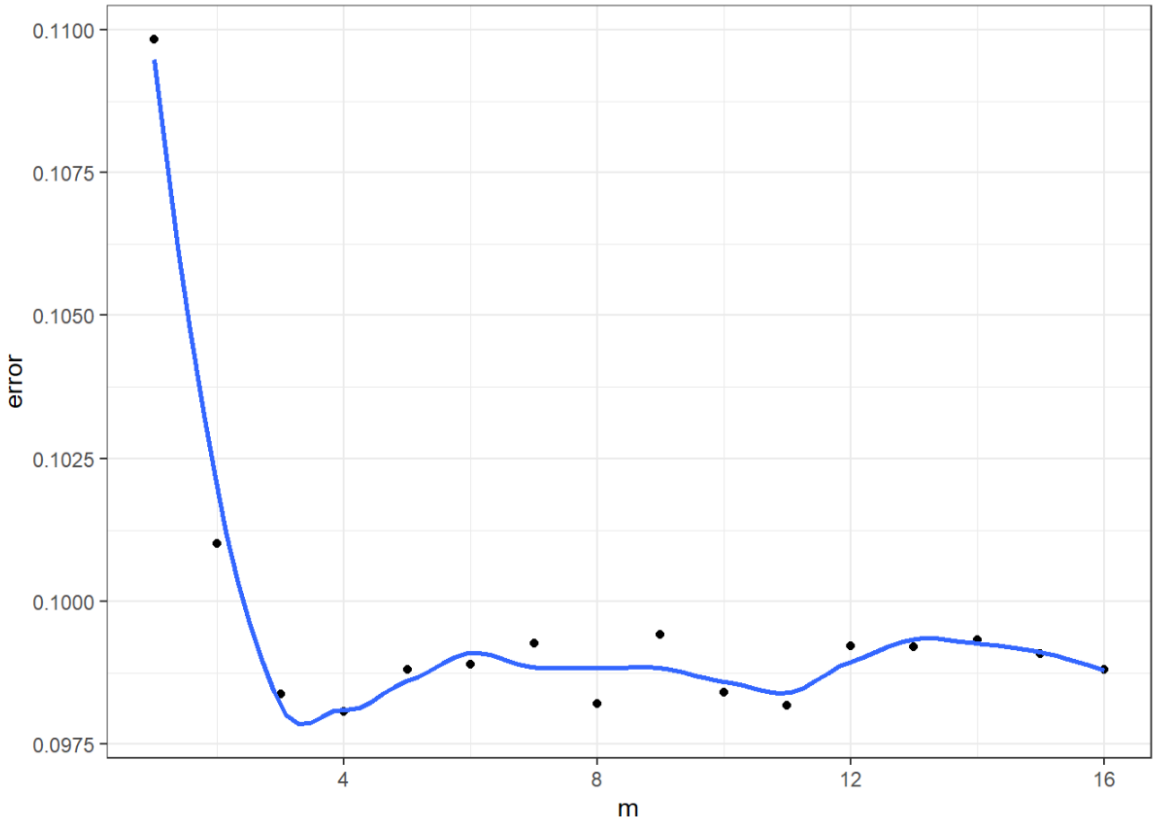
The first decision the tree makes is based on whether the duration of a certain parameter is less than 622. If yes, it goes to the left child node; if no, it goes to the right child node.

In the first level node, If ‘poutcome’ is ‘failure’, ‘unknown’, or ‘other’, it moves to the left child node; otherwise, it moves to the right child node. If the ‘duration’ is less than 828, it goes to the left child nod; otherwise, it moves to the right child node.

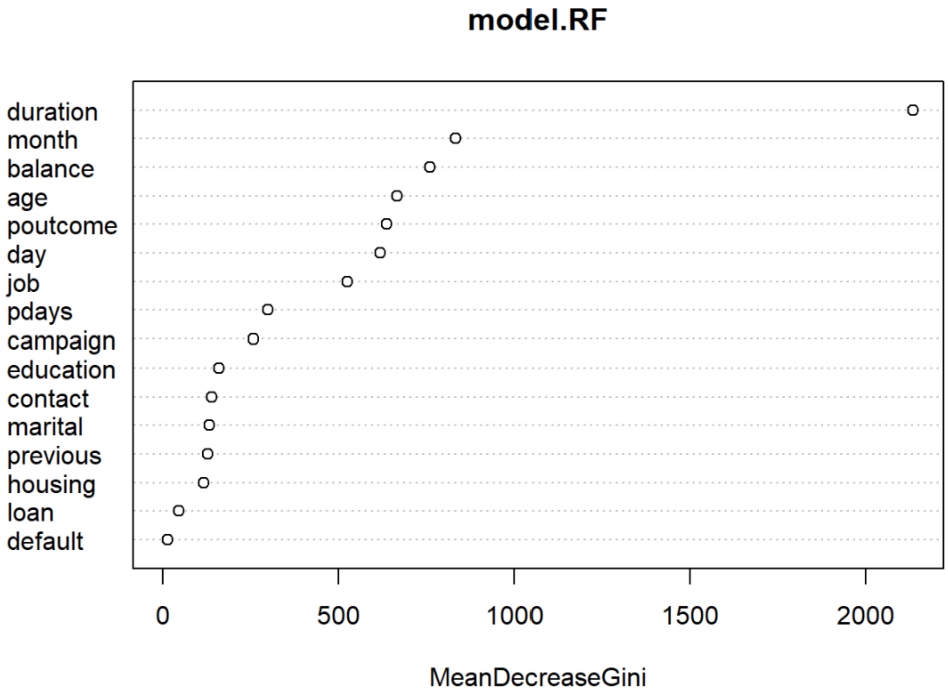
There is some example path to interpretation, if the ‘duration’ is less than 522, ‘poutcome’ is “success” at the same time, and ‘duration’ is greater than 163, this customer may have 2% to subscribe.

In this decision tree, we know that ‘duration’, ‘poutcome’, ‘contact’, and ‘month’ are the important parameters to affect whether the customer subscribes or not.

6.2.5 Random Forest



Minimum error = 4 to find the best



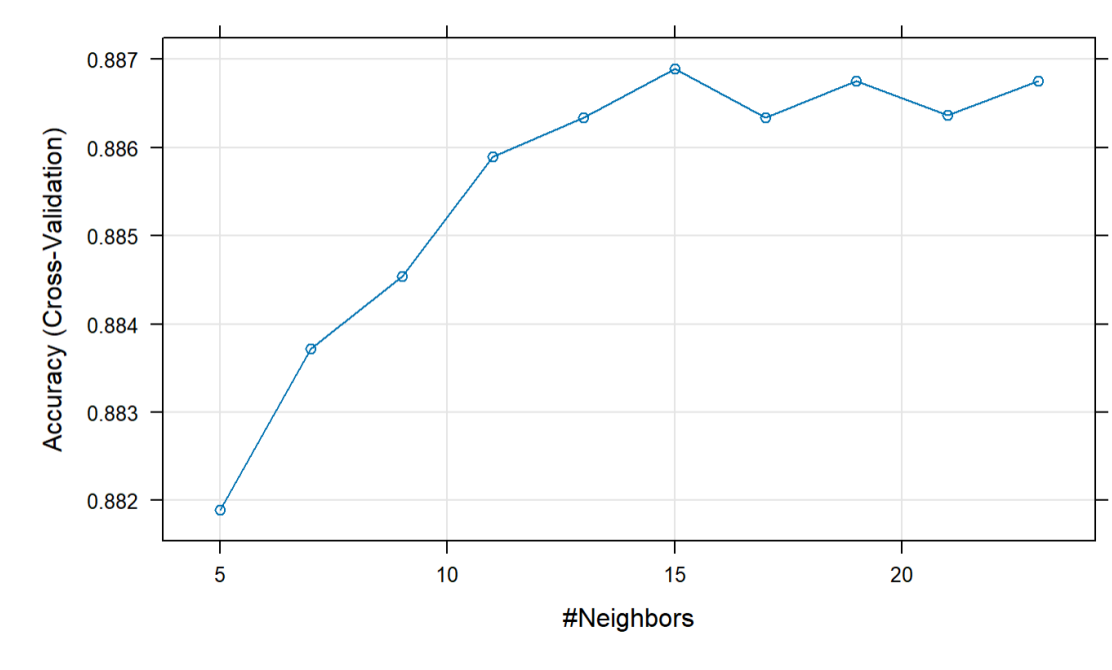
It shows the importance of each predictor variable in the model, measured by the Mean Decrease in the Gini index. Higher values of 'MeanDecreaseGini' indicate that the variable is more important for the classification task. The 'duration' of the call is the most important variable for predicting the target variable, whether a customer will subscribe.

The variable importance plot indicates which features the Random Forest model finds most relevant for making predictions. In this case, ‘duration’, ‘month’, and ‘balance’ are the top three important features.

6.2.6 KNN

```
set.seed(208)
knn_model_cv <- train(y ~ ., data = train_data, method = "knn", trControl = train_control,
tuneLength = 10)
print(knn_model_cv)
```

```
k-Nearest Neighbors

36170 samples
   16 predictor
    2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 32552, 32553, 32554, 32553, 32553, 32553, ...
Resampling results across tuning parameters:

  k   Accuracy   Kappa
   5  0.8818910  0.2944588
   7  0.8837157  0.2864649
   9  0.8845452  0.2791848
  11  0.8858998  0.2759073
  13  0.8863423  0.2762661
  15  0.8868954  0.2704266
  17  0.8863423  0.2602685
  19  0.8867569  0.2581967
  21  0.8863698  0.2505298
  23  0.8867569  0.2484841

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 15.
```



Accuracy increases as k increases from 5 to 15. After k = 15, the accuracy fluctuates slightly but remains relatively stable. The optimal K-NN model is achieved with k = 15, providing the highest cross-validation accuracy of approximately 88.69%.

6.3 Model Evaluation

Model	F1 score	Accuracy
Logistic Regression	0.46	90.24%
Tune SVM Polynomial	**0.50**	**91.10%**
Decision Tree	0.48	90.41%
Random Forest	0.56	90.63%
KNN	0.34	88.55%

The SVM Polynomial Tune model is the best choice given its superior performance in accuracy with a relatively high F1 Score. This model balances precision and recall effectively while also achieving the highest overall accuracy, making it a reliable and robust model for your classification task.