Cost Reduction & Process Optimization in Retail Banking

Nikhil Sharma

August 2019

Agenda

01 Business Understanding & Impact

02 Process Flow

03 Data Overview

04 Exploratory Analysis

05 Data Cleaning, Preprocessing & Feature Engineering

06 Data Partitioning & Class Balancing

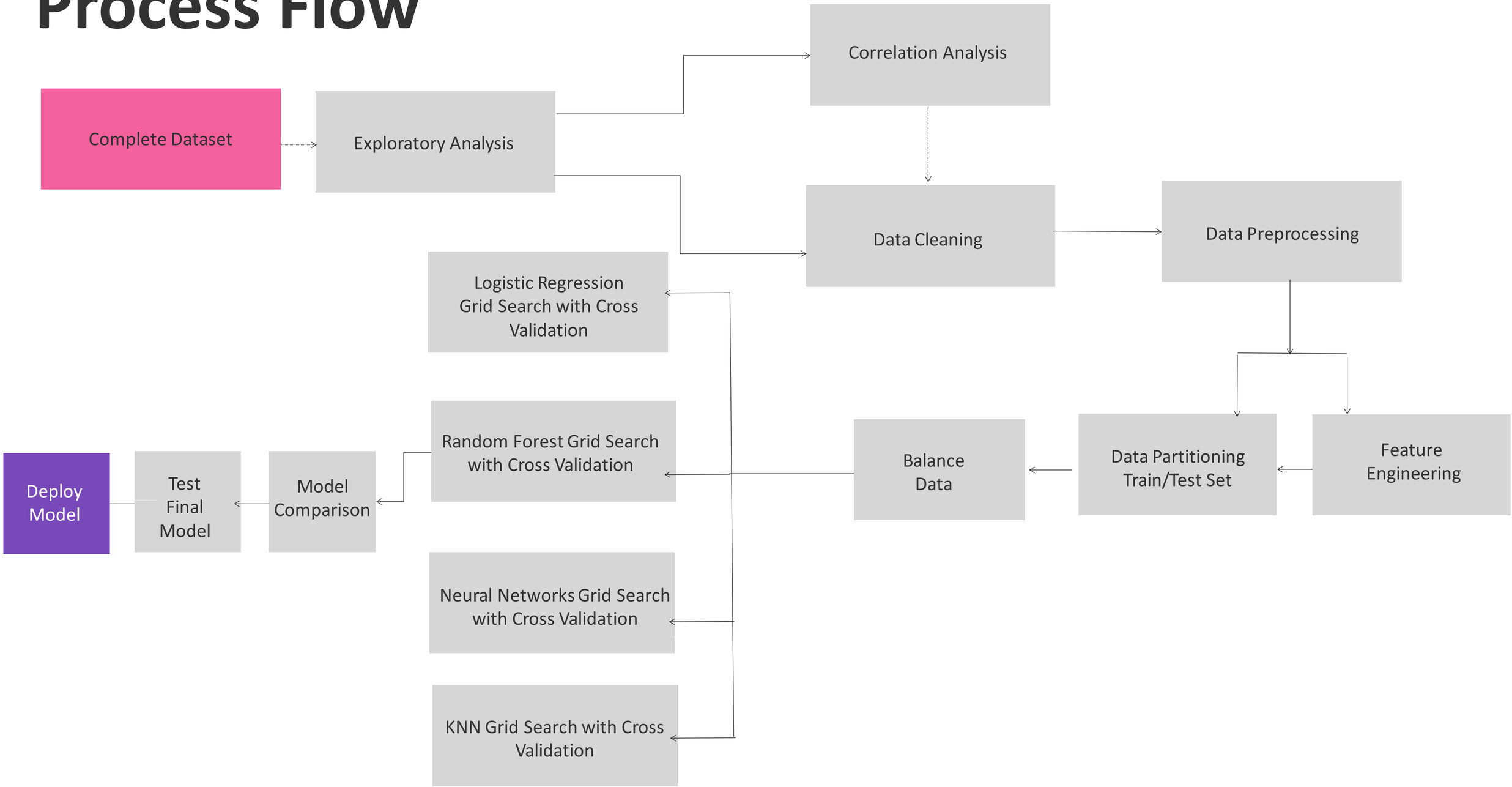
07 Model Development & Results

08 Recommendations

Business Understanding & Impact

- A Bank in Europe is looking to optimize its sales & telemarketing strategy by addressing their revenue decline, which is linked to a decrease in the number of term deposit subscriptions.
- In this project we will build a predictive model for imbalanced data classification that identifies customers who have a higher probability of subscribing to a term deposit.

Process Flow



Data Overview

Customer Demographics

- Age
- Job
- Marital status
- Education level

Personal Banking

- Credit in default.
- Average account balance.
- Housing loan.
- Personal loan.

Current Marketing Campaign

- Contact communication type.
- Last contact day of month.
- Last contact month of year.
- Last contact duration (seconds).

Previous Marketing Campaign

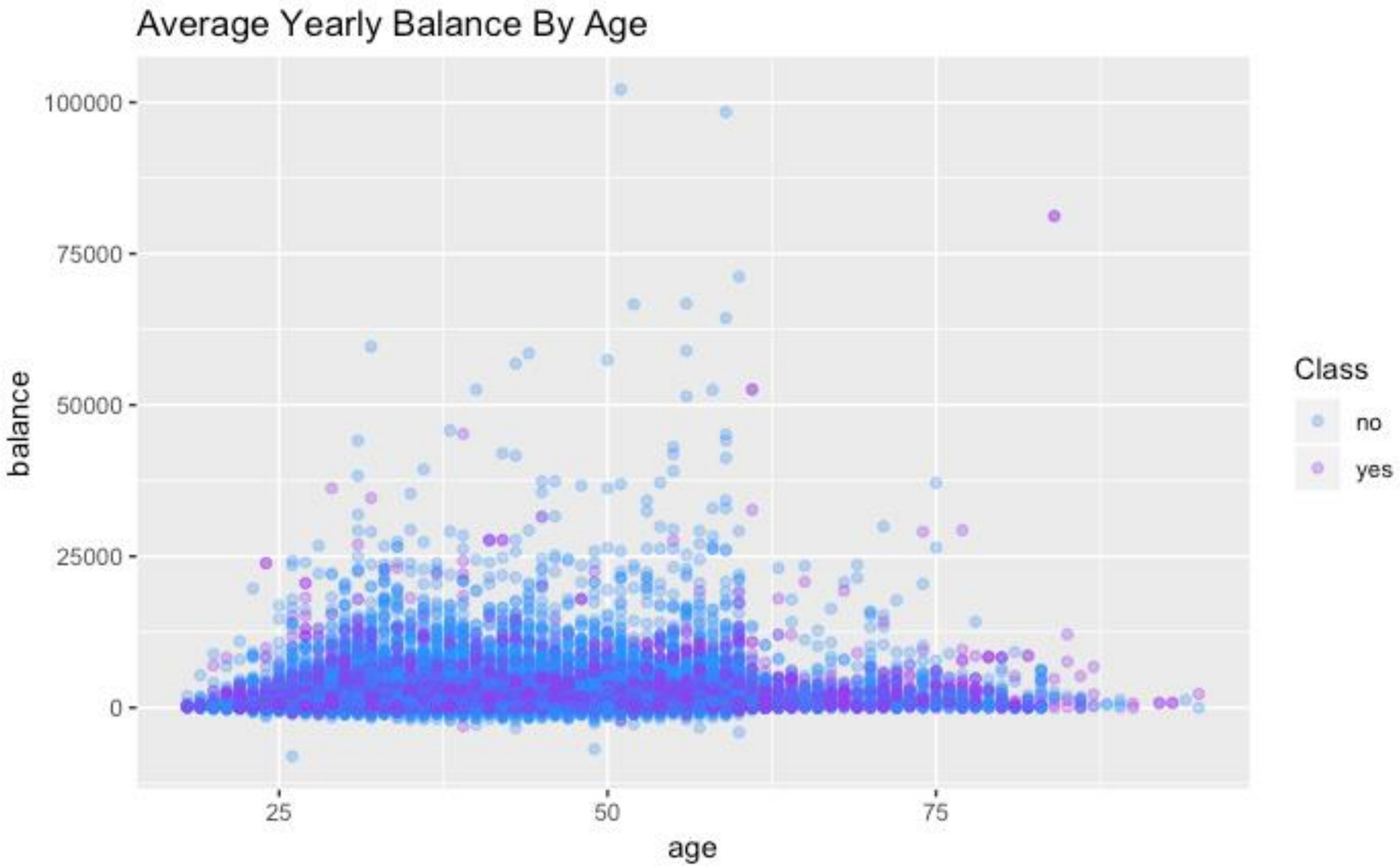
- Number of days after client was last contacted from previous campaign.
- Number of contacts performed before current campaign.
- Outcome of previous marketing campaigns.

Response Variable

- Subscription to a term deposit.

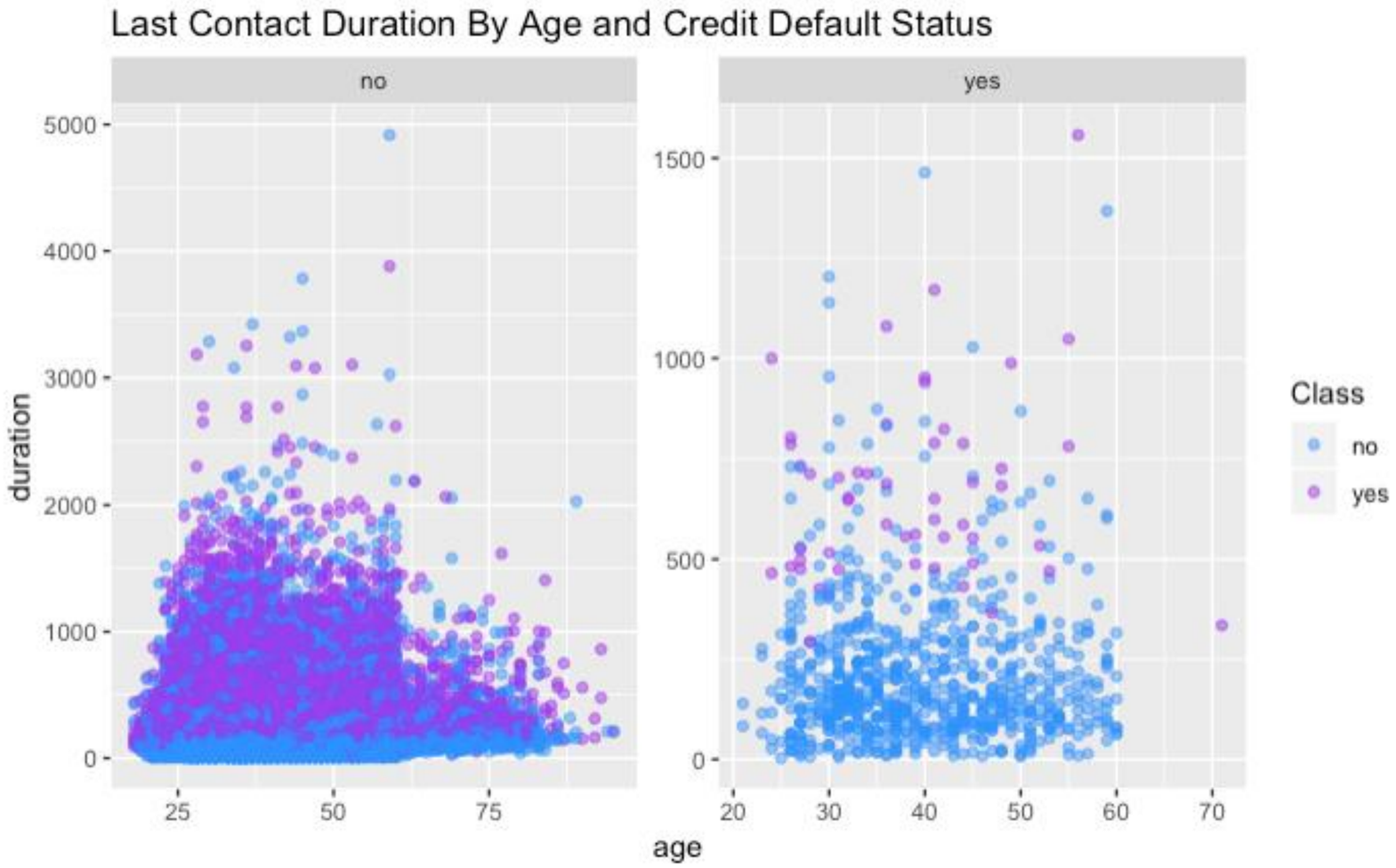
Exploratory Analysis: *Customer Demographics*

Most customers spent less than 33 minutes on the phone with a Financial Advisor regardless of their education or employment level.



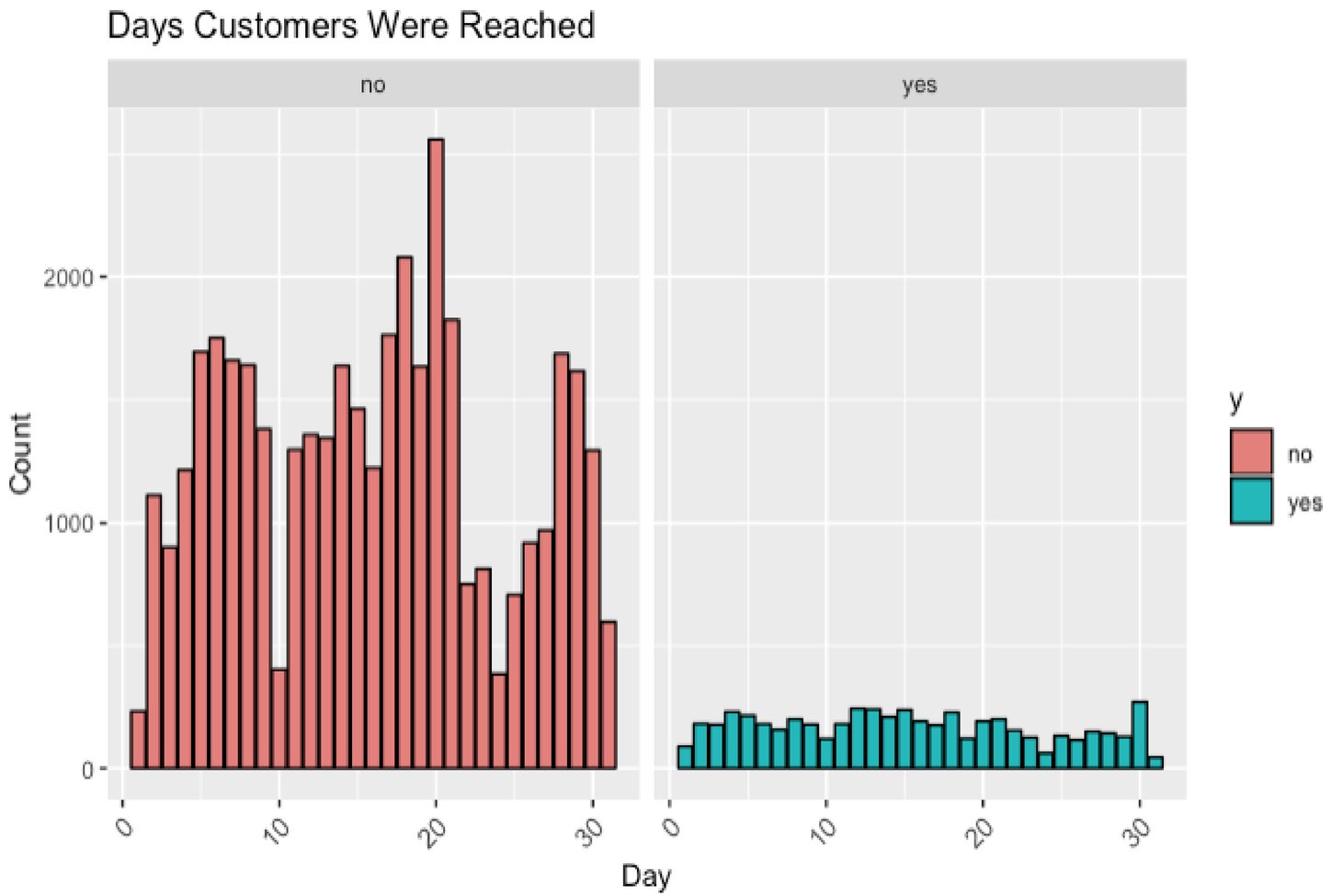
Exploratory Analysis: *Personal Banking*

If a customer had a personal or housing loan, they spent less time on the phone with a Financial Advisor, and if they did have a lengthy conversation, they subscribed to a term deposit.



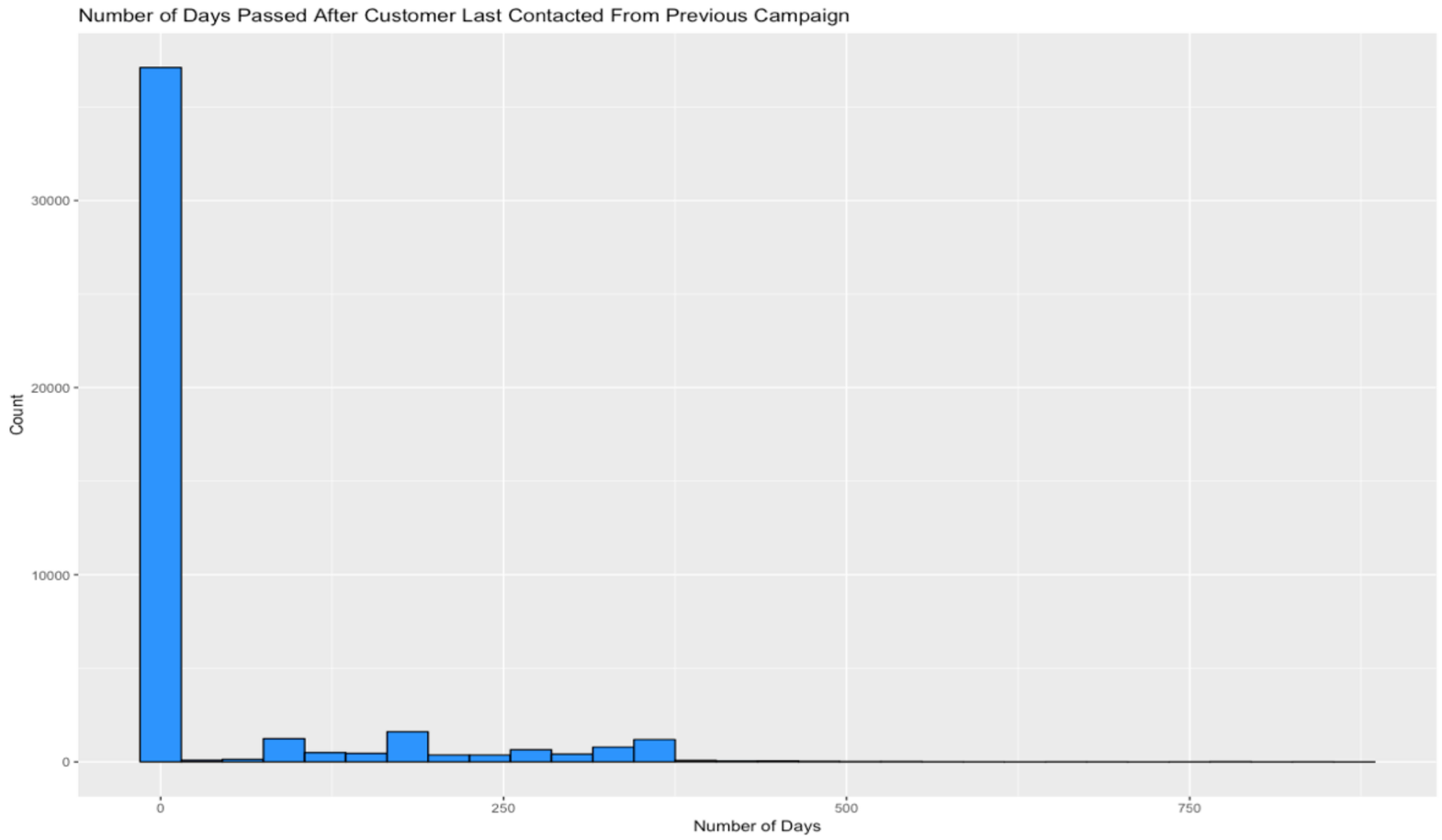
Exploratory Analysis: *Current Marketing Campaign*

Most customers were reachable in the middle of the month or towards the end of the month. 47% of customers that were last contacted by the Bank in December subscribed to a term deposit.



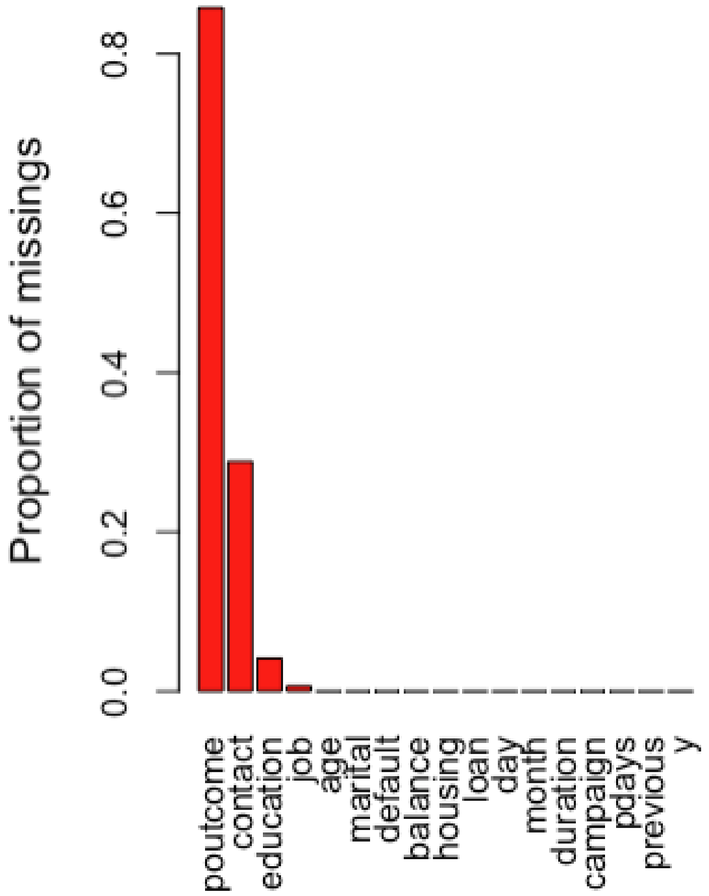
Exploratory Analysis: *Previous Marketing Campaign*

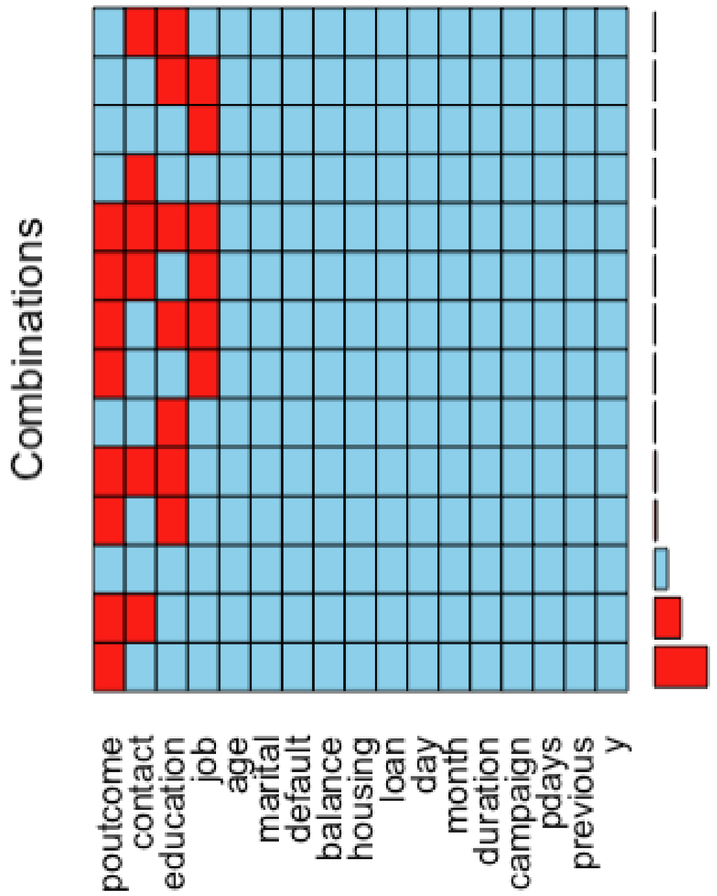
Average number of days the customer was last contacted after the previous campaign was 40 days. 75% of customers were not contacted by the Bank prior to the campaign.



Data Cleaning: *Missing Values & Duplicates*

Multivariate Imputation by Chained Equations (MICE) was used to impute 53,964 missing values. There were zero duplicates in the dataset.

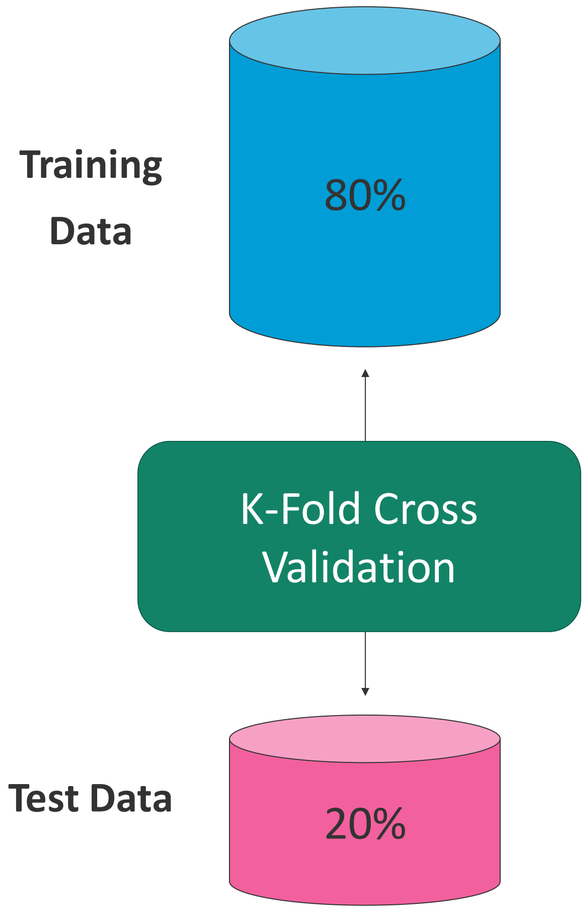




Data Preprocessing & Feature Engineering

- Categorical values for the variable ‘job’ were bucketed into ‘unemployed’ and ‘employed’ to better identify records in a shared a group to enhance algorithmic efficiency.
- Numeric variables were then scaled and centred. Scale transform calculates the standard deviation for each feature and divides each value by that output. Meanwhile, center transform calculates the mean for each feature and subtracts it from each value.
- One-Hot Encoding was used to treat categorical variables, where every unique value is added as a feature in the data set as a binary predictor (0,1).

Data Partitioning



Class Balancing

- The dataset is highly imbalanced with over 80% of customers being non-subscribers. SMOTE was used to balance the training data set. If imbalanced data is trained on the classifier, the results may be biased in favour of the majority class because there weren't enough data points to learn about the population of the minority class. Here were some of the options:
  - **Oversampling:** Minority class is duplicated until the class balance is achieved.
  - **Under-sampling:** Removes the random instances of the majority class. Although the results may be strong, the model suffers from a significant loss of data points and ends up not generalizing well enough.
  - **Synthetic Minority Over-Sampling Technique (SMOTE):** Balances the skewed class distribution of positives and negatives.

Performance Metrics

01

ROC

Used to evaluate how well a model can distinguish between classes when predicting. Higher the AUC (Area Under the Curve) or closer to 1, the stronger the model is at predicting the binary classes correctly.

02

Recall or Sensitivity

Indicates the percentage of correct positive classifications (true positives) from respondents that are positive. In other words how many customers were correctly classified as term deposit subscribers by the model.

03

Precision

Indicates the percentage of correct positive classifications (true positives) from respondents that were predicted as positive. In other words, how many customers were correctly classified as non-subscribers and subscribers, respectively.

Confusion Matrix

	ACTUAL VALUES	
PREDICTED VALUES	**YES**	**NO**
YES	TRUE POSITIVE	FALSE POSITIVE
NO	FALSE NEGATIVE	TRUE NEGATIVE

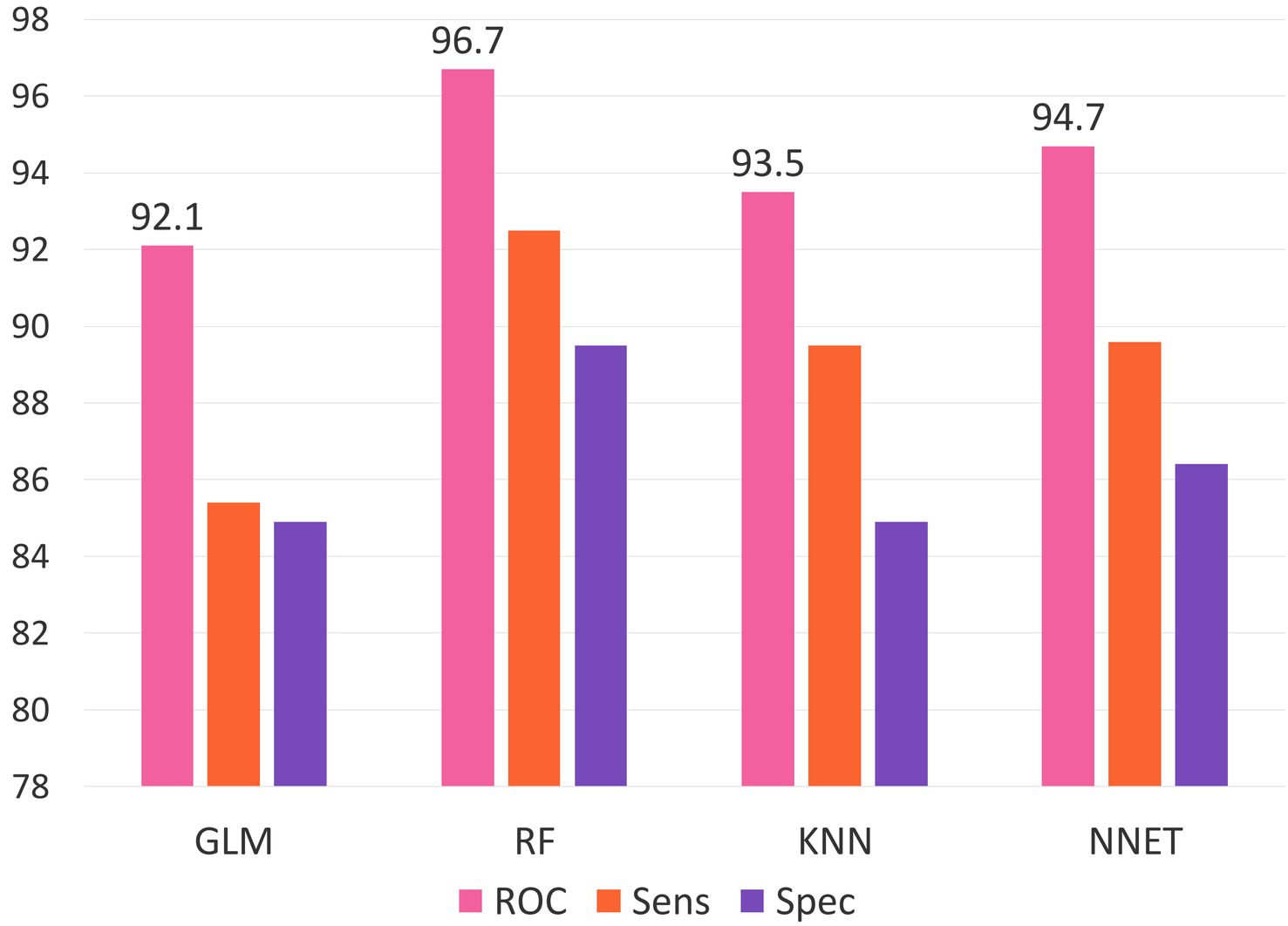
Cross-Validation Evaluation

- We need to be able to answer what is the cost of incorrectly classifying a customer who is a non-subscriber as a subscriber (false positive or Type I Error cost = marketing) versus what is the cost of failing to identify a customer who is a subscriber as a non-subscriber (false negative = revenue).
- When it comes to Bank marketing, false positives are inexpensive compared to false negatives or Type II, which could lead to sales opportunities.

Cross-Validation Results

After 10-fold cross validation Random Forest Model had the highest ROC average over 30 resamples, including the highest sensitivity, also known as true positive rate or recall.

Model Performance



Model Tuning: *Types of Feature Selection Methods*

Wrapper Method

- “Evaluates multiple models using procedures that add and/or remove predictors to find the optimal combination that maximizes model performance,” (Kuhn, 2019).

Filter Method

- “Evaluates the relevance of the predictors outside the predictive models and subsequently model only the predictors that pass some criterion,” (Kuhn, 2019).

Model Tuning: *Recursive Feature Elimination*

False positives are less costly compared to false negatives because the Bank would rather pay the marketing cost it would take to call the occasional customer who chooses not to subscribe rather than missing out on actual sales opportunities.

- **Recursive Feature Elimination** fits a model using all variables and calculates the importance or rank for each variable. The worst performing variables are eliminated at each iteration and the model is retrained to compute the importance for each variable. The feature selection process is repeated until the best subset of predictors are selected based on the highest variable importance with 10-fold cross- validation.
- The original model that included all 28 predictors performed better:
  - Able to better distinguish between classes.
  - High sensitivity means few false negatives (incorrectly predicting subscriber as a non-subscriber).
  - Low specificity means many false positives (incorrectly predicting non-subscriber as a subscriber).

Test Results Performance Metrics

PRECISION	AUC - ROC	RECALL
0.464	0.931	0.840

Model Deployment Evaluation

- Percentage of how strong the model is in distinguishing between classes remains >90%.
- Model is able to achieve a >80% true positive rate which means few false negatives (incorrectly predicting subscriber as a non-subscriber)

High False Positive Rate & Low False Negative Trade-Off

- The model achieves a low precision rate which indicates a high false positive rate i.e. how many customers were correctly classified as non-subscribers and subscribers, respectively.
- However, the model achieved a >80% true positive rate which means few false negatives (incorrectly predicting subscriber as a non-subscriber).
- False positives are less costly compared to false negatives because the Bank would rather pay the marketing cost it would take to call the occasional customer who chooses not to subscribe rather than missing out on sales opportunities.

Recommendations

- While the dataset does not provide the cost of each phone call, the efficiency of the Bank's telemarketing strategy was improved with the assumption that it will result in time and cost reduction of advisors making phone calls to customers who are likely to subscribe to a term deposit.
- The length of call was the most important feature followed by whether the client was known to the bank; the day the call was made, and how much money was in the client's account and the age of the a client.