Analysis of Bank Marketing Dataset

Annie Nguyen

I. Introduction

This report presents an analysis of a dataset derived from direct marketing campaigns conducted by a Portuguese banking institution. The dataset contains information about the marketing campaigns, including client data and the outcomes of the campaigns.

II. Dataset Investigation

1. Background

The dataset used in this analysis is derived from direct marketing campaigns conducted by a Portuguese banking institution. These campaigns primarily relied on phone calls, and often multiple contacts with the same client were necessary to determine whether the client would subscribe to a bank term deposit ('yes') or not ('no').

The dataset contains 41188 examples and consists of 20 input features, ordered by date from May 2008 to November 2010. The classification goal of this dataset is to predict whether a client will subscribe to a term deposit.

The dataset includes both bank client data and information related to the last contact of the marketing campaign. Bank client data includes features such as age, job, marital status, education, credit default status, housing loan status, personal loan status, contact communication type, and more. Other attributes capture the context of the last contact, previous campaign outcomes, and social and economic indicators.

Several categorical attributes contain missing values coded as "unknown."

Citation Information:

Researchers Sérgio Moro, Paulo Cortez, and Paulo Rita created this dataset in 2014. It has been publicly available for research purposes and was analyzed in the paper titled "A Data-Driven Approach to Predict the Success of Bank Telemarketing" published in Decision Support Systems in 2014.

Data Source

Data Research Paper

2. Explanation of Variables

Input variables:
+ 1 - age (numeric)
+ 2 - job : type of job (categorical: "admin.","blue-collar","entrepreneur","housemaid","management","retired", "self-employed","services","student","technician","unemployed","unknown")
+ 3 - marital : marital status (categorical: "divorced","married","single","unknown")

+ 4 - education (categorical: “basic.4y”,“basic.6y”,“basic.9y”,“high.school”,“illiterate”,“professional.course”, “university.degree”,“unknown”)
+ 5 - default: has credit in default? (categorical: “no”,“yes”,“unknown”)
+ 6 - housing: has housing loan? (categorical: “no”,“yes”,“unknown”)
+ 7 - personal: has personal loan? (categorical: “no”,“yes”,“unknown”)
+ 8 - contact: contact communication type (categorical: “cellular”,“telephone”)
+ 9 - month: last contact month of year (categorical: “jan”, “feb”, “mar”, . . . , “nov”, “dec”)
+ 10 - day_of_week: last contact day of the week (categorical: “mon”,“tue”,“wed”,“thu”,“fri”)
+ 11 - duration: last contact duration, in seconds (numeric)
+ 12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
+ 13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
+ 14 - previous: number of contacts performed before this campaign and for this client (numeric)
+ 15 - poutcome: outcome of the previous marketing campaign (categorical: “failure”,“nonexistent”,“success”)
+ 16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)
+ 17 - cons.price.idx: consumer price index - monthly indicator (numeric)
+ 18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)
+ 19 - euribor3m: euribor 3 month rate - daily indicator (numeric)
+ 20 - nr.employed: number of employees - quarterly indicator (numeric)

Output variable:
+ deposit - has the client subscribed a term deposit? (binary: “yes”,“no”)

3. Type transformation and Missing Values

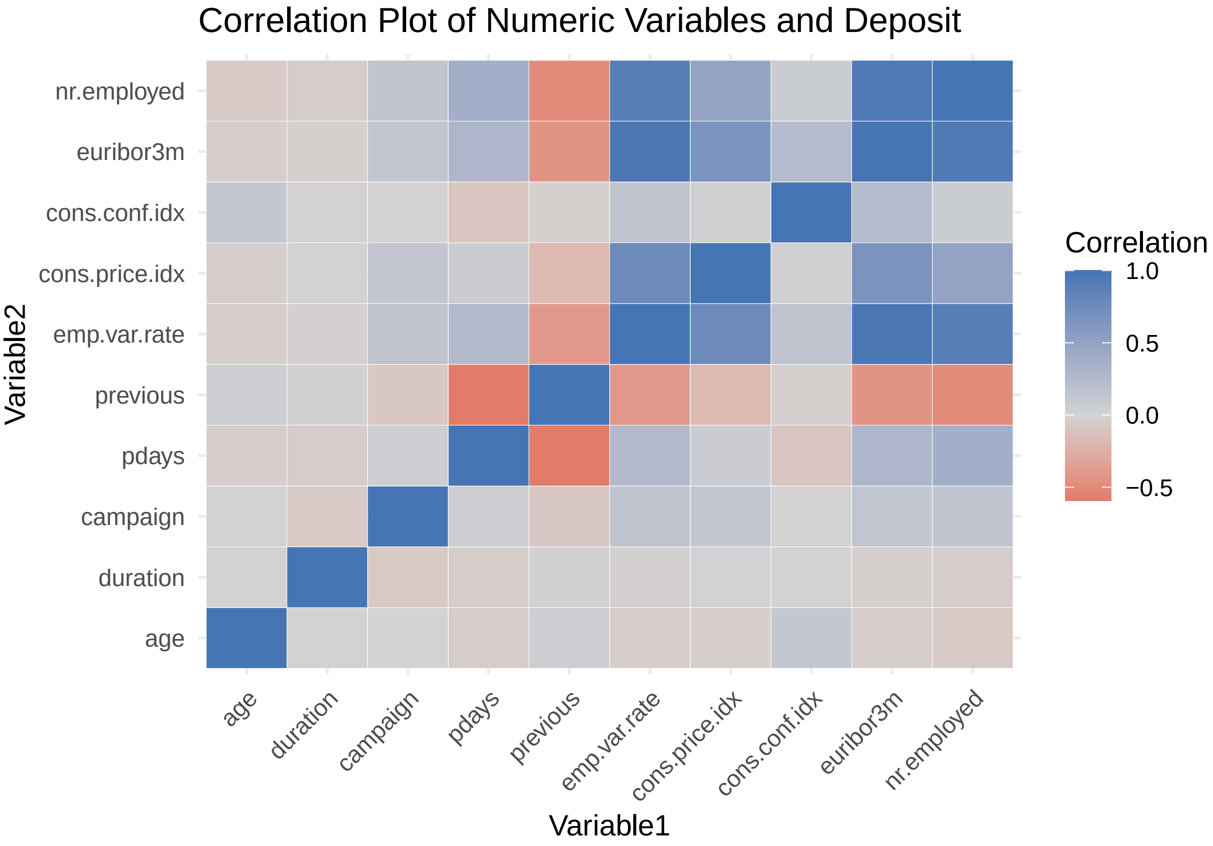
Upon investigating the data, I saw that there are some issues. The most notable one is missing values, which occurs many times in the data set. These values were coded as “Unknown”.

Variable	Percentage
age	0.0000000
job	0.8012042
marital	0.1942313
education	4.2026804
default	20.8725842
housing	2.4036127
personal	2.4036127
contact	0.0000000
duration	0.0000000
campaign	0.0000000
pdays	0.0000000
previous	0.0000000
poutcome	0.0000000
emp.var.rate	0.0000000
cons.price.idx	0.0000000
cons.conf.idx	0.0000000
euribor3m	0.0000000
nr.employed	0.0000000
deposit	0.0000000

The analysis of missing values indicates that the highest missing percentage is observed in the ‘default’ variable, reaching up to 20%. While not overly substantial, these missing values may introduce bias or noise into our analysis. Therefore, for the sake of this project's integrity, I will proceed by removing rows containing "unknown" values from the dataset.

Furthermore, there's an issue with the categorical variables having numerous disorganized levels. To facilitate visualization and analysis, I'll transform these variables into factor levels. This transformation will enhance the organization and clarity of our categorical data, facilitating more effective visualization and interpretation.

I want to further examine if there is correlation between the numeric variables



The correlation plot reveals strong positive correlations between several pairs of variables:

- "nr.employed" and "euribor3m" exhibit a high positive correlation.
- "nr.employed" and "emp.var.rate" are also highly positively correlated.
- Similarly, "euribor3m" and "emp.var.rate" demonstrate a strong positive correlation.

Consequently, to mitigate the risk of overfitting in subsequent modeling processes, it is prudent to remove "nr.employed" and "emp.var.rate" from the dataset. By eliminating these highly correlated variables, I can enhance model interpretability and generalizability, ensuring more robust and reliable predictions.

It is also stated in the dataset file that 'duration' highly affects the output target (e.g., if duration=0 then y="no"). Thus, I would drop it in order to have a realistic predictive model.

In this section, I'm examining the relationships between pairs of categorical variables within our dataset by employing Cramér's V statistic. This would give valuable insights into the associations between different categorical variables. This analysis helps to better understand how these variables relate to each other, providing crucial information for predictive modeling and decision-making processes.

Table 2: Cramér's V values

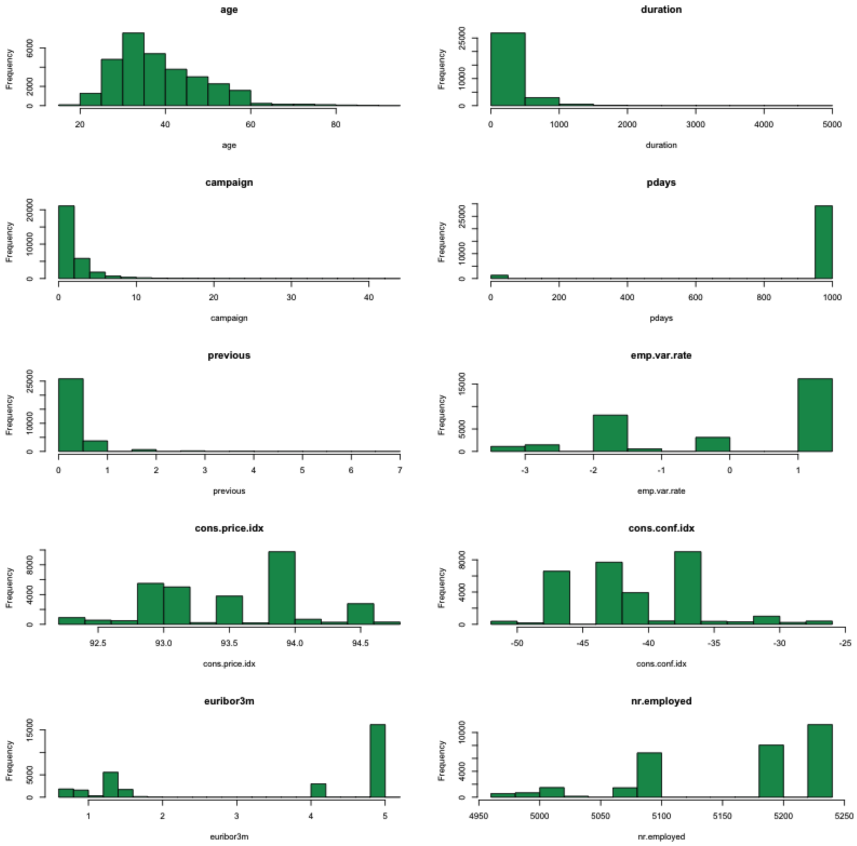
	job	marital	education	default	housing	personal	contact	poutcome	deposit
job	NA	0.20779	0.38540	0.02443	0.02202	0.02796	0.10805	0.09720	0.15476
marital	0.20779	NA	0.13615	0.00855	0.00985	0.00497	0.06917	0.02570	0.04224
education	0.38540	0.13615	NA	0.01592	0.02533	0.01346	0.10183	0.03760	0.06155
default	0.02443	0.00855	0.01592	NA	0.00415	0.00427	0.00695	0.00704	0.00378
housing	0.02202	0.00985	0.02533	0.00415	NA	0.04699	0.08057	0.02517	0.01007
personal	0.02796	0.00497	0.01346	0.00427	0.04699	NA	0.00902	0.00461	0.00503
contact	0.10805	0.06917	0.10183	0.00695	0.08057	0.00902	NA	0.23108	0.14386
poutcome	0.09720	0.02570	0.03760	0.00704	0.02517	0.00461	0.23108	NA	0.32302
deposit	0.15476	0.04224	0.06155	0.00378	0.01007	0.00503	0.14386	0.32302	NA

The Cramer matrix also indicates that there is little to no correlation between the categorical variables. Thus, no further action is needed.

4. Variables Investigation

Numeric variables

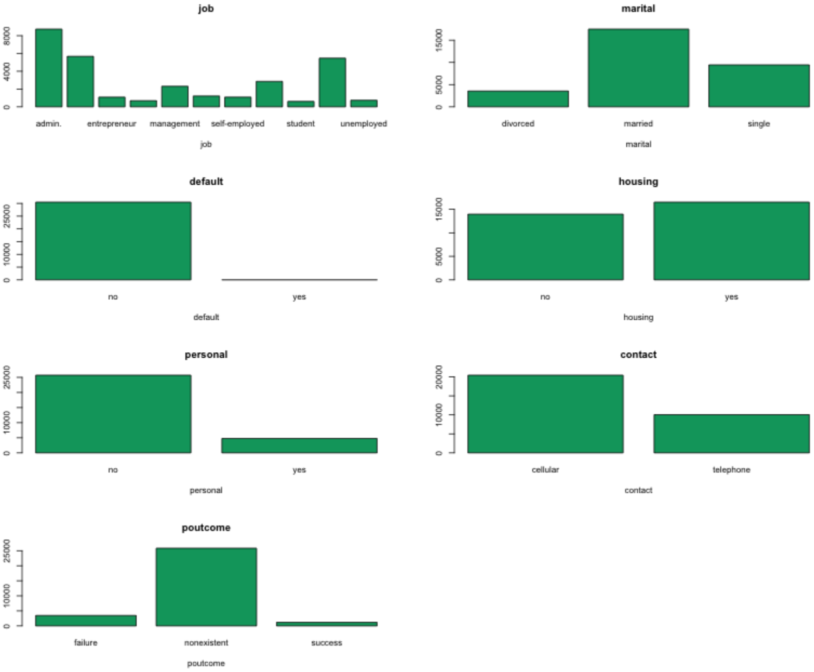
I first investigate the numeric variables by plotting a histogram distribution for each of them.



From this graph, Age is the only variable that might have a normal distribution.

Categorical variables

Then I want to look at the categorical variables

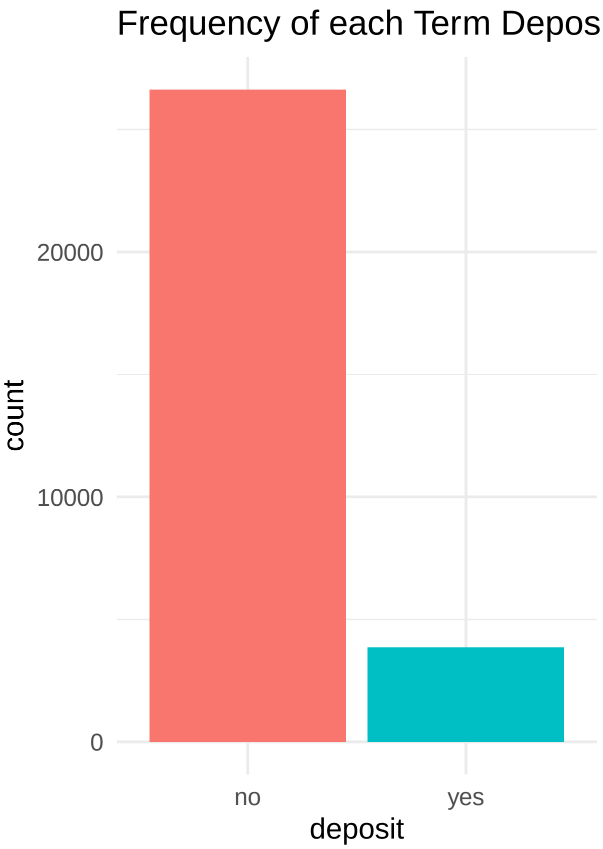


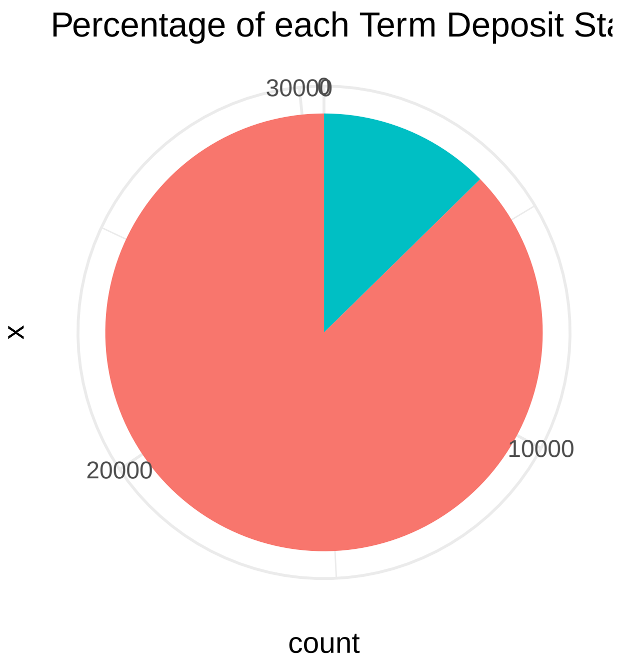
There is a moderate balance betweens groups in ‘housing’, a somewhat unbalance between groups in ‘contact’ and ‘marital’. However, groups in ‘default’ and ‘personal’ are heavily concentrated for ‘no’ in comparison to ‘yes’. In the case of ‘default’, since there is only 3 ‘yes’ observations, we will proceed to drop it to prevent zero variances. In the case of ‘personal’, it is not too concerning, but worth considering for model interpretation, since it indicates that the majority of the population doesn’t have neither credit in default nor personal loan. It is also worth noted that the population of people who work in ‘admin’ field is the biggest out of all jobs.

Due to the considerable number of unknown values within the poutcome variable, I will omit it from the analysis. The presence of these unknown entries may introduce bias or noise, potentially undermining the reliability of our findings. Therefore, to ensure the integrity and robustness of our analysis, poutcome was excluded from further consideration.

Target variable ‘deposit’

The provided visualizations, including a bar plot and a pie chart, offer a clear depiction of the distribution of the target variable ‘deposit.’ These graphical representations succinctly illustrate the proportion of positive (‘yes’) and negative (‘no’) outcomes within the dataset, providing valuable insights into the balance and prevalence of term deposit subscriptions.





The examination of the target variable “deposit” uncovered an inherent imbalance within the dataset, characterized by a notable disparity in the frequency of outcomes. Specifically, there exists a higher prevalence of negative outcomes, signifying instances where clients did not subscribe to a term deposit, in contrast to positive outcomes indicating successful term deposit subscriptions. This imbalance underscores the need for careful consideration and appropriate handling to mitigate its potential impact on modeling and analysis outcomes.

5. Resampling

Minority case, which is ‘yes’, must be have the factor level of 1.

With approximately 26,000 observations labeled as ‘no’ and only 3,000 labeled as ‘yes’, this class imbalance posed potential obstacles to both model training efficiency and predictive performance.

After careful consideration and weighing the available options, I opted to employ downsampling as our preferred method for addressing the class imbalance. Downsampling involves reducing the size of the majority class (‘no’ in our case) to match the size of the minority class (‘yes’). This decision was made based on several key factors and considerations:

- Computational Efficiency: Given the large size of the dataset, downsampling offers a practical solution to reduce computational complexity and training time. By decreasing the number of observations in the majority class, we can streamline the training process and alleviate potential computational burdens associated with handling a heavily imbalanced dataset.
- Improved Model Performance: Downsampling ensures a more balanced distribution of classes in the training data, which can help mitigate biases and enhance the model’s ability to generalize to unseen data. By providing a more equitable representation of both classes, downsampling contributes to more effective learning of underlying patterns and relationships within the data, ultimately leading to better predictive performance.

- Prevention of Overfitting: Downsampling reduces the risk of overfitting, a common concern in imbalanced datasets where models may exhibit a tendency to prioritize the majority class due to its prevalence. By creating a more balanced training set, downsampling helps prevent the model from being overly influenced by the dominant class, leading to more reliable and robust predictions on new data.

Below is the table of 'deposit' values after downsampling.

Var1	Freq
0	3859
1	3859

III. Models

To train models, a random seed (05102023) is set to ensure reproducibility of results. Subsequently, a random sample comprising 80% of the downscaled bank data is selected for training purposes, while the remaining 20% constitutes the test dataset.

Training models

In most models below, the predictors are preprocessed through centering and scaling to have mean 0 and standard deviation 1, to enhance model performance and convergence. Cross-validation with 10 folds is employed for robust model evaluation, ensuring reliable estimates of predictive performance and mitigating the risk of overfitting.

1. Logistic Regression Model

The logistic regression model predicts customer subscription to term deposits by analyzing various factors. It's a fundamental tool known for its interpretability and efficiency, making it ideal for understanding customer behavior in banking marketing campaigns.

```r
set.seed(05102024)

# Logistic Regression
log_model <- train(deposit_numeric ~ .,
                   data=train,
                   method="glm",
                   family=binomial(link=logit),
                   preProcess=c("center", "scale"),
                   trControl = trainControl(method="cv", 10))
```

2. K-Nearest Neighbors (KNN) model

The K-Nearest Neighbors (KNN) model is a versatile approach for predicting customer subscription to term deposits. Unlike parametric methods, KNN makes predictions based on the similarity of instances in the feature space. This non-parametric approach is particularly useful when the underlying data distribution is complex or unknown. In this analysis, the KNN model is employed to leverage the inherent patterns within the dataset to make accurate predictions about customer behavior in banking marketing campaigns.

```r
# K-Nearest Neighbors (KNN)
set.seed(05102024)

knn_model <- train(deposit_numeric ~ .,
  data = train,
  method = "knn",
  preProcess = c("center", "scale"),
  trControl = trainControl(method = "cv", 10))
```

3. Random Forest

The Random Forest (RF) model, a powerful ensemble learning technique, is utilized in this analysis to predict customer subscription to term deposits. RF combines the strengths of multiple decision trees to improve predictive accuracy and robustness. By leveraging the collective wisdom of individual trees, RF mitigates overfitting and enhances generalization performance. The RF model below utilizes the random forest algorithm with out-of-bag (OOB) samples for validation. Tuning is performed on the number of variables randomly sampled (mtry) and the number of trees in the forest (ntree) = 50.

```r
set.seed(05102024)

# Random Forest
rf_model <- train(deposit_numeric ~ .,
                   data=train,
                   method="rf",
                   trControl=trainControl("oob"),
                   tuneGrid=data.frame(mtry=10:18),
                   ntree=50)
```

4. Neural Network

The Neural Network (NN) model, a powerful algorithm in predictive modeling, offers a sophisticated approach to forecasting customer subscription outcomes in term deposit campaigns. This model employs a complex network of interconnected neurons to capture intricate patterns and relationships within the data. . To optimize its architecture and performance, the model undergoes tuning using a grid search approach, exploring various combinations of network size and decay parameters. Additionally, the final value used for the model was decay = 0.06.

```r
set.seed(05102024)

# Neural Network
nn_model <- train(deposit_numeric ~ .,
                    data = train,
                    method = "nnet",
                    trControl = trainControl("cv", 10),
                    tuneGrid = expand.grid(size = 1:6, decay = c(0, 0.06)),
                    trace = FALSE)
```

5. Gradient Boosting Machine

The Gradient Boosting Machine (GBM) model, offers a robust solution for predicting customer subscription outcomes in term deposit campaigns. By iteratively optimizing a sequence of decision trees, the GBM model enhances predictive accuracy and robustness. Preprocessing steps ensure standardized input features, facilitating efficient model training. Leveraging its ability to capture complex interactions, the GBM model drives strategic decision-making for banking institutions. Additionally, the tuning parameters used for the model was .n.trees=500, .shrinkage=0.01, .interaction.depth=1, .n.minobsinnode=10.

```r
set.seed(05102024)

# Gradient Boosting Machine
gbm_model <- train(deposit_numeric ~ .,
                   data = train,
                   method = "gbm",
                   preProcess = c("center", "scale"),
                   trControl = trainControl(method = "cv", number = 10),
                   verbose = F,
                   tuneGrid=data.frame(.n.trees=500, .shrinkage=0.01,
                                       .interaction.depth=1, .n.minobsinnode=10))
```

Model evaluation

After constructing our predictive models, we proceeded to assess their performance using a variety of metrics. Our primary focus was on the Misclassification Rate (MCR), which indicates the proportion of incorrect predictions regarding customer deposit behavior. Accuracy, which complements MCR, provides an overall measure of correct predictions.

Alongside MCR and Accuracy, we examined Precision, Recall, and the F1 Score. Precision measures the accuracy of positive predictions by calculating the ratio of true positives to all predicted positives. Recall evaluates the model's ability to identify actual positives among all positive instances. The F1 Score, a combination of Precision and Recall, offers a balanced assessment, particularly valuable when Precision and Recall carry differing degrees of importance.

In our evaluation, we considered all five metrics to determine the best-performing model. In the context of predicting bank deposits, overlooking genuine deposit instances (false negatives) can have significant repercussions, potentially impacting financial outcomes. Therefore, prioritizing recall becomes crucial as it emphasizes capturing as many true deposit instances as possible, even at the cost of misclassifying some non-deposit instances.

Conversely, false positives, where instances are incorrectly predicted as deposits, also carry adverse consequences, including unnecessary expenses and resource allocation. In scenarios where the costs associated with false positives are substantial, precision becomes more critical. Precision underscores the importance of accurately identifying only those instances truly representing deposits, thereby mitigating the adverse effects of misclassification.

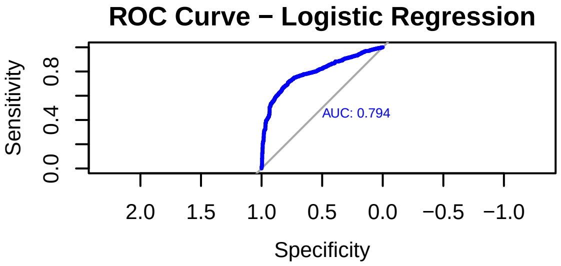
Given our focus on predicting bank deposits, finding a balance between false negatives and false positives was essential. While Recall prioritizes capturing all genuine positives, Precision emphasizes minimizing false positives. The F1 Score serves as a comprehensive measure, guiding our selection of the most suitable model by considering the trade-offs between Precision and Recall.

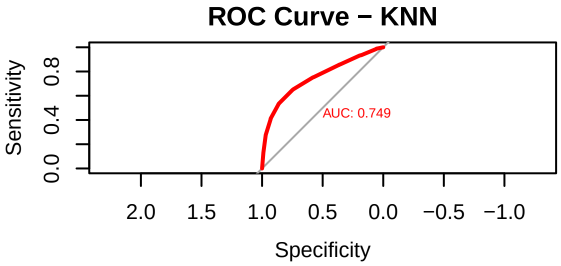
Below is the table for the models statistics.

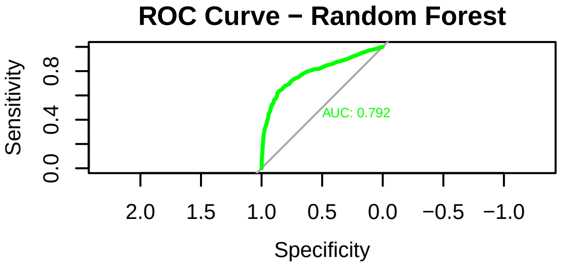
Table 4: Model Statistics - Performance

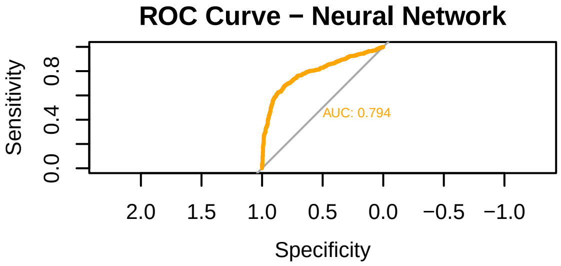
Model	Accuracy	MCR	Precision	Recall	F1
Logistic Regression	0.7415803	0.2584197	0.7345361	0.7470511	0.7407407
KNN	0.6975389	0.3024611	0.6481959	0.7216643	0.6829599
Random Forest	0.7344560	0.2655440	0.7010309	0.7534626	0.7263017

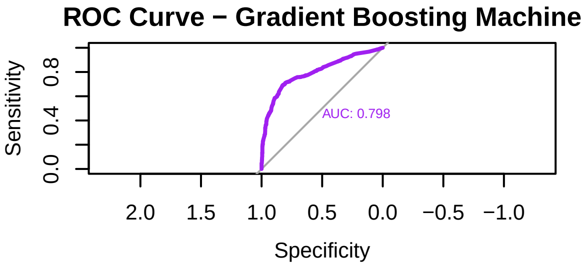
Model	Accuracy	MCR	Precision	Recall	F1
Neural Network	0.7415803	0.2584197	0.6585052	0.7922481	0.7192118
Gradient Boosting Machine	0.7461140	0.2538860	0.7268041	0.7580645	0.7421053


ROC Curve − Logistic Regression


ROC Curve − KNN


ROC Curve − Random Forest


ROC Curve − Neural Network


ROC Curve − Gradient Boosting Machine

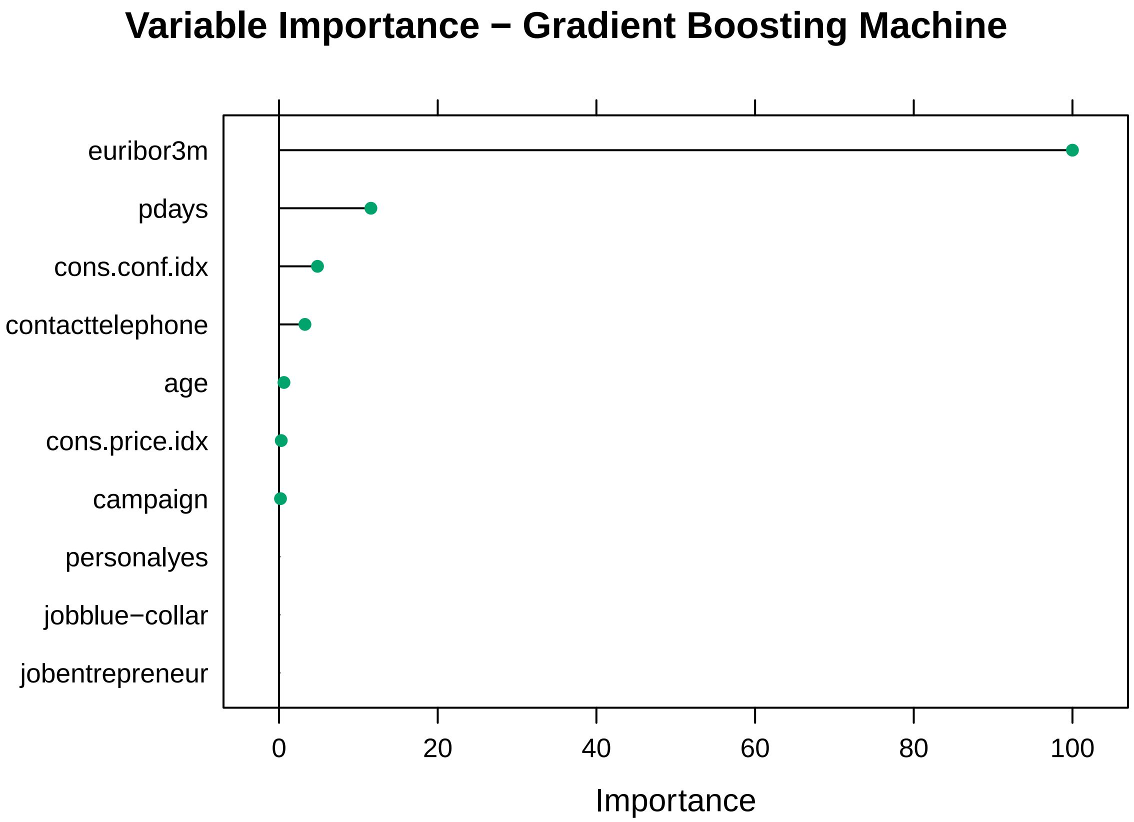
Considering both performance metrics and computational efficiency, the Gradient Boosting Machine (GBM) and Logistic Regression models emerge as top contenders for predicting term deposit subscriptions in our dataset.

While the Neural Network model demonstrates competitive performance metrics, its computational demands may limit its practical applicability, especially in resource-constrained environments. KNN model also offer viable alternatives, albeit with slightly lower accuracy compared to the ensemble methods.

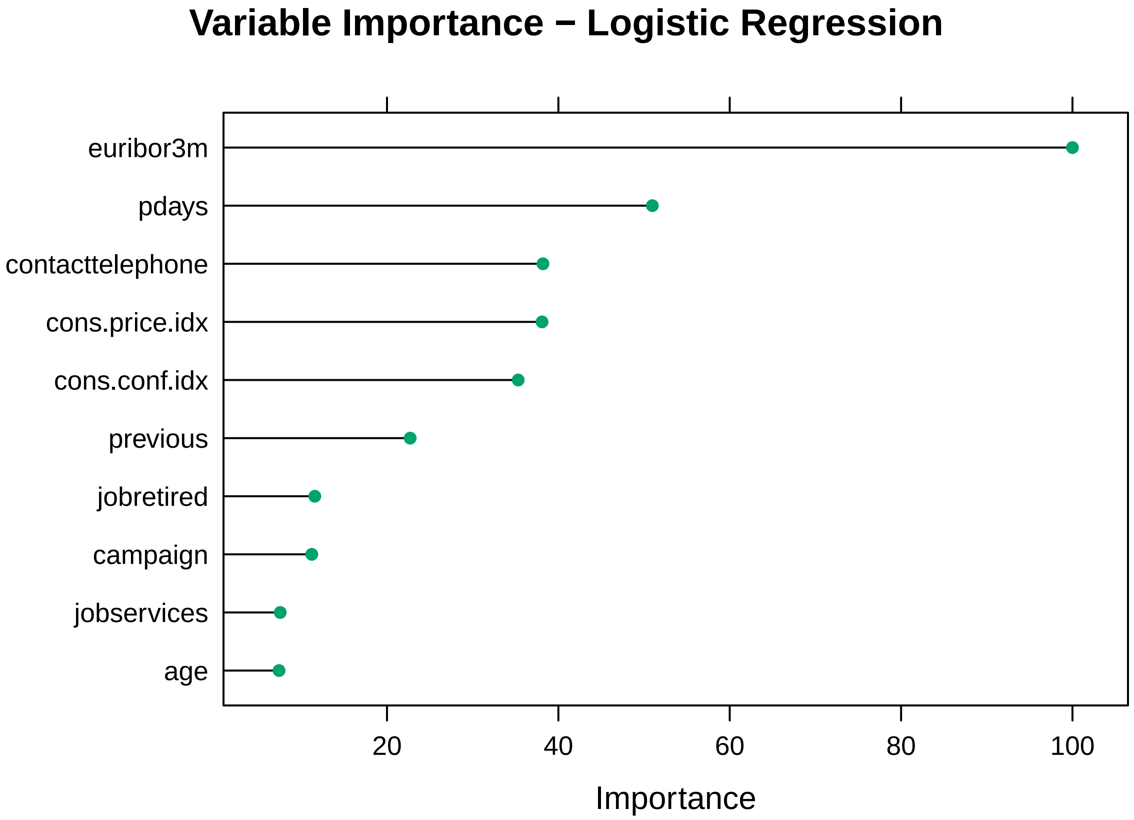
GBM and Logistic emerges as the top performers across all metrics, boasting the highest accuracy, precision, recall, and F1 score among the models considered. With an accuracy of 74.61% and a misclassification rate of 25.39%, GBM demonstrates robust predictive capabilities. Logistic Regression also delivers competitive results, achieving an accuracy of 74.15% and demonstrating balanced precision and recall. Their superior precision underscores their abilities to minimize false positive predictions, crucial in financial decision-making contexts.

Variable Evaluation

Examining the 15 variables used for training and testing is pivotal to discerning the most effective predictors for determining customer subscription to a term deposit. These key features offer valuable insights into the factors that drive the model's predictions.



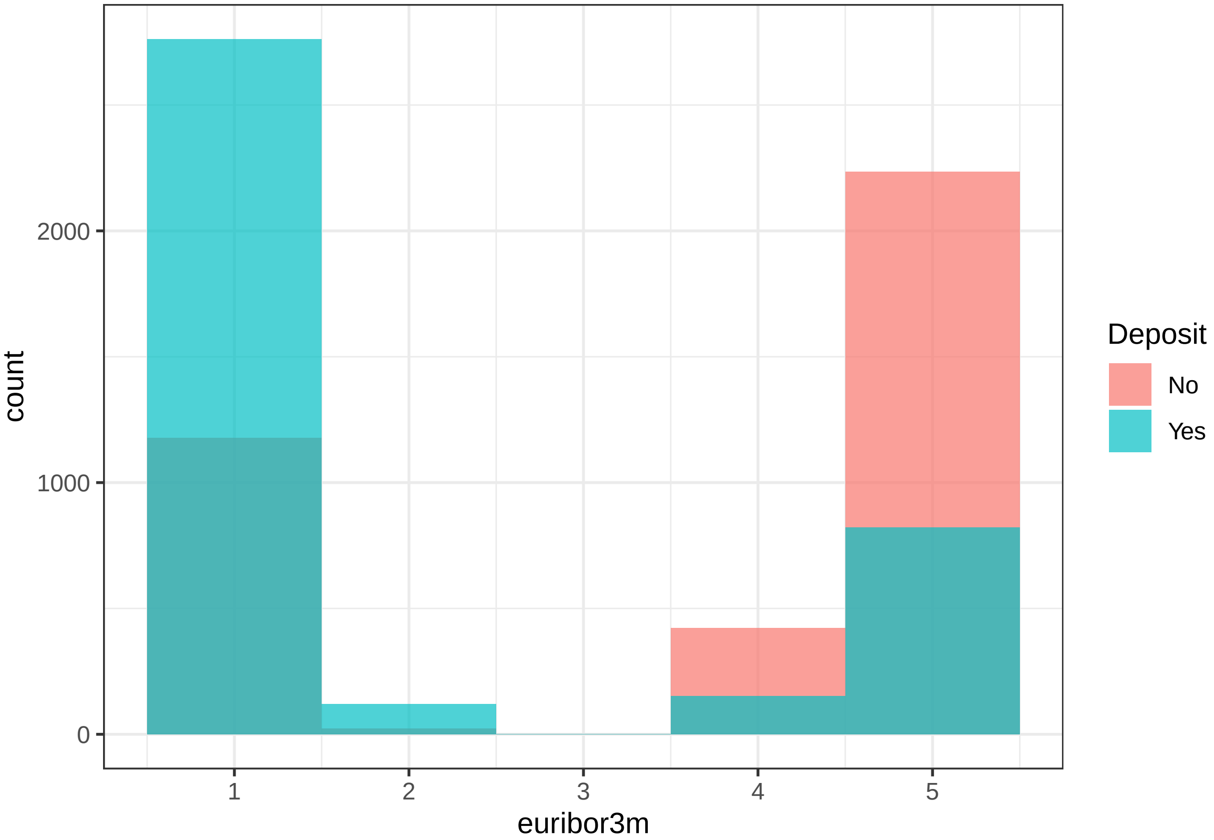
Variable Importance − Gradient Boosting Machine



Variable Importance − Logistic Regression

From the plots above, it is shown that the logistic model has more balance between variables importance that that of the GBM model. Additionally, the logistic model is also the more computationally efficient out of the two models, with little trade-offs.

The plots also reveal that 2 variables, namely euribor3m, and pdays are among the most influential factors. To gain deeper insights, the distribution of term deposit subscriptions is visualized across euribor3m.



The plot shows that observations concentrated on euribor3m = 1 and euribor3m = 5.

IV. Conclusion

The analysis revealed that Logistic Regression emerged as the preeminent model, displaying the highest predictive accuracy while maintaining a balanced precision and recall. This underscores the effectiveness of ensemble methods in discerning intricate patterns within intricate datasets, surpassing singular models like Random Forest or KNN.

Additionally, the variable importance analysis provided valuable insights into the significant predictors influencing term deposit subscriptions. Economic indicator euribor3m was identified as key influencer of subscription outcomes. This highlights the critical role of economic context and temporal dynamics in shaping customer decision-making processes.

Moreover, the acknowledgment of class imbalance within the 'deposit' target variable underscores the necessity of employing appropriate resampling techniques during model training. Techniques like downsampling can address this imbalance, thereby reducing skewed predictions and improving generalization performance.

These findings offer strategic direction for banking institutions aiming to refine their marketing strategies. By leveraging predictive analytics, institutions can tailor marketing campaigns to target specific customer segments more effectively, optimize resource allocation, and maximize return on investment.

The demonstrated efficacy of machine learning techniques emphasizes their practical utility in real-world business contexts. Through advanced analytics, organizations can derive actionable insights from their data, facilitating data-driven decision-making and cultivating a competitive edge in the marketplace.