MGT7177: Statistics for Business

Assignment 2

Title : Decoding Term Deposit Subscriptions: Advanced Modelling and Hypothesis Validation in Banking Analytics

Name: Dhanush Mathighatta Shobhan Babu

Student ID: 40412492

Word Count:2180

Table Of Content

Table of figures

1. Introduction

1.1 Overview and Problem Statement

The efficacy of the banking industry hinges on its ability to comprehend and fulfil consumer requisites, particularly in domains such as term deposit enrolments. The banking industry predominantly derives its income from clients' long-term deposits (Rony et al., 2021). Gaining a comprehensive understanding of client characteristics is crucial for banks in order to enhance product sales. This analysis investigates an intricate dataset comprising diverse client attributes, marketing strategies, and economic factors. The objective is to comprehend the variables that impact term deposit subscriptions. The primary challenge lies in understanding the complex interplay between customer demographics, campaign specifics, and economic indicators to effectively steer future marketing initiatives. The objective of this analysis is to identify the variables that impact clients' choices to subscribe to term deposits. This will be done by examining a dataset that includes various attributes such as age, occupation, campaign connections, and economic indicators. The primary concern is to ascertain the pivotal elements that impact client choices, allowing the bank to formulate targeted marketing strategies and offer tailored counsel to frontline staff. This will result in enhanced consumer involvement and increased rates of term deposit subscriptions. The aim of this investigation is to establish a link between theoretical understanding and practical observations derived from real-world data, leading to tangible solutions and outcomes in the banking sector.

1.2 Literature review

Title of the paper	Year of publication	Author	Model used, accuracy or conclusion
Term Deposit Subscription Prediction Using Spark MLlib and ML Packages	2019	Phan Duy Hung Tran Duc Hanh Ta Duc Tung	**1. Decision Tree:** <u>MLlib (Spark):</u> Started at 71% detection accuracy, improved to ~72% after balancing. <u>ML package (Spark):</u> Reached ~81% accuracy after tuning depth. **2. Random Forest (RF):** <u>MLlib (Spark):</u> Pre-normalization: ~73-75% accuracy. post-normalization: ~76-79% accuracy. <u>ML package (Spark):</u> Pre-normalization: ~82% accuracy. post-normalization: ~85% accuracy.

			3. Gradient Boosting (GBT): MLlib (Spark): Pre-normalization: ~69-87% accuracy. post-normalization: ~78-79% accuracy. ML package (Spark): Pre-normalization: ~82% accuracy. post-normalization: ~85% accuracy.
Statistical Decision Research of long-term deposit subscription in banks based on Decision Tree	2019	Guo Junfeng, Hou Handan	The study predominantly utilises the **Decision Tree algorithm** to evaluate the elements that impact consumers' long-term deposit subscriptions in the banking sector. The key findings emphasise that the number of employees, duration, and month have a crucial impact on client subscriptions, greatly restricting their extent. This impact improves the efficiency of banks, with the number of employees being the most significant factor.
Identifying Long-Term Deposit Customers: A Machine Learning Approach	2021	Mohammad Abu Tareq Rony, Md. Mehedi Hassan, Eshtiak Ahmed, Asif Karim, Sami Azam, D.S. A. Aashiqur Reza	**Models and Metrics Used:** **1.Logistic Regression (LR):** Accuracy: 0.9064 Sensitivity: 0.9905 Specificity: 0.2222 **2.Random Forest (RF):** Accuracy: 0.9016 Sensitivity: 0.9795 Specificity: 0.2667 **3. Support Vector Machine (SVM):** Accuracy: 0.8810 Sensitivity: 0.9615 Specificity: 0.2456

			4.K-Nearest Neighbors (KNN): Accuracy: 0.8991 Sensitivity: 0.9877 Specificity: 0.1778 **5.Multilayer Perceptron (MLP):** Accuracy: 0.8732 Sensitivity: 0.9455 Specificity: 0.1674
Machine Learning Performance on Predicting Banking Term Deposit	2022	Nguyen Minh Tuan	**The models used and their accuracy:** **1.Long-Short Term Memory (LSTM):** 90.3% Gated **2.Recurrent Unit (GRU):** 90.8% **3.Bidirectional Long-Short Term Memory (BiLSTM):** 90.5% **4.Bidirectional Gated Recurrent Unit (BiGRU):** 90.1% Simple **5.Recurrent Neuron Network (SimpleRNN):** 89.2%
Long-term deposits prediction: a comparative framework of classification model for predict the success of bank telemarketing	2019	Ahmad Ilham, ,Laelatul Khikmah, Indra,Ulumuddin, and Ida Bagus Ary Indra Iswara	The models used: **1.Decision Tree (DT)** - Accuracy: 90.00% **2.Naïve Bayes (NB)** - Accuracy: 87.18% **3.Random Forest (RF)** - Accuracy: 89.05% **4.K-Nearest Neighbors (K-NN) -** Accuracy: 88.23% Support **5.Vector Machine (SVM) - Accuracy:** 91.07%

			6.Neural Network (NN) - Accuracy: 88.59% **7.Logistic Regression (LR) -** Accuracy: 89.05%
Bank Deposit Prediction Using Ensemble Learning	2021	Muhammed J. A. Patwary, S. Akter, M. S. Bin Alam, A. N. M. Rezaul Karim	**Models and Metrics Used:** **1.Neural Network (NN):** Accuracy: 94.87% Sensitivity: 95.52% Specificity: 98.23% Error Rate: 5.13% **2.Support Vector Machine (SVM):** Accuracy: 89.76% Sensitivity: 85.2% Specificity: 96.88% Error Rate: 10.24% **3.Naive Bayes (NB):** Accuracy: 88.23% Sensitivity: 84.32% Specificity: 94.20% Error Rate: 11.77%
Prediction of Client Term Deposit Subscription Using Machine Learning	2023	Muskan Singh, Namrata Dhanda, U. K. Farooqui, Kapil Kumar Gupta & Rajat Verma	**Models and Accuracy obtained:** **Logistic Regression:**89.083728% **Support Vector Machine:**88.593997% **Random Forest Classifier:** 90.726698% **Decision Tree Classifier:** 90.426540%
Finding The Best Techniques For Predicting Term Deposit Subscriptions (Case Study UCI Machine Learning Dataset)	2022	Lila Setiyani, Ayu Indahsari , Rosalina, Tjong Wansen	**Models and Metrics Used:** **1.XGBoost:** Accuracy: 91.73% Recall: 91.73% Precision: 90.91% F-Score: 91.07%

			Overall top performer across all metrics. **2.Random Forest (RF):** Accuracy: 91.44% Recall: 91.44% Precision: 90.66% F-Score: 90.9% Strong performance, close to XGBoost. **3.Logistic Regression (LR):** Accuracy: 91.2% Recall: 91.2% Precision: 90.09% F-Score: 90.22% Consistent high performance. **4.K-Nearest Neighbors (KNN):** Accuracy: 90.6% Recall: 90.6% Precision: 89.93% F-Score: 90.2% Solid performance but slightly lower than the top contenders. **5.Support Vector Machine (SVM):** Accuracy: 89.6% Recall: 89.6% Precision: 87.58% F-Score: 87.2% Good accuracy but relatively lower precision. **6.Decision Tree (DT):** Accuracy: 88.79% Recall: 88.79% Precision: 88.84% F-Score: 88.82% Moderate performance compared to other models. **7.Naive Bayes (NB):** Accuracy: 84.78%

			Recall: 84.78% Precision: 88.35% F-Score: 86.18% Lower accuracy compared to other models, but higher precision.
Factors determining bank deposit growth in Turkey: an empirical analysis	2020	Ibrahim Nandom Yakubu and Aziza Hashi Abokor	**1.Serial Correlation (Breusch–Godfrey Test):** F-Statistic: 1.871 Probability Value: 0.169 **2.Heteroscedasticity (Breusch–Pagan Test):** F-Statistic: 1.530 Probability Value: 0.116 **3.Normality (Jarque–Bera Test): Test Statistic: 1.802** Probability Value: 0.406 **4.Ramsey RESET Test:** F-Statistic: 3.661 Probability Value: 0.064 The majority of the diagnostic tests indicate that the model meets basic assumptions regarding serial correlation, homoscedasticity, and normality. However, the Ramsey RESET Test suggests a slight potential for omitted variable bias, although the significance is not high.
Applying Machine Learning to the Development of Prediction Models for Bank Deposit Subscription	2021	Sipu Hou, Zongzhen Cai, Jiming Wu, Hongwei Du, Peng Xie	**Models and Metrics Used:** **1.Naive Bayes:** Accuracy: 86.54% Sensitivity: 89.79% **2.Decision Tree:** Accuracy: 91.79% Sensitivity:96.55% **3.Random Forest:** Accuracy: 91.89% Sensitivity: 96.58%

			4.Support Vector Machine (SVM): Accuracy: 91.72% Sensitivity;97.21% **5.Neural Network:** Accuracy: 88.86% Sensitivity:99.99%

1.3 Hypothesis assumed

SL NO	Hypothesis assumed
1	**1.Personal_loan VS Subscribed** H0: There is no relationship between Personal_loan and Subscribed H1: There is a relationship between Personal_loan and Subscribed
2	**2.pdays VS Subscribed** H0: There is no relationship between pdays and Subscribed H1: There is a relationship between pdays and Subscribed
3	**3. Occupation VS Subscribed** H0: There is no relationship between Occupation and Subscribed H1: There is a relationship between Occupation and Subscribed
4	**4. Credit default VS Subscribed** H0: There is no relationship between Credit default and Subscribed H1: There is a relationship between Credit default and Subscribed
5	**5.Campaign VS Subscribed** H0: There is no relationship between Campaign and Subscribed H1: There is a relationship between Campaign and Subscribed

2. Methodology

2.1 Analytical tasks and approach

The process involves several tasks, such as data pre-processing, hypothesis creation and testing, data visualisation, statistical association measurement, regression analysis, and model evaluation. The approach outlined is the usual methodology commonly utilised in Machine Learning models. The authors (Ghosalkar & Dhage, 2018) and (Manasa et al., 2020) employed a comparable approach.This can also be denoted as a modification of CRISP-DM (Schröer et al., 2021).
The following flowchart offers a succinct summary of the analytical tasks utilised.

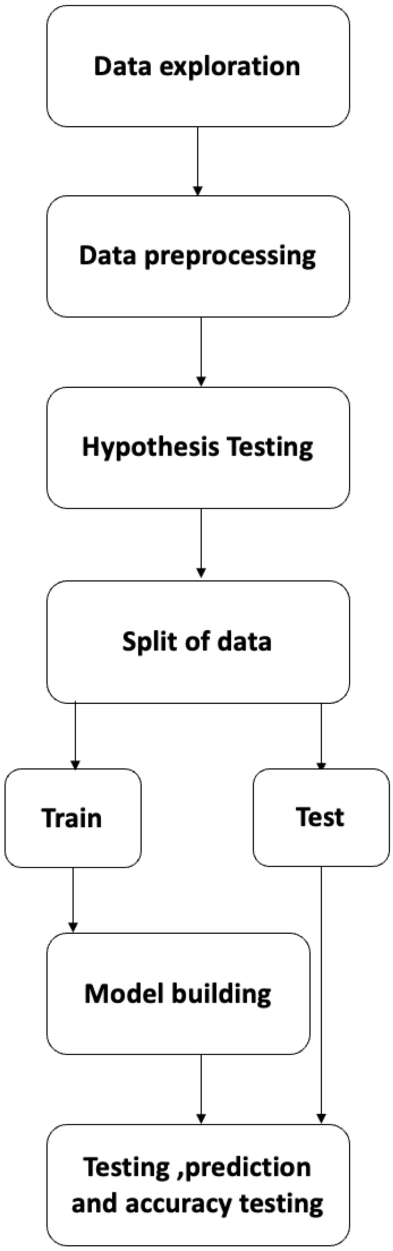
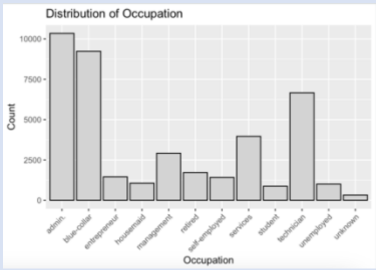
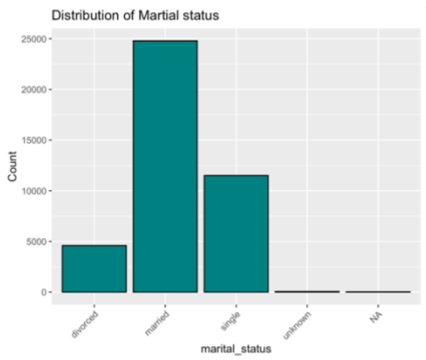
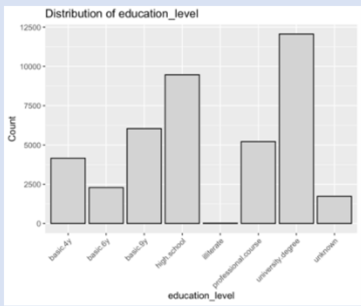


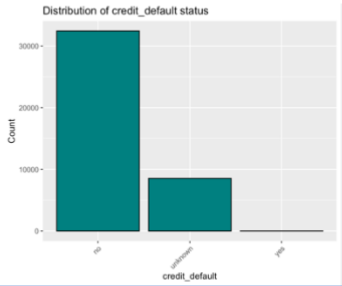
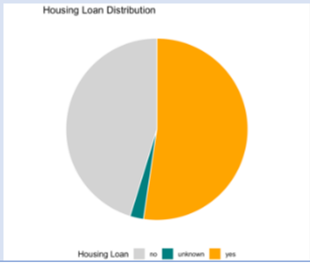
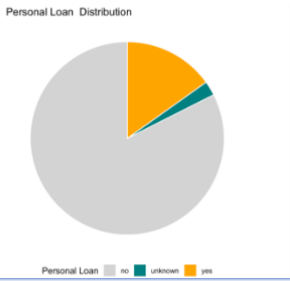
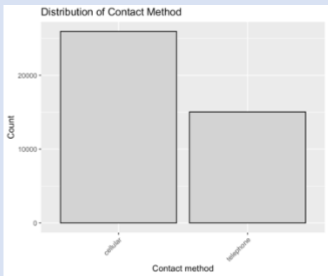
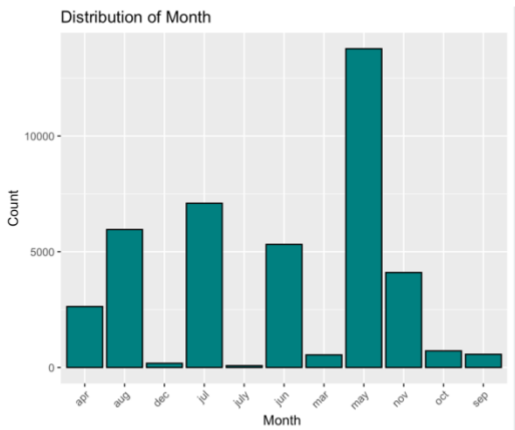
Fig 2.1Flow chart of Methodology

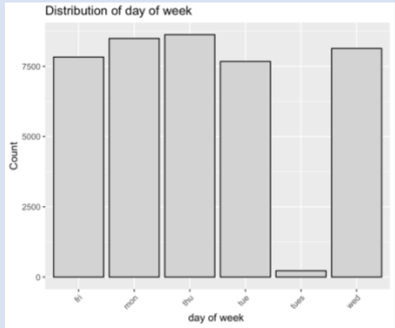
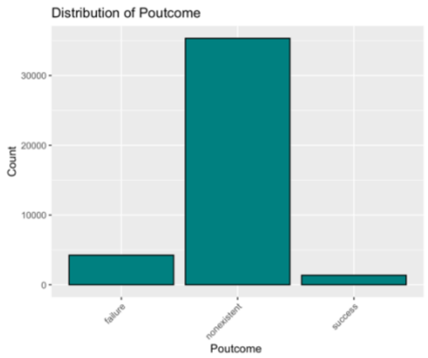
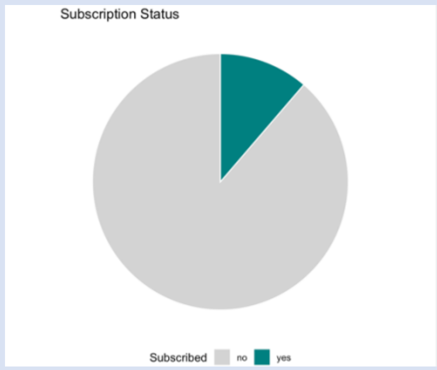
2.2 Data exploration and data quality assessment

Once the business challenge has been comprehended, the subsequent task is to have a thorough grasp of the data. Examining the provided data, describing its characteristics, and evaluating its quality are essential tasks in this phase (Schröer et al., 2021).The dataset comprises 21 variables, featuring 11 categorical factors including client attributes, contact methods, and campaign outcomes ('occupation,' 'marital_status,' 'contact_method,' etc.). Additionally, 10 numerical indicators encompassing age, campaign duration, economic indices, and contact history ('age,' 'contact_duration,' 'emp_var_rate,' etc.) are present. Among these, 'subscribed' serves as the target variable, delineating term deposit subscriptions. The descriptive statistics for the variables are shown below:

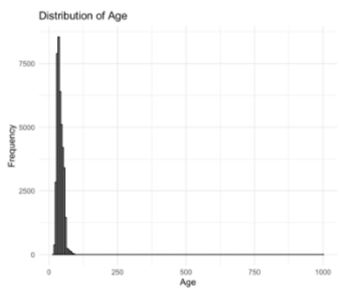
Categorical variables:

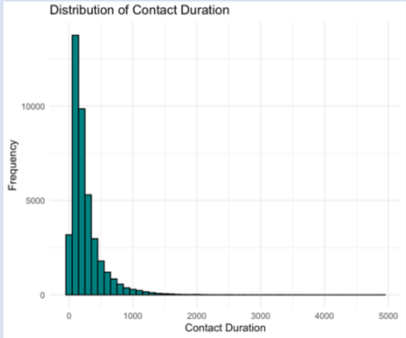
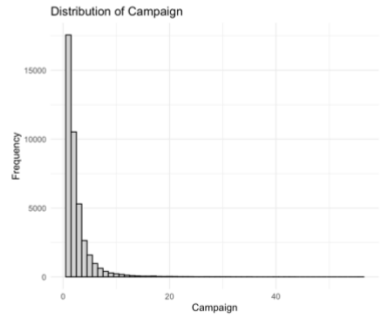
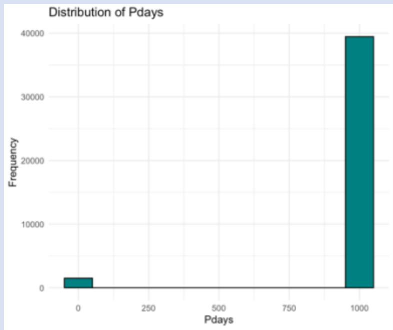
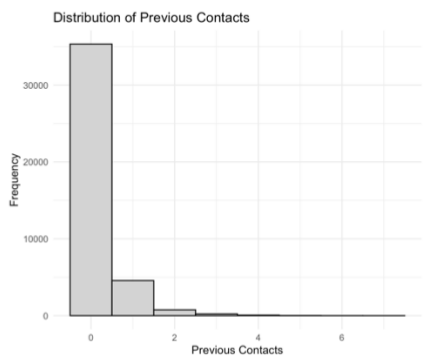
Sl No	Data Variable Name	Distinct values	Visualization
1	**Occupation**	**1."Admin" 2."blue-collar" 3 ."entrepreneur" 4. "housemaid" 5 ."management" 6 "retired" 7 ."self-employed" 8. "services" 9. "student" 10 ."technician" 11 ."unemployed" 12. "unknown"**	
2	**Marital_Status**	**1."divorced " 2 ."married" 3." Single" 4 ."unknown" 5." NA"**	
3	**Education_level**	**1."basic.4y" 2."basic.6y" 3."basic.9y" 4."high.school" 5."illiterate" 6."professional.course" 7."university.degree" 8."unknown"**	

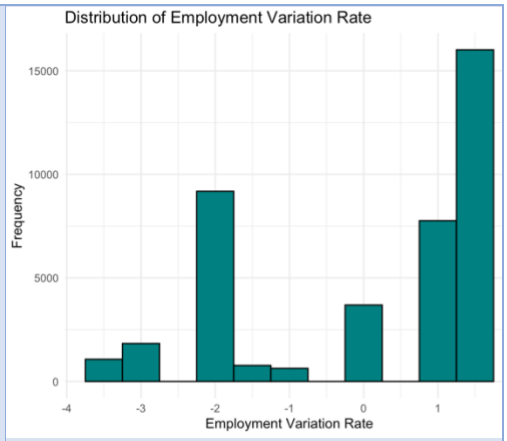
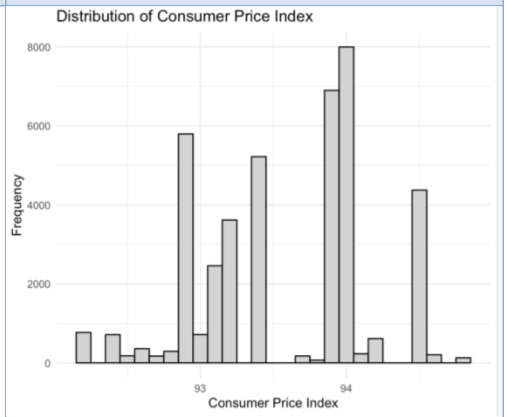
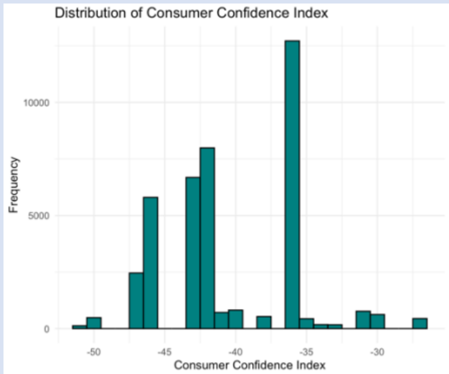
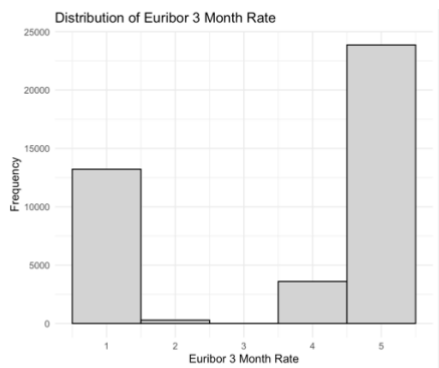
4	**credit_default**	**1."no"** **2."unknown"** **3."yes"**	
5	**housing_loan**	**1."no"** **2."unknown"** **3."yes"**	
6	**personal_loan**	**1."no"** **2."unknown"** **3."yes"**	
7	**contact_method**	**1."cellular"** **2 ."telephone"**	
8	**month**	**1."apr"** **2 ."aug"** **3."dec"** **4."jul"** **5."July"** **6."jun"** **7."mar"** **8."may"** **9."nov"** **10."oct"** **11."sep"**	

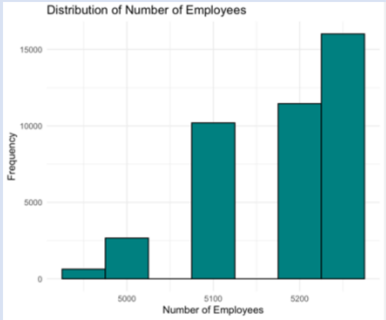
9	**day_of_week**	**1."fri"** **2."mon"** **3."thu"** **4."tue"** **5."tues"** **6."wed"**	
10	**poutcome**	**1."failure"** **2."non-existent"** **3."success"**	
11	**Subscribed (Target Variable)**	**1."no"** **2."yes"**	

Numerical variables:

SL NO	Data variable Name	Min	Median	Mean	Max	Visualization
1	**ID**	**1**	**20485**	**20485**	**40969**	**-**
2	**age**	**17.00**	**38.00**	**40.07**	**999.00**	

3	**contact_duratin**	**0.0**	**180.0**	**258.5**	**4918.0**	
4	**campaign**	**1.000**	**2.000**	**2.565**	**56.000**	
5	**pdays**	**0.0**	**999.0**	**962.3**	**999.0**	
6	**previous_contacts**	**0.0000**	**0.0000**	**0.1739**	**7.0000**	

7	emp_var_rate	-3.40000	1.10000	0.07484	1.40000	
8	cons_price_idx	92.20	93.80	93.58	94.77	
9	cons_conf_idx	- 50.80	- 41.80	- 40.53	- 26.90	
10	euribor_3m	0.634	4.857	3.614	5.045	

11	**n_employed**	**4964**	**5191**	**5167**	**5228**	

2.3 Variable Selection

We have selected a total of 15 variables. The selection of these variables has been.
based on three factors:

• Derived from a hypothesis
• Supported by research papers
• Established by logical reasoning

The variables that are deduced from a **hypothesis** are:

1. personal_loan
2. pdays
3. occupation
4. credit_default
5. campaign

The variables that are examined in **research papers:**

1. Age - (Hung et al., 2019)
2. marital_status - (Guo & Hou, 2019)
3. education_level - (Ilham et al., 2019)
4. day_of_week- (Setiyani et al., 2022)
5. month- (Singh et al., 2023)

The variables that are determined using **deductive reasoning:**

1.poutcome: Past campaign outcomes guide future strategies, pivotal for improving subscription rates through optimized approaches and better-targeted campaigns.

2. housing_loan: Financial commitments like housing loans impact decisions. Understanding this influence aids in gauging how obligations sway term deposit behaviours.

3.contact_method: Identifying effective communication channels optimizes campaigns, tailoring methods for higher client engagement and subscription rates.

4. nr_employed: Economic indicators reflect market conditions. Changes in employment rates may affect consumer behaviour, influencing term deposit decisions.

5. Euribor_3m: Market interest rates impact savings choices. Analysing this variable uncovers correlations between rates and term deposit subscriptions, revealing client responses to rate changes.

2.4 Data quality issues

We have just focused on resolving the data quality concerns pertaining to the factors that have been indicated previously. The table displays the data quality concerns.

SL NO	Data variable name	Data type	Outlier/Data quality issues
1	Age	Numeric	An age value of 999 stands as an illogical outlier within the dataset, presenting an implausible scenario.
2	marital_status	Categorical	The 'marital_status' variable has 23 missing values that can be resolved through either removal or imputation using the mode.
3	month	Categorical	The dataset includes entries for month as "July" and "jul", which are both referring to the same month. In order to maintain uniformity with the other three-letter abbreviations for months, we choose to utilise "jul".
4	day_of_week	Categorical	The dataset consists of entries for the day of the week, represented as both "tues" and "tue," both referring to the same day. To align with the consistent format used for other days of the week, we opt to use "tue."

2.5 Addressing data quality issues

Resolve the issue of inadequate data quality through the process of data cleansing. Create derived characteristics based on the selected model from the first phase. The optimal approach for these processes is contingent upon the model utilised (Schröer et al., 2021).

1. Age

To remove outliers we need first calculate the upper bound and lower bound Through Inter-Quartile Range (IQR), an outlier x can be detected

if: x < Q1 − 1.5 * IQR (IQR)
OR

Q3 + 1.5 (IQR) < x
where:
Q1 = 25th percentiles
Q3 = 75th percentiles
IQR = Q3 − Q1 (Truong et al., 2020) .

We utilised the filter function from the "dplyr" library to eliminate the outliers.

However, under this scenario, it is only necessary to exclude the value "999" as it is irrational and beyond the realm of human lifespan. Other age values remain valid, as individuals can continue to construct term deposits until their death without any limitations.

Fig 2.5.2: box plot of age before and after data cleaning

2. Marital_status

As previously stated, the 23 NA values are filtered using the filter function in the 'dplyr' library.

```
> count(data,marital_status)
# A tibble: 5 × 2
  marital_status     n
  <chr>          <int>
1 divorced        4597
2 married        24791
3 single         11501
4 unknown           57
5 NA                23
>
```

After

Cleaning

```
> count(newdata,marital_status)
# A tibble: 4 × 2
  marital_status     n
  <chr>          <int>
1 divorced        4597
2 married        24789
3 single         11501
4 unknown           57
>
```

Fig 2.5.2: before and after data cleaning of Marital_status

3. month

The month of July has been renamed to "jul" using the mutate function in the dplyr package.

```
# A tibble: 11 × 2
   month     n
   <chr> <int>
 1 apr    2629
 2 aug    5957
 3 dec     182
 4 jul    7095
 5 july     79
 6 jun    5315
 7 mar     545
 8 may   13762
 9 nov    4094
10 oct     716
11 sep     570
```

After

Cleaning

```
# A tibble: 10 × 2
   month     n
   <chr> <int>
 1 apr    2629
 2 aug    5957
 3 dec     182
 4 jul    7174
 5 jun    5315
 6 mar     545
 7 may   13762
 8 nov    4094
 9 oct     716
10 sep     570
```

Fig 2.5.3: before and after data cleaning of Month

3. day_of_week

The variable "day_of_week" contains two entries, "tue" and "Tue". In order to ensure consistency, we have renamed "tues" to "tue" using the "mutate" function in the "dplyr" library.

```
  day_of_week     n
  <chr>       <int>
1 fri          7821
2 mon          8485
3 thu          8614
4 tue          7667
5 tues          224
6 wed          8133
```

After

Cleaning

```
# A tibble: 5 × 2
  day_of_week     n
  <chr>       <int>
1 fri          7821
2 mon          8485
3 thu          8614
4 tue          7891
5 wed          8133
```

Fig 2.5.4: before and after data cleaning of day_of_week

2.6 Hypothesis Testing

The most prevalent hypothesis test often entails evaluating the null hypothesis in comparison to the alternative hypothesis.

H0: There is no relationship between X and Y versus the alternative hypothesis is
Ha: There is some relationship between X and Y (James et al.).

1.Personal_loan VS Subscribed

H0: There is no relationship between Personal_loan and Subscribed
H1: There is a relationship between Personal_loan and Subscribed

The statistical analysis, through Pearson's Chi-squared test (X-squared = 0.94188, df = 2, p = 0.6244), scrutinized the association between Personal_loan and Subscribed variables. The obtained p-value of 0.6244 surpasses conventional significance levels. This substantial p-value fails to provide ample evidence to refute the null hypothesis (H0: no relationship between Personal_loan and Subscribed), indicating inconclusive evidence of any discernible association between these variables. Therefore, the analysis supports the retention of the null hypothesis.

Fig 2.6.1: Bar plot for Personal_loan VS Subscribed

2.Pdays VS Subscribed

H0: There is no relationship between pdays and Subscribed
H1: There is a relationship between pdays and Subscribed

A Welch Two Sample t-test was performed to investigate the correlation between the variables "pdays" and "Subscribed." The test indicated a substantial disparity in averages ($t = 32.2$, $df = 4735.8$, $p < 2.2e-16$) and a noteworthy range of certainty (180.3095 to 203.6888). The null hypothesis (H0) that there is no relationship is rejected, and the alternative hypothesis (H1) that a relationship exists is highly supported. This indicates a definite association between the number of days since the client was last contacted ("pdays") and their subscription status.

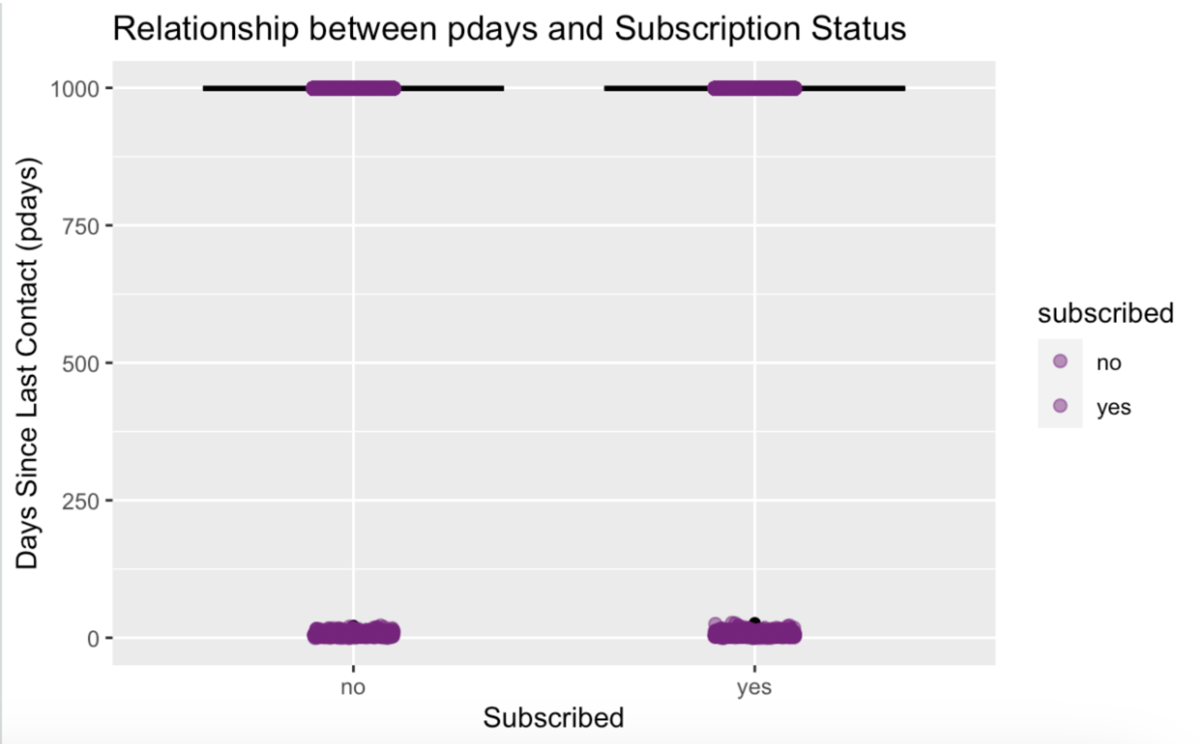


Fig 2.6.2: box for Pdays VS Subscribed

3. Occupation VS Subscribed

H0: There is no relationship between Occupation and Subscribed
H1: There is a relationship between Occupation and Subscribed

The Pearson's Chi-squared test was performed to comprehensively analyse the potential association between the variables "Occupation" and "Subscribed." The test yielded a chi-squared value of 955.67, with 11 degrees of freedom and a p-value less than 2.2e-16. The remarkably small p-value (< 2.2e-16) indicates a significant deviation from the assumption of no correlation (H0: no relationship). Therefore, the test strongly confirms the alternative hypothesis (H1: a relationship exists) about the connection between Occupation and Subscribed, indicating a considerable correlation between these variables.

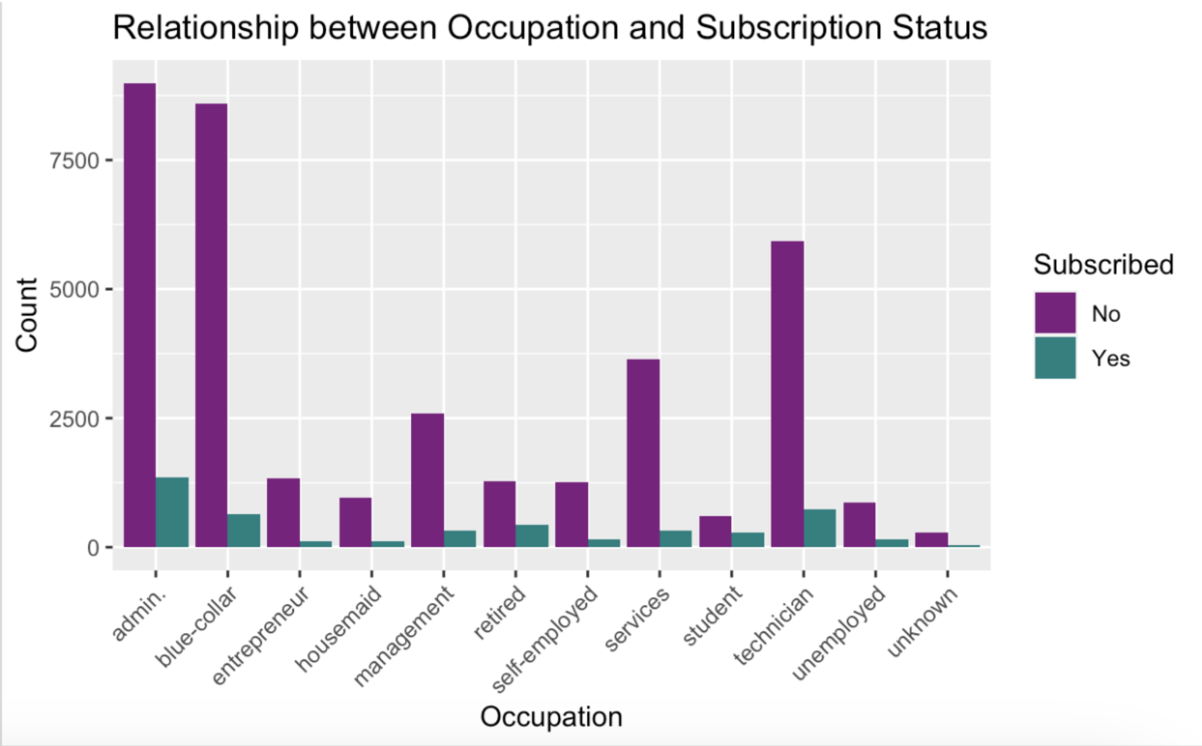


Fig 2.6.3: Barplot for occupation VS Subscribed

4. Credit default VS Subscribed

H0: There is no relationship between Credit default and Subscribed
H1: There is a relationship between Credit default and Subscribed

The Pearson's Chi-squared test was used to systematically assess the potential relationship between "Credit default" and "Subscribed." The test yielded an X-squared value of 409.3, with 2 degrees of freedom and a p-value of less than 2.2e-16. The test's p-value is extremely low (< 2.2e-16), indicating strong evidence against the premise of no association (H0: no relationship). Therefore, compelling evidence substantiates the alternative hypothesis (H1: a relationship exists) between Credit default and Subscribed, demonstrating a significant and noteworthy correlation between these variables.

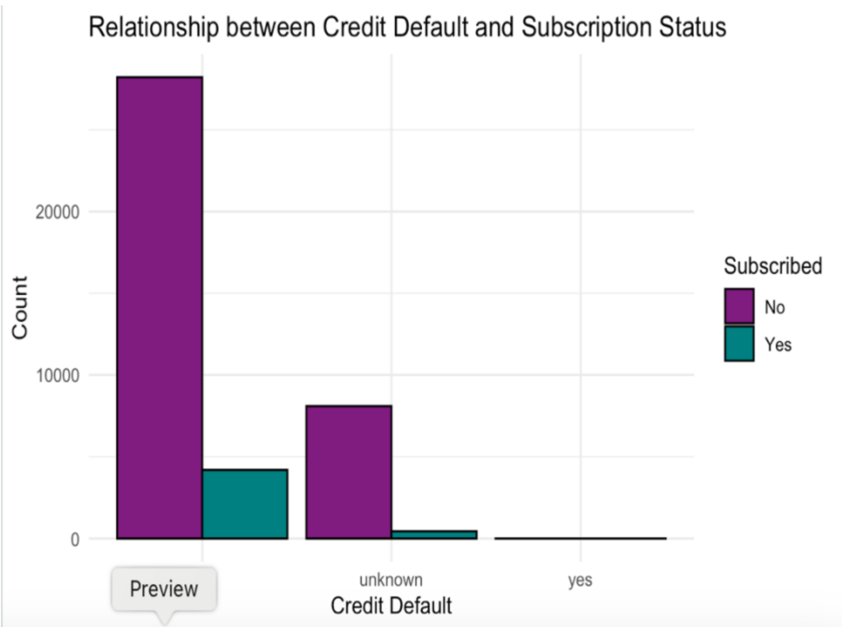


Fig 2.6.4: Barplot for Credit default VS Subscribed

5.Campaign VS Subscribed

H0: There is no relationship between Campaign and Subscribed
H1: There is a relationship between Campaign and Subscribed

The Welch t-test (t = 20.201, df = 8660.4, p < 2.2e-16) shows a statistically significant association between the Campaign and Subscribed variables. The mean for the "No" group is 2.63 and the mean for the "Yes" group is 2.05, with a 95% confidence interval ranging from 0.524 to 0.637. The compelling evidence refutes the null hypothesis (H0) and provides support for the alternative hypothesis (H1), indicating a significant association between these variables.

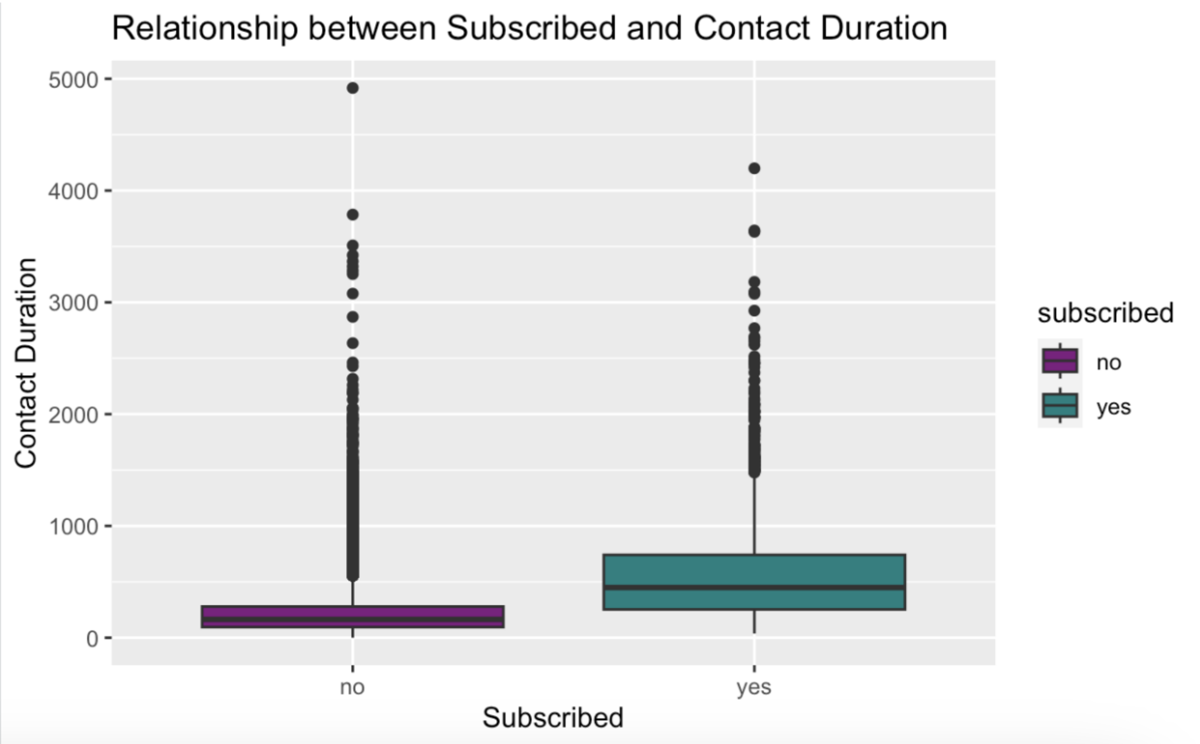


Fig 2.6.5: Boxplot for Campaign VS Subscribed

2.7 Regression Model Techniques

Logistic regression assesses binary outcomes like disease presence. A single predictor yields a simple logistic model, while multiple predictors, combining categorical and continuous variables, constitute a multivariable logistic model (Nick & Campbell, 2007).

The multivariable model incorporates a linear combination of the predictors. Let's examine a scenario involving three predictor variables, namely X1, X2, and X3(Nick & Campbell, 2007).
. The logarithm of the ratio of probabilities can be expressed as

$$\log \text{odds}(Y = 1 \mid X_1, X_2, X_3) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3.$$

2.8 Model building

We have utilised the forward method to build our model. A forward election rule begins with no explanatory factors and gradually includes variables, one by one, based on their statistical significance, until there are no more variables that show statistical significance (Smith, 2018).

In total we have built 3 models:

Model 1 : based on variables derived from hypothesis
Model 2: based on variables derived from hypothesis +Literature review
Model 3: Adding all the variables (hypothesis + Literature review + Logical reasoning)

3. Results and Discussion

3.1 Presentation of Key Outputs

Model 1 : based on variables derived from hypothesis

Model_1, predicting 'subscribed', achieved 89.47% accuracy and Kappa 0.2425, signifying moderate agreement. High sensitivity (98.55%) contrasted with lower specificity (18.25%). Key predictors were 'pdays', 'occupation' ('blue-collar', 'retired', 'student'), 'credit_default' ('unknown', 'yes'), and 'campaign'. Pseudo-R^2 values (Hosmer-Lemeshow: 0.119, Cox-Snell: 0.081, Nagelkerke: 0.159) indicate moderate fit, potentially limited in capturing 'subscribed' outcomes.

Model 2: based on variables derived from hypothesis +Literature review

Model_2, an extended logistic regression, included 'age', 'marital_status', 'education_level', 'day_of_week', and 'month'. With 89.58% accuracy and Kappa 0.2354, it showed high sensitivity (98.82%) and lower specificity (17.17%). Key predictors were 'age', 'marital_status' ('single', 'married'), 'month' ('mar', 'may', 'nov', 'sep'), and 'credit_default' ('unknown'). Pseudo-R^2 values—Hosmer-Lemeshow R^2 (0.158), Cox-Snell R^2 (0.106), Nagelkerke R^2 (0.208)—suggested better fit, but variability within 'subscribed' outcomes might still challenge the model.

Model 3: based on variables derived from hypothesis +Literature review + Logical reasoning

Model_3 logistic regression achieved 90.78% accuracy and Kappa 0.443. It excelled in 'no' predictions (sensitivity: 97.36%) but less in 'yes' (specificity: 39.20%). Key predictors include 'contact_duration', 'poutcome' ('nonexistent', 'success'), 'month' ('mar', 'may', 'jun', 'aug', 'nov', 'dec'), 'housing_loan' ('unknown'), 'n_employed', 'euribor_3m', and 'credit_default' ('unknown'). The expanded variables enhanced fit (Pseudo-R^2: Hosmer-Lemeshow 0.411, Cox-Snell 0.252, Nagelkerke 0.497), yet class imbalance in 'yes' predictions might pose challenges.

Model_1 achieved an accuracy of 89.47% but had imbalanced sensitivity (98.55%) and lower specificity (18.25%). Model_2, with added socio-demographics, slightly enhanced accuracy (89.58%) and sensitivity (98.82%), but specificity remained at 17.17%. Model_3, the most comprehensive, reached 90.78% accuracy, 97.36% sensitivity, and 39.20% specificity. However, there might be challenges predicting 'yes' due to class imbalance.

3.2 Presentation of Key Outputs of all models

Logistic Regression Models Summary

	Dependent Variable: Subscribed		
	subscribed		
	Model 1	Model 2	Model 3
pdays	-0.003^{***}	-0.002^{***}	-0.001^{***}
occupationblue-collar	-0.452^{***}	-0.218^{***}	-0.267^{***}
occupationentrepreneur	-0.321^{***}	-0.221^{*}	-0.239^{*}
occupationhousemaid	-0.180	-0.111	-0.012
occupationmanagement	-0.134^{*}	-0.147^{*}	-0.068
occupationretired	0.701^{***}	0.421^{***}	0.219^{*}
occupationself-employed	-0.052	-0.071	-0.134
occupationservices	-0.317^{***}	-0.145^{*}	-0.119
occupationstudent	0.889^{***}	0.716^{***}	0.348^{***}
occupationtechnician	-0.132^{**}	-0.085	0.029
occupationunemployed	0.126	0.105	0.016
occupationunknown	0.198	0.174	0.371
credit_defaultunknown	-0.770^{***}	-0.624^{***}	-0.340^{***}
credit_defaultyes	-8.661	-8.283	-6.031
campaign	-0.089^{***}	-0.074^{***}	-0.049^{***}
age		0.006^{**}	0.002
marital_statusmarried		0.128^{*}	0.075
marital_statussingle		0.276^{***}	0.106
marital_statusunknown		0.586	0.463
education_levelbasic.6y		0.060	0.071
education_levelbasic.9y		-0.063	-0.019
education_levelhigh.school		0.029	0.006
education_levelilliterate		0.120	0.020
education_levelprofessional.course		0.065	0.075
education_leveluniversity.degree		0.194^{**}	0.188^{*}
education_levelunknown		0.133	0.096
day_of_weekmon		-0.156^{**}	-0.152^{**}
day_of_weekthu		0.021	-0.008
day_of_weektue		0.062	0.082
day_of_weekwed		0.074	0.113
monthaug		-0.799^{***}	0.509^{***}
monthdec		0.802^{***}	0.460^{**}
monthjul		-0.759^{***}	0.479^{***}
monthjun		-0.598^{***}	0.509^{***}
monthmar		1.095^{***}	1.411^{***}
monthmay		-1.117^{***}	-0.698^{***}
monthnov		-0.888^{***}	0.088
monthoct		0.580^{***}	0.459^{***}
monthsep		0.469^{***}	0.061
poutcomenonexistent			0.523^{***}
poutcomesuccess			1.024^{***}
housing_loanunknown			-0.122
housing_loanyes			0.010
contact_duration			0.005^{***}
n_employed			-0.007^{***}
euribor_3m			-0.348^{***}
Constant	0.768^{***}	0.654^{***}	35.624^{***}
Observations	32,756	32,756	32,756
Log Likelihood	-10,183.180	-9,732.996	-6,807.582
Akaike Inf. Crit.	20,398.370	19,545.990	13,709.160
Significance Level	*p<0.1; **p<0.05; ***p<0.01		

Fig 3.2: Output of all models using stargazer.

3.3 Plot of Key Outputs of Best Model (model 3)

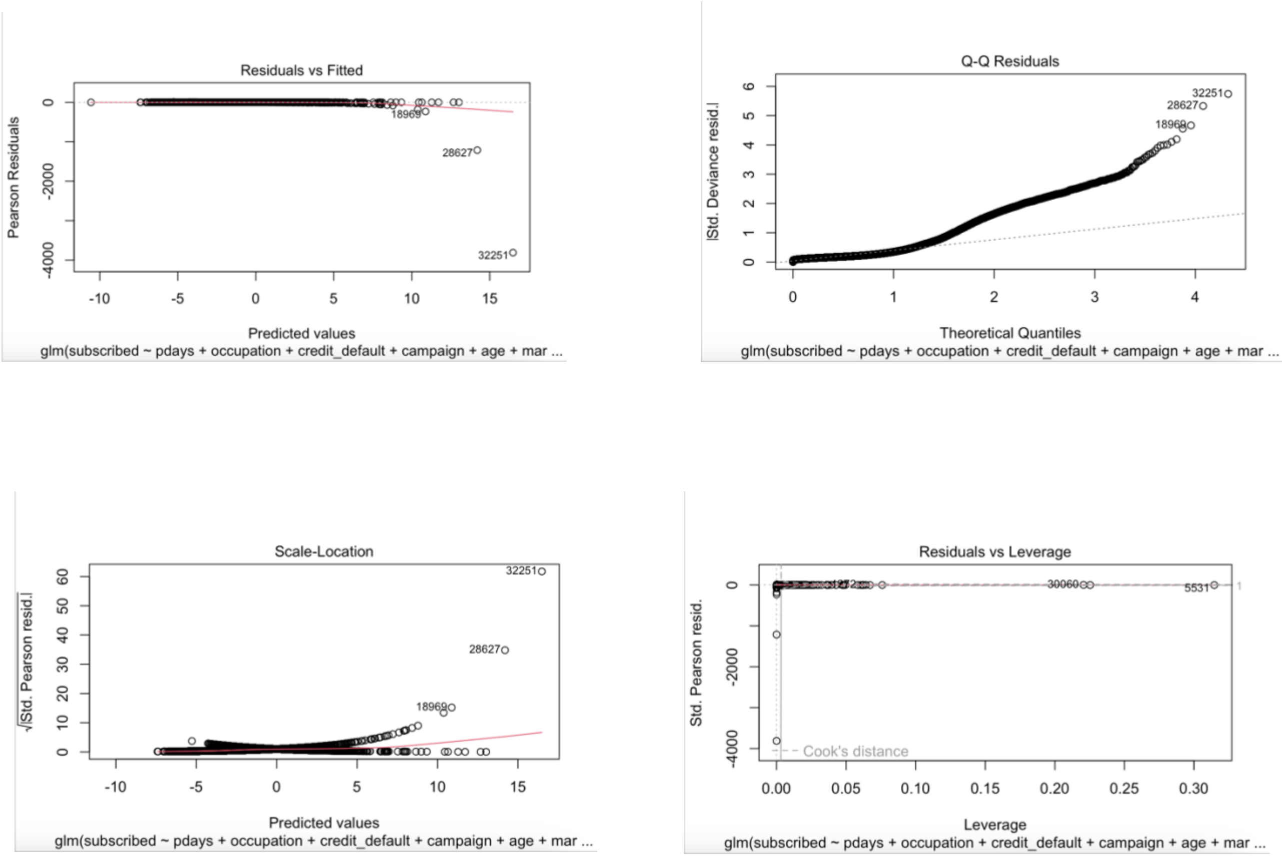
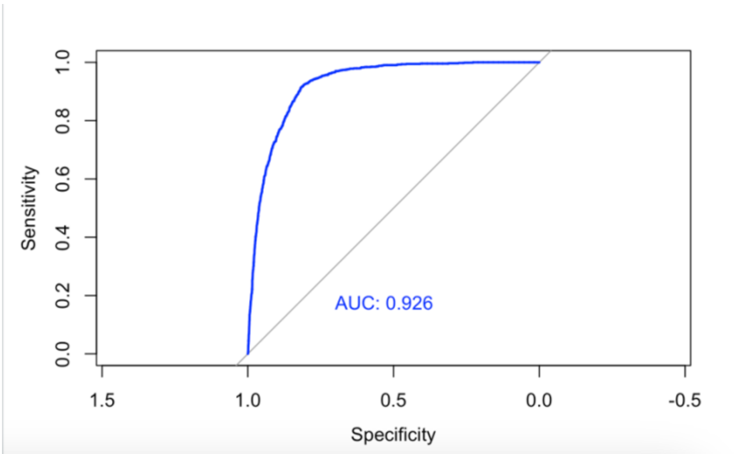


Fig 3.3.1: Plot of model (3)

Diagnostic plots for a generalized linear model show influential outliers (observations 30060 and 5531) and signs of heteroscedasticity, with deviations from normality in the Q-Q plot, particularly for observations 189690, 286270, and 322510.



The ROC curve displayed indicates a high level of model performance, with an AUC (Area Under the Curve) of 0.926. This suggests that the model has a strong ability to differentiate between the positive and negative classes, with high sensitivity and specificity.

Fig 3.3.2: Roc curve of model (3).

3.4 Model Assumptions

The Variance Inflation Factor (VIF) gauges multicollinearity among regression predictors. Key findings in model 3 are:

1. **pdays**: Moderate multicollinearity (VIF: 9.85) hints at correlation with others.
2. **occupation**: Shows mild multicollinearity (VIF: 5.78) among its categories.
3**. Moderate VIF (above 5):** poutcome, n_employed, euribor_3m.
4. **Low VIF (around 1):** credit_default, campaign, age, marital_status, education_level, day_of_week, month, housing_loan.

Variables like pdays and certain others exhibit moderate multicollinearity, while most predictors show minimal correlation with others in the model.

4.0 Reflective Commentary

4.1 Further Steps

The primary objective in improving predictive models is to enhance data quality and resolve issues related to multicollinearity. Addressing outliers, filling in missing values, and optimising models to mitigate overfitting are crucial. In addition, the integration of sophisticated algorithms such as ensemble methods and deep learning could enhance the precision of predictions. Improving the interpretability of the model is achieved by doing feature importance analysis and optimising thresholds to address class imbalances, resulting in a more accurate prediction. Continuous learning is crucial to understand the subtle details of new methodologies and make major contributions to data-driven achievements in this rapidly changing field.

4.2 Learnings and Future Aspiration

Exploring CARET, GLM, TIDYVERSE, GGPLOT, and CAR paved the way for me to become a proficient business analyst specialising in advanced supervised learning. The recent experiences with various machine learning strategies have ignited a strong motivation to enhance predictive models and categorization techniques. With the goal of enhancing understanding for better decision-making, I strive to utilise a wide range of sophisticated algorithms. I am driven by a desire for continuous learning, immersing myself in real-world data, and grasping important concepts like as overfitting. My objective is to make a significant contribution by challenging limits in this ever-changing domain, motivated by a strong enthusiasm for improving data-driven methods.

5.0 References

Sure, I'll help you alphabetize the list of references you provided:

- Ghosalkar, N.N. and Dhage, S.N. (2018) ‘Real estate value prediction using linear regression’, 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) [Preprint]. DOI Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8697639 [Accessed on 15th December 2023]].

- Guo, J. and Hou, H. (2019) ‘Statistical Decision Research of long-term deposit subscription in banks based on decision tree’, 2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS) [Preprint]. Doi: available at https://ieeexplore.ieee.org/document/8669650?denied [Accessed on 24th Dec 2023].

- Hung, P.D., Hanh, T.D. and Tung, T.D. (2019) ‘Term deposit subscription prediction using spark MLlib and ML Packages’, Proceedings of the 2019 5th International Conference on E-Business and Applications [Preprint]. doi:Available at : https://dl.acm.org/doi/pdf/10.1145/3317614.3317618 [Accessed on 20th November 2023].

- Ilham, A. et al. (2019) ‘Long-term deposits prediction: A comparative framework of classification model for predict the success of bank telemarketing’, Journal of Physics: Conference Series, 1175, p. 012035. doi:Available at : https://iopscience.iop.org/article/10.1088/1742-6596/1175/1/012035/pdf [Accessed on 2nd December 2023].

- James, G. et al. (no date) An introduction to statistical learning: With applications in R. Springer.

- Manasa, J., Gupta, R. and Narahari, N.S. (2020) ‘Machine learning based predicting house prices using regression techniques’, 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) [Preprint]. doi:Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9074952 [Accessed on 30th December 2023]].

- Patwary, M.J. et al. (2021) ‘Bank deposit prediction using ensemble learning’, Artificial Intelligence Evolution, pp. 42–51. doi:Available at : https://ojs.wiserpub.com/index.php/AIE/article/view/880/591 [Accessed on 3rd December 2023].

- Rony, M.A. et al. (2021) ‘Identifying long-term deposit customers: A machine learning approach’, 2021 2nd International Informatics and Software Engineering Conference (IISEC) [Preprint]. doi:Available at https://ieeexplore.ieee.org/abstract/document/9672452 [Accessed on 25th Dec 2023].

- Setiyani, L. et al. (2022) 'Finding the best techniques for predicting term deposit subscriptions (Case Study UCI Machine Learning Dataset)', 2022 IEEE International Conference on Sustainable Engineering and Creative Computing (ICSECC) [Preprint]. doi:Available at : https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10331379 [Accessed on 25th December 2023].

- Singh, M. et al. (2023) 'Prediction of client term deposit subscription using machine learning', Lecture Notes in Electrical Engineering, pp. 83–93. doi:Available at : https://link.springer.com/chapter/10.1007/978-981-99-2710-4_8 [Accessed on 4th December 2023].

- Smith, G. (2018) 'Step away from stepwise', Journal of Big Data, 5(1). doi:Available at: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-018-0143-6 [Accessed on 22nd December 2023].

- Truong, Q. et al. (2020) 'Housing price prediction via Improved Machine Learning Techniques', Procedia Computer Science, 174, pp. 433–442. doi:Available at: https://www.sciencedirect.com/science/article/pii/S1877050920316318?via%3Dihub [Accessed on 5th December 2023].

- Tuan, N. (2022) 'Machine learning performance on predicting banking term deposit', Proceedings of the 24th International Conference on Enterprise Information Systems [Preprint]. doi:10.5220/0011096600003179 Available at https://www.scitepress.org/PublishedPapers/2022/110966/110966.pdf [Accessed on 25th Dec 2023].

- Yakubu, I.N. and Abokor, A.H. (2020) 'Factors determining bank deposit growth in Turkey: An empirical analysis', Rajagiri Management Journal, 14(2), pp. 121–132. doi:10.1108/ramj-05-2020-0017 Available at https://www.emerald.com/insight/content/doi/10.1108/RAMJ-05-2020-0017/full/pdf?title=factors-determining-bank-deposit-growth-in-turkey-an-empirical-analysis [Accessed on 26th Dec 2023].

- Hou, S. et al. (2021) 'Applying machine learning to the development of prediction models for Bank Deposit Subscription', International Journal of Business Analytics, 9(1), pp. 1–14. doi:10.4018/ijban.288514 Available at https://www.igi-global.com/pdf.aspx?tid=288514&ptid=278205&ctid=4&oa=true&isxn=9781683182870 [Accessed on 1st Jan 2023].

- Schröer, C., Kruse, F. and Gómez, J.M. (2021) 'A systematic literature review on applying CRISP-DM process model', Procedia Computer Science, 181, pp. 526–534. doi:Available at:[

https://www.sciencedirect.com/science/article/pii/S1877050921002416 [Accessed on 19th December 2023].

- Nick, T.G. and Campbell, K.M. (2007) ‘Logistic regression’, Topics in Biostatistics, pp. 273–301. doi:Available at : https://link.springer.com/protocol/10.1007/978-1-59745-530-5_14#citeas [Accessed on 7th January 2024].

6.Appendix

6.1.R code

```
#load libraries neccessary
library(tidyverse)
library(readxl)
library(dplyr)

#check the current working directory
getwd()

#set the working directory
setwd("/Users/dhanush/Desktop/Bussiness analytics /Stas/Ass 2")

#import the given
#attaching the provided data named "term.xlxs" using "read_excel" function and attaching it
to variable named data
data <- read_excel("term.xlsx")

#decprtive analysis
summary(data) # provided the summary of the whole data
unique(data) # gives us the uniques value in the data
names(data) # provides us the name of all the variables given

#1)descpritve analysis of ID
summary(data$ID) # summarises ID
sum(is.na(data$ID))#finds the total sum of null values in ID

#2)descrptive analysis of age
summary(data$age)#gives mean mode median of the data
sum(is.na(data$age))

# Create a histogram for 'age'
ggplot(data, aes(x = age)) +
  geom_histogram(binwidth = 5, fill = "lightgray", color = "black") +
  labs(title = "Distribution of Age",
       x = "Age",
       y = "Frequency") +
  theme_minimal()


#3)descpritive analysis of occupation
summary(data$occupation)#tells it is catergorical variables
unique(data$occupation)#gives the unique entries
count(data,occupation)#cout of each occurrence
```

```
# Load necessary libraries
library(ggplot2)

# Create a bar plot for 'occupation'
ggplot(data = data, aes(x = occupation)) +
  geom_bar(fill = "lightgray", color = "black") +
  labs(title = "Distribution of Occupation",
       x = "Occupation",
       y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotating x-axis labels for better
readability

#4)descpritive analysis of marital status
summary(data$marital_status)#tells it is catergorical variables
unique(data$marital_status)#gives the unique entries
count(data,marital_status)#cout of each occurrence and there is error NA

# Create a bar plot for 'marital_status'
ggplot(data = data, aes(x = marital_status)) +
  geom_bar(fill = "#008080", color = "black") +
  labs(title = "Distribution of Martial status ",
       x = "marital_status",
       y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotating x-axis labels for better
readability


#5)descpritive analysis of education_level
summary(data$education_level)#tells it is catergorical variables
unique(data$education_level)#gives the unique entries
count(data,education_level)#cout of each occurrence

# Create a bar plot for 'education_level'
ggplot(data = data, aes(x = education_level)) +
  geom_bar(fill = "lightgray", color = "black") +
  labs(title = "Distribution of education_level",
       x = "education_level",
       y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotating x-axis labels for better
readability

#6)descpritive analysis of credit_default
summary(data$credit_default)#tells it is catergorical variables
unique(data$credit_default)#gives the unique entries
count(data,credit_default)#cout of each occurrence

# Create a bar plot for 'credit_default'
```

```
ggplot(data = data, aes(x = credit_default)) +
  geom_bar(fill = "#008080", color = "black") +
  labs(title = "Distribution of credit_default status ",
       x = "credit_default",
       y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotating x-axis labels for better
readability

#7)descpritive analysis of housing_loan
summary(data$housing_loan)#tells it is catergorical variables
unique(data$housing_loan)#gives the unique entries
count(data,housing_loan)#cout of each occurrence

#pie chart for "housing_loan"
ggplot(data = data, aes(x = "", fill = housing_loan)) +
  geom_bar(width = 1, color = "white") +
  coord_polar("y", start = 0) +
  scale_fill_manual(values = c("#D3D3D3", "#008080", "#FFA500")) +  # Light grey, teal, and
another color of choice
  labs(title = "Housing Loan Distribution",
       fill = "Housing Loan",
       x = NULL, y = NULL) +
  theme_void() +
  theme(legend.position = "bottom")

#8)descpritive analysis of personal_loan
summary(data$personal_loan)#tells it is catergorical variables
unique(data$personal_loan)#gives the unique entries
count(data,personal_loan)#cout of each occurrence

#pie chart for "personal loan"
ggplot(data = data, aes(x = "", fill = personal_loan)) +
  geom_bar(width = 1, color = "white") +
  coord_polar("y", start = 0) +
  scale_fill_manual(values = c("#D3D3D3", "#008080", "#FFA500")) +  # Light grey, teal, and
another color of choice
  labs(title = "Personal Loan  Distribution",
       fill = "Personal Loan",
       x = NULL, y = NULL) +
  theme_void() +
  theme(legend.position = "bottom")


#9)descpritive analysis of contact_method
summary(data$contact_method)#tells it is catergorical variables
unique(data$contact_method)#gives the unique entries
count(data,contact_method)#cout of each occurrence
```

```
# Create a bar plot for 'contact_method'
ggplot(data = data, aes(x = contact_method)) +
  geom_bar(fill = "lightgray", color = "black") +
  labs(title = "Distribution of Contact Method",
       x = "Contact method",
       y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotating x-axis labels for better
readability

#10)descpritive analysis of month
summary(data$month)#tells it is catergorical variables
unique(data$month)#gives the unique entries
count(data,month)#cout of each occurrence

# Create a bar plot for 'Month'
ggplot(data = data, aes(x = month)) +
  geom_bar(fill = "#008080", color = "black") +
  labs(title = "Distribution of Month ",
       x = "Month ",
       y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotating x-axis labels for better
readability

#11)descpritive analysis of day_of_week
summary(data$day_of_week)#tells it is catergorical variables
unique(data$day_of_week)#gives the unique entries
count(data,day_of_week)#cout of each occurrence

# Create a bar plot for 'day_of_week'
ggplot(data = data, aes(x = day_of_week)) +
  geom_bar(fill = "lightgray", color = "black") +
  labs(title = "Distribution of day of week ",
       x = "day of week ",
       y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotating x-axis labels for better
readability

#12)descpritive analysis of poutcome
summary(data$poutcome)#tells it is catergorical variables
unique(data$poutcome)#gives the unique entries
count(data,poutcome)#cout of each occurrence

# Create a bar plot for 'poutcome'
ggplot(data = data, aes(x = poutcome)) +
  geom_bar(fill = "#008080", color = "black") +
  labs(title = "Distribution of Poutcome ",
```

```
      x = "Poutcome ",
      y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotating x-axis labels for better
readability


#13)descpritive analysis of subscribed
summary(data$subscribed)#tells it is catergorical variables
unique(data$subscribed)#gives the unique entries
count(data,subscribed)#cout of each occurrence

#pie chart for subscribed
ggplot(data = data, aes(x = "", fill = subscribed)) +
  geom_bar(width = 1, color = "white") +
  coord_polar("y", start = 0) +
  scale_fill_manual(values = c("#D3D3D3", "#008080")) +  # Light grey and teal colors
  labs(title = "Subscription Status",
      fill = "Subscribed",
      x = NULL, y = NULL) +
  theme_void() +
  theme(legend.position = "bottom")

#14)descrptive analysis of contact_duration
summary(data$contact_duration)#gives mean mode median of the data
sum(is.na(data$contact_duration))

#plot of contact_duration
ggplot(data, aes(x = contact_duration)) +
  geom_histogram(binwidth = 100, fill = "#008080", color = "black") +
  labs(title = "Distribution of Contact Duration",
      x = "Contact Duration",
      y = "Frequency") +
  theme_minimal()

#15)descrptive analysis of campaign
summary(data$campaign)#gives mean mode median of the data
sum(is.na(data$campaign))

#plot of campaign
ggplot(data, aes(x = campaign)) +
  geom_histogram(binwidth = 1, fill = "lightgray", color = "black") +
  labs(title = "Distribution of Campaign",
      x = "Campaign",
      y = "Frequency") +
  theme_minimal()

#16)descrptive analysis of pdays
```

```
summary(data$pdays)#gives mean mode median of the data
sum(is.na(data$pdays))

#plot of days
ggplot(data = data, aes(x = pdays)) +
  geom_histogram(binwidth = 100, fill = "#008080", color = "black") +
  labs(title = "Distribution of Pdays",
       x = "Pdays",
       y = "Frequency") +
  theme_minimal()

#17)descrptive analysis of previous_contacts
summary(data$previous_contacts)#gives mean mode median of the data
sum(is.na(data$previous_contacts))

#hist of previous_contacts
ggplot(data = data, aes(x = previous_contacts)) +
  geom_histogram(binwidth = 1, fill = "lightgrey", color = "black") +
  labs(title = "Distribution of Previous Contacts",
       x = "Previous Contacts",
       y = "Frequency") +
  theme_minimal()

#18)descrptive analysis of previous_contacts
summary(data$emp_var_rate)#gives mean mode median of the data
sum(is.na(data$emp_var_rate))

#plot of emp_var_rate
ggplot(data = data, aes(x = emp_var_rate)) +
  geom_histogram(binwidth = 0.5, fill = "#008080", color = "black") +
  labs(title = "Distribution of Employment Variation Rate",
       x = "Employment Variation Rate",
       y = "Frequency") +
  theme_minimal()

#19)descrptive analysis of cons_price_idx
summary(data$cons_price_idx)#gives mean mode median of the data
sum(is.na(data$cons_price_idx))

#plot of emp_var_rate
ggplot(data, aes(x = cons_price_idx)) +
  geom_histogram(binwidth = 0.1, fill = "lightgrey", color = "black") +
  labs(title = "Distribution of Consumer Price Index",
       x = "Consumer Price Index",
       y = "Frequency") +
  theme_minimal()
```

```
#20)descrptive analysis of cons_conf_idx
summary(data$cons_conf_idx)#gives mean mode median of the data
sum(is.na(data$cons_conf_idx))

#plot of cons_conf_idx
ggplot(data = data, aes(x = cons_conf_idx)) +
  geom_histogram(binwidth = 1, fill = "#008080", color = "black") +
  labs(title = "Distribution of Consumer Confidence Index",
       x = "Consumer Confidence Index",
       y = "Frequency") +
  theme_minimal()


#21)descrptive analysis of Euribor_3m
summary(data$euribor_3m)#gives mean mode median of the data
sum(is.na(data$euribor_3m))

#plot of Euribor_3m
ggplot(data, aes(x = euribor_3m)) +
  geom_histogram(binwidth = 1, fill = "lightgrey", color = "black") +
  labs(title = "Distribution of Euribor 3 Month Rate",
       x = "Euribor 3 Month Rate",
       y = "Frequency") +
  theme_minimal()

#22)descrptive analysis of n_employed
summary(data$n_employed)#gives mean mode median of the data
sum(is.na(data$n_employed))

ggplot(data = data, aes(x = n_employed)) +
  geom_histogram(binwidth = 50, fill = "#008080", color = "black") +
  labs(title = "Distribution of Number of Employees",
       x = "Number of Employees",
       y = "Frequency") +
  theme_minimal()


#data cleaning

#1)age
attach(data)#attaching data varible for ease of typing
summary(age)#summmary of age
sum(is.na(age))#to find out if there are any null values

#box plot before data cleanning
ggplot(data)+
```

```
  geom_boxplot(aes(x=age),outlier.colour = "red")
# only one outlier 999 is found out

newdata <- data %>%
  filter(age !=999)
#box plot after data cleaning
ggplot(newdata)+
  geom_boxplot(aes(x=age),outlier.colour = "red")

#2)marital status
#summary statstics before cleaning data
summary(data$marital_status)
count(data,marital_status)

#There are 23 NA values need to replace them with the most occuring variable
newdata<-newdata %>%
  filter(!(is.na(marital_status)))

#summary statstics after cleaning data
summary(newdata$marital_status)
count(newdata,marital_status)

#barplot of variable marital_status
ggplot(newdata) +
  geom_bar(aes(x = marital_status, fill = marital_status))

#3)housing_loan
summary(newdata$housing_loan)#gives us the summary statstics
count(newdata,housing_loan)#gives the count of values of column housing_loan
sum(is.na(newdata$housing_loan))#gives us the sum of na if present

#barplot of variable housing_loan
ggplot(newdata) +
  geom_bar(aes(x = housing_loan, fill = housing_loan))


#4)occupation
summary(newdata$occupation)#gives us the summary statstics
count(newdata,occupation)#gives the count of values of column occupation
sum(is.na(newdata$occupation))#gives if na value is present

#barplot of variable occupation
ggplot(newdata) +
  geom_bar(aes(x = occupation, fill = occupation)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
#5)education level
summary(newdata$education_level)#gives us the summary statstics
count(newdata,education_level)#gives the count of values of column education level
sum(is.na(newdata$education_level))# tell us if there is na value present

#barplot of variable education level
ggplot(newdata) +
  geom_bar(aes(x = education_level, fill = education_level)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

#6)personal loan
summary(newdata$personal_loan)#gives us the summary stastics
count(newdata,personal_loan)#gives unique values or count of column of personal_loan
sum(is.na(newdata$personal_loan))#tell us if there is na value present

#barplot of variable personal loan
ggplot(newdata) +
  geom_bar(aes(x = personal_loan, fill = personal_loan))

#7)month
summary(newdata$month)#gives us the summary statstics
count(newdata,month)#gives the count of values inside the col of month
# there are two jul and july we need to keep jul
sum(is.na(newdata$personal_loan))

#barplot of variable month
ggplot(newdata) +
  geom_bar(aes(x = month, fill = month))

#rename the july to jul as others are also july
newdata <- newdata %>%
  mutate(month = ifelse(month == "july", "jul", month))

summary(newdata$month)#gives us the summary statstics
count(newdata,month)#gives us the count of month

#8)days_of_week
summary(newdata$day_of_week)#gives us the summary stastics
count(newdata,day_of_week)#gives the count of values inside col of day_of_week
#here there is a two variables tue and tues we need change Tues to tue
sum(is.na(newdata$day_of_week))# tells if there is na values present

#rename tues to tue
newdata <- newdata %>%
  mutate(day_of_week = ifelse(day_of_week == "tues", "tue", day_of_week ))

summary(newdata$day_of_week)#gives us the summary stastics
```

```
count(newdata,day_of_week)#gives the count of values inside col of day_of_week

#barplot of day_of_week
ggplot(newdata) +
  geom_bar(aes(x = day_of_week, fill = day_of_week))

#9)campaign
summary(newdata$campaign)#gives us the summary stastics
count(newdata,campaign)#gives the count of values  inside col of campaign
sum(is.na(newdata$campaign))#tells us if there is na values present

#box plot of campaign
ggplot(newdata)+
  geom_boxplot(aes(x=campaign),outlier.colour = "red")

#10)pdays
summary(newdata$pdays)#Gives us summary stats
count(newdata,pdays)#gives us the values inside pdays
sum(is.na(newdata$pdays))#tell is there any na values present

#boxplot of pdays
ggplot(newdata)+
  geom_boxplot(aes(x=pdays),outlier.colour = "red")

#11)poutcome
summary(newdata$poutcome)#gives us the summary stats
count(newdata,poutcome)#gives us the count of values in poutcome
sum(is.na(newdata$poutcome))#tells us is there is any na values

#barplot of poutcome
ggplot(newdata)+
  geom_bar(aes(x=poutcome,fill=poutcome))

#12) Credit_deafult
summary(newdata$credit_default)#gives us the summary stats
count(newdata,credit_default)#gives us the count of values in credit_default
sum(is.na(newdata$credit_default))#tells if there is any null vales

#barplot of credit_deafult
ggplot(newdata)+
  geom_bar(aes(x=credit_default,fill=credit_default))

#13)Contact_method
summary(newdata$contact_method)#gives us the summary stats
count(newdata,contact_method)#gives us the count of values in contact_method
sum(is.na(newdata$contact_duration))#tells if there is any null vales
```

```
#plot of contact method
ggplot(newdata)+
  geom_bar(aes(x=contact_method,fill=contact_method))

#14)n_employed
summary(newdata$n_employed)#gives us the summary stats
count(newdata,n_employed)#gives us the count of values in N_employed
sum(is.na(newdata$n_employed))#tell us if there is any null values

#plot of n_empployed
ggplot(newdata)+
  geom_histogram(aes(x=n_employed,bins=10,fill=n_employed))

#15)Euribor_3m
summary(newdata$euribor_3m)#gives us the summary stats
count(newdata,euribor_3m)#gives us the count of values in Euribor_3m
sum(is.na(newdata$euribor_3m))#tell us if there is any null values

#plot of Euribor_3m
ggplot(newdata)+
  geom_histogram(aes(x=euribor_3m))

#convert all as factor to numeric
newdata <- newdata %>%
  mutate_if(is.character,as.factor)


#Hypothesis Testing

#Personal loan, Pday, Occupation, Credit default, campaign


#1) Personal Loan

chisq.test(newdata$personal_loan, newdata$subscribed)
#Accepting the null hypothesis

# Create a contingency table
cont_table <- table(newdata$personal_loan, newdata$subscribed)

# Convert the contingency table to a dataframe for ggplot
df_for_plot <- as.data.frame(cont_table)

# Rename the columns for clarity
names(df_for_plot) <- c("PersonalLoan", "Subscribed", "Count")

# Create the bar plot
```

```
library(viridis)

# Define a color palette suitable for a white background
color_palette <- viridis_pal(option = "A", direction = 1)(3)  # Adjust the number '3' based on
the number of categories

# Plot with the chosen color palette
ggplot(df_for_plot, aes(x = PersonalLoan, y = Count, fill = Subscribed)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  labs(title = "Bar Plot of Personal Loan and Subscription Status",
       x = "Personal Loan",
       y = "Count") +
  scale_fill_manual(values = color_palette)



#2)pdays
t.test(pdays ~ subscribed, data = newdata)
#Rejecting the null hypothesis

#plot
ggplot(newdata, aes(x = subscribed, y = pdays, fill = subscribed)) +
  geom_boxplot(color = "black", fill = "#008080") +  # Green color for the boxplot
  geom_point(position = position_jitterdodge(), alpha = 0.5, size = 2, color = "#800080") +  #
Purple color for the swarm plot
  labs(x = "Subscribed", y = "Days Since Last Contact (pdays)",
       title = "Relationship between pdays and Subscription Status")

#extra violin graph
#appendix
ggplot(newdata, aes(x = subscribed, y = pdays, fill = subscribed)) +
  geom_violin() +
  labs(x = "Subscribed", y = "Days Since Last Contact (pdays)",
       title = "Relationship between pdays and Subscription Status")

#3) Occupation

chisq.test(newdata$occupation, newdata$subscribed)

ggplot(newdata, aes(x = occupation, fill = subscribed)) +
  geom_bar(position = "dodge") +
  labs(x = "Occupation", y = "Count", title = "Relationship between Occupation and
Subscription Status") +
  scale_fill_manual(values = c("#800080", "#008080"),
                    name = "Subscribed",
                    labels = c("No", "Yes")) +
```

```
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotating x-axis labels for better
readability


#Rejecting the null hypothesis

#4) Credit Default

chisq.test(newdata$credit_default, newdata$subscribed)

ggplot(newdata, aes(x = credit_default, fill = subscribed)) +
  geom_bar(position = "dodge", color = "black") +
  labs(x = "Credit Default", y = "Count", title = "Relationship between Credit Default and
Subscription Status") +
  scale_fill_manual(values = c("#800080", "#008080"),
                    name = "Subscribed",
                    labels = c("No", "Yes")) +
  theme_minimal()


# Rejecting the null hypothesis


#5) Campaign VS subscribed

t.test(campaign ~ subscribed, data = newdata)
# Assuming your dataset is named 'data'
library(ggplot2)

# Creating a boxplot with specified colors
ggplot(newdata, aes(x = subscribed, y = contact_duration, fill = subscribed)) +
  geom_boxplot() +
  labs(x = "Subscribed", y = "Contact Duration", title = "Relationship between Subscribed and
Contact Duration") +
  scale_fill_manual(values = c("#800080", "#008080"))

# Rejecting the null hypothesis

#model building
library(caret)

#set seed
set.seed(40412492)

#split the data set in training and test

index <- createDataPartition(newdata$subscribed,times = 1 ,p =0.8 ,list = FALSE)
```

```
train_data <- newdata[index,]
test_data <- newdata[-index,]

#Create to function to calculate R2
logisticPseudoR2s <- function(LogModel) {
  dev <- LogModel$deviance
  nullDev <- LogModel$null.deviance
  modelN <- length(LogModel$fitted.values)
  R.l <- 1 - dev / nullDev
  R.cs <- 1- exp ( -(nullDev - dev) / modelN)
  R.n <- R.cs / ( 1 - ( exp (-(nullDev / modelN))))
  cat("Pseudo R^2 for logistic regression\n")
  cat("Hosmer and Lemeshow R^2 ", round(R.l, 3), "\n")
  cat("Cox and Snell R^2 ", round(R.cs, 3), "\n")
  cat("Nagelkerke R^2 ", round(R.n, 3), "\n")
}

#1)using only hypothesis variables

model_1 <- glm(subscribed ~ pdays+occupation+credit_default+campaign,data =
train_data,family ="binomial")

#predictions of model 1
pred_1 <- predict(model_1,test_data,type = "response")
head(pred_1)
class_pred_hypo_1 <- as.factor(ifelse(pred_1 >.5 ,"yes","no"))
postResample(class_pred_hypo_1,test_data$subscribed)
confusionMatrix(data = class_pred_hypo_1,test_data$subscribed)


#to find the summary of model_1
summary(model_1)

#to find the R2 Model_1
logisticPseudoR2s(model_1)

#2)Hypothesis variables+ literature backed variables
model_2 <- glm(subscribed ~
pdays+occupation+credit_default+campaign+age+marital_status+education_level+day_of_
week+month,data = train_data,family ="binomial")
pred_model_2 <- predict(model_2,test_data,type = "response")
head(pred_model_2)
class_pred_hypo_2 <- as.factor(ifelse(pred_model_2 >.5 ,"yes","no"))
postResample(class_pred_hypo_2,test_data$subscribed)

confusionMatrix(data = class_pred_hypo_2,test_data$subscribed)
```

```
#to find the summary of model_2
summary(model_2)

#to find the R2 Model_2
logisticPseudoR2s(model_2)


#3)All variables
model_3 <- glm(subscribed ~
pdays+occupation+credit_default+campaign+age+marital_status+education_level+day_of_
week+month+poutcome+housing_loan+contact_duration+n_employed+euribor_3m,data =
train_data,family ="binomial")
pred_model_3 <- predict(model_3,test_data,type = "response")
head(pred_model_3)
class_pred_hypo_3 <- as.factor(ifelse(pred_model_3 >.5 ,"yes","no"))
postResample(class_pred_hypo_3,test_data$subscribed)

confusionMatrix(data = class_pred_hypo_3,test_data$subscribed)


#to find the summary of model_1
summary(model_3)

#to find the R2 Model_3
logisticPseudoR2s(model_3)

install.packages("stargazer")
library(stargazer)

stargazer(model_1, model_2, model_3,
        type = "html", title = "Logistic Regression Models Summary",
        out = "models_summary.html",
        column.labels = c("Model 1", "Model 2", "Model 3"),
        dep.var.caption = "Dependent Variable: Subscribed",
        model.numbers = FALSE,
        notes.label = "Significance Level",
        report = "vc*")

# View generated HTML file
browseURL("models_summary.html")

install.packages("pROC")  # Install pROC package if not installed
library(pROC)


# Predict probabilities for test data
```

```
pred_probs <- predict(model_3, test_data, type = "response")

# Create ROC curve
roc_curve <- roc(test_data$subscribed, pred_probs)

# Plot ROC curve
plot(roc_curve, col = "blue", lwd = 2, print.auc = TRUE, print.auc.y = 0.2, print.auc.x = 0.7)

#plot model 3

plot(model_3) #ploting model graph

#Assumptions
library(car)

vif(model_3)
```

6.2 Extra Visualisation

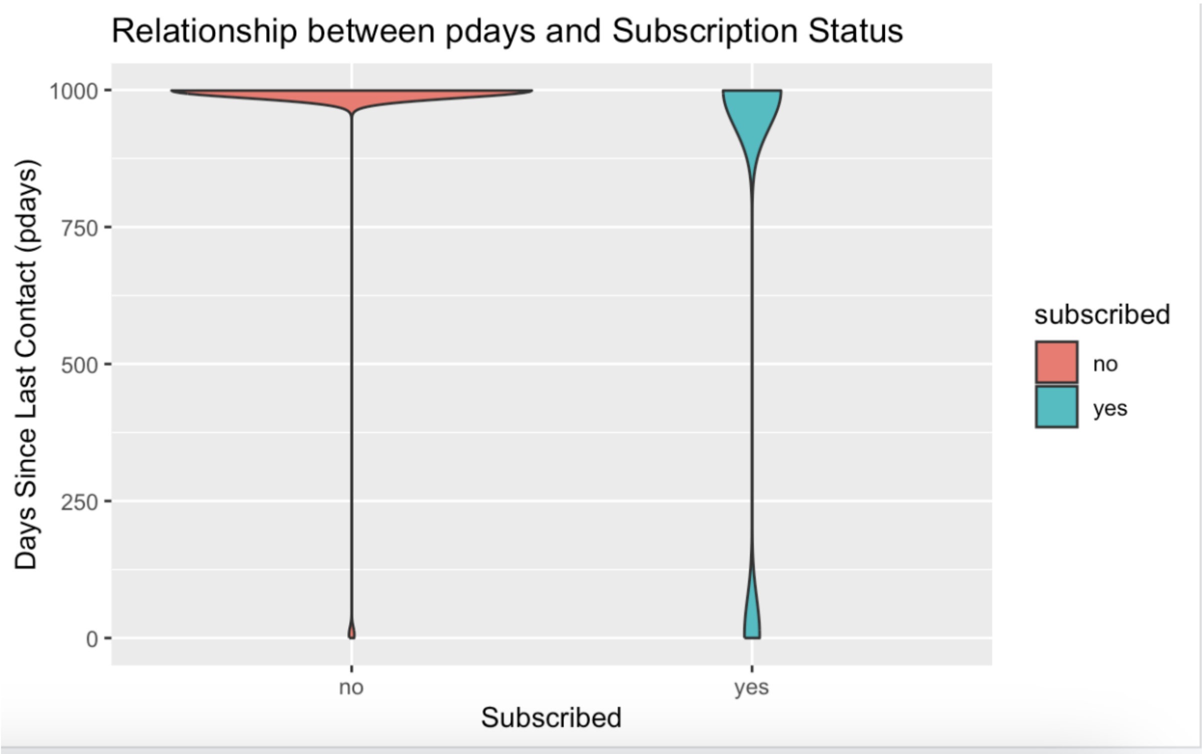


Fig 6.2: Violin graph of pdays and Subscription Status