Improving marketing strategies: A comprehensive analysis and logistic regression model for bank term deposit subscription

Jaime Rubio Diaz

40425150

2200 words

Table of content

1. Introduction

This project aims to develop a logistic regression model in R to predict subscriptions to a term deposit, utilizing a dataset derived from a bank telemarketing campaign. At the same time, a descriptive analysis of the data will be carried out to identify underlying patterns. The primary goal is to uncover influential factors associated with customers' subscription to term deposits. This will provide valuable insights to the bank's marketing team, enabling them to optimize marketing strategies and enhance communication effectiveness.

A term deposit is a fixed-term financial investment with a predetermined duration, usually with a financial institution, that restricts withdrawals until the end of the term and usually offers higher interest rates compared to easily accessible savings accounts (Chen, 2022). Numerous variables are associated with the likelihood of subscribing to a term deposit, the following are the expected relationships.

H1: There will be a positive relationship between the **age** and the **likelihood of subscribing** to the term deposit.

Age demonstrates a significant correlation with investment behaviour, younger individuals, particularly those aged 16-25, tend to prefer short-term investments, such as trading, while older age groups show a clear preference for long-term investments (Charles and Kasilingam, 2013). Given that term deposits can range from one month to up to ten years, it is anticipated that the age groups 26-35 and 35-45 will be the predominant customers for this type of investment.

H2: There will be a positive relationship between **married** individuals and their **likelihood of subscribing** to the term deposit.

Targeting married individuals could be particularly effective for these telemarketing campaigns, as they prioritize securing their family's future, including planning for their children's education, their own retirement, and preparing for unexpected events (Asare-Frempong and Jayabalan, 2017). The stability often associated with married life is therefore expected to be a crucial factor in the decision to subscribe to a term deposit.

H3: There will be a positive relationship between the **consumer confidence index** and the **likelihood of subscribing** to the term deposit.

The CCI, an annual survey that assesses consumers' views on the economic situation, has shown a clear trend: an increase in the CCI tends to correlate with an increase in consumer spending and investment activities (Kandah, 2023). This suggests that higher consumer confidence may lead to a greater propensity for long-term financial commitments, such as term deposits.

H4: There will be a negative relationship between the **Euribor 3-month rate** and the **likelihood of subscribing** to the term deposit.

The Euribor is a European benchmark interest rate. Typically, higher interest rates lead to increased borrowing costs, prompting consumers to save rather than spend; in contrast, lower interest rates often encourage spending and investment (Maverick, 2023). This inverse relationship suggests that as the Euribor 3-month rate rises, the attractiveness of subscribing to term deposits may decrease.

2. Methodology

This data mining project involves five key tasks. (1) Data Understanding, (2) Data Cleaning, (3) Descriptive Analysis, (4) Model Development and (5) Evaluation of Model Performance using test data.

The data set, named “term”, includes a target variable ("subscribed”), which indicates whether a customer subscribed to the term deposit after the marketing campaign.

Customers Characteristics		
ID	Unique identifier for each customer	Numerical
age	Age of the customer	Numerical
occupation	Occupation or job title of the customer	Categorical
marital_status	Marital status of the customer	Categorical
education_level	Educational qualifications of the customer	Categorical
credit_default	Presence of credit default	Binary (Yes/No)
house_loan	Presence of a housing loan	Binary (Yes/No)
personal_loan	Presence of a personal loan	Binary (Yes/No)
Most Recent Contact Characteristics		
contact_method	Method of last contact	Categorical
month	Last contact month	Categorical
day_of_week	Last contact day of the week	Categorical
contact_duration	Duration of the last contact	Numerical
Additional Variables		
campaign	Number of contacts during the campaign	Numerical
pdays	Number of days after the previous campaign	Numerical
previous_contacts	Number of contacts before this campaign	Numerical
poutcome	Outcome of the previous campaign	Categorical
Economic and social variables		
emp_var_rate	Employment variation rate	Numerical
cons_price_idx	Consumer price index	Numerical
cons_conf_idx	Consumer confidence index	Numerical
Euribor_3m	Euribor 3 month rate	Numerical
nr_employed	Number of employees	Numerical
Target Variable		
subscribed	Wheter the customer subscribed or not	Binary (Yes/No)

Table 1. Variables in the dataset

The analysis of the machine learning models' results, both pre- and post-data cleaning, indicates that the data cleaning process is crucial for enhancing the overall quality and clarity of the models (Chai, 2020).

In data analysis, a point that significantly deviates from the common range of a variable is identified as an 'outlier', according to Osborne and Overbay (2019), there is a predominant argument in favour of modifying them.

Variable	Detection Method	Outliers Eliminated	New Range
age	Box Plot	2	17-95
campaign	Box Plot	18	1-33

Table 2. Outliers

There are 23 missing values in the marital status variable, which have been converted to the most frequently occurring level ('married'). Missing values originating from the removal of outliers are excluded from the model, as techniques to impute them are tested but do not improve the accuracy of the final model.

Values that were stored incorrectly		
Variable	**Incorrect Value**	**Corrected Value**
month	“july”	“jul”
day_of_week	“tues”	“tue”

Table 3. Incorrect values

For this study, a multiple logistic regression model will be employed. A logistic regression model is a statistical approach that seeks to establish the connection between a categorical dependent variable (usually binary, the presence of a disease or not) and an independent variable (Campbel and Nick, 2007).

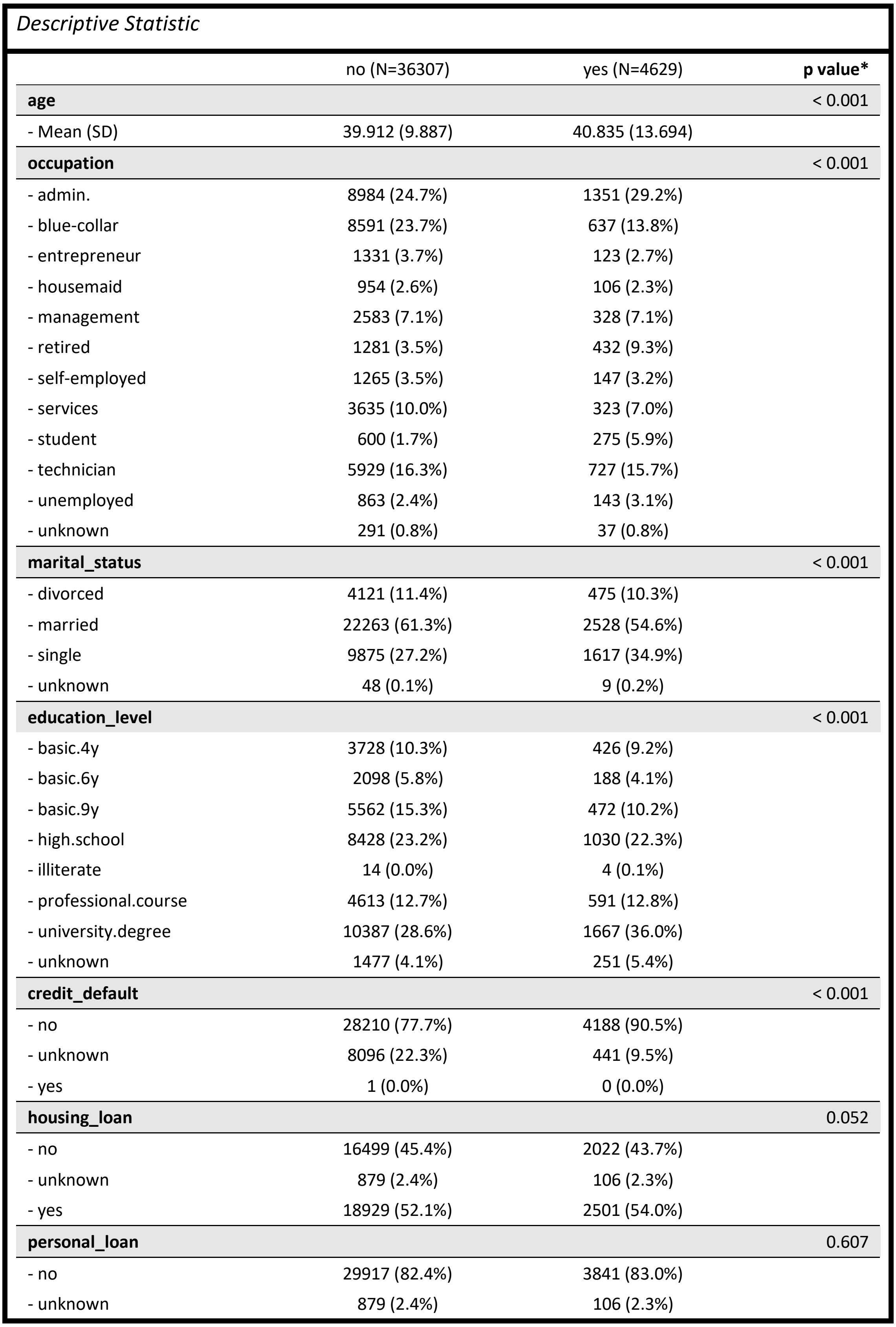
To build the model in R, the data will be split into train (80%) and test (20%). The model will be built using the training dataset, and its accuracy will be evaluated using the test dataset. Throughout the process of constructing multiple models to identify the most accurate one, Pseudo R-squared and AIC will be employed to measure their performance. Once the best model is selected, there are some metrics to evaluate its accuracy with the test dataset:

Accuracy	Percentage of the observations that were predicted correctly
Kappa	A measure of accuracy that assesses the proportion of values classified correctly, considering the probability of a random match (Freeman and Moisen, 2008).
Confusion Matrix	A matrix showing what the model predicted compared to actual values.
AUC-ROC Curve	The curve visualizes the balance between the probability of true positives and false positives (Gorsevski et. al., 2006).

Table 4. Metrics to evaluate model performance

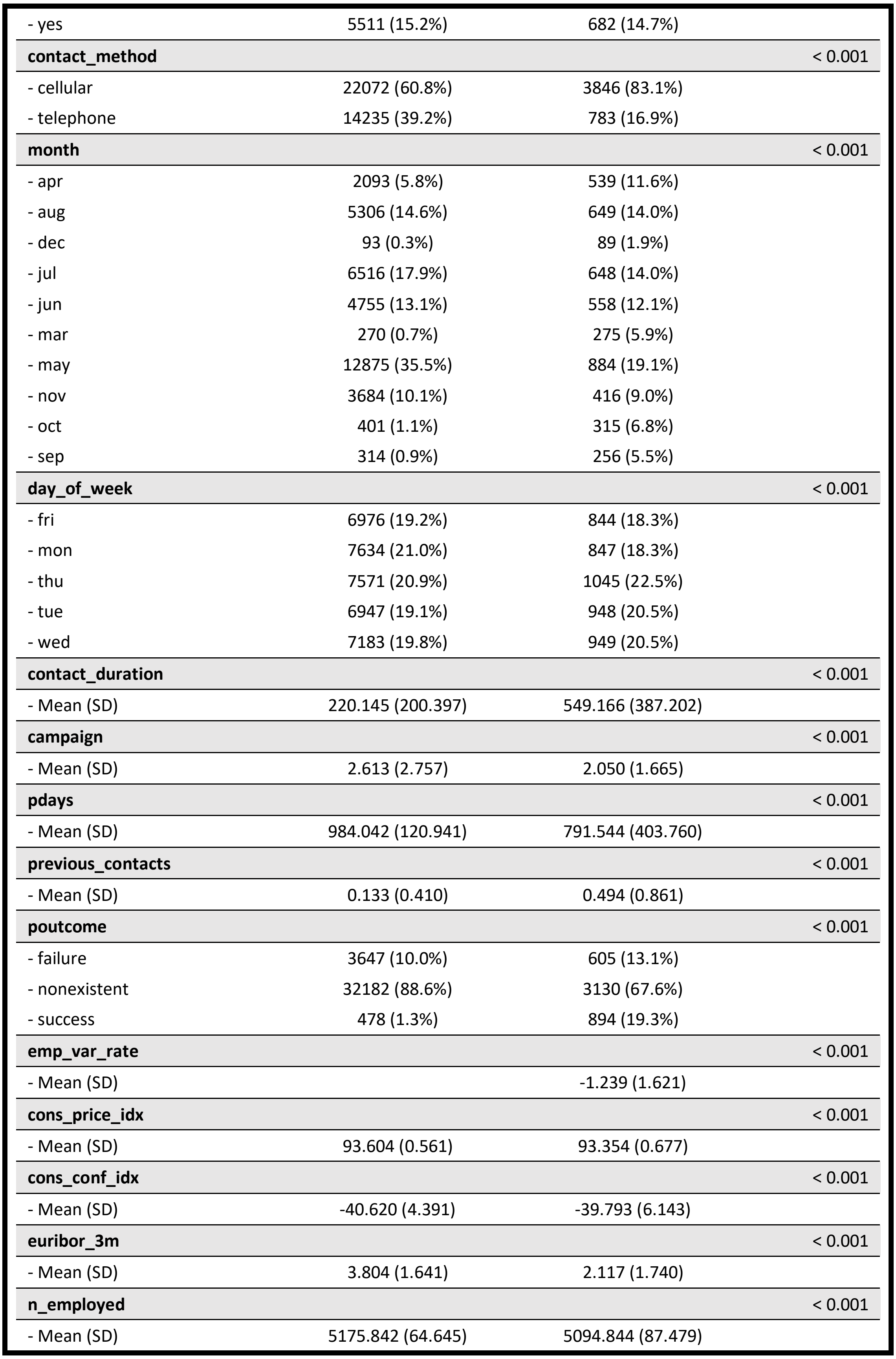
3. Results

Descriptive Analysis



Descriptive Statistic

	no (N=36307)	yes (N=4629)	**p value***
age			< 0.001
- Mean (SD)	39.912 (9.887)	40.835 (13.694)	
occupation			< 0.001
- admin.	8984 (24.7%)	1351 (29.2%)	
- blue-collar	8591 (23.7%)	637 (13.8%)	
- entrepreneur	1331 (3.7%)	123 (2.7%)	
- housemaid	954 (2.6%)	106 (2.3%)	
- management	2583 (7.1%)	328 (7.1%)	
- retired	1281 (3.5%)	432 (9.3%)	
- self-employed	1265 (3.5%)	147 (3.2%)	
- services	3635 (10.0%)	323 (7.0%)	
- student	600 (1.7%)	275 (5.9%)	
- technician	5929 (16.3%)	727 (15.7%)	
- unemployed	863 (2.4%)	143 (3.1%)	
- unknown	291 (0.8%)	37 (0.8%)	
marital_status			< 0.001
- divorced	4121 (11.4%)	475 (10.3%)	
- married	22263 (61.3%)	2528 (54.6%)	
- single	9875 (27.2%)	1617 (34.9%)	
- unknown	48 (0.1%)	9 (0.2%)	
education_level			< 0.001
- basic.4y	3728 (10.3%)	426 (9.2%)	
- basic.6y	2098 (5.8%)	188 (4.1%)	
- basic.9y	5562 (15.3%)	472 (10.2%)	
- high.school	8428 (23.2%)	1030 (22.3%)	
- illiterate	14 (0.0%)	4 (0.1%)	
- professional.course	4613 (12.7%)	591 (12.8%)	
- university.degree	10387 (28.6%)	1667 (36.0%)	
- unknown	1477 (4.1%)	251 (5.4%)	
credit_default			< 0.001
- no	28210 (77.7%)	4188 (90.5%)	
- unknown	8096 (22.3%)	441 (9.5%)	
- yes	1 (0.0%)	0 (0.0%)	
housing_loan			0.052
- no	16499 (45.4%)	2022 (43.7%)	
- unknown	879 (2.4%)	106 (2.3%)	
- yes	18929 (52.1%)	2501 (54.0%)	
personal_loan			0.607
- no	29917 (82.4%)	3841 (83.0%)	
- unknown	879 (2.4%)	106 (2.3%)	



- yes	5511 (15.2%)	682 (14.7%)	
contact_method			< 0.001
- cellular	22072 (60.8%)	3846 (83.1%)	
- telephone	14235 (39.2%)	783 (16.9%)	
month			< 0.001
- apr	2093 (5.8%)	539 (11.6%)	
- aug	5306 (14.6%)	649 (14.0%)	
- dec	93 (0.3%)	89 (1.9%)	
- jul	6516 (17.9%)	648 (14.0%)	
- jun	4755 (13.1%)	558 (12.1%)	
- mar	270 (0.7%)	275 (5.9%)	
- may	12875 (35.5%)	884 (19.1%)	
- nov	3684 (10.1%)	416 (9.0%)	
- oct	401 (1.1%)	315 (6.8%)	
- sep	314 (0.9%)	256 (5.5%)	
day_of_week			< 0.001
- fri	6976 (19.2%)	844 (18.3%)	
- mon	7634 (21.0%)	847 (18.3%)	
- thu	7571 (20.9%)	1045 (22.5%)	
- tue	6947 (19.1%)	948 (20.5%)	
- wed	7183 (19.8%)	949 (20.5%)	
contact_duration			< 0.001
- Mean (SD)	220.145 (200.397)	549.166 (387.202)	
campaign			< 0.001
- Mean (SD)	2.613 (2.757)	2.050 (1.665)	
pdays			< 0.001
- Mean (SD)	984.042 (120.941)	791.544 (403.760)	
previous_contacts			< 0.001
- Mean (SD)	0.133 (0.410)	0.494 (0.861)	
poutcome			< 0.001
- failure	3647 (10.0%)	605 (13.1%)	
- nonexistent	32182 (88.6%)	3130 (67.6%)	
- success	478 (1.3%)	894 (19.3%)	
emp_var_rate			< 0.001
- Mean (SD)		-1.239 (1.621)	
cons_price_idx			< 0.001
- Mean (SD)	93.604 (0.561)	93.354 (0.677)	
cons_conf_idx			< 0.001
- Mean (SD)	-40.620 (4.391)	-39.793 (6.143)	
euribor_3m			< 0.001
- Mean (SD)	3.804 (1.641)	2.117 (1.740)	
n_employed			< 0.001
- Mean (SD)	5175.842 (64.645)	5094.844 (87.479)	

Table 5. Descriptive Statistics

* The p-value for numerical variables was calculated using ANOVA, while for categorical variables, the chi-squared test (chisq.test) was employed.

Statistical significance is found in variables such as age, occupation, marital status, education level, and credit default. However, variables like 'housing_loan' and 'personal_loan' do not show statistical significance. This finding contradicts the assertions of Colaianni et al. (2016), who suggested that loan status is a major factor influencing subscription decisions. When analysing the occupations of the subscribers, a notable majority fall into the category of 'Admin', with 29.2% (Figure 1). Regarding the marital status, a significant portion of subscribers were married, accounting for 54.6% (Figure 2), which corroborates H2. Finally, the majority of subscribers, 90.5%, had no credit default history (Figure 4). The presence of a default often indicates financial instability or a history of poor credit management, which could reduce a person's willingness to commit money to a term deposit.

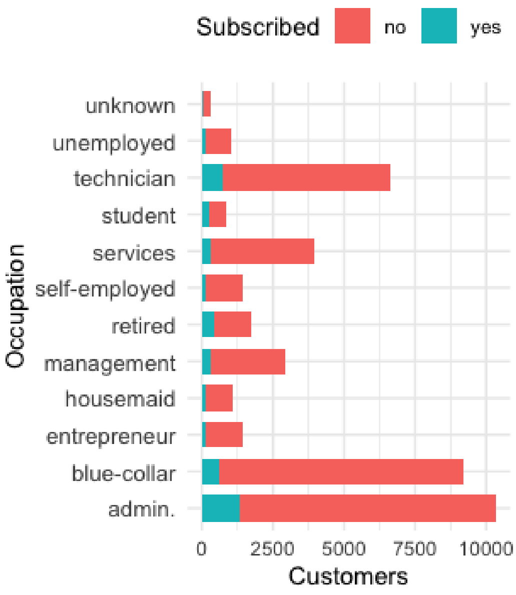


Figure 1. Occupation

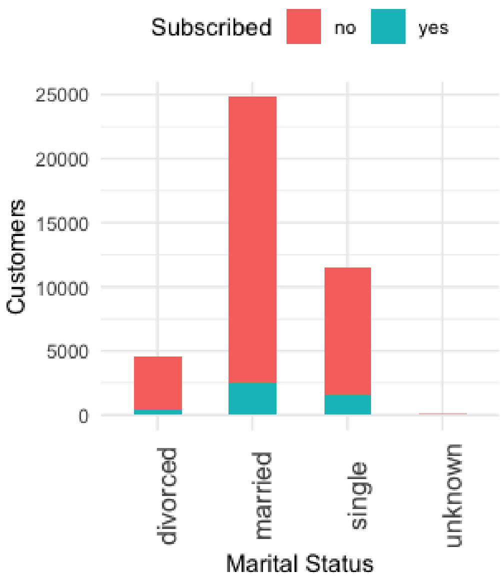


Figure 2. Marital Status

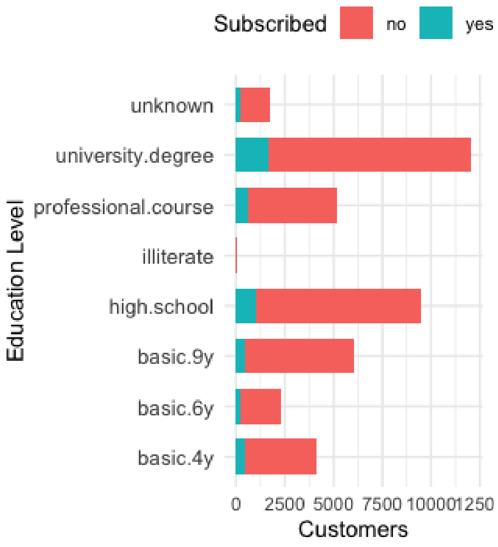


Figure 3. Education Level

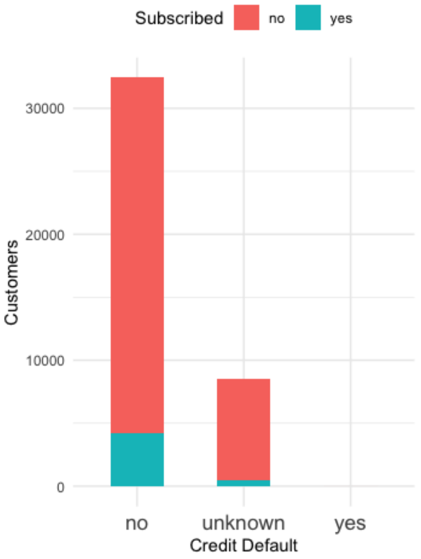


Figure 4. Credit Default

The analysis reveals that all variables pertaining to the last customer contact during the marketing campaign, as well as other variables, exhibit statistical significance. Focusing on the method of contact, the data indicates that most customers, 60.8% of non-subscribers and 83.1% of subscribers, were contacted via cellular (Figure 8). Additionally, a significant proportion of subscribers were successfully contacted in May (19.1%), August (14.0%), and July (14.0%) (Figure 6). The data also demonstrates that subscribers, on average, had fewer contacts (mean of 2.05) compared to non-subscribers (mean of 2.613) (Figure 7). Finally, by examining the results of the previous campaigns ("poutcome"), it is found that most of new subscribers (67.6%) in this campaign were customers who had not been contacted in previous campaigns.

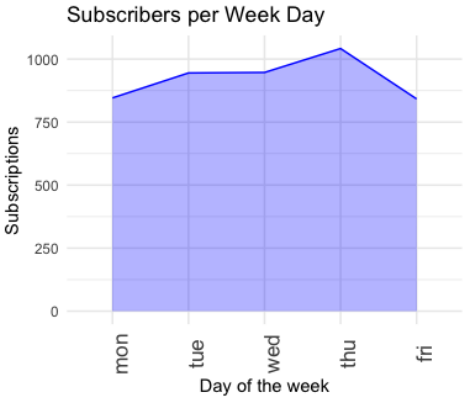


Figure 5. Day of the week

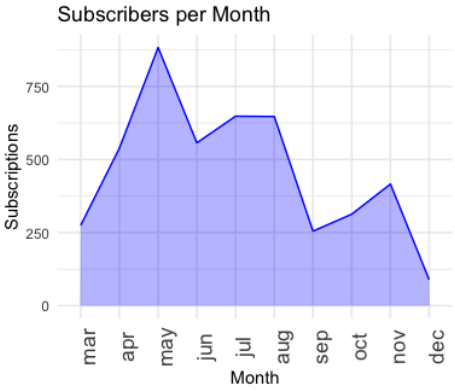


Figure 6. Month

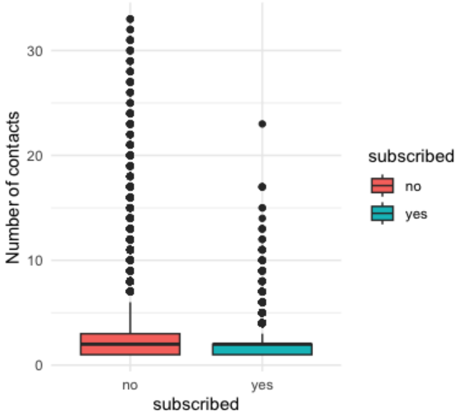


Figure 7. Number of contacts during campaign

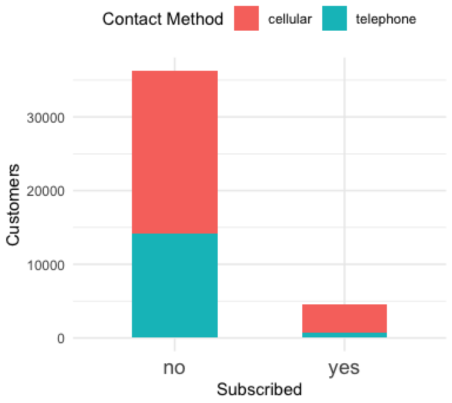


Figure 8. Contact Method

All of the economic and social variables demonstrate statistical significance. It is observed that subscribers are more likely during periods of declining employment rates. Regarding the consumer price index (cons_price_idx), subscribers had an average index of 93.354, slightly lower than the average of 93.604 for non-subscribers. In terms of the consumer confidence index (cons_conf_idx), subscribers were exposed to a higher confidence index during the campaign, aligning with H3. Finally with respect to the 3-month Euribor rate (euribor_3m), subscribing customers had an average rate of 2.117, which is lower than the average rate of 3.804 for non-subscribers.

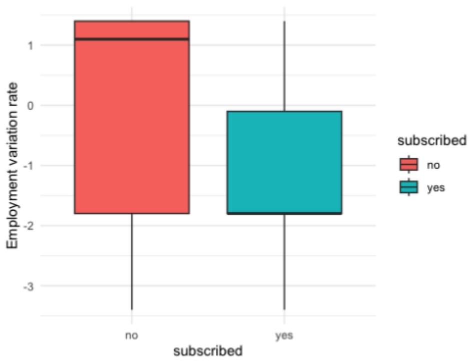


Figure 9. Employment variation rate

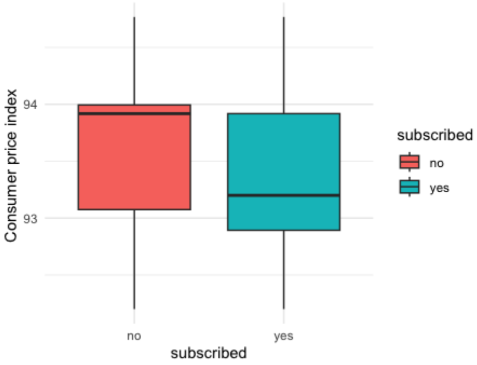


Figure 10. Consumer Price Index

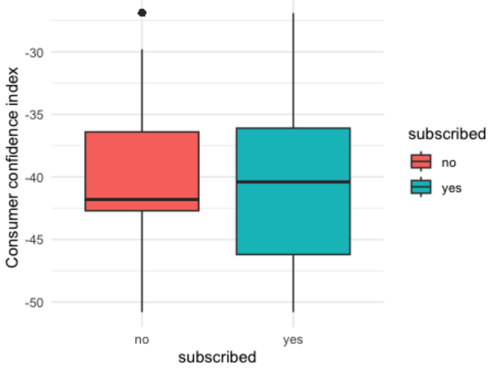


Figure 11. Consumer Confidex Index

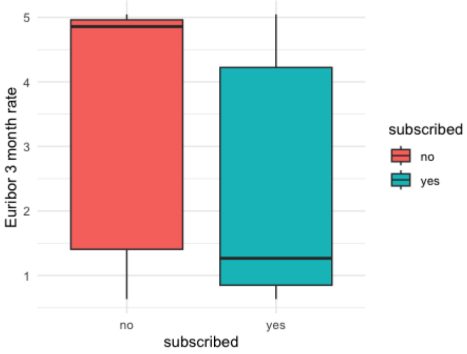


Figure 12. Euribor

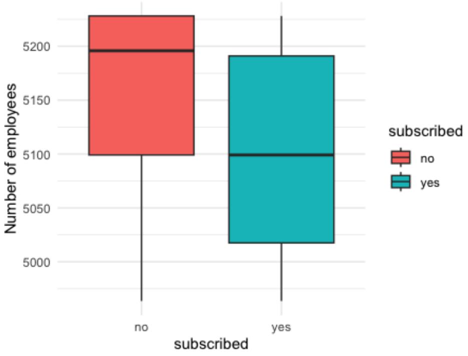


Figure 13. Number of employees

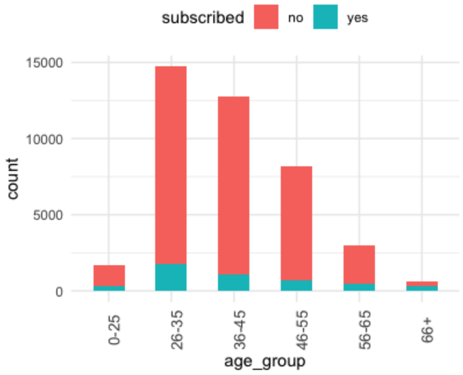


Figure 14. Age group

An age group is formed to examine Hypothesis 1 more closely. As illustrated in the Figure 14, the age groups that were contacted most frequently, and consequently had the highest number of subscribers, were those aged 26 to 45. This observation aligns well with H1.

Model Development

```
Call:
glm(formula = formula4, family = "binomial", data = train)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -2.774e+01  4.042e+00  -6.864 6.69e-12 ***
day_of_weekmon        -1.797e-01  6.362e-02  -2.824 0.004740 **
day_of_weekthu         8.393e-02  6.129e-02   1.369 0.170855
day_of_weektue         8.098e-02  6.346e-02   1.276 0.201947
day_of_weekwed         1.557e-01  6.272e-02   2.482 0.013059 *
occupationblue-collar  -3.100e-01  6.180e-02  -5.015 5.30e-07 ***
occupationentrepreneur  -2.449e-01  1.208e-01  -2.027 0.042641 *
occupationhousemaid    -1.837e-01  1.357e-01  -1.354 0.175868
occupationmanagement   -9.733e-02  8.132e-02  -1.197 0.231361
occupationretired       2.426e-01  8.162e-02   2.972 0.002956 **
occupationself-employed -1.139e-01  1.130e-01  -1.008 0.313439
occupationservices     -2.265e-01  7.847e-02  -2.886 0.003901 **
occupationstudent       2.376e-01  1.023e-01   2.322 0.020228 *
occupationtechnician   -1.023e-01  6.107e-02  -1.675 0.093939 .
occupationunemployed    6.228e-02  1.176e-01   0.530 0.596292
occupationunknown      -1.490e-01  2.277e-01  -0.654 0.512857
contact_methodtelephone -2.455e-01  6.045e-02  -4.062 4.87e-05 ***
campaign              -4.194e-02  1.004e-02  -4.175 2.97e-05 ***
monthaug               4.103e-01  8.415e-02   4.876 1.08e-06 ***
monthdec               9.265e-01  1.897e-01   4.885 1.04e-06 ***
monthjul              4.686e-01  8.584e-02   5.458 4.80e-08 ***
monthjun               3.375e-01  8.670e-02   3.892 9.93e-05 ***
monthmar               1.209e+00  1.168e-01  10.353  < 2e-16 ***
monthmay              -5.807e-01  7.098e-02  -8.181 2.82e-16 ***
monthnov               1.551e-01  8.995e-02   1.724 0.084754 .
monthoct               6.832e-01  1.081e-01   6.318 2.64e-10 ***
monthsep               3.969e-01  1.180e-01   3.364 0.000769 ***
euribor_3m            -5.153e-01  1.689e-02 -30.507  < 2e-16 ***
cons_price_idx         3.072e-01  4.338e-02   7.081 1.44e-12 ***
pdays               -1.456e-03  7.235e-05 -20.126  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 23133  on 32759  degrees of freedom
Residual deviance: 18473  on 32730  degrees of freedom
AIC: 18533

Number of Fisher Scoring iterations: 6
```

Table 6. Logistic Regression Model Output

The table above shows the best performing model among all those constructed. This model holds the lowest AIC (Table 14) and the highest p-pseudo R^2 (Table 15).

Customers contacted on a Wednesday (OR=1.16) are more likely to subscribe than those contacted on a Friday. However, compared to Friday, those who were contacted on a Monday

(OR=0.83) are less likely to subscribe. The odds ratio (OR) quantifies the probability of an outcome occurring in relation to a specific exposure and it is commonly used in logistic regression (Szumilas, 2010). For instance, in this context, customers contacted on Wednesday have 1.16 times higher odds of subscribing compared to those contacted on Friday.

Compared to customers working as administrators, those employed in blue-collar roles (OR=0.74), services (OR=0.78), as entrepreneurs (OR=0.81), or as technicians (OR=0.90) are less likely to subscribe. In contract, customers who are retired (OR=1.32) or student (OR=1.33) show a significantly higher likelihood of subscribing compared to administrators.

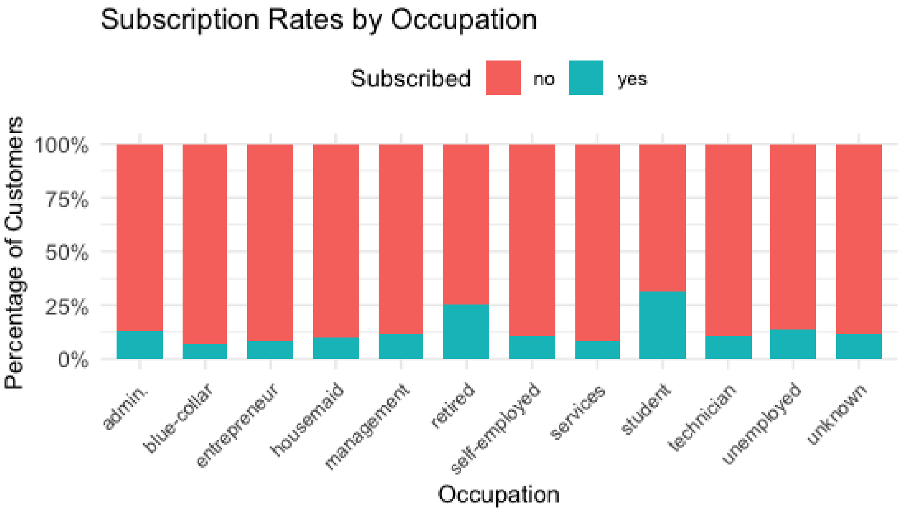


Figure 15. Subscription Rates by Occupation

Customers contacted by telephone (OR=0.78) are less likely to subscribe compared to those contacted by cellular.

The fewer contacts made during the campaign (OR=0.95), the more likely a customer is to subscribe to the term deposit. Similarity, the fewer days that have passed since the customer was last contacted (OR=0.99), the higher the likelihood of subscription. These findings corroborate the results of Choi and Choi (2022), who observed similar relationships using a random forest algorithm.

Compared to customers who were contacted in April, those contacted in May (OR=0.55) are less likely to subscribed. In contrast, customers contacted in August (OR=1.5), December (OR=2.52), July (OR=1.59), June (1.40), March (OR=3.35), November (OR=1.16), September (OR=1.48), and October (OR=1.98) are significantly more likely to subscribe than those contacted in February. While this finding contradicts the observations made by Choi and Choi (2022, it is supported by Xie et. al. (2023), who identified the month of contact, day of the week, and occupation as the three most influential variables in determining the likelihood of subscription.

The consumer price index (OR=1.36) is positively correlated with subscription likelihood, indicating that as the index rises, the likelihood of subscribe rises. However, the Euribor 3-month rate (OR=0.59) has a negative relationship with subscription likelihood, meaning that a higher Euribor 3-month rate reduces the probability of a customer subscribing to the term deposit. Similarly, Moro et al. (2014), utilizing a neural network (NN) approach, also found a negative correlation between the Euribor rate and the likelihood of subscription.

Assumptions Checking

Standardized Residuals			Assumption
> 1.96	1453	Less than 5% of the data	✓
> 2.58	209	Less than 1% of the data	✓
> 3	2	More than O	✗

Table 7. Residuals Assumption Checking

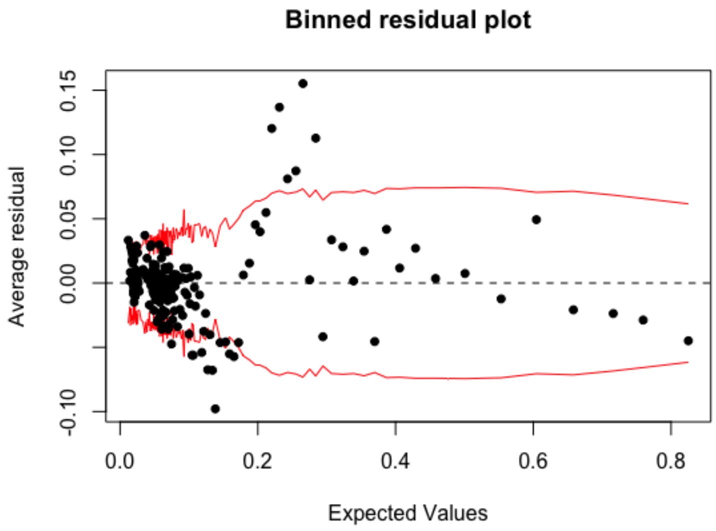
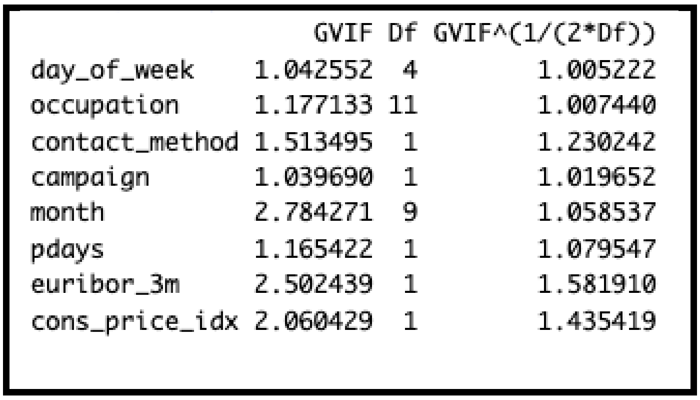


Figure 16. Binned Residuals Plot

The shaded red areas represent the range where approximately 95% of the observations are expected to be found. While not all values fall within this red area, less than 5% of the observations lie outside these boundaries.



```
                    GVIF Df GVIF^(1/(2*Df))
day_of_week     1.042552  4        1.005222
occupation      1.177133 11        1.007440
contact_method  1.513495  1        1.230242
campaign        1.039690  1        1.019652
month           2.784271  9        1.058537
pdays           1.165422  1        1.079547
euribor_3m      2.502439  1        1.581910
cons_price_idx  2.060429  1        1.435419
```

Figure 17. Multicollinearity

There are no variables in the dataset with Generalized Variance Inflation Factor (GVIF) values exceeding the threshold of 10. This threshold is commonly used as a benchmark to evaluate the presence of multicollinearity.

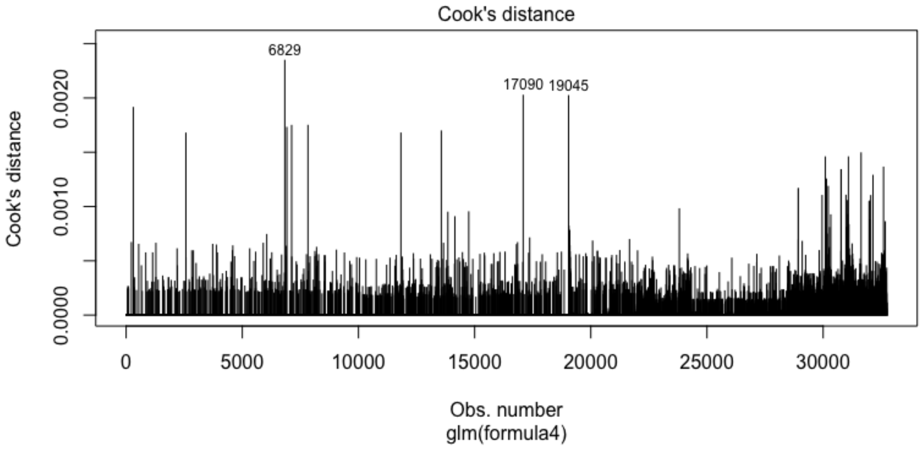
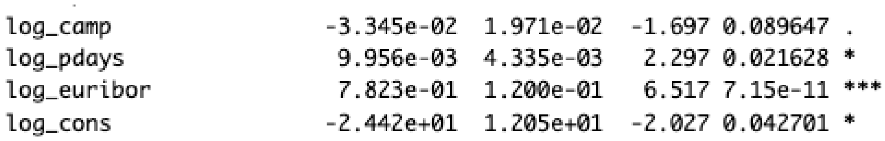


Figure 18. Cook's distance

When examining the influential cases within the model, no value has a Cook's Distance greater than 1. However, there are 8535 instances with a leverage higher than 0.0009.



```
log_camp              -3.345e-02  1.971e-02  -1.697 0.089647 .
log_pdays              9.956e-03  4.335e-03   2.297 0.021628 *
log_euribor            7.823e-01  1.200e-01   6.517 7.15e-11 ***
log_cons              -2.442e+01  1.205e+01  -2.027 0.042701 *
```

Table 8. Linearity of the Logit

Lastly, it appears that the linearity of the logit assumption is violated, as the log of numerical variables are statistically significant for the model. This violation leads to reduced confidence in the model's generalizability to the population from which the sample was drawn.

Model Performance

Accuracy	Kappa	Sensitivity	Specificity	AUC-ROC
90.05 %	0.303	0.23	0.98	79 %

Table 9. Model Evaluation

While an accuracy of 90.05% may seem impressive, it is only slightly better than a random guess, considering that more than 88% of the data represent 'no' subscriptions. This accuracy is better than that presented with a logistic regression model by Asare-Frempong and Jayabalan (2017). However, their dataset did not contain economic and social variables, which have been decisive in this study. On the other hand, Jiang (2018) obtained a higher accuracy of 92.03% with the same model for the same purpose.

A Kappa of 0.30 offers critical insight into the model's performance, especially in the context of an imbalanced dataset. The model's sensitivity of 0.23 indicates a limited ability to correctly identify True Positives.

	Reference	
Prediction	**no**	**yes**
no	7162	713
yes	101	213

Table 10. Confusion Matrix

The confusion matrix suggests that the model is quite effective in identifying non-subscribers. However, the model is not effective in identifying subscribers.

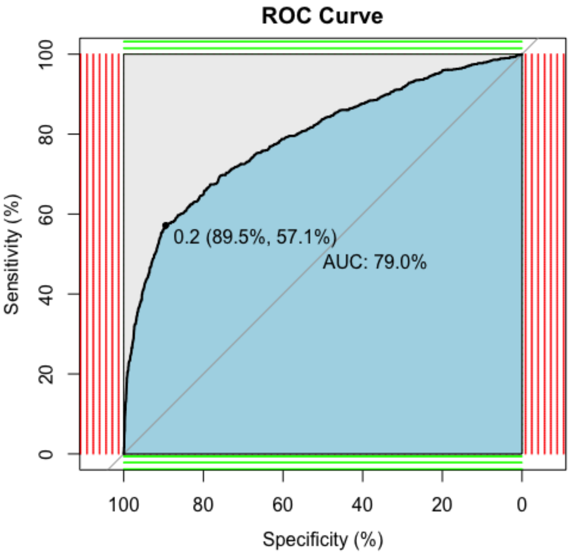


Figure 19. AUC-ROC

4. Discussion and Recommendations

Through this model H4 is confirmed by indicating a negative relationship between the Euribor 3-month rate and subscription likelihood. However, H1 and H2 are rejected, as they do not improve the model's accuracy. While H3 did improve accuracy, the variable was removed due to high multicollinearity.

Monday is the least effective for making contacts, possibly due to people readjusting to work routines after the weekend. On the other hand, Wednesday is considered the most effective day. Subscription preferences vary by month, with December and March being the most favourable and with May being the least favourable.

Entrepreneurs, who need more liquidity, and lower-income workers are less likely to invest in term deposits. Conversely, the simplicity of term deposits appeals to students and retirees, making them prime targets. Therefore, tailored plans are recommended for students and retirees.

It is suggested to intensify marketing activities when the Euribor index shows a decreasing trend, and when the consumer price index shows a high trend.

5. Conclusion and Limitations

This project has consisted of developing a multiple logistic regression model to predict term deposit subscriptions, using data from bank telemarketing campaigns.

Although logistic regression is a robust method for binary classification problems, it has limitations. Other algorithms, such as decision tree or random forest may offer better predictive accuracy in certain cases. For this data set as can be seen Random Forest offers an improvement in the detection of subscribers.

Imbalance Dataset	Accuracy	Kappa	Sensitivity	Specificity
Logistic Regression	0.90	0.303	0.23	0.98
Decision Tree	0.90	0.275	0.19	0.99
Radom Forest	0.89	0.34	0.29	0.97

Table 11. Comparison of classification models

The major challenge in this project is the unbalanced nature of the dataset. This imbalance may result in a model biased towards the prediction of the majority class, resulting in an underperformance of the classification of the minority class. Techniques such as SMOTE, which is a technique to generate synthetic data can help mitigate this problem.

SMOTE	Accuracy	Kappa	Sensitivity	Specificity	AUC-ROC
Logistic Regression	71.81 %	0.43	0.72	0.71	79.4 %
Decision Tree	90%	0.80	0.92	0.88	94.82%
Radom Forest	92%	0.85	0.90	0.96	96.82%

Table 12. Comparison of the classification models after balancing the dataset with SMOTE

6. Reflective Summary

I have developed the ability to build classification and regression algorithms in R during this semester. This experience has deepened my interest in data science, inspiring me to specialise in the whole process and not just data analysis. My favourite aspect has been developing machine learning models and investigating existing techniques in journals to improve their accuracy, such as using SMOTE to balance datasets. This exploration of advanced methodologies has further fuelled my passion for this field.

I look forward to the second semester to apply these concepts in Python and explore the realm of unsupervised machine learning, broadening my experience and knowledge in this fascinating area of data science.

References

Asare-Frempong, J. and Jayabalan, M. (2017) ‘Predicting customer response to Bank Direct Telemarketing Campaign’, 2017 International Conference on Engineering Technology and Technopreneurship (ICE2T) [Online]. doi:10.1109/ice2t.2017.8215961. Available at: https://ieeexplore.ieee.org/document/8215961 Accessed: 22/12/2023

Campbell, K.M. and Nick, T.G. (2007) ‘Logistic Regression’, in W.T. Ambrosius (ed.) Topics in biostatistics. New Jersey: Human Press, pp. 273–303. Available at: https://link.springer.com/protocol/10.1007/978-1-59745-530-5_14 Accessed: 27/12/2023

Chai, C.P. (2020) ‘The importance of data cleaning: Three visualization examples’, CHANCE, 33(1), pp. 4–9. doi:10.1080/09332480.2020.1726112. Available at: https://www.tandfonline.com/doi/full/10.1080/09332480.2020.1726112?casa_token=CsRZRAVr-iAAAAAA%3AVvpFqNZXHFt22EolGbnIMSsihW0GozGUog3smoBV6LLuYXEvQ47ete-mW5A1as6bC6yjNew0vlA Accessed: 27/12/2023

Charles, Mr.A. and Kasilingam, Dr.R. (2013) ‘Does the investor’s age influence their investment behaviour?’, Paradigm, 17(1–2), pp. 11–24. doi:10.1177/0971890720130103. Available at: https://journals.sagepub.com/doi/epdf/10.1177/0971890720130103 Accessed: 23 December 2023

Chen, J. (2022) Term deposit: Definition, how it’s used, rates, and how to invest, Investopedia. Edited by G. Scott. Available at: https://www.investopedia.com/terms/t/termdeposit.asp (Accessed: 22 December 2023).

Choi, Y. and Choi, J. (2022) ‘How does machine learning predict the success of bank telemarketing?’, *Research Squared*, pp. 1–13. doi:10.21203/rs.3.rs-1695659/v1.

Colaianni, G., Magdangal, J. and Mitchell, M. (2016) FACTORS DETERMINING TERM DEPOSIT PURCHASES [Preprint]. Available at: https://support.sas.com/resources/papers/proceedings17/2029-2017.pdf Accessed: 22/12/2023

Freeman, E.A. and Moisen, G.G. (2008) ‘A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa’, Ecological Modelling, 217(1–2), pp. 48–58. doi:10.1016/j.ecolmodel.2008.05.015. Available at: https://www.sciencedirect.com/science/article/pii/S0304380008002275?casa_token=-9ZAWdC597YAAAAA:R4htOWaxoSu2rKHfVqeNMTr0rDw6t5yi7EeWyIA-V-xNUUtfwZH7NxKGzXbE24lE07MNuHUy Accessed: 28/12/2023

Gorsevski, P.V. et al. (2006) ‘Spatial prediction of landslide hazard using logistic regression and ROC analysis’, Transactions in GIS, 10(3), pp. 395–415. doi:10.1111/j.1467-9671.2006.01004.x. Available at: https://web-p-ebscohost-com.queens.ezp1.qub.ac.uk/ehost/pdfviewer/pdfviewer?vid=0&sid=98b9600b-cd5c-4f49-8bf3-d5d7aaa36f6e%40redis Accessed: 28/12/2023

Jiang, Y. (2018) ‘Using Logistic Regression Model to Predict the Success of Bank Telemarketing’, International Journal on Data Science and Technology, 1(4), pp. 35–41. doi:10.11648/j.ijdst.20180401.15. Accessed: 02 January 2024

Kandah, A. (2023) ‘Consumer Confidence Index and spending trends’, Jordan Times. Available at: https://jordantimes.com/opinion/adli-kandah/consumer-confidence-index-and-spending-trends (Accessed: 23 December 2023).

Maverick, J.B. (2023) *Do changes in interest rates affect consumer spending?*, *Investopedia*. Available at: https://www.investopedia.com/ask/answers/071715/how-do-changes-interest-rates-affect-spending-habits-economy.asp (Accessed: 23 December 2023).

Moro, S., Cortez, P. and Rita, P. (2014) ‘A data-driven approach to predict the success of bank telemarketing’, *Decision Support Systems*, 62, pp. 22–31. doi:10.1016/j.dss.2014.03.001. (Accessed: 23 December 2023).

Osborne, J.W. and Overbay, A. (2019) ‘The power of outliers (and why researchers should ALWAYS check for them)’, Practical Assessment, Research, and Evaluation, 9(6). doi:https://doi.org/10.7275/qf69-7k43. Available https://scholarworks.umass.edu/pare/vol9/iss1/6 (Accessed: 27 Dic. 2023).

Szumilas, M. (2010) ‘Explaining Odds Ratios’, PubMed Central, (19), pp. 227–229. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2938757/#:~:text=What%20is%20an%20odds%20ratio,the%20absence%20of%20that%20exposure. Accessed: 03/01/2024

Xie, C. et al. (2023) ‘How to improve the success of bank telemarketing? prediction and interpretability analysis based on machine learning’, Computers &amp; Industrial Engineering, 175. doi:10.1016/j.cie.2022.108874. Available at: https://www.sciencedirect.com/science/article/pii/S0360835222008622 Accessed: 02 January 2024

Appendix 1

```
         (Intercept)        day_of_weekmon        day_of_weekthu
        8.948729e-13         8.355365e-01         1.087551e+00
       day_of_weektue        day_of_weekwed   occupationblue-collar
        1.084345e+00         1.168460e+00          7.334772e-01
occupationentrepreneur    occupationhousemaid    occupationmanagement
        7.827690e-01         8.321733e-01         9.072574e-01
    occupationretired occupationself-employed     occupationservices
        1.274546e+00         8.923751e-01          7.973455e-01
    occupationstudent    occupationtechnician   occupationunemployed
        1.268150e+00         9.027615e-01          1.064256e+00
    occupationunknown contact_methodtelephone              campaign
        8.615533e-01         7.822975e-01         9.589265e-01
           monthaug             monthdec             monthjul
        1.507259e+00         2.525736e+00          1.597691e+00
           monthjun             monthmar             monthmay
        1.401386e+00         3.349669e+00          5.595191e-01
           monthnov             monthoct             monthsep
        1.167723e+00         1.980295e+00          1.487169e+00
          euribor_3m        cons_price_idx             pdays
        5.973433e-01         1.359546e+00          9.985449e-01
```

Table 13. Coefficients of model 4

	Dependent variable:			
	subscribed			
	(1)	(2)	(3)	(4)
age	0.004**	0.004*	0.006***	
	(0.002)	(0.002)	(0.002)	
day_of_weekmon				-0.180***
				(0.064)
day_of_weekthu				0.084
				(0.061)
day_of_weektue				0.081
				(0.063)
day_of_weekwed				0.156**
				(0.063)
occupationblue-collar	-0.487***	-0.387***		-0.310***
	(0.060)	(0.073)		(0.062)
occupationentrepreneur	-0.379***	-0.372***		-0.245**
	(0.120)	(0.120)		(0.121)
occupationhousemaid	-0.181	-0.130		-0.184
	(0.132)	(0.137)		(0.136)
occupationmanagement	-0.141*	-0.173**		-0.097
	(0.079)	(0.080)		(0.081)
occupationretired	0.262***	0.301***		0.243***
	(0.095)	(0.098)		(0.082)
occupationself-employed	-0.190*	-0.199*		-0.114
	(0.110)	(0.111)		(0.113)
occupationservices	-0.342***	-0.266***		-0.226***
	(0.075)	(0.079)		(0.078)

occupationstudent	0.424***	0.467***		0.238**
	(0.099)	(0.103)		(0.102)
occupationtechnician	-0.081	-0.081		-0.102*
	(0.058)	(0.066)		(0.061)
occupationunemployed	0.131	0.183		0.062
	(0.111)	(0.113)		(0.118)
occupationunknown	-0.124	-0.134		-0.149
	(0.213)	(0.216)		(0.228)
marital_statusmarried	0.029	0.029	-0.008	
	(0.063)	(0.063)	(0.065)	
marital_statussingle	0.176**	0.163**	0.171**	
	(0.071)	(0.072)	(0.073)	
marital_statusunknown	0.322	0.294	0.287	
	(0.472)	(0.471)	(0.533)	
education_levelbasic.6y		0.002		
		(0.110)		
education_levelbasic.9y		-0.110		
		(0.088)		
education_levelhigh.school		-0.021		
		(0.084)		
education_levelilliterate		1.263*		
		(0.727)		
education_levelprofessional.coure		0.074		
		(0.093)		
education_leveluniversity.degree		0.140*		
		(0.084)		
education_levelunknown		0.178		
		(0.110)		
contact_methodtelephone				-0.246***
				(0.060)
euribor_3m	-0.521***	-0.520***		-0.515***
	(0.011)	(0.011)		(0.017)
cons_conf_idx	0.049***	0.048***		
	(0.003)	(0.003)		
credit_defaultunknown			-0.377***	
			(0.063)	
credit_defaultyes			-7.676	
			(119.468)	
cons_price_idx				0.307***
				(0.043)
pdays				-0.001***
				(0.0001)
monthaug			0.198**	0.410***
			(0.079)	(0.084)
monthdec			0.551***	0.927***

			(0.192)	(0.190)
monthjul			0.434^{***}	0.469^{***}
			(0.082)	(0.086)
monthjun			0.146^{*}	0.337^{***}
			(0.081)	(0.087)
monthmar			1.273^{***}	1.209^{***}
			(0.115)	(0.117)
monthmay			-0.668^{***}	-0.581^{***}
			(0.070)	(0.071)
monthnov			-0.303^{***}	0.155^{*}
			(0.085)	(0.090)
monthoct			0.516^{***}	0.683^{***}
			(0.107)	(0.108)
monthsep			0.532^{***}	0.397^{***}
			(0.119)	(0.118)
campaign			-0.045^{***}	-0.042^{***}
			(0.010)	(0.010)
poutcomenonexistent			0.272^{***}	
			(0.059)	
poutcomesuccess			2.031^{***}	
			(0.085)	
emp_var_rate			-0.480^{***}	
			(0.015)	
Constant	1.372^{***}	1.254^{***}	-2.725^{***}	-27.742^{***}
	(0.189)	(0.211)	(0.131)	(4.042)
Observations	32,760	32,760	32,760	32,760
Log Likelihood	-9,855.494	-9,845.500	-9,355.313	-9,236.360
Akaike Inf. Crit.	19,746.990	19,741.000	18,750.630	18,532.720
Note:			*p<0.1; **p<0.05;	***p<0.01

TRUE

Table 14. Comparative of the models

	Model 1	**Model 2**	**Model 3**	**Model 4**
Nagelkerke R^2	0.196	0.197	0.249	0.262

Table 15. Comparative of R^2

SUBSCRIPTION RATES OF THE CATEGORICAL VARIABLES IN THE MODEL

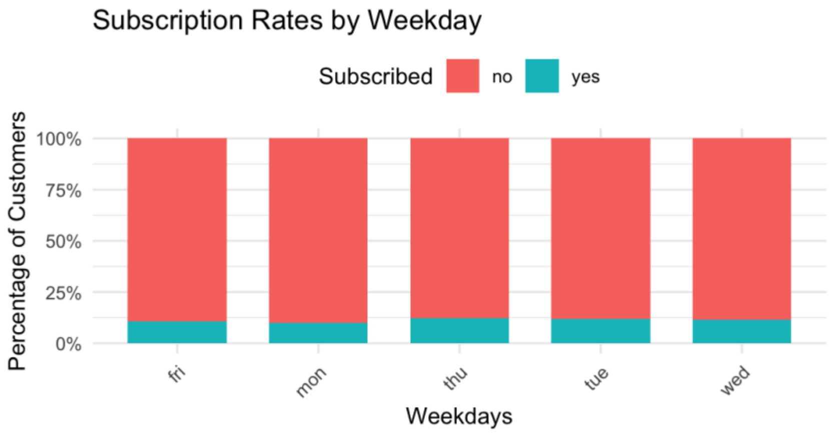


Figure 20. Subscription Rates by weekday

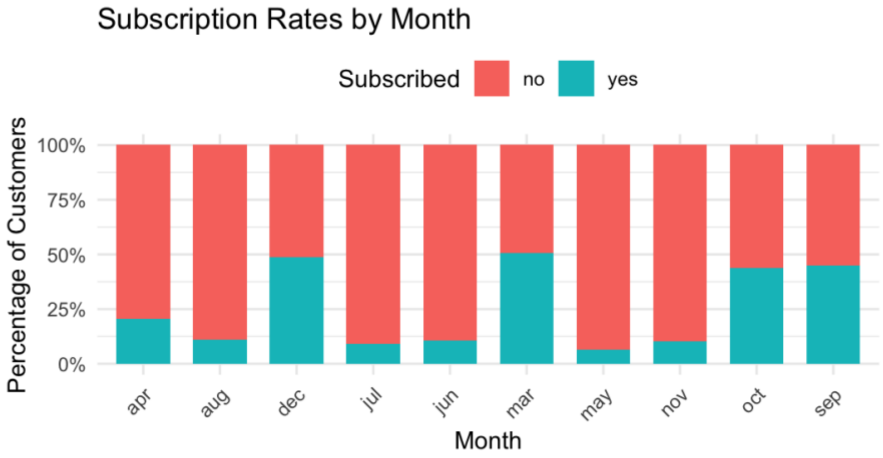


Figure 21. Subscription Rates by month

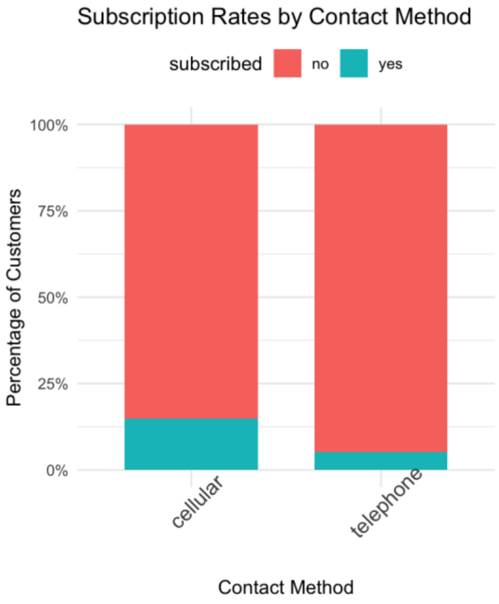


Figure 22. Subscription Rates by Contact Method

Appendix 2

```
library(ggplot2)
library(readxl)
library(tidyverse)
library(caret)
library(corrplot)
library(mice)
library(VIM)
library(pROC)
library(car)
library(stargazer)
library(arsenal)
library(arm)

#open the data set and save it as "term"
term <- read_excel("/Users/jaimerd/Desktop/master/Statistics Business/Assigment2 /term.xlsx")

##data cleaning ------

#convert categorical variables into factors
term <- term %>% mutate_if(is.character,as.factor)

summary(term)

##numeric variables
data_numeric <- term[sapply(term, is.numeric)]
data_numeric

#age
hist(term$age)
boxplot(term$age)
term %>% count(age>100)
term$age[term$age >100] <- NA
summary(term$age)

#try to impute outliers but the accuracy does not improve
term$age[term$age >100] <- mean(term$age)
term$age[term$age >100] <-median(term$age)

#contact duration (not important will not be used for the model)
hist(term$contact_duration)
boxplot(term$contact_duration)
summary(term$contact_duration)

#campaign ()
table(term$campaign)
boxplot(term$campaign)
term %>% count(campaign>33)
term$campaign[term$campaign > 33] <- NA
summary(term$campaign)

#pdays
hist(term$pdays)
table(term$pdays)
boxplot(term$pdays)
```

```
summary(term$pdays)

#previouscontacts
hist(term$previous_contacts)
table(term$previous_contacts)

#var rate
hist(term$emp_var_rate)
table(term$emp_var_rate)

#cons price index
hist(term$cons_price_idx)
table(term$cons_price_idx)
boxplot(term$cons_price_idx)

#cons conf index
boxplot(term$cons_conf_idx)
table(term$cons_conf_idx)
summary(term$cons_conf_idx)

#Euribor 3 month
boxplot(term$euribor_3m)

#number of employees
boxplot(term$n_employed)

##categorical variables
#month
table(term$month)
levels(term$month)[levels(term$month) == "july"] <- "jul"

#day of the week
table(term$day_of_week)
levels(term$day_of_week)[levels(term$day_of_week) == "tues"] <- "tue"

##missing values = converted them to the level with the highest frequency
summary(term$marital_status)
term$marital_status[is.na(term$marital_status)] <- 'married'

#omit missing values
term <- na.omit(term)

sumsummary(term)

##Descriptive Statistics and correlations ------

#Set parameters for the table, such as the type of correlation test for numerical and categorical
variables.
configuration <- tableby.control(
  test = T,
  total = FALSE,
  numeric.test = "anova", cat.test = "chisq",
  numeric.stats = c("meansd"),
  cat.stats = c("countpct"))

descriptive_stats <- tableby(subscribed ~ .,
                    data = term,
                  control = configuration)

report <- summary(descriptive_stats, title = "Descriptive Statistic", text= TRUE)
write2word(report, file = "Descriptive_Statistics.docx")


#bar plot showing the occupation
ggplot(term, aes(x = occupation, fill = subscribed)) +
  geom_bar(position = "stack", width = 0.7) +
  labs(
      x = "Occupation",
      y = "Customers",
      fill = "Subscribed") +
  theme_minimal() +
  theme(axis.text.y = element_text(size = 10),legend.position = "top") +
  coord_flip()

#Subscription Rates by occupation
ggplot(term, aes(x = occupation, fill = subscribed)) +
  geom_bar(position = "fill", width = 0.7) +
  scale_y_continuous(labels = scales::percent_format()) +
  labs(
    x = "Occupation",
    y = "Percentage of Customers",
    fill = "Subscribed",
    title = "Subscription Rates by Occupation"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "top")


#bar plot marital status
ggplot(term, aes(x = marital_status, fill= subscribed)) +
  geom_bar(position= "stack", width = 0.5) +
  labs(
      x = "Marital Status",
      y = "Customers",
      fill = "Subscribed") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle= 90, size=12),legend.position = "top")

#bar plot education level
ggplot(term, aes(x = education_level, fill= subscribed)) +
  geom_bar(position= "stack", width = 0.7) +
  labs(
    x = "Education Level",
    y = "Customers",
    fill = "Subscribed") +
  theme_minimal() +
  theme(axis.text.y = element_text(size = 10),legend.position = "top")+
  coord_flip()


#bar graph credit default
ggplot(term, aes(x = credit_default, fill = subscribed)) +
  geom_bar(position = "stack", width = 0.5) +
  labs(
    x = "Credit Default",
    y = "Customers",
    fill = "Subscribed"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(size = 13),legend.position = "top")



#subscriptions per month
#Create a data frame with the number of subscribers and group them by month in order.
month_subs <- term %>%
  filter(subscribed == "yes") %>%
  mutate(month = factor(month, levels = c("mar", "apr", "may", "jun", "jul", "aug", "sep", "oct","nov",
"dec"))) %>%
  group_by(month) %>%
  summarise(subscriptions = n())

#line graph with subscribers per month
ggplot(month_subs, aes(x = month, y = subscriptions, group=1)) +
  geom_area(fill = "blue", alpha = 0.3) +
  geom_line(color = "blue", size =0.5) +
  labs(title = "Subscribers per Month",
       x = "Month",
       y = "Subscriptions") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90 ,size = 13))

#bar graph with the subscription rates by month
ggplot(term, aes(x = month, fill = subscribed)) +
  geom_bar(position = "fill", width = 0.7) +
  scale_y_continuous(labels = scales::percent_format()) +
  labs(
    x = "Month",
    y = "Percentage of Customers",
    fill = "Subscribed",
    title = "Subscription Rates by Month"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),legend.position = "top")

#subscriptions per weekday
#Create a data frame with the number of subscribers and group them by weekday in order.
week_subs <- term %>%
  filter(subscribed == "yes") %>%
  mutate(day_of_week = factor(day_of_week, levels = c("mon", "tue", "wed", "thu", "fri"))) %>%
  group_by(day_of_week) %>%
  summarise(subscriptions = n())

#line graph with subscribers per weekday
ggplot(week_subs, aes(x = day_of_week, y = subscriptions, group=1)) +
  geom_area(fill = "blue", alpha = 0.3) +
  geom_line(color = "blue", size =0.5) +
  labs(title = "Subscribers per Week Day",
       x = "Day of the week",
      y = "Subscriptions") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, size = 13))

#Subscription Rates by Weekday
ggplot(term, aes(x = day_of_week, fill = subscribed)) +
  geom_bar(position = "fill", width = 0.7) +
  scale_y_continuous(labels = scales::percent_format()) +
  labs(
    x = "Weekdays",
    y = "Percentage of Customers",
    fill = "Subscribed",
    title = "Subscription Rates by Weekday"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),legend.position = "top")


#box plot campaign for subscribers
ggplot(term, aes(x= subscribed, y= campaign, fill=subscribed))+
  geom_boxplot()+
  labs(y = "Number of contacts") +
  theme_minimal()

#contact method
#create data frame for customers contacted by cellular
cellular <- term %>%
  filter(contact_method == "cellular") %>%
  group_by(subscribed) %>%
  summarise(cellular_cont = n())

#bar graph for customers contacted by cellular
ggplot(term, aes(x = subscribed, fill = contact_method)) +
  geom_bar(position = "stack", width = 0.5) +
  labs(
    x = "Subscribed",
    y = "Customers",
    fill = "Contact Method"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(size = 13),legend.position = "top")

#Subscription Rates by contact method
ggplot(term, aes(x = contact_method, fill = subscribed)) +
  geom_bar(position = "fill", width = 0.7) +
  scale_y_continuous(labels = scales::percent_format()) +
  labs(
    x = "Contact Method",
    y = "Percentage of Customers",
    title = "Subscription Rates by Contact Method"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, size=12),
    legend.position = "top"
  )
```

```
#Blot plot emp_var_rate
ggplot(term, aes(y = emp_var_rate, x = subscribed, fill= subscribed)) +
  geom_boxplot() +
  labs(y = "Employment variation rate") +
  theme_minimal()

#Box plot cons_price_idx
ggplot(term, aes(y = cons_price_idx, x = subscribed, fill= subscribed)) +
  geom_boxplot() +
  labs(y = "Consumer price index ") +
  theme_minimal()

#Box plot cons_conf_idx
ggplot(term, aes(y = cons_conf_idx, x = subscribed, fill= subscribed)) +
  geom_boxplot() +
  labs(y = "Consumer confidence index  ") +
  theme_minimal()

#Box plot euribor_3m
ggplot(term, aes(y = euribor_3m, x = subscribed, fill= subscribed)) +
  geom_boxplot() +
  labs(y = "Euribor 3 month rate") +
  theme_minimal()

#Box plot n_employed
ggplot(term, aes(y = n_employed, x = subscribed, fill= subscribed)) +
  geom_boxplot() +
  labs(y = "Number of employees") +
  theme_minimal()

#Categorize ages into groups and create a bar plot of subscriptions by age group
term$age_group <- cut(term$age, breaks = c(0, 25, 35, 45, 55, 65, Inf), labels = c("0-25", "26-35", "36-45", "46-55", "56-65", "66+"))
ggplot(term, aes(x=age_group, fill = subscribed)) +
  geom_bar(position= "stack", width = 0.5) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, size=10))+
  theme(legend.position = "top")


##Model Development -----

levels(term$subscribed)

#omit missing values
term <- na.omit(term)

#split the data into train (80%) and test (20%) data set.
set.seed(40425150)
index <- createDataPartition(term$subscribed, p=0.8, list=FALSE)
train <- term[index,]
test <- term[-index,]

#1 model
formula1 = subscribed  ~ age + occupation + marital_status + euribor_3m + cons_conf_idx
```

```
model1 <- glm(formula = formula1 , data = train, family= "binomial")
summary(model1)

#2 model
formula2 = subscribed  ~ age + occupation + marital_status + education_level + euribor_3m +
cons_conf_idx
model2 <- glm( formula = formula2, data = train, family= "binomial" )
summary(model2)

#3 model
formula3 = subscribed  ~ age + marital_status + credit_default + month + campaign + poutcome +
emp_var_rate
model3 <- glm(formula = formula3, data = train, family= "binomial")
summary(model3)

#4 model
formula4 = subscribed ~ day_of_week + occupation + contact_method + campaign + month +
euribor_3m + cons_price_idx + pdays
model4 <- glm(formula = formula4 , data = train, family = "binomial")
summary(model4)
exp(model4$coefficients)

#Pseudo R^2
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")
}
#Source: Field et al. (2012)

logisticPseudoR2s(model1)
logisticPseudoR2s(model2)
logisticPseudoR2s(model3)
logisticPseudoR2s(model4)

#table with all the models
stargazer(model1, model2, model3, model4, type = "html", out =
          "models.html")

##assumptions checking -----

#residuals
resid(model4)
train$standarisedResiduals <- rstandard(model4)
train$studentdResiduals <- rstudent(model4)
sum(train$standarisedResiduals > 1.96)
sum(train$standarisedResiduals > 2.58)
sum(train$standarisedResiduals > 3)
plot(model4)
summary(train$standarisedResiduals)
```

```r
table(train$standarisedResiduals)

#binned residual plot
binnedplot(fitted(model4),
           residuals(model4, type = "response"),
           col.pts = 1,
           col.int = "red")



#influential cases
train$cook <- cooks.distance(model4)
sum(train$cook > 1)

train$leverage <- hatvalues(model4)
sum(train$leverage > 0.0009)

cooks.distance(model4)

plot(model4,which=4)

#multicollinearity
vif(model4)

#linearity of the logit (log of numerical variables)
train$log_camp <- log(train$campaign)*train$campaign
train$log_pdays <- log(train$pdays)*train$pdays
train$log_euribor <- log(train$euribor_3m)*train$euribor_3m
train$log_cons <- log(train$cons_price_idx)*train$cons_price_idx



formula_linea = subscribed ~ day_of_week + occupation + contact_method + campaign + month + pdays
+ euribor_3m + cons_price_idx + log_camp + log_pdays +log_euribor +log_cons
model_check <- glm(formula = formula_linea , data = train, family = "binomial")
summary(model_check)




##Predictions with test data -----

predictions <- predict(model4,test, type = "response")
predictions
class_pred <- as.factor(ifelse(predictions > 0.5, "yes", "no"))
postResample(class_pred, test$subscribed)
conf_matrix <- confusionMatrix(class_pred, test$subscribed, positive = "yes")
conf_matrix

#ROC Curve
r <- multiclass.roc(test$subscribed, predictions, percent = TRUE)
roc <- r[['rocs']]
r1 <- roc[[1]]
plot.roc(r1,
         print.auc=TRUE,
         auc.polygon=TRUE,
         grid=c(0.1, 0.2),
        grid.col=c("green", "red"),
        max.auc.polygon=TRUE,
        auc.polygon.col="lightblue",
        print.thres=TRUE,
        main= 'ROC Curve')

#Source: finnstats, R-bloggers, 2021



##Fixing the imbalance data set ------

remotes::install_version("DMwR", version="0.4.1")

library ("DMwR")

smote_dataset <- as.data.frame(term)

#balancing the data set
smote <-
  SMOTE(
    form = subscribed ~ .,
    data = smote_dataset,
    perc.over = 400,
    perc.under = 100
  )

set.seed(40425150)
index <- createDataPartition(smote$subscribed, p=0.8, list=FALSE)
train_smote <- smote[index,]
test_smote <- smote[-index,]

#logistic regression with balanced data
formula_smote = subscribed ~ day_of_week + occupation + contact_method + campaign + month +
euribor_3m + cons_price_idx + pdays
model_smote <- glm(formula = formula_smote , data = train_smote, family = "binomial")
summary(model_smote)

#Predictions with test data

predictions_smote <- predict(model_smote,test_smote, type = "response")
predictions_smote
class_pred_smote <- as.factor(ifelse(predictions_smote > 0.5, "yes", "no"))
postResample(class_pred_smote, test_smote$subscribed)
conf_matrix_smote <- confusionMatrix(class_pred_smote, test_smote$subscribed, positive = "yes")
conf_matrix_smote

#ROC Curve
r_smote <- multiclass.roc(test_smote$subscribed, predictions_smote, percent = TRUE)
roc_smote <- r_smote[['rocs']]
r1_smote <- roc_smote[[1]]
plot.roc(r1_smote,
        print.auc=TRUE,
        auc.polygon=TRUE,
        grid=c(0.1, 0.2),
        grid.col=c("green", "red"),
        max.auc.polygon=TRUE,
        auc.polygon.col="lightblue",
        print.thres=TRUE,
        main= 'ROC Curve')





#improving accuracy with decision tree ------
library(rpart)

#imbalance data set
#decision tree with all variables except ID and contact duration
tree <- rpart(subscribed ~ age + occupation + marital_status + education_level + credit_default +
housing_loan + personal_loan + contact_method + month + day_of_week + campaign + pdays +
previous_contacts + poutcome + emp_var_rate + cons_price_idx +cons_conf_idx + euribor_3m +
n_employed, data = train, method = "class")
print(tree)
predictions_tree <- predict(tree, test, type = "class")
predictions_tree
postResample(predictions_tree, test$subscribed)
cm_tree <- confusionMatrix(predictions_tree, test$subscribed)
cm_tree

#smote
#decision tree with all variables except ID and contact duration
tree_smote <- rpart(subscribed ~ age + occupation + marital_status + education_level + credit_default +
housing_loan + personal_loan + contact_method + month + day_of_week + campaign + pdays +
previous_contacts + poutcome + emp_var_rate + cons_price_idx +cons_conf_idx + euribor_3m +
n_employed, data = train_smote, method = "class")
print(tree_smote)
predictions_tree_smote <- predict(tree_smote, test_smote, type = "class")
predictions_tree_smote
postResample(predictions_tree_smote, test_smote$subscribed)
cm_tree_smote <- confusionMatrix(predictions_tree_smote, test_smote$subscribed)
cm_tree_smote

#ROC Curve
predictions_tree_1 <- predict(tree_smote, test_smote, type = "prob")
probabilities_tree <- predictions_tree_1[, "yes"]
roc_curve_tree <- roc(test_smote$subscribed, probabilities_tree)
auc(roc_curve_tree)

#improving accuracy with random forest -------

library(randomForest)

#imbalance data set
#random forest with all variables except ID and contact duration
rf <- randomForest(subscribed ~ age + occupation + marital_status + education_level + credit_default +
housing_loan + personal_loan + contact_method + month + day_of_week + campaign + pdays +
previous_contacts + poutcome + emp_var_rate + cons_price_idx +cons_conf_idx + euribor_3m +
n_employed, data = train)
rf
pred_rf <- predict(rf,test)
cm_rf <- confusionMatrix(pred_rf,test$subscribed,positive = "yes")
```

```
cm_rf

#smote
#random forest with all variables except ID and contact duration
rf_smote <- randomForest(subscribed ~ age + occupation + marital_status + education_level +
credit_default + housing_loan + personal_loan + contact_method + month + day_of_week + campaign +
pdays + previous_contacts + poutcome + emp_var_rate + cons_price_idx +cons_conf_idx + euribor_3m
+ n_employed, data = train_smote)
rf_smote
pred_rf_smote <- predict(rf_smote,test_smote)
cm_rf_smote <- confusionMatrix(pred_rf_smote,test_smote$subscribed,positive = "yes")
cm_rf_smote

#ROC Curve
predictions_rf_1 <- predict(rf_smote, test_smote, type = "prob")
probabilities_rf <- predictions_rf_1[, "yes"]
roc_curve_rf <- roc(test_smote$subscribed, probabilities_rf)
auc(roc_curve_rf)
```