Logistic regression model to predict term deposit: Application of Down sampling Vs Base Data

Devane Risby

Contents

1 Introduction

Term deposits (TD) represent a fixed-term investment, involving the placement of funds into a financial institution's account, accessible for withdrawal only after a predetermined period (Chen, J. 2022). Opting for a TD fundamentally involves a personalized evaluation of one's circumstances and risk tolerance. Nevertheless, discernible patterns emerge in customer traits that allow us to forecast the likelihood of a customer subscribing to this service.

This report explores the implementation of a logistic regression model with the specific goal of predicting the probability that a customer subscribes to a TD. Leveraging insights from a comprehensive dataset comprising 40,927 observation, this analysis incorporates a diverse range of customer-specific details, including age, housing loan status, education, and more. Additionally, it encompasses social and economic variables such as employment variations, consumer price index, etc.

1.1 Literature Review

Logistic regression is increasingly becoming a cornerstone in the financial industry, with Broby, D. (2022) highlighting its significance through a comprehensive study. Specifically, within the domain of TD outcome classification, logistic regression has emerged as a widely utilized technique. Notably, Trivedi, N.K. et al. (2020) applied a logistic regression model to forecast subscription rates for TDs, achieving an impressive overall accuracy of 91.0%. This high accuracy underscores the model's suitability and effectiveness for its intended purpose. TD have been previously predicted at an accuracy of 85% (Jiang, E. et al. 2022) setting a benchmark for the study.

The impact of data balancing on model quality is multifaceted. Mooijman, P. et al. (2023) demonstrated that this process effectively enhances model quality by rectifying biases towards majority classes, thereby elevating overall classification performance and increasing precision. Down sampling, a prevalent method used to balance target categories, was examined in Lee, W. and Seo, K.'s (2022) study. Their research revealed a substantial enhancement in the overall model quality due to down sampling. However, they observed that while the classification performance improved significantly, it eventually reached a plateau, indicating a limit to the improvement.

Social characteristics, such as the unemployment rate, have a notable impact on consumer spending. Ganong, P. and Noel, P. (2022) highlight a significant negative correlation between the unemployment rate and consumer spending. Hassan, A. *et al* (2010) note that age impacts saving rate with notable concentration towards working age and a significant plateau at either end. Cole, S.A., Paulson, A.L. and Shastry, G.K. (2012) finds a direct positive relationship between education level and saving rate.

1.2 Hypothesis

H_1	Age has a positive correlation of TD Rate
H_2	Occupation is a significant indicator of TD Rate
H_3	Employment variation will have a positive correlation with TD Rate
H_4	Education level is a significant indicator of TD Rate
H_5	Housing Loan is a significant indicator of TD Rate

2. Methodology

A logistic regression model will be used to predict if a record will have a TD based on 11 independent variables and 1 target variable. The data will be split 80% training and 20% test. A total of 3 logistic regression models will be evaluated, (1) contains only the hypothesis variables, (2) contains revised variables and (3) will utilize downsampling to balance the target variable.

In R, the logistic regression model will be constructed using supplementary libraries and tables to augment the model and facilitate visualizations. Below is the logistic regression formula for Model 1. Predictions will be determined such that if P is greater than 0.5, it will be classified as 'Yes,' and if P is less than 0.5, it will be classified as 'No'.

$$P(subscribed) = \frac{1}{1 + e^{-(z)}}$$

Z = $$\beta_0 + \beta_1 \times \ Age + \ \beta_2 \times Occupation + \ \beta_3 \times Employment\ Variation\ Rate \\ + \ \beta_4 \times Education\ Level + \ \beta_5 + House\ Loan$$

2.1 Model Selection

The sequential regression method will be employed, involving the systematic addition and subtraction of predetermined hypotheses to a base model to assess the overall accuracy of the models. Subsequently, a p-value selection will be conducted to identify additional variables for expanding the model, following the guidance of Stoltzfus, J.C. (2011). This approach aims to streamline the model's scope while emphasizing the inclusion of variables with significant impact.

2.2 Data Quality Assessment

The data set contain 10,932 NA values – when viewing unknown as NA – making up 26.66% of the total dataset. A significant portion of these unknowns can be related to the data collection method of telecommunication with a unprepared respondent. It was determined that treating all unknowns as NA would impact the overall data quality. After data cleaning, the dataset was reduced to 39,936 observations, a reduction of 2.5%.

2.2.1 Variable Exclusion

‘Credit_default’ and ‘poutcome’ were found to possess significant missing data. This was primarily due to skewed data distribution — ‘Credit_default’ had only one 'yes' instance, while ‘poutcome’ contained a substantial number of 'unknown' values, rendering it an invalid feature. These variables were removed to reduce the number of records requiring alterations.

2.2.2 Extreme Data

The 'Age' attribute contained instances with a value of 999, prompting the removal of any ages exceeding 100 from the dataset. Contact duration contained a high skew of 3.26 with a range of 4,918, it was determined that any value above 1,000 would be removed to limit the spread of the variable reducing the skew to 1.55.

Additionally, the 'pdays' feature exhibited extreme values of 999, indicating that the customer had never been contacted. To accommodate the nature of this variable, it was transformed into a binary factor distinguishing between contacted and non-contacted instances, ensuring a balanced representation within the data.

2.2.3 Imputation

The ‘housing_loan’ and ‘personal_loan’ variable initially included 986 values of “unknown” as well as occupation contained 313, which were viewed as NA. The ‘mice’ package was utilized to predict and substitute for these missing values to enhance the datasets completeness. This approach aimed to preserve the data’s integrity by reducing the necessity for removing incomplete records (Buuren, S. van and Groothuis-Oudshoorn, K. (2011).

2.2.4 Missing Data

The column 'marital_status' had 80 NAs, constituting only 0.19% of the entire dataset. Consequently, after careful consideration, the decision was made to remove these records, deeming it an appropriate course of action.

2.2.5 Duplicate Factors

‘day_of_week’ contained both ‘tue’ and ‘tues’, these values where combined into 1 factor.

2.3 Model Quality Assessment

The assessment of the model's quality follows Hosmer, D.W., et al. (2013) criteria, emphasizing overall accuracy, the ROC curve, Cohen’s Kappa, precision, and sensitivity. Model comparison hinges on the Akaike Information Criterion (AIC) value, guiding the selection model variables.

3. Results

Figure 1 illustrates an imbalanced dataset, comprising 35,878 "No" compared to only 4,058 "Yes". Upon comparing subscription by ‘Age’, minor bias is observed, leaning slightly towards older age ranges, with "Yes" instances showing a slightly higher average age of 41.08, contrasting with the "No" at 39.92, figure 2..

Figure 1: Subscription Distribution

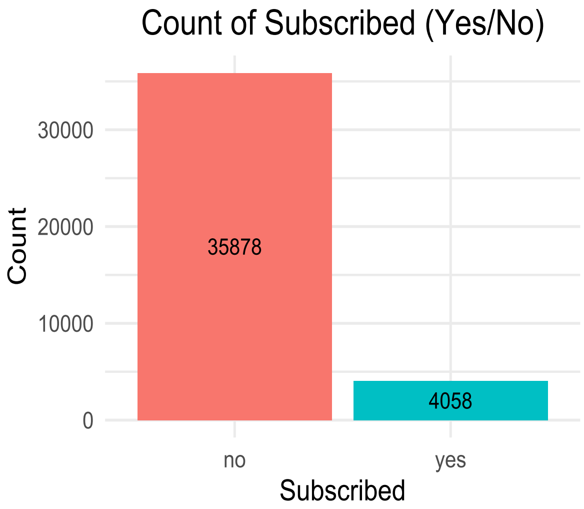
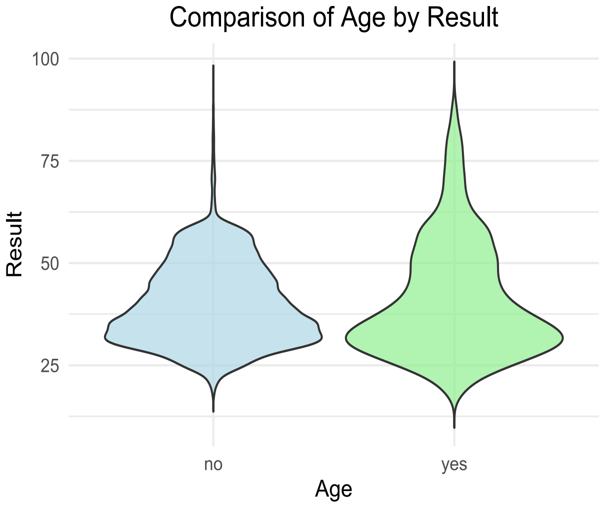


Figure 2: Age Distribution



The 'emp_var_rate' exhibits a moderate negative skew of -0.71. Surprisingly, the figure illustrates that when the employment variation rate is negative, over 40% of total subscriptions for 'Yes' occurred—a counterintuitive observation. Analysing occupations reveals that professional categories like 'admin' and 'technicians' constitute a substantial portion of TDs, in contrast to categories such as 'self-employed' or 'housemaid'.

Figure 4: Employment Variation by Subscription Proportion

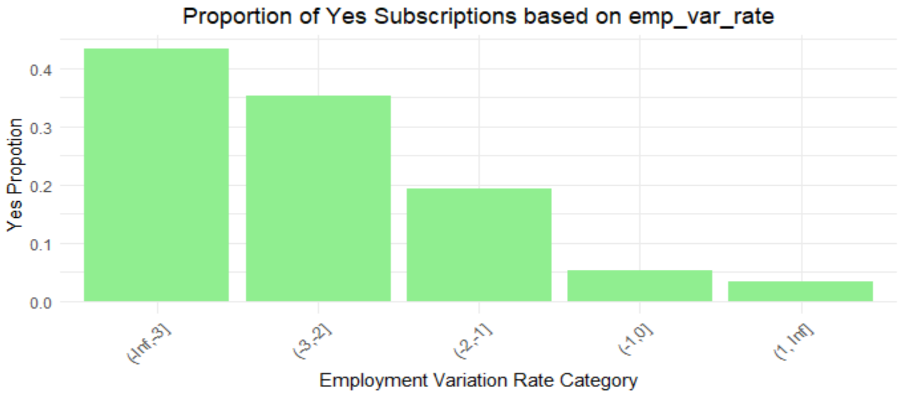
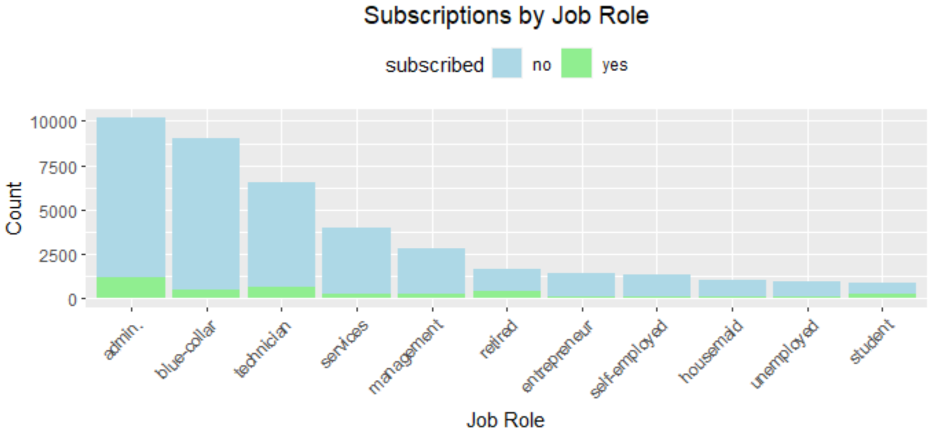


Figure 5: Subscriptions by Occupation



3.1 Logistic Regression Model Assessment

The initial logistic regression model 1, as shown in the table 1, resulted in an AIC of 17,725. Upon the inclusion of additional variables based on a selection method, this decreased to 15,901. The increases in variables enhanced the model. 'Emp_var_rate', 'cons_price_idx', and 'cons_conf_idx' where all significant indicators of TD rate (below 0.001).

TABLE 1: LOGISTIC REGRESSION MODELS

	Dependent variable:		
	subscribed		
	(1)	(2)	(3)
AGE	0.004*	-0.003	-0.005
	(0.002)	(0.002)	(0.003)
HOUSING_LOANYES	-0.056		
	(0.040)		
EMP_VAR_RATE	-0.598***	-0.796***	-0.818***
	(0.013)	(0.023)	(0.044)
CONS_PRICE_IDX		1.147***	1.117***
		(0.058)	(0.116)
CONS_CONF_IDX		0.028***	0.032***
		(0.005)	(0.009)
OCCUPATIONBLUE-COLLAR	-0.509***	-0.252***	-0.203*
	(0.079)	(0.083)	(0.115)
OCCUPATIONTECHNICIAN	-0.058	-0.016	0.042
	(0.069)	(0.074)	(0.106)
OCCUPATIONSERVICES	-0.300***	-0.139	-0.106
	(0.084)	(0.089)	(0.124)
OCCUPATIONMANAGEMENT	-0.224***	-0.050	0.129
	(0.083)	(0.089)	(0.132)
OCCUPATIONRETIRED	0.379***	0.230**	0.377**
	(0.102)	(0.111)	(0.178)
OCCUPATIONENTREPRENEUR	-0.466***	-0.163	-0.138
	(0.127)	(0.132)	(0.177)

OCCUPATIONSELF-EMPLOYED	-0.345***	-0.176	-0.275
	(0.119)	(0.126)	(0.175)
OCCUPATIONHOUSEMAID	0.105	0.014	-0.145
	(0.137)	(0.150)	(0.211)
OCCUPATIONUNEMPLOYED	0.161	0.014	0.126
	(0.120)	(0.131)	(0.207)
OCCUPATIONSTUDENT	0.658***	0.213*	0.190
	(0.103)	(0.115)	(0.196)
CONTACT_METHODTELEPHONE		-0.596***	-0.491***
		(0.072)	(0.109)
MONTHAUG		0.217**	0.347*
		(0.108)	(0.194)
MONTHDEC		0.424**	0.134
		(0.198)	(0.384)
MONTHJUL		0.204**	0.258*
		(0.097)	(0.154)
MONTHJULY		-10.853	-12.140
		(108.308)	(160.659)
MONTHJUN		-0.169*	-0.369**
		(0.093)	(0.145)
MONTHMAR		1.294***	1.152***
		(0.120)	(0.223)
MONTHMAY		-0.554***	-0.653***
		(0.077)	(0.119)
MONTHNOV		-0.368***	-0.450***
		(0.099)	(0.153)
MONTHOCT		0.077	0.349
		(0.126)	(0.235)
MONTHSEP		-0.038	0.033
		(0.135)	(0.271)
CAMPAIGN		-0.065***	-0.075***
		(0.012)	(0.015)
PDAYSNOT CONTACTED		-1.373***	-1.442***
		(0.074)	(0.168)
EDUCATION_LEVELHIGH.SCHOOL	-0.254***	-0.098	0.006
	(0.058)	(0.063)	(0.092)
EDUCATION_LEVELBASIC.9Y	-0.438***	-0.221**	-0.066
	(0.081)	(0.087)	(0.120)
EDUCATION_LEVELPROFESSIONAL.COURSE	-0.085	-0.035	-0.141
	(0.072)	(0.078)	(0.113)
EDUCATION_LEVELBASIC.4Y	-0.222**	-0.126	-0.017
	(0.089)	(0.096)	(0.139)
EDUCATION_LEVELBASIC.6Y	-0.378***	-0.131	0.025
	(0.115)	(0.121)	(0.167)
EDUCATION_LEVELUNKNOWN	0.046	-0.015	-0.078
	(0.100)	(0.109)	(0.165)
EDUCATION_LEVELILLITERATE	0.689	1.067	0.309
	(0.738)	(0.691)	(1.053)
CONSTANT	-2.365***	-106.703***	-101.459***
	(0.095)	(5.428)	(10.872)
OBSERVATIONS	31,950	31,950	6,494
LOG LIKELIHOOD	-8,841.640	-7,915.907	-3,258.630
AKAIKE INF. CRIT.	17,725.280	15,901.810	6,587.261
NOTE:			*p**p***p<0.01

Table 2 shows a comprise of the models. Model 2 attained the highest accuracy of 91.31% without displaying significant bias toward positive or negative results. Its McNemar's Test yielded a P-value of <2.2e-16, reinforcing its significance as a predictor. While balanced Model 3 exhibited a slightly higher Cohen's Kappa (0. 3902) than Model 2 (0. 3446), suggesting improved predictive capability, it notably achieved a significantly higher specificity of 0. 6942 compared to Model 2 (0. 2639). The ROC curve, figure supports the accuracy of the model limited drop of between sensitivity and specificity. A multicollinearity check was conducted, table 3, which outputted insignificant values.

Table 2: Logistic Regression Statistics

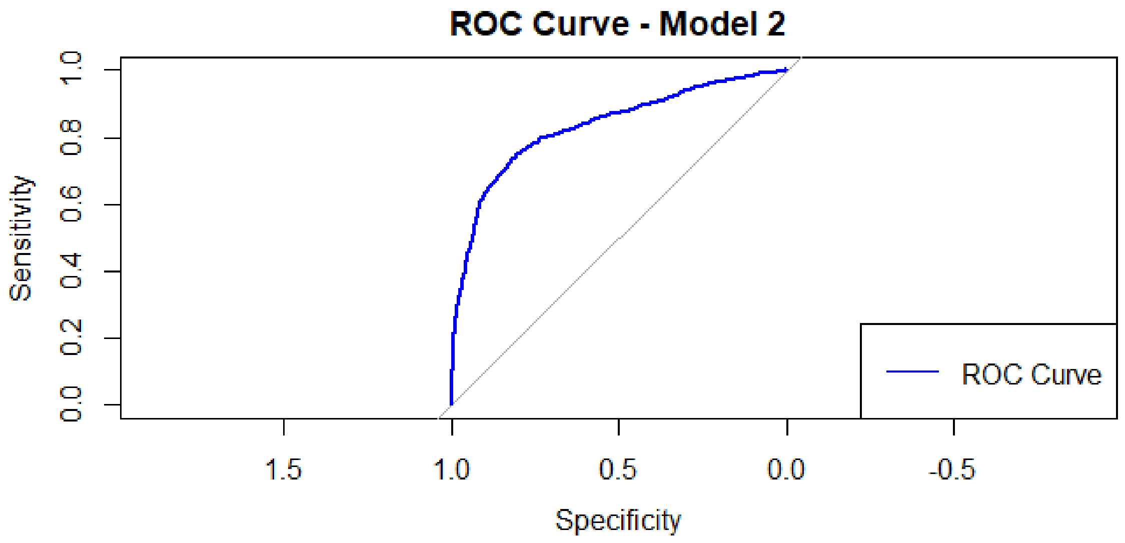
	Model 1	Model 2	Model 3
Accuracy	0.8989	0.9131	0.8422
95% CI	(0.8901, 0.9035)	(0.9067, 0.9192)	(0.834, 0.8502)

Mcnemar's Test P-Value	<2e-16	<2.2e-16	<2e-16
Cohen's Kappa	0.3181	0.3446	0.3902
Sensitivity	0.99526	0.9865	0.8590
Specificity	0.02713	0.2639	0.6942
Pos Precision	0.90050	0.9222	0.9613
Neg Precision	0.39286	0.6881	0.3575

Table 3: Multicollinearity

Variable	GVIF	DF	GVIF(1/(2*DF)
Age	1.897405	1	1.377463
Occupation	5.210711	10	1.086038
Contact Method	1.742502	1	1.320039
Month	4.670189	10	1.080107
Campaign	1.034190	1	1.016951
Pdays	1.154879	1	1.074653
Employment Variation Rate	3.403695	1	1.844911
Consumer Price Index	3.258726	1	1.805194
Consumer Confidence Index	2.62572	1	1.537391
Education Level	3.091801	7	1.083965

Figure 6: Model 2 ROC Curve



3.2 Hypothesis Testing

3.2.1 H_1 Age

Age displayed marginal significance (p = 0.08) within the model 1, exhibiting a positive relationship at 0.003734. When isolated in a T test against the target variable, it demonstrated a highly significant relationship (p = 5.77e-07), indicating its influence beyond the model and suggesting interactions with other variables. Age isn't a significant indicator of TD rate when assessed from the model 2.

3.2.2 H_2 Occupation

Occupation displayed a range of significance depending on category, 'Blue-collar' and 'Self-employed' having significant P values below 0.05 but 'Unemployed' and 'Management had P values that where extremely high of 0.6 and 0.9 respectively showing no relationship. A Chi-squared test obtained a P value of 2.2e-16 suggesting that there is an overall relationship.

3.2.3 H_3 Employment Variation Rate

The EVR exhibited an extremely low p-value of <2e-16, confirming its significance with an intercept of 7.9593e-01 and supporting the hypothesis. In a similar vein, a T test also yielded a significant p-value. Interestingly, the mean value for the 'Yes' category was -1.42, indicating a negative relationship compared to the mean for the 'No' category, which stood at 0.24.

3.2.4 H_4 Education Level

A fisher test was conducted on education level which achieved a p-value of <2.2e-16 indicting that there is a strong correlation. The logistic model 1 showed the 3:7 had a significant p-value. However, none where significant in model 2 with only ‘illiterate’ achieve a p value of 0.08519, still above the 0.05 cut off of significance.

3.2.5 H_5 Housing loan

Housing loan was shown to be an insignificance indicator of TD rate with a p-value of 0.159. This variable was removed to form the final model due to its insignificance.

4. Discussion

Initially, downsampling to balance the model resulted in a decrease in overall accuracy, indicating an unfavourable outcome. However, the balanced model showcased improved predictive capabilities, evident in its higher Cohen Kappa (0.3902) and specificity (0.6942). These metrics indicate its effectiveness in predicting both positive and negative cases, rendering it a robust general predictor.

Model 2, leveraging the significantly imbalanced dataset, benefitted from the larger data volume, enhancing its accuracy in predicting positive TD rates. Exploring alternative balancing methods like ROSE, SMOTE, or upsampling could potentially further enhance the model's performance. Nonetheless, considering an accuracy score of 91%, deemed satisfactory for this report, further adjustments may not be imperative matching contemporary results (Jiang, E. et al. 2022). An in-depth comparative study, exploring additional machine learning techniques such as Random Forest or Gradient Boosting, might yield a superior predictive model. Dhankhad, S., Mohammed, E., and Far, B. (2018) identified improvements through their research, showcasing the potential enhancements offered by these methodologies.

The use of a logistic regression model to assess variables with a non-linear relationship (Harrell, F.E. 2010) has been proven to be complex and often ineffective. Showcased by the ‘Age’ variable having an extremely low intercept of -0.0002544. Due to Age having a bell curved shape the model is a poor predictor if its effects.

Education level might indeed have an influence on term deposit rates, as indicated by Gray, D., Montagnoli, A., and Moro, M.'s (2021) findings suggesting that individuals with higher education levels tend to have a deeper understanding of financial circumstances. This understanding could potentially lead to a non-linear relationship, where those with higher education levels demonstrate greater consideration of external factors when deciding on term deposit subscriptions effecting both positive and negative outcomes.

5. Conclusion

The logistic regression model demonstrated remarkable effectiveness in predicting TD rates with an impressive overall accuracy of 91%. Although implementing downsampling didn't enhance the overall accuracy, it notably improved the model's generalization by narrowing the gap between sensitivity and specificity, refining its balance in performance metrics.

Hypotheses H_1 and H_5 were found to be unsupported, indicating that they are not significant indicators of TD rates, with no observed correlation for housing loans and age. Conversely, H_2 and H_3 emerged as significant predictors of TD rates. However, the non-linear impact of variable H_1 hindered its overall usability within the model. Consider employing a binning technique to potentially enhance its utility. H_4 overall significance can't be determined, variation in categories significant implied some factor were good indicator but other are not. A greater analysis is required.

References

Broby, D. (2022) 'The use of predictive analytics in Finance', *The Journal of Finance and Data Science*, 8, pp. 145–161. doi:10.1016/j.jfds.2022.05.003.

Buuren, S. van and Groothuis-Oudshoorn, K. (2011) '**mice**: Multivariate imputation by chained equations in*r*', *Journal of Statistical Software*, 45(3). doi:10.18637/jss.v045.i03.

Chen, J. (2022) *Term deposit: Definition, how it's used, rates, and how to invest, Investopedia*. Available at: https://www.investopedia.com/terms/t/termdeposit.asp (Accessed: 03 January 2024).

Cole, S.A., Paulson, A.L. and Shastry, G.K. (2012) 'Smart money: The effect of education on financial behavior', *SSRN Electronic Journal* [Preprint]. doi:10.2139/ssrn.1317298.

Dhankhad, S., Mohammed, E. and Far, B. (2018) 'Supervised machine learning algorithms for credit card fraudulent transaction detection: A comparative study', *2018 IEEE International Conference on Information Reuse and Integration (IRI)* [Preprint]. doi:10.1109/iri.2018.00025.

Ganong, P. and Noel, P. (2022) 'Reproduction of "consumer spending during unemployment: Positive and normative implications"', *AMERICAN ECONOMIC REVIEW* [Preprint]. doi:10.48152/ssrp-t4bd-4e87.

Gray, D., Montagnoli, A. and Moro, M. (2021) 'Does education improve financial behaviors? quasi-experimental evidence from Britain', *Journal of Economic Behavior &amp; Organization*, 183, pp. 481–507. doi:10.1016/j.jebo.2021.01.002.

Harrell, F.E. (2010) *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. New York, NY: Springer.

Hassan, A., Salim, R. and Bloch, H. (2010) 'Population age structure, saving, capital flows and the real exchange rate: A survey of the literature', *Journal of Economic Surveys*, 25(4), pp. 708–736. doi:10.1111/j.1467-6419.2010.00665.x.

Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (2013) *Applied logistic regression David W. Hosmer, Stanley Lemeshow, Rodney X. Sturdivant*. Hoboken, NJ: Wiley.

Jiang, E., Wang, Z. and Zhao, J. (2022) 'Prediction of term deposit in Bank: Using Logistic Model', *BCP Business &amp; Management*, 34, pp. 607–614. doi:10.54691/bcpbm.v34i.3071.

Lee, W. and Seo, K. (2022) ‘Downsampling for binary classification with a highly imbalanced dataset using active learning’, *Big Data Research*, 28, p. 100314. doi:10.1016/j.bdr.2022.100314.

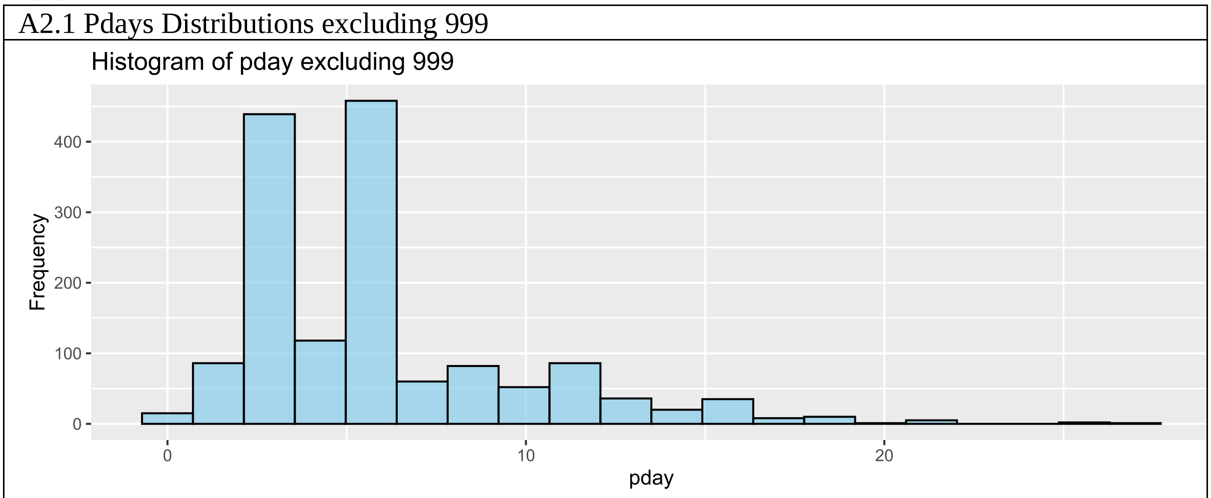
Mooijman, P. *et al.* (2023) ‘The effects of data balancing approaches: A case study’, *Applied Soft Computing*, 132, p. 109853. doi:10.1016/j.asoc.2022.109853.

Stoltzfus, J.C. (2011) ‘Logistic regression: A brief primer’, *Academic Emergency Medicine*, 18(10), pp. 1099–1104. doi:10.1111/j.1553-2712.2011.01185.x.

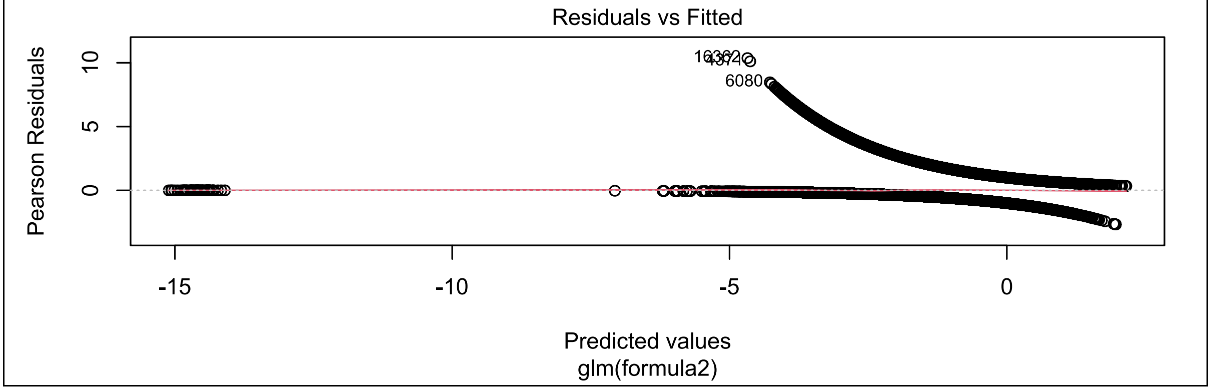
Trivedi, N.K. *et al.* (2020) ‘KFCM-based direct marketing’, *Rising Threats in Expert Applications and Solutions*, pp. 495–502. doi:10.1007/978-981-15-6014-9_57.

Appendix 1

A2.1 Pdays Distributions excluding 999



A2.2 Residual Vs Fitted


Residuals vs Fitted
Pearson Residuals
Predicted values
glm(formula2)