



Assessment #3

Telecom Marketing Campaign

Camilo Andres Uribe Guerra
ID: 25416518
MSc. Data Science for Innovation

Table of Contents

Table of Contents	2
1. Introduction	3
Problem Statement	3
Rationale.....	3
Project aims and objective	3
2. Methodology	4
Methods overview.....	4
Methods detail	4
3. Results.....	5
Key Findings.....	5
In depth results.....	5
4. Conclusions	8
5. References.....	9
6. Appendix	10
A. Data Dictionary.....	10
B. Confusion Matrix Results for RF Best Model.	11
C. Logistic Stats Model Analysis (No tuning)	12
D. Feature Engineering Binned Creation	13

1. Introduction

Problem Statement

In the competitive Australian telecommunications sector, marketing campaigns are vital for customer retention and profitability. TelecomSyd company aims to identify customer segments likely to respond positively to campaigns to optimize future strategies.

Rationale

Effective marketing boosts profitability and engagement. By leveraging past and current campaign data, TelecomSyd seeks to allocate resources efficiently and enhance outcomes through predictive insights.

Project aims and objective

The goal of this research is to build predictive models to estimate customer responses to TelecomSyd's marketing campaigns. The main research questions include:

1. Which predictive model—parametric or non-parametric—best estimates customer response?
 - H0: Both model types perform equally.
2. Do customer demographics or behaviors influence model performance?
 - H0: Demographic or behavioral features do not significantly affect performance.
3. Do past campaign outcomes impact future predictions?
 - H0: Previous outcomes do not significantly influence future predictions.

The objectives are:

- Develop and compare parametric and non-parametric models for customer response prediction.
- Identify key demographic and behavioral features affecting campaign success.
- Analyze the influence of previous campaign outcomes on future predictions.

2. Methodology

Methods overview

The project followed a structured approach involving EDA, data preprocessing, feature selection, and model development. Both parametric (Logistic Regression) and non-parametric (Random Forest) models were used, with hyperparameter tuning and cross-validation to identify the best-performing model for further analysis.

Methods detail

Feature Selection and Engineering

Feature selection involved evaluating numerical features using biserial correlation and categorical features with Cramér's V. The duration variable was excluded, as it only becomes known after campaign interactions and cannot be used for future predictions. Variables such as `cons.price.idx` and `emp.var.rate` were also excluded due to multicollinearity, being redundant with `euribor3m` and `nr.employed`. Additional feature engineering included binning previous contact counts and campaign interactions to enhance interpretability.

Data Transformation for Modeling

Data transformation included standard scaling for numerical features to standardize their ranges and one-hot encoding for categorical variables to ensure they were suitable for modeling. The processed data was then split into training, validation, and test sets, maintaining class distribution using stratified sampling.

Model Development

An initial logistic regression model was created using a basic stats model setup to assess its statistical properties, convergence, and characteristics without hyperparameter tuning, establishing a foundation for the parametric model. Both logistic regression and random forest classifiers underwent hyperparameter tuning through grid search with cross-validation, utilizing maximum likelihood estimation (MLE) for fitting the parametric model (Scikit-learn Developers, 2024) and employing ensemble methods for the non-parametric model. Precision and F1 score were the main evaluation metrics, with an emphasis on precision to accurately identify responsive customers, crucial for targeted marketing.

Model Comparison and Feature Importance

Models were compared based on cross-validation scores. The evaluation highlighted the best model, which was further analyzed for feature importance to guide business insights.

Note: The final set of features and parameters used for modeling are provided in the appendix for reference.

3. Results

Key Findings

1. The initial logistic regression model did not converge, indicating challenges in model fitting without hyperparameter tuning (Appendix C).
2. The random forest model outperformed logistic regression in precision (36.6% validation precision), supported by the Mann-Whitney U test which showed a significant difference in precision ($p < 0.05$) using 10 folds for best model.
3. Economic indicators (nr.employed, euribor3m, and cons.conf.idx) were the most influential features in predicting campaign success, aligning with the importance of financial metrics. In contrast, customer demographic and behavioral features had less impact.
4. The final random forest model demonstrated consistency across training, validation, and testing datasets (Figure 1), though the overall precision indicates further data refinement and acquisition is needed for better performance.
5. Feature importance analysis confirmed the relevance of previous campaign outcomes (pdays and poutcome_success), supporting the third hypothesis regarding their influence on future campaign responses.

In depth results

Initial Logistic Regression Analysis

The initial logistic regression model was assessed to understand the statistical behavior and overall convergence. The model did not converge (Appendix, Table 1), indicating potential limitations in fitting the data accurately without hyperparameter tuning. Some coefficients, such as pdays ($p < 0.001$) and nr.employed ($p < 0.001$), were found to be statistically significant, while others like job_blue-collar and education_high.school were not significant. The full list of coefficients and p-values can be found in the regression results table in the appendix.

Model Results with Hyperparameter Tuning

Logistic Regression The best logistic regression model was obtained with the following parameters: {'C': 1, 'fit_intercept': False, 'penalty': 'l2', 'solver': 'lbfgs'}. This model achieved:

- F1 Score (Training): 45.5%
- Precision (Training): 35.6%
- F1 Score (Validation): 43.1%
- Precision (Validation): 33.4%

Random Forest The optimal random forest model was tuned with parameters: {'bootstrap': True, 'max_depth': 5, 'min_samples_leaf': 2, 'min_samples_split': 4, 'n_estimators': 100}. This model produced:

- F1 Score (Training): 47.6%
- Precision (Training): 38.3%
- F1 Score (Validation): 45.5%
- Precision (Validation): 36.6%

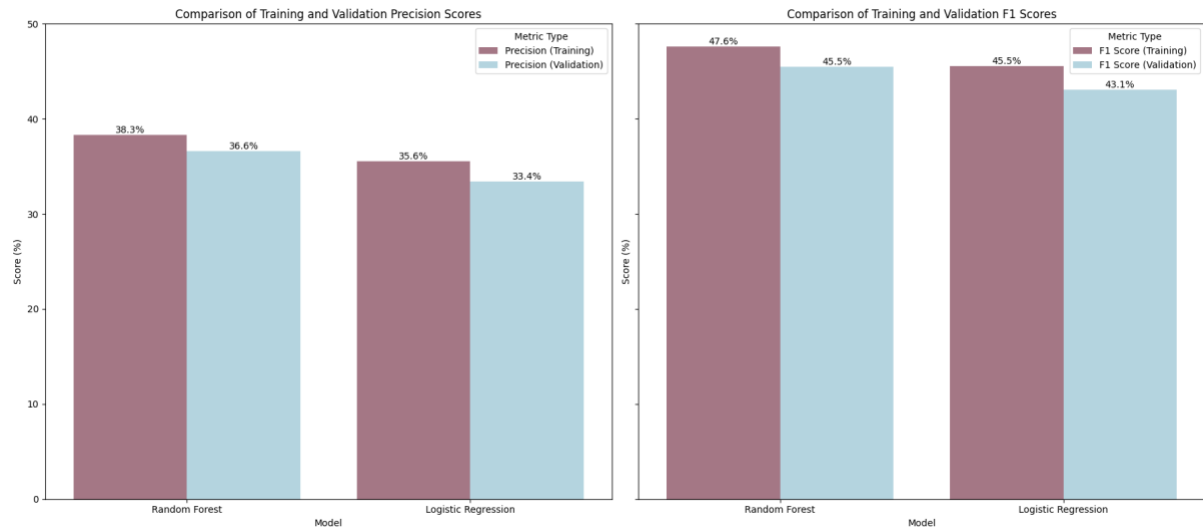


Figure 1. Best Models Performance Indicators using GridSearchCV for Logistic Regression and RandomForest across Training and Validation Datasets.

Mann-Whitney U Test for Model Comparison

The Mann-Whitney U test was conducted to compare the precision and F1 scores of the two models:

- Precision - U statistic: 80.0, p-value: 0.025 (significant difference)
- F1 Score - U statistic: 69.0, p-value: 0.162 (not significant)

Based on these results, the random forest model was chosen for its higher precision, with significant evidence supporting its superior performance over logistic regression.

Model Consistency

Figure 2 illustrates the training, validation, and testing results for precision and F1 scores across the Random Forest model, showing that it maintained consistent performance across datasets. The final confusion matrix for the training set (Appendix B) further verifies the performance.



Figure 2> Final Random Forrest Model consistency performance indicators across all splitting datasets.

Feature Importance Analysis

The feature importance analysis (Figure X) revealed that nr.employed was the most influential variable, highlighting the role of employment metrics. Financial indicators such as euribor3m and cons.conf.idx were also significant, confirming the impact of economic conditions. Variables related to campaign interactions, such as pdays and poutcome_success, showed moderate relevance, partially supporting the third hypothesis regarding the influence of previous campaign outcomes. Additionally, months like March, May, and October emerged as potential periods to target for campaigns. Most customer demographic and behavioral features demonstrated minimal impact, supporting the second hypothesis that these characteristics do not significantly affect model performance.

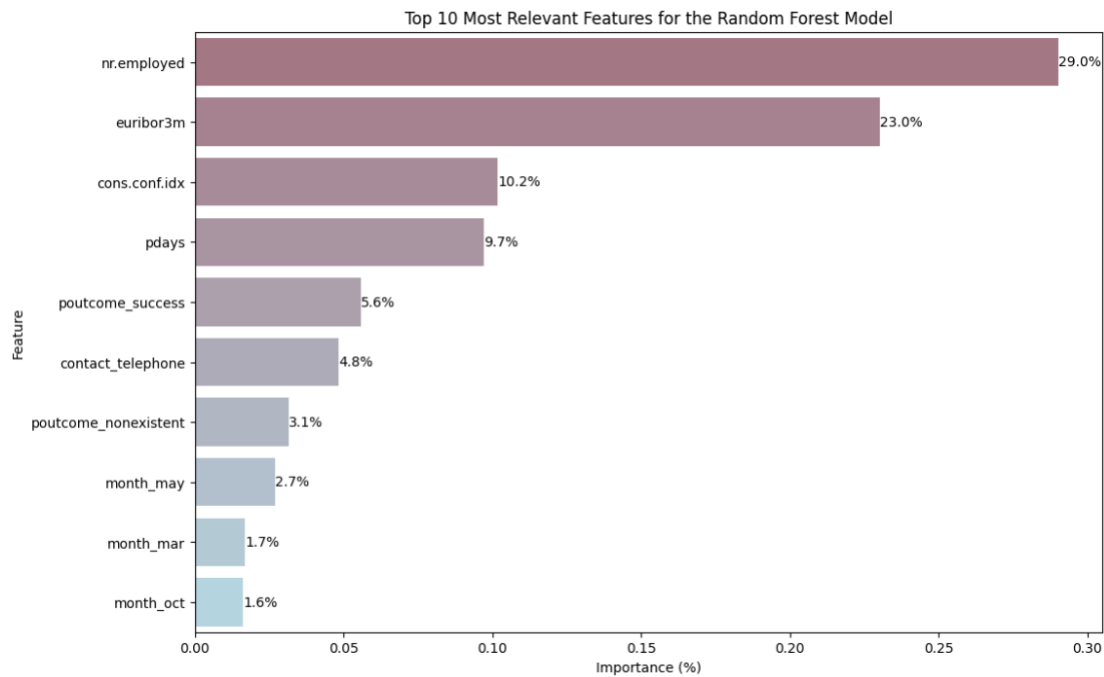


Figure 3. Feature Importance for Random Forrest best model (% by feature)

4. Conclusions

- The Random Forest model outperformed Logistic Regression, achieving precision (35.8%) and F1 score (44.9%) on testing dataset. However, the overall precision was still low, indicating a need for further experiments, data refinement, and data collection, showing this model is not suitable for business cases.
- Key features impacting predictions included economic indicators such as the number of employed individuals, the 3-month Euribor rate, and the consumer confidence index, underscoring the significant influence of broader economic conditions on campaign responses.
- Previous campaign interactions, such as the number of days since the client was last contacted and the success outcome of prior campaigns, held moderate importance but can influence results in future campaigns. These factors contribute to defining successful marketing targets based on past campaign experiences.

5. References

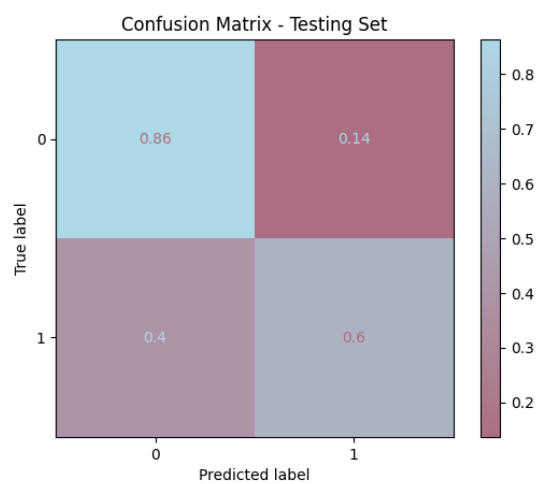
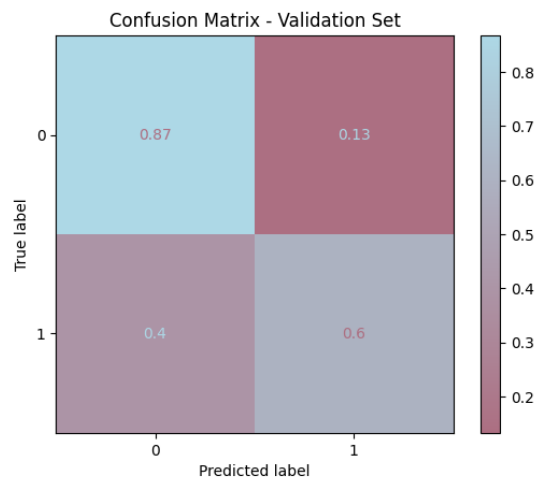
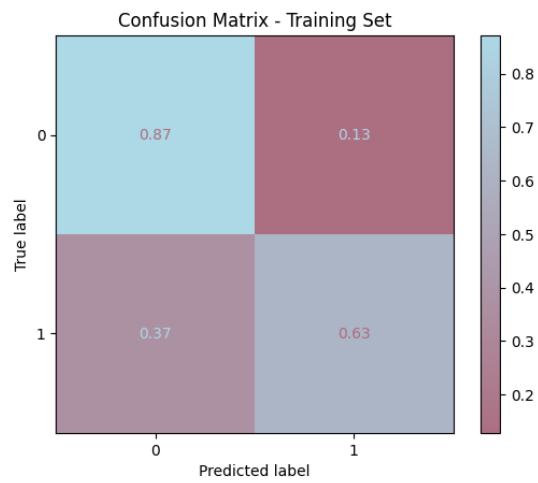
- Gao, W., Ding, Z., Cui, T., & Tao Cui. (2022). Construction of Digital Marketing Recommendation Model Based on Random Forest Algorithm. *Security and Communication Networks*, 2022, 1–9. <https://doi.org/10.1155/2022/1871060>
- Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector. *IEEE Access*, 7, 60134–60149. <https://doi.org/10.1109/ACCESS.2019.2914999>
- Scikit-learn Developers. (2024). *LogisticRegression*. Scikit-learn: Machine Learning in Python. https://scikit-learn.org/1.5/modules/generated/sklearn.linear_model.LogisticRegression.html

6. Appendix

A. Data Dictionary

Variable Name	Description
age	Age
job	Type of job
marital	Marital status
education	Level of education
default	Has credit in default
balance	Average yearly balance
housing	Has a housing loan
loan	Has a personal loan
contact	Contact communication type
day	Day of contact
month	Month of contact
duration	Last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y="no"). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known.
campaign	Number of contacts performed during this campaign and for this client
pdays	Number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
previous	Number of contacts performed before this campaign and for this client
poutcome	Outcome of the previous marketing campaign
emp.var.rate	employment variation rate - quarterly indicator (numeric)
cons.price.idx	consumer price index - monthly indicator (numeric)
cons.conf.idx	consumer confidence index - monthly indicator (numeric)
euribor3m	euribor 3 month rate - daily indicator (numeric)
nr.employed	number employed - quarterly indicator (numeric)
y	Did the client subscribe to a Telecom plan?

B. Confusion Matrix Results for RF Best Model.



C. Logistic Stats Model Analysis (No tuning)

Parameter	Value
No. Observations	26356
Model	Logit
Df Residuals	26310
Method	MLE
Df Model	45
Pseudo R-squ.	0.2169
Log-Likelihood	-7264
converged	FALSE
LL-Null	-9275.8
Covariance Type	nonrobust
LLR p-value	0

Coefficient	Coef	Std Err	z	P> z	[0.025	0.975]
const	-1.9468	nan	nan	nan	nan	nan
pdays	-0.224	0.047	-4.799	0	-0.315	-0.13
cons.conf.idx	0.0909	0.025	3.69	0	0.043	0.139
euribor3m	-0.0518	0.081	-0.64	0.522	-0.21	0.107
nr.employed	-0.7669	0.072	-10.656	0	-0.908	-0.63
job_blue-collar	-0.1164	0.086	-1.349	0.177	-0.286	0.053
job_entrepreneur	0.018	0.131	0.137	0.891	-0.238	0.274
job_housemaid	-0.129	0.162	-0.794	0.427	-0.447	0.189
job_management	0.0069	0.092	0.075	0.94	-0.174	0.188
job_retired	0.2885	0.103	2.792	0.005	0.086	0.491
job_self-employed	-0.0452	0.128	-0.354	0.724	-0.295	0.205
job_services	-0.0923	0.093	-0.993	0.321	-0.274	0.09
job_student	0.0696	0.124	0.563	0.573	-0.173	0.312
job_technician	0.0052	0.078	0.067	0.947	-0.148	0.158
job_unemployed	-0.0816	0.143	-0.569	0.569	-0.362	0.199
job_unknown	-0.0025	0.253	-0.01	0.992	-0.498	0.493
marital_married	-0.039	0.073	-0.531	0.596	-0.183	0.105
marital_single	0.0456	0.08	0.571	0.568	-0.111	0.202
marital_unknown	0.3784	0.459	0.825	0.409	-0.521	1.277

education_basic.6y	0.0712	0.131	0.546	0.585	-0.185	0.327
education_basic.9y	-0.0557	0.103	-0.54	0.589	-0.258	0.147
education_high.school	0.0014	0.1	0.014	0.989	-0.194	0.197
education_illiterate	1.0354	0.791	1.309	0.191	-0.515	2.586
education_professional.course	-0.0242	0.111	-0.219	0.827	-0.241	0.193
education_university.degree	0.0598	0.1	0.599	0.549	-0.136	0.256
education_unknown	0.1542	0.131	1.175	0.24	-0.103	0.411
default_unknown	-0.3165	0.073	-4.335	0	-0.46	-0.17
default_yes	-15.6819	8019.341	-0.002	0.998	-15700	15700
contact_telephone	-0.4245	0.073	-5.827	0	-0.567	-0.28
month_aug	-0.1013	0.112	-0.901	0.368	-0.322	0.119
month_dec	0.0233	0.229	0.102	0.919	-0.425	0.471
month_jul	0.2382	0.103	2.319	0.02	0.037	0.44
month_jun	0.223	0.098	2.268	0.023	0.03	0.416
month_mar	0.9031	0.136	6.627	0	0.636	1.17
month_may	-0.6775	0.081	-8.331	0	-0.837	-0.52
month_nov	-0.3364	0.112	-3.013	0.003	-0.555	-0.12
month_oct	-0.2053	0.147	-1.393	0.163	-0.494	0.083
month_sep	-0.4841	0.157	-3.078	0.002	-0.792	-0.18
previous_1 (Contacted Once)	-0.4686	nan	nan	nan	nan	nan
previous_2-3 (Few Contacts)	-0.6066	nan	nan	nan	nan	nan
previous_4+ (Many Contacts)	-0.8453	nan	nan	nan	nan	nan
campaign_2-3 Contacts	0.0047	0.048	0.098	0.922	-0.089	0.098
campaign_4-5 Contacts	-0.0766	0.08	-0.953	0.341	-0.234	0.081
campaign_6-10 Contacts	-0.3529	0.115	-3.061	0.002	-0.579	-0.13
campaign_11+ Contacts	-1.4928	0.361	-4.131	0	-2.201	-0.78
poutcome_nonexistent	-0.0264	nan	nan	nan	nan	nan
poutcome_success	0.6118	0.244	2.509	0.012	0.134	1.09

D. Feature Engineering Binned Creation

