Assignment 3

36103 Statistical Thinking for Data Science

Santiago Cardenas Carmona

ID: 25033147

Spring 2024

University of Technology Sydney

Table of contents

Problem formulation

A telecommunications company aims to predict customers' responses to a marketing campaign using machine learning (ML) instead of relying on arbitrary business rules. This approach allows the company to focus its efforts on customers who are more likely to purchase the new product, thereby reducing the costs of contacting a large pool of customers. Additionally, insights from the models can be used to identify relevant customer features to leverage engagement and retention strategies.

This project will explore two different types of ML models: parametric and non-parametric, comparing their performance and business implications in the current context. Both models will be evaluated based on the recall and F1 score, considering that the former assesses the model's capacity to identify the positive class by measuring the proportion of actual positives correctly identified, whereas the latter combines the model's precision and recall into a single metric, offering a balanced measure of both accuracy and completeness in identifying the positive class.

Data understanding and preparation

The dataset consists of demographic, economic, temporal, and other indicators related to a telecommunications company's marketing campaign for a new subscription service. It contains 21 columns and 41,180 rows, with no missing values. The key variable, 'y', indicates whether a customer subscribed. The data dictionary is provided in Table 1 (Appendix).

Fifteen duplicate records were removed before applying the models. Categorical and numerical features were tested against the two target classes (0 for non-adopters, 1 for adopters). A Chi-squared test was applied to categorical features, and only those with a statistically significant effect on campaign success (95% confidence) were selected, resulting in the removal of the 'loan' feature.

For numerical features, 'duration' was excluded as it is known only after the campaign outcome. The Mann-Whitney U test was used to compare adopters and non-adopters, and features where the null hypothesis of similar distribution was rejected at 5% significance were retained. Correlation analysis led to the removal of highly correlated features:

‘euribor3m’, ‘emp_var_rate’, and ‘cons_price_idx’. Features with low variability, such as ‘previous’, were also excluded.

The final numerical features were scaled to prevent bias in parametric models. Categorical features were one-hot encoded for nominal features (e.g., marital status, job) and label encoded for ordinal features (e.g., month, day, education level).

The dataset was split into training (70%) and validation (30%) sets using stratified sampling based on the ‘y’ variable to ensure similar proportions of adopters and non-adopters. This balanced distribution ensured fair model training and performance comparison on both seen and unseen data.

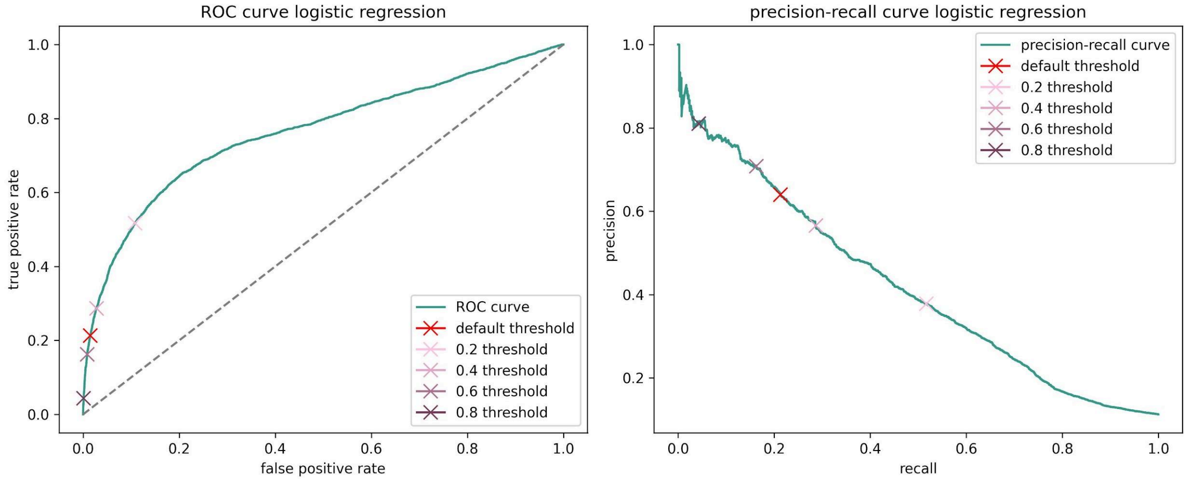
Modelling

This project will evaluate the performance of logistic regression and extra trees models using recall, F1 score, and confusion matrix results on both the training and validation datasets. The model that demonstrates higher and more consistent performance across each metric will be selected.

The logistic regression (LR) model uses a logistic function to estimate a discrete variable based on a set of features, where the function predicts the probability of a record belonging to a class, ranging from 0 to 1. The model was fitted on the training dataset using maximum likelihood estimation (MLE) to calculate the coefficients of each feature. The p-values of the features suggest that the coefficients for ‘age,’ ‘retirees,’ ‘unknown marital status,’ ‘positive credit default,’ and ‘non-existent previous contact’ are statistically zero (with a confidence level of 95%), implying that these features do not provide relevant information for predicting campaign success. The probability threshold to predict the positive target class was tuned using the ROC and precision-recall curves, as shown in Figure 1.

The extra trees (ET) model was implemented as the non-parametric alternative. This ensemble algorithm is similar to random forest, where the mode of multiple decision trees determines the class of a given record. However, in the ET model, the trees use the entire dataset instead of bootstrap samples, and the cut-off points for splitting nodes are chosen randomly. Multiple hyperparameter configurations were tested, and the best configuration (optimised for the highest F1 score) was achieved using 3-fold cross-validation.

Figure 1. ROC and precision-recall curves for logistic regression



Evaluation

The results from Table 2 shows that the best performance was achieved by the ET model, which delivered a consistent 0.51 F1 score and 0.58 recall both on training and unseen data. Even though the performance improvements of the ET model compared to LR are substantial, 15% higher F1 score and 20% higher recall, the best configuration of ET still struggles to identify just over 40% of the actual positive responses to the marketing campaign (Figure 2).

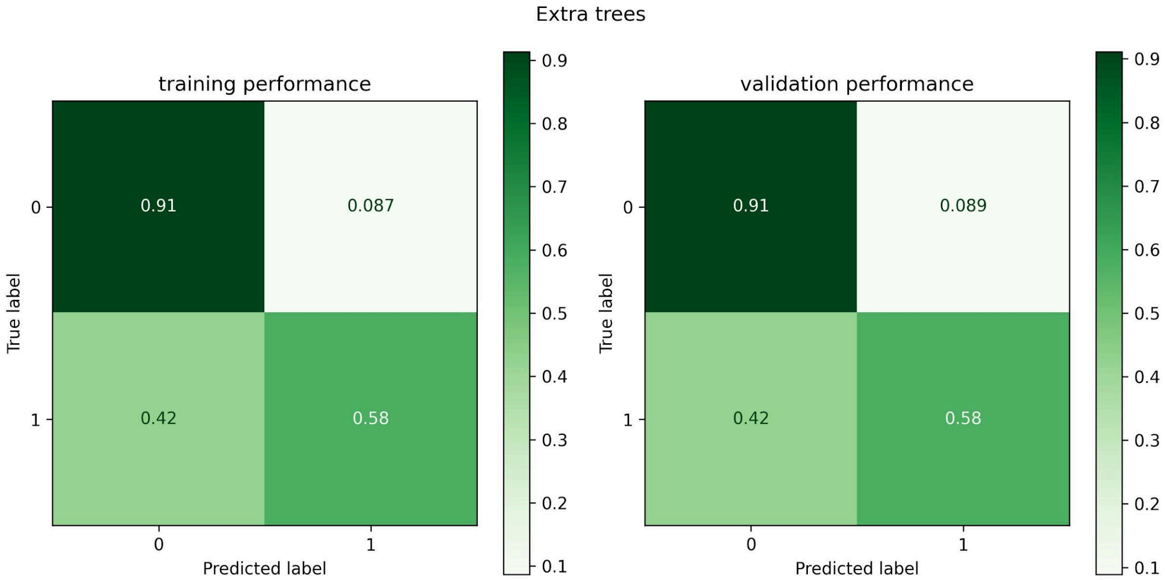
The struggle to improve recall without increasing the false positives might be attributed to the target variable imbalance found in the dataset, where only 11% of the records corresponds to positive outcomes on the marketing campaign. This issue makes it difficult to capture the underlying pattern of new product adopters by the algorithm.

If the best model were to be implemented, the efforts put into contacting customers will be reduced by 85% (assuming the sample has the same distribution of adopters to non-adopters as that of the original dataset), therefore reducing the campaign costs. However, the company will be capturing only around 60% of the adopters.

Table 2. Model performance comparison

Model	**F1 score**		**Recall**	
	Training	**Validation**	**Training**	**Validation**
LR	0.4373	0.4519	0.4750	0.5039
ET	0.5112	0.5077	0.5785	0.5794

Figure 2. Confusion matrix of ET model



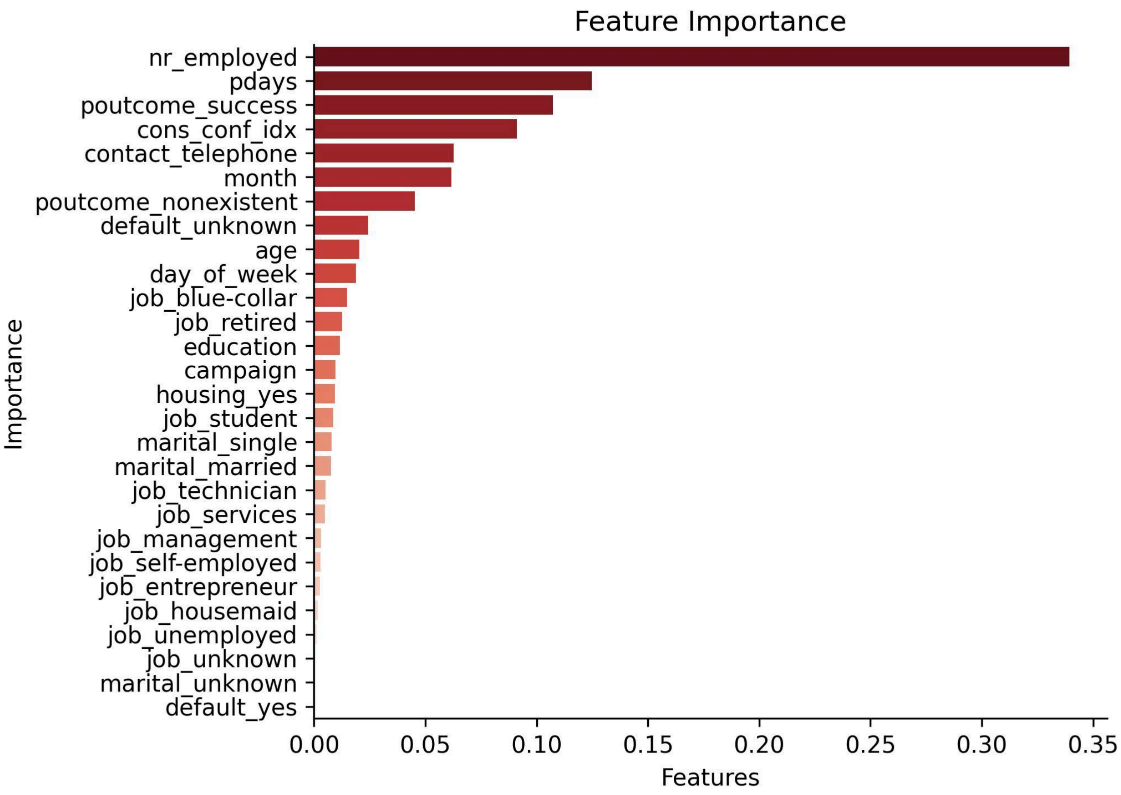
Conclusions

The experiments conducted with both parametric and non-parametric models show that the best model for predicting customers' responses to the marketing campaign is the Extra Trees (ET) model, as it achieved higher F1 score and recall compared to logistic regression (Table 2). However, even the best model configuration consistently fails to identify more than 40% of the actual positive responses to the campaign. This shortcoming implies that, while the business can narrow down the customers to target and reduce campaign costs, the company might incur a significant opportunity cost by missing customers who would have purchased the service if the model were deployed in its current stage.

Regarding the models used during experimentation, it is worth noting that parametric models were easier to tune and allowed for a broader interpretation of results. This advantage of parametric models might be important when addressing non-technical stakeholders. However, handling the greater complexity of non-parametric models yielded a positive payoff in performance, even though they narrow the model's implications to business impact.

Nevertheless, ET feature importance provides valuable insights into the determinants of campaign effectiveness. The top three features used to build the model's trees were the employment indicator, the number of days since the last contact, and the outcome of previous campaigns. Although the employment distribution by campaign outcomes shows significant overlap between the target classes (Appendix, Figure 4), the coefficient of this feature in the logistic regression model (and its significance level) suggests that higher levels of this indicator are associated with a decreased probability of success. Additionally, a successful outcome in a previous campaign increases the likelihood of success in the current campaign (Appendix, Figure 5 and Table 3).

Figure 3. Feature importance from ET model



Appendix

Table 1. Data dictionary

Column name	Definition
age	Age
job	Type of job
marital	Marital status
education	Level of education
default	Has credit in default
housing	Has a housing loan
loan	Has a personal loan
contact	Contact communication type
day_of_week	Day of contact
month	Month of contact

duration	Last contact duration, in seconds. This feature is known after
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
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
cons.price.idx	consumer price index - monthly indicator
cons.conf.idx	consumer confidence index - monthly indicator
euribor3m	euribor 3 month rate - daily indicator
nr.employed	number employed - quarterly indicator
y	Did the client subscribe to a Telecom plan? 0 = no, 1 = yes

Table 3. Logistic regression summary

Variable	**Coef**	**Std Err**	**z**	**P>|z|**	**[0.025**	**0.975]**
age	-0.0451	0.025	-1.785	0.074	-0.095	0.004
campaign	-0.1227	0.030	-4.126	0.000	-0.181	-0.064
pdays	-0.1531	0.043	-3.579	0.000	-0.237	-0.069
cons_conf_idx	0.2397	0.019	12.364	0.000	0.202	0.278
nr_employed	**-0.6250**	**0.021**	**-29.324**	**0.000**	**-0.667**	**-0.583**
education	-0.1862	0.014	-13.017	0.000	-0.214	-0.158
month	-0.1000	0.010	-9.618	0.000	-0.120	-0.080
day_of_week	-0.0356	0.014	-2.559	0.010	-0.063	-0.008
job_blue-collar	-0.8411	0.065	-12.884	0.000	-0.969	-0.713
job_entrepreneur	-0.5766	0.127	-4.535	0.000	-0.826	-0.327
job_housemaid	-0.5408	0.142	-3.807	0.000	-0.819	-0.262
job_management	-0.2640	0.086	-3.086	0.002	-0.432	-0.096
job_retired	-0.1238	0.106	-1.168	0.243	-0.331	0.084
job_self-employed	-0.3070	0.119	-2.579	0.010	-0.540	-0.074

job_services	-0.7669	0.081	-9.441	0.000	-0.926	-0.608
job_student	-0.2410	0.116	-2.074	0.038	-0.469	-0.013
job_technician	-0.3320	0.062	-5.395	0.000	-0.453	-0.211
job_unemployed	-0.3038	0.129	-2.353	0.019	-0.557	-0.051
job_unknown	-0.8604	0.250	-3.443	0.001	-1.350	-0.371
marital_married	-0.5784	0.056	-10.320	0.000	-0.688	-0.469
marital_single	-0.5360	0.067	-8.037	0.000	-0.667	-0.405
marital_unknown	-0.9288	0.477	-1.949	0.051	-1.863	0.005
default_unknown	-0.4249	0.067	-6.385	0.000	-0.555	-0.294
default_yes	-0.0514	2.771	-0.019	0.985	-5.482	5.380
housing_yes	-0.1684	0.040	-4.181	0.000	-0.247	-0.089
contact_telephone	-0.8850	0.056	-15.937	0.000	-0.994	-0.776
poutcome_nonexistent	**-0.0135**	**0.054**	**-0.247**	**0.805**	**-0.120**	**0.093**
poutcome_success	**0.7265**	**0.232**	**3.133**	**0.002**	**0.272**	**1.181**

Figure 3. Confusion matrix of logistic regression model with a positive class probability threshold of 0.225

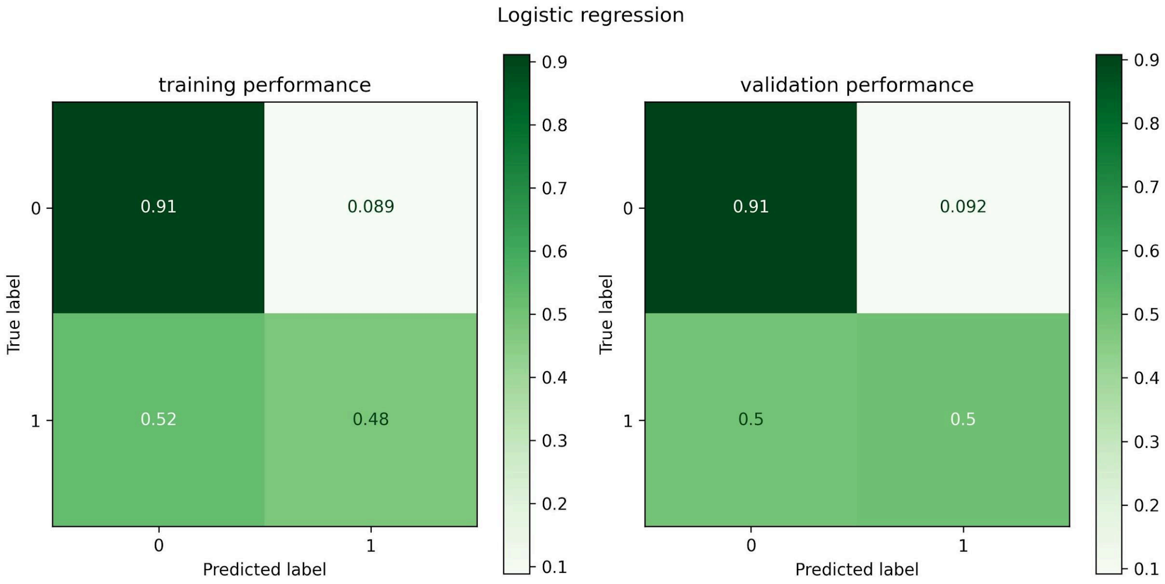


Figure 4. Distribution of success by employment indicators

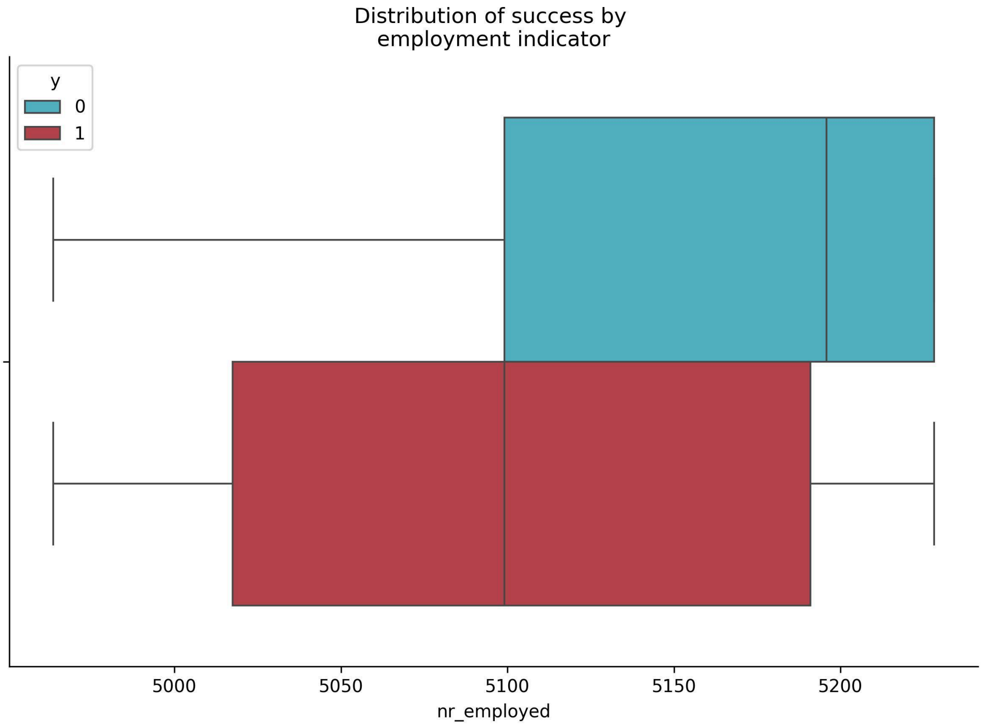
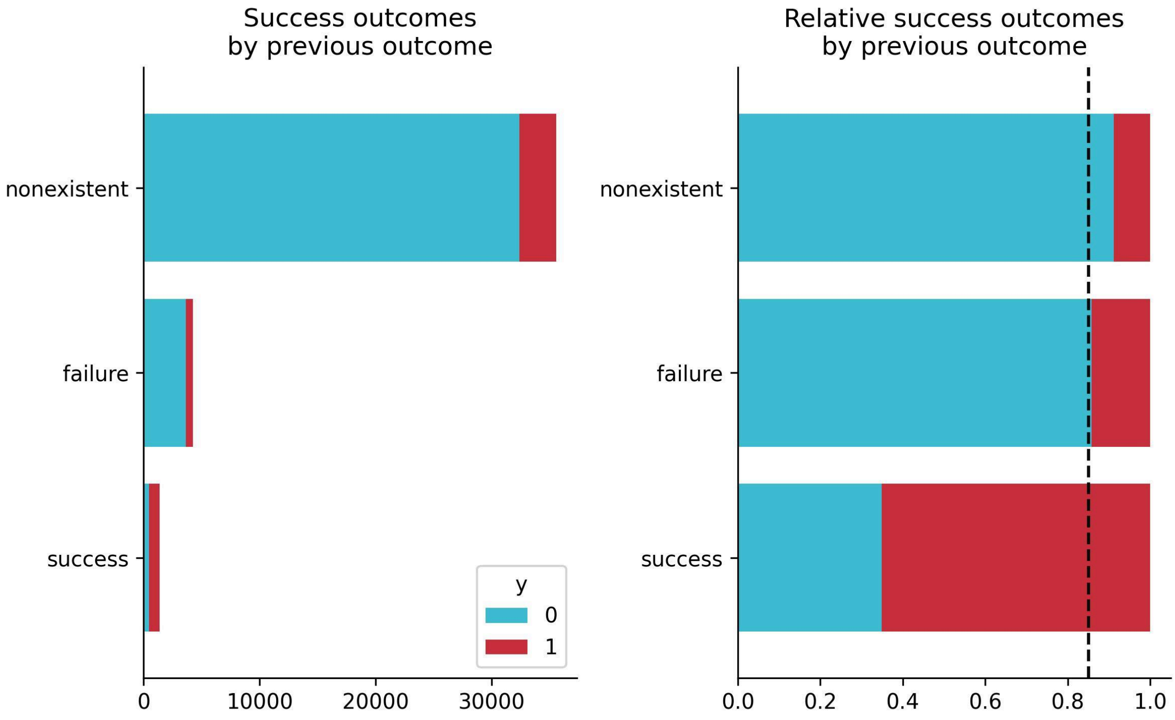


Figure 5. Distribution of success by previous contact outcome


Success outcomes
by previous outcome
nonexistent
failure
success
y
Relative success outcomes
by previous outcome
nonexistent
failure
success