Lead Conversion Analysis and Prediction

FOR TERM DEPOSIT ACCOUNT OPENINGS

Femi Onafalujo | Capstone 3: Springboard Data Science | September 27, 2021

Executive Summary

Selling value to customers has always been a tough proposition as a company needs to convince customers of the value provided by a product at the proposed price. This project focused on how data analysis and machine learning could be used to inform the process. Specifically, this project analyzed data provided by a Portuguese bank retrieved from the UCI machine learning repository that included customer information and the result to a request to open a term-deposit account. A term-deposit account is a fixed-term account, where the bank makes profit by earning interest on the deposited cash in excess of interest paid to the client during the term. Customer information included such features as the age, job and marital status. Furthermore, such economic information as the prevailing interest rates and unemployment rates were provided. This information was analyzed with respects to their relationship with a positive response. The duration of a phone call stood out to be a dominant factor during the analysis, whereby longer phone calls yielded higher positive response; however, this the duration of a phone call could not be used to develop a model.

Three algorithms were optimized to determine the best model. They included the logistic regression, random forest and XGBoost algorithms. The XGBoost algorithm provided the highest area under the precision-recall curve score of 0.48, and this model was chosen as the go-forward model. Nonetheless, the model's performance on unseen data was poor to question the quality of the model. Perhaps, more information would be useful in creating a better model

Contents

Introduction

Upselling is a widely adopted strategy that companies use to extract more profit from their customers. Tapping into an existing customer base for more business is better than going into the market and convincing new customers of the merits of your company and products. However, while seeking to extract more value from your existing clients, companies know that not all clients have the same needs. It therefore makes sense for companies to target clients that are more likely to respond to a marketing request, as not to waste company resources.

This was the case of a Portuguese banking institution that wanted existing clients to sign-up for a term-deposit account. A term-deposit account is a bank account that holds the money of clients for a fixed term (without withdrawal) in return for interest payments according to contract terms and conditions. The bank makes a profit if it can earn a return in excess of the interest payments. The fixed term ensures that the money is available to the bank within a certain period.

The Portuguese company initiated a direct marketing campaign (phone calls) to drive their sales efforts by offering term-deposits to their clients, at times, more than once. The result of their marketing campaign was made available as a csv file.

PROJECT OBJECTIVE

The project goal is to explore and analyze the factors that influence a positive response to a solicitation request in an effort to develop a model that predicts the likely response.

DATASET

The dataset required for analysis was available as a csv file, and consisted of 45,211 instances and 21 features, the last feature (y) being the response to the campaign (a 'yes' or a 'no').

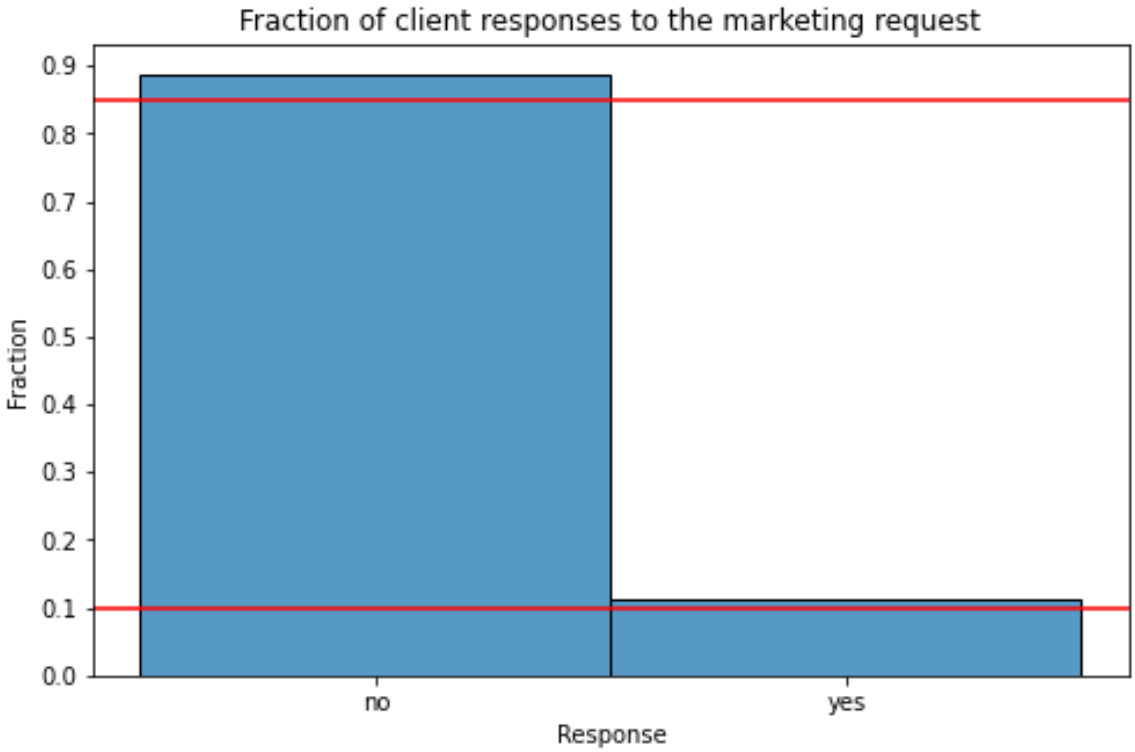
Source : http://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Data Wrangling

Data wrangling was straightforward: The dataset was imported to a notebook as a csv file, columns were renamed to more appropriate formats, and certain columns required regex commands to correct some of their values.

Exploratory Data Analysis

The focus of exploratory data analysis was to review the relationships between features and the response to the solicitation request (being either a 'yes' or 'no'). It was clear from the onset that the distribution of the response was imbalanced: Over 88% of the responses from the solicitation request was 'no', and just over 11% of the responses was 'yes'. This is illustrated in **figure 1** below. This result is not abnormal in the world of marketing.



The positive response rate of 11% therefore set the expectation for the response rates anticipated from the features in the dataset.

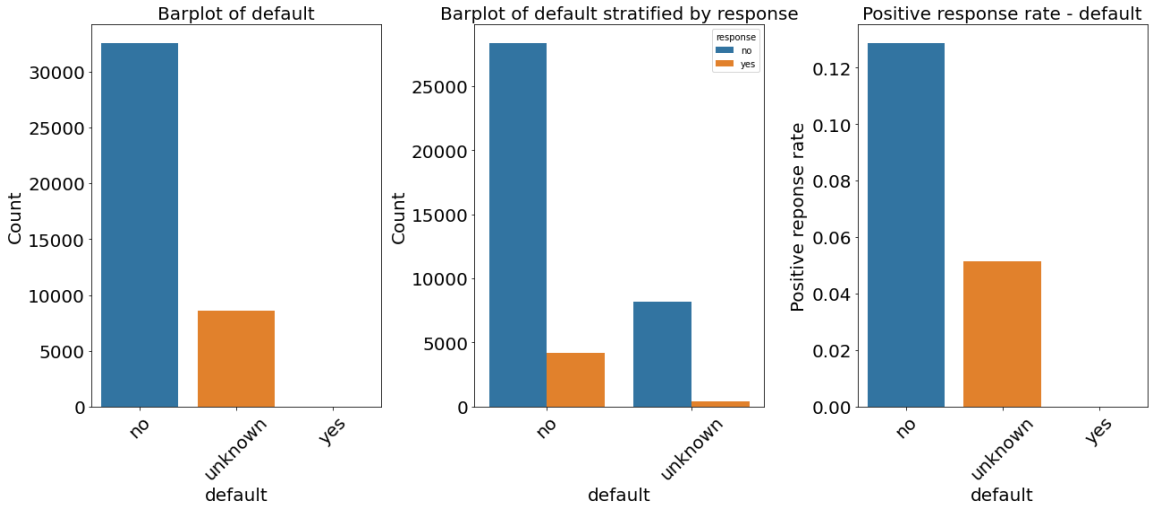
The response rate for each client category was therefore calculated and compared to the other categories within its group. For instance, in the 'job' group, the response rates of categories within the group (such as 'admin', 'blue-collar', and 'technician') were computed and analyzed.

This approach of computing the response rates for categories necessitated the conversion of quantitative values to categories. For each numerical feature, appropriate ranges were determined to allow for the computation of their associated aggregated response counts that was thereafter used to calculate the response rates. For instance, the boundaries set for the 'age' feature were - 10, 20, 30, 40, 50, 60, 70, 80, 90, 100. Categories were considered optimal if they were well represented in the dataset and had a relatively high response rate

when compared to other categories in its group. Notable observations are discussed in the following section.

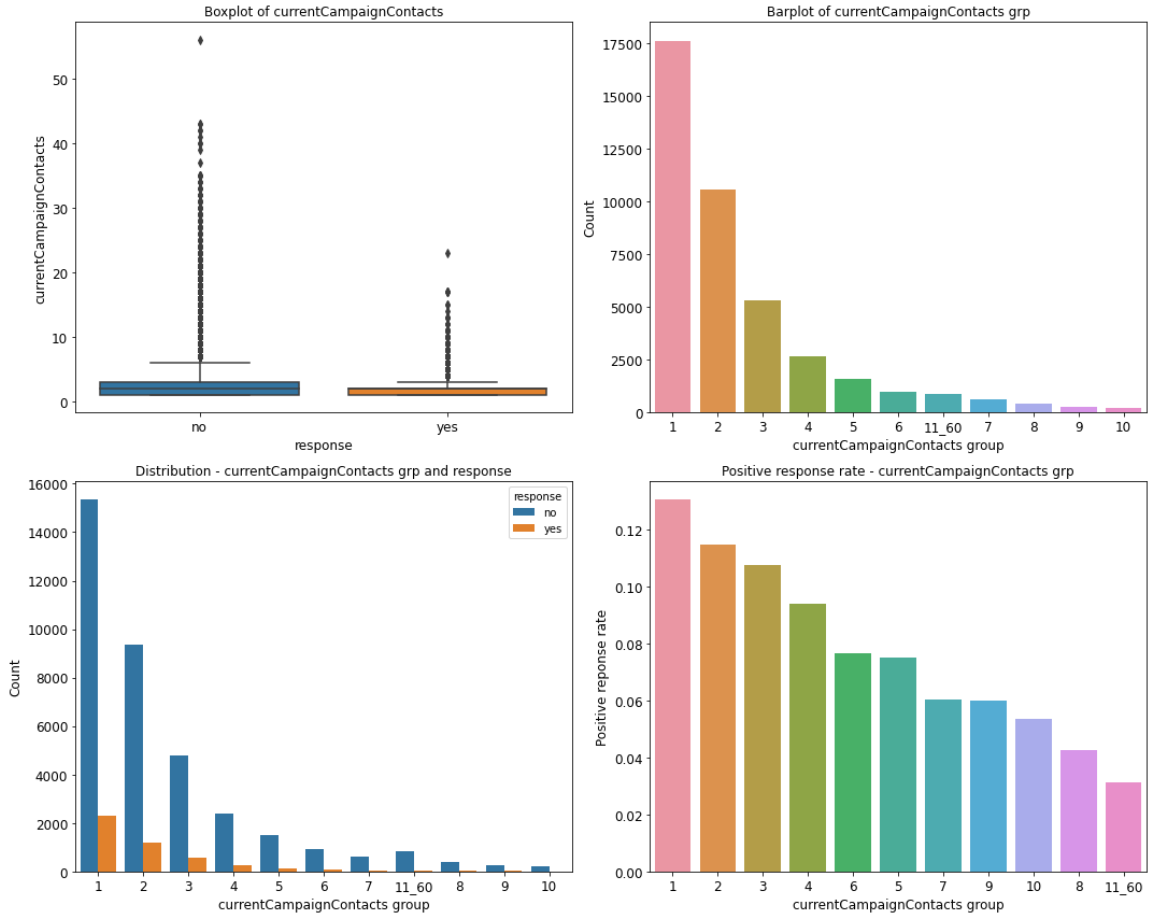
DEFAULT

Clients that had not defaulted on their loans were more than twice as likely to favorably respond to a solicitation request than other categories in its group (i.e., 'default), as shown in figure 2 below.



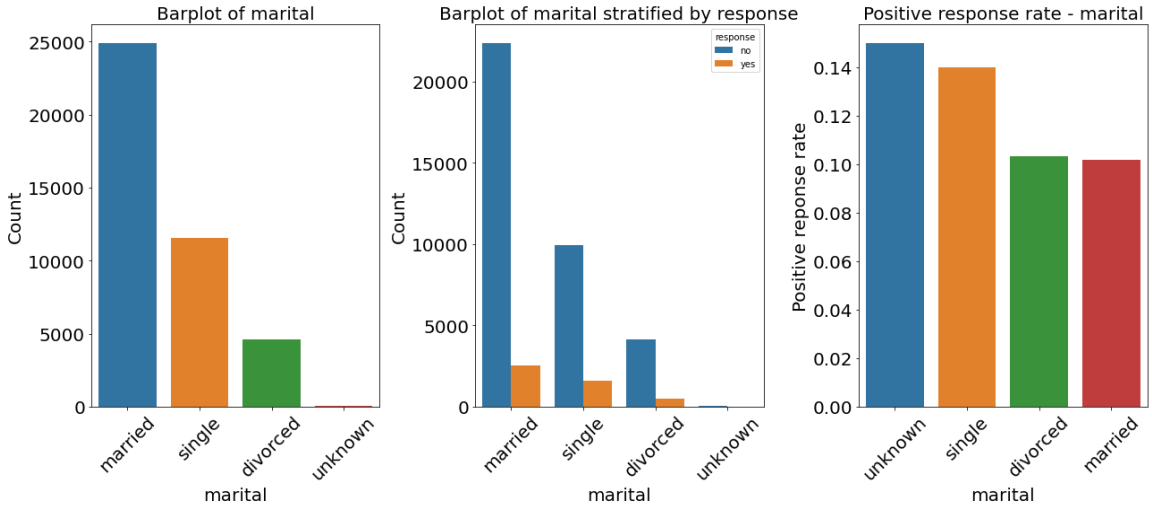
CURRENT CAMPAIGN CONTACTS

Similarly, clients that were contacted at most once during the current campaign were more likely to positively respond to a solicitation request than those who were contacted more times, as shown in figure 3 below.



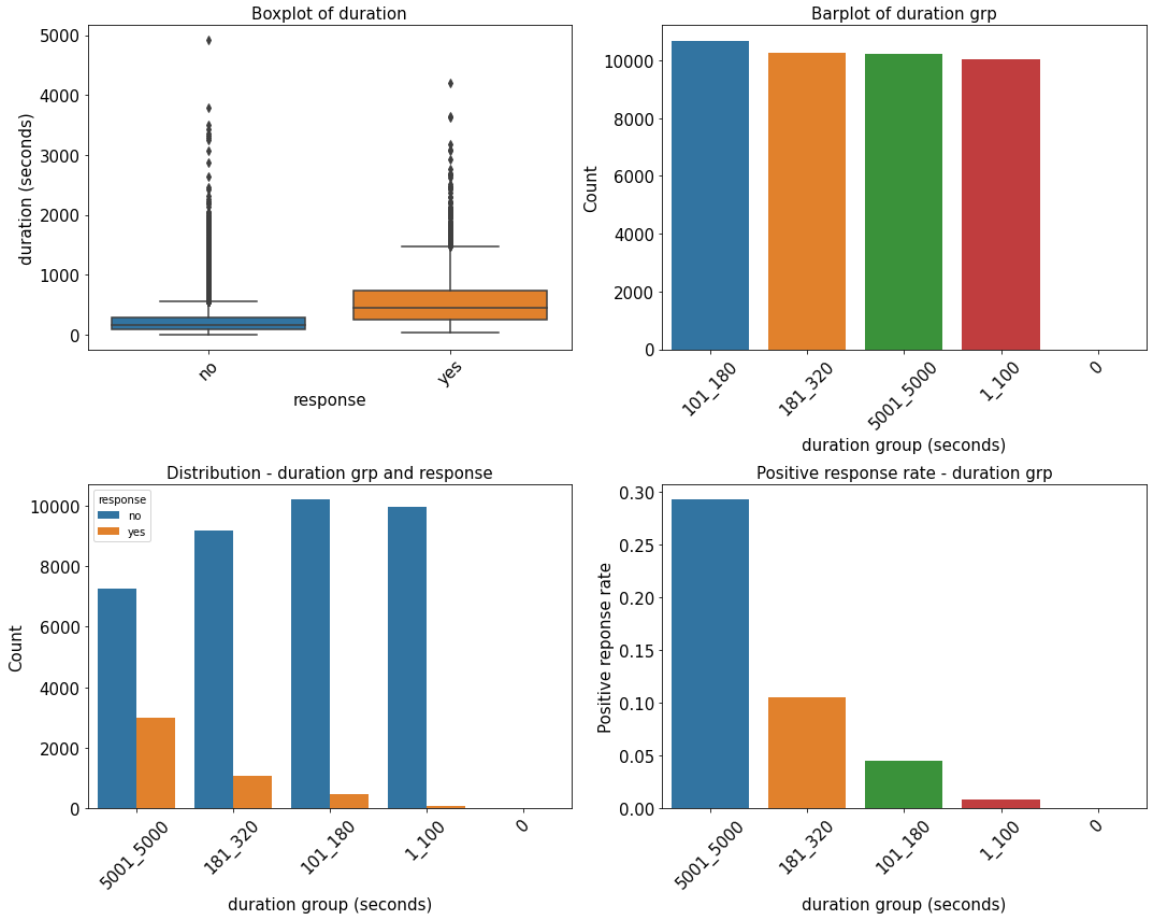
MARITAL STATUS

Another category that was well represented in the dataset and had a comparatively high response rate was being single. Single clients were more likely to favorably respond to a solicitation request than divorced or married clients, as shown in fig 3.



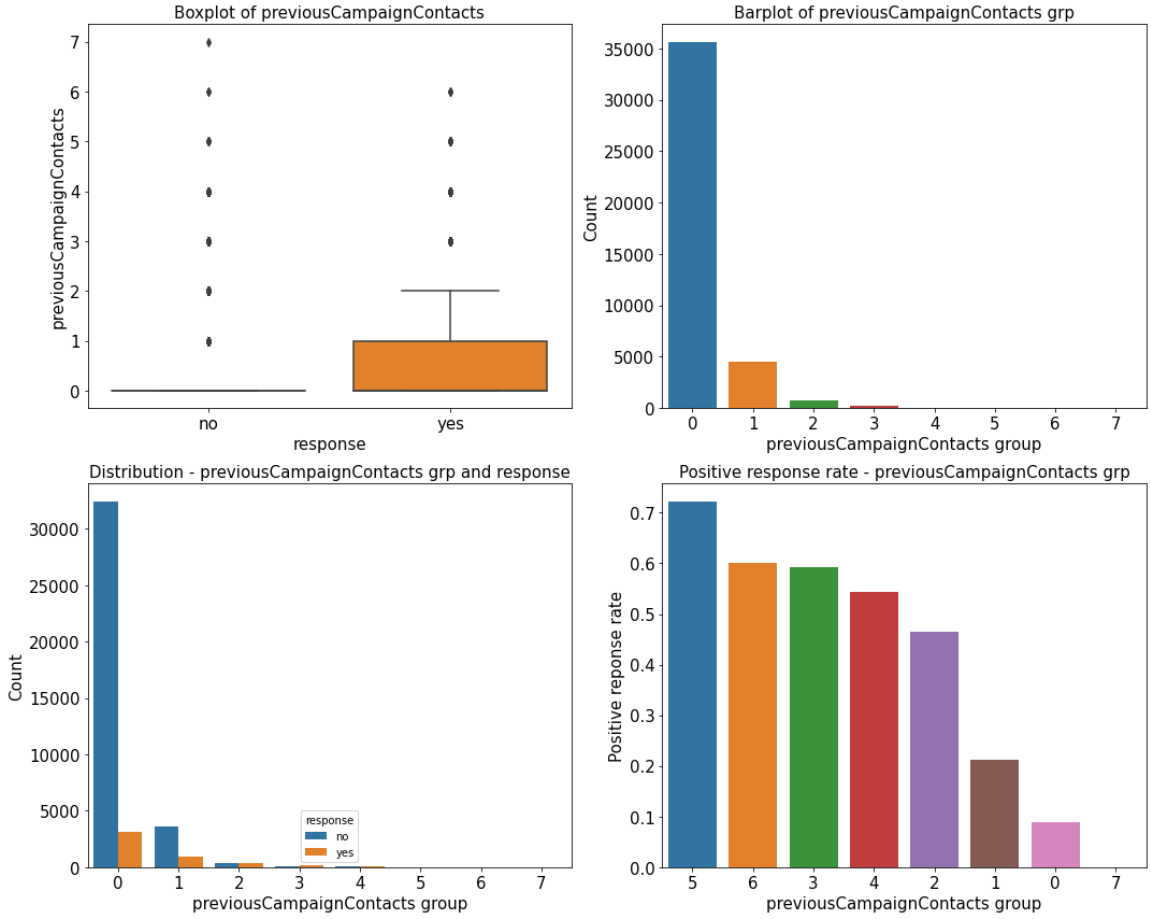
DURATION

There are typically more favorable phone calls that last between 253 seconds (~4 min. 30s) and 741 seconds (~12min. 30s) than any other period, as shown in the boxplot of figure. This period is perhaps long enough for a client to be interested in the offering and understand its details, while being short enough for the client to decide. In general, phone calls that last over 320 seconds (~5 min. 30s) are more likely to lead to a positive response than shorter calls, as indicated in the positive response rate chart of figure.



PREVIOUS CAMPAIGN CONTACTS

There are typically more clients who favorably respond to the solicitation request when they hear it for the first time than clients who hear the request multiple times. This is illustrated in the boxplot and bar plot of figure. A client that does not accept the offer when it is first presented is likely not interested in the offer.



Data pre-processing and Training Data Development

DATA PRE-PROCESSING

Pre-processing was uncomplicated. Fortunately, the dataset had no missing values. However, the 'duration' feature had to be dropped because the duration of a phone call is typically not known before the call is made. Including this feature for modelling would lead to data leakage. The numerical features were scaled and the categorical features dummy encoded.

TRAINING DATA DEVELOPMENT

Application set

Twenty random samples from the pre-processed dataset were assembled as a dataset called the application-set. The intent of the application set was to visually inspect the results of the eventual model because the application-set mirrored the type of information a marketing manager would have. The eventual model was then used to determine the

clients that would accept an offer and compared to the actual response values. This will be discussed in an upcoming section.

Training and test set

The data was divided into a training and test set at a 70% / 30% ratio. This division also factored the imbalanced nature of the response. The training set was used to create a model, while the test set was used to assess the model's performance. Each set was further divided into a dataframe of features and an array of the target feature (response), such that there were four datasets: X_train and y_train representing the features and response for the training set; and X_test and y_test representing the features and response of the test set.

METRICS AND MODEL ASSESSMENT STRATEGY

Metrics

It was important that our model prioritized the prediction of positive responses, indicating that a client opened a term-deposit account. This meant that the recall score or the f2 score, rather than the precision score or an f1 score, was likely the more important metric because it assessed the selection of true positives. Nonetheless, the metric also had to be threshold invariant. This elevated the area under the precision recall curve to the preferred metric (AUC-PR).

Model assessment strategy

A 5-fold cross-validation approach was used to simultaneously train and test folds of the training set to yield 5 cross-validation scores. A mean and standard deviation was computed from the array of scores to provide a sense for the typical performance of the model and the bounds of this performance. The average precision was computed for each model undergoing the cross-validation evaluation. The r2 scores were used to review the different modalities of a particular algorithm, while the r2 and mae scores were used for final model selection. The choice of five folds was a judgement call based on the size of the dataset, computational capabilities of the processor and the belief that the scores from five folds of the training dataset were sufficient to reveal overfitting concerns.

The final assessment of a model was conducted with the test set, where the final metrics provided an indication of the model's performance on unseen data. Other metrics such as the area under the receiver operating characteristics curve (AUC-ROC), precision, recall, and f2 score were also computed. Consideration was given to all metrics for the selection of a final model.

BASELINE MODEL CREATION AND ASSESSMENT

A logistic regression algorithm was used to create a baseline model after the pre-processing and training data development steps. The model was trained on the training

set and assessed with the test set. No cross-validation scores were computed for the model. The overall model performance on the training set and test set is shown in table 1.

Model	Class	Recall	F1	F2	AUC-PR	AUC-ROC
Logistic regression	0	0.99	0.95		0.48	0.81
	1	0.25	0.37	0.29		

The wide gap between the baseline AUC-PR score of 0.48 and the AUC-ROC score of 0.81 illustrates the imbalanced nature of the dataset. The AUC-PR score emphasizes the assessment the positive class (a 'yes' to the request), while the AUC-ROC score equally assesses the classification of the positive and negative class (a 'no' to the request). The overrepresentation of the negative class gives a more favorable AUC-ROC score.

EXTENDED MODELLING PLAN

With a baseline AUC-PR score of 0.48, the following activities were performed to develop a better model.

- Feature selection: From exploring the data, 'cci' and 'currentCampaignContacts' were shown to have a lower correlation coefficient with the response than other features. There was therefore an opportunity to drop features to improve models.
- Algorithms: A variety of algorithms were trialed to improve performance.
- Hyper-parameter tuning: Algorithm parameters were tuned to improve performance.
- Resampling methods: Various sampling methods were adopted to appropriately represent the under-represented class in the hopes that doing so would improve performance.

Model Optimization, Selection and Application

MODEL OPTIMIZATION

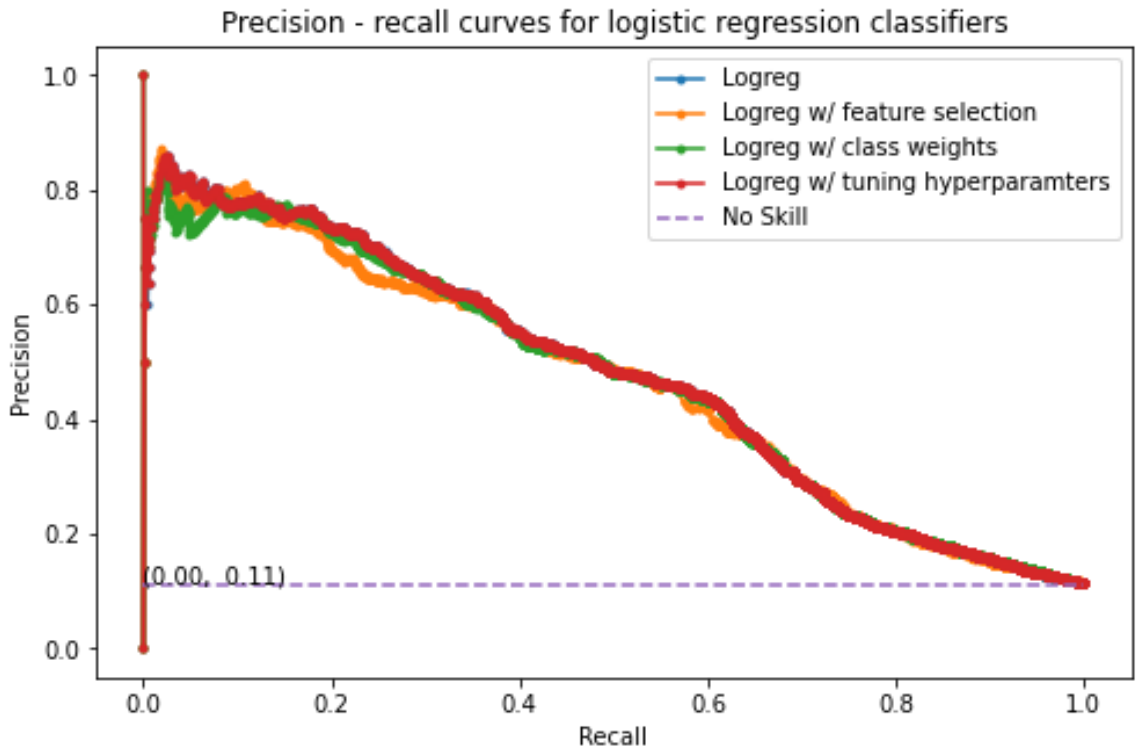
The logistic regression, random forest and XGBoost algorithms were used to create various models: Each algorithm was subject to feature selection, class weighing and hyperparameter tuning to determine if a better performing model could be derived. Thereafter, the best performing models from each algorithm were trained on a more balanced dataset to determine if resolving the data imbalance concern of the dataset improved performance.

Logistic regression

The logistic regression model yielded a baseline AUC-PR test score of 0.48. Feature selection, class weighing and hyperparameter tuning did not improve the result as shown in table below.

Logistic regression model - logreg	AUC-PR CV scores	AUC-PR test scores
Logreg	0.45	0.48
Logreg w/ class weights	0.45	0.48
Logreg w/ hyperparameter tuning	0.45	0.47
Logreg w/ feature selection	0.41	0.47

The precision-recall plots for the logistic regression models are shown in figure below.

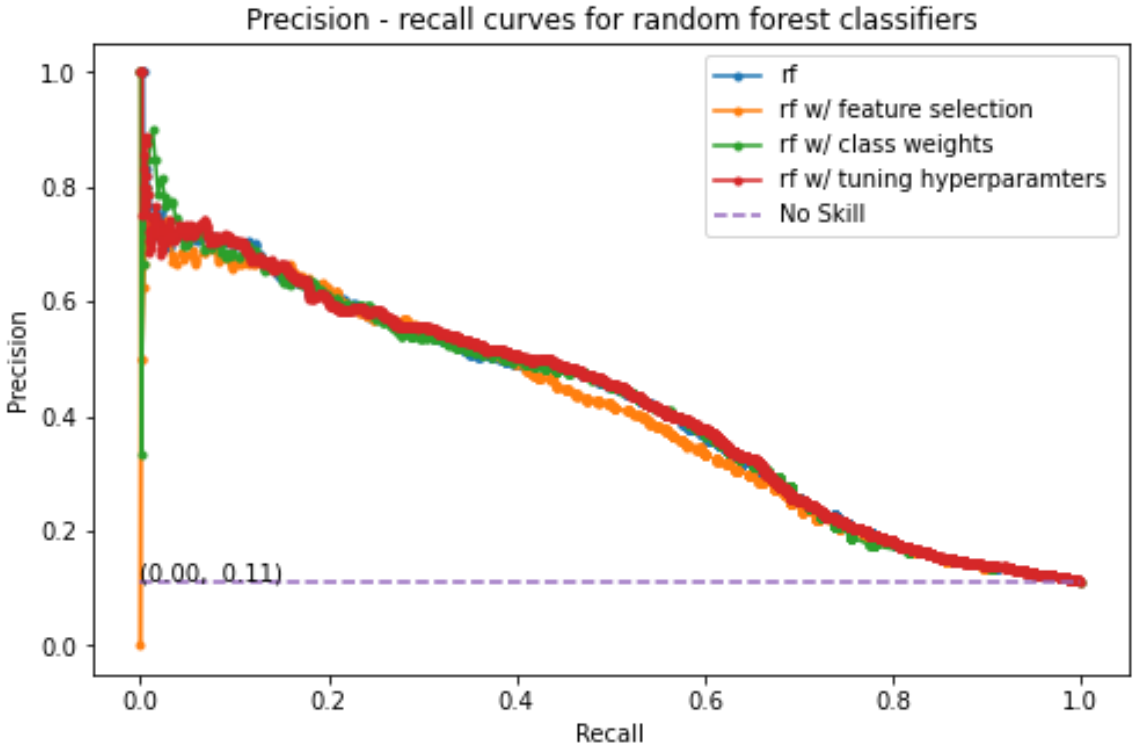


Random forest

The random forest model performed generally worse than the logistic regression model. Hyperparameter tuning only slightly improved the random forest score from a base model AUC-PR test score of 0.41 to 0.42 as shown in table.

Random forest model -rf	AUC-PR CV scores	AUC-PR test scores
rf w/ hyperparameter tuning	0.41	0.42
rf	0.40	0.41
rf w/ class weights	0.40	0.41
rf w/ feature selection	0.39	0.41

The precision-recall plots for the random forest models are shown in figure below.

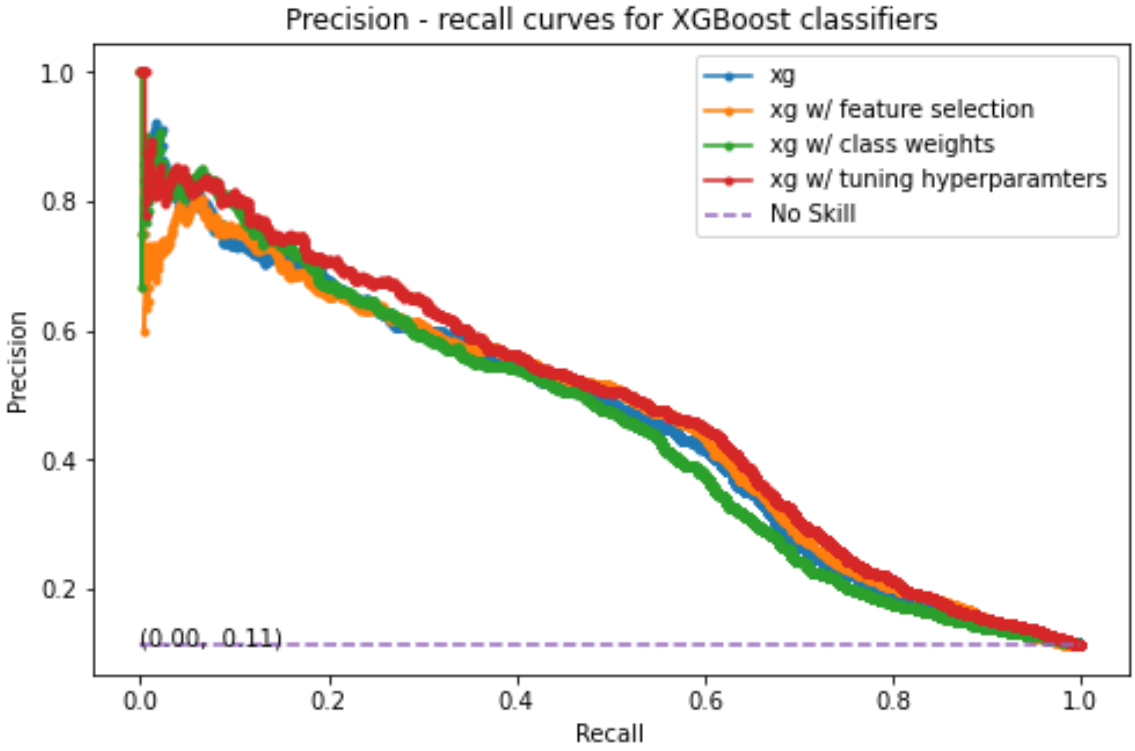


XGBoost

Hyperparameter proved useful in boosting the XGBoost classifier's performance: The AUC-PR test score increased from 0.46 to 0.48 from the base XGBoost model, as shown in table.

XGBoost model - xg	AUC-PR CV scores	AUC-PR test scores
xg w/ hyperparameter tuning	0.46	0.48
xg w/ feature selection	0.44	0.46
xg	0.43	0.46
xg w/ class weights	0.42	0.45

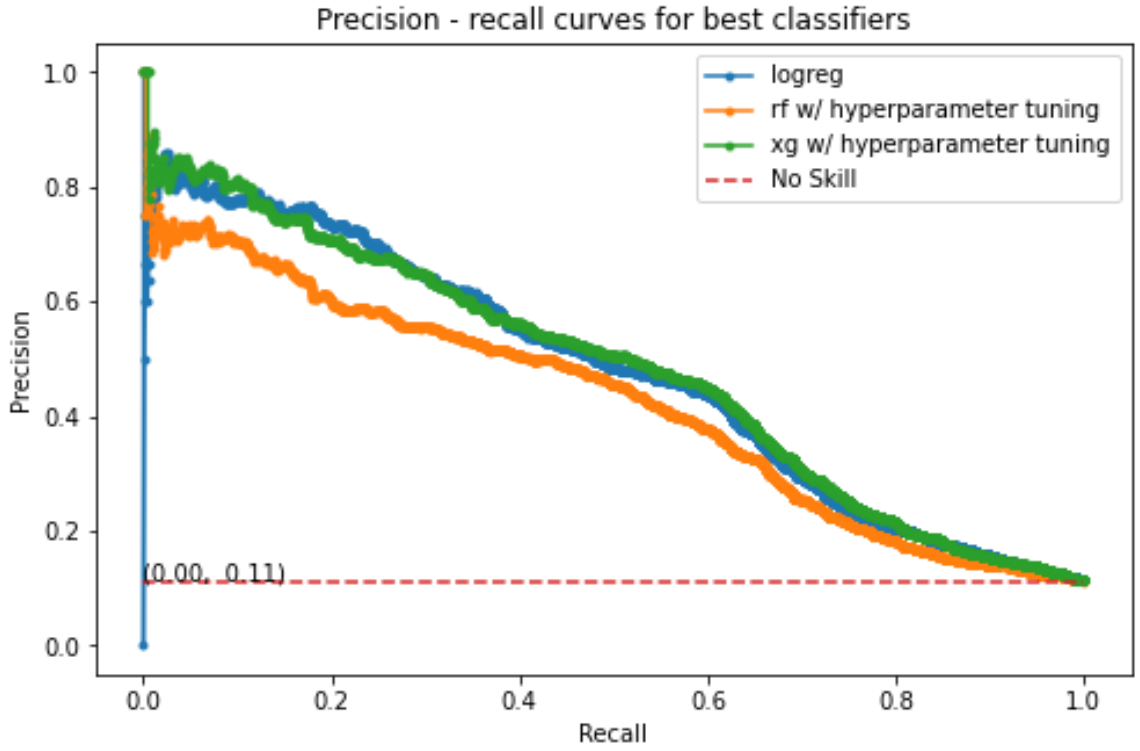
The precision-recall plots for the XGBoost models are shown in figure below.



Comparing the best models from each algorithm

The XGBoost model performed slightly better than the logistic regression model as shown in table and figure. Tuning was effective on the tree-based models versus the logistic regression model.

Best models	AUC-PR CV scores	AUC-PR test scores
xg w/ hyperparameter tuning	0.46	0.48
Logreg	0.45	0.48
rf w/ hyperparameter tuning	0.41	0.42



Training models on balanced data

The best models from each algorithm was trained on data balanced by different techniques, including random oversampling, random undersampling, Synthetic Minority Oversampling Technique (SMOTE) and SMOTE with undersampling. The best results came from SMOTE with undersampling, whose results are shown in the table below. Oversampling the data necessarily leads to overfitting of the dataset as suggested in the large differences between the cross validation AUC-PR scores on the training set and AUC-PR scores on the test set. Nonetheless, choosing a more balanced dataset did not improve the performance of the models.

SMOTE undersampling	AUC-PR CV scores	AUC-PR test scores
xg w/ hyperparameter tuning	0.93	0.48
Logreg	0.72	0.47
rf w/ hyperparameter tuning	0.94	0.41

MODEL SELECTION

The performance of the best models from each algorithm is shown in table.

Model	Class	Recall	F1	F2	5-fold CV : AUC-PR	AUC-PR	AUC-ROC
XGBoost w/ tuning	0	0.98	0.95		0.46	0.48	0.81
	1	0.27	0.39	0.31			
Logistic regression	0	0.99	0.95		0.45	0.48	0.81
	1	0.25	0.37	0.29			
Random forest w/ tuning	0	0.97	0.94		0.41	0.42	0.78
	1	0.31	0.40	0.34			

	Best score
	Tied for best score

The XGBoost model with hyperparameter tuning performed better than the other models in terms of the cross validation scores for the area under the precision recall curve on the training set. However, the random forest with hyperparameter tuning performed better than the other models in terms of the recall, F1 and F2 scores for the test set, while the XGBoost and logistic regression models were tied in terms of area under the precision-recall for the test set and the area under the receiver operator characteristic curve for the test set.

The XGBoost model was selected as the preferred model because its better cross validation performance implied that the model would perform better on unseen data than the other models.

MODEL APPLICATION

The preferred model (XGBoost with bespoke parameters) was trained on the entire dataset, save the application set. The 5-fold cross-validation scores on this model was observed to be substantially worse than cross validation scores observed over the train set: the mean area under the precision-recall curve was 0.09 ± 0.06, which is actually worse than the no-skill estimator of 0.11. This raised the first flag with regards to the model.

The model was used to predict the responses of the 20 clients in the application set and obtained the following results. None of the clients were predicted to give a positive response even though the four of the original 20 clients accepted the offer. We show the first 10 client results

	1	2	3	4	5	6	7	8	9	10
Age	39	29	50	40	34	29	28	30	54	43
Response	No	No	No	No	No	No	Yes	No	Yes	No
Predicted	No	No	No	Yes	No	No	No	No	No	Yes

Conclusion and Recommendations

Judging human behavior is hard. This project set out with the ambitious task of predicting the likelihood of a lead accepting an offer to open a term-deposit account with available data. The factors influencing a positive response were reviewed and qualified. Even though there were not clear factors, we observed that the duration had a relationship with the response, with phone calls lasting over 5 minutes favoring a higher likelihood of a response. Nonetheless, this feature could not be modeled because the duration of a phone call cannot be known before the phone call. Machine learning shed more light on the data. The XGBoost classifier performed the best and balancing the dataset did not yield improved results. In predicting actual client information with our model, the model's performance was subpar, suggesting that additional data was required.

Assumptions, Limitations and Opportunities

For this project, we want to err on the side of predicting positive cases, meaning that we will favor a probability threshold for selecting positive cases. Accordingly, it is very likely that our precision suffers, and we get a higher false positive rate.

There are costs associated with making wrong predictions. This means that we will be contacting clients that are unlikely to favorably respond to the marketing campaign. There should therefore be a cost / benefit analysis done to determine the trade-off between precision and recall when choosing a model.