Assessment 3 Report on Marketing Campaign Success Prediction

Author: Chengyang Zhou (24873630)
36103 Statistical Thinking for Data Science
Date: 04 Nov 2023

Table of Contents

Introduction

This research examines the application of statistical learning models to predict the success of marketing campaigns targeted at various customer segments. By analyzing a dataset with various customer attributes, the research aims to identify responsive customer segments and derive strategies to enhance the efficacy of future marketing campaigns. This research presents a comparative analysis of two models including a parametric model of Logistic Regression and a non-parametric model of Random Forest Classifier in order to forecast campaign outcomes and derive actionable insights.

Problem Formulation

Analysis Context

The telecommunications sector is a competitive industry with rapid changes in consumer preferences and technological advancements. In this dynamic market, the efficiency of marketing campaigns directly correlates with customer acquisition and retention rates. Our analysis concentrates on understanding the factors that contribute to the success of marketing initiatives and predicting customer responsiveness.

Specific Problem

The core problem is to predict which customers are likely to respond positively to a marketing campaign. This entails a binary classification task where the target variable is the success of the marketing campaign which is denoted as 'yes' for success and 'no' for nonsuccess.

Pertinent Questions and Hypotheses

The study seeks to address the following questions:

1. What customer attributes contribute significantly to the prediction of marketing campaign success?
2. Can we predict the outcome of a marketing campaign with reasonable accuracy using these attributes?
3. What insights can be derived from the model to improve the efficiency of marketing strategies?

Justification for Model Selection

Logistic Regression was chosen as a parametric model due to its well established reputation for providing clear interpretability in binary classification problems. This model is particularly advantageous when the goal is to not only predict outcomes but also to understand the effect of each independent variable on the likelihood of a positive campaign outcome. The logistic model's coefficients offer a direct estimation of the log odds of the response, and the model training uses Bayesian Estimation as a robust approach for parameter estimation. This method allows for the incorporation of prior distributions on parameters, accounting for existing knowledge or beliefs about the parameters' possible values. Training the model on scaled features ensures that the coefficient estimates are not unduly influenced by the scale of different variables, allowing for a fair comparison of feature importance across the model. (F et al., 2011, #)

The Random Forest Classifier represents a non-parametric approach, chosen for its proficiency in modeling complex and non-linear relationships without the need for a predefined form of the model. This capacity makes it particularly adept at capturing interactions between variables, which can be pivotal in understanding the multi-faceted nature of customer responses to marketing stimuli. Initially, the model was applied with default hyperparameters to establish a baseline understanding of feature relevances. Subsequent hyperparameter tuning was informed by a combination of cross-validation performance and Bayesian optimization techniques, which iteratively refine the search for optimal parameters by treating the hyperparameter space probabilistically. This strategy is well-suited to Random Forest, as it accommodates the high-dimensional and potentially complex interaction space within the model's structure. (F et al., 2011, #)

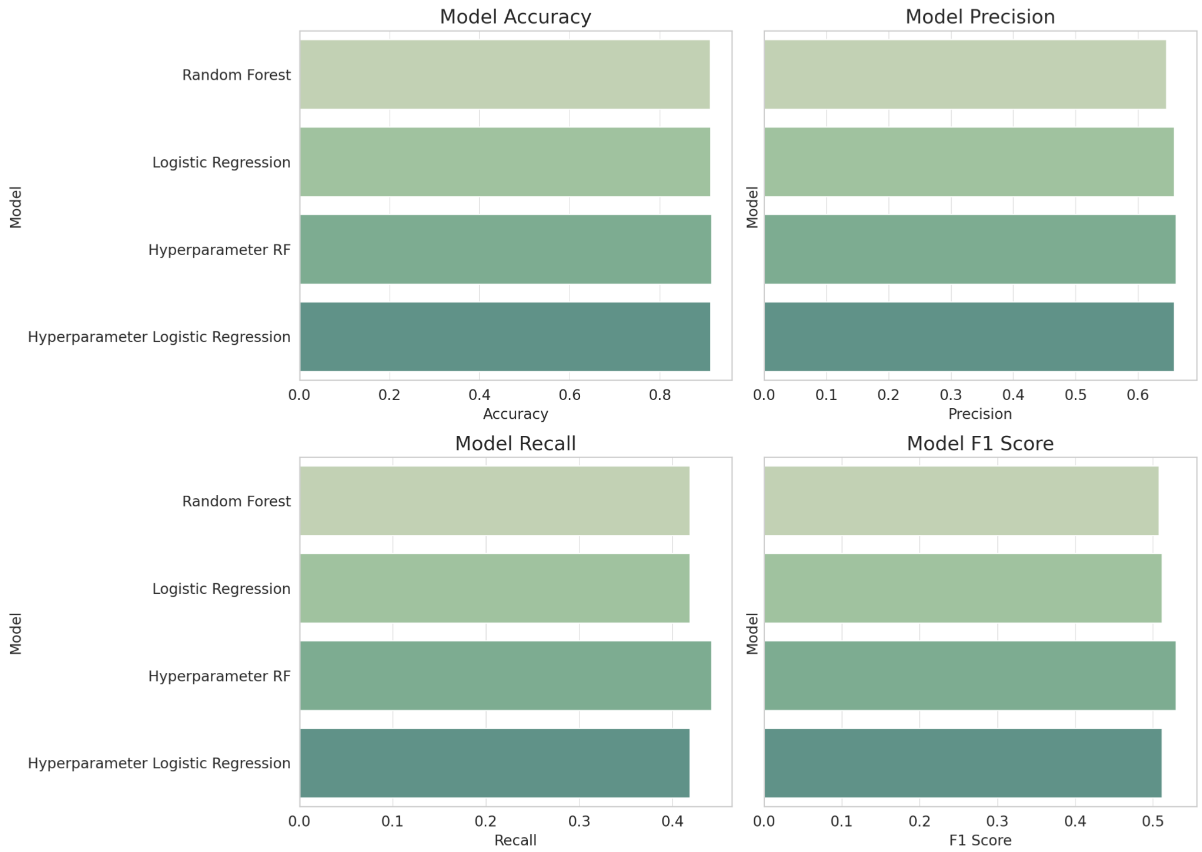
Both models were trained using a train-test split approach that ensures validation through unseen data. The Random Forest model used an initial unscaled dataset to preserve the inherent distribution of the data, while Logistic Regression required scaled data due to its sensitivity to variable scales.

Comparative Analysis of the Models
The models were assessed using accuracy, precision, recall, and F1 score. Cross-validation was used for hyperparameter tuning to ensure that the model generalization was not biased by the choice of the train-test split.

Results:

	Random Forest	Logistic Regression	Hyperparameter tuned Random Forest	Hyperparameter tuned Logistic Regression
Accuracy	0.9112	0.9126	0.9142	0.9126
Precision	0.6455	0.6579	0.6611	0.6579
Recall	0.4184	0.4184	0.4417	0.4184
F1 score	0.5077	0.5115	0.5296	0.5115

The Random Forest model achieved an accuracy of 91.12%, with a precision of 64.55% and an F1 score of 50.77%. Post-tuning, it performed better in terms of accuracy, precision, recall, and F1 score indicating the model has achieved a better balance in correctly predicting positive outcomes while reducing overfitting and misclassifications.



The Logistic Regression model, following hyperparameter optimization has achieved a cross-validation score of 91.07% and mirrored the Random Forest's test accuracy with 91.26%. It maintained a precision of 65.79% which is slightly higher than the non-tuned Random Forest, and matched it in recall at 41.84%. The fine-tuned Logistic Regression model exhibited a precision equal to its Random Forest counterpart with both maintaining the same recall, suggesting their equal strength in identifying true positives while the logistic model benefits from a straightforward interpretation and ease of implementation.

Estimation Methods

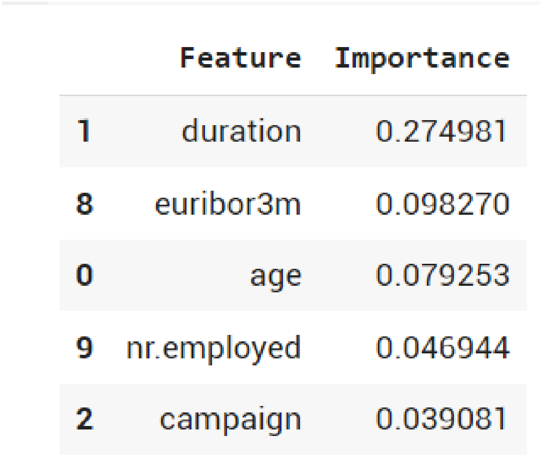
Logistic Regression typically uses Maximum Likelihood Estimation (MLE) to find the best fitting model to predict the probability of the outcome. However, Bayesian estimation methods can also be applied to logistic regression by using software packages that support Bayesian statistics. In this research, the logistic regression model was developed with standard MLE methods that are inherent to the logistic regression function in common statistical software. The coefficients obtained through MLE reflect the log odds of the campaign's success with given the customer features.

Parametric Estimates and Interpretations

Parametric models such as the logistic regression we used in this assessment provide coefficients for each input feature which represent the log odds effect on the outcome with all other variables held constant. The Logistic Regression model provided the following insights:

- ➔ The ‘C’ parameter value of 1 suggests a balance between complexity and regularization, avoiding overfitting while maintaining model accuracy.
- ➔ An ‘l2’ penalty indicates that feature weights were regularized to prevent overreliance on any single attribute.

Business Insights Extraction



	Feature	Importance
1	duration	0.274981
8	euribor3m	0.098270
0	age	0.079253
9	nr.employed	0.046944
2	campaign	0.039081

1. The most predictive attributes identified by the Random Forest can inform the marketing team on using customer characteristics such as ‘duration’ to focus their efforts.
2. The choice between models could guided by the marketing campaign's tolerance for false positives versus the need to minimize false negatives. The Hyperparameter-tuned Random Forest model with its higher precision is better suited for campaigns aiming to minimize false positives. In contrast, if the campaign's goal is to reduce false negatives, both models' similar recall rates offer comparable reliability.
3. With over 91% accuracy for both models implies that the dataset contains strong signals that can forecast the success of a marketing campaign. This suggests that the current data collection methods are effective and should be maintained or even expanded to refine future predictions.

Conclusions

Our study confirms that statistical models are valuable tools for predicting marketing campaign success with Logistic Regression and Random Forest providing a comprehensive view of customer behavior. However, these models do have their constraints. Logistic Regression may overlook complex patterns, and Random Forest models can be computationally heavy. Future improvements could include experimenting with advanced modeling techniques, enriching the dataset diversity and refining model training to enhance predictive performance could further bolster campaign performance.

References

F, P., G, V., & A, G. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825--2830. Retrieved November 5, 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

F, P., G, V., & A, G. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825--2830. Retrieved November 5, 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Data Source:https://canvas.uts.edu.au/courses/28112/files/5430422?wrap=1

Data 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?

Code Appendices

File name:
STAT-AT3.ipynb