# MapleTrust Bank: Marketing Campaign Optimization Using Machine Learning

Grace Abunyie
*Business Insights and Analytics*
*Humber Institute of Technology and Advanced Learning*
Toronto, ON, Canada
n01545785@humber.ca

Riya Johnson
*Business Insights and Analytics*
*Humber Institute of Technology and Advanced Learning*
Toronto, ON, Canada
n01530216@humber.ca

Shaquille Laing
*Business Insights and Analytics*
*Humber Institute of Technology and Advanced Learning*
Toronto, ON, Canada
n01533161@humber.ca

Soumil Kapri
*Business Insights and Analytics*
*Humber Institute of Technology and Advanced Learning*
Toronto, ON, Canada
n01542107@humber.ca

Tanishka Patil
*Business Insights and Analytics*
*Humber Institute of Technology and Advanced Learning*
Toronto, ON, Canada
n01537564@humber.ca

***Abstract*—This paper explores the synergy between machine learning (ML) and marketing strategies, focusing on optimising direct marketing for term deposit enrolment at MapleTrust Bank. Leveraging insights from Garcia-Serrano et al. (2018), Silva et al. (2019), Demir et al. (2018), and Wickham (2016), the study employs algorithms like Logistic Regression, Decision Tree, K-Nearest Neighbours, and Random Forest.**

**The research details meticulous data cleaning and preprocessing, emphasising target variable renaming, feature selection, and one-hot encoding for categorical variables. Descriptive statistics and visualisations offer insights into the dataset. Machine learning models are evaluated using metrics such as accuracy, F1 score, precision, recall, and balanced accuracy. The study recommends considering multiple metrics and business implications for model selection.**

**The outcomes aspire to enrich MapleTrust Bank's marketing strategy by delivering actionable insights into customer behaviour. While Random Forest emerges as the top performer based on the F1 score, the paper underscores the need for a comprehensive consideration of metrics and business implications before the final model selection. This research underscores the symbiotic relationship between ML expertise and marketing acumen for navigating complexities in consumer behaviour.**

***Keywords*— customer behaviour, data preprocessing, direct marketing, f1 score, machine learning, marketing analytics, predictive modelling, random forest.**

## I. INTRODUCTION

The realm of Machine Learning (ML) has increasingly become essential in deciphering complex patterns and making informed decisions, particularly in the context of marketing campaigns. The success of these campaigns can now be predicated on the intelligent analysis and predictions made by robust ML algorithms, which hinge on the fundamental steps of data preprocessing and visualization. The abstract at hand offers a glimpse into the pivotal methodologies that contribute to the preparation of datasets, the intricacies involved in understanding their hidden structures, and the application of advanced ML algorithms to predict customer behaviour—an area of growing interest and significant market value.

In this paper, we delve into the critical procedures of data handling, explore the art of visual data analysis, and scrutinise a spectrum of ML algorithms adopted in the domain of marketing for predicting customer responses. We leverage the insights from Garcia-Serrano et al. (2018) [1] and Silva et al. (2019) [2], which underscore the importance of precise data preparation, and Demir et al. (2018) [3] and Wickham (2016) [4], who emphasize the need for rigorous exploratory data analysis through visualization techniques. The methods they propose buttress the datasets before subjecting them to ML algorithms, thus ensuring a reliable foundation for subsequent predictive models.

Building upon this groundwork, we analyze the application of various ML algorithms such as Logistic Regression, Linear Regression, Random Forest, and Naive Bayes. This provides a comprehensive examination of how preprocessing, visualization, and advanced predictive algorithms converge to illuminate the path toward understanding consumer responses, thereby empowering more strategic and successful marketing initiatives.

## II. LITERATURE REVIEW

Previous studies highlight the importance of performing thorough data analysis and preprocessing. Garcia-Serrano et al. (2018) [1] and Silva et al. (2019) [2] provided valuable insights into addressing missing data, handling categorical variables, and normalizing numerical features. We saw their strong emphasis on preprocessing, which ensures the dataset's suitability for Machine Learning algorithms, laying the foundation for accurate predictions.

Visualization techniques play a pivotal role in understanding the dataset's intricacies. The works of Demir et al. (2018) [3] and Wickham (2016) [4] highlight the significance of visual exploratory data analysis through techniques such as scatter plots, histograms, and correlation matrices. Effective visualization aids in uncovering patterns and relationships, guiding subsequent modelling steps. We used this concept to guide our approach in this analysis.

The dataset has been a focal point for various machine learning algorithms in predicting customer responses for marketing campaigns. Logistic Regression, recognized for binary classification tasks, has been a popular choice for predicting term deposit subscriptions, while Linear Regression has shown adaptability in predictive modelling.

Several Machine Learning algorithms are commonly employed in predicting customer responses for marketing campaigns. Random Forest, an ensemble learning algorithm, has been widely applied across diverse applications, including marketing analytics as demonstrated by Liaw and Wiener (2002) [5] and Cutler et al. (2007) [6]. Naive Bayes, known for its simplicity and efficiency, has also shown promise in predicting customer responses, as indicated by Rish (2001) [7].

These works provide valuable benchmarks for the current project, showcasing different approaches to tackle the predictive challenge presented by the dataset.

## III. PROJECT OBJECTIVES

The overarching goal of this project is to employ machine learning methodologies for the optimization of direct marketing strategies employed by MapleTrust Bank, a leading Canadian financial institution. The specific focus lies in telephone-based marketing efforts aimed at promoting term deposit enrolment among clients. The primary objective is to construct a robust predictive model capable of accurately determining whether a client is likely to subscribe to a term deposit (categorised as "Yes" or "No"). The successful development and implementation of such a model are anticipated to empower MapleTrust bank with a refined understanding of client behaviour, enabling a more targeted and effective approach in its direct marketing endeavours. This initiative aims to contribute to the bank's overall marketing strategy, enhancing customer engagement and bolstering the success rate of term deposit enrolments through data-driven decision-making. The project's outcomes are expected to provide actionable insights, facilitating a more efficient allocation of resources, and yielding improvements in marketing effectiveness for the institution.

## IV. ABOUT THE DATASET

This dataset originally pertained to the outcomes of telephone-based direct marketing campaigns conducted by a Portuguese banking institution [8].

In the context of this report, the dataset will represent telephone-based direct marketing interactions by MapleTrust Bank, a fictional Canadian Institution, aiming to predict whether clients will enrol in a term deposit denoted by the variable 'y'.

| | |
|---|---|
| Age | Age of the client. |
| Job | Occupation of the client. |
| Marital | Marital status of the client. |
| Education | Education level of the client. |
| Default | Whether the client has credit in default (yes/no). |
| Balance | Client's account balance. |
| Housing | Whether the client has a housing loan (yes/no). |
| Loan | Whether the client has a personal loan (yes/no). |
| Duration | Duration of the last contact in seconds. |
| Campaign | Number of contacts performed during this campaign. |
| Poutcome | Outcome of the previous marketing campaign (unknown, failure, success) |
| Y (Target variable) | Whether the client subscribed to a term deposit (yes/no). |

Table 1: Description of variables

## V. DATA CLEANING AND PREPROCESSING

### *A. Data Cleaning*

The initial step in our data analysis process involves the meticulous cleaning of the dataset to ensure the reliability and accuracy of our subsequent analyses.

*1) Renaming Target Variable:* In our analysis, the target variable originally labelled as 'y' was renamed to 'deposit' to better reflect its significance. This variable serves as a determinant, indicating whether a client will enrol in the term deposit.

*2) Removing Unwanted Features:* To ensure faster processing of our algorithms, we removed unwanted features that were not relevant to our predictive model such as the ‘Day’ column.

*3) Checking for Missing Values:* We checked for any potential missing values in our dataset using the 'isnull()' function. The analysis revealed that our dataset exhibits no instances of missing values, eliminating the need for any specific handling procedures during the subsequent processing steps.

### *B. Data Pre-processing*

The pre-processing phase involves transforming raw data into a structured and suitable format for analysis. This phase is essential for enhancing the performance of machine learning models and ensuring that they can generalise well to new, unseen data.

*1) Encoding Categorical Variables:* Categorical variables, such as 'job,' 'marital,' 'education,' and 'contact,' were encoded using one-hot encoding. This transformation ensures that the categorical variables are appropriately represented as numerical features, allowing the machine learning models to interpret and utilise them effectively.

*2) Feature Scaling:* Numerical features such as ‘balance’ and ‘duration’ were standardized to a common scale using StandardScaler() function. Standardisation ensures that all features contribute equally to the model training process and prevents the dominance of certain variables due to differences in scale.

## VI. DESCRIPTIVE STATISTICS & VISUALIZATION

To gain an initial understanding of the dataset, we employed the .describe() function in Python, which provides key summary statistics for each numerical column. The following table presents an overview of the central tendencies, dispersions, and other relevant metrics:

| **Statistics** | **Mean** | **Standard Deviation** | **Minimum** | **Maximum** | **Median** |
|---|---|---|---|---|---|
| **age** | 41.1701 | 10.5762 | 19 | 87 | 39.0 |
| **balance** | 1422.6578 | 3009.6381 | -3313 | 71188 | 444.0 |
| **duration** | 263.9613 | 263.9613 | 4 | 3025 | 185.0 |
| **campaign** | 2.7936 | 3.1098 | 1 | 50 | 2.0 |
| **pdays** | 39.7666 | 100.1211 | -1 | 871 | -1.0 |
| **previous** | 0.5426 | 1.6936 | 0 | 25 | 0.0 |

Table 2: Summary Statistics for all Quantitative Variables

These statistics offer insights into the central tendencies and variability within the dataset. For instance, the mean age is approximately 41.17 years, with a minimum age of 19 and a maximum of 87. The average account balance is 1422.7 units, with a standard deviation of 3009.6.

The following bar graph illustrates the distribution of people who enrolled in a term deposit versus those who did not. This visualization provides an initial glimpse into the balance between positive and negative instances within the target variable.

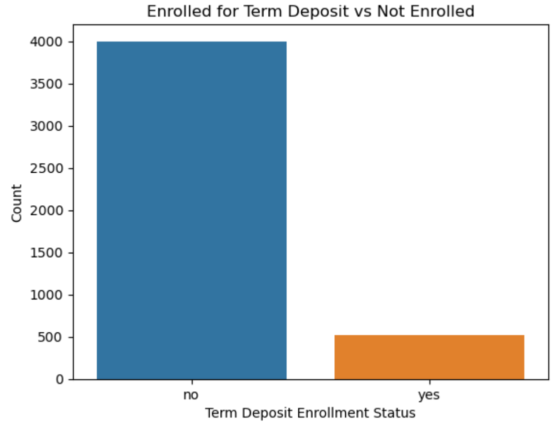


Fig 1. Enrolment status of original dataset.

To further understand the relationship between categorical variables and the target variable 'deposit,' we created bar graphs for variables such as loan status, job type, marital status, etc. Each bar graph showcases the distribution of deposits within different categories, shedding light on potential patterns or dependencies.

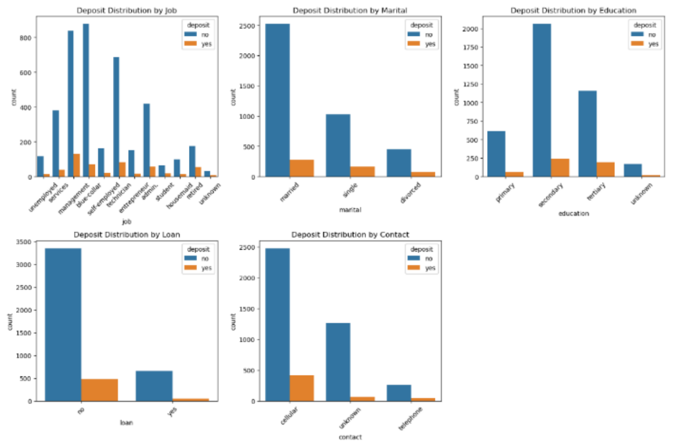


Fig. 2. Plotting the categorical variables with respect to target variable – deposit.

The last visualization illustrates the campaign outcome (term deposit enrolment - yes or no) for each month. This provides insights into seasonal patterns and campaign effectiveness over time.

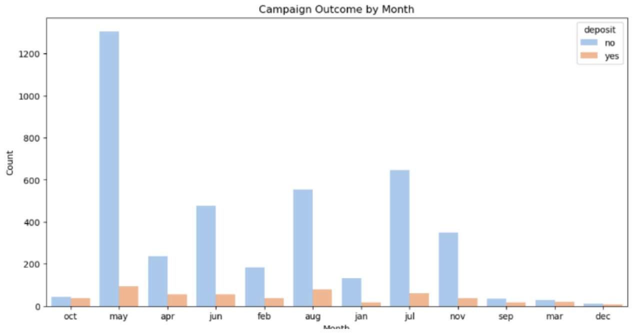


Fig. 3. Campaign Outcome by Month

## VII. MACHINE LEARNING MODELS

Here's a summary of the algorithms we used along with a brief explanation of each one:

### A. *Logistic Regression (LR)*

Logistic Regression is a linear model for binary classification that predicts the probability of an instance belonging to a particular class. It's based on the logistic function and is widely used for its simplicity and interpretability.

### B. *Decision Tree (DT)*

Decision Trees are non-linear models that make decisions based on a tree-like graph of decisions. They split the dataset into subsets based on the most significant attributes at each node.

### C. *K-Nearest Neighbors (KNN)*

K-Nearest Neighbors is a non-parametric algorithm that classifies a data point based on the majority class of its k-nearest neighbours. It's simple and intuitive.

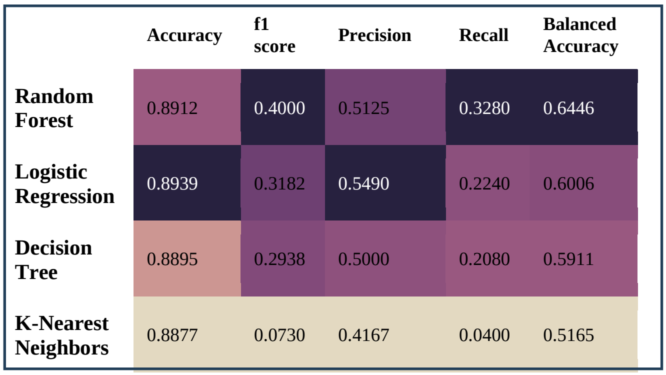
### D. *Random Forest (RF)*

Random Forest is an ensemble method that builds multiple decision trees during training and merges their predictions to improve accuracy and control overfitting.

For each algorithm, we trained the model using the train_evaluate_model function, which fits the model on the training set and evaluates its performance on the testing set. The evaluation metrics include accuracy, F1 score, precision, recall, and balanced accuracy.

## VIII. RESULTS OF ALGORITHM

The initial results of our algorithm were as follows.



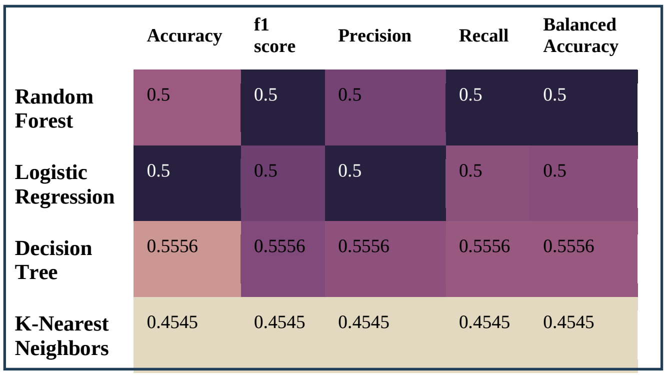
| | Accuracy | f1 score | Precision | Recall | Balanced Accuracy |
|---|---|---|---|---|---|
| **Random Forest** | 0.8912 | 0.4000 | 0.5125 | 0.3280 | 0.6446 |
| **Logistic Regression** | 0.8939 | 0.3182 | 0.5490 | 0.2240 | 0.6006 |
| **Decision Tree** | 0.8895 | 0.2938 | 0.5000 | 0.2080 | 0.5911 |
| **K-Nearest Neighbors** | 0.8877 | 0.0730 | 0.4167 | 0.0400 | 0.5165 |

Table 3: Model results for original data.

Here, based on accuracy alone we could say that Logistic Regression performed the best but as we saw in the visualisations, our target variable has very skewed data. There are 4000 records for '0' which indicates not enrolled and just 521 records for '1' which indicates enrollment in term deposits.

To further test our algorithms, we tested the performance of our machine learning models using a carefully crafted dummy dataset. The dataset was designed to address the imbalanced class distribution present in the original data, ensuring a more equitable representation of both positive and negative outcomes.

After doing the necessary preprocessing of renaming target variable, encoding categorical columns to numeric and standardising the dataset, we split the dataset into training & testing. As the dataset was more balanced this time, we achieved lower accuracy for all the models and hence used f-1 score as the metric to evaluate the model results.



| | Accuracy | f1 score | Precision | Recall | Balanced Accuracy |
|---|---|---|---|---|---|
| **Random Forest** | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 |
| **Logistic Regression** | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 |
| **Decision Tree** | 0.5556 | 0.5556 | 0.5556 | 0.5556 | 0.5556 |
| **K-Nearest Neighbors** | 0.4545 | 0.4545 | 0.4545 | 0.4545 | 0.4545 |

Table 4: Model Results for Dummy Data

## IX. CONCLUSION

The F1 score is a metric that balances precision and recall. It is particularly useful in situations where there is an imbalance between the classes, as it considers both false positives and false negatives.

Looking at the F1 scores for each algorithm:

- Decision Tree: F1 Score = 0.293
- Random Forest: F1 Score = 0.4
- Logistic Regression: F1 Score = 0.318
- K-Nearest Neighbors: F1 Score = 0.073

Among these, the **Random Forest** model has the highest **F1 score (0.4)**. This suggests that, based on F1 score alone, the Random Forest algorithm is performing relatively better in terms of finding a balance between precision and recall.

In conclusion, based on the F1 score, Recall & Balanced Accuracy the Random Forest algorithm appears to be the best performer among the evaluated models. However, it is recommended to consider a combination of metrics and thoroughly understand the business implications before making a final decision.

We anticipate that the practical application of our model will streamline MapleTrust Bank's resources more effectively, paving the way for higher returns on marketing investments and improved satisfaction among its clientele. Overall, our findings reinforce the importance of blending machine learning expertise with marketing acumen to navigate the complexities of consumer behaviours.

## ACKNOWLEDGMENT

The authors acknowledge the valuable contributions from Garcia-Serrano, Silva, Demir, Wickham, Liaw, Wiener, Cutler, and Rish, whose works formed the foundation of this research.

## REFERENCES

[1] Garcia-Serrano, A., et al. (2018). Data preprocessing in data mining. A review. Information Sciences, 431-432, 54-67.

[2] Silva, S., et al. (2019). Handling missing data in marketing datasets: A systematic literature review. Expert Systems with Applications, 129, 13-33.

[3] Demir, F., et al. (2018). The evolution of data visualization: A historical perspective. Computers in Human Behavior, 86, 73-84.

[4] Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer.

[5] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R News, 2(3), 18-22.

[6] Cutler, D. R., et al. (2007). Random forests for classification in ecology. Ecology, 88(11), 2783-2792.

[7] Rish, I. (2001). An empirical study of the naive Bayes classifier. In IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence (Vol. 3, pp. 41-46).

[8] Moro,S., Rita,P., and Cortez,P.. (2012). Bank Marketing. UCI Machine Learning Repository. https://doi.org/10.24432/C5K306.