Classification project

Group 64

[1]

1. Introduction

Imagine a busy morning at some bank's call center, where phone calls come in and out, with sales reps hoping to convince customers to open a new fixed deposit in a few short minutes of conversation. Meanwhile, a couple of kilometers away at the fashion brand's headquarters, the marketing team monitors the click-through and open rates of promotional emails, expecting each one to entice customers to click on the link and further explore the new season's exclusive offers. For both organizations, successful marketing campaigns mean not only higher revenues but also deeper customer relationships and brand longevity. However, not every call or every email will get a response. To make these efforts more effective, is it possible to 'anticipate' customer responses with the available data?

In this study, historical marketing data will be used to predict future customer responses to marketing campaigns. By analyzing bank telemarketing calls and fashion store email promotions, our aim was not only to discover the factors that drive positive customer responses but also to find models that could help both companies optimize their marketing decisions. By analyzing the underlying patterns behind this data, accurate, data-driven insights for future marketing campaigns are expected to be provided, making every call and every email more targeted and more likely to deliver successful results.

Initially, the historical marketing data provided by both companies was sorted and analyzed. Then, various machine learning models were constructed to predict the probability of marketing success. The performances of models will be evaluated and compared. The best performance models will be selected to profile the main relevant variables affecting customer response to enhance the customer response rate of the company's future-oriented campaigns.

After an experimentation, a stacked model with Logistic Regression and XGBoost achieved the highest accuracy for predicting a subscription of bank clients after receiving a call. With features study, macroeconomic factor like euribor 3 month rate, characteristics of customers e.g., level of education, marital status, and the approach the bank contacts customer such as day of week, number of contacting during the campaign plays a crucial role in providing important information to the model.

For fashion store, a Stack model combining Logistic Regression with L2 Regularization and LightGBM with L2 Regularization yields the best results in terms of minimizing loss. Specific feature groups like promotional engagement, purchase frequency, spending habits, product interest were used in the models as it is founded to influence predictive abilities.

2. Problem formulation and Objectives

In today's competitive market, improving the success of marketing campaigns through accurate customer insights has become a core challenge for banks and fashion brands. Banks' marketing teams endeavor to reach out to customers over the phone to promote their time-deposit products, while fashion brands aim to engage customers through email promotions. Despite the huge investment in marketing resources, only a few customers ultimately respond, thus identifying potentially highly responsive customers is crucial to optimize resource allocation and improve marketing efficiency.

The use of machine learning is widely spreading around the world. According to Domingos (2012), machine-learning algorithms are used in many business applications. In addition, a study by Ngai and Wu (2022) claims that machine learning is highly attractive to be used in marketing aspects. The reason is it can be utilized to predict consumer behavior, and enhance companies' decision-making by mining insights from large amounts of data.

This project aims to leverage machine learning to create predictive models to analyze and predict customer response behavior in telephonic and email marketing campaigns. By processing bank telemarketing and fashion brand email marketing data, the following key questions are aimed to be answered:

- What are the key elements that contribute to customer response decisions?
- How could the success of future marketing campaigns be predicted based on historical data?

To achieve these goals, a variety of machine learning algorithms will be implemented. Then, the performances of models will be assessed. This project will tackle two following challenges:

1) Predicting whether a customer will respond to a campaign:

1. We acknowledge the use of ChatGPT [https://chat.openai.com/], Gemini [https://gemini.google.com/] to help suggest grammatical, word choice improvements in the report.

Multiple machine learning models will be conducted to predict whether each customer will respond positively to marketing campaigns. The analysis will be based on data from the bank and fashion store. Addressing this challenge will enhance the ability of these organizations to identify potential customers and allocate marketing resources more effectively in future campaigns.

2) Gain insights that will help optimize marketing strategies:
In addition to predictions, at least three valuable insights from the data are expected to be extracted to help the companies develop a deeper understanding of what customer segments are more responsive to marketing campaigns. With these insights, companies can optimize their marketing strategies and improve the success of their campaigns.

From a decision theory perspective, firms face uncertainty in every single promotion activity. For example, should resources be invested in a specific customer? Will this customer respond to our marketing campaign? Therefore, the core issue is how to make the optimal decision to maximize the return on marketing investment (ROI) given limited resources.

This is a predictive problem, as companies need to predict the future behavior of their customers based on historical data. Machine learning is a robust tool to help businesses solve this type of problem. With machine learning models, businesses can find the factors that best influence customer response behavior in a large amount of previous data, increasing the probability of response. This approach allows companies to focus their marketing budgets on customer segments that are more likely to positively respond, thereby increasing the success rate of marketing campaigns and reducing unnecessary investment. Ultimately, the goal is to provide companies with valid tools for decision-making and help them achieve a higher ROI.

This analysis is based on the following assumptions:

1) Representativeness of historical data:
It is assumed that historical marketing data from banks and fashion brands is representative of the future marketing environment. It means that customer response patterns to phone or email promotions will continue to be valid in the future. However, this assumption might fail if there are significant changes in market trends or shifts in consumer behavior. This change is likely to lead to data drift, where statistical properties of the data change over time, resulting in potentially incorrect predictions and reduced model accuracy. To mitigate this risk, it is suggested that clients continuously monitor model performance and track key indicators of data drift. Periodically refreshing the model with updated data to ensure that the model remains effective in capturing current consumer response patterns.

2) Independence and Completeness of Variables:
It is assumed that all the variables used, such as customer characteristics, historical behavioral data, etc., play an independent role in customers' response decisions. Furthermore, due to the limited number of observable variables, other factors not included in the model will not be considered. If this assumption is not fully valid, it is likely to limit the performance of the model.

Moreover, a loss matrix is used to evaluate the performance of the model. This method is straightforward and effective for our binary classification task and helps us to assess the correctness of our predictions and optimize the model. In this project, a loss function tailored to evaluate the model's performance in predicting customer responses to phone and email marketing in a bank and fashion store context is presented. The loss matrices are shown in Table 1 and Table 2.

TABLE 1. LOSS MATRIX OF BANK

Actual/ Predicted	Subscribe	Not Subscribe
Subscribe	0	1
Not Subscribe	2	0

In the bank dataset's scenario, customer responses are collected via phone calls, which involve high investments (time and hiring costs). Therefore, minimizing calls to non-responders is more important. This is shown in the higher penalty for false positives (2), where the model incorrectly predicts a non-responder to be a responder, leading to wasted costs.

TABLE 2. LOSS MATRIX OF FASHION COMPANY

Actual/ Predicted	No response	response
No response	0	1
Response	3	0

In the fashion store dataset's context, customer responses are collected via email marketing, where the outreach cost is low. Thus, the goal is to identify as many potential responders as possible. In other words, occasionally sending marketing emails to non-responders is not as costly as not sending emails to potential responders. This is shown in the higher penalty for false negatives (3), where a model misses identifying a potential responder.

By employing these loss functions, we align the model's objectives with the strategic goals of our marketing campaigns. The distinct penalties for false positives and false negatives ensure that the model prioritizes preventing significant resource wastage and fostering customer trust.

The difference in loss function of the bank and the fashion store highlights that loss functions are context-specific, as the potential impacts of mispredictions vary significantly across different businesses. Ultimately, our approach aims to enhance marketing effectiveness while maintaining strong customer relationships, tailored to the unique dynamics of the business environments.

3. Data understanding

3.1. Bank dataset

In Bank datasets, there are 21 variables including the response. Below items are the data issues and meaningful insights related to the customer's profile and outcome of approaching clients.

- Missing value issues: According to the data dictionary provided by the clients, there are 22 variables: 21 variables for predictors and 1 variable for the response. However, after exploring the dataset, there are 21 variables consisting of 20 inputs and 1 output. This indicates the missing values which is the variable 'balance', representing average yearly balance. We believe that this predictor is missing due to entry errors.
- Outliers: We visualized the distribution of numerical variables using box plots. As shown in the box plot below, only the Last contact duration, in seconds ('duration') and number of days that passed by after the client was last contacted from a previous campaign ('pdays') variables exhibited significant outliers, while the other variables do not show significant outliers. To handle the outliers in the duration variable, we addressed this issue in the feature engineering section.
- New Column for y: After performing initial analysis, we discover that an output of each observation is collected as 'yes' or 'no'. Nevertheless, to perform modeling (Logistic Regression, K nearest neighbor, Decision Tree), they require the response to be numerical form. Consequently, we transform this variable from 'yes' to '1' and 'no' to '0' before further analysis. This mean that if the client subscribe a term deposit, the response will show as 'yes'.
- Key findings
  - Univariate:
    - Continuous Variables: The analysis revealed that none of the continuous variables followed a normal distribution. This non-normality indicates that traditional assumptions of normality may not hold for these predictors, which could impact the modeling approach. Recognizing this, transformations may be considered for certain variables to improve model performance.
    - Discrete Variables: The discrete variables, particularly 'age' and 'previous', exhibited highly right-skewed distributions. This skewness suggests that most customers fall into younger age brackets and have fewer previous contacts, highlighting the need for careful handling of these variables to avoid bias in model predictions.
    - Categorical Variables: In examining the categorical variable 'day_of_week', it was found to have a low impact on the target variable 'y' (subscription to term deposits). This insight suggests that the timing of the contact (specific days of the week) may not significantly influence customer behavior and can potentially be excluded from further analysis.
  - Bivariate Relationship:
    - Logistic regression analysis was performed to capture the relationships between predictors and the response variable. Notably, the variable 'previous' showed a significant relationship with the response, indicating that customers who had more prior contacts with the bank were more likely to subscribe to term deposits. This finding makes 'previous' a strong candidate predictor for the final model, as it highlights the importance of previous interactions in influencing customer decisions.
    - A crosstab plot was utilized to explore the relationship between the response (y) and the discrete and categorical variables ('age', number of contacts performed during this campaign and for this client ('campaign'), 'pdays', 'previous', contact communication type ('contact'), 'job', marital status ('marital'), level of education ('education'), a client has credit in default ('default'), a client has a housing loan ('housing'), a client has a personal loan ('loan'), month on contact ('month'), day of contact ('day of week'), and outcome of the previous marketing campaign ('poutcome'). We discover the interesting pattern between y and 'campaign', 'previous', 'contact', 'default', and 'poutcome'. These variables will be incorporated into the modeling process as candidate predictors due to their observed associations with subscription behavior.
  - Correlation Analysis:
    - Mutual Information (MI) Analysis: The MI analysis highlighted that the top five variables with the highest MI values were all continuous variables representing macroeconomic factors, such as employment variation and consumer confidence. This suggests that these macroeconomic indicators play a crucial role in predicting customer behavior and should be considered in the model.
    - Pearson Correlation Analysis: The analysis indicated that 'emp.var.rate', 'cons.price.idx', 'euribor3m', and 'nr.employed' were highly

correlated with one another. Given their interdependence, including all these variables simultaneously in the model could lead to multicollinearity issues, potentially distorting the model's interpretation and stability. Therefore, it is recommended to select only a subset of these variables based on their individual predictive power and relevance.

- Additionally, the variable 'previous' exhibited a high negative correlation with the response variable 'y', although this correlation was not as strong as those observed among the continuous variables. This negative correlation suggests that as the number of previous contacts increases, the likelihood of subscription may decrease, possibly indicating diminishing returns on repeated contact efforts.

3.2. Fashion store dataset

In a fashion store dataset, 45 variables were analyzed to determine their impact on whether a customer responds (RESP = 1) or does not respond (RESP = 0) to a marketing campaign. The analysis revealed following data quality issues, along with insights related to spending behavior and customer responses to marketing campaigns.

- PC_CALC20 is missing from the data dictionary. Even though its specific purpose or description is unclear, its nature needs to be analyzed. Based on the distribution shown in Appendix, Figure 3, it appears to be a discrete variable.
- Product Category fraction issue: All variables representing the fraction of spending on product category (PSWEATERS, PKNIT_TOPS, PKNIT_DRES, PBLOUSES, PJACKETS, PCAR_PNTS, PCAS_PNTS, PSHIRTS, PDRESSES, PSUITS, POUTERWEAR, PJEWELRY, PFASHION, PLEGWEAR, PCOLLSPND) cannot add up to 1. In some cases, the total was less than 1. In this case, it can be assumed that there were some unobserved categories. For example, if the total was 0.6, it was assumed that the remaining 0.4 represented spending on unrecorded categories. Conversely, when the total was more than 1, it was assumed that there were data entry errors. Therefore, the analysis will be focused on individual category behaviors.
- Percentage variables inconsistency in formatting:
  - Gross Margin Percentage (GMP) and Markdown percentage on customer purchases (MARKDOWN) were in decimal form (0.xx).
  - Percent of returns (PERCRET) had mixed formats, including values like 0.03, 1.29, and 40.92. It was assumed that some values were recorded as percentages (e.g., 10% as 10), while some were recorded as decimals (e.g., 0.1). This inconsistency in formatting will be tackled in feature engineering.
- Skewness and data transformation: Almost every continuous variable (except GMP) has the distribution of skewed to the right. Therefore, a transformation is needed. Yeo-Johnson Transformation was implemented during feature engineering as the data contains zeros.
- Outliers analysis: Outliers were detected in the fraction of spending on specific categories, particularly in PSWEATERS, with some values reaching 1 as shown in Appendix, Figure 4. However, since this variable represents a fraction of spending in a specific product category, it is possible that some customers might just spend a large portion on this category. In addition, all other product category variables had maximum values of 1. Therefore, adjustments were not required here.
- Comparative Analysis Between Responders (RESP = 1) and Non-Responders (RESP = 0)
  - Customers who responded to the campaigns demonstrated a significantly higher average value in frequency of purchases (FRE), average monthly spending (MON), and spending across product categories (AMSPEND, PSSPEND, CCSPEND, AXSPEND, TMONSPEND, OMONSPEND, SMONSPEND) as shown in Appendix, Table X. This trend suggests that customers who are more engaged and have substantial spending history seem to be more likely to respond to marketing campaigns.
  - There is a substantial difference in the average number of product classes purchased (CLASSES) between the two groups of customers as shown in Appendix, Table X. Customers who responded to the campaign appear more likely to purchase a wider variety of product classes compared to those who did not respond.
- Discrete and binary variables connection to response: Customers with a higher number of marketing promotions on file (PROMOS), a history of responding to campaigns (RESPONDED), and a greater number of promotions mailed in the past year (MAILED) seems more likely to be the one responding to the current campaign.
- Bivariate Relationship (Association of continuous variables with RESP): As shown in Appendix, Figure 5, MON, AMSPEND, PSSPEND, CCSPEND, AXSPEND, TMONSPEND, OMONSPEND, SMONSPEND, PREVPD, MARKDOWN, RESPONSERATE, and PERCRET all show a positive association with campaign response. Conversely, variables like AVRG, FREDAYS, HI, and LTFREDAY show a negative association with response.

- The Mutual Information (MI) analysis: MI values indicate the strength of the association between each predictor and response (RESP). As shown in Appendix, Table X, It was revealed that FREDAYS, LTFREDAY, and FRE have the top three highest MI values, making them interesting to use as predictors for the model.
- Correlation analysis: Correlations were analyzed in 2 aspects:
  - Correlation between continuous variables and a response variable: None of the continuous variables are significantly correlated with a response variable. The strongest correlations occurred between RESPONSERATE and RESP. (0.33)
  - Correlation between the continuous variables itself: As there are large numbers of variables, calculating correlations between all pairs of variables would be excessive. Therefore, only the pairs of variables with a correlation above 0.7 or below -0.7 were filtered as this range of numbers is considered a strong correlation according to Ratner (2009). This calculation identifies variable pairs that should be avoided to be included in the same model as it might introduce multicollinearity issues.

4. Feature Engineering

Feature engineering is a critical component of the machine learning process, as it involves the creation, transformation, and selection of features that enhance model performance. Studies have shown that effective feature engineering can lead to better performance compared to using raw data alone (Hastie, Tibshirani, & Friedman, 2009). The following outlines the approaches and processes for feature engineering that are considered suitable for the variables contained in the two datasets.

4.1. Handling Outliers

Handling categorical variables is considered a crucial step in preparing data for statistical and machine learning analysis, as numerical input is required by many algorithms. Categorical variables must be encoded numerically to ensure compatibility with these algorithms. This numeric coding is also seen to enhance model interpretability by clarifying relationships between variables, providing insights into how different groups affect the outcome. Furthermore, mathematical operations are facilitated, and potential interactions and non-linear relationships within the data can be captured more effectively.

- Bank dataset: During the EDA, numerous outliers with a value of 999 in the 'pdays' variable were observed, representing the number of days since the last customer contact. Given that -1 indicates no previous contact according to the data dictionary, all 999 values, which were inconsistent with the expected range, was replaced with -1 to accurately reflect the data's meaning. By doing this, the integrity of the dataset is maintained, ensuring that the model does not misinterpret these values as valid data points.
  Another critical insight from the EDA involved the duration variable, which captures the duration of the last contact made with the customer. Analysis revealed a strong correlation between longer call durations and an increased likelihood of the customer subscribing to a term deposit. This relationship is significant; however, it poses a risk of target leakage. Target leakage occurs when a feature used in training the model provides information about the target variable that would not be available at the time of prediction (Hastie, Tibshirani, Friedman, 2009). In this case, the duration of the call is likely influenced by the outcome of the conversation—if a customer is more engaged or interested, the call may naturally extend longer. This means that including duration could unfairly bias the model, leading to inflated performance metrics that do not generalize well to unseen data. To prevent target leakage and ensure that the model's predictions are based solely on features that would be available at the time of making a marketing call, the duration variable was excluded from the model.

4.2. Handling Categorical Variables

Both datasets contain several categorical variables, which are essential for capturing the diversity in customer characteristics and their potential influence on the success of the marketing campaign. However, machine learning algorithms typically require input features to be in a numerical format, necessitating a suitable encoding strategy for categorical variables. To convert categorical variables into a numerical format, dummy encoding (also known as one-hot encoding) was utilized in both band and fashion dataset. This method transforms each category within a categorical variable into a new binary variable (0 or 1), allowing machine learning models to interpret these variables effectively.

- Bank Dataset: The bank dataset contains several categorical variables, such as 'default', 'contact', 'poutcome' and so on.
  - Variable Example: The 'default' variable indicates whether a customer has credit

in default, with three possible responses: 'yes', 'no', and 'unknown'. In the context of predicting subscription to a term deposit, having credit in default may negatively impact a customer's likelihood to subscribe, as it reflects financial reliability. For instance, a customer with 'no' default may have a higher probability of subscribing compared to those with 'yes' or 'unknown'. The dummy encoding process for this variable involves creating two new binary variables using the drop_first option to avoid the dummy variable trap:

- default_yes: 1 if the customer has credit in default, otherwise 0.
- default_unknown: 1 if the customer's status is unknown, otherwise 0.

For a record indicating a customer with no credit in default, the values would be: default_yes = 0 and default_unknown = 0. This implies that the customer's status is 'no' regarding credit in default.

* Sparse Category Handling: 'Job' and 'education' contain many categories, some of them with low representation. To manage these variables effectively, sparse category encoding was applied initially. For instance, 'job' contains many categories that reflect various employment types. The categorization into three types based on meaning and distribution was performed:

  - Limited Income: This category includes jobs that typically offer lower financial returns, such as 'student' and 'retired'.
  - Low Income: This category encompasses jobs with generally lower wages, such as 'blue-collar' professions.
  - Others: This category contains all remaining job types that do not fit into the first two categories.

  By reducing the number of unique categories through sparse category encoding, the model is better equipped to identify patterns associated with income levels and their impact on the likelihood of subscribing to a term deposit. Following the application of sparse category encoding, dummy encoding was utilized for both 'job' variable. The same sparse category processes were applied to 'education' variable in the bank dataset. After that, all categorical and binary variables were used dummy encoding, enabling the model to capture relevant information effectively while maintaining a manageable feature set.

- Fashion store dataset:
  * During the EDA, the categorical variable found is 'VALPHON,' which indicates whether a customer has a valid phone number ('Y' for Yes and 'N' for No).
  * For communication purposes, analyzing this information helps in validating the accuracy of our contact database and allows for direct outreach opportunities, such as sending reminders, updates therefore, facilitates the response rates.
  * As most machine learning algorithms require numerical input, by transforming 'VALPHON' (Y/N) into a dummy variable (1 for Yes, 0 for No), it can be effectively used in models as dummy variables help maintain consistency across categorical variables in the dataset.
  * The categorical variable 'VALPHON' was first transformed into new dummy variables to facilitate its incorporation into our analysis. To mitigate the risk of multicollinearity, we subsequently dropped one of the dummy variables. Following this step, we removed the original 'VALPHON' column from the dataset. We then merged the retained dummy variable, designated as 'VALPHON_Y' back into the original dataset. This process resulted in the creation of a new feature, where 'VALPHON_Y' is represented numerically: it takes a value of 0 for "No" and 1 for "Yes".

4.3. Yeo-Johnson Transformation and Standardization

The Yeo-Johnson transformation is a valuable technique in data preprocessing, particularly for handling non-normally distributed data. It effectively addresses issues of skewness in the data, facilitating better modeling of relationships between variables and reducing the impact of outliers (Yeo & Johnson, 2000). Since many skewness are detected in EDA part, it is critical for transforming our dataset to improve the quality and interpretability of data before modeling.

- Bank dataset: Bank dataset: During the Exploratory Data Analysis (EDA), it was observed that both the 'cons.price.idx'(Consumer Price Index,

CPI) and the 'cons.conf.idx'(Consumer Confidence Index, CCI) displayed clear skewed distributions. Specifically, the CPI ranged from 92.201 to 94.767, indicating a narrow band of values concentrated around the lower end, while the CCI ranged from -50.8 to -26.9, indicating a broader negative range with values skewed towards lower confidence levels. These distributions can affect the performance of machine learning models, as many algorithms assume that input features are normally distributed.

  - The Yeo-Johnson transformation was chosen due to its ability to handle both positive and negative values, making it suitable for variables like the CCI, which contains negative values. By reducing skewness, the transformation helps ensure that the model can more accurately capture the relationships between economic conditions and customer behavior. For example, a more normally distributed CCI may allow the model to better understand how fluctuations in consumer confidence correlate with subscription rates to term deposits.
  - Following the transformation, StandardScaler was used to standardize the features. StandardScaler rescales the data so that each feature has a mean of 0 and a standard deviation of 1. This step is critical in a financial context, where the impact of various economic indicators needs to be evaluated on a comparable scale. After applying StandardScaler, the distributions of the CPI and CCI were centered around zero, facilitating better interpretation and model performance in subsequent analyses.

- Fashion dataset: When examining the distribution of continuous numerical variables, it is observed that a significant number exhibit right skewness. To enhance the quality of the analysis and improve the performance of predictive models, a transformation is necessary. Given that our dataset includes numerous zeros, the Yeo-Johnson transformation is deemed the most appropriate choice, as it effectively stabilizes variance and facilitates a more normal distribution of the data.
  - Following the transformation, we find that several key features, including 'AVRG', 'MON', 'CCSPEND', 'HI', and 'LTFREDAY', exhibit normalized distributions. These transformed variables have been shown to have interesting patterns that can provide valuable insights during the analysis. After that, we conducted standardization by using StandardScalar as standardization ensures that all features contribute equally to the analysis. This is crucial when our features have different units or scales, as it prevents features with larger ranges from disproportionately influencing the model (Hastie, Tibshirani, Friedman, 2009).
  - As mentioned in data understanding part, PERCRET have inconsistencies in formatting. As the bigger proportion of PERCRET values were less than 1; therefore the values that are greater than 1 were standardized by dividing by 100.

4.4. Binning for Chosen Continuous Variables

In both the Bank and Fashion datasets, continuous variables exhibit a large number of sparse values. Therefore, converting these continuous variables into categorical bins is necessary. This transformation can lead to improved model performance, particularly with algorithms that favor categorical inputs, such as decision trees. Binning can effectively capture non-linear relationships within the data (Friedman & Popescu, 2008).

- Bank dataset: By implementing the binning strategies, the model is better positioned to understand the relationships between these important variables and the likelihood of customers subscribing to term deposits. These transformations enhance the interpretability of the model and allow for more targeted marketing strategies based on customer profiles.
  - emp.var.rate: indicates the change in employment from the previous quarter, serving as a key indicator of economic health.
    - Binning Process: The emp.var.rate was categorized into two groups: positive (indicating economic growth) and negative (indicating potential recession).
    - Rationale: Binning this variable allows the model to capture significant shifts in economic conditions. A positive employment variation rate typically suggests an expanding economy, which may correlate with higher customer confidence and a greater likelihood of subscribing to term deposits. Conversely, a negative rate could reflect economic downturns,

which may deter customers from making long-term financial commitments.

* euribor3m: reflects the interest rates set by European banks and serves as a benchmark for lending rates across the Eurozone.
  - Binning Process: A threshold of 2% was established, categorizing the rate as low (below 2%) or high (above 2%).
  - Rationale: This binning enables the model to assess the impact of interest rate fluctuations on customer behavior. Low interest rates can incentivize customers to invest in term deposits, as they seek better returns compared to other savings options. Conversely, high rates may suggest tighter monetary policy and increased borrowing costs, potentially reducing customer willingness to subscribe to term deposits.
* age: a significant demographic factor influencing financial behavior and product suitability.
  - Binning Process: The age variable was divided into three categories: young (below 35), middle-aged (35 to 55), and older (above 55).
  - Rationale: Different age groups exhibit distinct financial needs and behaviors. Younger individuals may be more inclined to take risks or explore investment opportunities, while middle-aged customers often focus on savings and stability. Older customers may prioritize security and income generation in retirement. By binning age into these categories, the model can more effectively capture these behavioral differences and their influence on subscription likelihood.
* pdays: indicates the number of days since a customer was last contacted in a previous campaign.
  - Binning Process: Given the observation that a large proportion of customers had never been contacted before, this variable was binned into two categories: contacted before and not contacted before.
  - Rationale: This binning simplifies the analysis by highlighting the difference between customers who have prior interaction with the bank and those who do not. Customers who have been contacted previously may have a better understanding of the bank's offerings, potentially making them more responsive to marketing efforts. In contrast, those who have never been contacted may require different strategies to engage them effectively.

– Fashion store data:
  * In the next step, we identify two key features that may significantly impact the results of our marketing campaign and help detect potential customers with a positive response: 'MARKDOWN' and 'RESPONSERATE'. Both features are normalized to a range between 0 and 1, representing percentages.
    - MARKDOWN: This variable indicates the discount applied to products. It is evident that customers are generally more responsive to discounts. Notably, there are many sparse values in the 'MARKDOWN' variable, with the majority falling between 0 and 0.5 (representing discounts of 0
  * To mitigate the risk of losing valuable information due to sparsity and to enhance the model's accuracy, we propose performing binning on the MARKDOWN variable. This approach will allow us to categorize the data more effectively. Given that customers appear to be more attracted to marketing emails or campaigns featuring discounts, we will create two categorical variables based on binning: "No Discount" and "Discounted." This transformation will facilitate the model's ability to accurately learn patterns across the range of values.
    - The response rate from last year, represented by 'RESPONSERATE,' can significantly influence this year's campaign. If certain demographics or customer segments exhibited higher response rates, similar groups can be prioritized in the current campaign to optimize engagement. This feature allows for the leveraging of information, as last year's response rate may highlight customer preferences and behavioral trends that can be utilized in the current campaign. For

instance, if a specific product or discount type resulted in a higher response, analogous offers can be incorporated this year. Consequently, it is believed that 'RESPONSERATE' could provide valuable insights.

* A similar situation is observed with 'MARKDOWN,' where the majority of values are 0, indicating that customers did not respond positively to our marketing emails. However, a small proportion of values indicates positive responses, and the limited number of these values raises the risk of missing critical information during the model learning process. To address this, 'RESPONSERATE' was binned to enhance the significance of smaller values, ensuring that all ranges are equally represented. The response rate was rearranged into five categories: "Unlikely to Respond," "Unlikely," "Moderate," "Likely," and "Super Likely." This categorization is expected to enable the model to better capture high response rates.
* Finally, dummy transformation was applied once again to convert the categorical variables into numerical, facilitating the model's learning process.

4.5. Interaction Terms

In the context of learning concepts, the notion of irrelevant variables is particularly pertinent when considering their potential to interact with other variables. While these variables may not exhibit significant main effects on their own, their interactions with relevant predictors can uncover important insights that would otherwise remain obscured (James, G., Witten, D., Hastie, T. & Tibshirani, R., 2021). Additionally, creating interaction terms or aggregating features can provide clearer insights into the relationships between variables and the target outcome (Molnar, 2020). To enhance model performance and capture complex relationships between predictors, interaction terms were explored. For this purpose, ElasticNet regression with cross-validation (cv5) was employed.

- Bank dataset: By implementing ElasticNet regression, the top 3 interaction terms that could potentially enhance the model's predictive power was identified:
  * pdays and nr.employed: This interaction may indicate how the number of days since last contact affects the employment status, influencing the likelihood of subscription to term deposits.
  * nr.employed and month_march: This term explores how the employment levels interact with marketing efforts conducted in March, a month that may have specific seasonal effects on customer behavior.
  * nr.employed and emp.var.rate_positive: This interaction reflects how employment levels relate to positive employment variation, suggesting that increases in employment may have different impacts on customer subscription behavior depending on the overall economic sentiment.

  However, after integrating these interaction terms into the model, a decline in overall performance was observed. Several factors contributed to this outcome:

  * Overfitting: The introduction of additional interaction terms may have led to overfitting, where the model became too complex and tailored to the training data rather than generalizing well to unseen data. This is particularly concerning in a marketing context, where the goal is to predict customer behavior accurately in a broader population.
  * Increased Dimensionality: Adding interaction terms increased the dimensionality of the feature space, which can complicate the learning process for the model. With too many features, the model may struggle to identify the most relevant predictors, leading to poorer performance.
  * Noise Introduction: Some of the interaction terms (like pdays and nr.employed) may have introduced noise rather than useful information. If the interaction effects are not strong or consistent, they can dilute the signal provided by the primary features, reducing the model's predictive ability.
- Fashion store dataset:
  * The generation of interaction terms is a critical endeavor in modeling complex relationships among multiple variables. Given the presence of 47 variables in the dataset, the exhaustive exploration of all possible interaction combinations is not only time-consuming but also computationally inefficient. Therefore, a more pragmatic approach involves the strategic generation of interaction terms followed by a systematic selection process to identify the three most significant ones.
  * Initially, interaction terms are generated from the original features to capture potential interactions that may not be ap-

parent when considering features individually. After that we fit an ElasticNet model, which combines Lasso and Ridge regression, to the interaction terms. By using cross-validation (with 5 folds), we enhance the model's robustness and help prevent overfitting while selecting features. Ultimately, interactions are prioritized based on the absolute values of their coefficients, allowing for the identification of the three most significant interactions. This approach facilitates the pinpointing of feature pairs that exert the greatest influence on the response variable as follows: MON and LTFREDAY, DAYS and RESPONSERATE, and LTFREDAY and DAYS.

5. Methodology

To build a machine learning model for predicting the success of marketing campaigns, generalization or an ability of a learning algorithm to predict new data is an important concept. Therefore, an evaluation strategy is crucial for achieving this goal. Before any feature engineering or model training, each dataset will be split into a 70% training set and a 30% validation set. This split ensured that the models were trained on a significant portion of the data, while the validation set provided an independent dataset to evaluate the model's generalization performance.

5.1. Feature Selection

Before estimating any models, feature screening based on information we collected through the Exploratory Data Analysis (EDA) was performed.

- Bank Dataset: Features were selected based on the following criteria:
  - Using Mutual Information (MI), continuous variables are considered as potential predictors. However, to avoid multicollinearity, we excluded the following variables: 'emp.var.rate', 'cons.price.idx', and 'nr.employed' which show high correlation with 'euribor3m'.
  - In terms of nominal variables, the analysis shows that 'campaign', 'previous', 'contact', 'default', and 'poutcome' have an interesting relationship with the response.
  - Adding interaction terms to capture the combined effect of two or more variables and use the top three interaction terms.

  However, after estimating various models with selected features and/or interaction terms, the performance of the refinement models is worse than the full model. As a result, the following analysis will focus on the full model without 'emp.var.rate', 'cons.price.idx', and 'nr.employed' to avoid the multicollinearity issues.
  To sum up, the predictors used to estimate the model after data transformation and feature engineering are contact, campaign, pdays, cons.conf.idx, euribor3m_low, age_middleaged, age_older, job_low, job_other, marital_married, marital_single, marital_unknown, education_others, education_professional, default_unknown, default_yes, housing_unknown, housing_yes, loan_unknown, loan_yes, month_aug, month_dec, month_jul, month_jun, month_mar, month_may, month_nov, month_oct, month_sep, day_of_week_mon, day_of_week_thu, day_of_week_tue, day_of_week_wed, poutcome_nonexistent, poutcome_success, previous_None, and previous_Once.

- Fashion store Dataset: Features were selected based on the following criteria:
  - Continuous variables with positive or negative associations with response variables based on plots.
  - Discrete variables with relationships with response variables.
  - Variables with the top 3 highest mutual information (MI) scores, which indicate the high strength of association between the predictor and the response.

  To avoid multicollinearity issues, variables from highly correlated pairs (above or below the threshold of 0.7 and -0.7) like MON, TMONSPEND, OMONSPEND, SMONSPEND were removed.

  Additionally, the variables RESPONSERATE and MARKDOWN were binned to improve the model's predictive power as mentioned in the feature engineering part. After comparison, binning these features leads to better models in terms of loss. Thus, they were included in the final predictors. In addition, the top three interaction terms created during the feature engineering process were included to capture complex relationships and enhance model performance. (MON and LTFREDAY, DAYS and RESPONSERATE, and LTFREDAY and DAYS)

The following features were chosen as final predictors:
PROMOS, RESPONDED, FREDAYS, FRE, LTFREDAY, VALPHON, GMP, CC_CARD, PERCRET, PSSPEND, AVRG, AMSPEND, CCSPEND, MARKDOWN_Binned_Discounted, RESPONSERATE_Binned_Unlikely, RESPONSERATE_Binned_Moderate, RESPONSERATE_Binned_Likely, RESPONSERATE_Binned_Super Likely, AXSPEND, PREVPD, HI, STYLES, DAYS, MON LTFREDAY, DAYS RESPONSERATE, LTFREDAY DAYS.

5.2. Modeling

For each dataset, various modeling techniques have been implemented, focusing on three types of models: linear models, tree-based models, and model stacking. As this is a classification problem, one of the models would be Logistic Regression, followed by tree-based models like Decision Trees and Random forests. Bagging and boosting techniques (LightGBM, CatBoost, and XGBoost) were also applied. Finally, potential models were selected for model stacking. Regularization techniques, including Lasso (L1), Ridge (L2), and Elastic net, were incorporated across models to reduce the risk of overfitting and enhance performance. Additionally, hyperparameter tuning processes were included to optimize the models' predictive power.
After implementing the models, the top performers for each model type were identified. The main criterion for model selection is minimizing loss. Additionally, sensitivity and precision were important factors that should be considered. As mentioned in the problem formulation, the priorities will differ between contexts.

1) Bank Context: Avoid false positives, or incorrectly predicting a non-responder as a responder.
2) Fashion Store Context: Detect as many potential responders as possible.

These priorities are reflected in the loss matrices shown in Table 1 and Table 2.

5.2.1. Model.

- **Model Benchmark:** Logistic regression models will be selected as benchmark models for both datasets. The reason is Logistic Regression is simple and interpretable. It will serve as a baseline, allowing for comparison with more complex models to see if additional complexity and techniques provide meaningful improvements.
- **Linear Model:** Logistic Regression with various techniques added is conducted since it is an interpretable linear model commonly used for binary classification tasks, which suits well with goals of identifying potential responders to marketing campaigns. Additionally, from the EDA of fashion store data, it was observed that many variables have a sigmoidal (S-shaped) relationship when plotted against the response variable (as shown in Appendix, Figure 5). Having a sigmoidal shape is a key indicator for using Logistic Regression as it is a core characteristic of this type of model.
  - Bank Dataset: *Logistic Regression and Logistic Regression with L2 Regularization*
    To improve the model performance, both L1 and L2 were applied to the model. However, the result of the model with L1 and L2 appeared to be worse and equal to the original model (AUC 0.776 and 0.803 respectively while the AUC of Logistic Regression is 0.803). This indicates that our unregularized Logistic Regression model is stable and all predictors are not highly correlated and provide non-redundant information. This also reaffirms that all predictors except the ones that are highly correlated with 'euribor3m' should be the optimized set of predictors.
  - Fashion store Dataset: *Logistic Regression with L2 Regularization*
    To prevent overfitting, both L1 and L2 were included. The outcome is that using L2 leads to slight lower loss than L1 (0.195 compared to 0.196), suggesting that predictors in the model might be correlated. L2 regularization is robust in handling those correlations by shrinking coefficients smoothly rather than removing them, as L1 would. It can be implied that the predictors included do contribute some importance in the model.
- **Tree-Based Method** As this is a marketing-related dataset, customer behavior can vary significantly among different segments. There could be non-linear relationships. Thus, Decision Trees were chosen aiming to capture such relationships. (Zollanvari, 2023)
  - Bank Dataset: *Random Forest*
    Both Decision Tree and Random Forest were conducted; the Random Forest outperforms the Decision Tree. It was

introduced to overcome the problems of high correlation among the trees, leading to higher performance. Number of trees (n_estimators=100), number of candidate split variables (max_features = 3) and minimum node size for each tree (min_samples_leaf = 5) are setting. This is aimed at striking a balance between computational efficiency and accuracy.

* Fashion store Dataset: *Decision Trees* Both of Decision Trees and Random Forest were conducted. After comparison, Decision Trees performed better. Configuration and Regularization Choice: To prevent overfitting, where trees become too complex and capture training data too well rather than the underlying pattern, restrictions were implemented, including setting minimum samples leaf (min_samples_leaf) = 5 and maximum leaf nodes (max_leaf_nodes) = 10.

– **Model Stacking** Stacking is an ensemble technique that combines models to leverage individual strengths and improve overall performance. In this case, different sets of models were stacked and evaluated. The choices of the combinations to stack are from the model's individual performance. Logistic Regression is used as a meta-model throughout all the stacking choices as it is a simple and interpretable model, potentially providing a smooth interpretation of the predictions from base models. (Brownlee, 2021).

  * Bank Dataset: *A stacked model combining Logistic Regression with XGBoost* - Based on the performance of various individual models, XGBoost and Logistic Regression are the top two best performance models with loss 0.105 and 0.107 respectively. When combining both models together with Logistic Regression as a meta model, the highest model performance with loss 0.102 was achieved. Although, other potential combinations such as including regularization in the stacked model were conducted, the performance of the models did not improve. Therefore, the stacked model with the combination of Logistic Regression and XGBoost are the best combination producing the best accuracy. This indicates that the strength of the XGBoost in capturing nonlinear relationships strengthens the prediction accuracy with the simplicity and interpretability of the Logistic Regression.

  * Fashion store Dataset: *A stacked model combining Logistic Regression with L2 Regularization and LightGBM with L2* Regularization, using Logistic Regression as the meta-model. - Logistic Regression individual performance (minimize loss) is quite solid, when including L2 (loss = 0.195). LightGBM was included as another base model due to its strength in capturing non-linearity. It was found that applying L2 to LightGBM improved its performance by reducing the loss. The fact that this model stacking performs well can indicate that data may comprise a mix of linear and non-linear relationships. Configuration and Regularization Choice: After trying L1 and L2 techniques to generalize the model, it turned out that L2 yielded better results. This result is consistent with the fact that using L2 with Logistic regression yielded better results comparing to L1.

– **Other Models**

  * Bank Dataset: Other models and techniques were experimented to find the best possible model. K-Nearest Neighbors (KNN), Bagging, and various boosting techniques are performed. For KNN,K neighbors = 30 was selected as the optimal number of neighbors, determined through cross-validation, and used Euclidean distance to find the nearest neighbors for classification. The outcome of KNN model outperforms Logistic Regression in terms of loss. However, the AUC score is worse. Meanwhile, XGBoost outperforms other models in all metrics if the stacked models are not considered. This highlights the importance non-linear data patterns which were detected and learned by XGBoost (Hashcham, 2023).

TABLE 3. COMPARISON OF MODEL LOSS, PRECISION AND AUC

Model	Loss	Precision	AUC
Logistic Regression	0.107	0.534	0.803
KNN	0.106	0.540	0.803
Bagging	0.137	0.402	0.753
CatBoost	0.106	0.540	0.803
LightGBM	0.111	0.511	0.790
XGBoost	0.105	0.540	0.807

  * Fashion store Dataset: Various models have been conducted, including KNN, Bagging, Boosting techniques, with performance metrics presented in Appendix,

Table 10. All boosting models (XGBoost, LightGBM, CatBoost) outperform the Logistic Regression in terms of minimizing loss. (shown in Table X). However, sensitivity and precision of those models are lower than that of Logistic regression, implying that they may not perform well in identifying responders.

TABLE 4. COMPARISON OF MODEL LOSS AND SENSITIVITY

Model	Loss	Sensitivity
Logistic Regression	0.196	0.656
Bagging	0.207	0.655
XGBoost	0.193	0.633
LightGBM	0.172	0.530
CatBoost	0.189	0.594

5.2.2. Hyperparameter Optimization. Given computational constraints, only selected hyperparameters and selected models were tuned by using GridSearchCV and Optuna algorithms.

- Bank Dataset: Only Decision Tree was used during the process. A grid search was employed to optimize the hyperparameters for the Decision Tree classifier. The following hyperparameters were identified as the best combination based on cross-validation performance:

  1) Maximum Depth: 5
  2) Minimum Samples per Leaf: 1
  3) Minimum Samples per Split: 2

  These settings balanced the complexity of the tree and helped prevent overfitting, leading to better generalization on the validation set.

- Fashion store Dataset: Decision Trees, Random Forests, and LightGBM were models used for this process. The hyperparameters tuned are as follows:

  1) Decision Trees: Cost-complexity pruning was implemented. A hyperparameter 'ccp_alphas' was tuned. The objective of this is to control the complexity of the tree, and prevent overfitting by pruning branches. By tuning the ccp_alphas parameter, which controls the tree's complexity, the risk of overfitting will be reduced. To select an optimal ccp_alphas, a scoring method of 'negative log loss' which measures predictive probabilistic accuracy was used. The model using optimal ccp_alphas found by GridSearchCV algorithms is expected to have better balancing in accuracy and complexity. The optimized ccp_alphas is 0.001.
  2) Random Forests: Followings are hyperparameters that have been tuned:

     * criterion: The function used to measure the quality of splits ('gini' or 'entropy'). As different measures can lead to different predictive strengths, tuning criterion leads to the model adapt to the data better.
     * min_samples_leaf: The minimum number of samples required at a leaf node. Tuning this hyperparameter ensures that the model does not become overly complex and overfitting.
     * max_features: The maximum number of features considered when splitting. By tuning this, the diversity of trees in the forest will be controlled, leading to reduced risk in overfitting.

     The goal is to reach optimal balance between the model performance and complexity. Followings were identified as the best hyperparameters:

     a) Criterion: gini
     b) min_samples_leaf: 15
     c) max_features: 12

  3) LightGBM: Tuned hyperparameters were:

     * n_estimators: The number of boosted trees to fit.
     * num_leaves: The maximum number of leaves per tree for base learners. Tuning maximum number of leaves ensures that complexity of the model is controlled.

     Due to computation time constraints, the learning rate was fixed at 0.01, as that is one of a default value in machine learning [Brownlee, 2019] Followings were identified as the best hyperparameters:

     a) n_estimators: 1600
     b) num_leaves: 90

6. Results

To ensure reliable model performance, a validation set approach on the training data was performed, and model metrics were compared. By doing so,

a more comprehensive understanding of our models' generalization ability and robustness can be derived. The models were evaluated based on multiple performance metrics, including Loss, Sensitivity, Specificity, Precision, and Area Under the Receiver operating characteristic Curve (AUC). These metrics helped us understand each model's strengths and weaknesses in predicting the outcome of the marketing campaign. Specifically, Loss function and AUC are two metrics that will be mainly focused on, as Loss function determines the error between the prediction output and the provided target value. The model's performance will be evaluated by comparing the distance between the prediction output and the target values. Smaller loss means the model performs better in yielding predictions closer to the target values (Huynh, 2023). For AUC, it tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1 (Narkhede, 2018).

6.1. Bank dataset

Table 3 illustrates the performance of each model.

TABLE 5. COMPARISON OF MODEL PERFORMANCE METRICS

Model	Loss	Sensitivity	Specificity	Precision	AUC
Logistic Regression	0.107	0.376	0.958	0.534	0.803
Random Forests	0.109	0.453	0.947	0.520	0.801
XGBoost	0.105	0.433	0.953	0.540	0.807
Stacking: Logistic Regression and XGBoost	0.102	0.347	0.968	0.578	0.804

Interpretation of the Results

- **Benchmark Model:** The Logistic Regression model performed well in predicting the subscription of the clients with loss 0.107 and AUC 0.802. With AUC of 0.802, deploying the model will help the business reach potential customers more effectively compared to random chance.
- **Random Forest:** The Random Forest demonstrates a slightly worse performance with a higher loss 0.109, lower precision 0.520 and lower AUC 0.797 than the benchmark model, indicating poorer performance in all metrics. This means that if the bank deploys this model, it could lead to the potential loss in terms of cost and time spending on the non-potential customers. This also suggests that applying the benchmark model is a better option if considering the Random Forest.
- **XGBoost:** The XGBoost outperforms Logistic Regression and Random Forest with a lower loss of 0.105, higher precision of 0.540 and higher AUC of 0.807. This demonstrates the importance of non-linear relationship information has on the model performance which is accordance with the discovery during the EDA and modeling process that there are various showing non-linear relationships with the response. In addition, the fact that the full model (all variables (excluding highly correlated variables) outperforms the selected variables model supports the significance of non-linear patterns. To sum up, using all features (both linear and non-linear relationships) leads to better predictive performance, as each feature contributes significant insights to the model.
- **Stacked Model:** The stack model, combining the strength of Logistic Regression and XGBoost, presents the best prediction result with the lowest loss of 0.102, the highest precision of 0.577 and a strong AUC of 0.804, compared to other models, including the benchmark model. The AUC (0.804) indicates that using the machine learning model is better than random chance (Terra, 2024). With the fact that the model provides the best performance, this could help the bank maximize costs saving associated with contacting non-potential customers and can gain revenue from the potential ones.

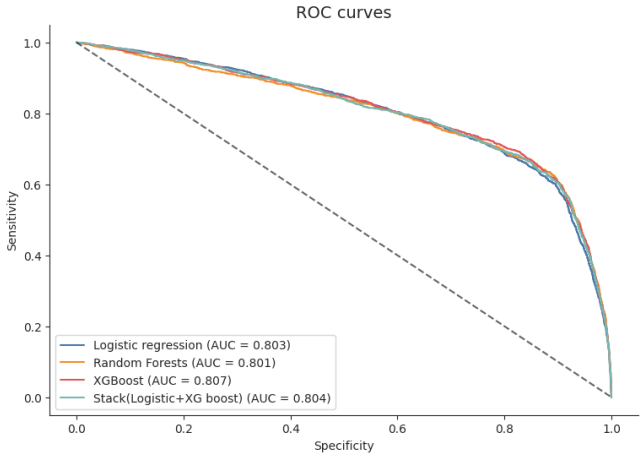

Figure 1. ROC Curve of Bank Dataset

ROC Curve presented in Figure 1 demonstrates the balance between sensitivity and specificity of each selected model. As mentioned above, the AUC score of XGBoost presents the best performance of a binary classifier; however, the scores of these four selected models are slightly different. When considering other important metrics such as loss and precision, the stacked model consisting of XGBoost and Logistic Regression yield the best in these two perspectives. Therefore, it is recommended that bank deploys the stacked model, as it is the best option in terms of cost-saving from misprediction and yielding high prediction performance.

6.2. Fashion store dataset

The performance metrics summarized in Table X show insights of each model: the benchmark Logistic Regression model, linear, tree-based, and stacked model.

TABLE 6. COMPARISON OF MODEL PERFORMANCE METRICS

Model	Loss	Sensitivity	Specificity	Precision	AUC
Logistic Regression	0.196	0.656	0.834	0.440	0.852
Decision Tree	0.194	0.571	0.852	0.435	0.825
Logistic Regression (with L2)	0.195	0.655	0.835	0.441	0.852
Stacking Logistic Regression (with L2) and LightGBM (with L2)	0.167	0.568	0.886	0.498	0.852

Interpretation of the Results

- **Benchmark Model:** The Logistic Regression model has a solid performance. It can identify non-responders quite well, seeing from high specificity. However, there is still a room for improvements in sensitivity and precision, which currently indicates that the model might not perform well in identifying and predicting responders.
- **Decision Tree:** The Decision Tree model slightly reduces loss compared to the benchmark model. The specificity also increases. However, it shows a lower sensitivity. This means that while it may correctly identify non-responders, it could have problems identifying responders. This makes the Decision Tree less suitable as sensitivity is more prioritized according to the business goals of identifying responders.
- **Logistic Regression with L2 Regularization:** Implementing L2 regularization to the Logistic Regression model yields minor increases in sensitivity and precision over the benchmark, with a slight reduction in loss. This indicates that regularization helps the model to become more generalized. However, the improvement is still low. Thus, more alternative techniques can be explored.
- **Stacked Model:** The stacked model, which combines Logistic Regression with L2 regularization and LightGBM with L2 regularization, demonstrates the best performance across loss, specificity, and precision. However, its sensitivity is lower than that of the benchmark model. This finding indicates that while ensemble techniques (combining models) improve the model's overall performance by reducing loss, further explorations and adjustments are necessary to enhance the model's ability to capture responders and improve sensitivity.

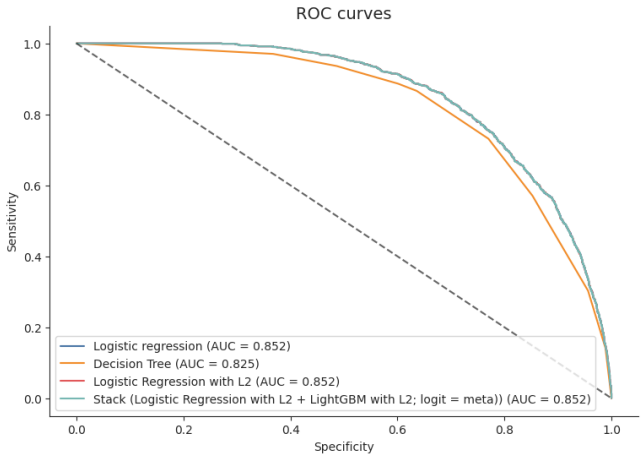

Figure 2. ROC Curve of Fashion store dataset

ROC Curve presented in Figure 2 shows how each model balances its specificity and sensitivity. AUC scores of Logistic Regression, Logistic Regression with L2, and the Stacked model (Logistic Regression with L2 and LightGBM with L2) are all equal at 0.852. However, as mentioned above, the The Decision Tree has the lowest AUC (0.825), highlighting its comparative weakness in overall performance.

7. Data mining

Through our analysis of the bank and fashion store marketing datasets, valuable insights that can enhance future campaign strategies and decision-making were uncovered. Key patterns likely to drive engagement and increase campaign success were

identified by examining customer responses across various factors.

7.1. Bank dataset

Firstly, according to the classification of the Decision Trees we find that the first prerequisite for the occurrence of 0, 1 classification is whether the month is October. This condition alerts us that during certain specific times of the year, the bank may need to adjust its marketing strategy to improve customer engagement. It is recommended to analyze the factors contributing to the low response, such as seasonal influences or competitive activities, and to explore more attractive promotional tools to optimize customer response. (Figure 3)

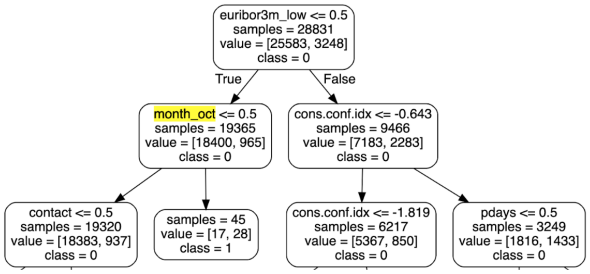

Figure 3. Decision tree of Bank data

Secondly, we observe that bank customers' response to events is highly correlated with the 3-month EURIBOR rate and consumer confidence, demonstrating the impact of the economic environment on customer behaviour. Lower EURIBOR rates typically promote more active participation, while higher consumer confidence reflects customers' optimistic expectations about the economy. Based on these indicators, banks can develop more targeted marketing strategies, increase advertising when economic conditions are favourable, and adjust strategies in time to cope with market volatility.

Finally, in our analysis of customer response, we found that different contact times (e.g., 'day of contact') had less impact on customer response rates. This finding suggests that marketing teams should not focus too much on the exact time of contact when developing their strategy, but rather focus on other more influential variables. For example, contact type and pre-campaign outcome ('poutcome') showed a significant association with customer response.

7.2. Fashion store dataset

The diversity of purchases plays an important role in determining the likelihood of customer responses to email marketing campaigns. It was found that respondents purchased a significantly larger variety of products (variable 'CLASSES') than non-respondents. (Appendix, Table X). It is recommended that diverse product recommendations need to be highlighted in the marketing strategy to attract this group of customers.

Moreover, based on the Logistic Regression, it was revealed that the coefficients for 'Lifetime average of days between visits' (LTFREDAY) and 'Average amount spent per visit' (AVRG) are negative (Appendix, Figure X), indicating that customers with lower frequency of visits or lower average amount spent are less likely to respond to the advertisement. This may imply that the purchasing habits of such customers do not fit well with the promotions or offers promoted in the email adverts, resulting in a lower likelihood of response.

However, the positive interaction coefficients for 'Total net sales' (MON) and 'Lifetime average of days between visits' (LTFREDAY) (Appendix, Figure X) suggest that in certain circumstances, response to an advert may increase if the customer's overall spending is high despite a low frequency of visits. It seems likely that high-value but low-frequency customers may still be motivated to respond to highly relevant promotions when they receive them. Therefore, optimizing email contents or adding personalized promotions for high-value but low-frequency customers might deliver better results.

Overall, the success of advertising depends on a combination of external and internal factors. On the one hand, advertising is more likely to be successful in a favorable economic environment; on the other hand, we need to focus not only on whether the customer responds to the advertisement but also on the value created after the response. In the case of fashion store data, insights from the analysis suggest that customers who purchase a diverse range of products are likely to be potential email marketing responders. Meanwhile, low-frequency but high-value customers should be addressed, potentially through exclusive, personalized promotional campaigns.

Therefore, it is recommended that clients should further segment different customer groups based on purchasing habits. Then, personalized campaigns can be tailored to increase the likelihood of customer engagement.

References

[1] Brownlee, J. (2021). *Stacking Ensemble Machine Learning With Python*. Machine Learning Mastery. https://machinelearningmastery.com/stacking-ensemble-machine-learning-with-python/

[2] Brownlee, J. (2019). *How to Configure the Learning Rate When Training Deep Learning Neural Networks*. Machine Learning Mastery.

https://machinelearningmastery.com/learning-rate-for-deep-learning-neural-networks/
[3] Domingos, P. (2012). *A few useful things to know about machine learning*. Communications of the ACM. https://homes.cs.washington.edu/ pedrod/papers/cacm12.pdf
[4] Friedman, J. H., & Popescu, B. E. (2008). *Predictive learning via rule ensembles. The Annals of Applied Statistics*, ResearchGate. https://www.researchgate.net/publication/23418298_Predictive_Learning_via_Rule_Ensembles

[5] Hachchan, A. (2023). *XGBoost: Everything you need to know*. Neptune.ai. https://neptune.ai/blog/xgboost-everything-you-need-to-know
[6] Hastie, T., Tibshirani, R., & Friedman, J., (2009). *The elements of Statistical learning: data mining, inference, and prediction*, Springer. https://link.springer.com/book/10.1007/978-0-387-84858-7
[7] Huynh, N. (2023). *Understanding loss functions for classification*. Medium. https://medium.com/@nghihuynh_37300/understanding-loss-functions-for-classification-81c19ee72c2a
[8] James, G., Witten, D., Hastie, T. & Tibshirani, R. (2021). *An introduction to Statistical learning) (Second Edition)*. New York: Springer.
[9] Narkhede, S. (2018). *understanding AUC-ROC Curve*. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5
[10] Ngai, E. W. T., & Wu, Y. (2022). Machine learning in marketing: A literature review, conceptual framework, and research agenda. Journal of Business Research, 145 https://doi.org/10.1016/j.jbusres.2022.02.049
[11] Molnar, C. (2020). *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/
[12] Ratner, B. (2009). The correlation coefficient: Its values range between +1/-1, or do they? Journal of Targeting, Measurement and Analysis for Marketing, 17(2), 139–142. https://doi.org/10.1057/jt.2009.5
[13] Terra, J. (2024). *What is a ROC curve and how to use it in performance modeling*. Simplilearn. https://www.simplilearn.com/what-is-a-roc-curve-and-how-to-use-it-in-performance-modeling-article
[14] Yeo, I. K., & Johnson, R. A. (2000). *A new family of power transformations to improve normality or symmetry*. JSTOR. https://www.jstor.org/stable/2673623
[15] Zollanvari, A. (2023). *Decision Trees. In: Machine Learning with Python*. Springer, Cham. https://doi.org/10.1007/978-3-031-33342-2_7

Appendix

1. Bank dataset

TABLE 7. COMPARISON OF MODEL PERFORMANCE METRICS

Model	Loss	Sensitivity	Specificity	Precision	AUC
Logit_L1	0.108	0.311	0.966	0.539	0.776
Logit_L2	0.107	0.375	0.959	0.534	0.803
KNN	0.106	0.369	0.960	0.540	0.803
Decision Tree	0.116	0.493	0.934	0.486	0.797
Bagging	0.137	0.448	0.916	0.402	0.753
CatBoost	0.106	0.418	0.955	0.540	0.803
LightGBM	0.111	0.413	0.950	0.511	0.790
Stacking: Logistic Regression and Random Forest	0.103	0.330	0.969	0.577	0.803
Stacking: Logistic Regression with L1 and XGBoost	0.102	0.348	0.968	0.578	0.804
Stacking: Logistic Regression with L2 and XGBoost	0.102	0.348	0.968	0.578	0.804

2. Fashion store Dataset

TABLE 8. MEAN VARIABLE COMPARISON: RESPONDERS VS. NON-RESPONDERS TO THE CAMPAIGN

Variable	RESP = 1 Mean	RESP = 0 Mean	% difference
FRE	10.928	3.896	180.464
MON	935.815	380.681	145.827
AMSPEND	31.758	10.724	196.140
PSSPEND	330.328	111.470	196.338
CCSPEND	519.537	240.505	116.020
AXSPEND	54.512	18.231	199.009
TMONSPEND	200.744	67.951	195.424
OMONSPEND	76.484	23.128	230.697
SMONSPEND	474.571	172.258	175.501
CLASSES	11.611	6.257	85.583

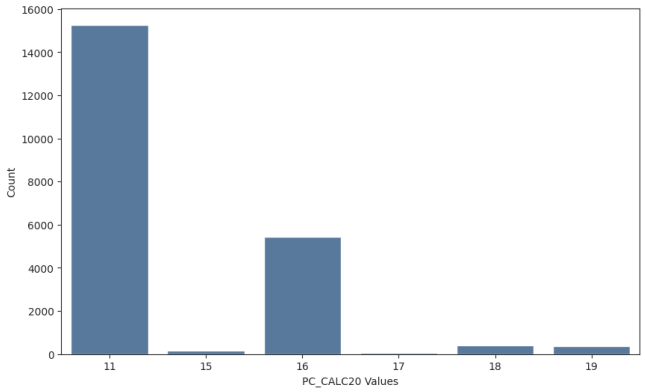

Figure 4. Distribution of PCCAL_20 variable

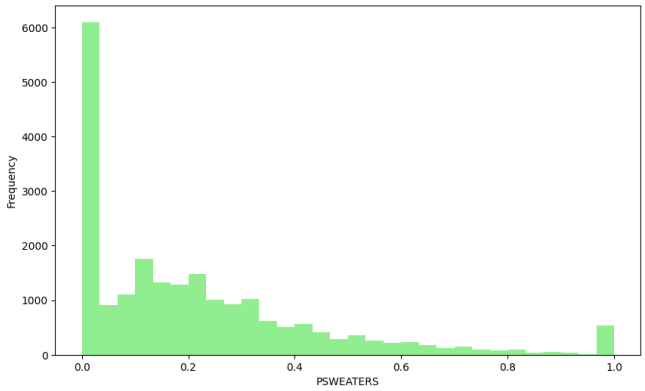

Figure 5. Distribution of PSWEATERS variable

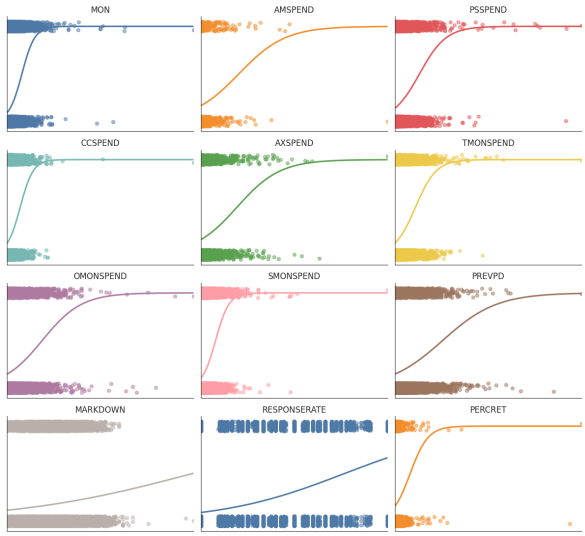

Figure 6. Continuous variables with potential relationships with the response variable

TABLE 9. VARIABLES WITH TOP 10 MI VALUES

Variable	MI
LTFREDAY	0.104
FRE	0.082
STYLES	0.073
FREDAYS	0.065
CLASSES	0.064
MON	0.056
RESPONSERATE	0.054
RESPONDED	0.054
SMONSPEND	0.052
COUPONS	0.046

TABLE 10. COMPARISON OF MODEL PERFORMANCE METRICS

Model	**Loss**	**Sensitivity**	**Specificity**	**Precision**	**AUC**
KNN	0.211	0.523	0.842	0.398	0.786
Random Forests	0.195	0.634	0.839	0.440	0.844
Bagging	0.207	0.655	0.821	0.422	0.838
XGBoost	0.193	0.633	0.841	0.442	0.849
LightGBM	0.172	0.530	0.887	0.483	0.842
CatBoost	0.190	0.594	0.854	0.447	0.842
Stacking: Logistic Regression and Random forest	0.168	0.560	0.886	0.495	0.851
Stacking: Logistic Regression and Bagging	0.168	0.570	0.885	0.496	0.852
Stacking: Logistic Regression and XGBoost	0.167	0.567	0.885	0.496	0.852
Stacking: Logistic Regression and LightGBM	0.169	0.570	0.883	0.492	0.852
Stacking: Logistic Regression and CatBoost	0.167	0.570	0.885	0.497	0.852

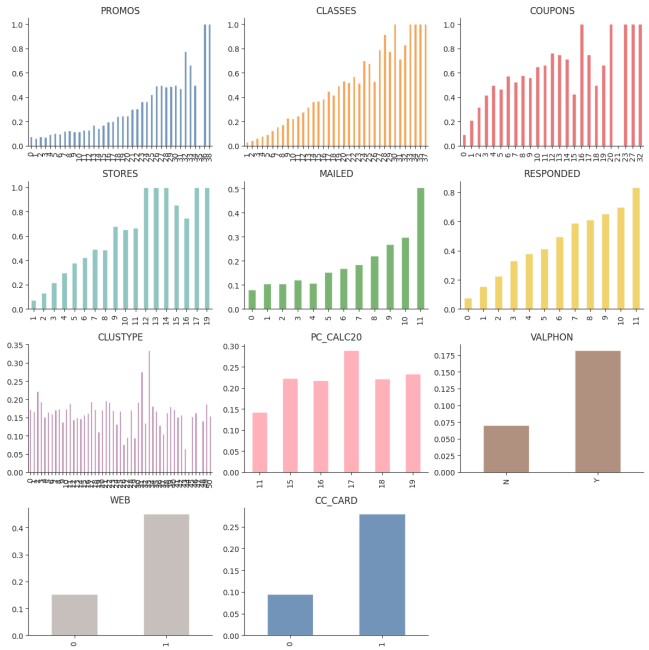

Figure 7. Distribution of discrete and binary variables among customers who respond to the campaign

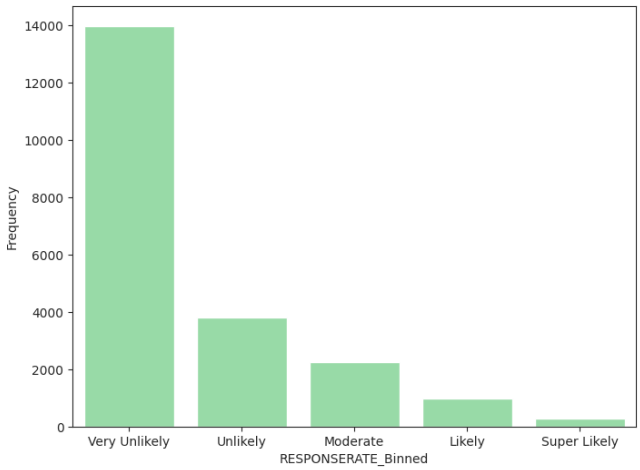

Figure 9. Binned RESPONSERATE variable

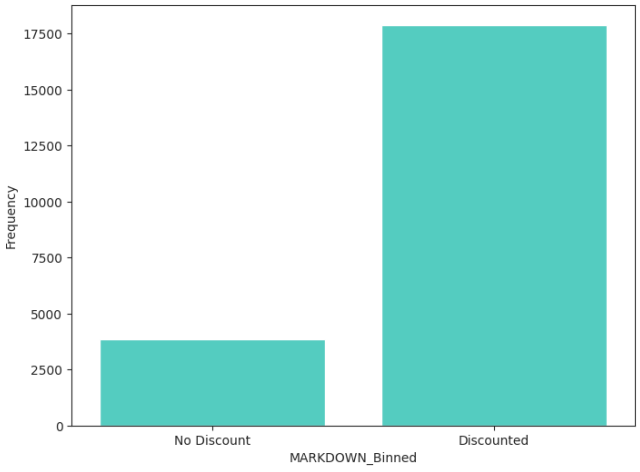

Figure 8. Binned MARKDOWN variable

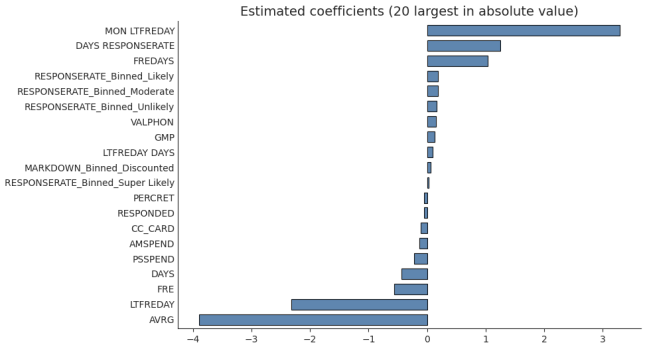

Figure 10. Estimated coefficients from Logistic Regression Model