DSCI 4411: Fundamentals of Data Mining - Dr. Seif Eldawlatly

Data Mining Project Report

Bank Marketing Dataset

Association Rule Mining and Classification Analysis

Farida Simaika - 900201753

Habeeba Hossam - 900191525

Katia Gabriel - 900202272

Table of Contents

Introduction

Data mining has gained increased popularity with the availability of huge amounts of information and data in our digital age. Data mining encompasses a wide variety of tools and techniques that are responsible for extracting important insights.The essence of data mining relies on its ability to discover underlying patterns and trends that the expert-eye fails to capture. Accordingly, this field empowers businesses as it transforms raw data into knowledge that helps in effectively making informed and data-driven decisions. It is therefore imperative that companies leverage data mining techniques to reach their full potential and exploit potential opportunities. Data mining tasks include classification, clustering, association rule mining and anomaly detection that should be conducted based on the nature of the dataset at hand and the objective of the analysis.

In the following report data mining techniques will be applied to help the bank marketing sector. Using the bank marketing dataset, the data mining tools will help in empowering the marketing segment in elevating their efficiency and effectiveness in the workplace.

Data Description

The Bank Marketing data set contains some of the most relevant information about the outcomes of direct marketing campaigns of a Portuguese banking institution. These campaigns relied on direct marketing by communicating their product to their target segment through direct phone calls and captured the characteristics of their contacted clients and their previous behavior in the given dataset. Accordingly, the dataset contains 16 predictor variables and encompasses 45,211 observations. The target variable, which is the variable to be predicted, is the "y" variable, which indicates whether the contacted client agreed to subscribe to a term deposit or not. In the following, a detailed description of all the features present in the dataset will be provided. The data set has been retrieved from the following source: https://data.world/data-society/bank-marketing-data.

Variable	Description
age	Numeric variable that captures the age of the respective contacted client during the marketing campaign.
job	Categorical variable that provides information about the types of jobs that contacted clients have. The variable has 11 job types which are: admin, blue-collar, entrepreneur, housemaid management, retired, self-employed, services, student, technician , unemployed & unknown
marital	Categorical variable that gives information about the marital status of the contacted

	clients. The variable captures 4 statuses, which are:divorced, married, single & unknown. The divorced status stands for both divorced or widowed.
education	Categorical variable that describes the education level of the contacted clients. The variable captures 4 classes: secondary, tertiary, primary & unknown.
default	Categorical variable that indicates whether the client has credit in default. The variable has 3 classes: “yes” if the client has credit in default, and “no” if the client has no credit in default.
balance	Numerical variable that provides information about the average yearly balance of the contacted clients in euros.
housing	Categorical variable that indicates whether the client has a housing loan or not. The variable has 3 classes, with “yes” if the client has a housing loan, “no” if the client does not rely on a housing loan.
loan	Categorical variable that provides information to the bank institution about whether the contacted client has a personal loan or not. The variable has 3 classes: “yes” if the client has credit in default, “no” if the client has no credit in default.
contact	Categorical variable that indicates the mean of communication that was used to contact the clients. The variable has 3 levels: cellular, telephone and unknown.
month	Categorical variable that states the month in which the client was last contacted. The variable has 12 classes, one for every month of the year.
day	Numerical variable that indicates the last day of the month the client was last contacted. This variable captures only the weekdays and disregards weekend days.
duration	Numerical variable that states the duration of the last call of the respective client. The

	duration is given in seconds.
campaign	Numerical variable that indicates the number of contact made during this campaign for each client.
pdays	Numerical variable that describes the number of days that have already passed by after the client was last contacted from the previous campaign. If the client was not previously contacted then it is given the value “-1”.
previous	Numerical variable that indicates the number of contacts made before this campaign for each client.
poutcome	Categorical variable that states the outcome of previous marketing campaigns of the respective contacted clients. The variable has 4 classes: success, failure, others and unknown.
y	Categorical variable that indicates whether the contacted client has accepted to subscribe to a term deposit: “yes” if the client has accepted the term deposit, “no” if the client has not accepted the term deposit.

Problem Identification

Data mining plays an important role in extracting meaningful insights from data. In this project, data mining techniques were leveraged to uncover hidden patterns. These hidden patterns might not be apparent through conventional analysis methods. Two key data mining techniques were applied on this dataset: Predictive analysis (through classification) and Association Rule Discovery (through the Apriori algorithm).

Predictive Analysis techniques (classification algorithms) were employed to enable the bank to predict client behavior accurately. Various classification models were tested on the dataset. Their performance was evaluated using diverse metrics. In this dataset, predictive analysis aids in determining the likelihood of a client subscribing to a term deposit. This task will help the bank in identifying the customers who are likely to accept the term deposits.

Association Rule Mining provides another dimension to the analysis. The Apriori algorithm identifies relationships between client attributes and their responses to marketing campaigns. By uncovering these association rules, the bank will gain insights into the combinations of client characteristics that correlate with positive outcomes (subscription to term deposits). This knowledge will enable the bank to optimize resources. They will focus efforts on clients with specific attributes that are more likely to result in successful term deposits subscriptions.

Data mining tasks such as predictive analysis and association rule mining help banks make better strategic decisions. This comprehensive analysis enables banks to tailor marketing strategies to target customers that are more likely to yield positive outcomes.

Methodology

The project methodology follows six main procedures, in order to come up with a comprehensive analysis. These parts encompass data exploration, data preparation and preprocessing, the development of classification models, the evaluation of these models, the development of the apriori algorithm and its evaluation. All previously mentioned steps will be executed using Python Programming Language.

The data exploration phase will examine all variables in the dataset to obtain a preliminary understanding of each variable and identify underlying patterns or trends that could require further investigation in the course of the analysis. Various descriptive statistics of numerical variables will be performed to capture the essence and characteristics of each variable. During the exploration part some of the issues in the respective variables could rise up, which is necessary to identify to handle in the preprocessing phase.

As a next step the dataset will undergo a data preprocessing procedure to meet the requirements of both classification models and association rule mining and ensure accurate and reliable results. This phase will recognize the existence of missing data as well as deal with inconsistent data input, variables with no explanatory power and unknown values. Moreover, various encoding, discretization and scaling techniques will be applied to deal with categorical variables and numerical variables, respectively. After this phase the data should be ready to be inputted into the respective predictive and association rule models.

Following the extensive data preprocessing phase, the classification model development phase comes into play. Before developing the models, the dataset will be split into a training set

(80%) and testing set (20%). This phase will rely on the development of various classification models, which are kNN, Logistic Regression, Naive Bayes, Decision Trees, Random Forest, Perceptron, Multiple Layer Perceptron, and SVMs, in an attempt to accurately predict whether a specific client will subscribe to the bank term deposit or not. Each model will be hypertuned to maximize the predictive results and will further be evaluated through different performance metrics to assess the performance of the model and be able to compare it with all other models. This comparison will recognize the best performing model in accurately predicting potential clients that will subscribe to the term deposit.

Lastly, the discretized dataset will be inputted to the apriori algorithm. As a first step both the support and the confidence thresholds will be determined. These thresholds will have a direct impact on the frequent itemset and association rule generation, which will be varied across the analysis to try and capture the most relevant rules that help in determining the target variable. As a last step the generated rules will be examined to determine their relevance and capture the underlying patterns detected by apriori.

Exploratory Data Analysis

The exploratory data analysis (EDA) phase involved an examination of each feature within the dataset. This section of the report combines the key highlights derived from the exploration of individual features.

Age

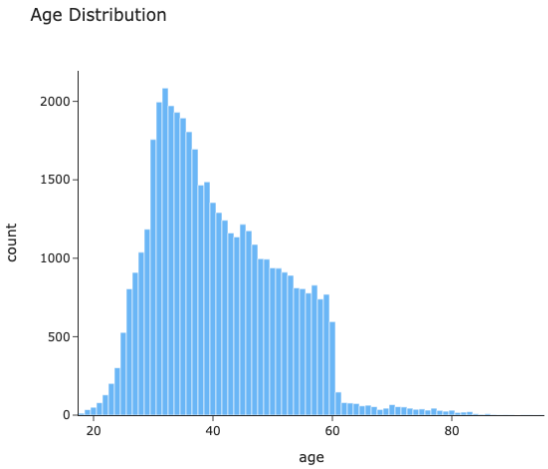


Figure 1

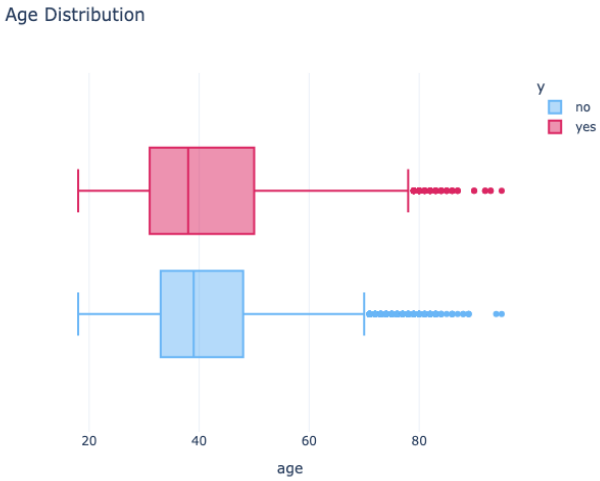


Figure 2

The age variable shows the age range of the contacted clients and whether this variable has a direct impact on the successful subscription to the term deposit or not. The statistical distribution of the age variable is left skewed, with the majority of data points lying between the ages 30 and 40 (Figure 1). The minimum age captured in the contacted clients is 18 years, while the maximum age is 95 years. The median age of clients that did not subscribe to the bank term deposit is 39, while the age of those who subscribed to the bank term deposit is 38.

Job

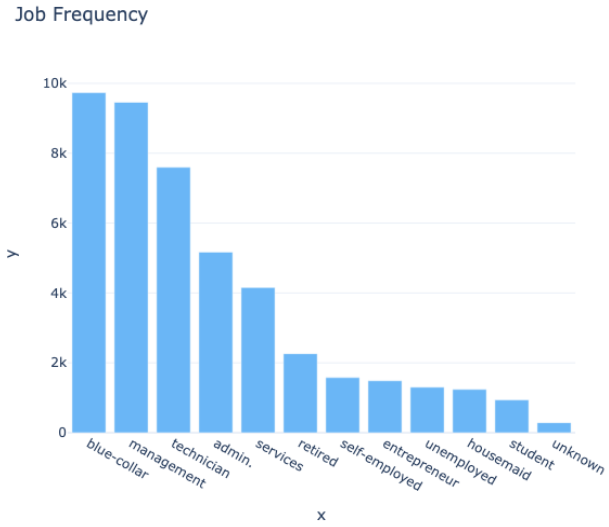


Figure 3

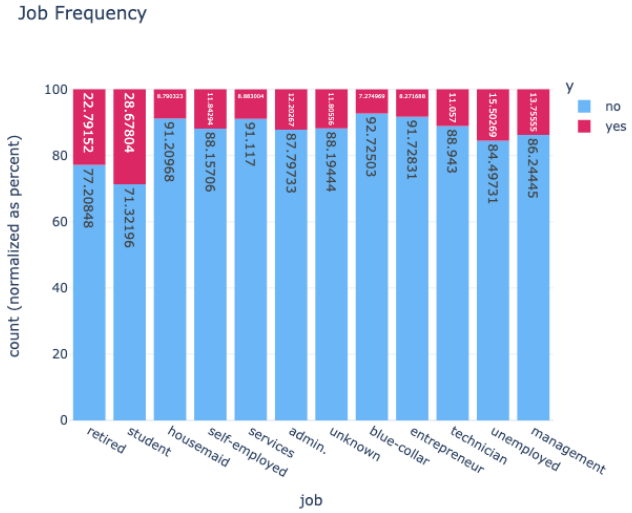


Figure 4

This feature variable captures the different job types that contacted clients have in an attempt to assess the client's behavior based on their job type. The most frequent job types are blue-collar, followed by management and technicians. The least frequent job types are unemployed, housemaids and students (Figure 3). It is evident that students are most likely to subscribe to bank term deposits, followed by retired clients and unemployed clients (Figure 4).

Marital Status

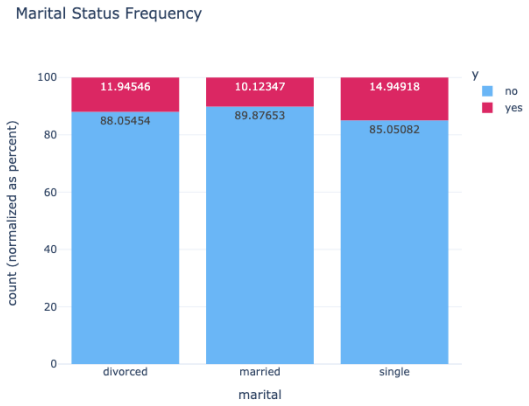


Figure 5

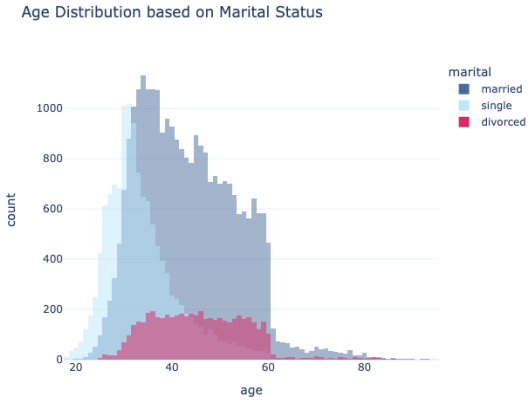


Figure 6

This variable shows the marital status of the bank's client base. It becomes clear that the majority of the clients are married, while divorced or widowed clients represent a minority. Out of the three marital statuses captured in the dataset, single clients are the most frequent clients that subscribe to the bank term deposits, followed by divorced and married clients. Single clients

range between the ages 25 and 35, while the married and divorced are spread out over the whole age range. This suggests that the younger generation are more likely to subscribe to the term deposit.

Education

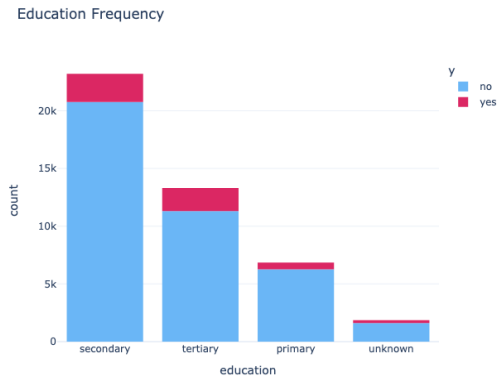


Figure 7

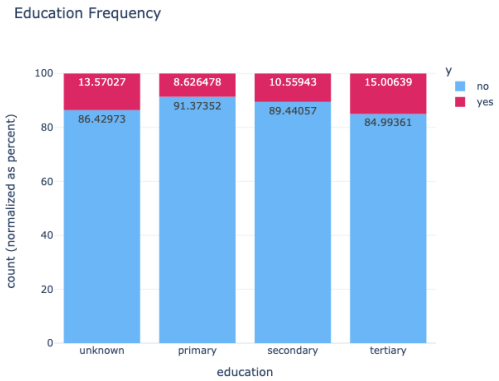


Figure 8

The education variable presents the education levels of the various clients. The most prevalent education level among the clients is secondary education, followed by tertiary and primary education. Among the various education levels, it becomes clear that clients with tertiary education are more likely to accept the bank terms deposit subscription. The second most frequent education level is unknown, which is a bit problematic and needs further investigations.

Default

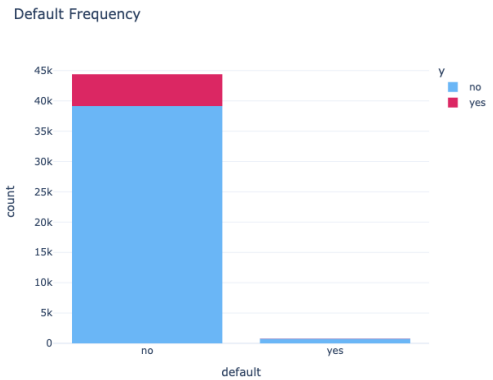


Figure 9

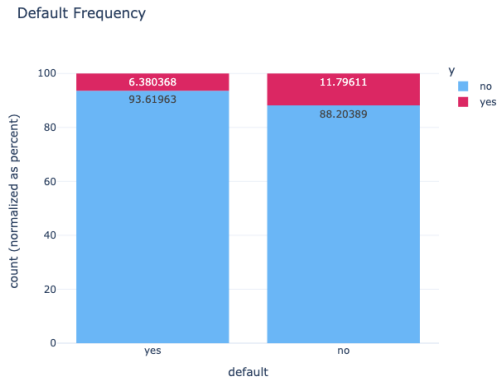


Figure 10

The default variable tries to indicate whether the failure of clients in repaying their loans has a direct impact on their behavior concerning the subscription to bank term deposits. It is

evident that the majority of clients do not have credit in default. Moreover, clients who do not have credit in default are more likely to accept the term deposit.

Balance

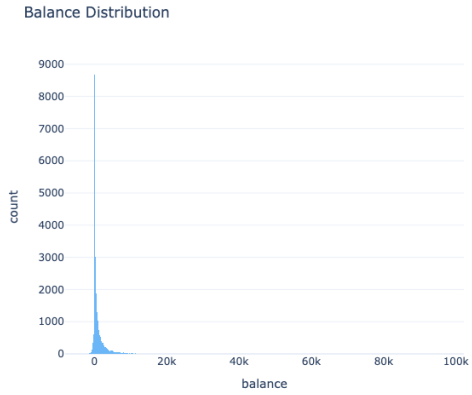


Figure 11

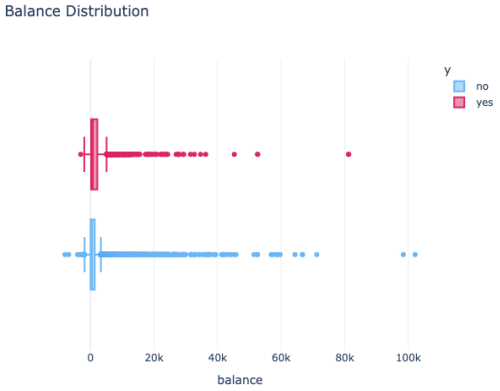


Figure 12

This feature variable indicates whether clients that have higher average yearly balance tend to accept the term deposits or not. It appears from the statistical distribution of the variable that the data is highly left skewed, with the majority of data points lying between 72 and 1428 euros. The minimum balance is -8019 euros, while the maximum balance is 102,127 euros. The median balance of clients that did not subscribe to the bank term deposit is 417, while the balance of those who subscribed to the bank term deposit is 733. This might indicate that clients with higher average yearly balance are more likely to subscribe to the term deposit.

Housing

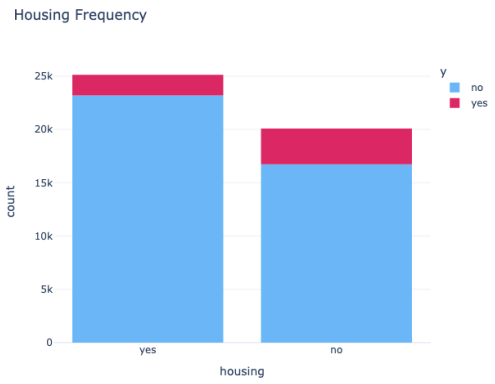


Figure 13

Figure 14

This variable indicates that the majority of the bank's client base rely on housing loans. However, clients who do not have housing loans are more likely to respond to the campaign.

Loan

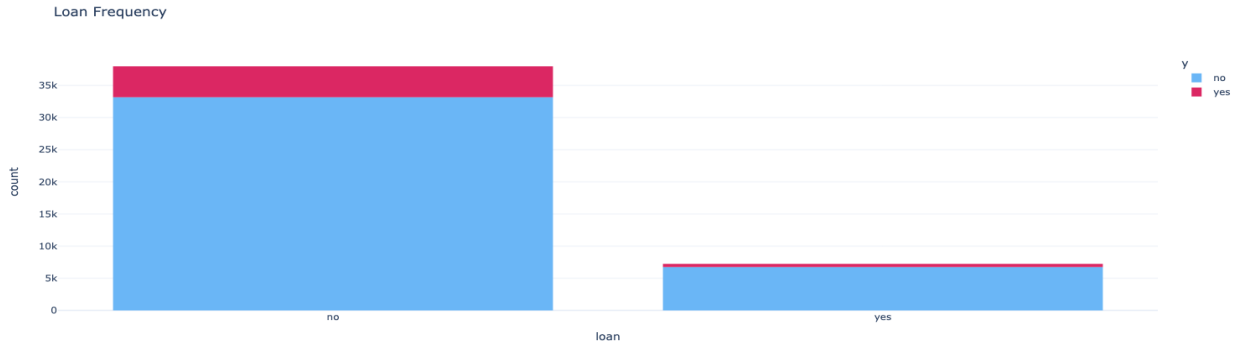


Figure 15

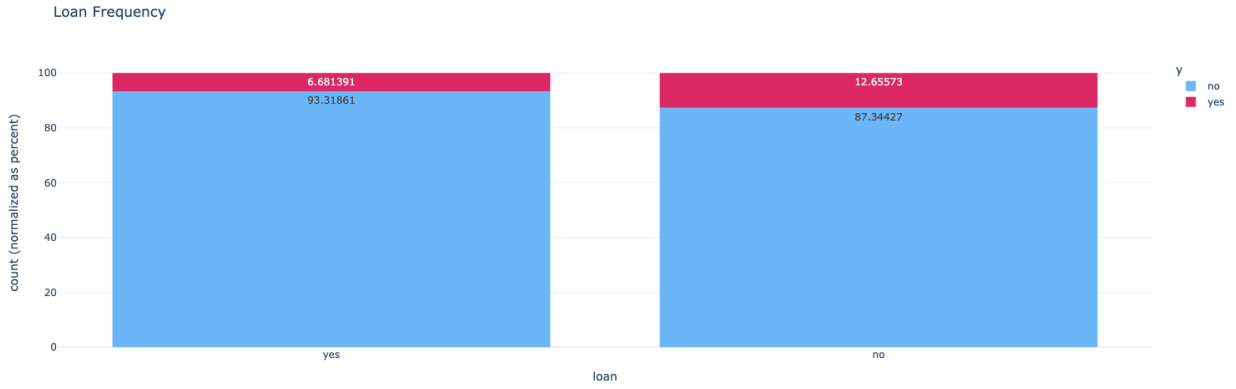


Figure 16

There are 37,967 instances of 'loan_no' and 6,760 instances of 'loan_yes'. For customers with loans, 6% accepted the term deposit, while nearly 12% of customers without loans accepted the term deposit offering.

Contact

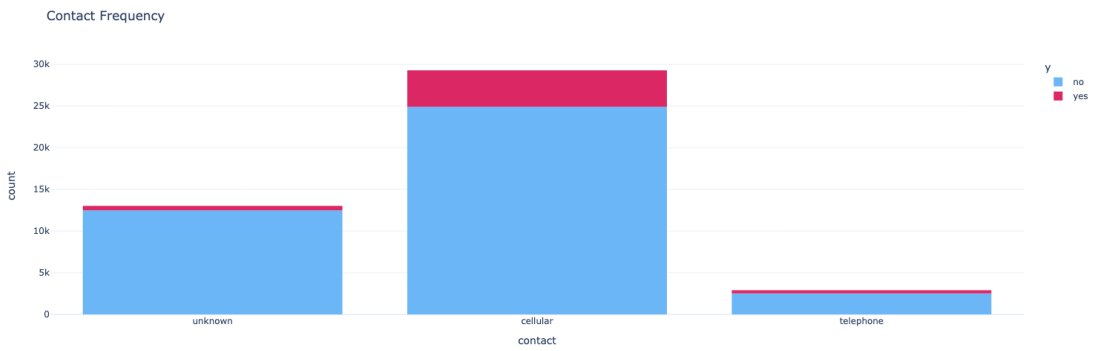


Figure 17

The 'contact' variable includes 29,285 instances of 'cellular,' 13,020 instances of 'unknown,' and 2,906 instances of 'telephone.' About 14% of those contacted via 'cellular' accepted the marketing offer.

Duration

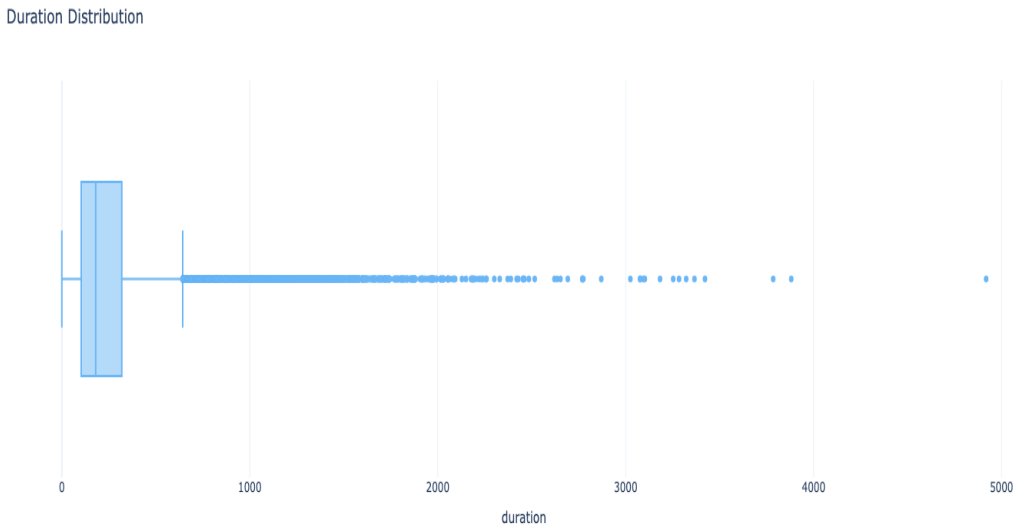


Figure 18

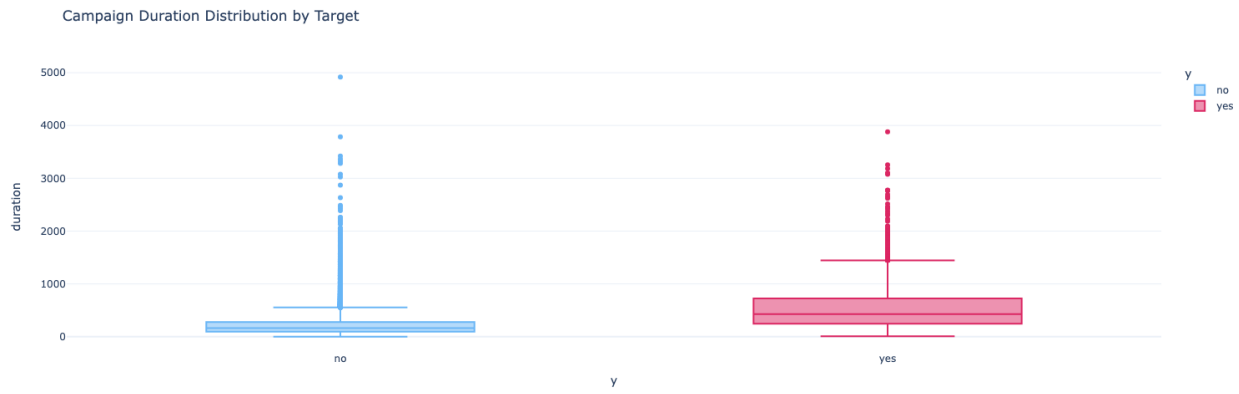


Figure 19

The average call duration is approximately 258 seconds (4.3 minutes), with a minimum duration of 0 seconds and a maximum duration of 4,918 seconds (1 hour and 21 minutes). The median call duration for clients who didn't accept the offer is 164 seconds (2.5 minutes). Those who accepted had a median call duration of 426 seconds (7.1 minutes). This suggests that, on average, successful marketing campaigns have longer call durations.

Previous Outcomes

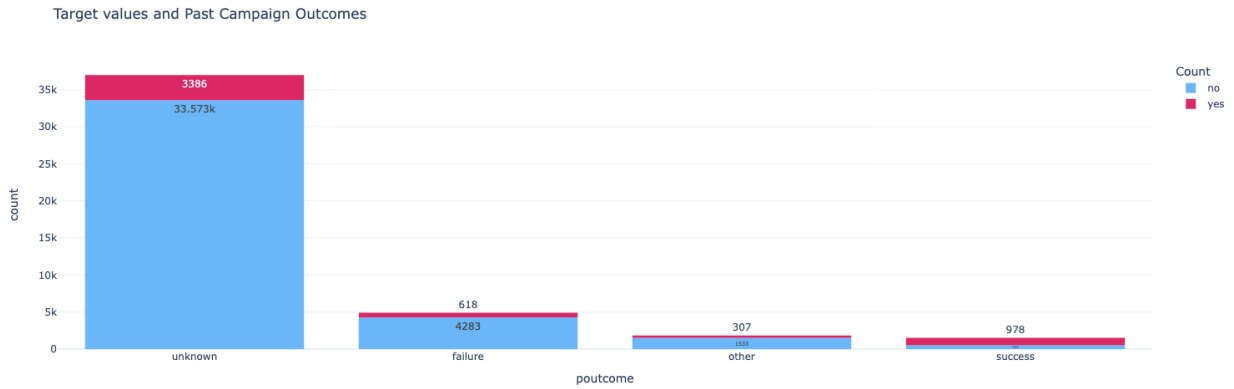


Figure 20

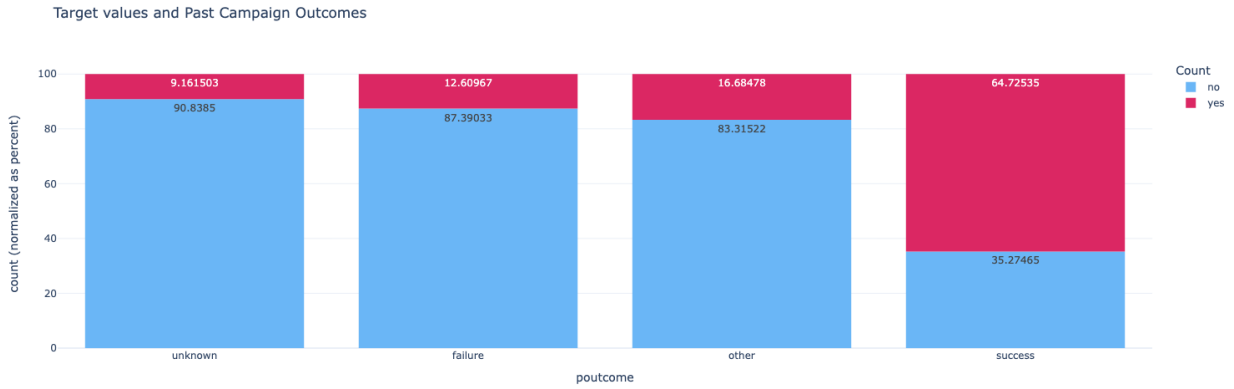


Figure 21

The 'poutcome' variable categorizes previous marketing outcomes into four groups: 'unknown,' 'failure,' 'other,' and 'success.' Among those who experienced successful past campaigns, 64% accepted the current marketing offer. The large number of 'unknown' values requires processing.

P-Days

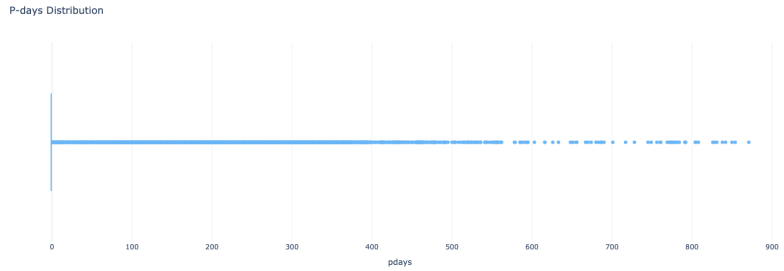


Figure 22

The average number of days since the last contact is approximately 40.2 days. The maximum duration since last contact reaches 871 days (almost 2 years). Notably, a significant portion of clients, as indicated by the 25th, 50th, and 75th percentiles have not been previously contacted (-1). Missing information about the last contact will require preprocessing.

Previous

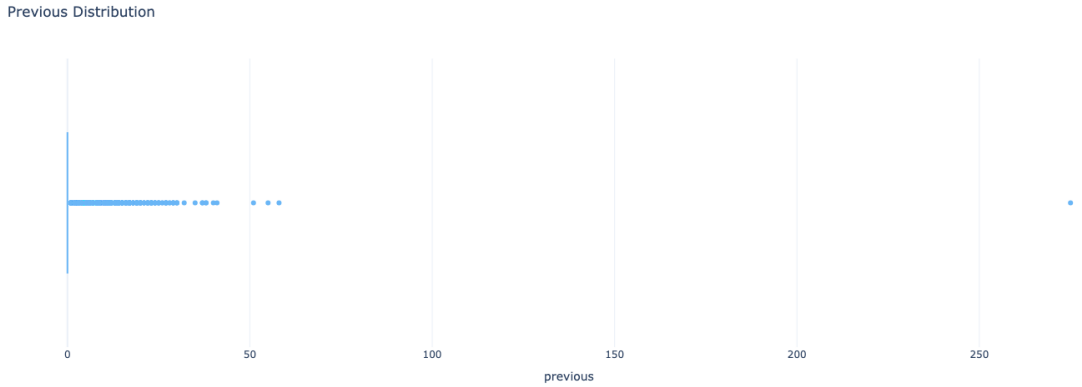


Figure 23

The average number of previous contacts is approximately 0. Half of the clients were not previously contacted. The maximum number of previous contacts reaches 275. Notably, the 25th, 50th, and 75th percentiles, all being 0, suggest that a considerable portion of clients had no previous contacts. This feature has a significant amount of missing information.

Month

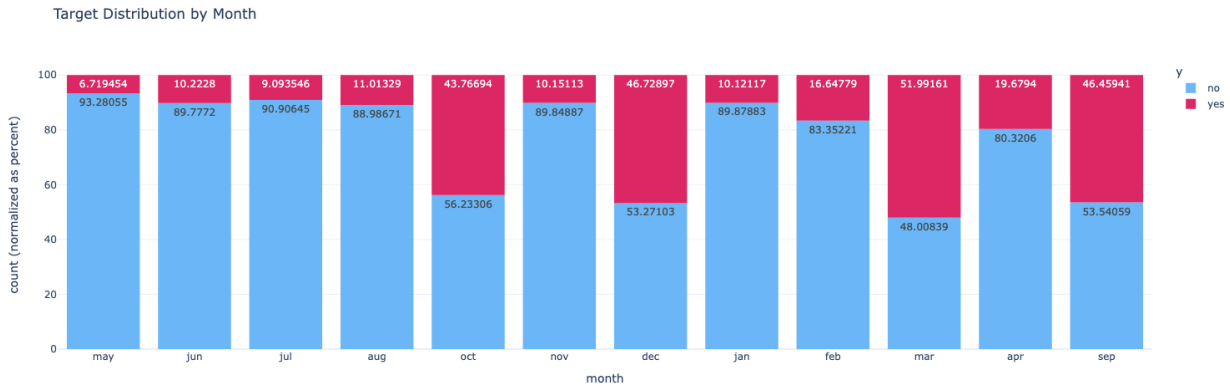


Figure 24

Acceptance and refusal rates of bank term deposits vary across different months. Months such as March, December and September had high acceptance rates. Almost half of the marketing campaigns conducted during the month of March were accepted by customers. On the other hand, the month of May had the lowest campaign acceptance rate of 6%.

Data Preprocessing

Before performing any predictive or associative analysis, it is essential to make sure that the data is clean and preprocessed. All variables were analyzed and dealt with to ensure that the data is of good quality and would yield reliable results. The preprocessing procedures were tailored to address the nature of both the predictive analysis and association rule mining parts.

Data Preprocessing addressing the Predictive Analysis Part

The dataset was first checked for the existence of missing values and it appeared that the data contains no missing values. However, when checking the unique values of each categorical variable, it became clear that some of the variables have "unknown" values. The variables that contain "unknown" values are: "job", "poutcome", and "education and were dealt with appropriately, as described in the following.

The "job" variable

The job variable contains 288 observations with "unknown" variables. Since the amount of "unknown" values is very small and constitutes around only 0.63% of all data points, these observations were dropped from the dataset.

The "poutcomes" variable

Almost 82% of the poutcomes variable contains "unknown" values. Accordingly, the outcomes variable has no explanatory power since it is mostly considered unknown. With that being said, the poutcomes column was dropped entirely.

The "education" variable

The education variable contains 1857 observations with "unknown" values. The "unknown" values constitute around 4.1% of all data points in the education variable. Accordingly, all observations with "unknown" values were dropped.

After dealing with the "unknown" values in the dataset, some of the columns were dropped such as the "previous", "campaign" and "duration" variables. The previous variable was

dropped because 90% of the variable is 0, meaning that 90% of the clients were not previously contacted before the current marketing campaign. Furthermore, the “campaign” variable was dropped because it contained inconsistent data points. Moreover, the data description suggested that the “duration” variable should be dropped while performing any predictive analysis because the duration of the call cannot be determined until the call, in which the client decides to subscribe to the term deposit or not, ends. Accordingly, the duration of the call is provided after the target variable is determined.

Handling Numerical Variables

As a next step the numerical attributes were normalized using the Minimum and Maximum scaling. This normalization brings all values within the range between 0 and 1 to offset the effect of varying scales and outliers that might significantly affect classification models that rely heavily on distance metrics.

Handling Categorical Variables

Encoding techniques were carried out to deal with categorical variables and transform them into numerical values. This is necessary to meet the requirements of various machine learning models that require the data to be numeric. Accordingly, the categorical attributes and symmetric attributes were converted into discrete "items". Each distinct attribute-value pair was transformed into a new binary item. For example, the nominal attribute 'Education' was replaced by binary items such as Education = primary, Education = secondary and Education = tertiary. One of the numerical features, which is “pdays” was converted into a categorical variable, giving it the value “0” if the person was not previously contacted and the value “1” if the client was contacted, regardless of the number of days since the client was last contacted from the previous campaign.

Handling Class Imbalance

The dataset had a severe class imbalance in the target variable (y_yes: 2,344 and y_no: 38,282). SMOTE (Synthetic Minority Oversampling Technique) was applied on the training data to oversample the minority class. This is necessary to ensure that the classification models are not biased towards the majority class while trying to minimize the error of prediction. In that way, both classes are equally represented in the dataset and makes the result of the classification models credible.

Data Preprocessing addressing the Apriori Association Rule Mining Part

Several preprocessing steps were taken to prepare the data for association rule mining. As mentioned above, the categorical and symmetric variables were also encoded, transforming each distinct attribute-value pair into a new binary item. This transformation was necessary to satisfy the requirements of association rule mining algorithms such as Apriori. Moreover, the class imbalance observed in the target variable was dealt with using SMOTE. This step was crucial as class imbalance affects the support levels of items. The minority class items might not achieve the required support threshold due to their scarcity. This can lead to their exclusion from frequent itemsets. In other words, imbalance can bias the association rules towards the majority class. The generated rules will predominantly reflect patterns present in the majority class. By addressing the class imbalance, we prevented skewed insights. As a result of applying SMOTE, the algorithm did not favor the majority class. Both classes were considered equally.

Lastly the preprocessing step that was done to respond specifically to the requirements of the apriori algorithm is the discretization of the continuous variables.

Discretization of Continuous Variables

Discretization was applied to convert continuous variables into categorical values. Continuous variables were discretized using the equal-width binning approach. Each continuous variable was divided into intervals of the same width across the range of values.

A lot of factors were involved to determine the number of bins for each continuous variable. First, the distribution of these variables was considered. In addition, the banking industry standards were also taken into account. For example, the 'Age' variable was categorized into four distinct bins (minors, young adults, middle-aged adults and seniors). Minors encompass individuals under 18 years old. The group of young adults usually ranges from late teens to early 30s. The middle-aged adults segment covers individuals in their 30s to 50s. Finally, seniors are individuals aged 60 and above. The choice of bins aligns with both the observed distribution of age values and industry standards within the banking sector. After discretization, the continuous variables underwent one-hot-encoding.

Apriori Association Rule Mining

In the context of bank marketing data, association rule mining holds significant importance. This data mining technique can help uncover relationships between customer attributes and their decision to subscribe to a term deposit. It can reveal patterns such as certain demographic groups being more inclined to subscribe. Banks can tailor their marketing strategies more effectively by identifying these association rules. This can result in increased customer satisfaction and loyalty. Customers will receive offers aligned with their preferences and needs. The Apriori algorithm was implemented to find the association rules.

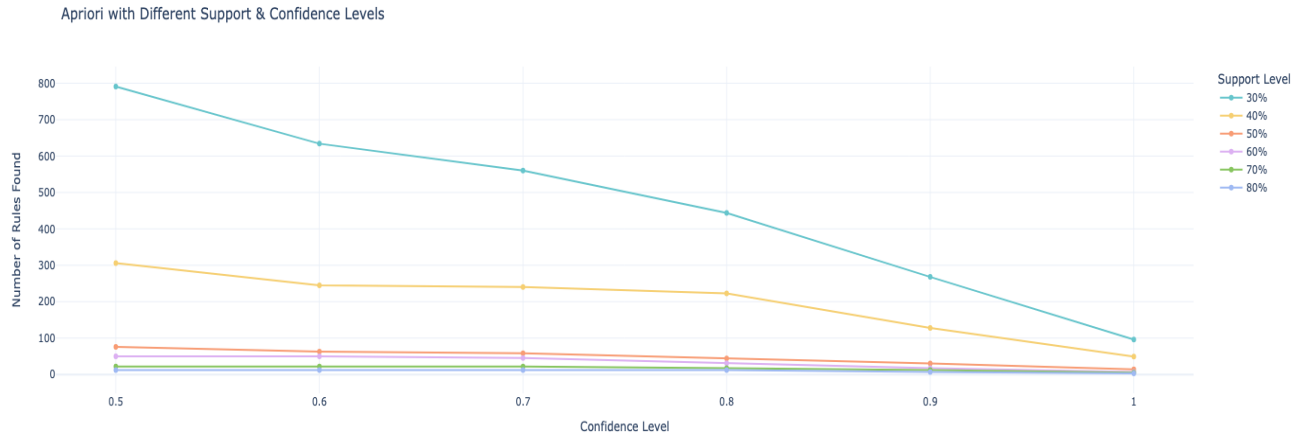
Applying the Apriori Algorithm

The Apriori algorithm was applied to the binarized data. The algorithm involves a two-step process. Firstly, it identifies frequent itemsets. Frequent itemsets are combinations of items that frequently occur together in the data. These itemsets meet a specified support threshold. The algorithm then proceeds to generate association rules based on these itemsets. The strength of an association rule can be measured in terms of its support and confidence. The

support determines how often a rule is applicable to a given dataset. The confidence determines how frequently items in Y appear in transactions that contain X.

Choosing support and confidence thresholds:

The Apriori algorithm generates association rules that fulfill both the support and confidence thresholds. To select the optimal thresholds, different visualizations were used.



The graph shows the relationship between confidence levels and the number of rules generated across various support thresholds. We have used this graph as a tool to help in determining the optimal combination of support and confidence thresholds. From the graph, a support and confidence values of 40% and 80% respectively seemed reasonable. Over 223 rules were generated using these thresholds.

Antecedent	Consequent	Support	Confidence
(young_adults)	(default_no)	0.5	0.98
(marital_married)	(loan_no)	0.54	0.98
(default_no,housing_no)	(loan_no)	0.51	0.87
(housing_no,	(default_no)	0.52	0.92

contact_cellular)			
(contact_cellular)	(default_no, loan_no)	0.52	0.83

From the rules, we can derive meaningful insights. Young adults tend to have a higher probability of not having defaults. Married individuals often correlate with not having housing loans and defaults. When there are no housing loans and defaults, there's a lower chance of having a loan. Bank representatives usually contact via cellular individuals with a low likelihood of defaults and having a loan.

Finding associations that include the target variable ‘y’ is the primary goal of our analysis.

No rules featured the target variable as a consequent when the support threshold surpassed 40%. Consequently, it was necessary to reduce the support and confidence thresholds to 40% and 50% respectively. These adjustments allowed us to generate reliable association rules that included our target variable.

Results:

Support and confidence thresholds of 40% and 50% yielded 99 frequent itemsets and 306 rules. Among these rules, 100 included the target variable as the consequent. This table showcases some of the association rules where the consequent includes the target variable.

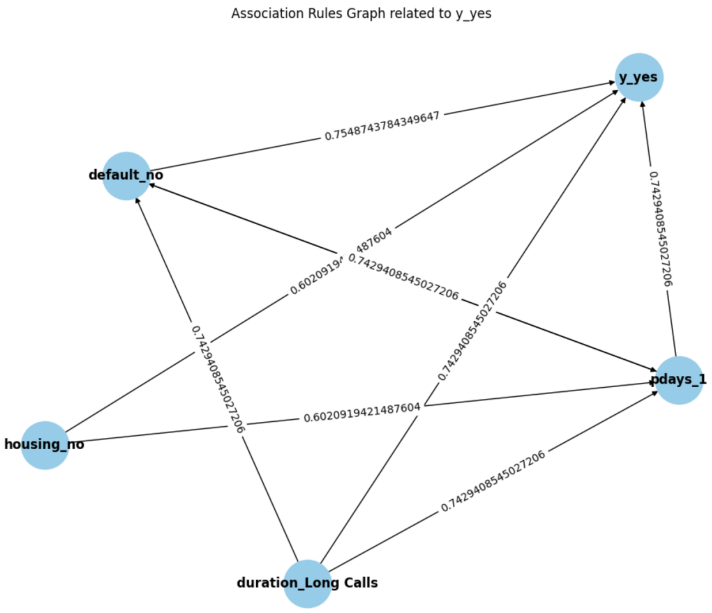
Antecedents	Consequents	Support	Confidence
(marital_married)	(y_no)	0.41	0.55
(marital_married, default_no)	(y_no)	0.53	0.55
(default_no, marital_married, pdays_1)	(y_no)	0.43	0.55

(housing_no, contact_cellular, pdays_1)	(y_yes)	0.40	0.51
(housing_no)	(y_yes)	0.45	0.52
(contact_cellular)	(y_yes)	0.41	0.56
(duration_LongCalls)	(y_yes)	042	0.59

The bank can potentially utilize these findings. Married individuals are inclined to refuse the term deposit. In addition, married individuals with no defaults tend to refuse the term deposit as well. The bank might tailor its marketing strategies differently for married individuals. Specific incentives and strategies should be explored to increase the success rate of marketing strategies among married couples. Short-duration contacts and marital status also align with a higher probability of refusing term deposits. Clients without defaults, contacted via cellular means and without housing loans are more inclined to accept term deposits. Moreover, individuals without housing loans tend to accept the offer. The bank should capitalize on these associations for effective marketing strategies.

To improve the reliability of our findings, the support and confidence thresholds were increased. By setting a support threshold of 30% and confidence level of 60%, twelve rules were found. These rules were exclusively associated with the target variable ‘y_yes’. Despite using SMOTE to balance the dataset, these rules did not include associations with ‘y_no’. The absence of associations with 'y_no' might suggest that the characteristics defining a negative response are not strongly tied to other features in the data.

Visualization of the association rules



This network graph illustrates relationships within bank marketing data variables. The edges denote these associations. The confidence of the rules is displayed on the edges. From the rules, we can derive meaningful insights. The table below presents a sample of the derived association rules obtained from the dataset using a support threshold of 30% and a confidence threshold of 60%.

Antecedents	Consequents	Support	Confidence
(duration_Long_Calls)	('y_yes')	0.31	0.75
(default_no, housing_no)	('y_yes')	0.32	0.61
(duration_Long Calls, pdays_1)	('y_yes')	0.31	0.75
(default_no, duration_Long_Calls)	('y_yes')	0.31	0.75
(default_no, housing_no, p_days1)	('y_yes')	0.3	0.62
(default_no, duration_Long Calls, pdays_1)	('y_yes')	0.3	0.76

The consistent appearance of 'default_no' within these rules suggests a significant trend. Individuals without previous defaults have a higher inclination towards accepting term deposit offers. The bank should try to capitalize on these findings and target these individuals. Clients who experience long marketing calls are also more likely to accept marketing campaigns. Moreover, when clients have long calls and had previous contact, they are also more likely to say yes to term deposits. Clients with no defaults that engage in long calls are more inclined to accept marketing offers. Finally, clients who were previously contacted and had no defaults or housing loans are likely to accept term deposits. Understanding these patterns help the bank focus its efforts on engaging these specific client profiles effectively

Classification Models

Experimental Set

Splitting the dataset into train and test sets

Before applying any of the classification models, the dataset was split with a ratio of 80:20. 80% of the dataset was used for training, while the remaining 20% were left for testing.

Parameters Hypertuning

Models with adjustable parameters, such as KNN, Logistic Regression, Decision Trees, Random Forest, Perceptron, and Multilayer Perceptron, underwent evaluation using 5-fold cross-validation. The chosen evaluation metric was accuracy, and the best parameters were selected for each classifier.

Confusion matrix

The confusion matrix design is structured with the real labels on the x-axis, indicating whether the customer accepted the campaign (1) or did not accept the campaign (0). On the y-axis, we have the predicted labels. Specifically, True Positive denotes instances where customers accepted the campaign, and the model correctly predicted so. True Negative

represents cases where the customer did not accept the campaign, and the model accurately predicted this. False Positive occurs when the model incorrectly predicted the customer would accept the campaign, but in reality, they did not. Lastly, False Negative indicates instances where the customer did accept the campaign, but the model predicted otherwise.

Evaluation Metrics

All specified metrics, including Accuracy, Precision, Recall, F1-Score, and False Negative Ratio, have been computed for each model. Notably, special attention is directed toward two key metrics. By focusing on maximizing Recall, we aim to capture as many positive instances as possible. Additionally, we emphasize the importance of minimizing the False Negative Ratio (FNR). This metric specifically addresses instances where actual positives are incorrectly predicted as negatives, crucially reducing the risk of overlooking customers who accept the campaign.

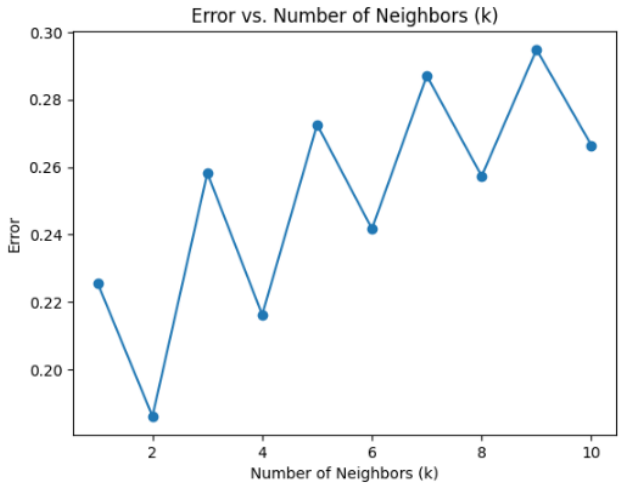
KNN

Model Overview

KNN makes predictions by identifying the k-nearest data points in the training set to a given input and then assigning the majority class.

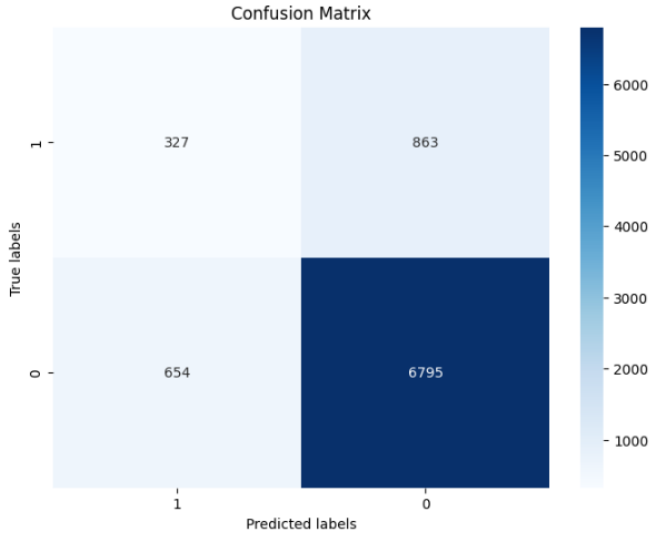
Model Details

To determine the optimal K value for the given dataset, we experimented with values ranging from 1 to 10, plotting the number of neighbors (K) against the error. As observed from the below line plot, the smallest error is at K = 2, which is used to fit a model with this parameter.



Results

The following confusion matrix provides an excellent overview of the classification results. The number of properly identified customers who accepted the campaign is 327, while those who refused it are 6,795. Those who rejected the campaign but were predicted as accepting it are 863. The most alarming statistic is that of customers who were predicted as refusing but actually accepted the service, with a count of 654. These represent missed opportunities for the bank.



The following machine learning evaluation metrics are considered: accuracy (0.8), precision (0.2), recall (0.30), F1-score (0.27), and false negative rate (0.6). Accuracy has a higher statistic, as it tends to yield higher scores in cases of class imbalance, which is evident in the dataset. It is noteworthy that both precision and recall have relatively low scores due to the fact that the number of customers accepting calls is a minority compared to the other data class.

Logistic Regression

Model Overview

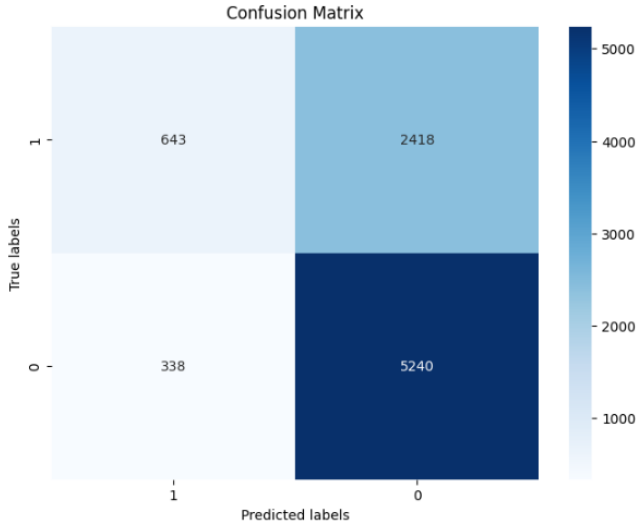
Logistic Regression aims to model the probability that an instance belongs to a particular class.

Model Details

After performing a grid search using cross-validation, the optimal hyperparameters for the Logistic Regression model were determined. The grid search considered different combinations of regularization strength (C) and penalty type. The best-performing model was found to have a regularization strength of (0.1) and utilized a (l2) penalty.

Results

The presented confusion matrix provides valuable insights into the classification results. Notably, the count of customers who accepted the campaign and were accurately predicted has almost doubled compared to the previous classifier. Additionally, it is noteworthy that the number of customers not interested, yet predicted as accepting the campaign, has significantly increased. Most notable, the count of interested customers who were misclassified decreased by nearly half, which translates into less lost chances. This detailed examination enhances our understanding of how the classes were classified and highlights the changes introduced by the new classifier.



In the context of machine learning evaluation metrics, the following key indicators provide valuable insights into the model's performance: Accuracy (0.6), Precision (0.2), Recall (0.6), F1-score (0.3), and False Negative Rate (FNR) (0.3). Notably, it is observed that enhancements have been made to Recall and False Negative Rate, in their respective definitions. Recall reflects the model's ability to capture positive instances effectively. A smaller FNR is preferred, as it indicates improved accuracy in terms of correctly identifying positive instances and reducing false negatives.

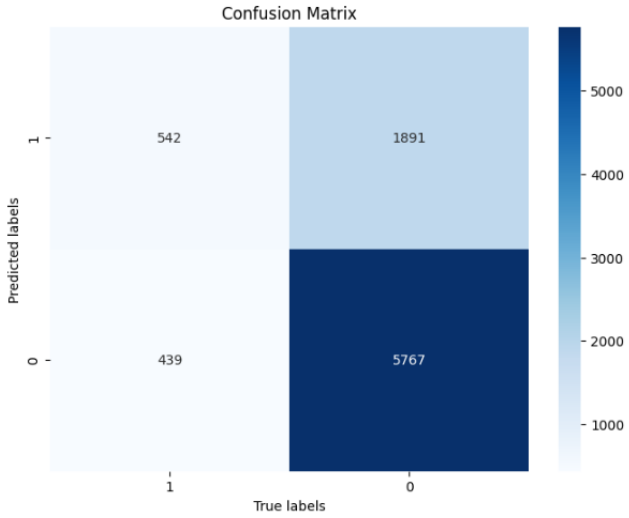
Naive Bayes

Model Overview

A Naive Bayes classifier is a probabilistic machine learning algorithm based on Bayes' theorem. The classifier makes the "naive" assumption of feature independence given the class.

Results

The confusion matrix below provides excellent insights into how the model classifies customers. As observed, the number of predicted customers accepting the campaign and behaving accordingly has decreased from the previous model, now standing at (542). Meanwhile, the majority group of customers refusing the campaign and predicted as such has increased to (5767). Once again, the most crucial category has seen an increased to (439), making it challenging for the bank to reach the utmost of customers who are, in fact, interested.



The machine learning metrics that reflect on the model's performance include Accuracy (0.7), Precision (0.2), Recall (0.5), F1-score (0.3), and False Negative Rate (FNR) (0.4). These statistics show little deviation from those obtained in the previous model. However, the subtle changes are not in favor of the crucial metrics for our specific problem.

Decision Trees

Model Overview

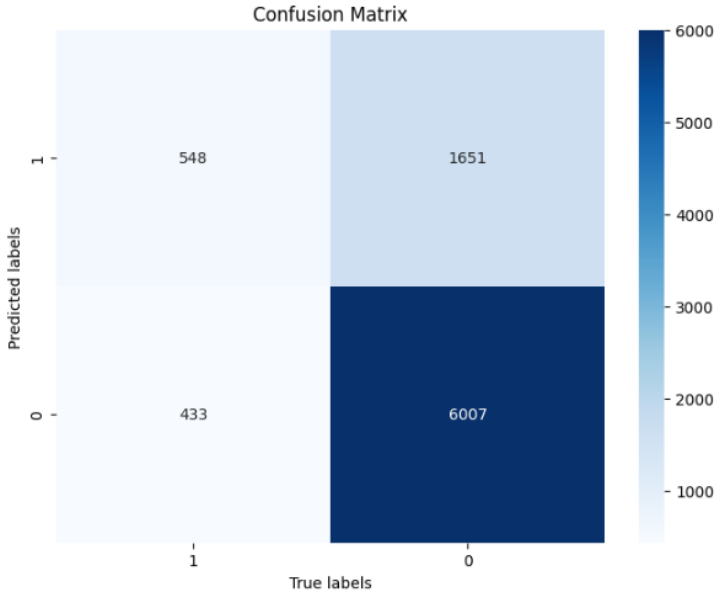
Decision Trees are part of a family of models that recursively partition the input space based on feature conditions to make predictions.

Model Details

In the process of fine-tuning the Decision Tree classifier, a grid search was conducted to explore various hyperparameter combinations. The hyperparameters considered include 'max_depth' with options (3, 5, 7, 10), 'min_samples_split' with choices (2, 5, 10), and 'min_samples_leaf' with values (1, 2, 4). The classifier's performance was assessed using 5-fold cross-validation, and the evaluation metric was accuracy. The Decision Tree model with the best performance was achieved with 'max_depth' set to (10), 'min_samples_split' set to (1), and 'min_samples_leaf' set to (20). This tuned model, determined through the grid search, represents the configuration that maximizes accuracy on the training data.

Results

The included confusion matrix highlights how the classifier predicts each class and the true labels. The figures show that there is no significant difference from the previously observed patterns of the last classifier. The number of customers who do not accept the campaign and were predicted as such still dominates among the four categories. It is followed by the count of customers who did not accept the campaign but were not predicted as such. Next is the number of customers who did accept the campaign and were predicted as such. Finally, the crucial statistics pertain to the number of accepting customers who are predicted otherwise.



Consistently evaluate the same machine learning performance evaluation metrics for comparability reasons: Accuracy (0.7), Precision (0.2), Recall (0.5), F1-score (0.3), and False Negative Rate (FNR) (0.4). As noticeable there is no change from the previous model.

Random Forest

Model Overview

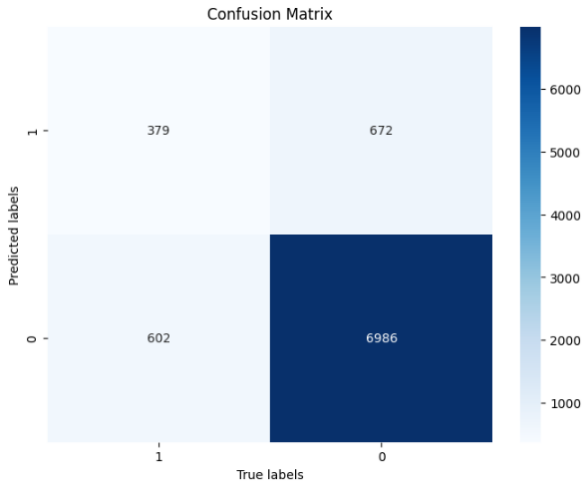
By combining the strengths of various trees and introducing randomness in both feature selection and data sampling, Random Forest proves to be a robust solution for classification problems. The ensemble nature of Random Forest mitigates overfitting by aggregating predictions from multiple decision trees, each trained on a random subset of features and samples. This approach enhances the model's ability to capture complex patterns and relationships within the data. The inherent diversity in the trees contributes to a more accurate and generalizable classifier, making Random Forest particularly effective in handling high-dimensional datasets and providing reliable predictions across various scenarios.

Model Details

Optimizing the performance of a Random Forest Classifier, a comprehensive grid search was conducted over a range of hyperparameter combinations. The parameter grid included variations in the number of trees in the forest ('n_estimators'), the maximum depth of the trees ('max_depth'), the minimum samples required to split a node ('min_samples_split'), and the minimum samples required at each leaf node ('min_samples_leaf'). By systematically exploring these configurations through cross-validated assessments, the grid search identified the set of hyperparameters that maximized the model's accuracy on the training data. The best-performing RandomForestClassifier was achieved with the following parameters: 'max_depth' set to None, 'min_samples_leaf' set to (1), 'min_samples_split' set to (2), and 'n_estimators' set to (100). This tuned model, as determined through the grid search, is expected to yield optimal results for the data at hand.

Results

The confusion matrix provides a valuable opportunity to understand the outcomes of the classifier. As observed, the distribution patterns remain consistent; however, a notable observation is that the number of customers predicted as accepting the campaign and are indeed interested, which is (379), is relatively low. Nonetheless, it still remains higher than the corresponding category in the KNN classifier, which recorded (327).



In terms of performance metrics, the following were evaluated: Accuracy (0.8), Precision (0.3), Recall (0.3), F1-score (0.3), and False Negative Rate (FNR) (0.6). When comparing and contrasting these performance metrics, it is notable that the low recall and high false negative rate are quite alarming.

Perceptron

Model Overview

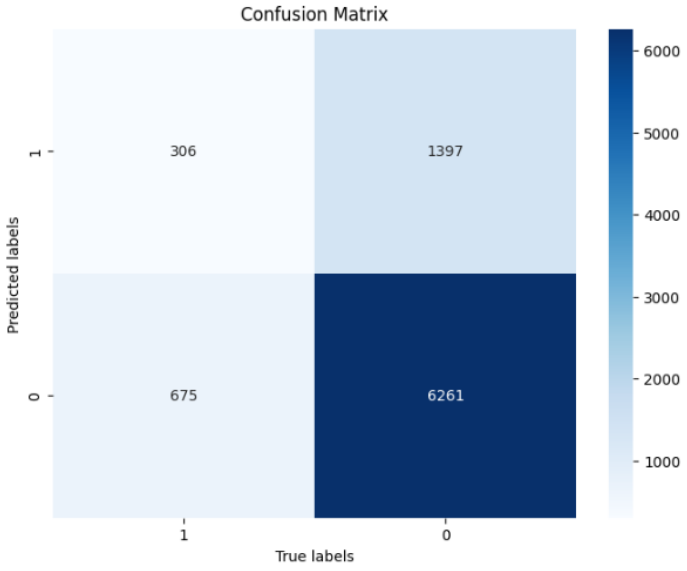
Perceptron is a simple, single-layer neural network used for binary classification. It takes input features, applies weights, sums them up, and passes the result through an activation function.

Model Details

With the aim of optimizing the performance of the perceptron model, a grid search was conducted over a range of hyperparameter combinations. The explored hyperparameters included the learning rate ('alpha') with options (0.0001, 0.001, 0.01), the maximum number of iterations ('max_iter') with choices (100, 500, 1000), and the initial learning rate ('eta0') with values (0.1, 0.01, 0.001). Through a systematic cross-validated assessment, the grid search identified the set of hyperparameters that maximized the model's performance on the training data. The best-performing perceptron was instantiated with these optimal parameters: 'alpha' set to 0.0001, 'eta0' set to 0.1, and 'max_iter' set to 100. This finely tuned perceptron was then trained on the training data, and its predictive capabilities were demonstrated on the test data.

Results

The below confusion matrix focuses on the distribution of predicted values vs true labels, revealing a consistent pattern for the perceptron model. Significantly, the number of customers who accept the campaign but are not predicted as such is extremely high. This count surpasses those observed in all previously mentioned classifiers and those yet to be discussed.



In terms of performance metrics, the following were evaluated: Accuracy (0.7), Precision (0.1), Recall (0.3), F1-score (0.2), and False Negative Rate (FNR) (0.6). The general pattern observed in these results is very similar to previously tried models. Notably, the impact of false negatives is evident in the remarkably high false negative rate.

Multiple Layer Perceptron

Model Overview

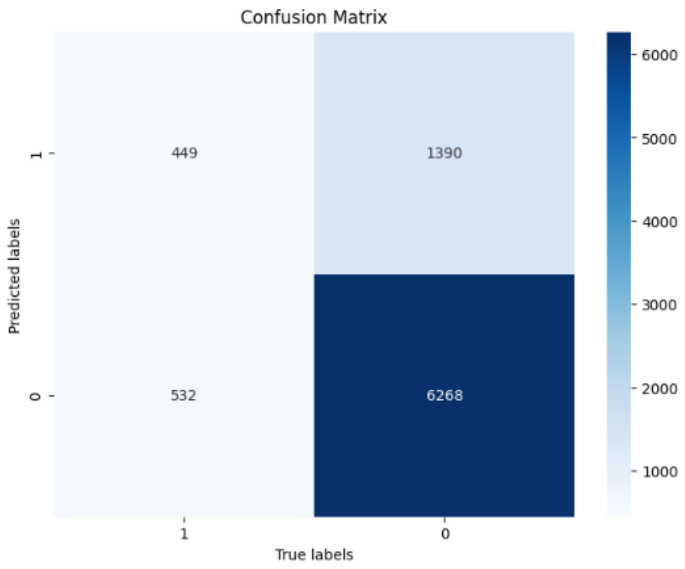
A Multiple Layer Perceptron (MLP) is a type of artificial neural network with multiple layers, including an input layer, one or more hidden layers, and an output layer. Each layer contains nodes (neurons) connected to nodes in the adjacent layers, and each connection has associated weights. Training involves adjusting the weights through backpropagation and optimization algorithms to minimize the error between predicted and actual outputs. The addition of hidden layers allows MLPs to model nonlinear relationships, enhancing their capability compared to single-layer perceptrons.

Model details

In the pursuit of optimizing the Multilayer Perceptron (MLP) classifier's performance, an extensive grid search was conducted over a range of hyperparameter combinations. The hyperparameters explored included different configurations for the hidden layers ('hidden_layer_sizes'), activation functions ('activation'), optimization solvers ('solver'), and the maximum number of iterations ('max_iter'). The grid search aimed to identify the set of hyperparameters that maximized the model's accuracy on the training data. The best-performing MLP model was found to have the following optimal parameters: 'activation' set to 'relu,' 'hidden_layer_sizes' set to (100, 100), 'max_iter' set to 100, and 'solver' set to 'adam'. This finely tuned model, as determined through the grid search, was then applied to make predictions on the test data, showcasing its improved performance for the classification task at hand.

Results

The following confusion matrix provides an excellent overview of the classification results. The number of properly identified customers who accepted the campaign is 449, while those who refused it are 6,268. Those who rejected the campaign but were predicted as accepting it are 1390. The most alarming statistic is that of customers who were predicted as refusing but actually accepted the service, with a count of 532. These represent missed opportunities for the bank.



The machine learning evaluation metrics taken into consideration include accuracy (0.7), precision (0.2), recall (0.4), F1-score (0.3), and false negative rate (0.5). The relatively high accuracy observed in the model is influenced by the data imbalance, where there are more instances of people rejecting the campaign than those accepting it, resulting in a high count of true negatives (TN).

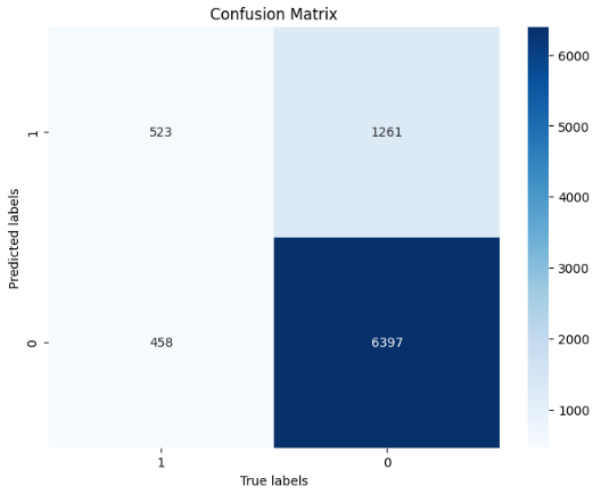
SVM

Model Overview

SVM can handle high-dimensional data effectively and is particularly useful in scenarios where the data is not linearly separable. SVM works by finding the optimal hyperplane that best separates different classes in the feature space. The hyperplane is determined by maximizing the margin, which is the distance between the nearest data points of the two classes. It employs a kernel trick to map the input features into a higher-dimensional space, allowing it to capture complex relationships and patterns.

Results

The below confusion matrix provides detailed insights into the distribution of class classifications and true labels. The observed pattern aligns with the common occurrence of the greatest count in the number of customers who did not accept the campaign and were predicted correctly (6397), followed by the second-largest count corresponding to people who refused the campaign but were not predicted accurately (1261). The third category encompasses those who accepted the campaign and were correctly predicted (523), while the fourth category comprises individuals predicted as not accepting but, in reality, accepted (458). This last group represents a potential risk, as it involves instances where the model may miss opportunities for the bank.



The machine learning evaluation metrics taken into consideration include accuracy (0.8), precision (0.2), recall (0.5), F1-score (0.3), and false negative rate (0.4).

Results

In the experimentation with six machine learning models, including KNN, Logistic Regression, Naive Bayes, Decision Tree, Random Forest, Perceptron, Multiple Layer Perceptron, and SVM, each model underwent evaluation using five key metrics: Accuracy, Precision, Recall, F1-score, and False Negative Ratio. Upon reviewing the results in the table, it is evident that Logistic Regression stands out as the model that maximizes Recall while concurrently

minimizing the False Negative Ratio. This indicates that Logistic Regression is particularly effective in capturing a high proportion of positive instances, crucial for scenarios where minimizing the risk of false negatives is important.

	Accuracy	*Precision*	*Recall*	*F1- score*	*False Negative Ratio*
KNN	0.82	0.27	0.33	0.30	0.66
Logistic Regression	0.68	0.21	**0.65**	0.31	**0.34**
Naive Bayes	0.73	0.22	0.55	0.31	0.44
Decision Tree	0.75	0.24	0.55	0.34	0.44
Random Forest	0.85	0.36	0.38	0.37	0.6
Perceptron	0.76	0.17	0.31	0.22	0.6
Multiple Layer Perceptron	0.77	0.24	0.45	0.31	0.54
SVM	0.89	0.60	0.17	0.24	0.4

Conclusion

Our exploration of the bank marketing dataset was aimed at two objectives. Our primary goal was to predict term deposits subscriptions. In addition, we wanted to gain insights into the combinations of client characteristics that correlate with term deposit subscriptions.

Our predictive analysis involved implementing various classifiers on the dataset. Our chosen metrics for evaluation were the recall and the FNR (False Negative Rate). The FNR and recall are crucial metrics because they specifically focus on correctly identifying customers who would accept the campaign but might be misclassified by the model. The FNR represents missed

opportunities for the bank. The recall calculates the model's ability to capture all positive instances correctly. Logistic regression emerged as the best model based on our chosen metrics. The model has effectively minimized False Negatives (FNR) and maximized the recall.

Apriori analysis uncovered impactful client attributes linked to term deposit acceptance. Clients without housing loans or previous defaults had a higher inclination towards accepting term deposits. Additionally, longer phone call durations often indicated a higher chance of term deposit acceptance. The bank should capitalize on these insights and refine its marketing strategies. The marketing team should target individuals with these specific attributes to potentially optimize the bank's resources.

Our analysis has provided predictive insights for the bank and uncovered hidden patterns. The bank should leverage these insights to create more impactful marketing campaigns.