WEST TEXAS A&M UNIVERSITY

Paul & Virginia Engler College of Business

CIDM 6355 FINAL PROJECT:

PREDICTING CUSTOMER ACCEPTANCE OF NEW TERM DEPOSIT THROUGH TELEMARKETING

Team 2: Data Miners

Chidinma Aniekwena

Chukwuebuka Dibie

Fougnigue Sefon

Mariam Adegbindin

Tokoni Forun

TABLE OF CONTENTS

I/ EXECUTIVE SUMMARY

This final project report encapsulates the culmination of an in-depth study aimed at predicting customer acceptance of new term deposit offers through telemarketing campaigns by leveraging data mining techniques. The study was conducted by Team 2, Data Miners, at West Texas A&M University, Paul & Virginia Engler College of Business, within the CIDM 6355 course.

At the heart of this investigation is the dataset sourced from the UC Irvine Machine Learning repository, which records customer responses to term deposit offers via phone calls, along with extensive demographic and behavioral attributes. The dataset was meticulously prepared, ensuring quality and usability through the elimination of redundancies and outliers, and addressing missing values with careful consideration.

Five predictive models were built using both R and RapidMiner: Decision Trees, Naïve Bayes, Logistic Regression, Gradient Boosting Machines (GBM), and Neural Networks. Each model was thoroughly evaluated based on accuracy, precision, recall, and F-measure to identify the model best suited to the objectives of telemarketing campaigns.

The Logistic Regression model in RapidMiner and Decision Trees in R emerged as top performers. These models not only demonstrate strong predictive power but also offer insights into customer behavior. This enhances the bank's ability to target potential subscribers efficiently, hence maximizing marketing ROI and reducing associated costs.

Deployment strategies for the models involve their integration into the current banking systems, continuous monitoring, and regular updates to adapt to changing customer patterns. The study, while comprehensive, acknowledges certain limitations such as the static nature of the historical data and the inherent challenges in interpreting complex models like Neural Networks.

In conclusion, the study underscores the transformative potential of data mining in the banking sector's marketing strategies. The results offer a strategic edge in an intensely competitive market. Future research directions include real-time data analysis and the exploration of hybrid models for enhanced predictive accuracy and interpretability. The study also opens up critical discussions on the ethical implications of data mining in maintaining customer trust and privacy.

II/ INTRODUCTION

From the early days of the banking sector, effective communication with potential customers has been crucial, with telemarketing emerging as a vital strategy for promoting banking products such as term deposits. Telemarketing, a form of direct marketing, involves contacting prospective clients via telephone to offer services or products, capitalizing on the personal touch of voice communication to enhance customer engagement and response rates. For banks, this technique has become increasingly important, serving not just to boost sales but also to establish and maintain customer relationships.

With telemarketing, banks, and financial institutions are tasked with the dual challenge of effectively engaging potential customers while understanding and adapting to their preferences and behaviors. In this competitive environment, accurately predicting customer behaviors

emerges as a crucial opportunity (Infobase, 2022). To seize this, we plan to employ data mining techniques aimed at predicting customer responses to new term deposit offers via telemarketing. This approach is part of our broader strategy to leverage data mining to drive business intelligence. We aim to unveil the underlying patterns that influence acceptance rates. This project not only helps in refining marketing approaches but also aligns with the broader goal of Data Miners: to discover meaningful patterns and rules and empower organizations with the knowledge to make informed decisions.

Our dataset was retrieved from the UC Irvine Machine Learning repository (https://archive.ics.uci.edu/dataset/222/bank+marketing). It contains data on customer responses to new term deposits via phone calls, along with comprehensive demographic and behavioral attributes. It provides a rich foundation for analyzing customer behavior towards bank term deposits.

To achieve our objective, we will undertake a data analytics process, examining the characteristics of customers who have either accepted or declined the telemarketing offer. Our initial step involves acquiring the dataset from the UC Irvine Machine Learning Repository and preparing it to ensure cleanliness and usability. We will scrutinize the dataset for highly correlated attributes, eliminating those deemed redundant, and conduct a thorough review of summary statistics to verify the absence of missing data and address any outliers. Our data will then be segmented using the holdout method, allocating 70% for training and the remainder for testing the models. We will employ various modeling techniques, including Decision Trees, Naïve Bayes, Logistic Regression, Gradient Boosting Machines, and Neural Networks, to construct our predictive framework. After developing these models, we will build a confusion matrix and assess their performance using critical metrics such as accuracy, recall, and precision. This comprehensive evaluation will enable us to discern the strengths and weaknesses of each model in accurately forecasting customer responses to telemarketing campaigns.

III/ DATA DESCRIPTION: EXPLORING THE VARIABLES

The Bank Marketing dataset, curated by Paulo Cortez and Sérgio Moro in 2012, serves as a valuable resource for understanding consumer behavior and marketing strategies in the banking sector. This dataset captures the outcomes of direct marketing campaigns conducted by a Portuguese banking institution between May 2008 and November 2010. Its primary objective is to predict whether clients will subscribe to a term deposit, providing insights into customer decision-making processes and enabling targeted marketing efforts.

Comprising a total of 45,211 records, the dataset offers a comprehensive view of various demographic and behavioral attributes that influence clients' subscription decisions. These attributes encompass a wide range of client characteristics, including age, occupation, marital status, and educational background. Additionally, the dataset includes financial indicators such as credit default status, average yearly balance, and loan status (both housing and personal loans), shedding light on clients' financial profiles and risk appetites.

Moreover, the dataset provides detailed information on the timing and nature of marketing interactions, including the type of contact, the day and month of the last contact, and the duration

of the contact in seconds. These attributes offer valuable insights into the effectiveness of different communication channels and the optimal timing for engaging with clients.

Furthermore, the dataset incorporates features related to the frequency and outcomes of previous marketing campaigns, such as the number of contacts made during the current campaign, the elapsed time since the last contact from a previous campaign, and the success or failure of past marketing efforts. Analyzing these attributes can help identify patterns in client responsiveness and refine marketing strategies accordingly.

At the core of the dataset lies the target variable 'y,' which indicates whether a client subscribed to a term deposit following the marketing campaign. Understanding the factors that influence this binary outcome is essential for optimizing marketing campaigns, improving customer engagement, and ultimately driving business growth.

Figure 1: Dataset Description

	Attribute	Attribute Type	Description
Demographic	Age	Numeric	Customer's age
	Marital Status	Categorical	marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
	Education	Categorical	(categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
	Job	Categorical	type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
Financial	Housing	Binary	has housing loan?
	loan	Binary	has personal loan?
	Credit (default)	Binary	has credit in default?
	Balance	Integer	average yearly balance
Communication	Month	Date	last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
	Duration	Integer	last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.
	Campaign	Integer	number of contacts performed during this campaign and for this client (numeric, includes last contact)
	Contact Frequency	Categorical	contact communication type (categorical: 'cellular','telephone')
	Day of week	Date	last contact day of the week

Communication Outcome	Pdays	Integer	number of days that passed by after the client was last contacted from a previous campaign (numeric; -1 means client was not previously contacted)
	Previous	Integer	number of contacts performed before this campaign and for this client
	Poutcome	Categorical	outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
	Y	Binary	has the client subscribed a term deposit?

IV/ DATA PREPARATION: ENHANCING THE DATA

Data preparation, often the most labor-intensive and time-consuming phase in data mining, is crucial for ensuring data quality before proceeding to model building, evaluation, and deployment. Our dataset comprised 17 variables, including one dependent variable (Y), which indicates customer subscription to our bank term deposits, and 16 independent variables. We prepared our data to be qualitative, understanding that subpar data quality would undermine the utility of our results. This phase involved careful checks for missing attributes, outliers, and other data inconsistencies.

Missing attributes

The dataset might appear to have no missing values at first glance. However, upon reviewing the summary statistics in RapidMiner, we identified that some variables were labeled as “unknown,” effectively representing missing values. These attributes include Job, Education, Poutcome, and Contact[1]. We considered various techniques for handling these, such as replacing missing values with the mode, imputing based on the distribution of known values, or removing the records. Ultimately, we decided to retain 'unknown' as a separate category of its own because its absence could convey meaningful information for the telemarketing process. In some cases, not knowing a client's education level, job or contact could be informative for the telemarketer.

Outliers

To detect outliers, we decided to use the three-sigma rule also known as the"68-95-99.7 rule." This rule considers that approximately 68% of the data falls within one standard deviation of the mean, about 95% falls within two standard deviations, and around 99.7% falls within three standard deviations. We considered data points falling more than three standard deviations from the mean as outliers. This rule helped us identify extreme values across our seven numeric variables: Age, Balance, Day, Duration, Campaign, Pdays, and Previous. Notably, we chose to retain all age data, viewing each age group as a significant market segment. For Balance, 745 records were removed; however, for Day which represents the day of the month on which contact was made we kept all records. We considered that removing what seemed to be outliers would distort the dataset by eliminating valid and potentially important observations. Duration outliers were adjusted by removing 964 records, while for Campaign, Pdays, and Previous, we removed 840, 1723, and 582 records respectively. After cleaning, the total number of records

removed was 4682. Our dataset was reduced to 40,529 records, accounting for overlaps in outlier identification across variables.

Figure 2: Statistical Summary for Identifying Outliers

Attributes	First Quartile	Third Quartile	Min	Max	Avg	Deviation	Three-sigma Lower Bound	Three-sigma Higher Bound
Age	33	48	18	95	40.936	10.619	9.079	72.793
Balance	72	1428	-8019	102127	1362.272	3044.766	-7772.026	10496.57
Day	8	21	1	31	15.806	8.322	-9.16	40.772
Duration	103	319	0	4918	258.163	257.528	-514.421	1030.747
Campaign	1	3	1	63	2.764	3.098	-6.53	12.058
Pdays	-1	-1	-1	871	40.198	100.129	-260.189	340.585
Previous	0	0	0	275	0.58	2.303	-6.329	7.489

Source: Team consensus

Enhancements and Corrections

We added an ID row to facilitate data handling. In RapidMiner, we assigned the role of 'ID' to this new attribute and 'label' to the Y attribute. We also adjusted the data types of 'Y', 'Default', 'loan', and 'housing' from polynomial to binomial to accurately reflect their binary nature.

Correlation matrix

Our examination of the correlation matrix among numerical variables revealed no significant concerns, with the highest correlation coefficient being 72% between Previous(number of contacts performed before this campaign and for this client) and Pdays(number of days that passed by after the client was last contacted from a previous campaign). This was considered negligible, as our threshold for concern was set at 85%.

Figure 3: Correlation matrix

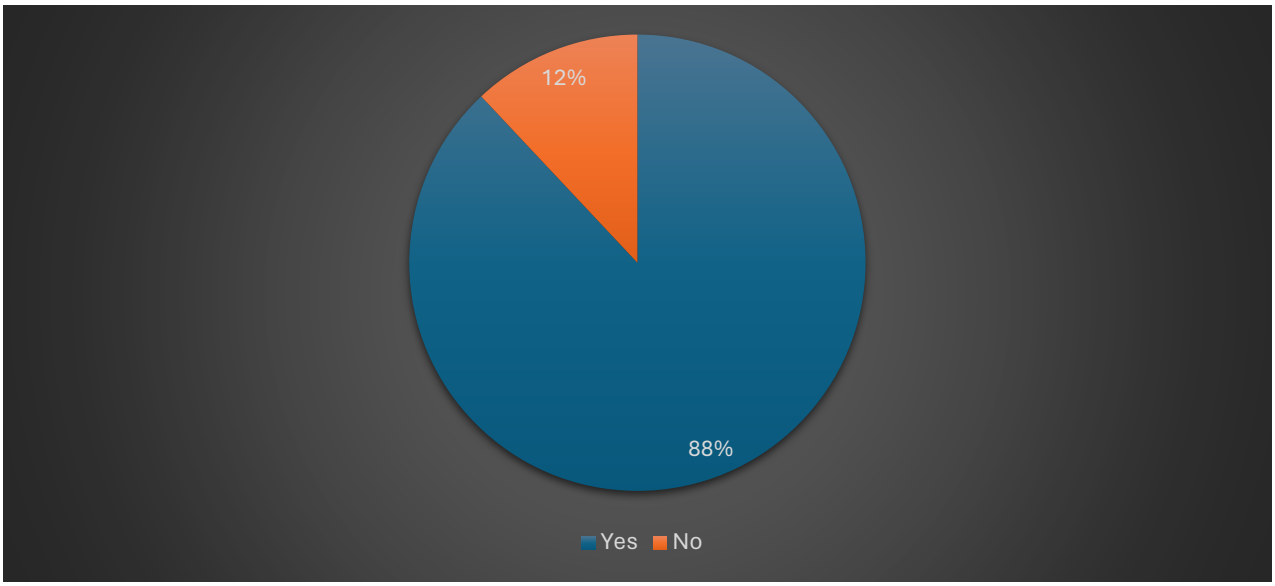
Attributes	age	balance	day	duration	campaign	pdays	previous
age	1	0.103	-0.011	-0.013	0.029	0.001	0.022
balance	0.103	1	0.010	0.033	-0.018	0.043	0.058
day	-0.011	0.010	1	-0.025	0.133	-0.062	-0.052
duration	-0.013	0.033	-0.025	1	-0.077	0.016	0.021
campaign	0.029	-0.018	0.133	-0.077	1	-0.089	-0.066
pdays	0.001	0.043	-0.062	0.016	-0.089	1	0.720
previous	0.022	0.058	-0.052	0.021	-0.066	0.720	1

Source: RapidMiner

Data Partitioning

For model training and testing, we used the holdout method, allocating 70% of the data (28,371 records) for training and the remaining 30% (12,158 records) for testing. The dataset exhibited an 88% 'Yes' (subscribed) to 12% 'No' (not subscribed) distribution, a balance maintained in both training and testing sets to ensure representativeness.

Figure 4: Proportion of Positive and Negative Responses in the Training and Testing Datasets



Source: Team Consensus

V/ MODELING: BUILDING THE PREDICTIVE FRAMEWORK

For our modeling, we decided to build five models: Decision Trees, Neural Networks, Logistic Regression, Naïve Bayes, and Gradient Boosting Machines (GBM). We are utilizing both R and RapidMiner for these tasks. Using both R and RapidMiner allows us to leverage the strengths of each tool: R for its extensive statistical capabilities and customizable modeling options, and RapidMiner for its user-friendly interface and robust data processing functionalities.

1/Decision tree

A decision tree is a classification method that uses a tree-like model of decisions and their possible consequences. It is constructed from a root node, internal nodes, and leaf nodes. Each internal node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label, a decision taken after computing all attributes (Navada et al., 2011). Decision trees are popular for their straightforward structure, which mimics human decision-making processes, making them easy to interpret and explain.

RapidMiner Implementation:

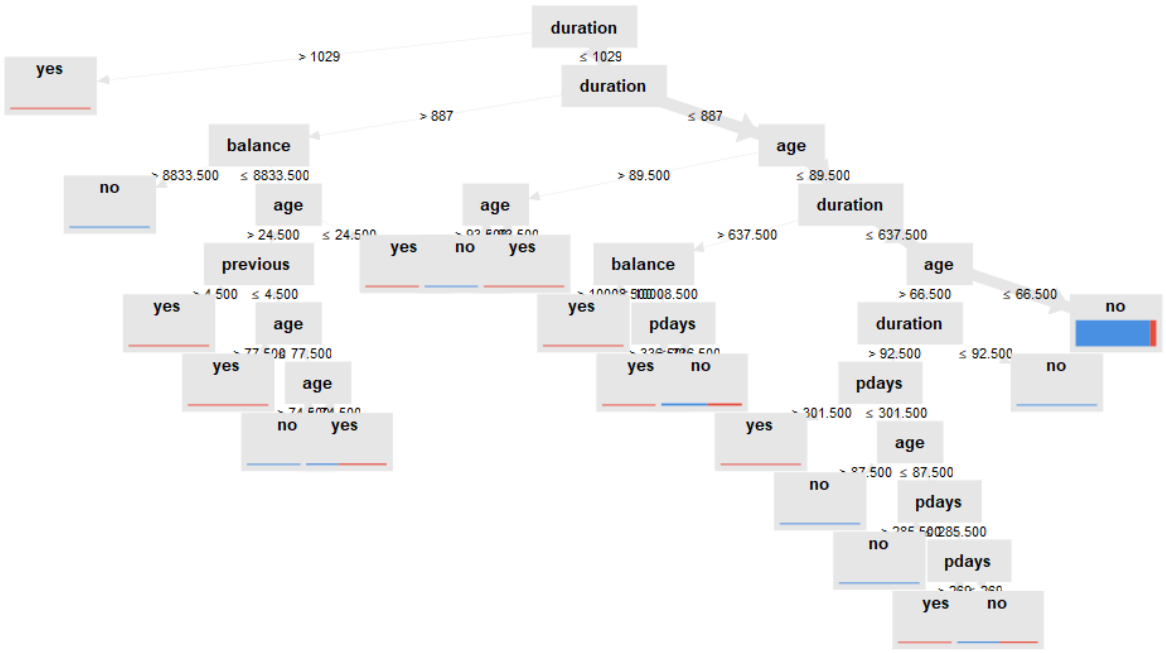
In RapidMiner, our decision tree predicts 10,756, instances of 'No' and 1,402 instances of 'Yes' for the subscription.

For our model, we chose the gain ratio as the criterion because it balances the size of the branches, thus reducing the bias towards attributes with more levels and providing a more normalized measure of information gain. We set a confidence of 0.25 to avoid overfitting, ensuring the tree is general enough for unseen data. A minimal gain of 0.08 and a minimal leaf size of 2 were selected to prevent the model from being too complex and to ensure that each decision path is supported by a sufficient number of cases. **(APPENDIX parameters)**

Key attributes identified by our decision tree in RapidMiner are:

- Duration: This is the most influential predictor. Longer call durations tend to increase the likelihood of subscription.
- Age: Specific age groups have distinct subscription behaviors, influencing the decision at various tree splits.
- Balance: Indicates financial stability, with higher balances often leading to subscription.
- Pdays: The time since the last campaign contact also affects the decision, with recent contacts being more significant.
- Previous: The number of prior contacts has a predictive value, influencing the likelihood of subscription.

Figure 5: Decision Tree Model in RapidMiner



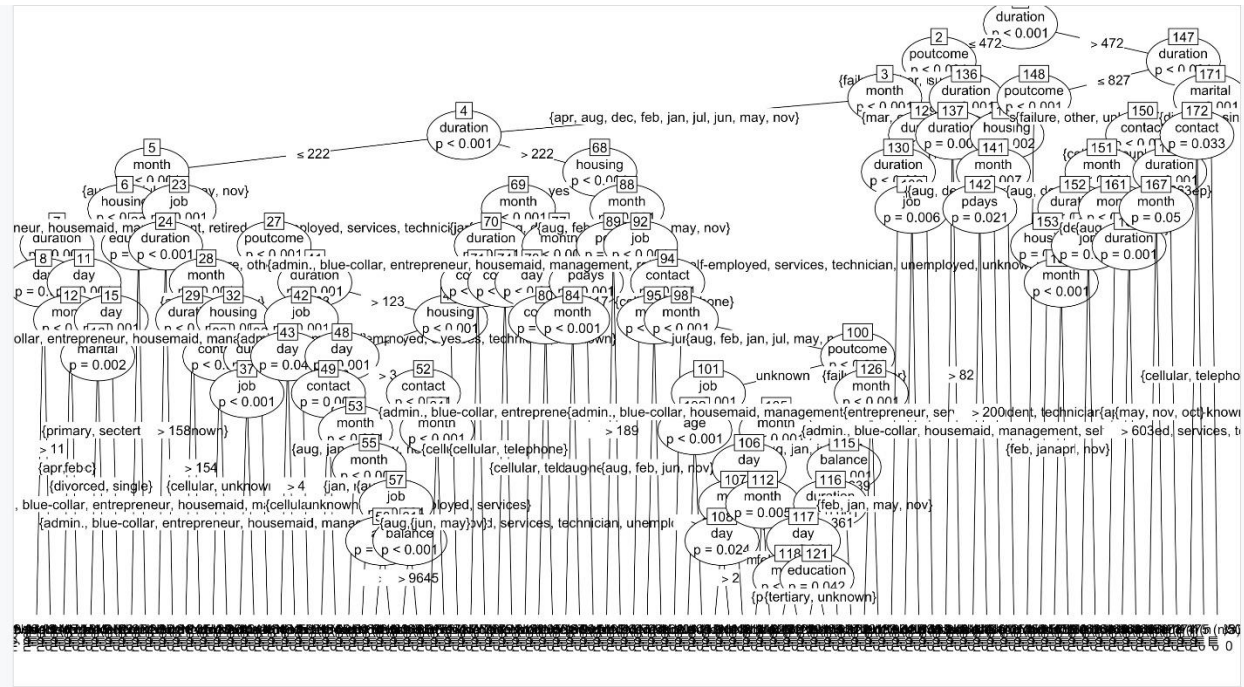
Source: RapidMiner

R Implementation:

In R, our decision tree model results in 1,358 'Yes' and 10,800 'No' predictions for subscription. The 'Poutcome' variable, representing the outcome of the previous marketing campaign, emerges

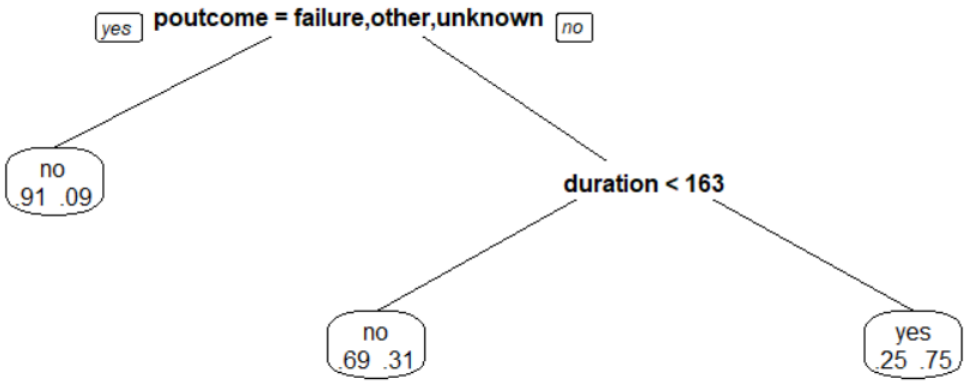
as the most significant predictor, suggesting that a positive experience with the bank increases the likelihood of a future subscription. The 'duration' of the call follows in importance, reinforcing the pattern seen in the RapidMiner model that longer interactions are more likely to result in positive outcomes.

Figure 6: Decision Tree in R



Source: R

Figure 7: Simplified Decision Tree in R



Source: R

In summary, both implementations highlight the significance of specific attributes in predicting subscription likelihood, with 'duration' consistently being a critical factor across both platforms.

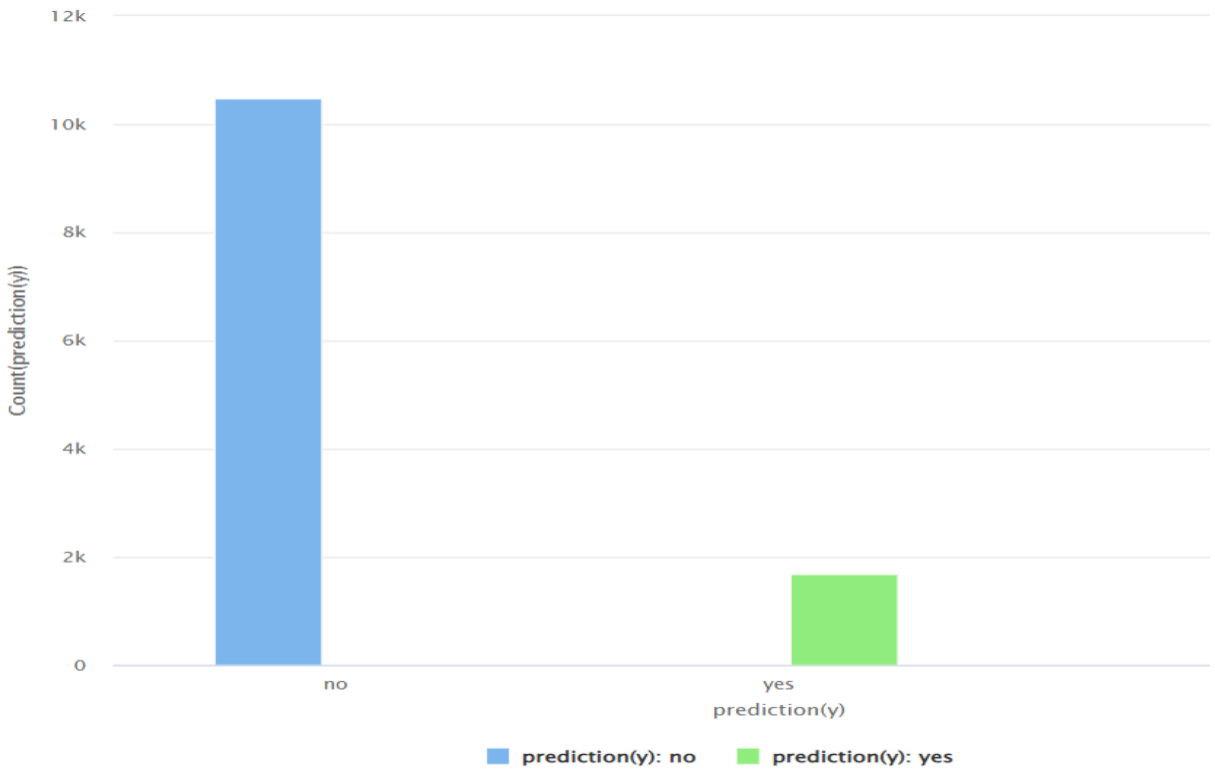
2/Naïve bayes

Naïve Bayes is a probabilistic classifier based on applying Bayes' theorem with the “naïve” assumption of independence between every pair of features. Despite its simplicity, Naïve Bayes can yield surprisingly accurate predictions (Kumar, 2022). The model calculates the probability of each class and the conditional probability of each class given each input value. This method assumes that the presence of a particular feature in a class is independent of the presence of any other feature, which is why it's termed "naïve."

RapidMiner and R Implementation:

In RapidMiner, our Naïve Bayes model predicts 10,471 instances of 'No' and 1,687 instances of 'Yes' for subscribing to the bank's term deposits. Similarly, in R, the model outputs 10,471 'No' and 1,687 'Yes' predictions. The attributes 'Duration' and 'Poutcome' appear to be particularly influential in both models. The notable differences in their means and standard deviations, or probabilities, across the two classes, suggest these variables are significant predictors in the context of term deposit subscriptions.

Figure 8: Predictive Outcomes of Term Subscription Decision Naïve Bayes in RapidMiner



Source: RapidMiner

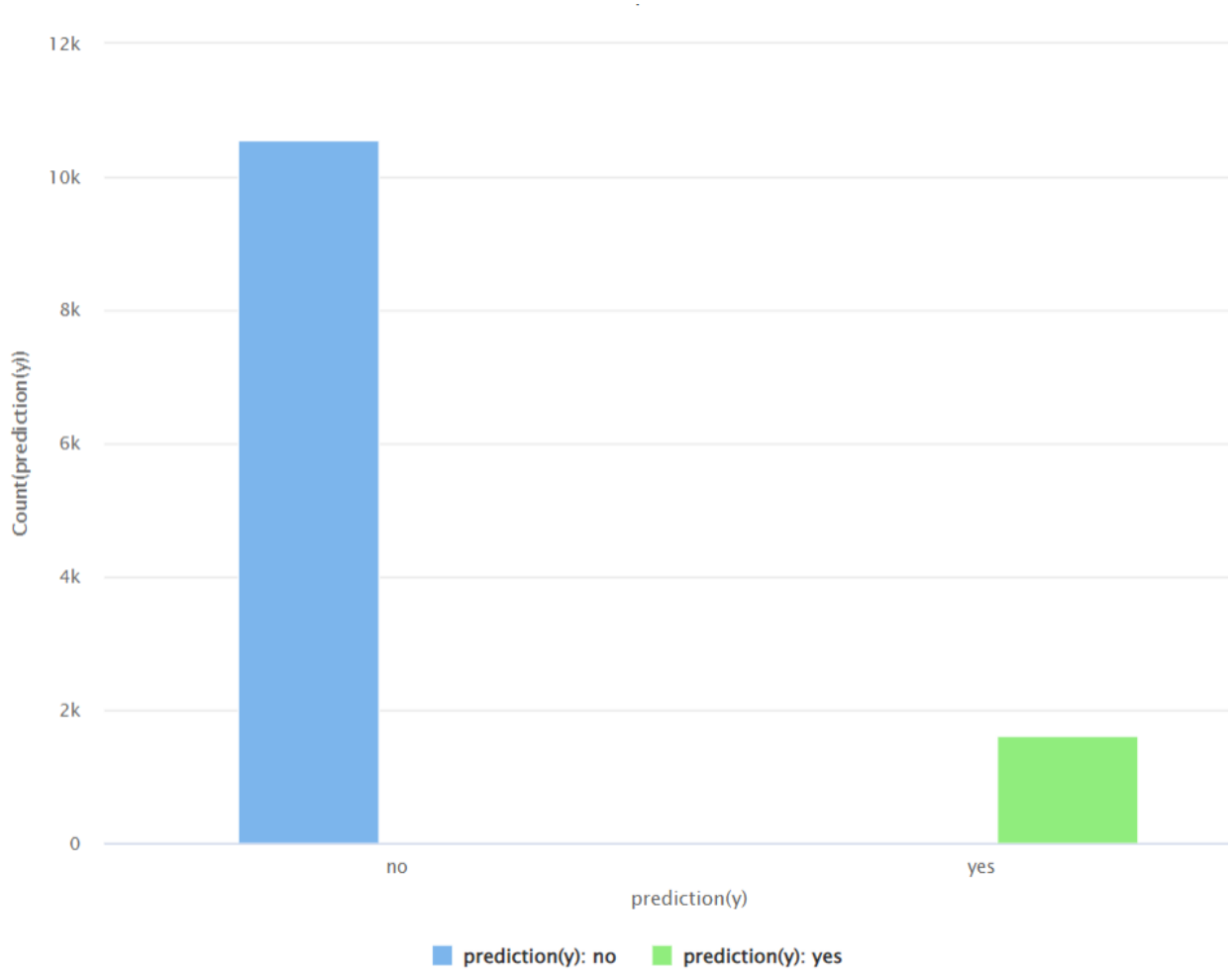
3/Logistic regression

Logistic regression is a statistical model that applies a logistic function to model a binary dependent variable.

RapidMiner (RM) Implementation

Our logistic regression model in RapidMiner predicted 10,551 'No' and 1,607 'Yes' instances for subscription.

Figure 9: Predictive Outcomes of Term Subscription Logistic Regression Model in RapidMiner



Source: RapidMiner

The influence of various attributes on the model's predictions can be determined by examining the standardized coefficients and the p-values.

Figure 11: Impact Assessment of Attributes on Model Predictions in RapidMiner

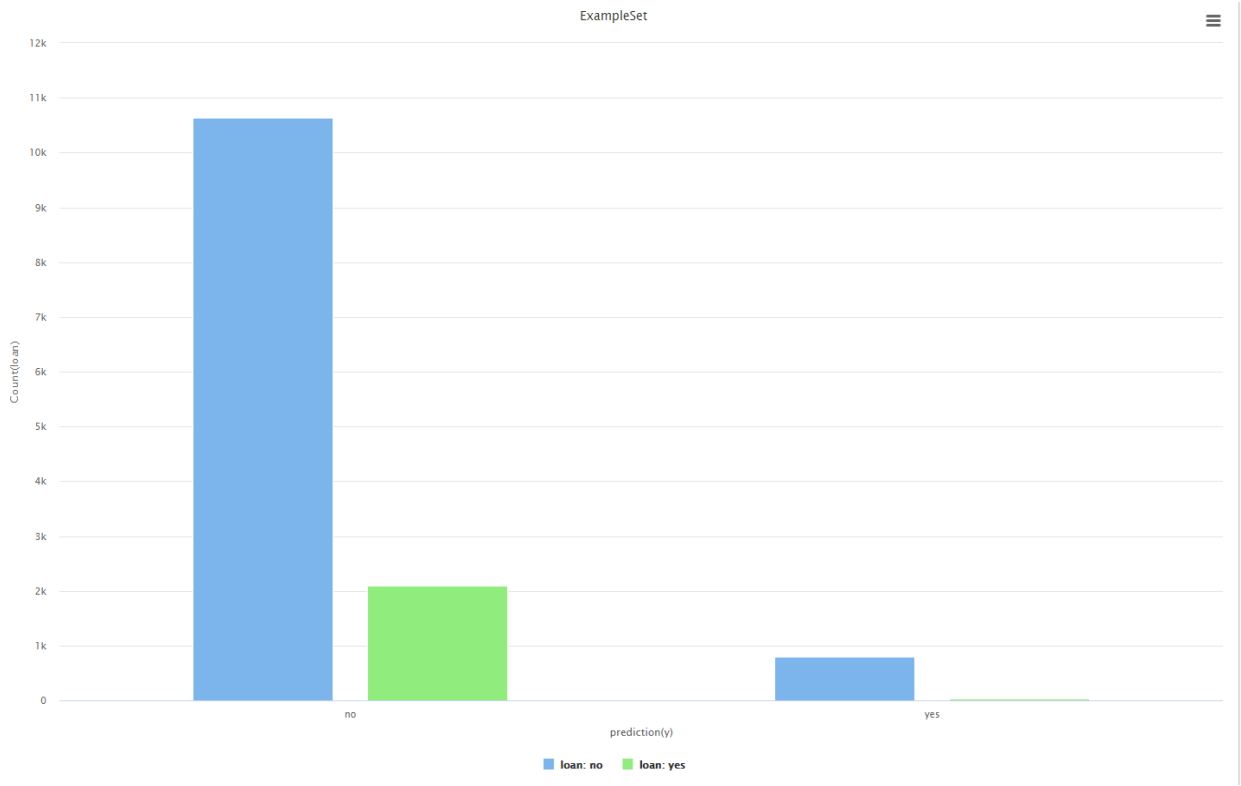
Attribute	Coefficient	Std.Coeff	Std. Error	z-value	p-value
job.unknown	0.14	0.14	0.32	0.43	0.67
job.management	0.19	0.19	0.16	1.18	0.24
job.technician	0.17	0.17	0.16	1.06	0.29
job.services	0.14	0.14	0.18	0.77	0.44
job.retired	0.71	0.71	0.18	3.91	0.00
job.admin.	0.35	0.35	0.17	2.09	0.04
job.blue-collar	0.10	0.10	0.17	0.60	0.55
job.self-employed	-0.07	-0.07	0.20	-0.33	0.74
job.unemployed	0.17	0.17	0.20	0.88	0.38
job.student	0.76	0.76	0.20	3.77	0.00
job.housemaid	-0.04	-0.04	0.23	-0.18	0.85
month.jun	0.79	0.79	0.10	7.68	0.00
month.jul	-0.71	-0.71	0.10	-7.21	0.00
month.aug	-0.59	-0.59	0.10	-5.94	0.00
month.oct	1.41	1.41	0.14	10.21	0.00
month.nov	-0.60	-0.60	0.11	-5.64	0.00
month.dec	1.25	1.25	0.22	5.68	0.00
month.jan	-1.00	-1.00	0.16	-6.18	0.00
month.feb	0.15	0.15	0.11	1.38	0.17
month.mar	2.06	2.06	0.15	13.83	0.00
month.apr	0.33	0.33	0.10	3.29	0.00
month.sep	1.29	1.29	0.16	8.14	0.00
education.unknown	-0.07	-0.07	0.12	-0.60	0.55
education.tertiary	0.30	0.30	0.07	4.41	0.00
education.primary	-0.16	-0.16	0.09	-1.81	0.07
poutcome.other	0.58	0.58	0.19	3.11	0.00
poutcome.failure	0.49	0.49	0.17	2.90	0.00
poutcome.success	2.76	2.76	0.16	17.29	0.00
marital.single	0.31	0.31	0.06	4.99	0.00
marital.divorced	0.21	0.21	0.08	2.69	0.01
contact.cellular	1.79	1.79	0.10	18.53	0.00
contact.telephone	1.75	1.75	0.13	13.11	0.00
default.yes	0.15	0.15	0.20	0.74	0.46
housing.no	0.74	0.74	0.06	12.54	0.00
loan.no	0.49	0.49	0.08	5.99	0.00
age	0.00	0.02	0.00	0.66	0.51
balance	0.00	0.07	0.00	2.91	0.00
day	0.01	0.07	0.00	2.57	0.01
duration	0.01	1.09	0.00	51.98	0.00
campaign	-0.11	-0.22	0.02	-6.77	0.00
pdays	0.00	-0.26	0.00	-5.39	0.00
previous	0.08	0.08	0.03	2.53	0.01
Intercept	-6.87	-5.60	0.24	-28.55	0.00

Source RapidMiner

- Duration: A high positive coefficient and a p-value of 0, indicating a significant positive influence on subscription likelihood.
- Contact Method (cellular/telephone): Both modes of contact show high positive coefficients and p-values of 0, signifying strong influence.
- Poutcome.success: A very high positive coefficient and a p-value of 0, highlighting its strong predictive power for positive responses.

- Months: March, October, December, and September show high positive coefficients and p-values of 0, which positively correlate with subscription outcomes, whereas July, August, and January show high negative coefficients, indicating a negative association.
- Job Categories (retired/student): These have high positive coefficients and very low p-values, suggesting higher subscription likelihood.
- Conversely, attributes like 'job.unknown', with a coefficient near zero and a high p-value, have less predictive significance.
- A higher proportion of people without a loan accepted the term deposit offer compared to those with a loan. The majority of respondents without a loan chose not to participate in the term deposit, while the acceptance rate was lower among those with a loan. This could be as a result of Loan repayment often seen as a more urgent financial goal and borrowers may prioritize reducing debt over saving.

Figure 10: Proportion of Customers with Loan to Term Deposit Subscription



Source: RapidMiner

R Implementation:

In R, the logistic regression outputs are 1,687 'Yes' and 10,471 'No' predictions. The model identifies contact type, specific months, job categories, housing status, loan status, duration of the last contact, and the outcome of the previous marketing campaign as significant predictors. For instance, having a housing loan (housing=yes) is associated with a negative impact on the

probability of subscribing, while a successful outcome in the previous campaign (Poutcome=success) greatly increases the probability of a subscription.

Figure 11: Impact Assessment of Attributes on Model Predictions in R

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.625e+00  2.727e-01  -9.627  < 2e-16 ***
age                1.944e-03  2.951e-03   0.659 0.509969
jobblue-collar    -2.505e-01  9.879e-02  -2.536 0.011212 *
jobentrepreneur   -3.508e-01  1.680e-01  -2.088 0.036795 *
jobhousemaid      -3.922e-01  1.847e-01  -2.124 0.033683 *
jobmanagement     -1.621e-01  9.889e-02  -1.639 0.101177
jobretired         3.635e-01  1.298e-01   2.801 0.005097 **
jobself-employed  -4.171e-01  1.541e-01  -2.707 0.006795 **
jobservices       -2.140e-01  1.150e-01  -1.860 0.062861 .
jobstudent         4.095e-01  1.470e-01   2.786 0.005329 **
jobtechnician     -1.775e-01  9.397e-02  -1.889 0.058891 .
jobunemployed     -1.773e-01  1.465e-01  -1.210 0.226259
jobunknown        -2.132e-01  2.941e-01  -0.725 0.468493
maritalmarried    -2.118e-01  7.877e-02  -2.689 0.007170 **
maritalsingle      9.668e-02  9.018e-02   1.072 0.283668
educationsecondary 1.556e-01  8.582e-02   1.813 0.069889 .
educationtertiary  4.571e-01  9.939e-02   4.599 4.25e-06 ***
educationunknown   8.099e-02  1.404e-01   0.577 0.563988
defaultyes         1.499e-01  2.016e-01   0.743 0.457252
balance            3.791e-05  1.301e-05   2.914 0.003567 **
housingyes        -7.417e-01  5.915e-02 -12.540  < 2e-16 ***
loanyes           -4.916e-01  8.207e-02  -5.991 2.09e-09 ***
contacttelephone  -3.966e-02  9.944e-02  -0.399 0.690010
contactunknown    -1.787e+00  9.644e-02 -18.530  < 2e-16 ***
day                8.561e-03  3.333e-03   2.569 0.010207 *
monthaug          -9.188e-01  1.071e-01  -8.580  < 2e-16 ***
monthdec           9.238e-01  2.232e-01   4.139 3.49e-05 ***
monthfeb          -1.758e-01  1.188e-01  -1.480 0.138843
monthjan          -1.322e+00  1.619e-01  -8.169 3.12e-16 ***
monthjul          -1.036e+00  1.052e-01  -9.851  < 2e-16 ***
monthjun           4.656e-01  1.258e-01   3.701 0.000215 ***
monthmar           1.733e+00  1.534e-01  11.294  < 2e-16 ***
monthmay          -3.267e-01  9.940e-02  -3.287 0.001014 **
monthnov          -9.259e-01  1.118e-01  -8.281  < 2e-16 ***
monthoct           1.087e+00  1.420e-01   7.655 1.94e-14 ***
monthsep           9.601e-01  1.635e-01   5.872 4.30e-09 ***
duration           5.707e-03  1.098e-04  51.988  < 2e-16 ***
campaign          -1.111e-01  1.641e-02  -6.769 1.30e-11 ***
pdays             -3.577e-03  6.639e-04  -5.388 7.14e-08 ***
previous           7.684e-02  3.035e-02   2.532 0.011355 *
poutcomeother      9.247e-02  1.397e-01   0.662 0.508196
poutcomesuccess    2.276e+00  1.158e-01  19.647  < 2e-16 ***
```

Source: R

4/Neural Networks

Neural networks are computational models inspired by the human brain's network of neurons. They consist of interconnected nodes or "neurons" arranged in layers (Zahner, 2000).

RapidMiner (RM) Implementation:

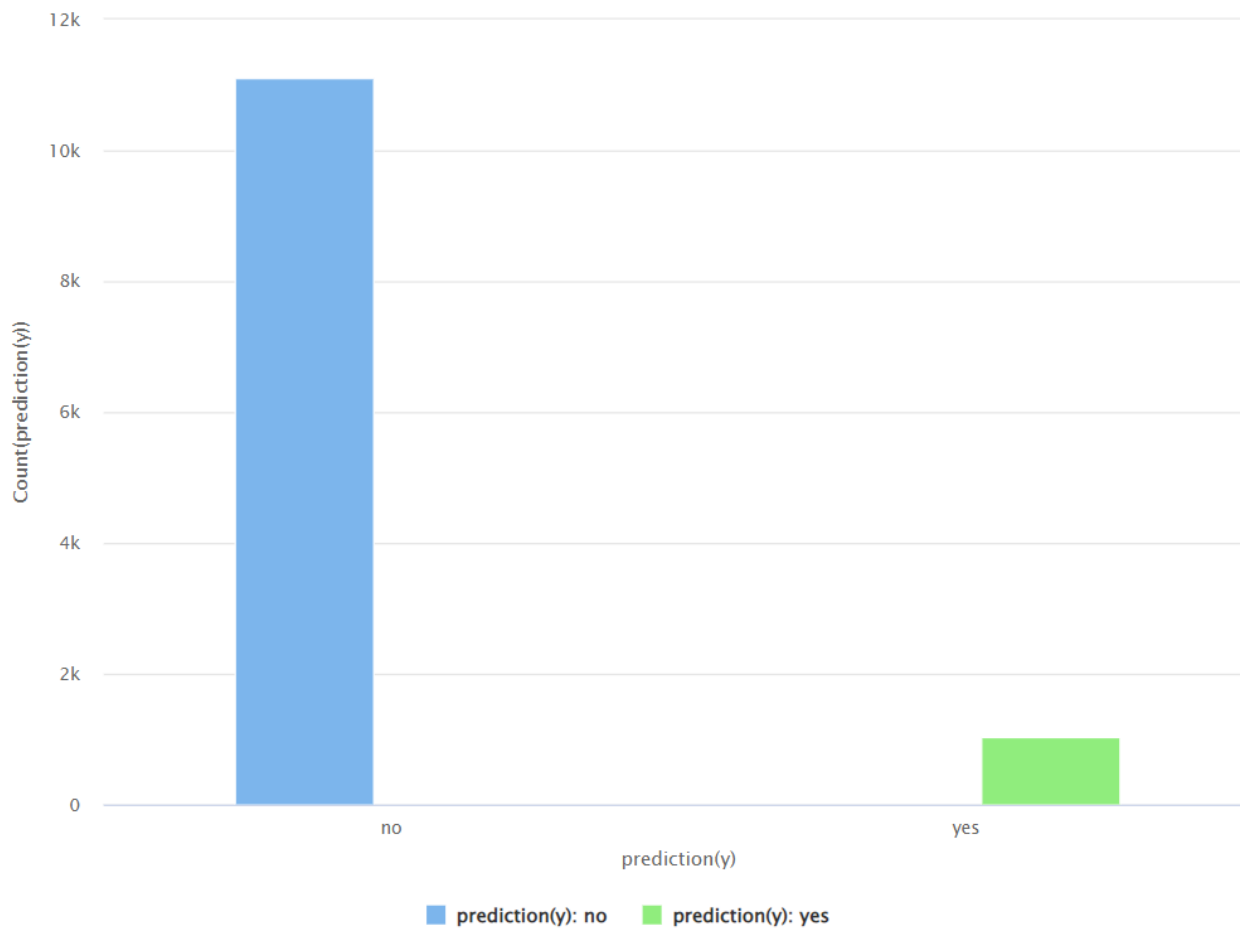
In our RapidMiner implementation, the neural network model predicted 10,818 'No' and 1,340 'Yes' outcomes. The model's complexity stems from its structure. In our case, it includes 52 inputs, corresponding to the different combinations of attributes and their possible values, particularly for nominal attributes, alongside the numeric attributes. This extensive input layer is necessary to capture the various factors that influence a customer's decision to subscribe to a term deposit.

Our network configuration comprises one hidden layer with 29 nodes. The number of nodes was chosen to balance the complexity of the model and the computational efficiency, allowing the network to learn the intricate patterns in the data without overfitting. These nodes, or neurons, in the hidden layer act as processors that weigh the input data, apply a transformation function, and pass the result to the next layer.

The model culminates in two output nodes corresponding to the binary classification outcomes ('Yes' or 'No'). These outputs are determined by aggregating the weighted signals from the hidden layer, passing them through an activation function that dictates the final prediction. The weights between the nodes are adjusted during the training process to improve the model's accuracy in predicting whether a customer will subscribe to a term deposit.

Understanding the exact reasoning behind each prediction in a neural network can be challenging due to its 'black box' nature, where the internal processing is not directly interpretable. However, the predictive performance of the model is often strong, making it a valuable tool in complex decision-making scenarios like customer behavior prediction.

Figure 12: Predictive Outcomes of Term Subscription Neural Network Model in RapidMiner



Source: RapidMiner

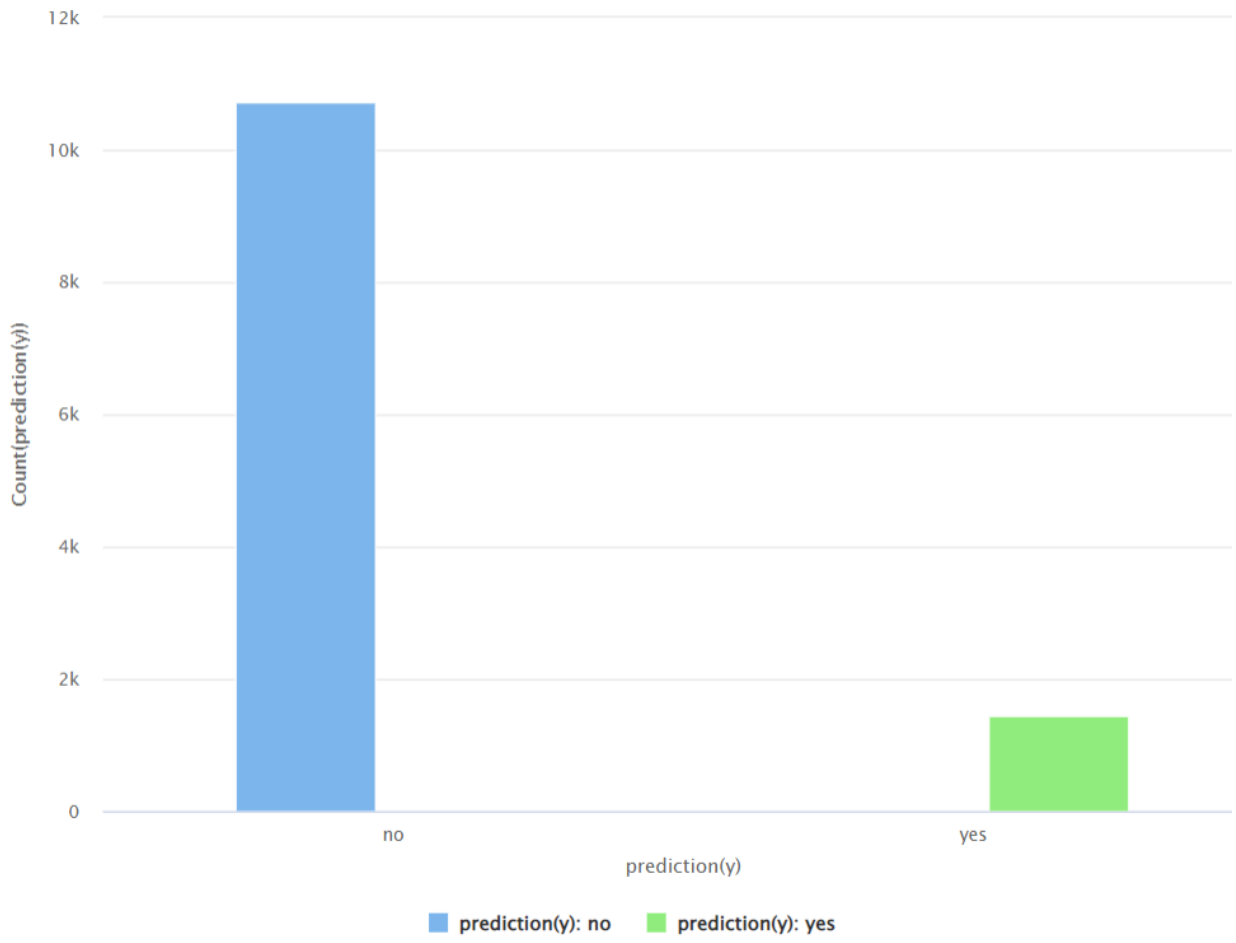
5/ Gradient Boosting Machines (GBM)

GBMs iteratively improve predictions by combining the strengths of multiple decision trees. This approach helps in reducing bias and variance, leading to improved model performance on complex datasets.

RapidMiner implementation:

In RapidMiner, our GBM model produced results with 10,723 'No' and 1,435 'Yes' predictions for the term deposit subscriptions.

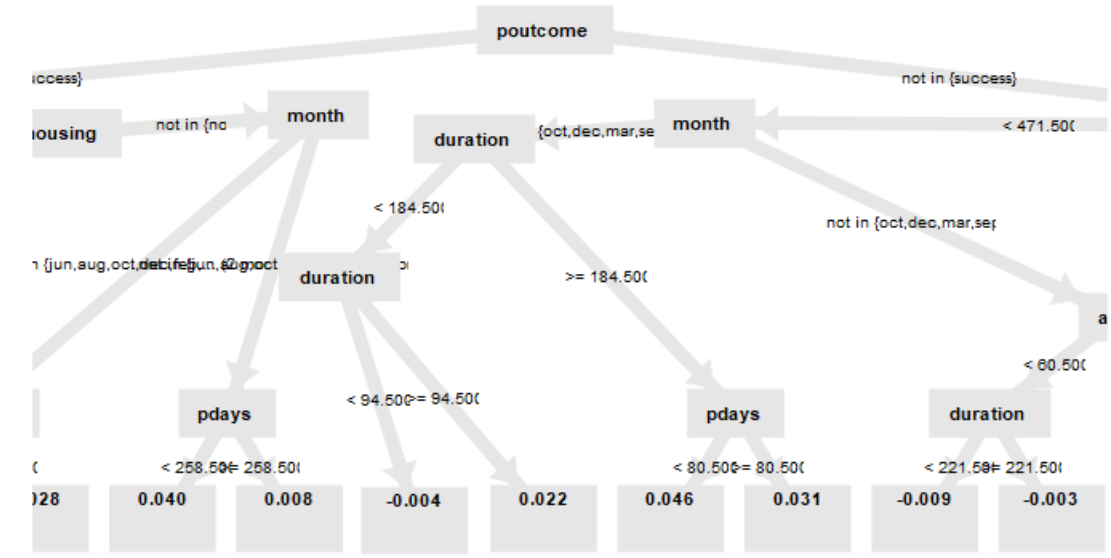
Figure 13: Predictive Outcomes of Term Subscription GBM Model in RapidMiner



Source: RapidMiner

We opted to create a model consisting of 50 trees. This number was chosen to balance the trade-off between model accuracy and the risk of overfitting. With too few trees, the model might not capture all the complexities of the data, whereas too many trees can lead to overfitting, where the model performs well on training data but poorly on unseen data. Each tree in the ensemble focuses on correcting the mistakes of the previous one, gradually improving the overall model's accuracy. By limiting the number to 50, we ensure that the model remains general enough to perform well on both the training and unseen test data, while still being sufficiently detailed to capture the underlying patterns in customer behavior regarding term deposit subscriptions.

Figure 14: Illustration of a tree used in GBM



Source: RapidMiner

VI/ EVALUATION: ASSESSING MODEL PERFORMANCE

In assessing the effectiveness of our predictive models, we use a confusion matrix and examine several key performance metrics: accuracy, precision, recall, and the F-measure. The choice of the best model depends on our specific campaign objectives, whether it be maximizing the number of correct predictions (accuracy), minimizing the cost of marketing unlikely customers (precision), expanding the reach to potential subscribers (recall), or achieving a balance between precision and recall (F-measure).

Figure 15: Key Performance Metrics for Model Evaluation

Accuracy	Capacity to correctly identify subscribers and non-subscribers
Precision	Capacity to be correct when predicting a customer will subscribe
Recall	Capacity to identify a high number of actual subscribers
F-measure	Balance of precision and recall

Source: Teams Consensus

Figure 16: Confusion Matrix for RapidMiner Models

	Decision Trees		Naïve Bayes		Neural Network		Logistic Regression		Gradient Boosted Model	
	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
No	10504	252	10424	877	10519	299	9958	593	10244	479
Yes	761	641	505	566	470	870	277	1330	336	1099
Accuracy	92.68%		89.11%		93.38%		95.47%		95.63%	
Precision	71.80%		57.40%		74.42%		69.18%		69.65%	
Recall	45.71%		70.08%		64.94%		82.77%		76.57%	
F-measure	50.86%		63.11%		69.36%		75.37%		72.95%	

Source: Teams Consensus

Figure 17: Confusion Matrix for R Models

	Decision Trees		Naïve Bayes		Logistic Regression	
	No	Yes	No	Yes	No	Yes
No	10572	228	10424	877	10424	877
Yes	667	691	505	566	505	566
Accuracy	90.69%		89.11%		89.11%	
Precision	75.19%		57.40%		57.40%	
Recall	50.87%		70.08%		70.08%	
F-measure	60.68%		63.11%		63.11%	

Source: Teams Consensus

Our Decision Trees in RM show lower performance in most metrics compared to other models, suggesting it may not be the best choice for this campaign. The one in R shows a substantial improvement in all metrics compared to its RapidMiner counterpart, especially in F-measure, indicating a better balance between precision and recall.

Naïve Bayes models are consistent across both platforms in terms of accuracy, precision, recall, and F-measure, indicating reliability. However, the lower precision compared to other models might be a concern if marketing resources are limited.

Neural Network in RM excels in accuracy and shows a balanced F-measure, marking it as a strong candidate for overall performance.

Logistic Regression shows a significant variance between R and RapidMiner implementations. The RapidMiner model outperforms in accuracy and F-measure, likely due to a better balance of recall and precision, suggesting it is a preferable choice for campaigns where both identifying subscribers and limiting marketing costs are important.

Gradient Boosted Machines in RM provide the highest accuracy, a solid indication of overall model performance. Its precision and recall are commendably high, but not the highest among the models.

Considering the marketing goals, if the bank aims to ensure that no potential subscriber is overlooked, the Logistic Regression model in RapidMiner should be considered because it has the highest recall. Conversely, if the costs of reaching out to non-subscribers are high, the model with the best precision, which is Decision Trees in R, would be optimal. In cases where both types of errors are costly, the model with the highest F-measure, indicating a balanced precision-recall trade-off, would be the most suitable choice. Given our metrics, the Logistic Regression model in RapidMiner stands out with the highest F-measure, suggesting it is the best model for campaigns where balance is critical.

VII/ DISCUSSION: IMPLICATIONS AND FUTURE DIRECTIONS

The importance of the study within the context of the rapidly evolving financial services industry cannot be overstated. By leveraging data mining techniques to predict customer responses to term deposit offers, this research contributes to the critical understanding of consumer behavior in the sector. Using data mining could also reduce the amount of time it takes to review customer data (Raj, 2015). The findings have the potential to revolutionize how banks and financial institutions approach telemarketing, moving from intuition-based to data-driven strategies.

Our results illuminate key factors influencing customer decisions, such as the duration of calls, the customer's job category, and the outcomes of previous interactions. These insights allow for a more nuanced understanding of the customer base, providing a pathway for more personalized and effective marketing strategies.

The strategy for implementing the models begins with a careful selection based on the specific objectives of the bank's telemarketing campaign. With the demonstrated performance metrics in mind, the deployment of models like Logistic Regression or Decision Trees in R, which show a balance of high precision and recall, can significantly increase the efficiency of the campaign.

Deployment and maintenance of these models will require a structured approach, involving the integration of the models into the bank's customer relationship management systems, continuous monitoring for performance, and periodic updating to account for emerging trends and patterns in customer behavior.

However, limitations exist within this study that must be acknowledged. The reliance on a historical dataset may not fully capture the dynamic nature of customer behavior, and the models may not account for unforeseen economic or social changes that affect customer decisions. Additionally, the 'black box' nature of models like Neural Networks makes it challenging to derive actionable insights on why certain predictions are made, which can hinder trust and acceptance among end-users.

Moving forward, further research should focus on the integration of real-time data analysis to keep the models current and accurate. The exploration of hybrid models that combine the interpretability of Decision Trees with the predictive power of methods like Gradient Boosting may also yield beneficial results. Finally, the examination of the ethical implications of data mining, particularly in terms of privacy and consent, is imperative in ensuring that these advancements are aligned with broader social values and norms.

VIII/ REFERENCES

Infobase. (2022). Acxiom. Retrieved from: https://www.acxiom.com/customer-data/infobase/

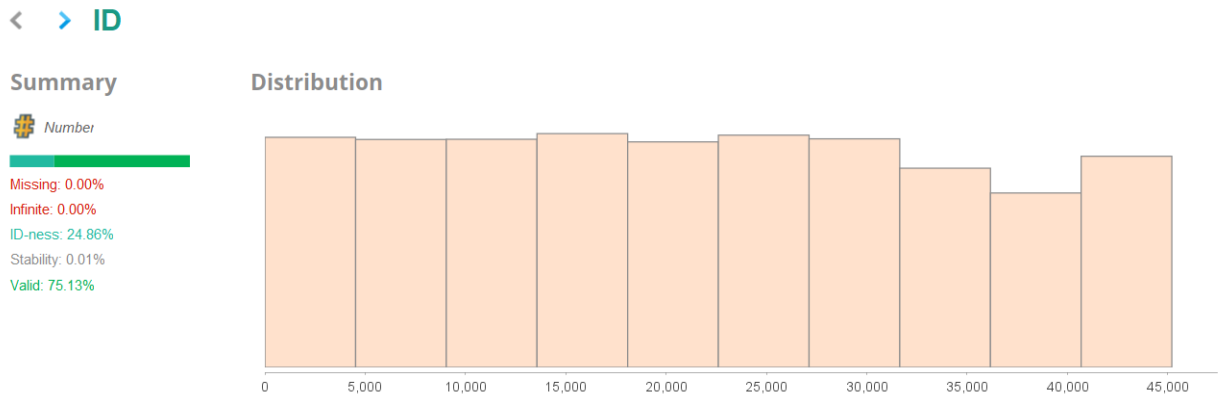
Kumar, N. (2022). *Naive Bayes Classifiers*. GeeksforGeeks. Retrieved from: https://www.geeksforgeeks.org/naive-bayes-classifiers/

Raj, S. (2015). *A Review of Data Mining Applications in Banking*. International Conference on Recent Advances in Electronics, Computer Science and Information Technology, Chennai, India. doi:10.4236/ce.2015.615165.

Zahner, D. A., & Micheli-Tzanakou, E. (2000). *Artificial Neural Networks: Definitions, Methods, Applications.* 1st Edition. CRC Press. Retrieved from: https://taylorandfrancis.com/chapters/edit/10.1201/9781420049770-2/artificial-neural-networks-daniel-zahner-evangelia-micheli-tzanakou

VIII/ APPENDIX

1-Summary statistics of attribute "ID"



Statistics

Name	Value
Minimum	1
Maximum	45210
Average	21866.377
Standard Deviation	12951.266

2-Summary statistics of attribute "Age"

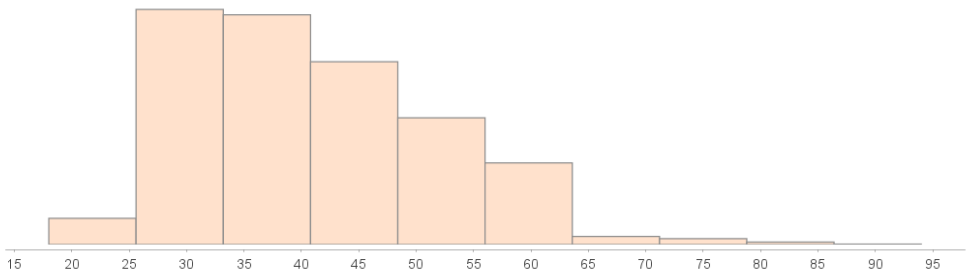
age

Summary

Number

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.18%
Stability: 4.80%
Valid: 95.02%

Distribution



Statistics

Name	Value
Minimum	18
Maximum	95
Average	40.980
Standard Deviation	10.629

3-Summary statistics of attribute “balance”

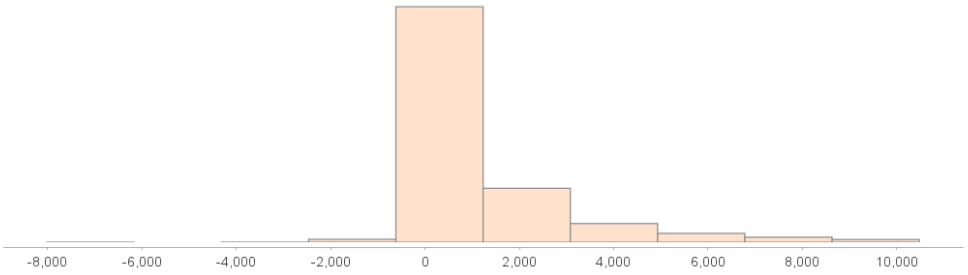
balance

Summary

Number

Missing: 0.00%
Infinite: 0.00%
ID-ness: 8.95%
Stability: 7.73%
Valid: 83.32%

Distribution



Statistics

Name	Value
Minimum	-8019
Maximum	10483
Average	1091.138
Standard Deviation	1730.698

4-Summary statistics of attribute “campaign”

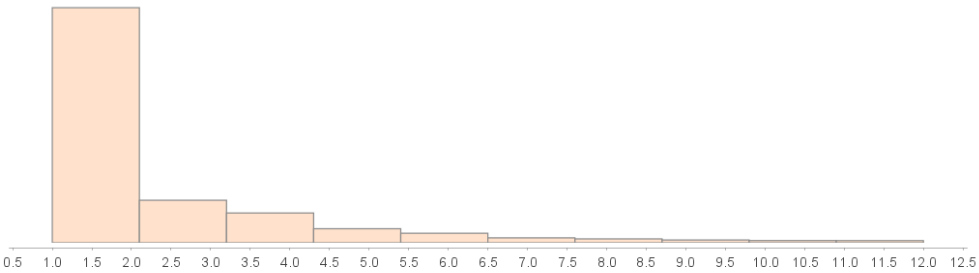
campaign

Summary

Number

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.03%
Stability: 39.58%
Valid: 60.39%

Distribution



Statistics

Name	Value
Minimum	1
Maximum	12
Average	2.461
Standard Deviation	1.950

5-Summary statistics of attribute “contact”

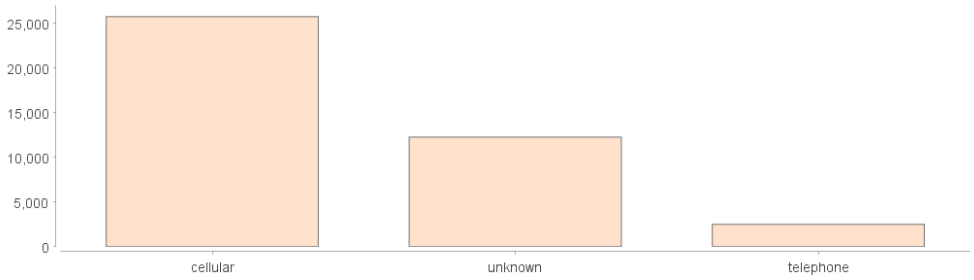
contact

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.01%
Stability: 63.82%
Valid: 36.17%

Top Values



3 Distinct Values:

Value	Count	Percentage
cellular	25,758	63.55%
unknown	12,268	30.27%
telephone	2,503	6.18%

6-Summary statistics of attribute “day”

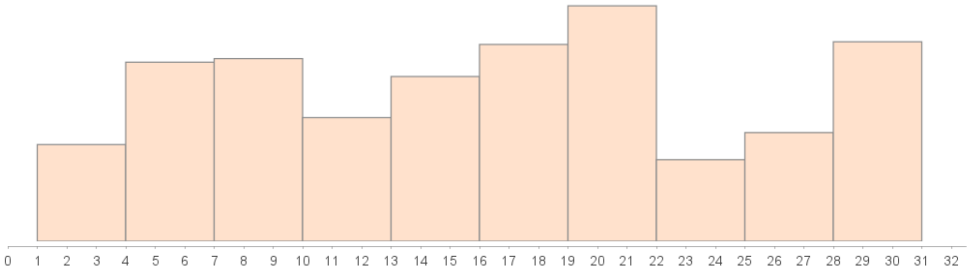
day

Summary

Number

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.08%
Stability: 6.15%
Valid: 93.77%

Distribution



Statistics

Name	Value
Minimum	1
Maximum	31
Average	15.817
Standard Deviation	8.363

7-Summary statistics of attribute “default”

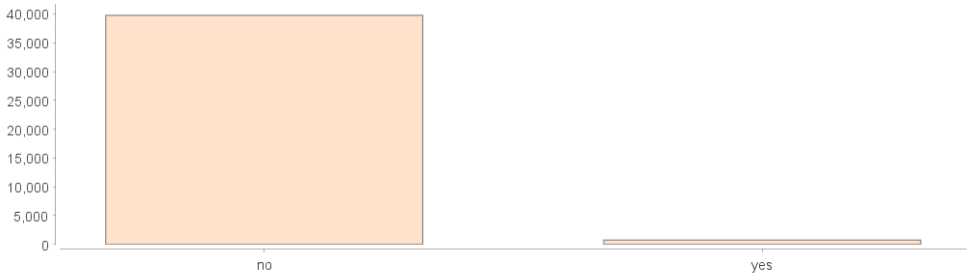
default

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.00%
Stability: 98.05%
Valid: 1.94%

Top Values



2 Distinct Values:

Value	Count	Percentage
no	39,775	98.14%
yes	754	1.86%

8-Summary statistics of attribute “duration”

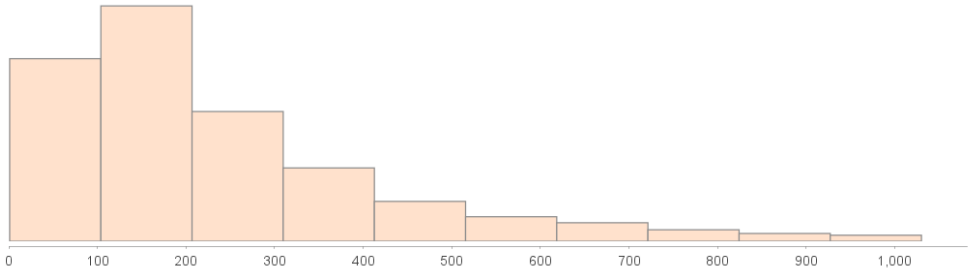
duration

Summary

Number

Missing: 0.00%
Infinite: 0.00%
ID-ness: 2.25%
Stability: 0.53%
Valid: 97.22%

Distribution



Statistics

Name	Value
Minimum	0
Maximum	1030
Average	236.305
Standard Deviation	190.678

9-Summary statistics of attribute “education”

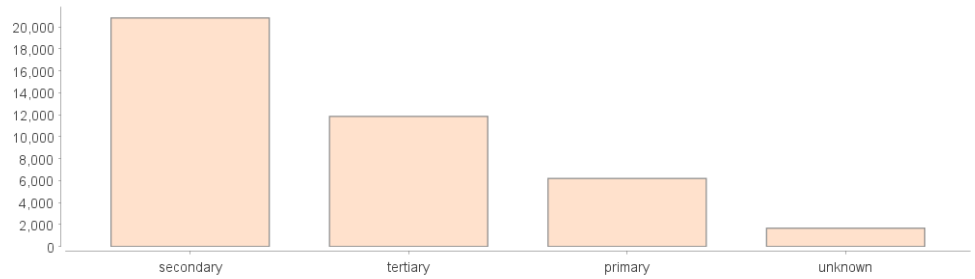
education

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.01%
Stability: 50.00%
Valid: 49.99%

Top Values



4 Distinct Values:

Value	Count	Percentage
secondary	20,810	51.35%
tertiary	11,849	29.24%
primary	6,200	15.30%
unknown	1,670	4.12%

10-Summary statistics of attribute “housing”

housing

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.00%
Stability: 54.00%
Valid: 45.99%

Top Values

2 Distinct Values:

Value	Count	Percentage
yes	22,149	54.65%
no	18,380	45.35%

11-Summary statistics of attribute “job”

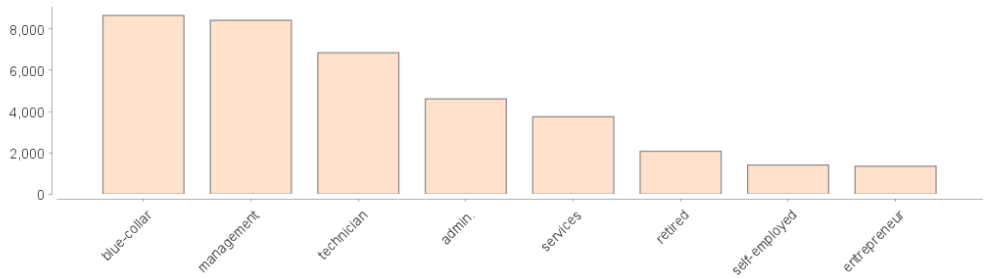
job

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.03%
Stability: 21.76%
Valid: 78.21%

Top Values



12-Summary statistics of attribute “loan”

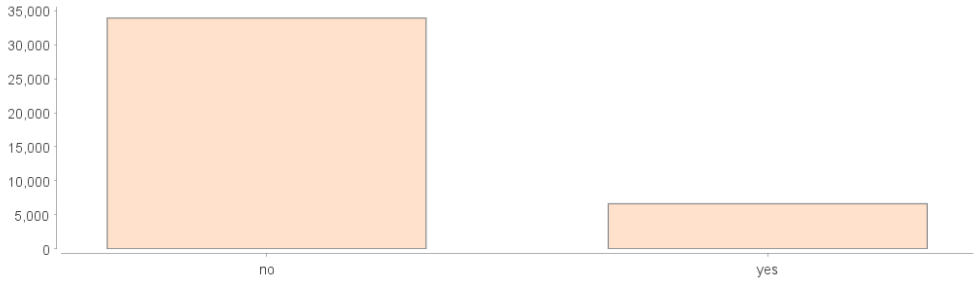
loan

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.00%
Stability: 83.76%
Valid: 16.23%

Top Values



2 Distinct Values:

Value	Count	Percentage
no	33,906	83.66%
yes	6,623	16.34%

13-Summary statistics of attribute “marital”

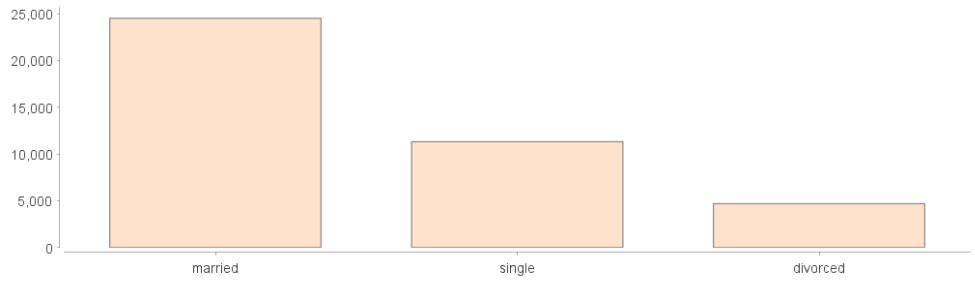
marital

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.01%
Stability: 60.48%
Valid: 39.52%

Top Values



3 Distinct Values:

Value	Count	Percentage
married	24,526	60.51%
single	11,319	27.93%
divorced	4,684	11.56%

14-Summary statistics of attribute “month”

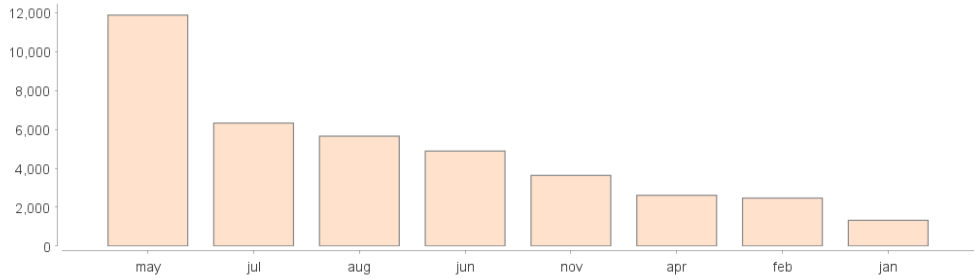
month

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.03%
Stability: 28.73%
Valid: 71.24%

Top Values



15-Summary statistics of attribute “pdays”

pdays

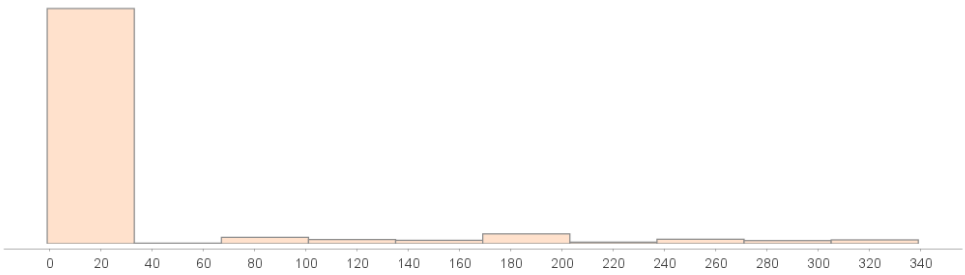
Data statistics

Summary

Number

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.70%
Stability: 85.69%
Valid: 13.62%

Distribution



Statistics

Name	Value
Minimum	-1
Maximum	339
Average	25.180
Standard Deviation	71.411

16-Summary statistics of attribute “poutcome”

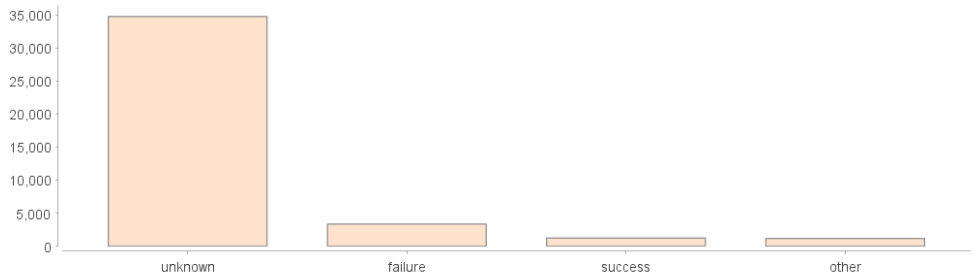
poutcome

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.01%
Stability: 85.69%
Valid: 14.30%

Top Values



4 Distinct Values:

Value	Count	Percentage
unknown	34,751	85.74%
failure	3,360	8.29%
success	1,239	3.06%
other	1,179	2.91%

17-Summary statistics of attribute “previous”

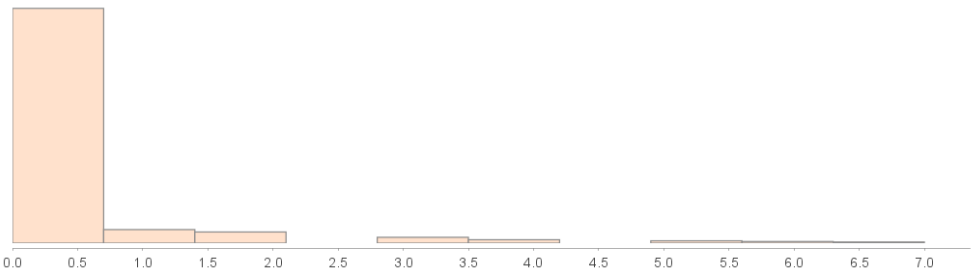
previous

Summary

Number

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.02%
Stability: 85.69%
Valid: 14.29%

Distribution



Statistics

Name	Value
Minimum	0
Maximum	7
Average	0.355
Standard Deviation	1.060

18-Summary statistics of attribute “y”

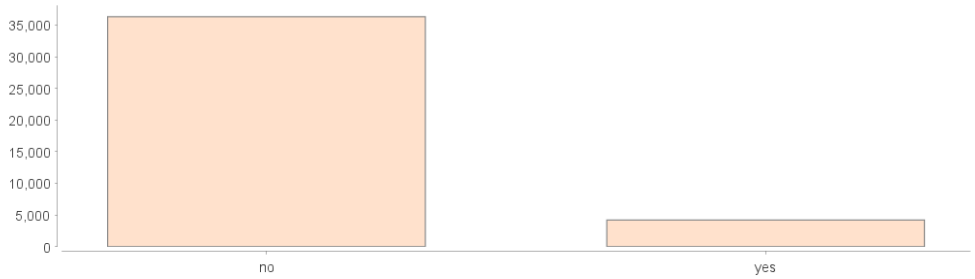
y

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.00%
Stability: 89.54%
Valid: 10.46%

Top Values



2 Distinct Values:

Value	Count	Percentage
no	36,328	89.63%
yes	4,201	10.37%

19-Python code for confusion matrix GBM Model

```python
import pandas as pd
from sklearn.metrics import confusion_matrix
file_path = 'C:/Users/sefon/OneDrive/Documents/WTAMU/Spring 2024/data mining/Class project/Final/results.xlsx'
df = pd.read_excel(file_path)
print(df.head())
true_labels = df['True answers'].values
predicted_labels_dt = df['Gradient Boosted Model - RM'].values
confusion_matrix_dt = confusion_matrix(true_labels, predicted_labels_dt)
print('Confusion Matrix for Gradient Boosted Model - RM')
print(confusion_matrix_dt)
```