Machine Learning: Bank Term Deposit Subscription

Students:
LE Thi Hoa - 12310380
TRAN Hai Linh - 12310487

Supervised by:
Assoc.Prof.Cazabet Rémy

17, December 2023

I. Introduction

The dataset is related to direct marketing campaigns of a Portuguese bank. This dataset, titled Direct Marketing Campaigns for Bank Term Deposits, is a collection of data related to the direct marketing campaigns conducted by a Portuguese banking institution. These campaigns primarily involved phone calls with customers, and the objective was to determine whether or not a customer would subscribe to a term deposit offered by the bank.

II. Dataset

The classification goal is to predict if the client will subscribe to a term deposit (variable y).

The dataset comprises 45,211 rows and 17 columns:

- Age (Numeric): The age of the client.
- Job (Categorical): The type of job the client is engaged in
- Marital Status (Categorical): The marital status of the client
- Education (Categorical): The highest education level attained by the client
- Default (Binary): Indicates whether the client has credit in default
- Balance (Numeric): The average yearly balance in euros.
- Housing Loan (Binary): Indicates whether the client has a housing loan
- Personal Loan (Binary): Indicates whether the client has a personal loan
- Contact (Categorical): The communication type used to contact the client
- Day (Numeric): The day of the month when the last contact was made.
- Month (Categorical): The month of the year when the last contact was made, using abbreviated names
- Campaign (Numeric): The number of contacts performed during the current campaign for this client.
- Duration (Numeric): The duration of the last contact in seconds.
- Pdays (Numeric): The number of days that passed since the client was last contacted in a previous campaign. (-1 indicates the client was not previously contacted.)

- Previous (Numeric): The number of contacts performed before the current campaign for this client.
- Poutcome (Categorical): The outcome of the previous marketing campaign
- Y (Binary): Indicates whether the client subscribed to a term deposit, with values "yes" or "no.

Summary statistics of numeric columns:

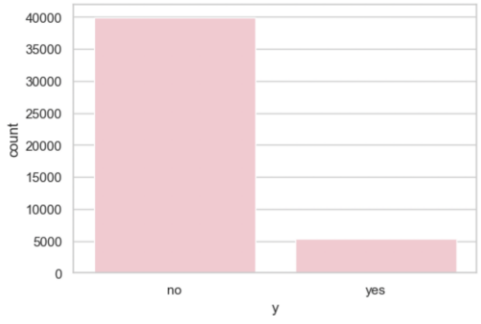
	age	balance	day	duration	campaign	pdays	previous
count	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000
mean	40.936210	1362.272058	15.806419	258.163080	2.763841	40.197828	0.580323
std	10.618762	3044.765829	8.322476	257.527812	3.098021	100.128746	2.303441
min	18.000000	-8019.000000	1.000000	0.000000	1.000000	-1.000000	0.000000
25%	33.000000	72.000000	8.000000	103.000000	1.000000	-1.000000	0.000000
50%	39.000000	448.000000	16.000000	180.000000	2.000000	-1.000000	0.000000
75%	48.000000	1428.000000	21.000000	319.000000	3.000000	-1.000000	0.000000
max	95.000000	102127.000000	31.000000	4918.000000	63.000000	871.000000	275.000000

The data exhibits significant variability in certain variables such as Balance and the number of Contacts in the Campaign. Information regarding the duration of calls appears to be diverse, ranging from very short to very long. Data pertaining to the number of days since the last call (Pdays) is predominantly around the average, but there are instances with a large number of days. For several variables, there is a substantial difference between the minimum and maximum values, indicating diversity in the data.

The number of unique values in each column of data type object (categorical) in the data set:

	job	marital	education	default	housing	loan	contact	month	poutcome	y
Count	12	3	4	2	2	2	3	12	4	2

Y Value



The target variable 'y' indicates whether customers subscribed to a term deposit. There are fewer instances of 'yes' compared to 'no' in the dataset.

III. Data Cleaning

The data consists of 45,211 records and 17 attributes. After checking for duplicate values and missing values, fortunately, there are no duplicates or missing values in the data.

Records with implausible values or errors are removed based on examining the distribution of each attribute and common sense to ensure the dataset's consistency and correctness.

After examining outliers in columns, the columns 'age,' 'duration,' and 'campaign' contain a considerable number of outliers, with 'duration' in particular showcasing a prominent distribution of outliers. To remove outliers in these columns, the Interquartile Range (IQR) method was applied.

The 'balance' column exhibits a substantial number of outliers. However, in the context of banking and financial data, these values might be normal if there are customers with exceptionally high account balances. Preserving this information could be vital to retain crucial insights about customers.

Duration represents the last contact duration in seconds (numeric). To ensure the data integrity and meaningful analysis, it was deemed necessary to exclude rows with a duration value less than or equal to zero.

Finally, after data cleaning, the dataset comprises 34,717 rows and 17 columns.

IV. Explore Data Analysis

For the convenience of analysis, we convert the values in the 'y' column to numeric format, where 'yes' is represented as 1, and 'no' is represented as 0

Convert the values in the 'pdays' column into two values yes-no: If 'pdays' is -1, it may signify that the customer has not been contacted (No), while other values may represent customers who have been contacted before (Yes)

Data distribution

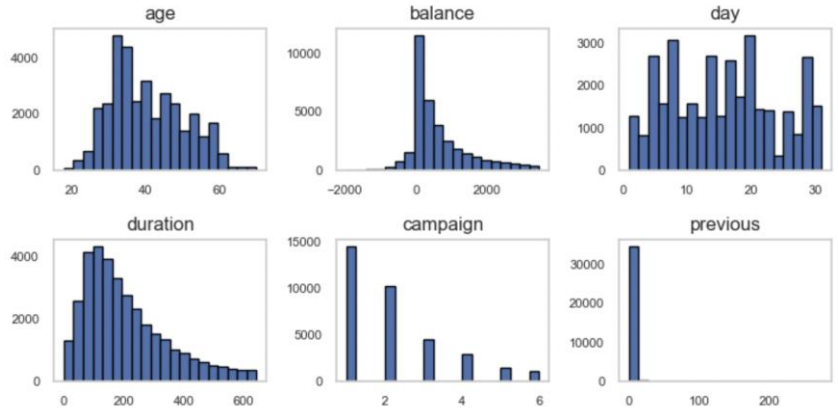
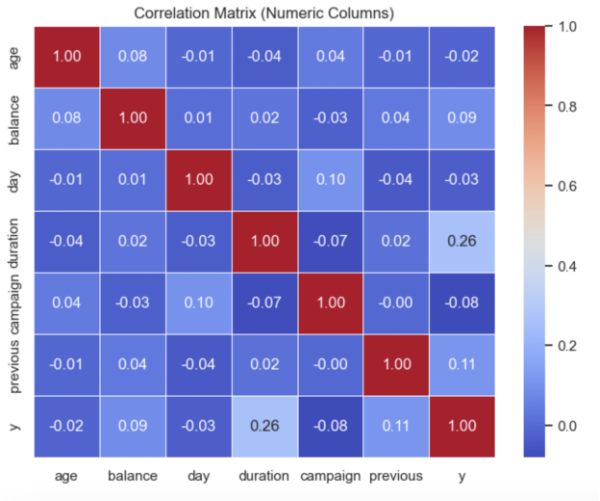


Figure 1: Distribution of numeric features

In figure 1, the age distribution is primarily centered between 32 and 48 years, with a significant concentration of customers in this range. There is a wide range of balance values, with a notable number of zero balances. Call volume increases over the weekend

(Friday to Monday), with a peak on the 20th. The duration distribution provides counts for different durations of calls. Most customers are part of one or two campaigns.

The degree of correlation between pairs of numeric variables



Then, we will investigate the correlation between continuous variables. As shown in figure 2, The **"duration"** and "**balance**" features stand out with a notable positive correlation, indicating its influence on the target variable.

Figure 2: Correlation between Continuous Attributes

Categorical Variables Impact on Subscription Registration

Please refer to the Appendix for figures

- Job vs. Y: Registrations are higher for management, technician, and admin. Conversely, blue-collar, services, and manual jobs exhibit lower registration rates.
- Marital vs. Y: Married individuals show a higher registration count compared to singles and divorced individuals.
- Education vs. Y: Those with tertiary education demonstrate a higher registration rate compared to other education groups.
- Default vs. Y: Clients with no defaults tend to have a higher registration count.
- Housing vs. Y: Customers without a housing loan exhibit a higher registration rate compared to those with a housing loan.
- Loan vs. Y: Customers without a personal loan tend to register more.
- Contact vs. Y: Registration through mobile (cellular) has a higher rate compared to other contact methods.
- Month vs. Y: There is variation in registration counts across months, with May having the highest registration count.
- Poutcome vs. Y: The "success" outcome of the previous marketing campaign has a higher registration rate.
- Pdays_bin: Data suggests that previous contact in a campaign (pdays_bin = 1) is associated with a higher likelihood of subscription (y = 1). Clients who were not previously contacted (pdays_bin = 0) are more likely to not subscribe (y = 0).

Please refer to the Appendix for additional figures illustrating the relationships between call duration and job, education, marital status with the 'y' value.

V. Feature Engineering

Feature engineering is the process of transforming raw data into useful features to get the most out of your data. Down below is a list to briefly summarize what we did:

- In the "poutcome" column, there are two values: "other" and "unknown". Replace the "unknown" values with "other".
- Encode categorical variables into integer values for machine learning model compatibility
- Apply a logarithmic transformation to the "duration" column. Use the natural logarithm function (log1p) on the values in the "duration" column to reduce the skewness of the distribution
- Normalizing numeric objects with StandardScaler helps normalize the range of numeric objects, ensuring that they are on the same scale.
- Apply an oversampling method to address class imbalance in the target variable.

VI. Feature Selection

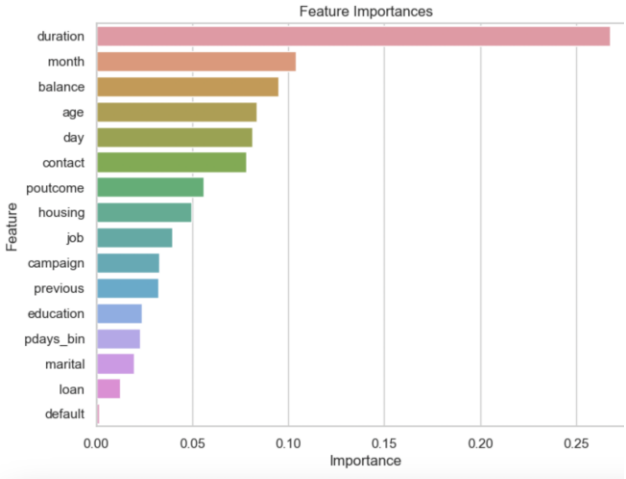


Figure 3: Plot of Feature Importances

We perform feature importance analysis using Random Forest Classifier. From figure 3, The "duration" feature has the most significant impact on the model, with the highest importance value. "month", "balance", "day" and "age" also contribute significantly to the predictions. Features such as "default" and "loan" have low importance values, indicating they have less influence on the model.

We perform the removal of columns that have little impact on the target variable, namely "loan" and "default".

VII. Model Selection

This section will build the models to predict if the client will subscribe to a term deposit (variable y). Model fit is evaluated using the train-test split method to randomly split sampled data into 80% training set and 20% testing set.

We experiment with a range of popular classification models: **Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), Decision Tree**

We utilize the following evaluation metrics: **Accuracy Rate** - The ratio of correct predictions on the dataset. **ROC AUC Score** - The area under the ROC curve, measuring the ability to discriminate between classes.

	Model	Accuracy_rate_train	Roc_auc_rate_train	Accuracy_rate_test	Roc_auc_rate_test
0	Logistic_reg	0.794122	0.794122	0.781826	0.788518
1	RandomForest	1.000000	1.000000	0.934764	0.708988
2	Knn	0.964580	0.964580	0.879032	0.757716
3	DesionTree	1.000000	1.000000	0.908842	0.692495

Table: Results of evaluating the performance of the models

The Logistic Regression model shows relatively stable performance on both the training and test sets. The Random Forest model seems to overfit the training data, as the test accuracy is significantly lower than the training accuracy. The substantial drop in ROC AUC on the test set is also a concerning sign. The KNN model performs well, but there is a slight decrease in accuracy on the test set compared to the training set. The ROC AUC also decreases, but it remains relatively stable. The Decision Tree model exhibits good performance on the training set, but there are signs of overfitting and a decrease in performance on the test set. The ROC AUC on the test set also drops significantly. Logistic Regression seems to be the most balanced model for this classification task

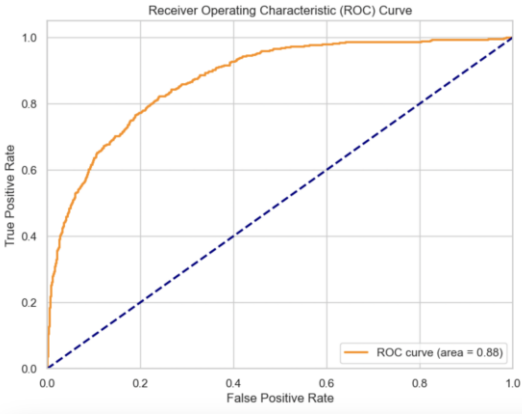


Figure 4: Evaluate Logistic Regression Model Performance Using ROC Curve

VIII. Conclusion

Based on the conclusions from analysis and modeling, duration is an important factor affecting the results. Some suggestions to improve marketing strategy:

- Focus on the timing of contacts: Enhance the communication strategy during peak subscription periods, such as weekends or around the 20th day of each month (May-July)
- Optimize the number of contact attempts within a campaign to maintain campaign effectiveness without causing inconvenience to customers.
- Intensify the use of mobile communication channels, particularly cellular, as they demonstrate higher subscription rates compared to other channels.
- Gain additional insights into customers' occupations, marital status, and education to develop a personalized marketing strategy based on their individual characteristics.