Bank Marketing (Campaign)

Neural Nomad

Problem Description

ABC Bank seeks to develop a machine learning model capable of predicting whether a customer will subscribe to their term deposit product based on past interactions. The objective is to identify customers with a higher likelihood of purchasing the product, enabling the bank to create a targeted marketing strategy. By leveraging predictive analytics, the bank can efficiently shortlist potential customers and optimize its outreach efforts.

Business Understanding

The implementation of a predictive machine learning model will provide insights into the characteristics of customers who are most likely to subscribe to the term deposit product. This will allow the bank to focus its marketing efforts on high-probability customers, leading to more efficient resource allocation, reduced operational costs, and improved profitability. By refining its customer targeting strategy, ABC Bank can enhance campaign effectiveness and maximize return on investment (ROI) for future marketing initiatives.

Data Understanding

The dataset utilized for this analysis, "bank-additional-full.csv," consists of 41,188 observations and 21 features. These features encompass a wide range of client demographic, financial, and marketing-related attributes, including:

Customer information: Age, job, marital status, education, credit default status, housing loan, and personal loan.

Contact details: Communication type, last contact month and day, contact duration, and number of previous contacts.

Marketing campaign metrics: Outcome of previous campaigns, employment variation rate, consumer price index, consumer confidence index, Euribor 3-month rate, and total number of employees.

Target variable ("y"): This binary variable indicates whether a customer has subscribed to a term deposit ("yes" or "no") and will serve as the key outcome variable for model training and evaluation.

This dataset will be used to train and validate the predictive model, ensuring it can accurately identify potential customers for the bank's future marketing campaigns.

Understanding the Dataset

Name	Type	About
age	Numeric	Age of the customer
job	Categorical	Customer's occupation
marital	Categorical	Customer's marital status
education	Categorical	Customer's education background
default	Categorical	If customer has credit in default
housing	Categorical	If customer has housing loan
loan	Categorical	If customer has personal loan
contact	Categorical	Customer's contact type
month	Categorical	Customer's last month of contact
day_of_week	Categorical	Customer's last weekday of contact

Understanding the Dataset

Name	Type	About
duration	Numeric	Customer's last contact duration(s)
campaign	Numeric	# of contacts during this campaign
pdays	Numeric	Number of days that passed by after the client was last contacted
previous	Numeric	Number of contacts performed before this campaign and for this client
poutcome	Categorical	Outcome marketing campaign
emp.var.rate	Numeric	Employment variation rate quarterly
cons.price.idx	Numeric	Consumer price index – monthly
cons.conf.idx	Numeric	Consumer confidence index – monthly
euribor3m	Numeric	Euribor 3 months rate – daily
nr.employed	Numeric	Number of employees - quarterly

TABULAR VIEW

Feature Name	Type	Data Type	# of Null or "Unknown"	# of Outliers	Comments
age	Numerical	int	0	0	Drop missing values
job	Categorical	str	330	0	Drop missing values
marital	Categorical	str	80	0	Drop missing values
education	Categorical	str	1731	0	*Two options: leave unknown as its own class or use a classification ML model on this feature to fill in the unknown data.*
default	Categorical	str	8597	0	
housing	Categorical	str	990	0	Replace with Mode
loan	Categorical	str	990	0	Replace with Mode
contact	Categorical	str	0	0	
month	Categorical	str	0	0	
year	Numerical	int	0	0	
day_of_week	Categorical	str	0	0	
duration	Numerical	int	0	1045	Using an upper bound defined as Q3+3*IQR to remove outliers
campaign	Categorical	str	0	0	
pdays	Numerical	int	0	0	
previous	Numerical	int	0	0	
poutcome	Categorical	str	0	0	
emp.var.rate	Numerical	float	0	0	
cons.price.idx	Numerical	float	0	0	
cons.conf.idx	Numerical	float	0	0	
euribor3m	Numerical	float	0	0	
nr.employed	Numerical	float	0	0	
y	Categorical	str	0	0	

Key Questions:

1.What challenges exist in the dataset?

1. Are there missing values (NA)?
2. Are there outliers or skewed distributions?
3. Any other inconsistencies that may affect analysis?

2.What strategies are being implemented to address these challenges?

1. How do you plan to handle missing values?
2. What techniques will you use to identify and manage outliers?
3. Why have you chosen these specific approaches?

Identified Issues in the Dataset

The dataset contains six categorical features with missing values: job, education, marital, default, housing, and loan. Additionally, the "duration" feature exhibits extreme outliers, with a mean of approximately 258 but a maximum value of 4,918, indicating significant variability. Furthermore, the dataset is highly imbalanced, with approximately 90% of instances classified as "No," making it challenging for a predictive classification model to accurately learn minority class patterns.

Approaches to Address Data Challenges

To manage missing values, a tailored strategy will be applied based on the severity and importance of each feature.

- Dropping missing values for features with a small number of unknown entries (job, marital).
- Replacing missing values with the most frequent category (housing, loan).
- Using a machine learning classification model to predict and fill missing values for the default and education features.

For handling outliers in numerical data, we will apply the interquartile range (IQR) method, setting the upper fence at Q3 + 3*IQR to retain approximately 97% of the original data while filtering extreme values.

To address data imbalance, appropriate evaluation metrics will be used to improve model performance.

- The AUROC curve will help assess model effectiveness in distinguishing between True Positives and False Negatives.
- Under-sampling of the majority class will be considered to balance representation in the dataset.
- During data splitting, we will ensure that the minority class is maintained in each fold rather than applying full randomization.
- Adjusting the rare-to-majority case ratio in the training data will help the model better learn patterns from the underrepresented class.

These methods aim to improve model robustness, minimize bias, and enhance prediction accuracy for term deposit subscriptions.

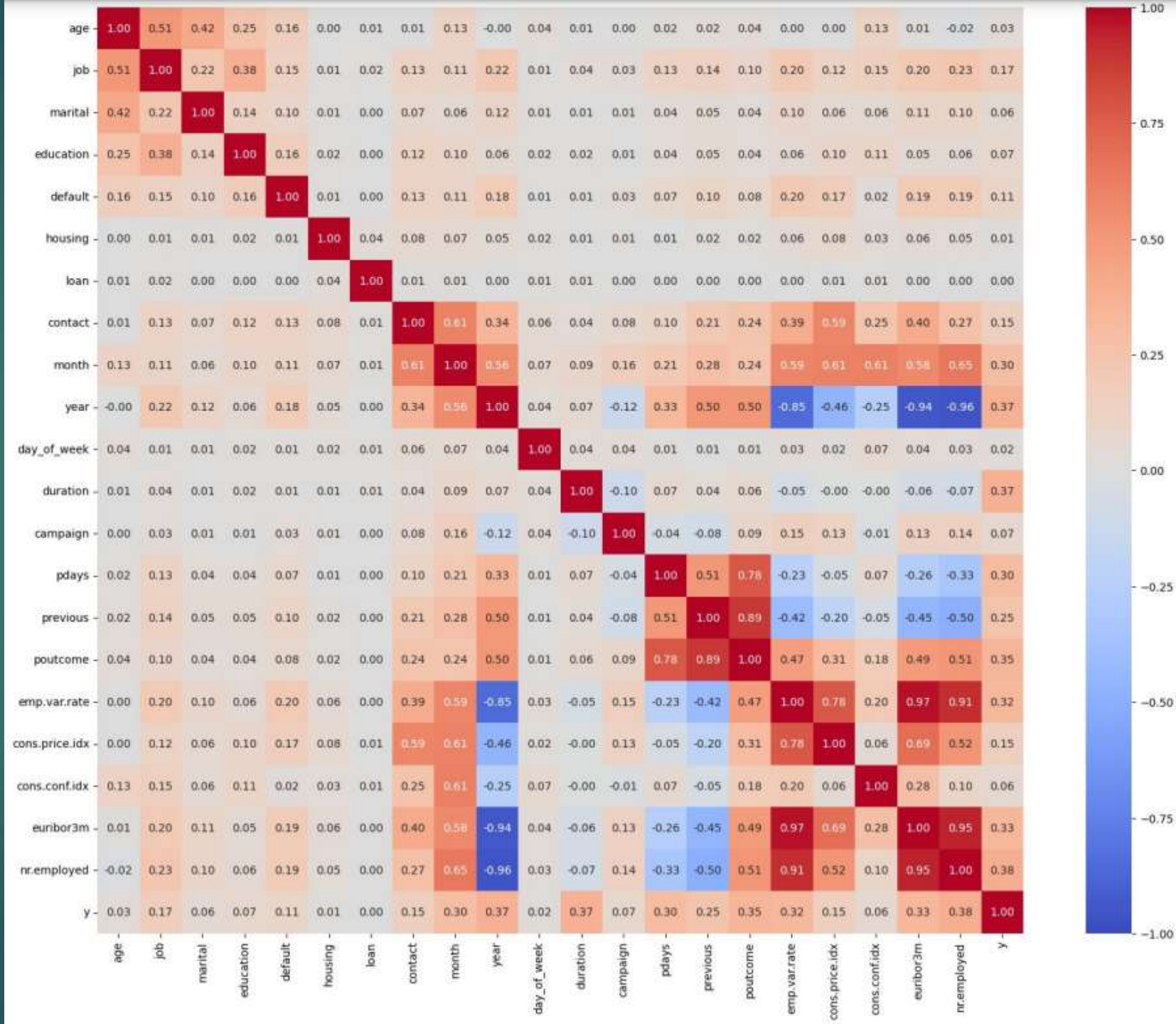
Exploring the Data

Data exploration involves analyzing the dataset to identify patterns, trends, and relationships among variables. This process helps uncover correlations, distributions, and anomalies that may affect model performance.

By using statistical and visualization techniques, we can determine which features are most relevant for training the machine learning model. Selecting the most important variables improves model accuracy, reduces complexity, and enhances overall predictive performance.

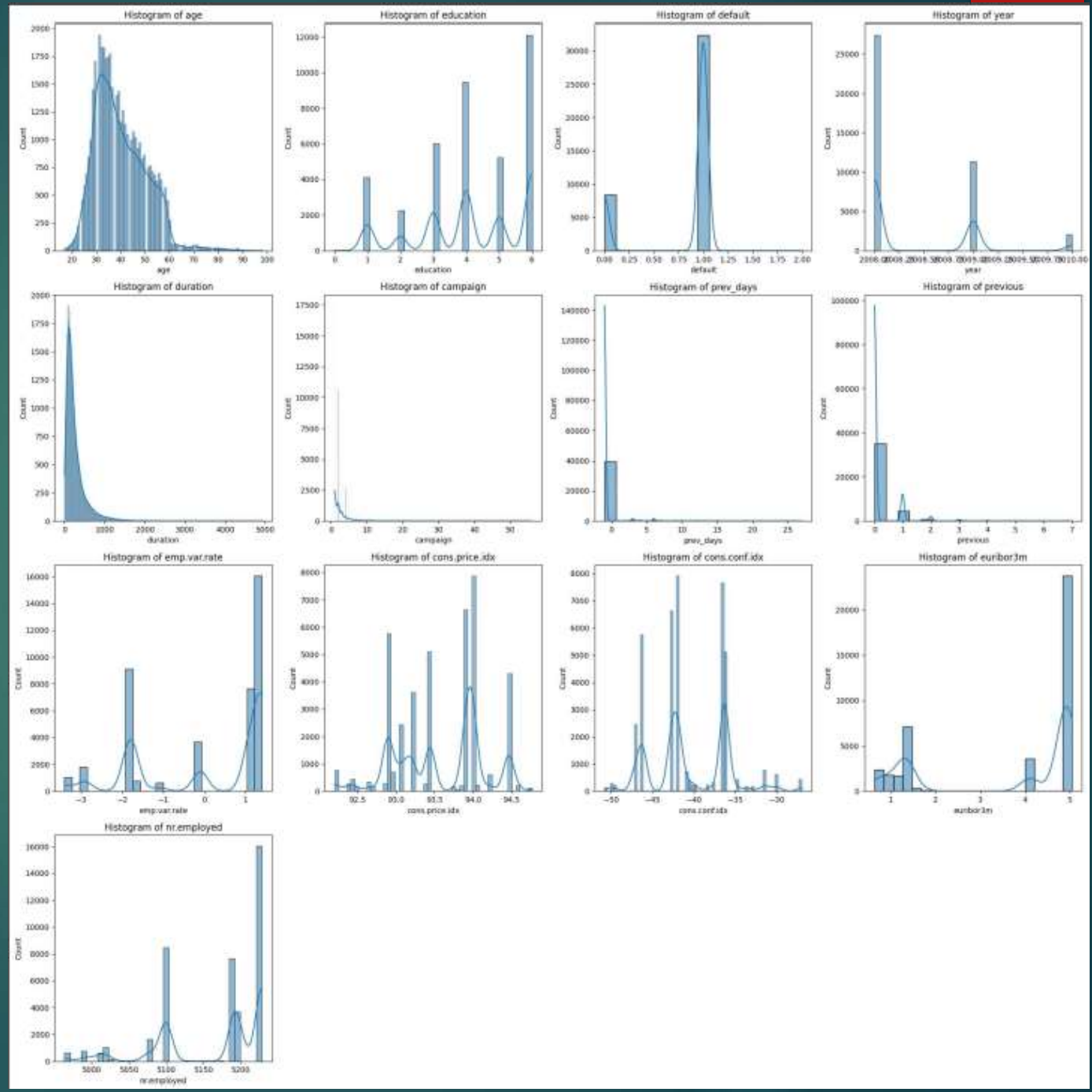
Correlation Heat Map

The figure on the right illustrates the correlation strength between different features within the dataset. This visualization helps in understanding the relationships among variables. For our analysis, the primary focus is on identifying features that have the strongest correlation with the target variable 'y,' as these are likely to be the most influential in predicting customer behavior.



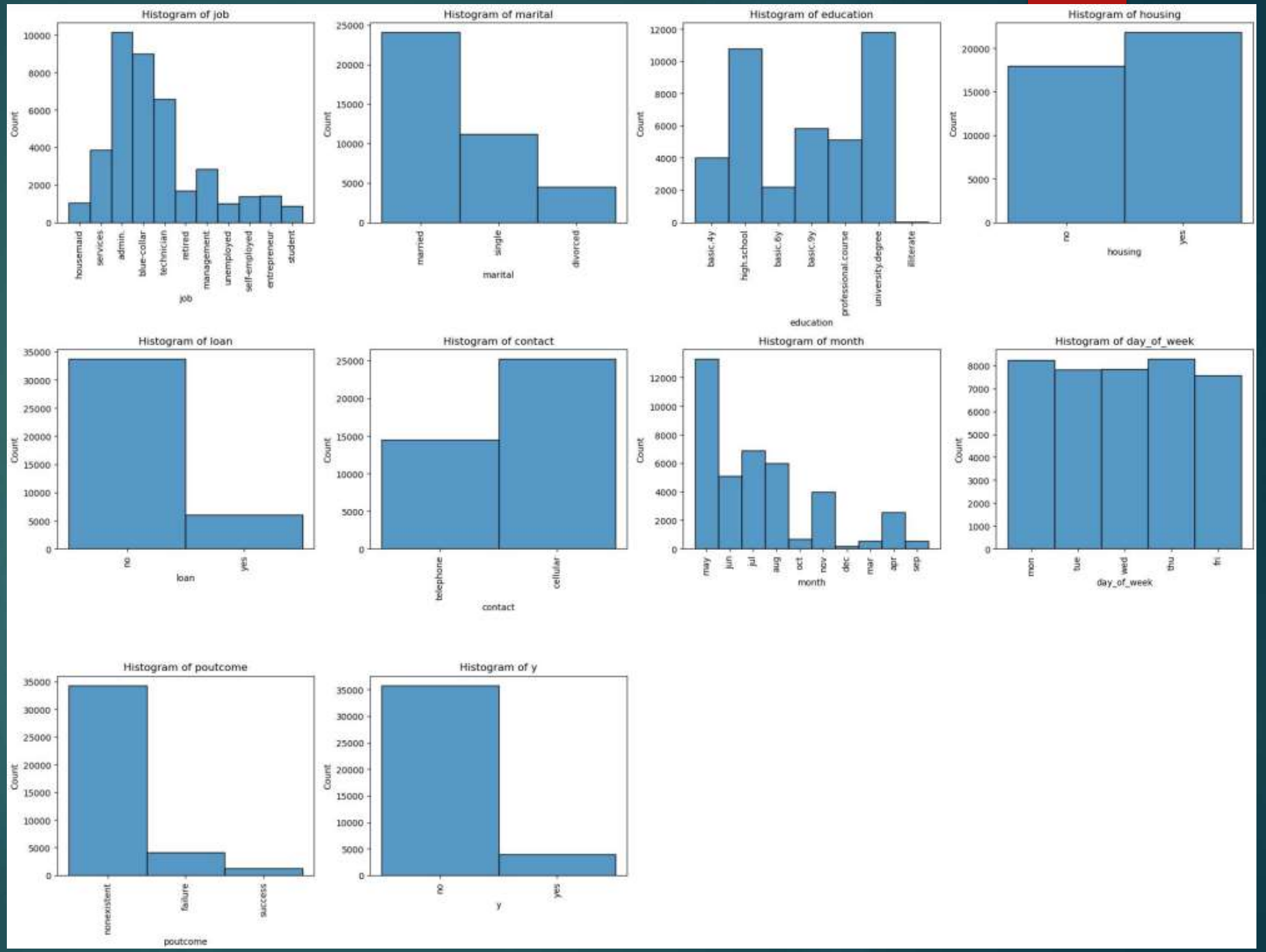
Numerical Features

To gain a clearer understanding of the range, distribution, and potential biases within each feature, we generated histograms for the entire dataset. The figures on the left provide a visual representation of the numerical data, helping to identify patterns, outliers, and overall data distribution.



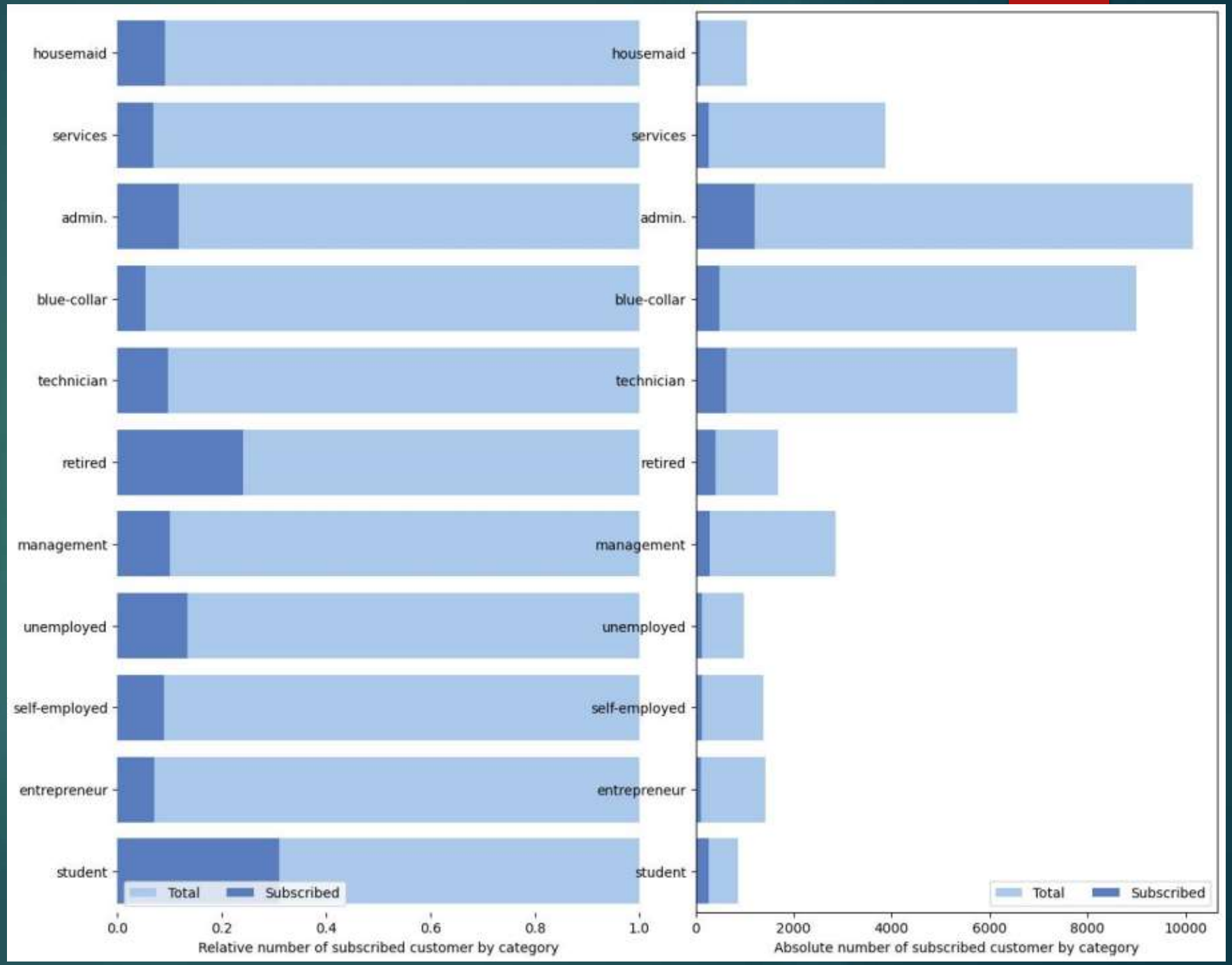
Categorical Features

Like the previous slide, these charts display histograms representing the categorical features in the dataset. A key observation is the significant imbalance in the target variable 'y,' where the 'yes' cases make up only about 10% of the total. This imbalance is crucial to consider when training the predictive model, as it may impact model performance and necessitate techniques to address class distribution.



Relative vs. Absolute Sales per Category

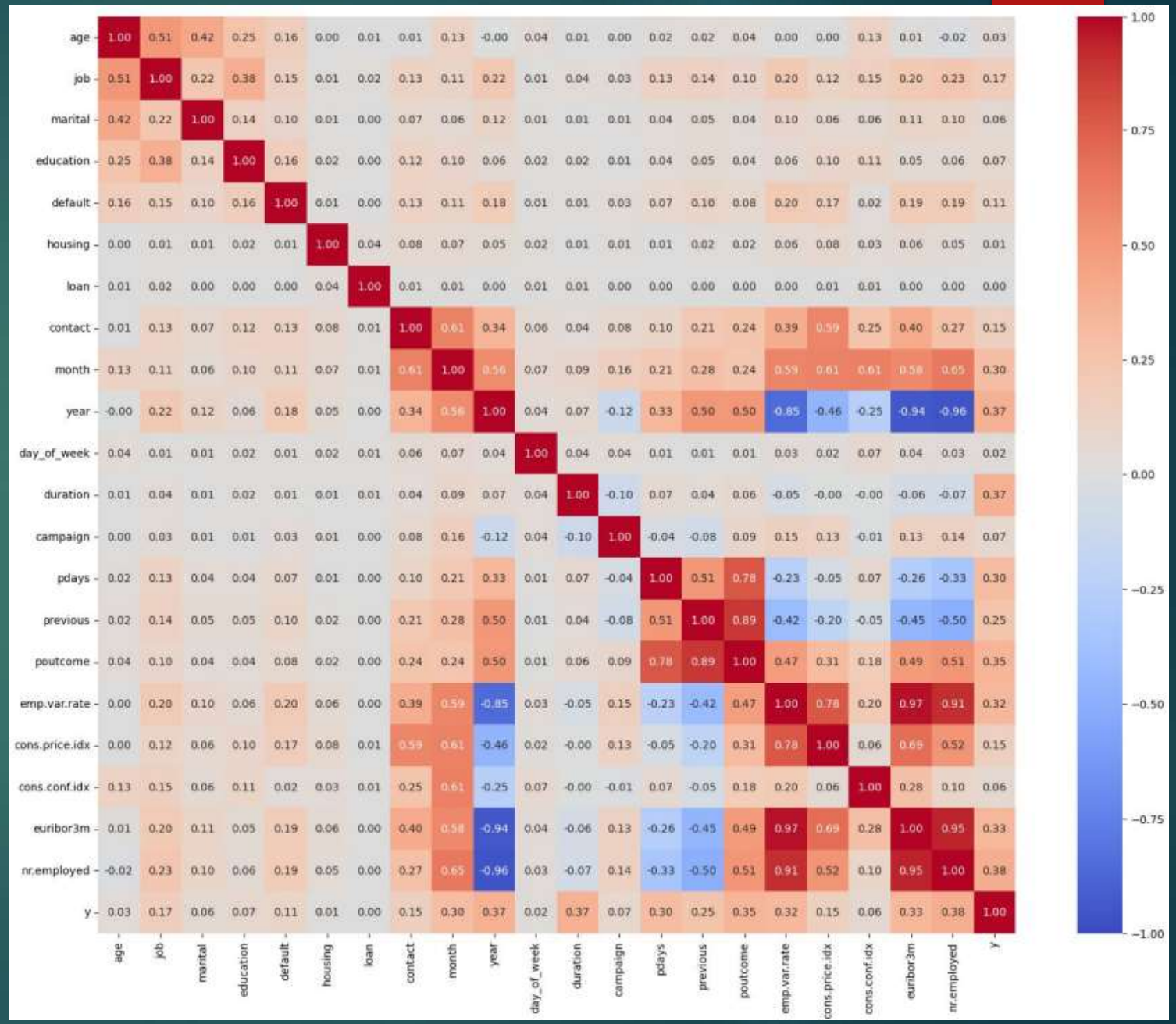
Analyzing categorical features in greater detail reveals categories that may be underrepresented in the dataset, such as 'student,' yet demonstrate a high relative sales conversion rate. This insight helps identify potentially valuable customer segments that, despite lower representation, contribute significantly to successful conversions.

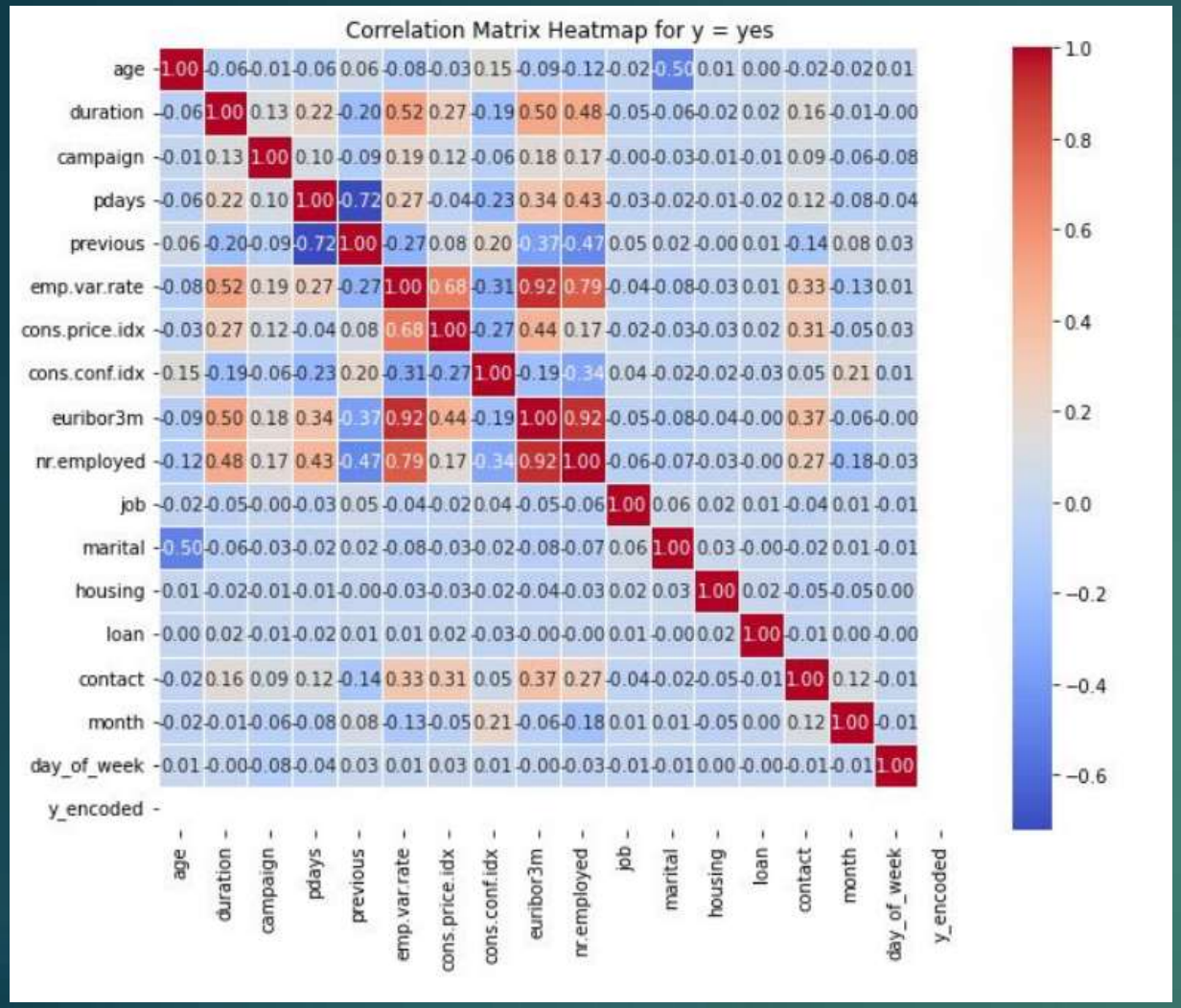


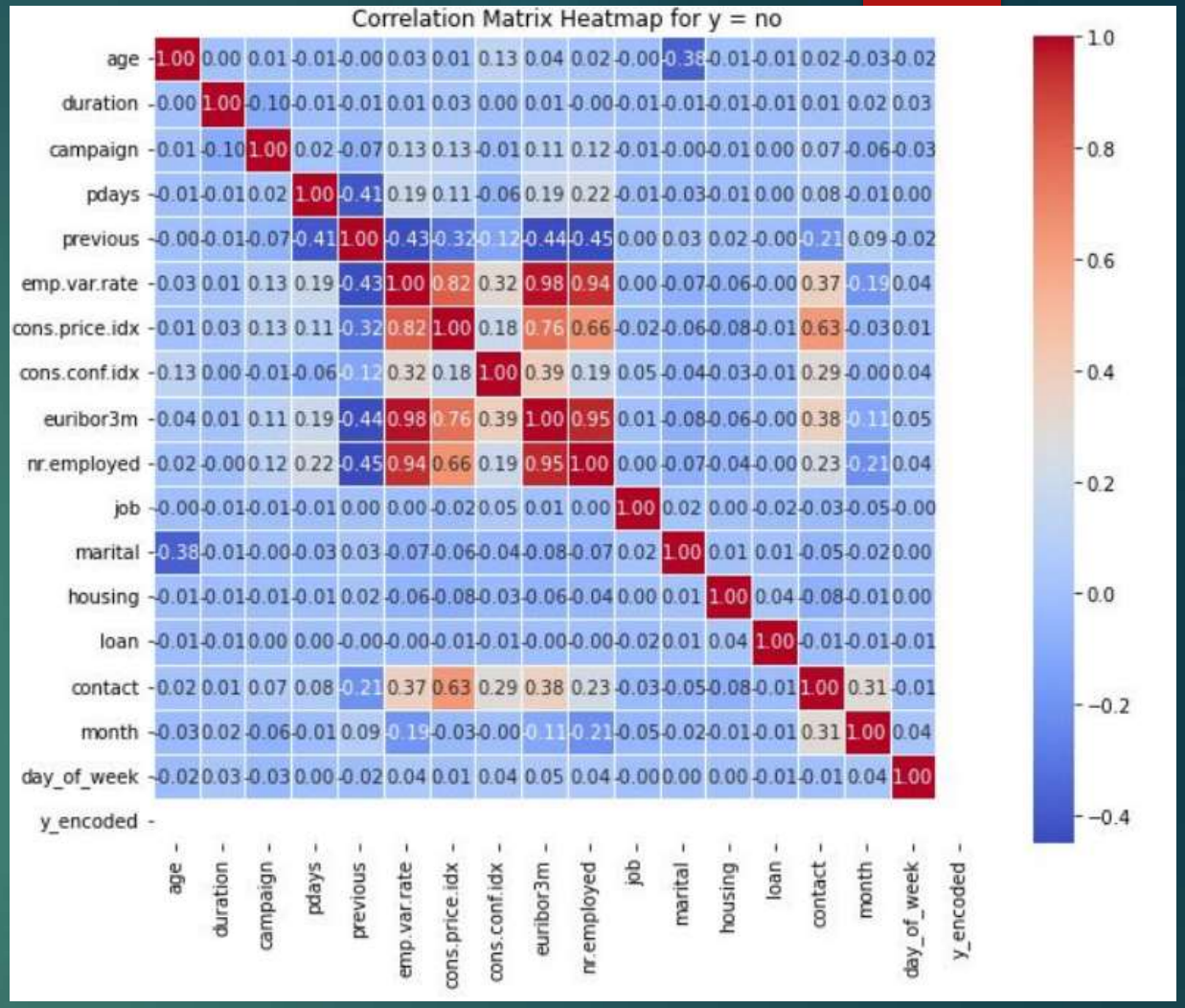
Recommendations

In the next section, I will identify the variables that exhibit the strongest correlation with the target variable 'y.' Based on the visualizations, we will provide insights and recommendations on how these key features influence customer behavior and their potential impact on predictive modeling.

From the heatmap, it is evident that most variables have a weak correlation with the target variable 'y,' with correlation values close to zero. However, certain features, such as 'year,' 'duration,' 'pdays,' and 'poutcome,' show a stronger relationship with 'y.' This outcome is reasonable, as variables like 'duration' play a significant role in influencing a customer's likelihood of subscribing to the term deposit.



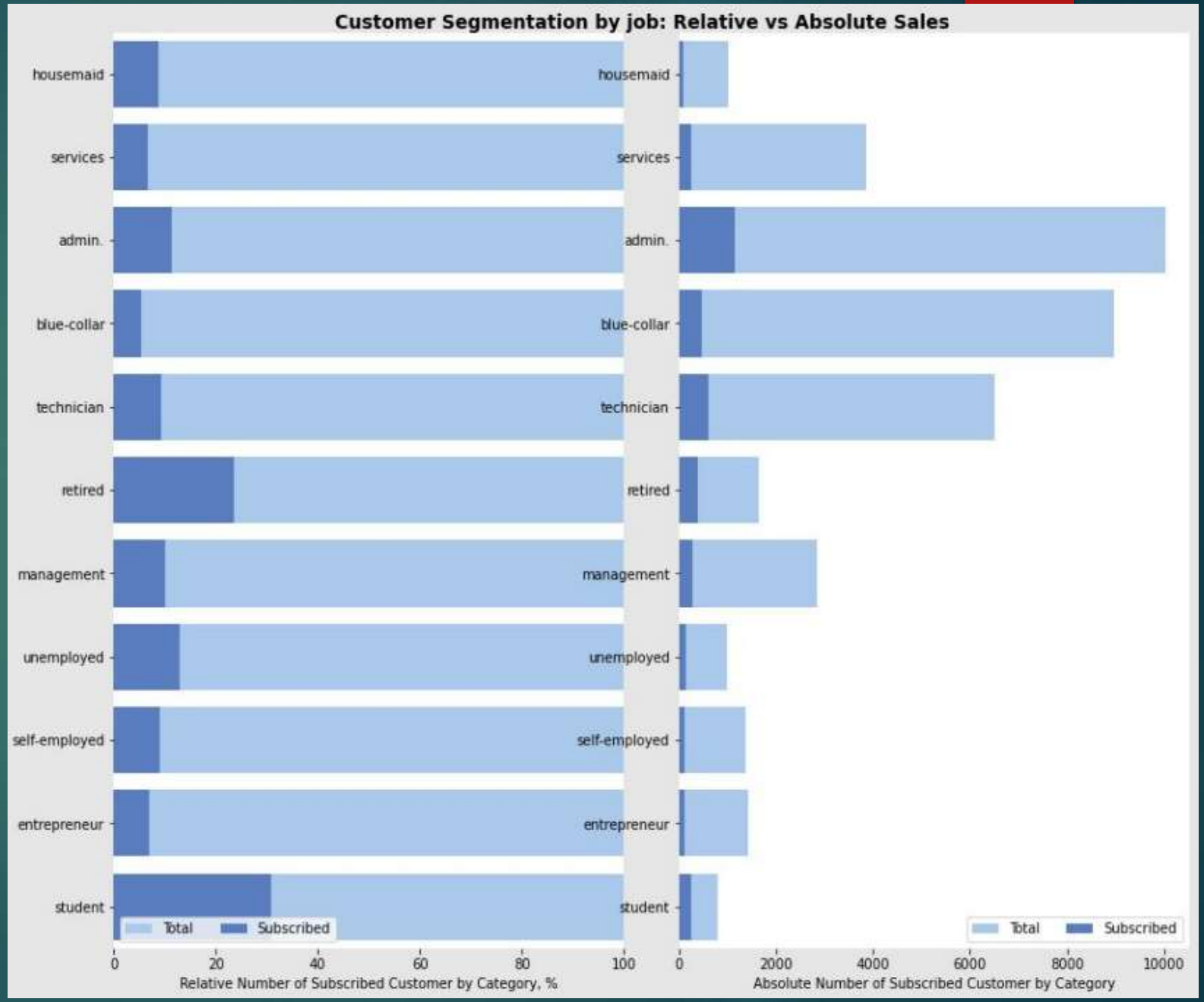




When splitting the data into 'yes' and 'no' cases, we observe that the correlation values remain close to zero for both groups. This indicates a weak relationship between most features and the target variable 'y' when analyzed separately.

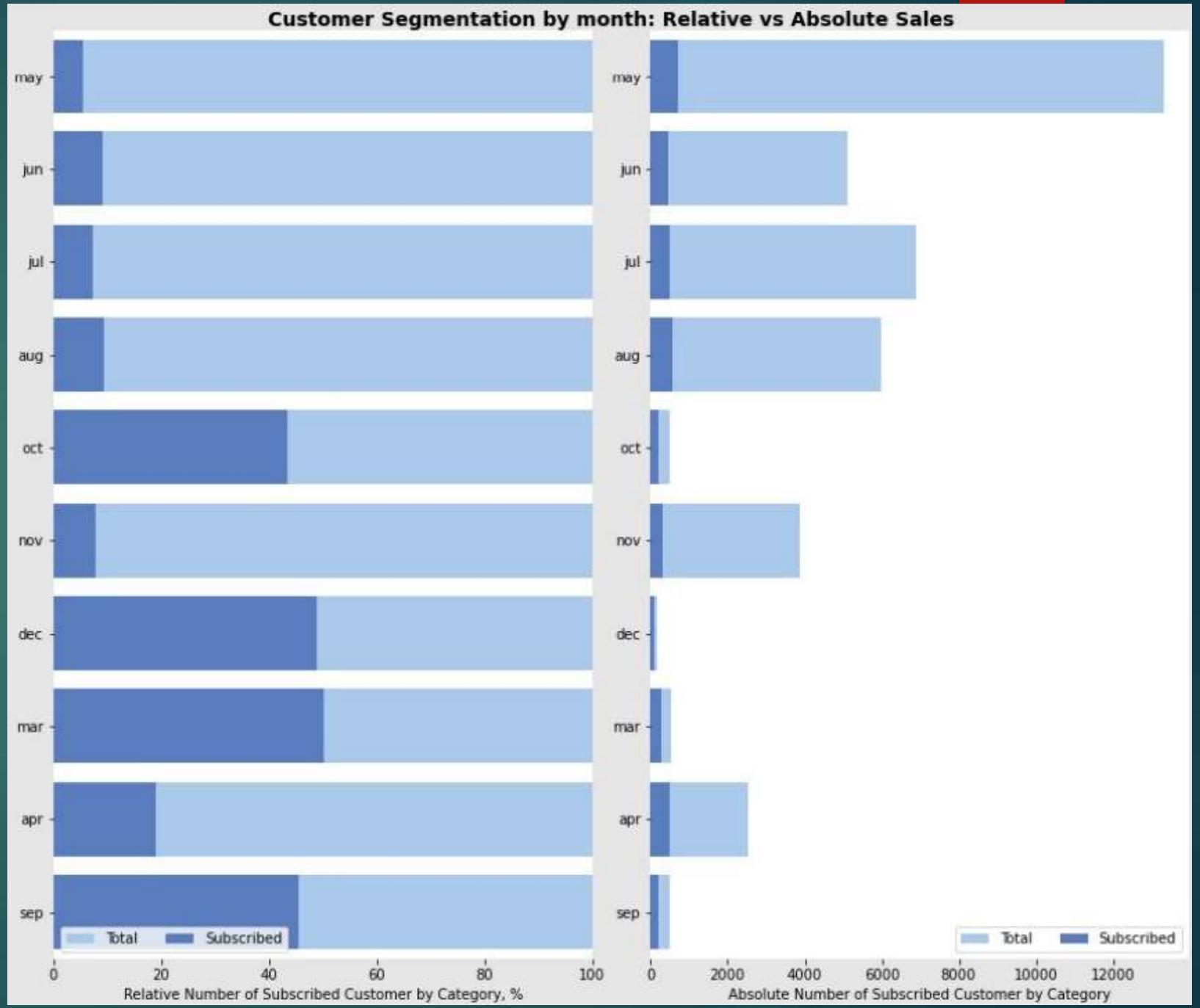
Improvements – Jobs

Analyzing the distribution of categorical variables reveals potential areas for improvement. Specifically, students and retirees have a higher relative subscription rate compared to other groups. However, their absolute number of subscriptions remains low due to fewer phone calls made to them. Increasing the number of calls targeted at these groups could potentially lead to a higher overall subscription rate.



Improvements – Months

Another area for improvement is the timing of calls. The data shows that months such as March, October, and December have relatively high subscription rates. However, the absolute number of subscriptions remains low due to fewer calls made during these months. Maintaining a consistent calling strategy throughout the year could help increase overall subscription rates by better leveraging high-conversion periods.



Recommended Models

Since the objective is to predict whether a customer will subscribe to a term deposit, a binary classification model is the most suitable approach. Below are some models that are well-suited for this problem, along with considerations for handling data imbalance and evaluating model performance.

•**Logistic Regression**: A simple and efficient model that is quick to train and provides interpretable results. However, it assumes a linear relationship between the features and the log-odds of the target variable, which may not always be valid.

•**Decision Trees**: These models are easy to interpret, can handle non-linear relationships, and automatically perform feature selection. However, they are prone to overfitting, particularly when the dataset has many features.

•**Random Forest**: An ensemble learning method that improves decision tree performance by reducing overfitting. It is effective for handling non-linear relationships but may require longer training time and can be more challenging to interpret.

•**Gradient Boosting Machines (XGBoost, LightGBM)**: These models provide high accuracy, handle non-linear relationships effectively, and manage missing data well. However, they are computationally expensive, require careful hyperparameter tuning, and may be harder to interpret compared to simpler models.

To address the **imbalance in the dataset**, techniques such as **undersampling, oversampling, and the use of appropriate evaluation metrics (such as AUROC)** will be implemented to ensure the model effectively differentiates between subscribed and non-subscribed customers.

Handling Imbalance

Imbalanced datasets are a common challenge in binary classification problems, and appropriate techniques must be applied to ensure model performance is not negatively impacted.

Resampling: The class distribution can be adjusted by **oversampling** the minority class, **undersampling** the majority class, or using a combination of both. While oversampling helps the model learn from more minority class instances, it can lead to **overfitting**. Conversely, undersampling reduces dataset size and may result in **loss of valuable information**.

Evaluation of Model

To accurately measure model performance, we will use **precision, recall, and F1-score** as key evaluation metrics. Given the class imbalance, relying on **accuracy alone** would be misleading, as it may not reflect the model's ability to correctly identify minority class instances. By focusing on precision and recall, we can ensure that the model effectively distinguishes between customers who are likely to subscribe and those who are not.