Bank Marketing (Campaign)

Suyog Nemade

PROBLEM DESCRIPTION:

ABC Bank aims to predict a customer's chance of enrolling in their term deposit product based on prior interactions. They want to develop a machine learning model to identify customers who are more likely to purchase the products. In other words, businesses want to choose customers who are more likely to buy the products.

Business Understanding:

After an ML prediction model has been created, we can next speculate about the kind of customers that are most likely to purchase the product. This will help the bank focus its marketing efforts on such customers in future marketing initiatives. By restricting the campaign's reach, the bank may save costs, preserve resources, and boost profit margins.

Data Understanding:

The "bank-additional-full.csv" dataset, which consists of 41188 observations and 21 features, includes information about clients' basic information, including age, job, marital status, education, credit in default, housing, and loan; marketing campaign information, including outcome, employment variation rate, consumer price index, consumer confidence index, euribor 3-month rate, and number of employees; and contact information, including contact communication type, last contact month, last contact day, last contact duration, and number of contacts. The objective variable y, which will also be utilized in subsequent predictions, is the response to the yes-or-no inquiry, "has the client subscribed a term deposit?"

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 month rate - daily
nr.employed	Numeric	number of employees - quarterly

ANOTHER VISUALIZATION

Feature Name	Type	Data Type	# of Null or "Unknown"	# of outliers	Comments
age	Numerical	int	0	0	
job	Categorical	str	330	0	Drop missing values
marital	Categorical	str	80	0	Drop missing values
education	Categorical	str	1731	0	
default	Categorical	str	8597	0	* Two options: leave unknown as its own class or use a use a classification ML model on this feature to fill in the unknown data.
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*IQ 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	

Two main questions:

1. Which data problems—such as the number of NA values, outliers, skewness, etc.—are present?

2. What methods are you trying to employ in your data gathering process to deal with problems like outliers, NA values, etc., and why?

WHAT ARE THE PROBLEMS IN THE DATA (NUMBER OF NA VALUES, OUTLIERS, SKEWED ETC)?

Six category features—job, education, marital status, default, home, and loan—have missing data. Outlier data may be found in one numerical attribute, "duration." In particular, the greatest value for "duration" is 4918, indicating the presence of outliers, whereas the mean is about 258. Additionally, because the goal variable for the predictive classification model skews around 90% to the "N" case, the dataset is often unbalanced.

What approaches are you trying to apply on your data set to overcome problems like NA value, outlier etc and why?

We will use a range of methods to deal with missing (NA) values, depending on how serious each column is and how it affects the dataset as a whole. removing the missing information for attributes that have less "unknown" data points ("marital" and "job"). substituting the most common category for "housing" and "loan" for the missing data. Additionally, the missing values for the "default" and "education" characteristics are filled in using an ML classification model. As previously noted, we may utilize an upper outer fence specified at 3IQ for the outlier numerical data (upper fence = Q3 + 3*IQR), where IQR is defined as the interquartile range. This will allow us to save 97% of the original data.

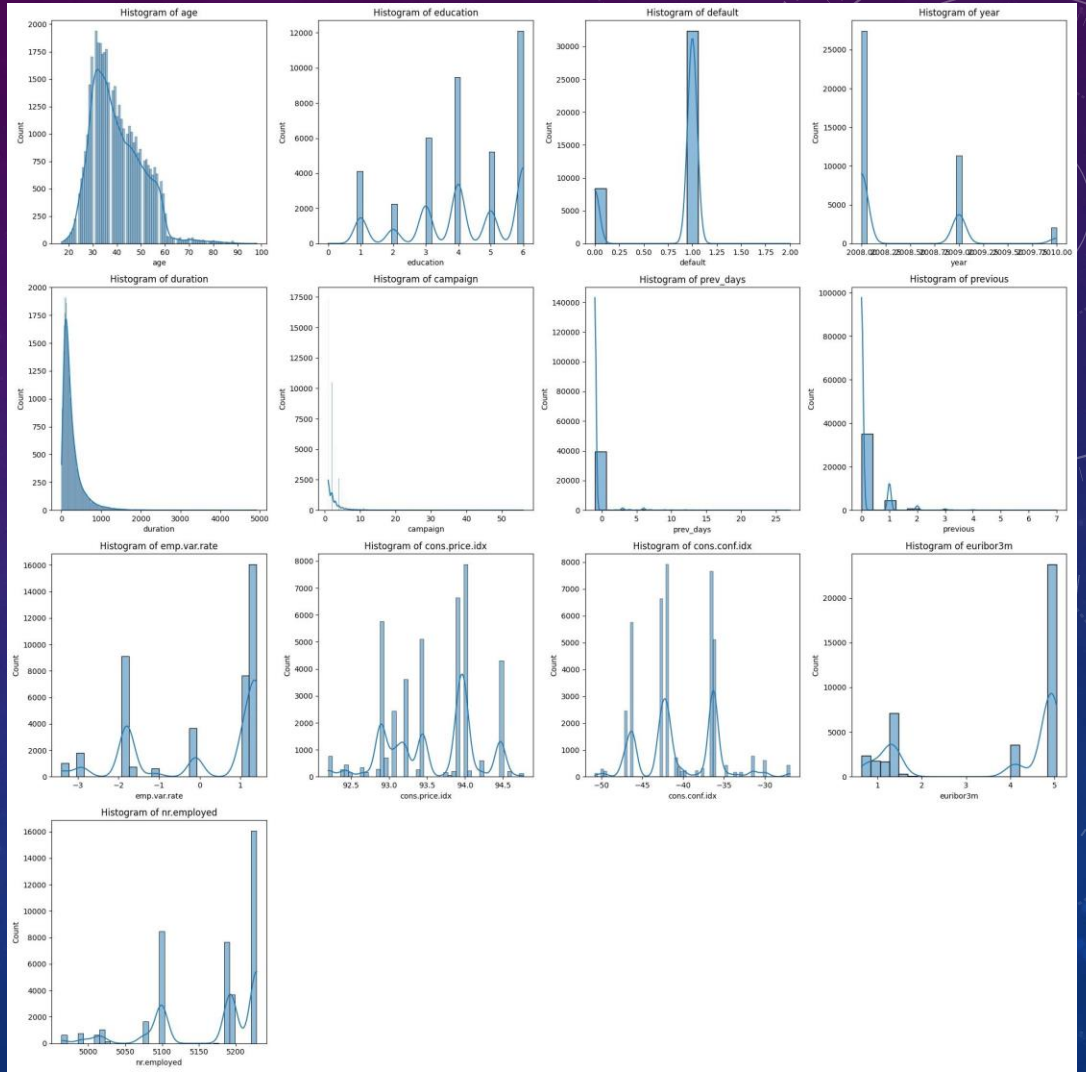
By selecting the appropriate assessment metric, we may assist in accounting for the imbalance associated with the target variable in the model. This will probably include utilizing the AUROC curve to determine which models yield the greatest True Positive and False Negative prediction outcomes for this data set. We may also think about under-sampling from the majority scenario because the dataset is sufficiently vast. Alternatively, we may make sure that the unusual instances are always retained and only randomly separated from the majority case when dividing the data during training, rather than randomly assigning folds. To over-represent the unusual instance for the model, we may also change the rare:majority case ratio in the training set.

DATA EXPLORATION

Uncovering patterns, trends, and correlations that allow us to select those features that are most relevant when training the ML model.

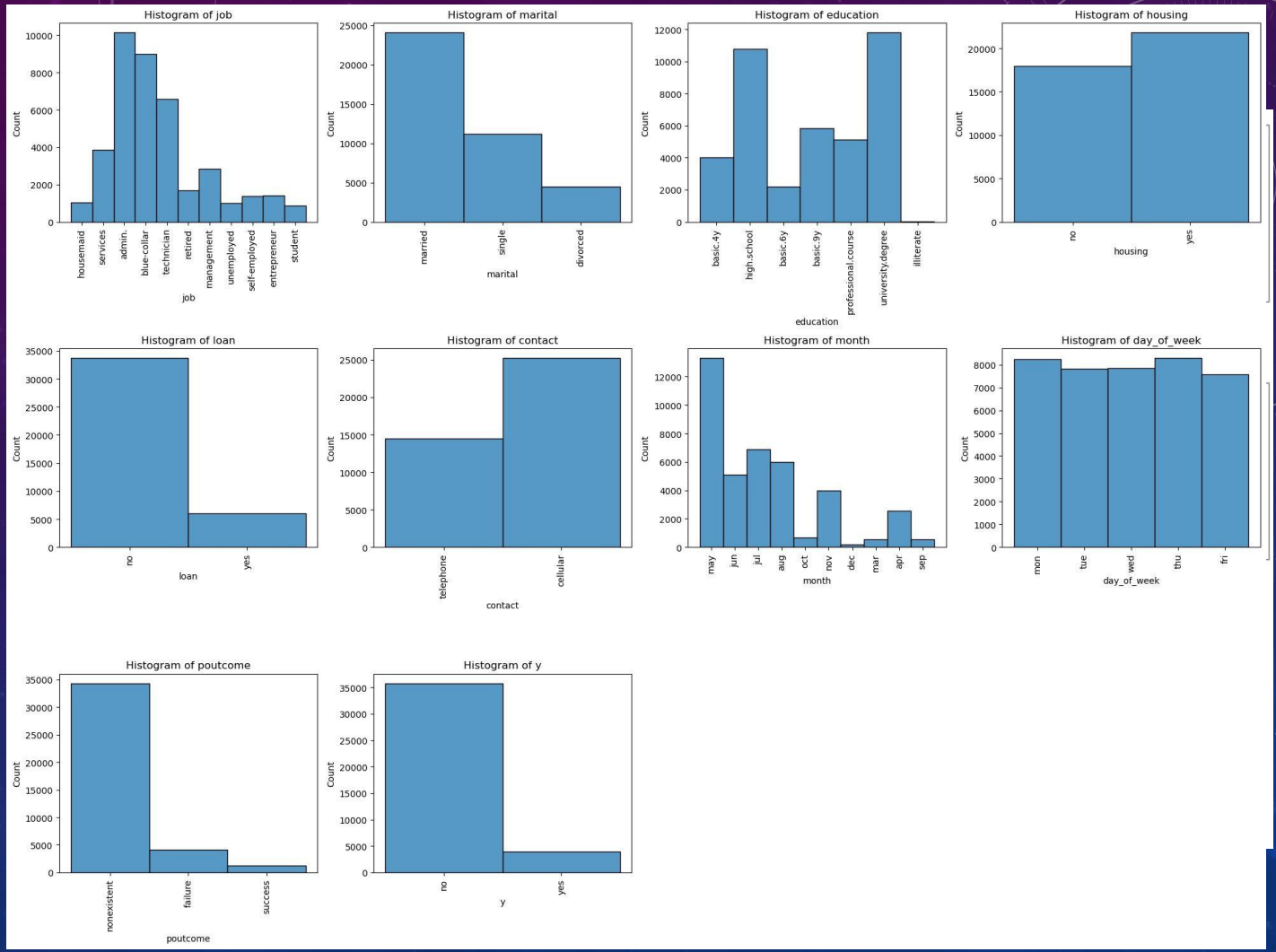
NUMERICAL FEATURES

We drew histograms for the complete data set in order to better see the range, spread, biases, etc. in each of our features. The numerical data is shown in the figures on the left.



CATEGORICAL FEATURES

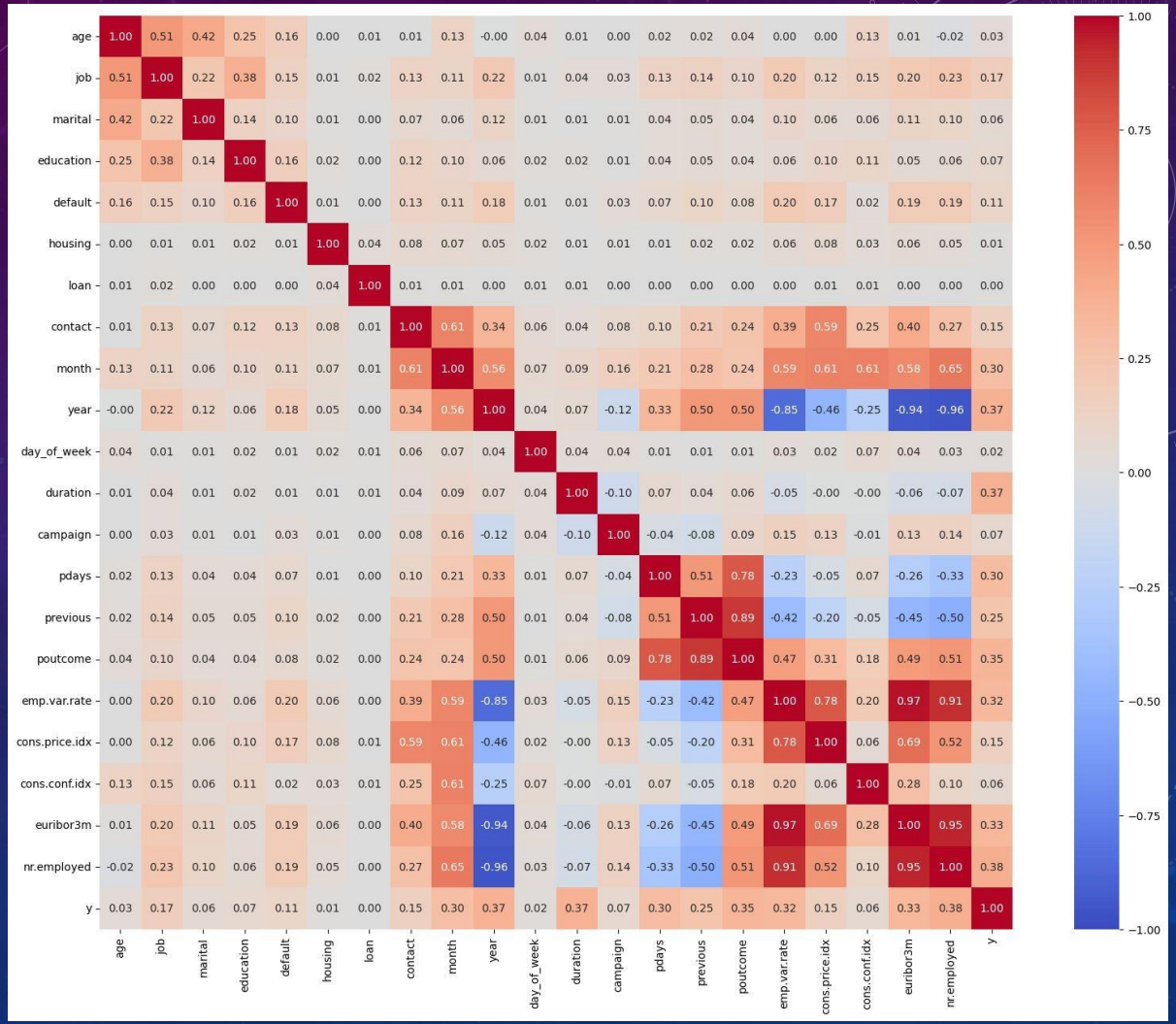
These charts are histograms of the data set's category characteristics, just like the one on the preceding slide. The significant bias in the target variable "y" is especially noteworthy. Of that feature, just around 10% are "yes" situations. When it comes time to train the predictive model, this information is crucial.

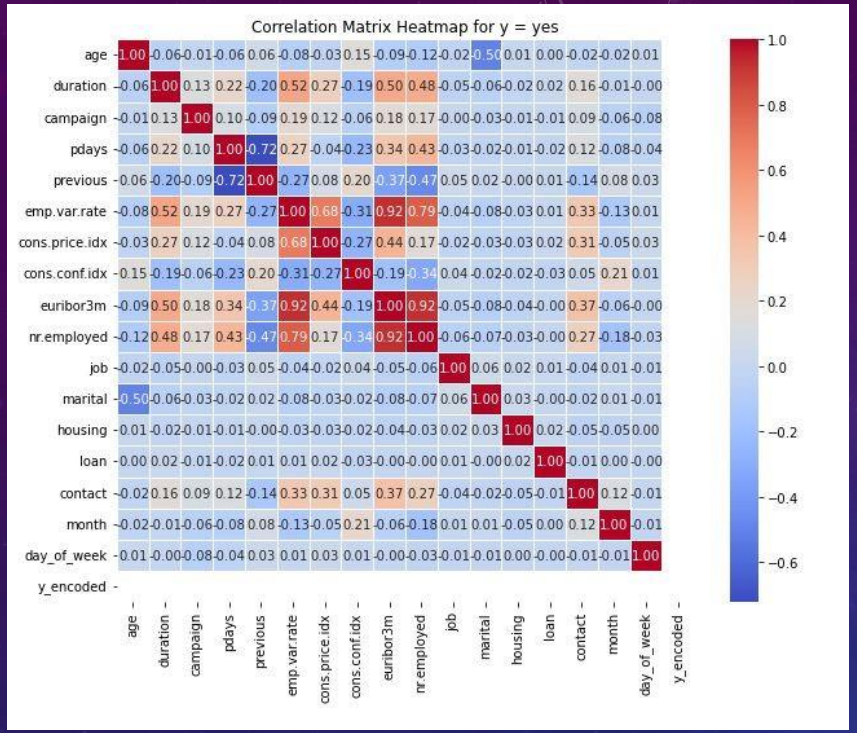


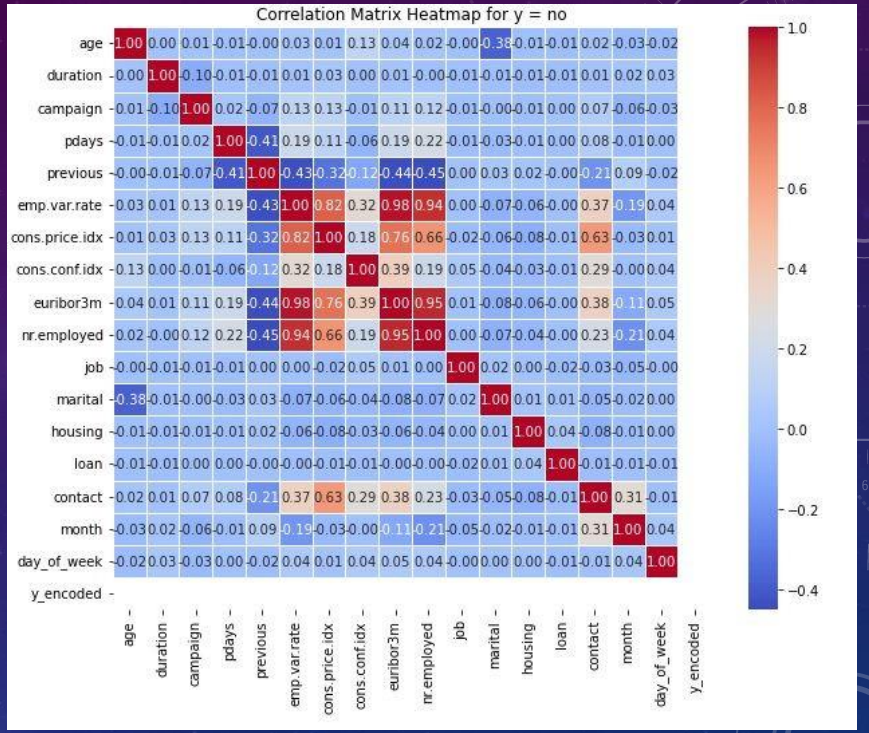
RECOMMENDATONS

In the following section, we are going to choose the variables that are most correlated with target y and make recommendations of them based on the visualizations.

The heatmap shows that, with values near zero, the majority of variables have little correlation with the target variable "y." "Year," "duration," "days," "poutcome," and so on are the ones that have a stronger correlation with y. Given that factors like "duration" have a significant impact on the terms' subscription, this conclusion makes considerable sense.



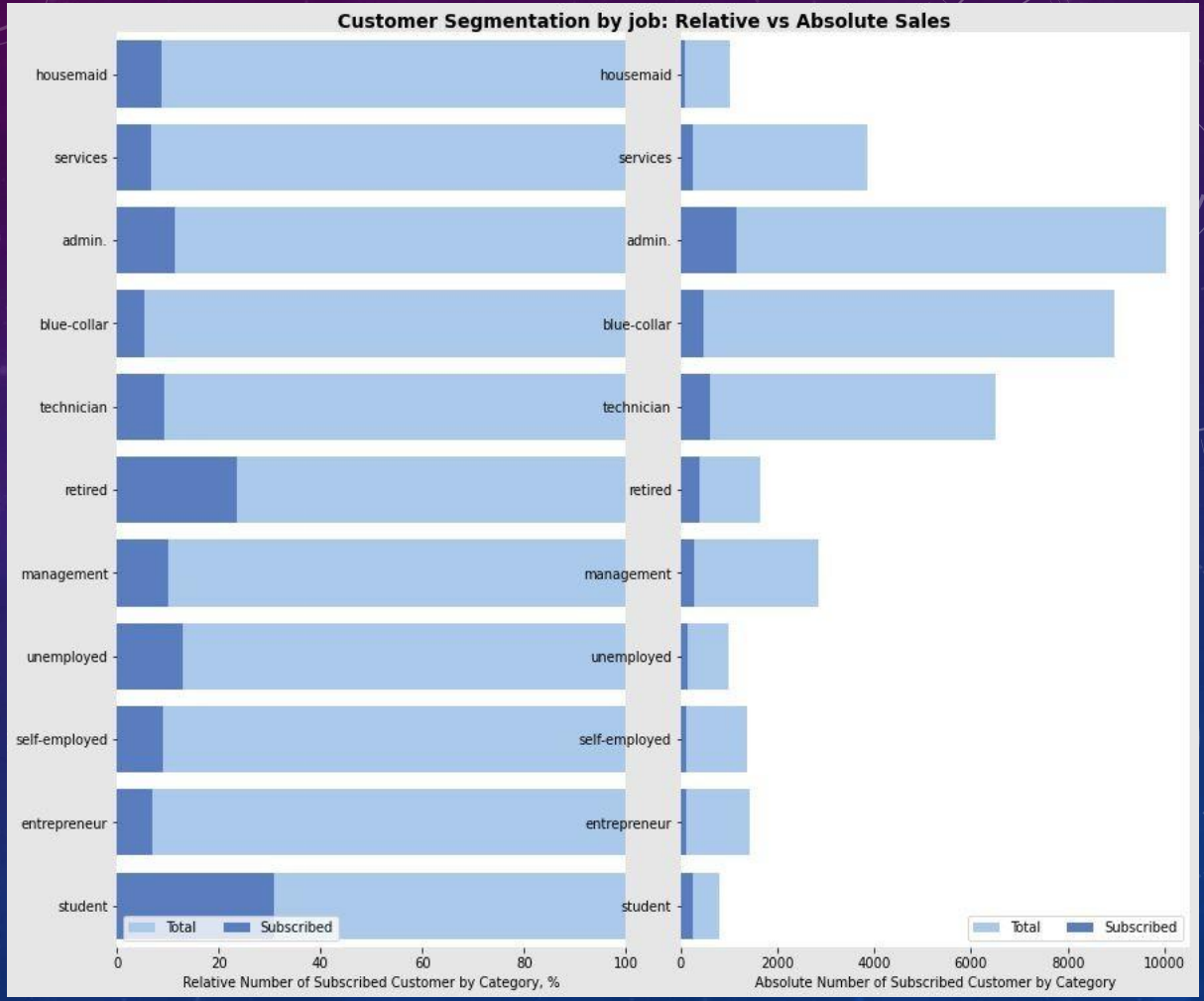




If we split into yes and no cases, we can see that the values close to 0 for both cases, indicating weak correlation to separated y.

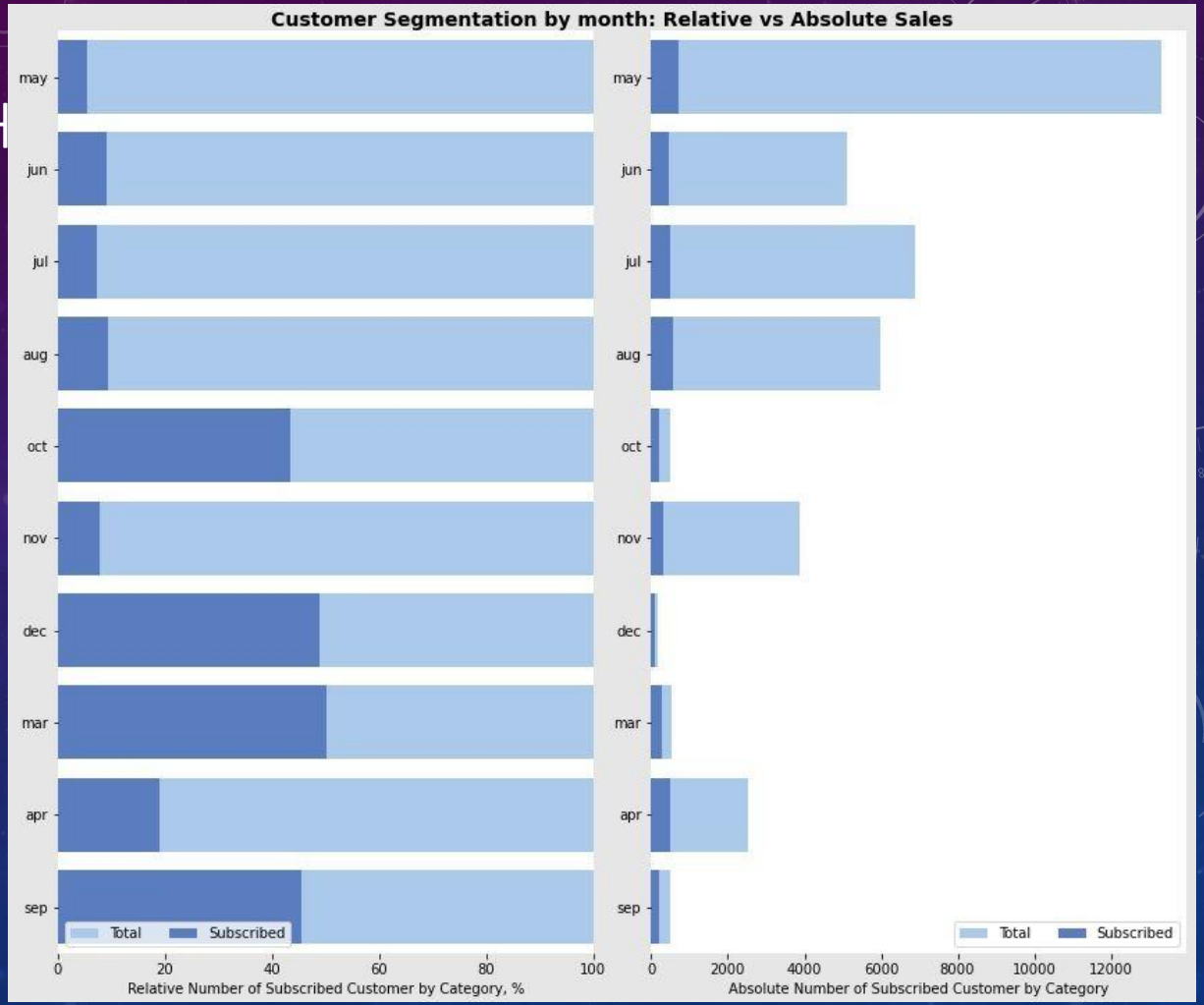
IMPROVEMENTS - JOBS

Numerous enhancements may be made, according to the distribution of categorical variables. In particular, pensioners and students subscribe to the terms more frequently than others in comparison, but their total number of subscriptions is significantly lower since they make fewer phone calls; hence, calling them more frequently might result in more subscriptions.



IMPROVEMENTS - MONTH

Calling months would be an additional area for improvement. As we can see, there are comparatively fewer subscriptions in March, October, December, and so on, but there are also fewer subscriptions overall because there are fewer phone calls during these months. As a result, making constant phone calls throughout the year will also result in a rise in subscriptions.



RECOMMENDED MODELS

A binary classification model that excels in making accurate predictions would be the optimal option, as our issue is to anticipate whether a consumer would acquire the term deposit or not.

We think the models listed below will work best for your issue. We will also describe our approach to testing our models and handling imbalance.
Logistic Reg/ession: This model is simple, easy to train, and produces very interpretable results. It does, however, make the potentially unfounded assumption that features and the target variable's log-odds have a linear relationship.

Conclusion Trees: These models can handle non-linear relationships and are interpretable. They also carry out feature selection in an implicit manner.

Random Forests are an ensemble technique that improves decision tree performance. They are good at handling non-linear interactions and are less prone to overfit. They might, however, take additional training time and be a little harder to comprehend.

G/adient Boosting Machines (XGBoost, LightGBM): These models are very precise, effective at handling missing data, and can deal with non-linear relationships. However, they demand careful parameter adjustment, take longer to train, and are more difficult to understand.

HANDLING IMBALANCE

- Imbalance are a normal problem of binary classification models so we need to handles these are seemed fit.

- **Resampling:** Adjust the class distribution by oversampling the minority class, undersampling the majority class, or using a combination of both. This helps create a more balanced dataset, but may lead to overfitting (oversampling) or loss of information (undersampling).

Evaluation of Model

We will use accuracy, recall, and F1-score as our assessment metrics to gauge how well our models are doing. In the context of our existing models, using accuracy as a metric would not produce trustworthy findings, especially when working with datasets that are unbalanced.