Bank Marketing(Campaign)

Data Go
Yuheng Chen, Terry Chou, Rishi Aluri, Justine Pile

Problem description

The objective of this project is to develop a predictive model for ABC Bank to determine the likelihood of customers subscribing to their term deposit product. By analyzing customer interactions with the bank and other financial institutions, the machine learning model will identify potential clients who are more likely to purchase the product. This will enable the bank to optimize their marketing efforts by focusing resources on customers with a higher probability of conversion, thus enhancing campaign efficiency and reducing costs. The project will assess model performance with and without using the "duration" feature while also addressing any data imbalance through suitable techniques.

Problem Statement:

ABC Bank wants to sell it's term deposit product to customers and before launching the product they want to develop a model which help them in understanding whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution).

Business understanding:

The objective of this project is to develop a predictive model for ABC Bank to determine the likelihood of customers subscribing to their term deposit product. By analyzing customer interactions with the bank and other financial institutions, the machine learning model will identify potential clients who are more likely to purchase the product. This will enable the bank to optimize their marketing efforts by focusing resources on customers with a higher probability of conversion, thus enhancing campaign efficiency and reducing costs. The project will assess model performance with and without using the "duration" feature while also addressing any data imbalance through suitable techniques.

Data Understanding

The key independent variables from these data are:
age, job, marital, education, default, balance, housing loan, contact, day, month, duration, campaign, pdays, previous, poutcome

- **age:** the age of the person
- **job:** the person's occupation (categorical: occupation name)
- **marital:** the person's marital status (categorical: 'married', 'single', 'divorced')
- **default:** whether or not the person has credit in default (yes/no)
- **balance:** average yearly balance (numeric)
- **housing loan:** whether or not the person has housing loan (yes/no)
- **contact:** communication type (categorical: 'unknown', 'cellular', 'telephone')
- **day:** last contact day of the month (numeric)
- **month:** last contact month of year (categorical: 'Jan', 'Feb', … , 'Dec')
- **duration:** last contact duration in seconds (numeric)
- **campaign:** number of contacts performed during this campaign and for this client
- **pdays:** number of days that passed by after the client was last contacted from a previous campaign (numeric)
- **previous:** number of contacts performed before this campaign and for this client (numeric)
- **poutcome:** outcome of the previous marketing campaign (categorical: 'unknown', 'failure', 'other', 'success')

Columns for **bank-additional.csv** and **bank-additional-full.csv** only:

- **emp.var.rate:** employment variation rate (numeric)
- **cons.price.idx:** consumer price index (numeric)
- **cons.conf.idx:** consumer confidence index (numeric)
- **euribor3m:** 3 month Euribor interest rate (numeric)
- **nr.employed:** the number of employees (numeric)

Questions We have

What are the problems in the data (number of NA values, outliers , skewed etc)?

- No duplicate rows (rows that are exactly the same across all columns) appears in **bank.csv**, **bank-full.csv**, and **bank-additional.csv**
- 12 duplicate rows appear in **bank-additional-full.csv**
- The age count graph (count vs age) is **right skewed** – the majority are between age 30 to 40. The number of people over 60 years old is very small
- The campaign count graph (count vs campaign) is right skewed – the higher the total number of campaigns, the smaller the total count.

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

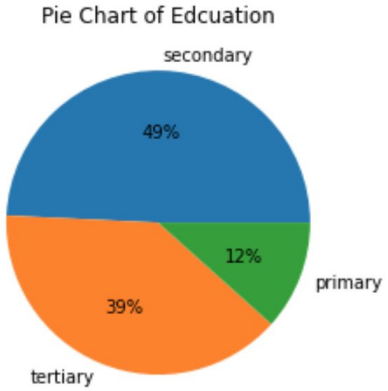
- EDA to determine NA values or other data that may be an issue
- Duplicate data can be removed from the data set
- Right skewed data in numeric columns such as age, balance, duration, campaign, previous will be fixed to normal distribution using log-transformation.
- Outliers can be identified through EDA methods including:
    - Statistical methods
    - Visualizations
- Machine learning algorithms can also be used to predict missing values

Bank-full.csv

Correlation between Features

The heatmap illustrates the relationships between each feature and both the target variable and one another. During our Exploratory Data Analysis (EDA) process, we will prioritize features that exhibit strong correlations with the target variable y.

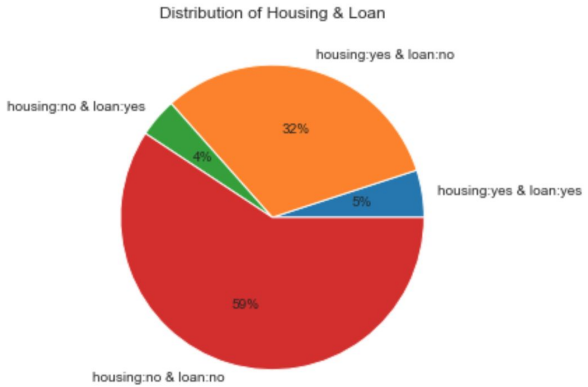
Distribution of Education



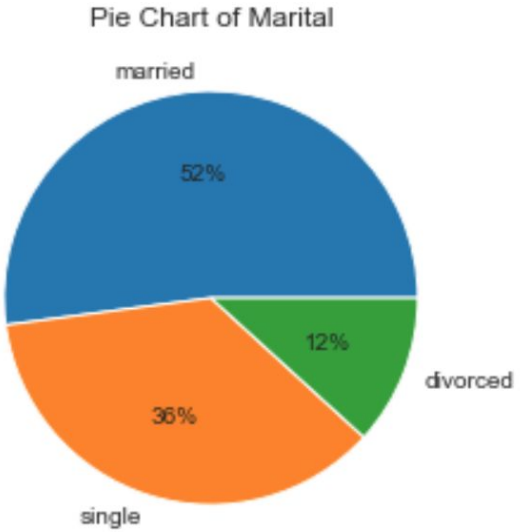
Percentage of ‘yes’ in education indicates that secondary education has more probability to subscribe, about 49%.

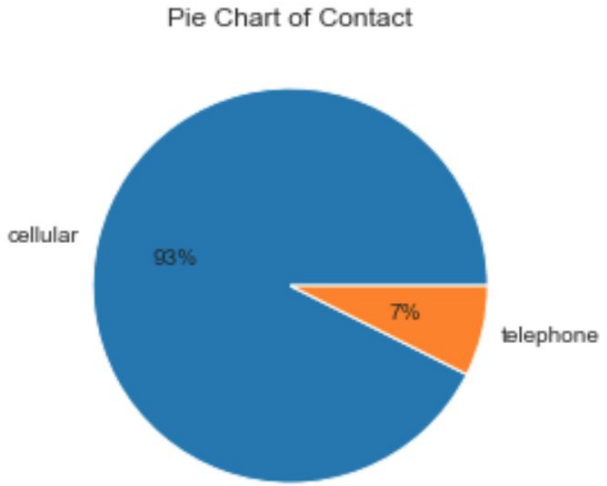
Distribution of Housing & Loan

Percentage of ‘yes’ in housing and loan. From the pie chart, the customers without any housing loans and personal loans are more likely to subscribe, about 59%.

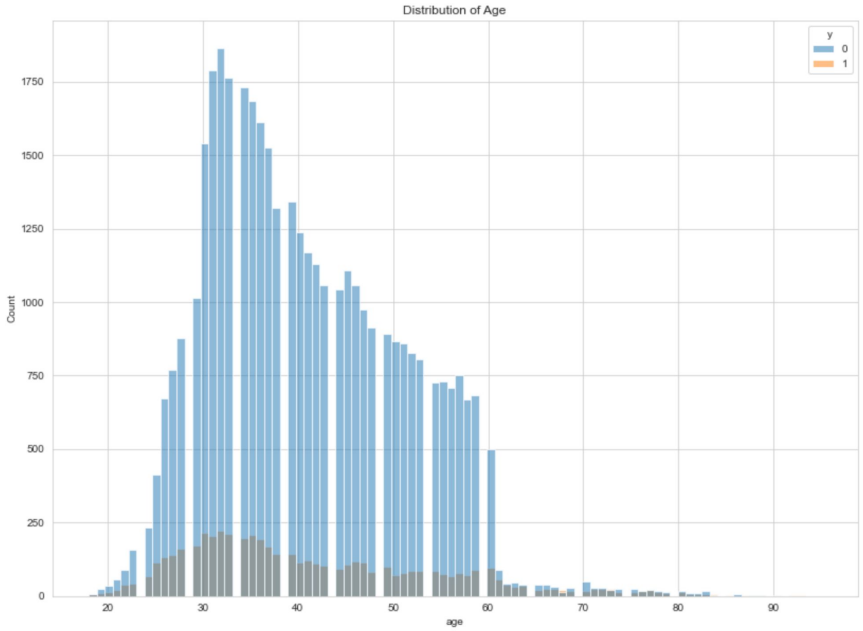


Distribution of Marital and Contact





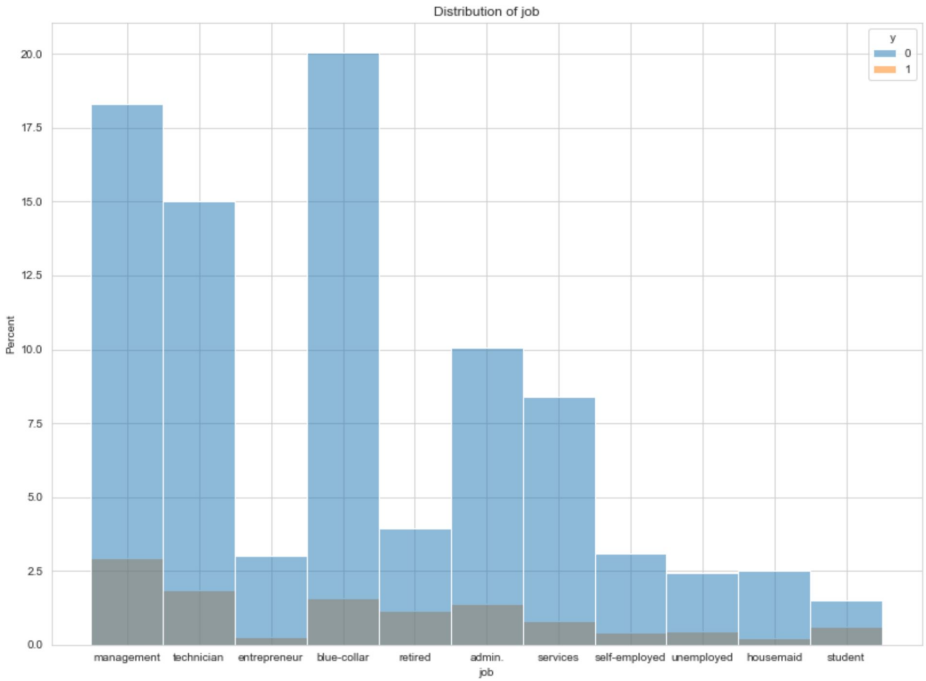
Distribution of Age

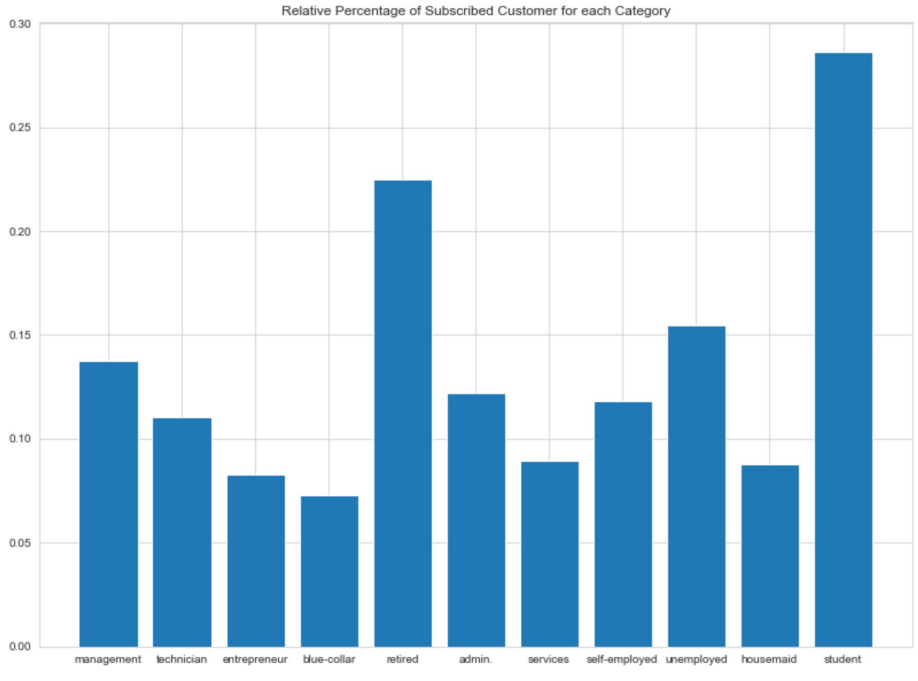


The customers about 30-40 years old are more likely to subscribe.

The bank can focus on age group to set up different kinds of promotions to target differnt age groups of customers.

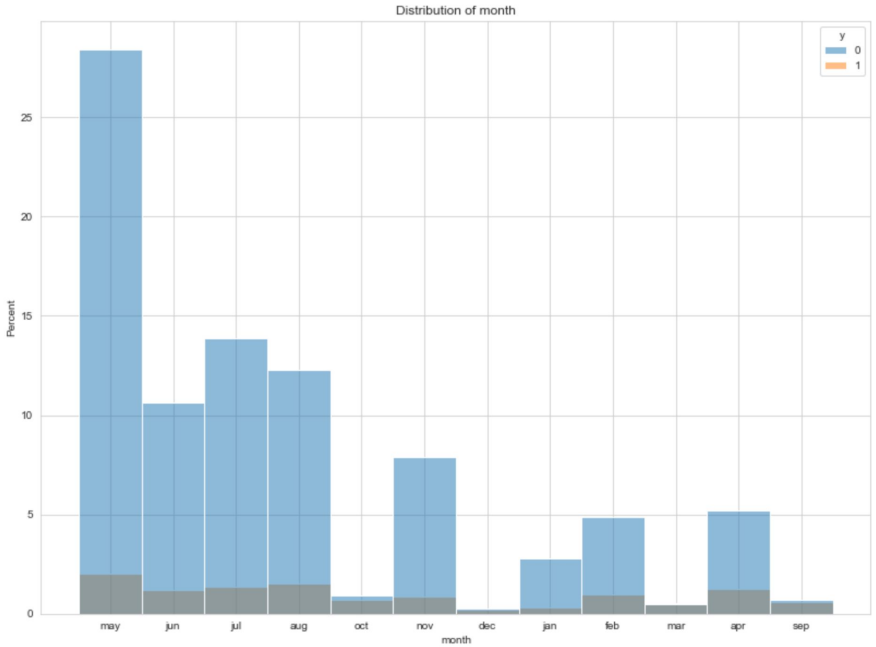
Distribution of Job

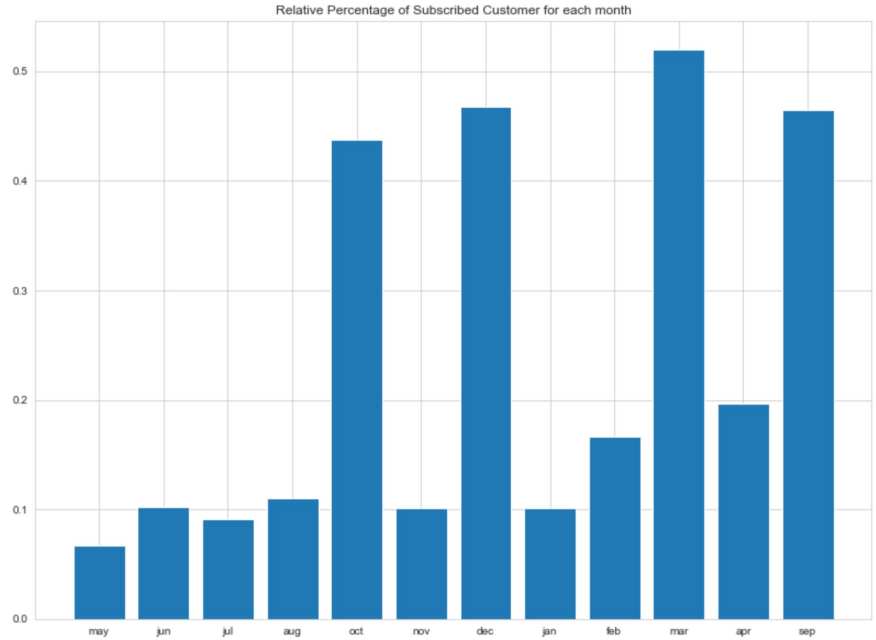




Recommendation: For each category in jobs, we can see that retired and student have the most relative percentage of subscribed customer;however, they have less absolute values of subscriptions because of fewer contacts. We can keep focus on those people and make more phone calls to increase the absolute values thus increase more subscriptions.

Distribution of Month





Recommendation: For each category in months, we can see that March, September, December and October have the most relative percentage of subscribed customer;however, they have less absolute values of subscriptions because of fewer contacts. We can keep focus on those months and make more phone calls to increase the absolute values thus increase more subscriptions.

Model Recommendations

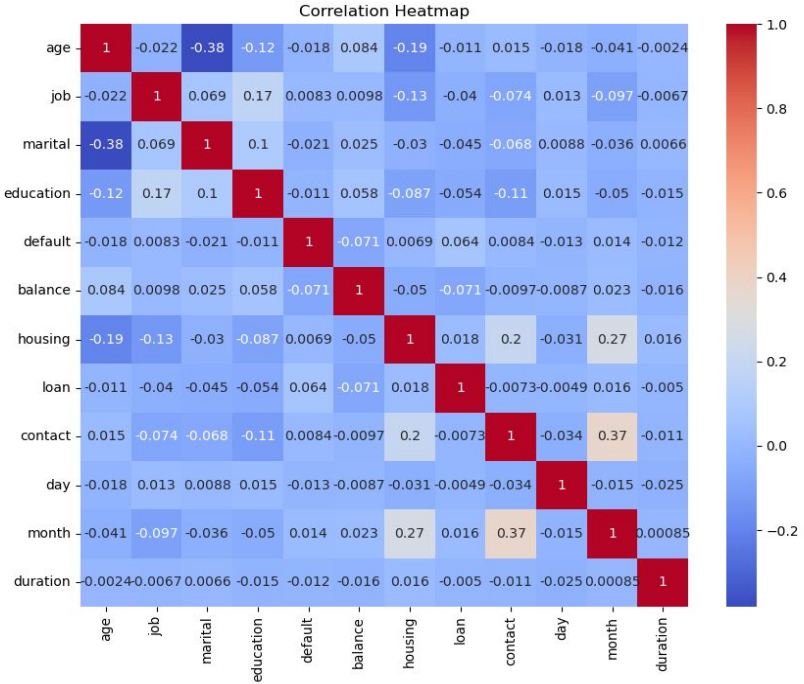
Since our goal is to decide whether customer will subscribe or not which is a binary classification, we can use binary classification models:

- Logistic Regression
- Random Forest
- Cart
- SVM

In order to make better predictions, the dataset may be resample to set an equal distribution of classes.

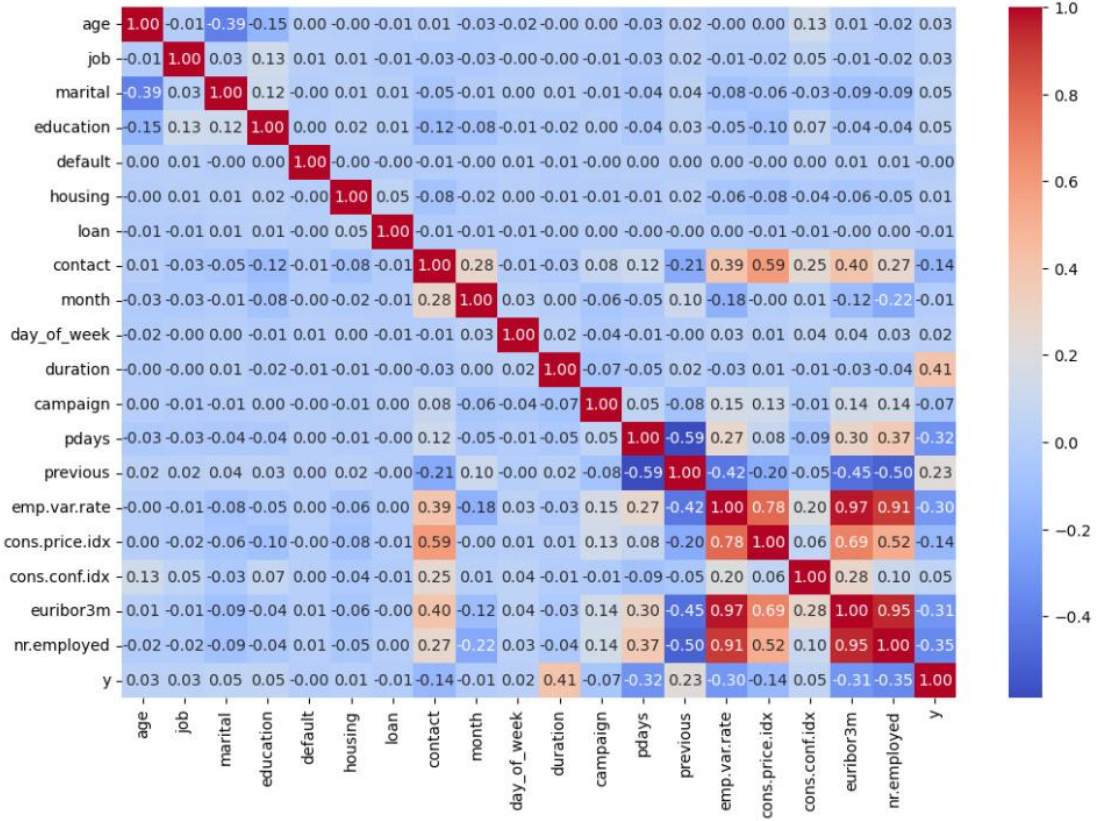
Bank.csv

Correlation heatmap of bank.csv features

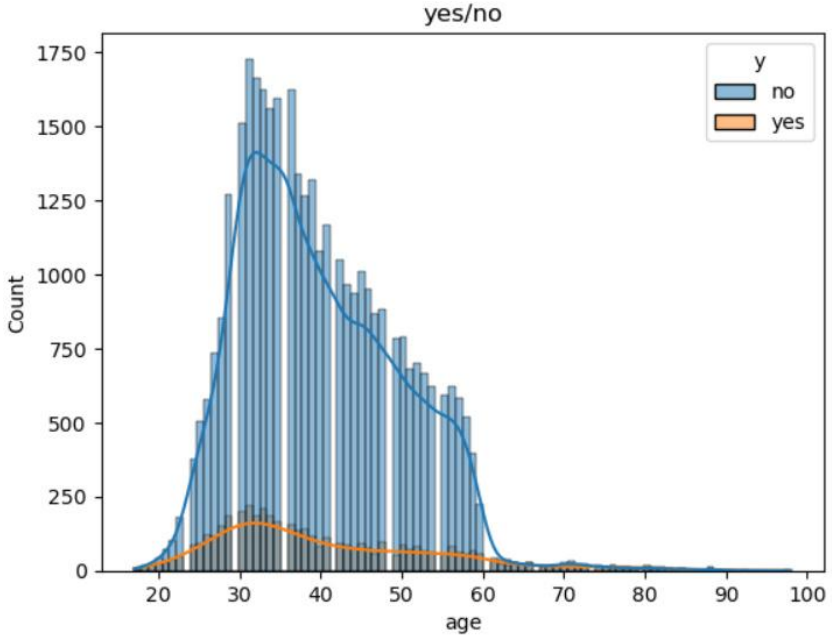


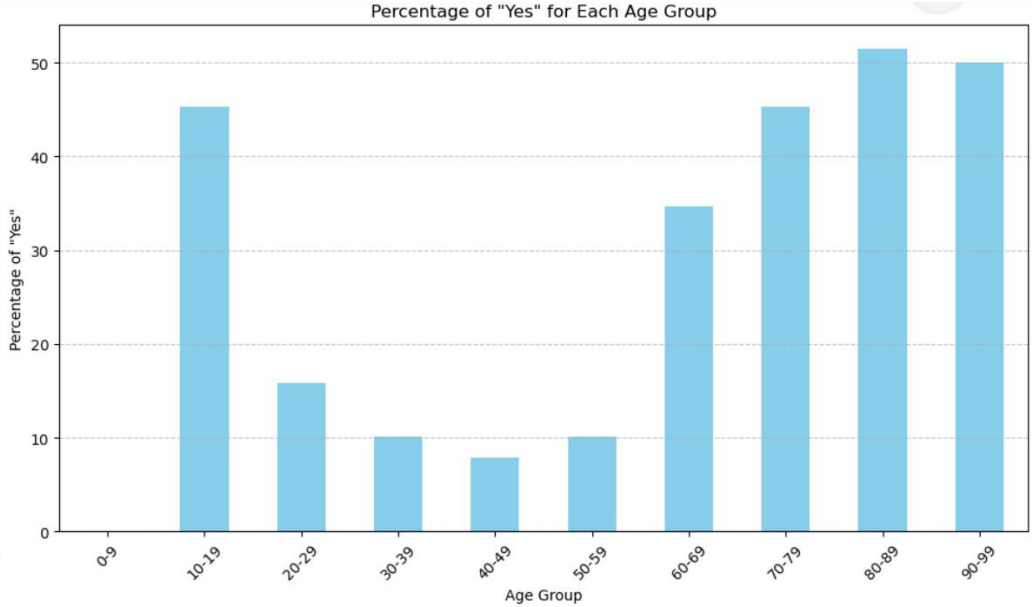
Bank-Additional-Full.csv

Correlations between Variables for this Data



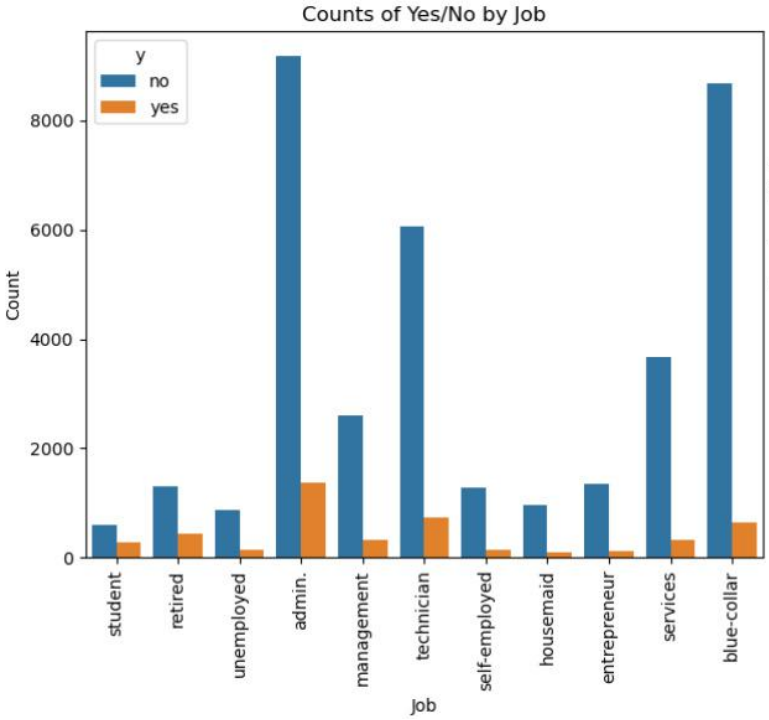
Age

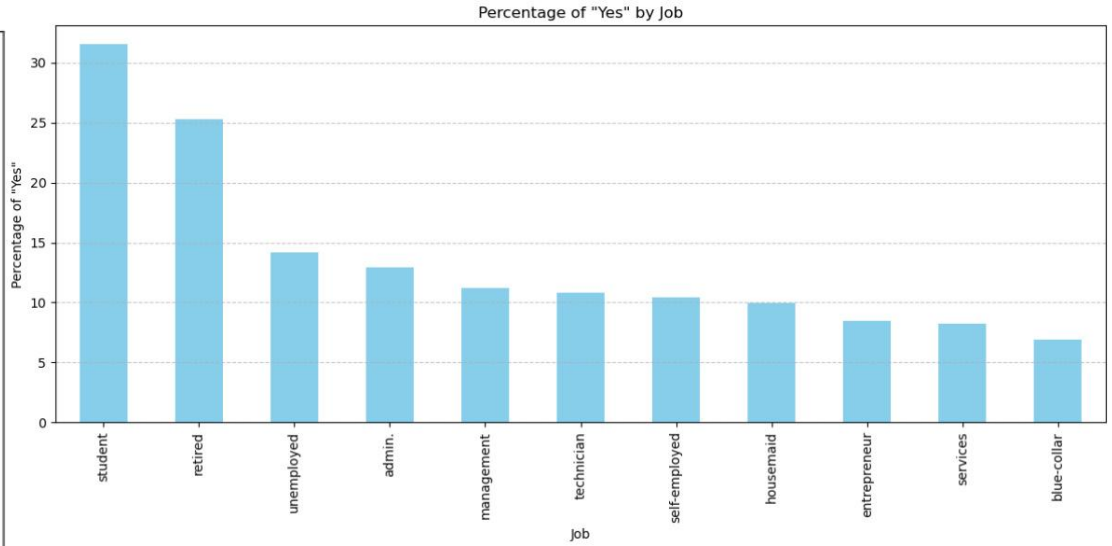




- Probability of purchase over 60 or below 20 years-old : over 30 %

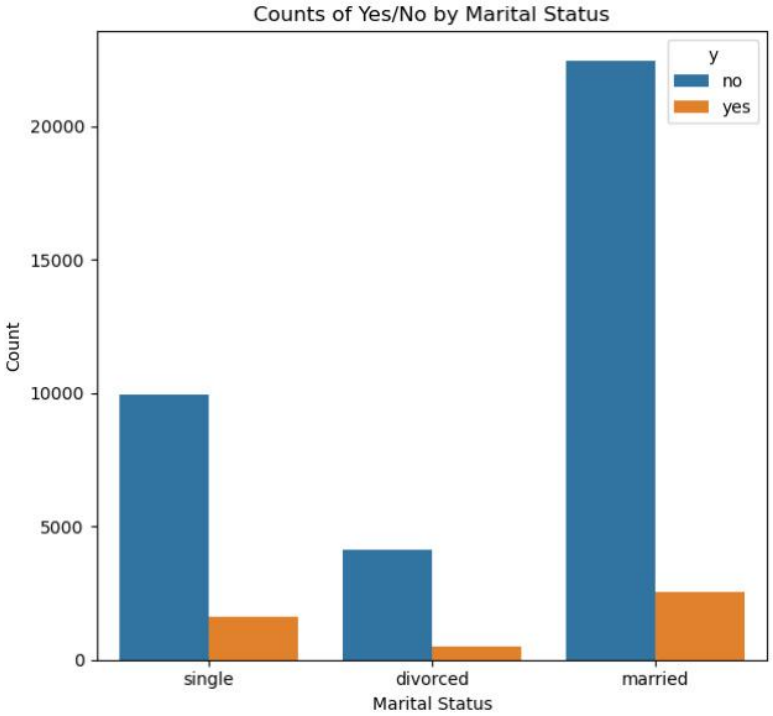
Job

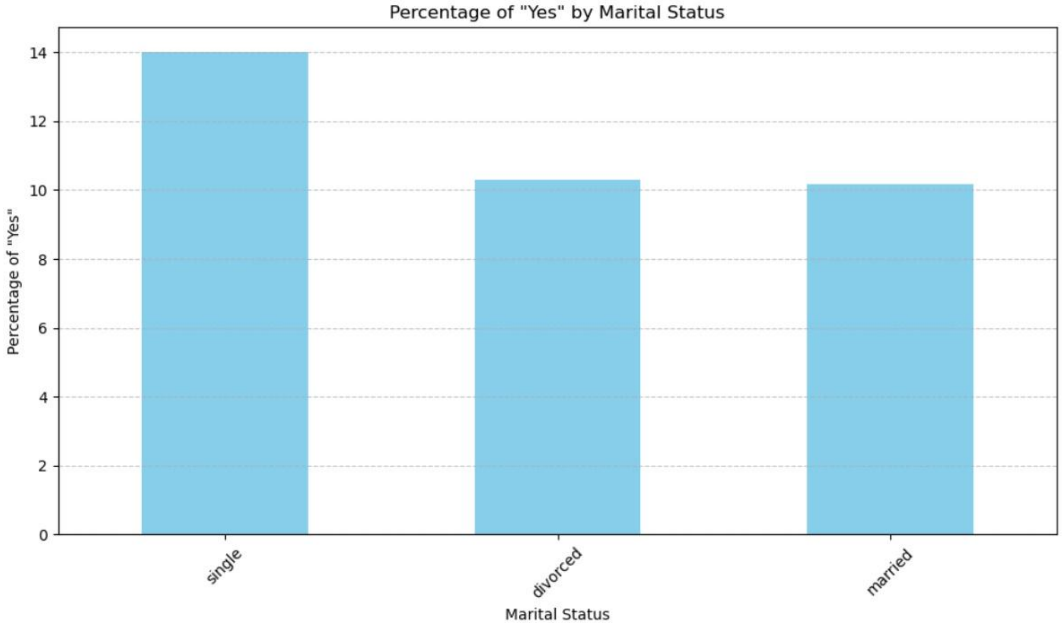




‘Retired’(25%) and ‘Students’(32%) have the highest probability of purchasing

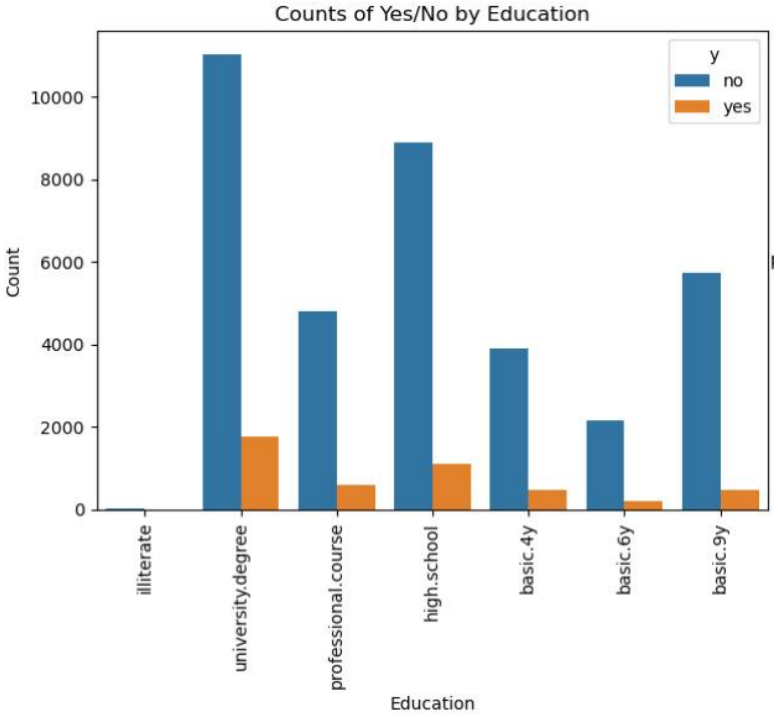
Marital Status

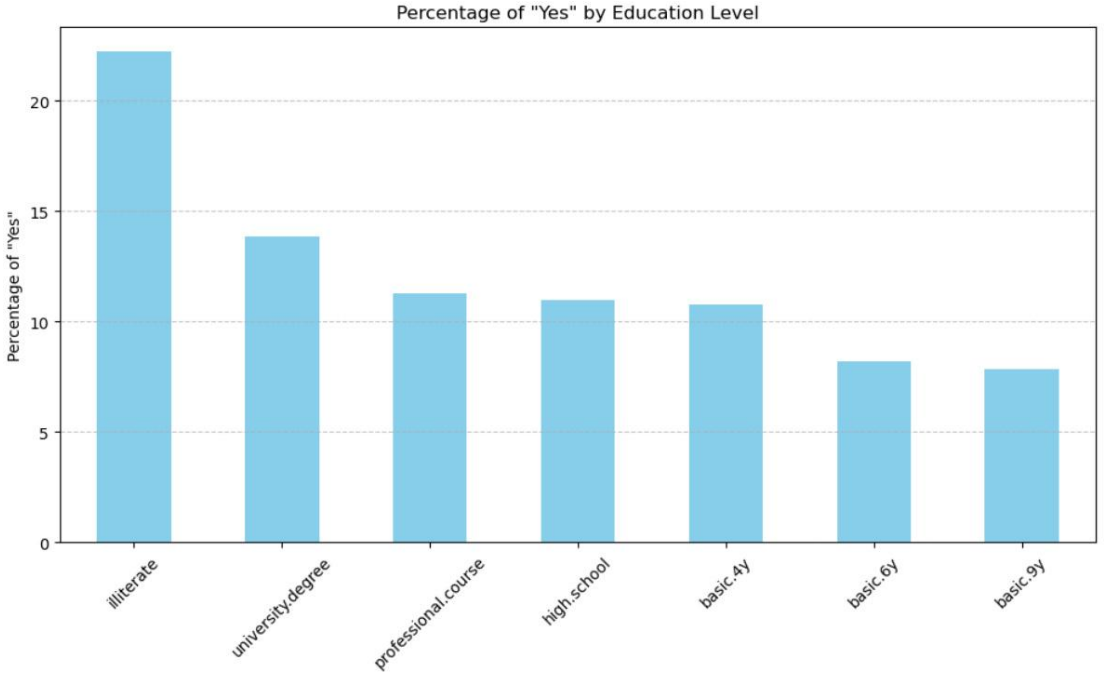




Customers who are ‘single’ has the highest chance of purchase (14%)

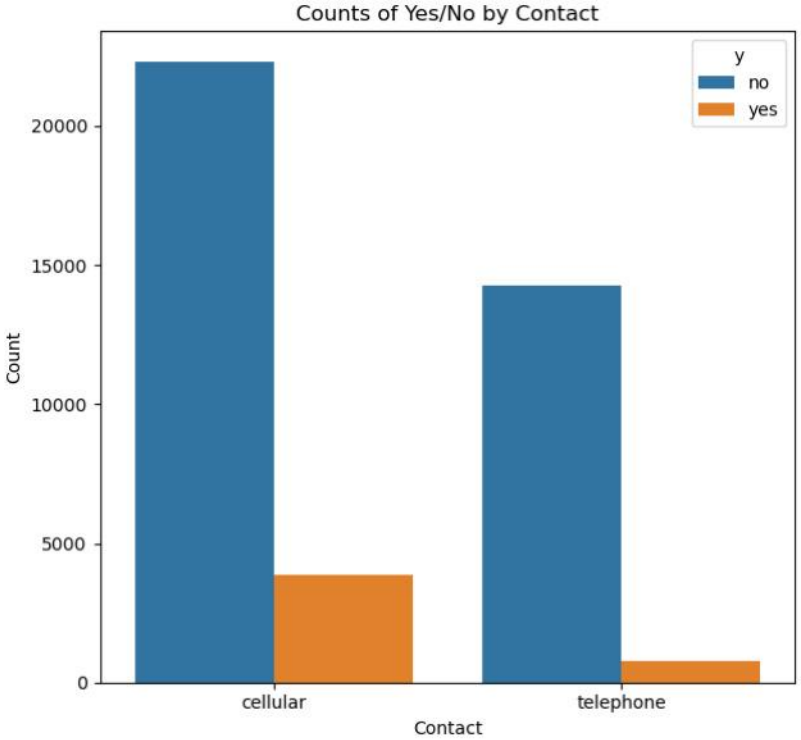
Education

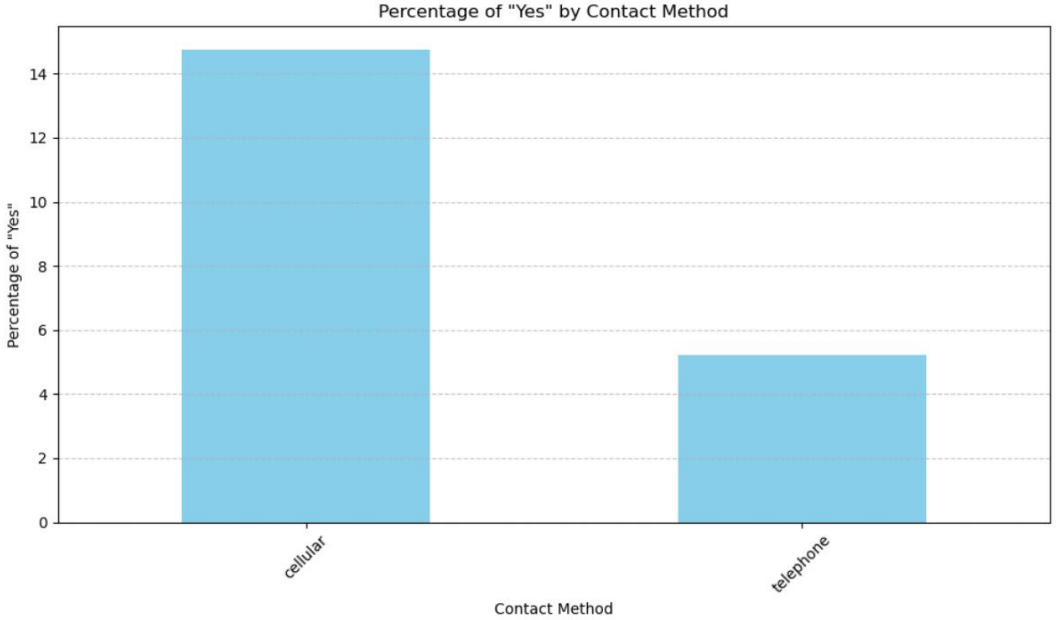




- Illiterate has too few total counts, not appropriate for consideration
- Customers with ‘university degree’ has the highest chance of purchasing (14%)

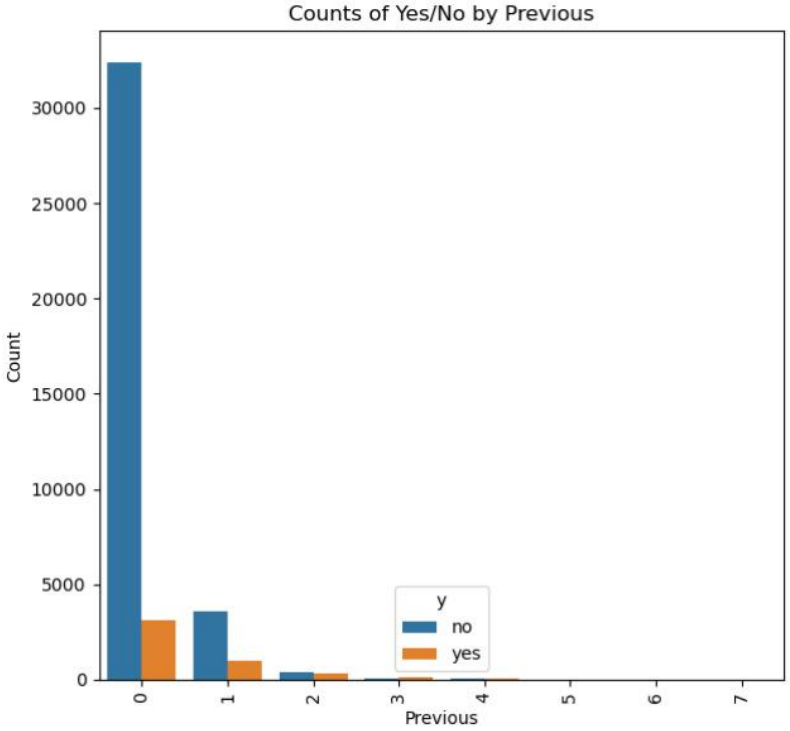
Contact Method

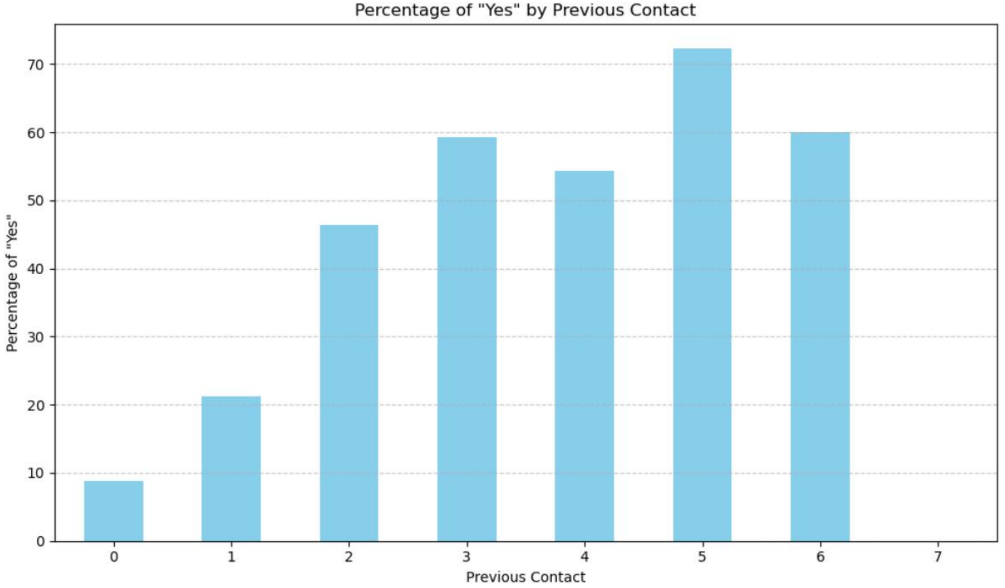




Customers with ‘Cellular’ as contact method has significantly higher chance of purchase (15%)

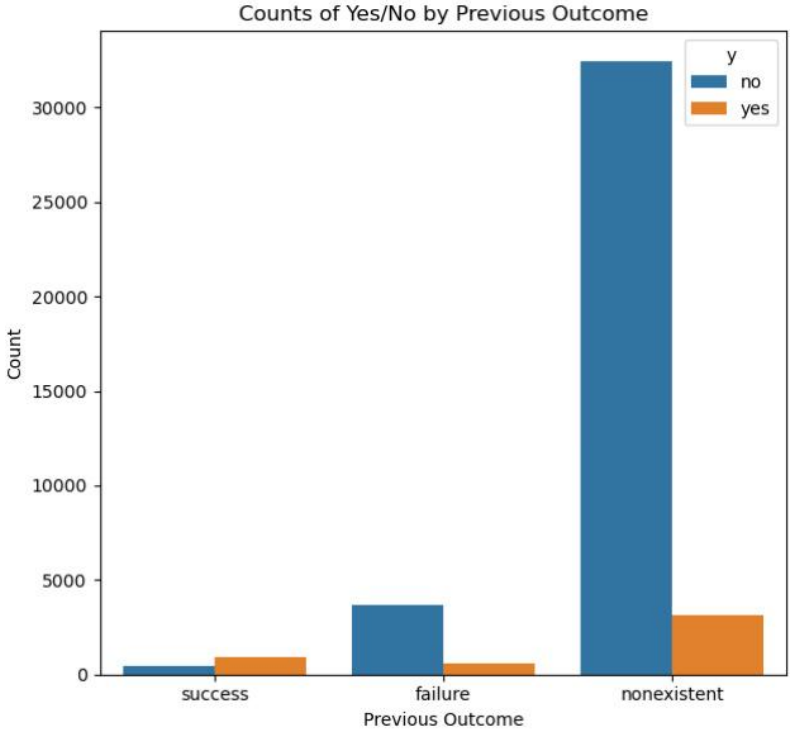
previous

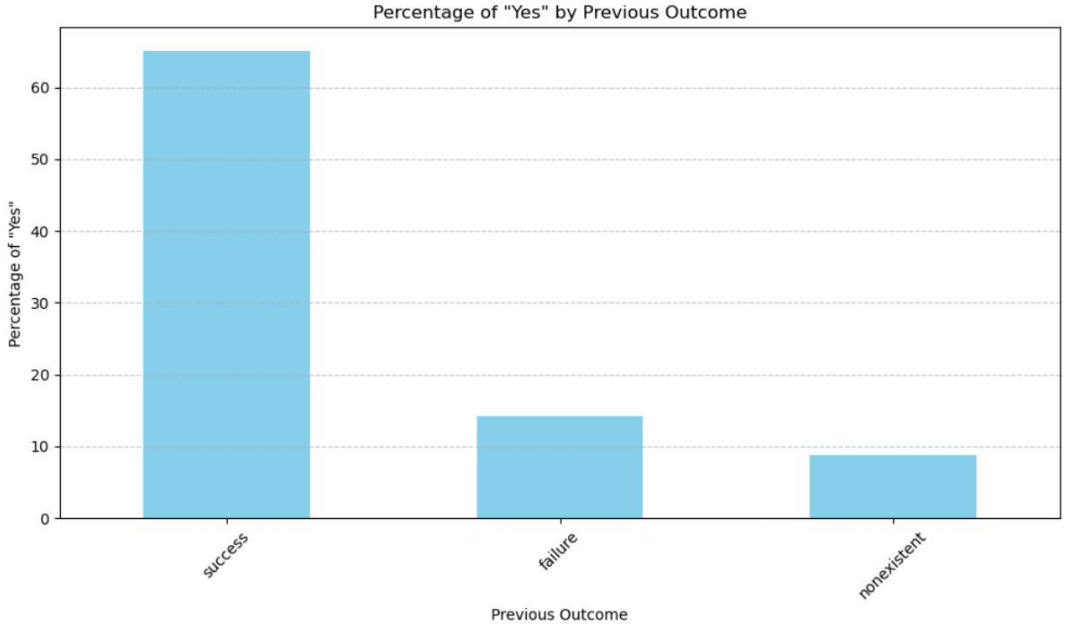




The probability of purchase increases as the number of contacts performed before the campaign on the customer increases in general

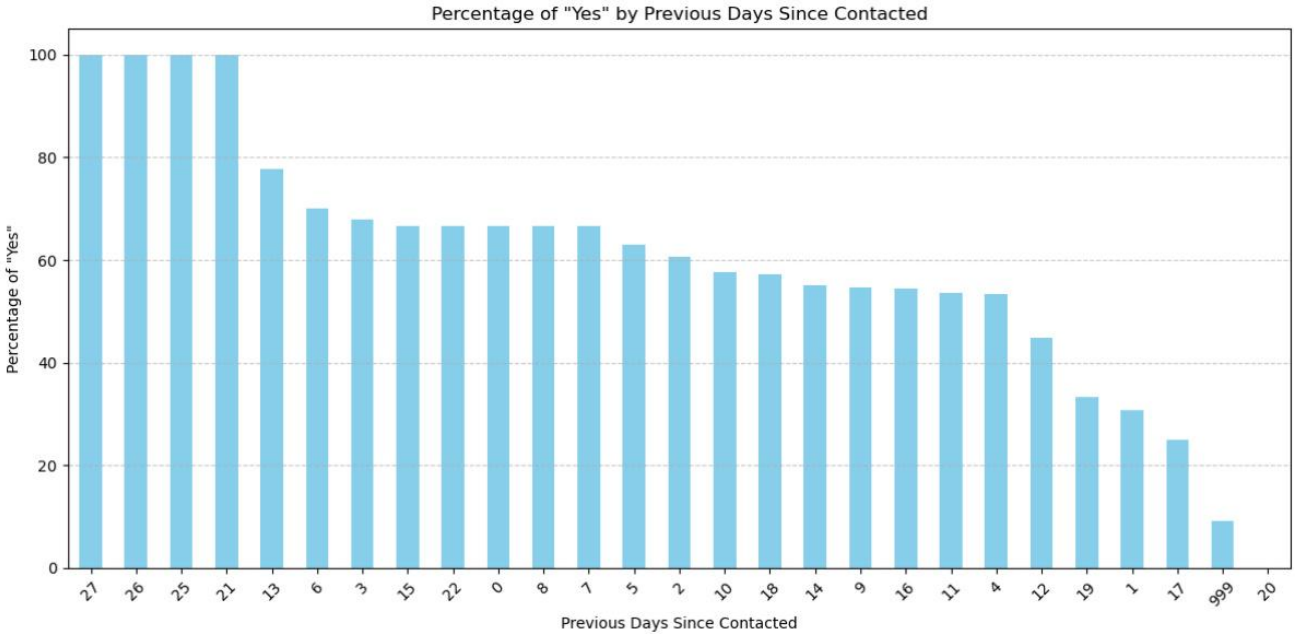
poutcome





Customers with positive experience with the previous campaign have much higher chance of purchasing (65%)

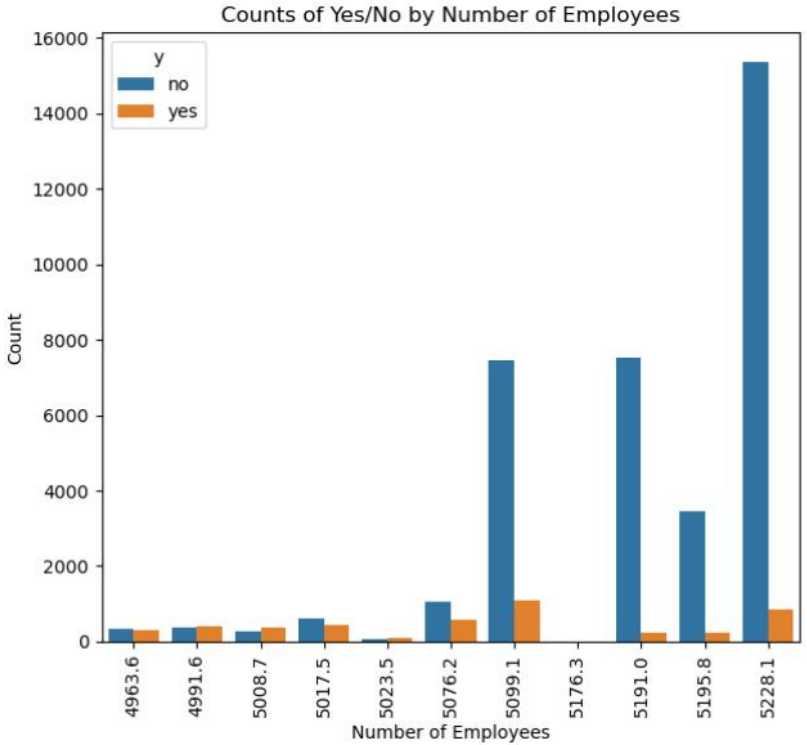
pdays

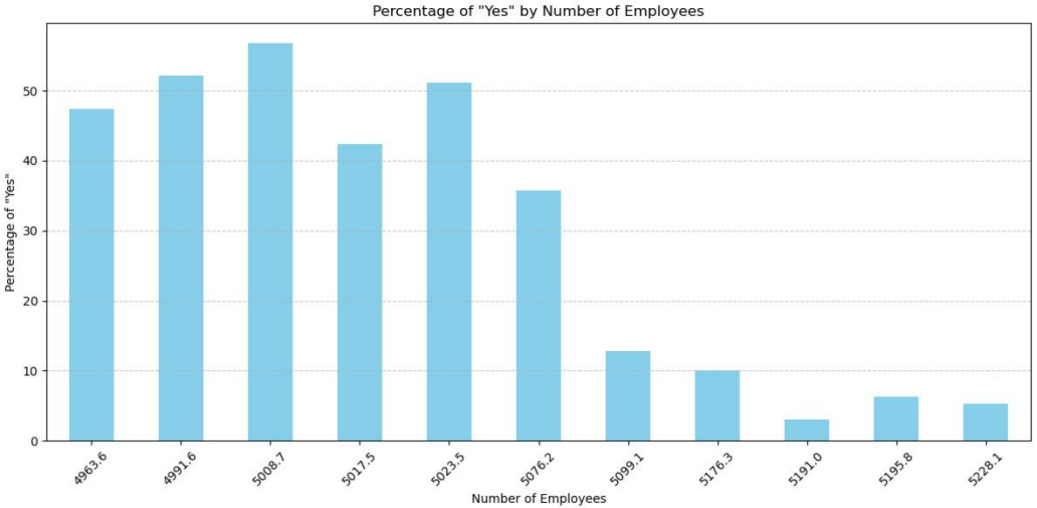


Customers who have never been contacted has significantly lower chance of purchasing the product

(pdays = 999 means the customer has never been contacted)

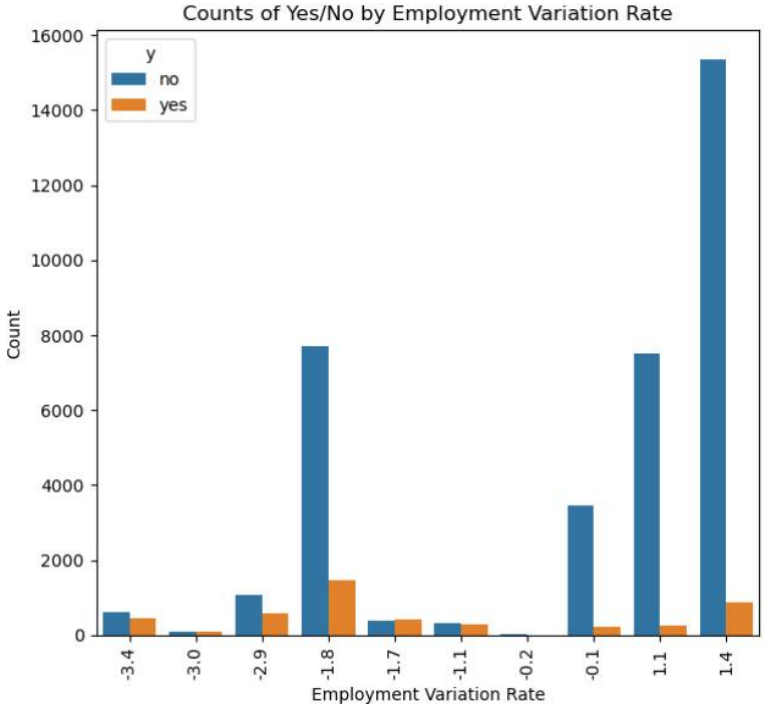
Number of Employees

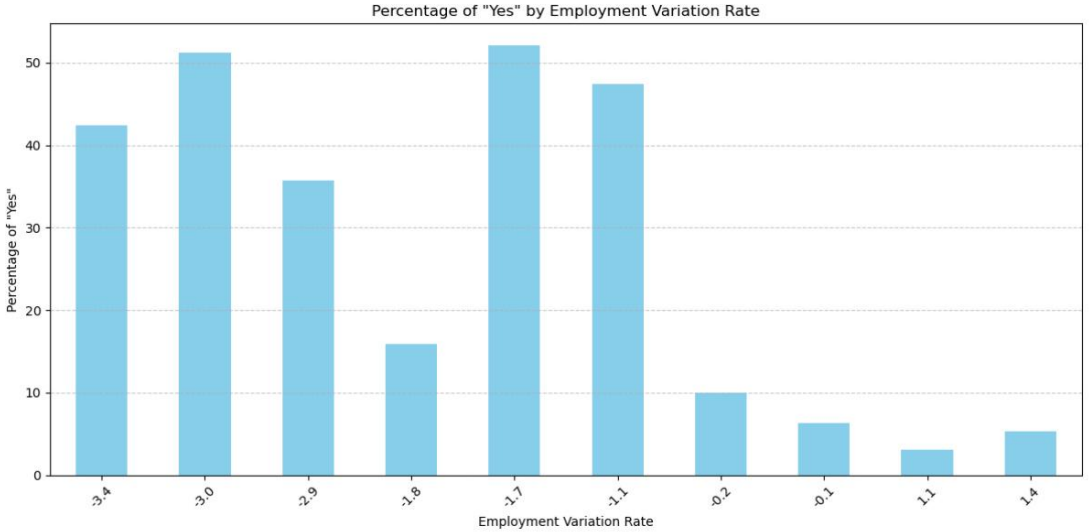




The higher the total number of people employed, the lower the probability of purchase for this product

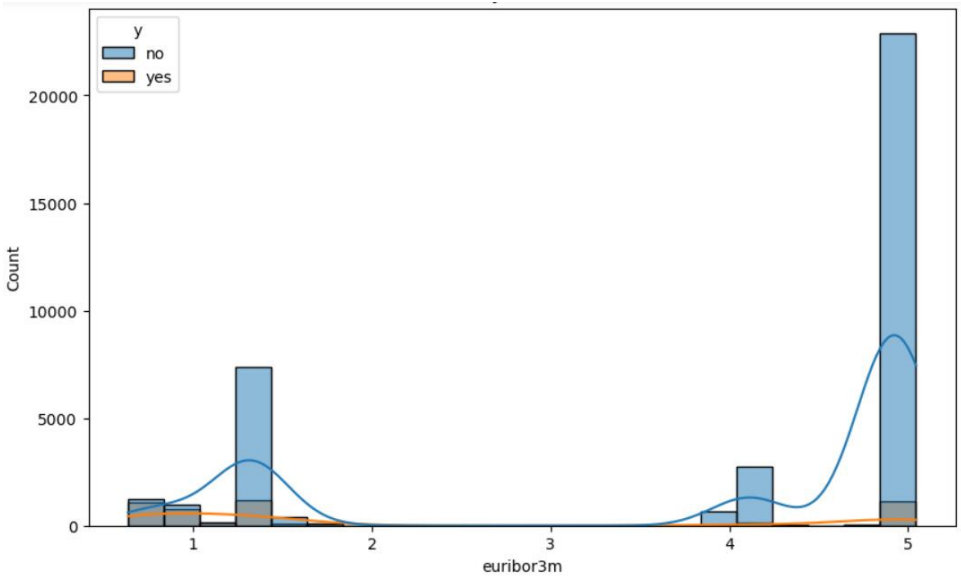
Employment Variation Rate





Generally, the probability of purchase is lower when the employment variation rate is higher

Euribor-3-month-rate

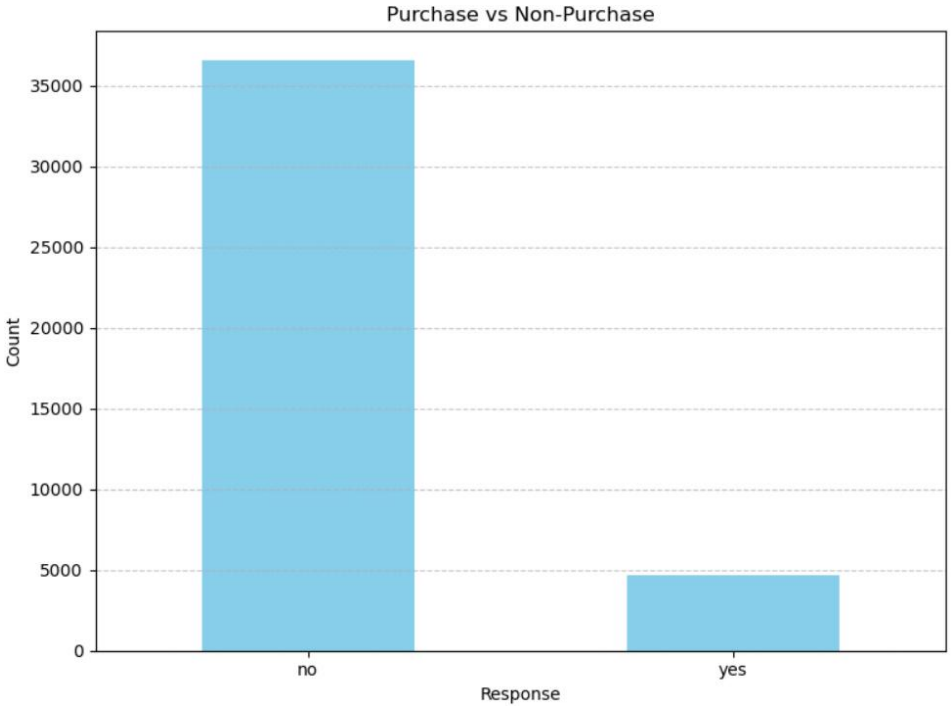


The interest rate at which a selection of European banks lend one another funds denominated in euros has negative correlation with the possibility of purchase

Model Suggestion

Random Forest

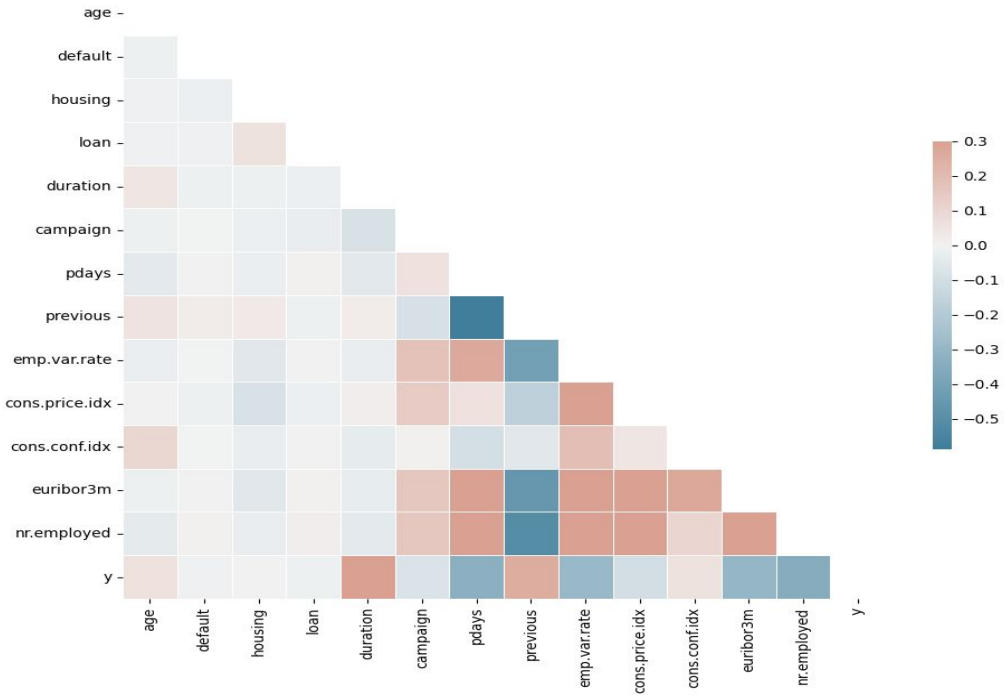
- **Handles Imbalanced Data**
- Robust to Overfitting
- Non-Linearity



The data is imbalance as the total number of ‘No’ is many more than the number of ‘Yes’ for target variable.

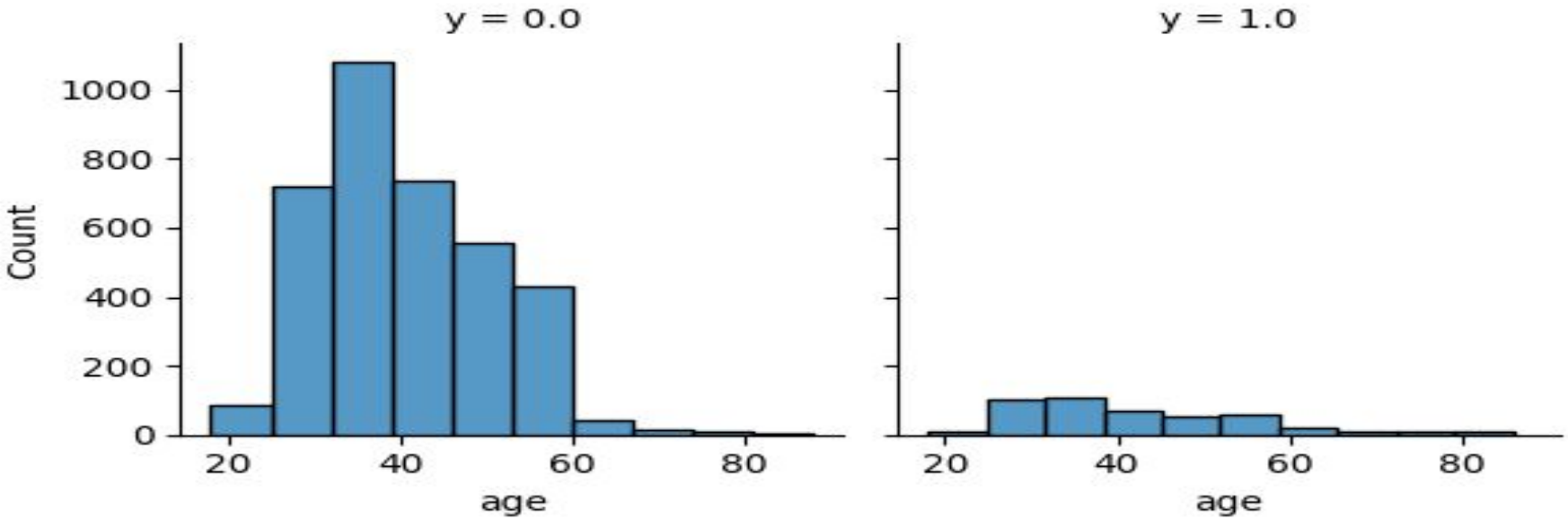
Bank-Additional.csv

Correlation of Independent Variables to Target



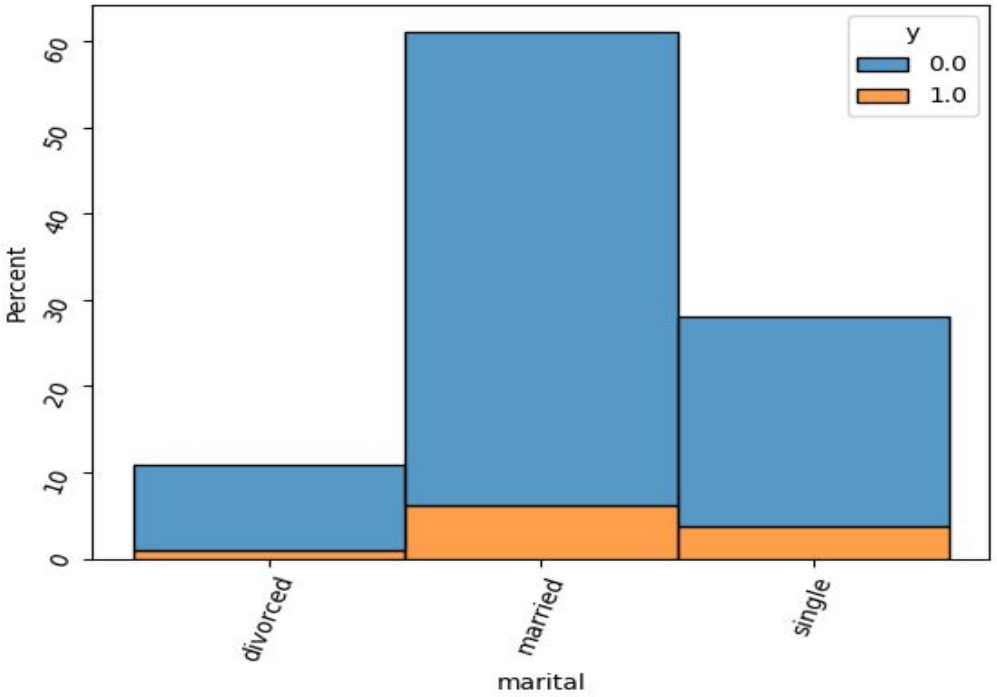
- Number of contacts performed in the 'Previous' campaign highly influency the current marketing campaign.
- Number of days 'pdays' of contact in-between campaigns has a negative correlation and thus stats that recently contacted customers can be influenced better to purchase the product.
- Similarly, 'duration' of last contact is highly correlated to probability of purchasing the product.
- The higher the total number of people employed 'nr.employed', the lower the probability of purchase for this product.
- Similarly, the employment variation rate ‘emp.var.rate’ and 3 month Euribor interest rate ‘euribor3m’ have a negative correlation to the probability of customer purchasing the product.

Influence of Age



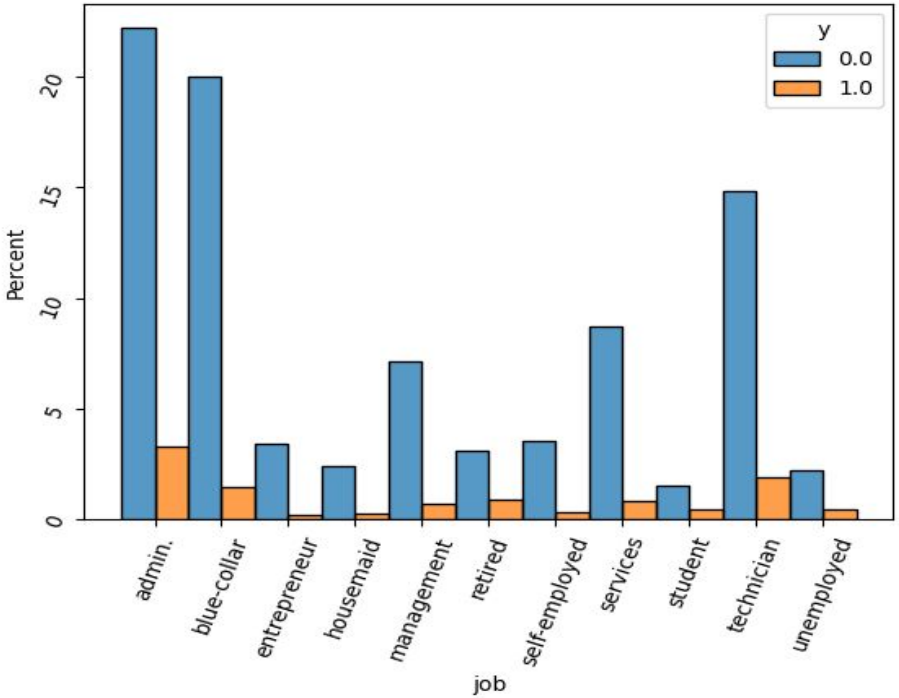
- Target the age group of 30 to 40 years old. This group has the highest percentage of people who have purchased the term deposit product.

Influence of marital status



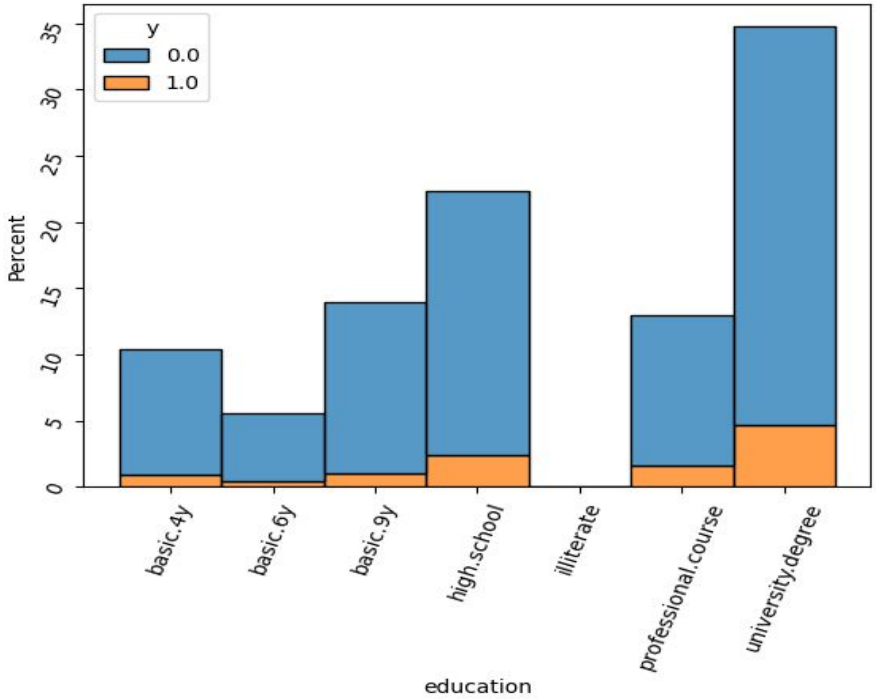
- Target single people. Single people are more likely to purchase the term deposit product than married or divorced people.
- 14% of Single people bought the product compared to 10% of married people and 9% of Divorced people.

Influence of Occupation



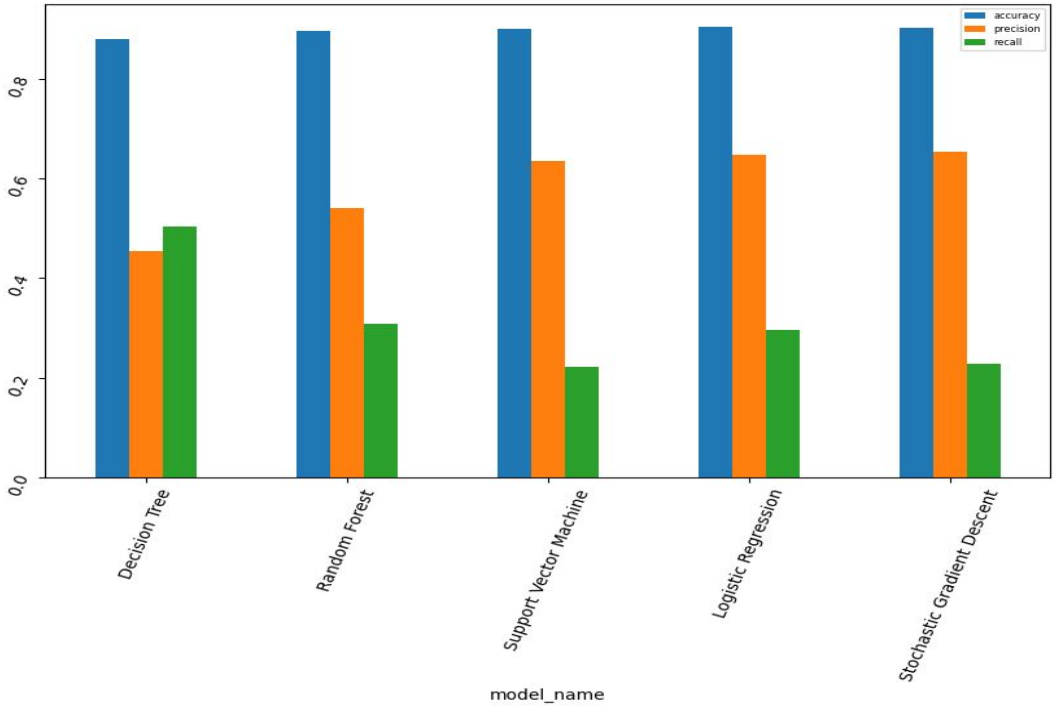
- Market the product more to retired people. Although the percentage of retired people who have purchased the product is lower than the percentage of people in the 30 to 40 age group, it is still significant.
- 25% of Retired people bought the product and the product should be marked to them more.

Influence of Education



- Target people with professional or university degrees. These people are more likely to have the financial means to purchase the term deposit product.
- People with professional course education - 15% bought the product and University degree education - 14% bought the product.

Model Recommendation



- Initial analysis suggests models such as Logistic Regression can provide better results.
- A voting classifier can be utilised to increase the precision and recall of the model outputs.
- Hyperparamter tuning is required using cross validation to improve the selected models performance.