G2M Case Study

Virtual Internship

14-Jan-2024

Background –G2M(bank marketing) case study

- ABC Bank wants to sell its 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).
- Objective : Bank wants to use ML model to shortlist customer whose chances of buying the product is more so that their marketing channel (tele marketing, SMS/email marketing etc) can focus only to those customers whose chances of buying the product is more.

The analysis has been divided into two parts:

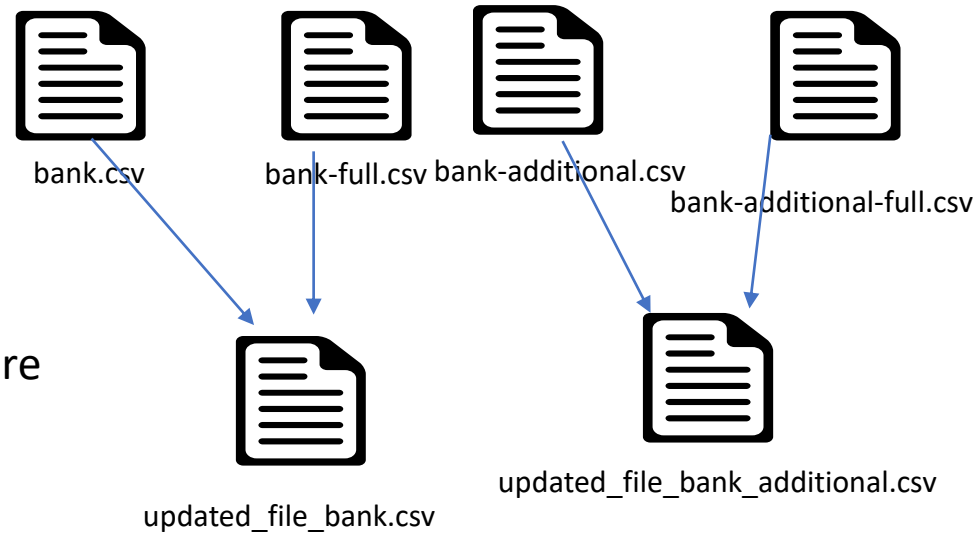
- Data Understanding
- Recommendations

Data Exploration

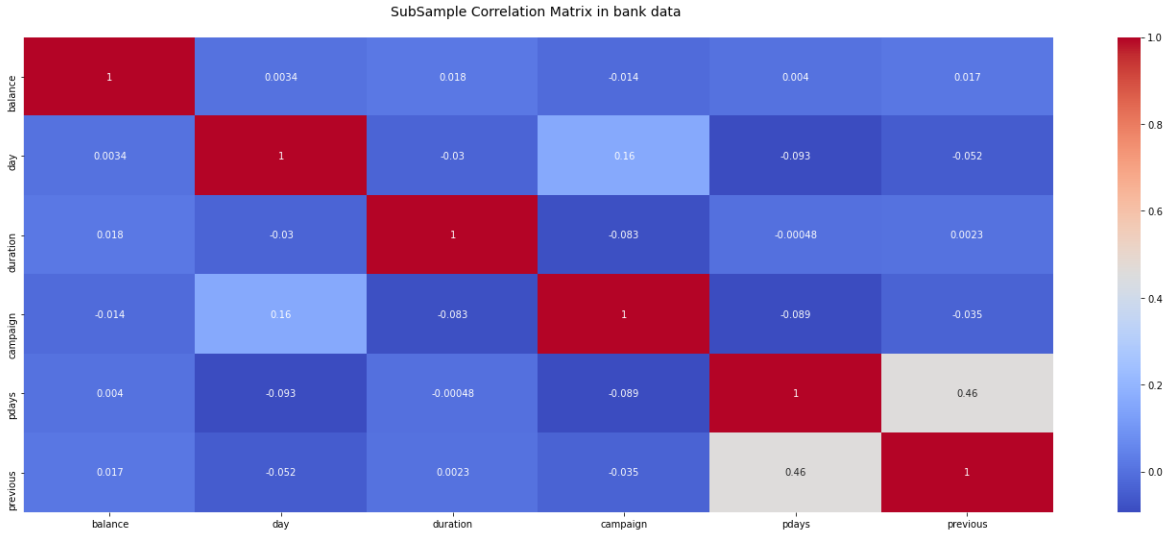
- 37 Features
- Timeframe of the data: 2014
- Total data points: 95,037

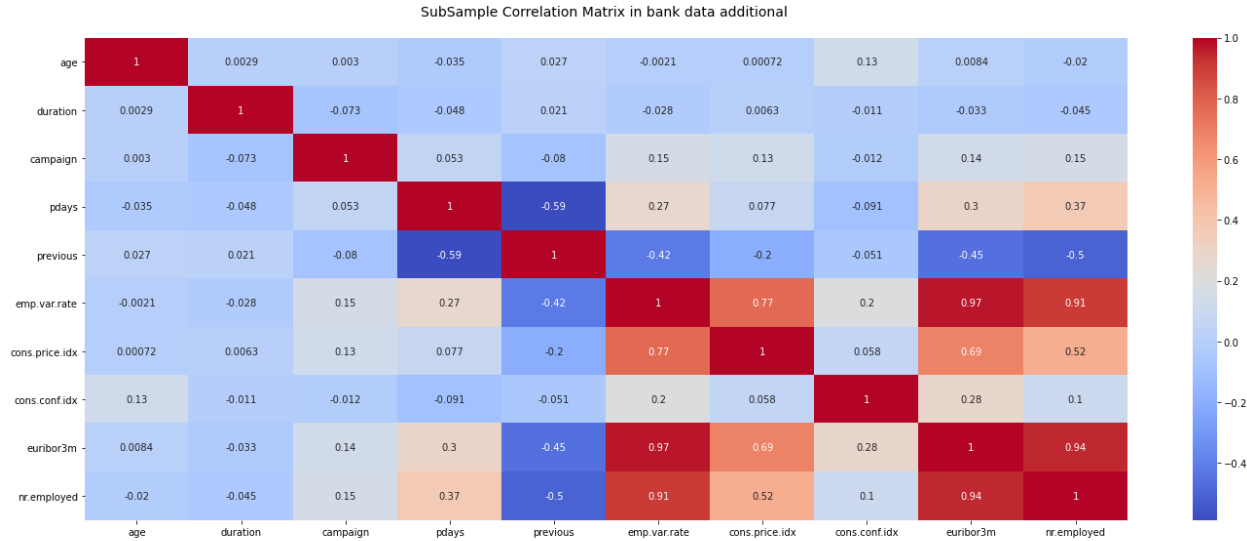
Hypotheses and assumptions:

- The bank-additional csv file would have more data that the original, with more rows and more detailed features
- There is a higher chance of selling the product to more educated, younger, familiar with current technology audience
- Married people are more likely to buy the product than single people
- By comparing the data between the original bank data and the additional bank data, we would notice that the prediction could be made by using only the additional bank data
- There is no strong correlation between the features which could influence the classification



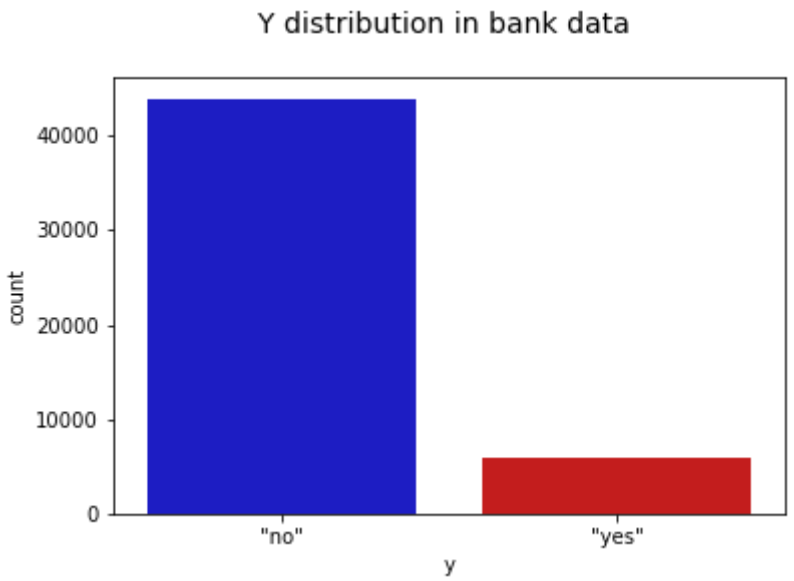
Correlation Matrix for both Datasets

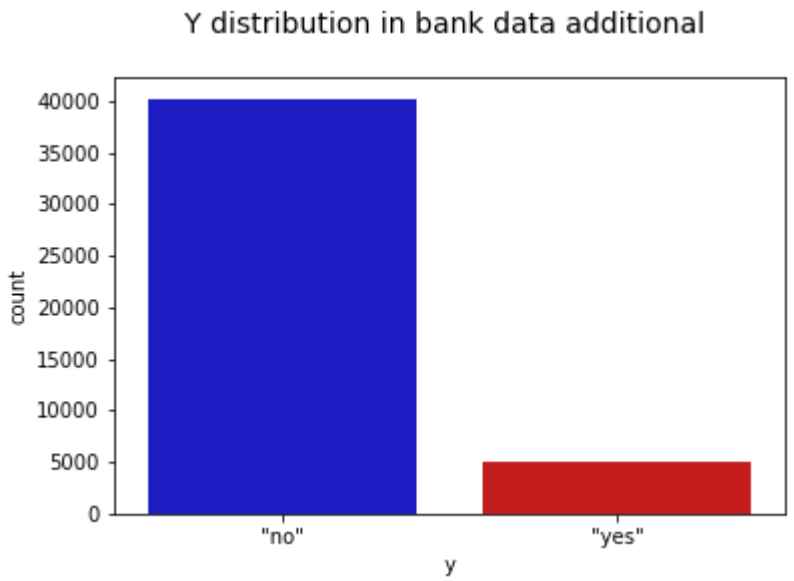




- In both datasets, the correlations between the features are not strong
- However, in the cases where the correlation is strong, those correlations do not influence the outcome of the classification

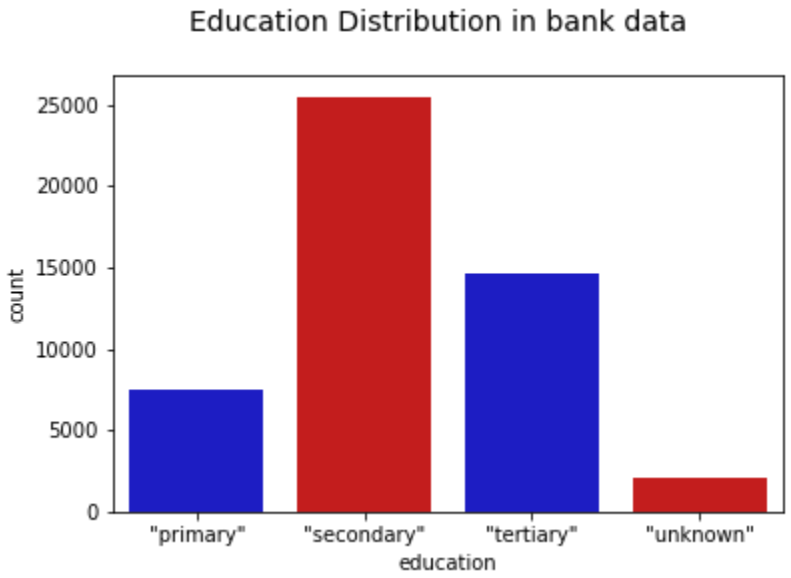
Y Distribution in both Datasets

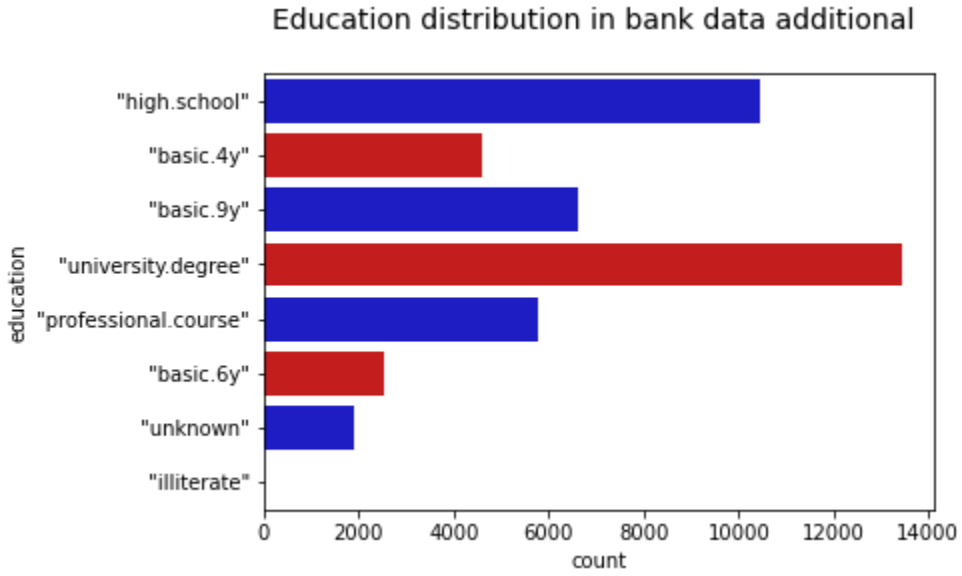




- The distribution of the Y feature in both datasets are very similar, with more rows in the original bank data

Education Distribution in both Datasets

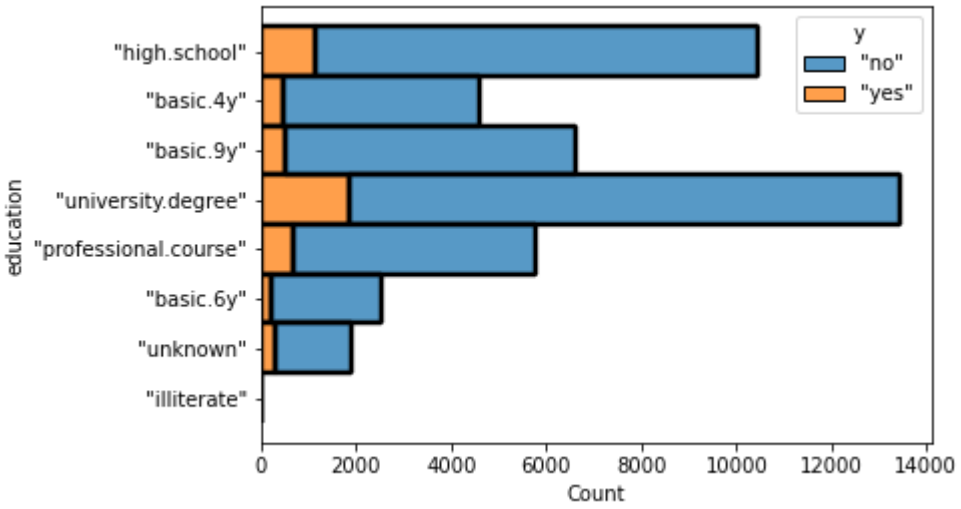




- As observed above, in the bank additional dataset, the data is more detailed, with on accurate educational description of the possible future clients
- With that in mind, the features in the bank additional dataset is chosen as the primary dataset in discussing data understanding
- However, it is important to take into consideration the original bank dataset
- The original bank dataset could be used in the comparison between the datasets to see if the similarities between them are constant throughout the datasets

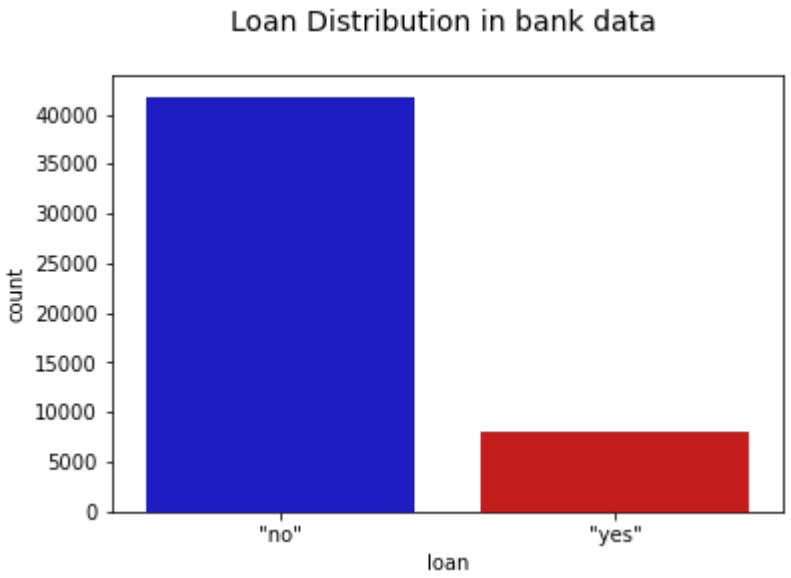
More Detailed Education Distribution

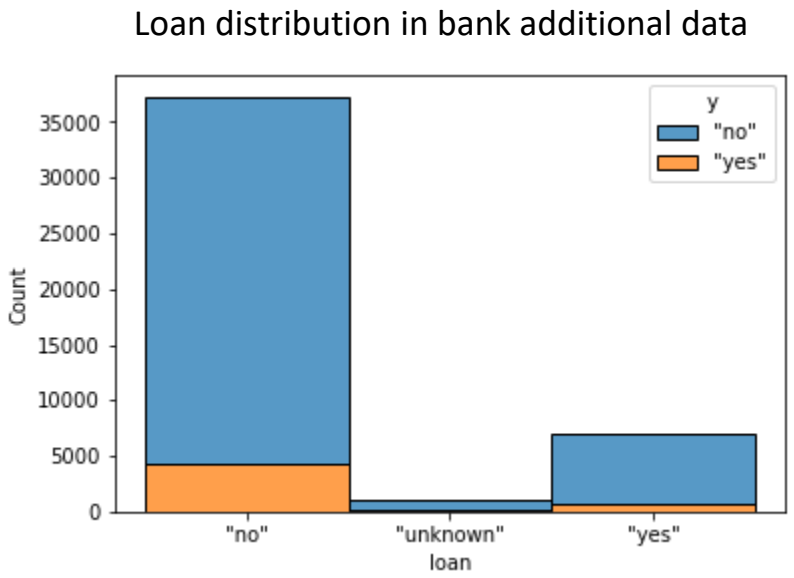
Education distribution in bank additional data



- As seen here, most of the clients have higher education background

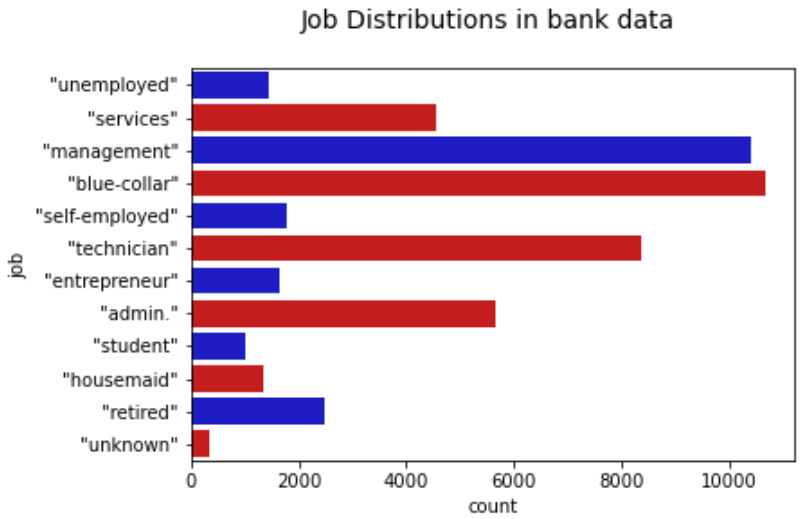
Loan Distribution in both Datasets

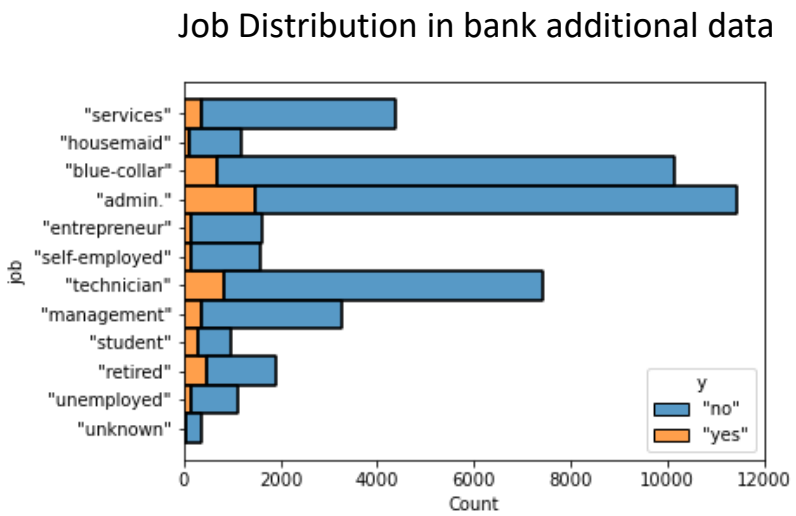




- As seen above, most of the customers do not have loans
- There are some unknowns in the dataset

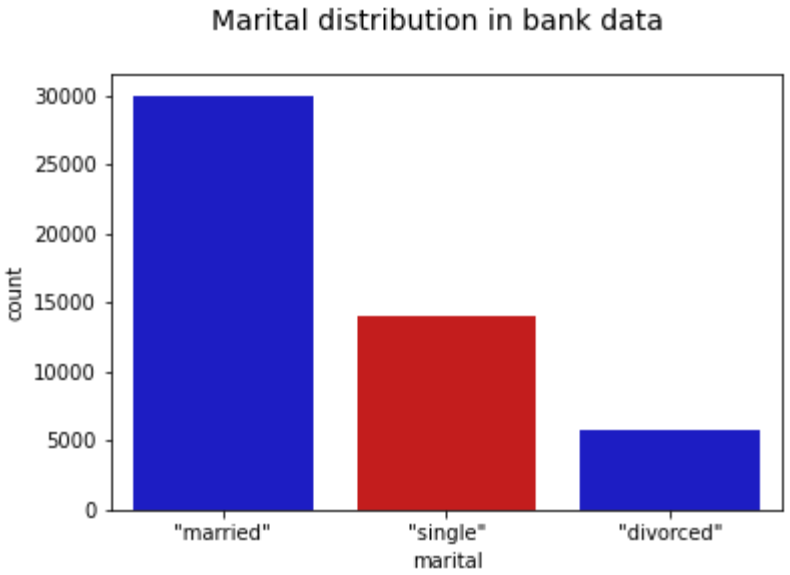
Job Distribution in both Datasets



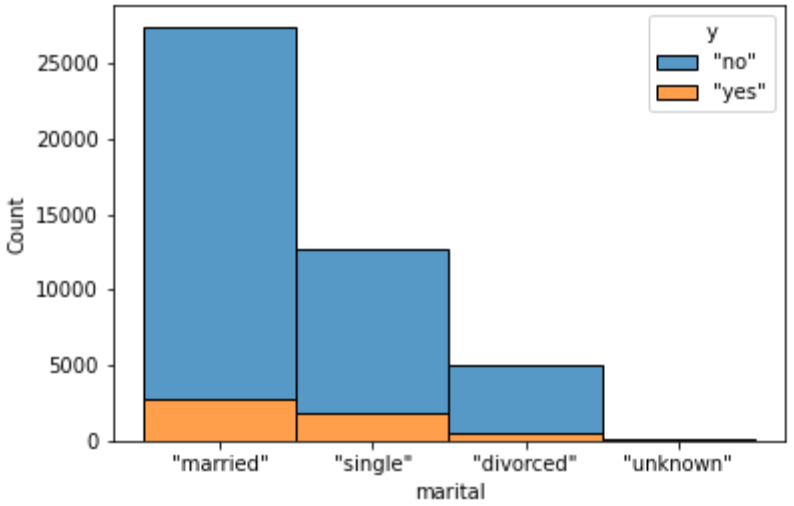


- Both datasets have similar details
- However, in the original, most of the clients have management and blue-collar backgrounds, while the bank additional have its clients with administration and blue-collar background

Loan Distribution in both Datasets

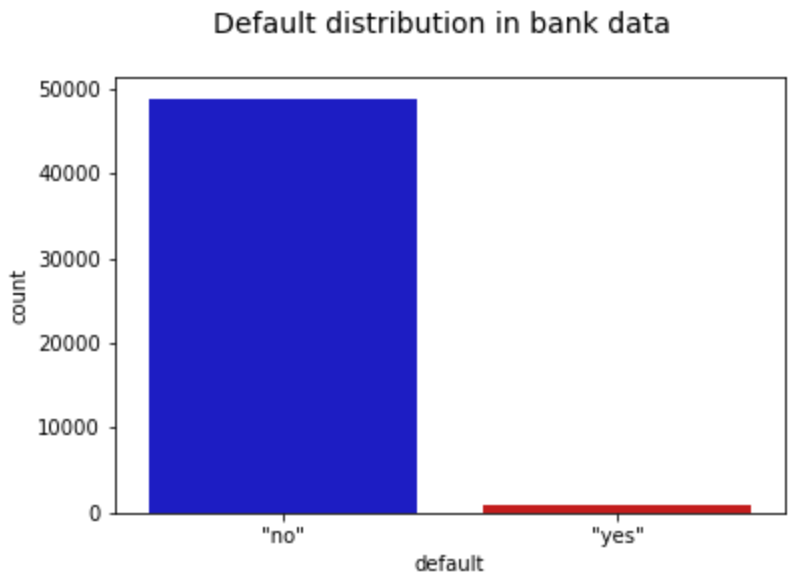


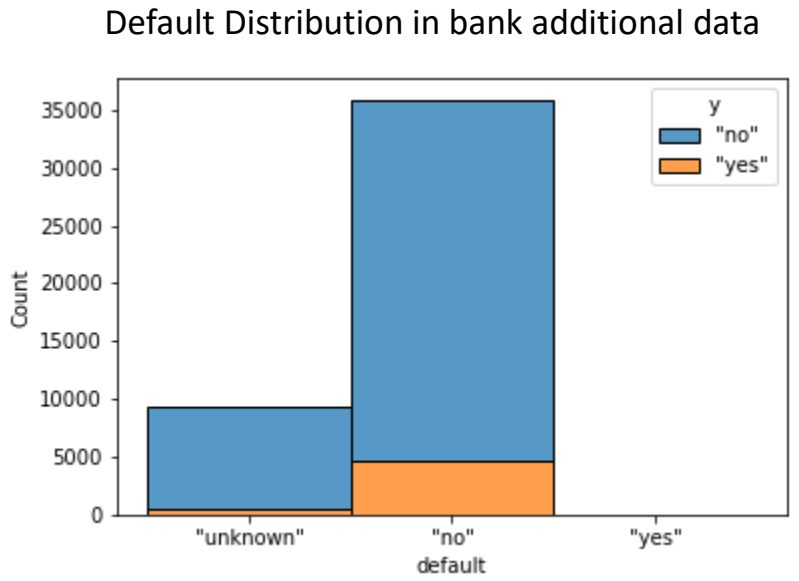
Marital Distribution in bank additional data



- Most of the clients are married in both datasets
- However, among the groups, single people are more likely to purchase the product

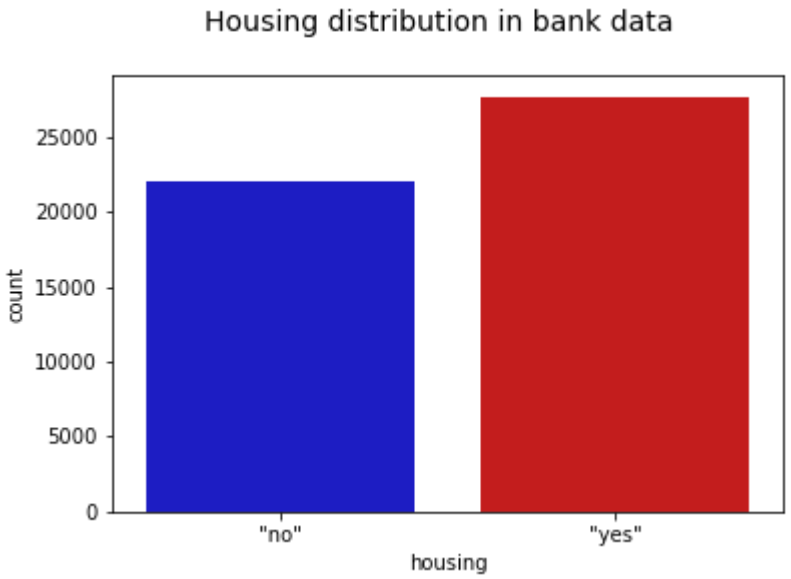
Default Distribution in both Datasets



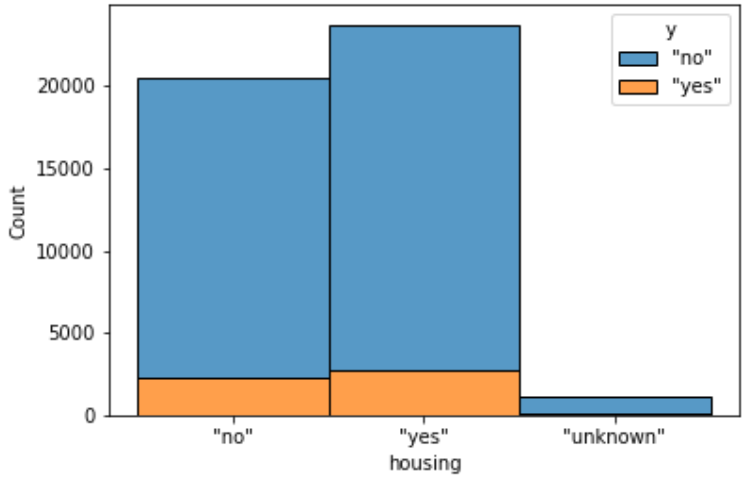


- In both datasets, the default count is barely noticeable

Profit Analysis

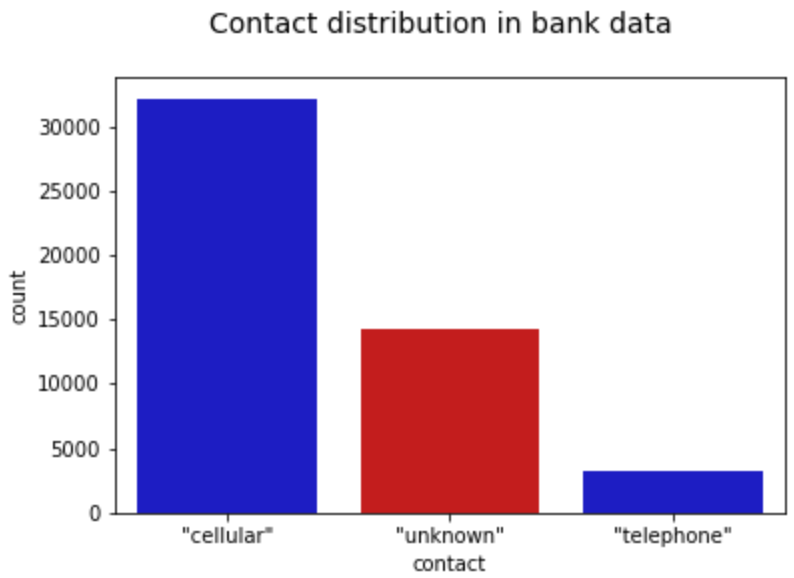


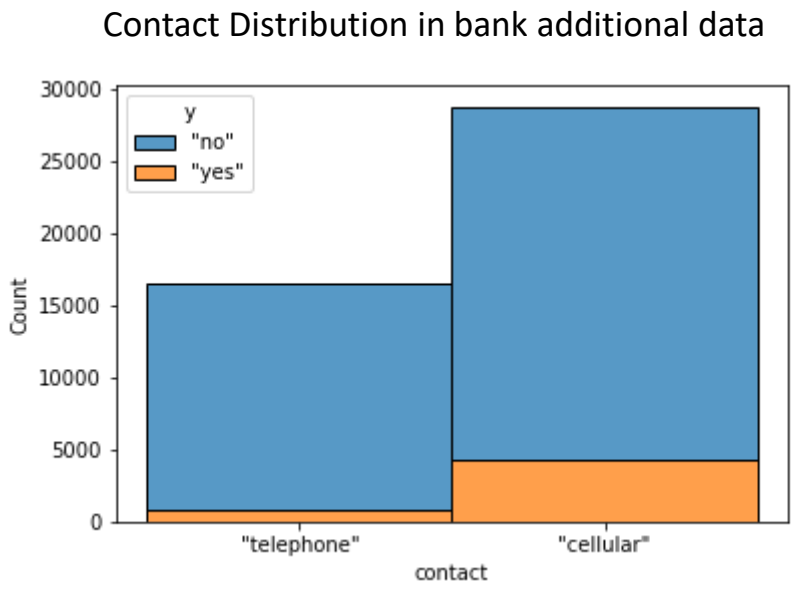
Housing Distribution in bank additional data



- The distribution of the Y feature in both datasets are very similar, with more rows in the original bank data

Contact Distribution in both Datasets

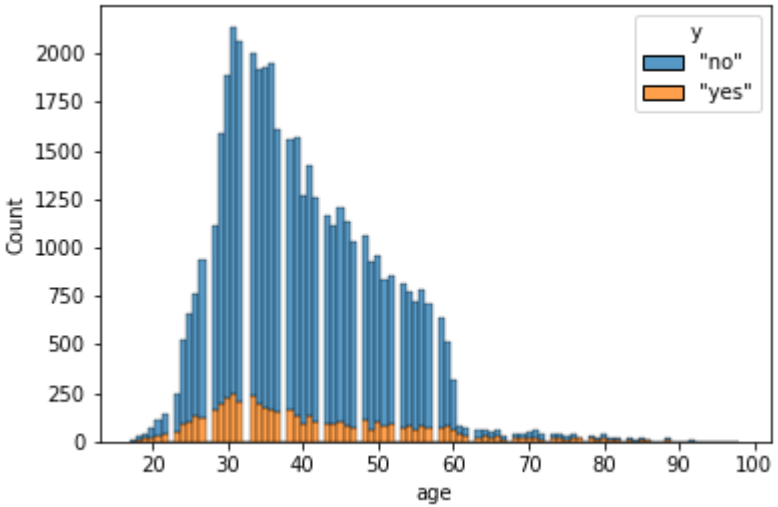




- As seen above, cellular is more popular in both datasets

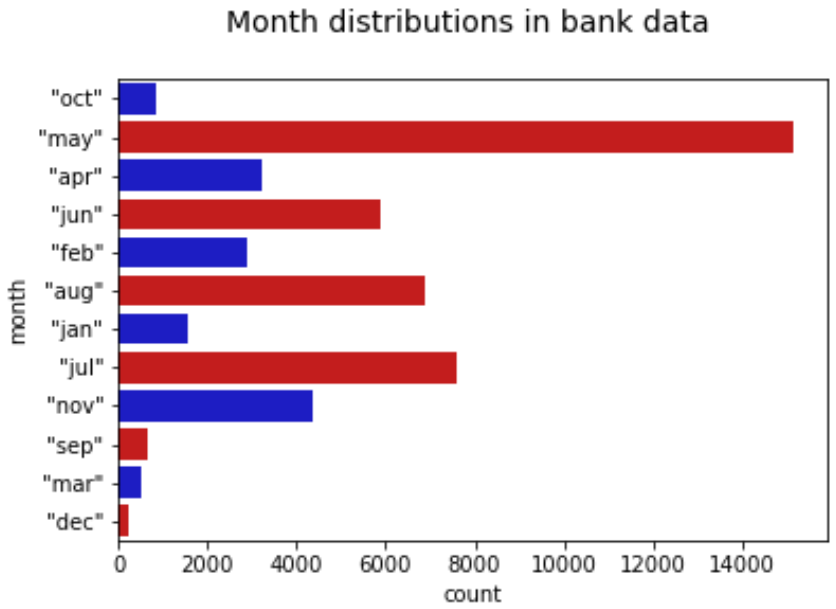
Age Distribution in both Datasets

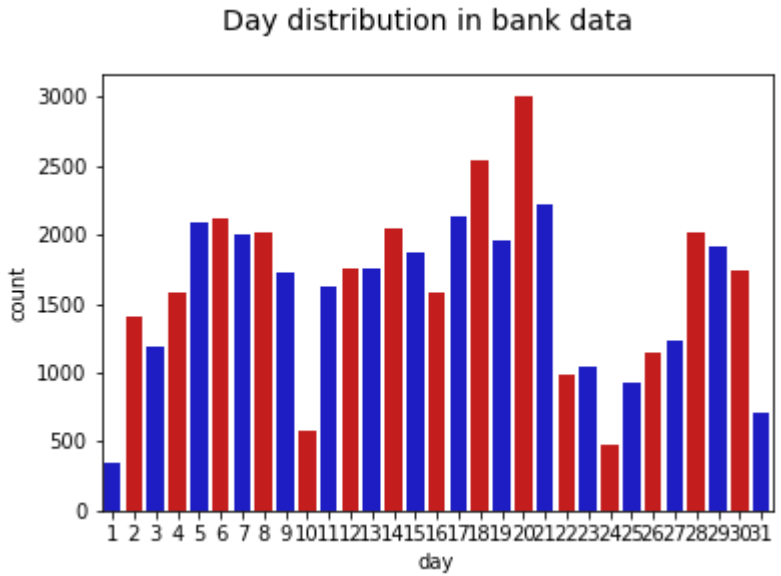
Age Distribution in bank additional data



- As seen here, most of the customers have their age within the range between 20 and 60
- The 30-to-40-year age group is the biggest group
- Most of the purchases come from the 20-to-40 age group

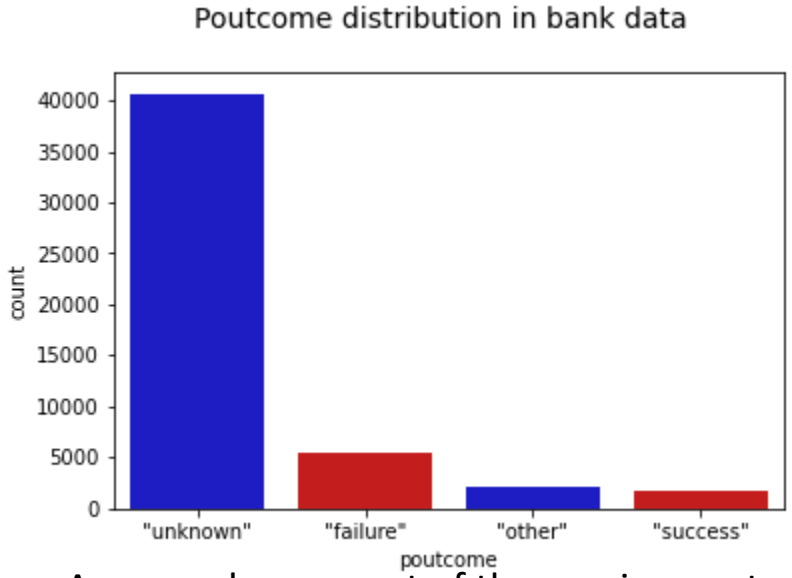
Month and Day Distribution in both Datasets

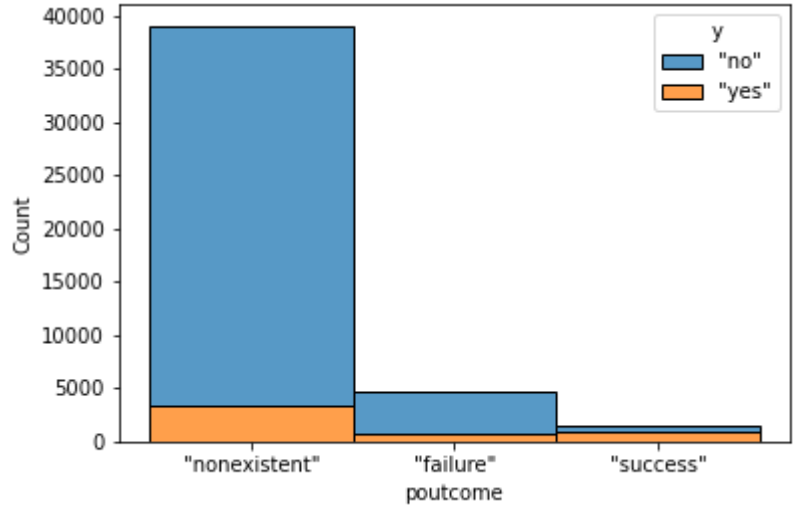




- As seen above, only the original bank dataset has the day and month features
- As notices, most of the purchases are made between May and August
- Also, most of the purchases are made in the middle of the month

Previous Outcome Distribution in both Datasets





- As seen above, most of the previous outcomes are unknown or nonexistent, depending on the dataset
- However, in the case of successful outcomes, the customers made a purchase

Hypothesis results

Based of testing the hypothesis, it has been found that:

- The bank-additional csv file did not have more data that the original. However, it has more detailed features
- There is a higher chance of selling the product to more educated, younger. When it comes to familiarity with current technology, it does not matter
- More married people bought the product. However, single people are more likely to buy the product
- By comparing the data between the original bank data and the additional bank data, we notice that the datasets are similar throughout identical features, the bank additional dataset contains more detailed data
- There is no strong correlation between the features which could influence the classification

Recommendations

We have evaluated both the cab companies on following points and found Yellow cab better than Pink cab:

- **Customer Reach :** Marketing should be made with the focus on cellular, rather than telephone, due to the fact most of the customers are using cellular as a contact option
- **Age wise Reach :** The product was purchases by clients from of age groups. However, it is very popular with the 20-to-40-year age group
- **Recommended models :** Logistic Regression, K Nears Neighbor, Support Vector Classifier, Decision Tree Classifier, Random Forrest

Thank You