Exploratory Data Analysis

Virtual Internship

Aryaman Srivastava

Agenda

- Executive Summary
- Problem Statement
- Approach
- EDA
- EDA Summary
- Recommendations
- Model Recommendation

Executive Summary

Perform an Exploratory Data Analysis on data of customers to determine which attributes make customers more likely to buy the product, and based on this information, which type of customers to shortlist.

Problem Statement

Background: 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)

Objective: Analyze the customer data to determine which demographics of customers are more likely to buy the product.

Approach

- The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.
- For this data, we have 4 files, all in csv format that correspond to the same large dataset. We can categorize the files into 2 categories: bank and bank-additional. The bank-additional files has 20 attributes while the bank files has only 16 attributes. By merging the two files in each category and performing outlier removal and filling all NA values with the mean, we have 66999 data points.
- Assumptions
  - All values for the month and day columns are valid months/days
  - The duration feature is expressed in seconds
  - Day of the week values are valid(MTWTFSS)
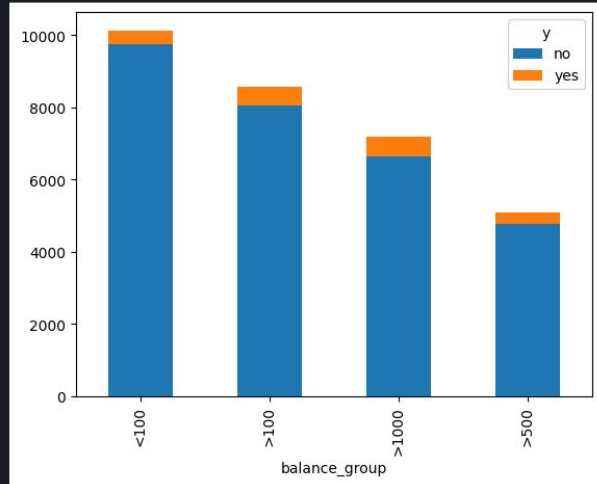
Occupation and Age Wise Analysis
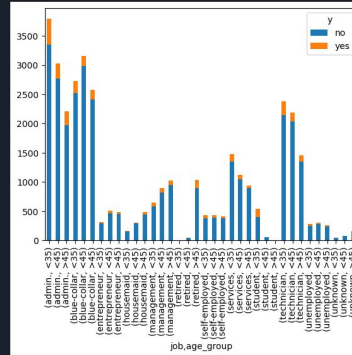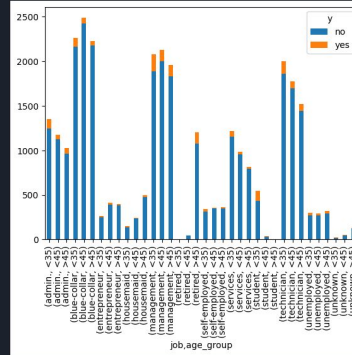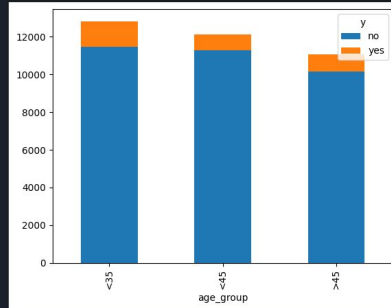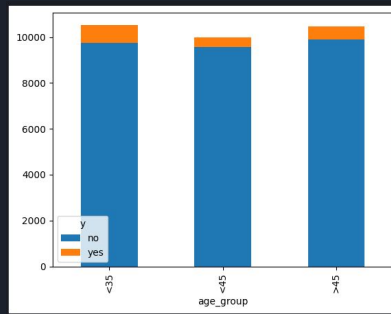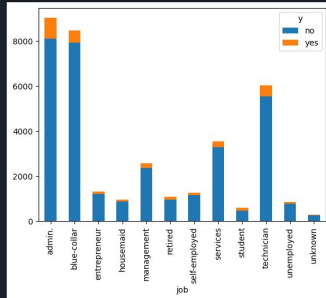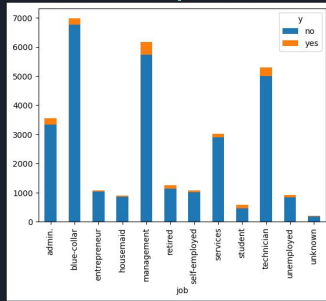
Clients who work in management, blue-collar, or technician and are older than 45 or younger than 35 buy the product the most

Marital status and Education level Analysis
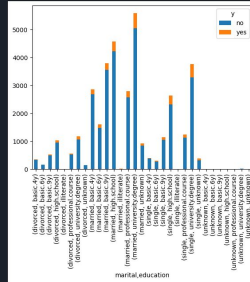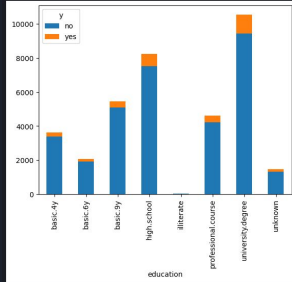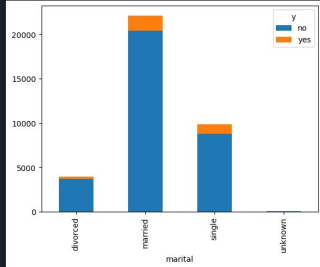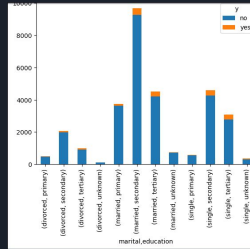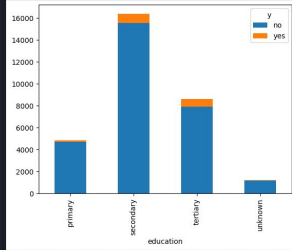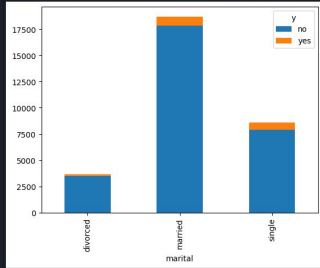
The most amount of people who bought the product were people who were married and at least had their secondary education
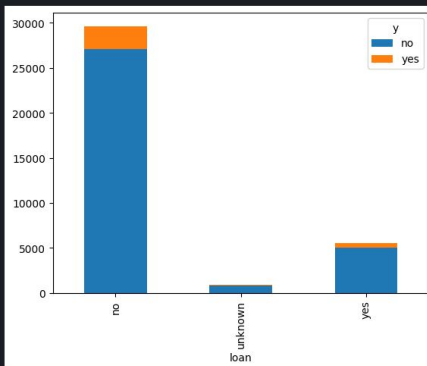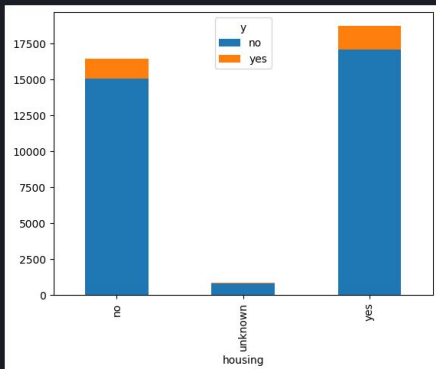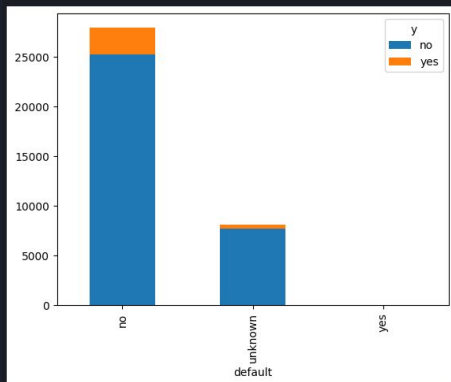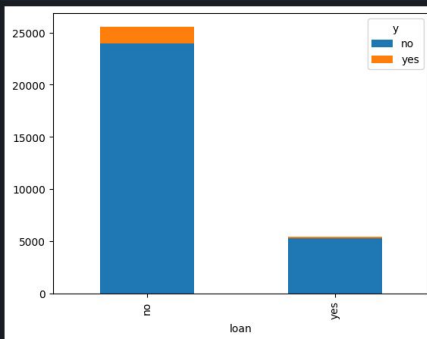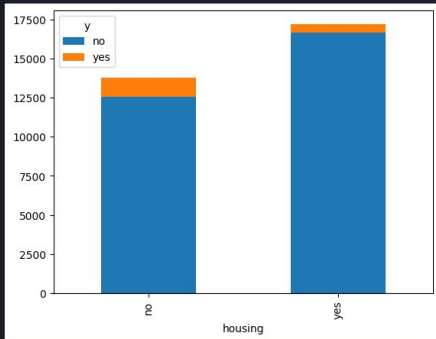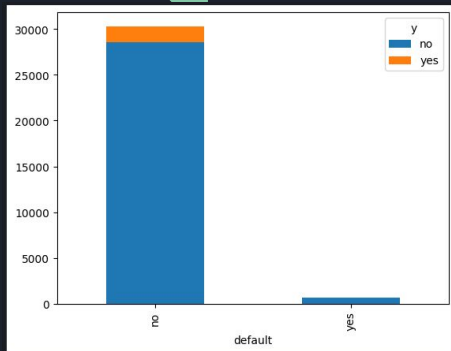
Balance analysis

Customers who had a balance greater than 1000 bought the most of the product, with customers who had a balance between 500 and 1000 and less than 100 buying the product the least
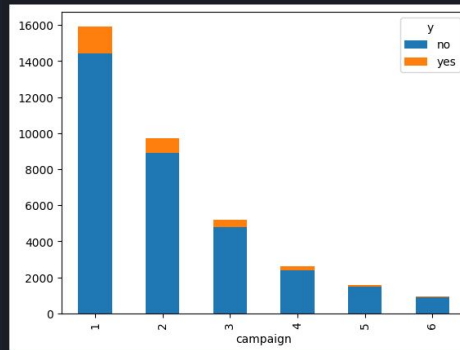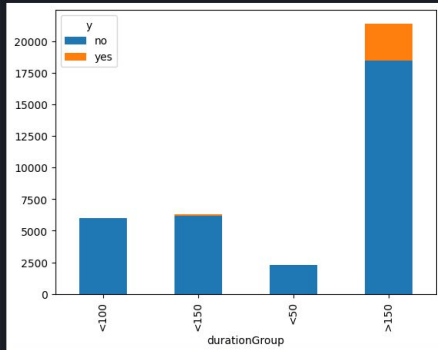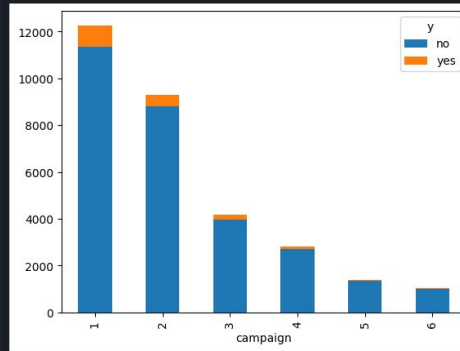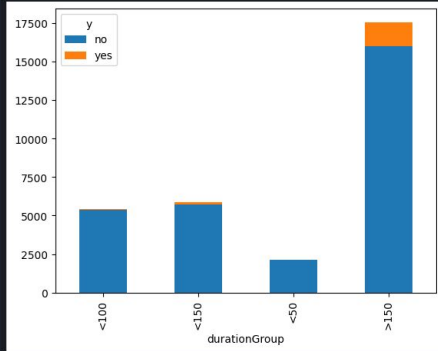
Loan/credit analysis

Clients who don't have a personal/housing loan and don't have credit in default are more likely to buy the product

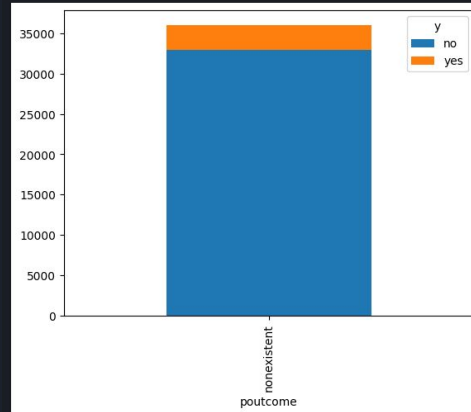# Campaign Analysis

Client's who had a small number of contacts and a longer duration of last contact are more likely to buy the product

Contact Analysis
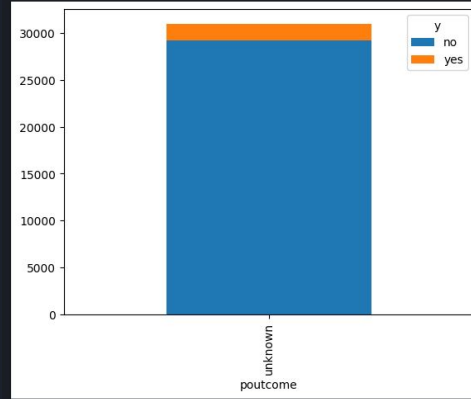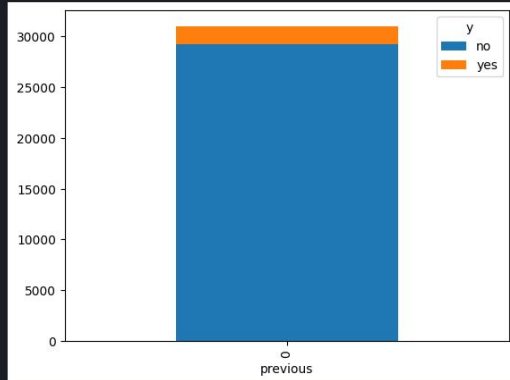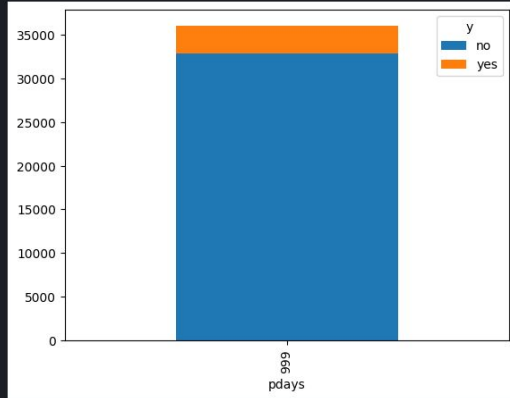
It seems that the outcome of the previous marketing campaign, number of contacts performed before the contact, and the number of days since the client was last contacted doesn't seem to affect the client's decision in buying the product or not

Index Analysis

Clients with a price index close to 93, a confidence index between -42 and -47, and employment variation rate -2.9, -1.8, and 1.4 are more likely to buy
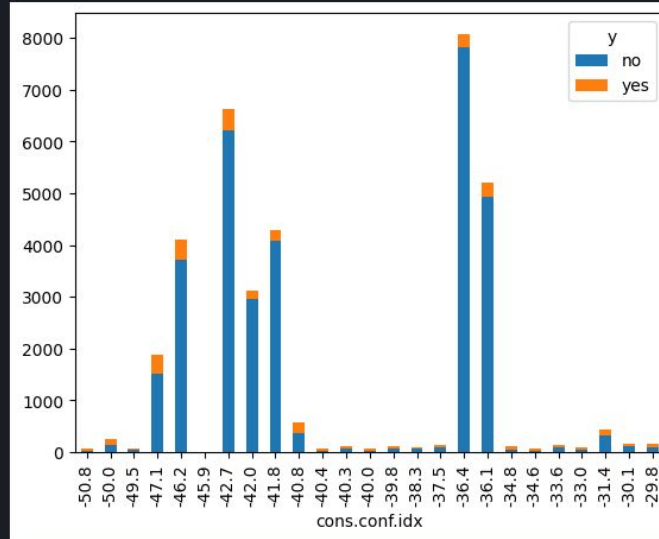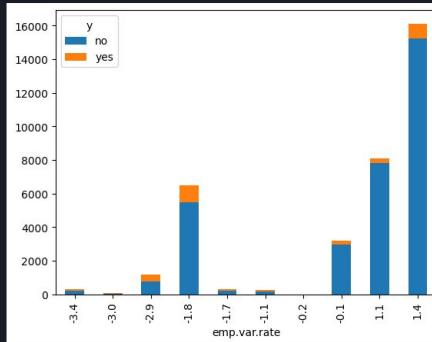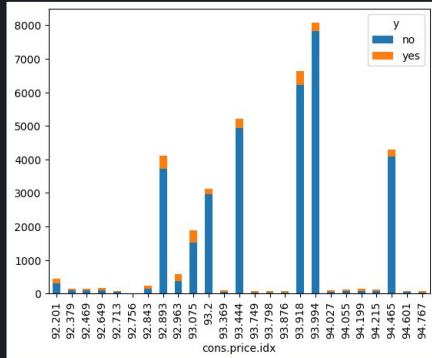
Indicator Analysis
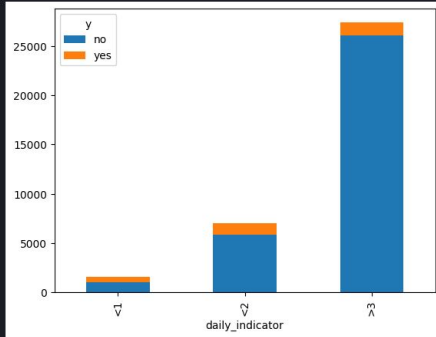
Clients with a higher 3 month rate and a number of employees above 5000 are more likely to buy the product.
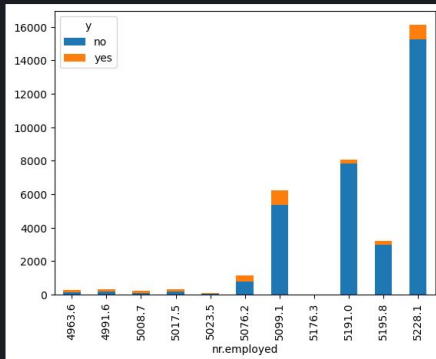
Final Recommendations

Based on the data and subsequent exploratory data analysis, I believe it would be best to shortlist clients who are less than 35 years of age or greater than 45 years of age, clients who are married and at least have had their secondary education. In addition, we should shortlist clients who work in management, blue-collar, or technician. Clients who have a balance higher than 1000 are more likely to buy the product. Clients who don't have any type of loan(personal or housing) and doesn't have credit in default are more likely to buy the product and should be shortlisted. Clients who were only contacted once and had a longer duration of contact should be shortlisted. The outcome of the previous marketing campaign, number of contacts performed before the contact, and the number of days since the client was last contacted doesn't seem to affect the likelihood that a client will buy the product. Clients with a higher 3 month rate and a number of employees above 5000 should be shortlisted. Clients with a price index close to 93 should be shortlisted. Clients with a confidence index between -42 and -47 should be shortlisted. Finally, clients with employment variation rate -2.9, -1.8, and 1.4 should be shortlisted

Model Recommendation

- Recommend using all attributes except for:
    - pdays
    - poutcome
    - Previous
    - These attributes seem to have no statistical significance on the buyer's decision
- As quite a few attributes are strings and not just numerics, I recommend a model for boosting where we can assign weights to different string values