Exploratory Data Analysis

Bank Marketing Case Study

11/10/2023

Jinghao Shen

Agenda

Problem Statement

Data Exploration

EDA

EDA Summary

Recommendations

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).

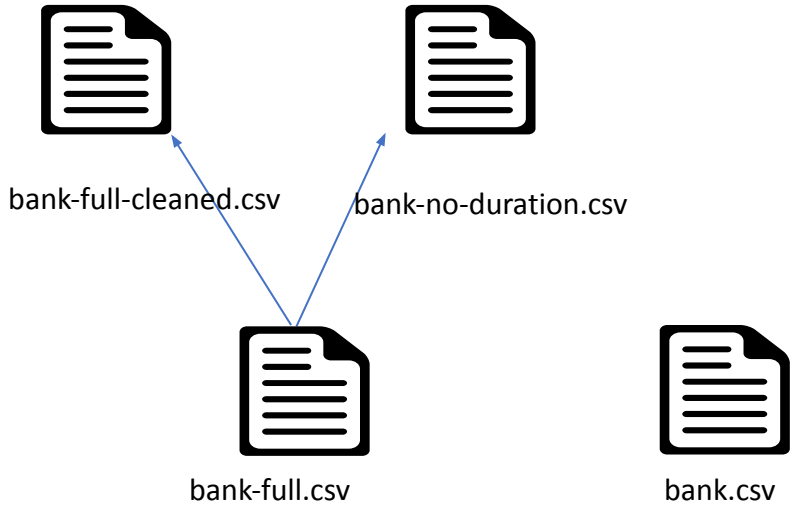
- Object: Develop an ML model to shortlist customer whose chances of buying the product is more so that their marketing channel can focus only to those customers whose chances of buying the product is more.

Data Exploration

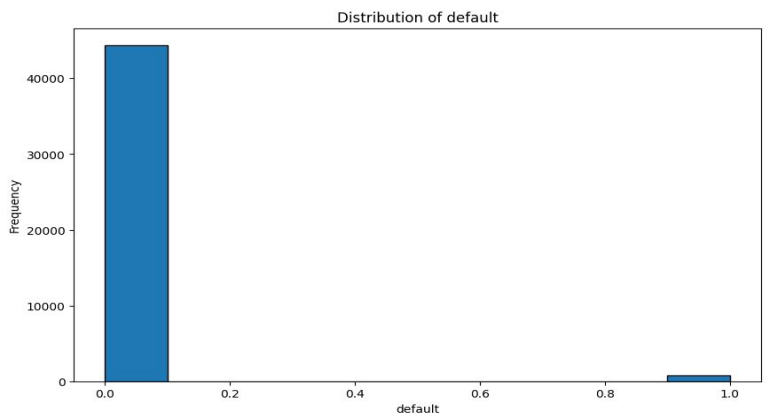
- Features: 16+1
- Time period of data: 2012
- Total number of observations: 45211

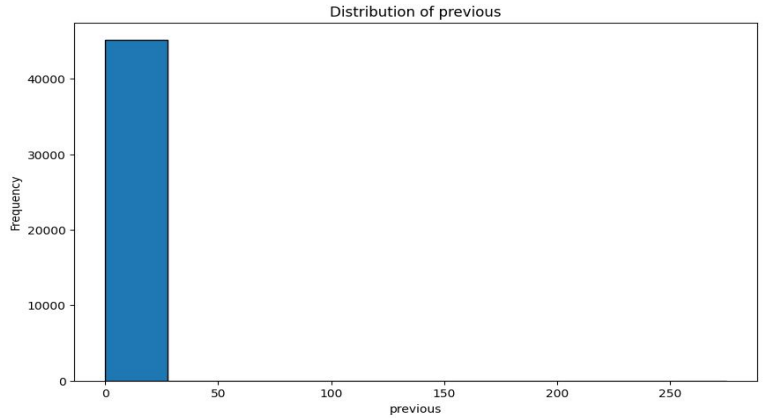
Transformations:

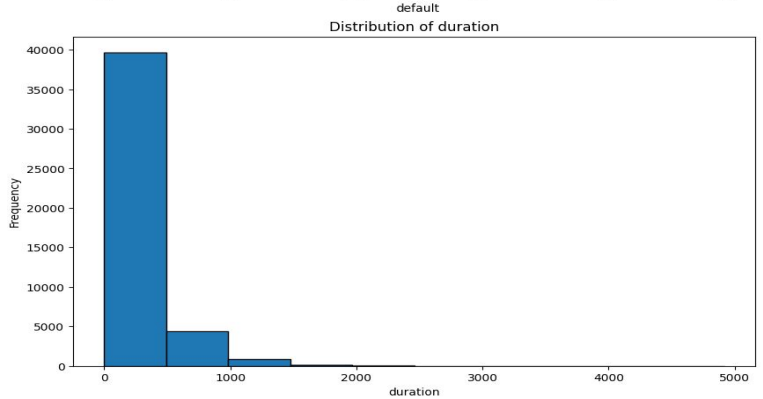
1. bank.csv with 10% of the sample randomly selected from bank-full.csv
2. bank-full.csv is transformed to bank-full-cleaned.csv that all non-numerical attributes is turned into numerical values.
3. bank-no-duration.csv removes ‘duration’ attribute from bank-full-cleaned.csv

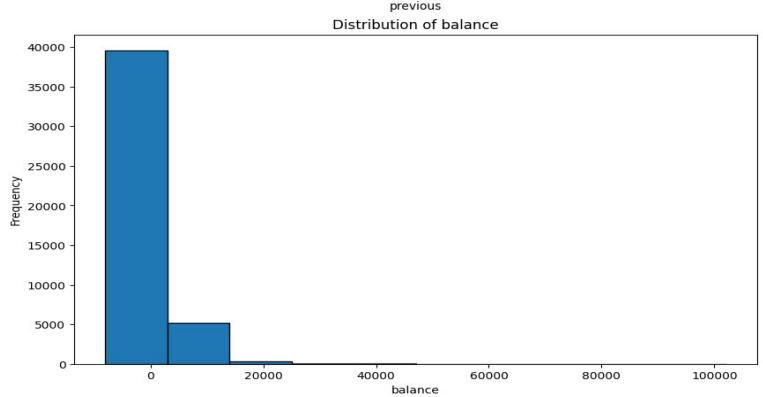


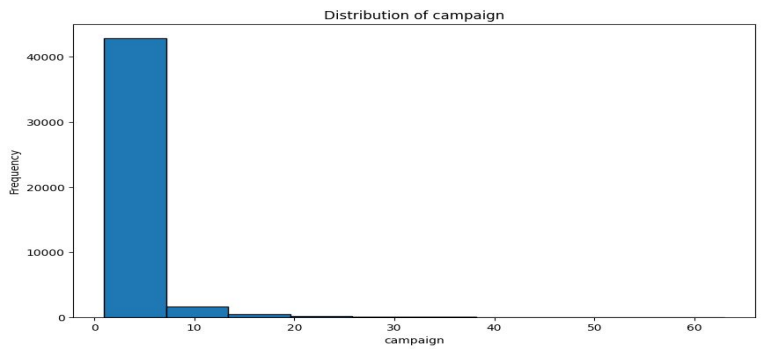
Attributes Analysis

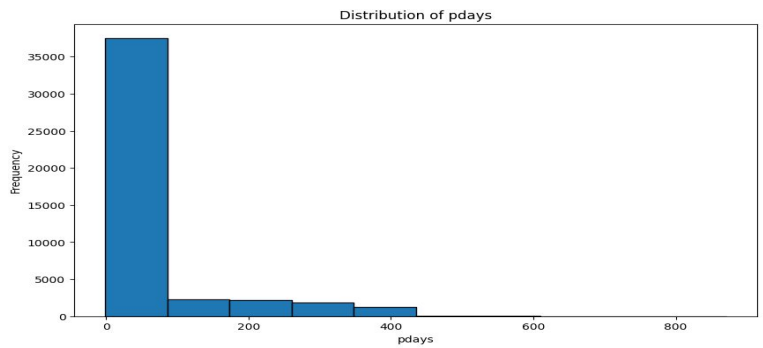






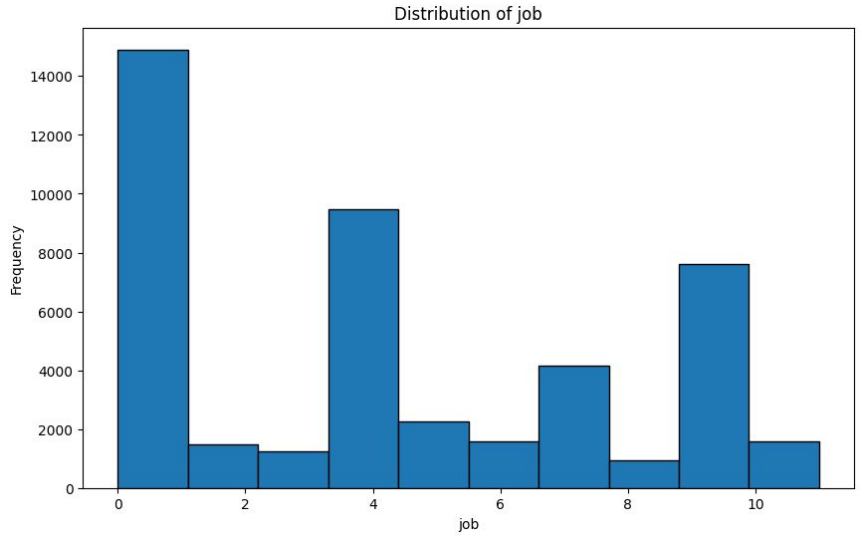


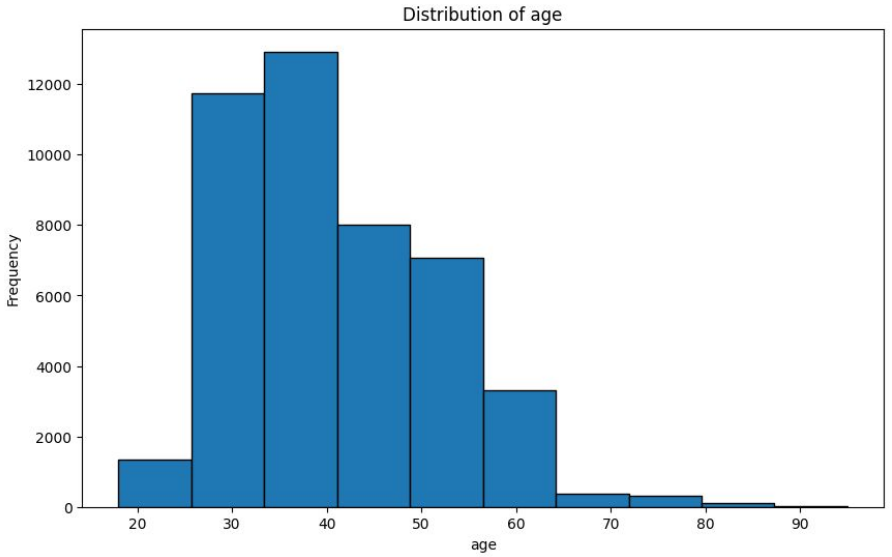


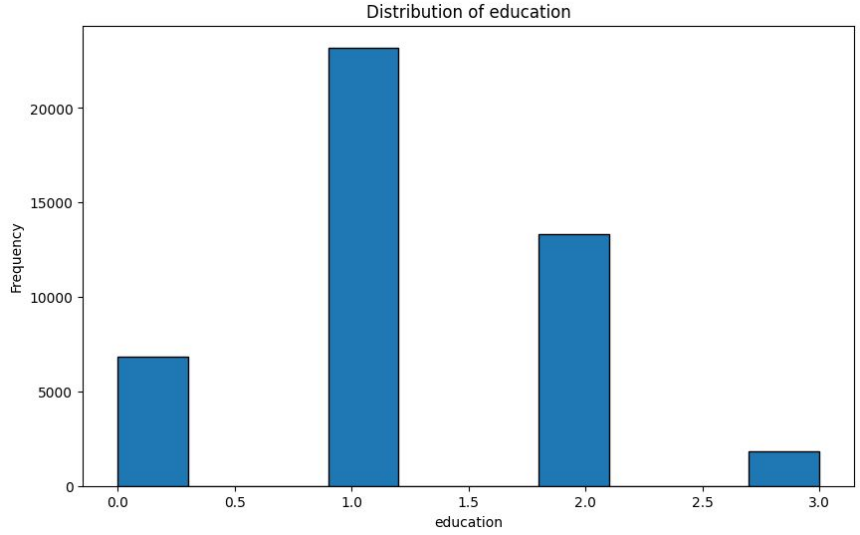


- We can observe long-tail distribution in several attributes.
  - Number of instances in head classes are way more than instances in tail classes.
- Missing values are replaced as ‘unknown’.

Attributes Analysis (client data)





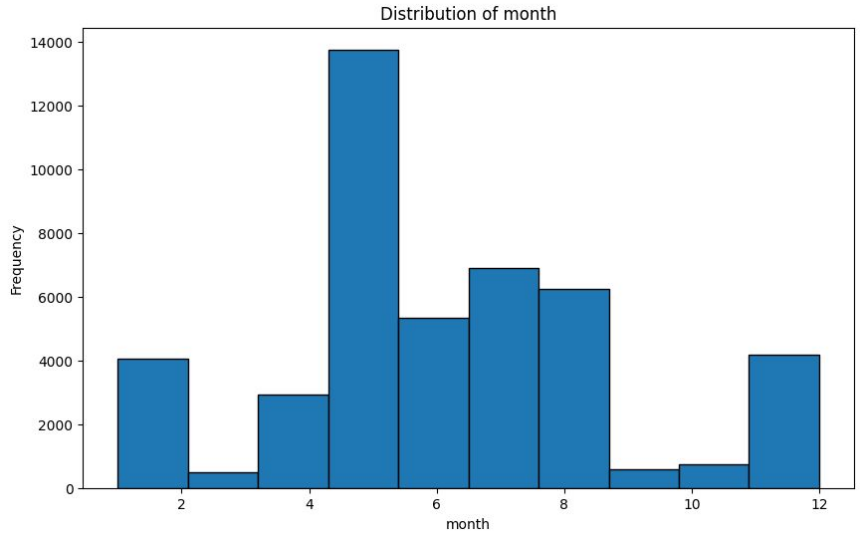


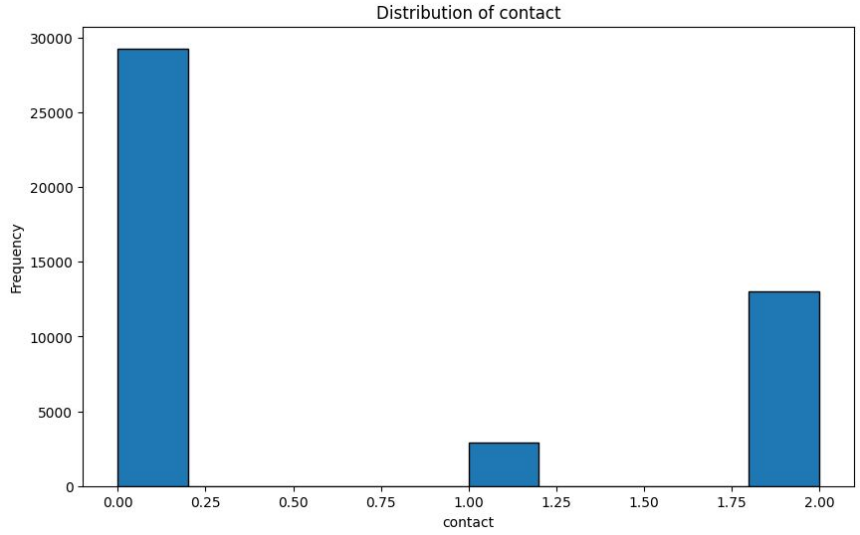
- Most client age falls within the range 30 to 40
- Top three client jobs are: blue-collar, management, and technician.
- Most client have an education level higher than secondary school.

Assuming: education:

- secondary: 1
- tertiary: 2
- primary: 0
- unknown: 3

Attributes Analysis (contact)





Most contact happened in May.

Less contact in March and September.

Cellular Phone is the most used contact approach.

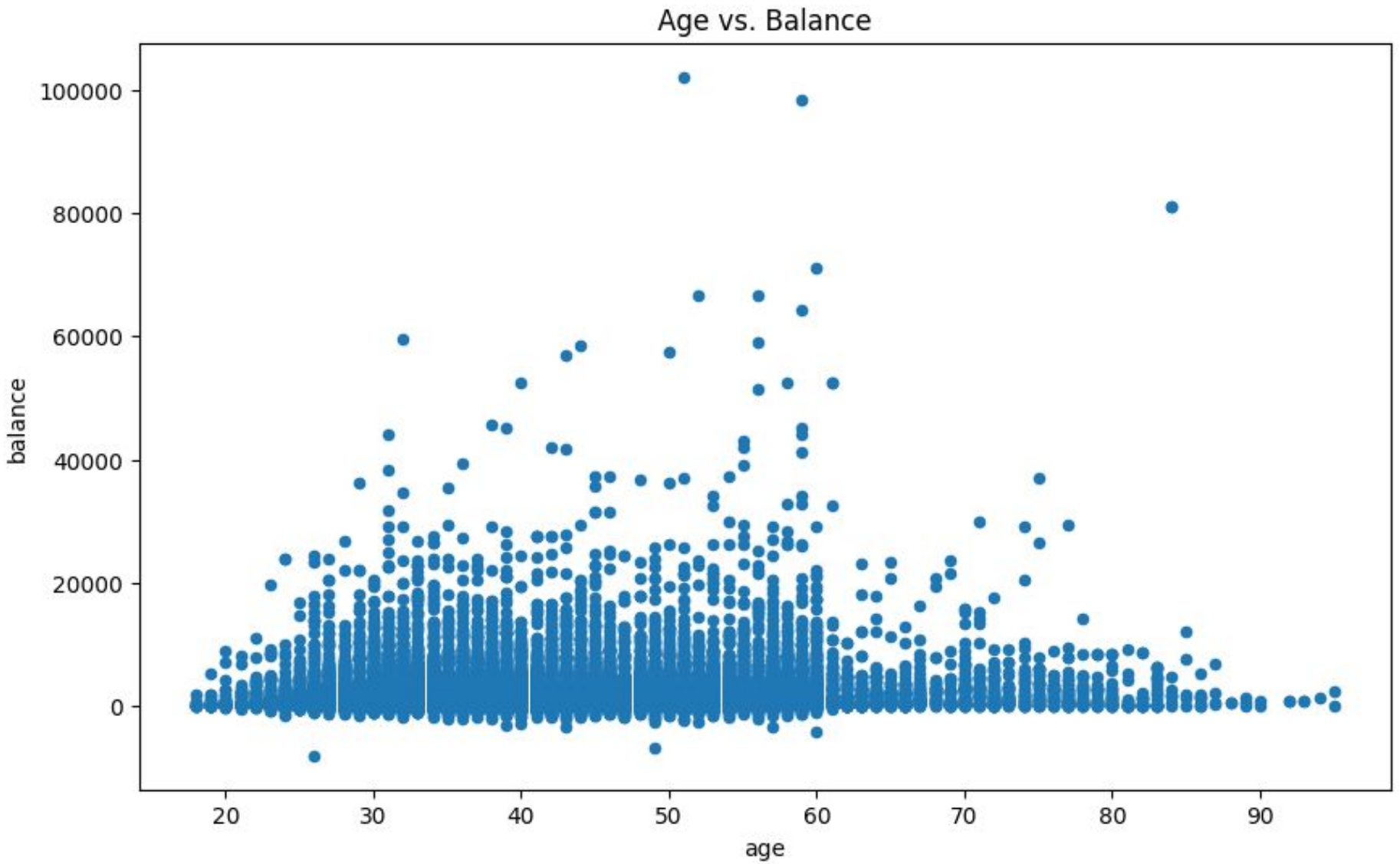
Assuming:

Cellular: 0

Telephone: 1

Unknown: 2

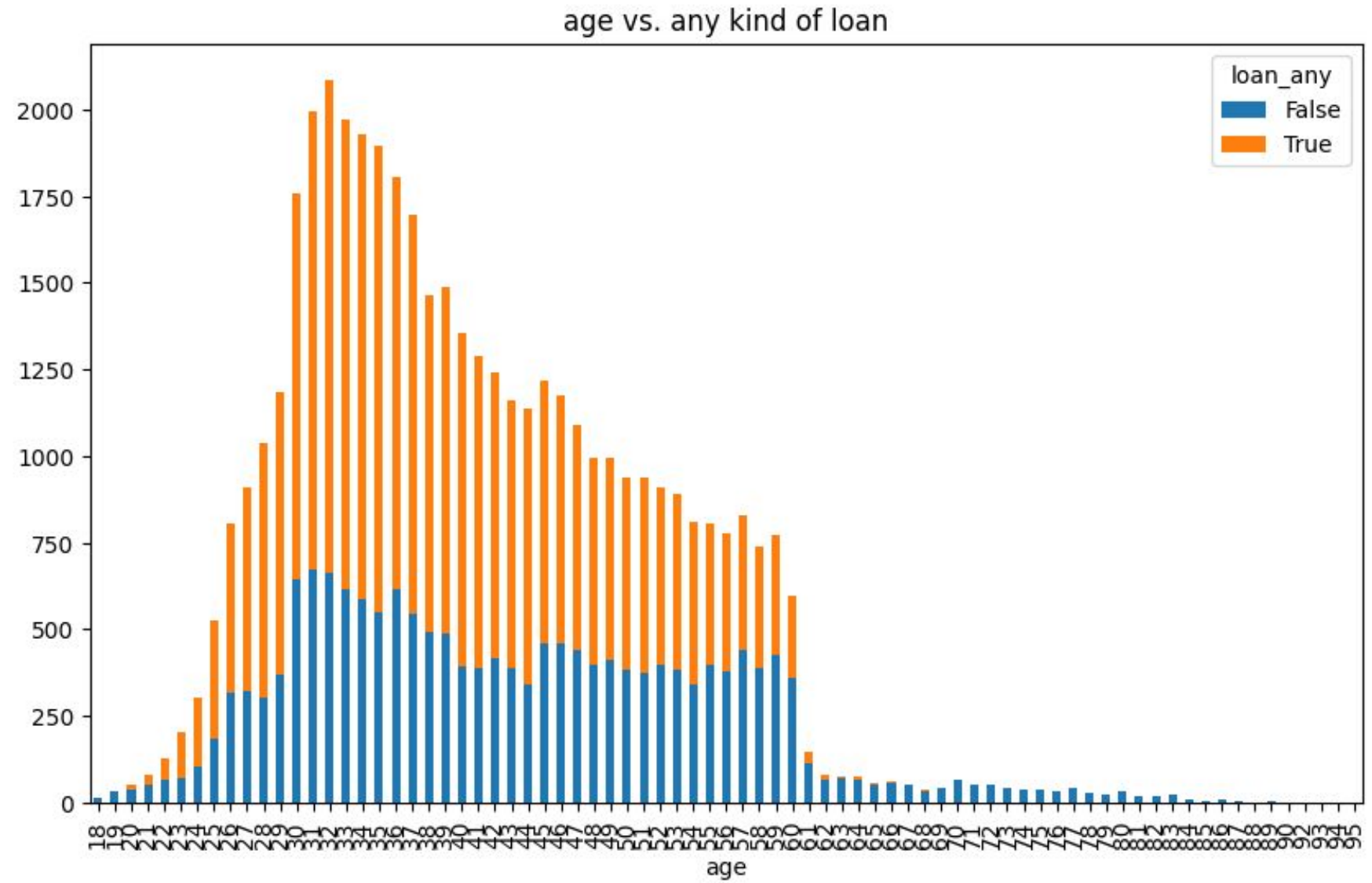
Customer Income based Analysis



The relationship between customer age and their balance is in normal distribution.

Customers at age 40 - 60 have the most balance.

Customer Loan based Analysis



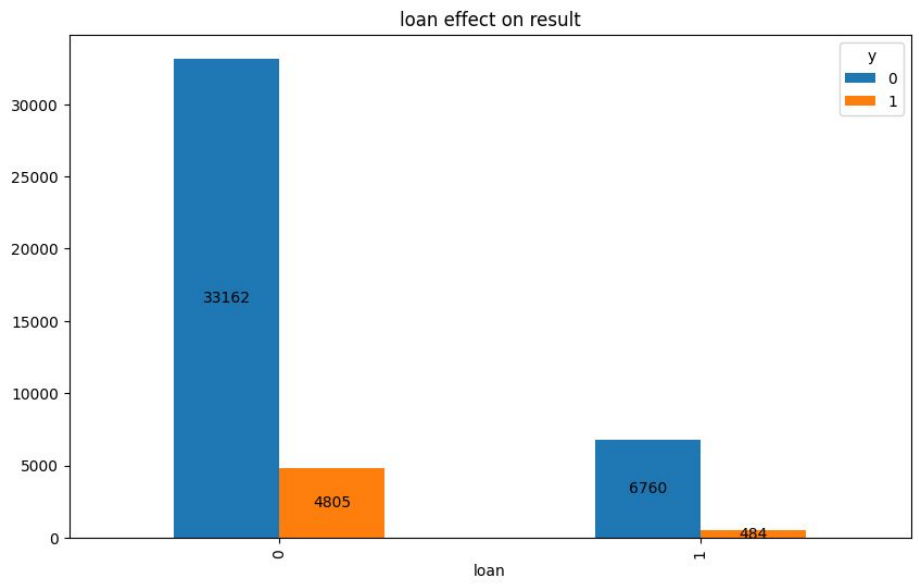
Over 50% of customer at age between 25 to 60 owns at least one kind of loan.

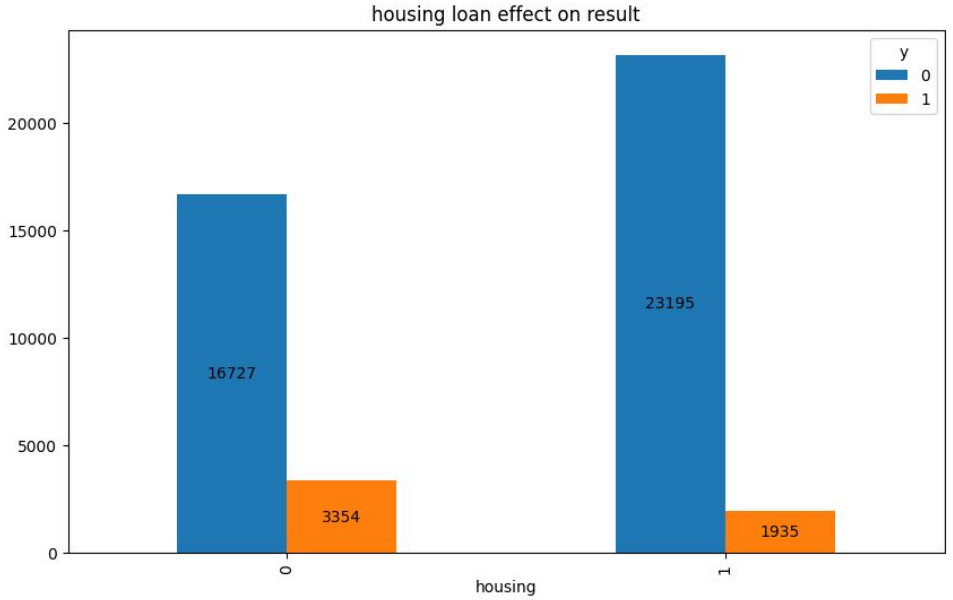
Customers over 65 year old barely own any loan.

Assuming:

If client has either housing loan or personal loan, loan_any will be True.

Customer Loan based Analysis





There are more people owning housing loan than people owning personal loan.

Clients without personal loan are 6% more likely to subscribe than those with personal loan.

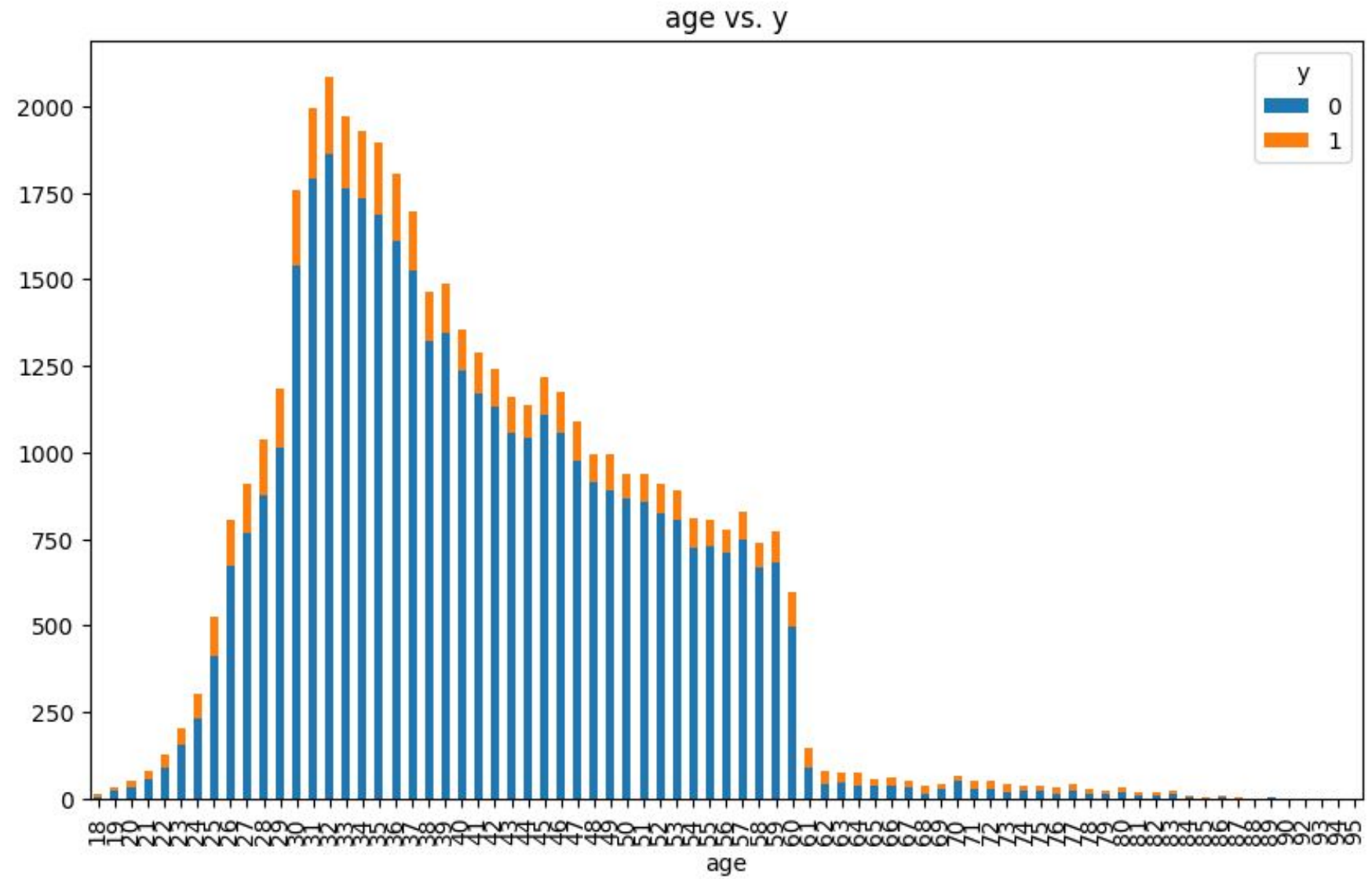
Clients without housing loan are 9% more likely to subscribe than those with housing loan.

Assuming:

Has loan: 1

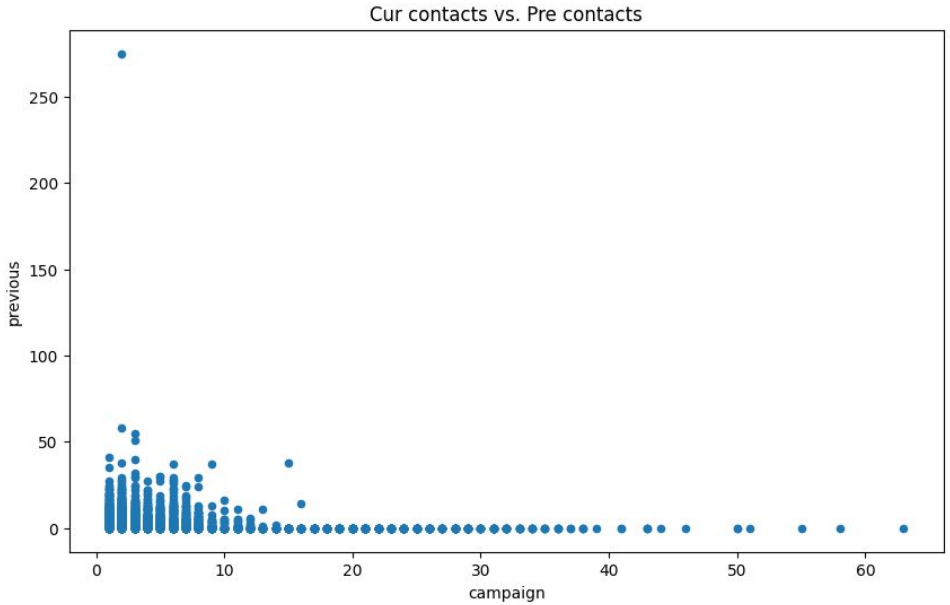
Don’t have loan: 0

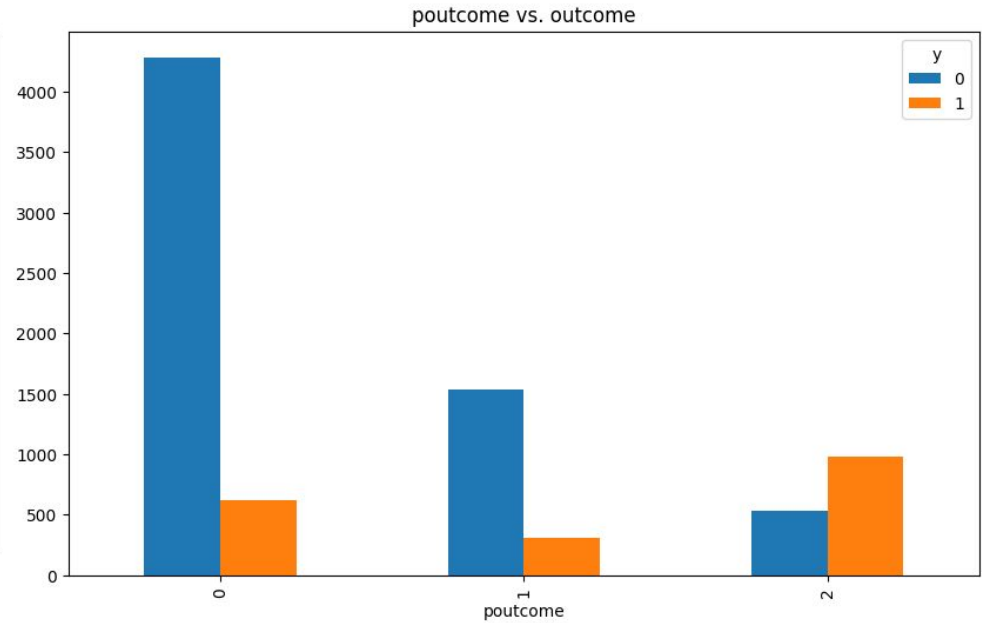
Customer age based Analysis



Customers at age between 30 to 37 have higher chance to subscribe.

Past Campaign based Analysis





Previous contact numbers has a negative relationship with current contact numbers.

Clients with high contact times in previous campaign are less likely to contact in current campaign.

However, clients subscribed in previous campaign are also likely to subscribe during current campaign.

We can observe a group of new clients with high number of contact during current campaign.

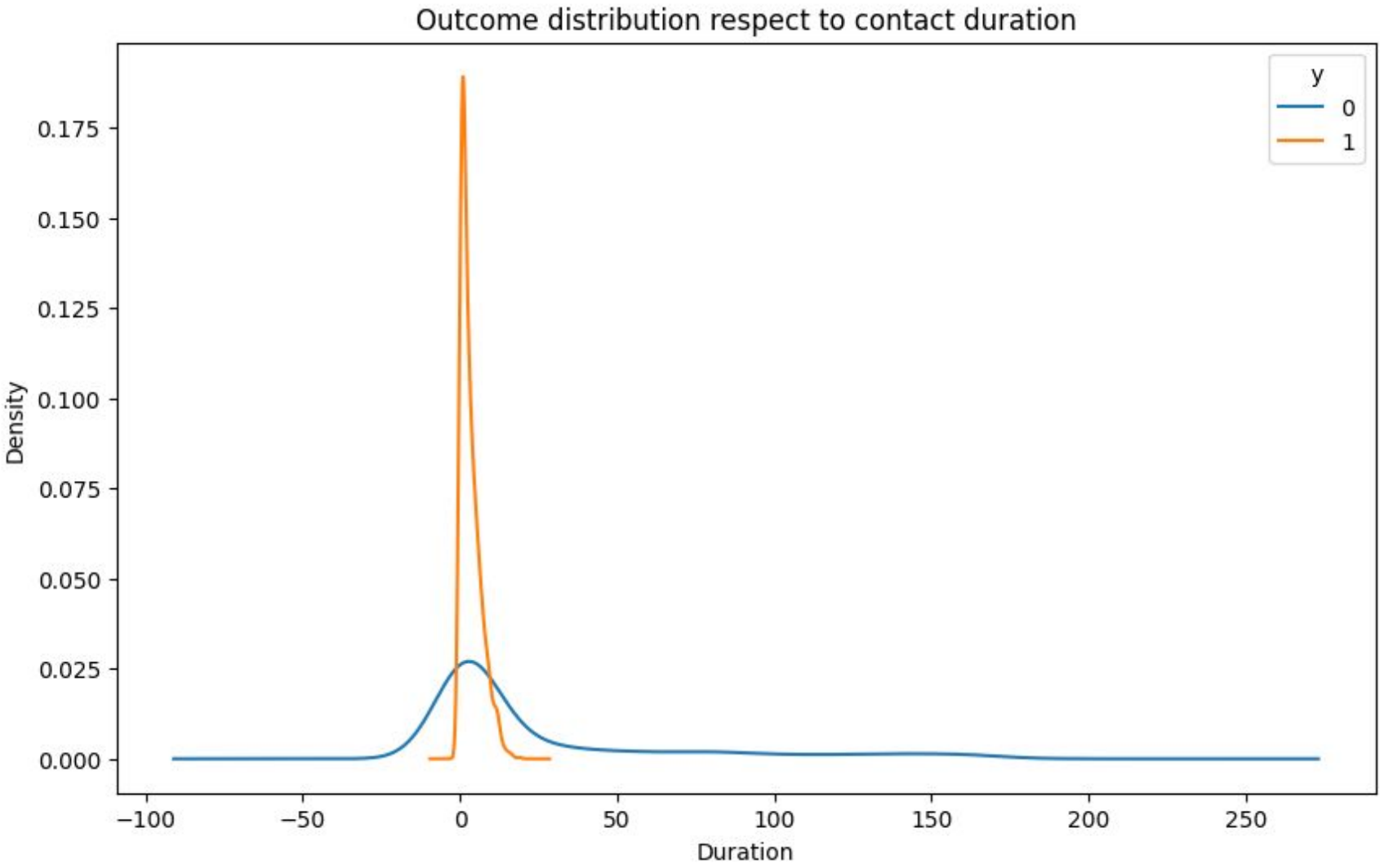
Assuming: poutcome:

success: 2

other: 1

failure: 0

Contact Duration effects on Outcome



Contact duration has a negative effect on the outcome.

Longer contact duration does not lead to higher subscribe rate.

Summary and Recommendations

- Data set: We do observe the presentation of long-tail distribution in several attributes, which may lead to bad model performance. Samples in head classes are over-represented and samples in tail classes are under-represented.
- Clients Age: We have most clients at age 32. They have the highest loan percentage among all ages and they have the best chance to subscribe the product.
- Contact: Number of contacts during previous campaigns has negative correlation with number of contacts during current campaign. We also found that longer contact duration won’t necessary mean higher chance of subscription.
- Previous Outcome: Clients who subscribed during previous campaigns have higher chance to subscribe in this campaign.

Summary and Recommendations

Recommendation:

Using logistic regression or K-NN model with standard scaling.

- K-NN can work well with imbalance data set
- For logistic regression, by adjusting class weighting, we can make it more sensitive to tail classes.
- Feature scaling normalize the range of independent variables. Standard scaling help achieve a normal distribution of data.

Thank You