Bank Marketing Campaign Case Study: Exploratory Data Analysis

Virtual Internship

Company: ABC Bank

Authors: Ammar Sidhu and Islom Pulatov

Date: 12/16/2022

Team Details

Group Name: AI Boys

Group Member ID	Name	Email	Country	College	Specialization
1	Islom Pulatov	islompulatov115@gmail.com	Poland	Epicode Global	Data Science
2	Ammar Sidhu	ammarsidhu@outlook.com	Canada	University of Toronto	Data Science

Github Repo: https://github.com/Islompulatov/Bank_marketing

AGENDA

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Correlation Analysis

Model Recommendations

Problem Description and Business Understanding

Problem Description:

- 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).

Business Understanding:

- 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, to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.
- The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

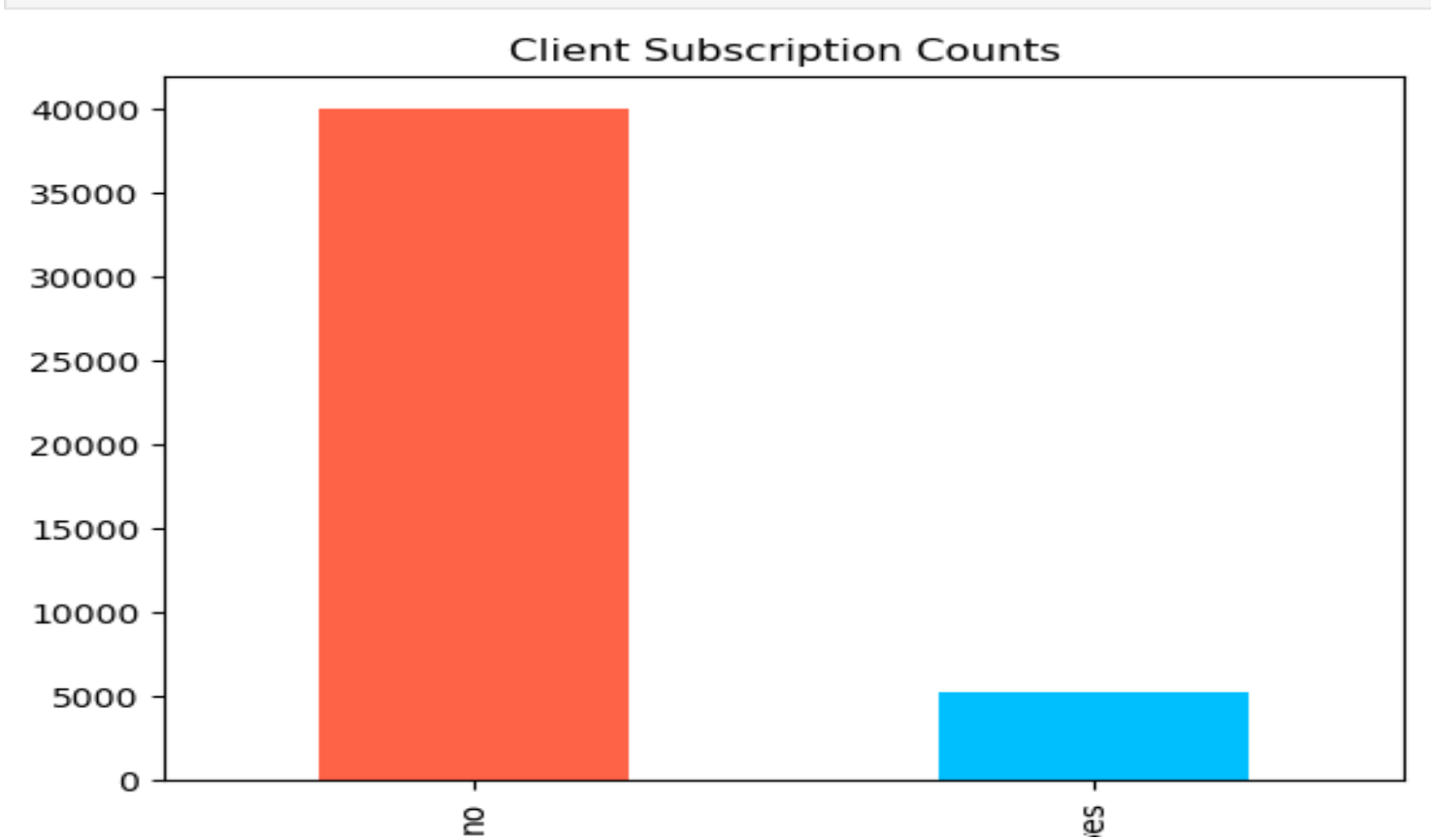
Why Machine Learning Models?:

- The ABC Bank wants to use machine learning models 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 on those customers whose chances of buying the product is more.

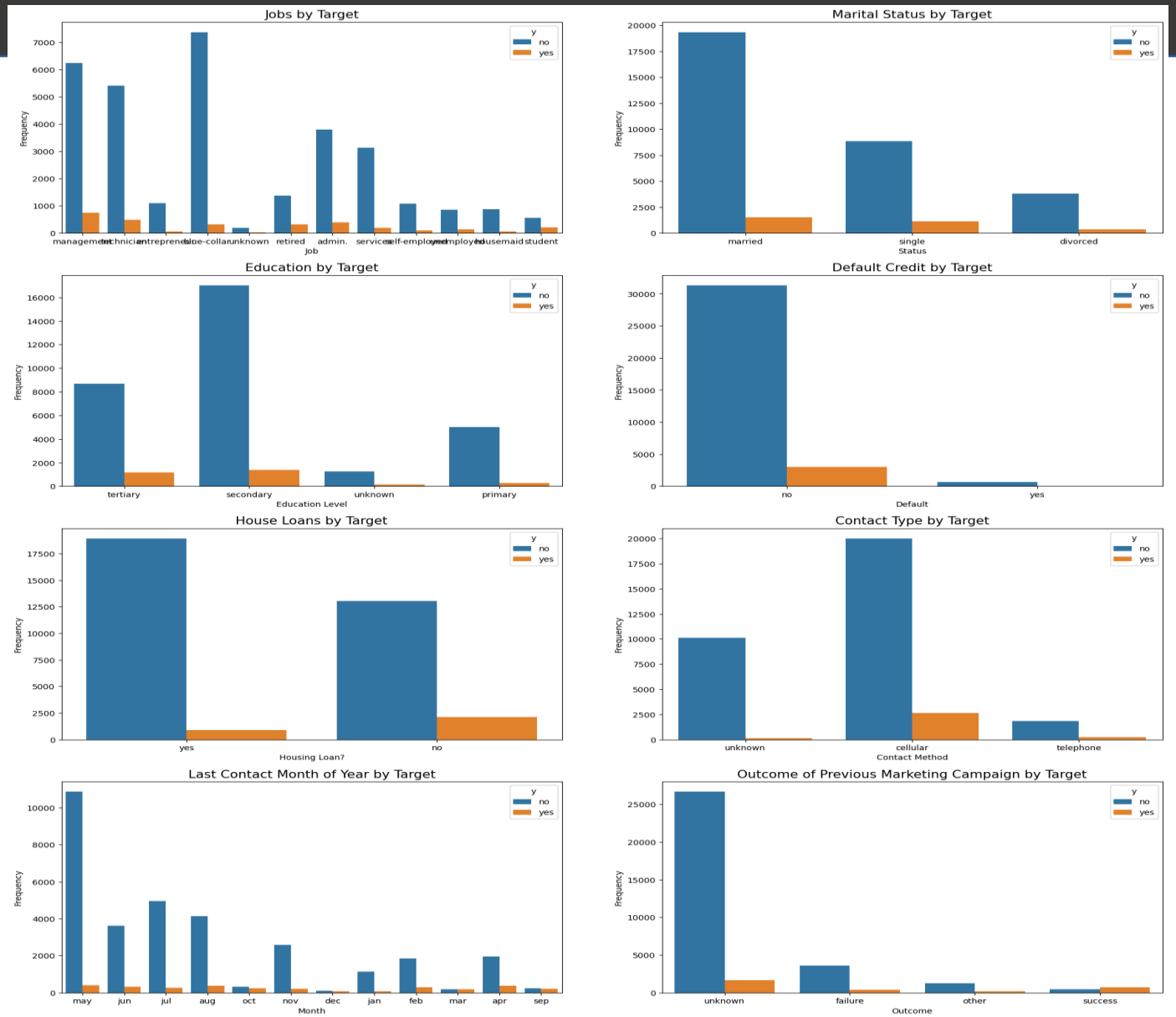
Data Cleaning

- Found **no missing values** in the dataset
- Found **no duplicate rows** in the dataset
- Handled outliers for the **7 Numerical Features** with two different approaches:
  (1) Do not drop outliers (Islom)
  (2) Drop outliers based on feature and context of the outliers using IQR Method (Ammar)
- 'Unknown' class for categorical variables were handled in two different ways:
  (1) As a unique class so not a missing value (Ammar)
  (2) Treated as a missing value and drop the corresponding row from data frame (Islom)

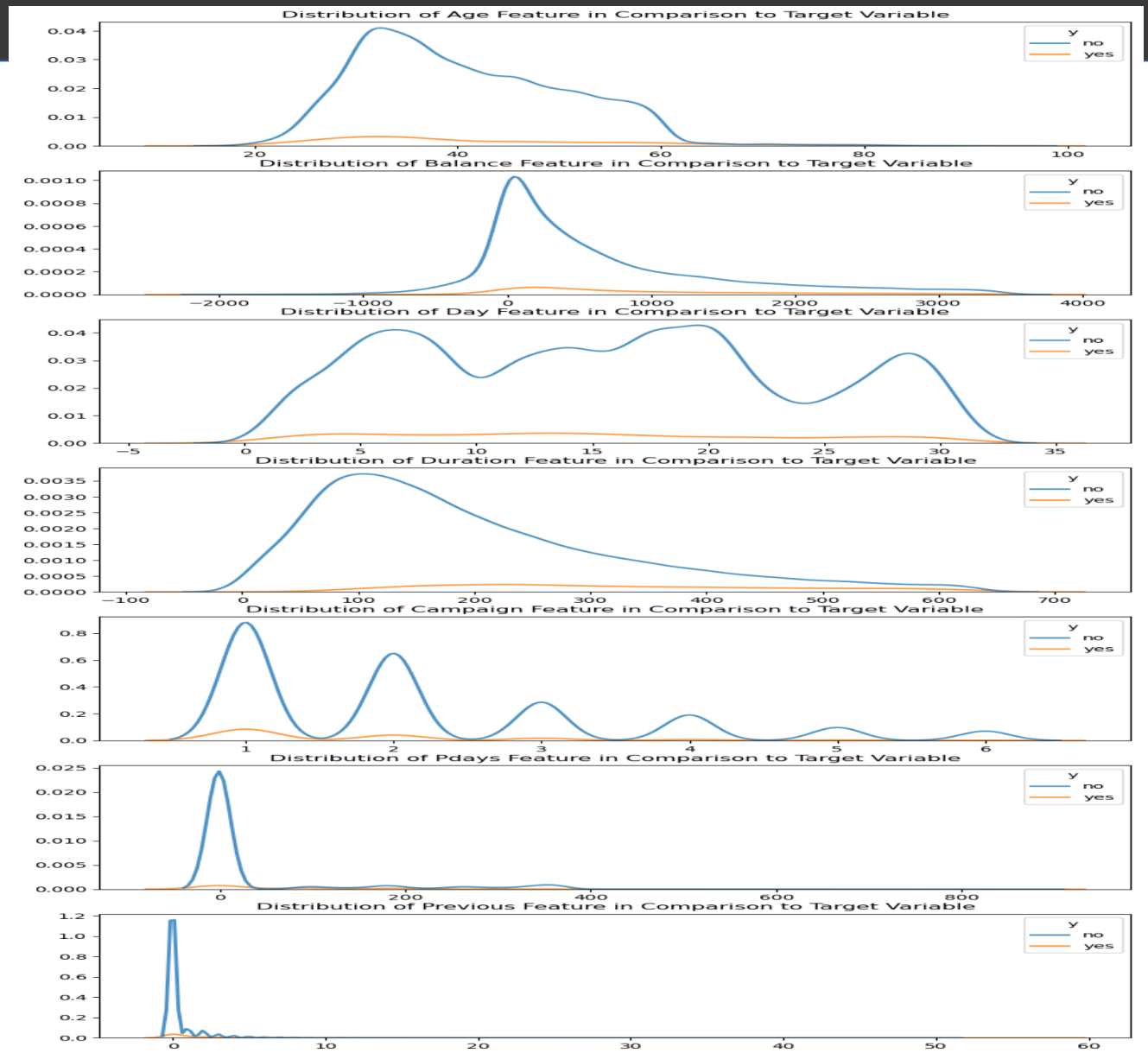
Client Subscriptions Counts – Target (y)

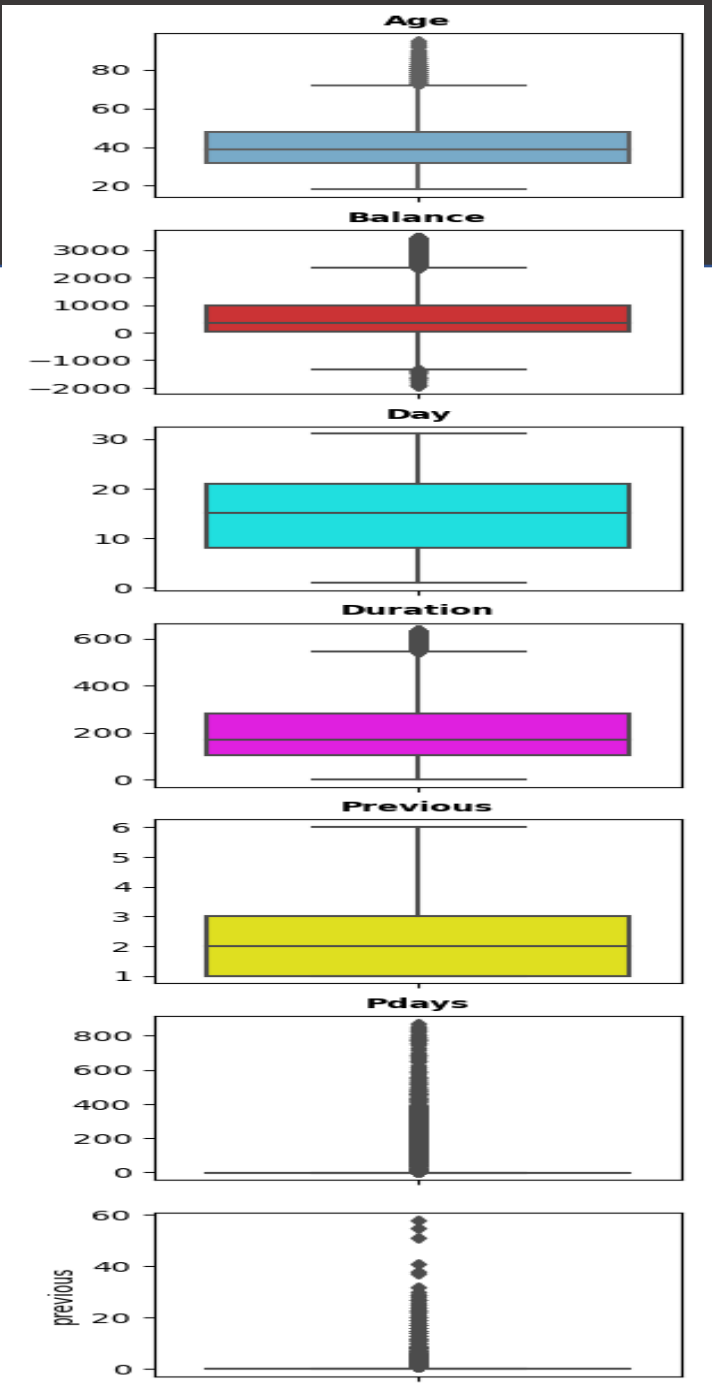


Visualizing Categorical Variables

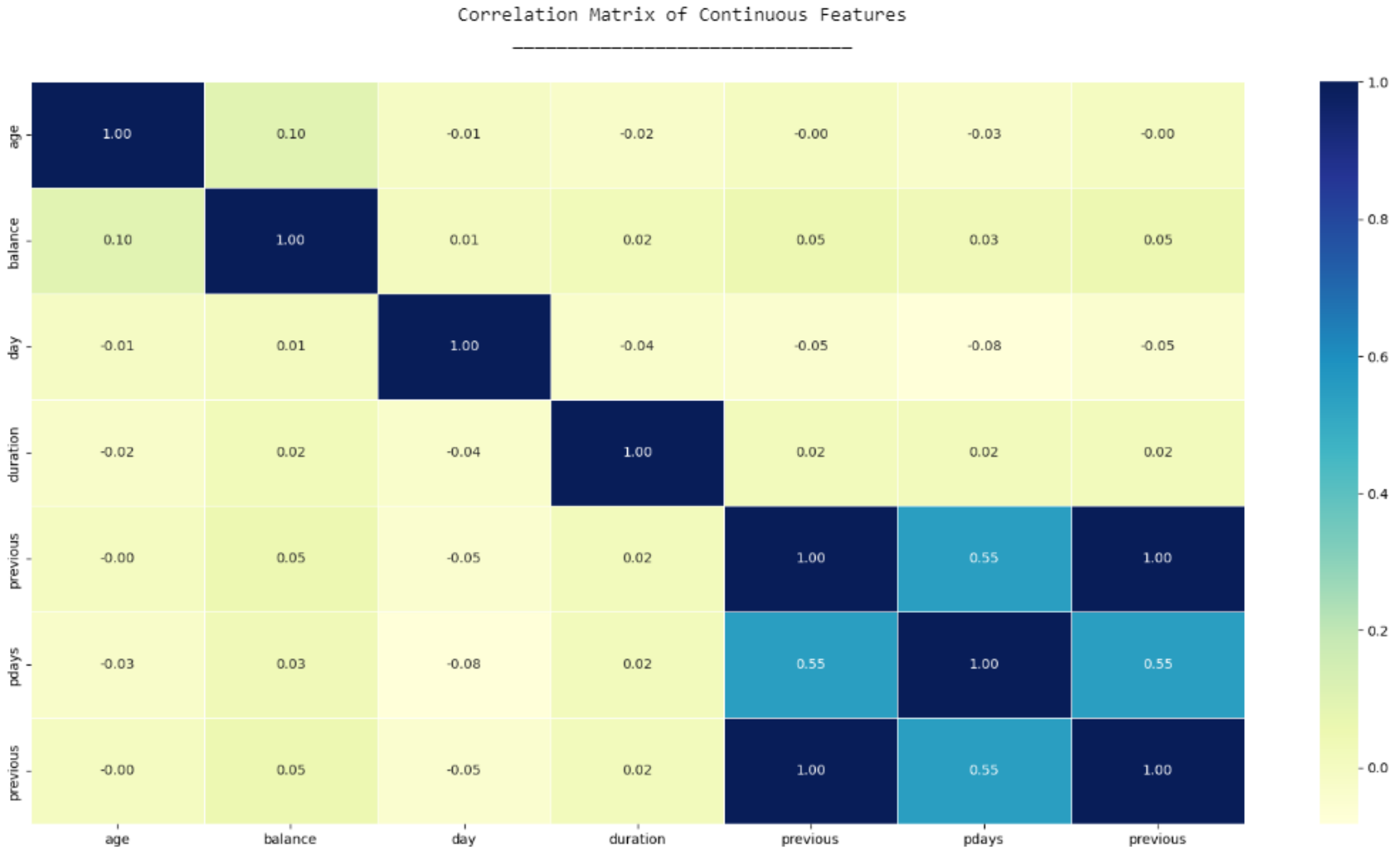


Visualizing Continuous Features





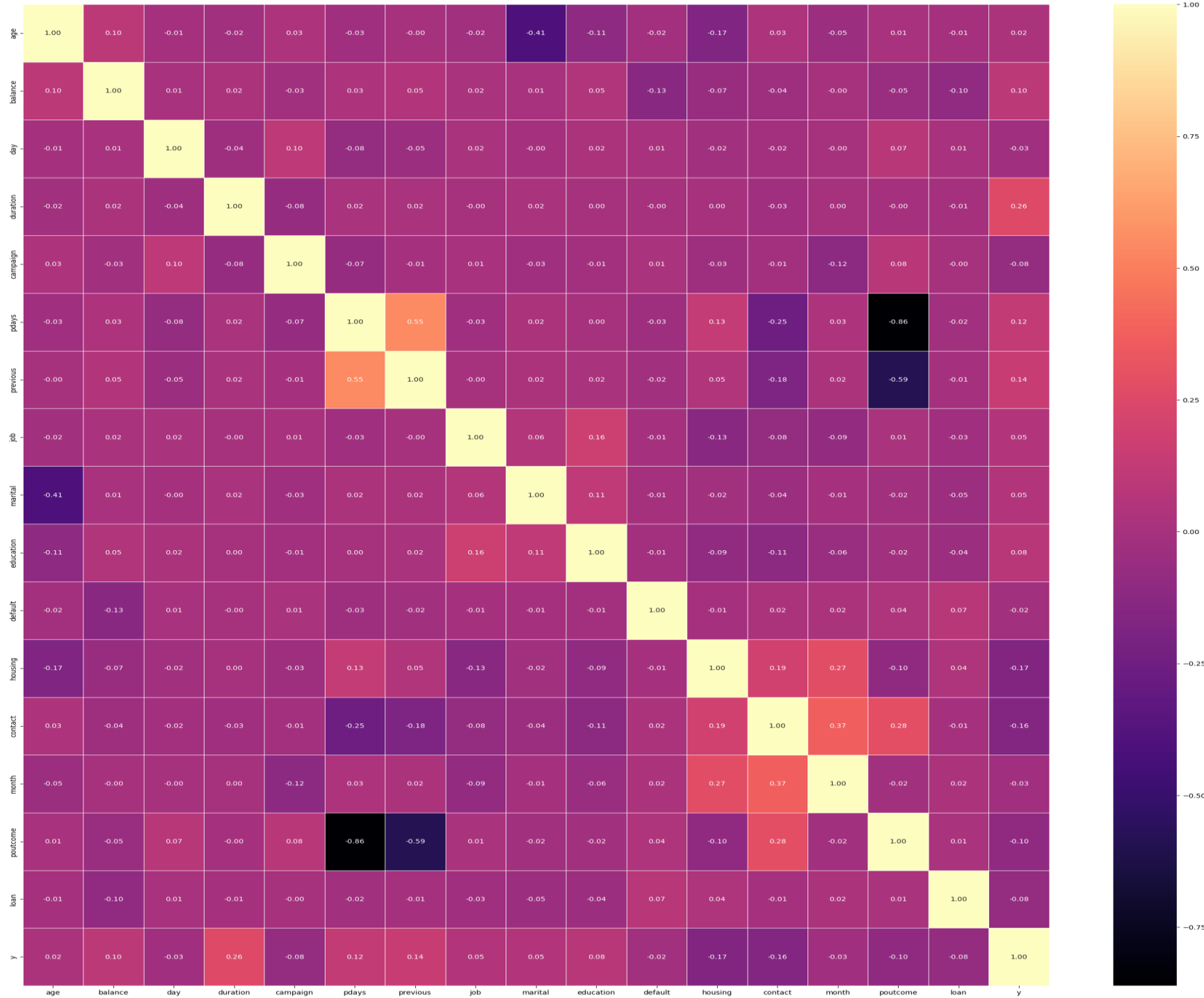
Correlation Analysis



Inference from Correlation Analysis:

- There is no multicollinearity between the numerical features.
- The only feature with a moderate correlation with the target - y is the duration feature.
- There is a strong correlation between the encoded poutcome feature, and the pdays feature.
- Some features are negatively correlated with each other.

Heatmap of all Features and Including One-Hot Encoded Categorical Features

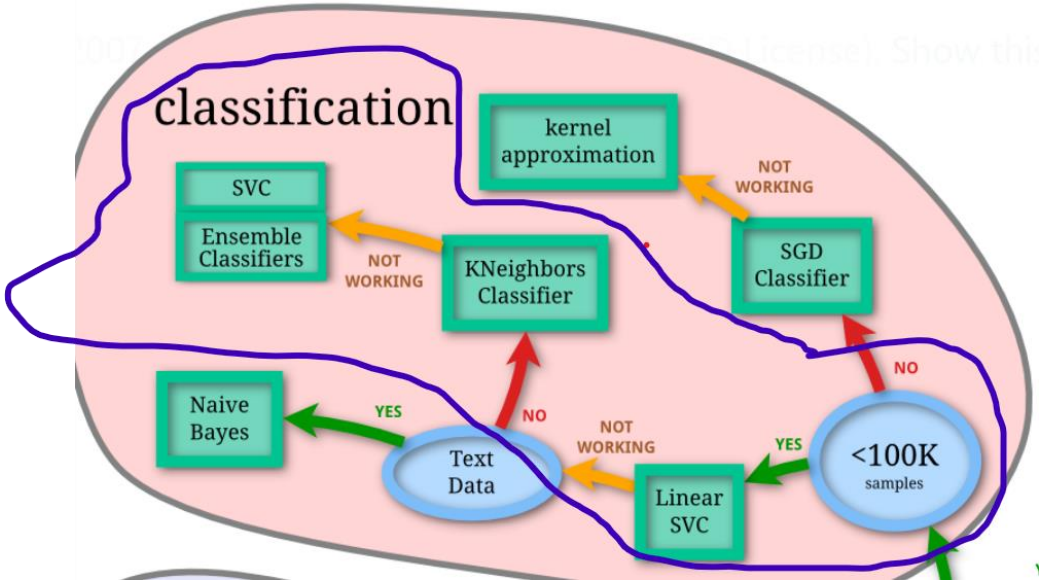


Conclusions from EDA

- There are **no NaN values** and **no duplicate values** in the dataset.
- The **target feature is imbalance**d as there are **more than 8x** the customers who subscribed vs. who dd not.
- Except the duration feature, all the other features have a **low correlation** with the target.
- **Most customers** are married, have loans, and work collar jobs.

Model Recommendations

- Test and Train Ensemble and Boosting Classification Models with Cost-Sensitive Learning (Ammar).
- Test and Train Ensemble and Boosting Classification Models with SMOTE (Islom).
- Tune Hyperparameters of the Best Performing Models.
- Compare Model Performances and check to see if dropping ‘unknown’ entries had an impact on accuracies.



Source: https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

Thank You