Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis on Bank Marketing Campaign

Shuran Fu Aug 11, 2022

Agenda

Executive Summary

Problem Statement

Data Cleaning

EDA

Recommendations

Executive Summary

- Business background
- Dataset details
  - 45211 rows
  - 17 features
    - customer own information: age, job, marital, education …
    - promotion contact information: contact, day, month, duration …

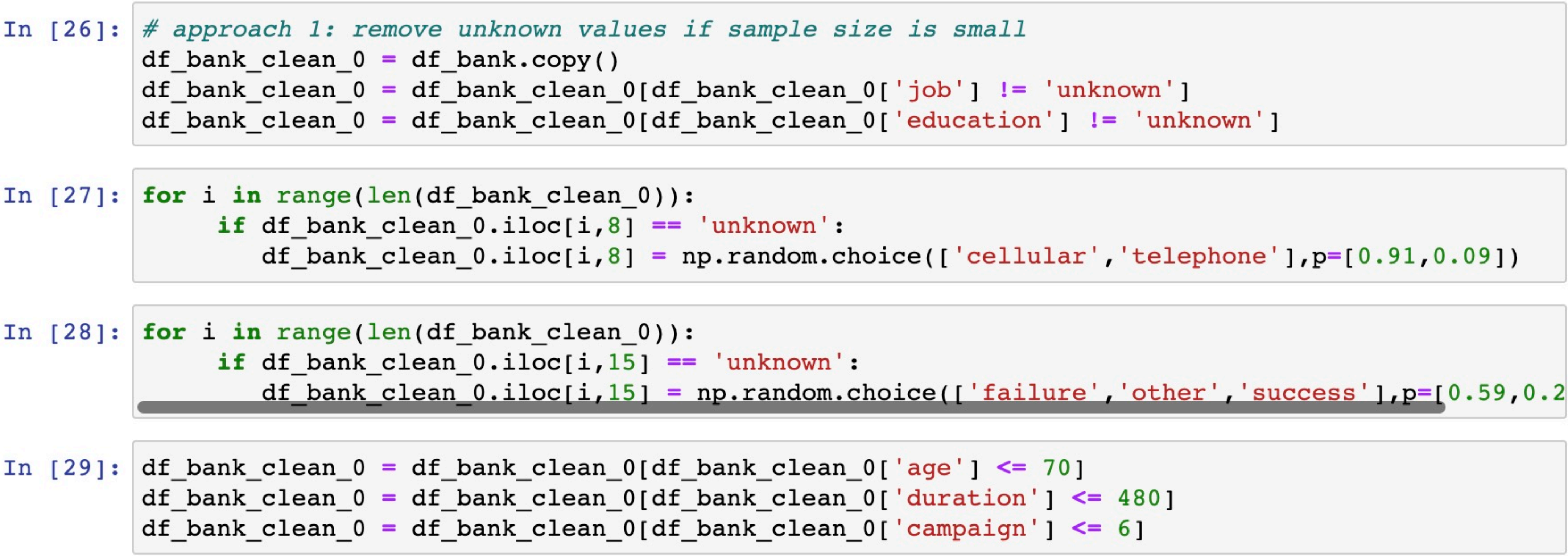
Problem Statement

- Develop a machine learning model to estimate whether a particular customer will buy a specific term deposit product or not based on the customer’s past interaction with bank or other Financial institution.
- y in dataset means whether the customer buy the product or not, so this is a supervised learning problem and we need to use classification model.
- data cleaning -> EDA -> Feature selection -> Model construction -> Performance analysis

Data Cleaning

- missing data
    - drop data randomly assign value
    - fill with mode value
- outliers

Data Cleaning



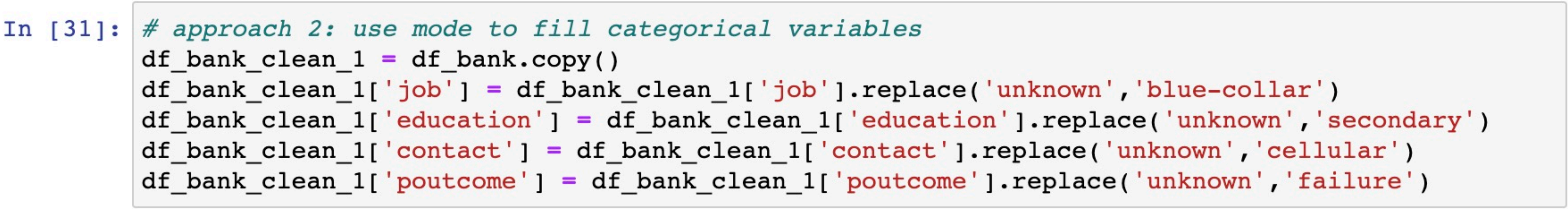
```
In [26]: # approach 1: remove unknown values if sample size is small
         df_bank_clean_0 = df_bank.copy()
         df_bank_clean_0 = df_bank_clean_0[df_bank_clean_0['job'] != 'unknown']
         df_bank_clean_0 = df_bank_clean_0[df_bank_clean_0['education'] != 'unknown']

In [27]: for i in range(len(df_bank_clean_0)):
                 if df_bank_clean_0.iloc[i,8] == 'unknown':
                     df_bank_clean_0.iloc[i,8] = np.random.choice(['cellular','telephone'],p=[0.91,0.09])

In [28]: for i in range(len(df_bank_clean_0)):
                 if df_bank_clean_0.iloc[i,15] == 'unknown':
                     df_bank_clean_0.iloc[i,15] = np.random.choice(['failure','other','success'],p=[0.59,0.2

In [29]: df_bank_clean_0 = df_bank_clean_0[df_bank_clean_0['age'] <= 70]
         df_bank_clean_0 = df_bank_clean_0[df_bank_clean_0['duration'] <= 480]
         df_bank_clean_0 = df_bank_clean_0[df_bank_clean_0['campaign'] <= 6]
```

Data Cleaning

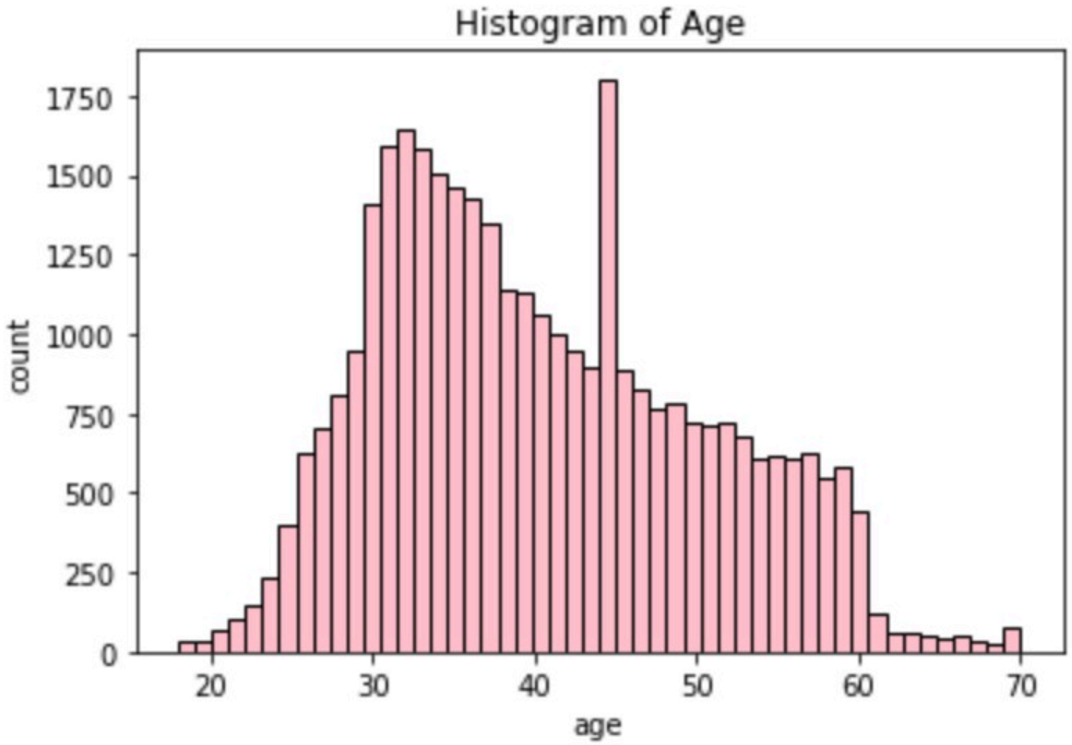


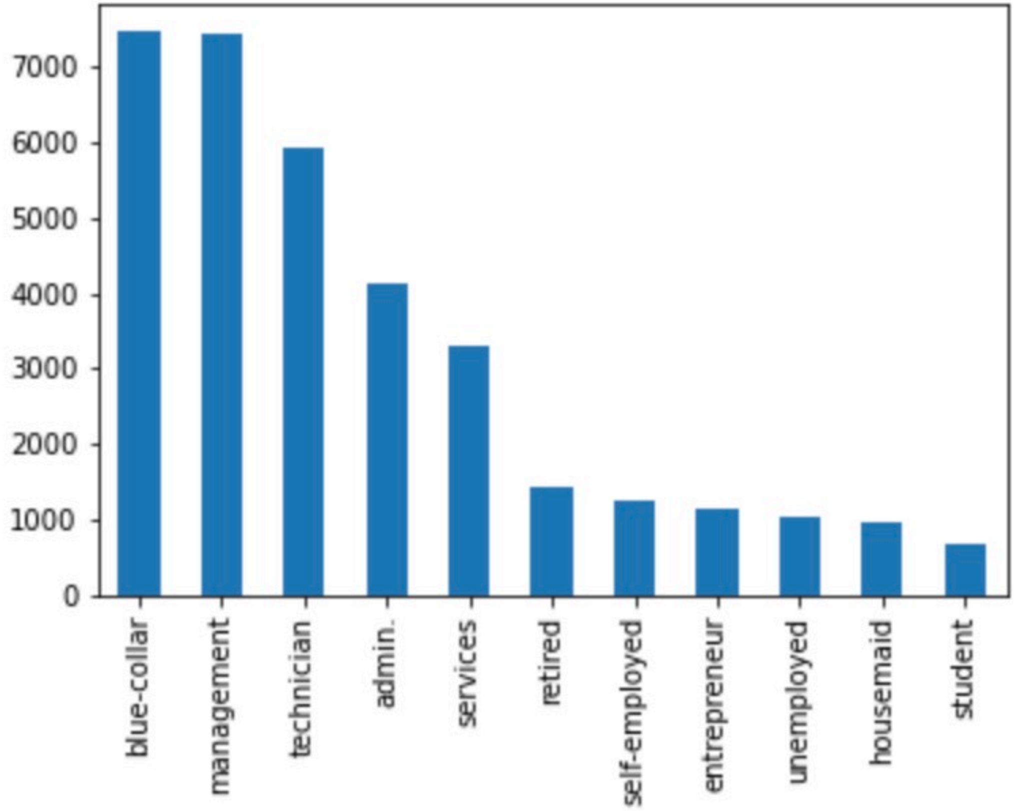
```
In [31]: # approach 2: use mode to fill categorical variables
         df_bank_clean_1 = df_bank.copy()
         df_bank_clean_1['job'] = df_bank_clean_1['job'].replace('unknown','blue-collar')
         df_bank_clean_1['education'] = df_bank_clean_1['education'].replace('unknown','secondary')
         df_bank_clean_1['contact'] = df_bank_clean_1['contact'].replace('unknown','cellular')
         df_bank_clean_1['poutcome'] = df_bank_clean_1['poutcome'].replace('unknown','failure')
```

- after data cleaning, we have 34727 and 36298 observations correspondingly.

EDA

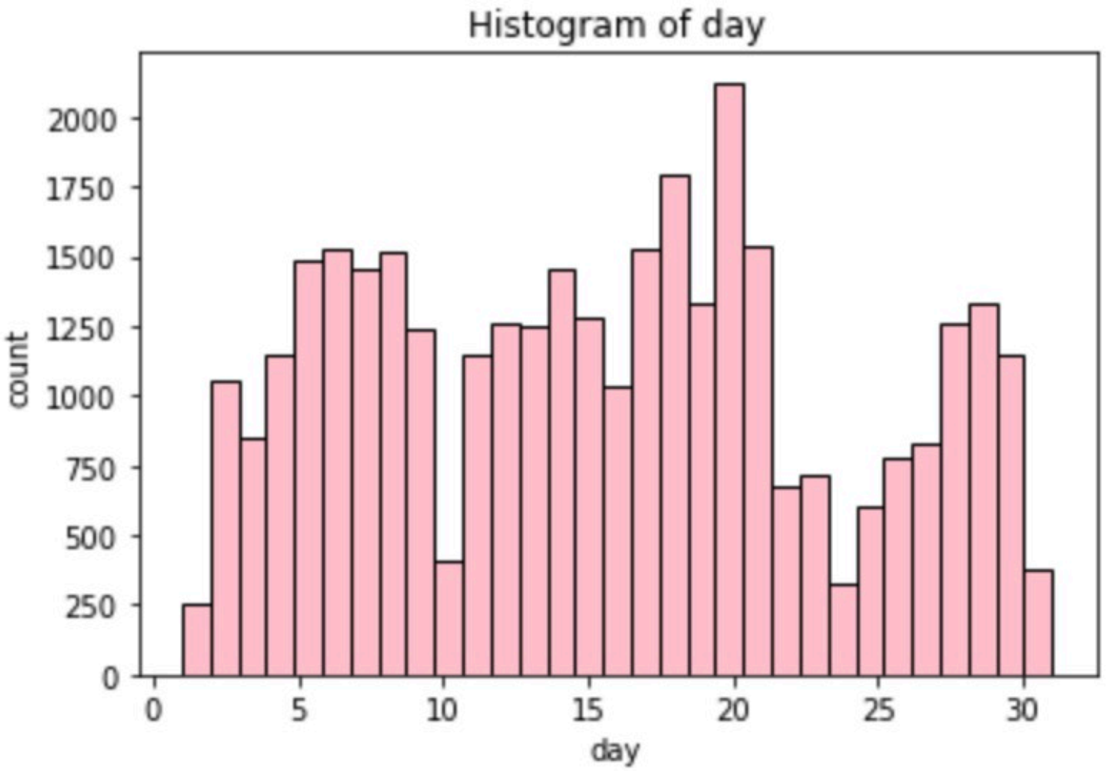
- Age distribution and job distribution

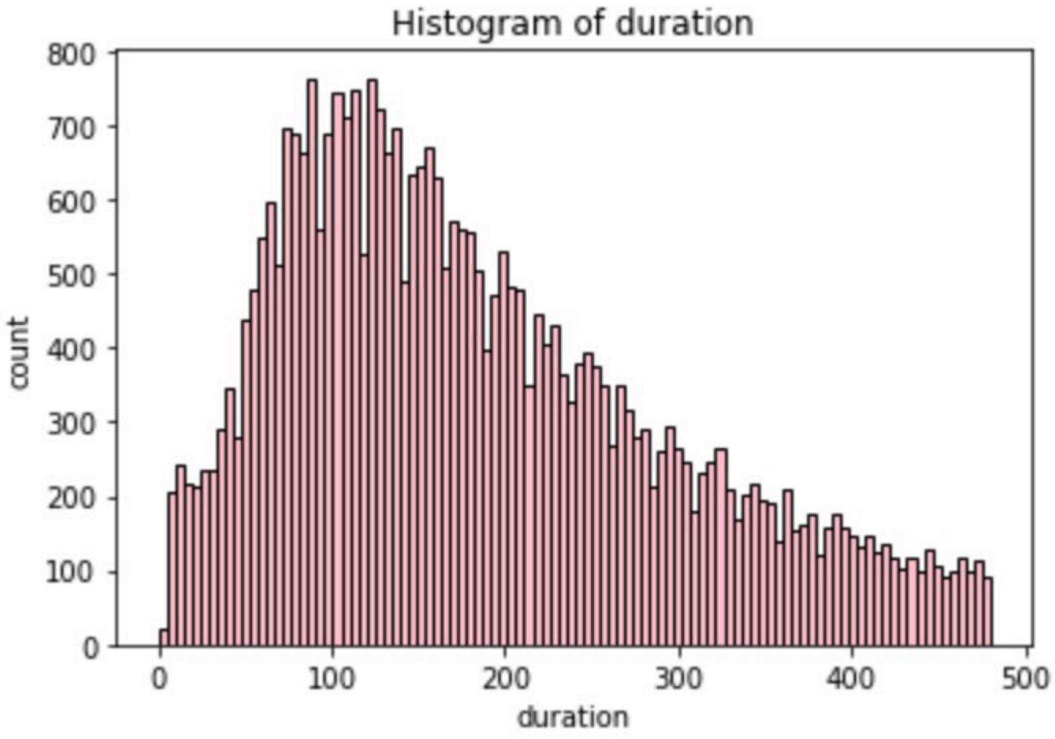




EDA

- day distribution and duration distribution

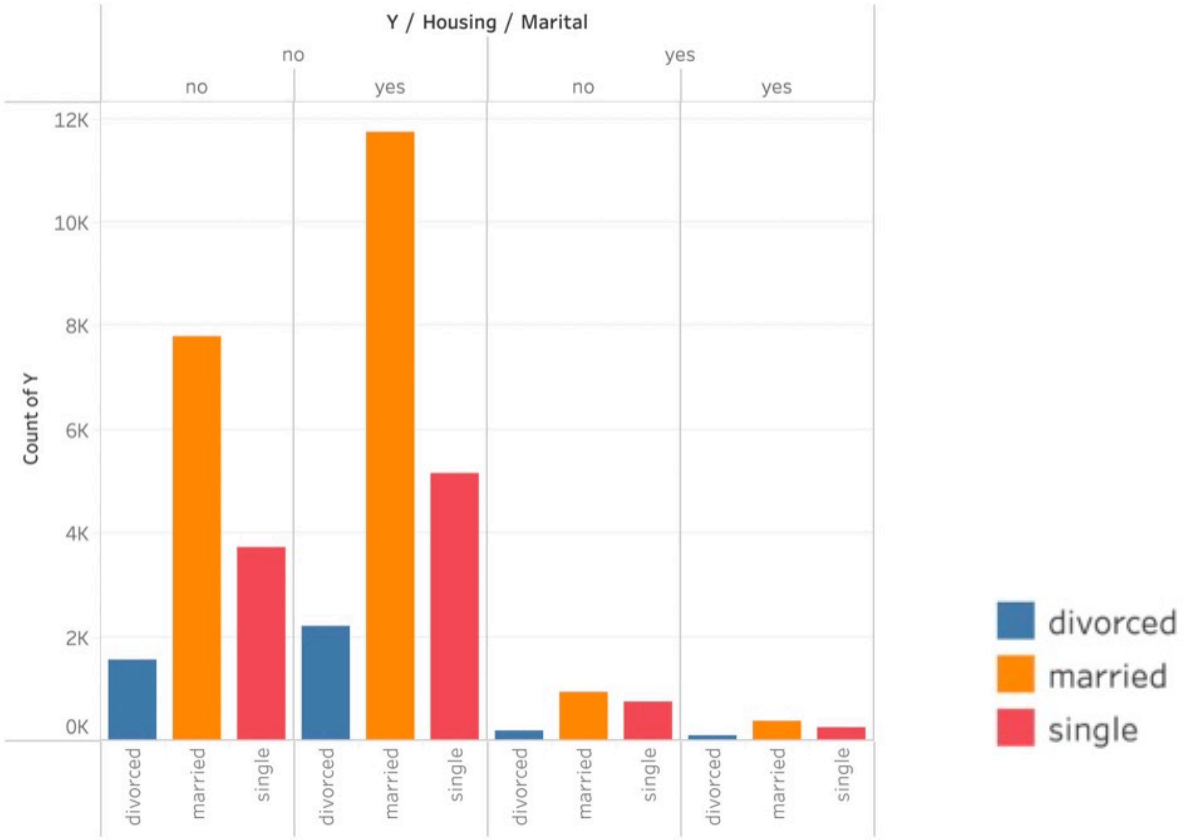




-

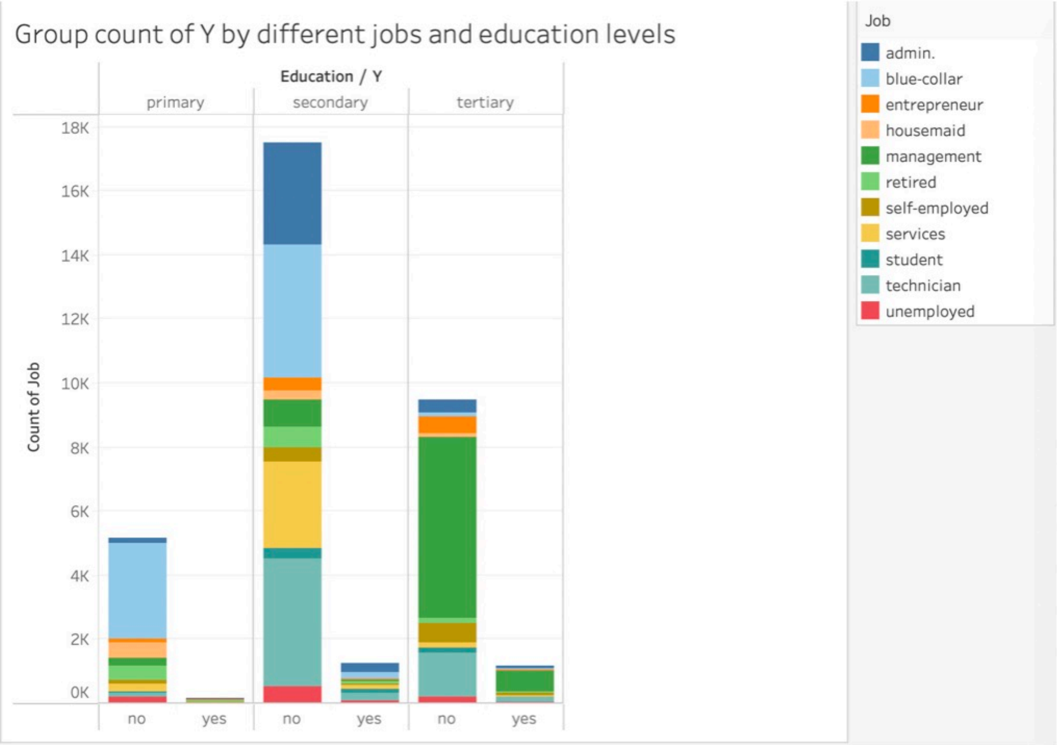
EDA and Recommendations

- Promotion should be focused on people who have married with no housing.



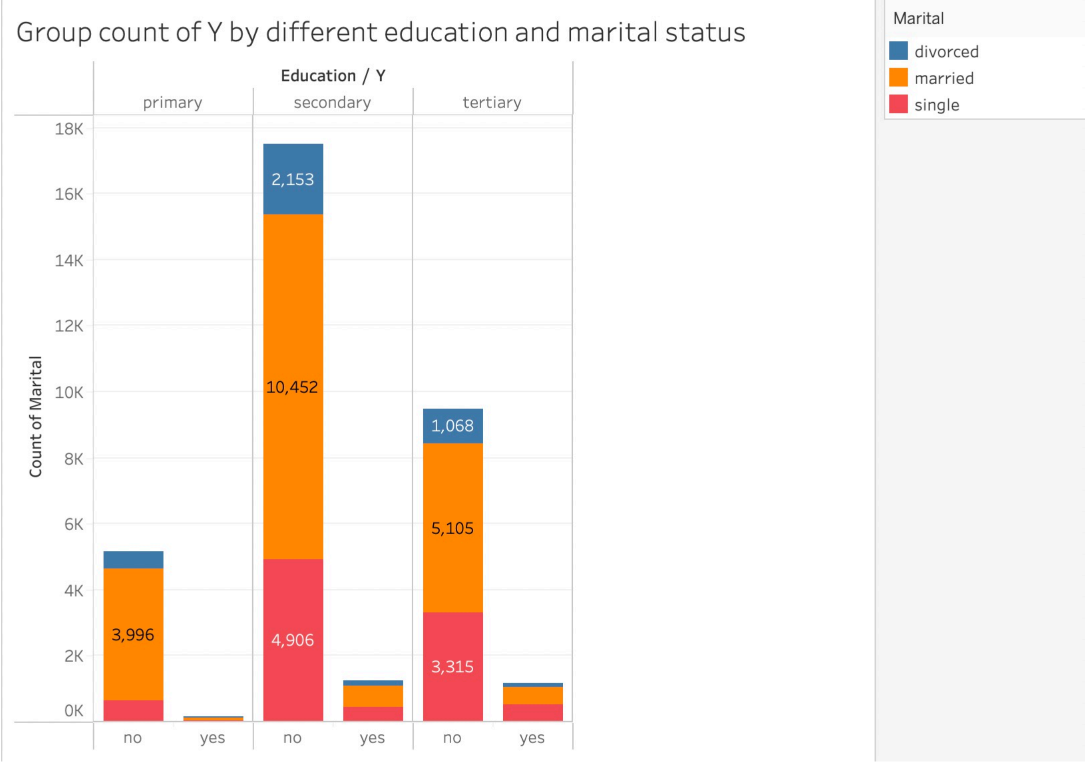
EDA and Recommendations

- Promotion should be focused on people who have management jobs with tertiary education level.



EDA and Recommendations

- Promotion should be focused on people who are single or married with higher education level.



Recommendations for Models

- Classification Models
    - Logistic Regression
    - Trees
    - Boosting Methods
    - Neural network

Data Glacier

Your Deep Learning Partner

Thanks !