Data Glacier
Your Deep Learning Partner

Exploratory Data Analysis

Bank Marketing Campaign

Maria Contractor
16 August 2023

Agenda

Problem Description

Business Understanding

Data Understanding

Cleansing Techniques

EDA

Recommendations

Proposed Model

Data Glacier

Your Deep Learning Partner

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). To solve this problem, we will need to predict whether or not the client will subscribe to a term deposit.

Data Glacier

Your Deep Learning Partner

Business Understanding

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. This will increase productivity and efficiency in selling their products.

Data Glacier
Your Deep Learning Partner

Data Understanding

The ABC Bank has data that includes bank client data and other attributes. The data includes 21 attributes/columns and has 41188 entries. Some of the attributes include age, job, marital status, housing, and other demographic information. It also includes economic data like price indexes and outcomes of previous campaigns. This data is either categorical or numeric. There is an output variable, which is a binary data value (Y/N).

Tabular data details: bank-additional-full.csv

Total number of observations	41188
Total number of files	1
Total number of features	21
Base format of the file	.csv
Size of the data	5.56 MB

Data Glacier
Your Deep Learning Partner

Cleansing Techniques

Imputing Categorical Data

- Focused on age, job, and education
- Each of these factors allowed me to decrease the amount of unknown variables by predicting where some of them may be placed based on the existing data.

Replacing ‘unknown’ with NaN variables

- Easier to read and adjust if necessary.

Data Glacier
Your Deep Learning Partner

Cleansing Techniques

Before:

education job	basic.4y	basic.6y	basic.9y	high.school	illiterate	professional.course	university.degree	unknown
admin.	77	151	499	3329	1	363	5753	249
blue-collar	2318	1426	3623	878	8	453	94	454
entrepreneur	137	71	210	234	2	135	610	57
housemaid	474	77	94	174	1	59	139	42
management	100	85	166	298	0	89	2063	123
retired	597	75	145	276	3	241	285	98
self-employed	93	25	220	118	3	168	765	29
services	132	226	388	2682	0	218	173	150
student	26	13	99	357	0	43	170	167
technician	58	87	384	873	0	3320	1809	212
unemployed	112	34	186	259	0	142	262	19
unknown	52	22	31	37	0	12	45	131

Cleansing Techniques

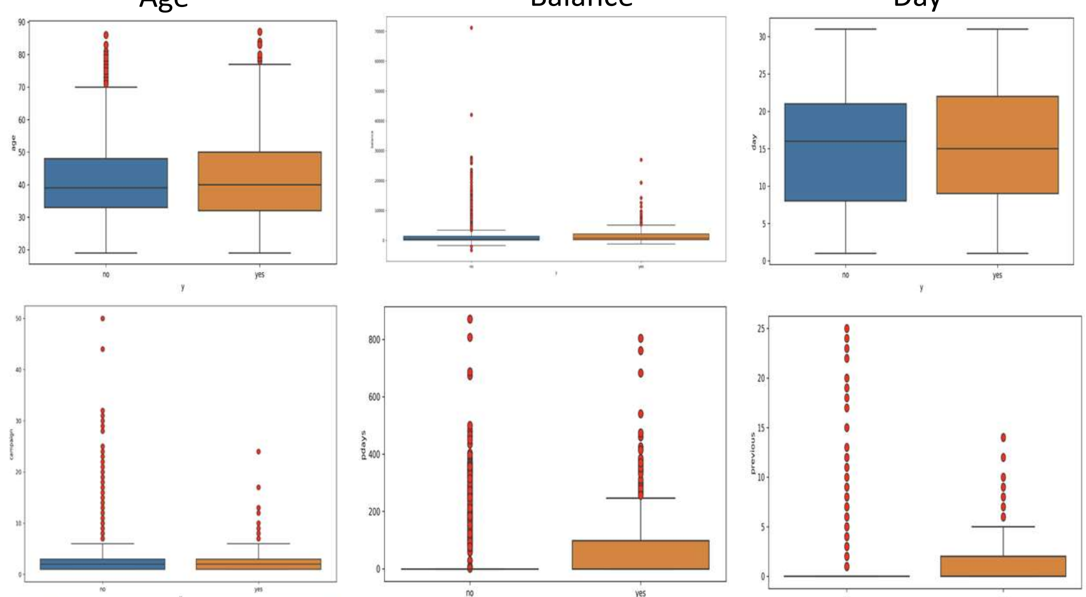
After:

education job	basic.4y	basic.6y	basic.9y	high.school	illiterate	professional.course	university.degree
admin.	77	151	499	3329	1	363	5750
blue-collar	2369	1447	3654	878	8	453	94
entrepreneur	137	71	210	234	2	135	610
housemaid	516	77	94	174	1	59	139
management	100	85	166	298	0	89	2186
retired	598	75	145	276	3	241	285
self-employed	93	25	220	118	3	168	765
services	132	226	388	2830	0	218	173
student	26	13	99	357	0	43	170
technician	58	87	384	872	0	3317	1809
unemployed	112	34	186	259	0	142	262

Data Glacier
Your Deep Learning Partner

EDA: Outlier Detection

Age Balance Day



Campaign Pday
s Previous

Each of the red dots represents an outlier in the numerical data. From the data, we can see that there are many outliers in most of the numerical categories in bank_add_full_data.

EDA: Outlier Detection

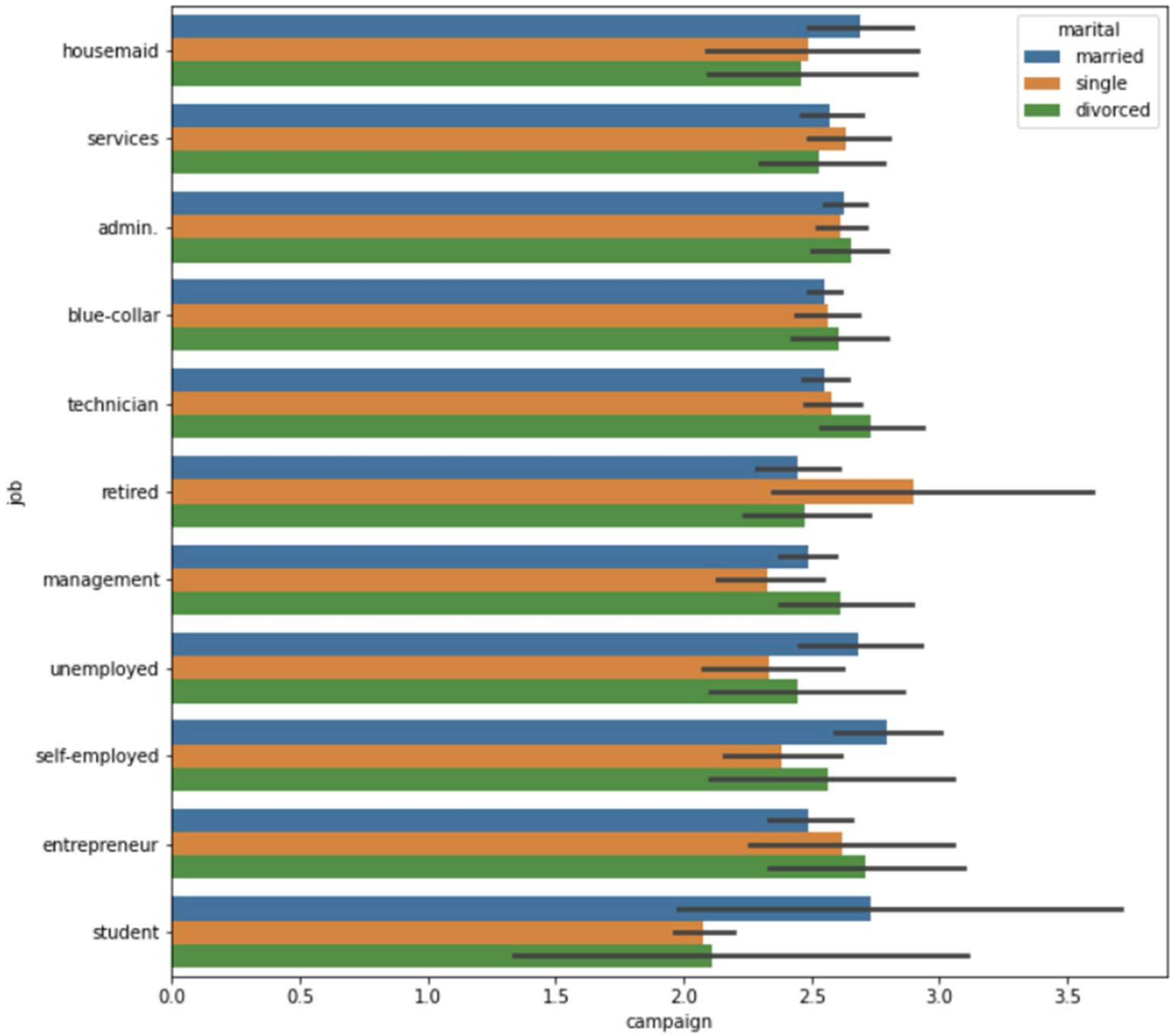
- The summary statistics of the numerical data show some of the maximums, which can also be used to detect outliers. For example, the max of campaign is 56, whereas the mean is about 2.6

	age	campaign	pdays	previous	emp.var.rate
count	41188.00000	41188.000000	41188.000000	41188.000000	41188.000000
mean	40.02406	2.567593	962.475454	0.172963	0.081886
std	10.42125	2.770014	186.910907	0.494901	1.570960
min	17.00000	1.000000	0.000000	0.000000	-3.400000
25%	32.00000	1.000000	999.000000	0.000000	-1.800000
50%	38.00000	2.000000	999.000000	0.000000	1.100000
75%	47.00000	3.000000	999.000000	0.000000	1.400000
max	98.00000	56.000000	999.000000	7.000000	1.400000

	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
count	41188.000000	41188.000000	41188.000000	41188.000000
mean	93.575664	-40.502600	3.621291	5167.035911
std	0.578840	4.628198	1.734447	72.251528
min	92.201000	-50.800000	0.634000	4963.600000
25%	93.075000	-42.700000	1.344000	5099.100000
50%	93.749000	-41.800000	4.857000	5191.000000
75%	93.994000	-36.400000	4.961000	5228.100000
max	94.767000	-26.900000	5.045000	5228.100000

Data Glacier

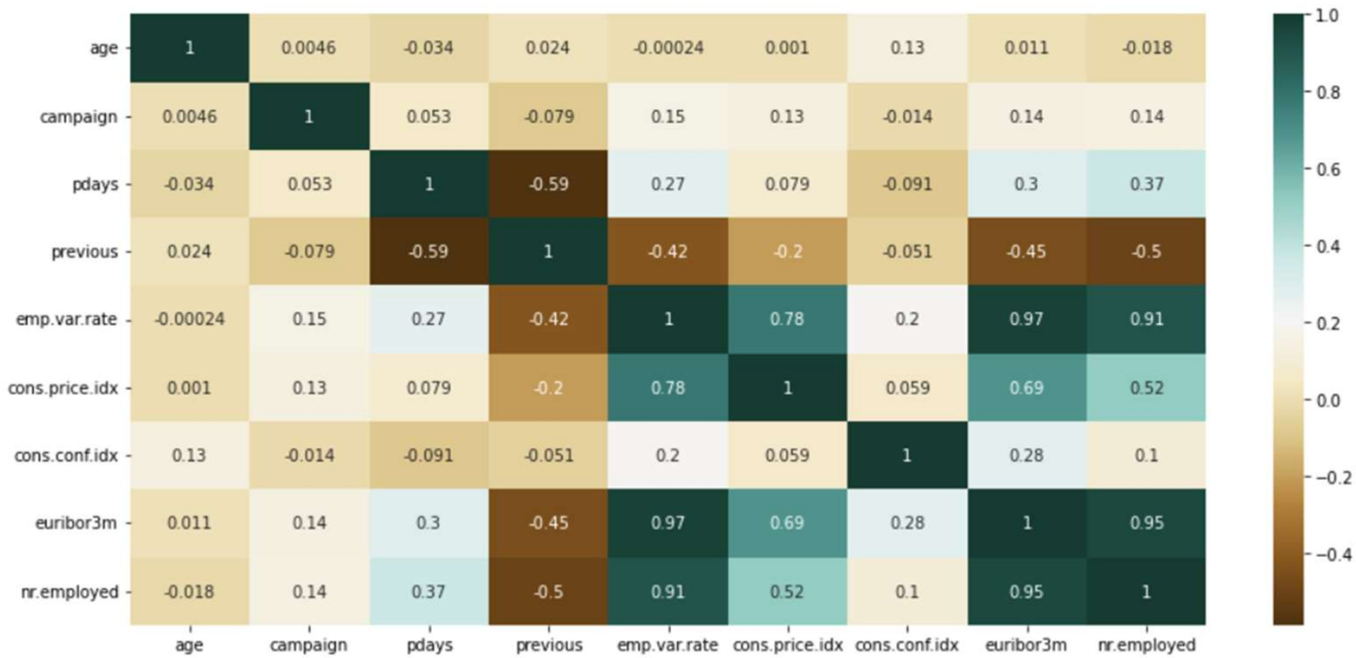
EDA: Marital Status, Job, and Campaign



Data Glacier
Your Deep Learning Partner

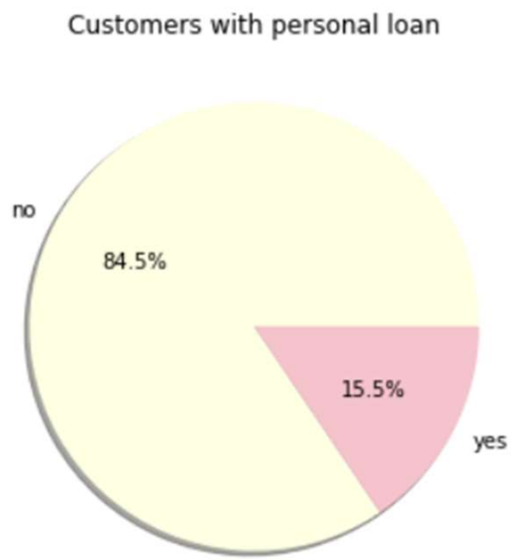
EDA: Correlation Heatmap

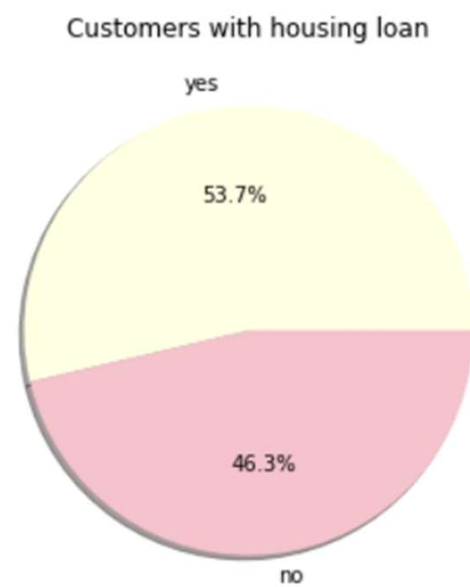
- We can see that there is a high correlation between employment variation rate and euribor 3 month rate, euribor 3 month rate and number of employees, and employment variation rate and number of employees.



EDA: Customer Loans

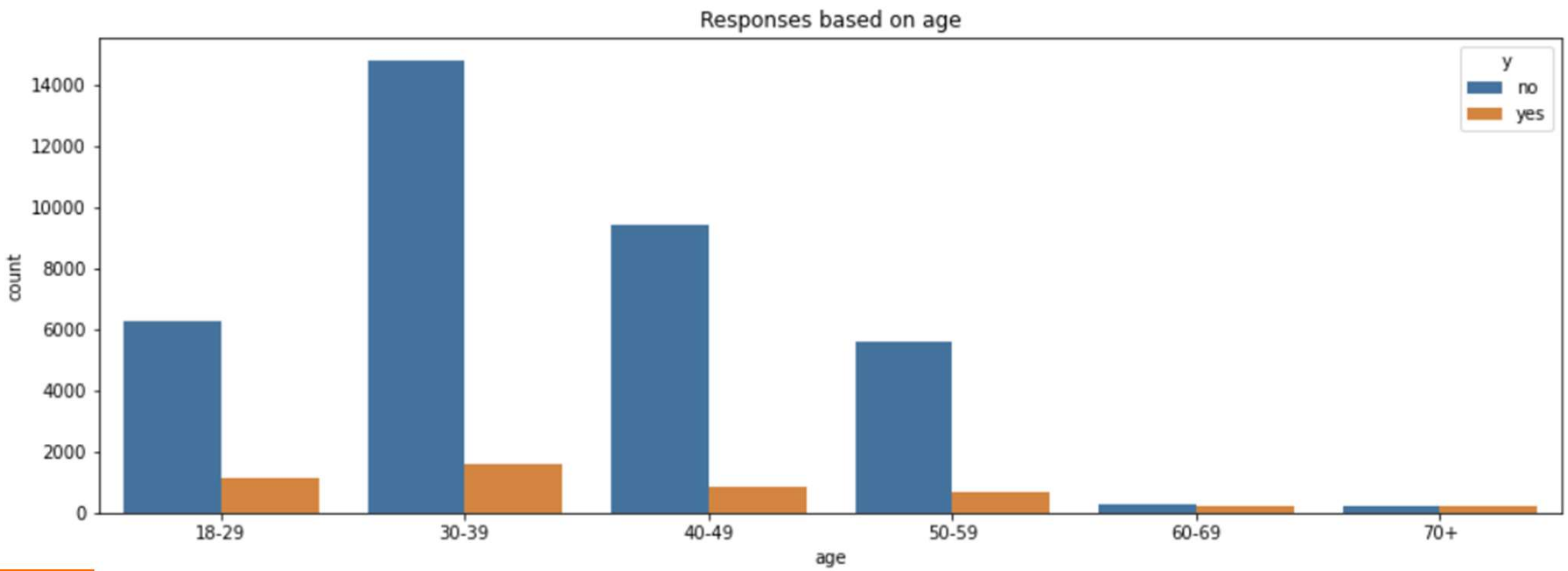
A majority of individuals have housing loans, while few have personal loans.





EDA: Age Group vs Subscriptions

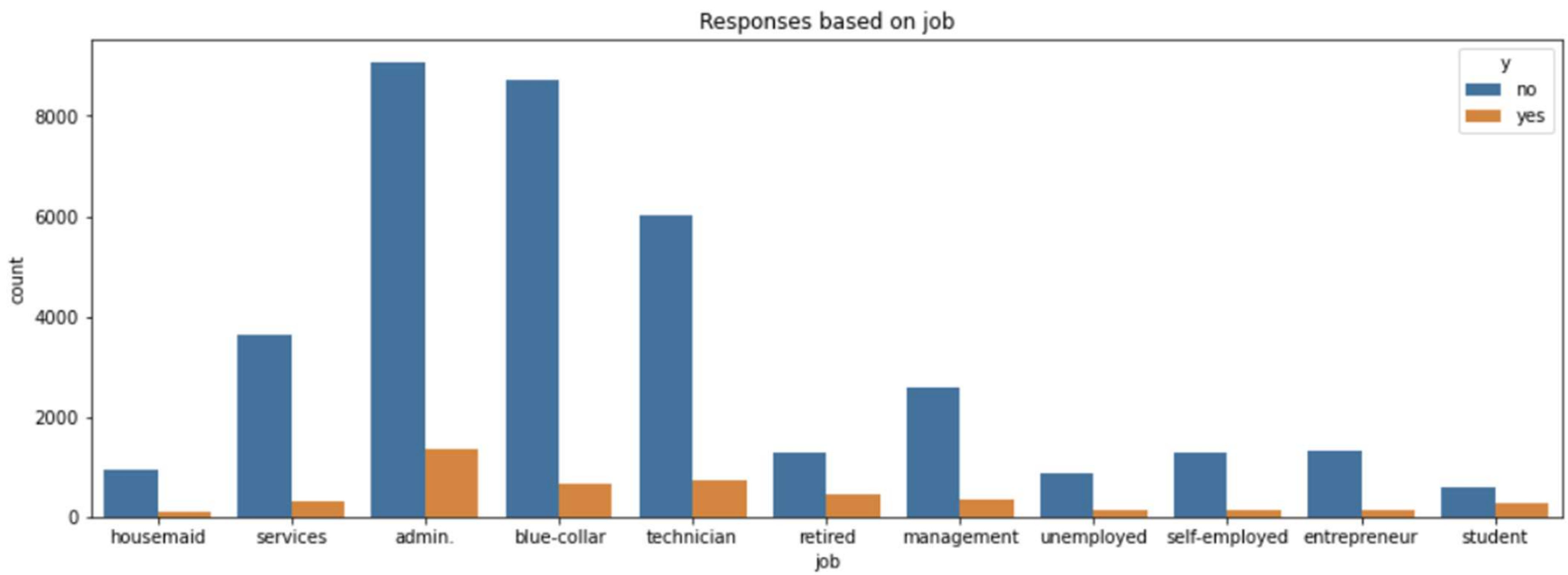
We can see that individuals who are in the age groups of 30-39 and 40-49 have received the greatest count, while those who are aged 60+ have received the least count.



Data Glacier

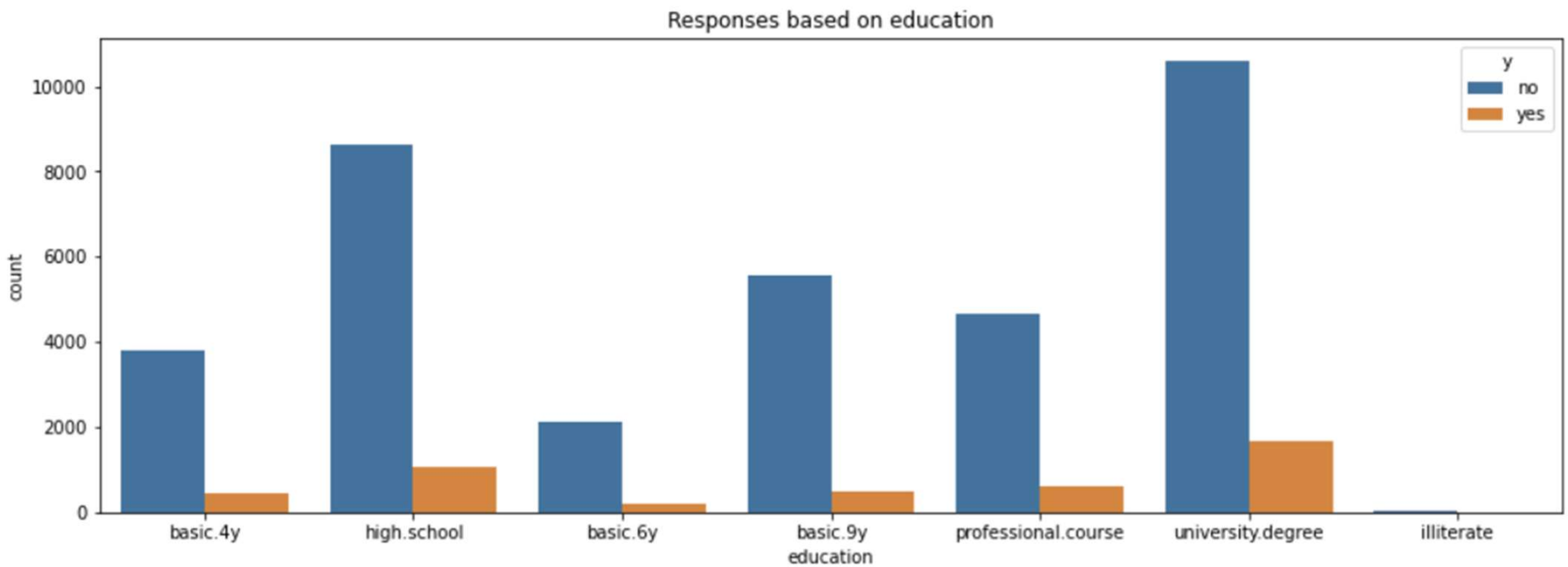
EDA: Job vs Subscriptions

Individuals who have jobs in either administration, blue collar, or technicians have been contacted the most. However, there is a much greater 'yes' outcomes for students.



EDA: Education vs Subscriptions

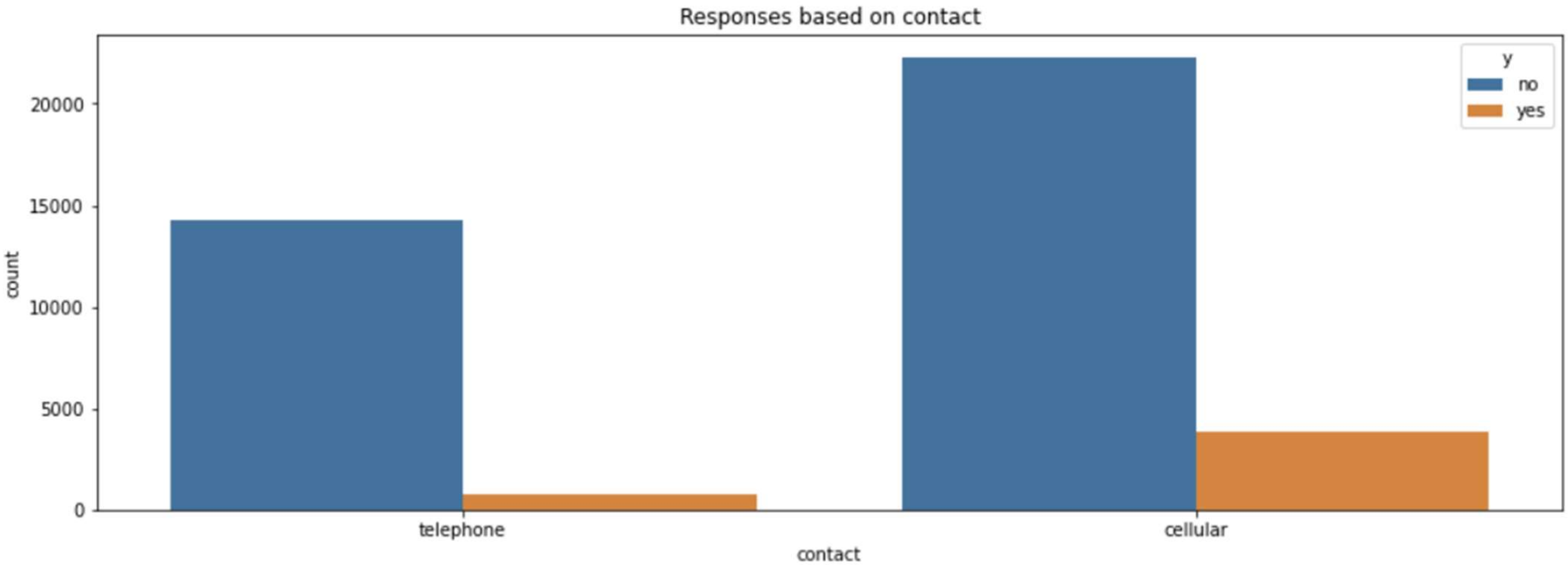
Individuals with a high school education or a university degree were amongst those who were contacted the most. They are also in the group of individuals who subscribed the most. However, those with a high school education had a greater subscription rate.



Data Glacier

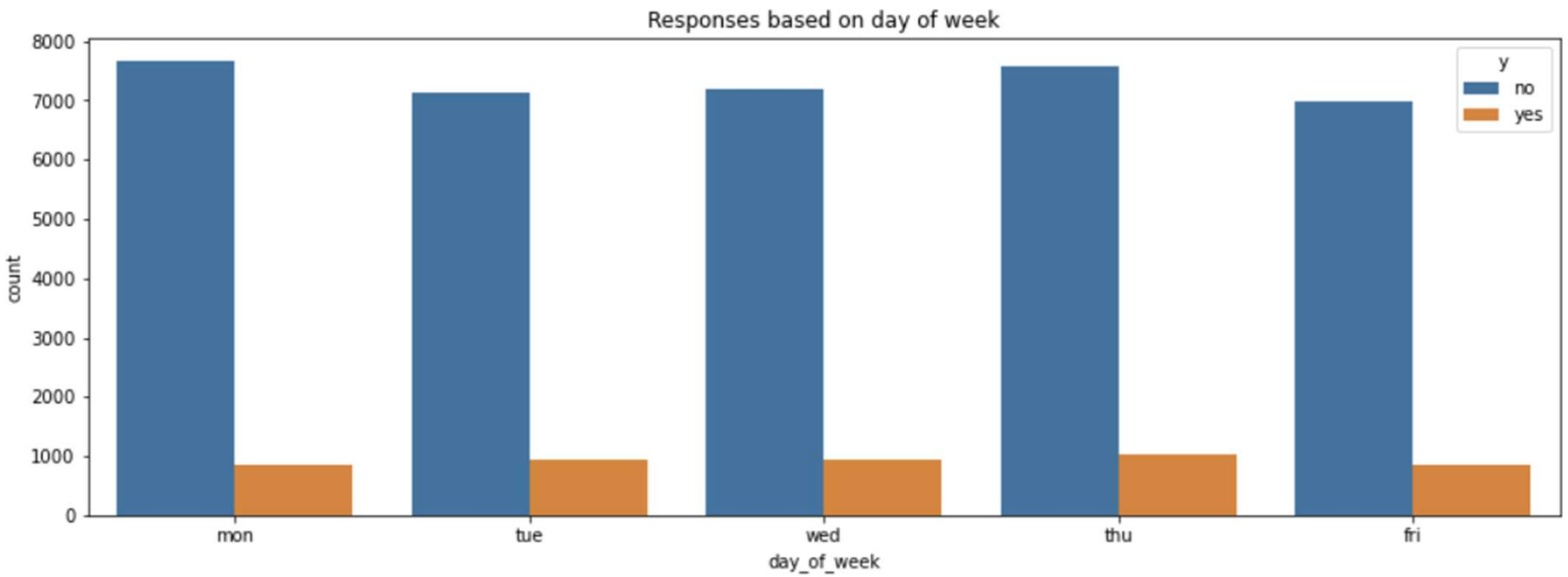
EDA: Contact Type vs Subscriptions

There is a higher rate of those who were contacted through cellular than telephone. Additionally, cellular roughly received more subscriptions



EDA: Day of Week vs Subscriptions

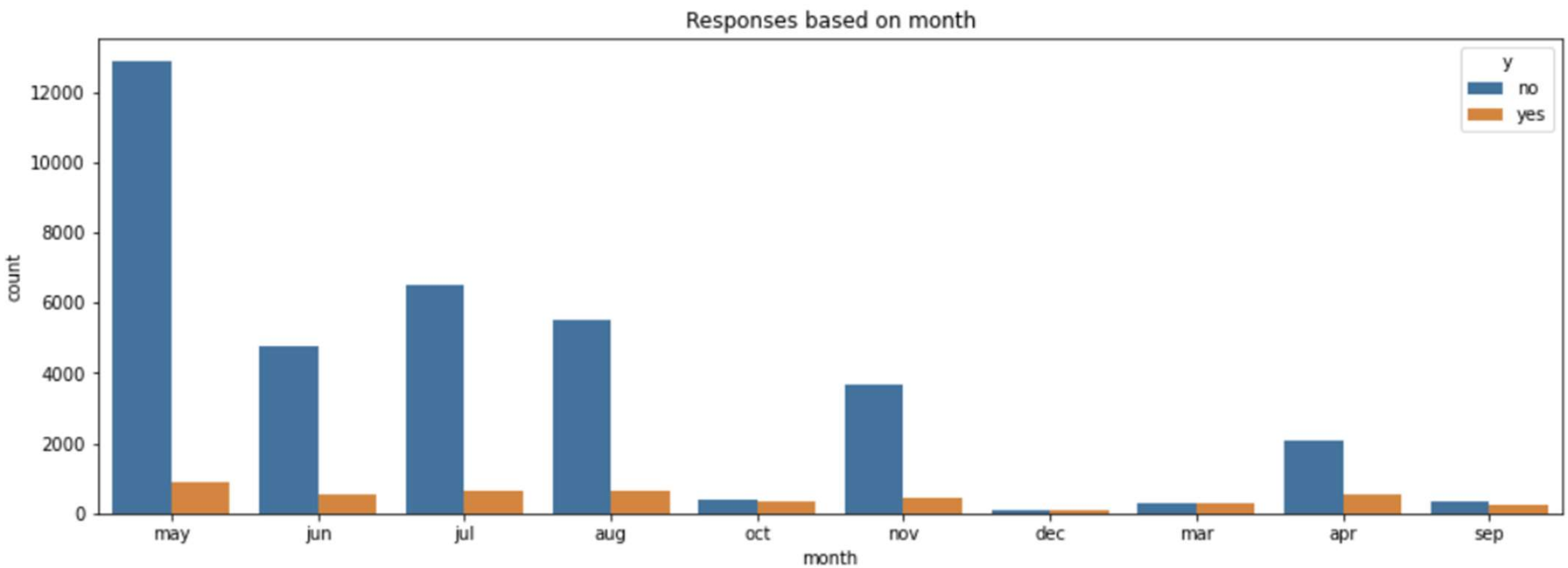
There was little difference of subscriptions based of the day of the week, however, Tuesdays and Thursdays seemed to be the most successful based on the graphs.



Data Glacier

EDA: Month vs Subscriptions

There was a large amount of contact between the months of may and august. This also resulted in higher subscription rates. Additionally, if there was consistent contact throughout the year.



Recommendations

- By looking at the data, it is evident that the data is skewed, favoring the option 'no' or not subscribing to a term deposit.
- We can see that individuals who are in the age groups of 30-39 and 40-49 have received the greatest count, so it is recommended to contact individuals in these age groups rather than others.
- A majority of individuals have housing loans, while few have personal loans.
- Individuals who have jobs in either administration, blue collar, or technicians have been contacted the most, so it is recommended to keep contacting those. However, there is a much greater 'yes' outcomes for students, so it is recommended to contact more of them to maximize subscriptions.
- Individuals with a high school education or a university degree were amongst those who were contacted the most. They are also in the group of individuals who subscribed the most. However, those with a high school education had a greater subscription rate, so it is recommended to contact more high school educated individuals.
- There is a higher rate of those who were contacted through cellular than telephone. Additionally, cellular received more subscriptions, so it is recommended to contact more individuals through their cellular device.
- There was little difference of subscriptions based of the day of the week, however, Tuesdays and Thursdays seemed to be the most successful based on the graphs.
- There was a large amount of contact between the months of may and august. This also resulted in higher subscription rates, so it is recommended to continue contacting individuals between those months. Additionally, if there was consistent contact throughout the year, it would be more efficient to measure which month had the highest subscription success rate.

Data Glacier
Your Deep Learning Partner

Proposed Model Building

- Logistic Regression
  - Binary classifications
- Decision Trees and Random Forest
  - Non-linear relationships
- Evaluation Metrics
  - Imbalances in data
- Gradient Boosting Algorithms
  - Classification and complex data

Data Glacier
Your Deep Learning Partner

Thank You,

Maria Contractor

Data Glacier

Your Deep Learning Partner