Group Name: BRAVO

Name	Email (registered with Data Glacier)	Country	College/Company	Specialization
Jackson Taylor	jacksonian.r.taylor@gmail.com	United States	Santa Clara University	Data Science
Balamurugan Purushothaman	balamurugan2001viruda@gmail.com	United Kingdom	University of Liverpool	Data Science
Nazrin Thanikattil Rafeeque	101nazrin@gmail.com	United Kingdom	University of Hertfordshire	Data Science
Gunjan Varyani	gunjanvaryani916@gmail.com	United States	University of the Cumberlands	Data Science

1. 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 helps 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).

The steps to solving this task include outlining the project, the initial data understanding and strategies to solve data problems, data cleansing and transformation, exploratory data analysis code, exploratory data analysis presentation and model recommendation, model selection and building, and presenting the final solution and code.

2. Data Understanding:

Section 1: Data feature meanings

age: age of the client

job: client’s job type

categories: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown'

Note: Not every type of job is listed. They fall into certain job categories.

marital: client’s marital status

categories: 'divorced','married','single','unknown'

Note: 'divorced' means divorced or widowed

education: client’s level of education

categories: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course',

'university.degree','unknown'

Assumptions:

- the highest education level achieved is used

- illiterate is the default if client has no education with any of the education categories

- if client is illiterate, 'illiterate' also replaces whatever level of education they may have

- education rank: ('illiterate','basic.4y','basic.6y','basic.9y','high.school', 'university.degree','professional.course')

'unknown' is unranked since it is best not to assume.

default: whether client has failed to meet the legal obligations of a loan or credit agreement

categories: 'no','yes','unknown'

yes – One or more loan or credit agreements are in default.

no – No loan or credit agreement is in default.

housing: client has housing loan or not

categories: 'no','yes','unknown'

loan: client has personal loan or not categories: 'no','yes','unknown'

contact: client’s contact communication type of the last contact

categories: 'cellular','telephone'

month: month of last contact to the client

categories: 'jan','feb','mar',...,'nov','dec'

day_of_week: day of week of last contact to the client

categories: 'mon','tue','wed','thu','fri'

duration: last contact duration, in seconds with client

Note: The duration is not known before a call is performed, so this data needs to be collected after the call. However, after the end of the call, y is obviously known since it should be on record that they accepted or rejected the product. This somewhat defeats the point of using a model with this feature.

campaign: number of contacts performed during this campaign and for this client (includes last contact)

pdays: number of days that passed by after the client was last contacted from a previous campaign (999 means client was not previously contacted)

previous: number of contacts performed before this campaign and for this client

poutcome: outcome of the previous marketing campaign for the client

categories: 'failure','nonexistent','success'

Assumption: This concerns the client itself, not the marketing campaign as a whole.

emp.var.rate: represents the percentage change in employment levels from one quarter of a year to the next.

Note: The difference is taken between the last completed quarter and the quarter before that.

cons.price.idx: (CPI) tracks the changes in the price level of a basket of consumer goods and services from one month to the next.

Note: The difference is taken between the last completed month and the month before that.

cons.conf.idx: monthly indicator that measures the degree of optimism or pessimism that consumers feel about the overall state of the economy

Note: Monthly is the frequency in which the index is updated and reported. The last complete month is used.

euribor3m: euribor 3 month rate - daily indicator: Euro Interbank Offered Rate for a three-month maturity with a daily indicator

Notes:

Euribor: average interest rate at which a large panel of European banks borrow funds from one another.

3-month: the maturity or the duration for which the funds are borrowed

Daily indicator: The rate of the most recent complete business day is used. The interest rate is calculated and published every business day.

nr.employed: number of employees at ABC Bank for the last quarter

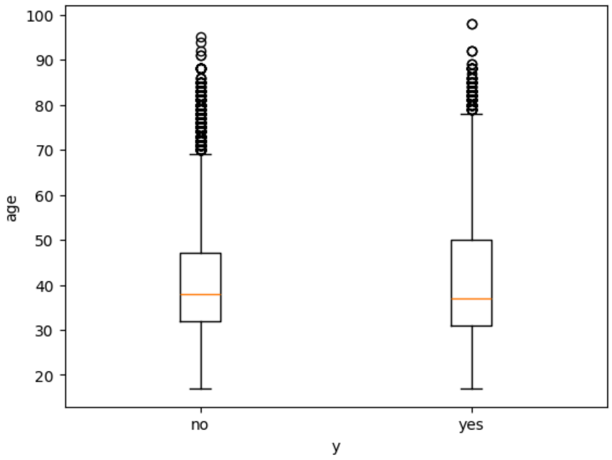
y: has the client subscribed a term deposit?

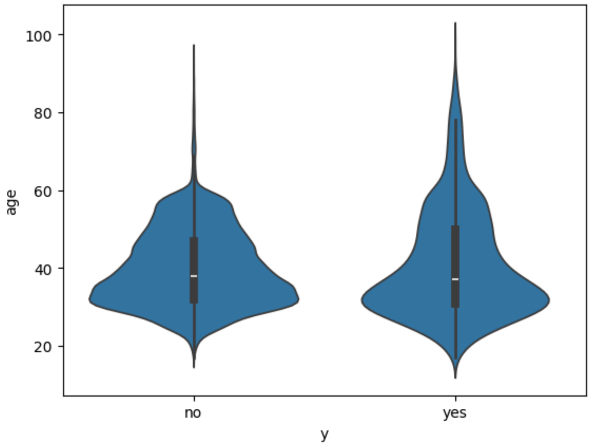
categories: 'yes','no'

Note: target feature

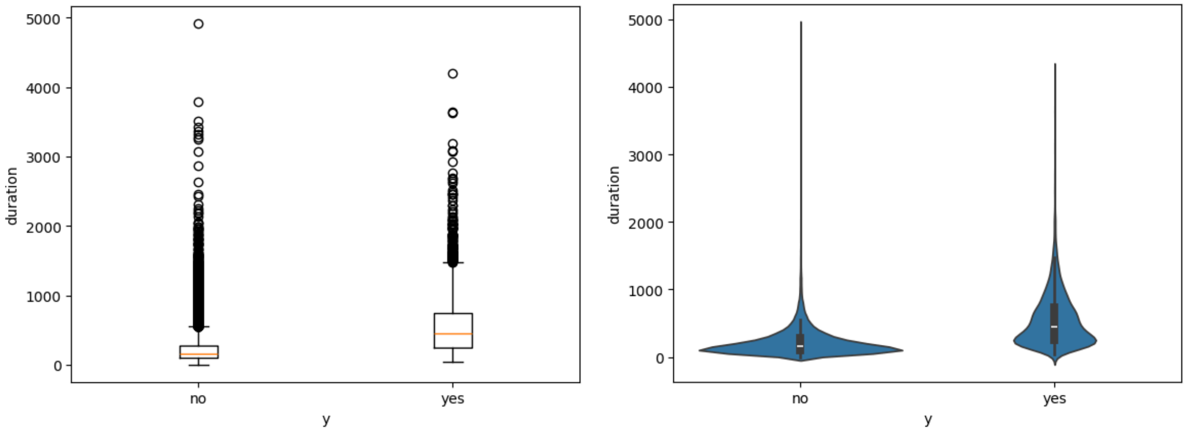
Section 2: **(target variable(y) X numerical variables) with box plot and violin plot**

age:

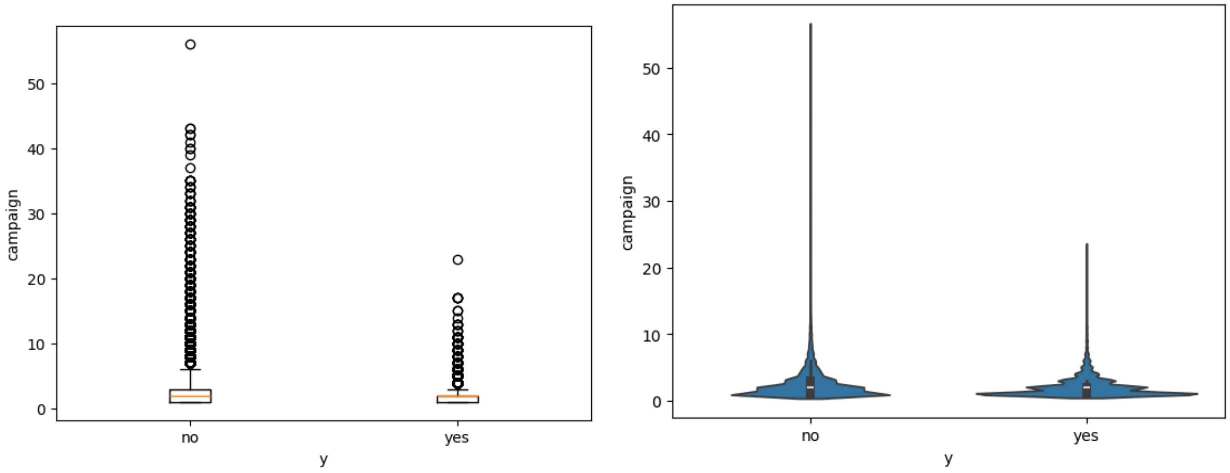




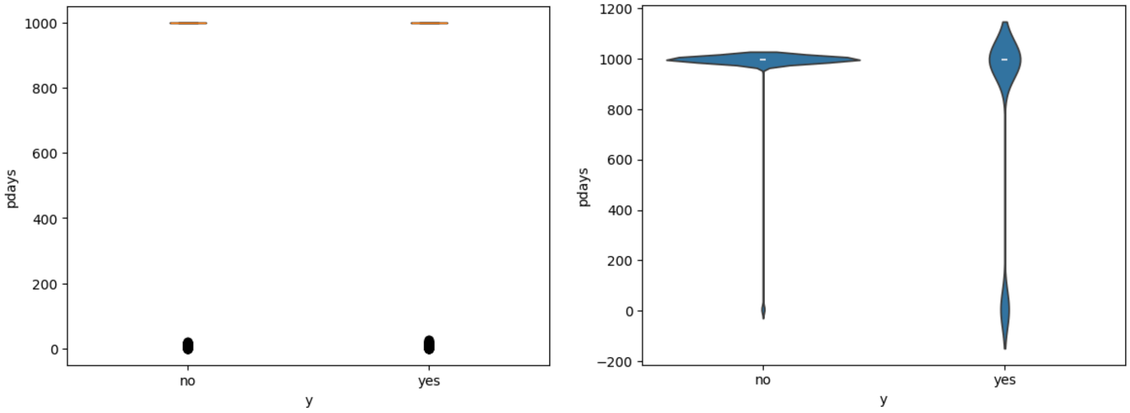
duration:



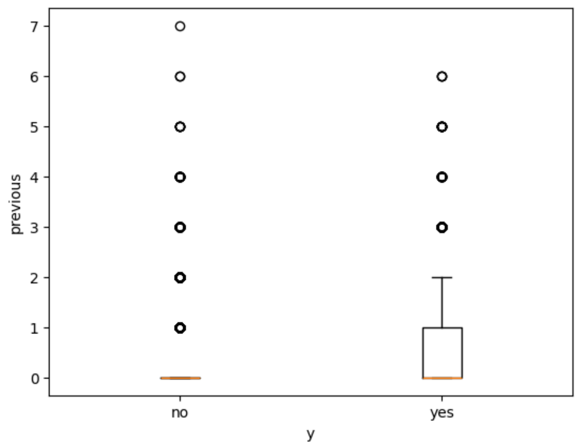
campaign:

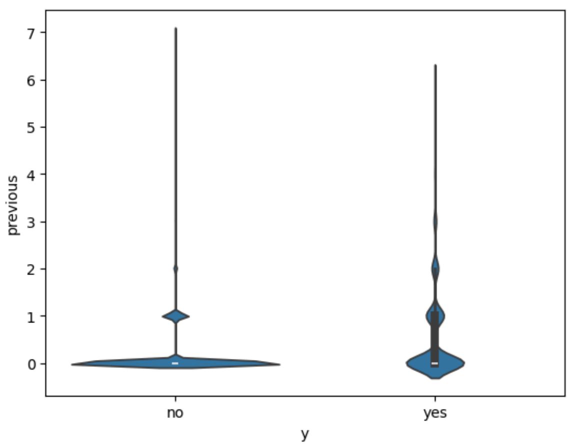


pdays:

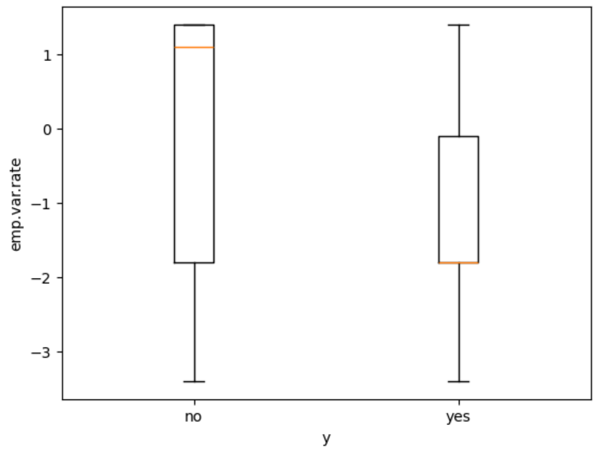


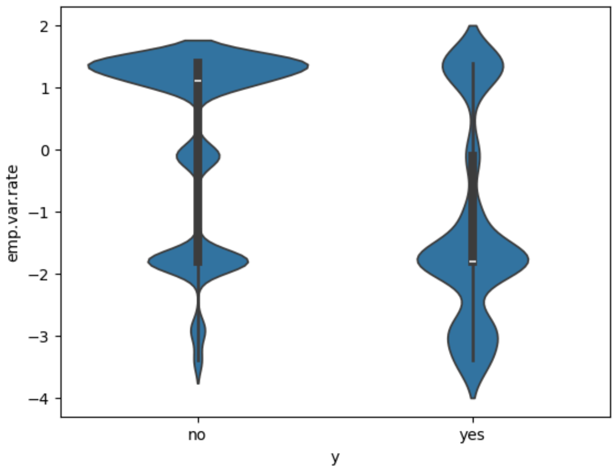
previous:



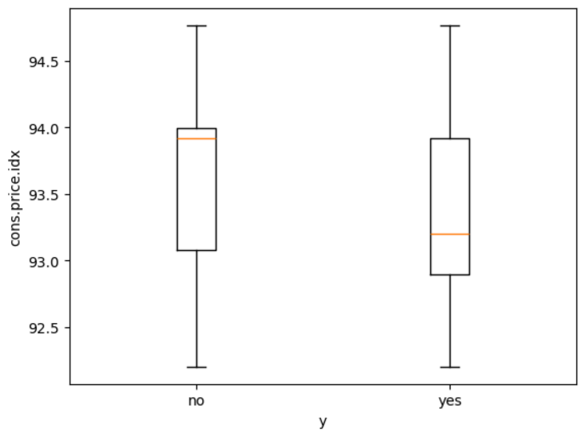


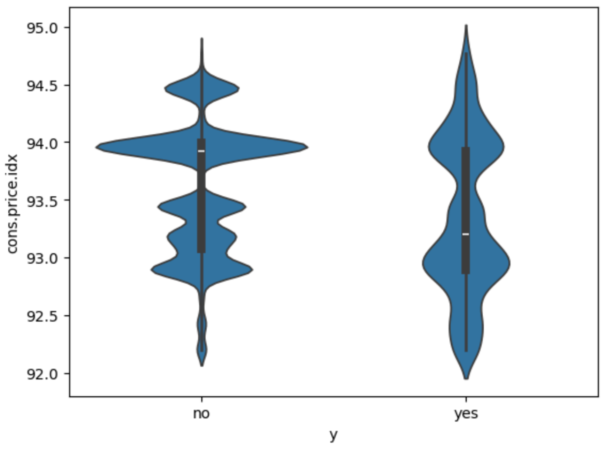
emp.var.rate:



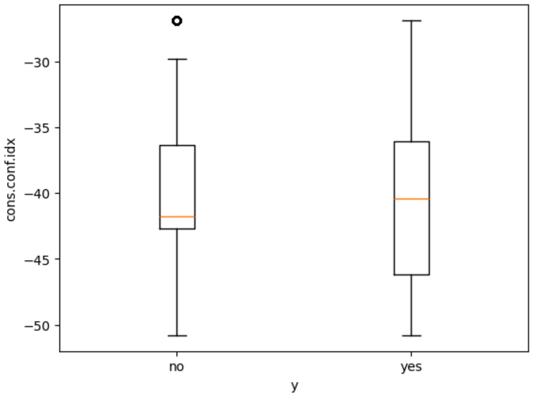


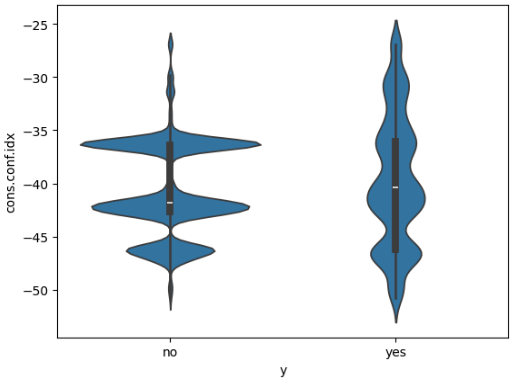
cons.price.idx:



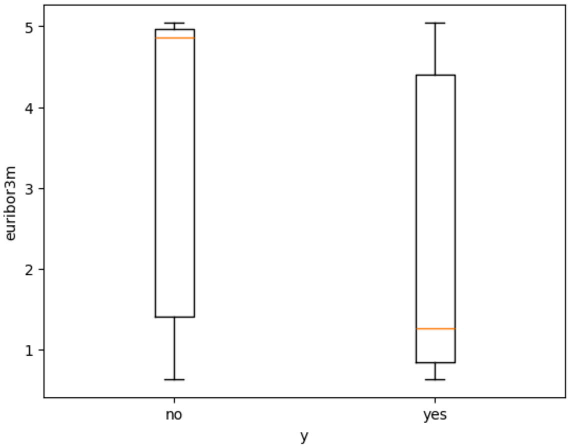


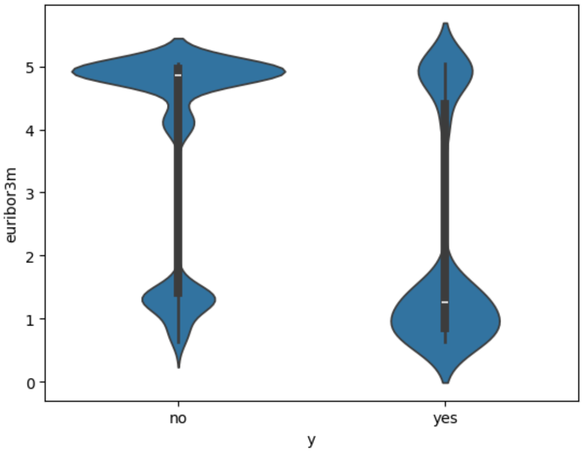
cons.conf.idx:



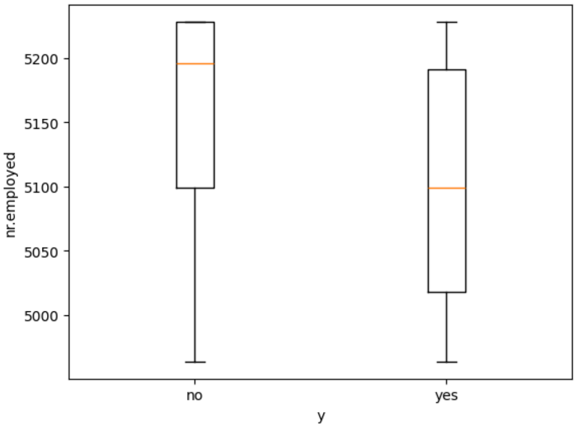


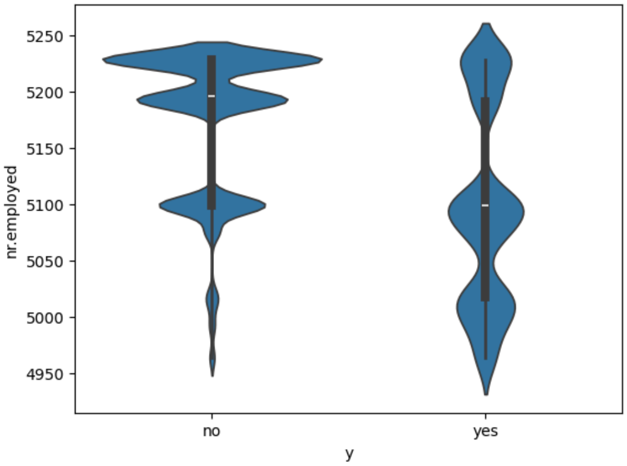
euribor3m:



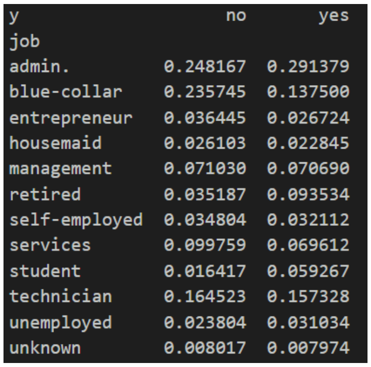


nr.employed:

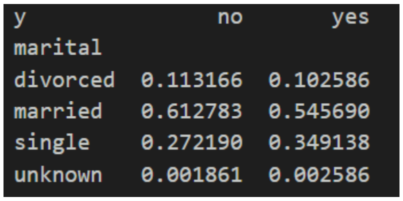




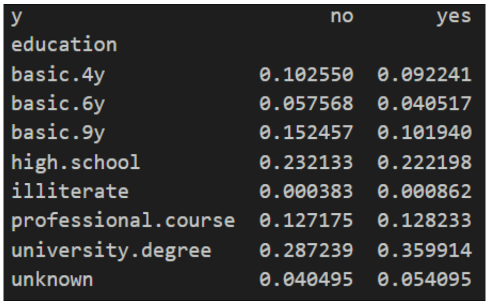
Section 3: (target variable(y) X categorical variables) with counts for each column value normalized by dividing by the total counts for the no and yes column respectively.



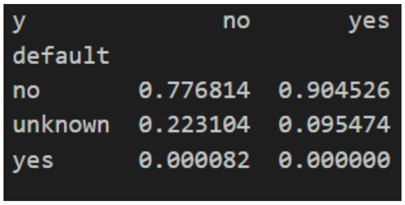
```
y                    no       yes
job
admin.         0.248167  0.291379
blue-collar    0.235745  0.137500
entrepreneur   0.036445  0.026724
housemaid      0.026103  0.022845
management     0.071030  0.070690
retired        0.035187  0.093534
self-employed  0.034804  0.032112
services       0.099759  0.069612
student        0.016417  0.059267
technician     0.164523  0.157328
unemployed     0.023804  0.031034
unknown        0.008017  0.007974
```



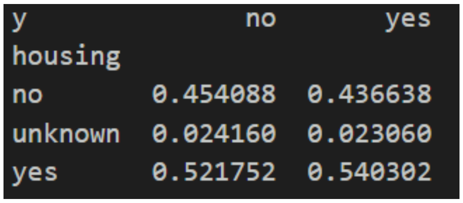
```
y               no       yes
marital
divorced  0.113166  0.102586
married   0.612783  0.545690
single    0.272190  0.349138
unknown   0.001861  0.002586
```



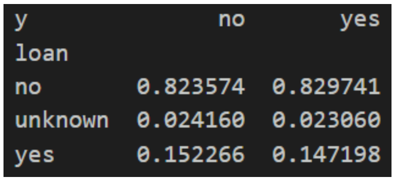
```
y                          no       yes
education
basic.4y             0.102550  0.092241
basic.6y             0.057568  0.040517
basic.9y             0.152457  0.101940
high.school          0.232133  0.222198
illiterate           0.000383  0.000862
professional.course  0.127175  0.128233
university.degree    0.287239  0.359914
unknown              0.040495  0.054095
```



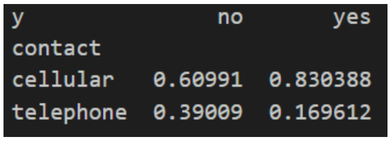
```
y              no       yes
default
no       0.776814  0.904526
unknown  0.223104  0.095474
yes      0.000082  0.000000
```



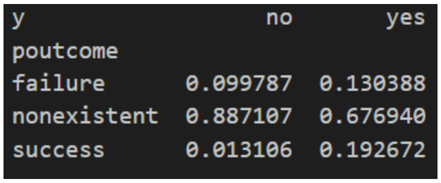
```
y              no       yes
housing
no       0.454088  0.436638
unknown  0.024160  0.023060
yes      0.521752  0.540302
```



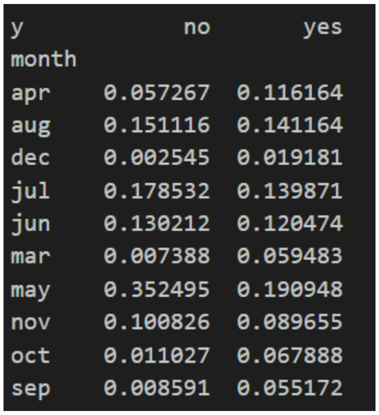
```
y              no       yes
loan
no       0.823574  0.829741
unknown  0.024160  0.023060
yes      0.152266  0.147198
```



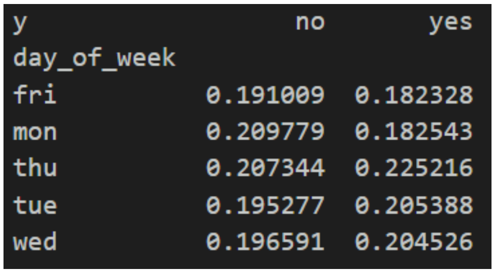
```
y               no       yes
contact
cellular   0.60991  0.830388
telephone  0.39009  0.169612
```



```
y                  no       yes
poutcome
failure      0.099787  0.130388
nonexistent  0.887107  0.676940
success      0.013106  0.192672
```

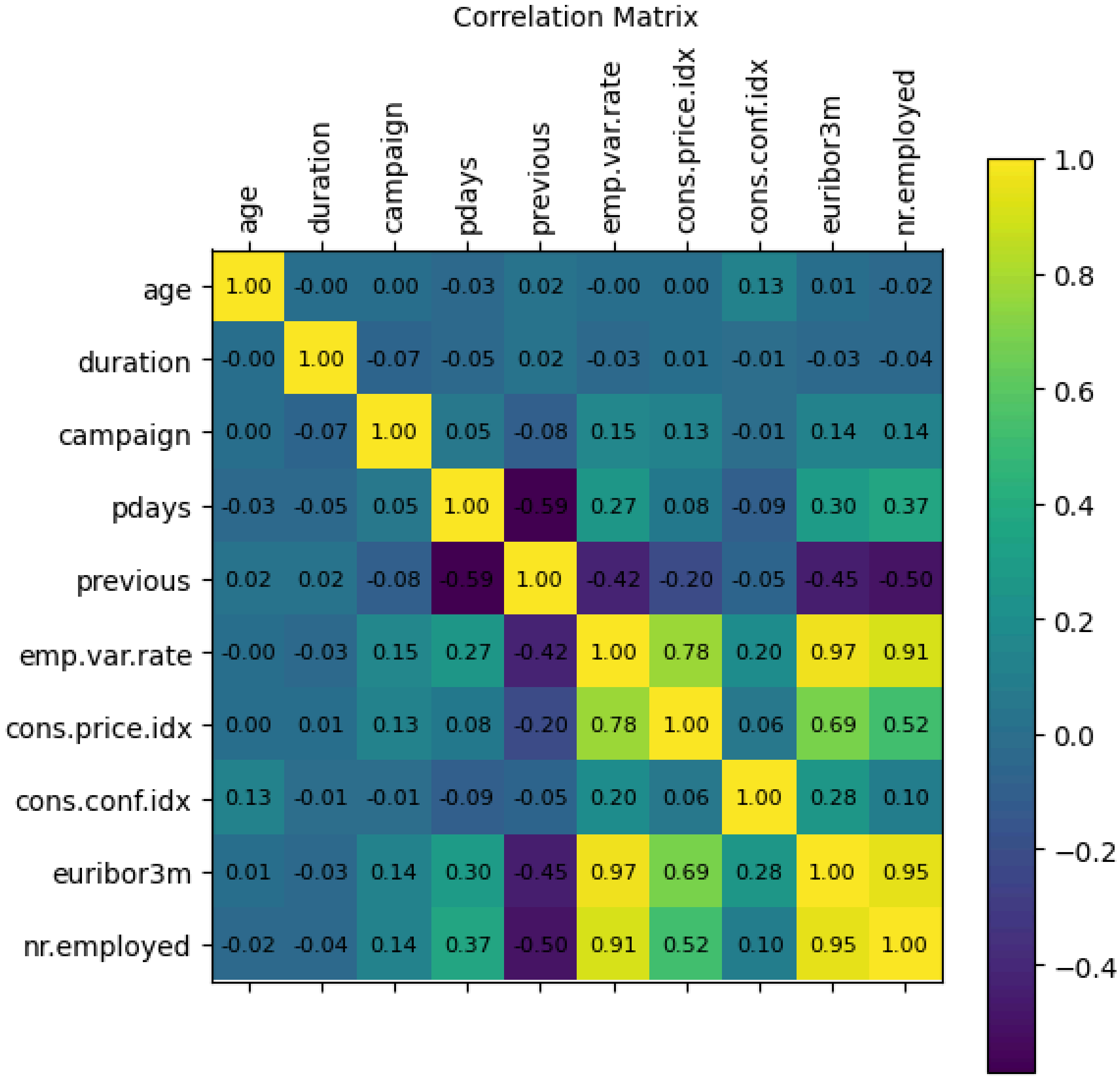


```
y            no       yes
month
apr    0.057267  0.116164
aug    0.151116  0.141164
dec    0.002545  0.019181
jul    0.178532  0.139871
jun    0.130212  0.120474
mar    0.007388  0.059483
may    0.352495  0.190948
nov    0.100826  0.089655
oct    0.011027  0.067888
sep    0.008591  0.055172
```



```
y                  no       yes
day_of_week
fri          0.191009  0.182328
mon          0.209779  0.182543
thu          0.207344  0.225216
tue          0.195277  0.205388
wed          0.196591  0.204526
```

Section 4: Correlation analysis



	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
age	1.00	-0.00	0.00	-0.03	0.02	-0.00	0.00	0.13	0.01	-0.02
duration	-0.00	1.00	-0.07	-0.05	0.02	-0.03	0.01	-0.01	-0.03	-0.04
campaign	0.00	-0.07	1.00	0.05	-0.08	0.15	0.13	-0.01	0.14	0.14
pdays	-0.03	-0.05	0.05	1.00	-0.59	0.27	0.08	-0.09	0.30	0.37
previous	0.02	0.02	-0.08	-0.59	1.00	-0.42	-0.20	-0.05	-0.45	-0.50
emp.var.rate	-0.00	-0.03	0.15	0.27	-0.42	1.00	0.78	0.20	0.97	0.91
cons.price.idx	0.00	0.01	0.13	0.08	-0.20	0.78	1.00	0.06	0.69	0.52
cons.conf.idx	0.13	-0.01	-0.01	-0.09	-0.05	0.20	0.06	1.00	0.28	0.10
euribor3m	0.01	-0.03	0.14	0.30	-0.45	0.97	0.69	0.28	1.00	0.95
nr.employed	-0.02	-0.04	0.14	0.37	-0.50	0.91	0.52	0.10	0.95	1.00

Section 5: Initial review of data features with visualizations, knowledge, and intuition

Note: The insights gathered here are based on first impressions. Confirmation and more rigorous analysis will be delivered later on.

From review of the visualizations at this stage there are certain features that don't seem to be helpful for model building.

Upon inspection of the visualizations of Section 5, there seems to be two features that standout as weak predictors to the target value. These features are age and campaign. When referring to the violin plot, they have the same shaped distribution for the yes and no values of the target feature. This indicates that there is little impact that these numerical features have on the target value.

After the normalization of the categorical frequencies that takes place in Section 6, you can observe the effect of the feature on the target value. If for all the values of a categorical feature, the proportion between the values of y (no and yes) are similar, then it would seem that the categorical feature is not a useful predictor of the target value. When reviewing the categorical features in the dataset there seems to be three features that standout in this way (housing, loan, and day_of_week).

By inspection of the correlation matrix in Section 7, the correlation between euribor3m and emp.var.rate as well as euribor3m and nr.employed are quite high. As part of feature selection, it is often appropriate to remove all but one of the independent features that are highly correlated to simplify the model training and improve performance of the trained model. In this case the correct action might be to remove both emp.var.rate and nr.employed.

There are also some features that are expected to perform highly as independent variables for a model.

Starting with strong numeric features: Excluding the four low impact numerical features determined above (age, campaign, emp.var.rate and nr.employed) there are three stand out numerical features from the rest. (Note: emp.var.rate and nr.employed are part of the four excluded features for simplicity, but instead, it might be better to only remove euribor3m or keep all 3 of these features(euribor3m, emp.var.rate, and nr.employed) regardless of correlations.) The standout features are cons.price.idx, cons.conf.idx, and euribor3m. When referring to the violin plots, they have significantly different shapes between values of the target feature (no and yes) which indicate strong predictive potential. One notable thing in common about these features is that they reflect the state of the economy and therefore they impact a vast range of clients. Furthermore, the data description indicates that duration is also a highly predictive independent feature, but it is less useful in practice (See Section 1).

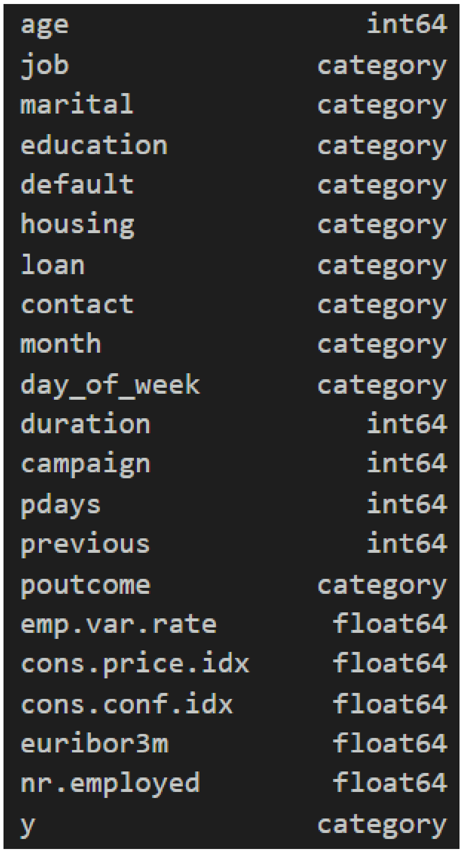
As for categorical features, upon inspection, default, contact, and poutcome are strong predictors of the target value. The notion that default and poutcome are useful predictors makes sense. default correlates with a client’s bank habits which can affect whether the client will buy the product. As for the poutcome feature, it makes sense that someone who bought the term deposit product from the previous campaign would buy it again. The contact feature is more surprising, but the type of phone system used does strongly indicate the likelihood of purchasing the product.

3. Data types and Data structure

Note: Data file used is bank-additional-full.csv

Note: It uses ; as the delimiter

These are the Pandas datatypes for each feature:



```
age                  int64
job               category
marital           category
education         category
default           category
housing           category
loan              category
contact           category
month             category
day_of_week       category
duration             int64
campaign             int64
pdays                int64
previous             int64
poutcome          category
emp.var.rate       float64
cons.price.idx     float64
cons.conf.idx      float64
euribor3m          float64
nr.employed        float64
y                 category
```

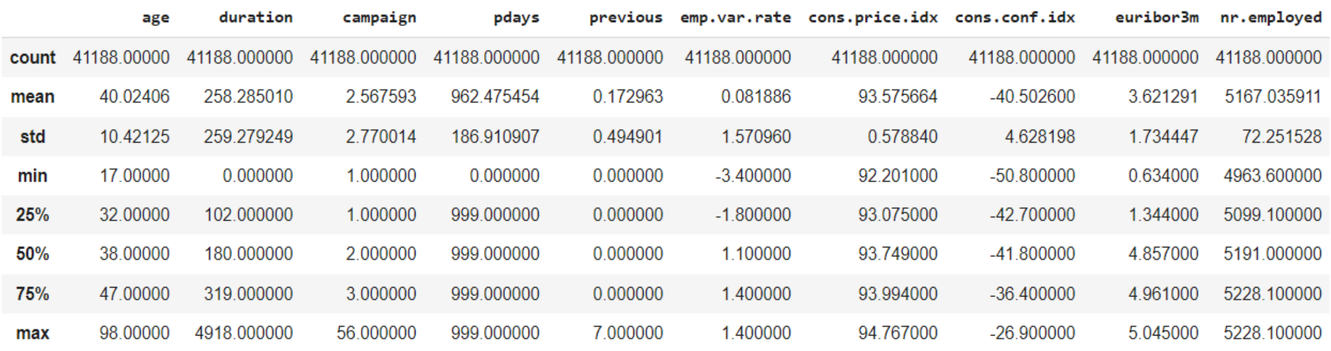
Note:

- contact and y could be considered binary variables.
- education, month, and day_of_week could be considered ordinal variables.

4. Problems within the Data

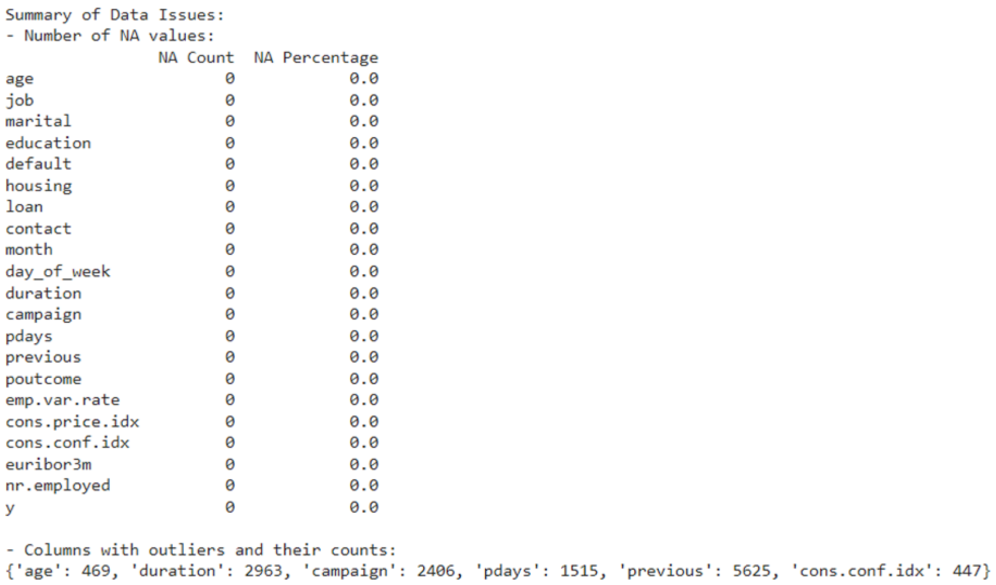
This section examines the data quality issues in the provided dataset, such as the number of missing values, outliers, and skewed distributions. The dataset has 41,188 records and includes a variety of parameters such as age, duration, campaign, etc.Following images indicates the consolidated statistical observations

Detailed Analysis



	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
count	41188.00000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000
mean	40.02406	258.285010	2.567593	962.475454	0.172963	0.081886	93.575664	-40.502600	3.621291	5167.035911
std	10.42125	259.279249	2.770014	186.910907	0.494901	1.570960	0.578840	4.628198	1.734447	72.251528
min	17.00000	0.000000	1.000000	0.000000	0.000000	-3.400000	92.201000	-50.800000	0.634000	4963.600000
25%	32.00000	102.000000	1.000000	999.000000	0.000000	-1.800000	93.075000	-42.700000	1.344000	5099.100000
50%	38.00000	180.000000	2.000000	999.000000	0.000000	1.100000	93.749000	-41.800000	4.857000	5191.000000
75%	47.00000	319.000000	3.000000	999.000000	0.000000	1.400000	93.994000	-36.400000	4.961000	5228.100000
max	98.00000	4918.000000	56.000000	999.000000	7.000000	1.400000	94.767000	-26.900000	5.045000	5228.100000

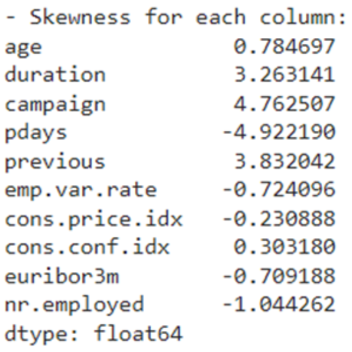
Figure 1



```
Summary of Data Issues:
- Number of NA values:
                NA Count  NA Percentage
age                    0            0.0
job                    0            0.0
marital                0            0.0
education              0            0.0
default                0            0.0
housing                0            0.0
loan                   0            0.0
contact                0            0.0
month                  0            0.0
day_of_week            0            0.0
duration               0            0.0
campaign               0            0.0
pdays                  0            0.0
previous               0            0.0
poutcome               0            0.0
emp.var.rate           0            0.0
cons.price.idx         0            0.0
cons.conf.idx          0            0.0
euribor3m              0            0.0
nr.employed            0            0.0
y                      0            0.0

- Columns with outliers and their counts:
{'age': 469, 'duration': 2963, 'campaign': 2406, 'pdays': 1515, 'previous': 5625, 'cons.conf.idx': 447}
```

Figure 2



```
- Skewness for each column:
age               0.784697
duration          3.263141
campaign          4.762507
pdays            -4.922190
previous          3.832042
emp.var.rate     -0.724096
cons.price.idx   -0.230888
cons.conf.idx     0.303180
euribor3m        -0.709188
nr.employed      -1.044262
dtype: float64
```

Figure 3

Missing Values

The dataset has been checked for missing values. It could be found that no data columns have missing values in it. This is a favorable element because it signifies that data collecting is complete.

Outliers

Outliers are data points that are significantly different from the expected range of values. They typically influence data analysis outcomes. In the given dataset the 10 numerical columns are subjected to outlier inspection by calculating the InterQuartile range from Q1 and Q3 to decide the lower and upper boundary of data values to decide the outliers existing in each column. Following columns contained outliers:

- Age: 469 outliers
- Duration: 2963 outliers
- Campaign: 2406 outliers
- Pdays: 1515 outliers
- Previous: The IQR is 0, indicating no variation.
- cons.conf.idx: 447 outliers

Categorical data: Categorical data, which represents discrete categories or labels, without a natural ordering or numerical distribution that allows for statistical analysis.These data columns where analyzed in excel for frequency distributions and discovered to be well-distributed, indicating high data quality.

Skewness

Skewness measures the asymmetry of the data distribution.From Figure :3 it could be inferred that the dataset demonstrates a variety of skewness features. Age, Duration, Campaign, and Previous are all positively skewed, with Campaign and Duration being significantly skewed, showing lengthy right tails and a concentration of low values. Pdays is extremely negatively skewed, with a long left tail and primarily high values offset by a few low outliers. Emp.var.rate, euribor 3m, and nr.employed have mildly negative skewed distributions, indicating greater values and fewer low outliers. Cons.price.idx and cons.conf.idx are slightly skewed, with the former negative and the latter positive, implying near-symmetry with tiny deviations. These skewness features reveal considerable asymmetries in data distributions, which affects data processing and interpretation.

Standard Deviation

The examination of standard deviation reveals how different components vary in the sample. Age does not vary significantly, indicating that ages are similar. However, Duration and Pdays have a broad range of values, implying that some data points may be distant from the average, influencing our results. Previous connections vary little, potentially limiting what we can learn from them. Economic indicators such as emp.var.rate, cons.price.idx, and euribor3m fluctuate moderately, reflecting economic ups and downs. Consumer confidence fluctuates moderately, showing fluctuations in how consumers feel about the economy. The number of employed fluctuates dramatically, reflecting changes in job levels. To better comprehend our data, we should focus on extreme numbers, particularly in Duration and Pdays, in order to make appropriate conclusions.

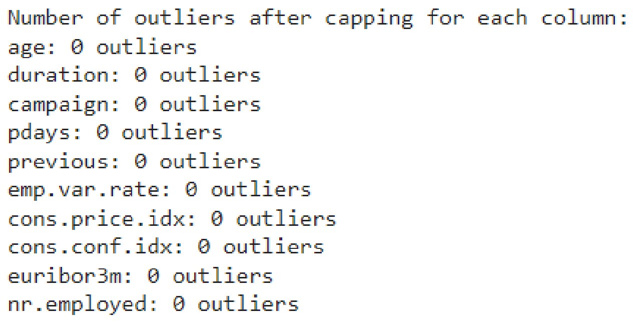
One other problem with this data is data imbalance in the target variable. The majority class is hugely in favor of no vs yes (36548 vs 4640). This can lead to issues later on with model development. Many models will likely bias the “no” class over the other if the data is used as it is. This can lead to Negative Predictive Value (NPV) being high and the Precision value being low assuming that negative is no and positive is yes. This means that the model will be much more effective at predicting a negative value correctly than predicting a positive value correctly.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives}+\text{False Positives}} \quad \text{NPV} = \frac{\text{True Negatives}}{\text{True Negatives}+\text{False Negatives}}$$

5. Approaches for Outliers and Skewness Management

Based on the above data analysis there exist a significant number of outliers in the majority of data columns along with skewness to the data points.As these can significantly impact the data analysis and modeling,data preprocessing is very crucial.

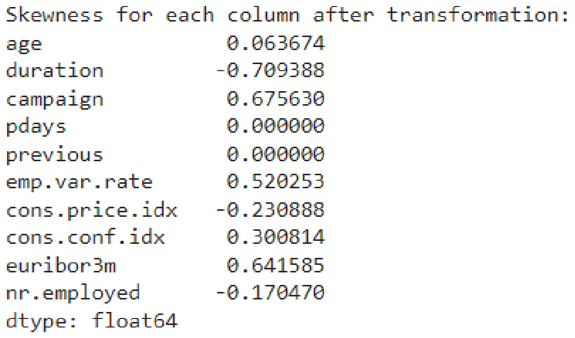
To address the outliers we could apply the **Capping** technique which replaces the extreme values with the nearest thresholds within the acceptable range.It allows to maintain the integrity of the dataset by retaining valuable information that would be relevant for analysis.Capping is performed by replacing least valued outliers with lower bound and and the most value with the upper bound of IQR.This effectively reduced the outliers to zero across all columns.



```
Number of outliers after capping for each column:
age: 0 outliers
duration: 0 outliers
campaign: 0 outliers
pdays: 0 outliers
previous: 0 outliers
emp.var.rate: 0 outliers
cons.price.idx: 0 outliers
cons.conf.idx: 0 outliers
euribor3m: 0 outliers
nr.employed: 0 outliers
```

As skewness within the data can impact the assumptions of statistical model .To address this issue data transformation has to be performed.Both positively and negatively skewed data has to be considered because positive skewness (tail to the right) could lead interpretations toward higher values, whereas negative skewness (tail to the left) do the opposite. There are several methods for data transformations like square root ,cube root,logarithm, Boxcox etc .In that **Logarithmic transformation** is an effective method .It has the capacity to compress larger values more than smaller ones, lowering the impact of extreme outliers and bringing the distribution closer to normality.

Positive skewness can be identified when skewness > 0.5 and are reduced using logarithmic transformations log(1+x) , while negative skewness can be identified when skewness <0.5 ,and are managed through reverse log transformations log(1+max(x)−x).Following figure represent the skewness after log transformation on data.



```
Skewness for each column after transformation:
age                0.063674
duration          -0.709388
campaign           0.675630
pdays              0.000000
previous           0.000000
emp.var.rate       0.520253
cons.price.idx    -0.230888
cons.conf.idx      0.300814
euribor3m          0.641585
nr.employed       -0.170470
dtype: float64
```

From the post transformation data it could be observed the skewness values across most columns significantly improved, resulting in a more symmetric distribution.After applying outlier and skewness reduction we could also observe that standard deviations have largely decreased. This decline implies that data distributions have grown more concentrated around their means and less dispersed. Attributes with formerly large standard deviations, such as 'duration' and 'nr.employed', now have significantly lower values, indicating a more homogeneous distribution of values following data adjustments.

To handle data imbalance, there are many approaches that can be tested later on. For now, the simple suggested method to use is Synthetic Minority Oversampling Technique(SMOTE), which can generate synthetic examples in the minority class to balance the dataset.