Data Science Intern at Data Glacier

Project: Bank Marketing (Campaign)

Week 9: Deliverables

- **Group Name:** LimitLess Team
- **Team member's details:**
  1. **Name:** Chitra S Chaudhari
     - **Email:** chitraksonawane@gmail.com
     - **Country:** USA
     - **Company:** Create & Learn LLC, GrandCircus Detroit
     - **Specialization:** Data Science
  2. **Name**: Nuri Öztürk
     - **Email:** ozturknuri8@gmail.com
     - **Country:** Turkey
     - **Specialization:** Data Science
- **Batch Code:** LISUM 17
- **Date:** 2 March 2023
- **Submitted to** Data Glacier
- **Github Repo link:** https://github.com/ChitraChaudhari/Bank-Marketing-campaign-

Problem description

ABC Bank wants to be able to predict which clients are most likely to subscribe to a term deposit. In this way, the bank wants to save time and money by running the marketing campaign more effectively and successfully.

Data understanding

In this project, we have been given 4 datasets:

- Bank
- Bank-full
- Bank-additional
- Bank-additional-full

We started data exploration with the bank dataset, which is the smallest one. And as we examined the data, we decided to drop two features that did not contain specific information for the target column. You can find more information about the features below. The features "default" and "poutcome" were removed. There are no lower outliers in the dataset and the upper outliers in the "balance" and “campaign” features were also removed. For “age” and “duration” features, we applied a logarithmic transformation to eliminate outliers. Our focus now is to understand the information contained in other datasets with additional features and prepare the bank dataset for model building. After creating our first model with the bank dataset, we will try to improve our model with other datasets.

Attribute information:

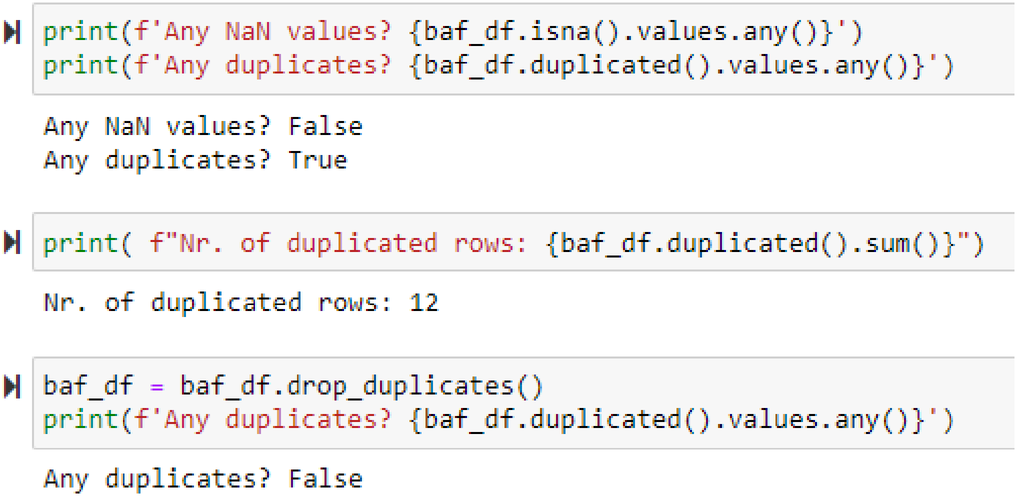
- Input variables:
  - Bank client data:
    - age (numeric)
    - job: type of job (categorical: "admin.", "blue-collar", "entrepreneur", "housemaid", "management", "retired", "self-employed", "services", "student", "technician", "unemployed", "unknown")
    - marital: marital status (categorical: "divorced", "married", "single", "unknown"; note: "divorced" means divorced or widowed)

    - education (categorical: "basic.4y", "basic.6y", "basic.9y", "high.school", "illiterate", "professional.course"," university. degree", "unknown")
    - default: has credit in default? (categorical):
        - 98.2% No
        - 1.8% Yes
        - 0% Unknown
    - housing: has a housing loan? (categorical: "no", "yes", "unknown")
    - loan: has a personal loan? (categorical: "no", "yes", "unknown")
- Related to the last contact of the current campaign:
    - contact: contact communication type (categorical: "cellular", "telephone")
    - month: last contact month of the year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
    - day_of_week: last contact day of the week (categorical: "mon", "tue", "wed", "thu", "fri")
    - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y="no"). Yet, the duration is not known before a call is performed. Also, after the end of the call, y is known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.
- Other attributes:
    - campaign: number of contacts performed during this campaign and for this client (numeric, includes the last contact)
    - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means the client was not previously contacted)
    - previous: number of contacts performed before this campaign and for this client (numeric)
    - poutcome: outcome of the previous marketing campaign (categorical)
        - 81.7% Unknown
        - 10.8% Failure
        - 4.1% Other
        - 3.3% Success
- Social and economic context attributes:
    - emp.var.rate: employment variation rate - quarterly indicator (numeric)
    - cons.price.idx: consumer price index - monthly indicator (numeric)
    - cons.conf.idx: consumer confidence index - monthly indicator (numeric)
    - euribor3m: euribor 3-month rate - daily indicator (numeric)
    - nr.employed: number of employees - quarterly indicator (numeric)
- Output variable (desired target):
    - y - has the client subscribed to a term deposit? (binary: "yes", "no")

Data cleansing and transformation

Duplicate & NaN Observations:

All four datasets were checked for duplicate rows. The bank-additional-full data got 12 duplicate rows. Using drop_duplicate() these duplicate rows dropped entirely. Also, none of the datasets contained NULL values.



```
print(f'Any NaN values? {baf_df.isna().values.any()}')
print(f'Any duplicates? {baf_df.duplicated().values.any()}')
```

```
Any NaN values? False
Any duplicates? True
```

```
print( f"Nr. of duplicated rows: {baf_df.duplicated().sum()}")
```

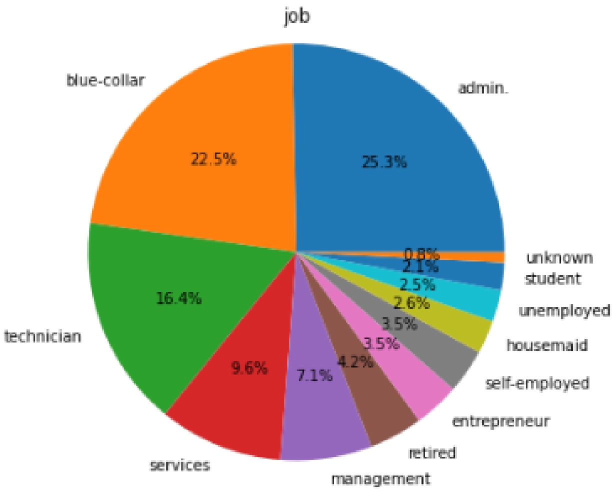
```
Nr. of duplicated rows: 12
```

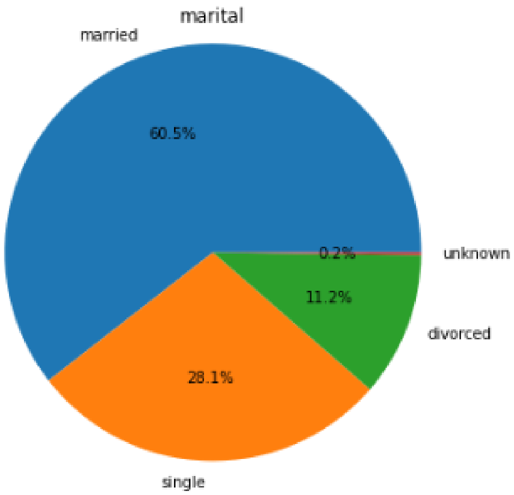
```
baf_df = baf_df.drop_duplicates()
print(f'Any duplicates? {baf_df.duplicated().values.any()}')
```

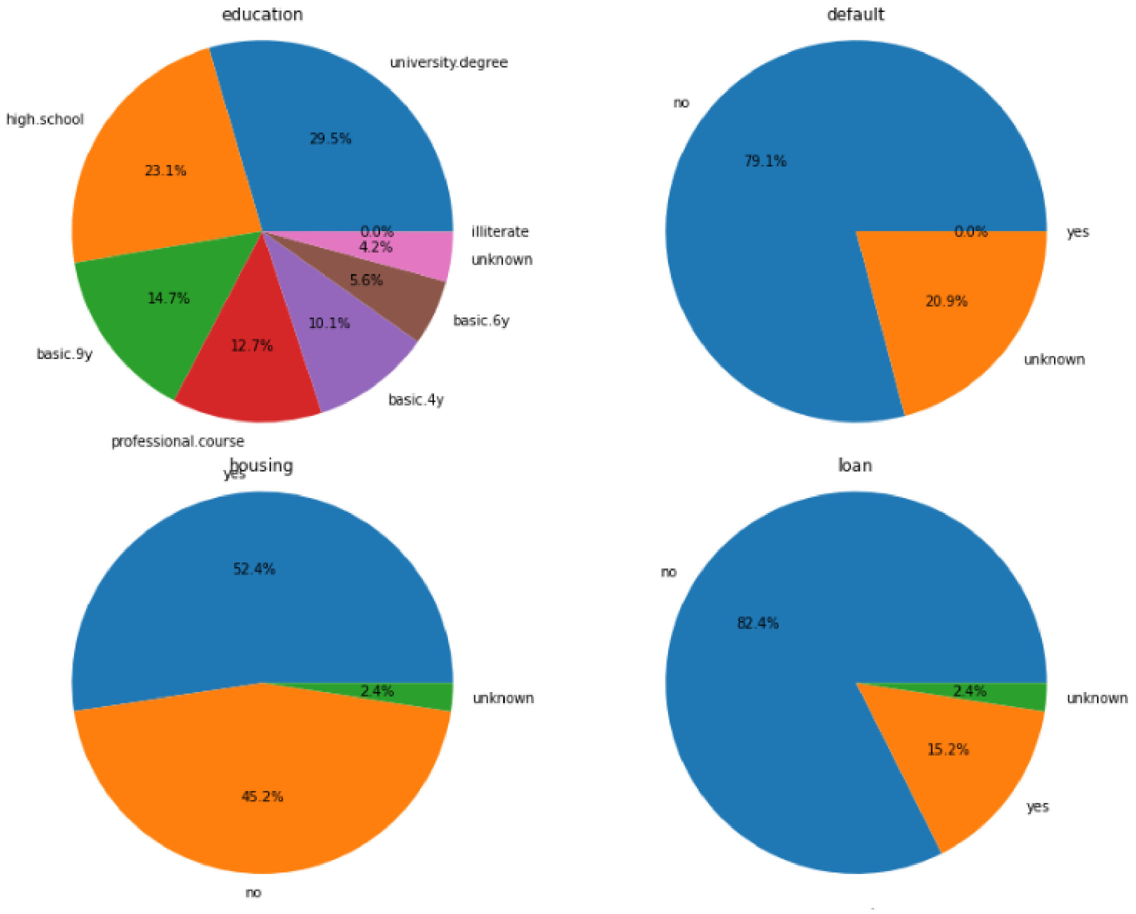
```
Any duplicates? False
```

Missing Attribute Values:

There are several missing values in some categorical attributes, all coded with the "unknown" label. As seen in the pie chart distribution of non-categorical variables, ‘jobs’, ‘marital’, ‘education’,’ default’, ‘housing’, and ‘loan’ contain unknown values.





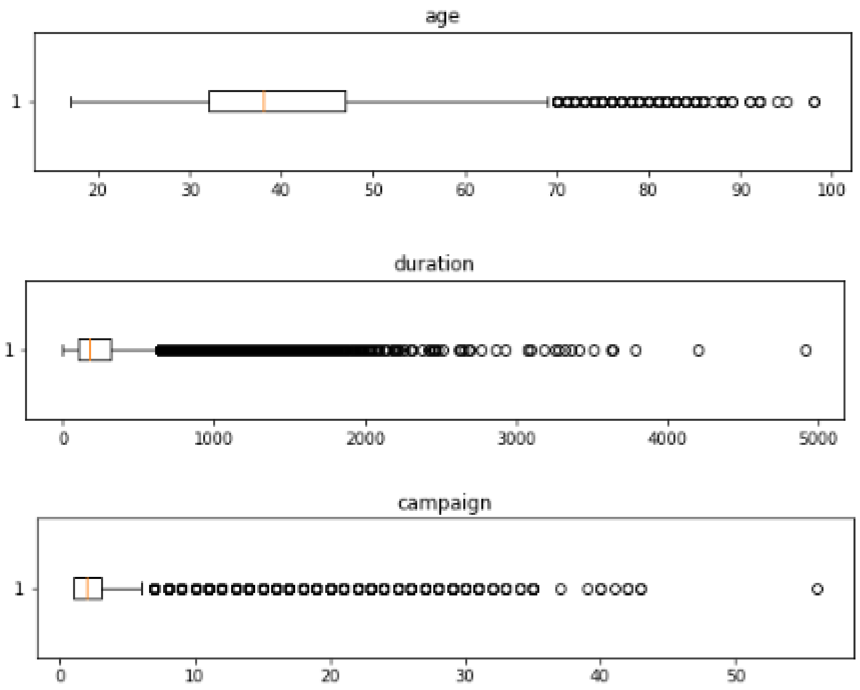


These missing values can be treated as a possible class label or using deletion or imputation techniques. We will be treating them as a Class label, for example, if you see a pie chart for the ‘default’ category there approximately 79% of clients have unknown responses. This may mean these clients do not want the bank to know their actual default status. So keeping this class can be a good addition to analyzing customer behavior.

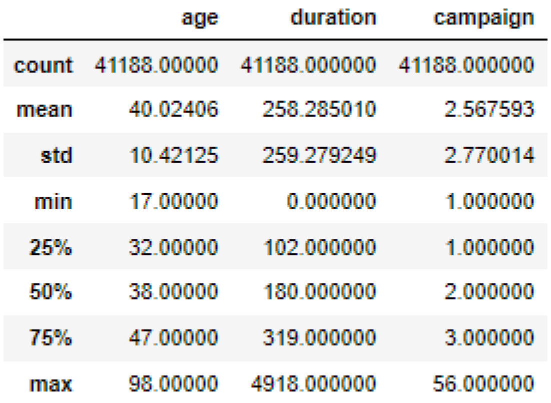
Outliers Detection:

Outliers are the values that are at an abnormal distance from the central distribution of data points.

Let's check for the distribution of numerical variables with the help of a boxplot and see if we can find any kind of outliers. The box plot for the ‘age’, ‘duration’, and ‘campaign’ feature shows there is a significant number of outliers in distribution.



For outliers treatment, we can totally drop the outliers from the dataset but that could lead to the loss of some important data, which can help in giving predictions from the model. We can further look at the statistical distribution of these variables to decide further.



	age	duration	campaign
count	41188.00000	41188.000000	41188.000000
mean	40.02406	258.285010	2.567593
std	10.42125	259.279249	2.770014
min	17.00000	0.000000	1.000000
25%	32.00000	102.000000	1.000000
50%	38.00000	180.000000	2.000000
75%	47.00000	319.000000	3.000000
max	98.00000	4918.000000	56.000000

As mentioned in the data description we will be dropping the duration column entirely. The maximum values for ‘age’ and ‘campaign’ are 98 and 56 respectively, and they don't seem unrealistic. Therefore the outliers in the data distribution of these features don't need to be dropped.