Data Glacier

Your Deep Learning Partner

Data Understanding

Virtual Internship

Felipe I. Crespo

Agenda

Data Understanding

Data Related

Features types

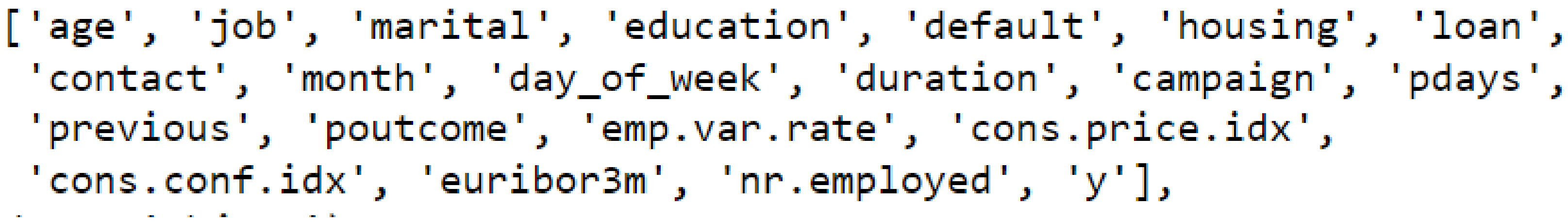
Problems

Problems Approach

Data Understanding

Features are divided into 4 groups:

1. Data related to clients (age, job, marital, education, default, housing and loan).
2. Data related with the last contact of the current campaign (contact, month, day_of_week and duration).
3. Data related with campaigns (campaign, odays, previous and poutcome).
4. Data related with socio economic context attributes (emp.var.rate, cons.Price.idx, cons.conf.idx, euribor3m, nr.employed).
5. TARGET: "y".



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

Data related to clients

1) age	(numeric)
2) job type of job	(categorical: 'admin.","blue- collar','entrepreneur','housemaid’, 'management', 'retired', 'self- employed', 'services', 'student’, 'technician','unemployed', 'unknown’)
3) marital: marital status	(categorical: 'divorced','married', 'single', 'unknown'; note: 'divorced' means divorced or widowed)
4) education	(categorical:'basic.4y', 'basic.6y','basic.9y', high school','illiterate’, 'professional.course','university.degree', 'unknown’)
5) default: has credit in default?	(categorical: 'no', 'yes', 'unknown’)
6) housing: has housing loan?	(categorical: "no', 'yes', 'unknown’)
7) loan: has personal loan?	(categorical: 'no','yes', 'unknown')

Data related to last contact

8) contact: contact communication type	(categorical: 'cellular', 'telephone’)
9) month: last contact month of year	(categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')10) day_of_week: last contact day of the week (categorical: 'mon', 'tue', 'wed', 'thu', 'fri’)
11) duration: last contact duration, in seconds	(numeric).

Data related to the campaing

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

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

14) previous: number of contacts performed before this campaign and for this client (numeric)

15) poutcome: outcome of the previous marketing campaign (categorical: 'failure', 'nonexistent','success')

Data related to social and economic context

16) emp.var.rate: employment variation rate - quarterly indicator (numeric)

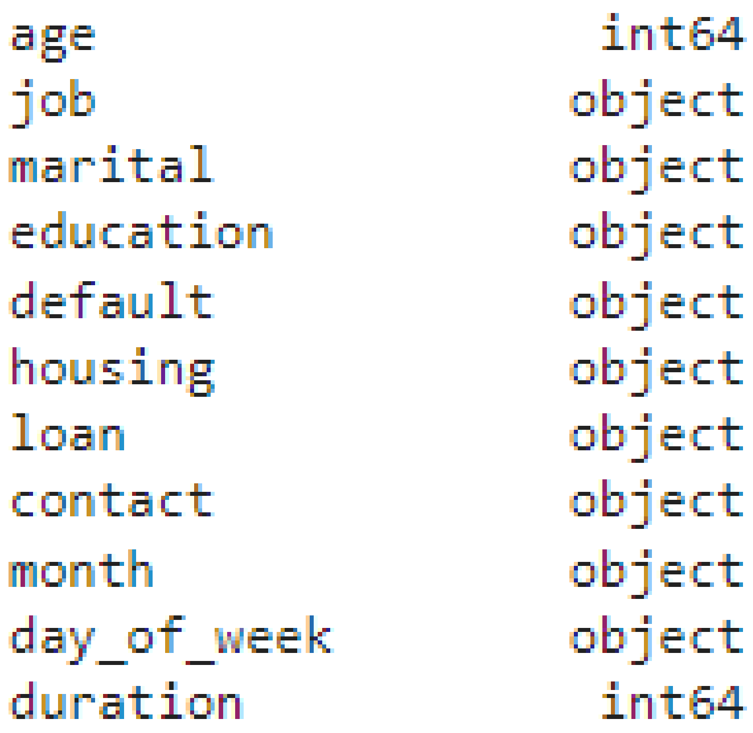
17) cons.price.idx: consumer price index - monthly indicator (numeric)

18) cons.conf.idx: consumer confidence index - monthly indicator (numeric)

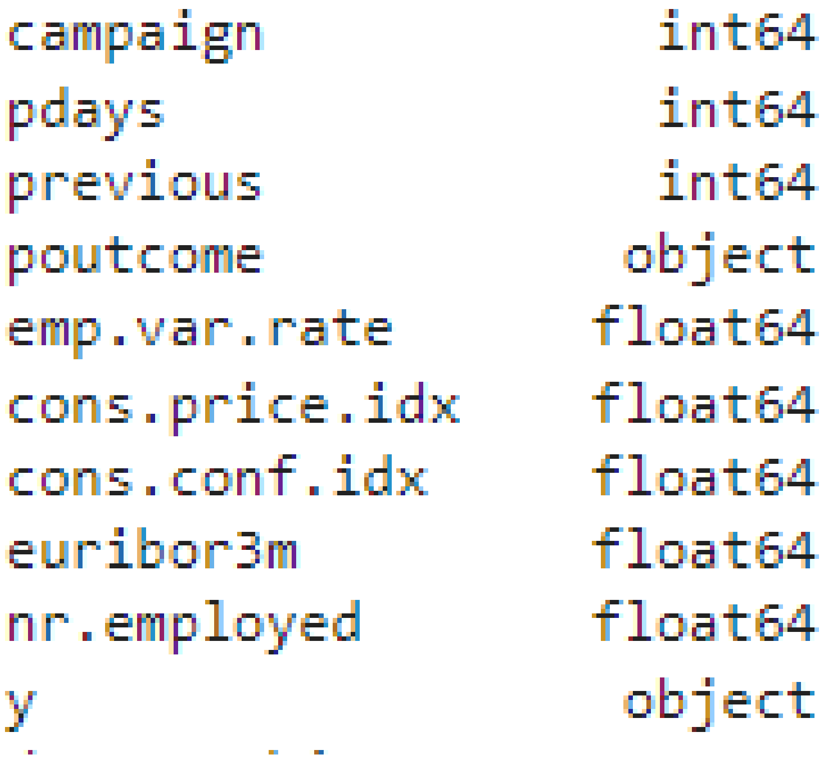
19) euribor3m: euribor 3-month rate - daily indicator (numeric)

20) nr.employed: number of employees - quarterly indicator (numeric)

Features types



```
age               int64
job              object
marital          object
education        object
default          object
housing          object
loan             object
contact          object
month            object
day_of_week      object
duration          int64
```



```
campaign            int64
pdays               int64
previous            int64
poutcome           object
emp.var.rate      float64
cons.price.idx    float64
cons.conf.idx     float64
euribor3m         float64
nr.employed       float64
y                  object
```

Problems

Some features in the data contain the string "unknown", which will be treated as null values. The affected features and their corresponding percentages are as follows: job (0.8%), marital (0.2%), education (4.2%), default (20.9%), housing (2.4%), and loan (2.4%).

The target variable, "y", is imbalanced, with 89% of the values being "no" and 11% being "yes".

The data contains a significant number of outliers in the features.

The feature "duration" is left-skewed, but it should be used with caution because the outcome variable is known after the end of the call. As such, it may not be a useful predictor. However, for now, the skewed nature of the feature is not a concern.

Problems Approach

To handle null values in the data, we will explore two options: imputation and deletion, and determine which approach is best for the model. Outliers will be treated with IQR and WOE.

For the unbalanced target variable, "y", we will consider two approaches. The first approach is passive, in which we will not generate anything new, but instead use the appropriate metrics for the model to avoid the impact of the imbalance (such as recall, F1 score, ROC, etc.).

The second approach involves applying SMOTE (Synthetic Minority Over-sampling Technique), which is a common solution for this type of problem in machine learning classification mode

Thank You

Data Glacier

Your Deep Learning Partner