Data Glacier

Your Deep Learning Partner

Week 8 deliverables

Bank Marketing (Campaign) - Group Project

Group Name - Bloodhounds,
Batch code - LISUM09,
Specialization: Data science.
Group member details:

- Name - Margarita Prokhorovich,
- email - marusya15071240@gmail.com,
- Country – Thailand,
- Submission date – 25 June, 2022

Problem description

Statement

- ABC Bank wants to sell it's 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).

- 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 save resource and their time (which is directly involved in the cost (resource billing))[1].

Data set problem statement

- A big part of customers are convinced of the effectiveness of an individual approach to service. In an Accenture Financial Services global study of nearly 33,000 banking customers spanning 18 markets, 49% of respondents indicated that customer service drives loyalty. By knowing the customer and engaging with them accordingly, financial institutions can optimize interactions that result in increased customer satisfaction and wallet share, and a subsequent decrease in customer churn[2].

- One of the challenges the banks encounter, is following:
How does a bank figure out what its customers specifically think about its services? Are their issues getting resolved? How satisfied are they with the experience? Why can't banks analyze customer care sessions to find real-time information about the customers and their pressing issues? Customer care records are very pointed and specific about the challenges the customer faces. In these cases building a model and detect patterns can help to improve customers' retaining and loyalty and reduce the churn[3].

1. Problem Statement. Data Science:: Bank Marketing (Campaign) -- Group Project. Data Glacier, URL
2. Top 10 Banking Industry Challenges — And How You Can Overcome Them. Hitachi Solutions, URL
3. Why Retaining Customers For Banks Is As Important As Winning New Ones. Forbes, URL

Data understanding

Characteristics

- Data Set Characteristics: Multivariate
- Attribute Characteristics: Real
- Associated Tasks: Classification
- Area: Business
- Date Donated: 2012-02-14

Source

- [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

Description

- The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed[4].

Goal

- The binary classification goal is to predict if the client will subscribe a bank term deposit (variable y).

Data set used

- bank-additional-full.csv with all examples, ordered by date (from May 2008 to November 2010).

- This data set is an updated bank-full.csv data set. The data is enriched by the addition of five new social and economic features/attributes (national wide indicators from a ~10M population country), published by the Banco de Portugal.

- It was found that the addition of the five new social and economic attributes (made available here) lead to substantial improvement in the prediction of a success, even when the duration of the call is not included.

4. Bank Marketing Data Set. UCI Machine learning repository, URL

Type of data

Categorical features

bank client data:

- job : type of job
- marital : marital status
- education
- default: has credit in default?
- housing: has housing loan?
- loan: has personal loan?

related with the last contact of the current campaign:

- contact: contact communication type
- month: last contact month of year
- day_of_week: last contact day of the week

other attributes

poutcome: outcome of the previous marketing campaign

Output variable - y - has the client subscribed a term deposit? (binary)

Numeric features

bank client data:

- age

related with the last contact of the current campaign:

- duration: last contact duration, in seconds

other attributes

- campaign: number of contacts performed during this campaign and for this client
- pdays: number of days that passed by after the client was last contacted from a previous campaign
- previous: number of contacts performed before this campaign and for this client
- poutcome: outcome of the previous marketing campaign

social and economic context attributes

- emp.var.rate: employment variation rate - quarterly indicator
- cons.price.idx: consumer price index - monthly indicator
- cons.conf.idx: consumer confidence index - monthly indicator
- euribor3m: euribor 3 month rate - daily indicator
- nr.employed: number of employees - quarterly indicator[5].

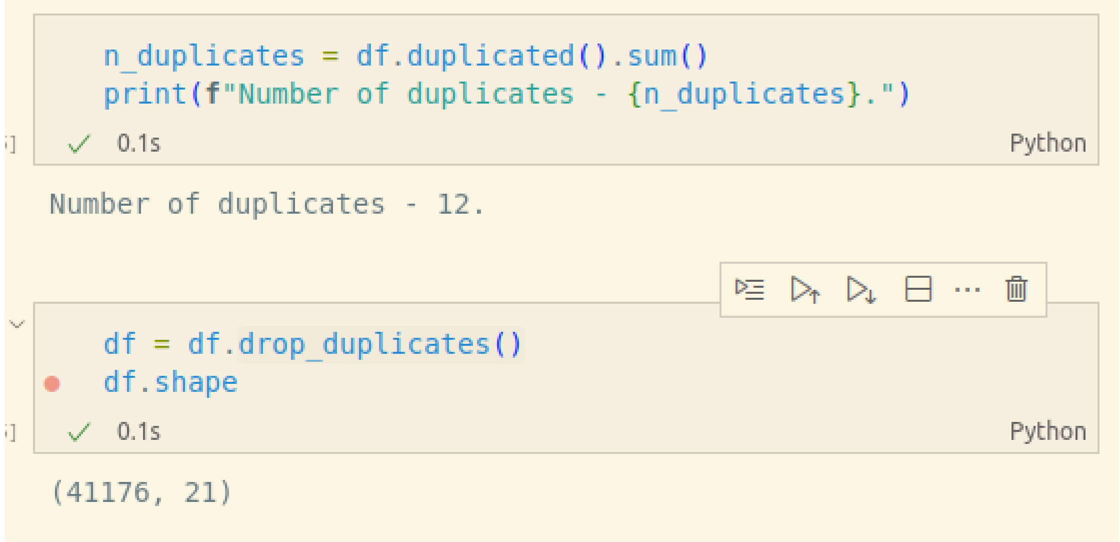
[5] Bank Marketing Data Set. UCI Machine learning repository, URL

Problems in the data and possible approaches

Approaches

Duplicate values

- Since we have duplicates in our data set, we need to delete them.
- We have 12 duplicates and remove them using drop_duplicates() method. This method deletes complete duplicates from the data set.



```python
n_duplicates = df.duplicated().sum()
print(f"Number of duplicates - {n_duplicates}.")
```

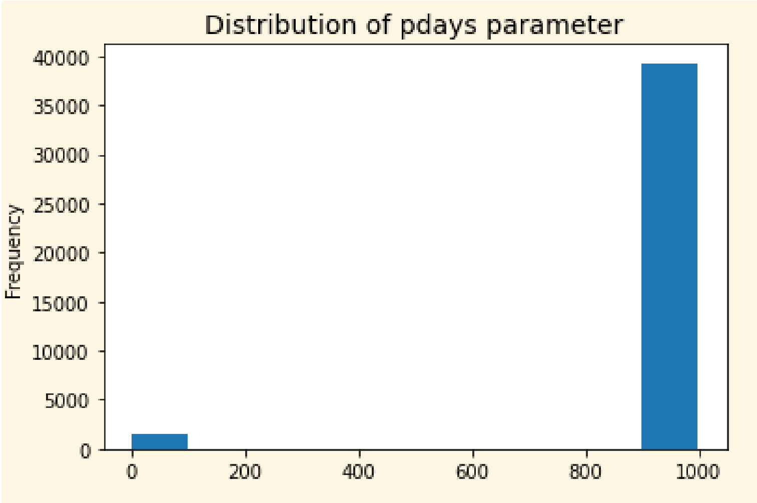
```
Number of duplicates - 12.
```

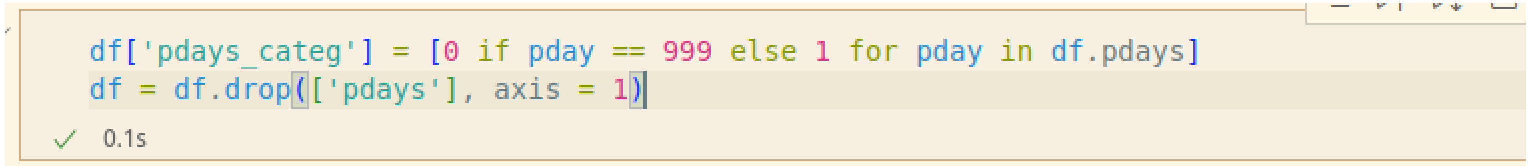
```python
df = df.drop_duplicates()
df.shape
```

```
(41176, 21)
```

Pdays parameter values

Pdays parameter has a lot of 999 values. 999 means client was not previously contacted. Since the variable is numeric, it can affect the interpretation of the model. Replacing 999s with 0 is also not effective since interpretation can be wrong - 0 days passed by after the client was last contacted from a previous campaign. Therefore it was decided to move from numeric variable to a binary one. It's also reasonable because except for 999s, another values lie in not large range.





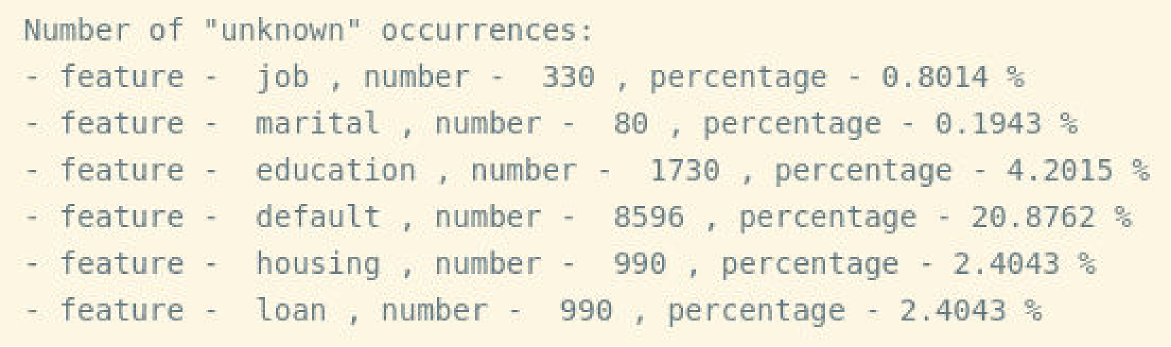
```python
df['pdays_categ'] = [0 if pday == 999 else 1 for pday in df.pdays]
df = df.drop(['pdays'], axis = 1)
```

Problems in the data and possible approaches

Approaches

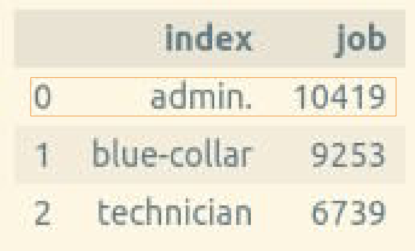
‘Unknown’ values

- Data set has no null values but some categorical values have values marked ‘unknown’. Following options can be considered:
- Delete all the rows with ‘unknown’ values or delete them only in certain columns
- Populate ‘unknown’ values with a major category.



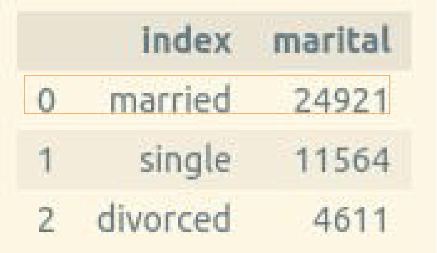
```
Number of "unknown" occurrences:
- feature -  job , number -  330 , percentage - 0.8014 %
- feature -  marital , number -  80 , percentage - 0.1943 %
- feature -  education , number -  1730 , percentage - 4.2015 %
- feature -  default , number -  8596 , percentage - 20.8762 %
- feature -  housing , number -  990 , percentage - 2.4043 %
- feature -  loan , number -  990 , percentage - 2.4043 %
```

The suggested approach is to populate the variables with the most common category. The exception is the default variable, since the proportion of unknowns is quite large, we will consider ‘unknown’ as a separate category.



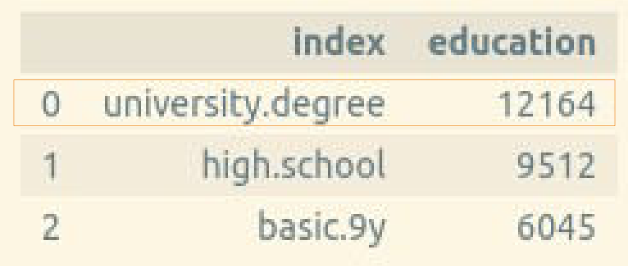
	index	job
0	admin.	10419
1	blue-collar	9253
2	technician	6739

Major category for job is admin.



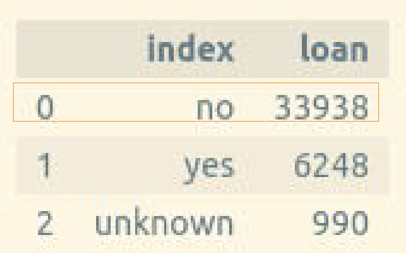
	index	marital
0	married	24921
1	single	11564
2	divorced	4611

Major category for marital is married.



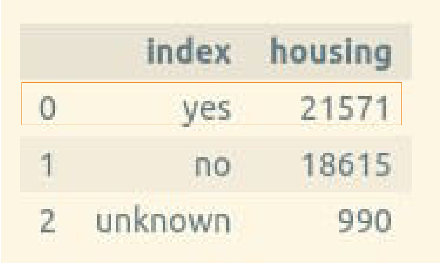
	index	education
0	university.degree	12164
1	high.school	9512
2	basic.9y	6045

Major category for education is university degree.



	index	loan
0	no	33938
1	yes	6248
2	unknown	990

Major category for loan is no (no loan).



	index	housing
0	yes	21571
1	no	18615
2	unknown	990

Major category for housing is yes (has housing credit).

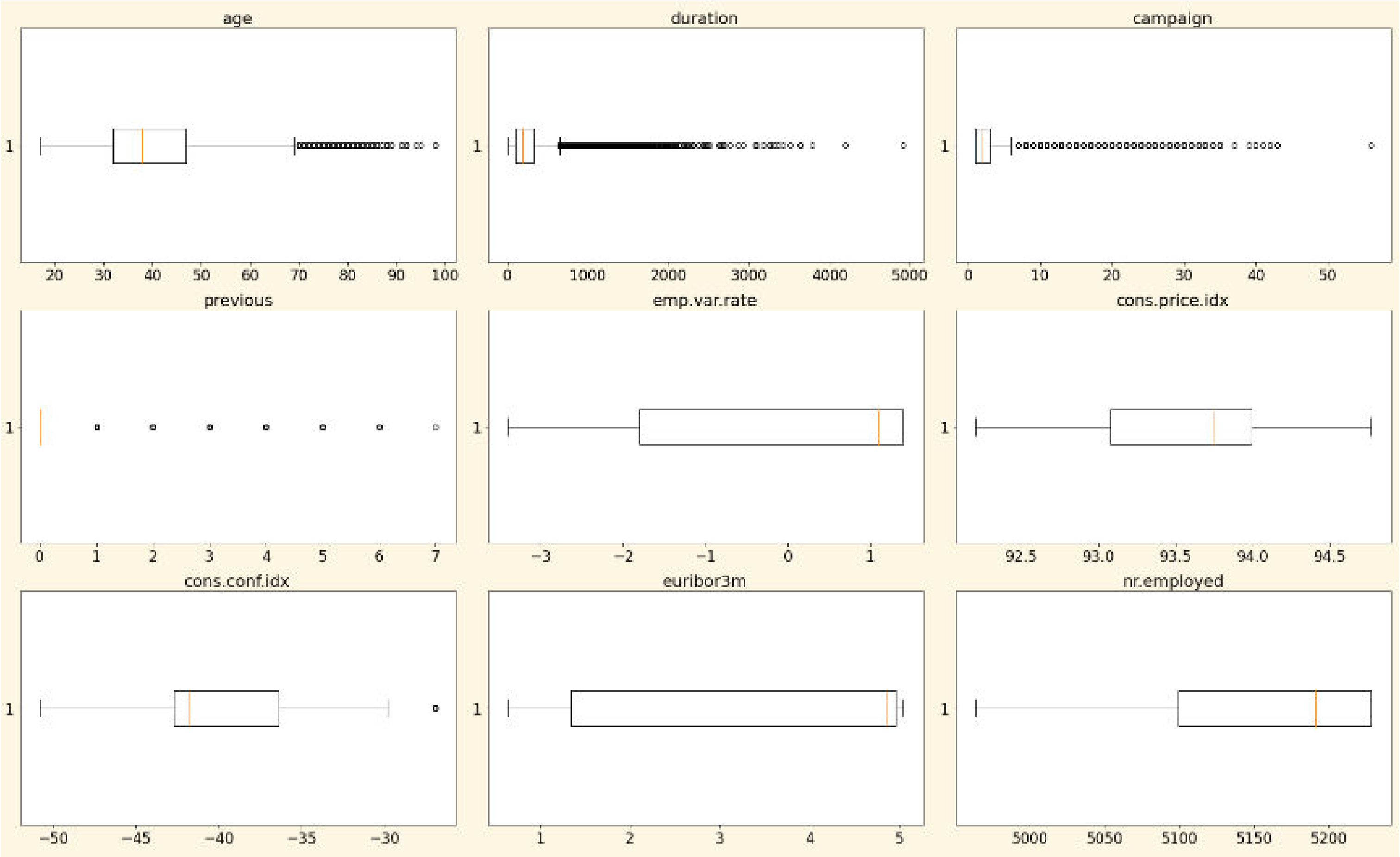
Problems in the data and possible approaches

Approaches

Outliers

- Since we detected outliers in some variables, we can either keep or remove them.
- Usually outliers cannot be removed without analyzing.
- We keep the outliers in age variable cause removing the oldest clients will affect the customer base understanding.
- We keep outliers in duration, campaign and previous cause they are just technical parameters and these outliers aren’t related to specific customer groups.
- Also we keep an outlier in consumer confidence index also for a reason that removing a customer with a higher confidence index will not display diversity of the customers.

Outliers graphs



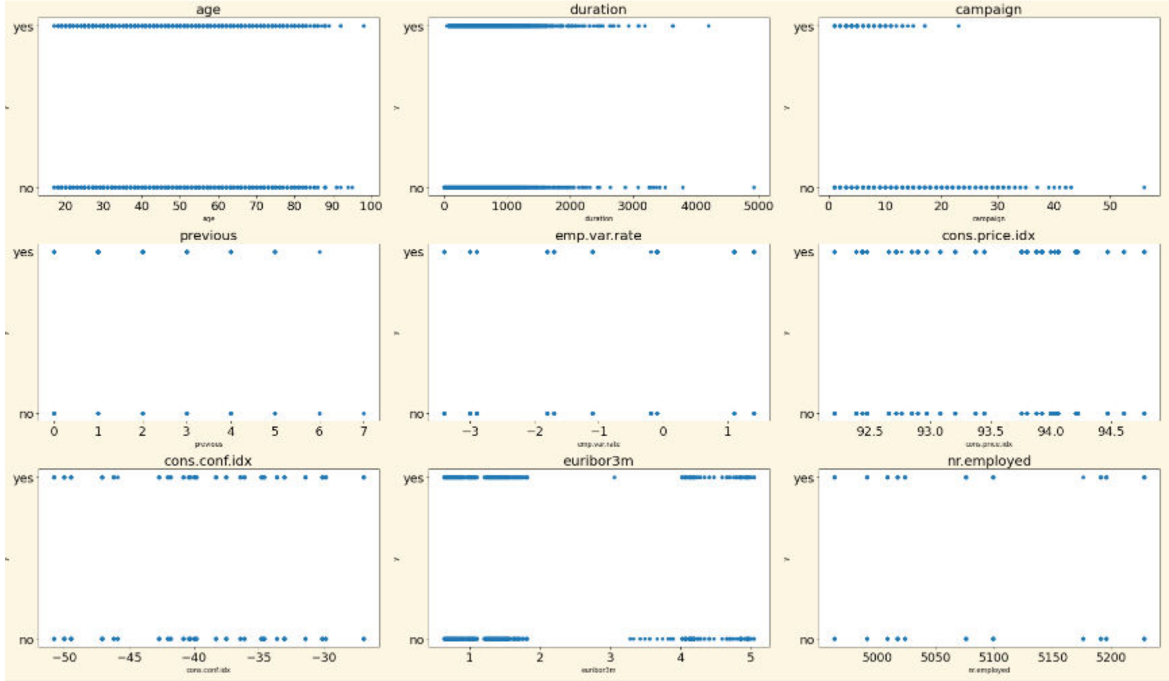
Problems in the data and possible approaches

Approaches

Outliers (continuation)

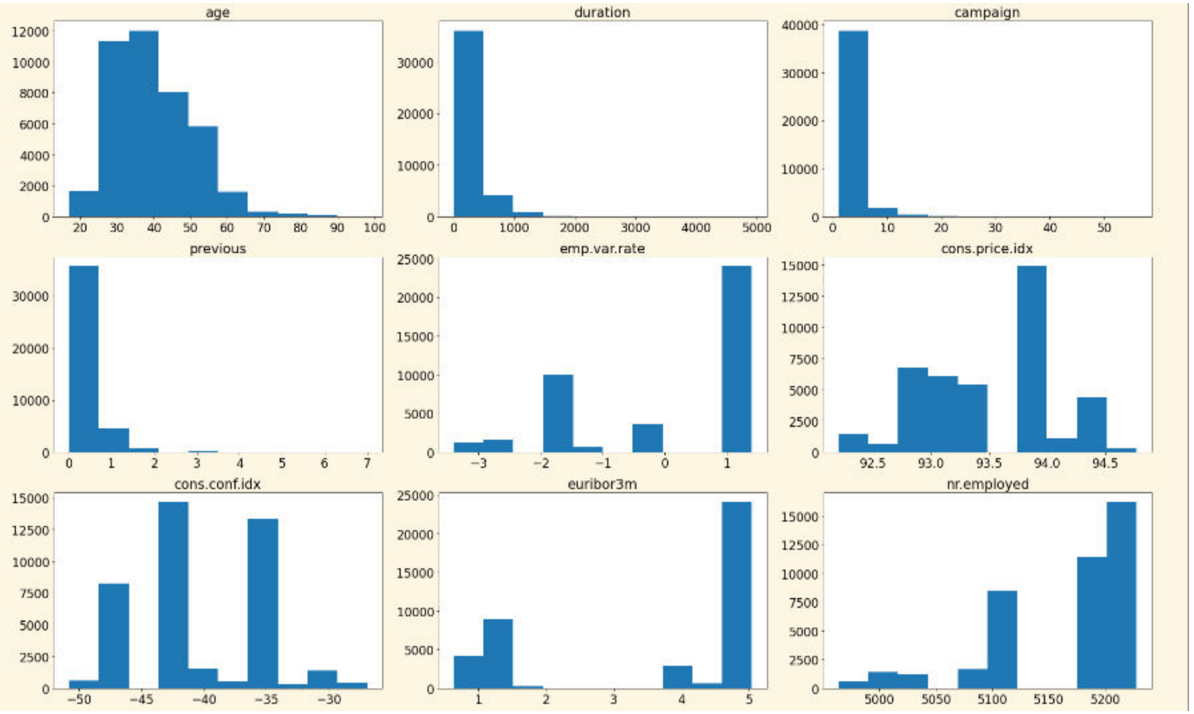
- Moreover, we can see that when plotting corresponding numeric variable and y variable, there’s no outliers that lie very far.

Variables values with respect to output variable



- We will try additional approaches of cleaning the data next week.

Variables distributions



Skewed distribution

- Each distribution is quite far from a bell shape the normal distribution has. We could transform some of them to normal by applying log function. However, it becomes harder to interpret the results, so we decided to keep the variables distributions in initial condition. This should be done as a last resort when the predictive power of the model is very poor.

Bank Marketing (Campaign) - Group Project

Data Glacier

Your Deep Learning Partner

Thank You