Data Glacier

Your Deep Learning Partner

Week #8 Deliverables

Team member details:

Group Name: Sophie The Scientist

Sophonie Sidrac, sophiesidrac@gmail.com, USA, Tennessee State University, Data Science

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).

Data Understanding

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. The data set involves customer profiles showing their age, job, marital status, education, housing, loan, and more.

What type of data you have got for analysis?

- **age** (**numeric)**
- **job** : type of job (**categorical**)
- **marital** : marital status (**categorical**)
- **education** (**categorical**)
- **default**: has credit in default? (**categorical**)

- **balance**: monetary balance in account (**numeric**)
- **housing**: has housing loan? (**categorical**)
- **loan**: has personal loan? (**categorical**)
- **contact**: contact communication type (**categorical**)
- **day**: last contact day of the week (**numeric**)
- **month**: last contact month of year (**categorical**)
- **duration**: last contact duration, in seconds (**numeric**).
- **campaign**: number of contacts performed during this campaign and for this client (**numeric**)
- **pdays**: number of days that passed by after the client was last contacted from a previous campaign (**numeric**)
- **previous**: number of contacts performed before this campaign and for this client (**numeric**)
- **poutcome**: outcome of the previous marketing campaign (**categorical**)
- **y** - has the client subscribed a term deposit? (**Boolean**)

What are the problems in the data (number of NA values, outliers , skewed etc)

- The data is semicolon-delimited instead of comma-delimited.
- Null values are noted as ‘unknown’ in data.
- The Booleans in the data are represented as y or n
- There is an imbalance in data. The distribution is heavily skewed to ‘no’. This could make the model inaccurate.
- The numerical columns are extremely positively skewed.
- There are also outliers present in one of the feature columns.

What approaches are you trying to apply on your data set to overcome problems like NA value, outlier etc and why?

- When reading csv file, (delimiter=’;’ and quotechar= ‘ “ ‘) must be applied to reload the data with the correct delimiter.
- To resolve the boolean issue, the data needs to be mapped to binary data as ‘1’ and ‘0’.

- To balance the skewness of labels in data, create a new data frame with a healthy balance of labels of the target variable.
- To resolve the positive skewness in numerical columns, transformative methods need to be applied.
- Lastly, I can use imputation methods to replace outliers with an estimated value.

Data intake Report:
Name: Bank Marketing Campaign (Data Science)
Report date: 03/14/2025
Internship Batch: LISUM41
Version:1.0
Data intake by: Sophonie Sidrac
Data intake reviewer: Sophonie Sidrac
Data storage location: https://github.com/1Sophani/DataGlacier-Internship/tree/main/Week%208

Tabular data details:

Total number of observations	42511
Total number of files	1
Total number of features	17
Base format of the file	.csv
Size of the data	4.4 MB

Submitted by: Sophonie Sidrac
Submitted to: Data Glacier
Submission Date: 03/14/25