Virtual Internship
Data Science
Data Intake Report

Group Name: LISUM19: Data science Group 1

Members:

No	Name	Email	Country	Specialization
1	Preeti Verma	vermapreeti.dataanalyst@gmail.com	Canada	Data Science
2	Thanuja	thanujayadav953@gmail.com	UK	Data Science
3	Abishek James	abishekjames1998@gmail.com	Ireland	Data Science

Name: Bank Marketing (Campaign)

Report date: 19-04-2023

Internship Batch: LISUM19

Data intake by: Preeti Verma

Data intake reviewer: Data Glacier

Data storage location: GitHub

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 in understanding whether a particular customer will buy the product or not (based on the customer's past interaction with the bank or other Financial Institution). This is an application of the company's marketing data.

Business Understanding:

The goal is to build a Machine Learning model that helps in predicting the outcomes of each customer's marketing campaign and analyzing which features impact the outcomes will help the company understand how to make the campaign more effective. Additionally, categorizing the customer group that subscribed to the term deposit helps to determine who is more likely to purchase the product in the future, thereby developing more targeted marketing campaigns.
This can be accomplished by using an ML model that shortlists the customers with a higher possibility of purchasing the product. So marketing such as telemarketing, SMS or email marketing can concentrate only on those customers. It will save time and resources by doing this.

Project Lifecycle

Deadline (Date/week)	**Plan and Deliverables**
19 April 2023(Week 7)	• Problem statement • Business understanding • Dataset collection
26 April 2023(Week 8)	• Data understanding • Data analysis - finding null values, and outliers. • Data processing
2 May 2023(Week 9)	Data cleaning and transformation
9 May 2023(Week 10)	EDA and Model Recommendation
16 May 2023(Week 11)	EDA Presentation and Proposed Modeling Technique

23 May 2023(Week 12)	Model Selection and Building the Model
30 May 2023(Week 13)	Final project report and code submission

Data Understanding:

- We were provided with 2 files namely bank_additional_full.csv, and bank_additional.csv
- The second dataset is a sample of the first dataset
- The data cover a period from May 2008 to November 2010.
- The 1st dataset (bank_additional_full.csv) contains 41188 records & 21 columns, whereas the 2nd dataset (bank_additional.csv) has 4119 records and 21 columns.
- Columns are not uniformed named for example "day_of_week", and "emp.var.rate".
- This needs to be modified for making it easier to work with
- Some of the columns' names were changed for easy understanding.
- All variables in both datasets have the right datatypes and all variable's unique values are fine.
- The mean age of the respondents is 40 years, the minimum age is 17years, and the maximum age is 98years
- The mean duration of the time taken talking to the respondent is 258.3 seconds, and the minimum time is 0 seconds. This will be discarded as it does not add value to our analysis
- The minimum number of contacts during this campaign was 1 contact and the maximum was 56 contacts
- pdays variable value 999 means the lapsed days before the person was conducted thus we will make this value to be zero.
- Some variables are right-skewed (e.g "age" variable) while others are left-skewed (e.g "no_emp")variables.
- Our aim is to predict whether the customer will buy "y" product or note based on other variable details.

Problems in Dataset:

- There are 12 duplicates values in the original dataset, we dropped them since they are minimal.
- The target variable is an imbalanced class, the "no" class has more observation than the "yes" class in both datasets. We will impute it to balance during modeling.
- Missing values were recorded as Unknown, we changed them to NaN values.

- The variable “job”, “marital”, “education”, “default”, “housing”, and “loan” has missing values.
- The missing values were imputed as follows
  a. “job”, marital, default variables were imputed by the mode of the column which is a categorical column.
  b. “education” variable was imputed with “N/A”(not applicable), this means we assumed the respondents didn’t have any formal education and thus were not applicable.
  c. “housing”, and “loan” variable missing values were dropped since they were minimal and will not affect our analysis.
- “age”, “duration”, “campaign”, “pdays”, “previous” and “cons.conf.idx” have outliers.
- We removed the outlier in the age variable since it was the only one at the extreme end with was altering the variable mean. Outliers on other variables we decided not to remove since they look genuine.
- We dropped the “contact”, and “duration” columns, they are not going to add value to our Analysis.
- Our columns have different formats, we changed them to standard form.