Week 8: Data Understanding

Group Name: Pattern Pros

Github Repository Link:

[DataGlacier/Group\_Project at main · danielkingswood/DataGlacier (github.com)](https://github.com/danielkingswood/DataGlacier/tree/main/Group_Project)

Team Member Details:

| Name | Email-ID | Country | University | Specialization |
| --- | --- | --- | --- | --- |
| Jay Panara | jay.panara@gmail.com | Canada | University of Waterloo | Data Science |
| Shreya Dwivedi | shreyad@usc.edu | USA | University of Southern California | Data Science |
| Sarah Sindeband | ssindeband2018@fau.edu | USA | Florida Atlantic University | Data Science |
| Daniel Kingswood | ddk727@gmail.com | UK | University of Bristol | 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 helps 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:

Data Source:

[Moro et al., 2011] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimarães, Portugal, October, 2011. EUROSIS.

bank-full.csv:

This data set is ordered by date and is broken down into four sections. The first section has columns pertaining to the specific client. The second section of columns has information related to the last contact of the current campaign. The third section of columns have information pertaining to previous marketing campaigns. The final column is whether or not the client subscribed to the term deposit (output variable).

bank.csv:

This file is 10% randomly selected from the bank\_full.csv file set aside for testing a

machine learning model.

Data type for analysis:

| **Column names** | **Data type** |
| --- | --- |
| age | numeric |
| job | categorical |
| martial | categorical |
| education | categorical |
| default | binary |
| balance | numeric |
| housing | binary |
| loan | binary |
| contact | categorical |
| day | numeric |
| month | categorical |
| duration | numeric |
| campaign | numeric |
| pdays | numeric |
| previous | numeric |
| poutcome | categorical |
| y | binary |

Problems in data:

* 5 columns have skewed values (balance, duration, campaign, pdays and previous)
* 4 columns have unknown values, 2 with large proportion (job, education, contact, poutcome)
* Outliers
* No client ID number

Approaches:

| Problem | Approaches | Why? |
| --- | --- | --- |
| Skewed values/ imbalanced dataset | Remove outliers, log transformation, normalize values to help balance the data. | The Tail region can act as an outlier for regression based models and cause a bias in the model. |
| Unknown values | For columns with a small amount of unknown values they could be removed.  For columns with a large amount of unknown values, imputation methods could be used or predict the missing values using either a regression or classification model. | Missing data or unknown values can lead to a reduced size of the data which leads to less efficient estimates from the model. Also if the values are left as unknown the model could find patterns between unknown values which is not helpful for accurate predictions. |
| No client identification number | Check for duplicate entries. This can be done by comparing each line, and if the line is exactly the same as another, it could be a duplicate entry. | To avoid duplicates |
| Outliers | Remove outliers, check if outliers are logical, or do further statistical tests to verify the outliers | Outliers can cause a decrease in normality(skewed data), cause a bias in models, have a significant impact on mean and standard deviation of data and can also cause problems during statistical analysis |