Data Glacier Project
Data Science: Bank Marketing (Campaign)

Team members:

Group	Names	E-mail	Country	College
DataWizz	Akhil Abraham	akhilabrahamuk@gmail.com	UK	Queen's University Belfast
	Pravallika Sheshabhatter	sbpravallika3@gmail.com	UK	
	Shiva Ramezani	shvramezani@gmail.com	US	Cal State Northridge

Problem Description:

The problem at hand is that ABC Bank wants to sell its term deposit product to customers. However, before launching the product, they want to develop a model that can predict whether a particular customer is likely to buy the product based on their past interactions with the bank or other financial institutions. The goal is to assist the bank's marketing efforts by targeting customers who are more likely to purchase the product. By focusing their marketing channels, such as telemarketing, SMS, or email marketing, on these potential customers, the bank can save resources, time, and costs associated with marketing to customers who are less likely to subscribe to the term deposit.

GitHub Repo link:

https://github.com/ShivaRamezani/Capstone/tree/master

Date:

07/02/2023

Cleansing and transformation:

Steps taken:
-Checking for NULL value: No Null value was found
-Distribution visualization
-Finding correlation between variables
-Detecting for outliers and removing the outliers using OLS regression.
-Handling the 'unknown' and 'nonexistent' values
-Balancing the dataset
-LabelEncoding the dataset

Observations:

Some relevant observations found were:

- No null values but "unknown" and "nonexistent" values are present which can be considered as missing values.
- Major customers fall under the age of 17-54.
- Longer duration of the call, higher chances of success in campaigns.
- Outliers were found in numeric data and using the OLS model we removed all the outliers
- Some values were not relevant, we might need to change them so we could change them to meaningful data (e.g. '999' meant the customers' were not contacted previously to '0').
- More insight can be found by performing bivariate or multivariate analysis.
- Imbalanced data can be solved by sampling the data.

Model Selected to predict whether the customers will subscribe to the term deposit or not:

- **Logistic Regression:** This model is easy to understand, can be trained rapidly, and offers highly interpretable results. Nonetheless, it presupposes a linear association between features and the log-odds of the target variable, which may not always be accurate.
- **Random forest:** An ensemble technique that improves decision tree performance, exhibits reduced overfitting tendencies, and excels in handling non-linear associations. Nevertheless, comprehending them could be somewhat more difficult, and their training duration might be longer.
- **XGBoost:** These models exhibit high accuracy, effective handling of non-linear relationships, and the ability to manage missing data. Nevertheless, they necessitate lengthier training periods, boast higher complexity in interpretation, and demand meticulous parameter tuning.
- **Feedforward Neural Networks**: Suitable for tabular data where features are independent of each other, like in traditional classification or regression problems.