Data Glacier Internship Final Project

Project : Bank Marketing

Name : Apoorv Ajay Thite

University : The Pennsylvania State University

Email : apoorvthite21@gmail.com

Country : United States of America

Specialisation : Data Science

Internship Batch : LISUM34

Date : 23rd July, 2024

Problem 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). 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 classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

ML Problem

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)). Develop model with Duration and without duration feature and report the performance of the model. Duration feature is not recommended as this will be difficult to explain the result to business and also it will be difficult for business to campaign based on duration.

My Approach

To address ABC Bank's objective of predicting whether a customer will subscribe to a term deposit based on past interactions, we will follow a structured machine learning approach. First, we will perform an exploratory data analysis (EDA) to understand the dataset, identify key features, and uncover any underlying patterns. We will then preprocess the data, handling missing values, encoding categorical variables, and normalizing numerical features to ensure the data is in an optimal format for modeling. Given the requirement to develop models with and without the 'duration' feature, we will create two versions of the dataset. We will train multiple classification algorithms, such as Logistic Regression, Decision Trees, Random Forest, and Gradient Boosting, to evaluate their performance on both versions of the dataset. Model performance will be assessed using metrics like accuracy, precision, recall, F1-score, and AUC-ROC to ensure a comprehensive evaluation. Additionally, we will employ techniques such as cross-validation and hyperparameter tuning to enhance model robustness and prevent overfitting. The model excluding the 'duration' feature will be particularly emphasized, given the bank's preference for more interpretable and actionable results. Finally, we will compare the performance of both models, providing insights into the impact of the 'duration' feature and making recommendations based on the model outcomes to optimize the bank's marketing strategies.

Project Lifecycle

- Week 7 : 19th July, 2024 ——————> Business Understanding
- Week 8 : 26th July, 2024 ——————> Data Understanding
- Week 9 : 2nd August, 2024 -——————> EDA
- Week 10 : 9th August, 2024 ——————-> Feature Engineering
- Week 11 : 16th August, 2024 ——————-> EDA Presentation and Modelling
- Week 12 : 23rd August, 2024 ——————-> Model Selection
- Week 13 : 30th August, 2024 ——————-> Final Project and Code