Data Science Internship at Data Glacier

Week 12

Project: Bank Marketing (Campaign)

Team Name: Data Science UVS

Name: Uday Vikram Singh
Email: udaykymset@gmail.com

Country: Scotland

University: University of Edinburgh

Specialization: Data Science

Batch Code: LISUM17

Date: 20/03/2023

Submitted to: Data Glacier

Problem Description

ABC Bank is planning to introduce a new term deposit product for its customers, and they need to develop a model that can predict whether a customer will buy their product or not. To achieve this goal, they want to analyse the past interaction of customers with the bank or other financial institutions. The bank wants to use this model to understand the customers' behaviour and preferences regarding term deposits, and to determine which customers are most likely to buy the product. This will help the bank to focus its marketing efforts on those customers who are more likely to buy, resulting in a more efficient and effective sales process. Ultimately, the bank hopes to increase its sales and revenue through the successful launch of its new term deposit product.

Approach

Although minimal data processing is required for the dataset we are handling, we have taken into account the following methods to mitigate potential issues if necessary:

- Identify relationships between variables
- Look at outliers and consider options for imputing or deleting
- Look at missing/unknown values
- Try to combine classes inside categorical columns so make the data less dense
- Log Transform/ Normalise/ Standardise numerical variables
- Remove features with very low variance (day_of_week)
- Use K-Best to see which features might be less important

Models Considered:

- Logistic Regression, Decision Trees, SVM (Linear and Non-Linear), KNN, Perceptrons, SGD Classifier, Ensemble Methods
- Train models with and without SMOTE and RandomUnderSampler, try different combinations.
- Compare metrics such as AUC-ROC, F1-score and Balanced Accuracy.

Model Chosen:

- Decision Trees are preferred in such classification problems with multiple categorical variables and where relationship between data is not clear. KNN, Perceptrons, SVM and algorithms like SGD take far too long to train and their performance is not much better.
- Logistic regression and Decision Trees are the better choices, but decision tree based algorithms perform better in this dataset.
- Ensemble Methods: Random Forest, AdaBoost, XGBoost, Extra Trees, Bagging, CatBoost, LGBM, RGF

Final Recommendation:

LGBM Classifier performs the best in terms of AUC-ROC, F1-score and Balanced Accuracy. This means that it has superior precision and recall. It is also very quick to train, so it is computationally cheap.

Github Repo link

Please visit the following link to see our project:

- https://github.com/gitkym/uvs_bank_marketing_project.git