Data Science - Bank Marketing Campaign

1. Group Information

Group Name: Datalux

Group Members: 3

Name	Email	Country	Uni/Company
Huu Thien Nguyen	nguyenhuuthien27296@gmail.com	Sweden	Skövde University
Asmaa Alqurashi	asmaa.idk@gmail.com	Saudi Arabia	Taif University
Deepak Rawat	deepakrawat68@gmail.com	Ireland	Dublin Business School

Specialisation: Data Science

Submitted to: Data Glacier canvas platform

Internship Batch: LISUM09

2. Problem Description

2.1. Introduction

A common approach for increasing business is to run marketing and selling campaigns. Companies use direct marketing to reach certain categories of clients to achieve a specific goal. Customer distant interactions may be centralised in a contact centre, making campaign administration easier. Technology allows us to reimagine marketing by optimising customer lifetime value through analysing accessible data and customer KPIs, allowing us to develop longer and closer relationships in line with company needs.

The ABC Bank wants to market its term deposit product to clients in this project. Before doing so, they want to construct a machine learning model that will assist them in determining whether a particular consumer would buy their product based on the customer's previous interactions with the bank or other financial institution.

To solve the problem mentioned above, the bank wants to use machine learning modelling to identify the customers who are more reluctant to buy their services so that their marketing channels will only focus on these customers, which in turn will save time and resources and finally leads to optimised cost for this campaign.

2.2. Background

ABC Bank wants to sell its term deposit product to customers. 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 the customer's past interaction with the bank or other Financial Institution). Bank wants to use the ML model to shortlist customer whose chances of buying the product is more so that their marketing channel (telemarketing, SMS/email marketing etc.) can focus only to those customers whose chances of purchasing the product is more.

3. GitHub Link / Data / Project cycle

The link for GitHub: https://github.com/AndrewNguyen27296/DataGlacier

The data is related to direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact with the same client was required to assess if the product (bank term deposit) would be ('yes') or not ('no') subscribed.
Size: 41188 records, 20 explanatory variables, and one binary response variable

The project's general view, along with the deadline, is described in the table below. The deadline is added accordingly to the requirements from Data Glacier's canvas page.

Task Name	Week 7	Week 8	Week 9	Week 10	Week 11	Week 12
Data understanding						
Data exploration						
Data preprocessing / Feature engineering						
EDA (Exploratory Data Analysis)						
Presentation						
Modeling						

Table 1: The project's timeline

4. Methods

In order to choose the appropriate model to provide the prediction for the term deposit, our methods followed the four steps to identify the correct techniques to fulfil the requirements. It included data exploration, feature engineering, model deployment, and predictive evaluation. This section aimed to be transparent, concise, and detailed for readers from non-technical perspectives to experts can comprehend and reproduce similar results.

A list of white-box ML models (logistic regression, a simple decision tree, and a Naive Bayes algorithm) and black-box ML models (ridge classifier, SVC, k neighbours classifier, gradient boosting, random forest, and neural network) was implemented to compare which model performs the best on this particular dataset

Multiple classification metrics were utilised to examine the model. It included accuracy, recall, and ROC-AUC. F1 scores were also included.

Accuracy is the metric that evaluates a classification machine learning model's performance by using the number of accurate predictions divided by the total number of predictions. It is the most frequently used statistic for assessing classifier tasks since it is simple to compute and apprehend. In order to understand the evaluation metrics of a classifier model, there are four indicators that readers need to comprehend.

TN / True Negative: the outcome was negative(0) and predicted negative(0)

TP / True Positive: the outcome was positive(1) and predicted positive(1)

FN / False Negative: the outcome was positive(1) but predicted negative(0)

FP / False Positive: the outcome was negative(0) but predicted positive(1)

The efficiency of a classifier to accurately detect all positive cases is measured by the recall, which is a metric of its correctness. It is calculated as the ratio of true positives to the sum of true positives and false negatives for each class.

$$\text{Precision} = TP / (TP + FP)$$

$$\text{Recall} = TP / (TP + FN)$$

The Receiver Operator Characteristic (ROC) is a binary classification problem evaluation metric. It's a probability curve that displays the true positive rate against the false-positive rate at different threshold levels, separating the signal from the noise. The Area Under the Curve (AUC) summarises the ROC curve that measures a classifier's ability to differentiate between categories. The AUC reveals how sufficiently the sample differentiates between positive and negative classes. The more significant the AUC, the better.

Predictive validity is assessed by the F1-score, which balances precision and recall, and is dependent on the accuracy of the model's prediction of consumer sentiment rating. The F1 score is a weighted harmonic average of precision and recall, with 1.0 being the highest and 0.0 being the lowest. Because precision and recall are factored into F1 scores, they are lower than accurate measurements. When comparing classifier models, utilise the weighted average of F1 rather than global accuracy as a rule of thumb.

$$\text{F1 Score} = 2*(\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision})$$

Since the data has an imbalance output, the technique of under-sampling...(fill in your methods) was implemented to provide a robust classification model.

5. Results

Thien

Comparing models – Undersampling – Feature Importance

The detailed metrics of white-box and black-box models are shown in the table below, in which the AUC_test is sorted in decreasing order. The black box functioned similarly to the white box in this case. However, it has a critical flaw: the core algorithm is incomprehensible. The inputs were undersampled to balance out the dependent variable.

Table: The result of the black-box and white-box classifications model applied in the project was sorted by the descending AUC test.

	Accuracy_train	Accuracy_test	Recall_train	Recall_test	ROC_AUC_test	F1_test	MCC_test
gradient_boosting	0.949841	0.880645	0.962903	0.908805	0.950144	0.886503	0.761832
mlp	0.888535	0.882258	0.974194	0.974843	0.945916	0.894661	0.776622
logistic	0.873408	0.890323	0.869355	0.899371	0.945416	0.893750	0.780491
random_forest	1.000000	0.883871	1.000000	0.927673	0.942855	0.891239	0.769810
knn	0.873408	0.838710	0.861290	0.814465	0.907930	0.838188	0.678831
SVC	0.773089	0.795161	0.740323	0.776730	0.890010	0.795491	0.591253
naive_bayes	0.792197	0.780645	0.691935	0.694969	0.889521	0.764706	0.573145
decision_tree	1.000000	0.837097	1.000000	0.830189	0.837280	0.839428	0.674338
ridge	0.880573	0.872581	0.874194	0.871069	0.000000	0.875197	0.745090

As indicated, the gradient boosting provided the highest accuracy train, test, and the ROC score. Additionally, the F1 score showed a harmonic between the recall and precision of the model. However, the classification contains a weakness of a black-box model. It cannot explain the outcome of a prediction. Therefore, if users strive for interpretability, logistic regression can be a suitable replacement.

In order to evaluate the influence of each feature on the classification model. The bar chart below was created to depict the feature important. The euribor3m, poutcome, and marital showed a highly effective score. The model was evaluated with all independent features without pre-processing. Therefore, with the appropriate feature engineering by using domain knowledge, a more powerful classification can be made for this project.

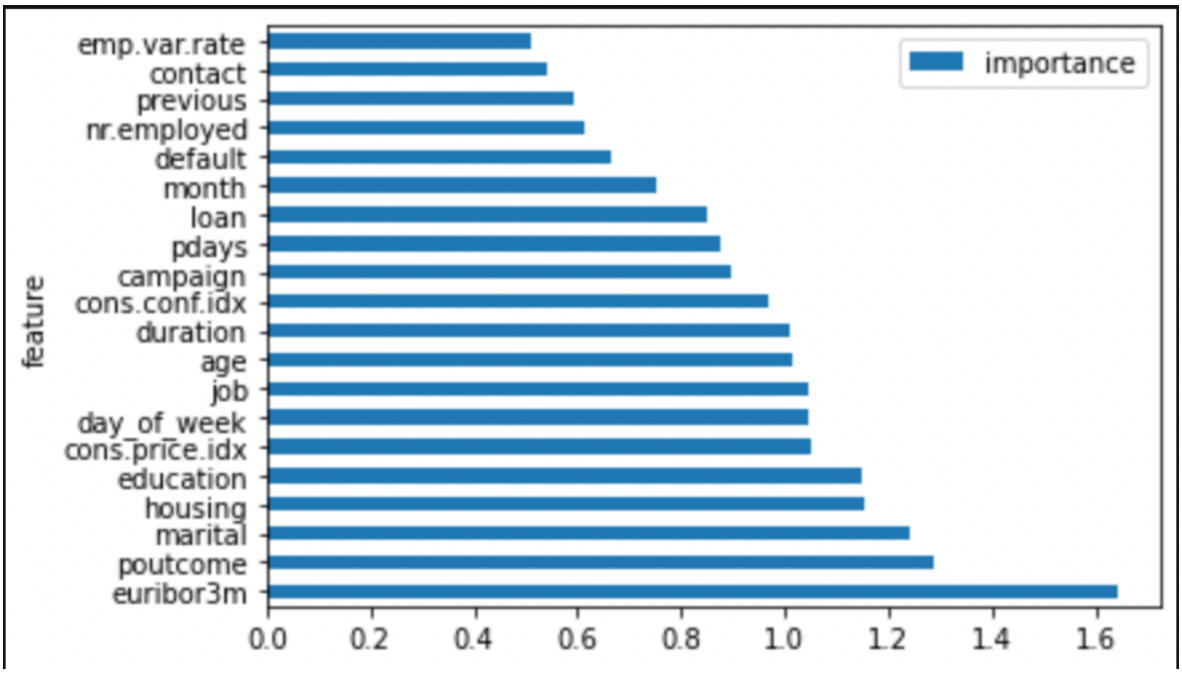

Figure: Feature importance comparison

6. Discussion

Overall, this is an excellent project to help understand the whole cycle of a data science project, from collecting the data, preprocessing, modelling, and evaluating results.

Despite the project's progress in implementing interpretation methods, it faced several challenges. As mentioned in the data exploration step, domain knowledge of banking is required because the data contain several features that need to be fully understood to make a feature engineering. If a more experienced analyst analysed the outcome, some interesting insights could be obtained to aid the business decision. For instance, deploying marketing campaign on primary client segment (subscribed term deposit customers), which are married/single, non-existent poutcome, and do not have loans.

.

7. Conclusion