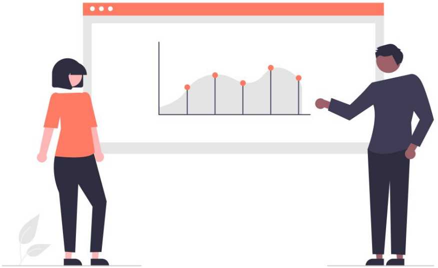
Final Project Report

Data Science - Bank Marketing Campaign

Data Glacier - Team Datalux

Asmaa Alqurashi
Deepak Rawat
Huu Thien Nguyen

PROJECT SECTION



1. Group Information
2. Problem Description
3. GitHub Repository
4. Methods
5. Results
6. Discussion

1. Team Introduction

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

The ABC Bank wants to market its term deposit product to clients in this project.

A machine learning model that will assist them in determining whether a particular consumer would buy their product.

Goal: Save the time and resources and finally leads to optimised cost for this campaign.

3. GitHub repository

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

4. Methods

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.

Since the data has an imbalance output, the technique of under-sampling was implemented to provide a robust classification model.

4. Methods

One of the methods was splitting the data to train-valid-test to evaluate the chosen models and prevent data lake by training the model and evaluate it using the training and validating sets.

Cross validation was used to evaluating the model, or hyperparameter, the model has to be trained from scratch, each time, without reusing the training result from previous attempts. The result of the cross validation give us the optimized model.

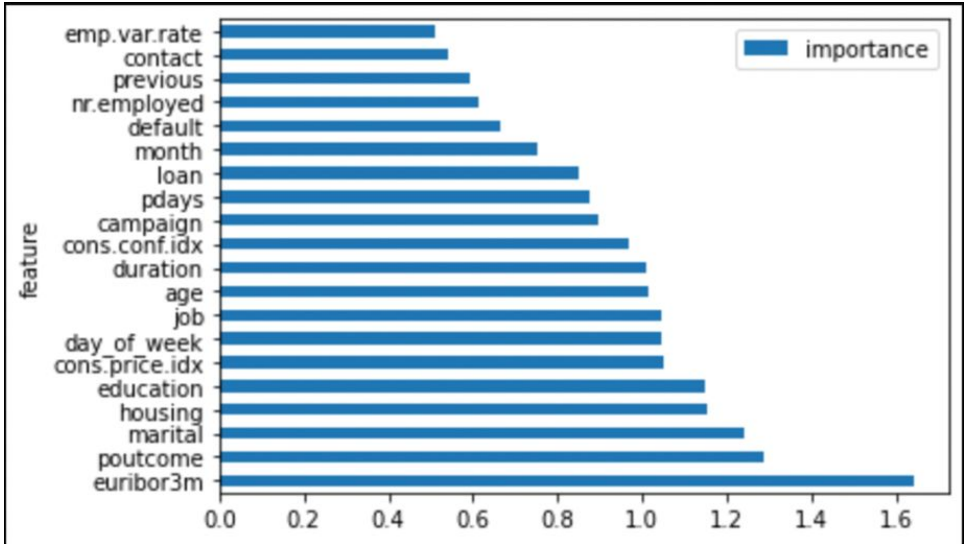
5. Results

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.

	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

5. Results

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.



5. Results

Excluding the “Duration” feature from the model

The features were engineered and the “Duration” were excluded and the data was split to 70% train, 15% validation, and 15% test.

First of all, the comparison of the models’ scores in the train and valid sets are represented in the following table:

	classifier	data_set	auc	accuracy	recall	precision	specificity	f1
0	KNN	train	0.796962	0.734085	0.603832	0.816548	0.858158	0.694262
1	KNN	valid	0.779379	0.741059	0.600858	0.834990	0.878398	0.698835
2	LR	train	0.796603	0.744438	0.632880	0.814638	0.855995	0.712348
3	LR	valid	0.798059	0.747496	0.632332	0.821561	0.862661	0.714632
4	SGD	train	0.792486	0.736557	0.646168	0.788759	0.826947	0.710379
5	SGD	valid	0.799965	0.748927	0.648069	0.811828	0.849785	0.720764
6	NB	train	0.771106	0.692522	0.491656	0.821798	0.893387	0.615236
7	NB	valid	0.779663	0.702432	0.496423	0.844282	0.908441	0.625225
8	DT	train	0.864211	0.784456	0.667800	0.871020	0.898640	0.755991
9	DT	valid	0.748045	0.719599	0.610873	0.780622	0.822604	0.685393
10	RF	train	0.812559	0.750309	0.631644	0.828201	0.868974	0.716690
11	RF	valid	0.794637	0.754649	0.642346	0.828413	0.866953	0.723610
12	GB	train	0.899603	0.820457	0.765760	0.859820	0.875155	0.810069
13	GB	valid	0.777319	0.721030	0.683834	0.738794	0.758226	0.710253

5. Results

Picking AUC performance indicator over other indicators since it is widely used and an easier metric to compare many models with.

All the algorithms have similar training AUC, but the ones that stood out are decision tree (DT) and gradient boosting (GB). Gradient boosting is considered the best metric to use because it has a higher AUC (0.89) than the other algorithms. At a threshold of 0.5, an AUC of 0.89 is good as it signifies that it is more than just a random guess towards a positive class.

After choosing the high score model let's specify the features' importance in GB. The following table shows the top 5 most important features in the model.

	importance
nr.employed	0.426433
cons.conf.idx	0.132930
euribor3m	0.101667
age	0.062249
campaign	0.029272

5. Results

As we choose Gradient Boosting Classifier as the best scoring model, we use cross validation to optimize the results.

Using RandomizedSearchCV we define the parameters and create an object to build the classifier with optimized hyperparameters

Finally we use the optimized model to evaluate it in the test set. Here is the results:

```
Gradient Boosting Classifier
Test:
AUC:0.795
accuracy:0.741
recall:0.620
precision:0.818
specificity:0.862
prevalence:0.500
f1:0.705
```

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.

Aloso, Excluding the “duration” feature had a big impact on the model score but the model can be improved to give a better result without the feature by training it in more data and focusing on the most important features.