Data Science Internship 2022
Data Science: Bank Marketing (Campaign)

Submitted by: Big Analytics
Group Members:

Name	Email	College/Company
Taimoor Razi	taimoor.r10@gmail.com	Middle East Technical University, Turkey.
Ogwu Augustine	ogwuaugust@gmail.com	University of Jos, Nigeria.
Akshar Chaklashiya	chaklashiya.akshar@gmail.com	Lambton College, Toronto.

Submitted to: Data Glacier
Due Date: 30th September 2022

Table of Contents

Project Lifecycle

Tasks

- Business Understanding
- Data understanding
- Exploratory data Analysis
- Data Preparation
- Model Selection & Model Building
- Performance reporting
- Deploy the model
- Converting ML metrics into Business metric and explaining result to business
- Presentation for non-technical persons.

Project Deadline

- 30th September 2022

Business Understanding

Problem Statement

ABC Bank wants to sell its 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).

Why ML Model

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)).

Data Understanding

Dataset Information

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).

Data Intake Report

Group Name: Big Analytics
Report date: 16-08-2022
Internship Batch: LISUM11: 30
Version: 1.0
Data intake by: Taimoor Razi
Data intake reviewer: NA
Data storage location: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Tabular data details:

Total number of observations	41188
Total number of files	1
Total number of features	21
Base format of the file	.csv
Size of the data	5.56 MB

Proposed Approach:

- The data is downloaded from the UCI Machine Learning Repository.
- The bank-additional-full has no null values but has 12 duplicates. These 12 duplicates were removed.
- There are some values labelled as “unknown” in categorical variables.

Attribute Information

Input variables:
bank client data:
1 - age (numeric)
2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
5 - default: has credit in default? (categorical: 'no','yes','unknown')
6 - housing: has housing loan? (categorical: 'no','yes','unknown')
7 - loan: has personal loan? (categorical: 'no','yes','unknown')
related with the last contact of the current campaign:
8 - contact: contact communication type (categorical: 'cellular','telephone')
9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.
other attributes:
12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
14 - previous: number of contacts performed before this campaign and for this client (numeric)
15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
social and economic context attributes
16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)
17 - cons.price.idx: consumer price index - monthly indicator (numeric)
18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)
19 - euribor3m: euribor 3 month rate - daily indicator (numeric)
20 - nr.employed: number of employees - quarterly indicator (numeric)
Output variable (desired target):
21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

Exploratory Data Analysis (EDA)

What type of data you have got for analysis?

Multivariate dataset with multiple numerical (continuous, discreet and temporal) and categorical variables present.

What are the problems in the data?

Duplicate values: One of the dataset (having 44K rows) has 12 duplicate values which are dropped.

Imbalanced target variable: Dataset is highly imbalanced dataset as the ratio of target variable value is 8:1.

“Unknown” Values seem to appear in some features which is basically a missing value put inside a category.

Duration variable: Duration is obtained after the call is made to the potential client so if the target client has never received calls, this feature is not very useful. Duration variable should be removed during the analysis

Outliers present in some of the variables

What approaches are you trying to apply on your data set to overcome problems and why?

Imbalance dataset: To deal with imbalance target variable oversampling technique will be applied before model preparation.

Missing values: Fortunately, there are no missing values in the dataset. However, after exploring we find that 5 of the categorical variables have an "unknown" value. Those are the only missing values which do not need to be dealt with for now as the "unknown category is already created for them. However, we are also considering to remove some of these unknown values after EDA.

Skewness: Transformations of features - log or normalize

Handling Categorical Data: Converting a few categorical values into numerical values by using One hot encoding - (ex. Default, housing, loan, contact). Converting temporal variables from categorical to numeric by using ordinal encoding - Month and week_of_day. Converting categorical target variable into numerical binary variable.

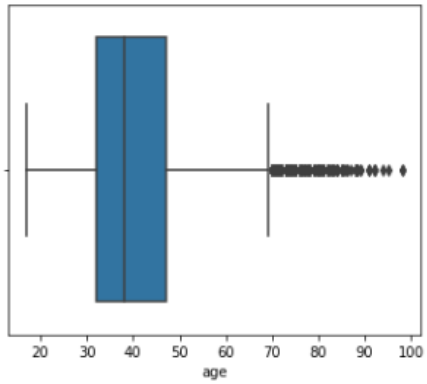
Handle Outlier: We have performed the visualization technique on numerical type variable to determine the outliers. Initially we have plot boxplot graph and then perform our analysis.

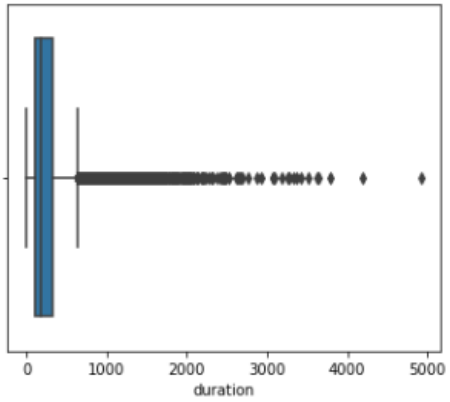
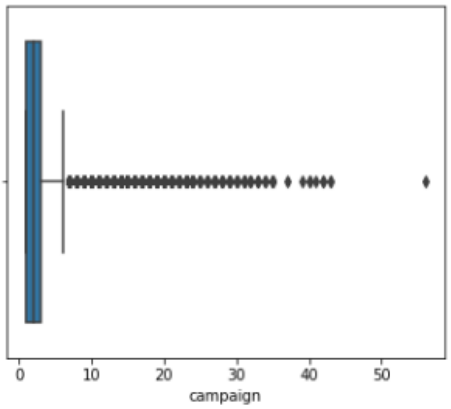
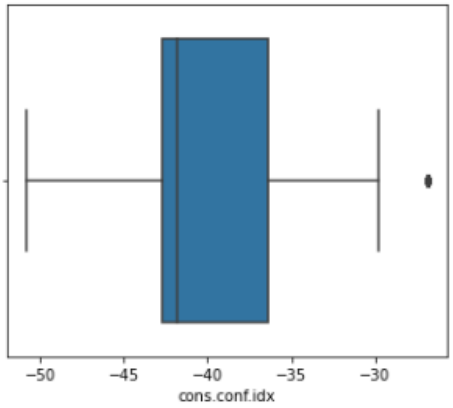
Using visualization (boxplot) we will determine which variables have outliers and then using IQR technique, we will remove/ round those values. (Q1 - 1.5IQR and Q3+1.5IQR)

We create common function to create a box plot graph.

```
fig,axes = plt.subplots(4, 3, figsize=(18, 20))
fig.suptitle('Boxplot graph')
count = 0;
for i in range(4):
    for j in range(3):
        if( count < len(numerical_var)):
            sns.boxplot(ax=axes[i,j], x=numerical_var[count], data=df)
            count = count + 1
```

Using able function we are creating 4 X 3 graph of box plot. And then we take decision based on that graph.

	As we see in the box plot graph of age, we can see there are many values after 70. But we have considered 95% as cutoff to truncate higher value. We used IQR technique to find upper 5% age group people and then we dropped those values.

	For the duration, by looking at boxplot we took 3500 as our threshold value and those value which are more than 3500, we replace it with 3500. This way we have handle duration value.
	For the campaign, we have decided to take 35 as threshold value. Same as duration, here also we are replacing value which is more than 35 to 35.
	For this feature, it is clearly visible that the last point is outlier. So we have replace it with -30 (upper value of upper whisker)

So these way we have handle the outlier in our numerical data sets.

Feature Engineering

Missing Data

Since our data were collected from phone call interviews, many clients refused to provide their personal information due to the privacy issue. The existence of missing data may blur the real pattern hidden in the data thus making it more difficult to extract information. Therefore, we chose two methods to deal with those missing data for different attributes.

Fortunately, there are no missing values in the dataset. However, after exploring we find that 6 of the categorical variables have an "unknown" value. The categorical variables with a small proportion of "unknown" values (less than 5%) can be replaced with the most frequent value of the column while the ones with high proportion of missing values (more than 5%) can be left as a separate "unknown" category. It turns out all the “unknown” values in columns were less than 5% and so we ended up filling the columns with the mode.

Numerical Variables Transformation

Logarithmic Transformation is applied to the "age" variable in order to get a more Gaussian-like distribution.
Yeo-Johnson transformation is applied to duration variable to obtain gaussian like distribution.

Categorical Feature Encoding:

Binary features, house and loan, values of “yes” and “no” are replaced with “1” and “0”.

Ordinal Encoding is done for education column

Imbalanced Data

The presence of imbalanced data may distort the algorithms and its predicting performance. This problem often happens in real world dataset, since people with some certain behaviors account for relatively smaller part. In this case, the responses in the training data are 90% “no” and 10% “yes”, which is surly a significantly imbalanced dataset. How to deal with this problem can be divided into two parts.

First, change the way of measuring algorithm's performance. As a traditional and common measurement of performance, the test accuracy rate can not be simply used here because the model will tend to fit the majority class better to improve the overall accuracy. However, we prefer to be more successful in identifying people who will subscribe a term deposit than the overall power of prediction. Therefore, we will use ROC (Receiver Operating Characteristic) curve and AUC (Area Under Curve) as the performance measurement.

Second, change the dataset using resampling method or apply different weights to the observations in objective function. As per the project requirements, oversampling (sample with replacement from the group with less data until the number equals to the larger group) will be used.

Machine Learning Models

Random Forest

The classification method we will try here is random forest. This method is good for prediction but a little bit difficult to interpret. Since we are facing the binary category, Random Forest is a good classification method to try. Random Forest will grow a big tree without trimming, then, take majority vote of the results of all the trees.

The process of this method is:

1. Take a sample of size n from the training dataset;

2. Randomly choose p variables from all the variables available;

3. Train a single big tree on the sample dataset and using p variables;

4. Repeat the step above B times;

5. Take a majority vote of the results for all of the B trees.

RandomForestClassifier class from sklearn is used to create the model. Hyperparameter tuning is done over n_estimators and max_depth of the ensemble through grid search with five-fold cross-validation.

Extreme Gradient Boosting

Popularly called XGBoost, this model works to accurately predict a target variable by combining the estimates of a set of simpler, weaker models. It is a tree based ensemble machine learning algorithm which is a scalable machine learning system for tree boosting. The model is instantiated using the XGBoostClassifier class. Hyperparameter tuning is carried out over learning_rate, n_estimators, max_depth, colsample_bytree, and subsample through random search with a 10-fold cross validation.

Logistic Regression

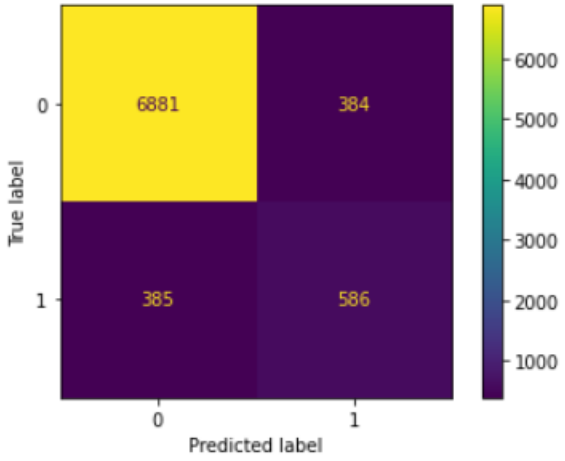
Logistic regression is one of the most popular Machine Learning algorithms, It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. For our example, we have set an upper limit to train our model up to 3000 iteration, not to overfit our model. Then we used the GridSerchLogistic algorithm as our final training algorithm and the roc_auc matrix to measure the performance/accuracy of our model.

Results

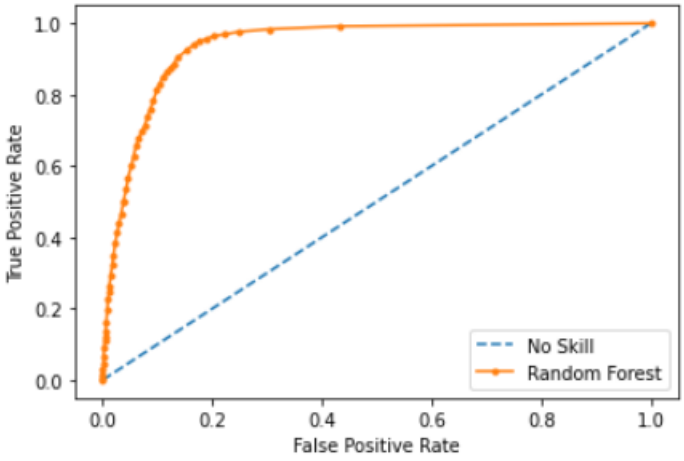
Random Forest

Since random forest classifier tends to be biased towards the majority class, we will use resampling to fix this problem. Oversampling method is used. The baseline accuracy is 88.86% which means if we prepare a model which outputs “no” for the term deposit then the model will still be correct 88.86% of the time since this many proportions of people are in the dataset who said “no” to term deposit. We need to build a model which should have at least more accuracy than the baseline model.

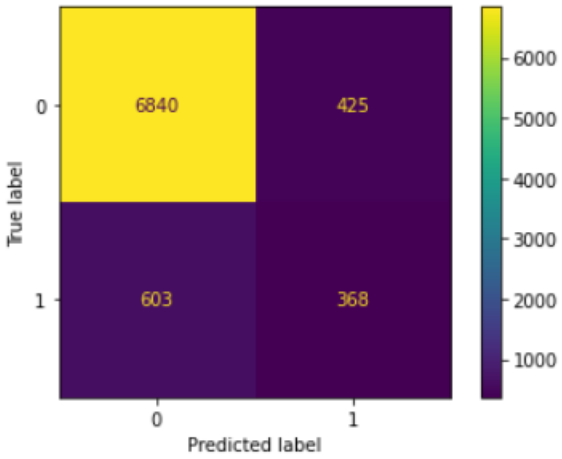
The test accuracy of 90.66% was obtained with a ROC AUC score of 93.8%. The confusion matrix was obtained as following.



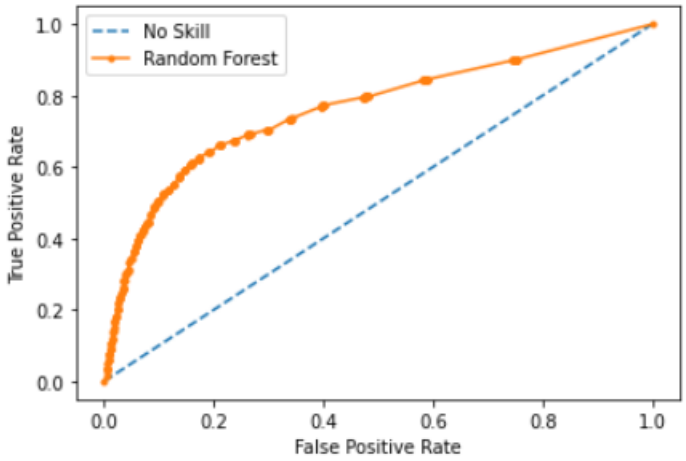
The ROC AUC curve is given below.



The model was also tested without the duration variable. The test accuracy fell down to 87.52 % and the ROC AUC score was found to be 76.2%. The confusion matrix was obtained as following.

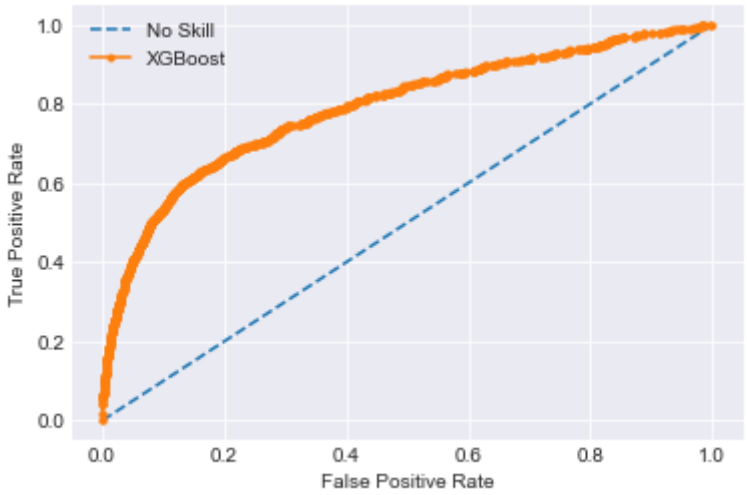


The ROC AUC curve is given below.



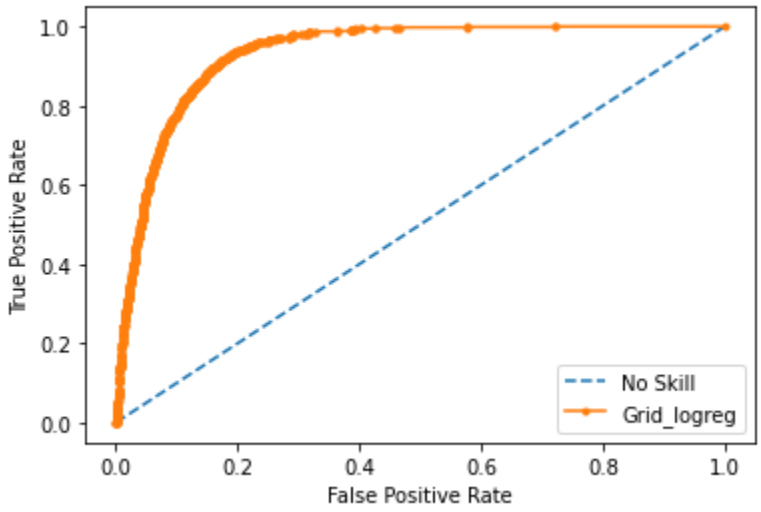
Extreme Gradient Boosting

A ROC AUC score of 62.9% was obtained without hyperparameter tuning. After hyperparameter tuning, A ROC AUC score of 60% was obtained with the curve shown below while the test and train accuracies are 82.4% and 78.9% respectively.



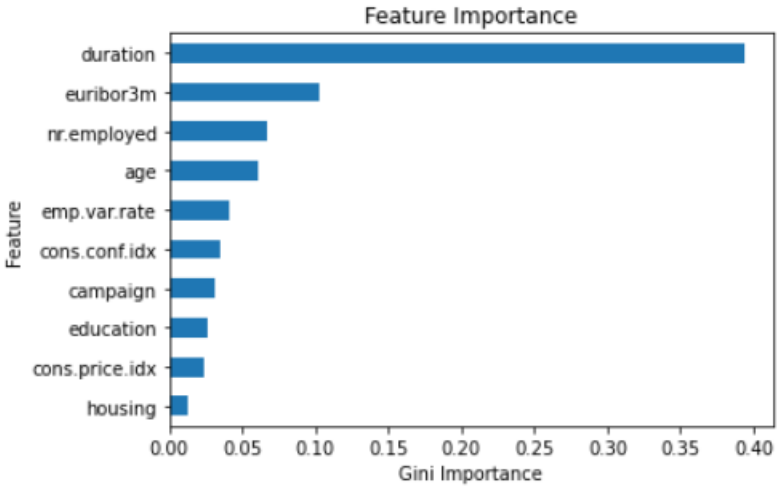
Logistic Regression

Using logistic Regression we got accuracy score for train dataset 0.879 and for test dataset 0.854. As we can see in graph that our true positive rate is sharpen at the initial stage and then steady graph. We got roc_auc score 0.864.



Conclusion

In the light of overall test accuracy and AUC, the best model is Random Forest. It has the most powerful prediction ability. Next, we need to find out which factors are most important and how these factors influence customers' decision. The bar chart below shows the feature importance for Random Forest model (with duration variable)".



According to the plot for both random forest and XGB we can tell that the most influential variables are duration, euribor3m, nr.employed, age, and emp.var.rate in descending order. "duration" has positive effect on people saying "yes". This is because the longer the conversations on the phone, the higher interest the customer will show to the term deposit. "nr.employed", which is the number of employees in the bank, has positive effect for turning people to subscribe the term deposit. This can be due to the fact that the more employees the bank have, the more influential and prestigious this bank is. "euribor3m" is another important variable, which denotes the euribor 3 month rate. This indicator is based on the average interbank interest rates in Eurozone. It also has positive effect since the higher the interest rate the more willingly customer will spend their money on financial tools. Employment variation rate (emp.var.rate) has negative influence, which means the change of the employment rate will make customers less likely to subscribe a term deposit. This makes sense because the employment rate is an indicator of the macroeconomy. A stable employment rate denotes a stable economic environment in which people are more confident to make their investment. Lastly, age is also an important factor and the marketing campaign should be based accordingly targeting different age groups. The old age group is the most vulnerable to saying "yes" for term deposit.

Therefore, if banks want to improve their lead generation, what they should do is to hire more people to work for them, improve the quality of conversation on the phone and run their campaigns when interest rates are high and macroeconomic environment is stable.

References

1. How to deal with outliers in python -. ProjectPro. (n.d.). Retrieved September 30, 2022, from https://www.projectpro.io/recipes/deal-with-outliers-in-python
2. *Detecting and treating outliers: How to handle outliers*. Analytics Vidhya. (2022, July 21). Retrieved September 30, 2022, from https://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/%20
3. Handling outliers in python. Handling Outliers in Python. (n.d.). Retrieved September 30, 2022, from https://www.datasciencesmachinelearning.com/2018/11/handling-outliers-in-python.html
4. Aleksandradeis. (2019, February 8). *Bank marketing analysis*. Kaggle. Retrieved September 30, 2022, from https://www.kaggle.com/code/aleksandradeis/bank-marketing-analysis
5. Sklearn.linear_model.logisticregression. scikit. (n.d.). Retrieved September 30, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html