IMBA PLATINUM DEPOSIT DATA ANALYSIS REPORT

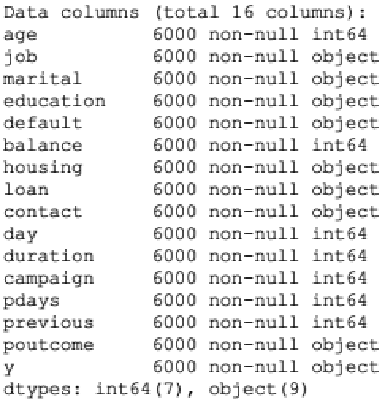
BUSI4390 Foundational Business Analytics
Coursework 2019 - 2020

Shiqi BAI
20219140

Section A: Summarization

Data Description

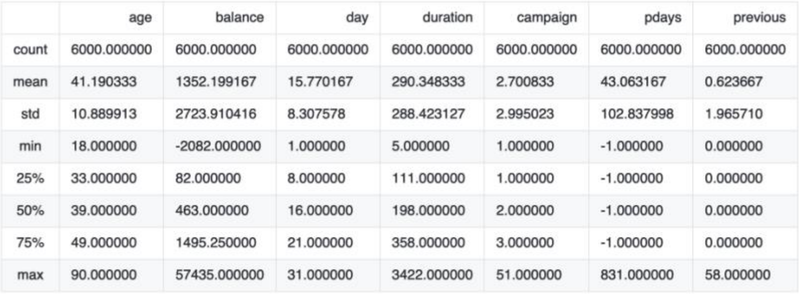
The project aims to implement a binary classification with python to predict if the client will subscribe the ‘IMBA Platinum Deposit’ or not. The dataset given is related to direct marketing campaign based on phone calls for the previous identical product of the IMBA enterprise. At the first of exploration data analysis, the functions df.head(), df.info(), de.describe() and df.shape() help us to get the quick overview of the whole dataset. The dataset for the project has 6,000 rows and 16 columns and there are no obvious missing values as shown in Figure 1.



```
Data columns (total 16 columns):
age          6000 non-null int64
job          6000 non-null object
marital      6000 non-null object
education    6000 non-null object
default      6000 non-null object
balance      6000 non-null int64
housing      6000 non-null object
loan         6000 non-null object
contact      6000 non-null object
day          6000 non-null int64
duration     6000 non-null int64
campaign     6000 non-null int64
pdays        6000 non-null int64
previous     6000 non-null int64
poutcome     6000 non-null object
y            6000 non-null object
dtypes: int64(7), object(9)
```

Figure 1: Basic description of the dataset

In the dataset of this case, the target variable ‘y’ contains ‘yes’ and ‘no’ two classes. The majority class ‘no’ consists 4,728 samples, while the minority class ‘yes’ consists 1,272. (i.e. 4,728 clients did not subscribe to the product in the previous marketing campaign, while 1,272 clients subscribed to the product). In other words, the imbalance issue addressed might influence the modelling. Moreover, the total 16 variables including 7 numerical variables and 8 categorical variables as well as 1 target variable which shows in Figure 1. For the further observation, the basic statistics description of numeric variables is shown below:



	age	balance	day	duration	campaign	pdays	previous
count	6000.000000	6000.000000	6000.000000	6000.000000	6000.000000	6000.000000	6000.000000
mean	41.190333	1352.199167	15.770167	290.348333	2.700833	43.063167	0.623667
std	10.889913	2723.910416	8.307578	288.423127	2.995023	102.837998	1.965710
min	18.000000	-2082.000000	1.000000	5.000000	1.000000	-1.000000	0.000000
25%	33.000000	82.000000	8.000000	111.000000	1.000000	-1.000000	0.000000
50%	39.000000	463.000000	16.000000	198.000000	2.000000	-1.000000	0.000000
75%	49.000000	1495.250000	21.000000	358.000000	3.000000	-1.000000	0.000000
max	90.000000	57435.000000	31.000000	3422.000000	51.000000	831.000000	58.000000

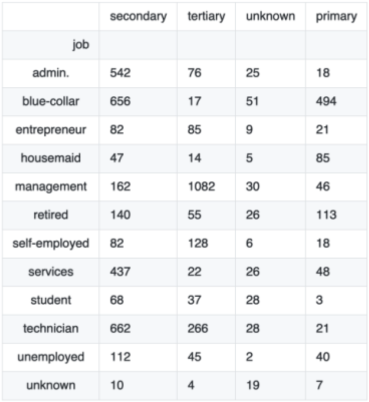
Table 1: Basic statistics information of numeric variables

Target customers’ characteristic inferred from Table 1:

- Middle age; debt-free; be willing to answer the calls; good relations maintained.

According to this dataset, age, degree of education, and marital status can reflect the differences in customer demand for specific term deposit products at different stages of life. The customer's work status, personal balance, and credit status reflect his economic status, which will also have a great impact on purchase intention. The bank's marketing means, timing of sales, length of sales time, and historical marketing status all reflect the bank's marketing strategy and marketing capabilities from different aspects.

Relationships between features



	secondary	tertiary	unknown	primary
job				
admin.	542	76	25	18
blue-collar	656	17	51	494
entrepreneur	82	85	9	21
housemaid	47	14	5	85
management	162	1082	30	46
retired	140	55	26	113
self-employed	82	128	6	18
services	437	22	26	48
student	68	37	28	3
technician	662	266	28	21
unemployed	112	45	2	40
unknown	10	4	19	7

Observation from Table 2: ‘Job’ and ‘Education’

With the exploration of features ‘Job’ and ’Education’, one hypothesis could be justified that a customer’s job will be influenced by their degree of education.

In addition, we discovered that there are some unknown values in both job and education data. Hence, we could infer the unknown values based on the hypothesis in the data cleaning phase.

Table 2: Cross tabulation (Yin, et al., 2013) of ‘job’ and ‘education’

Relationships between feature and output variable

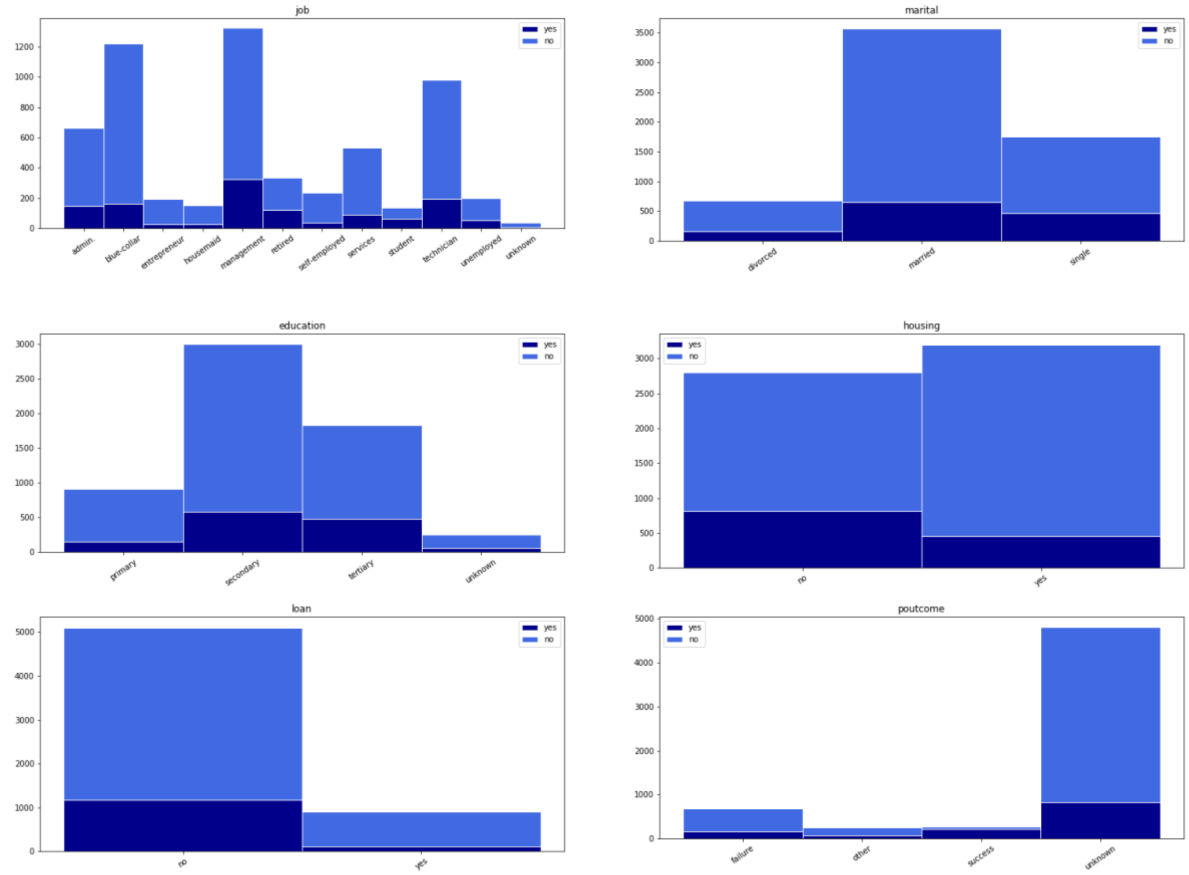


Figure 2: Bar charts of two classes proportion in each feature

Observation from Figure 2 (which are also supported by the quantitative analysis in Section B):

- Office workers, retired elderly, and students are more willing to buy such financial products than bosses, but blue-collar is weaker than bosses to buy these products.

- Compared with married customers, single or divorced people are more willing to purchase such products.

- Customers with a high degree of education and no pressure on mortgages or other loans are more likely to choose this term deposit product.

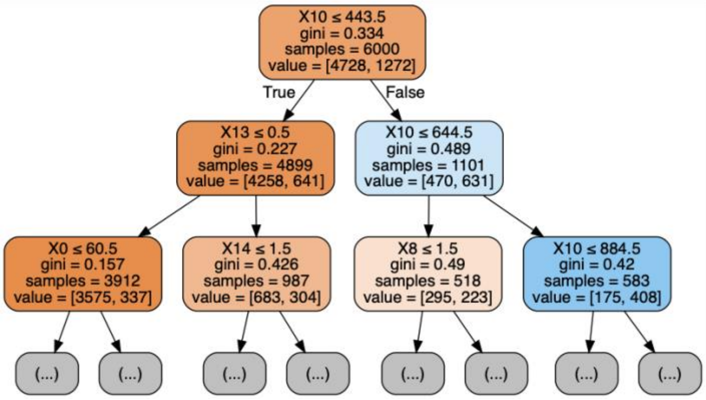
- The close relationship between the bank and customers in the past, and successful sales promotion effect before will increase the success rate of subsequent sales of other wealth management products to the customers.

- Personal loan pressure increases, or private wealth shrinks may cause customers' willingness to choose bank wealth management products to decline significantly.

- More sales time will increase the sales effect of the product, but frequent sales of the same financial products to the same customer will instead reduce the sales effect.

Section B : Exploration

Feature importance analysis based on a Decision Tree



Xₙ	**X0**	**X1**	**X2**	**X3**	**X4**	**X5**	**X6**	**X7**	**X8**	**X9**	**X10**	**X11**	**X12**	**X13**	**X14**
Feature	Age	Job	Marital	Education	Default	Balance	Housing	Loan	Contact	Day	Duration	Campaign	Pdays	Previous	Poutcome
Importance	**0.092**	0.028	0.019	0.023	0.002	0.131	0.041	0.016	0.030	**0.098**	**0.343**	0.031	0.032	**0.052**	**0.063**

Figure 3: The Decision Tree with the original dataset and importance of each feature

In this stage of exploratory data analysis, we applied a Decision Tree as a mean of feature selection (Liu & Motoda, 2007) to discover the important influencing factors. For this specific dataset given, it is shown in Figure 3 that ***duration, age, day, previous*** and ***poutcome*** are significant features impact customers whether subscribe the term deposit or not. The result could also further support the observations we discussed in Section A.

Correlation analysis

Since that categorical values exist in the dataset, we created the dummy variables from these features into the numerical values for the preparation of model building phase. Moreover, we built the heatmaps of dummy variables in Figure 4, in order to observe the correlation among the data and identify which features have good correlations with the target variable.

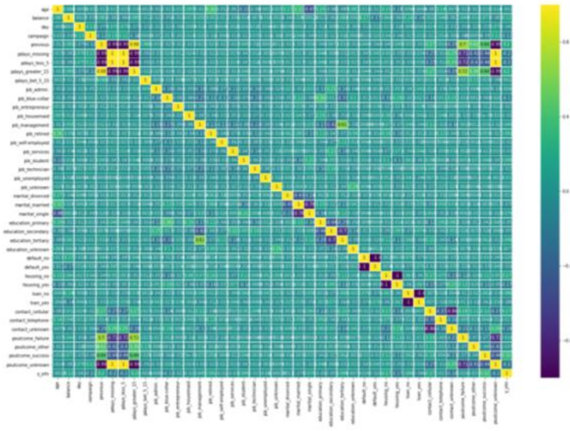


Figure 4: Heatmaps of correlation of variables

Observation from Figure 4:

- Well-connected (Last contact duration is long) middle-aged customers who have been contacted within a month and purchased the other products could be the target group.
- Our target variable has good correlation (Liu & Motoda, 2007) with ***default, housing, loan, poutcome*** and ***pdays***. We expect to regard these independent variables as significant while building the models.

Section C: Model Evaluation

Model evaluation strategy

Evaluation Combination	Model Performance
High recall + High precision	The class is perfectly handled
Low recall + High precision	Cannot detect the class well but is highly trustable when it does
High recall + Low precision	The class is well detected but the model also includes points of other classes in it
Low recall + Low precision	The class is poorly handled

Table 3: Evaluation criteria combine precision and recall

According to the bank’s CEO declaration, a trustable result of positive class detection (i.e. prediction that the customer will subscribe the term product) confirms to the company’s strategy. In other words, there are two types of wrong values to consider:

- False positive: customer will not subscribe the product, but the model predicts he will;
- False negative: customer will subscribe the product, but the model predicts he will not.

With reference to the Table 3 and the CEO’s perspective, false positive values should be as lower as possible to avoid unnecessary waste of human resources and time. Also, false negative values are also harmful, the higher value means that the marketing team might miss the potential customers to promote sales.

Confusion matrix, precision and recall could help a lot for evaluating the model performance. Furthermore, AUC score is chosen as the scoring metric since it has been established that for cases where classes are unbalanced (Longadge & Dongre, 2013), AUC score is a better evaluation criterion than the accuracy score. For each model we selected, five-fold cross-validation is performed over the training set. The mean AUC score is calculated for each set of selected parameters. The final model with tuned hyperparameters are selected based on the highest out-of-sample mean AUC score.

Model selection and training

In our case, we took three basic classification methods to implement the prediction:

i) Logistic Regression: compared with the conditional independence hypothesis of Naïve Bayes, Logistic Regression does not need to consider whether the samples are relevant. From the marketing insight, it's realizable that we could collect more data information in the future. Logistic Regression is worth the effort with the consideration of the future update and improvement in the model.

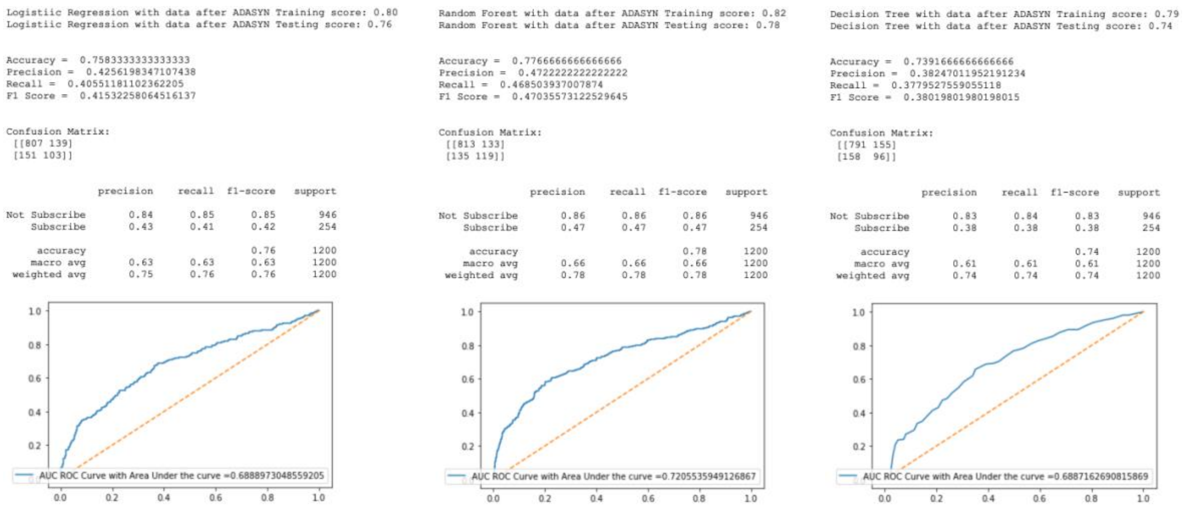
ii) Random Forest and Decision Tree: Decision Tree is easy to understand and explain as well as a non-parametric model. We do not need to worry about whether the outliers are linearly separable with the data. The main disadvantage of Decision Tree is overfitting thus we choose an ensemble learning algorism Random Forest as our candidate model.

	Train the model with the dummy variables				**Train the model with the data after ADASYN algorism to handle the imbalance data**			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Logistic Regression	0.8150	0.8076	0.1654	0.2745	0.7583	0.4256	0.4055	0.4153
Random Forest	0.8117	0.8684	0.1299	0.2260	0.7767	0.4722	0.4685	**0.4704**
Decision Tree	0.8183	0.7250	0.2283	0.3473	0.7392	0.3825	0.3780	0.3802

Table 4: Common scoring metric for different classifiers with the data before and after balancing

As the discussion in the first section, the original dataset is considered as an imbalance dataset with the majority negative samples. We applied the ADASYN algorism (Longadge & Dongre, 2013) to balance the proportion of the dataset to enhance the robustness of the model facing the real-world application. In this initial training stage, it’s obvious that the Random Forest classifier trained with the balanced data perform well in accuracy, precision, recall and F1 score than other models.

From the below results after training the models with their default parameters in Figure 5, it's evident that Random Forest is performing better based on False Positive and AUC than both the Logistic Regression and Decision Tree models. The next step is finding out the best parameters of the Random Forest by cross validation.



```
Logistiic Regression with data after ADASYN Training score: 0.80
Logistiic Regression with data after ADASYN Testing score: 0.76

Accuracy =  0.7583333333333333
Precision =  0.4256198347107438
Recall =  0.40551181102362205
F1 Score =  0.41532258064516137

Confusion Matrix:
 [[807 139]
 [151 103]]

              precision    recall  f1-score   support

Not Subscribe       0.84      0.85      0.85       946
    Subscribe       0.43      0.41      0.42       254

     accuracy                           0.76      1200
    macro avg       0.63      0.63      0.63      1200
 weighted avg       0.75      0.76      0.76      1200
```

```
Random Forest with data after ADASYN Training score: 0.82
Random Forest with data after ADASYN Testing score: 0.78

Accuracy =  0.7766666666666666
Precision =  0.4722222222222222
Recall =  0.468503937007874
F1 Score =  0.47035573122529645

Confusion Matrix:
 [[813 133]
 [135 119]]

              precision    recall  f1-score   support

Not Subscribe       0.86      0.86      0.86       946
    Subscribe       0.47      0.47      0.47       254

     accuracy                           0.78      1200
    macro avg       0.66      0.66      0.66      1200
 weighted avg       0.78      0.78      0.78      1200
```

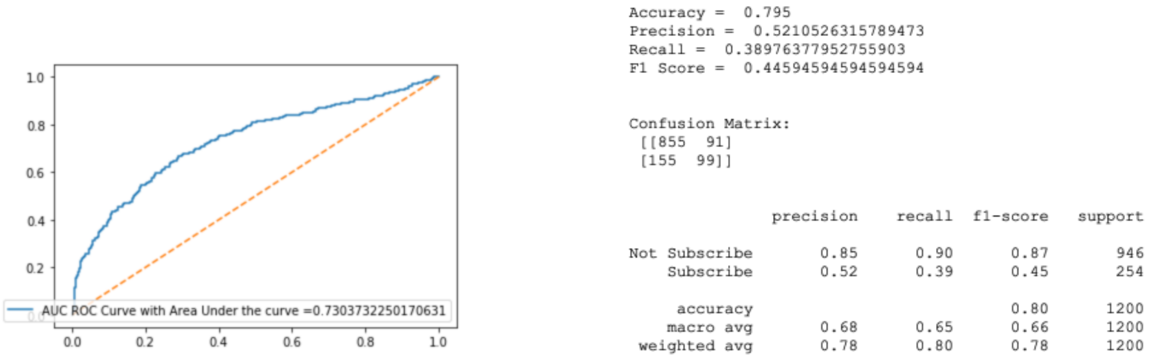
```
Decision Tree with data after ADASYN Training score: 0.79
Decision Tree with data after ADASYN Testing score: 0.74

Accuracy =  0.7391666666666666
Precision =  0.38247011952191234
Recall =  0.3779527559055118
F1 Score =  0.38019801980198015

Confusion Matrix:
 [[791 155]
 [158  96]]

              precision    recall  f1-score   support

Not Subscribe       0.83      0.84      0.83       946
    Subscribe       0.38      0.38      0.38       254

     accuracy                           0.74      1200
    macro avg       0.61      0.61      0.61      1200
 weighted avg       0.74      0.74      0.74      1200
```

1.0
0.8
0.6
0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0
AUC ROC Curve with Area Under the curve =0.6888973048559205
AUC ROC Curve with Area Under the curve =0.7205535949126867
AUC ROC Curve with Area Under the curve =0.6887162690815869

Figure 5: Classification report for three models

5-fold cross validation and model hyperparameters tuning

For Random Forest, two hyperparameters were tuned: minimum samples split (the minimum number of samples required to split an internal node) and minimum samples leaf (the minimum number of samples required to be at a leaf node). These two parameters help control the depth of the trees and thus help to control the model’s complexity. It is clear that the classifiers were sensitive to the hyper-parameter chosen. After gaining the best parameters, the result of tuned model is shown in Figure 6.



```
Accuracy =  0.795
Precision =  0.5210526315789473
Recall =  0.38976377952755903
F1 Score =  0.44594594594594594


Confusion Matrix:
 [[855  91]
 [155  99]]

                precision    recall  f1-score   support

Not Subscribe       0.85      0.90      0.87       946
    Subscribe       0.52      0.39      0.45       254

     accuracy                           0.80      1200
    macro avg       0.68      0.65      0.66      1200
 weighted avg       0.78      0.80      0.78      1200
```

Figure 6: Classification report for Random Forest classifier with tuned hyperparameters

Section D: Final Assessment

From the above results, the best out of model performance was obtained for the Random Forest classifier with the minimum number of samples required to split an internal node = 10 and the minimum number of samples required to be at a leaf node = 5. On the test data, the best AUC score achieved was 0.7304 and the False Positive was 91 samples. Recall the CEO’s opinion, the importance of the features was plotted in Figure 8 which are in terms of how greatly they influenced the coefficients. This also provides valuable information toward exploring which features contribute the most toward the model’s performance. For example, marketing team should pay more attention to customers’ characteristic than contact forms.

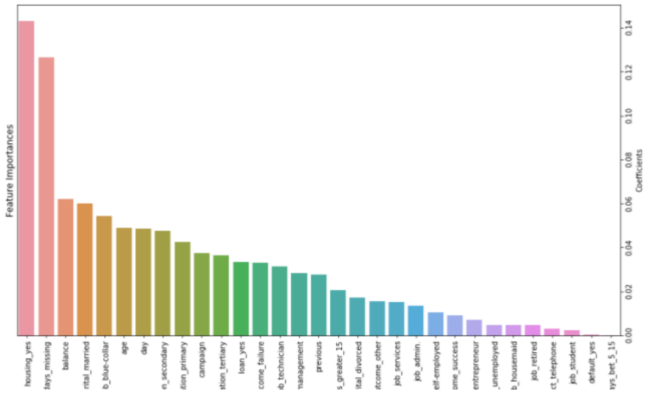


Figure 8: The importance of features

Section E: Model Implementation

PIPELINE		DESCRIPTION
Data collection		* Import the data from the files in folder and combine them into a dataframe.
Data manipulation/Data pre-processing	Handling the missing/unknown data	* Deal with the unknown values in ‘job’ and ‘education’ features; * Process the value = -1 in ‘pdays’ (convert the numerical data into categorical data).
	Selecting the important features	* Apply a decision tree with default parameters; * Create the dummy variables for further analysis of features as well as modelling phase;

		* Pearson correlation and heatmaps; * Drop the features.
	Oversampling the data	* Handle the imbalance data with ADASYN algorism; * Split the cleaning data into train and test data.
Train models	Logistic Regression Classifier	* Train the models with their default parameters using train data; * Predict the results with test data by the trained models; * Report the classification performance.
	Random Forest Classifier	
	Decision Tree Classifier	

Validate the model	Tuning the hyperparameters
	(Tuning of the models is done simultaneously with cross-validation in order to decide the parameterizations for each model)
		* Logistic Regression: 'C' and 'penalty';
		* Random Forest: 'min_samples_leaf' and 'min_samples_split';
		* Decision Tree: 'criterion' and 'max_depth'.

Test and select best model		* Report the classification performance after tuning. * Select the model with the best performance in this real-world case.
Instruction		* Ensure the test data file is in the same folder with our Jupyter Notebook file; * Follow the steps in README.md to make the prediction.

Table 5: Brief instructions on programming by Python

Section F: Business Case Recommendations

Based on the feature importance plot, some recommendations can be made to the bank's marketing team:

- The marketing team should collaborate with economic experts so that as soon as they have some signals indicating the consumer satisfaction goes up, they can expect more customers to subscribe for the product and should reach out to them pro-actively before the competitors do.

- 'Duration' should be considered as a significant factor. This makes intuitive sense since the longer duration shows that the customer is more interested in the product. Hence, the marketers should try to make the call engaging and increase the duration of the call.

- The marketing team should prioritise those customers to whom they previously reached out during previous campaigns. They are likely to subscribe for the term deposit.

Reference

Liu, H. and Motoda, H. eds., 2007. *Computational methods of feature selection*. CRC Press.

Longadge, R. and Dongre, S., 2013. Class imbalance problem in data mining review. *arXiv preprint arXiv:1305.1707*.

Yin, C., Hirokawa, S., Yau, J.Y.K., Hashimoto, K., Tabata, Y. and Nakatoh, T., 2013. Research trends with cross tabulation search engine. *International Journal of Distance Education Technologies (IJDET)*, *11*(1), pp.31-44.