LEHIGH UNIVERSITY

Introduction to Machine Learning

(ISE 364)

Classification Model for the Prediction of Whether a Client Will Subscribe to a Long-term Deposit Program

Final project Report

Instructor:

Dr.Louis Plebani

Team members:

Mohanad Khazaali

Fahim Rustamy

Table of Contents

List of Tables

List of Figures

Abstract

The purpose of this report is to find a machine learning algorithm to predict whether a client will subscribe to a term deposit and identify the important factors that influence the clients' decisions. Data visualization is performed to clean the data and evaluate the weight (influence) of each feature on the outcome of the customer's decision. The imbalance in the provided data is treated by resampling techniques before building the classification models. Five machine learning models (i.e Logistic Regression, K Nearest Neighbor, Random Forest, Support Vector Machine and Neural Networks by Keras) have been used to determine the best model that give the best prediction. The optimal model with highest prediction accuracy of 89 % was found to be the Neural Networks algorithm.

1- Introduction

Data classification and administration are important topics where company managements and marketing personnel are interested in the possibility of the desired outcomes based on their available recorded data from the past. Through this process, companies put value to the customers and create a good connection to get profit out of customer's decision. Factors that have highly influenced the decision of a client in the past are classified based on the machine learning algorithms and are used to predict the tendency of the new customers decisions. For this project 40182 rows of data are available from a company's interaction with the customers in the past showing whether the clients have subscribed to a long term deposit or not, will be used to train our machine learning algorithms. Profile of each customer has been recorded in this file containing 15 features as is shown in Table 1. All the categorical features are transformed to the numerical values using sklearn.presprocessing.LabelEncoder before performing any machine learning tasks.

Five machine learning algorithms, such as Logistic Regression, K Nearest Neighbor, Random Forest, Support Vector Machine and Neural Networks by Keras are executed by using Python program. First the original untouched data is used to determine the accuracy of our mentioned models and it was found out that the accuracy rate was about 90 % for all models. After inspection of the data, it was found out that our models predict only one class of outcomes (client would not subscribe). The reason behind that is that our models cleverly oversees the data and accordingly decides that the best prediction is the class that has occurred the most.

According to our conducted research, it was found that some features highly affect client's decision to subscribe for a long term deposit. As an example, majority of customers who were contacted after the previous campaign by the company, did subscribe for the deposit. Based on the data observations, majority of the customers, around 90 % responded 'no' to the subscription and only around 10 % decided to subscribe for the deposit. This clearly indicates that the training data is imbalanced and that regress the evaluation of the modal performance towards only one class of prediction (i.e. 'No'). To get a good correlation between the expected and predicted classes, this issue is required to be solved by using Resampling technique.

Table 1 Description of the features

Feature	type	Feature	Type	Feature	Type
age	Numerical	job	Categorical	loan	Categorical
campaign	Numerical	marital	Categorical	contact	Categorical
pdays	Numerical	education	Categorical	day_of_week	Categorical
cons_price_idx	Numerical	default	Categorical	P outcome	Categorical
cons_conf_idx	Numerical	housing	Categorical		

2- Objective

The objective of this report is to find a best classifier that would predict whether a client would subscribe to a long term deposit or no based on the past available records. Two sets of data was provided by Dr. Plebani for the project under the names of data.csv and future.csv. The former one contains information that would be used for training purposes and the later one will be used for the prediction. The following are the main objectives of this project:

1. Clean the data by preprocessing and fix any imbalances in the data
2. Find the best machine learning algorithm among the five commonly used classifiers for the provided training dataset (data.csv).
3. Predict whether the future users from the future.csv dataset would subscribe to the long term deposit or not.

3- The Summary

As mentioned in the previous section, the objective of this project is to find an optimal machine learning algorithm that would be able to predict whether the client would subscribe for the long term deposit or not. This report implemented the preprocessing on the data, tested the five machine learning algorithms and implement the optimal model to predict the clients' future decisions. Following are the summary of the report:

- Before performing any machine learning data processing is performed on the data. All the data are visualized to understand the unique categories of each feature. Then all the

categorical data is transformed to numerical values through LabelEncoder. Finally, the imbalance issue in the data is taken cared of using the SMOTE method.

- After performing the data processing, five machine learning algorithms were implemented on the data.csv dataset to determine the optimal algorithm to be used for predictions. Logistic regression, KNN, Random Forest, Support Vector Machine and Neural Networks are the algorithms used. Among them, the Neural Network model, provided the best accuracy of prediction and is used for the prediction in the future.csv dataset.
- Finally, the trained model using the optimal machine learning algorithm (Neural Networks) is used to predict the future decisions of client.
- It was found out that the there is no perfect solution to the machine learning problems. There is always room for the improvement by manipulating the data and using/trying various algorithms.

4- Data Preprocessing

Two types of data are provided for this project, training dataset and future dataset as was described before. Training dataset contains 40182 rows of data with 15 columns of features as well as the client's decision output (as shown in Table 2). On the other hand, the prediction dataset contains 1007 lines of data containing 15 features and we are supposed to predict the user's decision of whether they will subscribe for this long term deposit or not based on these features. For the data preprocessing purposes, the following tasks will be performed on the training dataset:

- Data Visualization
- Transforming the categorical data to numerical
- Dealing with imbalance data

Table 2 the first five rows of the training dataset

	age	job	marital	education	default	housing	loan	contact	day_of_week	campaign	pdays	poutcome	cons_price_idx	cons_conf_idx	prime_rate	y
0	56	housemaid	married	basic.4y	no	no	no	telephone	mon	1	999	nonexistent	93.994	-36.4	4.857	no
1	57	services	married	high.school	unknown	no	no	telephone	mon	1	999	nonexistent	93.994	-36.4	4.857	no
2	37	services	married	high.school	no	yes	no	telephone	mon	1	999	nonexistent	93.994	-36.4	4.857	no
3	40	admin.	married	basic.6y	no	no	no	telephone	mon	1	999	nonexistent	93.994	-36.4	4.857	no
4	56	services	married	high.school	no	no	yes	telephone	mon	1	999	nonexistent	93.994	-36.4	4.857	no

4.1 Data Visualization

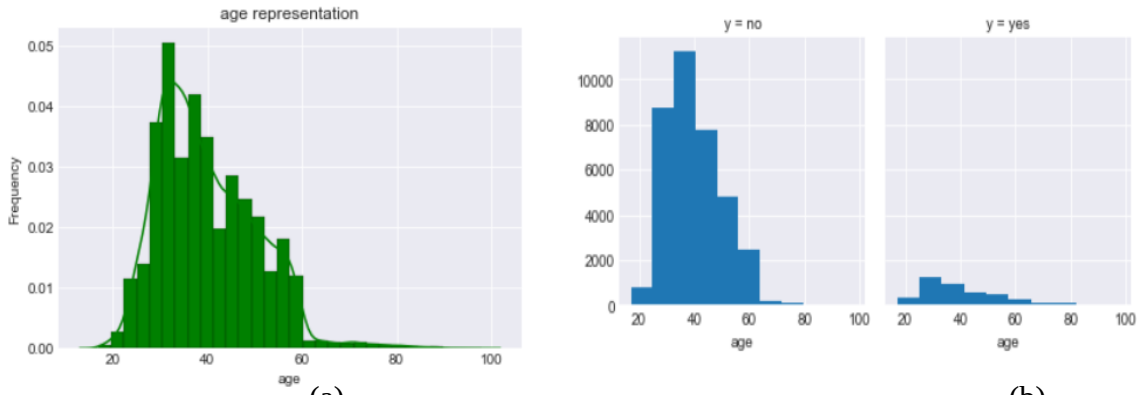
To understand the unique category and its weight within each feature and how it would effect on the output, data visualization is performed. *Figure 1*a shows the distribution of the client's age spectrum from the data that is available in the training dataset. *Figure 1*b shows the number of clients with different ages that agreed or disagreed to the long term deposit. *Figure 1* shows that the average age of the client who bought the long term deposit is a little higher than the one who did not buy it! (still does not give a good picture about the data). We need further visualization to get better understanding of our data.

Job description of the customers is another feature that has been recorded in the data.csv file and their count for each profession is plotted in *Figure 2*a with the admin and blue-collar workers being the majority. *Figure 2*b shows the percentage of people who said yes and no and it can be seen that this is an important feature for classification as with some of the profession, the probability of subscribing is relatively high.

Regarding other features, the same procedure mentioned above was applied to evaluate the importance of each category on the output as shown in *Figure 1* to Figure 12. It was found that other features (i.e., day of week, and Pdays (most are 999)) shown in Figure 5 and Figure 10 would not affect our model significantly. Therefore, they can be dropped from the model. This depends on the team's judgment, but to be accurate it is better to determine the information gain (IG) of the feature and evaluate its effect on the output.

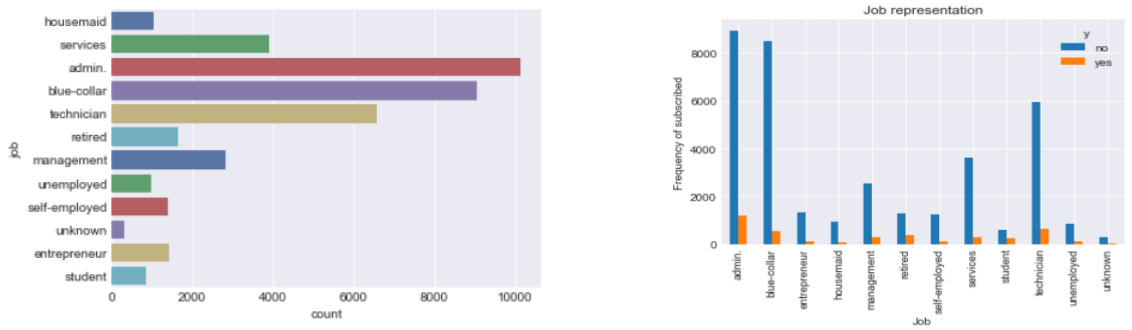
Summary of the Visualizations:

1. The job feature has a great impact on the output (see*Figure 2*)
2. The marital feature has an impact on our output (see *Figure 3*), but it is not significant.
3. The Education feature seems to have an obvious impact on our output (see Figure 4)
4. The day of the week and the number of days that passed by after the client was last contacted from a previous campaign (pdays) features seem to have a very weak impact on our output (see Figure 5 and Figure 10), so they were dropped from our model.
5. The contact feature seems to have an acceptable impact on our output (see Figure 12)
6. The housing, poutcome and loan features seem to have a good impact on our output, but with less weight compared to features in 1,2 and 3 (see Figure 7)



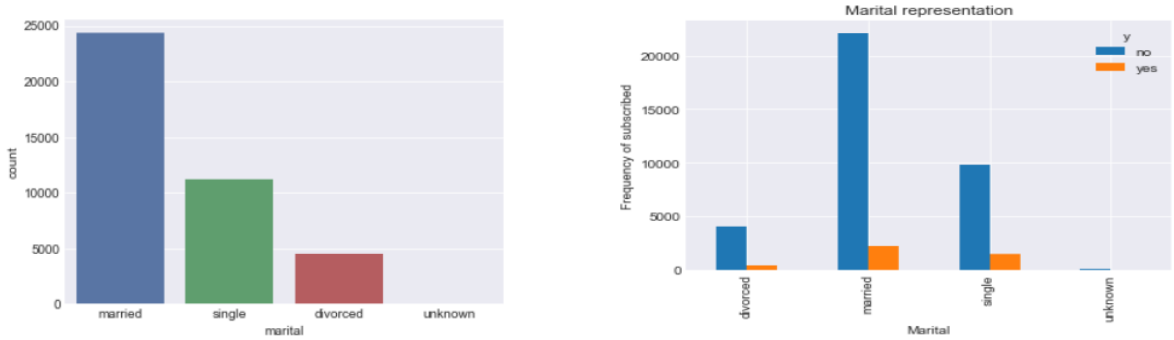
(a) (b)

Figure 1 (a) age distribution of customers (b) distribution of customers' age with Yes and No subscription.



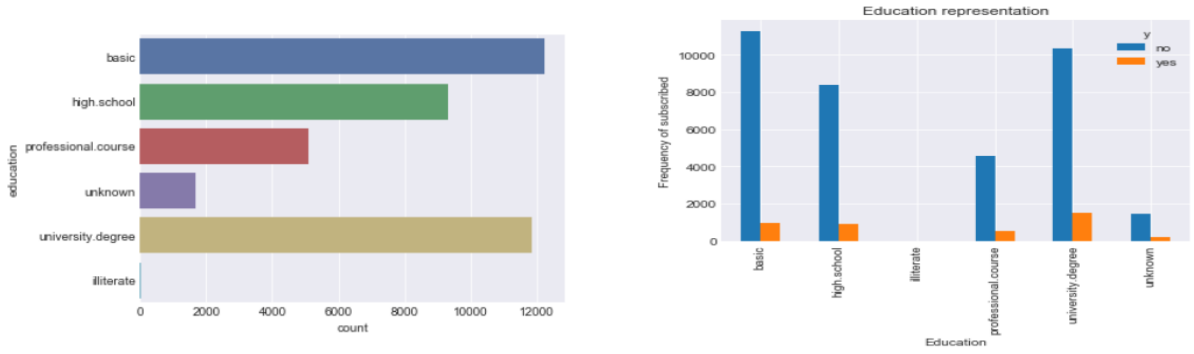
(a) (b)

Figure 2 occupation of customers in the record. (b) Number of people who said Yes and No for subscription based on their occupation.



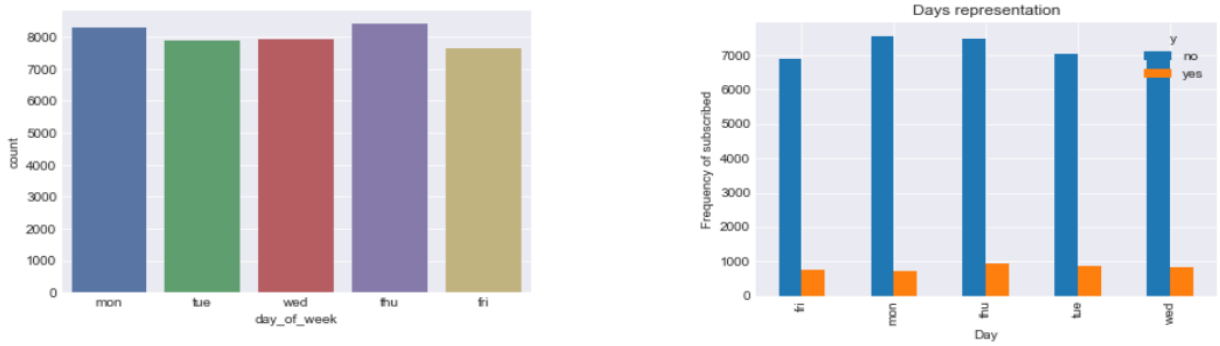
(a) (b)

Figure 3 Customers' marital status. (b) Number of customers who said Yes or No for subscription based on their marital status.



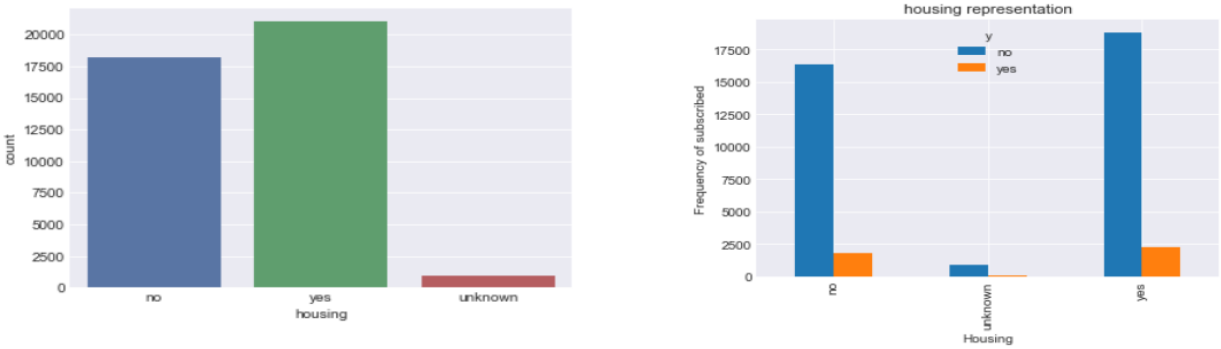
(a) (b)

Figure 4 Education of customers in record. (b) number of customers wo said Yes and No for subscription based on level of their education.



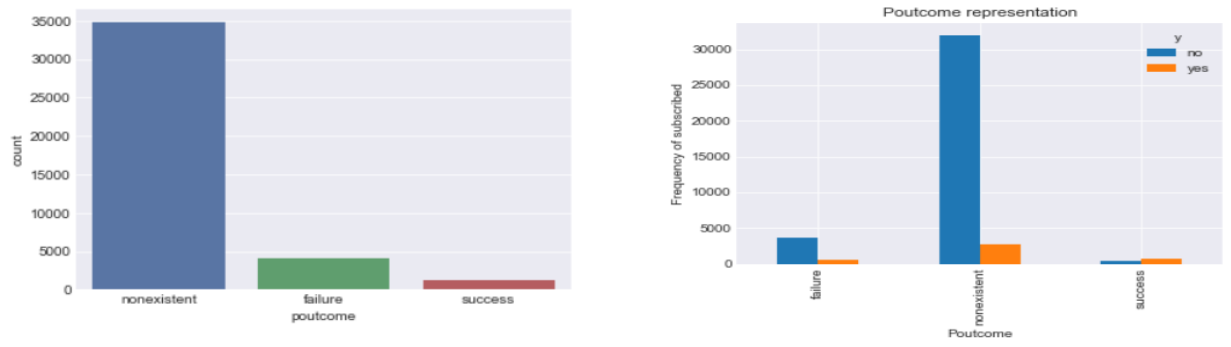
(a) (b)

Figure 5 days of the week that customer was contacted. (b) number of customers wo said Yes and No for subscription based on the days of the week.



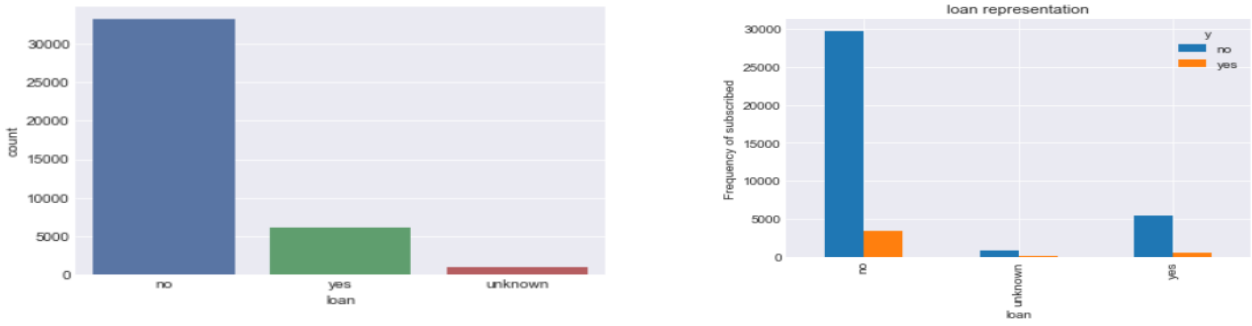
(a) (b)

Figure 6 number of customer with housing loan. (b) number of customers wo said Yes and No for subscription based on whether they have housing loan or not.



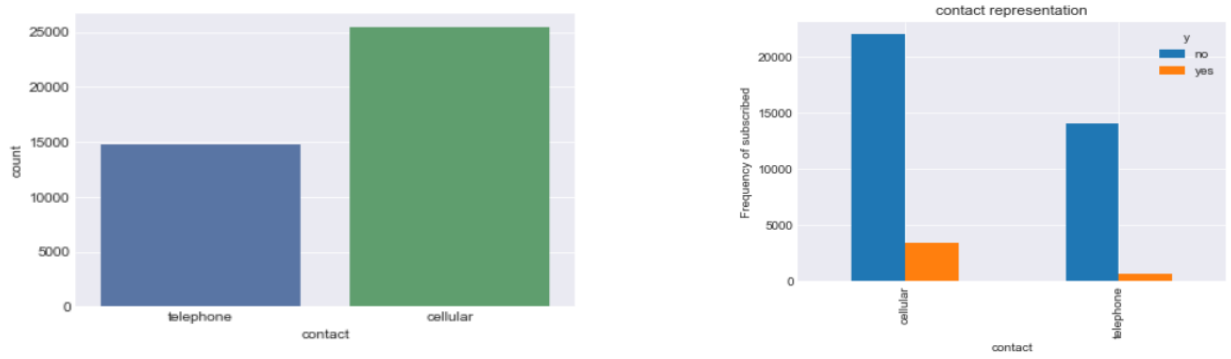
(a) (b)

Figure 7 Outcome of the customer in previous campaign. (b) Number of customers who said Yes and No for subscription based on their previous outcome campaign.



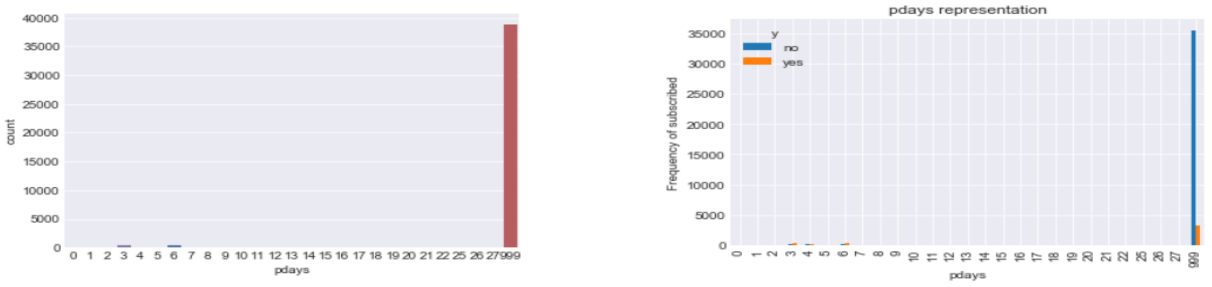
(a) (b)

Figure 8 number of customers who has personal loan. (b) number of customers wo said Yes and No for subscription based on whether they have personal loan or not.



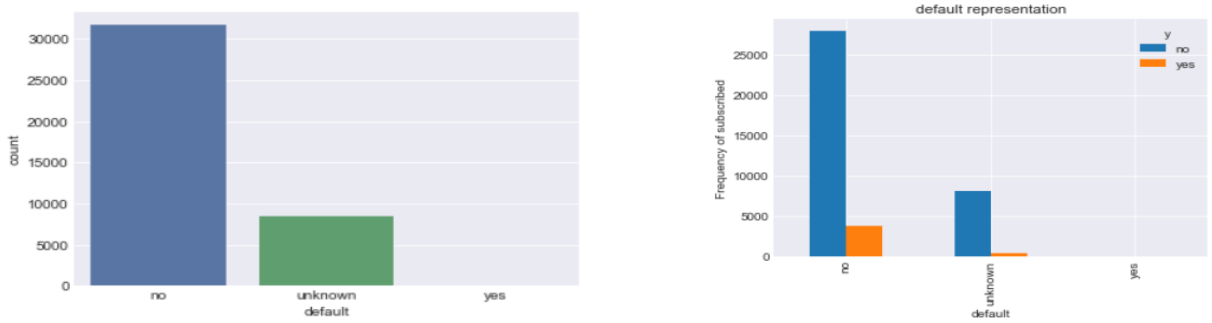
(a) (b)

Figure 9 customer's communication type. (b) number of customers wo said Yes and No for subscription based on the communication type they use.



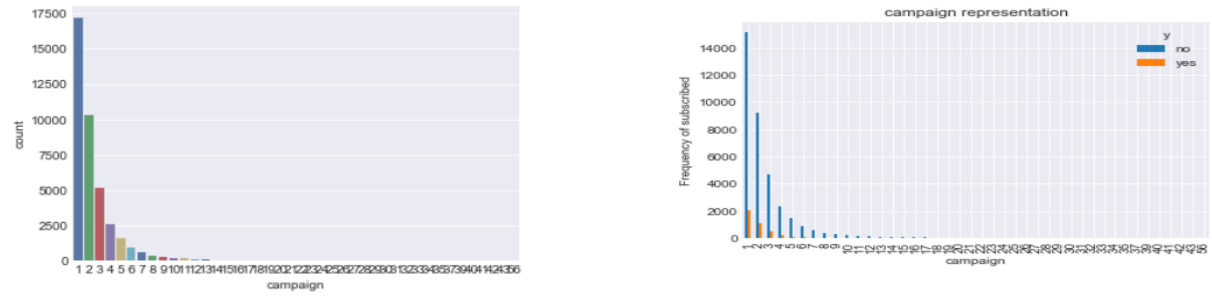
(a) (b)

Figure 10 number of days customer was contacted after the previous campaign (pdays). The 999 means, they weren't contacted. (b) Number of people who said Yes and No based on the pdays



(a) (b)

Figure 11 customer's credit by default. (b) number of customers wo said Yes and No for subscription based on whether they have default credit.



(a) (b)

Figure 12 Number of contacts during this campaign. (b) number of customers wo said Yes and No for subscription based on number of contacts during this campaign.

4.2 Transforming the categorical data to numerical

As can be seen from Table 1, there are nine categorical features in the data.csv dataset and in order to include them in our machine learning models, we will transform them to numerical values. For this purpose, we use LabelEncoder to transform these categorical values to numerical. This is a method that normalizes these labels such that they will have values between zero and the number of class minus one for each categorical label. Table 3 shows the features of the data.csv file after applying this method and it can be seen that all categorical labels are transformed to numbers which the model assigns to each class of data. After making all the elements of the data numerical, pairplot and heatmap of all the data can be plotted for comparisons. Figure 13 shows the heatmap plot for correlation between different features in the dataset.

Table 3 the first five lines of the data.csv after transforming the categorical values to numerical

	age	job	marital	education	default	housing	loan	contact	campaign	poutcome	cons_price_idx	cons_conf_idx	prime_rate	y
0	56	3	1	0	0	0	0	1	1	1	93.994	-36.4	4.857	0
1	57	7	1	1	1	0	0	1	1	1	93.994	-36.4	4.857	0
2	37	7	1	1	0	2	0	1	1	1	93.994	-36.4	4.857	0
3	40	0	1	0	0	0	0	1	1	1	93.994	-36.4	4.857	0
4	56	7	1	1	0	0	2	1	1	1	93.994	-36.4	4.857	0

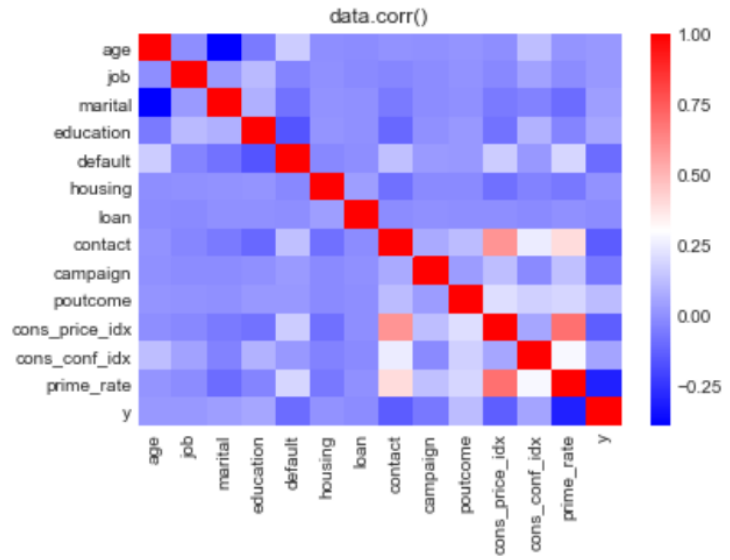


Figure 13 heatmap of correlation between features of the data.csv file

4.3 Dealing with imbalance data

Imbalance in the data refers to an issue with classification problems where the classes are nor represented equally. This happens when the number of observations in one class is significantly lower than those belonging to the other classes. In our current data.csv dataset, about 90 % of the client said 'No' and only about 10 % of the client responded 'Yes' to subscribe to the long term deposit which shows a big 9:1 ratio. This clearly shows that the data is imbalanced and is biased towards 'No' responses. To further demonstrate this problem, the original untouched data is used to determine the accuracy of our mentioned models using the logistic regression model and it was found out that the accuracy rate was about 90 % which tends to regress more towards the percentage 'No' responders. Figure 14 and Table 4 shows the confusion matrix obtained from this analysis. In these situation the prediction models developed using the conventional machine learning algorithms could be biased and inaccurate.

Table 4 confusion matrix of original data using the logistic regression classification model

Number of samples = 40181	Class0 Predicted	Class 1 Predicted
Class 0 Actual	TP=35763	FP=302
Class 1 Actual	FN=3510	TN=606

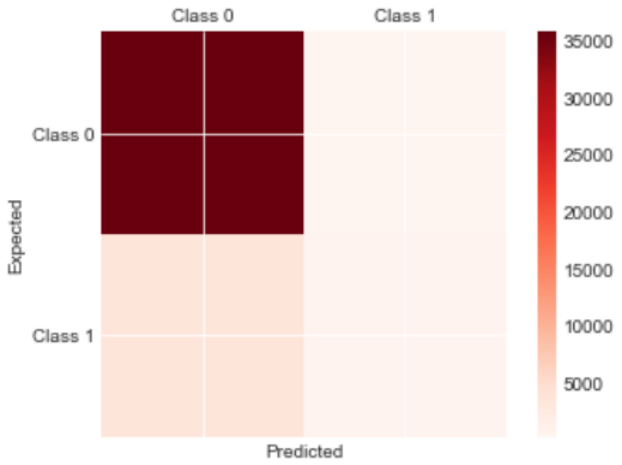


Figure 14 heatmap of the confusion matrix for the original data using the logistic regression classification model

To solve this issue, the following techniques can be used to remove the imbalance from the data:

- Oversample the data (increasing the number of "yes" in the sample)
- Undersample the data (decreasing the number of "no" in the sample)
- SMOTE (This object is an implementation of SMOTE - Synthetic Minority Over-sampling Technique)
- XG Boost

Oversampling and under-sampling are the methods of increasing the number of ‘Yes’ and decreasing the number of ‘No’ for the former one as is shown in *Figure 15*. For this project, the over sampling method was used to increase the number of ‘yes’ to match the majority group. Figure 16 shows the number of ‘yes’ and ‘no’ responders after performing oversampling on the ‘yes’ responses. To understand how it has impacted our data, a logistic model was run and the predicted accuracy was 0.7273 and the confusion matrix for this model is shown in Figure 17. It can be seen from this Figure that we have higher values in the diagonals compared to the original data shown in Figure 14.

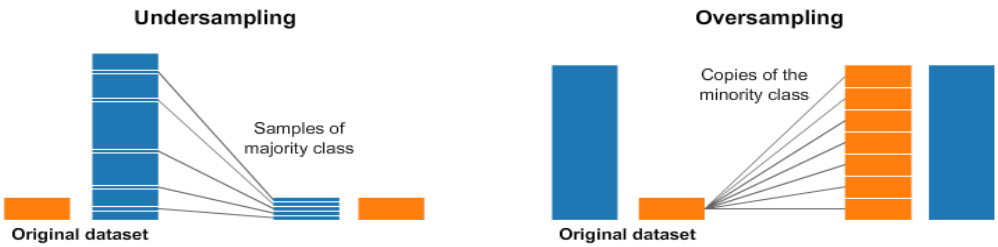


Figure 15 visual demonstration of undersampling and oversampling methods.

Source: https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets

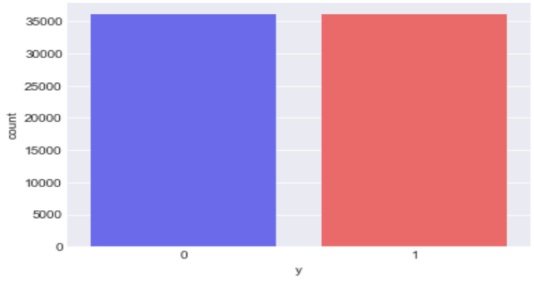


Figure 16 number of 'yes' and 'no' responses after oversampling to mitigate the imbalance issue from the data.

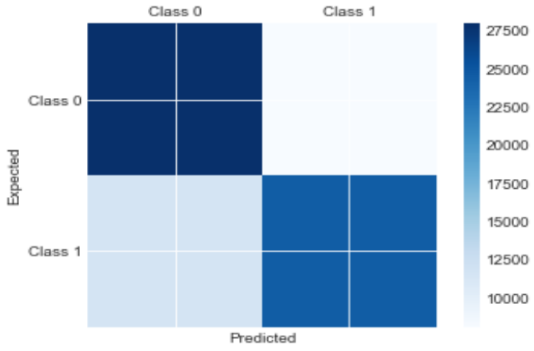


Figure 17 Confusion matrix after running the logistic regression model on the oversampled data

Synthetic Minority Over-sampling Techniques (SMOTE) is another method of treating the imbalance in the data. This technique is used to prevent the overfitting that occurs while adding replicas of minority samples to the main dataset. After applying this technique to the main model, the number of data increases from 40181 to 50504 samples, with half of them are the client who subscribed and the other half the ones that did not. After running the logistic regression on this new balanced sample, the accuracy of the prediction was 0.7724 and the confusion matrix is shown

in Figure 18. Our goal is to make the diagonal values of the confusion matrix high (the true positive and the true negative values) to get higher accuracy.

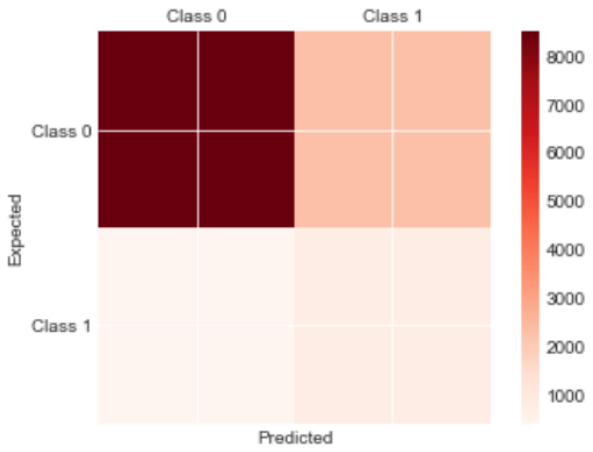


Figure 18 confusion matrix for the balanced data using the SMOTE method

Another technique to deal with the imbalance is the XG boost, which is a numerical optimization algorithm which minimizes the loss function using a gradient decent method. This technique was not implement in this project, but can be considered as an available option in the future.

After we performed all mentioned above resampling techniques to remove the imbalance in our data and applying the logistic regression model to evaluate each method, it was found that the SMOTE has led to a good accuracy in comparison to others. Therefore, we decided to use SMOTE for all performed models.

5- Methodology

The next course of action is to find the optimal machine learning algorithm with the best accuracy to be used for future predictions. To find that the following commonly used machine learning models will be evaluated each separately in the following sections:

1. Logistic Regression,
2. K Nearest Neighbor,
3. Random Forest,
4. Support Vector Machine, and
5. Neural Networks by Keras

5.1 Logistic Regression

Logistic regression is a simple but a powerful algorithm used for binary classification problems where it divides elements of a set in to two groups (i.e. 1 and 0). Even though its name has regression in it, but in reality it is a classification model. Linear regression cannot be used for problems containing binary groups as is shown in Figure 19a, therefore, logistic regression/sigmoid function is used to take input values 0 and 1 and transform them to values over the entire real number range shown in Figure 19b. The results of sigmoid function is presented as the probability of the sample belonging to class 1 or 0 based on Equation 1.

$$log\frac{p}{(1-p)} = z \;\Rightarrow\; p(z) = \frac{1}{1+e^{-z}}$$

$$z = w_0 + w_1x_1 + \cdots + w_nx_n = \mathbf{W}^{\mathrm{T}}\mathbf{X} \qquad \text{Equation 1}$$

$$\hat{y} = \begin{cases} 1 & if\ p(z) \geq 0.5 \\ 0 & otherwise \end{cases}$$

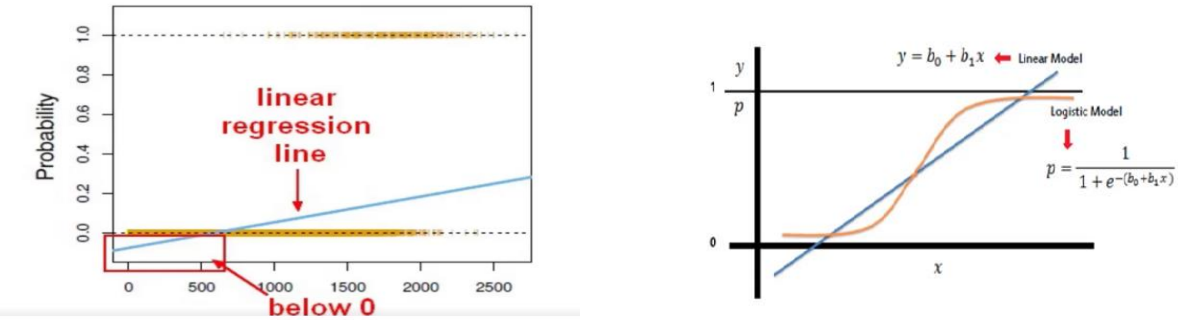


Figure 19 Linear regression vs Logistic regression for binary data

5.2 K-Nearest Neighbor

KNN is a very simple algorithm that works with any number of classes, works with only few parameters (K and distance metric), and is easy to add or remove data. It is a classification method where conditional distribution of Y given is X is estimated and then the new observation is classified to the class with the highest estimated probability. Given an arbitrary K and a new

observation x_0, this classifier determines K closest points to x_0 that we call as N_0 subset. Then the algorithm classify this new observation to a class which has majority values in the N0 dataset. This process is shown in Figure 20. Determining K value is very important in having the best classifier. Figure 20 shows that for the same amount of data, K=3 will classify to Class A and K=6 will classify as group B. For this algorithm different K values will be analyzed and the optimal value will considered in the classification.

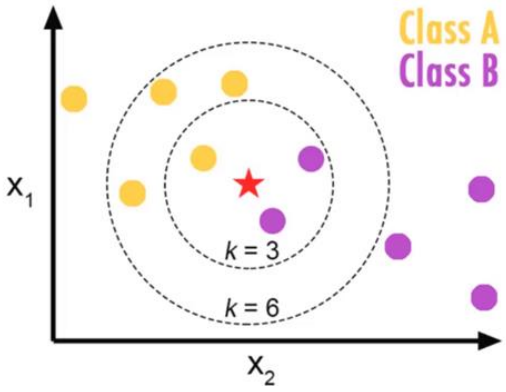


Figure 20 Graphical representation of the importance of K value

5.3 Random Forest

Random forest is an easy and flexible machine learning algorithm that produces the best results with high accuracy majority of the time. It is one of the most highly used machine learning models due to its simplicity and being applicable for both classification and regression problems. Like its name suggest, this model creates a forest from a large number of trees and make them random and then take majority vote for the results of all the trees. One of the best benefits of this method that is that there are enough number of trees in the model that classifier does not over-fit the data in random forest. Figure 21 shows an example of random forest with only two trees. The following steps are taken when creating random forest for a set of data:

1. Take a sample of size n randomly from training dataset
2. Choose p variables randomly from all existing variables
3. Create one big tree using the sample dataset with p variables
4. Repeat step three B number of times
5. Finally take majority votes for the results of all B trees

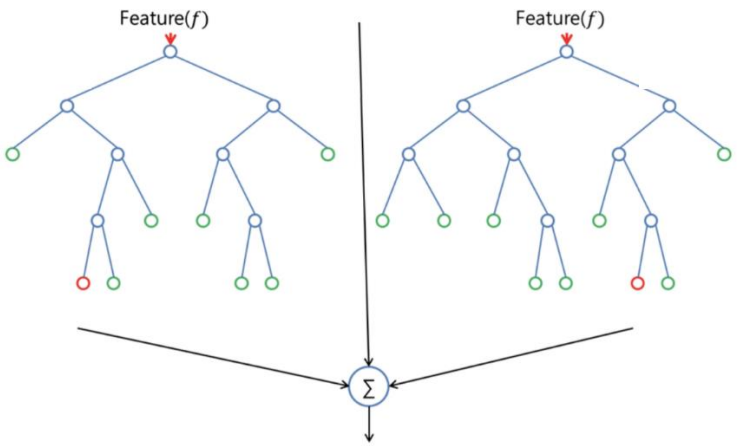


Figure 21 Two random trees showing how the combination of trees becomes random forest

Source: https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

5.4 Support Vector Machine

Support Vector Machine is a supervised machine learning algorithm where it outputs an optimal hyperplane which categorizes the new examples. In two dimensional space, this hyperplane is linear line in which each class lays on side of the line. The equation of the hyperplane in the multidimensional space has the form of Equation 2 where p represents the number of dimensions and B are called normal vectors orthogonal to the surface of the hyperplane. Sometimes when the data is not separable using a simple linear line, transformation can be applied by adding the z-axis as can be seen in Figure 22. Also, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p = 0 \qquad \text{Equation 2}$$

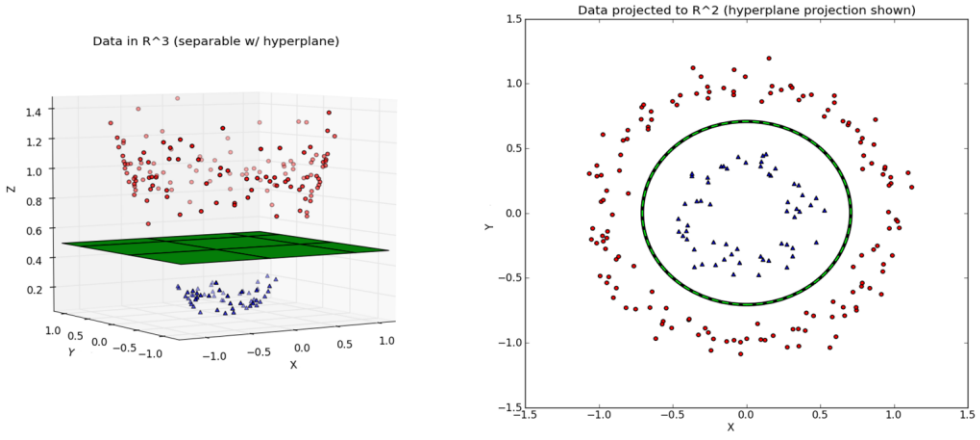


Figure 22 Elevating the power of the data into the z dimension

5.5 Neural Networks

Neural networks is a machine learning algorithm that is modeled to mimic human's brain and recognize patterns. Neural networks help to cluster and classify and can be considered as clustering and classification layer on top of the data that is stored and managed. Neural networks are typically organized in layers. Layers are made up of a number of interconnected 'nodes' which contain an 'activation function'. Each unite in one layer is known as perceptron which can compute a continuous output by using logistic function by using the output of the last layer as input. Patterns are presented to the network via the 'input layer', which communicates to one or more 'hidden layers' where the actual processing is done. Figure 23 shows a typical neural network algorithm pattern with softmax activation function that is used to predict multi classes output. Sigmoid function is used to predict the binary classes.

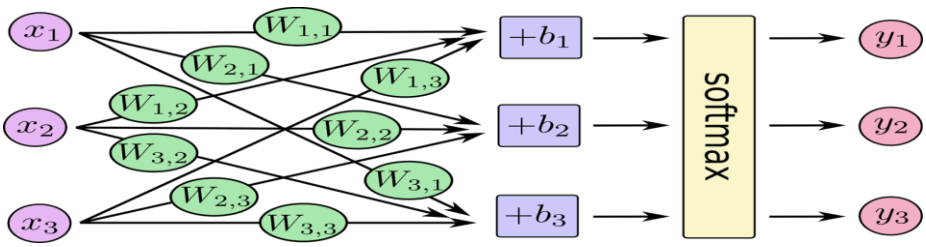


Figure 23 Graphical representation of neural network's layers and nodes

6- Results and Discussions

This section present the results of the five implemented machine learning algorithms described in the previous section using Python program. These models are created based on the balanced data from the SMOTE technique that was described in Section 4.3. Accuracy and the confusion matrix obtained from each model will be compared with each other at the end to find the best algorithm to be used to predict the client's decision in the future.csv dataset.

6.1 Logistic Regression

This section present the results of the logistic regression model based on the balanced training data.csv dataset. The following is the classification report and the confusion matrix for this model. Figure 24 shows graphical representation of the confusion matrix.

```
classification_report is:
              precision    recall  f1-score   support

           0       0.95      0.79      0.86     10813
           1       0.26      0.66      0.37      1242

   micro avg       0.77      0.77      0.77     12055
   macro avg       0.61      0.72      0.62     12055
weighted avg       0.88      0.77      0.81     12055

confusion_matrix is:
[[8497 2316]
 [ 428  814]]
```

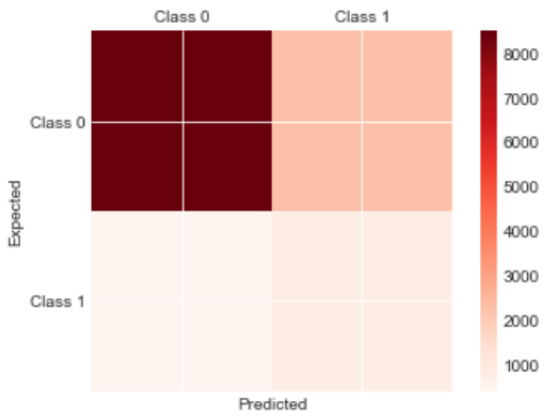


Figure 24 graphical representation of the confusion matrix for Logistic regression

To comprehend the accuracy range sensitivity, ROC is used. Overall accuracy is based on one specific cutpoint, while ROC tries all of the cutpoint and plots the sensitivity and specificity.

So when we compare the overall accuracy, we are comparing the accuracy based on some cutpoint. The overall accuracy varies from different cutpoint. The result of ROC for this logistic regression is shown in Figure 25. We can conclude that the selected metrics evaluation can lead to different results (i.e. accuracy score, classification report, and ROC).

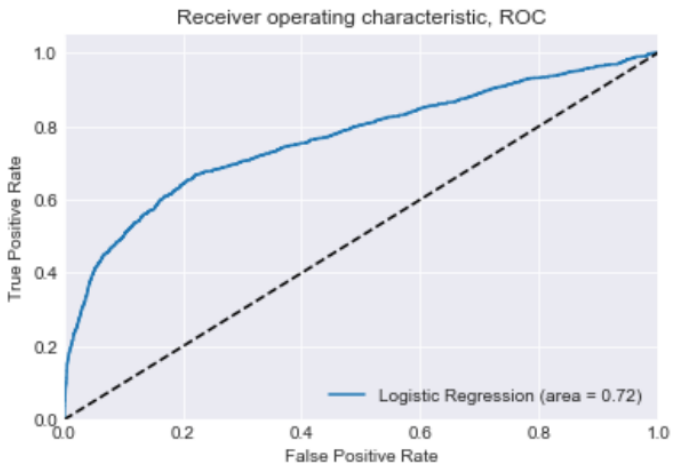


Figure 25 Receive Operating Characteristic (ROC) for the logistic regression model

It can be seen from previous figures that the accuracy is equal 88 % with the confusion metric while the accuracy is equal 72 % with the ROC method (the area under the curve).

6.2 K Nearest Neighbor (KNN)

For the KNN classification model, selection is K is an important aspect of the modeling. For this purpose, the plot of model error rate with respect of the value of K is plotted using the Elbow method as depicted in Figure 26. It was determined that with the K=1, the model provided the best accuracy (uncommon though!). It appears that we have inverse Elbow where the rate of error increased when the value of K gets larger. The interpretation is perhaps related to the oversampling of the data or probably due to the overfitting issue. More research and investigation are required to be performed.

The following are the classification report and the confusion matrix for the KNN model. Figure 27 shows graphical representation of the confusion matrix. The result of ROC for this model is shown in Figure 28.

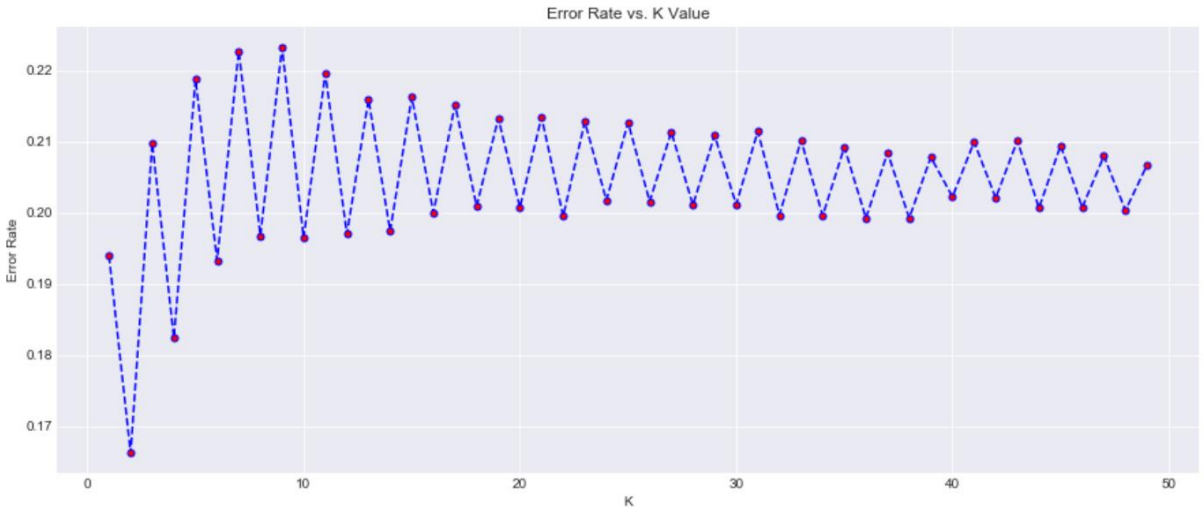


Figure 26 Error rate vs K value plot

```
classification_report is:
              precision    recall  f1-score   support

           0       0.92      0.86      0.89     10813
           1       0.22      0.34      0.27      1242

   micro avg       0.81      0.81      0.81     12055
   macro avg       0.57      0.60      0.58     12055
weighted avg       0.85      0.81      0.82     12055

confusion_matrix is:
[[9291 1522]
 [ 816  426]]
```

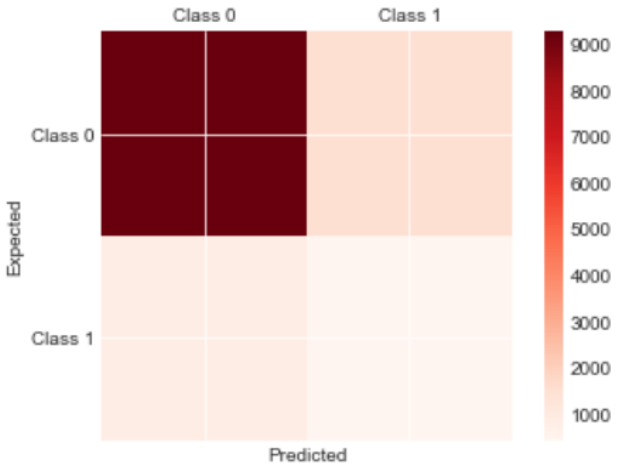


Figure 27 confusion matrix graphical representation of the KNN model

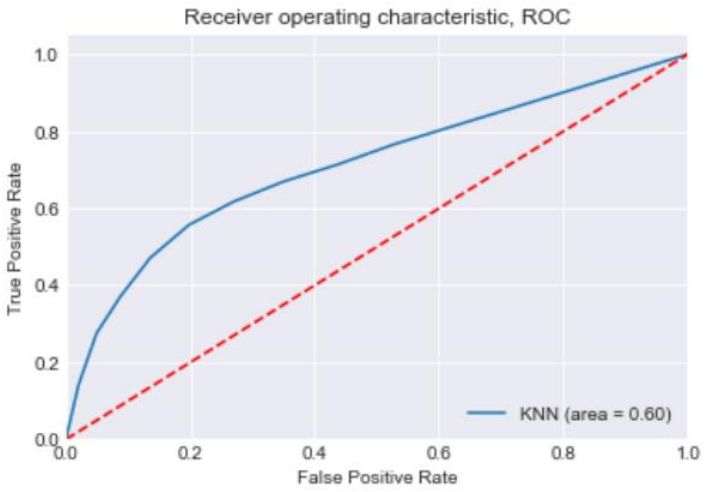


Figure 28 Receive Operating Characteristic (ROC) for the KNN model

6.3 Random Forest

The section describes the random forest method and it presents the results of the training this model with the balanced data.csv dataset. The following is the classification report and the confusion matrix for this model. Figure 29 shows graphical representation of the confusion matrix. The result of ROC for this random forest model is shown in Figure 30.

```
classification_report is:
              precision    recall  f1-score   support

           0       0.92      0.97      0.94     10813
           1       0.49      0.29      0.37      1242

   micro avg       0.90      0.90      0.90     12055
   macro avg       0.71      0.63      0.65     12055
weighted avg       0.88      0.90      0.88     12055

confusion_matrix is:
[[10437   376]
 [  879   363]]
```

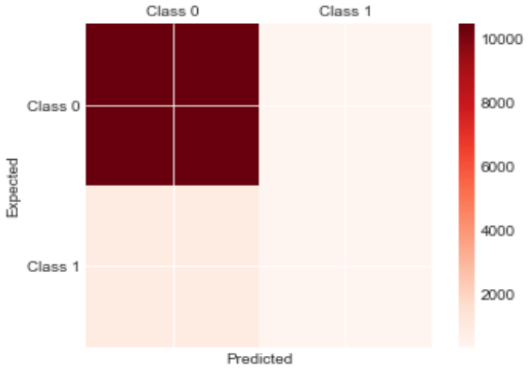


Figure 29 confusion matrix graphical representation of the random forest model

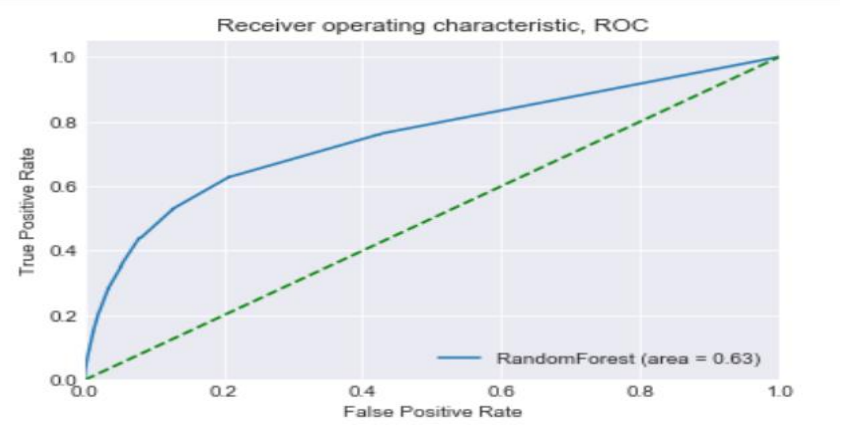


Figure 30 Receive Operating Characteristic (ROC) for the Random Forest model

6.4 Support Vector Machine (SVM)

The support vector machine is the fourth machine learning model implemented on the balanced data.csv. The following is the classification report and the confusion matrix for this model. Figure 31 shows graphical representation of the confusion matrix. The result of ROC for this model is shown in Figure 32.

```
classification_report is:
              precision    recall  f1-score   support

           0       0.94      0.84      0.89     10813
           1       0.27      0.50      0.35      1242

   micro avg       0.81      0.81      0.81     12055
   macro avg       0.60      0.67      0.62     12055
weighted avg       0.87      0.81      0.83     12055

confusion_matrix is:
[[9109 1704]
 [ 626  616]]
```

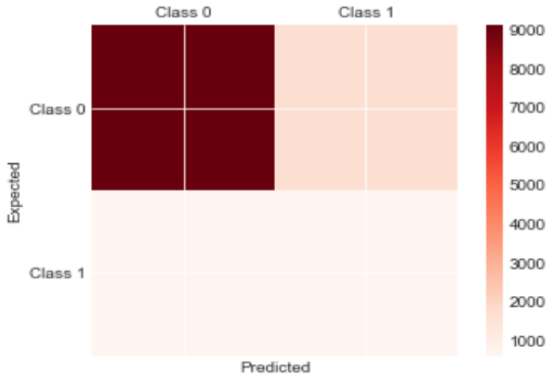


Figure 31 confusion matrix graphical representation of the SVM model

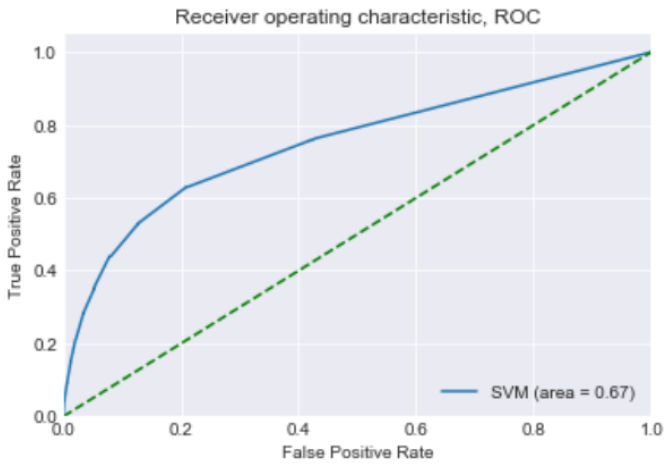


Figure 32 Receive Operating Characteristic (ROC) for the SVM model

6.5 Neural Networks Using Keras

This is powerful machine learning algorithm that has been implement on the balanced data.csv file. When we used this model, we implemented two types of activation functions: the rectified linear unit (Relu) and hyperbolic tan (tanh). Later, the sigmoid activation function is used to predict the output layer. It was found out that their prediction accuracy was almost the same. The following is the classification report and the confusion matrix for this model. Figure 33 shows graphical representation of the confusion matrix. The result of ROC for this Neural Networks model is shown in Figure 34.

```
classification_report is:
              precision    recall  f1-score   support

           0       0.94      0.90      0.92     10813
           1       0.38      0.53      0.44      1242

   micro avg       0.86      0.86      0.86     12055
   macro avg       0.66      0.72      0.68     12055
weighted avg       0.89      0.86      0.87     12055

confusion_matrix is:
[[9730 1083]
 [ 581  661]]
```

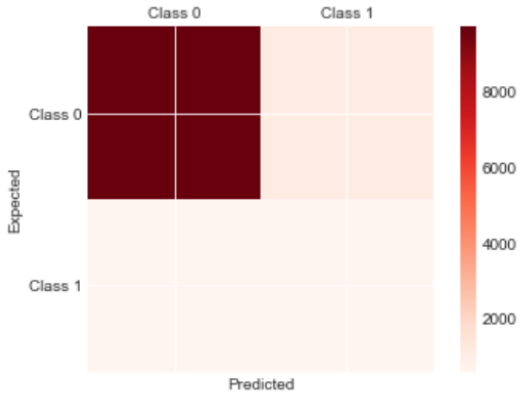


Figure 33 confusion matrix graphical representation of the Neural Networks model

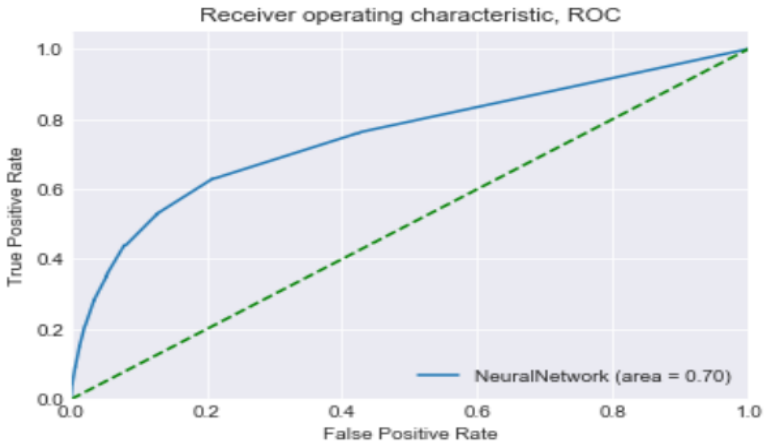


Figure 34 Receive Operating Characteristic (ROC) for the Neural Networks model

6.6 Summary of the Methods.

Table 1 shows the summary of the accuracy for the five implemented machine learning classifiers. It can be seen that the Neural Network model provides the best accuracy and will be used for prediction of future.csv file. Random forest gives a good prediction as well and could be an option to be used.

Table 5 accuracy of five machine learning algorithms

Model	**accuracy**	
	Classification precision	ROC
Logistic Regression	0.88	0.72
K Nearest Neighbor	0.85	0.60
Random Forest	0.88	0.63
Support Vector Machine	0.87	0.67
Neural Networks	0.89	0.70

6.7 The Future Prediction

In our model, we recall the future.csv as x_new and then we drop the same feature that we dropped in our original data which are days of week and pdays. All the categorical features in the future.csv dataset have been transformed to numerical value. Then the data was scaled to prepare it for neural network (the optimal model).

Since we are not certain about the output of the future, we feed two types of data into our neural network classifier.

The oversampled data (sm_data_x and sm_data_y) throgh SMOTE.

Since most of clients do not subscribe in our original data, we expect to have more ‘no’ answers compared to ‘yes’ in our future prediction. However, we got the opposite with more ‘yes’ answers compared to the ‘no’ answers as shown below.

```
output_future_reqn=pd.DataFrame(output_future_reqn)
print('The "yes" value is')
print((output_future_reqn[0]=='yes').sum())
print('The "no" value is')
print((output_future_reqn[0]=='no').sum())
```

```
The "yes" value is
818
The "no" value is
189
```

Original data before resampling (Data.cvs) with (data_train_x and data_train_y)

In this case, the future prediction match our expectation. We got more ‘no’ answers compared to the ‘yes’ answers as shown below.

```
output_future_req_originaln=pd.DataFrame(output_future_req_originaln)
print('The "yes" value is')
print((output_future_req_originaln[0]=='yes').sum())
print('The "no" value is')
print((output_future_req_originaln[0]=='no').sum())
```

```
The "yes" value is
288
The "no" value is
719
```

Since we do not have confidence about our oversampling data (we do not know exactly which features duplicated to increase the number of samples). That is why we used two types of data (feeding) in our predictions and also we don’t know the exact answer. The comparison of prediction for the future.csv dataset is shown in

Figure 35 in which ‘1 means yes’ and ‘0 means no’.

The categorical presentation of the futures data prediction (yes or no) is described below based on feeding the oversampled training data vs the original data to our Neural Network algorithm.

Future prediction with the oversampled training data

1	no
2	no
3	no
4	no
5	no
6	no
7	no
8	no
9	no
10	no
11	no
12	no
13	no
14	no
15	yes
16	no
17	no
18	no
19	no
20	no
21	yes
...	...
999	yes
1000	yes
1001	yes
1002	yes
1003	yes
1004	yes
1005	yes
1006	yes
1007	yes

Future prediction with the original training data

1	no
2	no
3	no
4	no
5	no
6	no
7	no
8	no
9	no
10	no
11	no
12	no
13	no
14	no
15	no
16	no
17	no
18	no
19	no
20	no
21	no
...	...
999	no
1000	yes
1001	no
1002	yes
1003	yes
1004	no
1005	no
1006	yes
1007	no

```
In [146]: X_new.head()
```

Out[146]:

	age	job	marital	education	default	housing	loan	contact	campaign	poutcome	cons_price_idx	cons_conf_idx	prime_rate	y
0	42	1	1	0	1	0	0	1	1	1	93.994	-36.4	4.857	no
1	41	4	1	0	0	0	0	1	2	1	93.994	-36.4	4.857	no
2	34	9	1	1	0	0	0	1	1	1	93.994	-36.4	4.857	no
3	54	5	1	1	1	0	0	1	1	1	93.994	-36.4	4.857	no
4	48	1	1	0	0	2	0	1	1	1	93.994	-36.4	4.857	no

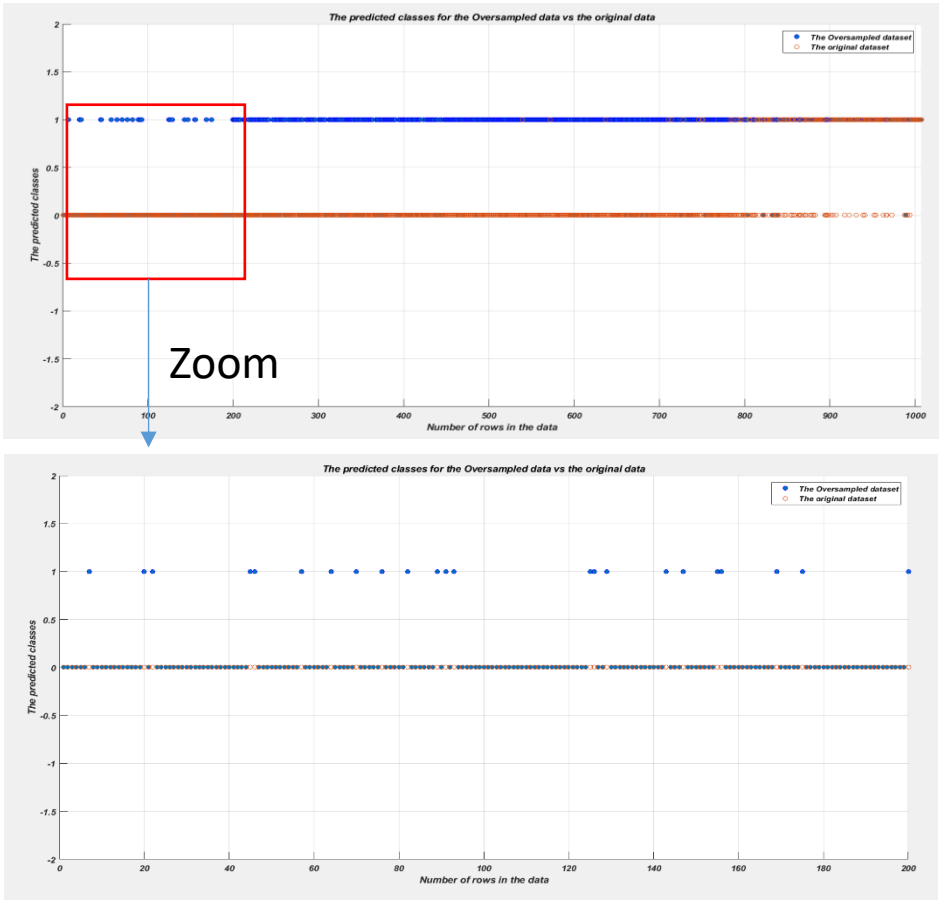


Figure 35 comparison of prediction using the oversampled and the original datasets.

7- Conclusions

It was determined that the neural networks provide most accuracy and it due to the type of the activation functions. The rectified linear unit and hyperbolic tan (tanh) have been used through the hidden layer and later the sigmoid function has been implemented to predict the binary classification.

Based on the provided data, the company should target the clients who have responded positively to the given features and take the benefits from that. The company can follow the information gain from each feature to evaluate their customers and find out whom are interested to subscribe for long term deposit.

In the prediction of the future data, we used predict_classes functions, but it would be helpful if the user use other types of function and compare the results.

8- Recommendations and Future works

In machine learning there is no right answer (No free lunch) and always with further investigation and using different classifiers, result can be improved. It would be nice if one can do deep study and understand the parameters and factors inside each algorithm rather than only just applying them. The accuracy can be more sensitive to some parameters compare to the others, therefore, we recommend a further parametric study and accuracy sensitivity analysis.

- We used different methods to deal with the unbalance data, but later we found that XGboost can be a good example to deal with the imbalances in data (tune the scale weight of the data).
- The accuracy improved when we used neural network classifier, but may be it can be achieved more accuracy with other machine learning classifiers which are not mentioned in our study.
- Since we do not know the future output, it would be great if one used the prediction data and upload it on the Kaggle competition and see the accuracy of the prediction.

Reference

- Chen Ji, Han Yu, Hu Zh & Lu Yi. 2014. Who will Subscribe A Term Deposit?
- Choong Alvin, 2017, Predictive Analytics in Marketing A Practical Example from Retail Banking
- Li Susan, 2017, Building A Logistic Regression in Python, Step by Step.
- How to handle Imbalanced Classification Problems in machine learning? Source: https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/
- How to Handle Imbalanced Classes in Machine Learning. Source: https://elitedatascience.com/imbalanced-classes
- James G., Witten D., & Robert. 2016. An Introduction to Statistical Learning with Applications in R,

Acknowledgements

The authors of this report would like to acknowledge Professor Louis Plebani for his continues support and putting his time and effort during class ISE 364. Finishing this project wouldn’t have been possible without his notes, comments and other class material.

Appendix

Python Jupyter Notebook