MODEL GENERATION TO IDENTIFY THE IDEAL CUSTOMER

BY: S M SAZID ISLAM

Problem Analysis:

In this problem, X enterprises is an organization that is expanding its business into the banking sector. With the goal of capital gathering, they planned to launch and market a financial product named as **"X Platinum Deposit"**. To take the campaign forward, the organization decided to make cold call the potential customers and make them buy this new product.

As an analyst, I have analyzed the historical data provided and generated a model to find the ideal customer of this financial product.

Section A: Summarization

This section includes a summarization of the dataset that was provided using statistical analysis. To begin with, the important functions has been imported, numpy as **np**, pandas as **pd**, seaborn as **sns**, matplotlib as **plt** are being used to make the statistical analysis.

A dataset of 4000 datapoints with 15 inputs has been given which are named as features in the python code. A sum of duplicated data is calculated using the **duplicated().sum()** function.

The first 10 rows of dataset are shown using the **head ()** function to get a brief idea of the data. To understand about the customers, **unique** function is used which let us know the job sectors and education level of the potential customers of the dataset. Since, we are concerned about the output (y), we made a value counts of the output. From there we can see that there are 846 customers who have agreed to buy the financial product and 3154 customers declined.

To make a detail statistical analysis, a number of graphs has been used to understand the relationship of the different variables with the output(y). The graphs are displayed below and explained accordingly:

Figure.1:

This figure shows the relationship of Output(y) with duration of the last conversation with the customer in seconds. This shows that higher number of customers accepted the deal when they had longer conversation. This means that those who had a long conversation regarding the product are more interested in purchasing the financial product rather than those who talked for less time. Hence, this can be stated that duration is a very impactful factor for successful sales of the product, and X enterprises should hire more qualified and experienced call center agents who can describe in detail about the financial product. In addition, as per the graph, it can be stated that divorced people will be more influenced with this sales call compared to the people with other marital statuses.

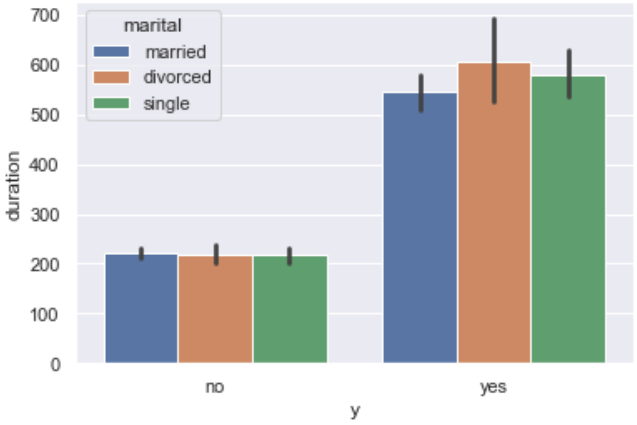


Figure 1: Duration Vs Output(y)

Figure.2:

This figure shows the relationship between balance and output results. The higher the balance at their bank, the more interested people are for such financial products. Hence, the amount of money a person belong is a very crucial factor which must be considered while marketing the product. X enterprises should make a list of the upper-class people and make conduct in person meeting to talk about the product.

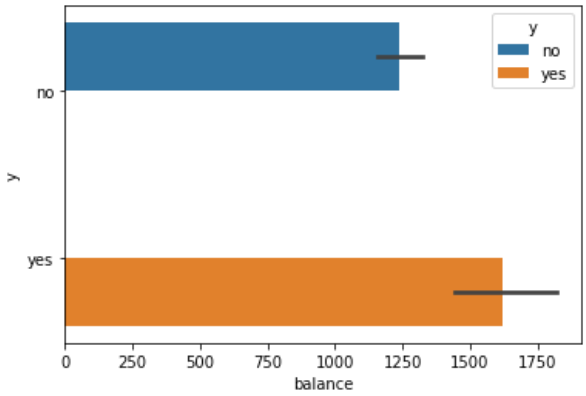


Figure 2: Balance Vs Output (y)

Figure.3:

Here, in this graph, this can be stated that higher the number of previous calls, the higher the number of potential customers agreed to buy the product. This depicts that making repetitive calls is very important in order to make a make a successful sales call.

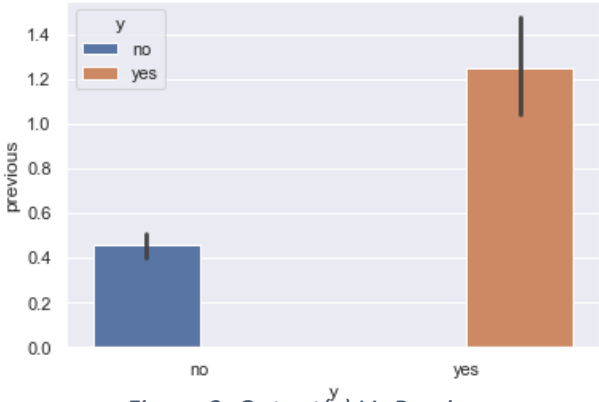


Figure 3: Output(y) Vs Previous

Figure.4:

In this graph, it is showing the relationship between people from different occupation and their output to the call, this can be stated that management people are the biggest group who agreed to go forward with the product, next comes technician, blue-collar and the people working in admin. Surprisingly, there are some students also who were positive with the product. Hence, this makes us understand that the job roles we must consider while selecting people to make calls.

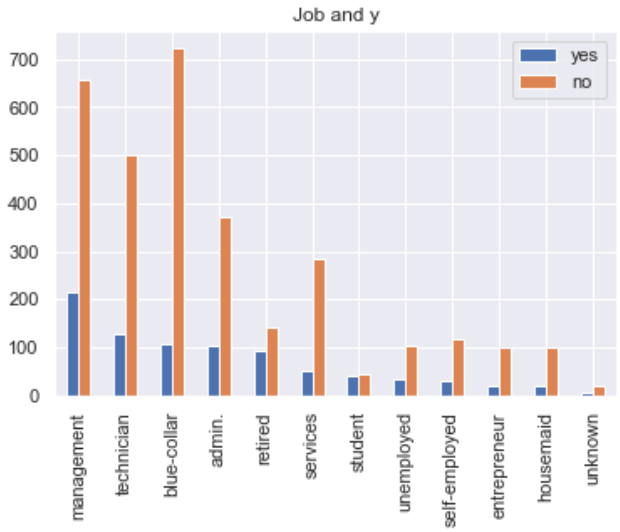


Figure 4: Count of y Vs Job

Figure.5:

This is the Correlation chart which is used to determine the relationship among the variables and to analyze the dependency of variables on each other. There are normally 3 types of correlation among variables, positive, negative and negative and zero correlation. From the graph, this can be evaluated that output(y) has a positive correlation of 0.48 with duration. This means that duration which is the last contact duration with a customer is impactful to a positive output. Higher the duration, higher the acceptancy of the customers.

On the other hand, there is strong negative correlation between pdays and poutcome. This means that higher the number of days passed after the last contact, it becomes more difficult to sell the product to the desired contact, poutcome decreases.

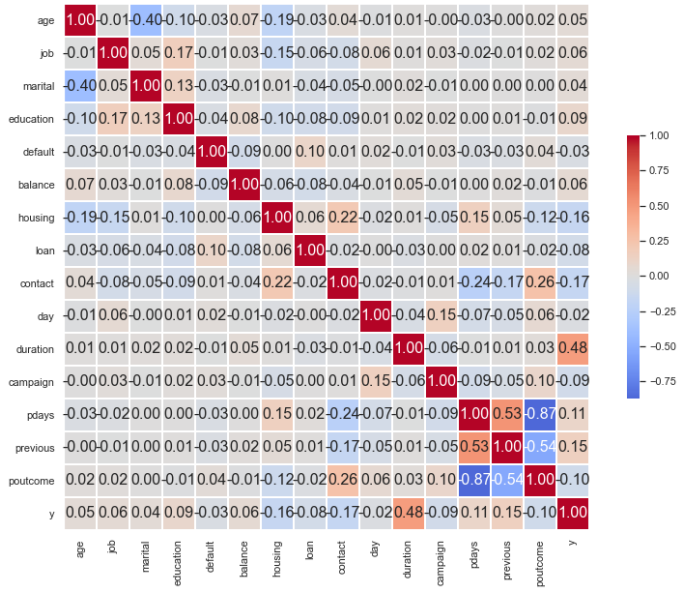
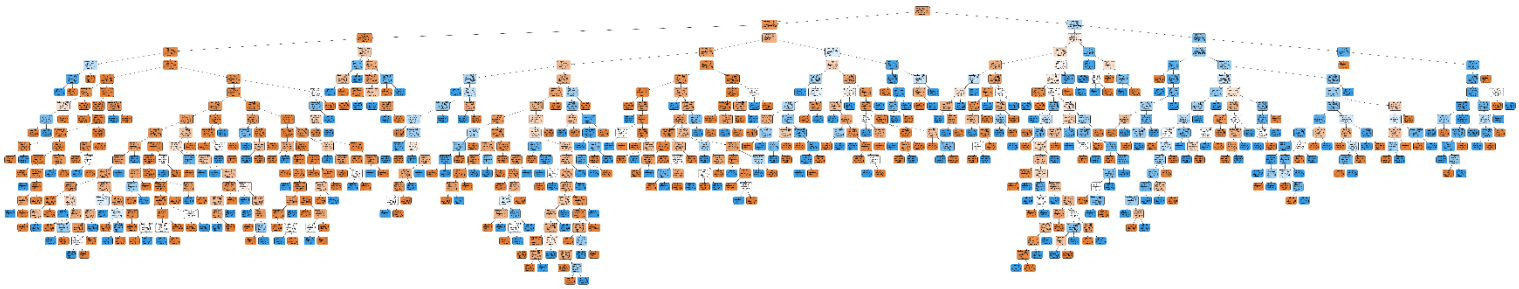


Figure 5: Correlation chart

Section B: Exploration



A Decision tree has been used to make a classification of the influencing factors in the data because it identifies the most significant variable and its value that gives the best homogenous sets of population. Here, Gini index concept has been used which measures the degree or probability of a particular variable being wrongly classified when chose randomly. Here, the root node is the duration node with a gini index value of 0.337 which is the quality of the split and this means 34% of the samples would go in one direction. The duration is <- 477.5, which means that every customer who had the call for less than 477.5 seconds will follow the True arrow to the left of the box and the rest will follow the False arrow to the right. The sample size is 3200 which is 80% of the total dataset as this is the first step.

In the next step, the root node splits into two other nodes called the decision nodes, the arrow to the True arrow is directed towards a box of 2654 customers. This means that 2654 customers had conversation for less than 477.5 seconds. However, the False arrow is directed towards a box of 546 customers which means that 546 customers had a conversation for more than 477.5 seconds. The splitting process goes on, the two boxes get further split into other variables such as job, previous, balance, age and the split go on until there is no more dataset to train. The important variables identified from the decision tree are duration, job, balance, previous, age and pdays because they are more important variables based on their feature importance. There are also some unimportant variables such as pdays, campaign, previous because these variables do not split much and terminates early.

The accuracy of the decision tree has also been checked in Python, the accuracy is 76% which indicates that the decision tree is performing well with this dataset.

Section C: Model Evaluation

The classification models those were used to train the dataset are **Decision Tree**, **Random Forests**, **Naïve Bayes Classifier** and **k-nearest neighbors** (KNN).

KNN: It is a kind of supervised learning classification model where the data is classified based on the nearest neighbor's characteristics. The exact number of neighbors to be considered is given input and based on that it makes the predictions. It is also called lazy learning algorithm because it does not have a training phase while it memorizes the training dataset. When the model is trained using our dataset, the accuracy is **81%** which shows that the model is working very well with the test data. The ratio of true positive and true negative defines the model accuracy. The recall value however is low compared to the other models of about **28%** because it considers all the data near its proximity.

Random Forest: It is a powerful and easily adaptable supervised machine learning algorithm that grows and combines a number of decision trees and creates a forest. It can be used for both classification and regression purposes. It is used here because it is a classification analysis and the variables are categorical. The accuracy of this model with the dataset provided is **84%** which implies that this is the best-fitted model. The recall value is **32%**, this means that the model can correctly identify the actual positives out of all the positive data in the dataset while having an accuracy level of 84%. Random forest can overcome the over-fitting problem by using random-sampling approach.

Naïve Bayes Classifier: The Naïve Bayes is a popular machine learning algorithm which works on the basis of probability. This model does not consider the unnecessary features of a given dataset to predict the result for which the this gives more accurate and valid result than other algorithms or models. Here, Gaussian Naïve Bayes Algorithm is used to train this dataset because it is a classification problem. The accuracy of this model is **77%** which is less than the other models used, however the recall value is **41%** which indicates that the model can more accurately consider the actual positives out of all the positive data.

Decision Tree: This is an algorithm model which breaks the data points into various decision nodes and attains a tree structure. There is a root node which split into child nodes and the tree keeps evolving until the data points at a specific child node is pure. The accuracy of this model is **76%** and the recall value is **45%**. This shows that the effectiveness of the model is not that high compared to the other classification models but its efficiency of positive data selection is better than other models.

Section D: Final Assessment

As all the different models are evaluated by training the same dataset in the models, finally the best model applicable for the particular historical dataset named "cwk_data_lixsi12.csv" has been identified. According to my analysis and results, the best

fitted model is the Random Forest model. The reasons behind selecting this model as the proper and well-fitted model are, first, the accuracy is the highest among all other models used, 84% of accuracy. Accuracy is a very important performance metric which is defined as the ratio of true positives and true negatives to all positive and negative observations. It is the percentage of times the model will correctly predict an outcome out of the total time it made predictions. Next, the recall value of 32% is good compared to the accuracy. Stating about its coherence with the case provided, the organization, X enterprises is using a big dataset and it will be very expensive and time consuming to cold call such a big number of potential customers. Using this model, it can be identified more precisely that an individual is a better prospect to contact. This will eventually allow the organization to work with more potential customers and success rate of marketing the financial product will be higher.

Section E: Model Implementation

This model has been designed for **X** enterprises to market a financial product called, **X Platinum Deposit**.

As per the analysis, best model selected is Random Forest, to test and run the model, a parameter grid has been created based on random search. Depth, features, sample leaf, sample split and estimators are considered. Next, the best model is created for Random forest and initiated and the data are fitted to the model. The outcome comes is 1215 fits. Next, the grid search estimator is used and using confusion matrix, the accuracy is found to be 83%. The model is them optimized as it is shown in the python file provided with this file.

Section F: Business Case Recommendations:

In a nutshell, the whole historical dataset has been analyzed using 4 different classification models, Decision Trees, Random Forests, Naïve Bayes Classifier and the k-nearest neighbor approach. According to the analysis of the historical data, Random Forest will be the proper classification model to be used for this dataset, it will help the organization to keep the expense low while reaching the potential candidates with this product. As recommendation, the organization, X enterprises should collect similar datasets from other financial institutions and make data analysis which will allow the organization to reach higher number of potential customers.

References:

1. 2022. [online] Available at: <https://vitalflux.com/accuracy-precision-recall-f1-score-python-example/> [Accessed 13 January 2022].