Foundational Business Analytics
Coursework 2021

Overview:

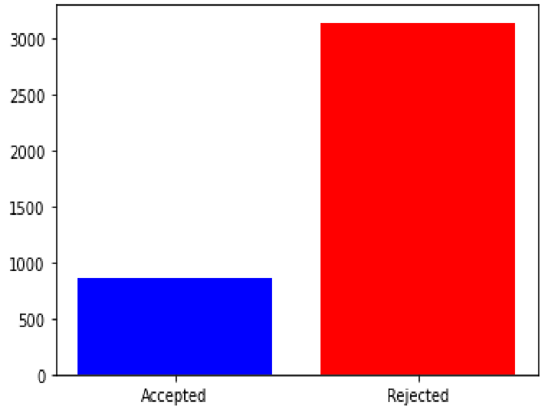
N/LAB Enterprises is about to launch a financial product, **N/LABs platinum deposit**, to attract deposits from customers. The deposit is held for a year and a certain interest rate is offered for keeping the money in the enterprise. The main challenge for the company is to target a customer base and reach out to the potential buyers. In this regard, the data base of its predecessor is available to the N/LAB. As the product is almost similar but branded in a different way, the available record of the previous phone calls to customers is going to be handy in not just sifting non-serious buyers, but also in reaching out to potential customers. Therefore, our goal is to use the best models to predict who is going to buy our product form the previous conversation with the consumers.

Section A: Summarization

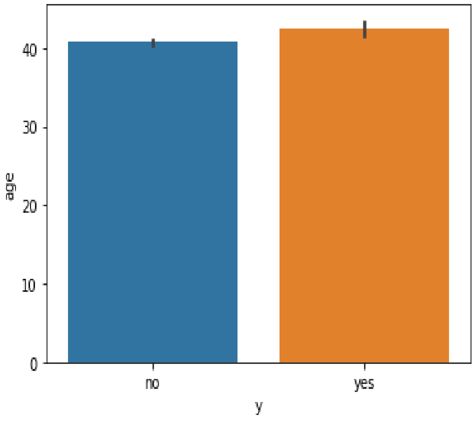
The dataset we have is organized into inputs and result. Characteristics for each caller include his/her age, job profession, marital status, level of education, default status, balance in the bank account, housing loan history, personal loan record, medium of contact, previous call day of the month, number of contacts during campaign, number of days passed-by to previous call, and prior number of contacts performed before this campaign, and the past outcome of the campaign. Based on the success and failure of the call(outcome), the data shows following result.

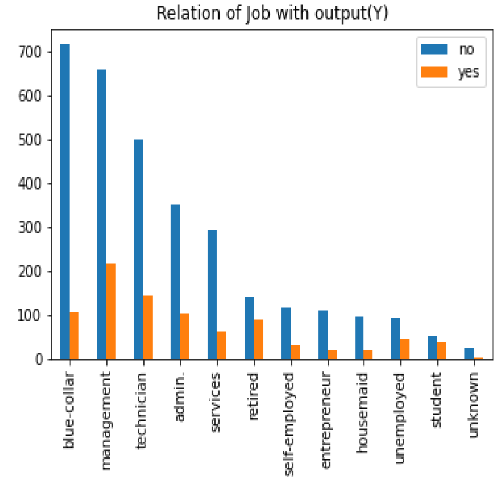
This bar plot shows that out of 4000 calls, only 864 callers agreed to participate in the financial product of the N/LAB's predecessor. The segregation in success and failure of previous campaign is important to set expectation for seeking prospective customers of N/LAB platinum deposit.

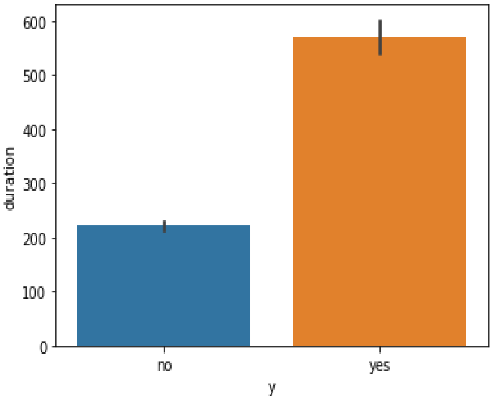
It will be interesting to analyze who said yes and no in the previous campaign. Understanding the factors behind success and failure of previous campaign, we have created plots of each input with outcome.

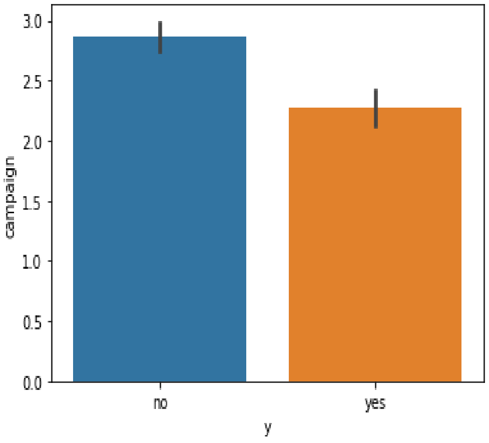


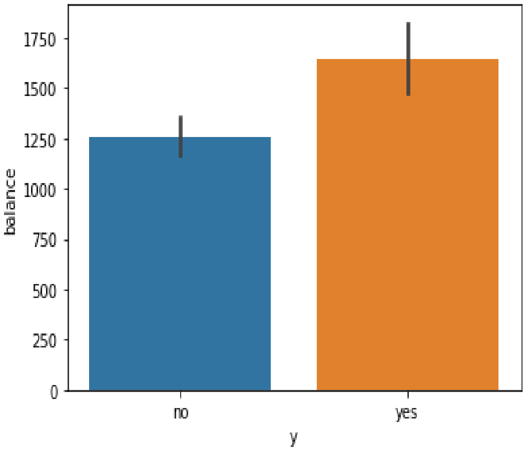
Some input variables are more important with respect to outcome than others. For instance, if we look at the relationship of callers age with the outcome of the campaign, it is not that much different. Those who have declined and those who have accepted the deposit request seem to be of the same average age. However, this can be further investigated though other techniques. It also depends on the magnitude of the impact. Nevertheless, plotting output against profession tells more significant story. People from specific profession had more representation in the last data. For example, management professionals were more in numbers who agreed to lend money to N/Lab's forerunner for the interest rate. Thus, N/Lab should focus more on management individuals while campaigning for selling its financial product. Conversely, most decliners of the product were from blue-collar professions. This indicator tells the important characteristics of people who could be neglected to avoid hefty advertisement and other campaign costs. Similarly, the below bar plot shows that the calls which lasted lesser than 200 seconds were turned outcome to be more in negatives than positives. Thus, longer the call, higher the chances that the customer will agree to invest in the financial product. Another important factor among them is the medium of the call. Those who were contacted through cellular medium are more likely to subscribe to the financial product. Count plot also shows that the people with secondary education were more interested in service than those with higher and tertiary education. Furthermore, close analysis of previous campaign reveals that those who were contacted more times are not likely to endorse to the financial service. Nonetheless, those having bank balance are more likely to buy N/Lab platinum while looking at the pattern form last data. Thus, they should be targeted in an effective way. Another important revelation of the data is about previous contacts before this campaign. More frequent calls can be regressive to these customers as previous data shows a negative relationship between subscription and calls. From bar plot below, it is visible that the calls which were made after a long gap(pdays) fetch more customers than those who were frequently time and again. N/Lab should take care in launching its campaign as it may offend come buyers with frequent calls.

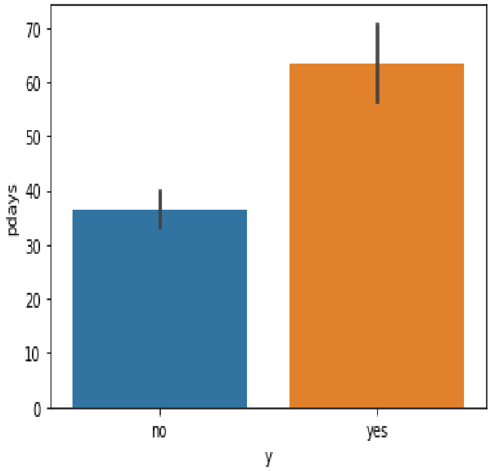




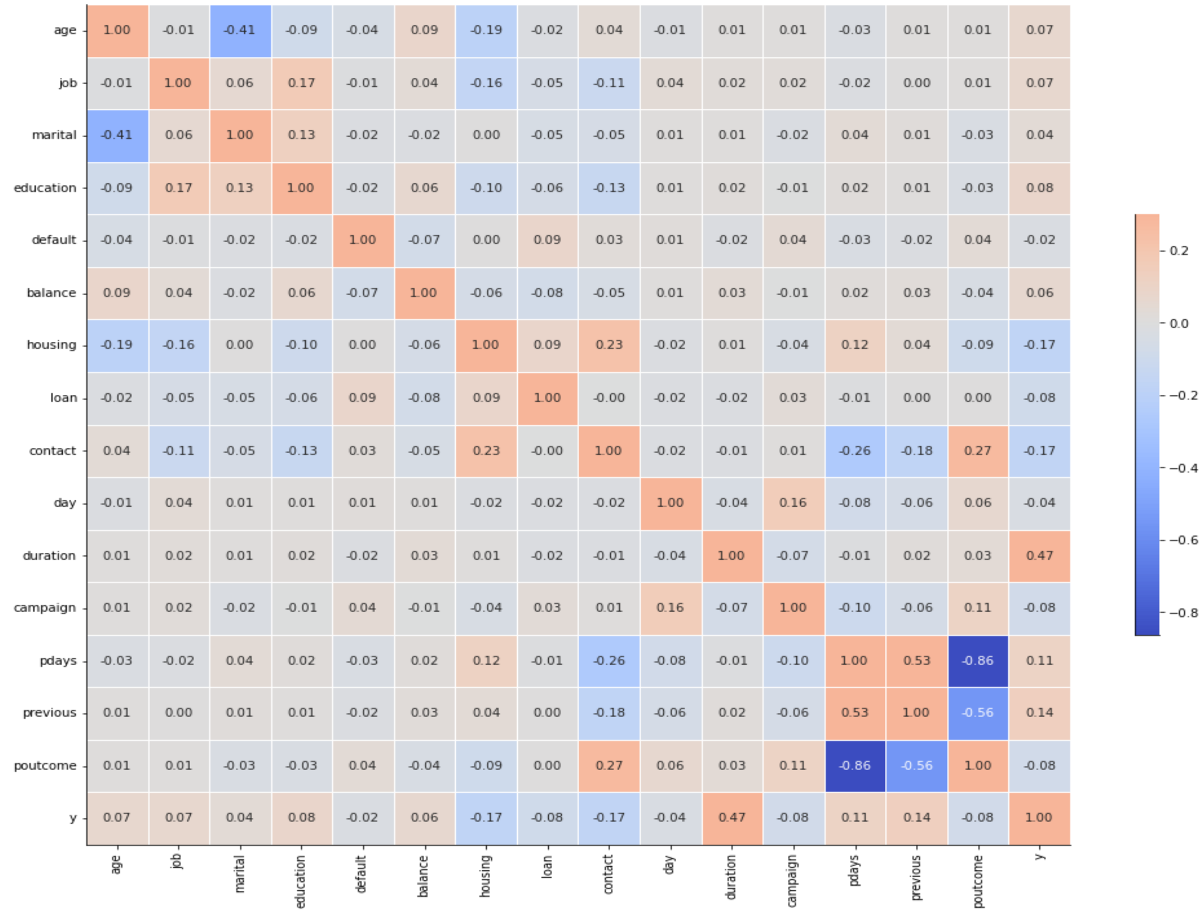








The correlation among different variable reveals a lot about patterns in the data. For instance, the duration and output(y) are highly correlated with a correlation of 0.47. Similarly, pdays and poutcome are negatively correlated- the more days passed to previous contact, more likely the negative outcome. However, Pdays and prior of calls have a positive correlation of 0.53. This shows that the more the number of days passes by after the client was last contacted in the previous campaign, more will be prior number of contacts. Thus, there will be less calls to the persons who have already been called multiple times.

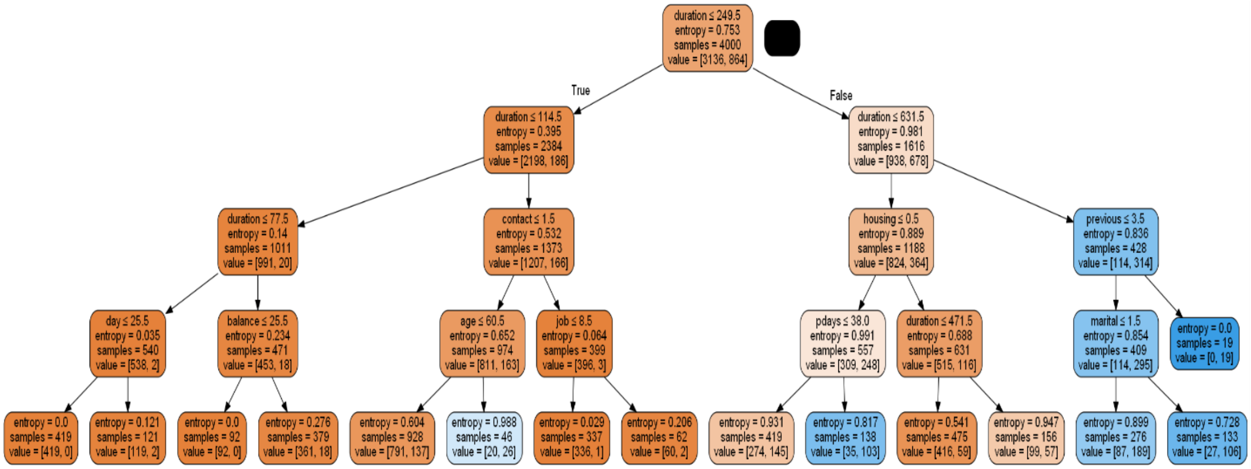


	age	job	marital	education	default	balance	housing	loan	contact	day	duration	campaign	pdays	previous	poutcome	y
age	1.00	-0.01	-0.41	-0.09	-0.04	0.09	-0.19	-0.02	0.04	-0.01	0.01	0.01	-0.03	0.01	0.01	0.07
job	-0.01	1.00	0.06	0.17	-0.01	0.04	-0.16	-0.05	-0.11	0.04	0.02	0.02	-0.02	0.00	0.01	0.07
marital	-0.41	0.06	1.00	0.13	-0.02	-0.02	0.00	-0.05	-0.05	0.01	0.01	-0.02	0.04	0.01	-0.03	0.04
education	-0.09	0.17	0.13	1.00	-0.02	0.06	-0.10	-0.06	-0.13	0.01	0.02	-0.01	0.02	0.01	-0.03	0.08
default	-0.04	-0.01	-0.02	-0.02	1.00	-0.07	0.00	0.09	0.03	0.01	-0.02	0.04	-0.03	-0.02	0.04	-0.02
balance	0.09	0.04	-0.02	0.06	-0.07	1.00	-0.06	-0.08	-0.05	0.01	0.03	-0.01	0.02	0.03	-0.04	0.06
housing	-0.19	-0.16	0.00	-0.10	0.00	-0.06	1.00	0.09	0.23	-0.02	0.01	-0.04	0.12	0.04	-0.09	-0.17
loan	-0.02	-0.05	-0.05	-0.06	0.09	-0.08	0.09	1.00	-0.00	-0.02	-0.02	0.03	-0.01	0.00	0.00	-0.08
contact	0.04	-0.11	-0.05	-0.13	0.03	-0.05	0.23	-0.00	1.00	-0.02	-0.01	0.01	-0.26	-0.18	0.27	-0.17
day	-0.01	0.04	0.01	0.01	0.01	0.01	-0.02	-0.02	-0.02	1.00	-0.04	0.16	-0.08	-0.06	0.06	-0.04
duration	0.01	0.02	0.01	0.02	-0.02	0.03	0.01	-0.02	-0.01	-0.04	1.00	-0.07	-0.01	0.02	0.03	0.47
campaign	0.01	0.02	-0.02	-0.01	0.04	-0.01	-0.04	0.03	0.01	0.16	-0.07	1.00	-0.10	-0.06	0.11	-0.08
pdays	-0.03	-0.02	0.04	0.02	-0.03	0.02	0.12	-0.01	-0.26	-0.08	-0.01	-0.10	1.00	0.53	-0.86	0.11
previous	0.01	0.00	0.01	0.01	-0.02	0.03	0.04	0.00	-0.18	-0.06	0.02	-0.06	0.53	1.00	-0.56	0.14
poutcome	0.01	0.01	-0.03	-0.03	0.04	-0.04	-0.09	0.00	0.27	0.06	0.03	0.11	-0.86	-0.56	1.00	-0.08
y	0.07	0.07	0.04	0.08	-0.02	0.06	-0.17	-0.08	-0.17	-0.04	0.47	-0.08	0.11	0.14	-0.08	1.00

Section B: Exploration

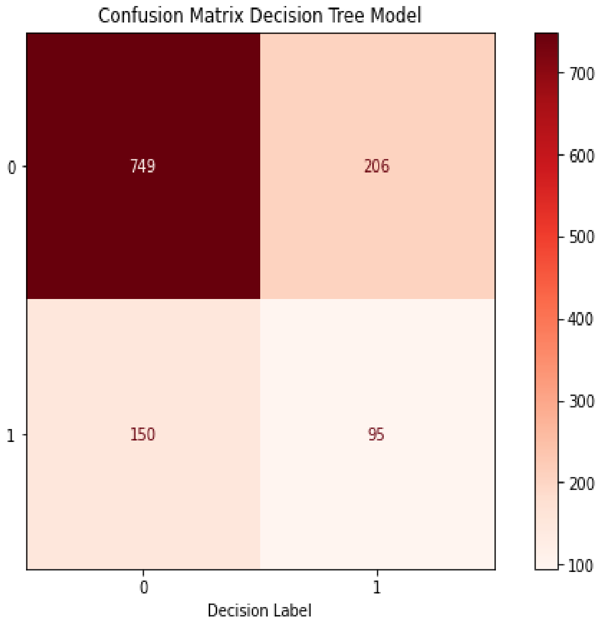
After analyzing basic patterns in data, we move towards supervised learning where we applied machine learning algorithms. Decision tree is one of the widely used algorithm in machine learning for classification problems. It works on a pre-defined target. The algorithm identifies relatively important variables based on their similarity in a uniform set of population. Decision tree uses different concepts such as Gini Index, entropy and information gain to filter decision variables based on their importance in the data set.

When we applied decision tree to our data set for potential buyers of N/Lab platinum, its root node turned out to be duration. It is noteworthy that in section(A) above, we saw a high correlation between the output the call duration. Following the same relevance, decision tree also starts with duration and follows an order of importance for other decision variables. Thus, as seen in the plot above, the more the duration of call with the customer, higher the chances that he/she will invest in the financial product. Therefore, N/Lab can analyze the call record of this data and check what fascinated to the customers. It can hire people from call centers and follow the successful conversations. It can be seen from the decision tree diagram below that balance has high gini scores in the succeeding nodes after the root point. The decision tree also gives high gini scores to the balance nodes. Not surprisingly, it is shown in the bar above in part one that higher the bank balance of the caller, the more likely he/she is to invest in the financial deposit. Job profession also seems to be important. Since the original tree was too large to interpret, we used pruning with the help of entropy (with max depth 4) to ease our understanding of the result. As pointed out above, the contact medium also played an important role in segregating the potential customers from those who are highly unlikely to buy the product. Certain professions such as management and technicians are more likely to invest than blue-color workers and students. Thus, the N/Lab can cut it input cost by avoiding people based on their professions.



There are combinations of data in the decision tree. Duration, housing, contact medium, balance, previous number of contacts before the campaign, marital status, and previous outcome seem to be synchronized and more important. However, variables such as campaign, default, marital status, and the day of the month the individual was last contacted seem to be redundant.

Evaluation of this decision tree model was also done in python to check its validity. After testing, its accuracy came out to be 70.33 per cent. Accuracy is calculated by dividing correct number of predictions by total number of predictions from the test data. Confusion matric of decision tree also reveals important facts about the precision of the model. From our tested data, decision tree identified true rejecters-those who said no to depositing money in the last campaign. However, there were 206 false negatives too-the model incorrectly considered 206 non-buying customers as buying customers. Similarly, even though the model rightly identifies 95 persons form our data set to be those who have bought the investment plan, it also incorrectly selects 150 who did not buy the product-i.e., false negative. Recall, another indicator of assessing the model, tells the percentage of true classifiers. For decision tree, it turned out that 38 percent of the deniers were truly identified by the model.

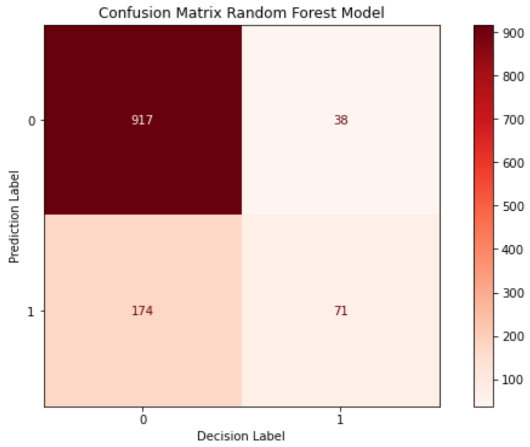


Section D: Model Evaluation

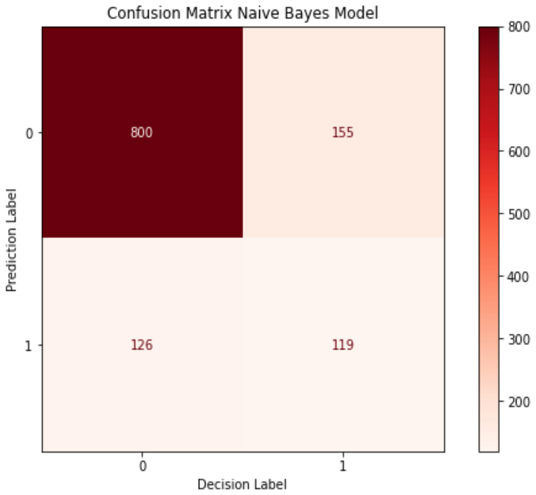
We have used three basic machine learning models – Random Forest, Naive Bayes and K-nearest – to check our training data. For better results, continuous data is scaled and categorizes data is encoded using sklearn's preprocessing tool. Following that, we split the data into training data and testing data. The ratio is set 30 to 70 for training and testing data, respectively.

Random Forest is one of the most effective models being used for predictive analysis. This model is based on a technique called begging. This method reduces variance by mixing the result of multiple classifiers on different subsamples of the same data collection. It works in three steps: creating multiple data sets, building multiple classifiers, and combining classifiers. Simply, it makes multiple tress and do not take all the data for the modelling. More importantly, it also takes care of overfitting.

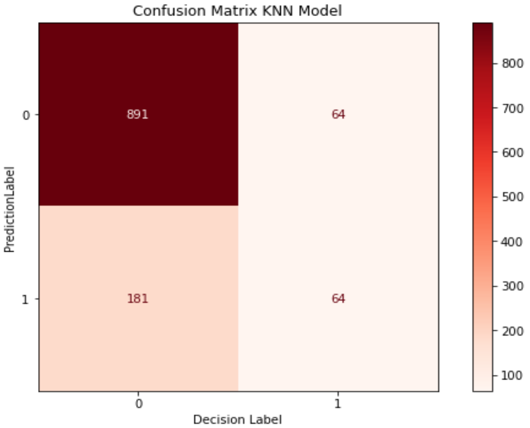
Random forest is widely preferred over decision tree and other models. The reason is that it takes care of the overfitting issue and give us more accurate results. After applying random forest to our data, it came out to be the best model. Accuracy score for random forest is 82.3 percent as shown below. Confusion matric also shows the same trend as it truly identified true positives and true negatives(diagonal values). While testing data, it finds out 917 true decliners who did not accept to buy the deal. Compared to decision tree model, random forest has just identified 38 non-buying customers as buying customers (false positives in the 2nd quadrant)



Third algorithm used on our data is Naïve Bayes. Given a set of conditions, this model uses classification technique to determine the probability of an outcome. This gaussian naïve bayes is based on continuous distribution described by mean and variance. Accuracy on our data turned out to be 76 percent for Naïve Bayes. Confusion matric shows it gave 800 true false values-correctly predicted who has declined to invest with N/Lab's predecessors. However, It is less accurate than the random forest as it predicts 155 false positives



Lastly, we apply K-nearest algorithm on our data set. It used k as the number of nearest neighbors for assigning a label to a data point. This model was second best to random forest as its accuracy was 79 percent. Confusion matric of this model depicts that it correctly identified 891 true positives. It incorrectly identified false negatives to be 181 while random forest predicted 174 false negatives.



Section D: Final Assessment

Based on accuracy scores, random forest is the winner. It is superior to decision tree because it is good at predicting on a new data set, where decision tree does not do well. However, Naïve Bayes identifies the success(yes) rate more accurately, i.e., false negative in the confusion matric interpretation terminology. It correctly predicts 119 customers from the test data who had agreed to deposit with the N/lab's predecessor.

Section E: Model Implementation

Optimizing our best selected model, Random Forest in this case, will require the library called GridSearchCv. We import it form sklearn while using this command sklearn.model_selection. Afterwards, we give parameters to the grid-man_depth,min_sample_split, min_sample_leaf,n-estimator.. Max_depth parameter shows the depth of each decision tree in the random forest. The min_sample_leaf parameter tells the minimum number of samples require to split at internal leaf node. The min_sample_leaf specifies the minimum number of samples required to be at the leaf node. The n_estimator_parameter specifies the number of trees in the forest of the model. Setting parameter are followed by the base random forest classifier. Then we initiate a grid search model and fit the grid search to the data. Finally, we test our best grid and see if the optimized model had more accuracy. After checking our optimized random forest model, our accuracy improved marginally to 82.5 percent form 82.3 percent of the base version.

Section F: Business Case Recommendation

Machine learning models used for analyzing the available data provide insightful information to the N/Lab. It sifts important variables from redundant ones. As N/lab will reach out to potential customers, it can focus more than the customers with long call duration. Furthermore, it can reduce its research cost, both by phone call bills and call agents 'salary, by eliminating certain professionals who are less likely to deposit money for the interest gain. More specifically, it can neglect students, entrepreneurs, and self-employed people as they less likely to buy N/Lab platinum. Form the decision tree, it can identify the pattern in age, education, pdays, loan, and employment groups and reach out to those segments who have higher chances of acceptability for buying N/lab platinum deposit.

References:

Meinert, R. (2019). *Optimizing Hyperparameters in Random Forest Classification*. [online] Medium. Available at: https://towardsdatascience.com/optimizing-hyperparameters-in-random-forest-classification-ec7741f9d3f6#:~:text=2.%20max_depth%3A%20The%20max_depth%20parameter%20specifies%20the%20maximum [Accessed 13 Jan. 2022].

DeZyre. (n.d.). *How to find optimal parameters using GridSearchCV in ML in python* -. [online] Available at: https://www.projectpro.io/recipes/find-optimal-parameters-using-gridsearchcv#:~:text=To%20get%20the%20best%20set%20of%20hyperparameters%20we [Accessed 13 Jan. 2022].

Springboard Blog. (2021). *Decision Tree Implementation in Python with Example*. [online] Available at: https://www.springboard.com/blog/data-science/decision-tree-implementation-in-python/.

G, E. (2018). *k-Neighbors Classifier with GridSearchCV Basics*. [online] Medium. Available at: https://medium.com/@erikgreenj/k-neighbors-classifier-with-gridsearchcv-basics-3c445ddeb657#:~:text=kNN%20in%20a%20GridSearchCV%20Some%20of%20the%20most [Accessed 13 Jan. 2022].