National College of Ireland

Project Submission Sheet – 2019/2020

School of Computing

Student Name:	Sarath Chandra Gollamudi		
Student ID:	x19193581		
Programme:	Data Analytics	**Year:**	JAN-2020
Module:	Domain Applications of Predictive Analytics		
Lecturer:	Vikas Sahni		
Submission Due Date:	23/08/2020		
Project Title:	Prediction of Bank Term Deposit Subscription		
Word Count:	3649		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: Sarath Chandra Gollamudi

Date: 12-08-2020

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Prediction of Bank Term Deposit Subscription

Sarath Chandra Gollamudi
School of Computing
National College of Ireland
Dublin, Ireland
x19193581@student.ncirl.ie

***Abstract*— In today's world, direct marketing is indeed an immersive tactic introduced by several financial institutions to offer term deposits through the telephone to customers. All marketing strategies for banks are based on massive digital data from consumers. The scale of this collection of data is difficult for an analyst to analyze manually and bring up a useful knowledge that can improve the effectiveness of business decisions. Machine learning algorithms are performing good in prediction and in the success of telemarketing campaigns. The bank telemarketing campaign success rate can be increased by predicting customer behavior in advance with the help of customers response data of the marketing calls. This research analysis proposed a random forest algorithm to predict the likelihood of customer subscription to term deposit which helps banks to take necessary decisions accordingly. The dataset utilized for this analysis is provided by the Portuguese bank for research analyses. The result of this analysis has shown that the proposed algorithm is performing well in predicting the likelihood of customers. In addition, a detailed analysis of the data is performed with visualizations using attributes like marital, education, age, etc., to understand the hidden story of the data which helps the business leaders to take decisions.**

Keywords—Bank Telemarketing, Term Deposit Subscription, Visualization, Random Forest, Predictive Analytics

I. INTRODUCTION

Enterprises have main aspects to promote items and/or services: Public lobbying, Targeted or aimed at the general indiscriminate population, Advertising, and targeting a range of common contacts. Nowadays, in a globally competitive environment, the mass campaign responses are usually very small according to a campaign survey. Additionally, targeted marketing relies on assumable targets which will be anxious to a specific product or service and such campaigns are more attractive because of their effectiveness [1]. In the present era, new technologies have had a profound impact on the marketing sector with specific requirements for specific goals. Thus, many industries consider target marketing campaigns rather than conventional marketing strategies, which are far less successful when confronted with demands from a rapidly evolving consumer environment due to intense competition [2]. The effective marketing strategies which include telemarketing are in many banks and insurance firms' which is a key strategy for engaging with their consumers and implementing new businesses [3]. Moreover, Large consumer information gathered was very relevant and useful for the target marketing campaign as in telemarketing for example [3]. Because of its inaccessibility, telemarketing is quantified direct marketing through a call centre that consists of an interactive strategy for requesting potential customers via telephone, email, and social media to sell the products or services directly [4]. However, there are some drawbacks as well, for example due to privacy intrusion it can trigger a negative attitude towards the banks.

The objective of telemarketing is to concentrate on the value of the prospective information, to try to understand the actions of customers and to predict the potential customers who may be more likely to be a customer with machine learning algorithms. With the development of information technology and the potential availability of databases, banks have collected enormous amounts of data that are stored in different ways, but it becomes very difficult to identify valuable information or knowledge to make a specific decision from data. During this point, a new aspect of data mining came into being which makes it possible to identify the meaningful value from data. Depending on the selected attributes of the features, different machine learning algorithms can be employed from one case to another [5], [6]. Data mining is helpful in making the huge amount of data obtained by CRM software system for a better understand. [1] In recent times several firms, particularly banks and financial firms, have identified the value and importance about their customer information. It is machine learning that minimizes the cost of buying, finds the most effective and profitable marketing, and offers effective solutions for many other organizations that need them. As is obvious data mining solution having many implementations all over industry sectors.

A substantial increase in processing power and memory has allowed vast volumes of data to be collected and analysed. As a result, there was a massive growth of methods for the discovery of knowledge. Choosing an effective data analysis method is not a simple job and the problem is still relevant. Machine Learning approaches have become the basis for insightful data analysis. It is a study of data science that aims to build automated machines with the aid of artificial intelligence that can enhance their function making a mockery of the gained experience and develop new insights [7]. On the other hand, with a view to maximizing data mining predictive accuracy level, analysis in strategic management and artificial intelligence is a level of consistency in strengthening the competitive classification models and effective algorithm parameter tuning. The classification methods have been thoroughly evaluated in substantial reference studies, using pre-processed datasets to assess the impact on predictive performance and computational effectiveness [8].

The main aim of this project is to solve the problems mentioned above by doing the predictive analytics of the

customer term deposit subscription of the bank telemarketing campaign data which helps the banks to take a right decision about the customers and strategies. On the other hand, to develop a predictive classification machine learning algorithm or model by using the past and current data of the customers and to train the algorithm to asset the hidden patterns in the data which helps to the prediction with more accuracy.

This project report is structured as follows; Section I is introductory to a selected domain, Section II includes the hypothesis of the project, Section III contains the related research work done in this domain, Section IV and V explains the project analysis and respective findings. Section VI and VII end the project and discusses potential research in this area.

II. Hypothesis

Its highly useful to establish a hypothesis for any project. The null and alternate hypothesis for this project analysis is described as follows:

The null hypothesis is that "there is no significant association between the Dependent Variable **deposit** and Independent variables such as employment, marital, etc."

The alternate hypothesis of the analysis is that "there is a significant association between the Dependent Variable **deposit** and Independent variables such as employment, marital, etc."

	Expected results		
Prediction		No	Yes
	No	TN (true negative)	FN (false negative)
	Yes	FP (false positive)	TP (true positive)
Classification error: (FP+FN)/(Number of Instances)			

Table-1: Confusion Matrix

The machine learning algorithms will be applied on this bank telemarketing campaign dataset to verify whether the defined hypothesis is true or not true.

III. Related Work

Organizations face various challenges and they need to handle all of those for the organization’s growth. The bank telemarketing, in this respect, depends on the huge data of the customer that helps to determine it and make the decision. In earlier research of the last couple of years, several researchers have based their work on predicting bank telemarketing success with data mining algorithms. Below are some of the related works.

The bank telemarketing campaign's success is, to look at the quality of the potential customer information, trying to understand the behaviour of the customers and to predict the expected customers who might have a higher likelihood of being a customer using data mining techniques. In this paper [9] C. S. T. Koumetio with his colleagues has done research on a telemarketing campaign dataset by applying different machine learning algorithms like Naïve Bayes, Decision Tree, Artificial Neural Network, Support Vector Machine, and a new classification technique. After a thorough analysis, they have compared the results of the new technique with existing algorithms, and the results have shown the proposed technique performs well than other algorithms with a good f-score.

Over time, the growing range of advertising campaigns has reduced its effect on the public at large. Additionally, economic demands and competitive pressure have led business owners to spend with a strict and comprehensive choice of contacts on direct campaigns. In the paper [1] R. M. S. Laureano and his fellow colleague have conducted research on Portuguese bank telemarketing data based on CRISP-DM methodology in 3 iterations. The different machine learning algorithms like Naïve Bayes, SVM, and Decision Tree were used on the data and the results have shown that the SVM performs well over other algorithms with an AUC of 0.938.

In banks, large amounts of data were documented about their clients. Such information could be used to establish and maintain direct customer interaction and link to target them individually for different goods or banking offers. The chosen customers are usually approached directly to promote the new product/service via direct contact, mobile phone, fax, and e-mail, this type of campaign is called a direct campaign. In this paper [3] A. M. Elsayad and his colleagues performed an analysis on telemarketing dataset to increase the effectiveness of the campaign with deep learning and machine learning methods which are Multilayer Perception Neural Network and Ross Quinlan decision tree (C5.0). The results showed that C5.0 predicts with the best accuracy, specificity, and sensitivity over MLPNN.

Decision tree learning is one of the most successful approaches to building the classification models which has become a baseline for several other classification techniques. Decision trees are created through recursive partitioning that intends to split the variable space until the dependent variable reaches the threshold degree of quality within each input space. In the paper [7] B. B. D. Grzonka and G. Suchacka compared the results of different algorithms like bagging, boosting, and random forest with the decision tree algorithm of research conducted on banks marketing campaign data. The decision tree performs well with a classification error of 0.159 over other techniques.

To help strategic decision-making, corporate data analysis faces the task of systematic information exploration across large data streams. Although work in industrial engineering, direct marketing, and data science centres on the design and analysis of machine learning techniques, there has been no thorough investigation into the relationship of data mining with the preceding data pre-processing stage. In this paper [8] S. F. Crone, S. Lessmann, and R. Stahlbock examined the

effect on the classifier efficiency of decision trees, neural networks, and SVM of various pre-processing techniques along with statistical analysis. The results of the analysis showed that SVM has the highest prediction accuracy.

The three main tasks to be carried out in consumer marketing are customer segmentation, market targeting, and positioning strategy. Market segmentation is clearly a much more important objective for being successful in the market. In this paper [5] R. Vaidehi has performed an analysis of the Portuguese bank telemarketing data by using the machine learning algorithms Decision Tree, Rule-based Modelling, and Generalized Linear Model. After a thorough analysis of the data the results have shown the Rule-based model performs well than the remaining techniques with an accuracy rate of 91.04%.

Telemarketing has slowly become one of the most used marketing platforms in direct marketing, for its low cost and simple interaction. So, telemarketing companies need to evaluate consumer information and must make decisions using predictive models to identify those customers who are likely to react to direct campaigning. In the paper [10] Che, J., and Zaho. S. along with their colleagues conducted research on telemarketing data using the machine learning methods t-SNE which is a feature extraction method and traditional SVM method. The final output showed that the SVM performs well over the other one with good prediction accuracy.

In the above all papers, most of the researchers have covered all the machine learning methods for the classification problem. In this analysis, Random Forest is applied on the bank telemarketing campaign dataset.

IV. METHODOLOGY

This predictive analysis was conducted based on the CRISP-DM methodology to understand the data, recognizing patterns, and to evaluate the final output for taking the business decisions. Below is the flow chart of methodology steps which are involved in the process.

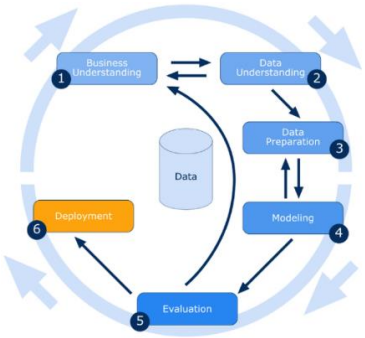


Fig-1: CRISP-DM

A. *Business Understanding:*

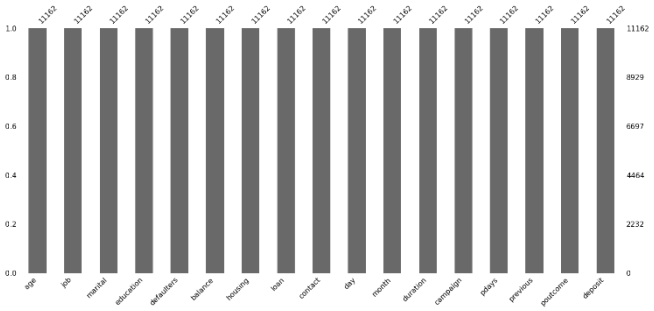
Business Understanding is the first phase of the process where we need to understand what the consumer needs to accomplish. In this phase there are different strategies one must consider are determining business objectives, deciding data mining goals, assessing the situation, and producing the project plan. With this project analysis, the business leaders can take a quick decision about the customers who are willing and not willing for term deposit subscriptions based on the target column deposit. The objective of this analysis to predict the customer likelihood of term deposit subscriptions.

B. *Data Understanding:*

This analysis has performed with the 'Bank Marketing Campaign Dataset' which is given for analysis purposes. This dataset is composed of 11162 records with 17 features, and the dependent variable is 'deposit' with which term deposit subscription likelihood to be predicted. On the other hand, the remaining all other features like the job, martial, education, and age could be used as independent variables.

C. *Data Preparation:*

To start with data exploration and applying machine learning algorithms first we need to make sure that there are no missing values in the data. If there are any missing values in the data, the prediction results will not be so accurate. These missing values will be handled by the imputation technique which means replacing the NA's with mean or median of that respective column. In this dataset there are no missing values as we can see in the below graph.



Graph-1: Missing Values

D. *Modelling:*

Random Forest method is used to predict the likelihood of customer term deposit subscription, as discussed in the above sections this technique gives a good prediction accuracy. The result and the visualization of this analysis is easy to understand and very useful for all the business leaders to take necessary decisions about customers. This is a significant reason to use a random forest algorithm for predictive model implementation.

This analysis is performed using the python language and the visualizations are generated with Plotly, Matplotlib, and Seaborn. In this dataset, the target variable is 'deposit' with which we can identify whether the customer marked as 'Yes' or 'No' after the call. There are 52.62% of people who did not open term deposits and 47.38% of people who subscribed for a term deposit. Below is the pi-chart of the same and the

subscription percentage of people with respect to their educational background.

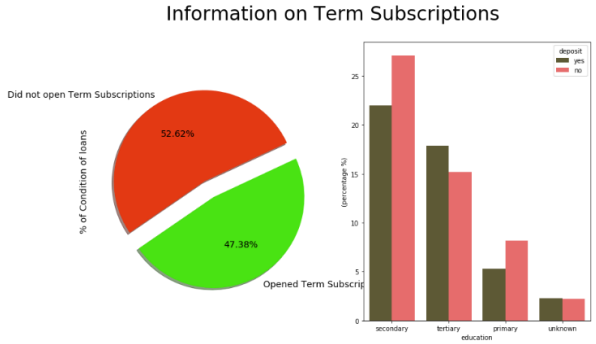


Fig-2: Information on Term Deposit Subscriptions

The random forest generates a decision tree with all the independent variables which contributes to the prediction of the dependent variable. The following tree diagram is a small tree from the entire random forest tree structure of this analysis.

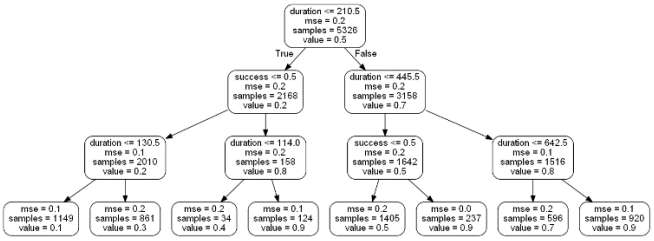


Fig-3: Random Forest Tree Structure

From the above tree structure, the business leaders can easily identify which factors are contributing more to the likelihood of customer subscription. For instance, in this analysis it shows that the more duration a customer is talking with an executive the more chances for the subscription.

E. Evaluation:

Model evaluation can be performed using different assessment metrics. The efficiency of the model created is verified by using the confusion matrix as shown in the below figure. The accuracy of this model is 79.97% with an F1 score of 0.79. On the other hand, the precision and recall are at 0.80 and 0.79, respectively.

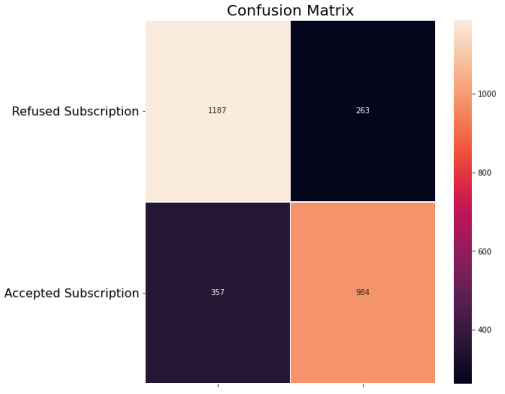


Fig-4: Random Forest Confusion Matrix

F. Deployment:

This project is about academic education, it will establish and provide a detailed record of all steps. This paper includes all references for any other research to be carried out in the future.

V. Result

By evaluating the outcome of the random forest classifier, the business leader will know whether the customer should subscribe to the term deposit. The machine learning method will be useful in finding the reason for the customers who are not subscribing, and necessary steps can be taken after the prediction. There are various features that are influencing the target variable 'deposit'. On the other hand, by verifying the relation between all the variables it will be useful in making decisions. As mentioned in the project design document the exploratory data analysis is performed on all the attributes like Age, Marital, Education, Job, Duration, and balance, etc. to know the influencing features and relationship among the independent variables (IVs). Below is the correlation matrix of all the numeric IVs. We could see that there is no high correlation between variables so we can use all variables for our analysis.

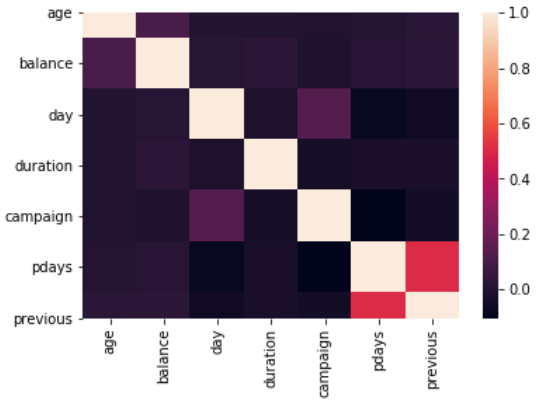


Fig-5: Correlation Matrix

A. Balance and Job:

In the below figure we could see that retired people are maintaining more balance (>80000) in the bank account followed by blue-collar jobs and self-employed people in second and third place. This means that targeting the retired people will give more chances for term deposit subscription.

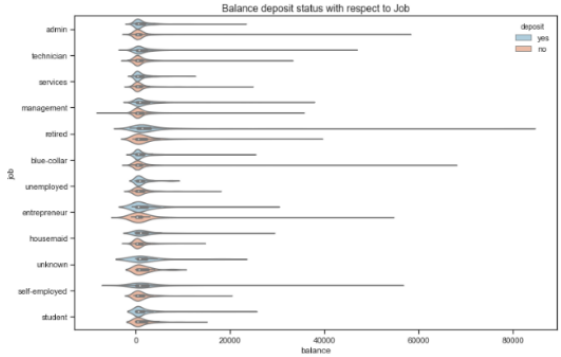


Fig-6: Balance deposit status with respect to job

B. Duration, Age and Balance:

In the below 3D interactive plot, we could see the relationship between all 3 variables (x-axis: balance, y-axis: duration, Z-axis: age). The chart shows that different age group people and the duration of the calls they are talking with the customer executives and how much balance they have in their bank account. For instance, we could see that people under the age of 20 are having a balance of 348 euros. This means that a business leader can easily identify whom to target for term deposit subscription.

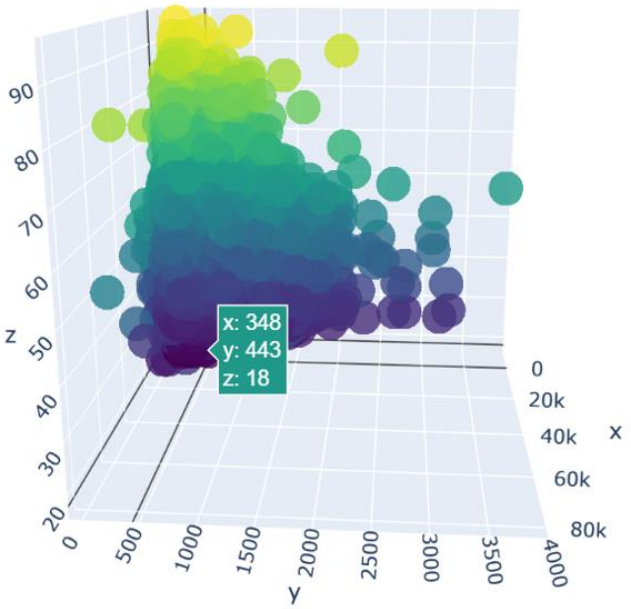
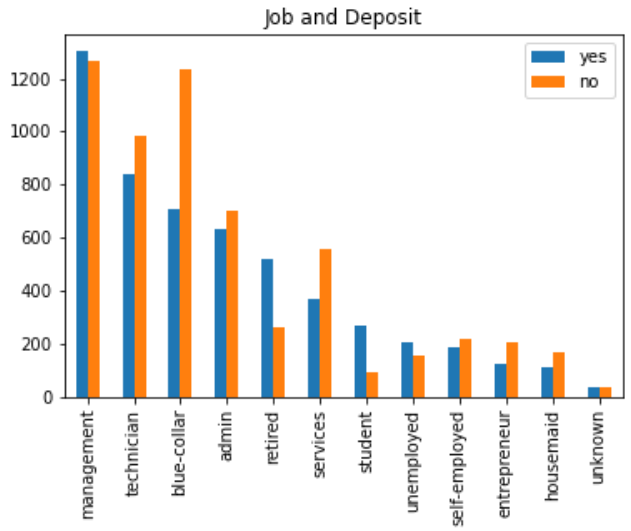


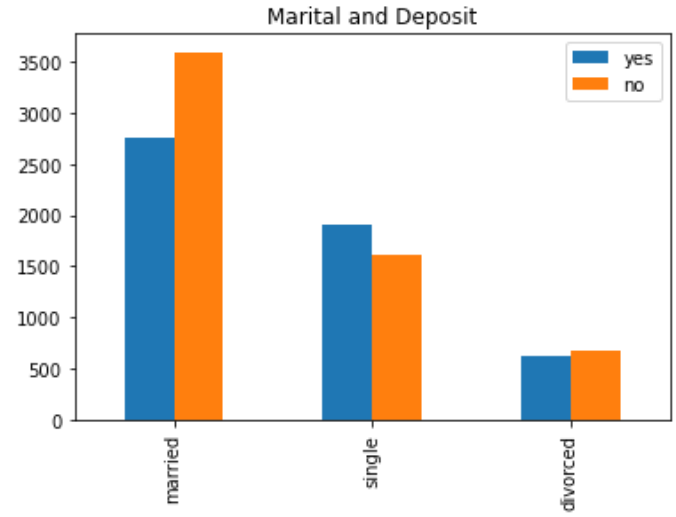
Fig-7: Relation between variables Age, Duration and Balance

C. Job, Martial, Education, and Deposit:

From the below graphs we could tell that according to our dataset: Customers with 'blue-collar' and 'services' jobs are less likely to subscribe for a term deposit. On the other hand, Married customers are less likely to subscribe for a term deposit.



Graph-2: Job and Deposit



Graph-3: Marital and Deposit

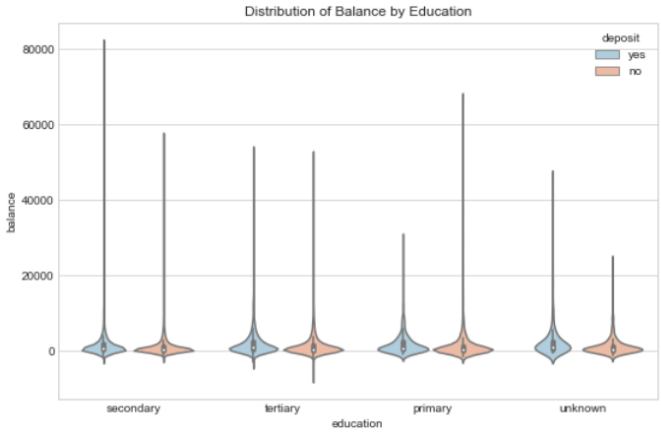


Fig-8: Education and Balance

In the above fig-8, we could see that secondary education people are showing more interest in term deposit subscription. On the other hand, the primary education people are likely to say ‘no’ for the subscription. This means that with the help of these visualizations an organization could easily understand which group of people they need to target.

All the above visualizations and dashboards are generated using python language. The business leaders or analysts can prepare different strategies for attracting the customers for term deposit subscription based on the prediction results.

D. Hypothesis Result:

This project analysis shows that the target variable and independent variables have a significant relation. The random forest uses the features to build an algorithm to find out whether the customer would subscribe term deposit. We are rejecting the null hypothesis and accepting the alternate hypothesis since there is a significant relationship between the variables.

VI. Conclusion

This research focused on the topic of predicting consumer term deposit subscription. Every financial institution needs to

recognize the possibilities of customer subscriptions to sustain the organization reputation and performance. The motivation of this project is to predict the likelihood of a customer term deposit subscription by building a machine learning algorithm with historical data of the telemarketing campaign and finding the hidden patterns in the data which could be helpful for the business leaders to make decisions. Moreover, as discussed in the related work section, there are a couple of researches done on this with different algorithms. A Random Forest algorithm is used for this analysis which for tracking the likelihood of customers. All the visualizations and algorithm analysis are performed using Python language. The time taken for training the algorithm was very less, this means that the random forest is reliable for this analysis. The evaluation metrics of the results have shown an acceptable performance as discussed in section IV.

The work in this analysis has identified many hidden patterns of the data which could be very useful for the decision making process about the customer's likelihood. Moreover, different offers and schemes can be given to attract customers based on prediction and EDA results.

VII. Future Work

This research analysis is performed on a medium records dataset, a larger dataset could be used to classify the important features for further studies. Various machine learning can also be implemented to get more insight into the features that affect the target variable.

References

[1] S. Moro and R. M. S. Laureano, "Using Data Mining for Bank Direct Marketing: An application of the CRISP-DM methodology", Eur. Simul. Model. Conf., no. 1, pp. 117-121, 2011.

[2] C. T. Su, Y. H. Chen and D. Y. Sha, "Linking innovative product development with customer knowledge: a data-mining approach", Technovation, vol. 26, no. 7, pp. 784-795, 2006.

[3] H. A. Elsalamony and A. M. Elsayad, "Bank Direct Marketing Based on Neural Network", Int. J. Eng. Adv. Technol, vol. 2, no. 6, pp. 392-400, 2013.

[4] C. Vajiramedhin and A. Suebsing, "Feature selection with data balancing for prediction of bank telemarketing", Appl. Math. Sci, vol. 8, no. 114, pp. 5667-5672, 2014.

[5] R. Vaidehi, "Predictive Modeling to Improve Success Rate of Bank Direct Marketing Campaign", IJMBS, vol. 9519, pp. 2230-2232, 2016.

[6] E. W. T. Ngai, L. Xiu and D. C. K. Chau, "Application of data mining techniques in customer relationship management: A literature review and classification", Expert Syst. Appl, vol. 36, no. 2 PART 2, pp. 2592-2602, 2009.

[7] B. B. D. GRZONKA and G. SUCHACKA, "Application of selected supervised classification methods to bank marketing campaign d", Inf. Syst. Manag, vol. 5, pp. 36-48, 2016.

[8] S. F. Crone, S. Lessmann and R. Stahlbock, "The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing", Eur. J. Oper. Res, vol. 173, no. 3, pp. 781-800, 2006.

[9] C. S. T. Koumétio, W. Cherif and S. Hassan, "Optimizing the prediction of telemarketing target calls by a classification technique," 2018 6th International Conference on Wireless Networks and Mobile Communications (WINCOM), Marrakesh, Morocco, 2018, pp. 1-6, doi: 10.1109/WINCOM.2018.8629675.

[10] Che, J. , Zhao, S. , Li, Y. and Li, K. (2020) Bank Telemarketing Forecasting Model Based on t-SNE-SVM. Journal of Service Science and Management, 13, 435-448. doi: 10.4236/jssm.2020.133029.