Analysis of Customer Behaviour with Machine Learning Models.

Abdul Azadh Abdul Saleem
M.Sc in Data Analytics
School of Computing
National College of Ireland
Dublin, Ireland.
x18203621@student.ncirl.ie

Abstract— **Advancements have been happening at a rapid rate across the various type of organizations. It is always important for an organization to analyze their customers, who plays a vital role in their development. Every organization must undergo the changes according to the changes in the current market and the customers' interest. It becomes an important aspect in companies to analyze their customer at a regular rate to make sure that the they are satisfied with their service, which is called as Customer Retention. Categorizing the customers based on the past records helps to company to take necessary steps on the target group of customers. First part of this analysis is about the analyzing the customer behavior from the data obtained because of a marketing campaign of a Bank. Support Vector Machine model has been adapted for this analysis. Second part covers the prediction of default credit card customer with Logistic Regression and KNN classifier. Lastly, the churn of the telecom customer is analyzed with Decision Tree and Random Forest. And their results were compared to evaluate the performance of the model on the datasets considered.**

Keywords—Analysis, Prediction, Churn, Credit Card Defaulter, Bank Marketing, Logistic Regression, KNN, SVM, Decision Tree, Random Forest, Evaluation, Accuracy.

I. INTRODUCTION

In this rapid industrializing world, it is an important aspect of a firm to analyze the customers. Customer Relationship Management plays a key role in determining the company's position in the market, as customer plays an important role in determining the company's success among its competitors in the market. The concept of data mining has gained its importance in CRM in recent times, because of the availability of huge customer related data which has to be analyzed to obtain the key insights to help to board of management in their decision making. Classification is an important data mining technique used for this purpose. The keys concepts of CRM are 1. Identification of potential customers, 2. Customer Attraction, 3. Retention of Customer, 4. Customer Development. Along with these concepts Customer Segmentation also has an equal importance, Customer management. In this project, the among these main factors in CRM, Selection of customers has been conducted with Bank Marketing dataset [21], which is a record of past data of telephonic marketing campaign, used to analyze whether the Customer will accept the long-term deposit plan of the Bank. Secondly, another factor of CRM which is Customer Development is done with Credit Card defaulter dataset [22], where the future bill payment of customers was analyzed with the past data. Lastly, Customer retention has been analyzed with the customer churn dataset [23], where the churn of customer is analyzed. This analysis has been conducted to find solution for the following questions.

1. How to predict Customers' behavior based the past records of a company's marketing campaign?
2. What is the accuracy of predicting the future bill payment of customer using their past records?
3. How to predict the customer churn using the various aspects available in the telecom industry?

KDD methodology has been adapted for this analysis, the process flow from selection of data, cleaning and wrangling, transformation, and data mining in order to extract the insights which are required for this study.

II. LITERATURE REVIEW

This article [1] discusses about the application of Data mining techniques in Customer Relationship Management of an organization. This research has been done over nine hundred journals about application of data mining in customer analysis and resulted that the between the years 2000 and 2006 there are eighty-seven articles published about the application of data mining models in CRM domain. This research resulted that Retention of customer has the high importance among the four CRM dimensions. And few articles talk about customer analysis using Decision tree and Neural Networks helps to obtain the segment of customers who are profitable for the organization. This article also infers that Classification models are mostly applied in Customer management to predict their future behaviors

The authors in this topic [2] discusses about the Segmentation of Customers which is an important term in Customer Management. In this study they have merged the advantages of BP neural network and Principal Component Analysis to conduct the customer segmentation for a retail business. PCA has been used in this analysis since there are many for customer behavior with very less independence, using this technique here resulted in few indicators which does not affect the quality of the raw observations. Then BP neural network model is conducted with the observations obtained from the indicators resulted from the PCA analysis, this helped to increase the efficiency of the network and better ability to forecast the Customer behavior. There are some limitations observed in this study that are the indicators of Customer segmentation from raw data is not proper enough to perform the PCA analysis.

This article [3] explains about the importance of Data mining and Knowledge Discovery in Databases in obtaining the required essential information from the large data which is growing in size. This paper provides an overview of increasing multi-disciplinary research area, basic overview of some techniques and their applications. Data mining problems like parallel computing and high performances were also discussed.

1. Prediction of Customer behavior in Bank Marketing

This study explains about the [4] importance of using the Data mining tools in the Banking sector in order to compete with its peers. The challenging factors discussed here include capturing data, storage of data and obtaining necessary information for the huge data of the firm. Also discusses about the use of Data mining tools for separation of customers into different individual segments depending on their various characteristics. This study resulted that, usage of Data warehousing, which helps to the combine the data from many sources of the bank firm and storing it in a particular format that to could help to exploration of the data using data mining methods.

In this research [5], the authors addresses about imbalanced data which could affect the quality of the prediction of the analysis. This issue has a large impact on the data classification and predict as the model considered the class which has the majority, ignoring the other class. In order to handle this data imbalance issue, the one method from Oversampling (SMOTE) and one method from Under sampling (Tomek links) are used on the raw data and the resulting data from this process is analyzed with Support Vector Machine model. This study resulted that the accuracy obtained from running SVM model on the data obtained from combined execution of SMOTE and Tomek links is greater than, the accuracy of the model obtained from data obtained by these sampling techniques individually.

[6] is study conducted in the real data obtained about the credit card customers of the Chinese bank. It involves customer segmentation by building prediction models like neural network, prediction tree, classification and regression. This resulted that Decision tree model has the highest accuracy among the four in classifying and predicting the customer behavior.

[7] is about an approach to predict the success rate of telemarketing calls for selling a deposit plan of a retail bank. Initially 150 features were obtained for this analysis which has been reduced to 22 with the help of feature selection. Four models were considered in this analysis for predicting the success rate of telemarketing calls, out of which neural network model presented the better accuracy than Logistic regression, Support vector machine and Decision tree models. Evaluation metrics like area of receiver operating characteristic curve (AUC) and area of LIFT cumulative curve are evaluating the performance of the models used in this study.

The author [8] presented in this article about the benefit of using the Customer lifetime value(LTV) in telemarketing campaign. About twelve LTV values of the customers were tested using realistic rolling windows scheme, under forward selection method which resulted in five valid new LTV features. The result obtained by this Data-driven LTV approach using neural network, helped to classify the customers based on the previous campaign.

2. Credit card defaulter prediction.

This study [9] validate the heuristic approach to develop a model which was trained with the past data and test on the recent transactions to predict that the customer will be default in bill payment for the consecutive month. The main purpose of this approach is to calculate the risk factor from the latest transaction data and combining this with the pre-computed factor from the past data. But this approach is finally resulted slightly less accurate than the Machine learning approach which is done on the same dataset utilized in this analysis. Future of this analysis is to improve the performance of heuristic approach so that it outperforms the Machine Learning approach in prediction.

This research [10] is aimed in predicting the customers in Taiwan, who are default in payment using different data mining techniques. This study the compared the accuracy of probability of customer default among the six data mining methods considered here. This study is presented with Sorting Smoothing Method in estimate the probability of default. The six method considered for this analysis is K-nearest neighbor classification, Logistic Regression, Discriminant analysis, naïve Bayesian classification, and Artificial Neural Networks, out of which Artificial Neural network is proved to performing well than the other models in predicting the customers default payment in future, with accuracy of 96%.

The author [11] has conducted this study to predict the credit card delinquency, on a huge dataset containing the information from six large banks. Three classification models have been used for the analysis, they are regularized logistic regression, test decision tree and random forest, out of which decision tree and random forest performs better than Logistic regression. And it is also observed that there is heterogeneity in the risk factors and sensitivity across banks, which makes no single model to perform better in identifying the drivers of delinquency across the banks.

This study [12] has discussed about the implementation of data mining models in classification of credit card defaulter, with Time efficiency as the key factor. Since training a model with huge data takes a large amount of time, Adaptive boosting method has been utilized to predict the default credit card client and it is compared with Support Vector Machine, naïve Bayes, Random Forest (RF) and Extreme Machine Learning (ELM). This study resulted that Adaboost has very less run-time, produces high accuracy and provides the classification model stablity.

3. Prediction of Customer Churn.

This paper [13] is about the study of attrition of customer in European financial service company. As a part of Customer Relationship Management predictors of incidence of customer churn is investigated here. It is suggested that changes in environment, demographic characteristics and stimulating the continuous and interactive relationship with customer are the major concern in retention of customer. This study has used the method of Cox proportional hazard method for this research, because of the time-dependent nature of most of the covariant. The total population considered in this analysis was made to undergo random sampling with the help of location axis data.

This study [14] attempts to understand the three important aspects of Customer outcome, they are partial defection, evolution of customers' profitability and future purchase. Two types of random forest technique are utilized in this analysis, Binary classification using Random forests and Regression forest. This study resulted that Random forest technique provides the Goodness of fit for the analysis, then Logistic regression and ordinary linear regressions. This study suggest that historical customer behavior is important to produce more future purchase and profitability.

The authors [15] in this study explores the application of Rough Set Theory in application of prediction of customer churn. The performance of different rules generation algorithms like Genetic Algorithm, Covering Algorithm, LEM2 Algorithm and Exhaustive Algorithm are evaluated. This study resulted that RST classification along with GA performs better than the rest. Limitations like class imbalance and detection and elimination of outliers were proposed to be addressed in their future research.

[16] studies about the important aspect of telecom industry which is Customer churn. The main aim of this study is to develop a Machine Learning model to identify the factors causing the customer churn and to take necessary measure to reduce it. The model developed in this analysis has to ability to run the machine learning techniques in big data platforms. The Social Network Analysis features has been utilized in this model which has further improved the performance of the model, which is evaluated using AUC evaluation metrics. The model was tested through Spark environment. This model has undergone experimentation with four algorithms deployed into them, they are Gradient Boosted Machine Tree, Extreme Gradient Boosting, Random Forest and Decision tree, out of which XGBOOST performs better in Customer Churn prediction.

[17] discusses about the Retention and Acquisition of customers since the churn of customers ultimately results in the reduced profit of the organization. The paper review about sixty-one journals to study the advantages and disadvantages of the prediction of customer churn model in telecommunication industry.

Another study [18] discusses about the importance of using the data mining techniques in analyzing the customer behavior in Financial sectors. This study explains about the main attributes of CRM, like customer acquiring, retaining and increasing the profitable customers. This analysis has utilized Decision tree algorithm (J48), Support Vector Machines and naïve Bayes as the data mining techniques to predict the customer churn.

The author [19] has studied about the Customer churn, based on availability heuristic concept, which involves with the identification of influence of the word exposure in online. This analysis examines the churn of customer through the words identified from previous studies, techniques like Decision tree graphing, logistic regression, partial least square model and neural networks models. This study identified the prediction rate more likely as obtained from the analysis done based on the customer data. Limitations of this study is identified as it mainly focuses on the macro perspective neglecting the individual customer tendency.

This work [20] attempts to create data processing approach for churn prediction in a telecom firm that faces loss due to churn of customers. The author has proposed a hybrid classification technique which is compared with the classification and clustering. Hybrid Decision tree and Logistic regression classifier were implement but it was observed that both takes more time to process, so rectify this hybrid FURIA (Fuzzy unordered rule induction algorithm) along with Fuzzy C-Means clustering was proposed for predicting the churn of customer.

III. Methodology

There are various methodologies across the industries utilized to analyze their performance in the market, one among such methodology is Knowledge Discovery in Databases methodology (KDD) which involves selection of correct data for the analysis, Pre-processing and transforming the data in order eliminate the discrepancies like noise, missing values, etc., then Data mining where the patterns of the cleaned data were interpreted using various machine learning models and Evaluation of the outcome in order to obtain the essential knowledge from the data. The study has adapted this KDD for the analysis of three large datasets analyzed below.

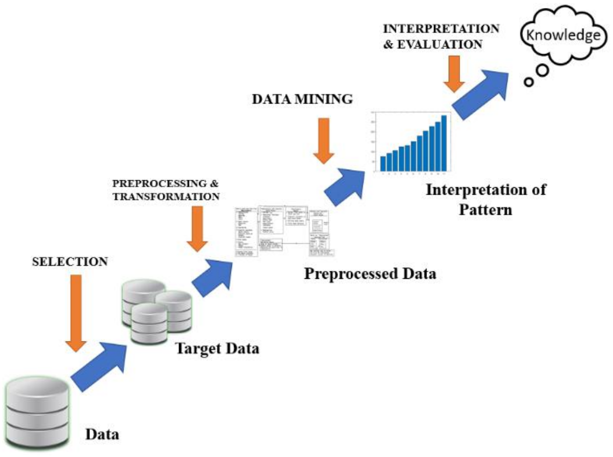
INTERPRETATION & EVALUATION
Knowledge
DATA MINING
Interpretation of Pattern
PREPROCESSING & TRANSFORMATION
Preprocessed Data
SELECTION
Target Data
Data

Fig.1 Process flow of KDD Methodology.

1. Prediction of Customer Behavior in Bank Marketing.

A. Data Selection.

The dataset [21] chosen for this analysis has the details of direct campaign of Portuguese Banking institution, which is available for public in UCI Machine Learning Repository. Dataset contains information regarding telephonic campaign of this banking institution, for their long-term deposit plan. And the attribute to be predicted is the result of this campaign whether the customer has subscribed the bank’s long-term deposit plan, which is provided as ‘yes’ or ‘no’. This dataset consists of totally more than forty thousand observations with 20 attributes which could help for this analysis.

B. Data pre-processing and transformation:

The Dataset is imported in csv format and its initial structure id analyzed and checked for the presence of null values which could affect the quality of the prediction. Then, all the variables were converted into numeric in order to make a correlation plot of the data for analyzing their contribution for the prediction as shown in Fig.2. It has been identified that the attributes ‘month’ and ‘day_of_the_week’ were insignificantly contributing for the analysis, which are then eliminated from the analysis.

The data is normalized using scale transformation in order to avoid the miscalculation due the values with greater numerical range. Since we have utilized Support Vector

Machine algorithm for analyzing this data, feature scaling helps to improve its performance. Further the data is split into train and test data with the proportion of 75:25 respectively, for the evaluation of the model in prediction of the Customer behavior.

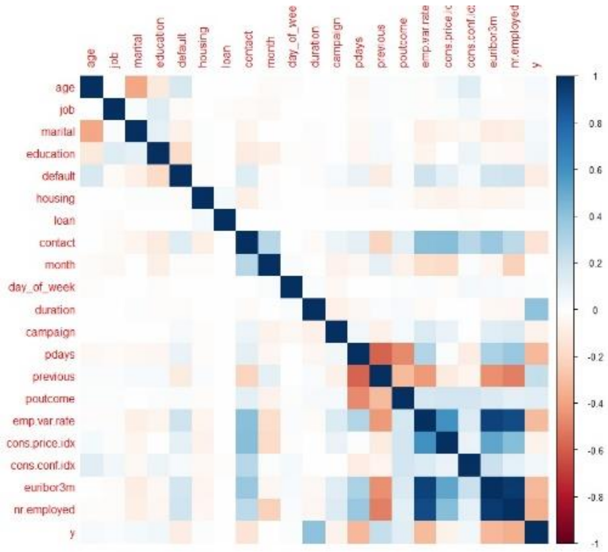

Fig.2 Correlation matrix of Banking Marketing dataset

C. Data mining method

i. Support Vector Machine

The Data mining techniques utilized for analyzing this data is Support Vector Machine method. Since SVM plots the variables on a plane spreading across and classify them based on the linear line drawn, then dataset having values of different ranges which were transformed using Scale transformation. The training set of data separated, is used for training the SVM prediction model. The Kernel here considered for this SVM model is linear.

2. Credit card defaulter prediction

A. Dataset Selection

The dataset [22] considered for this analysis has the information about the past payment pattern of the credit card customer, which helps the credit card service providers to predict the customer's future payment. This dataset collective has the details of the six months payment record of the customer including their financial credit information. This dataset has been extracted from UCI public repository for this project. Dataset is huge, with thirty thousand instances and 24 attributes to predict the future payment of the customer.

B. Data Pre-processing and Transformation

The dataset is initially analyzed for the presence of Null values. The ID column is removed then, since it does not contribute for the prediction of customer behavior. The name of the dependent column has been changed from 'default_payment_next_month' to 'Result' for naming convenience. The data is analyzed for Correlation of the predictor with the result value using Correlation matrix, shown in Fig.3. And it has been found that about eight attributes including 'AGE','EDUCATION', 'BILL_AMT1', 'BILL_AMT2', 'BILL_AMT3', 'BILL_AMT4', 'BILL_AMT5', 'BILL_AMT6', do not contribute for the prediction, which has been removed from analysis.

KNN is the other method chosen to analyze this data. The continuous variables present ion the data are normalized, since the KNN prediction relies on the Euclidian distance calculation for prediction. The normalized data is further over sampled in order to remove the imbalance in the data which could affect the accuracy. The data has been separated in the proportion that three out of four parts of the data (75%) has been used to train the prediction model and the remaining part (25%) is used to test the prediction model.

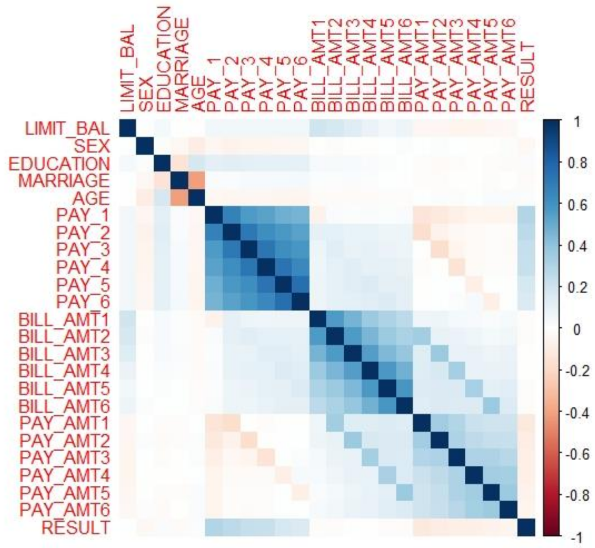

Fig.3 Correlation matrix of Credit card defaulter data

C. Data mining method

i. Logistic Regression

Since the data has the dichotomous values to be predicted, Binomial Logistics Regression has been used for this analysis. This model is used to analyze the occurrence of every class in the dependent variable with the categorical and continuous variables in the data. The data taken for training the prediction model is trained with generalized linear model function. And the predicted model is evaluated with the test data, in order to obtain the prediction accuracy of the model.

ii. K Nearest Neighbor

KNN method works on plotting all the observations of the independent attributes from the past data and the new test data is plotted against them. The search of nearest neighbors happens with the K values, and the method used here to find the neighbors which is nearest is Euclidean distance. The training data is processed with this method with the help of CARET function. Tune length is here assigned to 10 for the prediction of best value of k, with better kappa value & accuracy.

3. Prediction of Customer Churn

A. Data Selection

The data [23] used for this part of analysis is obtained from Kaggle, a public repository. This dataset contains many attributes that defines different factors in telecom industry, where customer churn is the target variable. Prediction of whether the customer will churn or not is the main purpose of this analysis. This dataset has about more than a lakh of instances with 100 attributes.

B. *Data Pre-processing and Transformation*

The data has been imported initially and then explored for the presence of NA values, and it resulted positive there was NULL values in both categorical and continuous variables in the data. The categorical variables with missing values were removed and few other continuous variable columns like ‘recv_vce_Mean’ and ‘crclscod’ were removed since they do not contribute for this analysis. After this there are eighteen other continuous variables with very fewer missing values, which were replaced by the mean values of the attribute. Since, we have about 99 independent variables for predicting the dependent variable, Principle Component Analysis has been conducted. The Predictor variables are separated into continuous and categorical variables, from which the PCA analysis has been conducted on the Continuous variables. And later, the categorical variables were combined with the variable resulting from the PCA analysis. PCA analysis has been conducted with procomp function, then the output is visualized to find the optimum value. It has been there are five PCA factors that contribute for this analysis, based on the analysis on Variance and Bias. The data resulted from the PCA analysis is then over sampled in order to reduce the data imbalance which could affect the quality of prediction. After sampling, then dependent variable was found to be having weightage for both the individual groups to be predicted. Further feature selection has been done with the help of Variance Importance Plot provided in the Fig.4, in order to improve the performance of the Random Forest analysis. The variables selected with the help of Variance importance plot is shown in Fig.5. After Feature selection the data has been splitted in the proportion of 70 and 30, for training the model and data for testing the predicted model, respectively.

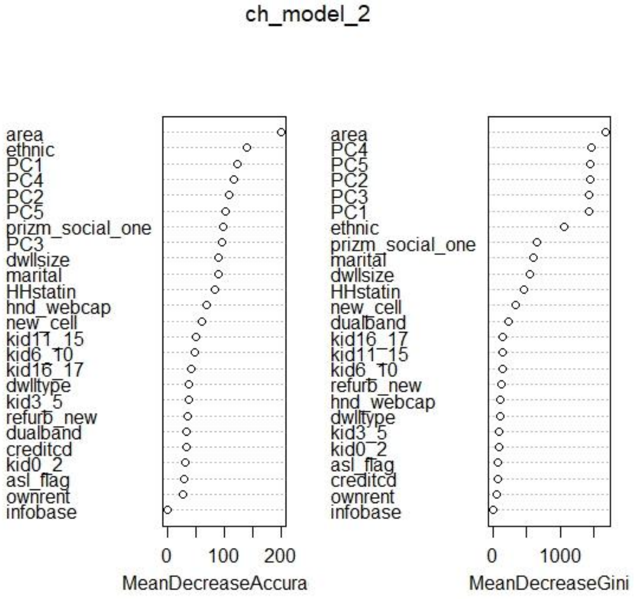

Fig.4 Variance Importance Plot before Feature Selection

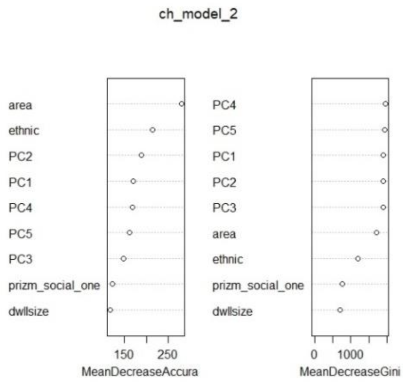

Fig.5 Variance Importance Plot after Feature Selection

C. *Data Mining Method*

i. Decision tree

The Rpart package is used for the conducting the Decision tree model on the training data, where the pattern of the data has been identified. Then the training model is utilized for the prediction on the test data. The tree plot of decision tree has been done with the help of rpart.plot function, which has been given in the Fig.6 below. Post this, selection of significant feature contributing for the analysis has been performed and the attributes are sort and the model is trained with only keeping these feature as independent variables, but this doesn’t seem to provide any noticeable change in the accuracy of the model.

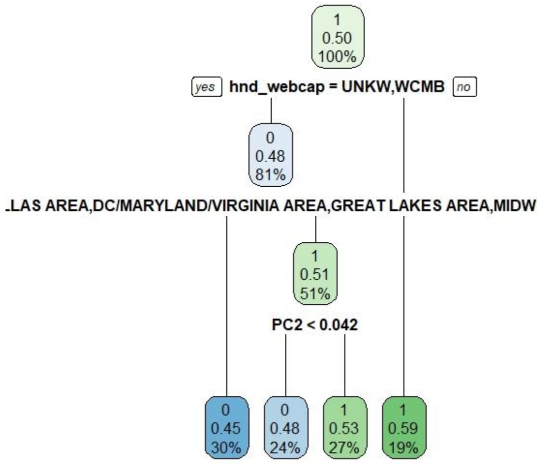

Fig.6 Decision tree plot.

ii. Random Forest

This method is a tree-based classifier that utilizes more than one decision tree to perform the analysis. The same test and train data have been utilized for this classifier. Initially the model is trained, and the prediction is analyzed with the help of Confusion matrix. Then the performance of the model is tuned with ntree as 500 and 6 as mtry to check the accuracy. And further feature selection has been done with the help of Variable Importance plot, and the attributes which are contributing significantly were filtered, this resulted in increasing the accuracy of the model.

IV. Evaluation

1. Prediction of Customer Behavior in Bank Marketing

This analysis has been carried out using the Support Vector Machine Classification. The Model is trained with the Training data which has the 75% percentage of the data. And tested with the remaining 25% data which helps to identify the accuracy of the model in predicting whether the Customer will subscribe to the long-term deposit plan of the Bank. It has been identified that prediction accuracy of the trained model is 90%. And further the Evaluation metrics were identified as Cohen's Kappa to be 0.36, Recall value 0.92, Precision value 0.98 and Area Under Curve to be 0.63. And the Receiver Operating characteristic curve is also plotted for the SVM as shown in the Fig.7.

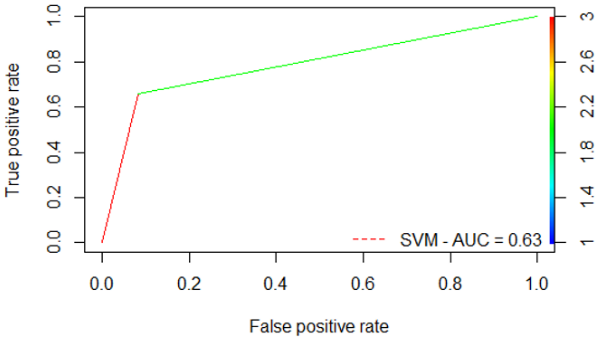

Fig.7 ROC curve of SVM classification.

2. Credit card defaulter prediction.

Credit card defaulter dataset has been analyzed with two models and their evaluation metrics were compared for both the models. Initially the data was trained with the Binomial Logistic Regression and then by the KNN classifier. After scale transformation the dataset is separated in to training and testing data and the Logistic regression model is trained with the training data. The trained model is then used to analyze the test data, to obtain the accuracy of prediction. Then a confusion matrix is obtained from which the accuracy, Recall, Precision, F1 value were identified for the model. Logistic Regression model has presented 81% accuracy in predicting the Credit card customers who will be defaulter for the next month, based on the past credit card bill payment pattern. Other Evaluation metrics that are identified for this model are Recall which is 0.97, Precision is 0.82, F1 value is 0.88 and AUC value is 0.76. Secondly, the dataset was trained with KNN model. It has been identified that the value K=5 has produced the high accuracy for this model. From the confusion matrix of this model accuracy of this model in predicting is found to be 74%, and other Evaluation metrics are Recall value as 0.62, Precision as 0.81, F1 as 0.70 and AUC value is 0.74. The comparison of evaluation metrics of these two models was shown in Fig.8. And the ROC curve has been combined and plotted for these models as shown in Fig.10.

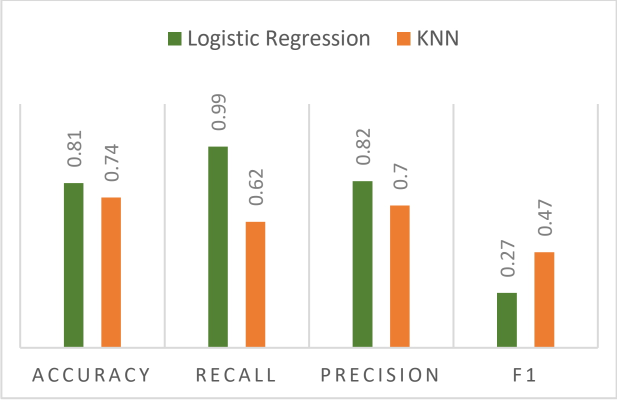

Fig.8 Comparison of Evaluation metrics for Logistics Regression & KNN

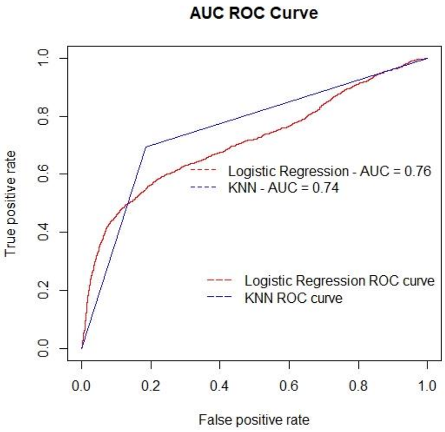

Fig.10 ROC curve – Logistics Regression & KNN

3. Prediction of Customer Churn

The dataset containing the information about the churn of customer of telecom industry is analyzed with Decision tree and Random forest models. Initially the raw data has been reduced to less predictor attributes with the help of PCA analysis and the resulting data is oversampled in order to overcome the class imbalance in the target variable. Initially the Decision tree has shown only 55% accuracy in prediction, whereas the Random forest produce 56% accuracy. Then feature selection is done for both the models by analyzing the Variance Importance of both the model. The accuracy of Decision tree for the resulting prediction is reduced to 53% where the Random forest has shown considerable increase in the accuracy to 69%. Evaluation metrics considered here are Accuracy, Kappa value, Area under curve and Sensitivity. Decision tree could maximum produce 55% accuracy in predicting the customer churn for this dataset, and Kappa value is 0.07, Sensitivity of the model is 0.50 and Area under curve is 0.54. Random forest could ably boost its accuracy to 69%, with the help of feature selection and with Cohen's Kappa value 0.39, Sensitivity as 0.64 and AUC 0.70. The ROC plot has been plotted for these models before and after feature selection performed, with the ROCR package as shown in the Fig.12 and Fig.13 respectively.

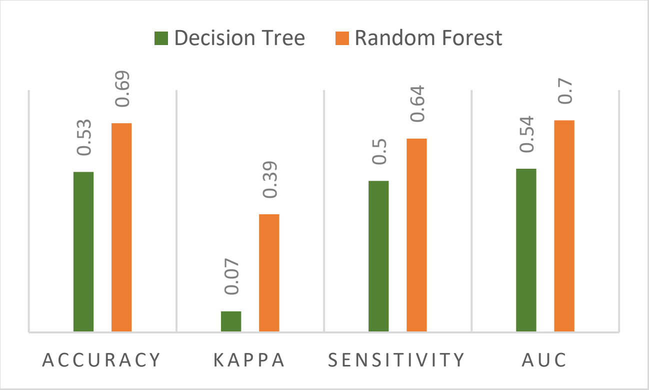

Fig.11 Comparison of Evaluation metrics of Decision Tree & Random Forest.

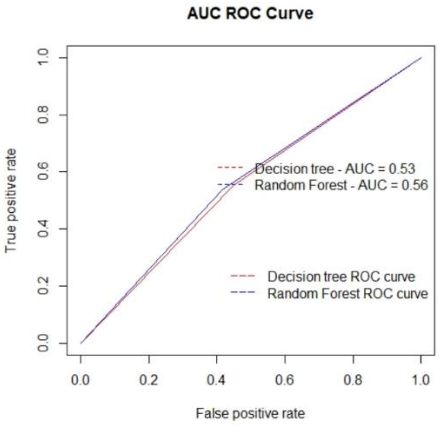

Fig.12 ROC curve of Decision tree & Random forest before Feature selection.

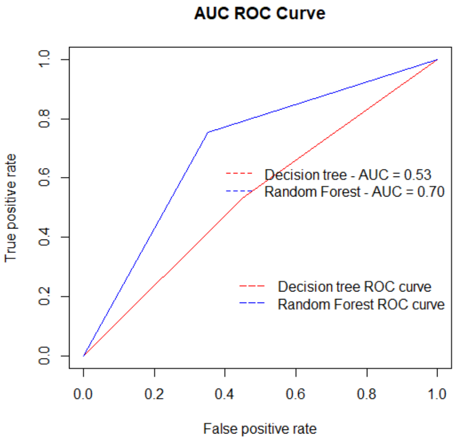

Fig.12 ROC curve of Decision tree & Random forest after Feature selection.

V. Conclusion

In this research the five models applied on the three dataset has provided their accuracy, out of which Support Vector Machine model has provided the highest accuracy of 90% on predicting the customer behavior in Banking Marketing campaign whereas the Decision tree has provided the least accuracy in predicting the Customer churn. The Logistic Regression model has performed better than KNN classification in predicting the Defaulter customer in Credit card payment. On Telecom customer churn dataset, initially the Decision Tree and Random Forest produced the output around the same accuracy, on further enhancement Random forest outruns the Decision Tree in predicting the churn of the customer with 69% accuracy. Since the datasets were taken from the publicly available repository the noise and discrepancies in the data were removed as the part of pre-processing which has helped in the analysis. Future scope of these analysis can be done with real-time data collected cleaned and analyzed with these models.

VI. Reference

[1] 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 Systems with Applications*, vol. 36, no. 2 PART 2. Elsevier Ltd, pp. 2592–2602, Mar. 01, 2009, doi: 10.1016/j.eswa.2008.02.021.

[2] H. Minghua, “Customer segmentation model based on retail consumer behavior analysis,” in *Proceedings - 2nd 2008 International Symposium on Intelligent Information Technology Application Workshop, IITA 2008 Workshop*, 2008, pp. 914–917, doi: 10.1109/IITA.Workshops.2008.225.

[3] U. Fayyad and P. Stolorz, “Data mining and KDD: Promise and challenges,” *Futur. Gener. Comput. Syst.*, vol. 13, no. 2–3, pp. 99–115, Nov. 1997, doi: 10.1016/s0167-739x(97)00015-0.

[4] K. I. Moin and Q. B. Ahmed, “Use of Data Mining in Banking,” *Int. J. Eng. Res. Appl.*, vol. 2, no. 2, pp. 738–742, 2012, Accessed: May 02, 2020. [Online]. Available: www.ijera.com.

[5] H. Sain and S. W. Purnami, “Combine Sampling Support Vector Machine for Imbalanced Data Classification,” in *Procedia Computer Science*, Jan. 2015, vol. 72, pp. 59–66, doi: 10.1016/j.procs.2015.12.105.

[6] W. Li, X. Wu, Y. Sun, and Q. Zhang, “Credit card customer segmentation and target marketing based on data mining,” in *Proceedings - 2010 International Conference on Computational Intelligence and Security, CIS 2010*, 2010, pp. 73–76, doi: 10.1109/CIS.2010.23.

[7] S. Moro, P. Cortez, and P. Rita, “A data-driven approach to predict the success of bank telemarketing,” *Decis. Support Syst.*, vol. 62, pp. 22–31, Jun. 2014, doi: 10.1016/j.dss.2014.03.001.

[8] S. Moro, P. Cortez, and P. Rita, “Using customer lifetime value and neural networks to improve the prediction of bank deposit subscription in telemarketing campaigns,” *Neural Comput. Appl.*, vol. 26, no. 1, pp. 131–139, Sep. 2014, doi: 10.1007/s00521-014-1703-0.

[9] S. R. Islam, W. Eberle, and S. K. Ghafoor, “Credit Default Mining Using Combined Machine Learning and Heuristic Approach,” *Indian J. Psychiatry*, vol. 58, no. 4, p. 372, Jul. 2018, Accessed: May 02, 2020. [Online]. Available: http://arxiv.org/abs/1807.01176.

[10] I. C. Yeh and C. hui Lien, “The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients,” *Expert Syst. Appl.*, vol. 36, no. 2 PART 1, pp. 2473–2480, Mar. 2009, doi: 10.1016/j.eswa.2007.12.020.

[11] F. Butaru, Q. Chen, B. Clark, S. Das, A. W. Lo, and A. Siddique, “Risk and risk management in the credit card industry,” *J. Bank. Financ.*, vol. 72, pp. 218–239, Nov. 2016, doi: 10.1016/j.jbankfin.2016.07.015.

[12] “ResearchGate.” https://www.researchgate.net/profile/Haifeng_Wang38/pu

blication/319689046_Real_Time_Credit_Card_Default_Classification_Using_Adaptive_Boosting-Based_Online_Learning_Algorithm/links/59b991e1458515bb9c48a3f8/Real-Time-Credit-Card-Default-Classification-Using-Adaptive-Boosting-Based-Online-Learning-Algorithm.pdf (accessed May 02, 2020).

[13] D. Van den Poel and B. Larivière, "Customer attrition analysis for financial services using proportional hazard models," in *European Journal of Operational Research*, Aug. 2004, vol. 157, no. 1, pp. 196–217, doi: 10.1016/S0377-2217(03)00069-9.

[14] B. Larivière and D. Van Den Poel, "Predicting customer retention and profitability by using random forests and regression forests techniques," *Expert Syst. Appl.*, vol. 29, no. 2, pp. 472–484, Aug. 2005, doi: 10.1016/j.eswa.2005.04.043.

[15] A. Amin *et al.*, "Customer churn prediction in the telecommunication sector using a rough set approach," *Neurocomputing*, vol. 237, pp. 242–254, May 2017, doi: 10.1016/j.neucom.2016.12.009.

[16] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," *J. Big Data*, vol. 6, no. 1, Dec. 2019, doi: 10.1186/s40537-019-0191-6.

[17] N. Hashmi, N. A. Butt, and M. Iqbal, "Customer Churn Prediction in Telecommunication A Decade Review and Classification." 2013.

[18] M. Kaur, K. Singh, and N. Sharma, "Data Mining as a tool to Predict the Churn Behaviour among Indian bank customers." 2013.

[19] E. B. Lee, J. Kim, and S. G. Lee, "Predicting customer churn in mobile industry using data mining technology," *Ind. Manag. Data Syst.*, vol. 117, no. 1, pp. 90–109, 2017, doi: 10.1108/IMDS-12-2015-0509.

[20] A. S. Choudhari and M. Potey, "Predictive to Prescriptive Analysis for Customer Churn in Telecom Industry Using Hybrid Data Mining Techniques," in *Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation, ICCUBEA 2018*, Jul. 2018, doi: 10.1109/ICCUBEA.2018.8697532.

[21] "UCI Machine Learning Repository: Bank Marketing Data Set." https://archive.ics.uci.edu/ml/datasets/Bank+Marketing (accessed May 03, 2020).

[22] "UCI Machine Learning Repository: default of credit card clients Data Set." https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients (accessed May 03, 2020).

[23] "Telecom customer | Kaggle." https://www.kaggle.com/abhinav89/telecom-customer/data (accessed May 03, 2020).