Project Report

***Abstract*— Machine Learning is the process of teaching a computer to derive insights from data and in the process improve its performance. In this paper, I present the comparison between different machine learning models on data belonging to marketing and financial sector. I have analyzed machine learning models on various parameters and studied the effect of imbalance classes on learning models.**

I. Introduction

A) Bank Direct Telemarketing

It has become relevant in recent years for companies to exploit the database of direct marketing as a strategy to increase their profit. This is important for understanding and adapting to the consumer's immediate needs. Banking domain is no such exception. Marketing department in banking industry is responsible for coming up with campaigns that attracts customer and give their bank a competitive edge against other banks. Therefore, it is vital for them to analyze customer data and launch campaigns for target customer with high success rate. A well-executed direct marketing campaign can result in high returns by indicating the number of positive responses achieved. Thus, data mining and application of machine learning models can significantly improve success rate. In this respect, I have analyzed the two data mining models; SVM and Logistic regression to predict the response of the customer. The main objective of this study is to identify the variation in accuracy with application of different technique and determine the predictive capabilities of the models.

B) Default of Credit Card

In a developed monetary system, risk prediction has always been the primary objective than risk management. Risk prediction is used in many financial domains to limit the impact of volatile factors. And since a vast majority of population is dependent on credit card, there has been dramatic rise in both the quantity of credit cards and the debt incurred by the cardholders. To avoid the delay in repayment and significantly drop the number of delinquent payments it is necessary to establish efficacious risk prediction system. Hence, comes the need to analyse the defaults. Many data mining techniques has been applied by researchers to predict such results. Since fault records constitutes of a very small in portion in data, in this analysis I tend to observe the significant change in prediction performance of models using both balanced and imbalanced data.

C) Marketing Customer Value

To thrive in this highly competitive economy, it has become essential for organizations to understand the customer needs and develop strategies revolving around it. It is a major issue that organizations are facing are churning of customers. And relatively it is five times costlier to acquire a new customer than retaining one. Thus, organizations have shifted their focus towards data mining, to identify the pattern in customer behavior which might lead to churning in future. So that strategies can be developed to prevent such situation well in advance. To analyse the customer behavior and their attributes I have opted use Random forest method and tried to analyze variation caused by the data in the model.

II. Related Work

This study compares and evaluates the performance of different machine learning model on different cases. This section reviews the theoretical aspect and problems of the machine learning model.

A. Bank Direct TeleMarketing

Charles and Chenghui applied data mining on Direct Marketing dataset to determine its importance [2]. During the process they encountered problems due to high imbalanced nature of the response variable and indicated predictive accuracy not a good measure to evaluate model performance. They applied Naïve Bayes and Decision tree algorithm C4.5 after using ada-boost as an ensemble technique. And used Lift and AUC as the criterion to evaluate and compare models. They concluded that both Naïve Bayes and C4.5 as an efficient method to apply on huge datasets.

Hany applied multiple machine learning models; multilayer perception neural network (MLPNN), tree augmented Naïve Bayes (TAN), Logistic regression, and Decision tree C5.0, on the Portuguese Bank Marketing dataset [3]. And evaluated these models on accuracy which he mentioned is a deterministic parameter on selecting the model along with sensitivity and specificity. In his paper he concluded that C5.0 performed little better than other models.

Moro, Laureano and Cortez used CRISP-DM methodology on the Portuguese Bank Marketing dataset. Their goal was to determine which model can predict positive response of a campaign [4]. They used Naïve Bayes, Decision Tree and SVM as their models. And based on evaluation parameters; Lift and AUC, determined that SVM achieved better predictive performance with AUC value of 0.938.

Sagarika followed the KDD approach towards the direct marketing dataset [5]. He applied Naïve Bayes and Decision tree as classification technique and K Means Clustering to segment customer. He evaluated the classification methods on F measure, Lift and AUC. The objective of his study was to compare the prediction performance when a balance and an unbalanced data was used. Results concluded that AUC value (0.939) enhanced after balancing the data.

Tuba and Songul used KDD methodology and applied two feature selection methods; Information Gain (IG) and Chi Square, on Naïve Bayes classifier [6]. Their goal was to improve the efficiency of the campaign. They evaluated the Naïve Bayes classifier on measures like F measure, precision

and recall. And concluded that reduction in features the performance of classification improved.

Anatoli followed CRISP-DM approach on the bank marketing dataset [7]. He used sensitivity, specificity and AUC, and indicated that they are more relevant than accuracy as the data is highly imbalanced. He performed comparative analysis of Neural Net (NN), logistic regression, Naïve Bayes, linear and quadratic discriminant analysis at multiple levels of data saturation. And concluded that NN provides the best result at all levels of saturation except poorly saturated data.

Feature selection a method to select only those predictors that significant for the learning model is studied by Chakarin and Anirut [8]. They carried out different feature scaling using different algorithms and evaluated the model on misclassification rate and prediction rate.

B. Default of Credit Card

Application of machine learning algorithm in predicting credit default is a prominent discussion among researchers.

Amar, Adlar and Andrew proposed that manipulating of certain feature of in consumer bank-account activity can significantly increase the prediction accuracy of credit default [9]. They constructed nonparametric forecast model by manipulating customer transaction and credit bureau data in a way that improved the linear regression R^2's of 85%.

According to the study Ravinder and Rinkle carried out on Australian credit card data [10]. They assessed 8 machine learning methods using leave one out cross validation approach since K- neighborhood and kernel density estimation cannot be calculated for 10-fold cross validation. And evaluated on parameters; misclassification rate, specificity and sensitivity. Support vector Machine and Genetic Programming model were found to be best methods to predict loan applicant as the misclassification rate achieved were lowest in these cases.

Yeh and Lien examined six machine learning models on Taiwan credit card dataset [11]. They split the data into two sets train and validation set. They determined that error rate is not sensitive to the accuracy of classification models and thus evaluated models using lift chart. As per the results KNN performed better than other models in training data with an area ratio of 0.68. But in validation, Artificial Neural Networks (ANN) turned out to perform better with an area ratio of 0.57. Therefore, they concluded ANN to be the best fit in predicting defaults.

To verify the effectiveness in predicting the customer ability to repay the loan, Huseyin and Bora applied five learning models on Turkish credit card data [12]. The models were then evaluated on the misclassification rate achieved for Type 1 and Type II errors. They concluded that decision tree CART algorithm performed better for Type I error, but Neural Network achieved better predictive accuracy for Type II errors.

Tsungnan and Mingmin reported low predictive accuracy due to imbalance of delinquent cases in the dataset [12]. They countered the problem by implementing grey incidence analysis and fuzzy decision tree. The proposed methods increased the average accuracy from 0.82 to 0.86 and 0.89 respectively.

In [14] Gustavo et al. mentions the problems which occur due to imbalance nature of data. They applied various sampling methods and evaluated their performance in efficiently predicting the minority class. Results revealed that oversampling (general methods) and the combination of SMOTE with TOMEK and ENN achieved results. In addition, they mentioned about analyzing the ROC curve obtained by the classifiers as an extension to their work, which is being used in this paper as evaluation parameter for classification models.

In [15] Faith proposed the use of binning methods to discretize continuous variables in our data. Her goal was to improve Naïve Bayes classifier accuracy and used Adult dataset for application of models. And concluded that by discretization of all the continuous variables classification accuracy of Naïve Bayes improves significantly from 27171 to 27704 correctly classified cases.

C. Marketing Customer Value

Xia and Jin mentioned in [16] on the predictive ability of machine learning methods on accurately determining customer churn. In their research they applied SVM using multiple kernels and compared its performance with other models. They evaluated the models on accuracy rate, hit rate, coverage rate and lift coefficient. And concluded that among all the models SVM RBF kernel achieved good prediction precision.

Xie et al in [17] applied improved balanced random forest (IBRF) on Chinese bank data. Their aim was to compare the effectiveness IBRF against standard random forest models. They used accuracy rate and lift gain chart to evaluate the models and concluded that IBRF performed significantly better than it counterpart in predicting the churn in customer.

By emphasizing on customer behavior and attributes Ibrahim et al aimed to predict churn customer [18]. They tested different learning techniques like clustering, classification and association rule. And concluded that preciseness and accuracy of method as an important of selecting a model.

III. Data Mining Methodology

For the project Data has been taken and according to the research question a targeted data is figured. This targeted data is then put in to preprocessing, after cleaning the data set a preprocessed data is generated. On the preprocessed data machine learning algorithm is applied and evaluated on different parameters.

A. Bank Direct Telemarketing

Direct telemarketing data of Portuguese Bank is used for this analysis from UCI Machine Learning repository. The data is primarily made of 21 columns which contains 10 continuous variables and 11 categorical variables. And it contains total 41,188 observations.

1) Data Preprocessing:

- Direct Marketing data was first imported into R Studio and required libraries were also install.

- Checked the dimensions of data and looked for any missing value.
- Plotted histogram plot for numerical variable to check skewness in data and plotted vertical bar plots to understand the different levels and their frequency.
- Calculated the proportion of a level in their respective column to determine their significance and refrain from removing them. Observed few columns hold no significance hence opted to remove them (duration, default and poutcome).
- Many columns contained multiple levels which can easily be combined to represent a single level, therefore combined them and encoded the categorical levels into numbers. And also converted the data type of columns accordingly.
- Performed feature selection to drop non-significant predictors using Boruta method. Output of Boruta function concluded to drop housing and loan columns from the dataset.
- Transformed the continuous variables using scale function to normalize the data.
- Split the data into train and test set in 70:30 ratio respectively.
- Performed rose sampling to balance the output variable in the train dataset to counter the issues caused by imbalanced class.

2) Data Modeling:

Two machine learning model; SVM and Logistic regression have been applied on this dataset to predict the whether a successful response would be achieved from a customer.

Support Vector Machine (SVM) is a machine learning method which can be used for both classification and regression. It differentiates between classes by creating hyperplane. SVM Kernel are basically a set of mathematical equations. In this analysis the model was trained using three SVM Kernels; Linear, Radial and Radial Basis function (RBF).

Logistic Regression is a regression method which is applied when the output variable is binary. It determines the relationship between the binary output variable and one or more input variables which can either continuous or categorical. Logistic regression basically predicts the probability of an outcome.

B. Default of Credit Cards

The main objective of this analysis is to compare the effectiveness of Naïve Bayes and Decision Tree models in predicting the delinquent payments. To facilitate this study Taiwan's customer default payment data is used from UCI Machine Learning repository. The dataset consists of 12 categorical and 11 continuous predictor columns, with default payment being the output variable. And a total of 30000 observations is present in this data.

1) Data Preprocessing

- Default of credit card data is first imported into R Studio and required libraries are imported.
- Checked the data types, dimensions of data and also checked for missing values.
- Using CrossTable function determined the proportion of each variable in the column and dropped unknown and levels with very low proportion in the column.
- Analyzed each column individually and accordingly merged levels.
- Reencoded levels and converted data types of columns.
- Discretized continuous variable to apply classification methods.
- Split the data in 70:30 ratio for train and test set respectively.
- Performed rose sampling to balance the output variable in the train dataset to counter the issues caused by imbalanced class.

2) Data Modeling:

Two machine learning model; Naïve Bayes and Decision Tree have been applied on this dataset to predict the fault in credit card payments.

Naïve Bayes is classification methods which used when a high number of categorical variable is present in the data. It is a widely used classifier which is represented by following Bayesian equation:

$$P(A|B) = (P(B|A) P(A)) / P(B)$$

Decision Tree likes the name suggest form a tree like structure of tests and outcomes. The nodes in the decision tree indicates the test on an attribute and outcome of these tests is represented by branches. Decision tree can take both categorical and numerical inputs. It is a simple algorithm which can be used to provide output of complex issues.

C. Marketing Customer Value

To analyse the performance of random forest in predicting the response of customer IBM Watson Marketing Customer Value Dataset has been taken. The dataset consists various predictors variable describing the attribiutes of a customer. The dataset consists of 9135 observations, 23 predictor variables and 1 categorical response variable.

1) Data PreProcessing:

- CSV file was imported into R studio and checked for dimensions.
- Checked for missing values and plotted the variables to get a better understanding of the nature and distribution of variables.
- Performed label encoding on categorical variables and used boruta function to select significant features to ease the processing strain on data models.

- Split the data into train and test set in 80:20 ratio and transformed the numerical variable using scale function.
- Response the variable in the data is imbalance, hence used Rose sample method to balance the classes of response variable.

2) Data Modeling:

Random Forest is an ensemble technique that creates multiple number of decision trees. It is can be used for both classification and regression. It works better than other classification methods in terms that it takes less time when computing relatively large datasets has been used to train the model. And unlike most classification it is very versatile. For this dataset I have analyzed the performance of the model in two scenarios; without balanced class and with balance class.

EVALUATION

A) Bank Direct Telemarketing

S.No.	Evaluation Measures				
	Models	***Accuracy***	***Sensitivity***	***Specificity***	***AUC***
1	Logistic Regression	0.818	0.947	0.345	0.789
2	SVM RBF Kernel	0.858	0.945	0.421	0.748
3	SVM Linear Kernel	0.828	0.945	0.358	0.739
4	SVM Radial Kernel	0.821	0.938	0.338	0.713

Table 1

The machine learning models were evaluated on mainly 4 parameters; accuracy, sensitivity, specificity and area under the curve (AUC). It can be observed from table that SVM RBF kernel has slightly better predictive accuracy. And Logistic regression model is the most efficient in predicting the true positive class (sensitivity value of 0.947). But since we are concerned about identifying the response of the customer which is the minority class. Accuracy and sensitivity cannot be taken as the only factor in determining the model. The minority class that is success in response is most precisely identified by SVM RBF kernel (specificity value of 0.421).

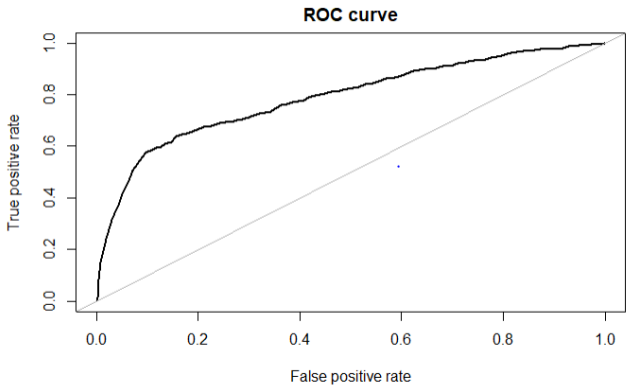

Fig1. ROC Curve Logistic Regression

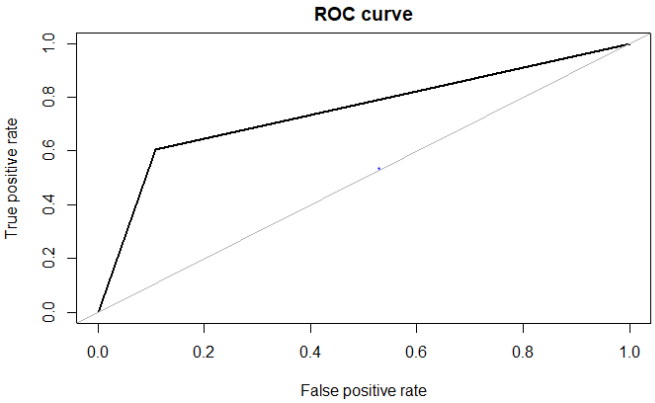

Fig 2 ROC Curve of SVM RBF Kernel

Fig 1 represents the ROC curve of Logistic regression which has been observed to have the highest AUC values of 0.789.

B) Default of Credit Card

Classification method, Naïve Bayes and Decision tree were analyzed for this dataset. And evaluated on measures mentioned in Table 2. Since the data is heavily imbalanced measures apart from accuracy were included, since accuracy can easily be affected by majority class. Since in this scenario we are more interested in correctly predicting the minority class specificity value for all the models was also calculated.

S.No.	Evaluation Measures				
	Models	***Accuracy***	***Sensitivity***	***Specificity***	***AUC***
1	Imbalanced Naive Bayes	0.684	0.778	0.221	0.50
2	Rose Sampled Naïve Bayes	0.782	0.868	0.507	0.698
3	Imbalanced Decision Tree	0.822	0.958	0.338	0.686
4	Rose sampled Decision Tree	0.776	0.850	0.522	0.686

Table 2

Naive Bayes when trained with the imbalanced data achieved low accuracy of 0.684 in comparison of when sampled data through Rose method was used, then the model performed significantly better and achieved an accuracy of 0.782. In

contrast accuracy of decision tree dropped from 0.822 to 0.776 when sampled data was used for training the model. On comparison of sampled models and imbalanced models it is evident that sampled models accurately predicted more minority class as the specificity value is higher than imbalanced models.

Area under the curve value of 0.686 is achieved through both Decision Tree models, which is explained by decrease in sensitivity and increase of specificity. But a significant increase in AUC value is observed when sampled data is used. The AUC value increased from 0.50 to 0.698.

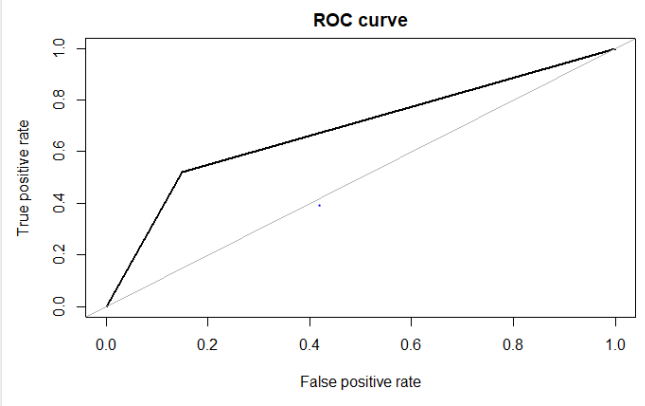

Fig 3 ROC Curve: Rose Sampled Decision Tree

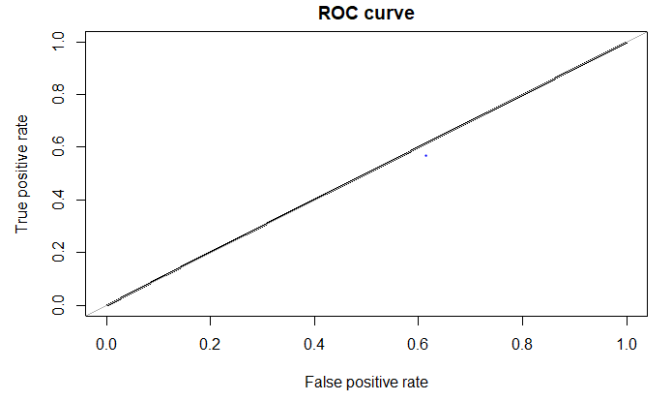

Fig 4 ROC Curve: Imbalanced Naïve Bayes

C) *Marketing Customer Value*

S.No.	Evaluation Measures				
	Models	***Accuracy***	***Sensitivity***	***Specificity***	***AUC***
1	Imbalanced Random Forest	0.923	1.00	0.653	0.956
2	Balanced Random Forest	0.923	1.00	0.6517	0.956

Table 3

It can be observed that in both cases the parameters; Accuracy, Sensitivity, specificity and AUC is almost same. Even without the application sampling a highly predictive model that can predict minority class is achieved. High accuracy value of 92% Is achieved by both models.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1426  139
         1    0  262
```

Fig 5 Confusion Matrix: Imbalanced Random Forest

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1425  140
         1    0  262
```

Fig 6 Confusion Matrix: Balanced Random Forest

From Fig 5 & Fig 6 it is evident that imbalanced nature of data does not affect Random forest.

CONCLUSION & FUTURE WORK

A) *Bank Direct Telemarketing*

Direct marketing is an essential marketing strategy which focuses on a target customer rather than the whole population. Hence, the customer data needs to be thoroughly analyzed to target campaign towards customers that would reciprocate in a positive way. In my analysis I found that the logistic regression performed better than standard support vector machine using linear and radial kernels. And achieved an accuracy of 81%. But since scope of the analysis is to determine the minority class that is, a positive response. Specificity and AUC parameters should also be weighed. Hence from the output achieved, it can be proposed that SVM model RBF kernel performed better in predicting the response due to its high specificity value of 0.421 and AUC value of 0.748. Many techniques like ada-boost and ensembling techniques can be applied in this analysis to increase the predictive capabilities of the model. Also, since I have only tested rose sampling method to balance the classes other sampling methods like SMOTE and under sampling can be tried to analyze the variation caused through them.

B) *Default of Credit Cards*

In order to decrease the amount of delinquent payment and reduce the losses incurred by lending credits for a long period of time it is vital to analyze the previous records of customers before lending credits. To facilitate this process machine learning researcher has proposed different models with different predictive capabilities. To analyze this global issue, I have opted Naïve Bayes and Decision Tree methods. From the results it was revealed that decision tree performed better when unbalanced data was used for analysis. But when a balanced data was used for training models Naïve Bayes classifier showed a significant improvement. While the AUC value of decision tree remained the same. Since in this analysis binning was done for running the Naïve Bayes it is uncertain how decision tree would have performed when the data constituted only of categorical predictors. Also, the use of complicated machine learning models like random forest and neural network can be provide better results for this issue.

C) *Marketing Customer Value*

Customer behavior is essential factor that is being widely analyzed nowadays to reduce churn rate. Since a lot competition has increased and retaining a customer has become a lot harder than it used to be, it is essential for companies to analyzed customer demographics and identify the pattern that might cause a customer to avail services from competitors. I applied random forest to identify the response of customers and concluded that irrespective of nature of data (balanced or imbalanced) random forest achieved high accuracy and an AUC value of 95.6%. This analysis can further be extended by studying the effects on complex machine learning models and performing a comparative analysis.

REFERENCES

[1] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (Eds.). Advances in Knowledge Discovery and Data Mining. MII Press, Mento Park. 1996. [Accessed 5 December 2019].

[2] Ling, X. & Li, C. Data Mining for Direct Marketing: Problems and Solutions. Proceedings of the 4th KDD conference, AAAI Press, 73±79. 1998. [Accessed 5 December 2019].

[3] Elsalamony, H. A. Bank direct marketing analysis of data mining techniques, International Journal of Computer Applications 85(7). 2014.

[4] Moro, S., Laureano, R. and Cortez, P. Using data mining for bank direct marketing: An application of the crisp-dm methodology, Proceedings of European Simulation and Modelling Conference-ESM'2011, Eurosis, pp. 117–121. 2011. [Accessed 6 December 2019].

[5] Prusty, S. Data mining applications to direct marketing: identifying hot prospects for banking product. Web data mining (ECT 584), Spring. DePaul University, Chicago. 2013. [Accessed 6 December 2019].

[6] T. Parlar and S. Acaravci, "Using Data Mining Techniques for Detecting the Important Features of the Bank Direct Marketing Data", International Journal of Economics and Financial Issues, 2017. Available: https://dergipark.org.tr/en/download/article-file/365990 [Accessed 7 December 2019].

[7] Nachev, A. (2015). Application of data mining techniques for direct marketing. Computational Models for Business and Engineering Domains. Available at: http://www.foibg.com/ibs_isc/ibs-30/ibs-30-p09.pdf. [Accessed 7 December 2019].

[8] C. Vajiramedhin and A. Suebsing, "Feature selection with data balancing for prediction of bank telemarketing", Applied Mathematical Sciences, vol. 8, pp. 5667-5672, 2014. Available: 10.12988/ams.2014.47222. [Accessed 12 December 2019]

[9] A. Ehsan Khandani, A. J. Kim and A. W. Lo, "Consumer credit-risk models via machine-learning algorithms", Journal of Banking and Finance, 2010. Available:https://dspace.mit.edu/bitstream/handle/1721.1/66301/Consumer%20Credit%20Risk.pdf?sequence=1&isAllowed=y. [Accessed 13 December 2019].

[10] R. Singh and R. Aggarwal, "Comparative Evaluation of Predictive Modeling Techniques on Credit Card Data", International Journal of Computer Theory and Engineering, pp. 598-603, 2011. Available: 10.7763/ijcte.2011.v3.377 [Accessed 14 December 2019].

[11] Y. I. Cheng & L.Che-hui. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, pp 2473-2480, 2009. Available: https://doi.org/10.1016/j.eswa.2007.12.020 [Accessed 14 December 2019].

[12] H. Ince & B. Aktan. A comparison of data mining techniques for credit scoring in banking: A managerial perspective. Journal of Business Economics and Management. 2009. Available: https://www.tandfonline.com/doi/pdf/10.3846/1611-1699.2009.10.233-240?needAccess=true [Accessed 14 December 2019].

[13] T. Chou and M. Lo, "Predicting Credit Card Defaults with Deep Learning and Other Machine Learning Models", International Journal of Computer Theory and Engineering, vol. 10, no. 4, pp. 105-110, 2018. Available: 10.7763/ijcte.2018.v10.1208 [Accessed 14 December 2019].

[14] G. Batista, R. Prati and M. Monard, "A study of the behavior of several methods for balancing machine learning training data", ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, p. 20, 2004. Available: 10.1145/1007730.1007735 [Accessed 14 December 2019].

[15] F. Kaya, "Discretizing Continuous Features for Naive Bayes and C4.5 Classifiers". Available: https://pdfs.semanticscholar.org/680d/e400a534028b548c2297b8e8ddd904ebbd56.pdf. [Accessed 14 December 2019].

[16] G. XIA and W. JIN, "Model of Customer Churn Prediction on Support Vector Machine", Systems Engineering - Theory & Practice, vol. 28, no. 1, pp. 71-77, 2008. Available: 10.1016/s1874-8651(09)60003-x [Accessed 15 December 2019].

[17] Y. Xie, X. Li, E. Ngai and W. Ying, "Customer churn prediction using improved balanced random forests", Expert Systems with Applications, vol. 36, no. 3, pp. 5445-5449, 2009. Available: 10.1016/j.eswa.2008.06.121 [Accessed 15 December 2019].

[18] I. M.M.Mitkees, S. M Badr and A. Ahmed Ibrahim Bahgat ElSeddawy, "Customer Churn Prediction Model using Data Mining techniques", International Computer Engineering Conference (ICENCO), 2017. Available: https://www.researchgate.net/publication/323135664_Customer_churn_prediction_model_using_data_mining_techniques. [Accessed 15 December 2019].