DOI: 10.21917/ijsc.2020.0291

PREDICTIVE ANALYSIS OF CUSTOMER CHURN IN TELECOM INDUSTRY USING SUPERVISED LEARNING

Shreyas Rajesh Labhsetwar
Department of Computer Engineering, Fr. C. Rodrigues Institute of Technology, India

Abstract

Customer acquisition and retention is a key concern for several industries and is particularly acute in fiercely competitive and fast growth businesses. Retaining a loyal customer is far more important than acquiring a new one, thus making customer churn one of the critical concerns for big corporations. Finding factors triggering customer churn is vital to implement necessary remediation to pre-empt and cut back this churn. This research focuses on implementing machine learning (ML) algorithms to identify potential churn customers, categorise them based upon usage patterns, and visualize the analysis results. Results show that Extra Trees Classifier, XGBoosting Algorithm and Support Vector Machine have the best churn modelling performance, particularly for 80:20 dataset distribution with average AUC scores of 0.843, 0.787 and 0.735 respectively and low false negatives. The research demonstrates that ML algorithms can successfully predict potential customer churn and help in devising customer retention programmes.

Keywords:

Customer or Client Retention, Customer Churn, Telecommunication Industry, Machine Learning

1. INTRODUCTION

Telecom industry is constantly evolving and innovating. Due to the ever-increasing competition among the corporations, an increased importance is being given to targeted promoting methods for customer churn management. Modern day customers expect the best services at affordable rates. In case they are not satisfied, they quickly switch to another telecom network. Companies must find innovative ways to predict potential customer churn in order to prosper in such a competitive market.

Customer churn is defined as the proportion of customers who stopped using a particular company's products or services during a definite time frame.

Mathematically,

$$C(T) = \frac{A(T)}{B(T)} \times 100 \tag{1}$$

where,

C represents the churn % for a time frame T,

$A(T)$ represents the total number of customers after time T,

$B(T)$ represents the total number of customers before time T.

Few statistical studies have shown the impact of customer churn on the industry. Research [1] showed that a 1% increase in customer retention strategies equipped with adequate increase in corresponding budget may decrease the churn rate by up to 5% and lead to vast increment in revenues earned. Research [2] showed that cost of retaining a loyal customer is much lesser than that of acquiring a new one. Furthermore, once a company starts losing its customers, the rapid decline in revenues makes it financially infeasible to initiate new retention programs. Hence, predictive analysis of customer retention is an absolute necessity in all businesses, especially in the case of Telecommunications Industry.

Machine learning algorithms provide the best solution to predictive analysis of customer churn. Companies can use such ML pipelines to initiate retention strategies on those customers who are classified as likely targets of churn. In this research, multiple classification algorithms are employed and their results are compared to discover the most accurate algorithm for prediction of customer churn in businesses. The various algorithms used include Logistic Regression, Gaussian Naive Bayes, AdaBoost Classifier, XGB Classifier, SGD Classifier, Extra Trees Classifier, and Support Vector Machine. Also, multiple evaluation parametric are used to determine the best overall performing model for the prediction of churn at the earliest stages.

2. RELATED WORKS

Significant research has been performed for prediction of customer churn in telecom companies in the past decade, where most efforts focus on binary classification. Before the evolution of machine learning, several data mining approaches were used to analyse the data. Much research has been performed at the European financial services companies by investigating the predictors of churn and the impact of customer relationship management (CRM) on churn rates. Several international organizations have developed standards regarding recognition and sharing of the best set of practices in customer service in order to reduce customer attrition.

Tsai and Lu [6] combined two different ML models for customer churn prediction. They combined Back Propagation Artificial Neural Networks and Self Organizing Maps. The first process consisted of data reduction by filtering out unrepresentative data using a hybrid model combining two neural networks. Thereafter, the data was classified using Self Organised Maps (SOM). The performance of proposed hybrid model was evaluated using three different kinds of testing sets; the general testing set and two fuzzy testing sets. They demonstrated that the hybrid model outperforms a single neural network baseline model in terms of prediction and classification accuracy [6].

Ghorbani et al. [7] focused on creating a custom classification model called locally linear model tree (LOLIMOT) using a combination of Artificial Neural Networks, Fuzzy modelling, and Tree based models. They showed that the model outperformed similar classification algorithms including artificial neural networks, decision trees and logistic regression. Also, LOLIMOT achieved accurate classification even in extremely unbalanced datasets [7].

Verbeke et al. [8] focused on performing classification using two data mining and extraction methods; AntMiner Plus and algorithm learning by bootstrapped approval (ALBA). AntMiner Plus was based on the principles of Ant Colony Optimization which also allowed to include domain knowledge by imposing monotonicity constraints on the final rule-set. ALBA combined the high prediction accuracy of non-linear Support Vector Machine (SVM) model with the comprehensibility of the rule-set model. The results showed that ALBA improved the learning of classification techniques thereby resulting in comprehensible models with increased accuracy and performance. AntMiner+ resulted in accurate, comprehensible, and justifiable models, unlike other modelling techniques [8].

Yeshwanth et al. [9] proposed a custom hybrid model by combining two classification algorithms: genetic programming and C4.5 Decision Tree. They used various game theory techniques to understand the community effect of churn. The games were designed with different levels of accuracy and execution time. The most important rules (if-then rules) were obtained from reinforced results to optimize the nodal network parameters including coverage, recharge methods, plans and packages [9].

3. DATASET

The dataset used in this research work is BigML churn in Telecom's Dataset from UCI Machine Learning Repository [3]. The dataset is extremely large and contains detailed information of all the parameters which are extremely important for predictive churn analysis.

The dataset consists of 3334 instances with 21 attributes. The rich set of attributes presented by the dataset helped in identifying customer churn more effectively. It consists of both churned and not-churned customer types.

The tool used for dataset visualisation is Tableau Public Software. The data is input in CSV format and then visualised using various visual elements like charts, graphs, and maps, facilitating in understanding trends, outliers, and patterns in data [21], [22].

Table.1. Dataset Attributes

Feature	Description
State	Categorical, 50 states (USA) and Columbia District
Account Length	Integer representing the active duration of an account
Area Code	Categorical Variable
Phone Number	Customer phone number
International Plan	Binary, International Plan activated or not
Voice Mail Plan	Binary, Voice Mail Plan activated or not
Number of Vmail Messages	Integer representing number of Voice Mail Messages
Total Day Minutes	Total usage in minutes during day time
Total Day Calls	Total calls made during day time
Total Day Charge	Total charge incurred during day time
Total Evening Minutes	Total usage in minutes during evening time
Total Evening Calls	Total calls made during evening time
Total Evening Charge	Total charge incurred during evening time
Total Night Minutes	Total usage in minutes during night time
Total Night Calls	Total calls made during night time
Total Night Charge	Total charge incurred during night time
Total International Minutes	Total usage in minutes for International calls
Total International Calls	Total international calls made
Total International Charge	Total charge incurred for International calls
Customer Service Calls	Number of Customer Service Calls made
Churn	Binary, Churned or Not

4. METHODOLOGIES

4.1 MACHINE LEARNING

Machine learning is a branch of Artificial Intelligence referring to systems, which learn to solve a set of prediction or detection tasks, based on learning from data. Machines can be trained in three broad ways of supervised, unsupervised, and reinforcement training methodologies [18] [19]. ML algorithms enable analysis of massive quantities of information that is otherwise not possible or extremely time-consuming for manual analysis by humans. The performance of ML systems can be quantitatively assessed using different error metrics. As a result, ML algorithms and artificially intelligent systems have found tremendous applications in several fields such as computer vision, agricultural industry, healthcare systems, communication systems, text and speech analysis, and so on.

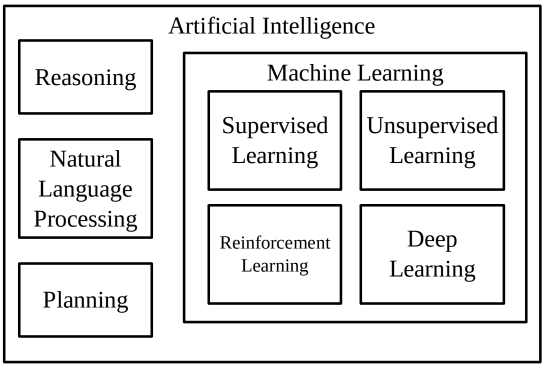

Fig.1. Overview of Artificial Intelligence

Supervised Learning Algorithms are trained on externally supplied instances to generalise a hypothesis, enabling prediction on future instances.

The primary goal of a supervised system is to estimate the parameters of a defined model using labeled sample dataset. Having a labeled dataset $(x_1,y_1),(x_2,y_2),...,(x_N,y_N)$, where x_k is the k^{th} tuple from the training dataset, and $y_k \in \{1,2,...,C\}$ is the corresponding label (target), a supervised algorithm attempts to best fit itself into the training data samples [4].

Once the parameters of a model are all set, it is possible to predict the output label of a query sample *xq*. The entire supervised learning process essentially includes three stages of training, validation, and testing. Accordingly, the labeled dataset is partitioned into three subsets. The first step involves feeding the set of training samples to an estimator to build a predictive model. In the validation phase, the hyper-parameters of the model are tuned using the validation set. Finally, the performance of the model is measured on the testing set. Another extremely important aspect which needs to be addressed in machine learning algorithms is overfitting. It makes the model fit extremely tightly to the training dataset, due to which, it does not end up generalising the hypothesis, and makes incorrect predictions on new and unseen data. The efficiency of a supervised system for making correct predictions on the new data depends on how large is the number of labeled samples of training data compared to the parameters of the predictive model.

However, in many practical scenarios, the available dataset is often un-labeled. Unsupervised algorithms are those which learn to extract patterns from a set of unlabeled data. These patterns help in clustering the data points into multiple classes.

Reinforcement algorithms learn through interactions with their environments. The learning procedure consists of a series of trials, which receive a feedback of an error or a delayed reward, helping the system distinguish the desired situation.

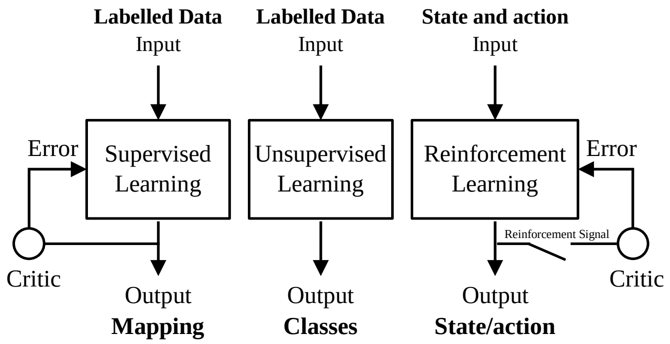

Fig.2. Models for Machine Learning

Predictive models are further categorised to two types of classification and regression. Classification methods are used to predict discrete categories or target classes for an input object using its defined attributes. Whereas, in regression problems, the target is a continuous variable. The task of regression is to build a model, capable of finding the output value of a new input variable.

5. DATA PREPROCESSING

Step 1: *Importing the required libraries*

The entire research is performed on Python 3.7.3 programming language. Python has earned the title of one of the most popular languages for machine learning tasks due to its vast collection of libraries. The machine learning libraries used in this research are:

- NumPy
- Matplotlib
- Pandas
- SkLearn

Step 2: *Importing the Dataset*

The dataset is present in CSV format consisting of tabular data stored in plain text. The read_csv() method of pandas library is used to create a data frame of the given dataset.

Step 3: *Handling Missing Values*

The dataset chosen for this research had been prepared extremely carefully and did not contain any missing values.

Step 4: *Encoding Categorical Data*

Machine learning algorithms only deal with numerical values and not label values. Hence, the following attribute values are label encoded and converted to Boolean values:

- International Plan
- Voicemail Plan
- Churn

Step 5: *Splitting the dataset into train and test set*

The telecom dataset is partitioned into two subsets, one for training the ML models called training set and the other for evaluating the performance of the trained model called test set.

The dataset initially consists of 3334 instances. The split ratio is chosen to be 70:30 resulting in 2333 instances in the training set and 1001 instances in test set.

Step 6: Feature Scaling

Since most ML algorithms use Euclidian Distance between two data points in their computations, features possessing high variance in magnitudes lead to inconsistency in calculations. Hence, all the features are scaled using z-score normalisation technique.

6. MACHINE LEARNING ALGORITHMS

6.1 LOGISTIC REGRESSION

Logistic regression is a classification algorithm used when the independent variable is dichotomous (binary) [10]. It estimates the probability of an object belonging to a class.

The algorithm involves finding the best fit β parameters for:

$$y = \begin{cases} 1 & \beta_0 + \beta_1 x + \varepsilon > 0 \\ 0 & \text{else} \end{cases} \tag{2}$$

where ε represents the error distributed by the standard logistic distribution.

The logistic function is a sigmoid function, which takes any real input *t*, where $t \in$ Real Values, and outputs a binary value. The algorithm can also be interpreted as taking input log-odds and having output as the probability. The standard logistic function given as $\sigma : \mathbb{R} \to (0,1)$ is defined as follows:

$$p(x) = \sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}} \tag{3}$$

where, $p(x)$ is interpreted as the probability of the dependent variable *Y* equaling a success/case rather than a failure/non-case.

6.2 GAUSSIAN NAIVE BAYES

Naive Bayes classifiers are a set of classification algorithms based on Bayes' Theorem of conditional probability, which finds the probability of an event occurring given the probability of another event that has already occurred. The mathematical representation is given in Eq.(4):

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)} \tag{4}$$

where,

y represents the class variable (output label), and

X is a dependent feature vector, i.e., $X=(x_1,x_2,x_3,\ldots,x_n)$.

The Gaussian Naive Bayes algorithm assumes continuous values associated with each feature to be distributed according to the Gaussian distribution/Normal Distribution [11].

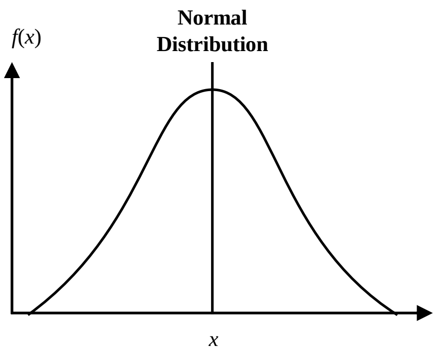

Fig.3. Normal Distribution

In Gaussian Naive Bayes Algorithm, the conditional probability is modified as:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right) \tag{5}$$

6.3 ADABOOST CLASSIFIER

Adaptive Boosting Classifier is an ensemble classifier, combining the outputs of multiple weak classification algorithms, allowing for better predictive performance as compared to any single constituent algorithm [12].

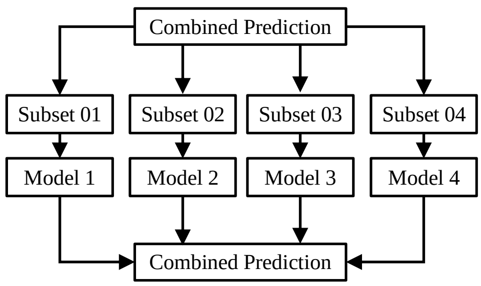

Fig.4. AdaBoost Classification Flow

The overall training set is divided into multiple random subsets, which act as the training sets for the constituent (weak) classifiers. After training a weak classifier, adaboost assigns a weight to each training tuple as well as to the weak classifier. Misclassified training tuples get assigned higher weight values so that they participate in the training of the next classifier as well with higher probability.

Weak classifiers with more accurate classification performance get assigned higher weight values enabling them to have more impact in the final outcome. The mathematical representation is given in Eq.(5):

$$F_T(x) = sign\left(\sum_{t=1}^{T} \theta_t f_t(x)\right) \tag{6}$$

where, f_m stands for the m^{th} classifier, and θ_m is the corresponding weight.

6.4 XGB CLASSIFIER

Extra Gradient Boosting Classifier is a decision tree-based ensemble algorithm that uses a gradient boosting framework. It works on the principle of boosting weak classifiers (Classification and Regression Trees) using the architecture of Gradient Descent [13].

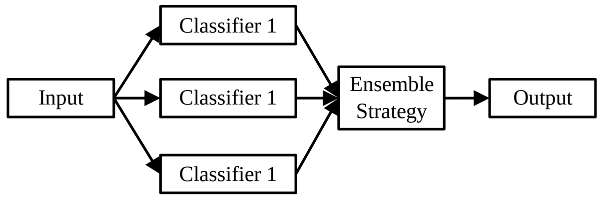

Fig.5. Parallel Ensemble

XGBoost differs from AdaBoost Classification algorithm primarily due to the inter model independence existing in XGB. Also, XGBoost employs a parallel ensemble technique, wherein multiple learners are produced during the training phase in a parallel manner [14].

6.5 SGD CLASSIFIER

Stochastic Gradient Descent involves random selection of a few samples of data from the whole data set for each training iteration. It is a stochastic (random) approximation of gradient descent algorithm [15]. In Gradient Descent, the term batch denotes the total number of samples to be selected from the dataset for calculating the gradient during each iteration.

Stochastic gradient descent (SGD) performs a parameter update for each tuple in the training set $x(i)$ and label $y(i)$ as:

$$\theta = \theta - \eta \cdot \nabla\theta J(\theta; x(i); y(i)) \tag{7}$$

Hence, in SGD, the gradient of the cost function is computed for each single tuple at each iteration instead of the gradient calculation for a batch of data.

In SGD, since only one tuple from the dataset is chosen at stochastically at each iteration, where the path taken by the algorithm to reach the minima is noisier compared with typical Gradient Descent algorithm.

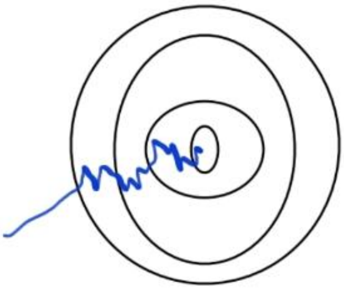

Fig.6. Minimum Cost Function Path

6.6 EXTRA TREES CLASSIFIER

Extra Trees Classifier, also called as Extremely Randomized Trees Classifier is another form of ensemble learning technique that aggregates the results of several de-correlated decision trees to give the most precise output [16].

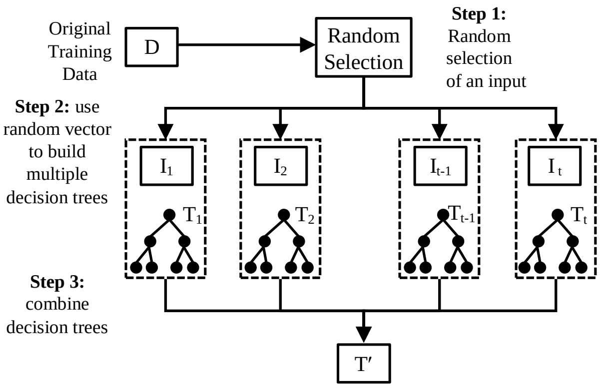

Fig.7. Random Trees Classification

Each constituent decision tree in the extra trees algorithm is constructed from the original training dataset. At each test node, each tree is given an input of a random sample of *k* features from the feature-set. Each constituent decision tree selects the best feature to split the data using a mathematical criterion such as the Gini Index [20]. This random permutation of features results in the creation of a set of de-correlated decision trees.

6.7 SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) is a supervised ML algorithm which solves both regression and classification problems [17]. The algorithm plots each data point in an n-dimensional space (n represents the number of features) where the value of each feature corresponds to the value of each coordinate.

Data classification involves finding the most optimal hyper-plane differentiating the classes perfectly.

7. PERFORMANCE METRICS

The performance of various ML classifiers was evaluated by computing their confusion matrix, and thereby determining the values of True Positive (*TP*), True Negative (*TN*), False Positive (*FP*), and False Negative (*FN*) are calculated.

- *ML* - Machine Learning
- *AUC* - Area Under the Curve
- *ROC* - Operating Characteristics
- *XGB* - Gradient Boosting
- *SGD* - Gradient Boosting
- *SVM* - Vector Machine
- *NN* - Network
- *SOM* - Self Organising Maps
- *ANN* - Artificial Neural Network
- *ALBA* - Algorithm Learning by Bootstrapped Approval

The various performance parameters used were:

Accuracy: The classification accuracy of an algorithm is directly proportional to the number of correctly classified samples (true positives and true negatives).

$$Acc = (TP+TN)/N \tag{8}$$

$$N = TP + TN + FP + FN \tag{9}$$

Sensitivity: The proportion of correctly classified positive samples is represented by the term sensitivity.

$$S = TP/(TP+FN) \tag{10}$$

Specificity or Recall: Specificity, also called as Recall value, measures the proportion of correctly classified negative samples

$$Spf = TN/(TN+FP) \tag{11}$$

Balanced Classification Rate (BCR): It is defined as the geometric mean (GM) of specificity and sensitivity.

$$BCR = \sqrt{Spf * S} \tag{12}$$

Precision: Precision measures the ratio of correctly classified positive samples to all the classified positive samples.

$$P = TP/(TP+FP) \tag{13}$$

F1-Score: It is defined as the harmonic mean (average) of precision and specificity/recall value.

$$F1_{score} = \frac{2 * P * Spf}{P + Spf} \tag{14}$$

Matthews's Correlation Coefficient (MCC): It is a measure of the quality of binary class classification. MCC is a correlation coefficient between the predicted and observed binary classifications.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{15}$$

AUC-ROC Curve: The ROC curve consists of sensitivity/ recall plotted against false positive rate (FPR) at various threshold values.

$$FPR = 1 - Spf \tag{16}$$

AUC is a measure of the area under ROC curve.

8. RESULTS

The aforementioned classification methodologies were implemented with Python 3.7.3, on BigML churn in telecom dataset [3] on laptop with 2.3GHz Intel Core i5 processor, 8GB RAM, and macOS Catalina platform.

Table.2. Classification Results

Performance Parameters	Logistic Regression	Gaussian Naïve Bayes	AdaBoost Classifier	XGB Classifier	SGD Classifier	Extra Trees Classifier	SVM
Accuracy	0.877014	0.864975	0.866992	0.926001	0.861004	0.937404	0.896513
Sensitivity	0.879132	0.915995	0.88924	0.925485	0.877997	0.935659	0.895615
Specificity or Recall	0.210144	0.507246	0.333333	0.376811	0.384057	0.405797	0.101449
BCR	0.429819	0.681641	0.544438	0.590536	0.580691	0.616188	0.301429
Precision	0.987238	0.923433	0.977958	0.994199	0.976798	0.995359	0.995359
F1-Score	0.321471	0.508359	0.403381	0.520463	0.333483	0.566632	0.176308
MCC	0.273201	0.403216	0.337228	0.651139	0.22433	0.709876	0.465488
AUC	0.569706	0.697224	0.608544	0.787099	0.564486	0.843911	0.73536

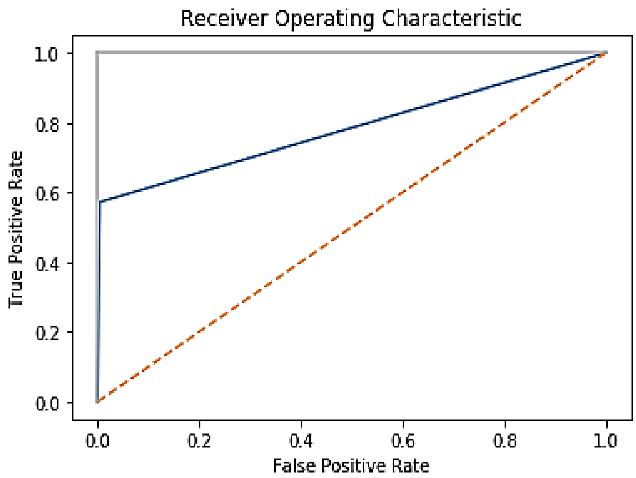

Fig.8. ROC Curve of Extra Trees Classifier

The results of this research do not necessarily imply any one particular algorithm as the best classifier across performance metrics. Even so, Extra Trees Classifier, XGBoost Classifier and SVM perform well, based on average AUC scores, with Extra Trees Classifier possessing the highest absolute AUC score of 0.8439. The Table.2 details the performance results of all the classifiers across all the performance metrics.

As discussed, additional metrics are used to further assess the performance of various classifiers to help determine the actual detection capabilities, particularly when the correct detection of real churn cases is more important than detecting non-churn ones, since the scale of retention programmes directly affects the overall cost. The tradeoff between Sensitivity and Specificity/Recall is analysed with the help of Balanced Classification Rate (BCR).

Extra Trees Classifier, XGBoost Classifier and SGD Classifier possess high BCR values of 0.6161, 0.5905, and 0.5806 respectively whereas SVM has a low BCR value of 0.3014. Gaussian Naive Bayes Classifier possesses the highest absolute BCR value of 0.6816. A balance between precision and specificity or recall can be achieved using the F1-Score. Extra Trees Classifier possesses the highest F1-Score of 0. 5666. It is followed by XGBoost Classifier with a score of 0.5204 and Gaussian Naive Bayes Classifier with a score of 0.5083. The lowest F1-Score is earned by SVM with a value of 0.1763

Since the classes are of different sizes, MCC is used as a balanced measure which accounts both true and false positives and negatives to give a value between 1 and -1. Highest MCC value of 0.7098 is achieved by Extra Trees Classifier. It is closely followed by XGBoost Classifier with a value of 0.6511.

9. CONCLUSION

Currently giant corporations including the telecommunication industry is battling with the problem of customer churn. It negatively impacts customer retention rate, making it even harder to acquire new customers due to dwindling revenues. Predictive analysis of customer churn can help tackle this menace by helping companies identify vulnerable customers and implement customer centric retention measures. Machine Learning algorithms provide the perfect solution to the proposed predictive analysis.

The beginning of this paper highlights the immense threat of customer churn to the telecom industry, supported by statistical data. Thereafter, a comprehensive study of related research work is presented. The paper investigated into the realm of supervised learning ML methodologies which were experimented on the above-mentioned dataset. The findings of the research conclude that Extra Trees Classifier is the best overall learner with an AUC score of 0.8439. XGB Classifier and SVM also provide great prediction capability to tackle the menace of customer churn haunting the telecom industry.

REFERENCES

[1] P. Kisioglu and I.Y. Topcu, "Applying Bayesian Belief Network Approach to Customer Churn Analysis: a Case Study on the Telecom Industry of Turkey", *Expert Systems with Applications*, Vol. 38, No. 6, pp. 7151-7157, 2011.

[2] B. Huang, M.T. Kechadi and B. Buckley, "Customer Churn Prediction in Telecommunications", *Expert Systems with Applications*, Vol. 39, No. 1, pp. 1414-1425, 2012.

[3] Bigml, "Churn in Telecom's Dataset", Available at: https://bigml.com/user/francisco/gallery/dataset/5163ad540c0b5e5b22000383

[4] J.E.T. Akinsola, "Supervised Machine Learning Algorithms: Classification and Comparison", *International Journal of Computer Trends and Technology*, Vol. 48, No. 3, pp. 128-138, 2017.

[5] S. Sperandei, "Understanding Logistic Regression Analysis", *Biochemia Medica*, Vol. 24, No. 1, pp. 1-12, 2014.

[6] C.F. Tsai and Y.H. Lu, "Customer Churn Prediction by Hybrid Neural Networks", *Expert Systems Applications*, Vol. 36, No. 10, pp. 12547-12553, 2009.

[7] A. Ghorbani, F. Taghiyareh and C. Lucas, "The Application of the Locally Linear Model Tree on Customer Churn Prediction", *Proceedings of International Conference of Soft Computing and Pattern Recognition*, pp. 472-477, 2009.

[8] W. Verbeke, D. Martens, C. Mues and B. Baesens, "Building Comprehensible Customer Churn Prediction Models with Advanced Rule Induction Techniques", *Expert Systems with Applications*, Vol. 38, 2011, pp. 2354-2364.

[9] V. Yeshwanth, V.V. Raj, and M. Saravanam, "Evolutionary Churn Prediction in Mobile Networks using Hybrid Learning", *Proceedings of 24th Florida Conference on Artificial Intelligence Research Society*, pp. 471-476, 2011.

[10] S.C. Bagley, H. White and B.A. Golomb, "Logistic Regression in the Medical Literature: Standards for Use and Reporting, with Particular Attention to one Medical Domain", *Journal of Clinical Epidemiology*, Vol. 54, No. 10, pp. 979-985, 2001.

[11] Irina Rish, "An Empirical Study of the Naive Bayes Classifier", *Proceedings of 17th International Joint Conference on Artificial Intelligence*, pp. 1-6, 2001.

[12] A. Tharwat, "AdaBoost Classifier: An Overview", Machine Learning Project, 2018.

[13] R. Santhanam, N. Uzir, S. Raman Sunil and S. Banerjee, "Experimenting XGBoost Algorithm for Prediction and Classification of Different Datasets", *Proceedings of National Conference on Recent Innovations in Software Engineering and Computer Technologies*, pp. 1-4, 2017.

[14] Matteo Re and Giorgio Valentini, "*Ensemble Methods: A Review*", Chapman and Hall Publisher, 2012.

[15] Thanh-Nghi Do, "Parallel Multiclass Stochastic Gradient Descent Algorithms for Classifying Million Images with Very-High-Dimensional Signatures into Thousands Classes", *Vietnam Journal of Computer Science*. Vol. 1, No. 2, pp. 107-115, 2014.

[16] Pierre Geurts, Damien Ernst and Louis Wehenkel, "Extremely Randomized Trees", *Machine Learning*, Vol. 63, No. 1, pp. 3-42, 2006.

[17] Theodoros Evgeniou and Massimiliano Pontil, "Support Vector Machines: Theory and Applications", *Proceedings of International Conference on Machine Learning and its Applications*, pp. 249-257, 2001.

[18] Mohssen Mohammed, Muhammad Khan and Eihab Bashier, "*Machine Learning: Algorithms and Applications*", CRC Press, 2016.

[19] F. Fabris, D. Magalhaes, J. Pedro and A. Freitas, "A Review of Supervised Machine Learning Applied to Ageing Research", *Biogerontology*, Vol. 18, No. 2, pp. 171-188, 2017.

[20] Haim Shalit, "Calculating the Gini Index of Inequality for Individual Data", *Oxford Bulletin of Economics and Statistics*, Vol. 47, No. 2, pp. 185-189, 2009.

[21] Sarah Murphy, "Data Visualization and Rapid Analytics: Applying Tableau Desktop to Support Library Decision-Making", *Journal of Web Librarianship*, Vol. 7, No. 2, pp. 465-476, 2013.

[22] M. Sadiku, A. Shadare, S. Musa, C. Akujuobi and R. Perry, "Data Visualization", *International Journal of Engineering Research and Advanced Technology*, Vol. 2, No. 12, pp. 11-16, 2016.