Forecasting Customer Churn in the Telecommunications Industry

Mr. Kritarth Gupta
Department of Computer Science
MPSTME, NMIMS University
Mumbai, India
kritarth.gupta118@nmims.edu.in

Mr. Atharva Hardikar
Department of Computer Science
MPSTME, NMIMS University
Mumbai, India
atharva.hardikar056@nmims.edu.in

Mr. Devansh Gupta
Department of Computer Science
MPSTME, NMIMS University
Mumbai, India
devansh.gupta379@nmims.edu.in

Dr. Shweta Loonkar
Computer Science Department
Visiting Professor
MPSTME, NMIMS
Mumbai, India
shwetaloonkar@gmail.com

***Abstract*— Data mining is a broad field that helps the company to combine statistics, databases, machine learning, and artificial intelligence. As the size of the company grows, so do such situations, making it impossible for a normal information system to manage such perilous scenarios. Due to this companies face significant income loss since customers are leaving the firm for unexplained reasons. It is well acknowledged that acquiring new clients is more cash intensive than maintaining existing ones and hence customer management is critically impacted by customer churn, which happens when a customer decides he no longer wants to keep in touch with the company. Traditional market research methodologies are challenging to support the churn problem. There is still much potential for improvement in churn forecast accuracy despite the development of several churn prediction tools that look at hundreds of parameters. Ultimately, this research will aid in the analysis of consumer behavior and the categorization of whether or not a client is churning through the use of a variety of data mining approaches to predict customer churn. Using a data set available on Kaggle's website, this study tested multiple classifiers on the problem of predicting customers' propensity to leave a company. In this study, we utilized Kaggle's online data set to predict customer churn behavior using several classifiers, including Random Forest, Logistic, J48, Stacking, ADA Boost, Decision Table, and Logit Boost, and observed that our model achieved 93.55 percent accuracy.**

Keywords— Churning of data, Big Data, Customer churn prediction, Artificial Intelligence

SECTION - I

I. INTRODUCTION

Data mining is a fundamental component of problem solving that is used to extract valuable data from large databases. It is a highly effective instrument in terms of rising business trends and approaches. It assists businesses in making sound business choices, boosting gains, and improving prolificacy. Data mining, in a nutshell, is the science of extracting useful information from large databases via various processes of data collection, cleaning, processing, analysis, and interpretation [1].

Customer churn refers to customer income or customers who leave a certain service and move to a rival. Customer churn has a significant influence on company success since it is directly related to essential business objectives. However, in the corporate sector, where the customer is the most valuable asset, decreasing customer turnover is the primary aim. As a result, customer satisfaction is the primary goal in maintaining loyal and long-term relationships with current customers [2]. So, customer churn management is the management of customers who leave one firm for another.

In order to get a competitive edge and maximize customer data mining is critical. Various data mining approaches are used to find high value and low risk consumers in order to maximize income. Using techniques like customer relationship management, market campaign management, customer clustering, and customer segmentation, these methods are also utilized to keep an eye on consumer transaction data. Data mining enables more personalized service and focuses on each individual customer's needs and interests. Which is incredibly time-saving, when every client is alerted with new and interesting goods and provides them with excellent happiness and customer is the major aim in customer retention management. Customer turnover has become a hot concern in the telecommunications industry [3]. Keeping current customers costs less than acquiring new customers. As a result, client turnover is very complicated in the telecommunications industry. Existing customers play an important part in improving marketing and sales via referrals since they are satisfied, trustworthy, and pleased with your brand or product. They are quite aware of the assistance you provide. To what extent their wants, interests, and anticipations are met. They are relied upon to use a potentially dangerous item with an optimistic approach. Even if they do not purchase further products, they become referral clients. However, nothing beats a sales lead from a satisfied and trusted customer. Such a referral consumer is also a motivation to increase marketing and sales effectiveness.

Existing customers are more likely than new customers to purchase the company's other products. This has a direct impact on the company's earnings and cash flow. Customers who return are less price sensitive. They may also result in sales recommendations and connections [4].

We conducted a series of experiments to test the effectiveness of several customer churn algorithms using the Kaggle data set, and we came to the conclusion that J48 is the preferable algorithm.

The portions of our paper are Section-I and Section-II. In the first part, we describe the prior work of other scholars. In the second section, we cover the source of the dataset, information about the dataset, the technique employed, the results and a discussion of the acquired results, and we conclude with our conclusions.

II. LITERATURE REVIEW

Our literature review examines current research as well as prior studies conducted by different researchers in the telecom business in order to anticipate customer turnover. Maintaining the Integrity of the Specifications

Muhammad Turki Alshurideh reviewed the situation of customer retention competition, growing client switching rates, and the character of distributing classifications. From several perspectives, he claimed that correct categorization for customer-supplier relationships alleviate client retention concerns [5]. In a related manner, the advantages of a link are critical to its beginning, development, and continued existence. As a result, it is important to investigate how customers feel about upkeep in relation to a variety of behavior-related issues, such as the influence study has paid less consideration to the impact on customers' purchase and upkeep choices of utility outcomes driven by prebehavioral prior enhancements. especially in some industries (e.g., the cell phone area).

In their study, Eunju et al. hypothesized that customer turnover might be identified using a decision tree data mining system known as the exhaustive Chi-square Automatic Interaction Detector (E-CHAID). According to their findings, an E-CHAID-based model might be useful for identifying probable churners and their churn behavior [06]. In Eva et al experiment's it was intended that certain consumers made plan suggestions while others did not. They discovered that 6% of consumers are attracted by plan suggestions, while 10% of customers apply for plan recommendations. The effect focused encouragement programs are accessed by the authors [7].

Vicente et al. suggested a novel life time model for a real-world customer turnover data set. Their model is an extension of the standard one-factor survival/cure rate model [8] to include additional variables. Semantics-driven subtractive clustering was the new method utilized by Wenjie et al (SDSCM). By comparison to SCM and FCM, they showed that SDSCM provides much stronger semantic clustering results [9].

The agent-based modelling and simulation (ABMS) application was presented by Chiozie et al. in the Bell Study as a novel approach to analyzing consumer loyalty. In order to learn what factors contribute to customers leaving their existing telecoms providers and how they act, ABMS is used. Data analysis also identifies location and mobile device preference [24].

Indepreet et al. examined the RFM (Recency Frequency Monetary) paradigm in their studies. As a consequence, clients are separated into three groups that are targeted by various plan offers based on comparable criteria. 3-V (data Volume, Variety, and Velocity) problems are addressed in their study [10]. Author, divide your consumers into subsets according to the structured and unstructured data you have gathered, and then focus on your most valuable, at-risk customers. Concentrating on certain types of customers allows you to create tailored products and services. In this article, author John Hadden [11] employed an innovative approach for predicting turnover consumers utilizing complaints data and three separate algorithms. For categorization, he utilized 202 records and 700 test records. He used a neural network, a decision tree, and regression to classify items, with the neural network yielding the best results (90 percent). Because different authors utilized different approaches based on significance and relevance, Saad et al. [12] used an online data source and three algorithms to evaluate the literature. In terms of attribute selection, he computes the p-value of the attributes for data balance by resampling the data set. Using a decision tree, he achieves an average of 75% accuracy by inserting five more derived attributes, after reaching 70% accuracy with the initial dataset.

Manpreet et al. [13] solely utilized the decision tree J48 method in their research and represented their results using graphs and other graphics. He made use of the same dataset provided from Kaggle. Furthermore, a clone of the dataset is accessible on the websites of 'SGI' and 'data mining consultant' [14], [15]. In this work, Akmal et al. [16] employed four vendors of Pakistani telecom operators, including Telenor, Zong, Warid, Ufone, and Mobilink. The author utilized an interview strategy using a tape recorder to collect data. This study discovered that customer turnover occurs for a variety of reasons, including price, call quality, network quality, spam messages, international roaming, and hidden costs. Following all of the references, the creator of the data set [17] has made our dataset findings public on Kaggle. He employed two algorithms, Logistic Regression and Random Forest, using a K-10-fold technique, with the accuracy of Logistic Regression being 87% and Random Forest being 93%.

SECTION – II

III. METHODOLOGY

This part was further broken into six sub sections based on various elements.

A. WEKA

Waikato University in New Zealand created WEKA, a data mining framework written in Java [18] that can be used to construct Machine Learning algorithms and tackle practical data mining problems. WEKA is state-of-the-art software that facilitates a wide variety of data mining tasks, including preprocessing, classification, regression, clustering, and association rules, all of which may be applied straight to data sets with the aid of a variety of user-friendly visualization tools [18]. For the most part, WEKA reads and writes CSV and ARFF (Attribute-Relation File Format) files. Attribute Name, Attribute Type, Attribute Values (Possible Data), and Attribute Data are only few of the numerous identifying labels used in ARFF.

B. DATA

The information used in this research came from the Kaggle online data store [17]. All of the information needed for the investigation is included in the dataset. Prior to deciding on the focal characteristics, a literature review of the relevant domain was performed. Data may be seen in Microsoft Excel, and we export the whole dataset as a CSV file for processing in WEKA, which only accepts CSV and ARFF files. There are a total of 21 features and 3333 recordings in the collection. There are no outliers or blanks in this dataset.

The nominal attributes* that we used are State, Area Code, Phone Number, Account Length, Voice mail plan, International Plan, Voice mail messages, Total day calls, Total day minutes, Total day charges, Total Evening minutes, Total Evening calls, Total Evening charges, Total Night Minutes, Total Night Calls, Total Night Charges, Total International Minutes, Total International Calls, Total International Charges, Customer Service Calls.

The class attributes** that we used is churn.

*Nominal Attribute: It provides a sufficient number of characteristics to separate one object from another.

**Class Attribute: This kind of attribute is a set of nominal values that can grow and change over time.

C. PRIOR PROCESSING OF DATA

Most real-world data sets are incomplete, hard to read, have missing values, or have noise or outliers. Data pre-processing is a way to fix these problems and turn the data into a format that is easier to read. Data pre-processing is a lot of work and involves many different steps, such as cleaning the data, changing it, separating it, etc. In our dataset, we have no missing value.

D. AIM OF THE DATASET

We choose our focus qualities and data collection based on the importance gleaned from the literature research. With a total of 3333 records, our dataset contains 20 nominal attributes (N: Nominal Attribute) and 1 class attribute (C: Class Attribute).

IV. EXPERIMENTAL WORK

All of the experiments and their procedures are described here.

A. ALGORITHMS USED

1) **J48:** J48 is a decision tree classification algorithm for machine learning that is based on Iterative Dichotomiser 3. It helps a lot to look at the data in groups and in a continuous way [19].
2) **Logistic Regression:** Logistic regression is an algorithm that is part of a technique called "Supervised Learning." It is used to predict the categorical dependent variable given a set of independent variables. Logistic regression predicts the outcome of a categorical dependent variable [19].
3) **Random Forest:** Random Forest is a classifier that takes the average of a number of decision trees on different subsets of a given dataset to enhance the predicted accuracy of that dataset. Instead than depending on a single decision tree, the random forest forecasts the final output based on the majority of predictions from each tree [19].
4) **Stacking:** In stacking, an algorithm takes the outputs of sub-models as input and tries to learn the optimal way to combine the input predictions to get a more accurate output prediction [19].
5) **ADA Boost:** AdaBoost, also known as Adaptive Boosting, is a Machine Learning approach used as an Ensemble Method. The most popular method used with AdaBoost is one-level decision trees, or decision trees with just one split. These are also known as Decision Stumps [19].
6) **LogitBoost:** LogitBoost is a method for boosting classification. Both LogitBoost and AdaBoost perform an additive logistic regression, so they are similar. The difference is that AdaBoost tries to reduce the loss as much as possible while LogitBoost tries to reduce the loss as much as possible [20].
7) **Decision Table:** Decision Table is nothing more than a tabular representation of all circumstances and actions. Decision Trees are always used if the processing logic is very complex and incorporates several circumstances. Conditions Stubs, Action Stubs, and rules are the primary building blocks used to construct the Data Table [19].

B. TRAINING MODEL USING ORIGINAL DATASET

In our first experiment, we want to determine the best algorithm for predicting customer attrition in the telecom industry. In our first experiment, we utilized the whole dataset without any changes for experimental work and used seven different methods for classification. Table 1, displays the results and time required by the classifier algorithms in Experiment 1.

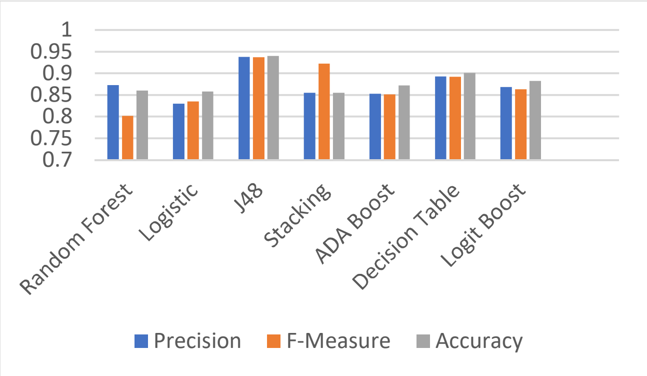
Table 1: Accuracy of Classifiers with Original Dataset

Algorithm	TPR	FPR	Acc	Precision	Recall	F-Measure	Time (s)
Random Forest	0.860	0.821	0.860	0.873	0.860	0.802	1.57
Logistic	0.859	0.660	0.858	0.830	0.859	0.835	0.48
J48	*0.940*	*0.273*	*0.940*	*0.938*	*0.940*	*0.937*	*0.73*
Stacking	0.855	0.855	0.855	0.855	0.855	0.922	0.01
ADA Boost	0.873	0.618	0.872	0.853	0.873	0.851	0.14
Decision Table	0.902	0.457	0.901	0.893	0.902	0.892	0.58
Logit Boost	0.883	0.586	0.882	0.868	0.883	0.863	0.14

ANALYSIS OF FIRST EXPERIMENT

In experiment 1, we construct two more measures of accuracy after developing the model (training the model with the whole dataset) using the original data set (to assess the accuracy level of the other techniques), as shown in Table 1. Precision and F-Measure both examine the time as well as the accuracy of the classifiers in experiment 1. According to the statistics in the preceding table, the J48 algorithm outperforms with an accuracy of 94.02%, and we also obtained the second highest accuracy level from the Decision Table algorithm with a fair percentage, taking nearly one second to build the classifier, whereas the Decision Table algorithm only takes 1.57 seconds to train the model. As a result, J48 and Decision Table have a 0.03% difference, implying that J48 performs irrationally better than Decision Table. The findings were also classified graphically, as shown in Fig.1

Fig 1: Accuracy Level of Classifiers in Experiment 1



C. TRAINING MODEL USING MODIFIED DATASET

After the first experiment, we discovered that certain algorithms in our experiment require additional time due to the vast quantity of data, so we considered another alternative, and the gap of the first experiment is the goal of our second experiment. To lower the runtime complexity, we will need to reduce the data set, which may be done by reducing attributes/features or records. Because record reduction is problematic in this scenario, we opt either feature subset selection or attribute minimization. After finishing this procedure, we receive the characteristics displayed in the data section above to create the classifiers, and the classifier results are provided in Table 2.

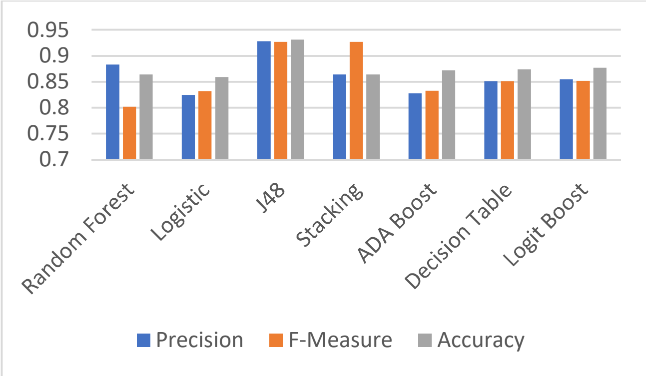
Table 2: Accuracy of Class

Algorithm	TPR	FPR	Acc	Precision	Recall	F-Measure	Time (s)
Random Forest	0.864	0.861	0.864	0.883	0.864	0.802	1.3
Logistic	0.860	0.713	0.859	0.825	0.860	0.832	0.61
J48	*0.931*	*0.336*	*0.931*	*0.928*	*0.931*	*0.927*	*0.27*
Stacking	0.864	0.864	0.864	0.864	0.864	0.927	0.05
ADA Boost	0.863	0.725	0.872	0.828	0.863	0.833	0.31
Decision Table	0.875	0.661	0.874	0.851	0.875	0.851	0.24
Logit Boost	0.877	0.667	0.877	0.855	0.877	0.852	0.72

ANALYSIS OF SECOND EXPERIMENT

According to the data in the preceding table, the J48 method and the Logit Boost algorithm outperform with accuracy of 93.1% and 87.7%, respectively, in experiment two. In terms of time complexity, J48 reduces the time although Logit Boost has a low time and practically the same accuracy level, thus we can claim that J48 performs irrationally better than Logit Boost. The findings were also classified graphically, as shown in Fig.2

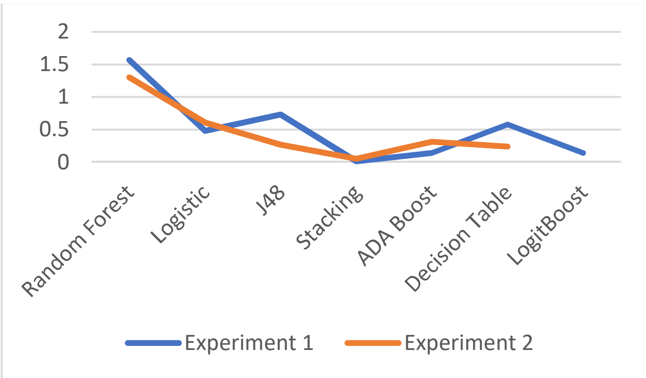
Fig 2: Accuracy Level of Classifiers in Experiment 2



D. COMPARISION OF BOTH EXPERIMENTS

We attempt a comparison of the results of two separate trials. As can be seen in Fig. 3, the WEKA classifier algorithms increase their accuracy as they are trained, and the J48 algorithm is the most time-efficient overall, making it the ideal choice for predicting customer churn in the telecom industry.

Fig 3: Time Complexity of both experiments



E. LIMITATIONS OF J48

The J48 algorithm is based on decision trees and hence shares some of the limitations of such methods. Some problems with this algorithm are as follows:

1) **Empty Branches:** When it comes to rule creation, the J48 method relies heavily on the construction of trees with meaningful values. On the other hand, these variables just serve to make the tree more extensive and convoluted without really aiding in the creation of any classes for the categorization jobs.

2) **Insignificant Branches:** The number of possible divisions in a decision tree is the same as the number of different attributes that you choose. However the truth is that not all of them are useful for sorting tasks. These unimportant branches not only make decision trees less useful, but they also make the problem of "overfitting" worse.

3) **Over Fitting:** When an algorithm display gets information with unusual features, this is called "over fitting." This makes the process distribution have a lot of fragmentations, which are statistically unimportant nodes. Most of the time, the J48 algorithm builds trees with branches that are "just deep enough to perfectly classify the training examples." This method works best with data that is free of noise. Most of the time, though, this method makes the training examples too much like the noisy data. There are currently two popular approaches to avoiding overfitting in decision tree learning. The first is that a tree's growth should be limited before it reaches the optimal height at which it can categorize training data with the greatest precision. Allowing the tree to match the training data too well and then pruning it is the second method.

F. ESCALATIONS IN J48

1) USING BETTER ENTROPY SCHEME: The most important part of putting together decision trees is choosing the best attribute to decide which level the current node is on. For each value in the training set, the best calculated entropy is used to figure out where each node should be placed. In general, training sets adhere to certain restrictions, such as include one or more samples from the same class, no samples, or a diverse selection of classes in the training set. These issues may lead the resulting decision tree to incorrectly represent the generality. Because of this, people have come up with different ways to figure out the entropy gain more accurately. A better way to figure out the entropy gain can make the J48 algorithm better.

2) USING NON-LINEAR HYBRID CLASSIFICATIONS: Most decision trees are made by partitioning from the top down. It finds the optimal linear split for maximizing entropy gain by iteratively exploring the training set. The resulting tree is often excessively large and a poor fit to the data, thus it must be trimmed by considering whether or not each intermediate node would be better as a leaf. After pruning, a leaf node's local set of training cases can get quite big, and taking the majority class may make the tree too general. From the basic decision tree learning algorithm, the J48 logic can be expanded so that a different kind of model can be used in any of the leaves instead of the majority rule. During post-pruning, a decision is made about whether or not to replace a simple leaf with a different type. Most of the time, the alternative leaf classifiers are added after the tree has been pruned. Before replacing a leaf node, the estimates of how wrong they are compared. The resulting hybrid algorithms take the strengths of both top-down, entropy-based decision tree induction and the corresponding alternative leaf models.

3) BY META LEARNING: In order to get the best model and results from data mining, it is usually necessary to set the parameters that the algorithm uses. Experiments show that when the right parameters are used, accuracy goes up by a lot. Changing the parameters of most data mining algorithms, on the other hand, comes with a problem. Finding the best parameters for this task could take a lot of time and work, or you could have to rely on assumptions that could make the results wrong. Most algorithms and tools used for data mining today need to be set up before they can be used. In other words, users must set the right values for the parameters before they can get good results or models. To find the right settings, users must have a certain amount of knowledge. To solve this problem, data mining can be used to find out how the algorithms worked in the past so that the parameters can be chosen better in the future based on how the algorithm worked in the past. Meta-learning is the study of rules-based methods that use meta-knowledge to create useful models and solutions by adapting machine learning and the data mining process. Some parameters in the decision tree model change how much pruning is done. By cutting down trees, the model can be made to work faster and classify things more accurately.

V. CONCLUSION

Customer churn is what happens when a customer decides to stop doing business with a company. Management of customer churn is an important part of customer management. To make more money and get a bigger share of the market, telecommunications companies are now focusing on finding customers who are valuable and likely to switch providers. Everyone knows that getting a new customer is more expensive than keeping an old one. Customers leave the company for unknown reasons, which is a problem. In our research, we try to find interesting patterns that can help us predict how customers in the telecom sector will behave when they decide to leave. We can sum up what we've learned by saying that the J48 outperforms with a 93.5% accuracy using 21 attributes and 3333 records from the Kaggle online Data Repository. Even though it has some flaws, the J48 algorithm is a useful tool for classifying data. It can be improved to make it a very accurate classification algorithm. Meta-learning is another big improvement to this algorithm. It uses old data and old runs of the algorithm to better classify the new data. More ways to improve this algorithm will need to be found in order to make it even more reliable and accurate.

REFERENCES

[1] Han, Jiawei, Jian Pei, and Micheline Kamber "Data mining: concepts and techniques" ISBN: 978-0-12-381479-1, 2011.

[2] Aggarwal, Charu C. "Data streams: models and algorithms", Vol. 31. Springer Science & Business Media, DOI 10.1007/978-0-387-47534-9, 2007.

[3] Agyekum, Kwame & Tchao, E. T. & Affum, Emmanuel, "Customer experience with mobile number portability implementation in ghana", International Journal of Emerging Technology and Advanced Engineering, 2013.

[4] Michalski, Ryszard S., Jaime G. Carbonell, and Tom M. Mitchell (EDS). "Machine learning: An artificial intelligence approach". Springer Science & Business Media, ISBN 978-3-662-12405-5, 2013.

[5] Alshurideh, Muhammad Turki. "Is customer retention beneficial for customers: A conceptual background?", Journal of Research in Marketing, [S.l.], v. 5, n. 3, p. 382-389, apr. 2016. ISSN 2292-9355.

[6] EunjuSuh Matt Alhaery, "Customer retention: reducing online casino player churn through the application of predictive modelling". UNLV Gaming Research & Review Journal, Volume 20 Issue 2, ISSN:1535-758.

[7] Ascarza E, Iyengar R, Schleicher M., "The perils of proactive churn prevention using plan recommendations: Evidence from a field experiment". Journal of Marketing Research. 2016 Feb;53(1):46-60.

[8] Cancho VG, Dey DK, Louzada F., "Unified multivariate survival model with a surviving fraction: an application to a Brazilian customer churn data". Journal of Applied Statistics. 2016 Feb 17;43(3):572-84.

[9] Bi W, Cai M, Liu M, Li G. "A big data clustering algorithm for mitigating the risk of customer churn". IEEE Transactions on Industrial Informatics. 2016 Jun;12(3):1270-81.

[10] Inderpreet Singh, Sukhpal Singh, "Framework for targeting high value customers and potential churn customers in telecom using Big Data Analytics", DOI: 10.5815/ijeme.2017.01.04.

[11] John Hadden, Ashutosh Tiwari, Rajkumar Roy, and Dymtr Ruta, "Churn Prediction using Complaints Data" World Academy of Science, Engineering and Technology 19 2006.

[12] Ahmed Qureshi, Saad & Saleem Rehman, Ammar & Qamar, Ali & Kamal, Aatif & Rehman, Ahsan. (2013). Telecommunication Subscribers' Churn Prediction Model Using Machine Learning. 8th International Conference on Digital Information Management, ICDIM 2013. 10.1109/ICDIM.2013.6693977.

[13] Manpreet Kaur, Dr. Prerna Mahajan, "Churn Prediction in Telecom Industry Using R", (IJETR), Volume-3, Issue-5, May 2015 ISSN: 2321-0869.

[14] Source of Data set: http://www.sgi.com/tech/mlc/db/

[15] Source of Data Set: http://www.dataminingconsultant.com/data/churn.txt

[16] Muhammad Akmal, "Factor Causing Customer Churn: A Qualitative Explanation of Customer Churns In Pakistan Telecom Industry", [MS Thesis], March 2017.

[17] Kaggle Data set: https://www.kaggle.com/arpitharavi/telecom-churnnotebook.

[18] WEKA: https://www.cs.waikato.ac.nz/~ml/weka/

[19] Geeks for Geeks: https://www.geeksforgeeks.org/building-a-machine-learning-model-using-j48-classifier/

[20] Github: https://oneapi-src.github.io/oneDAL/daal/algorithms/boosting/logitboost.html