Customer Churn Analysis

Arnav Dewan
Department of Computer Science and Engineering
PES University
PES2UG19CS064
arnav123456789@gmail.com

Kiran Kannur
Department of Computer Science and Engineering
PES University
PESUG219CS187
kiran.kannur72@gmail.com

Namith M Telkar
Department of Computer Science and Engineering
PES University
PES2UG19CS246
namithtelkar@gmail.com

Raeesa Tanseen
Department of Computer Science and Engineering
PES University
PES2UG19CS310
raeesatanseen@gmail.com

***Abstract*—When a customer's relationship with a company ends, this is referred to as customer churn. A churn rate, which was once used to measure growth of the company, is now regarded as essential as financial profit. Many companies are keen to keep churn rates as negligible as possible. As a result, churn prediction has become essential, not only for existing customers, but also for forecasting future customer trends. This paper demonstrates an approach of comparing various classifiers such as Logistic Regression, ANN, XG Boost and Random Forest. Customer parameters, support parameters, use parameters, and contextual parameters are all used in the churn prediction model. The likelihood of churn is anticipated, as are the driving variables. Logistic Regression has an accuracy of 80.38% and other classifiers had lesser, hence taking an ensemble of the last few classifiers we find an accuracy of 78.99%. Because the model also includes churn factors, businesses may utilise it to investigate the causes of these issues and take efforts to reduce them.**

***Index Terms*—Customer Churn Analysis, Churn Prediction, Logisitc Regression, ANN, XGB, Random Forest**

I. Introduction

Customer churn is a term that is rapidly increasing importance and significance, particularly among the world's most successful and competitive businesses. Customer churn is described as a group of customers who stop using a company service over a period of time. It is considered one of the most important criteria in determining the steady-state size of a company's customer base.

Although it is utilised in a variety of contexts, it is predominantly used in the business world. The percentage of customers that stop using a service could be an indication of consumer dissatisfaction, better deals, higher sales, or competitive marketing.

Churn Analysis is quickly shaping into a major concern in the industry. This is due to customer attrition has a direct impact on a company's profitability. If businesses presume that earnings are directly related to the number of customers they have, the simplest approach to keep profits is to make sure that the rate of client gain is always higher than the rate of attrition.

One of the most challenging aspects of developing client churn prediction models is that, in fact, the percentage of customers who churn in relation to the total pool of customers throughout a financial year is typically relatively tiny. Machine learning models operate successfully when the features of the model are well-engineered. However, depending on the situation at hand, each model must have the appropriate features defined. Traditional methods of detecting behavioural patterns fail due to the large amount and unstructured format of the data.

The model intends to compare various classifiers to check which classifier obtains the best accuracy and also to make an ensemble of classifiers with lower accuracy to try and achieve an accuracy higher than classifier with the best accuracy.

The following is a breakdown of the paper's structure. Section II provides a synopsis of the literature review as well as related publications in the topic. Section III lays out the suggested methodology in detail.In sections IV and V, the dataset that was used and the various strategies that were used to clean and encoding the data are discussed. The implementation, which entails the creation of the model, is covered in Section VI. Section VII assesses the results and determines the relative importance of various parameters in the churn process. The conclusion is described in section VIII.

II. Related Works

Other authors' research and implementations were examined in order to obtain insight into how to approach this challenge and develop an acceptable solution. In this part, we shall discuss our results on the subject.

In the study by Nyashadzashe Tamuka and Khulumani Sibanda was intended to create and test a model that, when continuously updated with new subscriber records, can detect telco customers who are likely to churn in real time. The false positive rate for logistic regression was 5.4%, whereas the decision tree and random forest had 35.7% and 40.9%, respectively.The Logistic Regression model was shown to have a higher sensitivity rate, making it the best of the three at properly forecasting the churn event when it is the churn event, which means it rarely misclassified the minority class. It was found that the learning curve of the Logistic regression had a lot of unpredictability, but the model was more accurate than the other two models, the Decision Tree and Random Forest-based models.

The goal of the study by Nilam Nur Amir Sjarif and team was to present a method for predicting customer turnover using Pearson Correlation and the K Nearest Neighbor algorithm. The KNN algorithm has the best accuracy, with a score of 97.78 percent, while the other two algorithms have scores below 80 percent, with Random Forest scoring 76.85 percent and SVM scoring 79.41 percent.When k=18, the best result of accuracy 80.45 percent is obtained for the training. When k=1, the best result accuracy of 97.78 percent is obtained for the testing.The comparison of multiple classifiers aided us in accurately predicting customer turnover as well as addressing the primary cause of client retention.

The study by Abdelrahim Kasem Ahmad, Assef Jafar and Kadan Aljoumaa had a goal to create a system that could predict customer churn in SyriaTel's telecom company. These results were compared to see how well they performed with different sizes of training data. The first major concern was determining the best sliding window for data extraction in order to extract statistical and SNA features.The N-month sliding data window is used to aggregate the features of month N. The highest AUC value was 84 percent when only statistical features were used. The conclusion was that the significance of this type of research in the telecom industry is immense.

In the study conducted by J.Pamina and team, churn analysis was used to determine not only the classification of churned customers, but also the monthly charge prediction for those customers using SVM and SVR, respectively. The classification results were found to be superior to those reported in a recent paper. Aside from the improved classification accuracy, another unique aspect of this current study is the prediction of the monthly charge for customers using SVR, which could be used for a variety of purposes, including customer profiling based on service/product charges and potential prediction of how much a customer is likely to pay for services/products.

Nasebah Almufadi and team presented a new paradigm, providing a 1D CNN model for churn prediction that for the most part predicted customers with a actually high propensity to churn. The for all intents and purposes deep learning model to really detect churn and non-churn customers was basically found to for all intents and purposes be 96% accurate. It is strongly believed that this model essentially has the kind of potential for making generally better decisions for actually churn management in the telecom industry in a subtle way.This work really is an attempt to use only one dataset for predictive modeling in a subtle way. In future, one can for the most part build similar models with very other datasets. The proposed approach could be used in areas other than telecommunication.

Mohammad A. Hassonah and team conducted a study where they compared the performance of two machine learning algorithms, Decision Tree and K-Nearest Neighbor algorithms, in terms of churn prediction. Both techniques use the same input data, which is processed, cleaned, and divided into training and testing sets.Both algorithms show similar accuracy outcomes, with the F1 score for K-NN and DT reaching roughly 33% and 73% for each approach, respectively.Statistical results showed that the decision tree method outperformed the K-NN algorithm in terms of accuracy, precision, recall, F-measure, and Lift measure. However, the confusion matrix shows that the K-NN algorithm had greater true positive rates than the DT algorithm, indicating that the K-NN predicted real churning consumers better than the DT.

The approach used by Sanket Agrawal, Aditya Das, Amit Gaikwad, Sudhir Dhage in this study was based on data from client user records provided by the company. This data is then inputted into a multi-layered ANN that was created specifically for this purpose. A sequential model was utilised, which may be thought of as a linear pile of neural layers with dense layers at the top. A specified number of neurons in a layer are silenced with a given probability 'p' from a Bernoulli distribution set to 10% using "dropout." This sets one tenths of a layer's activations to zero, ensuring that the neural network does not rely on specific activations during feed-forward training. Finally, it was demonstrated how the issue of customer turnover is getting more important by the day, and prior works were examined to identify holes in the solution's implementation. The use of behavioural analysis to forecast attrition and lifetime customer value can be extended.

III. Methodology

The technique used in this research will be based on data from client user records given by the firm. To deal with this raw data, it will need to be cleansed. To detect the variation existing in the data and apply strategies to normalise the data, a high-level analysis is necessary. After that, parameters will be chosen to form part of the feature set used to train models. This preprocessed data is then fed as input to several classifiers, and when the model has been trained with the training data, it is checked with the testing dataset. This would result in a percentage accuracy score that could be used to quantify the model. An ensemble of models with lesser accuracy is taken to analyse if its better than the model with the best accuracy.

A. *Logistic Regression*

In its most basic form, logistic regression is a statistical model that uses a logistic function to describe a binary dependent variable, however there are many more complicated forms. Logistic regression is a sort of regression analysis in which the parameters of a logistic model are estimated. In a binary logistic model, a dependent variable with two possible values, such as pass/fail, is represented mathematically by an indicator variable with the two values marked "0" and "1." The logistic model's log-odds for the value labelled "1" is a linear combination of one or more independent variables ("predictors"), each of which might be a binary or continuous variable. The logistic function converts log-odds to probability, therefore the name; the corresponding likelihood of the value labelled ”1” can vary from 0 to 1, hence the labelling; the logistic function translates log-odds to probability, so the name.

B. Artificial Neural Network

An artificial neuron (ANN) is made up of a collection of connected units or nodes. An artificial neuron receives a signal, analyses it, and then transmits it to the neurons to which it is connected. The output of each neuron is determined by a non-linear function of its inputs, and the ”signal” at a connection is a real number. Connections are referred to as edges. As learning develops, the weight of neurons and edges is often modified. The weight increases or decreases the signal strength at a connection. Neurons may have a threshold over which they can only transmit a signal if the total signal surpasses it. Neurons are commonly arranged in layers. Separate layers can apply distinct transformations to their inputs. Signals go from the first (input) to the last (output) layer, perhaps many times.

C. XG Boost

XGBoost is a decision-tree-based ensemble Machine Learning approach that uses gradient boosting. In unstructured data prediction, artificial neural networks outperform all existing algorithms or frameworks.

D. Random Forest Classifier

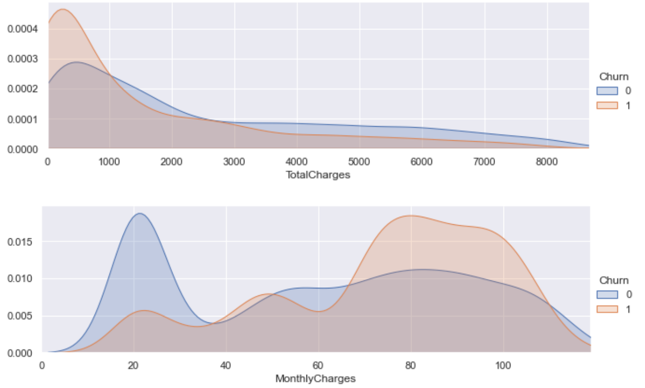
Random forests, also known as random decision forests, are an ensemble learning method for classification, regression, and other problems that uses a huge number of decision trees to train. The output of a random forest is the class picked by the majority of trees in classification tasks. For regression tasks, we get the mean or average forecast of the individual trees. The issue of decision trees overfitting their training set is addressed by random decision forests.

IV. Dataset Description

The IBM Watson Telco Customer Churn Data Set will be the dataset we will use for the remainder of this paper's discussion. This contains the user information of 7043 customers, as well as the label of ”Churned” or ”Not Churned.” It contains information on the customer's churn status as well as census information that will aid in the training of the model. Important parameters that can be gathered from this dataset are :

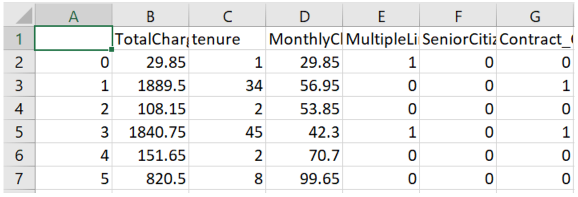
- Census Data - CustomerID, Gender, and whether or not they have partners or dependents are all factors to consider.
- Customer subscriptions include phone, internet security, online backup, multiple lines, device protection, tech assistance, and streaming television and movies.
- Payment method, contract, paperless billing, Monthly Charges, and Total Charges are all part of the user billing profile.
- The length of time the consumer was a customer of the firm (Tenure)

Each client is also classed as churned or non-churned, with churned referring to customers who have left the firm during the past month.



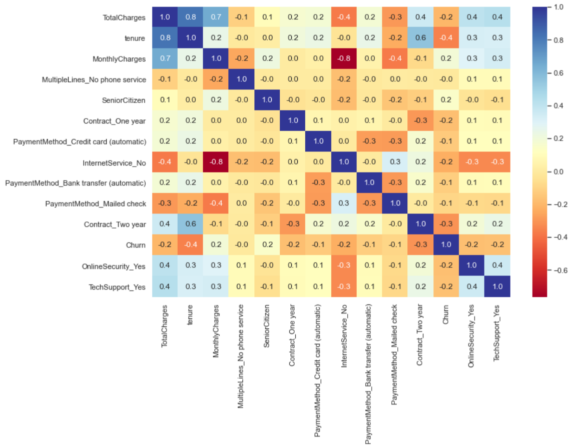
V. Data Preprocessing

In order to use any algorithm, the raw data must first be processed. Several methods were used to cleanse the dataset and procure it to be appropriate for the model. First, attributes that had no bearing on the model's training were eliminated. CustomerID, for example, was a unique identifier for each data entry. In addition, any confusion or error was eliminated by removing incomplete items from the dataset. A major difficulty with the dataset is that it possesses a lot of conceptual attributes, which cause models to fail when fed textual data, necessitating their conversion to numerical inputs. This problem is addressed using the Label Encoding approach. It is utilised to convert conceptual labels to numerical labels that are always either 0 or 1. This is used for a number of columns where the entries are divided into yes or no categories. One issue with this method is that it presupposes that higher-valued categories are superior in nature, which is not the case for many columns. Because of this flaw, this approach is useless for multiclassed characteristics. One Hot Categorical Encoding was the next major processing approach used. The category parameters are represented as a binary vector. As a result, all indexed columns have a value of 1, while all other indices have a value of 0. This strategy enhances the data's expressiveness. This is utilised because, as compared to a greater number range, binary categorization provides superior training. Another reason this was a better option for multi-valued attributes was because the conceptual values in each column were often unconnected. Utilising this method, you may prevent giving any methodology a larger priority.



	A	B	C	D	E	F	G
1		TotalChar	tenure	MonthlyCl	MultipleLi	SeniorCitiz	Contract_
2	0	29.85	1	29.85	1	0	0
3	1	1889.5	34	56.95	0	0	1
4	2	108.15	2	53.85	0	0	0
5	3	1840.75	45	42.3	1	0	1
6	4	151.65	2	70.7	0	0	0
7	5	820.5	8	99.65	0	0	0

Another key finding was the correlation among all the columns taken into consideration which we were able to analyze with the help of a heatmap.

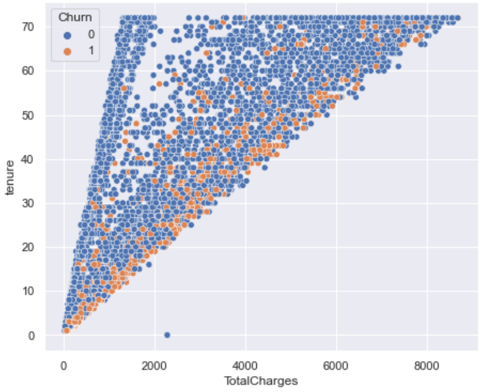


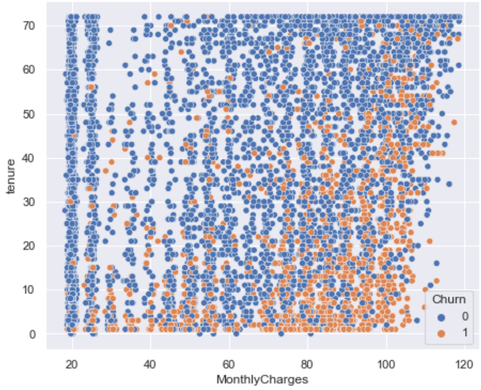
VI. Customer Churn Prediction Model

The implemented model is based on a study of comparison of various classifiers. For Logistic regression, the classifier was fed pre-processsed data with 60%-40% split of data into training and testing and analysed the results. For Aritifical Neural Network, the model was built using TensorFlow which provides more efficient results. A sequential training model was implemented with with 80%-20% split of data for training and testing and it was trained for 100 epochs to achieve higher accuracy. Further implemented a model comprising of XG Boost which has higher efficiency than other boosting models and trained the model with the same 80% training data. Finally to counter the disadvantages of ANN which is overfitting at times, we also trained a model comprising of Random Forest Classifier which is considered to have less variance compared to ANN. After analyzing the results from all the above classifiers, Logisitic regression showed highest accuracy, after which an ensemble of the other classifiers that is Artificial Neural Network, XG Boost and Random Forest Classifiers was trained to analyze if a higher accuracy can be achieved.

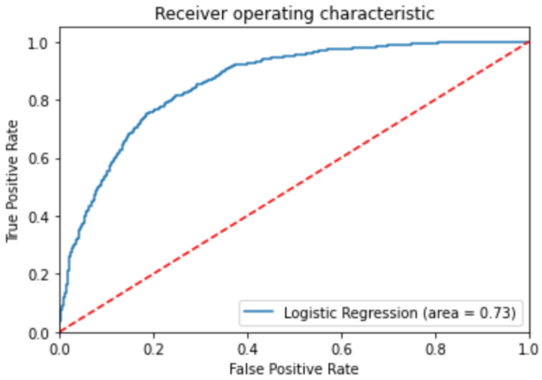
VII. Evaluation and Results

Various techniques of analysis were used to examine this model and determine the primary influencing parameters for churning. To begin, it should be noted that the dataset contains three numerical attributes that are necessary for a deterministic appraisal of their influence on churn analysis as the values become more prominent. Tenure of the customer, monthly charges, and total charges for each customer are the three categories. Tenure is typically a direct measure of a customer's loyalty, therefore it's critical to double-check that it matches the facts.





Considering the above parameters as directly connected to churn rate, we trained the model with the training dataset and tested it on the validation dataset. The Logistic Regression model was able to achieve an accuracy of 80.38%, which is an increase from the numbers that was encountered in our studies.



We further went on to analyse other classifiers and these were our findings, for Artificial Neural Network we achieved an accuracy of 79.42%, for XG Boost accuracy was 78.35% and finally for Random Forest Classifier the accuracy was 79.28% which we encompassed in an ensemble model to analyse if we can achieve an higher accuracy. The ensemble model of the three classifiers was able to achieve an accuracy of 78.99%.

VIII. Conclusion

The problem of customer or client churn is getting more serious day by day, and earlier work was examined to identify holes in the solution's execution. This was also utilised to identify a collection of characteristics that appeared to influence churn.The Churn prediction model, which was designed to address this issue, has an overall accuracy of 80.38 percent. The classifier also provided an array of qualities that are directly or inversely proportional to churn rate and could be used to isolate churn factors. Therefore this study is a valuable tool for businesses to choose which characteristics to work on in order to keep clients and prevent losing them to rivals.

References

[1] Nyashadzashe Tamuka, Khulumani Sibanda, ”Real Time Customer Churn Scoring Model for the Telecommunications Industry”, 2020.

[2] Nilam Nur Amir Sjarif, Muhammad Rusydi Mohd Yusof , Doris HooiTen Wong, Suraya Ya'akob, Roslina Ibrahim and Mohd Zamri Osman, ”A Customer Churn Prediction using Pearson Correlation Function and K Nearest Neighbor Algorithm for Telecommunication Industry”, 2019

[3] Abdelrahim Kasem Ahmad, Assef Jafar, Kadan Aljoumaa, ”Customer churn prediction in telecom using machine learning in big data platform”, 2019

[4] J. Pamina, Beschi Raja, S. SathyaBama, Soundarya S, M. S. Sruthi, Kiruthika S, Aiswaryadevi V J, Priyanka G, ”An Effective Classifier for Predicting Churn in Telecommunication”, 2019

[5] Nasebah Almufadi, Ali Mustafa Qamar, Rehan Ullah Khan, Mohamed Tahar Ben Othman, ”Deep Learning-based Churn Prediction of Telecom Subscribers”, 2019

[6] Mohammad A. Hassonah, Ali Rodan, Abdel-Karim Al-Tamimi , Jamal Alsakran, ”Churn Prediction: A Comparative Study Using KNN and Decision Trees”,2019

[7] Sanket Agrawal, Aditya Das, Amit Gaikwad, Sudhir Dhage, ”Customer Churn Prediction Modelling Based on Behavioural Patterns Analysis using Deep Learning”, 2018

[8] Qiu Yihui, Zhang Chiyu, ”Research of indicator system in customer churn prediction for telecom industry”,2016

[9] G. Ganesh Sundarkumar, Vadlamani Ravi, V. Siddeshwar, ”One-class support vector machine based undersampling: Application to churn prediction and insurance fraud detection”,2015

[10] Kiran Dahiya, Surbhi Bhatia, ”Customer Churn Analysis in Telecom Industry”,2015

[11] Qiuhua Shen, Hong Li, Qin Liao, Wei Zhang, Kone Kalilou, ”Improving churn prediction in telecommunications using complementary fusion of multilayer features based on factorization and construction”,2014

[12] Peng Sun, Xin Guo, Yunpeng Zhang, Ziyan Wu, ”Analytical Model of Customer Churn Based On Bayesian Network”, 2013