International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 04 | April - 2023 Impact Factor: 8.176 ISSN: 2582-3930

Churn Prediction Model Using Machine Learning

Prof. Aditi Malkar[1], Tanvi Visapurkar[2], Vedika Shetty[3], Vaishali Valvi[4], Sakshi Vaidya[5]
[1]Assistant Professor, *Department of Computer Engineering,*
[2]*BE Student, Department of Computer Engineering,*
[3]*BE Student, Department of Computer Engineering,*
[4]*BE Student, Department of Computer Engineering,*
[5]*BE Student, Department of Computer Engineering,*
[1,2,3,4,5]*MCT's Rajiv Gandhi Institute of Technology, Mumbai*

***Abstract*—Client churn is a major problem and one of the most important enterprises for large companies. Especially in the telecommunications industry, companies are trying to develop tools to predict implicit customer churn because of its direct impact on company profits. Thus, changing factors that increase client churn is important to take necessary conduct to reduce this churn. The main donation of our work is to develop a churn vaticination model which assists telecom drivers to prognosticate guests who are most likely subject to churn. The model developed in this work uses machine literacy ways on a big data platform and builds a new way of features engineering and selection. To measure the performance of the model, This work also linked churn factors that are essential in determining the root causes of churn. By knowing the significant churn factors from guests' data, CRM can ameliorate productivity, recommend applicable elevations to the group of likely churn guests grounded on analogous patterns, and exorbitantly ameliorate marketing juggernauts of the company.**

***Index Terms*—Churn Prediction, Machine Learning, ANN, Random Forest, Naïve Bayes**

I. INTRODUCTION

Consumers go through a complex decision-making process before subscribing to one of the many communication service options. The services handled by the Telecom merchandisers aren't largely discerned and number portability is commonplace. The mobile telephone assiduity churn is the analogous problem(2)(9)(12). client fidelity becomes an issue. Therefore, it is becoming increasingly important for operators to proactively identify factors that tend to churn and take preventative measures to retain guests. To calculate your probable monthly churn, start with the number of druggies who churn that month. Also, divide by the total number of stoner days that month to get the number of churns per stoner day. Also, multiply by the number of days in the month to get the monthly churn rate. It's set up that data mining ways are more effective in prognosticating consumer churn from the exploration conducted over the once many times(17). Creating an effective churn vaticination model is an essential exertion taking a lot of work right from determining applicable predictor variables(features) from the large volume of available client data to choosing an effective prophetic data mining fashion suitable for the point set.

To effectively predict churn, use a synthetic set of key metrics defined by your team to let you know when your customers are likely to churn so your business can take action is needed. The goal of churn prediction is to be able to answer questions such as "[X] customers will he leave us in X months?" Or you can understand a larger churn trend with "Are [X] customers renewing their subscriptions?"

A study conducted by McKinsey found that technology and SaaS companies with the highest performance and revenue growth also had higher retention and lower net revenue churn. The ability to predict churn before it happens allows organizations to take proactive measures to prevent existing customers from churning. Our Customer Success team seeks support from these high-risk customers and assesses their potentially unmet needs. Reminder emails that run more targeted re-engagement campaigns, such as Creating more focused customer education content, and Reassessing retention initiatives across large enterprises, such as pricing. Predicting churn not only identifies at-risk customers but also the pain points that lead to churn, helping to improve overall customer retention and satisfaction. Predicting churn can help prevent churn. Conversely, avoiding churn is a huge revenue stream for businesses. Predicting churn can help prevent churn.

Churn prediction is knowing which customers are most likely to leave or unsubscribe from your service. This is an important prediction for many companies. This is because it often costs more to acquire new customers than to retain existing ones. Once you've identified customers at risk of churn, you need to know exactly what marketing actions you should take with each customer to maximize their likelihood of retention. Different customers have different behaviors and preferences, and different reasons for canceling their subscriptions. Therefore, it is important to keep them on your customer list. You need to know which marketing efforts are most effective for individual customers and when they are most effective.

Existing research indicates that the main purpose is to use large amounts of communication data to identify valuable churn customers. However, the existing one has some limitations. A powerful obstacle model that gets in the way of this problem

real environment. It will be a lot of data occurs in the telecommunications department and data is missing values that lead to poor predictive model results. To address these issues, there are data pre-processing methods. It has been adjusted to remove noise from the data. Model to better classify your data and improve performance. However, feature selection is used in the literature. Many information-rich features are ignored in model development. Mainly in various fields Statistical methods are used that lead to bad results in predictive models. The model is used in existing studies This is not possible, although verified with benchmark data sets represents a true representation of the data and has no value. For decision-makers, to work around this limitation, some algorithm applies to the same data set and best classification, Selected for storage, intelligent mechanism, helping develop predictive models for automatic churn Predict and saving. Another big existing problem in the model is feature selection. any customer or group Percentage of customers with different reasons for churn.

II. LITERATURE REVIEW

According to [4], experimental results show that two ensemble learning methods**,** the Adaboost classifier and XGBoost classifier**,** have an AUC value of 84% for the churn prediction problem compared to other models, and a It indicates that it has maximum precision. The globalization and advancement of the telecom industry has led to an exponential increase in the number of telecom operators in the market and increased competition [9]. In this age of competition, **it** is imperative to maximize profits on a regular basis, and various strategies have been proposed to achieve this, such as acquiring new customers, upselling existing customers, and extending the retention period of existing customers. **.** Among all strategies, retaining existing customers is the most cost-effective compared to others. To adopt the third strategy, companies must reduce potential customer churn. H. Transfer of Customer from One Service Provider to Another Service Provider**.** The primary reason for termination is dissatisfaction with consumer service and support systems**.** The key to unlocking a solution to this problem lies in predicting which customers are at risk of churn**.** According to [3], this research helps to analyze customer behavior and build models to predict which customers want to churn. With the significant increase in the number of customers and companies using the telecom sector [1], the level of competition between companies has increased [2, 3]. All companies try to survive in this race through a number of strategies [4]. The main strategies are: 1**)** Resale of existing customers, 2) Extension of the customer retention period**,** 3) Acquisition of new customers. Businesses are concerned because they see it as a profit to retain or retain customers and it is cheaper to retain new customers than to attract them. All businesses try to retain customers by making them more loyal. Customers are great ambassadors in the marketplace as they can be used by companies to promote their products and services [5]. According to [2], a customer churn model for data analysis is provided in this study and validated by standard metrics. The obtained results show that using machine learning techniques improves the performance of the proposed churn model**.** In today's world, vast amounts of data are being generated by telecommunications companies at breakneck speeds**.** There are many telecom providers competing in the market to increase their customer share. Customers have multiple choices in the form of better and cheaper services. The ultimate goal of telecommunications companies is to maximize profits and survive in a highly competitive market [1]. Customer churn occurs when the majority of customers are dissatisfied with a carrier's service. This leads to service migration for customers who start switching to other service providers. There are many reasons for churn**.** Unlike postpaid customers, prepaid customers are not tied to a service provider and can be migrated at any time. Churning also impacts a company's overall reputation, leading to brand loss. Loyal customers who bring high profits to the company are rarely affected by competitors. These customers maximize the company's profits by recommending the company to friends, family and colleagues. Carriers consider changing policies when customer numbers fall below a certain level. This can significantly reduce your earnings**.** Churn prediction is important in the telecommunications industry as operators need to retain valuable customers and improve customer relationship management (CRM) management [5], [6]. The most difficult task for CRMs is retaining existing customers.

III. PROPOSED METHODOLOGY

A. Dataset Description

Most mobile operators have a history of customers who churn and those who continue to use their service. Using this historical information, you can create a machine learning (ML) model of carrier churn using a process called training. After training and testing the model, you can feed it random customer profile information to predict whether that customer will churn or stay.

TABLE I - DESCRIPTION OF THE DATASET

Sr. No.	Attribute Names	Attribute Description
01	State	The US state in which the customer resides, indicated by a two-letter abbreviation
02	Account length	the number of days that this account has been active
03	Area code	the three-digit area code of the corresponding customer's phone number
04	International plan	whether the customer has an international calling plan: yes/no
05	Voice mail plan	whether the customer has a voice mail feature: yes/no
06	Number v-mail messages	presumably the average number of voice mail messages per month
07	Total day minutes	the total number of calling minutes used during the day
08	Total day calls	the total number of calls placed during the day
09	Total day charge	the billed cost of daytime calls
10	Total eve minutes	the total number of calling minutes used during the evening
11	Total eve calls	the total number of calls placed during the evening
12	Total eve charge	the billed cost of evening time calls
13	Total night minutes	the total number of calling minutes used during the night
14	Total night calls	the total number of calls placed during the night
15	Total night charge	the billed cost of night time calls
16	Total international minutes	the total number of international minutes
17	Total international calls	the total number of international calls
18	Total international charge	the billed cost for international calls
19	Customer service calls	the number of calls placed to Customer Service
20	Churn	whether the customer left the service: true/false

B. Data Pre-processing

In order to increase the system's accuracy, the suggested study intends to create and build a method for predicting customer churn utilizing NLP and machine learning techniques. Next, we identify customer behavior patterns that change during forecasting [4]. It also evaluates the factors that reduce the accuracy of the churn prediction the most and finally evaluates and calculates the churn rate for the month and day. This helps improve the quality of service of the system. In this study, we proposed churn prediction from large-scale data. The system first processes a synthetic communication dataset containing imbalanced metadata. Apply data preprocessing, data normalization, feature extraction, or feature selection [17]. During this run, we used several optimization strategies to eliminate redundant functions that could generate high error rates during the run. A proposed system design for training and testing. After the system is complete, both phases describe the classification accuracy of the entire dataset.

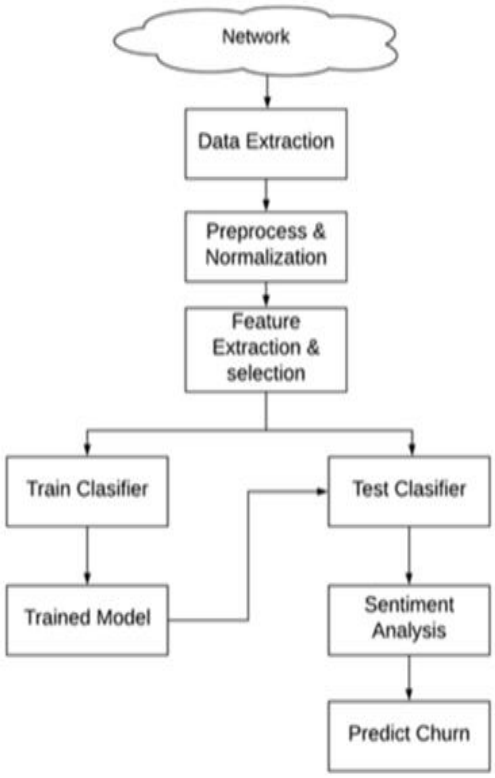

Fig.1 BLOCK DIAGRAM OF WORKFLOW

1) Data Extraction

The process of collecting or extracting various types of data from many sources is called data extraction. Consolidating, processing, and fine-tuning data allows it to be kept in a centralized area where it may be modified, and data extraction makes this feasible.

2) Pre-Processing:

Data Pre-processing – Before working on data it's necessary to pre-process it. Data may contain missing values that lead to poor results to avoid this, pre-processing is necessary. Filtering and noise removal: In the filtering process, unwanted and unnecessary features are removed and only keep useful features. In noise removal, null values, space, and missing characters are removed.

3) Feature Extraction:

Feature extractions remove the unwanted data from the data set and keep only accurate and complete data that should be processed.

C. Classification

1) Random Forest (RF)

Random Forest is a supervised machine learning algorithm commonly used in classification and regression problems. Build decision trees with different samples and for regression take majority votes for classification and averaging. One of the most important features of the random forest algorithm is its ability to handle datasets containing continuous variables, as in regression, and categorical variables, as in classification. Improves performance for classification and regression tasks.

2) Naïve Bayes (NB)

The naive Bayes algorithm is a supervised learning algorithm based on Bayes' theorem and is used to solve classification problems. It is primarily used in text classification with high-dimensional training datasets. Naive Bayes Classifier is one of the simplest and most effective classification algorithms that help you build fast machine learning models that can make fast predictions. It's a probabilistic classifier, meaning it makes predictions based on object probabilities.

3) Artificial Neural Network (ANN)

Artificial Neural Networks (ANNs) use learning algorithms that adapt (in a sense, learn) autonomously when given new inputs. This makes it a very effective tool for nonlinear statistical data modeling. Deep learning ANNs play an important role in machine learning (ML) and support a wider range of artificial intelligence (AI) technologies.

D. Evaluation Measures

To evaluate the effectiveness of our proposed model, we utilized K-fold cross-validation techniques to randomly split the dataset into k subsets to construct the training and test sets. During each iteration, k-1 subsets were used to train the model, while the remaining subset was used for testing. By repeating this process k times, we obtained the model's performance by averaging the test results of the independent k subsets. In this study, we used 10-fold cross-validation to reduce bias and variance. Various statistical measures, such as F1-score, precision and recall were considered to evaluate the model's performance. Typically, the confusion matrix summarizes the overall performance of any prediction model, and accuracy, F1-score, recall, and precision can be derived from it. From the confusion matrix accuracy, F 1 score, recall, and precision are intended as follows.

TABLE II. CONFUSION MATRIX

Actual	**Predicted**	
	Positive	*Negative*
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

$$\text{Accuracy} = \frac{(\text{TN} + \text{TP})}{(TN + TP + FN + FP)}$$

$$\text{Precision} = \frac{(\text{TP})}{(FP + TP)}$$

$$\text{Recall} = \frac{(\text{TP})}{(FN + TP)}$$

$$\text{F I - Score} = \frac{2\,(\text{TP})}{(\text{FN} + \text{FP} + 2\text{TP})}$$

TABLE III. CLASSIFICATION PERFORMANCE USING ALL FEATURES

Evaluation Measure	Random Forest Algorithm	Naïve Bayes Algorithm	ANN Algorithm
Accuracy	93.35	85.59	84.75
Precision	0.92	0.87	0.86
Recall	0.99	0.96	0.95
F1-score	0.96	0.91	0.91

IV. RESULTS AND DISCUSSION

Experiments were done in Eclipse. A 5-fold cross-validation technique was used to evaluate model performance. 70% of the dataset was used for training and the remaining 30% for testing the accuracy of the model. The classification phase was split into two steps and all features were used for classification.

The results shown below show that the Random Forest Classifier, Naive Bayes Classifier, and ANN algorithms achieved accuracies of 93.35%, 85.59%, and 84.75%, respectively.

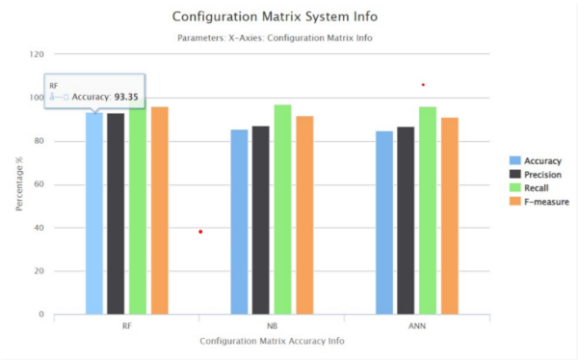

Fig.2 Performance Analysis Using All Features

V. CONCLUSION

Based on the experimental results, we can conclude that the Random Forest algorithm performs other algorithms in terms of accuracy. Machine learning and in-depth analytics to predict customer churn can help the company increase retention, save on retention costs, and even protect future revenue from churn. By introducing a customer churn model into your business, you can deter many consumers from churning and make more money as a result.

REFERENCES

[1] Abdelrahim Kasem Ahmad, Assef Jafar and Kadam Aljoumaa "Customer churn prediction in telecom using machine learning in big data platform" Springer (2019)

[2] IRFAN ULLAH, BASIT RAZA, AHMAD KAMRAN MALIK, SAIF UL ISLAM, SUNG WON KIM "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector" IEEE Access (2019)

[3] Essam Abou el kassem, Alaa Mostafa Abdelrehman, Shereen ali Hussein, Fahad kamal alsheref "Customer churn prediction model and identifying features to increase customer retention based on user generated content" IJACSA (2020)

[4] Praveen Lalwani, Manas Kumar Mishra, Jasroop Singh Chadha, Pratyush Sethi "Customer churn prediction system: a machine learning approach" Springer (2021)