2022 5th International Conference on Advances in Science and Technology (ICAST)

Feedback Based Telecom Churn Prediction Using Machine Learning

1st Dr. Madhura Phadke
Department of Computer Engineering
K J Somaiya Institute of Engineering and Information Technology
Mumbai, India
madhura.phadke@somaiya.edu

2nd Anirudh Bhattacharya
Department of Computer Engineering
K J Somaiya Institute of Engineering and Information Technology
Mumbai, India
anirudh.b@somaiya.edu

3rd Mink Shethia
Department of Computer Engineering
K J Somaiya Institute of Engineering and Information Technology
Mumbai, India
mink.shethia@somaiya.edu

4th Saumya Shah
Department of Computer Engineering
K J Somaiya Institute of Engineering and Information Technology
Mumbai, India
saumya.ss@somaiya.edu

***Abstract*—In an industry with stiff competition among individual organizations, Churn is an important factor to be considered by the company itself as well as by prospective customers. In the telecommunication industry, churn can be affected by multiple factors: a customer's preference, location, job, and so on. Thus, customer churn in the telecom industry is a widely studied subject. However, churn based on previous rates alone is not enough to predict future churn and there must be additional factors considered. We propose a method to overcome this inadequacy. By using the predictions generated by past data, a rough estimate of the churn rate for a certain time period can be generated, such as a quarter or a year. A Machine Learning algorithm can be used for the same to get the value of the prediction. This value can be further tweaked by incorporating customer feedback which can affect the churn rate. Thus, the value generated by the predictions and the feedback will be more accurate.**

Keywords—Customer churn, Machine Learning, Natural Language Processing, Web Scraping

I. Introduction

With the popularity of 5G increasing day by day and all of the major telecommunication companies scrambling to get their infrastructure altered for the widespread use of 5G, these companies must make sure that their ventures continue to be well funded for now as well as in the future. Thus, a major factor that these companies take into consideration is their rate of churn, which is the rate at which existing customers leave a company. If a company has a low churn rate, it could mean that its customers are unsatisfied with it's service. Consider the Indian telecommunication sector's giant: Reliance Jio. The first quarter of 2022 was particularly fruitful for them as their churn rate fell to 2% from 3.7% [18].

Thus, to ascertain the success or even the survival of a company, companies must aim to predict the churn rate of a certain time period so that they can accordingly alter policies and improve their service and increase their revenue. These predictions can be made by considering the factors which led to customers leaving the service in the previous time period and by analyzing the feedback given by previous customers.

When a customer gives negative feedback on a product or service, it is an indication that they may choose to leave said service in the near future. On review sites, it's noted that for Reliance Jio, there are a certain number of positive reviews and a certain number of negative reviews from customers of Reliance Jio. If a customer has posted a negative review between April and June 2022, it can be summarized that said customer intends to leave the company's services before the quarter ends. Thus, it can be concluded that the feedback given by customers is an important consideration when one tries to analyze the churn rates of any telecom company.

Thus, to build the system which will take into account previous factors affecting churn as well as customer feedback, an adequate dataset is needed, where attributes such as Age, Gender, Package, Services Used and other such details are listed, a machine learning model can be fit to predict the churn for the next time period solely based on these attributes. Now, if a web scraper is added to get reviews from retail sites and search engines, and a natural language processor is used to find the emotion behind the reviews, the positivity and negativity of the reviews can be ascertained and by the number of positive and negative reviews, the output from the machine learning model can be altered to give a more accurate prediction of the churn rate in the telecommunication industry.

II. Related Work

Methods such as Machine Learning algorithms, Deep learning algorithms, Probabilistic Graphical Models, and Ensemble models were used to predict customer churn in the Telecommunication industry.

The machine learning techniques which were used by other authors in [4], [5], [6], [13] were Random Forest (RF), Logistic Regression (LR), Hoeffding algorithm and Support Vector Machine (SVM) among others. These algorithms performed relatively well with the accuracy of the models ranging from 84 to 95. In [4], the authors concluded that the Random Forest and J48 algorithms performed the best when compared to the other algorithms when tested on the churn-bigml dataset. In [5], SVM and RF were used to fit a model on a dataset consisting of 3333 users, in a 70-30 training and

978-1-6654-9263-8/22/$31.00 ©2022 IEEE

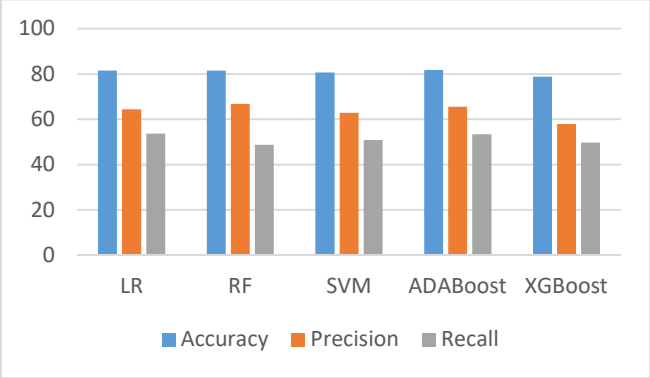
testing split respectively. Among SVM and RF individually, the RF algorithm performed better. The Hoeffding and Logistic Regression algorithms were compared in [6]. The LR algorithm was found to be 1.1% more precise than Hoeffding algorithm and the Recall of LR was greater by 2.04%. When the CART algorithm was fitted on a dataset given by Indian Telecommunication Service Providers, the model was 94.3% accurate. Additionally, when J48 was used, the model was 82.07% accurate. In [13], the Feature Selection method outperformed all of the other machine learning algorithms used.

The Deep Learning techniques which were used by other authors proved to be superior to the Machine Learning method by 5.77%. The Deep Learning algorithms also produced much higher metrics, as observed in papers [2], [8], [9], and [13]. The PPFCM-ANN, PBCPP, and CNN methods were used in [2] with the accuracy of the algorithms lying between 96 and 99 which are much higher than the values observed when the datasets were fitted with machine learning algorithms. When CPNN and fuzzyARTMAP were fit on a dataset in [9], the accuracies were 89.89% and 91.67% respectively.

Probabilistic Graphical Models were used to fit models on datasets in [3], [12], and [13], which yielded very good accuracies. In [3], the authors created a Bayesian Network which was found to be 69.8% accurate. According to certain factors, the authors categorized the motivation for customers to leave the service. Pronaya Bhattacharya *et al* used Naive-Bayes classification in their model [12], based on a dataset that had 3333 instances. The model was 97.89% accurate. Naive Bayes was also analyzed in [13] and it yielded an accuracy of 88.46% with feature selection and 86.80% without.

The averaging ensemble model, which was a combination of a few different algorithms, developed in [5] outperformed all of the individual models by at least 1% accuracy and the selective ensemble model achieved an accuracy of 93.9%, the highest observed in the study.

In order to find the metrics of more algorithms, IBM's WA_Fn-UseC_-Telco-Customer-Churn, as used in [1], and the churn-bigml-80 dataset were used, as in [4]. This analysis helped in knowing the effectiveness of different algorithms on the datasets used in previous papers. The first dataset contained information such as customer age, gender, the existence of partners and dependents, and finally their usage of the services provided by the telecommunication company. The dataset had 7043 rows and 21 columns. The data was then split into 75% data available for training and 25% for testing, after which the models were created. The Accuracy, Precision, and Recall metrics were used. The results were as such: The Logistic Regression algorithm had an accuracy of 81.4562%, a precision of 64.4257%, and a recall of 53.6130%. The Random Forest algorithm had an accuracy of 81.5699%, a precision of 66.7731%, and a recall of 48.7179%. The Support Vector Machine algorithm had an accuracy of 80.6598%, a precision of 62.8242%, and a recall value of 50.8158%. The ADABoost algorithm had an accuracy of 81.7974%, a precision of 65.6160%, and a recall value of 53.3799%. The XGBoost algorithm had an accuracy of 78.8964%, a precision of 57.8804%, and a recall value of 49.6503%. The results are shown in graphical form in Fig. 1:

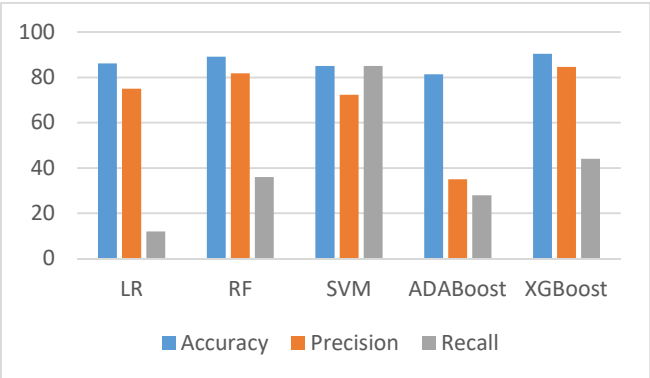


LR: Logistic Regression
RF: Random Forest
SVM: Support Vector Machine

Fig. 1: Analysis of algorithms on IBM's dataset.

On observation of Fig 1, the conclusion is that for this dataset the Adaptive Boosting classifier seems to be the best algorithm for this dataset.

The churn-bigml-80 dataset was then taken into consideration which contained information such as the customer's location, their contract details with the company, and also their activity. The data had 667 rows and 20 columns. The data was pre-processed and models were fit. In this case, the Logistic Regression algorithm was 86.2275% accurate, 75% precise, and had a recall value of 12%. The Random Forest algorithm was 89.2215% accurate, 81.8181% precise, and had a recall value of 36. The Support Vector Machine algorithm had an accuracy of 85.0299%, it was 72.3009% precise, and had a recall value of 85.0299. The ADABoost classifier was 81.4371% accurate, 35% precise, and had a recall value of 28% and the XGBoost algorithm was 90.4191% accurate, 84.6153% precise, and had a recall value of 28%. The results are shown in graphical form in Fig. 2.



LR: Logistic Regression
RF: Random Forest
SVM: Support Vector Machine

Fig. 2: Analysis of algorithms on churn-bigml

From Fig. 2, it was observed that the Support Vector Machine classifier seems to be the best algorithm for this dataset.

The workflow for the prediction is currently as such:

978-1-6654-9263-8/22/$31.00 ©2022 IEEE

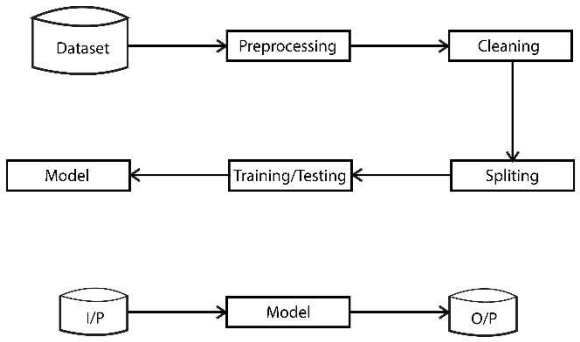


Fig. 3: Original workflow of the prediction

From the entirety of the research taken into consideration, it could be concluded that. that Probabilistic Graphical Models and Deep Learning Models are the best algorithms to predict the rate of customer churn, it's observed that when an algorithm like CNN is used on different datasets, there may be varying accuracy. Thus, if the dataset has problems, that too needs to be taken into account.

In order to find out the problems in datasets, in [16], the authors attempted to find the readiness of a dataset for machine learning. When the data was divided into "interconnects", they found that the first interconnect itself had 71.2653% bad data. Upon further analysis, there were more quality issues detected which would be harder to find for a non-specialist. Hence, while data may be pre-processed and cleaned, the dataset may still have underlying issues. Thus, with this in mind, the quality of the dataset needs to be considered before considering that an algorithm is the best for it.

An attempt to find the influences on customer churn was made in [7]. A social graph was built to discover a person's relationships which would alter in some way when a person churned. The authors assumed that restoring an influencer's pleasure will remove their churn influence on the social network. The study highlighted that there is a weakness in the current customer churn methods.

Many factors could affect the rate of customer churn in a certain city. If a company ceases its operations in a certain city, the churn rate for that quarter or the time period in consideration will be exceptionally high. There could also be a stoppage of certain services provided by the company which could displease some customers and would also result in churn. Additionally, a representative of the company could be involved in a controversial scenario which could also result in churn. A method to take into consideration must be developed in order to get the best prediction possible.

Thus, it's been determined that in order to get the best prediction possible, an additional factor needs to be introduced, which will help improve the accuracy of the prediction and yield better metrics overall.

III. Proposal

The proposed system is: the Adaptive Boosting Classifier using a decision tree classifier as the weak learner can be used to fit a model on IBM's Telco Customer Churn (WA_Fn-UseC_-Telco-Customer-Churn) dataset.

This dataset was chosen as it is quite popular with users; it has been used in over 900 notebooks that have been published on Kaggle, and it was also used in [1]. Additionally, when compared to the other supervised learning models fit on this dataset, the Adaptive Boosting Classifier performed the best with an accuracy of over 81% and that's the reason it was chosen for the machine learning model.

The dataset contains 21 features, out of which 19 are used for fitting the model. The features are crucial for fitting the model, and thus, there is no redundancy in the dataset, eliminating the need for dimensionality reduction.

As seen from [16], the dataset itself may have some issues, making the workflow shown in Fig. 3 problematic. Thus, to improve the prediction, another factor needs to be considered while considering the prediction of customer churn [7]. This factor could be the reviews left by customers about a particular telecommunication service provider.

The reviews left by customers are important for the prediction, as seen in [17] Thus if it's considered that during a quarter, a customer cancels their contract with a particular telecom company, they are likely to leave a review on sites such as Google and Yelp. A negative review will be an indicator of problems at the company, and would increase the churn rate by a certain percentage. On the other hand, if the review is positive, it shows that for the time period in question, the customers are satisfied and there will not be any increase in churn rate.

The total rate of Churn will be affected by the reviews taken into consideration. The calculation of customer churn will require the result from the machine learning model as well as customers' reviews left on the internet, which will not be considered while fitting the model as the customers' reviews are not mapped to the comments they leave on the internet. For the implementation of this system, a machine learning model, a web scraper, a natural language processor, and a database will be required. There are technologies available that can enable one to do web scraping. David Mathew Thomas *et al* used the XPath technique to develop a project called Scrapy [11]. The Beautiful Soup library can also get data from HTML files and save it to a database. Thus, the web scraper can get customer reviews from a website and they can be used for further analysis, by the natural language processor. In [14], the authors used BERT, RoBERTa, and XLnet models in order to find the emotion of a sentence. NLP technology can also be applied to find the emotion of a review, be it happiness or sadness.

The methodology of the proposed system is explained below:

The proposed system will use IBM's Telco Customer Churn dataset to fit a model using the Adaptive Boosting algorithm. The prediction from the model using new independent variables can be stored in MP in Table I. Next, the web scraper will take reviews from sites such as Google and Yelp and save them in Table II. The site, review, and star-value (if exist) can be saved in their respective columns. The Review column can be analyzed using the Natural Language Processor, the output of which will be saved in Alt. Alt will be the value by which, the predicted churn rate will increase or decrease. This column, Alt, can be mapped to Table I of the database, where it will affect the value of MP.

978-1-6654-9263-8/22/$31.00 ©2022 IEEE

For example: If the predicted rate of churn in Table I (MP) is 2.2% and Alt is 0.6%, the final prediction (FP) will be MP + Alt = FP = 2.8%. This will give the final prediction of customer churn for the telecommunication company in question. In Tables, I and II, the proposed schema for the database is given.

TABLE I. TABLE 1 OF DATABASE

	Database Columns		
No.	**MP**	**Alt**	**FP**
1	2.2	0.9	3.1
2	2.1	0.5	2.7
…			
n	2.5	0.0	2.5

MP: The prediction made by the machine learning model.
Alt: The percentage by which the prediction by the machine learning model should be altered.
FP: The final prediction, inclusive of the prediction made by the machine learning model and the customer's feedback.
Note: Values in the table are only for reference.

TABLE II. TABLE 2 OF DATABASE

	Database Columns			
No.	**Site**	**StarValue**	**Review**	**Alt**
1	TrustPilot	1	Bad	0.9
2	G2	3	Fair	0.5
…				
n	MouthShut	5	Good	0.0

Site: The site from which the review has been scraped
StarValue: The rating given by a customer (out of 5)
Review: Plaintext of the review given by the customer.
Alt: The percentage by which the prediction by the machine learning model should be altered.

Note: Values in the table are only for reference.

The proposed predictor system will work as such:

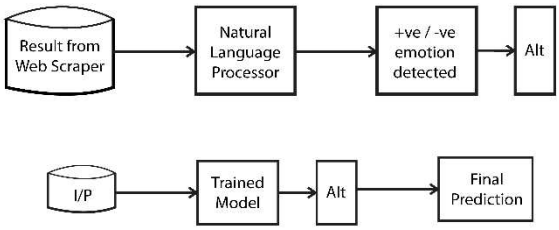


Fig. 4: Workflow of proposed system

Fig 4 shows the workflow of the prediction system. It shows that the data gathered from the web scraping application will be segregated into positive or negative emotion values based on the Natural Language Processor. This gives the value (Alt) by which the model prediction is affected. This, in turn, gives the Final Prediction.

Thus, with this prediction system in place, a more accurate prediction will be generated which will help better predict the churn rate of a telecom company.

IV. CONCLUSION AND FUTURE WORK

The churn rate of a telecommunication company is the rate at which existing customers leave the service, in a particular time period. The churn rate for a telecom company is important as a company needs to consider how they will cope with the loss of revenue and how much will their revenue be impacted by such leaving customers. Analysis of customer churn using various machine learning algorithms has been done in the past but for the most accurate prediction, customer feedback needs to be taken into account as well, because when customers are dissatisfied with a service, they tend to review the companies as such on popular sites. Thus, with a system consisting of a machine learning model, a web scraper, and a natural language processor to analyze the emotion in a review, the churn rate of a telecom company can be accurately predicted. With the best algorithm in use and the most reliable, unbiased sites being used for reviews, the accuracy of the prediction will increase.

In the future, we will be able to build a fully functional web-based application that can accurately predict the rate of customer churn for a given telecommunication company.

REFERENCES

1. Pan Tang (2020), Telecom Customer Churn Prediction Model Combining K-means and XGBoost Algorithm, *International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*
2. Priya Gopal, Dr. Nazri Bin MohdNawi (2021), A Survey on Customer Churn Prediction using Machine Learning and data mining Techniques in E-commerce, *IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*
3. Peng Sun, Xin Guo, Yunpeng Zhang, Ziyan Wu (2013), Analytical Model of Customer Churn Based On Bayesian Network, *2013 Ninth International Conference on Computational Intelligence and Security*
4. Irfan Ullah, Basit Raza1, Ahmad Kamran Malik, Muhammad Imran, Saif ul Islam, Sung Won Kim (2019), A Churn Prediction Model using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector *IEEEAccess*
5. Ahmad Hammoudeh, Malak Fraihat, Mahmoud Almomani (2019), Selective Ensemble Model for Telecom Churn Prediction, *IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*
6. Mr. Anurag Bhatnagar, Dr. Sumit Srivastava (2020), Performance Analysis of Hoeffding and Logisitic Algorithm for Churn Prediction in Telecom Sector, *International Conference on Computation, Automation and Knowledge Management (ICCAKM) Amity University*
7. Sara Motahari, Taeho Jung, Hui Zang, Krishna Janakiraman, Xiang-Yang Li, Kevin Soo Hoo (2014), Predicting the Influencers on Wireless Subscriber Churn, *IEEE WCNC'14 Track 4 (Services, Applications, and Business)*
8. Xin Hu, Yanfei Yang, Lanhua Chen, Siru Zhu (2020), Research on a Customer Churn Combination Prediction Model Based on Decision Tree and Neural Network, *IEEE 5th International Conference on Cloud Computing and Big Data Analytics*
9. Ramakanta Mohanty, Jhansi Rani K (2015), Application of Computational Intelligence to predict churn and non-churn of customers in Indian Telecommunication, *International Conference on Computational Intelligence and Communication Networks*
10. Ankit Verma, Chavi Kapoor, Abhishek Sharma, Biswajit Mishra (2021), Web Application Implementation with Machine Learning, *2nd International Conference on Intelligent Engineering and Management (ICIEM)*
11. David Mathew Thomas, Sandeep Mathur (2019), Data Analysis by Web Scraping using Python, *Third International Conference on Electronics Communication and Aerospace Technology [ICECA 2019]*
12. Pronaya Bhattacharya, Akhilesh Ladha, Ashwani Kumar, Ashwin Verma, Umesh Bodkhe, (2020) BaYcP: A novel Bayesian customer Churn prediction scheme for Telecom sector, *2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)*
13. Yakub K. Saheed, Moshood A. Hambali (2021), Customer Churn Prediction in Telecom Sector with Machine Learning and Information Gain Filter Feature Selection Algorithms, *2021 International Conference on Data Analytics for Business and Industry (ICDABI)*
14. F. A. Acheampong; H. Nunoo-Mensah; Wenyu Chen (2021), Recognizing Emotions from Texts Using an Ensemble of Transformer-Based Language Models, *International Conference on Wavelet Active Media Technology and Information Processing (ICWAMTIP)*
15. Seo-Hui Park, Byung-Chull Bae, Yun-Gyung Cheong (2020), Emotion Recognition from Text Stories Using an Emotion Embedding Model,

978-1-6654-9263-8/22/$31.00 ©2022 IEEE

International Conference on Big Data and Smart Computing (BIGCOMP)

16. Philip Hart; Lijun He, Tianyi Wang, Vijay S. Kumar, Kareem Aggour, Arun Subramanian, Weizhong Yan (2022), Application of Big Data Analytics and Machine Learning to Large-Scale Synchrophasor Datasets: Evaluation of Dataset 'Machine Learning-Readiness', *IEEE Open Access Journal of Power and Energy*
17. Jo Mackiewicz, Dave Yeats, Thomas Thornton (2016), The Impact of Review Environment on Review Credibility, *IEEE Transactions on Professional Communication*
18. Tanay Singh Thakur (2022), Reliance Jio Subscriber Churn Dipping, ARPU Rising a Good Thing, *TelecomTalk*

978-1-6654-9263-8/22/$31.00 ©2022 IEEE