International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 7
DOI: https://doi.org/10.17762/ijritcc.v11i7.7828
***Article Received**: 18 April 2023 **Revised**: 19 June 2023 **Accepted**: 30 June 2023*

Churn Identification and Prediction from a Large-Scale Telecommunication Dataset Using NLP

Dr. Vijay R.Sonawane[1], Mr.Abhinav S.Thorat[2], Dr.Jaya R.Suryawanshi[3], Dr.Ravindra G.Dabhade[4], Dr. Megharani Patil[5], Dr. Bhausaheb Musmade[6]
[1]MVPS's K.B.T.College of Engineering, Nashik, Maharashtra, India
[2] Amrutvahini College of Engineering, Sangamner, Maharashtra, India
[3]MVPS's K.B.T.College of Engineering, Nashik, Maharashtra, India
[4] Matoshri College of Engineering & Research Centre, Nashik, Maharashtra, India
[5] Thakur College of Engineering & Technology, Mumbai,Maharashtra, India
[6]D.Y Patil College of Engineering, Akurdi,Pune,Maharashtra, India
[1]vijaysonawane11@gmail.com, [2]abhinav.th1990@gmail.com, [3]suryavanshi.jaya@kbtcoe.org, [4]rgdabhade@gmail.com, [5]megharanitpatil@gmail.com, [6]bbmusmade@dypcoeakurdi.ac.in

Abstract— The identification of customer churn is a major issue for large telecom businesses. In order to manage the data of current customers as well as acquire and manage new customers, every day, a substantial volume of data gets generated. Therefore, it's crucial to identify the causes of client churn so that the appropriate steps can be taken to lower it. Numerous researchers have already discussed their efforts to combine static and dynamic approaches in order to reduce churn in big data sets, but these systems still have many issues when it comes to actually identifying churn. In this paper, we suggested two methods, the first of which is churn identification and using Natural Language Processing (NLP) methods and machine learning techniques, we make predictions based on a vast telecommunication data set. The NLP process involves data pre-processing, normalization, feature extraction, and feature selection. For feature extraction, we employ unique techniques like TF-IDF, Stanford NLP, and occurrence correlation methods, have been suggested. Throughout the lesson, a machine learning classification algorithm is used for training and testing. Finally, the system employs a variety of cross validation techniques and training and evaluating Machine learning algorithms. The experimental analysis shows the system's efficacy and accuracy.

Keywords- Churn Prediction, NLP (Natural Language Processing), Feature Extraction, Feature Selection, Machine Learning Classification.

I. INTRODUCTION

One of the major sectors in developed nations is now the telecommunications industry. Telecommunications companies face an unprecedented demand for the triple-play bundle, encompassing audio, video, and online services. The key challenge lies in delivering a superior customer experience, be it service utilization or seeking assistance from providers. Today, almost everyone has access to telecom packages offering seamless connectivity and a range of service options. [1]. The degree of competition increased due to technological advancement and increased operators [2]. Businesses are putting a lot of effort into surviving in this competitive sector by utilizing various techniques.

To enhance revenue, three primary tactics have been suggested [3]: Analysing customer acquisition, upselling, and retention strategies in terms of return-on-investment values reveals that the most lucrative approach is customer retention [4]. It has been demonstrated that retaining existing customers is not only more cost-effective than acquiring new ones but also simpler than upselling [5]. To implement this strategy successfully, businesses need to address the issue of customer churn, which refers to the movement of customers from one provider to another [5]. In highly competitive service sectors, customer churn poses a significant challenge. Conversely, early identification of customers likely to churn can present an opportunity for a substantial additional revenue stream. [6].

The number of consumers departing a business each year is known as the churn rate. Any business organization may encounter difficulties because of this. Therefore, one of the most important tasks for any business is anticipating customers who may wish to depart the company. Due to recent advancements in data analytics, numerous Customer Relationship Management (CRM) system types have been integrated as data analysis methods. They have since come to dominate many studies and practices. Such analytical techniques place a strong emphasis on customer-centric strategies rather than product-centric strategies. As a result, several new advertising opportunities have arisen due to altered customer-company relationships. Reducing Churn and keeping current clients are the most profitable marketing strategies that maximize shareholder value [7]. For a company to stay in business, its customers are still its most valuable asset. For businesses, a decline in clients is an unexpected occurrence.

Corporations must therefore look into customer profiles to segment their markets and make better decisions [8].

Customer turnover is one of the most pressing concerns in today's fiercely competitive industries, such as the telecom sector [9]. Given the exorbitant expenses associated with acquiring new customers, the telecom sector has shifted its attention to retaining existing ones. This approach not only leads to increased revenues but also results in lower marketing costs compared to customer acquisition [10]. Consequently, the accurate prediction of customer turnover has become a vital component of the telecommunications industry's strategic leadership and planning process [11].

One study claims that the yearly customer churn rate for the telecom sector ranges from 20 to 40% and that keeping existing customers rate is 5–10 times less than finding new ones [12]. Therefore, forecasting customer churn is 16 times cheaper than acquiring new clients. Furthermore, when the churn rate decreases by 5%, the profit rises from 25% to 85% [13]. It shows how crucial client attrition forecasting is for the telecom sector. For telecom companies to keep their current clients and lower customer churn, CRM is essential. Therefore, the CRM analyzers' prediction systems must be accurate. If analysts cannot accurately predict consumer attrition, no advertising can be run [14].

Consumer attrition can be addressed with the aid of data mining and machine learning methods, which have recently undergone significant advancements in data science. Personalized marketing tactics are essential for companies that want to strengthen consumer interactions and attract more loyal clients. To achieve this, it is essential to offer their customer service and sales representatives the ability to gather data on the company's users and train them to communicate effectively with all of their customers. Using AI techniques, such a task could be completed in the era of big data [10] without significantly taxing the customer support and sales teams. For businesses to successfully engage with customers and gain their trust, AI must be integrated into all business processes, including marketing, social media marketing, CRM, and sales. The works on churn prediction that have already been done have been done on customers from other industries, such as employee churn [15]. Still, most of them are on telecom information because this industry has the greatest losses from losing consumers. Some researchers have expressed worry about maximizing profits while attempting to estimate the churn rate, while others have emphasized improving prediction accuracy. The objective should be increasing forecast accuracy without sacrificing profit, which calls for less difficult techniques that may be employed without a significant outlay of money. This is the fundamental problem, though, that has not been thoroughly researched. This research project uses machine learning techniques to create a churn prediction model to address this problem.

II. LITERATURE SURVEY

Consumer Businesses often face customer attrition challenges as they struggle to predict when a consumer might churn. Acquiring new customers to compensate for the losses can be highly costly in terms of both time and resources. However, with effective customer churn forecasting, customer retention becomes more achievable and efficient, allowing organizations to retain valuable consumers. For telecom companies, losing customers to competitors can lead to substantial losses. Hence, leveraging customer churn prediction enables corporations to offer more effective programs, reducing customer churn and fostering stronger customer loyalty.

Numerous churn prediction models have been developed over several years utilizing a variety of methods. For churn prediction, the use of machine learning has grown in popularity. With the use of strong algorithms, machine learning can learn from previous observations to produce precise predictions. Numerous studies that compared various machine learning tools for churn prediction were provided [12]. This section explains various machine learning algorithms for churn prediction.

Over the years, several churn prediction models have emerged, employing diverse methodologies. Among these approaches, machine learning has gained significant popularity for churn prediction. By harnessing powerful algorithms, machine learning can analyze past data and generate accurate forecasts. Extensive research has been conducted, comparing various machine learning tools for churn prediction [12]. This section delves into a comprehensive explanation of different machine learning algorithms used for churn prediction.

- Mandak and Hanclova [14] applied logistic regression to predict customer turnover in European telecom carriers, utilizing demographic and service usage data.
- Ahmed et al. [15] employed metaheuristic-based churn prediction methods on a large orange telecoms dataset, developing a hybridized version of the Firefly algorithm as the classifier.
- Another work by Ahmed et al. [16] utilized a hybrid firefly-based classification to build a churn prediction model.
- Hoppner et al. [17] adopted the ProfTree decision tree to create churn prediction models using real-world datasets from multiple telecom providers.
- Faris [18] proposed a hybrid model based on feedforward neural networks and particle swarm optimization for churn prediction.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 7
DOI: https://doi.org/10.17762/ijritcc.v11i7.7828
***Article Received**: 18 April 2023 **Revised**: 19 June 2023 **Accepted**: 30 June 2023*

- In a separate study, Sjarif et al. [19] utilized the public Telco Customer Churn dataset from Kaggle and applied the Pearson correlation and KNN algorithm for churn prediction.

These diverse approaches showcase the ongoing efforts to enhance churn prediction methods in the telecommunications domain. Various machine-learning algorithms have been applied in the context of churn prediction within the telecom industry:

- Almufadi et al. [20] employed Convolutional Neural Networks (CNN) to forecast subscriber attrition in the mobile telecom sector.
- Li and Marikannan [21] utilized Particle Swarm Optimization (PSO) and an Extreme Learning Machine to predict churn in the telecommunications industry.
- Ullah et al. [22] adopted a combination technique to estimate churn in the telecom industry. They used the random forest approach to categorize churn and non-churn clients and identify relevant components in k-means clustering. Features such as call time, free calls, prices, and others were incorporated.
- Choudhari and Potey [23] conducted predictive analysis for customer attrition in the telecom business using a hybrid decision tree and logistic regression classifier. Additionally, they suggested a hybrid approach involving fuzzy unordered rule induction and fuzzy c-means clustering for predicting client attrition.

Table 4.1 summarizes these different machine-learning algorithms for churn prediction, showcasing the diversity of techniques utilized in this area of research.

TABLE I. COMPARISON OF VARIOUS MACHINE LEARNING ALGORITHMS FOR CHURN PREDICTION

Research Paper	Churn Prediction Algorithms	Dataset	Prediction Results
[13]	K-nearest neighbor, random forest, XGBoost	IBM Watson dataset (7043 instances, 21 attributes)	XGBoost achieved an accuracy of 79.8%
[14]	Logistic regression	Two real datasets of European Telecommunications (50,000 customers)	AUC: 0.9759
[15]	Hybrid firefly-based classification	Orange telecom dataset	Accuracy: 86.38%
[16]	Boosted-stacked learners, bagged-stacked learners	UCI Churn dataset	Boosted-stacked learners: 98.4%, Bagged-stacked learners: 97.2%
[17]	Novel classifier integrating EMPC metric	Real-life datasets from various telecom providers	N/A (Metric integration for churn prediction)
[18]	Particle swarm optimization, feedforward neural network	Discovering Knowledge in Data (US mobile operators)	Accuracy: 0.920
[19]	Pearson correlation, k-nearest neighbor (KNN) algorithm	Public Telco Customer Churn dataset	Training accuracy: 80.45%, Testing accuracy: 97.78%
[20]	Convolutional neural networks	Mobile Telephony Churn Prediction Dataset	96% accuracy
[21]	Particle swarm optimization (PSO), extreme learning machine (ELM)	Kaggle Telecommunication dataset	Training accuracy: 84.71%, Testing accuracy: 81.16%
[22]	Various classification algorithms, K-means clustering	GSM telecom service dataset, churn-bigml dataset	GSM dataset (Random Forest): 88.63%, Churn bigml dataset (J48): 91.91%
[23]	Hybrid decision tree, logistic regression classifier, FURIA with fuzzy c-means algorithms	Telecommunication dataset	N/A (Advanced hybrid approach for churn prediction)

III. PROPOSED METHODOLOGY

In this section, we delve into the churn prediction methodology utilizing machine learning algorithms. Figure 3.1 illustrates and elucidates the proposed model designed for churn prediction. The initial stage entails data preprocessing, encompassing data filtering and standardization to ensure a consistent format for subsequent feature selection. To find accurate values and predict customer churn, various algorithms such as Random Forest, XGBoost, and Logistic Regression are used in this system. The trained and tested dataset is used to implement the model, resulting in the greatest number of correct values.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 7
DOI: https://doi.org/10.17762/ijritcc.v11i7.7828
***Article Received**: 18 April 2023 **Revised**: 19 June 2023 **Accepted**: 30 June 2023*

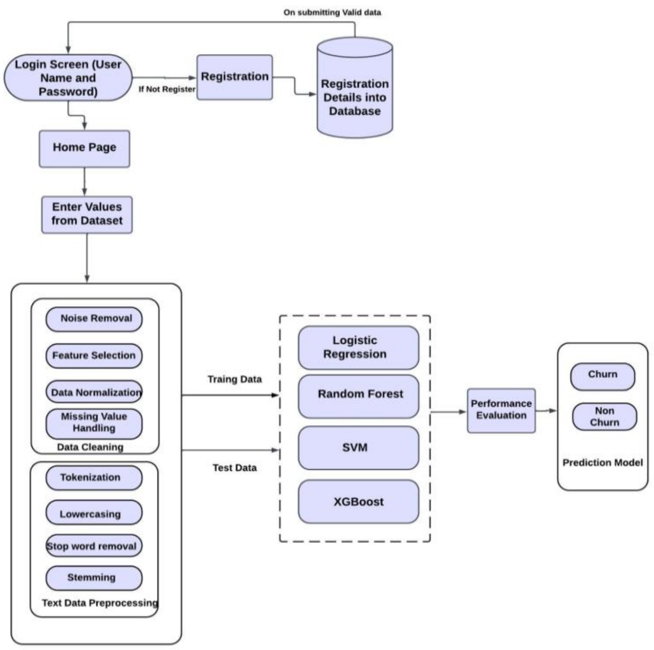


Figure 1. Proposed Model for Customer Churn Prediction

3.1 Data Preprocessing

A dataset comprises N rows, each representing a collection of features. These values may come in various formats and serve as the foundation for machine learning to comprehend the problem at hand. Datasets can be sourced from scraped information on the internet or purposefully created to reflect real-time inputs for specific issues. Each member of the dataset contains values for variables like height and object weight, with every value termed as a datum. As stated earlier, the dataset is pivotal in initiating this process. Table 3.2 showcases the Customer Data used in the preprocessing phase.

TABLE II. DATA PREPROCESSING

Attributes	Column	Non-Null Count	Dtype
1	customerID	non-null	object
2	gender	non-null	object
3	SeniorCitizen	non-null	int64
4	Partner	non-null	object
5	Dependents	non-null	object
6	tenure	non-null	int64
7	PhoneService	non-null	object
8	MultipleLines	non-null	object
9	InternetService	non-null	object
10	OnlineSecurity	non-null	object
11	OnlineBackup	non-null	object
12	DeviceProtection	non-null	object
13	TechSupport	non-null	object
14	StreamingTV	non-null	object
15	StreamingMovies	non-null	object
16	Contract	non-null	object
17	PaperlessBilling	non-null	object
18	PaymentMethod	non-null	object
19	MonthlyCharges	non-null	float64
20	TotalCharges	non-null	object
21	Churn	non-null	object

Data collection involved multiple sources, leading to diverse formats to represent single values, such as gender represented by M/F or Male/Female. Since machines comprehend only 0s and 1s, three-dimensional image data needs to be converted to a two-dimensional format, as shown in the dataset, to eliminate noise, null values, and size discrepancies. Cleaning the data can be achieved using Panda's tabular data approach for structured information and OpenCV for images, ensuring a streamlined and consistent dataset.

3.2 Data Preprocessing: Filtering and Noise Removal

Ensuring the data's quality is crucial as undesirable or null values may lead to unsatisfactory or less accurate results. The dataset is thoroughly examined to identify and retain only the most useful features, resulting in enhanced accuracy by including only relevant information.

3.3 Feature Selection and Engineering

Feature selection plays a vital role in extracting essential elements from the dataset based on domain knowledge. In this study, we carefully select features that contribute to performance improvement and decision-making while discarding less impactful ones. By focusing on valuable and highly predictive variables, the classification performance significantly improves, leading to a more efficient classification process.

3.4 Prediction and Classification

In the telecommunications industry, many techniques for predicting customer churn have been proposed. In this research, three modeling techniques are used as churn predictors. They are Random Forest (RF), Logistic Regression (LR), and XGBoost.

3.4.1 Random Forest

In the realm of customer subscription prediction, Random Forest emerges as a prominent choice. This ensemble learning method employs multiple Decision Trees to gauge whether a customer will cancel their subscription. The strength of Random Forest lies in its aggregation of numerous decision trees, each focused on predicting a specific class. By determining the class with the majority votes, the classifier for each customer is derived.

One of the key advantages of Random Forest over individual Decision Trees is its ability to mitigate sensitivity to training data. This is achieved through a technique called

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 7
DOI: https://doi.org/10.17762/ijritcc.v11i7.7828
***Article Received**: 18 April 2023 **Revised**: 19 June 2023 **Accepted**: 30 June 2023*

bagging, wherein random samples are drawn from the dataset to train decision trees, reducing the risk of overfitting.

The Random Forest approach not only produces more reliable and accurate results but also overcomes the limitation of correlated predictions found in individual Decision Trees. This improvement is due to the reduced correlation among predictions obtained from sub-trees.

In the Random Forest method, the algorithm is restricted to search for the best-split point from a randomly chosen subset of features (m), as opposed to exhaustively exploring all variables and values like the CART method.

The benefits of employing Random Forest are manifold. In the banking industry, it is particularly prized for its stability and accuracy in forecasting results, making it a valuable tool for distinguishing between loyal customers and those at risk of churn. Furthermore, Random Forest helps to alleviate the problem of overfitting by averaging predictions from multiple decision trees.

However, the complexity of the model can be a drawback, making it challenging to interpret the underlying basis for its predictions. Despite this limitation, Random Forest remains a compelling choice for its robustness and exceptional performance in customer subscription prediction.

3.4.2 Logistic Regression

In our pursuit of predicting customer churn, Logistic Regression emerges as a formidable tool. This supervised learning classification algorithm assesses the likelihood of a customer canceling their subscription.

Logistic Regression operates on the principle of setting a threshold for classification. By applying the sigmoid or logit function, which produces values between 0 and 1, the algorithm efficiently classifies customers into two categories: churned or not churned.

This binary classification algorithm is particularly well-suited for scenarios where we need to distinguish between customer-churned and non-churned cases. Its simplicity and interpretability make it one of the most accessible algorithms to comprehend.

The sigmoid function (∂) and its corresponding function value (z) are integral components of the logistic regression equation. The outcome is then determined based on whether the resulting value exceeds the threshold set for classification.

In the realm of churn prediction, Logistic Regression proves to be an invaluable tool, offering precision and clarity in its ability to identify customers likely to churn.

3.4.3 Optimal Model Performance and Efficient Execution with XGBoost

XGBoost is the preferred choice due to its exceptional execution speed and model performance. This advanced technique leverages ensemble learning, combining multiple algorithms to create a powerful single model. Notably, XGBoost facilitates parallel and distributed computing, optimizing memory utilization and enhancing overall efficiency. While the gradient boosting algorithm is used in XGBoost, it differs from other implementations in that it uses a sophisticated model to control overfitting, which improves speed. XGBoost can be used for regression and ranking in addition to categorizing data into "churn" or "no churn" categories. The feature selection process can be carried out automatically by XGBoost, and it has a parameter for randomization that helps to lessen the correlation between each tree.

IV. RESULT DISCUSSIONS

In this section, we assess the performance of the proposed churn model using machine learning algorithms. The dataset used for analysis is sourced from Kaggle and encompasses a comprehensive range of information, including:

- Customer demographic details: age, gender, occupation, education, location, marital status, and more.
- Customer account information: account type, credit limit, balance, payment history, and related data.
- Customer behavior information: frequency of purchases, time spent on the website, product usage patterns, etc.
- Customer feedback insights: customer ratings, surveys, reviews, and other feedback data.
- Marketing-related information: details about campaigns, offers, promotions, etc.
- Transaction-specific details: transaction history, purchase information, payment specifics, and more.

By analyzing this diverse dataset, we aim to evaluate the effectiveness of the churn model and its application of machine learning algorithms in predicting customer churn.The following metrics are used to evaluate the performance of the customer churn prediction: precision, recall, f-measure, and accuracy. It evaluates a predictive model's accuracy in predicting when a client will leave [24]. The information obtained using the confusion matrix is used to generate the four measures described above. Table 4.3 displays the confusion matrix representation.

ISSN: 2321-8169 Volume: 11 Issue: 7
DOI: https://doi.org/10.17762/ijritcc.v11i7.7828
***Article Received**: 18 April 2023 **Revised**: 19 June 2023 **Accepted**: 30 June 2023*

TABLE III. CONFUSION MATRIX

Predicted		
	Churners	**Non-Churners**
Churn	TP	FN
Non-Churn	FP	TN

To comprehend the evaluation criteria, we need to consider the following four terms:

i. True Positive (TP): This represents the proportion of customers correctly identified by the predictive model as churners within the churner category.

ii. True Negative (TN): This refers to the proportion of customers accurately classified by the predictive model as non-churners within the non-churner group.

iii. False Positive (FP): This denotes the number of customers who are actually non-churners but have been incorrectly labeled as churners by the predictive algorithm.

iv. False Negative (FN): This represents the number of customers who are churners but have been inaccurately classified as non-churners by the predictive model.

Furthermore, the model's accuracy (Acc) is calculated as the percentage of correctly predicted cases out of all possible predictions. It represents the ratio of all correct forecasts and can be defined as follows:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}$$

Precision (Pre): The precision metric evaluates the number of positive cases identified by the algorithm within the positive class. It quantifies the ratio of accurately predicted churners, and it can be computed as follows

$$Pre = \frac{TP}{TP + FP}$$

Recall (Rec): The recall metric helps identify the actual positive class samples that the model correctly predicted. It determines the ratio of true positives, or real churners, and can be calculated as follows:

$$Rec = \frac{TP}{TP + FN}$$

F-Measure: The F-measure, which is the harmonic mean of precision and recall, gauges the balance between accuracy and recall. It is defined as

$$F - Measure = \frac{2 \times Pre \times Rec}{(Pre + Rec)}$$

A value closer to one indicates a more effective integration of accuracy and recall by the classifier. The experimental results with the Random Forest algorithm are presented in table 4 Random Forest (RF) is a powerful classification algorithm known for efficiently handling nonlinear data. Compared to other techniques, RF yielded superior results in terms of accuracy and performance. Given the significance of accurate churn prediction, we favor techniques with improved accuracy. Similarly, we examined the results of using the Logistic Regression technique, Support vector machine and XGBoost as shown in table 5-7 respectively. Figure 2 shows the comparison of performance metrics for LR, RF, SVM and XGBoost algorithms. Additionally Figure 3 and Figure 4 feature importance according to random forest and XGBoost.

TABLE IV. CROSS VALIDATION SCORE FOR RANDOM FOREST CLASSIFIER

	0.0	**1.0**	**accuracy**	**macro avg**	**weighted avg**
precision	0.8346	0.6546	0.7992	0.7446	0.7875
recall	0.9081	0.4918	0.7992	0.7000	0.7992
f1-score	0.8698	0.5617	0.7992	0.7157	0.7892
support	1556.0000	551.0000	0.7992	2107.00	2107.0000

TABLE V. CROSS VALIDATION SCORE FOR LOGISTIC REGRESSION

	0.0	**1.0**	**accuracy**	**macro avg**	**weighted avg**
precision	0.8352	0.6432	0.7964	0.7392	0.7850
recall	0.9023	0.4973	0.7964	0.6998	0.7964
f1-score	0.8675	0.5609	0.7964	0.7142	0.7873
support	1556.0	551.00	0.7964	2107.0	2107.0

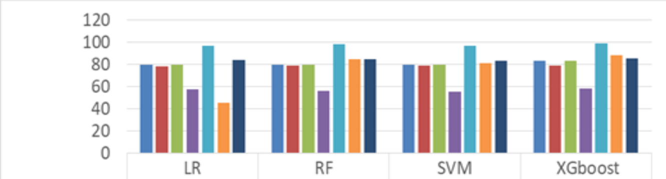
TABLE VI. CROSS VALIDATION SCORE FOR SUPPORT VECTOR MACHINE

	0.0	**1.0**	**accuracy**	**macro avg**	**weighted avg**
precision	0.8258	0.6739	0.7992	0.7498	0.7861
recall	0.9229	0.4501	0.7992	0.6865	0.7992
f1-score	0.8716	0.5397	0.7992	0.7057	0.7848
support	1556.00	551.00	0.7992	2107.00	2107.00

TABLE VII. CROSS VALIDATION SCORE FOR XGBOOST

	0.0	**1.0**	**accuracy**	**macro avg**	**weighted avg**
precision	0.8473	0.6182	0.8288	0.7274	0.7892
recall	0.8845	0.5162	0.8288	0.7005	0.8288
f1-score	0.8644	0.5668	0.8288	0.7126	0.7848
support	1556.00	551.00	0.8288	2107.00	2107.00

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 7
DOI: https://doi.org/10.17762/ijritcc.v11i7.7828
***Article Received**: 18 April 2023 **Revised**: 19 June 2023 **Accepted**: 30 June 2023*



	LR	RF	SVM	XGboost
Accuracy	79.64	79.92	79.92	82.88
Precision	78.5	78.75	78.61	78.92
Recall	79.64	79.92	79.92	82.88
F1 Score	57.37	56.3	54.95	57.89
Specificity	97.02	98.56	96.5	99.32
G-Mean	45.2	84.51	80.82	88.25
AUC	84.1	84.75	83.55	85.72

Accuracy Precision Recall F1 Score Specificity G-Mean AUC

Figure 2. Performance metrics comparison

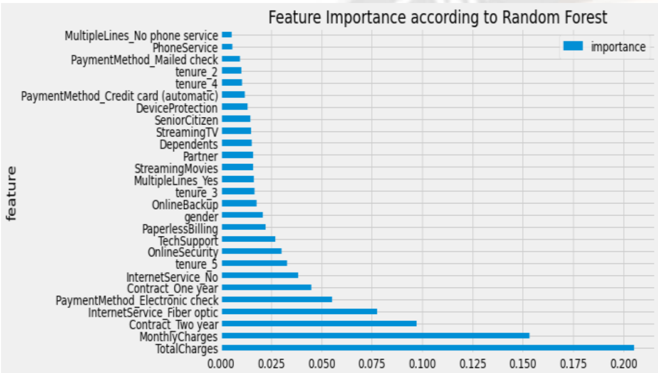


Figure 3. Feature importance according to Random Forest.

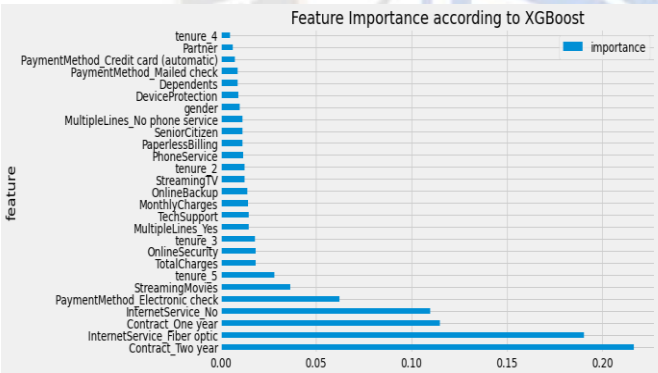


Figure 4. Feature importance according to XGBoost.

There could be several reasons for XGBoost performing so well. XGBoost has built-in regularization techniques that help prevent overfitting, which can be a problem when working with complex datasets. The dataset we used had 20 features and some of them had low correlation with the target label. For this complex dataset, XGBoost performed effectively. XGBoost is optimized for speed and efficiency, which allows it to train and make predictions quickly. As discussed before, accuracy measure is not enough in this case as the dataset used was imbalanced. Besides the XGBoost model, LG, and RF classifiers also performed excellently.

V. CONCLUSION

The designed churn prediction approach, incorporating eager learning and lazy learning techniques, has significantly enhanced the system's accuracy. XGBoost algorithm has proven to be instrumental in achieving superior performance compared to other traditional methods like Random Forest and Logistic Regression. The XGBoost algorithm's powerful ensemble learning capabilities have demonstrated their effectiveness in handling complex churn prediction tasks. XGBoost has surpassed the accuracy achieved by Random Forest and Logistic Regression. Furthermore, XGBoost's ability to handle imbalanced datasets, a common characteristic in churn prediction scenarios, has contributed to more reliable predictions. The incorporation of XGBoost has not only improved the accuracy of churn prediction but has also reduced computational time, making it a viable choice for real-time or large-scale churn prediction implementations. Overall, the integration of the XGBoost algorithm in the churn prediction approach has been a significant factor in the success of this project. The combination of eager and lazy learning techniques with XGBoost has enabled the system to identify customer behavior patterns, influential factors, and calculate churn rates more accurately and efficiently.

REFERENCES

[1] J.Z. Zhang, and C.W. Chang, "Consumer dynamics: Theories, methods, and emerging directions", Journal of the Academy of Marketing Science, Vol. 49, Page 166-196, 2021

[2] S.A. Banday, and S. Khan, "Evaluation Study of Churn Prediction Models for Business Intelligence", In Big Data Analytics, Auerbach Publications, Page 201-213, 2021

[3] A. Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo, S. Anwar, "Customer churn prediction in telecommunication industry using data certainty", Journal of Business Research, Vol. 94, Page 290-301, 2019.

[4] A.K. Ahmad, A. Jafar, K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform", Journal of Big Data, Vol. 6, Issue 1, Page 1-24, 2019

[5] D. Bell, and C. Mgbemena, "Data-driven agent-based exploration of customer behavior", Simulation, Vol. 94, Issue 3, Page 195-212, 2018 [6] Yang, L. J. (2011) US Patent 8,033,499B2.

[6] L.C. Cheng, C.C. Wu, and C.Y. Chen, "Behavior analysis of customer churn for a customer relationship system: an empirical case study", Journal of Global Information Management (JGIM), Vol. 27, Issue 1, Page 111-127, 2019

[7] P. Lalwani, M.K. Mishra, J.S. Chadha, and P. Sethi, "Customer churn prediction system: a machine learning approach", Computing, Page 1-24, 2022

[8] V. Kavitha, G.H. Kumar, S.M. Kumar, and M. Harish, "Churn prediction of customer in telecom industry using machine learning algorithms", International Journal of Engineering Research & Technology (IJERT), Vol. 9, Issue 5, Page 181-184, 2020

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 7
DOI: https://doi.org/10.17762/ijritcc.v11i7.7828
***Article Received**: 18 April 2023 **Revised**: 19 June 2023 **Accepted**: 30 June 2023*

[9] S.K. Routray, "Marketing strategy through machine learning techniques: A case study at telecom industry", International Journal of Innovation Engineering and Science Research, Vol. 5, Issue 3, Page 21-30, 2021

[10] S. Kim, and H. Lee, "Customer churn prediction in influencer commerce: an application of decision trees", Procedia Computer Science, Vol. 199, Page 1332-1339, 2022

[11] X. Hu, Y. Yang, L. Chen, and S. Zhu, "Research on a customer churn combination prediction model based on decision tree and neural network. In 2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), IEEE, Page 129-132, 2020

[12] Krishnan, V. G. ., Saradhi, M. V. V. ., Dhanalakshmi, G. ., Somu, C. S. ., & Theresa, W. G. . (2023). Design of M3FCM based Convolutional Neural Network for Prediction of Wheat Disease. International Journal of Intelligent Systems and Applications in Engineering, 11(2s), 203 –. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2523

[13] H. Jain, A. Khunteta, and S. Srivastava, "Churn prediction in telecommunication using logistic regression and logit boost", Procedia Computer Science, Vol. 167, Page 101-112, 2020

[14] J. Pamina, B. Raja, S. SathyaBama, M.S. Sruthi, and A. VJ, "An effective classifier for predicting churn in telecommunication", Journal of Advanced Research in Dynamical & Control Systems, Vol. 11, 2019

[15] J. Manďák, and J. Hančlová, "Use of Logistic Regression for Understanding and Prediction of Customer Churn in Telecommunications", Statistika: Statistics & Economy Journal, Vol. 99, Issue 2, 2019

[16] A.A. Ahmed, and D. Maheswari, "Churn prediction on huge telecom data using hybrid firefly based classification", Egyptian Informatics Journal, Vol. 18, Issue 3, Page 215-220, 2017

[17] M. Ahmed, H. Afzal, I. Siddiqi, M.F. Amjad, and K. Khurshid, "Exploring nested ensemble learners using overproduction and choose approach for churn prediction in telecom industry", Neural Computing and Applications, Vol. 32, Page 3237-3251, 2020

[18] S. Höppner, E. Stripling, B. Baesens, S. vanden Broucke, and T. Verdonck, "Profit driven decision trees for churn prediction", European journal of operational research, Vol. 284, Issue 3, Page 920-933, 2020

[19] H. Faris, "A hybrid swarm intelligent neural network model for customer churn prediction and identifying the influencing factors", Information, Vol. 9, Issue 11, 2018

[20] Prof. Barry Wiling. (2018). Identification of Mouth Cancer laceration Using Machine Learning Approach. International Journal of New Practices in Management and Engineering, 7(03), 01 - 07. https://doi.org/10.17762/ijnpme.v7i03.66

[21] N.N.A Sjarif, M. Rusydi, M. Yusof, D. Hooi, T. Wong, S. Yaakob, et al., "A customer Churn prediction using Pearson correlation function and K nearest neighbor algorithm for telecommunication industry", International Journal of Advances in Soft Computing and Its Applications, Vol. 11, Issue 2, Page 46-59 2019

[22] N. Almufadi, A.M. Qamar, R.U. Khan, and M.T.B. Othman, "Deep learning-based churn prediction of telecom subscribers", International Journal of Engineering Research and Technology, Vol. 12, Issue 12, Page 2743-2748, 2019

[23] Martínez, L., Milić, M., Popova, E., Smit, S., & Goldberg, R. Machine Learning Approaches for Human Activity Recognition. Kuwait Journal of Machine Learning, 1(4). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/146

[24] K.G. Li, and B.P. Marikannan, "Hybrid particle swarm optimization-extreme learning machine algorithm for customer churn prediction", Journal of Computational and Theoretical Nanoscience, 16(8), 3432-3436, 2019

[25] Leila Abadi, Amira Khalid, Predictive Maintenance in Renewable Energy Systems using Machine Learning , Machine Learning Applications Conference Proceedings, Vol 3 2023.

[26] Ullah, B. Raza, A.K. Malik, M. Imran, S.U. Islam, and S.W Kim, "A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector", IEEE access, Vol. 7, Page 60134-60149 ,2019

[27] A.S. Choudhari, and M. Potey, "Predictive to prescriptive analysis for customer churn in telecom industry using hybrid data mining techniques", In 2018 Fourth international conference on computing communication control and automation (ICCUBEA), IEEE, Page 1-6, 2018

[28] R. Liu, S. Ali, S.F. Bilal, Z. Sakhawat, A. Imran, A. Almuhaimeed, et al., "An Intelligent Hybrid Scheme for Customer Churn Prediction Integrating Clustering and Classification Algorithms", Applied Sciences, Vol. 12, Issue 18, 2022