IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429

Volume 4, Issue 2, October 2024

Impact Factor: 7.53

The Role of Predictive Analytics in Customer Churn Prevention Across Global Markets

Chetankumar Patel

Independent Researcher

chetankumar24.patel@gmail.com

Abstract: *Customer churn—when a customer stop buying a company's products or using its services—has a bad influence on the revenues. Some industries have churn rates of around 20-40% each year. The paper presents a strong model of estimating the customer turnover based on the Kaggle Telecom turnover Dataset. The proposed technique consists of extensive data preprocessing that gets rid of irrelevant features, fills in missing values, detects outliers, and corrects class imbalance with the help of the SMOTE-ENN method. Pearson correlation analysis and visualizations are very helpful in understanding the relationships between features and in reducing dimensions while retaining important information. On the other hand, Principal Component Analysis (PCA) eliminates some of the data's dimensions but retains the significant information. The accuracy (ACC), precision (PRE), recall (REC), F1-score (F1), ROC curve, and confusion matrix are some of the performance metrics used to train and evaluate an XGBoost classifier. The accuracy score is 98.42. The experimental findings show that the suggested model outperforms the existing machine learning approaches and has extremely good prediction performance. The study's findings validate the model's use in the creation of proactive customer retention strategies within the telecom industry by demonstrating its capacity to properly anticipate customer attrition.*

Keywords: Customer Churn, Class Imbalance Correction, Feature Relationships, Pearson Correlation Analysis, Churn Prediction.

I. INTRODUCTION

Products and services are becoming increasingly similar in today's business environment, where competition is heating up. Over the last few years, a shift has been observed in marketing campaigns towards products, to consumers [1]. Minimizing user turnover should be the top priority for the organization. A business spends a tonne of money developing a new user, what with all the shifting priorities in product pricing, marketing, and service improvement. Potential clients more swayed by advertising and PR [2]. Consequently, growing a company's client base and the market as a whole depend on keeping current clients happy and rewarding loyal users [3].

Customer churn is the rate of a total number of customers a company loses during a certain time period or under a certain contract. The majority of companies experience customer attrition, which worsen when customer data is scarce [4]. Companies struggle to anticipate when customers are likely to go due to this lack of understanding, which in turn causes customer relationship crises and churn. Higher profit margins are generated by maintained customers , making customer retention a far more lucrative strategy than client acquisition. Also, keeping existing clients costs five to 10 times as much as acquiring new ones. Therefore, it is vital for firms providing continuous services to reduce client turnover. In order to effectively address customer turnover, businesses must have a complete understanding of their customers' thoughts and behaviours. This enable them to anticipate and mitigate churn risk beforehand. So, this may be considerably improved with the use of technology analysis for churn detection.

One of the most common challenges that businesses face is keeping customers loyal [5]. A number of people in the banking, telecom, insurance, gambling, and academic sectors are interested in user churn forecasts. Churn rates have been increasing across a number of industries, including finance, where the cost of acquiring new users is more than five times more than the cost of retaining current customers [6]. Thus, for banks to enhance their core competitiveness, it is crucial to construct a model that can forecast subscriber attrition, provide early warning when churn is likely, and allow account managers to take effective actions to avoid churn.

Consumers' voluntary churn may be better understood with the use of predictive marketing analytics and churn analysis, which employ sophisticated machine learning algorithms. A fundamental principle in strategic marketing is the importance of happy customers. How many customers a telecom company quickly loses to rivals is determined by its churn rate. Telecom businesses continue to incur losses while trying to attract new consumers, despite the fact that it is easier to hold on to current ones. Therefore, in order to handle churn, first have churn prediction and client retention. Several techniques for identifying customers who are leaving have been developed throughout the years. The use of mathematical models to make predictions, however, has led to improved ACC in churn detection [7]. Subscription data security is compromised when churn characteristics from customer data are examined. Data 'value' analysis leads to better digital marketing and more money, but at the cost of consumers' privacy.

The service delivery phase in particular is still very much in its analog form, but that is about it. The TH industry relies heavily on human labour, thus companies in this sector need to invest in AI training if they want to make money off of AI. The capacity of an organization to appropriately choose, oversee, and utilize resources designated for artificial intelligence is known as AI competency. To effectively apply ML algorithms, it is crucial to first identify which characteristics adequately reflect the data [8]. Of paramount relevance in customer churn analysis are the behavioural patterns that indicate a client's tendency to churn. The many ML data and algorithm processing tools do not, however, fully support cognitive and behavioural characteristics. Modern churn prediction approaches rely on Machine Learning [9] and DL due to their ability to efficiently evaluate massive, multi-dimensional, and ever-changing consumer information [10]. On top of that, the intricacies of customer behavior are often oversimplified by the conventional methods of churn prediction, which rely on algorithms or statistical models.

A. Motivation and Contribution of Paper

Intensifying market competition and reduced product differentiation have made customer retention a major challenge, especially in the telecom sector. The loss of customers results in a substantial decrease in revenue, and the classical prediction techniques usually do not manage to detect the complicated customer behavior. Researchers turned to advanced ML techniques to identify clients most likely to churn in an effort to achieve early and accurate churn detection. This, in turn, allow the firms to implement the retention policies and thus make their businesses more sustainable. The contributions of the study are framed below:

- Analysis on Kaggle telecom churn data with various data collections available between different data generations.
- Data pre-processing data performed in a comprehensive manner, which consisted of the removal of unnecessary columns, the handling of missing values, and the detection of outliers.
- Applying SMOTE-ENN to rectify class imbalance and enhance data quality.
- Data dimensionality may be substantially reduced using principal components analysis (PCA).
- Implementation of an XGBoost classifier for robust churn prediction.
- Taken together, ROC, ACC, PRE, REC, and F1 make up a thorough performance evaluation.

B. Significance of the Study

Predicting customer turnover in the telecom business has never been easier than with the help of this study's methodical and effective methodology. An sophisticated machine learning model, together with data cleansing, class balancing, and dimensionality reduction, helps to uncover the essential patterns associated with customer attrition in the research. The precise prediction of churn permits the telecom industries to pre-emptively execute retention actions, lowering revenue loss, and enhancing the customer's satisfaction level, and also supports the making of decisions based on data.

C. Structure of Paper

The paper is organized as follows: In Section II, conduct a literature study on the topic of customer churn prediction in international markets. In section III, the process flow is discussed including data set, techniques, and model implementation. Section IV outlines the study results. Section V concludes the study and demonstrates the future direction.

II. LITERATURE REVIEW

The present literature review gives a detailed report on the new developments in predicting customer churn in various global markets. A concise overview of the studies under review is illustrated in Table I, which presents the methods used, performance results, major findings, recognized drawbacks, and recommended paths for further research.

H M et al. (2023) This method offers telecom operators a thorough examination of customer turnover by integrating customer segmentation with churn prediction. The experiments make use of four ML classifiers. One of the first uses of ML classifiers is to predict customer attrition. To narrow down the issues with imbalanced datasets, they apply SMOTE, the proposed Synthetic Minority Oversampling Scheme. The best technique, according to the experimental investigation, is Gradient Boosting Classifier, which has achieved an ACC of 95.13% [11].

Sharma and Desai (2023) This study's findings highlight the power of ML models to spot trends and patterns in customer behavior, paving the way for proactive measures to lower customer churn and boost happiness and loyalty. In terms of training ACC, XGBoost achieves 96.6%, while in terms of testing ACC, it achieves 100%. This work makes a substantial contribution to the creation of a reliable churn prediction system, which may assist organizations in identifying clients who are on the verge of leaving and implementing focused retention tactics [12].

Abbas, Usman and Qamar (2022) This study used a retail shop data set to do exploratory data analysis and feature engineering. LR, DT, XGboost, KNN, and RF are the five methods that have been utilized. The classification approaches' performance has been evaluated using PRE, ACC, AUC, F1, and REC. This study proves that the proposed technique may help management keep consumers by successfully forecasting customer turnover 73% of the time [13].

Kavyarshitha, Sandhya and Deepika (2022) supported by research that highlights the possibility of customer commitment. The evaluation relies on the conventional memory, SVM, KNN, and AI algorithms, all of which use authentication. This review also used ANN definitions and the TensorFlow and Keras frameworks. A beat model informative index is crucial to this investigation. All this information was found in Kaggle. Viewed as the correct model versus the correct one, the results are contrasted. Thus, the recollection of the Random computation is now front and centre. A respectable 87% were found. The definition was similarly low at 78.3% using option tree analysis [14].

Nagaraju and Vijaya (2021) detailed a process for implementing ML methods to aid such a company in addressing customer attrition. Applying real-world data from the XYZ Insurance customer turnover dataset kept in Indonesia, they evaluate several ML and DL methods. As a classifier, this article makes use of: Methods used in DT include forward selection, NB, and ANN. After ANN and NB, the strongest choices, DT with forward selection performs best with an ACC of 91.3111% and a 0.970 [15].

He et al. (2021) In the context of the customer churn prediction model, the influence on ACC, AUC, and F1 was assessed using the most prevalent predictive models, such as SVM, RF, KNN, and gradient boosting classifier. Experiments showed that RFs achieved a success rate of 94.29% and the Gradient boosting classifier a success rate of 95.32%. The greatest AUC for both the gradient boosting classifier and RF was 91%. Outperforming all others, the Gradient Boosting Classifier has the greatest F1 of 97.3% [16].

Table 1: Summary of Literature Review on Customer Churn Prediction Across Global Markets.

References	Methodology	Dataset	Key Findings	Challenges	Future Scope
H. M. et al. (2023)	Customer segmentation combined with churn prediction using four ML classifiers; SMOTE applied to handle class imbalance	Telecom customer dataset	Gradient Boosting Classifier achieved the best performance with 95.13% accuracy, demonstrating effectiveness in churn prediction for imbalanced datasets	Handling highly imbalanced data and ensuring model generalizability across different telecom environments	Integration of real-time churn prediction systems and hybrid models combining deep learning with ML

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 4, Issue 2, October 2024
Impact Factor: 7.53

Sharma & Desai (2023)	Machine learning–based churn prediction with comparative analysis of classifiers, including XGBoost	Customer behavioral dataset (domain not explicitly specified)	XGBoost achieved the highest performance with 96.6% training accuracy and 100% testing accuracy, proving strong predictive capability	Risk of overfitting due to extremely high testing accuracy; limited dataset diversity	Deployment of scalable churn prediction frameworks and validation on larger, real-world datasets
Abbas, Usman & Qamar (2022)	EDA and feature engineering with LR, RF, DT, KNN, and XGBoost	Retail store customer dataset	The proposed model achieved 73% accuracy, showing ML usefulness in retail churn prediction	Lower accuracy compared to telecom and service-based datasets; limited feature richness	Incorporation of advanced feature selection techniques and deep learning approaches
Kavyarshitha, Sandhya & Deepika (2022)	Comparative study of SVM, KNN, Decision Tree, ANN using Keras and TensorFlow	Kaggle customer dataset	Random Forest performed best with ~87% accuracy; Decision Tree showed the lowest accuracy at 78.3%	Dependence on benchmark datasets; limited discussion on imbalance and scalability	Exploration of ensemble and hybrid deep learning models with real-time datasets
Nagaraju & Vijaya (2021)	Integration of ML and DL methods: DT utilizing forward feature selection, NB, and ANN	XYZ Insurance customer churn dataset (Indonesia)	DT with forward selection achieved the highest accuracy of 91.31%, outperforming ANN and NB	Feature dependency and reduced interpretability in ANN models	Expansion to cross-industry churn datasets and use of explainable AI techniques
He et al. (2021)	Comparative evaluation of SVM, RF, KNN, and Gradient Boosting using performance metrics (Accuracy, AUC, F1-score)	Customer churn dataset (domain not explicitly specified)	Gradient Boosting Classifier outperformed others with 95.32% accuracy, 91% AUC, and 97.3% F1-score	Computational complexity and tuning requirements of ensemble models	Optimization of ensemble methods and application to large-scale, real-time customer data

III. METHODOLOGY

The suggested method includes acquiring the telecom churn dataset, performing pre-processing by eliminating superfluous columns, Missing value treatment and outliers detection, then the class distribution is equalized through SMOTE-ENN. XGBoost is trained and evaluated using ACC, PRE, REC, F1, and ROC measures. Next, principal component analysis (PCA) is used to reduce the feature space. A training set and a testing set are then created from the dataset. The process flow is shown in Figure 1.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 4, Issue 2, October 2024
Impact Factor: 7.53

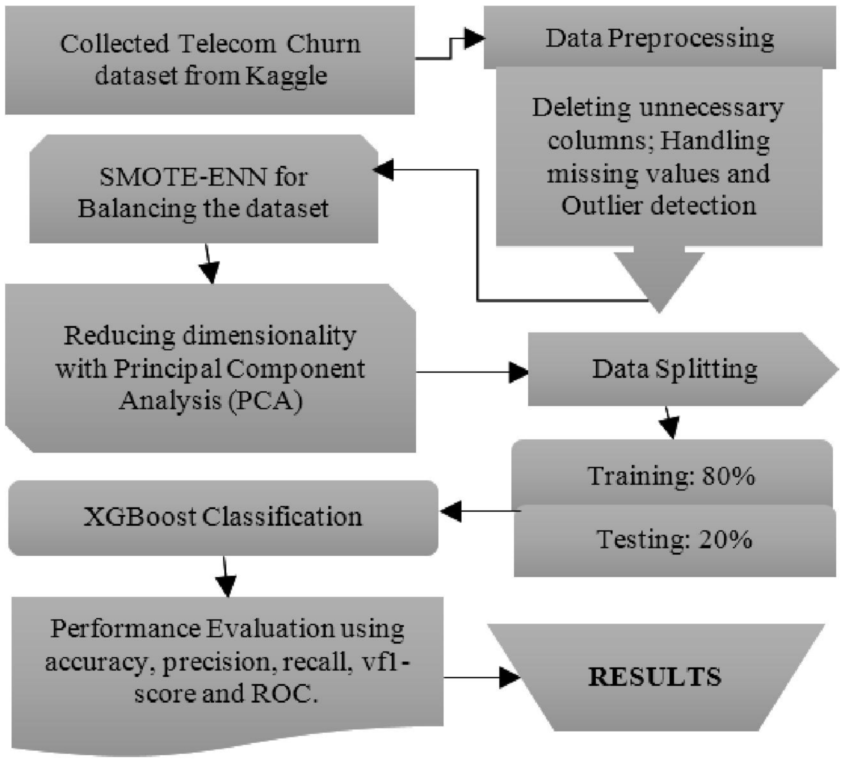

Fig. 1. Flowchart Diagram of the Customer Churn Prediction Across Global Markets

A wide range of developments, such as the flowchart, are a necessary tool to underpin step-by-step processes:

A. Data Gathering and Visualizations

This research work employs the Kaggle Telecom Churn Dataset, a dataset which has undergone cleaning and is ready for use in customer churn prediction. The churn prediction project counts up to a total of 4,000 records, of which 3,333 (20 features) are represented in the churn-bigml-80.csv file and 667 (20 features) in the churn-bigml-20.csv file, respectively for training and testing, and they all contain a binary indicator for churn and other features related to call behavior, usage patterns, and account information.

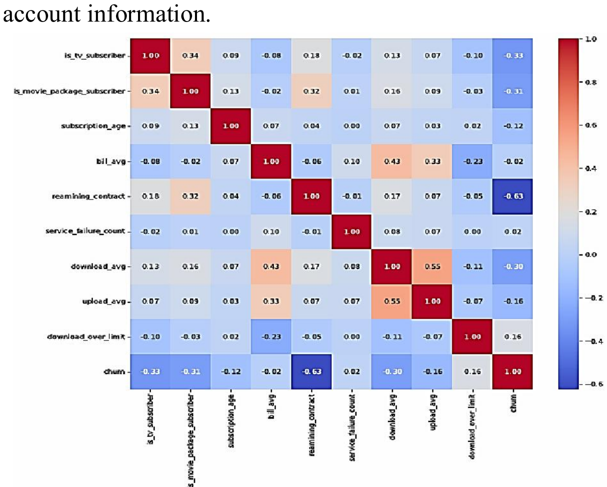

Fig. 2. Pearson Correlation Heatmap

Fig. 3.

DOI: 10.48175/IJARSCT-19900F

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 4, Issue 2, October 2024
Impact Factor: 7.53

Figure 2 shows a Pearson correlation heatmap where most features have weak to moderate relationships. Customer retention is lower among existing and long-term subscribers, since there is a moderate negative link between turnover and remaining contract length and data consumption, billing, and download and upload usage.

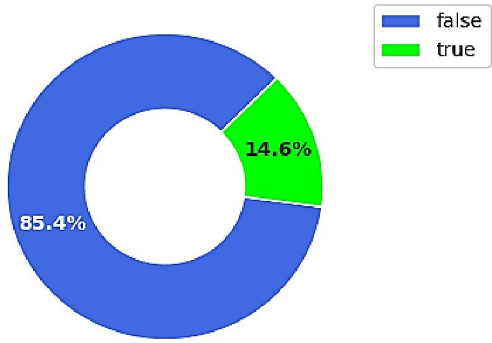

Fig. 4. Imbalanced Data Distribution

The churn variable's distribution is shown in Figure 3 using a donut chart. The majority of customers (85.4%) did not churn, while a smaller proportion (14.6%) churned. This indicates an obvious class imbalance in the data most of the non-churning customers far outnumber the ones churning.

B. Data Preparation

This is a necessary procedure for cleaning up the actual world of all the unnecessary and unclean data. This problem must be solved since the data is compiled from many sources. Cleaning and organizing data through classification and pre-processing makes it much easier to work with. To make data more usable, pre-processing and data categorization remove unnecessary details.

- **Deleting unnecessary columns:** To clean up the dataset, reduce noise and enhance the overall efficiency, unnecessary columns like customer IDs, duplicated indices and features with high redundancy or low value are removed, thus only relevant variables are left for churn analysis and modelling.
- **Handling Missing Values:** Imputation techniques allow us to efficiently handle missing data, such as replacing numerical data with mean or median and categorical data with mode. This results in a fully developed and trustworthy dataset, which is perfect for churn analysis and modelling.
- **Outlier Detection:** Utilized Tukey's gates to identify any numerical feature outliers in order to handle the issue of possible outliers in the dataset. The bounds are defined as Equation (1):

$$\text{Bounds} = Q_1 \pm 1.5\ \text{IQR} \quad (1)$$

In the experiment, no outliers were identified within the defined thresholds.

C. SMOTE-ENN for Data Balancing

In real life, a lot more people choose to keep doing business with a company than choose to leave. Since the foundation of ML and DL models is a balanced dataset, using imbalanced datasets directly cause models to be majority class fitted. Resampling algorithm SMOTE-ENN integrates two methods: ENN and synthetic minority over-sampling (SMOTE). Implementing SMOTE on imbalanced data enhances the model's capacity to detect subsets of churned customers in the customer churn prediction job. Using the SMOTE to synthesize samples does, however, contribute to the generation of certain noisy samples. The SMOTE-ENN resampling algorithm successfully removes noisy data by using the ENN approach to process the nearest neighbors of the samples for each majority class.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 4, Issue 2, October 2024
Impact Factor: 7.53

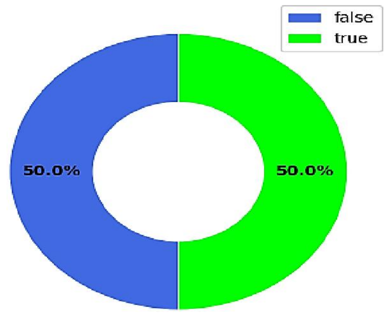

Fig. 5. Balanced Data Distribution

The donut chart presented in Figure 4 shows the classes being equally distributed, with 50 percent of the samples being labelled as "true" and the remaining 50 percent as "false." The abhorrent antithesis presented in this passage is between the world of reality and the world of illusion.

D. Dimensionality Reduction using Principal Component Analysis (PCA)

PCA is a way to examine the interrelationships of a system's variables, provided that all the relevant variables are known. Since the user does not have any prior information or expectations regarding the method's output, it is categorized as an unsupervised machine-learning approach. It categorizes the operational variables in order to identify the factors that have the most impact on them. Using PCA, one may determine which factors have the most impact on the dataset's target population. By reducing the number of input variables, the PCA method lowers the dimensionality of the issue. These are known as PCs, or primary components. Simplifying the issue while preserving as much ACC as feasible is the major objective of the PCA technique. Its goal is to reduce complexity by keeping most of the input feature information and preserving the dissimilar text values. PCA is a statistical method that separates input information that may be associated. The dimensionality problem is addressed by principal component analysis (PCA) [17] by the orthogonal projection of observational data onto lower-dimensional subspace PCs. As a result, the problem becomes one with fewer dimensions while maintaining the ACC.

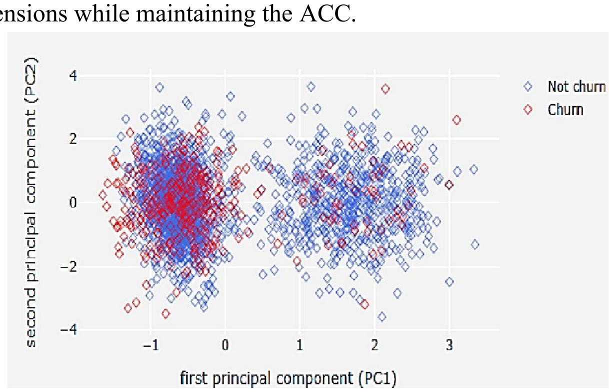

Fig. 6. Principal Component Analysis in the Dataset

A PCA scatter plot is shown in Figure 5, where the distributions of churn and non-churn customers are overlapping to some extent, with the first principal component being able to separate them to some degree, which suggests that the PCA alone is not sufficient to completely distinguish the churn behavior.

E. Data Splitting

Training, validation, and testing are the three sections that make up the dataset. For training and cross-validation, the telecom dataset was split into two parts: 80% (churn-80) and 20% (churn-20).

IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429 Volume 4, Issue 2, October 2024 Impact Factor: 7.53

F. Classification Model: Extreme Gradient Boosting (XGBoost)

A scalable, adaptable, and resource-efficient tool, the XGBoost technique was designed to address the shortcomings of earlier gradient boosting approaches [18]. while compared to other gradient boosting approaches, XGBoost stands out due to its innovative regularization strategy that controls overfitting. This approach makes XGBoost more efficient and resilient while fine-tuning the model. To make this method more consistent, and may modify Equations (2) and (3) by adding the following term to the loss function:

$$L(f) = \sum_{i=1}^{n} L(\hat{y_i}, y_i) + \sum_{m=1}^{M} \Omega(\delta_m) \quad (2)$$

with

$$\Omega(\delta) = \alpha|\delta| + 0.5_\beta||w||^2 \quad (3)$$

in where w stands for the value of every leaf, Ω depicts the regularization function, and |δ| signifies the quantity of branches.

G. Evaluation Parameters

When the conventional ACC metrics are utilized to measure the performance of classifiers on imbalanced datasets, they are not able to render a truthful image since they depend primarily on the correct classification of the majority class. In order to give a better evaluation the study stresses the use of PRE, REC and F1 coming from the confusion matrix.

A famous way to measure how well a classification model predicts the real classes is using the confusion matrix. When evaluating a model's ACC in classification, the Confusion Matrix is useful for more than just its aesthetic value. The four values can also be used for a number of computations. In order to find out where the model excels and where it falls short, these computations are all aimed at more precise assessments of the model.

- **TN:** stands for true negative. In this case, both the audience and the model agree that the consumers in question do not constitute churners.
- **FP:** stands for false positive. In this case, consumers are seen as non-churners, yet according to the model, they are actually churners.
- **FN:** stands for false negative. Although consumers are seen as churners, the model has labelled them as non-churners.
- **TP:** stands for true positive. Customers are watched and categorized as churners here.

The evaluation metrics for individual classes were calculated using standard classification parameters:

Accuracy: Accuracy rate of classification (ACC) is the ratio of correctly classified cases to the total number of occurrences. This is a total score that takes into account the model's overall performance. Here is how ACC is calculated in Equation (4):

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Precision: The amount of correct positive predictions made by a model and the number of real positive estimations make up precision. The method used to determine this is Equation (5):

$$\text{Precision} = \frac{TP}{TP+FP} \quad (5)$$

Recall/Sensitivity: The real positive rate is this. How many consumers were accurately identified as churners is shown by the rate. The following is the formula for sensitivity testing: Equation (6):

$$\text{Recall} = \frac{TP}{TP+FN} \quad (6)$$

F1-Score: One way to look at the F1 is as a weighted average of the two metrics, REC and ACC. An F1 may go as high as 1 and as low as 0. When calculating the F1, REC and ACC both contribute equally. Equation (7) is used to determine it

$$F_1 - \text{Score} = 2 \times \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

ROC Curve: A receiver operating characteristic (ROC) plot shows how well a model can make accurate classifications. The graph shows the FPR (x-axis) and the TPR (y-axis) as a correlation.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 4, Issue 2, October 2024
Impact Factor: 7.53

IV. RESULTS ANALYSIS AND DISCUSSIONS

The machine used for all the studies has 32.0 GB of RAM, an Intel Core i7 9850H × 12, Intel UHD Graphics 630, and NVIDIA Quadro T2000 graphics cards, as well as a 1.0 TB mechanical drive. The operating system is Ubuntu 24.04.2 LTS (Kernel 6.11.0-24-generic), coupled with GNOME 46, X11, and Python 3.10.16.

A. Result Demonstrations

The current state of the XGboost model's performance in predicting customer turnover across various markets is shown in Table II. The model demonstrates excellent predictive capability with an ACC of 98.42%. The commendable PRE (98.35%) and REC (97.32%) denote that the churned customers were accurately spotted with only a few incorrect predictions, and the F1 of 97.83% not only ensures but also illustrates the model's robustness and reliability through a perfect balance between PRE and REC.

Table 2: Model Performance for Customer Churn Prediction Across Global Markets

Metrics	XGBoost
Accuracy	98.42
Precision	98.35
Recall	97.32
F1-Score	97.83

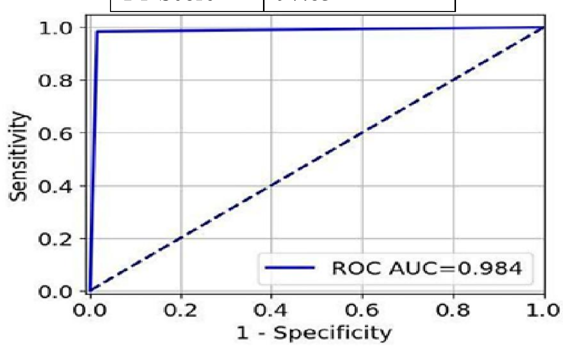

Fig. 7.ROC Curve Analysis

Figure 6 displays the ROC curve of the suggested model for predicting customer turnover. Its great sensitivity and specificity are indicated by its near proximity to the top-left corner. The outstanding capacity of the model to differentiate between churn and non-churn consumers is validated by the high AUC value of 0.984.

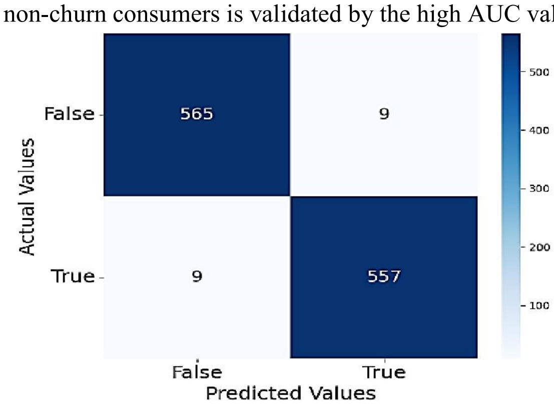

Fig. 8.Confusion Matrix of the XGBoost Model

Figure 7 shows the proposed model's confusion matrix; it demonstrates the model's balanced and strong predictive performance by showing that it correctly predicts a large number of times (565 true negatives and 557 true positives) with very few misclassifications (9 false positives and 9 false negatives).

B. Comparative Analysis

Table III compares the ACC of several ML models in predicting customer attrition on a worldwide market. With the best results in terms of ACC (98.42%), PRE (98.35%), REC (97.32%), and F1 (97.83%), the XGBoost model outperforms all of the others that were tested, including LR, SVM, and DT. Although LR and SVM get moderate scores and DT mentions only ACC, the XGBoost model discloses its strength and versatility in prediction, thus proving its worth in customer churn forecasting.

Table 3: Comparison in Customer Churn Prediction Across Global Markets

Model	Acc.	Pre.	Rec.	F1-Sc.
LR [19]	84.82	79.78	85.76	80.54
SVM [20]	79.92	78.61	79.92	54.95
DT [21]	91.22	-	-	-
XGBoost	98.42	98.35	97.32	97.83

C. Discussion

The discussion demonstrates that the proposed customer churn prediction model utilizing XGBoost excels in performance and reliability evaluation for all international markets. The model is able to distinguish between churn and non-churn customers without significantly reducing the ACC of detections or misclassifications, according to the assessment findings that are backed by ROC analysis and the confusion matrix. The results of the comparison demonstrate that XGBoost outperforms more conventional ML approaches, such as DTs, LR, and SVM models.

V. CONCLUSION AND FUTURE WORK

The development business is particularly vulnerable to customer turnover due to the inextricable relationship between the development business and the core engineering business. This research proves that machine learning–based customer churn prediction is very effective using the Telecom Churn Dataset, which mainly deals with real-world problems like class imbalance and intricate customer behavior patterns. The outcome of the experiments confirms that the suggested model based on XGBoost delivers very accurate and trustworthy churn forecasts, attaining a total correctness of 98.42% besides the high values of PRE, REC, and F1. It means that the model can recognize both churn and non-churn customers correctly, and at the same time, it can keep a balanced trade-off between FP and FN. The results of the ROC and the confusion matrix analyses not only emphasized the discriminating power and durability of the model but also confirmed it. The comparison with classical methods like LR, SVM, and DT plainly shows that among them, XGBoost has been the best one to use for revealing the nonlinear relationships and to detect the minute global behavior trends.

Future research could make this study even more comprehensive by using real-time customer information and time-based behavioural patterns to facilitate early churn detection. The incorporation of deep learning and explainable AI methods could not only increase the ACC of the predictions but also make the interpretations clearer, while the validation of the model through different industries and larger datasets would make it more widely applicable.

REFERENCES

[1] C. Patel, "A Survey of Data-Driven Customer Segmentation Methods for Targeted Marketing Campaigns," vol. 3, no. 3, pp. 154–162, 2023, doi: 10.56472/25832646/JETA-V3I7P119.

[2] J. Li, X. Bai, Q. Xu, and D. Yang, "Identification of Customer Churn Considering Difficult Case Mining," *Systems*, vol. 11, no. 7, 2023, doi: 10.3390/systems11070325.

[3] A. R. Bilipelli, "Visual Intelligence Framework for Business Analytics Using SQL Server and Dashboards," *ESP J. Eng. Technol. Adv.*, vol. 3, no. 3, pp. 144–153, 2023, doi: 10.56472/25832646/JETA-V3I7P118.

[4] Y. Macha, "A Review of Cloud-Based CRM Systems in Healthcare: Advances, Tools, Challenges, and Best Practices," *Int. J. Curr. Eng. Technol.*, vol. 12, no. 6, pp. 848–856, 2022, doi: 10.14741/ijcet/v.12.6.20.

[5] K. M. R. Seetharaman, "Digital Transformation in Retail Sales: Analyzing the Impact of Omni-Channel

Strategies on Customer Engagement," *J. Glob. Res. Math. Arch.*, vol. 10, no. 12, 2023, doi: 10.5281/zenodo.15280578.

[6] Y. Liu, M. Shengdong, G. Jijian, and N. Nedjah, "Intelligent Prediction of Customer Churn with a Fused Attentional Deep Learning Model," *Mathematics*, vol. 10, no. 24, 2022, doi: 10.3390/math10244733.

[7] K. Nagaraj, S. GS, and A. Sridhar, "Encrypting and Preserving Sensitive Attributes in Customer Churn Data Using Novel Dragonfly Based Pseudonymizer Approach," *Information*, vol. 10, no. 9, 2019, doi: 10.3390/info10090274.

[8] V. Pal, "Hybrid Quantum-Classical Machine Learning Architectures for Accelerated Drug Discover," *Int. J. Adv. Res. Sci. Commun. Technol.*, vol. 6, no. 2, pp. 1641–1653, 2021, doi: 10.48175/IJARSCT-6582M.

[9] S. Thangavel, S. Srinivasan, S. B. V. Naga, and K. Narukulla, "Distributed Machine Learning for Big Data Analytics: Challenges, Architectures, and Optimizations," *Int. J. Artif. Intell. Data Sci. Mach. Learn.*, vol. 4, no. 3, pp. 18–30, Oct. 2023, doi: 10.63282/3050-9262.IJAIDSML-V4I3P103.

[10] V. Verma, "LSTM-based to Predicting Stock Market Trends Using Machine Learning and Sentiment Analysis," vol. 8, no. 3, pp. 102–111, 2021.

[11] A. H. M, B. Thyagaraj, S. Tanisha, and S. C. B, "Customer Churn Prediction Using Synthetic Minority Oversampling Technique," in *2023 4th International Conference on Communication, Computing and Industry 6.0 (C216)*, IEEE, Dec. 2023, pp. 01–05. doi: 10.1109/C2I659362.2023.10430989.

[12] S. Sharma and N. Desai, "Identifying Customer Churn Patterns Using Machine Learning Predictive Analysis," in *2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*, 2023, pp. 1–6. doi: 10.1109/SMARTGENCON60755.2023.10442651.

[13] W. Abbas, M. Usman, and U. Qamar, "Churn Prediction of Customers in a Retail Business using Exploratory Data Analysis," in *Proceedings - 2022 International Conference on Frontiers of Information Technology, FIT 2022*, 2022. doi: 10.1109/FIT57066.2022.00033.

[14] Y. Kavyarshitha, V. Sandhya, and M. Deepika, "Churn Prediction in Banking using ML with ANN," in *Proceedings - 2022 6th International Conference on Intelligent Computing and Control Systems, ICICCS 2022*, 2022. doi: 10.1109/ICICCS53718.2022.9788456.

[15] J. Nagaraju and J. Vijaya, "Methodologies used for Customer Churn Detection in Customer Relationship Management," in *Proceedings of International Conference on Technological Advancements and Innovations, ICTAI 2021*, 2021. doi: 10.1109/ICTAI53825.2021.9673382.

[16] C. He, C. H. Q. Ding, S. Chen, and B. Luo, "Intelligent Machine Learning System for Predicting Customer Churn," in *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, 2021. doi: 10.1109/ICTAI52525.2021.00085.

[17] U. A. Korat and A. Alimohammad, "A Reconfigurable Hardware Architecture for Principal Component Analysis," *Circuits, Syst. Signal Process.*, vol. 38, no. 5, pp. 2097–2113, May 2019, doi: 10.1007/s00034-018-0953-y.

[18] J. Thomas, K. V. Vedi, and S. Gupta, "Enhancing Supply Chain Resilience Through Cloud-Based SCM and Advanced Machine Learning: A Case Study of Logistics," *J. Emerg. Technol. Innov. Res.*, vol. 8, no. 9, 2021.

[19] T. Xu, Y. Ma, and K. Kim, "Telecom Churn Prediction System Based on Ensemble Learning Using Feature Grouping," *Appl. Sci.*, vol. 11, no. 11, p. 4742, May 2021, doi: 10.3390/app11114742.

[20] V. R. Sonawane, A. S. Thorat, J. R. Suryawanshi, R. G. Dabhade, M. Patil, and B. Musmade, "Churn Identification and Prediction from a Large-Scale Telecommunication Dataset Using NLP," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 11, no. 7, pp. 39–46, Sep. 2023, doi: 10.17762/ijritcc.v11i7.7828.

[21] R. Krishnan, C. V Krishnaveni, and A. V. K. Prasad, "Telecom Churn Prediction using Machine Learning," *World J. Adv. Eng. Technol. Sci.*, vol. 7, no. 2, pp. 087–096, 2022.