JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Understanding Customer Attrition: A Machine Learning Perspective for Business Sustainability

[1]**Nithyashree T,**
Department of Computer Science and Engineering,
Sri Ramakrishna Engineering College, Coimbatore, India

[2] **Fiza S,** [3]**Abinayasri S,**
[4]Aravind C
Department of Computer Science and Engineering,
Sri Ramakrishna Engineering College, Coimbatore, India

Abstract*— Telecom operators face a big problem with customer churn, or attrition, which lowers customer satisfaction, reduces market share, and results in revenue loss. For telecom businesses to apply efficient retention tactics and keep a competitive edge in the ever-changing telecom market, it is critical to comprehend the root causes and trends of customer attrition. In an effort to improve company sustainability in the telecom sector, this study provides a thorough investigation of telecom customer attrition from a machine learning perspective. The study uses cutting-edge machine learning techniques applied to telecom data to examine the trends and causes of customer turnover. By using predictive modelling, we may determine the main causes impacting attrition and gain insights into the dynamics of client churn. We create reliable predictive models that can precisely predict customer attrition by conducting a methodical assessment of machine learning methods such as gradient boosting, random forests, and decision trees. This study advances predictive analytics in the telecom industry by using machine learning to evaluate telecom customer attrition. It also offers useful implications for the expansion and sustainability of businesses.

Keywords: Telecom, Customer Attrition, Customer Churn, Machine Learning, Predictive Analytics, Business Sustainability.

I. INTRODUCTION

The problem of customer attrition, in which users stop using services, which reduces revenue and market share, is a major issue in the telecommunications sector. For telecom operators hoping to preserve business sustainability, comprehending, anticipating, and resolving customer attrition becomes more and more important as technology advances and competition heats up. Our study explores this problem by using cutting-edge machine learning methods to identify the underlying causes and trends of customer attrition. By employing an extensive dataset that includes customer demographic data, consumption metrics (like talk time and charges), and customer contact indicators (like service calls), our goal is to extract meaningful insights that can guide successful retention tactics.

In recent years, the proliferation of competitor solutions, combined with changing customer expectations, has made it more difficult to retain subscribers in the telecommunications business. As a result, understanding the root causes and patterns of customer attrition has become critical for telecom operators looking to preserve a competitive advantage and assure corporate sustainability.

By employing feature selection and engineering techniques, we extract relevant features that capture the essence of customer churn dynamics. The core of the project involves the application of various machine learning algorithms, including logistic regression, decision trees, random forests, and gradient boosting machines, to construct predictive models. We aim to understand the nuances of customer attrition dynamics by examining the complicated interactions between these factors and giving telecom operators with invaluable insights to proactively address churn and foster long-term customer loyalty.

Furthermore, our research extends beyond analysis to real-world application, as we incorporate our predictive modelling efforts' results into an intuitive user interface. Using the Flask framework, we provide a streamlined user interface that gives telecom operators instant access to data on customer loss probability. This solution fills the knowledge gap between predictive analytics and real-world implementation, giving telecom operators a strong tool to predict customer attrition and take proactive steps to keep customers.

II. LITERATURE REVIEW

The use of machine learning approaches for predicting customer attrition in the telecom industry is examined in this article. The study seeks to effectively predict customer turnover by utilising four different machine learning classifiers and applying the Synthetic Minority Oversampling Technique (SMOTE) to overcome imbalanced dataset concerns [1]. M. Galal, S. Rady, and M. Aref's paper focuses on predicting client turnover in digital banking platforms. They present a classifier-based model for consumer profile data and compare various supervised classification methods such as KNN, Logistic Regression, AdaBoost, Gradient Boosting, and Random Forest [2]

The paper by Xiaowei Zhang and Juanqiong Gou explores the relationship between customer emotions and service purchases. Using customer complaints from an information service enterprise, the study establishes a warning model for customer churn based on emotions. Through data analysis techniques like data mining, the research aims to predict customer churn by identifying emotional indicators in complaints [4]. By defining churn customers, selecting relevant features, and training SVM, Adaboost, RandomForest, and Xgboost models, they aim to identify customers at risk of churn. The study aims to assist airlines in adopting personalized marketing strategies to maximize profits, address customer churn, maintain market share, and boost profitability [6].

III. PROPOSED SYSTEM

In order to predict telecom customer churn, the proposed system for this research project includes a thorough pipeline that includes data cleaning and preprocessing, exploratory data analysis (EDA), feature selection, model training and evaluation, and model deployment as a Flask web application. First, to guarantee data quality and consistency, the dataset which includes vital customer data including account duration, international plan status, voicemail messages, call statistics, and churn labels is carefully cleaned and preprocessed. Subsequently, feature selection attempts are guided by the insights obtained from the dataset's features and distributions through the application of EDA approaches. he most pertinent predictors for churn prediction are then found using feature selection techniques, such as assessing the impact of various attributes on the predictor variable. This system uses a range of machine learning methods, such as MLP Classifier, XGBoost, Decision Tree, and Random Forest, to forecast customer attrition in the telecom sector. The dataset is used to train and assess these algorithms in order to find out how well they predict churn.

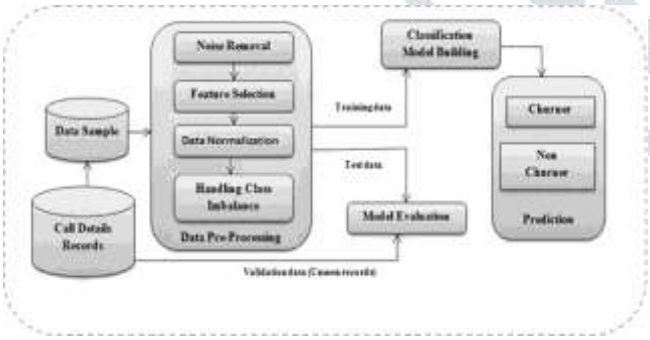

Fig 1: System Architecture

SYSTEM SPECIFICATIONS

2.1 SOFTWARE REQUIREMENTS

- Language: Python
- Libraries: pandas, NumPy, scikit-learn, XGBoost, seaborn, MGD_outliers, matplotlib
- Flask web framework
- HTML/CSS web interface
- Jupyter Notebook

IV. WORKING PROCESS

Procedure – Data collection

The first stage in the process is to obtain the Telecom Customer Churn dataset, which is the main source of data for the study. This dataset comes from the internal database of the telecom firm that is the subject of the investigation, or from a reliable third-party source. The dataset includes all of the pertinent customer data, including call statistics, churn labels, voicemail messages, international plan status, and account length. To guarantee the accuracy of the data for further analysis, it is essential to confirm the integrity and completeness of the dataset.

Procedure – Preprocessing the dataset

In order to achieve equitable representation of both churn and non-churn cases, we correct inherent imbalances in the dataset during the preprocessing step. This helps to prevent model bias and improve prediction accuracy. Unbalanced datasets can cause skewed model performance, when the model shows a bias towards predicting the majority class, leading to worse than ideal performance when it comes to identifying occurrences of the minority class, like churn cases. To accomplish this balance, methods like as resampling and Synthetic Minority Oversampling Technique (SMOTE) are used, either by duplicating instances of the minority class or by creating synthetic samples that correspond to the majority class. Preprocessing also includes data transformation and cleaning to improve data quality and guarantee feature consistency. This includes addressing outliers, inconsistencies, and missing values, all of which, if ignored, can negatively impact the performance of the model.

(EDA) is essential for comprehending the properties of the dataset and providing guidance for further modelling decisions. We can learn more about feature distributions, spot outliers, and investigate connections with the goal variable, churn, by using EDA. While bivariate analysis examines correlations between pairs of features and their link with churn, univariate analysis assists in identifying outliers and anomalies among individual features. This analysis helps with feature and model selection and offers insightful information about the data structure. Recursive Feature Elimination (RFE), which iteratively chooses the most instructive features for model training, is one feature selection technique used in preprocessing. We can improve prediction accuracy and lower the chance of overfitting by choosing the most pertinent characteristics, which will improve the model's capacity to generalise to new data.

Procedure – Model building and evaluation

The process of generating dependable predictive models for telecom customer churn prediction entails a number of crucial elements in the model building process. First, a collection of machine learning algorithms is chosen according to how well they fit the dataset's properties and the challenge at hand. A few of these algorithms are XGBoost, Decision Tree, Random Forest, K-Nearest Neighbours (KNN), Gradient Boosting, Support Vector Classifier (SVC), and Logistic Regression.

Subsequently, the chosen algorithms are implemented, creating a lexicon of models for assessment. The performance of each model is then determined by training and evaluating it using a cross-validation methodology. In particular, a 5-fold cross-validation technique is used,

in which each model is trained on four subsets (or folds) of the dataset and tested on the remaining fraction. Every subset serves as the test set once during the five repetitions of this process.

Because it strikes a compromise between recall and precision, the F1 score is employed as the evaluation metric during cross-validation to assess each model's performance. This makes it appropriate for imbalanced classification issues such as churn prediction. Furthermore, a scoring function is generated using the make_scorer function to guarantee the validity of the assessment.

For additional analysis, the cross-validation scores for every model are kept in a dictionary called cv_scores_models. The performance of each model over various dataset folds is represented by these scores. The distribution of cross-validation scores for each model is displayed in a boxplot, which allows for the visualization and comparison of the models' performances. This visualization makes it simple to compare the performance of the models and aids in determining which ones are the most promising for additional assessment. After training, the models are tested using the validation set to determine how well they perform using measures like receiver operating characteristic (ROC) curves, accuracy, precision, recall, and F1-score. This makes it possible to evaluate several machine learning methods and choose the best model or models for additional research.

In general, the process of developing a model entails employing cross-validation to systematically assess a variety of machine learning algorithms and choosing the top-performing models according to their F1 scores. By using an iterative process, the final predictive model is made to be strong, dependable, and appropriate for forecasting telecom customer attrition.

Procedure – Model deployment with Flask UI

The trained model is deployed to a web-based environment for practical use in the process of integrating the churn prediction model with a Flask UI. The Flask application loads the churn prediction model, which enables it to forecast fresh data that users submit via the user interface. Users can input customer information and receive churn forecasts through the user-friendly interface (UI) created for the model. It is possible to integrate feedback systems and visualizations to improve the user experience. The performance and dependability of the integrated system are guaranteed by extensive testing and certification. Through an easily navigable online interface, this smooth implementation enables telecom firms to leverage machine learning for customer retention initiatives.

Through the usage of Flask's lightweight web application framework, decision-makers can access the trained model via an intuitive interface. Users may input customer data and receive real-time attrition forecasts thanks to this seamless deployment, which supports proactive retention strategies and well-informed decision-making.

V. MODEL PERFORMANCE

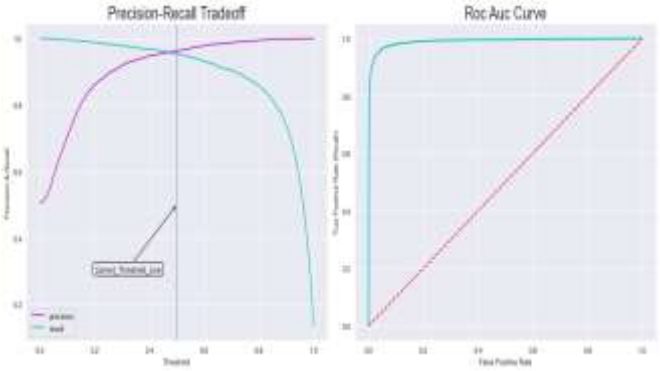

Fig 2: Performance of Random forest

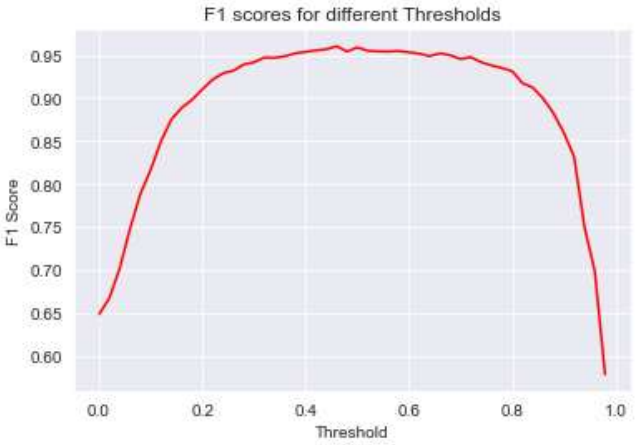

Fig 3: F1 score for Random forest

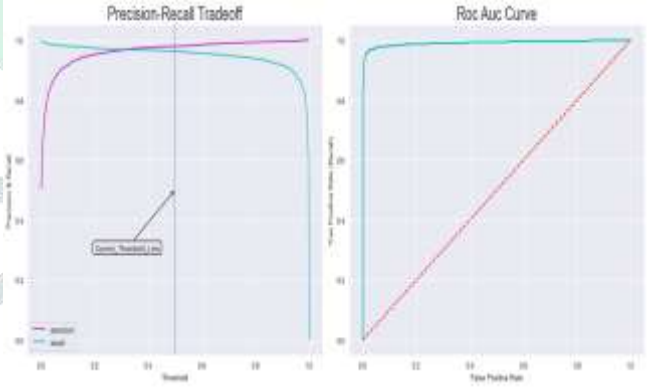

Fig 4: Performance of XGBoost

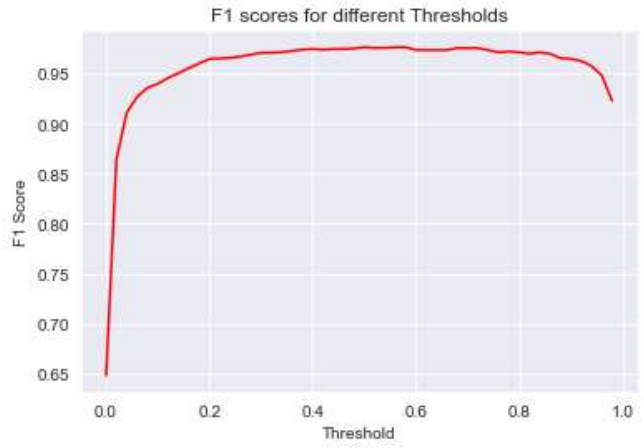

Fig 5: F1 score for XGBoost

VI. RESULT

Out of all the models that were assessed, logistic regression had comparatively lower F1 scores, which range from 0.7716 to 0.7951. Support vector classifiers (SVC) and K-nearest neighbours (KNN) both perform consistently and well; their F1 scores range from 0.8818 to 0.9163 and 0.8983 to 0.9392, respectively. The decision tree exhibits consistent performance, attaining F1 scores within the range of 0.9101 to 0.9351.

As more features are chosen, Random Forest and XGBoost continuously display higher F1 scores.Achieving higher F1 scores requires choosing the ideal number of characteristics. In this instance, choosing 10 characteristics gives both models their highest F1 scores.The higher scores across various feature selections suggest that XGBoost performs better than Random Forest in terms of F1 scores.These findings imply that the features chosen have a major influence on how well both models perform, and choosing the right features can increase the model's predictive power and accuracy of the target variable.These findings suggest that using XGBoost with 10 chosen features is advised since it regularly produces higher F1 scores..

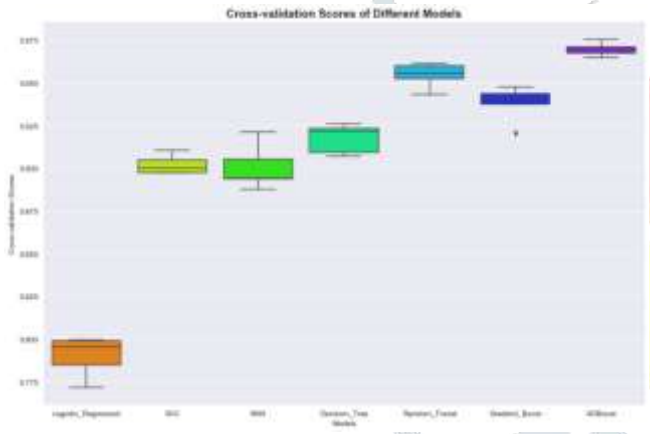

Fig 6: Cross validation scores

Fig 7: Input Interface for Churn Prediction Model

Fig 8: Prediction for Churned customer

Fig 9: Prediction for Non-Churned customer

VII. CONCLUSION

In conclusion, our project's churn prediction algorithm offers No-Churn Telecom a big chance to proactively keep clients and raise customer satisfaction levels. Through precise identification of high-risk clients, the organisation can employ tailored retention tactics and offer customised promotions to reduce customer turnover. Increased client loyalty, lower churn rates, and eventually higher business profitability are all possible outcomes of this strategy. Throughout the project, we followed the best standards in data science, which include careful feature engineering, thorough performance evaluation, rigorous data preprocessing, and judicious model selection. The aforementioned procedures emphasised the significance of comprehending organisational goals and customising analytical techniques to effectively tackle practical issues. The experiment also shown how effective XGBoost is as a strong tool for churn prediction jobs.

In the future, telecom churn prediction research will focus on data augmentation to create a larger dataset, sophisticated feature engineering to gain deeper understanding, and ensemble learning to combine models for higher accuracy. Investigating deep learning techniques such as RNNs and CNNs may improve predicted performance even more. Timely intervention is made possible by the real-time deployment of models coupled with telecom systems, and ongoing monitoring guarantees that the models remain relevant. Telecom firms are able to improve churn prediction, strengthen customer retention tactics, and cultivate long-term customer loyalty through the collection of varied data, use of sophisticated techniques, and implementation of models in production.

REFERENCES

[1] M, Aishwarya & T, Bindhiya & Tanisha, S & B, Soundarya & Shanuja, C. (2023). Customer Churn Prediction Using Synthetic Minority Oversampling Technique. 01-05. 10.1109/C2I659362.2023.10430989.

[2] M. Galal, S. Rady and M. Aref, “Enhancing Customer Churn Prediction in Digital Banking using Ensemble Modeling,” 2022 4th Novel Intelligent and Leading Emerging Sciences Conference (NILES), Giza, Egypt, 2022, pp. 21–25, doi:10.1109/NILES56402.2022.9942408.

[3] H. Karamollaoğlu, İ. Yücedağ and İ. A. Doğru, "Customer Churn Prediction Using Machine Learning Methods: A Comparative Analysis," 2021 6th International Conference on Computer Science and Engineering (UBMK), 2021, pp. 139-144, doi: 10.1109/UBMK52708.2021.9558876.

[4] Xiaowei Zhang and Juanqiong Gou, "Warning model of customer churn based on emotions," 2015 International Conference on Logistics, Informatics and Service Sciences (LISS), Barcelona, 2015, pp. 1-3, doi: 10.1109/LISS.2015.7369683.

[5] M. D. S. Rahman, M. D. S. Alam and M. D. I. Hosen, "To Predict Customer Churn By Using Different Algorithms," 2022 International Conference on Decision Aid Sciences and Applications (DASA), 2022, pp. 601-604, doi: 10.1109/DASA54658.2022.9765155.

[6] J. Ran and X. Cheng, "Airline Customer Value Analysis and Customer Churn Prediction Based on LRFMC Model and K-means Algorithm," 2021 2nd International Conference on Computer Science and Management Technology (ICCSMT), Shanghai, China, 2021, pp. 185-193, doi: 10.1109/ICCSMT54525.2021.00044.

[7] K. Kim and J. -H. Lee, "Bayesian Optimization of Customer Churn Predictive Model," 2018 Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS) and 19th International Symposium on Advanced Intelligent Systems (ISIS), Toyama, Japan, 2018, pp. 85-88, doi: 10.1109/SCIS-ISIS.2018.00024.

[8] P. Hemalatha and G. M. Amalanathan, "A Hybrid Classification Approach for Customer Churn Prediction using Supervised Learning Methods: Banking Sector," 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN), Vellore, India, 2019, pp. 1-6, doi: 10.1109/ViTECoN.2019.8899692

[9] J. Yang, "Design of E-commerce Customer Churn Prediction System Based on Data Mining Techniques," 2023 IEEE 3rd International Conference on Social Sciences and Intelligence Management (SSIM), Taichung, Taiwan, 2023, pp. 114-118, doi: 10.1109/SSIM59263.2023.10468983.

[10] F. Alhaqui, M. Elkhechafi and A. Elkhadimi, "Machine learning for telecoms: From churn prediction to customer relationship management," 2022 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT), Soyapango, El Salvador, 2022, pp. 1-5, doi: 10.1109/ICMLANT56191.2022.9996496.

[11] S. D. Kumar, K. Soundarapandiyan and S. Meera, "Comparative Study of Customer Churn Prediction Based on Data Ensemble Approach," 2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS), Chennai, India, 2023, pp. 1-10, doi: 10.1109/ICCEBS58601.2023.10449139.

[12] D. Azzam, M. Hamed, N. Kasiem, Y. Eid and W. Medhat, "Customer Churn Prediction Using Apriori Algorithm and Ensemble Learning," 2023 5th Novel Intelligent and Leading Emerging Sciences Conference (NILES), Giza, Egypt, 2023, pp. 377-381, doi: 10.1109/NILES59815.2023.10296608.