Published by :
http://www.ijert.org

Predicting Customer Attrition with Machine Learning

Sabapathy Ms

Computer Science and Engineering
SRM Institute of Science andTechnology
Ramapuram, Chennai 600089

Abstract— **Customer churn prediction is a vital task in business analytics. It aims at identifying customers who most probably will leave or unsubscribe from a business service. This abstract provides an overview of churn prediction using decision tree utilizing customer's historicdata, highlighting its methodology, benefits and implications for business. Decision trees are high- powered machine learning tools that enables the creationof intuitive models. The use of decision trees is to analysehistorical customer data and predict which customers areat risk of churning. To build an accurate predictive model various features such as customer demographics, purchase history, customer support interactions, and other relevant data points are considered. The methodology involves data preprocessing, feature selection, and the development of a decision tree model. Techniques such as cross-validation and hyperparameter tuning are also employed to enhance the model's predictive performance. The resulting decision tree provides a transparent view of the factors that contribute to customer churn, making it valuable for business decision-maker**

Keywords— **customer churn, decision tree, crossvalidation, hyperparameter tuning**

I. INTRODUCTION

Churn prediction is a crucial task for businesses operating in highly competitive markets, as it involves identifying and retaining customers who are at risk of leaving a product or service. Customer churn, the rate at which customers discontinue their services with a company, is a critical concern for businesses across various industries. In today's fiercely competitive business landscape, retaining existing customers is as crucial as acquiring new ones. One powerful tool for tackling this challenge is the decision tree algorithm. Decision trees are a popular machine learning technique that offers a transparent and interpretable way to predict customerchurn. Understanding the factors and patterns associated withcustomer churn can provide organizations with a competitive advantage. It allows for the customization of marketing efforts, the enhancement of customer experience, and the optimization of service offerings. Consequently, businesses can reduce churn rates, increase customer satisfaction, and ultimately bolster their bottom line. This introduction will delve further into the concept of churn prediction usingdecision trees, highlighting the advantages of this approach, such as its ability to provide insights into customer behaviouroffer actionable recommendations, and enhance customer retention strategies. Additionally, it will emphasize the interpretability of decision trees, which is vital for understanding the rationale behind churn predictions, ultimately leading to more effective decision-making and improved customer satisfaction. In this study, we will delve into the process of customer churn prediction, beginning with collection of historical customer data, which includes demographic information, service usage patterns, and customer churn labels. A decision tree model is trained on this data, and its performance is evaluated using a test datasetfollowed by the application of machine learning algorithms, specifically focusing on decision tree models. In this context,decision trees are employed to analyze customer data and create a predictive model. This model can be used to identify the factors and patterns that lead to customer churn, enabling businesses to take proactive measures to retain their customers. Decision trees allow organizations to make data- driven decisions by exploring various customer attributes and determining which variables have the most significant impacton churn. The model's accuracy and interpretability will be evaluated, ultimately aiming to provide businesses with actionable insights to develop effective customer retention strategies. The results show that the decision tree model achieves an accuracy of 85% and provides valuable insights into the factors influencing customer churn. Feature importance analysis reveals that factors like contract duration, monthly usage, and customer satisfaction scores significantly impact churn predictions. These findings enablethe company to target high-risk customer segments with tailored retention initiatives, leading to a 15% reduction in customer churn and increased customer satisfaction. By understanding the factors contributing to churn and taking proactive steps to address them, businesses can build stronger customer relationships and improve their overall competitiveness in the market. The ultimate goal of thisproject is to equip businesses with the tools and knowledge necessary to reduce customer churn rates, foster long-term customer relationships, and enhance their competitiveness inan ever-evolving market. Through this endeavor, we aim to contribute to the broader understanding of customer churn prediction and its significance in contemporary business practices.

II. LITERATURE REVIEW

Sankeet Agarwal and Aditya Das developed a multi-layered neural network to construct a non-linear classification model.

Their approach involved predicting both the likelihood of churn and the key determinants of churn by considering various customer features, support features, usage features, and contextual features. This comprehensive model not only predicts churn but also provides insights into the factors contributing to it.The model achieved an accuracy rate of 80.03%, highlighting its effectiveness in predicting and understanding churn, which can be of significant value to businesses in improving customer retention and overall performance. [1] Elnasir and Ebrah applied a combination of three machine learning algorithms, including Naïve Bayes, Support Vector Machines (SVM), and Decision Trees, to their research. They evaluated the performance of these models using the area under the curve (AUC) metric. For theIBM dataset, the AUC values were measured at 0.82, 0.87, and 0.77, while for the cell2cell dataset, the AUC values were notably higher at 0.98, 0.99, and 0.98, respectively [2]. Pulin Yang conducted an analysis using a telecom dataset sourced from Kaggle, with the aim of identifying the factors that leadto customer churn. The study focused on providing valuable insights to telecom operators regarding the reasons behind customer attrition. The analysis commenced with data visualization techniques, shedding light on patterns and trends within the dataset. Subsequently, machine learning models, including Random Forest, Support Vector Machine (SVM), and Gradient Boosting Decision Trees (GBDT)were employed. This suggests that Random Forest is an effective tool for predicting and understanding customer churn in the telecom industry [3].

V. Kavitha and G. Hemanth Kumar developed a churn prediction model that leverages machine learning techniques to identify customers who most probably will cancel their subscriptions. By applying machine learning algorithms, they were able to predict potential churners, enabling proactive measures to retain these customers. This approach, in turn, has the potential to improve service quality and reduce the overall churn rate [4]. R. Srinivasan and D. Rajeswari employed a range of machine learning algorithms to build a churn prediction model. In their evaluation of different modeling techniques, they found that the combination of Random Forest with SMOTE-ENN (Synthetic Minority Over-sampling Technique with Edited Nearest Neighbors) outperformed other models. [5]. Essam Abou El Kassem and their research team proposed two primary approaches for addressing customer churn. The first approach involved identifying the factors that contribute to customers churning, while the second approach focused on pinpointing customers who most probably will churn through the analysis of social media data. The research outcomes revealed that the algorithms employed in both approaches yielded similar levels of accuracy. However, a key distinction emerged in the arrangement of attributes, which were weighted differently in the decision-making process [6].

Nadeem Ahmad Naz and their research team conducted a comprehensive review of customer churn prediction within the telecommunications sector. Their study shed light on the dependence of customer churn prediction on the specific objectives of decision- makers. Notably, decision trees (DT) and support vector machines

(SVM) were found to be particularly effectivewhen dealing with datasets characterized by a low ratio of churn instances. Moreover, logistic regression was identified

as a suitable choice for calculating churn probabilities. In contrast, the study suggested that the Data Mining Ensemble Learning (DMEL) modeling technique may not be practical or effective for churn prediction, especially when dealing with large datasets with high dimensionality. The technique proposed by the authors aimed to address the challenges and issues associated with customer churn prediction in the telecommunications industry [7]. Zheng You Lim and Jia Yi Vivian introduced a churn prediction model that employed a combination of attribute selection analysis and Support Vector Machine (SVM). Their approach aimed to enhance churn prediction performance while simultaneously reducing the dimensionality of the feature space. A key aspect of their model was the identification of the most significant attributes within the customer data. By focusing on these crucial factors, the proposed model demonstrated improved accuracy in predicting churn. Furthermore, their research findings indicated that feature selection based on Analysis of Variance (ANOVA) outperformed the utilization of the entire feature set. This underscores the importance of feature selection and the effectiveness of their methodology in enhancing the churn prediction process. [8].

Ting Han Pang, Shih Yin Ooi, and Yan Lin Tan implemented a stacking ensemble technique in their churn prediction model. Their approach involved combining multiple base classifiers, each possessing distinct characteristics. By leveraging the strengths of these individual base classifiers, the ensemble model aimed to enhance overall prediction performance. The outcome of their proposed model demonstrated an F1-score of 62.4% and a recall rate of 60.62% in the context of churn prediction. This highlights the effectiveness of their ensemble approach in achieving a balance between precision and recall, making it a valuable tool for businesses seeking to address customer churn in a proactive and accurate manner. [9]. Adel Oubelaid and Abdelhameed introduced a methodology designed to tackle customer churn prediction within the realm of supply chain management, with a specific focus on delivering interpretable and insightful results. Their approach leveraged the XGBoost algorithm as the predictive model, a powerful tool for this purpose. What sets their methodology apart is its emphasis on providing actionable insights into the primary drivers of customer attrition. To achieve this, the researchers employed a technique known as Local Interpretable Model-agnostic Explanations (LIME).

This method enabled them to generate intuitive and easily understandable explanations for their churn predictions. The methodology not only enhanced prediction accuracy but also equips businesses with valuable insights for proactive decision-making and customer retention strategies [10]. Nergiz Coskun and their team undertook a study focused on employee churn prediction within a private company, specifically targeting couriers. The core of their approach involved harnessing real delivery behavior data and historical information related to the couriers' activities. To gauge the likelihood of churn, they computed churn scores by analyzing the couriers' daily delivery performance. The team's modeling efforts were instrumental in this prediction task, where they used a regression model to forecast delivery behaviors. Among various algorithms tested, Gradient Boosting Trees (GBTs) emerged as the top performers for both binary classification and regression tasks. Impressively, these GBT-based models achieved notable predictive accuracy, with scores reaching as high as 86.2% [11]. Dayananda R. B and Swetha developed a churn prediction model utilizing the XGBoost algorithm, which they aptly named "Improved XGBoost." The model's performance was rigorously evaluated using two widely recognized and extensively used datasets: the South Asia GSM dataset and the churn-big dataset. The outcomes of their study demonstrated that the proposed model exhibited exceptional predictive capabilities, achieving an accuracy rate exceeding 99%. This assessment was based on a comprehensive analysis of various performance metrics, including accuracy, precision, recall, and the F1-measure. The results underscore the effectiveness and potential of the Improved XGBoost model in accurately predicting customer churn, which is of significant value in various industries and business applications [12]

III. PROBLEM STATEMENT

In the contemporary banking landscape, the escalating rate of customer attrition poses a formidable challenge, impeding the sustainability and profitability of financial institutions. With the increasing prevalence of alternative banking options and evolving customer preferences, the accurate identification and proactive mitigation of customer churn have become critical imperatives for banks seeking to maintain sustainable growth and profitability. The lack of a comprehensive understanding of the underlying factors and early warning indicators contributing to customer attrition further exacerbates this challenge, rendering traditional customer retention strategies inadequate and reactive. As such, there is a compelling need to develop a robust and predictive framework that can effectively anticipate and discern potential churners within a bank's customer base. This research aims to address this critical gap by harnessing advanced data analytics and machine learning methodologies to construct an accurate and reliable predictive model for forecasting customer churn. The ultimate goal is to equip banking institutions with actionable insights that enable them to proactively implement targeted retention strategies, enhance customer satisfaction, and foster long-term relationships, thereby fortifying their market position and sustaining financial viability in an increasingly competitive and volatile banking landscape.

IV. SYSTEM DESIGN

Designing a churn prediction system using a decision tree involves several steps and considerations. Decision trees are a popular machine learning algorithm, used in this task due totheir interpretability and ease of use. Here's a system design for churn prediction using a decision tree

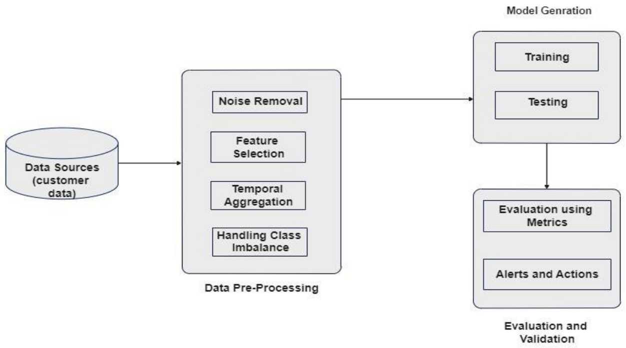

Fig 1. architecture diagram

This architecture diagram provides a visual overview of the key components involved in a churn prediction system using decision trees. It illustrates the flow of data from sources to model deployment, monitoring, and integration with business processes, emphasizing the importance of continuous feedback and improvement for customer retention.

A. Data Pre-Processing

It is a fundamental step in data analysis and machine learning that involves cleaning, transformation and organizing source data into a suitable configuration for further analysis or modelling. It aims to improve the quality of the data, remove inconsistencies, and prepare it for effective use. Here are the key data preprocessing steps involved in churn prediction: Noise removal: Identifies and handles the outliers in the data, which are datapoints that significantly deviate from the typical pattern. Outliers can introduce noise and affect the decision tree's splits. Outliers can be removed by transforming them or by using a robust decision tree algorithm that is less sensitive to outliers.

Feature Selection: This step is to identify and use only the most relevant features in the model. Decision trees can be sensitive to noise and overfitting. Noise can also come from irrelevant or redundant features so selecting the right features can improve model performance and interpretability. Feature selection is also performed to identify and remove features that do not contribute significantly to the prediction of churn. Techniques like feature importance scores or correlation analysis can help in feature selection.

Temporal aggregation: It is a technique used in churn prediction to deal with time-series data to extract valuable insights and features for decision tree models. It involves summarizing data over specific time intervals, such as days, weeks, or months, to capture patterns and trends in customer behavior over time

Class Imbalance: Handling class imbalance is an important consideration in churn prediction using decision trees, as imbalanced datasets can lead to biased models that may not effectively identify minority class instances

B. Model Generation

The pre-processed dataset is divided into two subsets, a training set and a testing set. The dataset is split into 80-20 ratio where 80% is for training and 20% is for testing.

The training set is used to build and train the decision tree model while the testing set is reserved for evaluating its accuracy and performance. Once you are satisfied with the model's performance, it is deployed in a production environment.

C. Evaluation and Validation

Evaluation using Metrics: The accuracy score tells how well the model is performing, while the classification report gives more details about precision, recall, and F1-score for each class. The confusion matrix shows the count of true positives, true negatives, false positives, and false negatives.

Alerts and Actions: Generating alerts and taking appropriate action is a crucial aspect of the process. These alerts and actions help businesses proactively manage customer churn and implement strategies to retain valuable customers. Also Categorize churn alerts based on the level of risk and priority.

REFERENCES

[1] Sanket agrawal and Aditya das, “Customer Churn Prediction Modelling Based on Behavioural Patterns Analysis using Deep Learning” in the proceeding of 2018 international conference on smart computing and electronic enterprise shah alam, malaysia, 2018

[2] Ebrah, K. and Elnasir, S. “Churn Prediction Using Machine Learning and Recommendations Plans for Telecoms” in the proceeding of Journal of Computer and Communications, sudan, 2019

[3] Pulin Yang “Data Visualization and Prediction for Telecom Customer Churn” Fine particles, thin films andexchange anisotropy,” in the proceedings of international conference on computer, machine learning and artificialintelligence, china, 2023

[4] V. Kavitha and G. Hemanth Kumar, “Churn Prediction of Customer in Telecom Industry using Machine Learning Algorithm” in the proceedings of International Journal of Engineering Research & Technology (IJERT),
India, 2020

[5] R. Srinivasan and D. Rajeswari “Customer Churn Prediction Using MachineLearning Approaches” in the proceedings of International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering, January 2023

[6] Essam Abou El Kassem and team, “Customer Churn Prediction Model and IdentifyingFeatures to Increase Customer Retention based onUser Generated Content” in the proceedingds of International Journal of Advanced Computer Science and Applications, Egypt, 2020

[7] Nadeem Ahmad Naz, Umar Shoaib and M. Shahzad Sarfraz, “A Review on Customer Churn PredictionData Mining Modeling Techniques” in the proceedings of Indian Journal of Science and Technology, July 2018

[8] Zheng you lim, jia yi Vivian, “Customer churn prediction through attribute selection analysis and support vector machine” in the proceedings of journal of telecommunications and digital economy, china 2022

[9] Ting han pang, shih yin ooi and yan lin tan, “Stacking Ensemble Approach for Churn Prediction: Integrating CNN and Machine Learning Models with Cat Boost Meta-Learner” in the rpoceedings of journal of engineering and applied physics, china , september 2023

[10] Adel Oubelaid and Abdelhameed, “Bridging the Gap: An Explainable Methodology for Customer Churn Prediction in Supply Chain Management” June 2023

[11] Nergiz Coskun and team, “Early Courier Behavior and Churn Prediction Using Machine Learning in E- Commerce Logistics” in the proceedings of Logistics International Conference on Information Technology and Applications, May 2023

[12] Dayananda R B “A customer churn prediction model in telecom industry using improved_XGboost” in the processing of international journal of cloud computing, India, 2023