Enhancing Customer Retention: A Machine Learning Approach for Churn Prediction

DR. MARRYNAL S EASTAFF[1], A. SATHIYA PRIYA[2], R. RANJITHA[3]
[1]*Associate Professor, Department of Computer Science with Cyber Security, Hindusthan College of Arts & Science, Coimbatore, India*
[2]*Assistant Professor, Department of Information Technology, Dr.N.G.P. Arts and Science College, Coimbatore, India.*
[3]*Hindusthan College of Arts & Science, Coimbatore, India*

Abstract— Using a diverse array of machine learning techniques, we conducted churn prediction and classification on Orange Telecom's Churn Dataset. This dataset comprises a churn label indicating whether a customer terminated their subscription, along with preprocessed customer activity data. For the project, the larger churn-80 dataset was utilized for training and cross-validation, while the smaller churn-20 dataset was reserved for final testing and performance evaluation. To predict potential customer churn, we undertook an in-depth analysis of the dataset, emphasizing the identification of user personas and feature significance. Parameter optimization was carried out using Grid Search across several classifiers, including Gradient Boosting models, Ensemble Trees, Decision Tree Classifiers, and Logistic Regression. Among these, the XGBoost Classifier achieved the highest ROC, while the LightGBM Classifier excelled in validation set performance.

Indexed Terms- Gradient Boosting models, Ensemble Trees, Decision Tree Classifiers, Logistic Regression.

I. INTRODUCTION

In today's dynamic Information Technology (IT) industry, businesses face the ongoing challenge of staying competitive while meeting the diverse needs of their customers. One powerful strategy to gain an edge in this competitive landscape is customer segmentation. By leveraging advanced data mining techniques, IT organizations can divide their extensive customer base into distinct segments, each defined by unique characteristics, behaviours and preferences. This analytical approach enables companies to gain deeper insights into customer needs optimize resource allocation, and design targeted strategies for customer acquisition and retention. As a result customer segmentation not only enhances operational efficiency but also fosters stronger customer relationships driving sustained growth and innovation in a highly competitive market.

1.1 Association Rule Mining

Association rule mining is a powerful data mining technique widely applied across various industries, including retail, market basket analysis, recommendation systems, and more, to uncover meaningful and insightful relationships within datasets. By identifying patterns and dependencies within the data, this method empowers businesses to make informed decisions and enhance their operations. At its core, association rule mining involves detecting frequent item sets and generating rules that explain the co-occurrence of items within a transactional or categorical dataset. These rules are invaluable for extracting hidden correlations that are not immediately apparent through simple observation. The primary goal of association rule mining is to uncover relationships or connections between items that occur together more frequently than random chance would suggest. This technique is a critical tool for deriving actionable insights from large and complex datasets, enabling organizations to optimize strategies and improve efficiency.

1.2 Customer Segmentation

Customer segmentation is a strategic approach used by businesses and organizations to divide their customer base into distinct groups based on shared characteristics, behaviors, and preferences. This data-driven methodology enables businesses to better understand their diverse audience, allowing them to create tailored products, services, and marketing strategies that address the specific needs of each

segment. By recognizing that no two customers are the same, businesses can enhance operational efficiency, improve customer satisfaction, and maintain a competitive edge in the market. Customer segmentation serves as a blueprint for targeted marketing campaigns, enabling organizations to allocate resources effectively and foster personalized, long-lasting customer relationships. Factors such as demographics, purchase history, geographic location, and psychographics are often used to define these segments. By leveraging these insights, companies can refine their customer service programs, optimize product offerings, and craft compelling marketing messages. Ultimately, this leads to stronger brand loyalty, higher retention rates, and greater overall business success.

1.3 Market Analysis

Market analysis is a vital process for making informed business decisions. It involves an in-depth examination of a specific market or industry to understand its potential, trends, and dynamics. By employing this analytical approach, businesses can gain valuable insights into market size, growth opportunities, competition, and consumer behaviour. These insights enable organizations to develop strategies, make sound investments, and adapt to constantly evolving market conditions. In today's fast-paced, technology-driven world, where consumer preferences are ever-changing, market analysis is an essential tool for businesses striving to stay competitive and relevant on a global scale.

Market analysis encompasses a range of activities, such as evaluating competitive landscapes, identifying potential risks and opportunities, and collecting and interpreting data on customer preferences and market demographics. The insights derived from this process serve as the foundation for critical business decisions, including market entry, product development, pricing strategies, and the launch of targeted marketing campaigns. By harnessing the power of market analysis, organizations can align their strategies with market demands, seize opportunities, and maintain their relevance in an ever-changing business environment.

II. LITERATURE REVIEW

Targeted marketing strategy [1] is a hot topic that has drawn a lot of interest from academics and industry, as suggested by Fahed Joseph et al. in this research. A popular method for examining the diversity of consumer purchasing behaviour and profitability is market segmentation. It is noteworthy that traditional models of market segmentation used in the retail sector are mostly descriptive in nature, do not provide adequate market insights, and frequently do not identify small enough categories. In order to process large amounts of data, this study also makes use of the dynamics present in the Hadoop distributed file system. Expectation-Maximization (EM) and K-Means++ clustering algorithms were used in three separate market segmentation studies utilizing modified best fit regression. The results were evaluated using cluster quality evaluation. The study's findings are as follows: (i) each consumer Lifetime Value (CLTV) segment's insight into consumer buying behaviour (ii) the clustering algorithm's performance in creating precise market segmentation. Based on the data, the average customer lifetime was found to be just two years, with a 52% churn rate. As a result, a marketing plan was created based on these findings and applied to department store sales.

In this study, [2] E. Ernawati et al. proposed that data mining (DM) is the process of knowledge extraction from data. Companies can utilize the information gathered from customer behaviour segmentation to help them define their target market and create a marketing plan. The most widely used behaviour segmentation model is the Regency Frequency Monetary (RFM) model. The RFM model works in tandem with DM in numerous customer-segmentation studies across multiple application domains. With so many techniques available in DM, choosing the right ones might help uncover insightful hidden patterns in client segments. The purpose of this research is to analyse and synthesize DM techniques that operate with the RFM model in order to provide a framework for consumer segmentation. A thorough examination of the literature covering the years 2015–2020 is used in this study. Among the seven DM techniques examined, grouping and visualization are the most often applied techniques. This study proposes a new framework for combining DM approaches with the

RFM based segmentation in the Geographic Information Systems (GIS) environment because of the expanded visualization function and the requirement for customers' geo-demographic data to be taken into consideration in the analysis.

In this research, SHULI WU et al. [3] argue that since keeping current customers is less expensive than acquiring new ones, recruiting new ones is no longer a wise business strategy in the telco sector. In the telecom sector, churn management becomes crucial. This paper attempts to present an integrated customer analytics framework for churn management, since there is a lack of research combining customer segmentation and churn prediction. The framework consists of six parts: factor analysis, churn prediction, customer segmentation, customer behaviour analytics, exploratory data analysis (EDA), and data pre-processing. In order to give telecom operators a comprehensive churn analysis and help them better manage customer attrition, this system combines churn prediction with the customer segmentation process. The tests using six machine learning classifiers use three datasets. First, several machine learning classifiers are used to forecast the customers' churn state. To address the issues with imbalanced datasets, the training set is subjected to the Synthetic Minority Oversampling Technique (SMOTE). The models are evaluated using the 10-fold cross-validation.

In this paper, Saumendra Das et al.[4] suggest that consumers are becoming more aware, knowledgeable, and involved in society. They adapt their habits and preferences to suit their demands. Understanding consumer demands is a critical component of marketing, since it allows a business to identify its most devoted clientele amidst this diversity. Customer segmentation is the idea of breaking heterogeneity up into homogeneous forms. A crucial component of marketing is customer segmentation, which enables businesses to manage a vast amount of customer data in an orderly way while fostering relationships with consumers. Comprehending the concealed knowledge of the client is a clever application of computational analysis, whereby precise data can be tailored to the client's preferences and tastes. We call this kind of computational analysis "data mining." This article examined consumer segmentation using data mining approaches in a methodical manner. It is an organized analysis of several segmentation-related data mining approaches, including supervised and unsupervised methods.

In this work, Angel Martín et al. [5] argue that location and navigation services based on global navigation satellite systems (GNSS) are necessary for high-precision positioning applications in real-time in significant economic sectors like mapping, civil engineering, precision agriculture, and transportation. Since the 1990s, the number of people using GNSS networks for real-time navigation has grown dramatically worldwide and has beyond earlier projections. Therefore, market segment trends can be examined by tracking the evolution of GNSS network users. In order to put this idea into practice, large amounts of navigation data must be processed over several years, and clients must be continuously monitored. The efficient management of large-scale GNSS user connections is the main goal of this research, which aims to gather statistics and analysis. Using big data architecture and data mining techniques for data analytics has been found to be the most effective method of testing the hypothesis. The findings show the dynamics of users across various market groups and the rising demand over time.

Retail marketers are always trying to find methods to make their campaigns more effective. Targeting customers with incentives that are most likely to draw them return to the business and spend more money and time there is one strategy to do this. One method of market segmentation is demographic market segmentation. A business creates groups within the broader market according to a number of predetermined standards. Among the often-used demographic segmentation factors are age, gender, marital status, occupation, education, and income. To elucidate the principle of segmentation applied to a Turkish supermarket chain, a sample case study has been conducted. Determining product and shopping habit reliance is the aim of this case study. Additionally, projected sales inform product and customer profile advertising.

IV. PROPOSED FRAMEWORK FOR CHURN PREDICTION USING MACHINE LEARNING

The proposed framework aims to build an effective churn prediction system for Orange Telecom by leveraging machine learning techniques. The approach involves multiple stages, including data loading, pre-processing classification, training, testing, and performance evaluation. The system is designed to handle challenges such as class imbalance and outliers while delivering actionable insights to reduce customer attrition. Below is a detailed breakdown of the framework, along with a diagram for clarity:

4.1 Load Data

This module initializes the process by loading the Orange Telecom Churn Dataset, which includes churn labels indicating subscription cancellations and processed customer activity data. This dataset serves as the foundation for model training, evaluation, and analysis.

4.2 Data Pre-Processing

This module focuses on visualizing the dataset to gain insights into its structure and distribution:

Feature Distribution: Most features exhibit a normal distribution, while some categorical variables display bimodal peaks.
Class Imbalance: The target variable (churn vs. non-churn) is imbalanced, which may impact model performance.
Outliers: Despite the presence of outliers, they are retained due to the dataset's manageable size.
To address these challenges, cross-validation is employed during model fitting to ensure robust predictions.

4.3 Classification – Predicting Potential Churn

This module implements classification techniques to predict customer churn. A range of machine learning models is explored, including CatBoost Classifier, Random Forest Classifier, XGBoost Classifier, Decision Tree Classifier, Logistic Regression. Dimensionality reduction techniques are also considered for visualizing and understanding the dataset.

4.4 Training and Testing

The classification models are trained on the churn-80 dataset and tested on the churn-20 dataset. This step is critical for:

Assessing Model Performance: Ensuring each model's ability to generalize on unseen data.
Selecting the Best Classifier: Identifying the model that delivers the most accurate churn predictions.

4.5 Performance Evaluation Using ROC AUC Metric

The ROC AUC metric is used to evaluate the classification models:

LGBM Classifier achieves the highest ROC AUC of 0.93 on the validation set.
On the test set, XGBoost Classifier achieves the best ROC AUC of 0.95, demonstrating its superior ability to predict customer churn accurately.
By leveraging these insights, Orange Telecom can enhance its customer retention strategies and effectively minimize attrition rates.

V. RESULT ANALYSIS

The result analysis of the Orange Telecom churn prediction project highlights the effectiveness of various machine learning models in forecasting potential customer attrition. Among the evaluated models, the LGBM Classifier emerged as the top performer, achieving an impressive ROC AUC score of 0.93 on the validation set. Other classifiers, including CatBoost, Random Forest, Decision Tree, XGBoost, and Logistic Regression, demonstrated competitive performance on the test set, with ROC AUC scores ranging from 0.867 to 0.955. These results underscore the robustness and precision of the proposed churn prediction system in identifying likely churners. Additionally, the analysis of feature importance and model performance provides Orange Telecom with valuable insights, enabling the development of targeted strategies to enhance customer retention and minimize attrition. By leveraging these findings, the company can make data-driven decisions to improve operational efficiency and strengthen customer relationships.

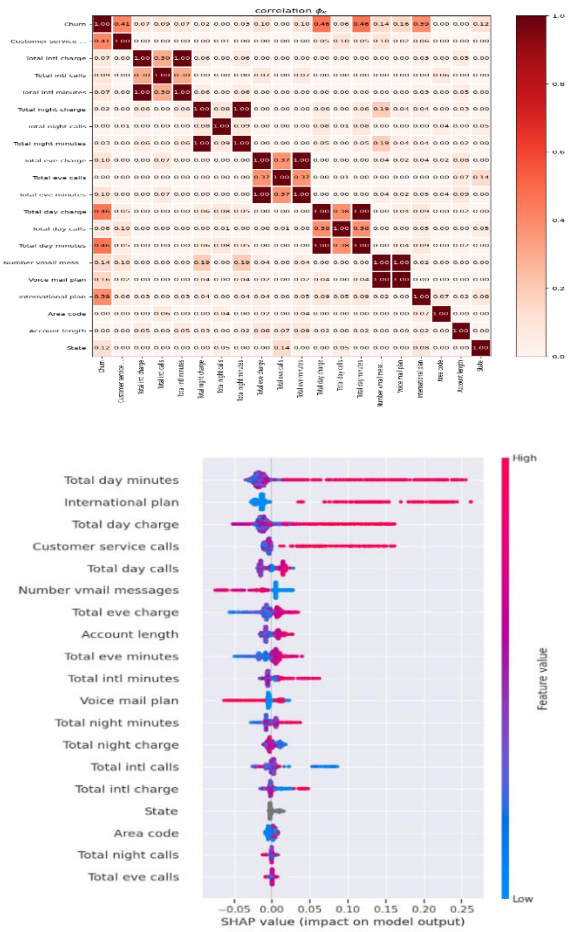

Figure 1. SHAP values

CONCLUSION

In conclusion, Orange Telecom's dataset was used for the churn prediction study, which shows how successful machine learning approaches are in predicting possible customer turnover. The project successfully determined that the LGBM Classifier was the top-performing model, achieving a ROC AUC of 0.93 on the validation set, and that the XGB Classifier had the best ROC AUC of 0.954 through extensive data analysis, model training, and evaluation. Further confirming their usefulness in churn prediction tasks were the encouraging outcomes that the chosen models—Logistic Regression, Decision Tree Classifier, Random Forest Classifier, XGB Classifier, and Cat Boost Classifier—showed on the test set. These results highlight the value of using ensemble approaches and thorough model evaluation to address customer attrition and offer telecom businesses insightful information to improve their client retention tactics.

FUTURE WORK

In order to improve the churn prediction system going forward, a number of directions for further research might be investigated. First, using more sophisticated feature engineering methods and looking at different data sources might enhance the performance of the model even more. Predictive accuracy may also be improved by applying sophisticated ensemble techniques like model stacking and blending. Additionally, examining the effects of integrating sentiment analysis and real-time data streams from social media platforms may enhance the accuracy of churn prediction and offer insightful information about consumer behavior.

REFERENCES

[1] CISCO, Forecast and Methodology for the Cisco Global Cloud Index, 2016–2021. The Effects of Data Mining and Clustering Techniques on Big Data Market Segmentation

[2] M. Y and Chen. Hao, "A review of RFM-based customer segmentation using data mining techniques," IEEE J. Sel. Community Areas, vol. 36, no. 3, March 2018, pp. 587–597.

[3] Three.I. R. J. Lopez, M. Roman, and M. Mambo, "Customer Segmentation and Churn Prediction Framework for Telco Business," IEEE Communications. Vol. Mag. Jan. 2018, 78, no. 2, pp. 680–698.

[4] G. El-Sayed et al., "State-of-the-Art Review: Customer Segmentation via Data Mining Techniques," IEEE Access, vol. 6, 2018, pp. 1706–1717.

[5] Kumari, R. M. Parizi, N. Kumar, S. Tanwar, S. Tyagi, and K. For market sector applications of differential global navigation satellite system (GNSS) services, R. Choo, "Big data architecture and data mining analysis" J. Webw. Computer. Vol. Appl. 128, 2019; pages 90–104.