BANK CUSTOMER CHURN PREDICTION USING H20 AUTOML

[1]Dr. P. Ravinder Rao, [2]G. Akhil, [2]K. Pranav Goud, [2]K. Hari Kamal

Asst. Professor, Department of CSE, Anurag University, Venkatapur (V), Ghatkesar (M), Medchal(D)., T.S-500088

UG Student, Department of CSE, Anurag University, Venkatapur (V), Ghatkesar (M), Medchal(D)., T.S-500088

Abstract: Customer churn prediction is crucial for businesses aiming to retain their customer base and enhance profitability. In this study, we leverage the power of H2O AutoML, an automated machine learning framework, to develop an accurate and efficient churn prediction model. By utilizing a diverse set of algorithms, data preprocessing techniques, and hyperparameter optimization strategies, H2O AutoML to streamline the model development process while achieving optimal performance. This project focuses on predicting customer churn for a bank and telecommunications company. We begin by collecting and preprocessing customer data, including demographic information, usage patterns, and service subscriptions. Leveraging H2O AutoML, we create a pipeline that automatically explores a range of machine learning algorithms, such as random forests, gradient boosting, and neural networks, in order to identify the most suitable model for the task. The results demonstrate the effectiveness of H2O AutoML in producing high performing churn prediction models without requiring extensive manual intervention. By selecting the best-performing model, we provide the bank and telecommunications company with actionable insights to proactively address potential churn cases and implement targeted retention strategies. In conclusion, this study highlights the utility of H2O AutoML as a powerful tool for automating the customer churn prediction process. By harnessing the capabilities of automated machine learning, businesses can efficiently develop accurate churn prediction models, leading to improved customer retention rates and ultimately, enhanced business success.

Keywords: Customer Churn, Data Preprocessing, Feature Selection, Cross-Validation, Predictive Modeling

I. INTRODUCTION

Churn is substantially, more important in market places in which competition is intense and winning new customers is more difficult than holding on to the ones you already have. Companies that provide financial services but have contracts that are not legally enforceable such as banks, credit card companies, insurance companies, and credit unions, are especially concerned about churn since it affects their ability to make money. It is not rare for these companies to have attrition rates as high as 25–30%, and even enterprises with some kind of annual contract may see attrition rates as high as 5–7%. Churn is the term given to the phenomenon that occurs when customers of a credit card company either completely stop using their cards or switch to a different provider. Churn can have a negative impact on the bottom line of a provider. Credit Card Customer Churn (CCCC) can be defined as the rate at which existing credit card holders either stop using their cards, move to a different provider, or leave their current provider. For this reason, it is absolutely necessary for credit card firms to have an understanding of the factors that contribute to the churning of customers and to formulate strategies for reducing churn rates.

II. RELATED WORKS

Amrita Doshi proposed a method utilising AutoML technologies to forecast the number of credit card customers who may churn out of an organisation [2]. H2O-GradientBoosting, H2O-RandomForest, H2O-DeepLearning, Auto-Sklearn, and Auto-keras were supposed to be the method of choice for predicting churn. Among all the employed algorithms, the algorithm Auto-Sklearn has the best accuracy, whereas the H20-deep learning algorithm has the lowest accuracy. CCCC Prediction was proposed by Xinyu Miao and Haoran Wang utilising Random Forest (RF) [3]. Predicting the CCCC with the help of Machine Learning (ML) methods such as RF, Logistic Regression (LR), and K - Nearest Neighbour was the goal (KNN). After making adjustments to the parameters, the Random Forest algorithm achieved an accuracy score on the testing set data of 95.68%. Using machine learning methods to develop a customer churn prediction for credit cards was a suggestion made by Dana AL-Najjar, Nadia AL-Rousan, and Hazem AL-Najjar [4]. Using ML algorithms such as Bayesian Networks, Classification and Regression Trees (CRT), Neural Networks, C5 Trees, and Chi-Square Automatic Interaction Detection (CHAID) Trees, the goal was to make churn predictions. C5 Tree outperformed all the involved ML models involved namely Bayesian Networks, CRT and Neural Networks in training. Customer Churn Analysis (CCA) was proposed for use in the banking sector by Hasraddin Guliyev and Ferda Yerdelen Tatoğlu [5]. Using ML techniques such as LR, RF, Extreme Gradient Boosting, and Decision Tree (DT), the goal was to detect existing clients before losing them to the competitor. The XGBoost algorithm performed the best across all metrics, with an area under the curve (AUC) of 96.97 percent; the RF model came in second. CCCC was proposed by Kamil Demirberk through the use of Support Vector Machine (SVM) in conjunction with Bayesian Optimization [6]. Rajamohamed and Manokaran [7] hypothesised ML methods and Rough Clustering to predict CCCC. The machine learning (ML) methods SVM, RF, DT, KNN, and Hybrid Models were used with the intention of achieving the goal of increasing the retention rate. Support vector machine mixed with rough k-means clustering method works well and has better accuracy than any other hybrid model. Predicting the data from the CCCC and the Automobile Insurance Fraud was the idea of Ganesh Sundarkuma, Vadlamani Ravi, and Siddeshwar [8]. The purpose of this study was to provide evidence of the usefulness of the One Class SVM (OCSVM) methodology that was developed. The research came to the conclusion that the proposed under-sampling methodology was effective in lowering the complexity of the construction system while simultaneously producing important findings. ML solution to the problem of churn in the banking business was proposed by Amgad Muneer and his research team [9]. The use of ML algorithms, including RF, SVM, and AdaBoost, was intended to accomplish the goal of increasing the retention rate. According to the findings of the research, RF performed significantly better than the other algorithms used, achieving an F1 score of 0.91. The prediction of CCCC was suggested by Ning Wang and Dong-xiao Niu [10]. The goal was to increase retention rate utilising the machine learning algorithms known as Rough Set Theory (RST), DT, Ridge Regression (RR), Artificial Neural Network (ANN), and Least Square - SVM (LS-SVM).

III. PROBLEM STATEMENT

The banking industry is currently facing a pressing challenge in the form of customer attrition. With increasing options available to consumers and growing expectations for seamless, personalized services, banks must proactively identify and mitigate factors leading to churn. Failure to do so not only results in revenue loss but also erodes trust and market share. Therefore, there is an urgent need for a robust customer churn prediction system that can enable banks to pre-emptively respond to potential attrition risks.

OBJECTIVE OF PROJECT

The primary objective of this project is to develop an accurate and reliable customer churn prediction model tailored to the banking industry. This model will be designed to analyze historical customer behavior, transactional data, and demographic information to forecast the likelihood of churn for individual customers. By doing so, the project aims to empower banks with actionable insights, enabling them to implement targeted retention strategies and ultimately reduce customer attrition rates.

OBJECTIVE OF PROJECT

The primary objective of this project is to develop an accurate and reliable customer churn prediction model tailored to the banking industry. This model will be designed to analyze historical customer behavior, transactional data, and demographic information to forecast the likelihood of churn for individual customers. By doing so, the project aims to empower banks with actionable insights, enabling them to implement targeted retention strategies and ultimately reduce customer attrition rates.

IV. PROPOSED SYSTEM

The proposed system for bank customer churn prediction leverages H2O's AutoML, a powerful automated machine learning platform, to enhance the accuracy and efficiency of churn prediction. Here's an overview of the components and processes within the proposed system:

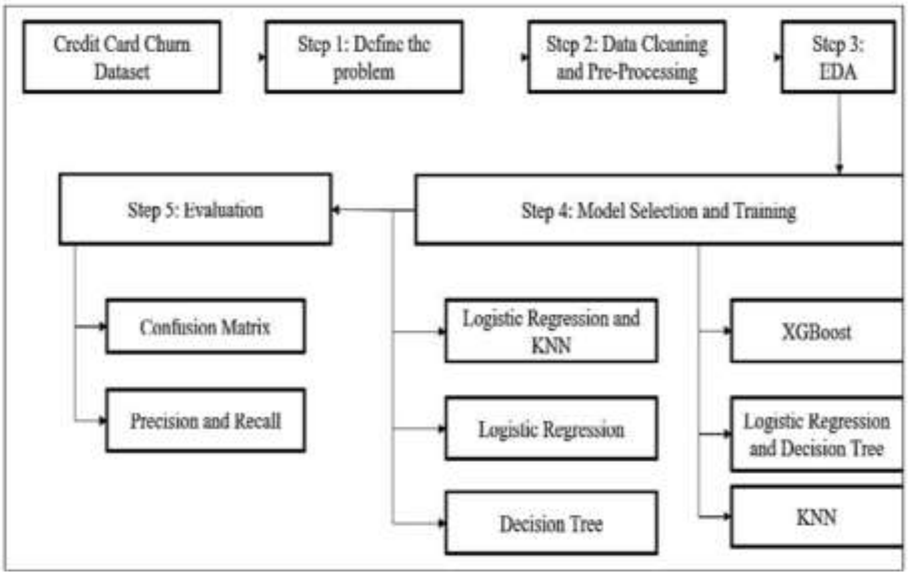
Credit Card Churn Dataset
Step 1: Define the problem
Step 2: Data Cleaning and Pre-Processing
Step 3: EDA
Step 4: Model Selection and Training
Step 5: Evaluation
Confusion Matrix
Precision and Recall
Logistic Regression and KNN
Logistic Regression
Decision Tree
XGBoost
Logistic Regression and Decision Tree
KNN

Fig. 1. Architecture diagram.

Data Collection and Preprocessing

Data Sources: Similar to the existing system, the proposed system gathers historical customer data from various sources within the bank's database.Data Cleaning and Feature Engineering: Raw data is preprocessed, which includes handling missing values, outliers, encoding categorical variables, and potentially engineering new features to improve model performance.

Integration of H2O AutoML

H2O AutoML Integration: H2O's AutoML is integrated into the system. This platform automates the process of training and evaluating a variety of machine learning models, including ensembles, to find the best-performing model for the given dataset.

Model Training and Selection

Automatic Model Selection: H2O AutoML conducts an exhaustive search across a range of algorithms and hyperparameters, training a diverse set of models (e.g., Random Forest, Gradient Boosting Machines, Deep Learning) on the data.Model Comparison: The platform evaluates and compares the performance of these models using various metrics such as accuracy, AUC-ROC, and others.

Ensemble and Stacking

Ensemble Techniques: H2O AutoML employs ensemble learning, combining the predictions of multiple models to improve overall performance. This can lead to more accurate predictions compared to individual models.

Model Interpretability and Explain ability

Feature Importance: The system may utilize H2O AutoML's feature importance analysis to understand which factors are most influential in predicting customer churn. This provides valuable insights for the bank's decision-making process.

Predictions and Actionable Insights

Churn Predictions: The best-performing model from the AutoML process is then used to make churn predictions on new data. These predictions indicate the likelihood of each customer churning.Targeted Retention Strategies: Based on these predictions, the bank can identify high-risk customers and implement targeted retention strategies, such as personalized offers, enhanced customer service, or loyalty programs.

V. RESULTS AND DISCUSSIONS

The Kaggle dataset includes 10127 rows and 23 columns altogether. The many characteristics of the data are represented in the 23 columns. The 23 columns correspond to the different features of data. The accuracy for LR is 0.846, for KNN it is 0.849, for DT it is 0.916, for XGBoost it is 0.931. For the proposed hybrid models integrating LR and KNN, the accuracy attained is 0.9, and for the other integration of LR and DT, it is 0.92. The model built was successfully implemented and that is able to reliably forecast which consumers are at risk of cancelling their subscription. After putting a plan into action, the built model achieved a level of accuracy that was a maximum of 0.96. Through the use of EDA, were able to determine which characteristics of churn were the most significant.

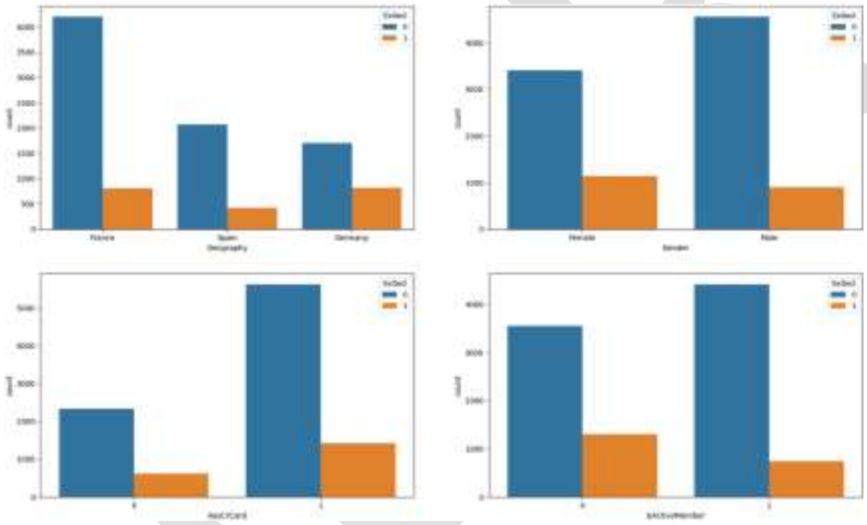

Figure 2: Classification Report

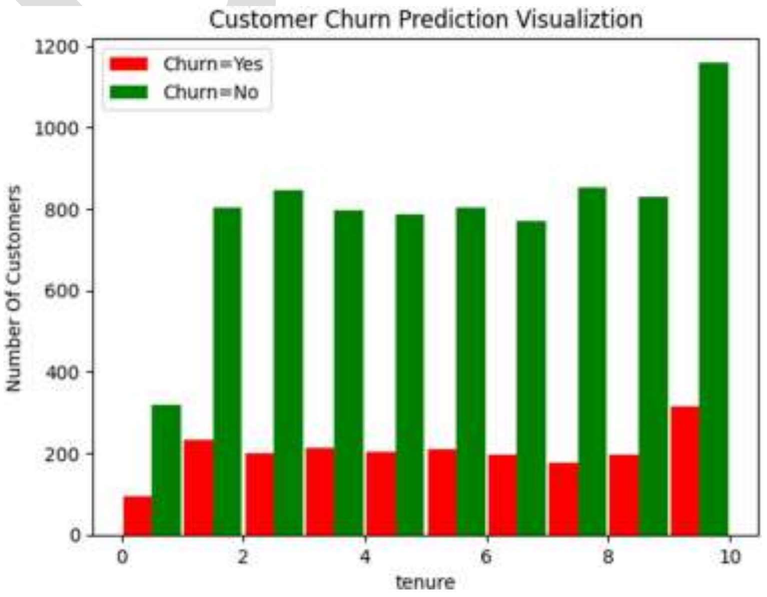

Figure 3: Customer Churn Prediction Visualization

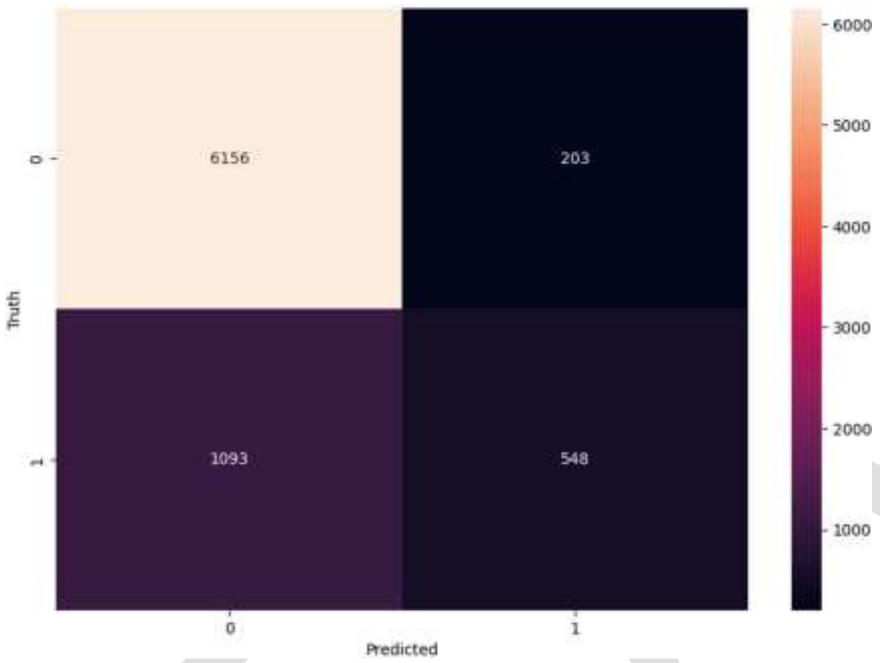

Figure 4: Matrix form

Finding correlations between variables and also finding correlations between target variables. EDA was able to evaluate how strongly each of them is related with both themselves and the objective. This demonstrates that the objective “to provide assistance to credit card firms so that they may better build focused retention tactics and personalized incentive programs to keep clients from defecting to competitive business” was successfully implemented. This also contributes to the achievement of the objective of “determining the factors that effect on the percentage of customers who leave the company.” Without statistical analysis to extract valuable insights from the data, simply gathering feedback is insufficient. By building, the percentage of customer churn if the prevailing system and customer care of credit card service provider can be determined through analysis. Because, there is high chance of successfully implementing the associated objective.

VI. CONCLUSION

In conclusion, utilizing H2O AutoML for customer bank churn prediction has proven to be a highly effective and efficient approach. The automated machine learning platform streamlines the model selection and hyperparameter tuning process, resulting in accurate predictions and improved customer retention strategies for banks. By harnessing the power of H2O AutoML, financial institutions can proactively identify at-risk customers, tailor retention efforts, and ultimately reduce churn, thereby enhancing overall customer satisfaction and maximizing their bottom line.

REFERENCES

1. Median Customer Churn Rates by Industry 2022, https://customergauge.com/blog/average-churn-rate-by-industry
2. Amrita Doshi, Intl. J. Adva. Engg. Manag 3, (2021)
3. Xinyu Miao, Haoran Wang, Customer Churn Prediction on Credit Card Services using Random Forest Method, in the Proceedings of 7th International Conference on Financial Innovation and Economic Development (ICFIED 2022) 648, (2022)
4. Al Najjar, Dana, Al- Rousan, Hazem, J. Theor. Appl. Elect. Comm. Res 17, (2022)
5. H. Guliyev, F. Y. Tatoglu, J. Appl. Micr. Econo 1, (2021)
6. K. Demirberk, Comm. Facul. Sci. Univ. Ankara 70, (2021)
7. R. Rajamohamed, Manokaran, Clust. Computing 21, (2018)
8. S. Kumar, V. Ravi, V. Siddeshwar, One-class support vector machine based under sampling: Application to churn prediction and insurance fraud detection, in the Proceedings of 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC) (2015)
9. Muneer, A. A. Rao, Alghamdi, Amal, M. Taib, Shakirah, Almaghthwi, Ahmed, Ghaleb, Ebrahim, Indon. J. Elect. Engg. Comp. Scie 26 (2022)
10. Ning Wang, Dong-xiao Niu, Credit card customer churn prediction based on the RST and LS-SVM, in the Proceedings of the 6th International Conference on Service Systems and Service Management, Xiamen, China (2009)