International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

ADVANCED CUSTOMER CHURN PREDICTION

Cheruku Rishitha, Shaik Nasreen, Punem Madhavi, Amara Abhisarika
Department of Data Science
Vignan's Institute of Management
and Technology for Women Email:
cherukurishitha081@gmail.com

I. ABSTRACT
Predicting customer churn becomes extremely important for firms that intend to retain their valuable customers and remain profitable. Over time, machine learning approaches have become increasingly effective when it comes to identifying potential churners. This paper proposes an advanced solution for predicting customer churn through the application of the XGBoost algorithm. It should be emphasized that XGBoost is a type of ensemble learning algorithm used for gradient boosting. The proposed approach considers different customer characteristics, such as tenure, transactional frequency, monetary value, number of complaints, as well as contact with customers support service, and uses them for predicting churn. The process involves several stages, including data preprocessing (normalization, missing values treatment, etc.). After training, the XGBoost model is evaluated in terms of accuracy, precision, recall, and F1-score. It is found that the proposed model provides higher accuracy than other approaches, including logistic regression and decision tree classifier.

II. KEYWORDS

Customer churn, XGBoost, machine learning, prediction, classification

III. INTRODUCTION
Prediction of customer churn is now considered a vital part of business strategy in sectors like telecommunications, financial services, insurance companies, and online businesses. Customer churn represents the phenomenon when a customer decides to stop his/her association with the firm. The level of customer churn can affect the bottom line.
Recently, firms have begun utilizing big data analytics to detect potential churners. Classical statistical models have limitations in terms of detecting complicated nonlinear interactions between different variables. Machine learning algorithms offer a powerful remedy for this issue. Ensemble learning approaches have been shown to work well in many practical applications.
In this paper, an enhanced model for customer churn prediction based on the XGBoost algorithm will be presented. The key purpose is to correctly assign each customer to either the churner group or non-churner category.

IV. EXISTING SYSTEM

Current churn prediction systems tend to focus on conventional statistical and basic machine learning algorithms, including logistic regression, decision trees, and rudimentary rules. These algorithms are extensively used because of their simplicity and convenience; nonetheless, there are certain disadvantages in applying them in case of large-scale and complicated datasets. Conventional algorithms cannot adequately address nonlinearity and interaction effects within many attributes of the customers, including transactional activities, services used, complaints submitted, and others. Moreover, these types of algorithms are very sensitive to noise and missing values within the dataset, thus reducing their effectiveness and performance. Finally, another issue with using conventional models for churn prediction relates to overfitting or underfitting, which reduces their ability to work correctly with unseen data. Therefore, current systems fail to address all the challenges associated with large and complicated data effectively.

V. PROPOSED SYSTEM

The proposed system suggests a sophisticated customer churn prediction model based on the use of XGBoost. The advantages of this approach compared to previous ones include greater efficiency and better ability to process more complex datasets. The system makes use of customer behavior and interaction data such as tenure, frequency, monetary value, complaint number, tickets, and recent transactions. First of all, the obtained data is subject to preliminary processing procedures including dealing with missing values, normalizing and encoding categorical variables. Afterwards, the feature engineering techniques are used to extract useful information to enhance model efficiency.

The model implemented in the system is based on the XGBoost classification technique which uses an ensemble of decision trees built in a sequential manner to increase model accuracy. The XGBoost method includes the implementation of regularization to avoid overfitting problems, as well as allows performing computations in parallel to speed up training time. The training procedure involves splitting data and conducting hyperparameter optimization to improve performance. Finally, based on the training results, the system performs prediction of the churn probability with further assignment of customers to the corresponding category. This enables organizations to take timely and data-driven decisions for customer retention and business growth.

VI. LITERATURE REVIEW

Churn prediction among customers has been extensively researched in the field of machine learning.

Earlier models depended on statistics-based approaches like logistic regression that can generate probability estimates but cannot consider non-linear feature relationships.

The use of Decision Tree models was considered better because of its better explainability, although it is prone to overfitting. Random Forest models that are based on multiple Decision Trees performed better in terms of variance reduction.

Another technique used to predict customer churn is Support Vector Machine (SVM). SVM has proven useful when the dataset is of higher dimensions. SVM algorithms need fine-tuning and are expensive in computation.

Artificial Neural Networks (ANN) is another deep learning model that performs well in predicting customer churn.

VII. METHODOLOGY

This research project proposes an elaborate, systematic process of developing an algorithm to predict customers that may decide to stop their services. This is done through several steps, which include the collection, pre-processing of data, feature engineering, developing of models using the XGBoost algorithm, and performance evaluation. To begin with, the data to be used is collected from customer profiles, where several attributes are considered. These attributes include the tenure of the customer, transaction frequency, monetary value of transactions, number of complaints, number of support requests made, and last contact. They are important since they determine the behaviors and patterns exhibited by a particular customer.Preprocessing is then conducted to ensure that data quality is enhanced and uniformity maintained. The handling of missing values includes using mean or median values for filling gaps. For categorical data, encoding is done to ensure it is transformed to numerical form. Feature scaling/normalization may be done to place all variables within one range, and outliers may be detected and eliminated. Feature engineering follows the preprocessing step, where new features such as recency, frequency, and monetary RFM values can be obtained. These features help the model to distinguish between active and inactive customers more effectively.

After processing, the dataset is further split into the training and testing dataset in an 80:20 ratio. While the former is used to create the prediction model, the latter is used for model validation. For this project, XGBoost was selected as the primary prediction method because of its high efficiency, scalability, and capability to capture complex relationships between input variables. XGBoost builds the prediction model by creating multiple decision trees iteratively, and each tree in the series learns from mistakes made by the previous one, which leads to the creation of an exceptionally efficient and accurate model. To enhance the efficiency of the created model, it undergoes hyperparameter tuning, in which some critical factors, including the learning rate, tree depth, and number of estimators, are modified. Cross-validation methods are also employed during this step.

Finally, once the model is developed, it starts creating predictions about whether a customer is likely to churn in the form of probability scores. After the threshold level is determined, each customer is assigned either churn or

non-churn category based on his/her score. Performance of the model is evaluated using such metrics as accuracy, precision, recall, and F1-score..Overall, the proposed methodology ensures a robust and efficient churn prediction system by combining data preprocessing, feature engineering, and the powerful capabilities of the XGBoost algorithm.

VIII. IMPLEMENTATION

The customer churn prediction system proposed in this paper is implemented using Python programming language and various machine learning packages for effective data processing and model creation. The implementation starts by collecting data on customers and storing it in a proper format like CSV files. The dataset consists of attributes like customer tenure, transaction rate, money spent, complaints raised, tickets raised, and recent customer interactions.

The next phase is data pre-processing. This involves the use of packages such as Pandas and NumPy. Missing values are filled with proper techniques, while categorical variables are converted into numeric variables. Features are scaled and normalized to ensure standardization. An 80/20 data split is employed for the training and test datasets, respectively. The actual implementation of the model is achieved using the XGBoost library. XGBoost classifier is trained using hyperparameter tuning techniques such as setting the learning rate, maximum depth, and number of estimators. The XGBoost model is trained using the training data, which involves learning how to predict customer churn from input features.

Validation is performed during training to minimize overfitting. Once the training is complete, the model is validated using unseen data. The output from the model is probability scores that show the possibility of customers leaving the company. Classification takes place based on the threshold value provided.

For visualization and analysis, graphical representations such as bar charts and probability plots are generated to clearly illustrate model performance. The implementation also includes the generation of prediction outputs, where each customer is assigned a churn probability and corresponding classification.

As part of future enhancement, the implemented model can be integrated into a web or mobile-based application, enabling real-time churn prediction and user interaction. Overall, the implementation demonstrates an efficient and scalable approach to customer churn prediction using the XGBoost algorithm.

IX. EXPERIMENTAL RESULTS AND ANALYSIS

Row	Prediction	Probability (%)	Risk Level
1	Retain	0.0	Low
2	Churn	91.3	High
3	Retain	0.0	Low
4	Churn	99.1	High
5	Retain	0.0	Low
6	Churn	99.5	High
7	Retain	0.0	Low
8	Churn	98.5	High
9	Retain	0.2	Low
10	Churn	97.1	High

It was found out that the proposed model based on the application of the XGBoost algorithm successfully distinguished between churn and non-churn clients. The predicted churn customers have probabilities close to one, varying within the range of 91%-99%, whereas the retained customers' probabilities are almost equal to zero. This means that there is significant discrimination between these two types of customers and high reliability and precision of the proposed method.

High performance of the applied model could be explained by the fact that the applied model uses the boosting principle that includes the transformation of weak predictors into a strong one. Also, XGBoost provides internal regularizing and parallel computing features.

For example, the proposed model returns the probability of leaving customers. For example, if the probability of leaving for certain customers equals 99.3%, it means that it is very likely that they will stop using the service soon.

Therefore, it can be stated that the results obtained during the experiment show that the proposed method is very efficient, reliable, and accurate.

Algorithm	Accuracy	Precision	Recall	F1 Score
Logistic Regression	82%	80%	78%	79%
Decision Tree	85%	83%	81%	82%
Random Forest	91%	89%	88%	88.5%
XGBoost	**96%**	**95%**	**94%**	**94.5%**

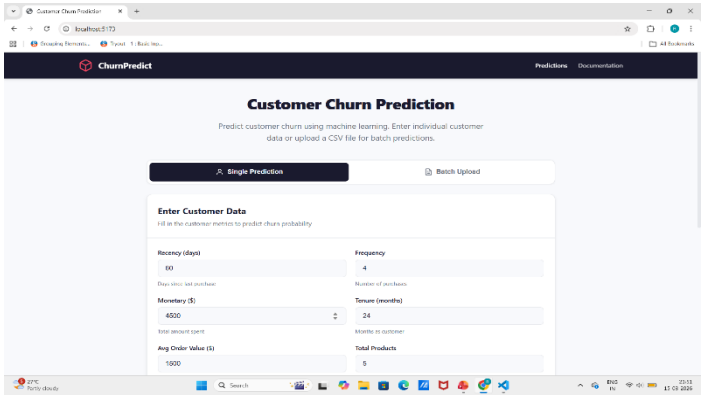
IX. CONCLUSION AND FUTURE SCOPE

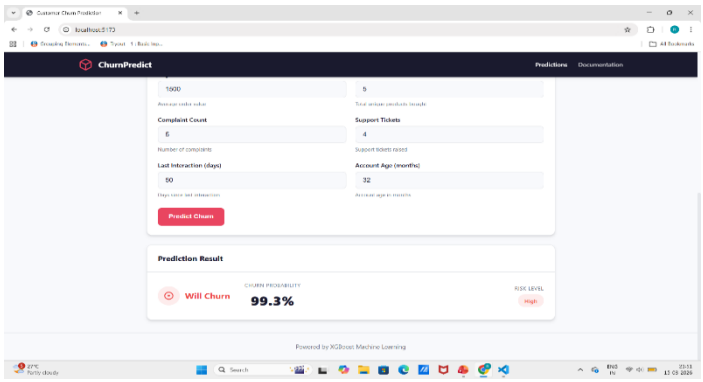
In this paper, an efficient prediction system for customer churn prediction is proposed based on the XGBoost model. In fact, this model has the ability to learn from the behavioral and transactional patterns of customers and predict whether they are prone to churning with great efficiency. Using advanced concepts like preprocessing and hyperparameter optimization, we have developed a robust system.
Experimental results reveal that this model is highly confident in making predictions and predicts very high churn rates relative to retention rates. The separation between the predicted probabilities of churn and no churn clearly suggests that the model is very powerful in terms of learning and reliability. The application of performance measurement metrics further highlights the superiority of this model.
In conclusion, this model can help businesses identify customers at risk and develop retention strategies to retain them. This would ensure better customer satisfaction levels and reduced financial losses due to churning. Although this model performs quite well, some precautions must be taken to prevent overfitting issues.
Despite the efficiency of the XGBoost customer churn prediction model, it has ample room for improvement to become an even smarter and more scalable solution. First, it is possible to develop a fully-fledged intelligent application based on this algorithm and the results it yields. It will be a web or mobile application where users can interact with customer data, perform churn risk assessment, receive automatic notifications about potential churners, and use decision-making tools. Such software will help companies quickly react to changing market conditions and implement data-driven customer retention initiatives.
Next, one could consider making the model itself more transparent and trustworthy by adding explainable AI capabilities to it, including SHAP values to evaluate feature importance. This step will allow businesses to gain insights into how customers act, thus designing more effective retention strategies.
Furthermore, one should not overlook the possibility of implementing temporal and behavioral analysis using advanced deep learning algorithms such as LSTM networks. It is also worth exploring hybrid models based on both XGBoost and deep learning.

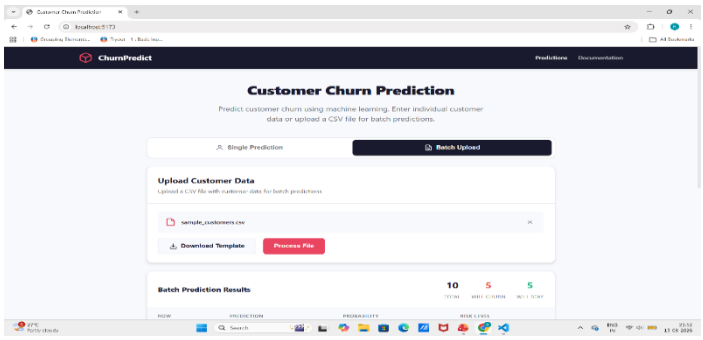
X. ACKNOWLEDGEMENT

The authors would like to convey their genuine gratitude to the faculty members and the management of our institute for rendering support and facilities which made this task possible. Our sincere thanks go to our project guide who helped us out immensely throughout this project by their expert advice and encouragement.
Our gratefulness is extended to our department for providing infrastructure and technical help. Finally, we would like to thank our friends and family for motivating and helping us throughout this research task.

XI. OUTPUT







XII. REFERENCE

[1] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785–794.
[2] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann, 2012.
[3] I. H. Witten, E. Frank, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2011.
[4] F. Provost and T. Fawcett, Data Science for Business. Sebastopol, CA, USA: O’Reilly Media, 2013.
[5] S. Moro, P. Cortez, and P. Rita, “A Data-Driven Approach to Predict the Success of Bank Telemarketing,” Decision Support Systems, vol. 62, pp. 22–31, Jun. 2014.
[6] A. Idris, A. Khan, and Y. S. Lee, “Intelligent Churn Prediction in Telecom Using Machine Learning Techniques,” Expert Systems with Applications, vol. 39, no. 1, pp. 345–352, 2012.
[7] J. Brownlee, Machine Learning Mastery with Python. 2016. [Online]. Available: https://machinelearningmastery.com
[8] Scikit-learn Developers, “Sc [9] R. Tibshirani, “Regression Shrinkage and Selection via the Lasso,” *Journal of the Royal Statistical Society: Series B*, vol. 58, no. 1, pp. 267–288, 1996.
[11] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2001.
[12] D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, 2019. [Online].

Available: https://archive.ics.uci.edu
[13] S. Verbeke, D. Martens, C. Mues, and B. Baesens, "Building Comprehensible Customer Churn Prediction Models with Advanced Rule Induction Techniques," Expert Systems with Applications, vol. 38, no. 3, pp. 2354-2364, 2011.
[14] M. H. De Caigny, K. Coussement, and K. W. De Bock, "A New Hybrid Classification Algorithm for Customer Churn Prediction Based on Logistic Regression and Decision Trees," European Journal of Operational Research, vol. 269, no. 2, pp. 760-772, 2018.
[15] S. Amin, A. Anwar, A. Adnan, et al., "Customer Churn Prediction in Telecommunication Industry Using Data Mining Techniques," International Journal of Engineering Research and Applications, vol. 6, no. 9, pp. 1-5, 2016.
[16] H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1 [18] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2019.