Grenze International Journal of Engineering and Technology, June Issue

Customer Churn Prediction Model and UI Dashboard Visualization

Roshan Lal[1], Jigyasu Rawal[2] and Vansh Agrawal[3]
[1-3]Department of Computer Science and Engineering, Amity University Uttar Pradesh, India
Email: rchhokar@amity.edu, rawal.jigyasu123@gmail.com, agrawalvansh41@gmail.com

***Abstract*— The increasing competition in sectors including telecommunications, banking, e-commerce, and subscription- based services has made customer churn prediction an important academic and commercial issue. Early detection of churn-prone consumers is a strategic advantage for businesses since retaining current customers is far less expensive than recruiting new ones. This study looks at the use of machine learning methods to forecast customer loss and pinpoint the main causes. Analysis is done using a structured churn dataset that includes billing, behavioural, demographic, and service-related characteristics. The research methodology follows a methodical pipeline that comprises of feature engineering, model creation, exploratory data analysis (EDA) to help understand the data patterns and anomalies, and data pretreatment. EDA highlights the churn patterns, that includes the contract type and the monthly charges. Data preprocessing and feature engineering used in the models help enhance the model quality, and the models are assessed using system of measurement appropriate for imbalanced data. Gradient Boosting was found to be the most effective for identifying high-risk churners. Tenure, contract type, and monthly fees are the most significant indicators, according to feature importance analysis. Overall, the study offers a methodology which is useful for integrating the churn predictions into customer relationships management systems and strike a balance between predictive accuracy and interpretability.**

I. Introduction

As a result, plummeting churn is a crucial constituent of a long-term competitive advantage rather than just counting as a defensive tactic as it directly impacts customer lifetime value. Data-driven methods have revolutionized churn management tactics within the last ten years. Due to their interpretability, traditional statistical models like logistic regression were first preferred. Nevertheless, these models are seen frequently falling short of capturing the complicated, non-linear correlations present in the consumer behaviour. The creation of complex models that can analyse vast amounts of consumer data and identify minute behavioural trends has been made possible by developments in machine learning. Predicting customer attrition is extremely important in all industries. Customers in the tele-communication industry are seen frequently switching the service providers due to factors such as bundled goods, pricing, or service quality. Within banking and financial facilities, churn can take the arrangement of investment relocation or account closures. Similarly in a subscription-based business platform including the Software As a Service providers, the streaming services, and online gaming, are particularly vulnerable ass the churn directly results in continuous revenue losses. Despite incredible advancements in the predictive techniques the existing models still continue to face challenges and the problems still exist.

Grenze ID: 01.GIJET.12.2.168_20

Since churners frequently make up a small portion of the customer base, class imbalance is nonetheless a significant problem. Furthermore, despite the great predicted accuracy of complicated models, managerial trust and adoption are hampered by their interpretability. In order to close the gap between technical modelling and business decision- making, this study compares several machine learning techniques for churn prediction with a focus on both performance and interpretability to achieve the accurate results for retaining the customer's attention by improving the recommendation strategies.

II. Literature Review

Data mining, machine learning, and marketing analytics have all done a lot of research on customer churn prediction. These models were valued for their ability to organise the data and helped in revealing the understandable patterns. Yet, there were characteristic limitations. Due to the assumed variables related to each other in a linear fashion, these models were struggling with complex or non-linear relationships that exists in the real-world data While they were excellent in highlighting the basic patterns, they sometimes were struggling with the more complicated dynamics of the data.

A. Initial Statistical Techniques

The survival analysis and the logistic regression were commonly used in the initial researches of churn to determine the probability of client churning. These models enabled managers to examine the impacts of such factors as tenure, billing issues, and the patterns of usage. Nevertheless, the shortcomings of linear assumptions were observed as data became larger and increasingly complex and thus reduced the effectiveness of predictions.

B. Machine Learning and Rule-based Techniques

Rule-based classifiers and decision trees are also popular methods of data mining that were used in the early 2000s. These models increased the accuracy of the prediction and gave human understanding of the decision criteria. Support Vector Machines and Artificial Neural Networks additionally enhanced performance in terms of non-linear data and high- dimensional data but their performance came at the cost of reduced interpretability.

C. Feature Engineering and Ensemble Learning

Random Forests, Gradient Boosting, XGBoost, and LightGBM are modes of ensemble; they are popular in the 2010s, as big data has risen in prominence. The models are more accurate and resistant to overfitting because they are a combination of multiple weak learners. Similar to RFM variables, interaction terms and behavioural aggregates, feature engineering, especially inclusion of these variables, was found to be a key challenge in improving the performance of a model.

D. Class Imbalance and Measures of Evaluation

Churn dataset is often not in balance with less than 20% of churners. In a bid to boost the recollection of the minority subset, analysts came up with algorithms such as SMOTE, under sampling, cost-sensitive learning, and threshold tuning. It substituted accuracy as the primary performance measure with such evaluation metrics as precision, recall, F1-score and AUC-ROC.

E. Deep Learning, Temporal and Behavioural Models of Model.

Recent studies include both temporal dependencies and network effects, and behavioural data. The autoencoders and embeddings of high-dimension categorical data also enhance the accuracy of prediction.

F. From Prediction to Action

Modern churn research is not concerned with prediction: on the contrary, it is concerned with actionable insights.

F. Emerging Trends

These developments demonstrate that dynamic and business-integrated decision-support systems are substituting the churn classification which is static.

III. Methodology

The methodology explains the systematic approach of creating prediction models and predicting client attrition. It uses data pre-treatment, exploratory analysis, feature engineering, model building and evaluation to ensure that the results are consistent and reproducible.

A. Data Preprocessing and Correction

A database consisting of a structured customer turnover supplied the dataset, including billing, service related, and demographic characteristics.

Preprocessing Included

- Managing missing values (removal or imputation).
- Categorical variable encoding (one-hot encoding).
- Scaling continuous features, such as monthly fees and tenure.
- Addressing class imbalance with the use of techniques such as SMOTE or class weighing. Through such actions, consistency and modelling preparedness on data was ensured.

B. Exploratory Data Analysis (EDA)

EDA was applied in order to identify trends and relationship among features and churn. The statistical summaries and visual figures inclined feature selection (correlation heatmaps, box plots and histograms). Key observations were used to model, including increased churn rates among month-to- month consumers and consumers with large monthly fee.

C. Feature Engineering

The input variables were later narrowed down with the feature importance ranking.

D. Model Development

The comparison method was used as a test of several categorization algorithms: 17. One of the models that can be interpreted as a baseline is the logistic regression. For better predictive performance, we used gradient boosting (XGBoost/LightGBM).
Patterns are more deeply identified with the help of neural networks. A 70 percent and a 30 percent stratified sampling was used to train and test the models respectively.

E. Evaluation Metrics

Preciseness alone was not enough considering the contrast in class. Precision, recall, and F1-score were used to evaluate the models in order to balance this. In order to assess discriminatory skills, ROC-AUC Score is to be used.
Confusion Matrix: it is a tool that displays the effectiveness of classification.
The two scores were AUC and F1-score that were used to decide which model was best.

F. Implementation Framework

The reproducibility was ensured using cross- validation and random seed control.

IV. Implementation and Analysis of Results

The findings of the churn prediction models devised in this paper are discussed in the Results and Discussion section. It provides a description of the model performance in a comprehensive way, compares different algorithms, determines the essential factors that lead to churn, and discusses the implications of these findings to the practical things in the industry and the study. It is supposed to demonstrate the practical utility of the models in guiding client retention strategies in the company in addition to its technical soundness.

A. Experimental Setup

The dataset was pre-processed, balanced, and stratified to split the dataset into a 70-30 train-test split in order to preserve the proportions of the classes. Each model was trained using Python software, such as scikit-learn, XGBoost, and TensorFlow. To ensure the presence of universality, performance was measured in terms of k-fold cross- validation. The measures of accuracy, precision, recall, F1-score, AUC-ROC, and confusion matrices were some of the measures that were utilized inscrutinizing misclassifications.

B. Comparison on Performance of the Model.

Logistic Regression

- Approximately 80% accuracy
- Moderate precision
- Recall: Very bad, which indicates that no attention was paid to churners.

- AUC: about 0.76 27As a benchmark, logistic regression performed quite well, showing that significant features such as contract type, tenure, and monthly charges are statistically significant. Yet, it did not have high predictive power because of its inability to reproduce complex non-linear trends.

Decision Trees

- Precision: Approximately around 83%
- Revision: Better than Logistic Regression. Interpretability is extremely high, complexity of rules is minimal (e.g., customers: those who have a month- to- month contract and are charged a lot might be churners).

Random Forests

- Precision: approximately 86%
- AUC: about 0.84
- Precision and Recall: balanced performance Random Forests was able to reduce overfitting by combining multiple trees together. The most important features in predicting turnover were tenure, monthly fees and type of contract, to the extent of feature importance.

Gradient Boosting (XGBoost, LightGBM)

- Precision: around 89–91%
- Accuracy: High, Recall: Strong and also minimises false negatives.
- AUC: between 0.90 and 0.93

They continue to be more accurate than the others, and XGBoost even proved to be the most robust churn predictor, hitting the right compromise between identifying the churners (recall) and maintaining the false alarm rate at a minimum (accuracy).

Neural Networks

- Approximately 90% accuracy
- AUC: approximately 0.92• Recall: In some ways, a little less than Gradient Boosting

It was overall in a good performance but not as efficient as the boosting models in picking up all the churners. The downside? It was a black box and more difficult to train, so we would prefer the network to operate behind the scenes to serve analytics teams, but not to make direct customer contact.

C. Confusion Matrix Insights

False negatives were the largest error of all models that comprised of missed churners. That would be an enormous blow since one would be missing revenue due to the lack of such customers. Due to this reason, the Gradient Boosting models were most impressive in magnifying false negative and this is an important strength of the model as far as the business perspective is concerned.

D. Importance Analysis of Features

Exploring the weights we discovered some uniform churn variables: Contract Type: the churn rates of customers on month-to-month plans were significantly higher in comparison with month-to-year and two-year contracts.
Tenure: the longer customers had been the less likely they were to migrate--loyalty was fostered through time.
Monthly charges: predictions of higher churn were higher with higher charges, meaning they are price sensitive.
Additional Entrance: the absence of the internet, technological support, or security online drove away people.
Payment Method: the use of electronic checks fragmented into a bigger churn rate compared to automatic systems.

E. Comparative Discussion

There was evident trade-off of the accuracy and explainability. The trees in decision and logistic regression are simpler to interpret, and not as precise. Random Forests and Gradient Boosting get performance with a boost, whereas neural networks are not transparent but can achieve state of the art accuracy. The reasonable solution that applies to industry is a compromise: predicting with Gradient Boosting and a less advanced model to discuss decisions with managers.

F. Business Implications

The goals of the focused retention initiatives comprise: a) Building Strength: students with learning disabilities possessing superior language abilities are able to deliver significant contributions to both the school and the community. The objectives of the focused retention initiatives include:

Focused Retention Initiatives

Ranked on churn risk, companies can prioritize interventions on customers:

- High risk churners: good retention incentives (discounts, reward loyalty).
- Medium-risk changers: individual or bundle communications.

Making Decisions Based on Profit

A high-value customer is much more valuable to save than a low-value customer despite their relatively equal of churning.

Integration of Operations

CRM systems should receive churn scores so that instant alerts are displayed to sales or support staffs and they can take timely actions.

G. Limitations of the Study

There are caveats associated with these results even though the results appear promising:

- Class Imbalance: The heavy imbalance still biases the models, despite the application of SMOTE.
- Interpretability Issue: Data: XGBoost and neural nets should have explanation tools (e.g., SHAP).
- Dynamic Customer Behaviour: Churn drivers change over time, and this requires continuous retraining of the models.

V. Future Directions

A. In future studies, it might be investigated

Real-Time Prediction: Predict churn in real time with the use of streaming data. Text and Sentiment Analysis: Improving models using complaints, call logs or social media conversation. Uplift Modelling: Determining the responsiveness of customers to retention initiatives. Fairness and Ethics: Making sure that the privacy laws are adhered to, and that predictions are not biased. Generally, the project indicates that machine learning can fuel a concrete churn prediction. Gradient Boosting ensemble methods strike just the right chord of precision, recall and AUC, yet the discussion makes a point of reminding us that practical implementation of churn management depends on alignment between model output with strategy, customer value and technical achievement – technical performance is not sufficient.

References

[1] Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2022). Customer churn prediction system: a machine learning approach. Computing, 104(2), 271-294.

[2] Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. Expert Systems with Applications, 39(1), 1414-1425.

[3] Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. Expert systems with applications, 38(3), 2354-2364.

[4] De, S., & Prabu, P. (2022). Predicting customer churn: A systematic literature review. Journal of Discrete Mathematical Sciences and Cryptography, 25(7), 1965- 1985.

[5] Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. Expert Systems with Applications, 36(3), 4626-4636.

[6] Rahman, M., & Kumar, V. (2020, November). Machine learning based customer churn prediction in banking. In 2020 4th international conference on electronics, communication and aerospace technology (ICECA) (pp. 1196-1201). IEEE.

[7] Wagh, S. K., Andhale, A. A., Wagh, K. S., Pansare, J. R., Ambadekar, S. P., & Gawande, S. H. (2024).Customer churn in telecom sector using machine learning techniques. Results in Control and, 14, 100342.

[8] Fujo, S. W., Subramanian, S., & Khder, M. A. (2022). Customer churn prediction in telecommunication industry using deep learning. Information Sciences Letters, 11(1), 24.

[9] Imani, M., Joudaki, M., Beikmohammadi, A., & Arabnia, H. R. (2025). Customer churn prediction: A systematic review of recent advances, trends, and challenges in machine learning and deep learning. Machine Learning and Knowledge Extraction, 7(3), 105.

[10] Chandrakala, D. (2016). A survey on customer churn prediction using machine learning techniques. International Journal of Computer Applications, 154(10).

[11] Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction intelecom using machine learning in big data platform. Journal of Big Data, 6(1), 1-24.

[12] Manzoor, A., Qureshi, M. A., Kidney, E., & Longo, L. (2024). A review on machine learning methods for customer churn prediction and recommendations for business practitioners. IEEE access, 12, 70434-70463.