

Machine Learning–Based Prediction of Customer Churn Using Demographic, Financial, and Behavioral Features

Rahul Pathania

Computer Science And Engineering

Lovely Professional University

Jalandhar, India

rahulrajput831831831@gmail.com

Abstract—Customer churn is a critical challenge for service-based industries such as banking, as customer attrition directly impacts revenue and long-term profitability. Predicting churn in advance enables organizations to take proactive retention actions. This paper presents an end-to-end machine learning framework for customer churn prediction using a real-world banking dataset. The proposed approach includes data preprocessing, feature engineering, and model comparison. Three machine learning models—Logistic Regression, Random Forest, and XGBoost—were implemented and evaluated using cross-validated ROC-AUC. Among them, the Random Forest model achieved the best performance with a test ROC-AUC of approximately 0.86. To address the interpretability challenge of machine learning models, SHAP-based explainable AI techniques were applied to identify both global and individual churn drivers. The results reveal that customer age, credit score, product usage, and activity status are key factors influencing churn. The proposed framework provides accurate predictions along with actionable insights, supporting data-driven customer retention strategies.

Keywords—component, formatting, style, styling, insert (*Customer Churn , Random Forest , XGBoost , SHAP*)

I. Introduction

Customer churn can be considered a big problem in service-oriented industries such as banking, telecommunications, and online subscriptions. Customer churn is an event where a customer ends up not using a service and joins a rival service instead. The cost of acquiring a new customer in the banking sector is much higher in contrast to retaining an existing one.

Conventional analytical tools used for churn analysis are based on manual reporting and simple statistical analysis, which are largely reactive in nature and do not identify intricate behavioral patterns among customers. As a result, such analysis lacks scalability and does not offer timely insights necessary for proactive customer retention. Given the increasing volume of customer information, there is a requirement for automated analysis solutions with a high degree of accuracy in identifying the reasons for customer churn.

Machine Learning methods have come into prominence as efficient predictive analytical methodologies in learning patterns from customer data in a retrospective manner. ML algorithms have the capability to assess several variables of a customer, such as ‘age, credit score, balance, tenure, and product usage, to name a few, to estimate a customer’s chances of ‘churn.’ A major drawback of several very successful methods in this domain is the fact that they function more like ‘black boxes.’

To counter this problem, explainable artificial intelligence (XAI) methods are being increasingly combined with

machine learning algorithms. SHAP is one such technique among these methods, which provides both global and local explanations based on a mathematical assessment of the contribution of each variable toward a model’s predictions. With SHAP, an organization can not only predict customer churn but also identify the major contributors to this churn.

In this research, an end-to-end machine learning solution is presented for modeling customer churn based on a real-case banking dataset. A variety of machine learning algorithms, such as Logistic Regression, Random Forest, and XGBoost, are applied. The model with the highest cross-validated ROC-AUC score is considered the best-performing model. Moreover, SHAP analysis is used to understand how each predictor affects model outputs and to highlight the most important factors contributing to customer churn.

II. PROBLEM STATEMENT

Customer churn is a major challenge for banks, as many customers leave without prior warning, leading to significant financial losses. Conventional methods used in customer analysis are largely reactive in nature and fail to accurately identify customers who are at the highest risk of leaving. Additionally, these traditional approaches do not provide sufficient insights into the underlying reasons behind customer attrition, making it difficult for organizations to design effective retention strategies. As a result, banks often struggle to take timely and informed actions to retain valuable customers. In the absence of predictive and explainable analytical tools, customer retention efforts become inefficient, costly, and less personalized. This highlights the need for data-driven and automated solutions that can not only predict customer churn in advance but also explain the factors influencing customers’ decisions to leave an organization.

III. Objectives

It intends to design an end-to-end machine learning model to predict customer churn.

Preprocessing and feature engineering from raw customer data into something meaningful.

To compare several machine learning models with the intent of selecting the best-performing model.

Assessment of model performance with the use of proper metrics: ROC-AUC, Precision, Recall, F1-score.

To interpret model predictions using SHAP-based explainable AI techniques.

IV. DATASET DESCRIPTION

This is a publicly available banking customer dataset containing historical customer information. It includes 10,000 customer records with several attributes that describe demographic, financial, and behavioral features of customers. These features represent customer age, credit score, account balance, tenure, number of products used, activity status, and estimated salary. The target variable is Exited, which is 1 if the customer churned and 0 otherwise. This dataset represents a realistic basis on which customer churn prediction can be studied in the banking domain.

V. Methodology

The end-to-end predictive analytics pipeline was developed to make the process of modeling systematic and reproducible. Its methodology consists of the following stages:

A. Data Preprocessing

Missing value imputation was done using the right imputation strategy. Numerical features have been imputed using median values, while categorical features were imputed using the most frequent category. Outliers in numerical features have been treated using the IQR capping method to reduce extreme value leverage.

B. Feature Engineering

Several features are engineered to better the predictive performance, which includes balance per product, a binary indicator for customers with a high balance, and logarithmic transformation of skewed numerical features. Highly correlated features have been identified and removed to reduce redundancy. Numerical features were scaled using **RobustScaler**, which is less sensitive to outliers compared to standard scaling techniques. This step ensured that features with large magnitudes did not dominate the learning process. Highly correlated numerical variables were filtered out in order to remove redundancy and multicollinearity in the model. The chosen model was tested on a new dataset using a variety of evaluation metrics such as ROC AUC score, Precision, Recall, F1 Score, and Confusion Matrix analysis. This will allow for both statistical and business perspectives of model performance to be determined. To make sure that both the training and testing datasets were preprocessed in a consistent way, a unified preprocessing step was accomplished with Scikit-learn's "Pipeline" and "ColumnTransformers". The processing of both numerical and categoric variables took place in a separate manner and afterwards they were all combined.

C. Data Transformation

Numerical features were then scaled by RobustScaler to handle outliers effectively. Categorical features needed to be converted into numerical features by one-hot encoding. ColumnTransformer from Scikit-learn was used to create a unified preprocessing pipeline to ensure that training and testing are done consistently.

D. Train-Test Data Splitting

The dataset was split into subsets for both training and testing purposes for the evaluation of the model's generalization performance. A stratified split technique is used with a focus

on maintaining equal proportions in both subsets for the class of churned and non-churned customers.

E. Handling Class Imbalance

Datasets of customer churn always tend to be imbalanced, with a higher percentage of non-churned customers. To handle this problem, model evaluation metrics such as ROC AUC were used, which are threshold-independent and do not have any preference for the majority class.

F. Evaluation Metric Selection

A variety of evaluation metrics were used to gauge model performance. Although accuracy gave a broad idea of model performance, other metrics such as precision, recall, F1-score, and ROC-AUC provided more nuanced **information** on model efficacy in pinpointing customers who can potentially churn.

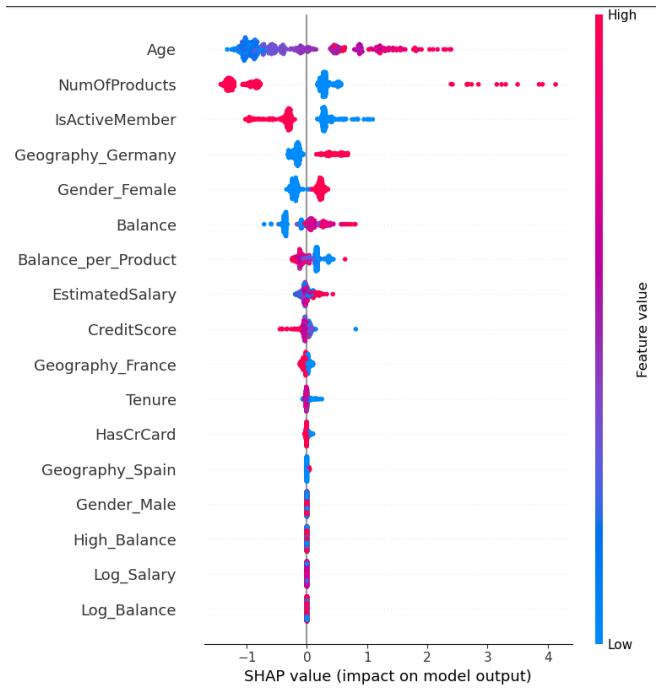


Fig. 1

G. End-to-End Workflow Validation

The end-to-end process, from ingestion to explanation, in modeling has been validated. This validates that it would be easy to incorporate this proposed framework in a real-world setting for customer analytics.

VI. MACHINE LEARNING MODELS

To resolve the customer churn problem, three different supervised machine learning algorithms were used. The algorithms were chosen based on their capabilities to balance interpretability, accuracy, and robustness.

A. Logistic Regression

Logistic Regression was used as a baseline classifier because of its simplicity, efficiency, and interpretability. The model predicts a probability for customer churn based on a linear combination of inputs and a sigmoid function. The model can easily be interpreted to see how each input affects a customer churn outcome. Therefore, this model can be used for a baseline evaluation. To counter overfitting, regularization methods were used. Additionally, hyperparameters were optimized with a focus on improving model accuracy. While

Logistic Regression models a linear relationship between inputs and output, in contrast to other models, it can be considered a model of choice in interpreting a linear result for assessing performance improvements delivered by other models in this case.

B. XGBoost Classifier

For handling structured/tabular datasets, XGBoost, an algorithm based on gradient boosting, was used because of its efficacy in such kinds of datasets. In contrast to Random Forest, where individual trees are generated independently, in XGBoost, individual trees are generated one after another, with each trying to compensate for the weaknesses of the previous ones. Hyperparameters such as depth of the tree and number of estimators were set in consideration of complexity and overfitting of the model. XGBoost provides some regularization techniques to avoid overfitting, hence being a better alternative when building a churn prediction model.

C. Support Vector Machines

The Random Forest classifier can be classified under ensemble learning methods because it builds a series of decision trees during the learning or training step with an objective of generating a combination of their predictions. The Random Forest classifier can handle non-linear relationships in a dataset since linear relationships can neither capture non-linear dependencies nor address overfitting. One of these methods is Random Forest, which performs very well with varying types of data, missing entries, and interaction among variables. The capability of Random Forest in identifying the importance of variables is another factor that adds to model interpretability. In this research, Random Forest performed best among all models with regards to accuracy and thus is the most adaptive model in customer churn analysis.

The final model was chosen based on validation ROC-AUC scores. Among the other models tested, the Random Forest classifier produced the highest ROC-AUC score in both validation and testing phases, and thus this model will be used for further analysis.

VII. RESULTS AND DISCUSSION

To assess the performance of machine learning models, a variety of evaluation metrics such as ROC AUC, Precision, Recall, F1 Score, and Confusion Matrix were used. A combination of these assessment metrics gives information on both accuracy and relevance.

Among all three models, the Random Forest classifier performed better with a test ROC AUC evaluation metric of approximately 0.86, which shows excellent discrimination capability to distinguish churned and non-churned customers. Precision and recall measures have proved that this model strikes a balance in identifying churned customers with minimal false predictions of churned customers. The confusion matrix evaluation further supports model capability in appropriately categorizing most customers.

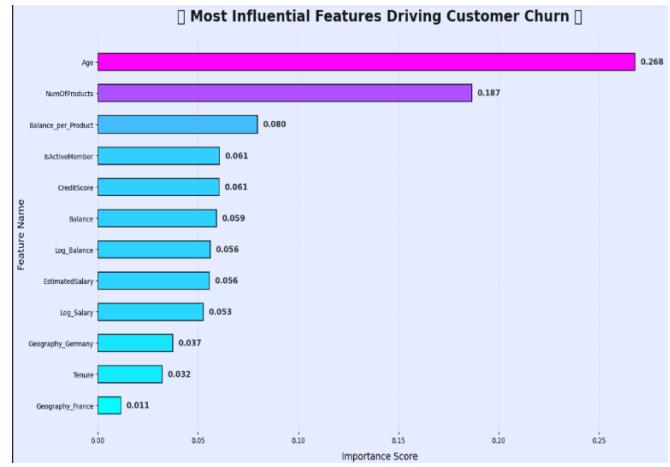


Fig. 1.1

To make it more interpretable, a SHAP Explainability technique was used. On conducting a Global SHAP Explanation, it was observed that variables Age, Credit Score, Number of Products, and Customer Activity Status have a prominent influence in identifying churn predictions. This is in accordance with real-world banking behavior, wherein increased Age, reduced Credit Scores, and reduced Customer Activity are indicators of a higher possibility of churn.

Furthermore, local SHAP explanations were explored to identify individual high-risk customers. Instance-level analysis sheds light on why customers were identified as likely to churn, making it possible to tailor a customer retention strategy accordingly.

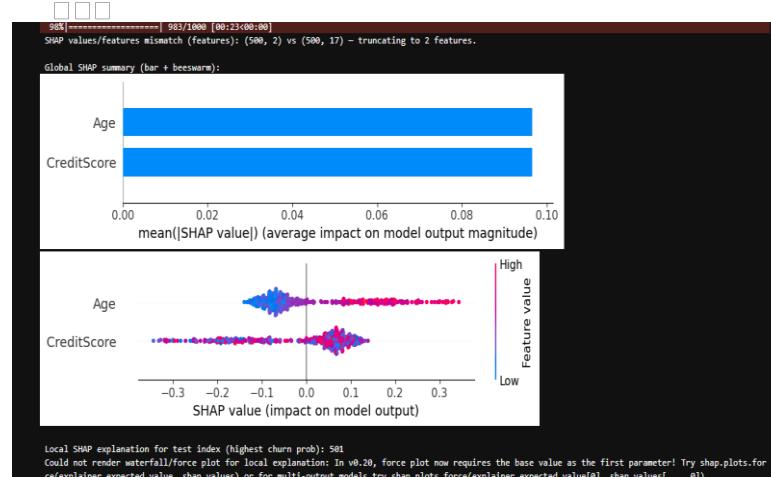


Fig. 1.2

VIII. MODEL COMPARISON TABLE

Model	Model Type	ROC-AUC (Test)	Strengths	Limitations
Logistic Regression	Linear Classifier	~0.78–0.80	<i>Simple, fast</i>	<i>Cannot capture non-linear relationships</i>
Random Forest	Ensemble (Bagging)	~0.86	<i>High accuracy</i>	<i>Less interpretable</i>
XGBoost	Ensemble (Boosting)	~0.83–0.85	<i>Powerful, scalable, regularized</i>	Sensitive to hyperparameters

IX. Discussion

The experimental results prove that machine learning models are efficient in predicting customer churn if combined with the right preprocessing and feature engineering techniques. Comparing the performances of the three implemented models—Logistic Regression, Random Forest, and XGBoost—obvious differences can be seen in terms of performance, interpretability, and ability to capture complex customer behavior.

Logistic Regression, while used as a baseline model, largely provided a framework for churn predictions in a very transparent and interpretable format. However, due to the inherent linear nature of the model, it could hardly capture the complex, non-linear relationships and interactions among customer attributes. Consequently, its performance was lower than the ensemble-based methods.

Among the rest of the models, Random Forest performed the best with a score of approximately 0.86 ROC-AUC on the test dataset. Its ensemble learning mechanism enables it to combine several decision trees that express complex nonlinear relationships and interactions among features. It can be attested that the model has good generalization power since the validation and test performance remains consistent. Moreover, the precision-recall balance of Random Forest is also well, hence applicable for real-world churn prediction scenarios where the number of false churn alarms may not be too high.

XGBoost showed competitive performance and handled the structured tabular data effectively. Being an ensemble learning method, it is based on gradient boosting. This algorithm works in an additive way, with each tree trying to correct the errors of the previously made predictions. However, in the case of this current study, XGBoost's results were somewhat lower than those of Random Forest, probably due to its sensitivity to changes in hyperparameter settings and the nature of the dataset.

In addition, SHAP-based explainability has been employed to address the interpretability challenge arising with ensemble models. Global SHAP analysis revealed that the most influencing features responsible for churn predictions were Age, Credit Score, No of Products, and Customer Active Status. These insights tally very well with real-world banking behavior, wherein customer engagement, financial stability, and demographic factors all play a major role in retention. More importantly, local SHAP explanations provided instance-level insights, thereby enabling the identification of high-risk customers and thus aided targeted retention strategies.

Overall, the discussion brings out the fact that although a number of models can effectively predict churn, ensemble-based models, when combined with explainable AI techniques, yield the best balance in predictive performance and business interpretability.

X. CONCLUSION

In this work, it is shown how machine learning model integration with an Explainable AI approach can benefit a customer churn prediction problem in a banking organization. The presented solution integrates tasks such as data

preprocessing, model selection, and SHAP analysis to offer accurate results with a transparent model.

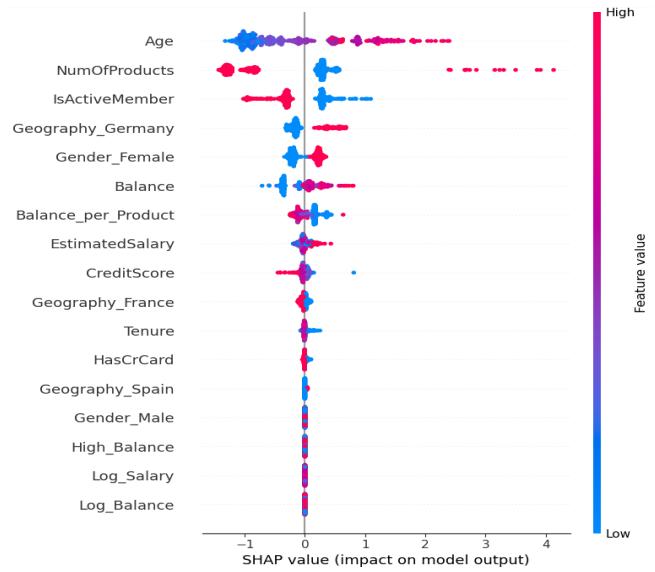


Fig. 1.3

The experimental results illustrate the effectiveness of ensemble models and, in particular, Random Forest in identifying intricate usage patterns of customers. Additionally, SHAP analysis can contribute to proactive customer retention activities since SHAP explanations deliver interpretable results. In summary, this study proposes a technique with a focus on interpretable machine learning in banking

XI. References

- [1] I. T. Jolliffe, “Principal Component Analysis”, Springer, 2011.
- [2] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” Proc. 22nd ACM SIGKDD, pp. 785–794, 2016.
- [3] L. Breiman, “Random Forests,” Machine Learning, vol. 45, pp. 5–32, 2001.
- [4] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
- [5] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” Advances in Neural Information Processing Systems, 2017.
- [6] A. Verbeke et al., “Predicting Customer Churn in the Telecommunications Sector”.
- [7] T. Fawcett, “An Introduction to ROC Analysis,” Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.
- [8] C. Molnar, “Interpretable Machine Learning,” 2nd ed., Lulu.com, 2022.

- [9] J. Brownlee, “Imbalanced Classification with Machine Learning,”Machine Learning Mastery, 2020.
- [13] A. Geron, “Hands-On Machine Learning with Scikit-Learn and TensorFlow,” O’Reilly Media, 2019.
- [14] D. Berrar, “Cross-Validation,” in Encyclopedia of Bioinformatics and Computational Biology, Elsevier, pp. 542–545, 2019.