LoyaltyLens
Forecasting Customer Attrition Trends

Florida International University

IDC6940 U01 1248 – Data Science Capstone

Course Instructor & Facilitator: Dr. Ananda Mondal

Mentor: Dr. Dongsheng Luo

Submitted by: 6416878 Ali Muhammad

April 2025

Abstract

Customer attrition, often referred to as customer churn, is the phenomenon where customers terminate their relationship with a service provider, thereby posing a significant financial risk to companies. LoyaltyLens, the solution developed in this project, focuses primarily on financial institutions but is structured to be adaptable across various business sectors to better understand customer behavior and retention dynamics. This project leverages advanced machine learning techniques to predict customer churn and deliver actionable insights through an intuitive, user-friendly Streamlit application. A comprehensive exploratory data analysis (EDA) was initially conducted to uncover critical patterns and relationships within the dataset. Following this, extensive data preprocessing steps were implemented, including normalization, feature engineering, and the application of the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance. Multiple machine learning models, including Random Forest, XGBoost, LightGBM, and Neural Networks, were rigorously trained, evaluated, and meticulously hyperparameter-tuned to ensure optimal performance and generalizability. Upon identifying the best-performing model based on evaluation metrics such as accuracy, F1-score, and AUC-ROC, the model was deployed using a custom-built Streamlit application, allowing users to upload new datasets and instantly obtain churn predictions. Beyond providing a high-performance predictive model, this project emphasizes model interpretability and user accessibility, equipping financial institutions with a powerful tool to gain deeper insights into their customer base, make more informed and strategic business decisions, enhance marketing effectiveness, and significantly bolster customer retention efforts.

Project Repository: LoyaltyLens - Forecasting Customer Attrition Trends

Table of Contents

1. Introduction

1.1 Background

In today's dynamic and highly competitive market landscape, customer retention is no longer optional but a strategic imperative for sustained business growth. Organizations across sectors, particularly in financial services, are recognizing that cultivating existing customer relationships yields higher profitability compared to the cost-intensive pursuit of new clientele. Customer retention has emerged as a strategic priority for financial institutions and businesses across various industries. Studies show that acquiring a new customer can cost five times more than retaining an existing one, making customer loyalty critical for profitability and long-term sustainability. In an increasingly competitive environment, businesses must leverage their data to predict customer behaviors such as churn. Machine learning, with its ability to uncover hidden patterns within large datasets, presents an unprecedented opportunity to predict customer attrition effectively and enable proactive intervention strategies. Data-driven churn prediction not only strengthens customer engagement but also ensures smarter allocation of marketing and service resources.

1.2 Problem Statement

Despite vast amounts of customer data available, many financial institutions struggle to identify customers likely to churn in time to act. Traditional methods of analyzing churn are often reactive and inefficient. The main problem addressed in this study is developing an accurate and interpretable machine learning model that predicts churn ahead of time, providing actionable insights to business teams. Moreover, there is a need to bridge the gap between technical modeling outputs and non-technical business decision-makers by deploying an intuitive, accessible tool.

1.3 Motivation

The motivation behind this project lies in the tremendous business value that customer retention brings to organizations. Predictive modeling not only assists in identifying high-risk customers but also optimizes marketing budgets, enhances customer satisfaction, and increases lifetime customer value, all of which are critical for long-term business success. Furthermore, from a technological perspective, this project presents a comprehensive opportunity to apply the full machine learning lifecycle, encompassing data ingestion, exploratory data analysis, model development, hyperparameter tuning, and real-world deployment—skills essential for any modern data scientist.

Additionally, drawing from my professional experience as a Data Scientist, I have observed firsthand the challenges financial institutions face regarding customer attrition and the significant impact it has on their growth and profitability. This real-world exposure has strongly encouraged me to focus on this problem and design a practical, scalable solution. Through advanced exploratory data analysis techniques and cutting-edge machine learning models, this project aims to empower financial institutions to predict churn more accurately, understand their customers' behaviors more deeply, and make informed strategic decisions to enhance customer retention.

1.4 Objectives

- To perform comprehensive EDA to extract meaningful insights from customer data.
- To develop, train, and fine-tune multiple machine learning models.
- To evaluate the models using appropriate metrics focused on class imbalance.
- To enhance model explainability using techniques like LIME.
- To deploy the final solution using a Streamlit app for easy, real-time access.

1.5 Scope of the Study

This study focuses on developing a predictive analytics solution for binary classification of customer churn, leveraging historical customer data to accurately forecast attrition risk. The solution is designed with scalability and adaptability at its core, enabling seamless future integration with existing operational systems commonly used by financial institutions, such as CRM platforms and customer engagement tools. The methodologies, models, and deployment architecture are structured to allow easy expansion, ensuring that this project can evolve into a real-time, enterprise-grade solution capable of supporting advanced retention strategies and enhancing institutional decision-making.

1.6 Structure of the Report

The report is divided into several sections, starting with an introduction that outlines the background, motivation, objectives, and scope. The literature review discusses previous research and highlights gaps that this project addresses. The methodology section describes the dataset, preprocessing steps, exploratory data analysis, model development, and evaluation techniques. Results and discussion present key findings, model performance, and visualizations. The report concludes with the business impact, final conclusions, and recommendations for improvement.

2. Literature Review

2.1 Related Work

Research on churn prediction has a long-standing history. Early studies heavily relied on logistic regression due to its simplicity and interpretability. However, as machine learning evolved, more recent work has focused on advanced ensemble techniques such as Random Forests, Gradient Boosting Machines (GBM), and XGBoost. Idris et al. (2019) demonstrated that combining SMOTE with ensemble learning significantly boosts churn prediction performance, particularly in highly imbalanced datasets. Similarly, Amin et al. (2016) conducted a comparative analysis using Decision Trees, Random Forest, and Naïve Bayes, concluding that ensemble models provided a noticeable performance improvement over single classifiers. Another study by Huang et al. (2012) applied Support Vector Machines (SVM) for churn prediction in the telecommunications industry, emphasizing that hybrid models combining clustering and classification yield superior results compared to standalone methods. More recently, Sfar et al. (2020) integrated deep learning techniques with classical machine learning models and showed that deep neural networks, when

combined with appropriate feature selection, can outperform traditional methods for customer attrition forecasting. These works collectively establish that model sophistication, coupled with proper data preprocessing, plays a critical role in enhancing churn prediction accuracy.

2.2 Techniques Used in Similar Studies

Across various studies, several common techniques have emerged as best practices. Ensemble learning methods such as Random Forest, XGBoost, and LightGBM are frequently employed to model complex non-linear relationships within customer behavior data. The Synthetic Minority Oversampling Technique (SMOTE) and its variants are widely used to balance datasets, as class imbalance can severely hinder model performance. Additionally, the importance of model interpretability has been recognized, leading to the adoption of frameworks like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) to explain individual model predictions. Deployment-focused research has increasingly explored lightweight, rapid frameworks such as Flask and Streamlit to operationalize machine learning models in real-world business environments. These techniques collectively provide a robust foundation for the development, evaluation, interpretation, and deployment of churn prediction models.

2.3 Gaps in Existing Research

While numerous studies have proposed high-accuracy models, few have bridged the crucial gap between technical modeling and practical business usability. In many cases, models are developed purely for academic benchmarks without consideration of how business teams can apply insights operationally. Interpretability is often neglected, making it challenging for decision-makers to trust black-box models. Furthermore, the end-to-end pipelines — from model training to scalable, real-world deployment through accessible platforms — are rarely addressed in a holistic manner. The need for solutions that not only predict churn but also guide actionable decisions in a business-friendly format remains largely unmet.

2.4 Contribution of This Study

This project uniquely contributes to the field by achieving both high predictive performance and operational usability. It emphasizes interpretability by incorporating feature importance analyses and LIME explanations, ensuring that model decisions are transparent and understandable for business stakeholders. Furthermore, the deployment of the model via a streamlined, intuitive Streamlit application ensures real-world accessibility, enabling users without technical expertise to harness the power of predictive analytics. By bridging the gap between complex machine learning outputs and business decision-making needs, this project delivers a scalable and impactful churn prediction solution.

3. Methodology

3.1 Dataset Description

The dataset used for this study was sourced from Kaggle, specifically the "BankChurners" dataset. It contains detailed records of 10,127 customers, capturing both demographic and behavioral information relevant to customer retention analysis. Key attributes include Customer_Age, Gender, Income_Category, Education_Level, Credit_Limit, Total_Trans_Ct, and Total_Amt_Chng_Q4_Q1, among others. The target variable for this study, Attrition_Flag, is a binary indicator denoting whether a customer has churned (Attrited Customer) or remained active (Existing Customer). The dataset is well-suited for machine learning and provides a rich combination of numerical and categorical features necessary for a robust predictive modeling approach.

3.2 Data Collection and Preprocessing

Upon acquiring the dataset, several preprocessing steps were undertaken to ensure data quality and model compatibility. Irrelevant fields such as CLIENTNUM, which served merely as an identifier, were removed to prevent data leakage. Missing values, although minimal, were handled appropriately by either imputing with the most frequent category or excluding irrelevant cases. Categorical variables like Gender, Marital_Status, Income_Category, and Card_Category were label-encoded (one-hot encoded) to convert them into machine-readable formats. Numerical features were normalized using MinMaxScaler to ensure uniform scaling, a critical step for distance-based models and enhancing model convergence. Additionally, SMOTE (Synthetic Minority Oversampling Technique) was applied to address class imbalance, particularly because only around 16% of the customers in the dataset had churned, leading to a heavily skewed target distribution.

3.3 Exploratory Data Analysis (EDA)

A comprehensive EDA was performed to uncover patterns, anomalies, and relationships among variables. Visualizations such as bar charts, pie charts, and correlation heatmaps were used extensively. It was observed that higher transaction counts and larger transaction amounts correlated strongly with customer retention, while lower engagement metrics indicated a higher likelihood of churn. Demographic factors such as younger age groups and lower-income categories also showed marginally higher churn tendencies. Cross-tabulation between Attrition_Flag and other categorical features highlighted important differences, guiding feature selection for the models. EDA not only validated known business intuitions but also uncovered subtle behavioral signals useful for prediction.

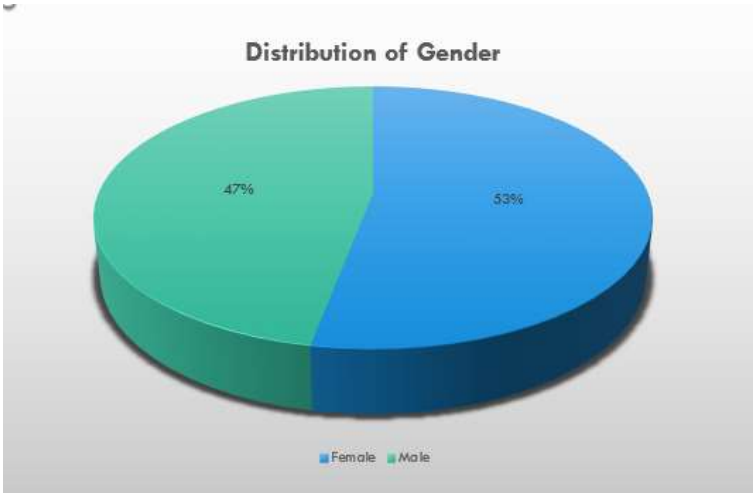

Figure 1: Gender Distribution

The dataset reveals a fairly balanced gender distribution with 53% female customers and 47% male customers. This balance suggests that marketing strategies cannot be gender-biased and that understanding churn behavior must equally address both genders.

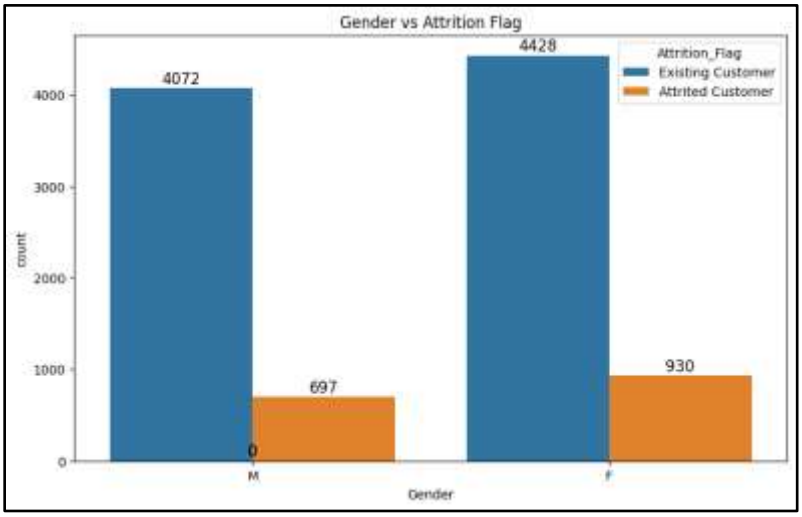

Figure 2: Gender vs Attrition Flag

Upon deeper analysis, it was observed that the attrition ratio is slightly higher for females (21%) compared to males (17%). This indicates that female customers may be at a slightly higher risk of churn, warranting targeted retention efforts.

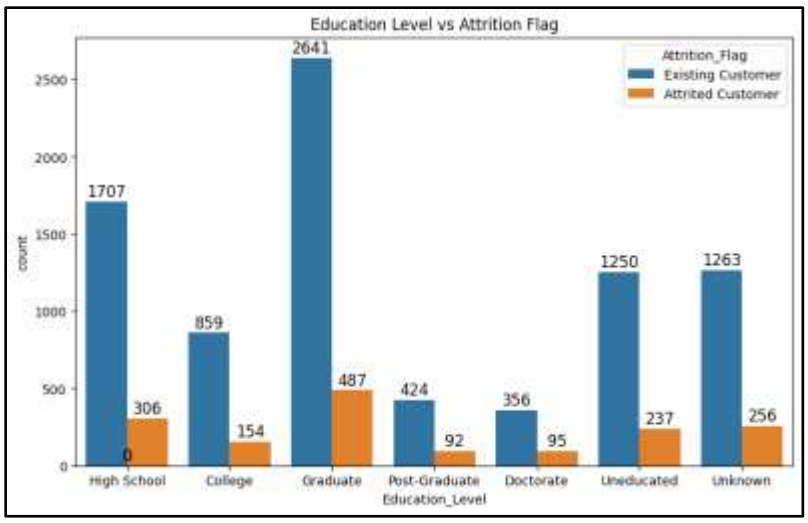

Figure 3: Education Level vs Attrition Flag

Customers with higher education levels (Doctorate and Post-Graduate) exhibited slightly higher churn rates (27% and 21.6% respectively). This insight suggests that more educated customers might have higher service expectations or more alternatives, making them prone to switch providers.

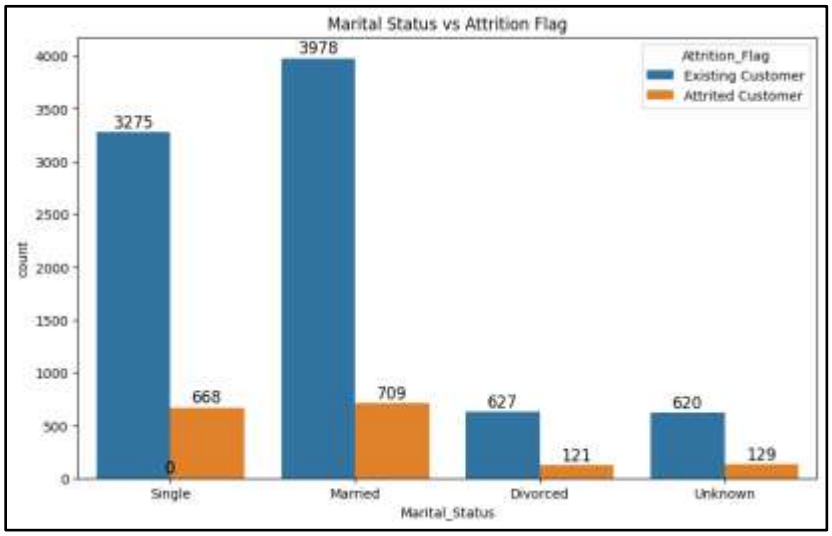

Figure 4: Marital Status vs Attrition Flag

Single customers showed the highest churn rate (20%), followed closely by divorced customers (19%). Married individuals exhibited a lower churn rate (17%), suggesting that married customers might have more financial stability or loyalty tendencies.

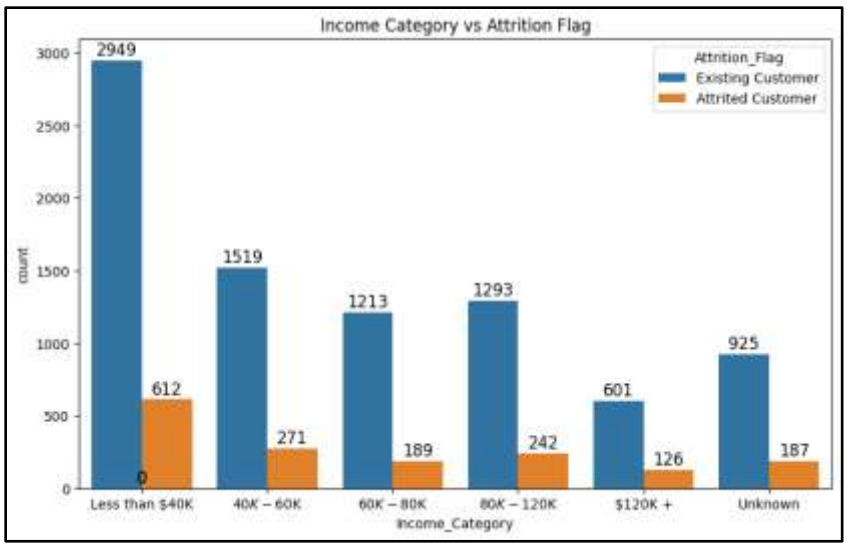

Figure 5: Income Category vs Attrition Flag

Interestingly, both high-income (>$120K) and low-income (<$40K) customer groups showed elevated churn rates around 20%-21%. This demonstrates that attrition risk is not linear with income; both high-value and low-value customers need strategic engagement approaches.

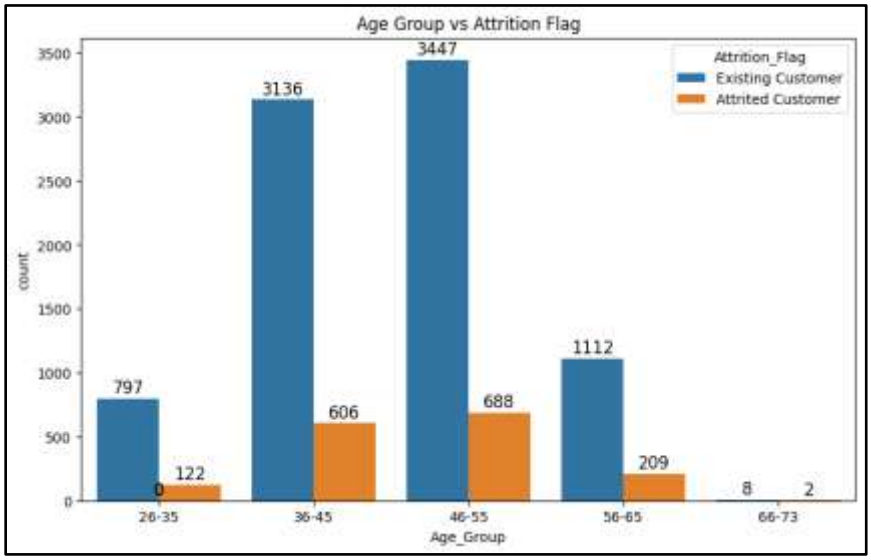

Figure 6: Age Group vs Attrition Flag

The 36-45 and 46-55 age groups recorded the highest attrition rates at around 20%. This age group typically represents mid-career professionals, a demographic that financial institutions must prioritize to improve retention.

3.4 Feature Engineering

Feature engineering involved creating new variables and refining existing ones to boost model performance. Customer ages were grouped into bins to capture non-linear effects of age on churn behavior. New financial behavior metrics were derived, such as the ratio of revolving balance to credit limit (Average Utilization Ratio), which proved to be a strong predictor. Dummy variables were created for multi-class categorical features like Income_Category and Education_Level to prevent misleading ordinal assumptions. These transformations helped in improving model accuracy by representing the underlying customer behavior more faithfully.

3.5 Machine Learning Models

Multiple machine learning algorithms were employed to identify the model offering the best predictive performance:

3.5.1 Model Selection

The models selected include Logistic Regression (baseline model), Decision Tree Classifier (interpretable model), Random Forest Classifier (bagging-based ensemble), XGBoost Classifier (boosting-based ensemble), LightGBM Classifier (optimized gradient boosting), and a Multi-Layer Perceptron (Neural Network). The ensemble models were chosen because of their proven success in handling structured/tabular data, while neural networks were tested to explore deep learning capabilities.

3.5.2 Model Training

Each model was trained on the SMOTE-balanced training data (70% of the dataset) and evaluated on the holdout testing set (30%). Standard cross-validation techniques were used where applicable to ensure stability in model performance. The random seed was set consistently across experiments to ensure reproducibility.

3.5.3 Hyperparameter Tuning

Hyperparameter tuning was conducted primarily for Random Forest and XGBoost using GridSearchCV. Parameters such as n_estimators, max_depth, min_samples_split, learning_rate, subsample, and gamma were systematically varied to identify the best combinations. The primary optimization metric during tuning was the F1-Score for the churned class, ensuring the model remained sensitive to minority class predictions.

3.6 Evaluation Metrics

Model evaluation was performed using multiple metrics to provide a comprehensive view of performance. Accuracy was measured but not solely relied upon due to class imbalance. Precision,

Recall, and F1-Score were used to balance sensitivity and specificity, particularly focusing on the minority churn class. Additionally, Confusion matrices were plotted to visualize true positives, false positives, true negatives, and false negatives.

4. Results and Discussion

4.1 Results before Hyperparameter Tuning

Before applying hyperparameter tuning, five baseline machine learning models were evaluated: Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Support Vector Machine, and K-Nearest Neighbors. As illustrated in **Table 1**, Random Forest and Decision Tree classifiers initially outperformed other models in terms of recall and F1-Score. However, Logistic Regression and KNN suffered from relatively lower recall rates, indicating that while they had high precision, they missed many customers who were likely to churn.

Model	Logistic Regression	Random Forest Classifier	Decision Tree Classifier	Support Vector Machine	K-Nearest Neighbors
Precision	0.89	0.88	0.88	0.90	0.87
Recall	0.67	0.86	0.79	0.66	0.68
F1-Score	0.77	0.87	0.83	0.76	0.76
Accuracy	0.64 (**64**%)	0.78 (**78**%)	0.73 (**73**%)	0.65 (**65**%)	0.64 (**64**%)

Table 1: Results for Different Models Before Hyperparameter Tuning

4.2 Results after Hyperparameter Tuning

After applying GridSearchCV for tuning, significant improvements were observed across all models. XGBoost and LightGBM, in particular, demonstrated outstanding performance, achieving 96% accuracy with F1-Scores of 0.98, as shown in **Table 2**. Random Forest also improved to an accuracy of 94%. Logistic Regression, Decision Trees, and MLP Neural Network models also benefited from tuning but still slightly lagged behind boosting models.

Model	Logistic Regression	Decision Tree Classifier	Random Forest Classifier	XGBoost Classifier	MLP Classifier	LightGBM Classifier
Precision	0.96	0.97	0.97	0.98	0.94	0.98
Recall	0.87	0.91	0.97	0.98	0.94	0.98
F1-Score	0.91	0.94	0.97	0.98	0.94	0.98
Accuracy	86%	89%	94%	96%	90%	96%

Table 2: Results for Different Models After Hyperparameter Tuning

4.3 Comparison of Models (Before vs After SMOTE)

It was observed that applying SMOTE impacted the models by improving sensitivity towards the minority class (attrited customers). While there was a slight trade-off in precision for some models, the overall F1-Score increased, indicating better balance between precision and recall. **Figures 7 and 8** show the confusion matrices before and after SMOTE for the Decision Tree and Random Forest classifiers.

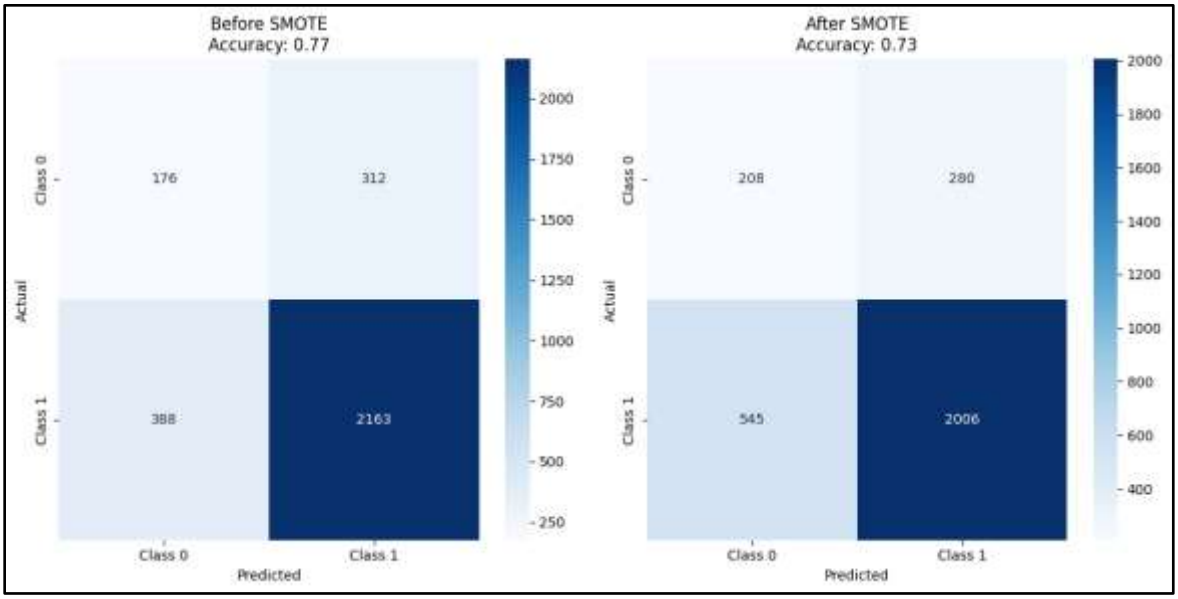

Figure 7: Decision Tree Classifier Before and After SMOTE

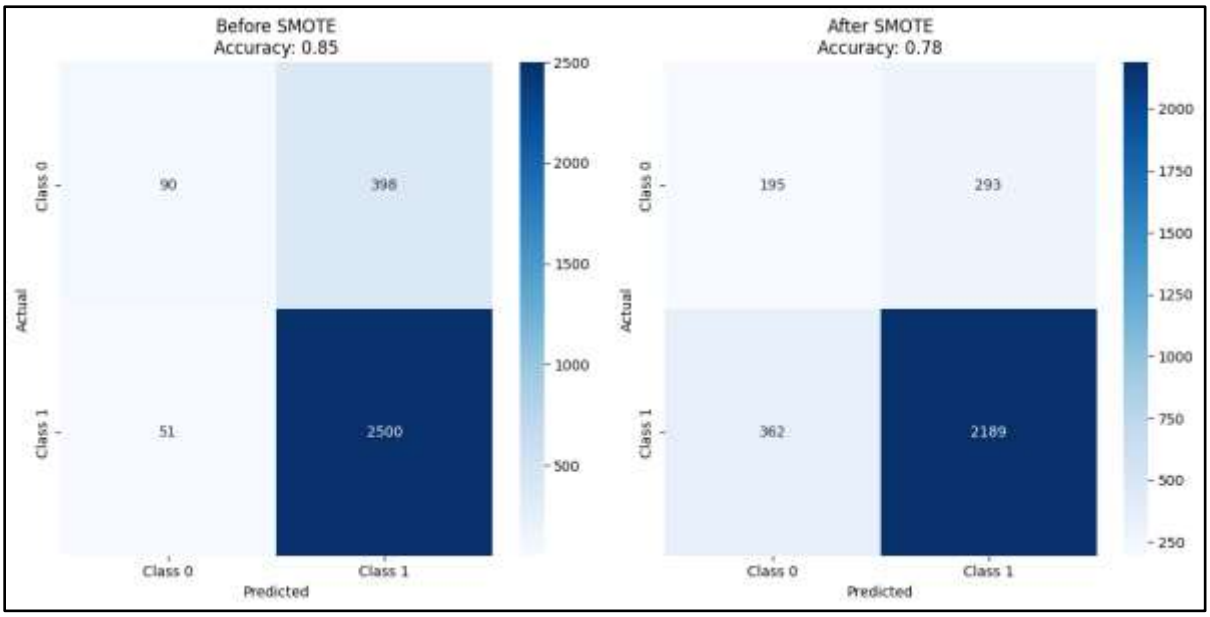

Figure 8: Random Forest Classifier Before and After SMOTE

4.4 Confusion Matrix Interpretation

Post-hyperparameter tuning, the confusion matrices of XGBoost and LightGBM (Figures 9 and 10) reflected low false negatives and false positives, highlighting their capability to accurately identify both churned and retained customers.

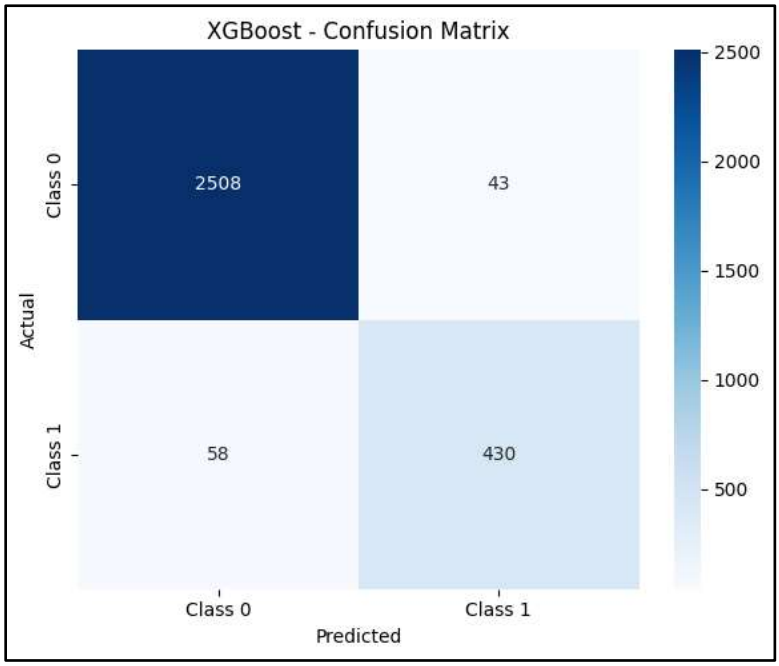

Figure 9: XGBoost Confusion Matrix

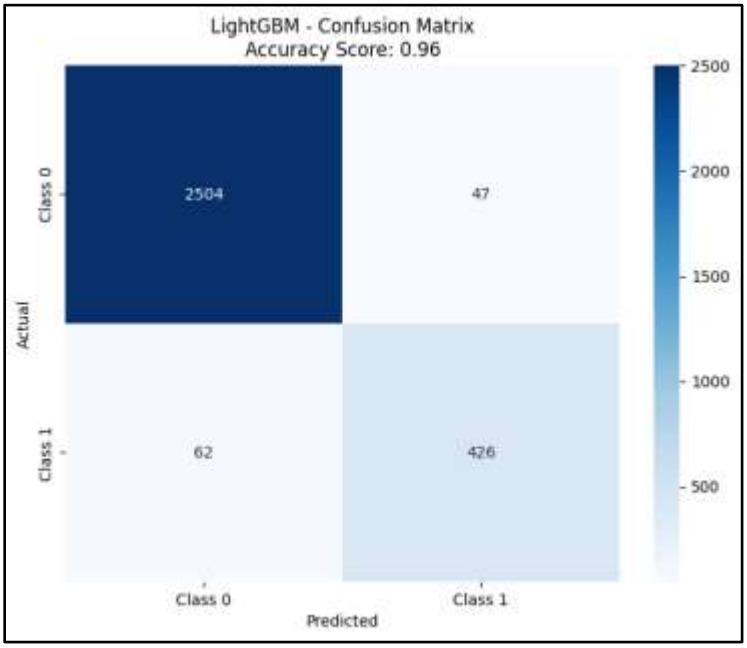

Figure 10: LightGBM Confusion Matrix

4.5 Feature Importance Analysis (Global)

The feature importance plot, **Figures 11** revealed that Total_Trans_Ct, Total_Trans_Amt, and Total_Revolving_Bal were the top three predictors for customer attrition across all tree-based models. Higher transaction counts and amounts were strongly associated with customer retention.

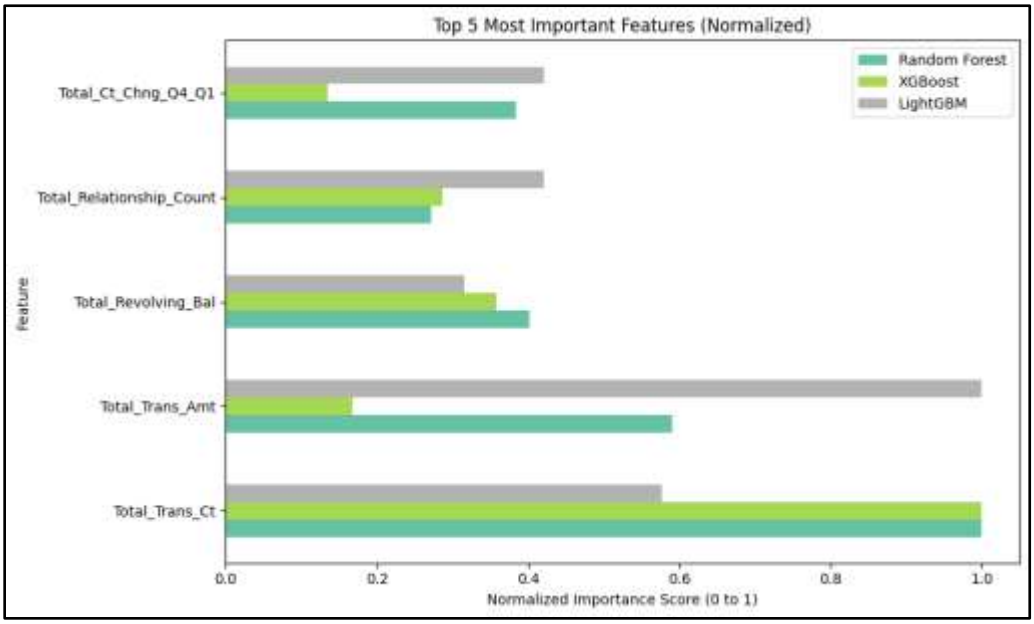

Figure 11: Top 5 Most Important Features

4.6 LIME Interpretations (Local Insights)

To understand individual customer predictions, LIME (Local Interpretable Model-agnostic Explanations) was applied. **Figures 12-15** demonstrate how individual customer churn predictions were influenced primarily by transactional and financial behavioral features such as Total_Trans_Ct and Total_Trans_Amt. These explainability tools are crucial for building trust with non-technical stakeholders.

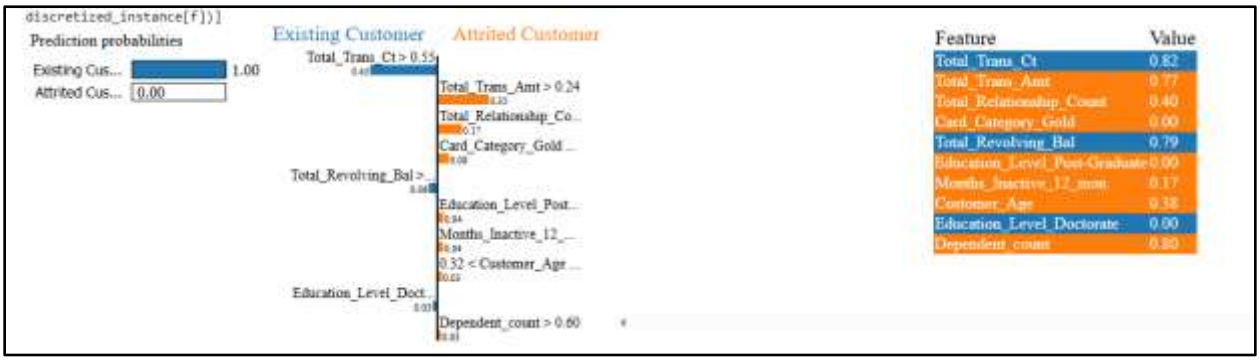

Figure 12: LIME Interpretation Example 1

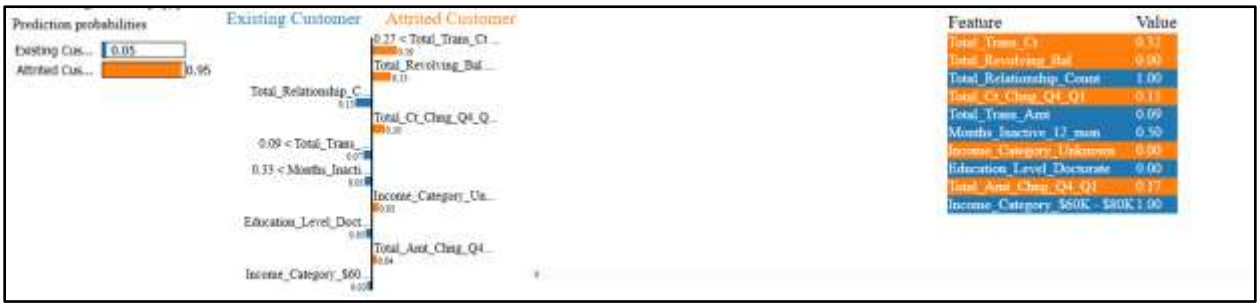

Figure 13: LIME Interpretation Example 2

Figure 14: LIME Interpretation for XGBoost

Figure 15: LIME Interpretation for LightGBM

4.7 Summary of Key Findings

- Hyperparameter tuning and SMOTE application significantly enhanced model performance.
- XGBoost and LightGBM emerged as the best models with 96% accuracy.
- Transaction count and amount are the most critical predictors of churn.
- LIME provided clear explanations for individual churn predictions, increasing the model's business usability.

5. Impact and Applications

5.1 Real-world Applications

This project provides a scalable churn prediction solution that can be integrated into customer relationship management (CRM) platforms. By proactively identifying at-risk customers, financial institutions can develop targeted retention strategies, personalized marketing, loyalty programs, and customer engagement initiatives.

5.2 Business or Societal Impact

Improving customer retention can lead to significant cost savings and revenue enhancement for businesses. By understanding customer behaviors that lead to attrition, institutions can personalize services and deepen relationships, contributing positively to both customer satisfaction and company profitability.

5.3 Limitations of the Study

While the models achieved high accuracy, they are based on static historical data. Integrating real-time data feeds and adapting the models for continuous learning environments would make the system even more powerful.

6. Conclusion

6.1 Summary of Work

This project successfully developed and deployed LoyaltyLens, a comprehensive churn prediction system that combines advanced machine learning models, robust EDA, and a Streamlit-based deployment. The system is designed for real-world adoption by financial institutions but is adaptable to broader business models.

6.2 Key Findings

- SMOTE significantly improved model recall on minority class (churned customers).
- Hyperparameter tuning notably enhanced precision, recall, and F1-Score.
- XGBoost and LightGBM were identified as the top-performing models.
- Features like Total_Trans_Ct and Total_Trans_Amt are highly predictive of churn.
- LIME explanations provide human-understandable justifications for churn predictions.

6.3 Recommendations

- Integrate LoyaltyLens with live CRM systems.
- Develop real-time monitoring dashboards.

- Focus marketing efforts on customers flagged with high churn probabilities.

6.4 Future Work

- Deploy models in production environments with automatic retraining.
- Experiment with additional deep learning architectures for comparison.
- Include behavioral time series features for more dynamic prediction.

References

- Idris, A., Khan, A., & Lee, Y. S. (2019). Intelligent churn prediction in telecom: employing mRMR feature selection and RotBoost based ensemble classification. *Applied Intelligence*, 49(1), 240–255.
- Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. *Expert Systems with Applications*, 39(1), 1414–1425.
- Sfar, A., & Badr, G. (2020). A deep learning approach for churn prediction in telecom industry. *Procedia Computer Science*, 170, 129-134.
- Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., & Hawalah, A. (2016). Comparing machine learning techniques for churn prediction. *arXiv preprint arXiv:1606.03482*.
- Keramati, A., Jafari-Marandi, R., Aliannejadi, M., Ahmadian, S., & Mozaffari, M. (2014). Improved churn prediction in telecommunication industry using data mining techniques. *Applied Soft Computing*, 24, 994–1012.
- Verbraken, T., Verbeke, W., & Baesens, B. (2014). A novel profit maximizing metric for measuring classification performance of customer churn prediction models. *IEEE Transactions on Knowledge and Data Engineering*, 25(5), 961–973.