BUSINESS REPORT

CUSTOMER CHURN

Presented by
Prabha Bharati

Table Of Contents

Introduction

An E-Commerce company (or DTH provider) is facing intense competition in the current market, resulting in difficulties retaining existing customers. The company is particularly concerned about account churn, where the loss of a single account can mean losing multiple customers. The goal is to develop a churn prediction model to identify accounts at risk of leaving and provide targeted offers to retain them.

Need of the Study/Project

The study is necessary for the churn because it has multifaceted impact and opportunity. It has business impact, social impact and business opportunity.

Business Impact:

1. **Revenue Loss**: Losing accounts results in a significant loss of revenue, as each account often comprises multiple customers.
2. **Customer Acquisition Costs**: Acquiring new customers is typically more expensive than retaining existing ones. Effective churn management can help reduce these costs.
3. **Customer Lifetime Value (CLV)**: By retaining more customers, the company can maximize the lifetime value of each account, improving long-term profitability.
4. **Market Position**: In a competitive market, high churn rates can damage the company's reputation and market share.

Social Impact:

1. **Customer Satisfaction**: Identifying and addressing the reasons behind churn can lead to improved customer satisfaction and loyalty.
2. **Employment**: Reduced churn can contribute to business stability and growth, which in turn can support employment within the company.
3. **Community Engagement**: Businesses with lower churn rates often have more resources to invest in community engagement and corporate social responsibility initiatives.

Business Opportunity:

1. **Data-Driven Decision Making**: Implementing a churn prediction model allows the company to use data analytics to make informed decisions about retention strategies.
2. **Personalized Marketing**: The model enables the creation of segmented offers tailored to specific customer needs, improving the effectiveness of marketing campaigns.
3. **Operational Efficiency**: Understanding the drivers of churn helps optimize customer service operations and resource allocation.
4. **Competitive Advantage**: Reducing churn can provide a competitive edge in the market, allowing the company to retain a larger customer base than its competitors.

By developing a churn prediction model, the company can not only improve its financial health but also contribute positively to its customers and the broader community. This study offers a valuable opportunity to leverage data analysis for business growth and social impact.

EXPLORATORY DATA ANALYSIS and Business Implication :

Uni-variate Analysis

Continuous Attributes:

- **Distribution and Spread:**
  - Visualizing histograms and summary statistics (mean, median, standard deviation) for attributes like Tenure, Service_Score, rev_per_month, etc., to understand their distribution across the data set.

- Using box plots to identify outliers and understand the range and distribution of each variable relative to key factors such as churn status.

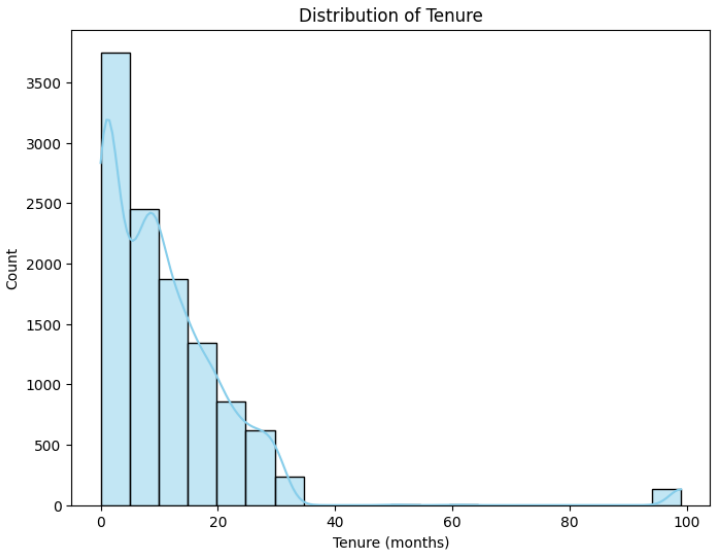

Fig. 01: Histogram for Distribution of Tenure

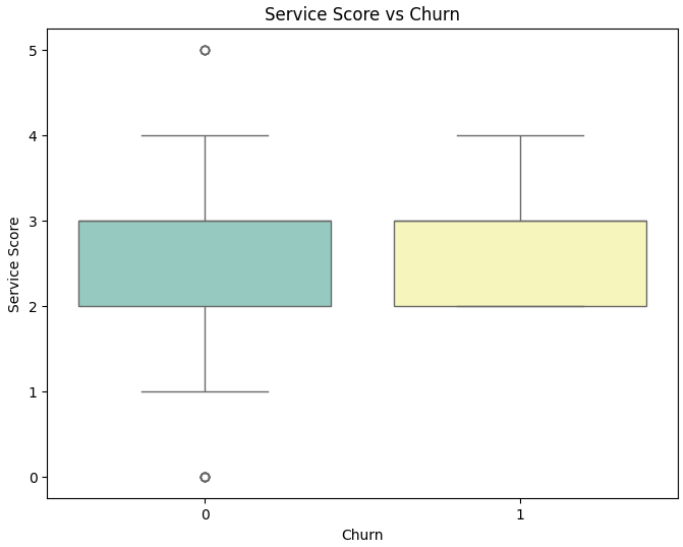

Fig. 02: Box-plot for Service Score Vs Churn

Categorical Attributes:

- **Distribution of Data:**
    - Utilizing count plots to visually depict the distribution of categories within variables like City_Tier, Payment, Gender, etc.

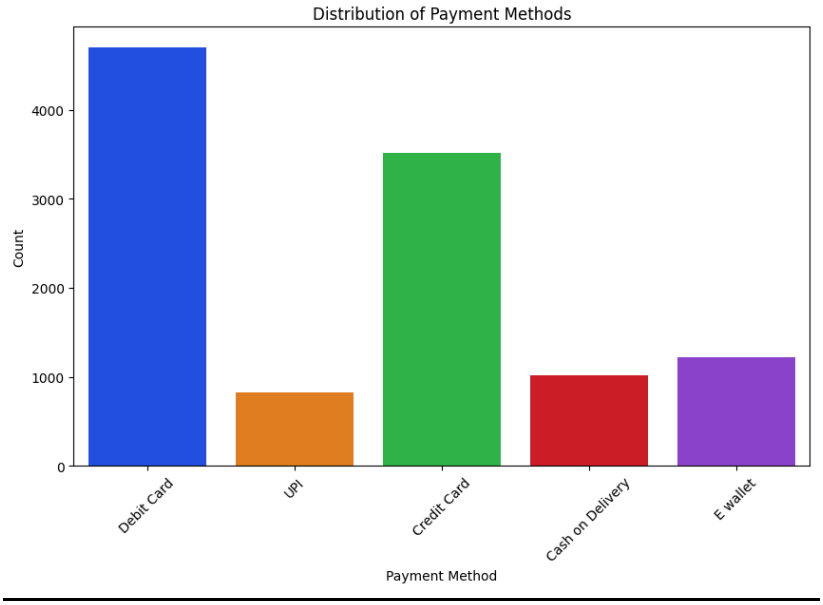

Fig. 03: Countplot for Distribution of Payment Methods

Bi-variate Analysis

Relationship Between Variables:

- **Correlations:**
    - Calculating correlation coefficients (e.g., Pearson correlation for numeric variables) to uncover relationships between continuous variables.

- Visualizing correlations using heat-maps to quickly discern which variables exhibit strong correlations, aiding in understanding potential interactions or dependencies.
- Using Violin Plot to show the distribution of a numeric variable across different categories.

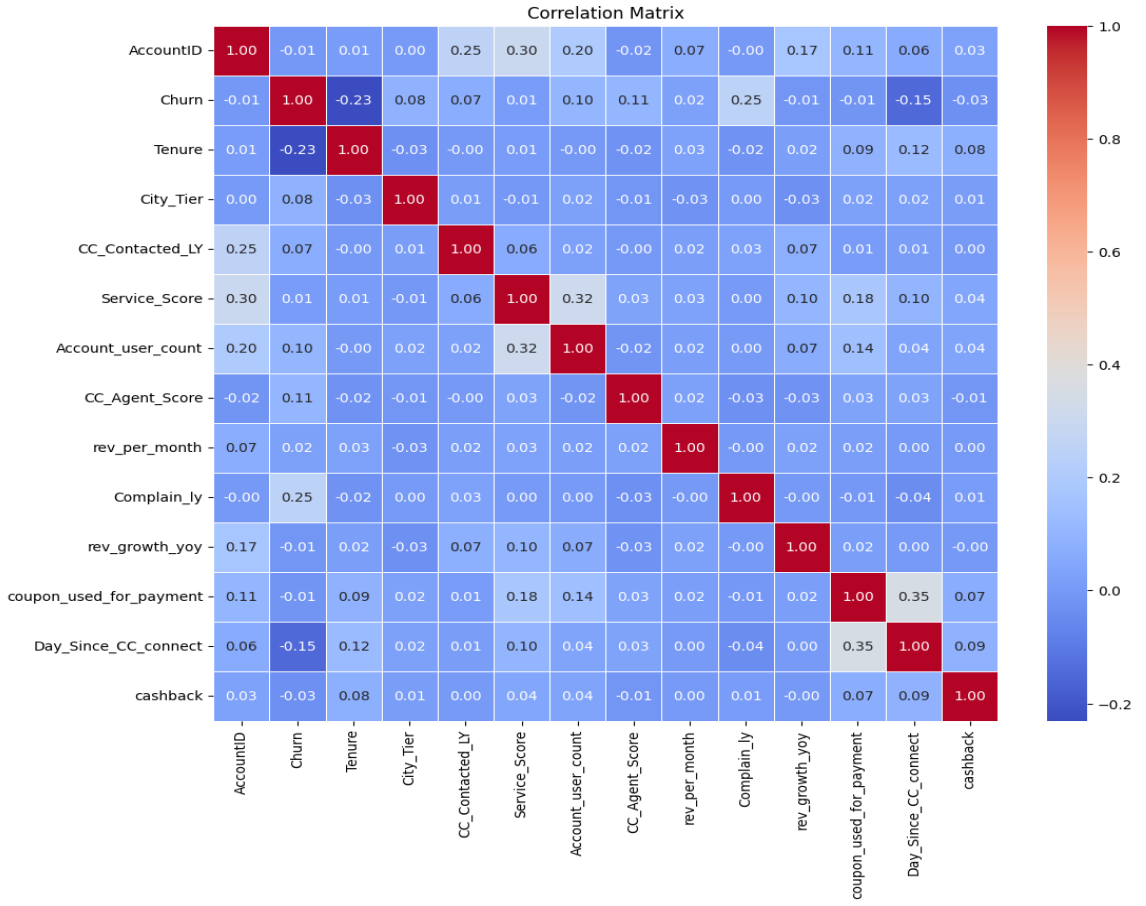

Fig. 04: Correlation Heatmap

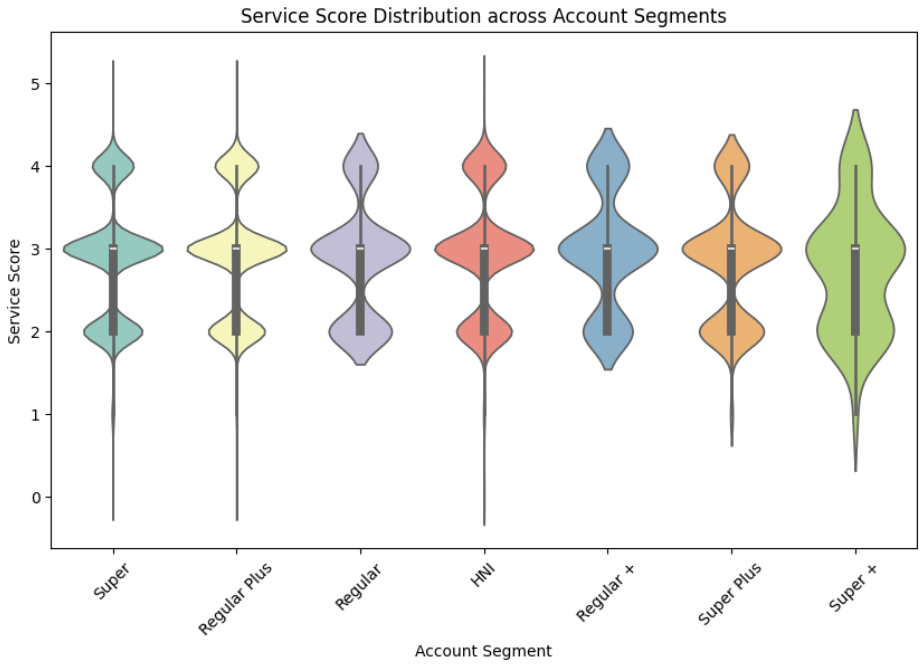

Fig. 04: Violin Plot showing distribution of a numeric variable across different categories.

Multi-variate Analysis

Relationship Between Variables:

Pairwise Scatter Plot Analysis:

- **Purpose:** To explore the relationships between all pairs of continuous variables in the dataset, providing a comprehensive view of how each numeric variable interacts with the others.
- **Method:** By creating a pairplot, scatter plots are generated for every pair of numeric variables. This visualization includes histograms on the diagonal to show the distribution of each individual variable. If a categorical variable is provided as hue, the plots will be colored based on this category, adding context to the relationships.
- **Interpretation:** The pairwise scatter plots enable the identification of patterns, correlations, and potential interactions between all pairs of numeric variables. By examining these plots, one can detect linear or non-linear relationships, clusters, or anomalies, and understand how variables relate to one another. This analysis aids in recognizing variable dependencies, multicollinearity, and any interesting trends or groupings within the data.

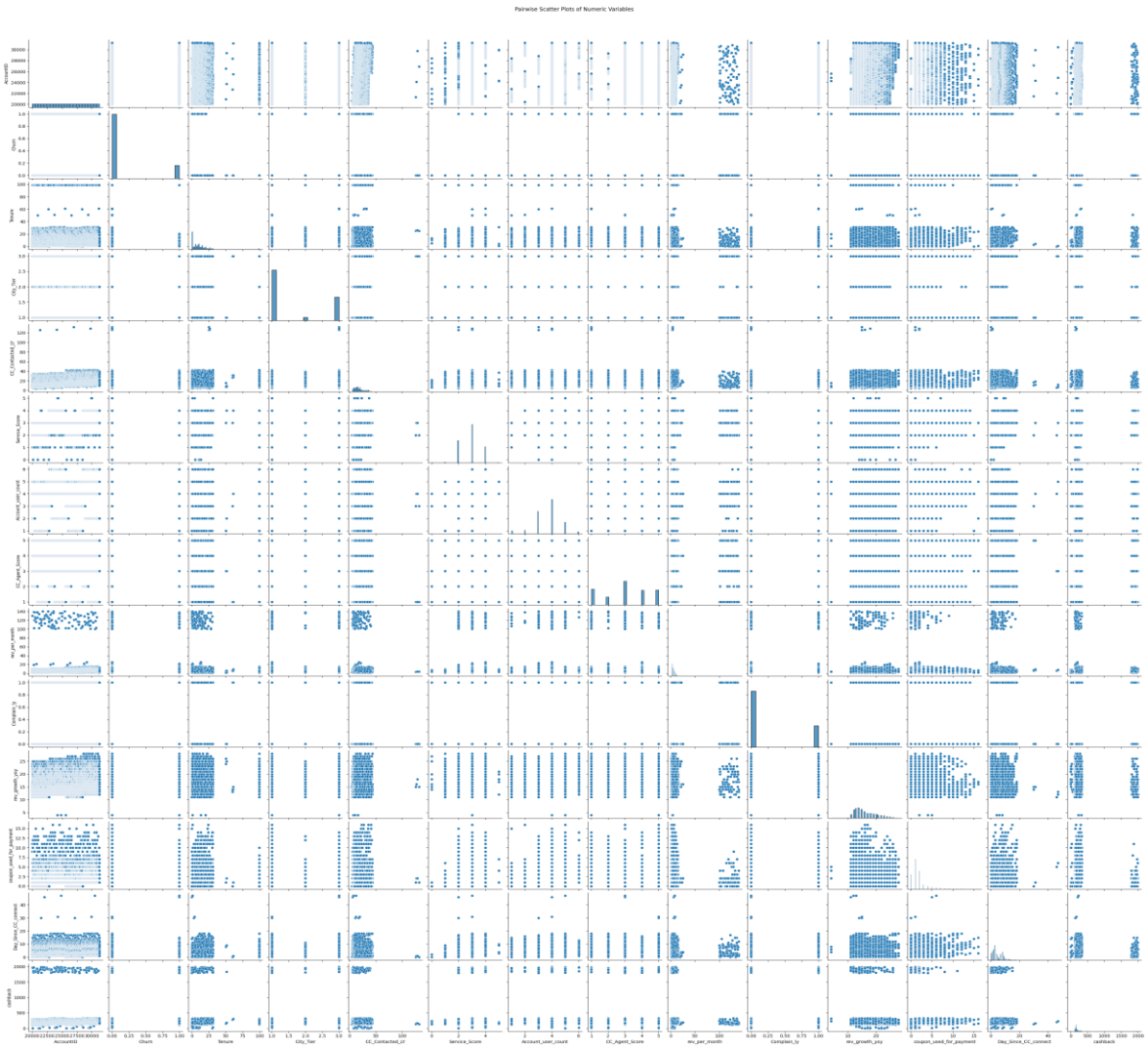

Fig. 05:Multivariate Analysis through Pairwise Scatter plot

Explanation :

Univariate Analysis Insights

Histograms and Box Plots:

- **Purpose:** These visualizations reveal the distribution and variability of continuous variables like Tenure and Service Score. Histograms provide insights into the frequency distributions, showing peaks and patterns across their ranges. Box plots complement this by identifying outliers and quartile ranges, essential for detecting anomalies or unusual patterns in the data.

Count Plots:

- **Purpose:** For categorical variables such as Payment methods and Gender, count plots illustrate the composition and prevalence of different categories within the data set. They offer a straightforward view of categorical variable distributions, helping to understand relative frequencies and proportions of each category.

Bivariate Analysis Insights

Correlation Heatmap:

- **Purpose:** The heatmap calculates and visually represents correlation coefficients between pairs of numeric variables. It plays a crucial role in identifying potential multicollinearity among predictors, guiding feature selection and model building. High correlations indicate possible redundant or highly influential variables that could impact model performance.

Violin Plots:

- **Purpose:** These plots show the distribution of a numeric variable across different categories, such as Service Score across various account segments. They provide deeper insights into how numeric variables vary within groups, facilitating visual comparisons of distributions. This helps in understanding differences or similarities across categories.

Multivariate Analysis Insights

Pairwise Scatter Plot Analysis:

- **Purpose:** It provides a comprehensive view of relationships between all pairs of continuous variables in the dataset. This analysis helps identify patterns, correlations, and interactions among variables, enabling insights into variable dependencies, potential clusters, and anomalies that inform model refinement and strategic decisions.

Data Cleaning and Pre-processing

In our data cleaning process, we focused on converting some columns to the correct type and handling missing values to ensure the data set is accurate and reliable. Here is a detailed explanation of the steps taken and the outcomes:

1. Converting Columns to Numeric Types

We found that several columns in our dataset, which should contain numerical data, were stored as text. These columns include:

- Tenure
- Account user count
- Revenue per month
- Revenue growth year-over-year
- Coupons used for payment
- Days since last customer care contact
- Cashback

We converted these columns to the correct numerical format. This step is crucial because it allows us to perform accurate calculations and analysis on these data points.

2. Handling Missing Values

We addressed missing values in two ways, depending on whether the data was numerical or categorical:

- **Numerical Data**: For columns with numbers, we filled in the missing values with the median value of each column. The median is the middle value and is less affected by extremely high or low values than the average.
- **Categorical Data**: For columns with categories (like 'City Tier' or 'Payment Method'), we filled in the missing values with the most common category. This method helps maintain the overall pattern of the data.

3. Verifying Data Cleaning

After cleaning the data, we checked to make sure there were no missing values left. We also confirmed the number of rows and columns in our dataset and summarized the key statistics for our numerical columns.

Summary Statistics for Key Numerical Columns

Here are some important statistics for a few key columns after cleaning:

Tenure (how long a customer has been with us):

- Total data points: 11,260
- Average tenure: 10.25 months
- Typical (median) tenure: 9 months
- Shortest tenure: 0 months
- Longest tenure: 37 months

Revenue per Month:

- Total data points: 11,260
- Average monthly revenue: $5.25
- Typical (median) monthly revenue: $5
- Lowest monthly revenue: $1
- Highest monthly revenue: $13

Cashback:

- Total data points: 11,260
- Average cashback: $177.20
- Typical (median) cashback: $165
- Lowest cashback: $74.50
- Highest cashback: $270.50

By converting columns to the correct types and addressing missing values, we have made sure our dataset is clean and ready for further analysis. This process helps ensure that the insights we derive from this data are accurate and reliable, providing a solid foundation for making informed business decisions.

Outlier Treatment Summary

In our analysis, we identified and treated outliers in four key areas of our dataset: Tenure, Customer Contacts Last Year (CC_Contacted_LY), Monthly Revenue (rev_per_month), and Cashback. Here's a simplified explanation of what was done and why it matters:

1. Identifying Outliers

Outliers are data points that differ significantly from other observations. They can distort results and lead to misleading insights. To ensure our analysis is accurate, we need to address these outliers.

2. Setting Boundaries

For each variable, we calculated acceptable ranges based on statistical methods. If a value fell outside these boundaries, it was considered an outlier.

3. Capping Outliers

Instead of removing outliers, we capped them at the maximum or minimum acceptable values. This means that any value that was too high was reduced to the maximum acceptable value, and any value that was too low was increased to the minimum acceptable value.

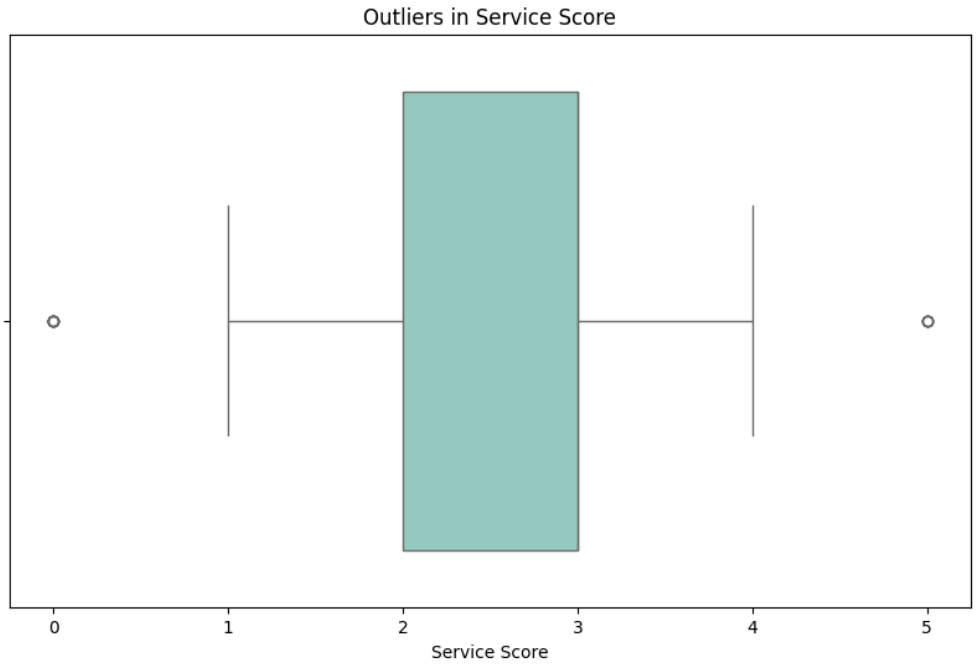

Fig. 05 : Presence of Outliers in Service Score

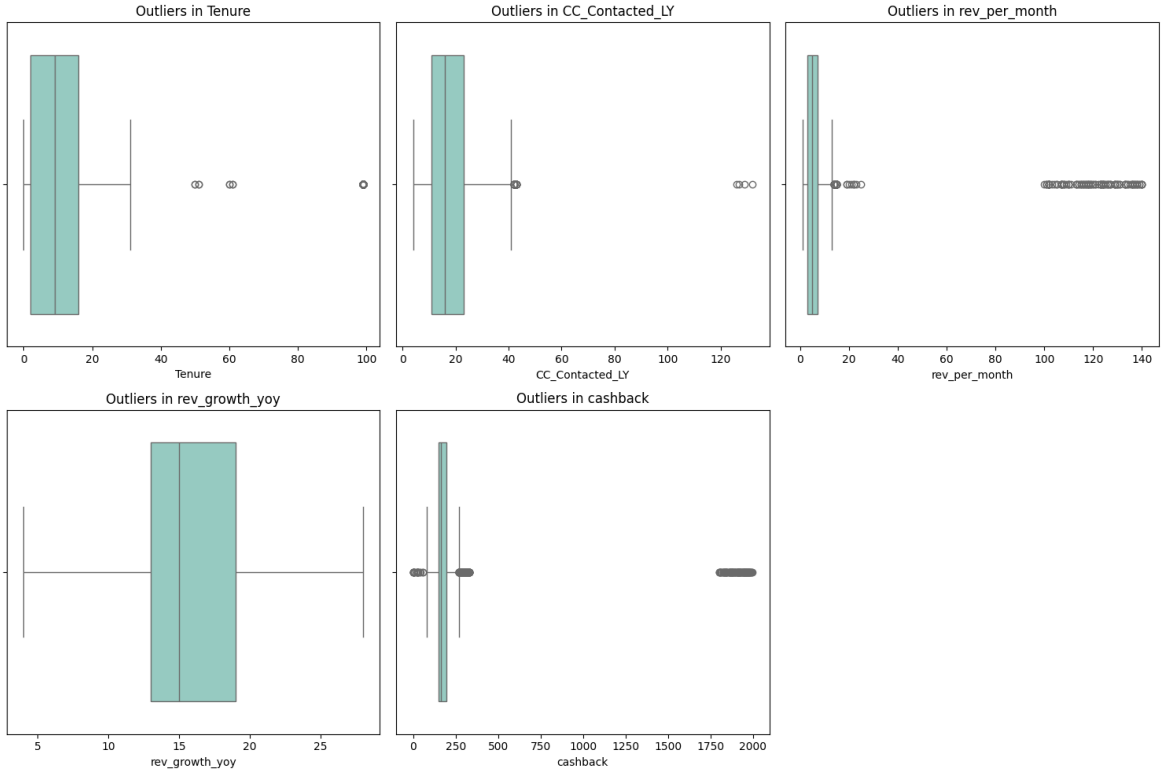

Fig. 06 : Presence of Outliers Across Categories

Results After Outlier Treatment

Tenure (Length of Customer Relationship)

- **Original Range**: 0 to much higher than acceptable
- **Adjusted Range**: 0 to 37 months
- **Impact**: This ensures that no customer is shown as having an excessively long tenure, which could skew average tenure calculations.

Customer Contacts Last Year (CC_Contacted_LY)

- **Original Range**: Low of 4 to a high beyond acceptable
- **Adjusted Range**: 4 to 41 contacts
- **Impact**: This helps us understand customer contact patterns without the distortion from excessively high contact counts.

Monthly Revenue (rev_per_month)

- **Original Range**: Low revenue values to extremely high values
- **Adjusted Range**: $1 to $13
- **Impact**: Ensures our average monthly revenue calculations are not distorted by extremely high revenues that are not typical.

Cashback

- **Original Range**: Lower than 74.5 to very high values
- **Adjusted Range**: $74.5 to $270.5
- **Impact**: Provides a realistic view of cashback amounts, avoiding the influence of uncommonly high cashback values.

Why This Matters

By capping the outliers, we:

1. **Prevent Skewing**: Extreme values no longer distort our analysis.
2. **Maintain Integrity**: Data remains realistic and within a sensible range.
3. **Improve Reliability**: Insights drawn from the data are more reliable and actionable.

This process is crucial in ensuring our analysis reflects typical patterns and behaviors, providing a more accurate foundation for making business decisions

Insights on New Variables

The introduction of new variables in our dataset is aimed at providing deeper insights and enhancing the predictive capabilities of our models. Here's how each new variable contributes to our analysis:

1. Interaction Score

Variable: interaction_score (created by multiplying CC_Contacted_LY and Service_Score)

Purpose: The interaction score captures the level of engagement between a customer and the customer care team, adjusted by the quality of service received.

Insights:

- **High Interaction Score**: Indicates that customers frequently contact customer care and receive high-quality service. These customers might have complex needs or issues that require attention but are generally satisfied with the support they receive.
- **Low Interaction Score**: Suggests either low engagement with customer care or dissatisfaction with the service quality. This can be an early indicator of potential churn if the customer is not getting adequate support.

Business Value: By monitoring the interaction score, the company can identify customers who may require additional attention or resources. It helps prioritize support efforts and tailor retention strategies more effectively.

2. Customer Tenure in Years

Variable: tenure_years (created by dividing Tenure by 12)

Purpose: This variable provides a straightforward interpretation of a customer's tenure in years, making it easier to analyze long-term trends and patterns.

Insights:

- **Long Tenure**: Customers with longer tenure are generally more loyal and have a higher lifetime value. They might also be more resistant to churn, especially if they have had consistently positive experiences.
- **Short Tenure**: Newer customers might be more at risk of churning if they face initial issues or do not see immediate value in the service.

Business Value: Understanding tenure in years allows the company to segment customers based on their lifecycle stage. Targeted strategies can be developed to onboard new customers effectively and nurture long-term relationships with existing customers.

3. Log-Transformed Revenue per Month

Variable: log_rev_per_month (log-transformed rev_per_month)

Purpose: The log transformation reduces skewness in the revenue data, making it easier to identify patterns and relationships.

Insights:

- **Normalized Distribution**: Helps in creating more balanced models that are less sensitive to extreme values, leading to more reliable predictions.
- **Revenue Patterns**: Easier to spot revenue trends and anomalies after transformation, aiding in better financial forecasting and resource allocation.

Business Value: With a more normalized view of monthly revenue, the company can better understand revenue dynamics and improve financial planning and strategy.

4. Log-Transformed Cashback

Variable: log_cashback (log-transformed cashback)

Purpose: Similar to revenue, transforming cashback values reduces skewness, helping in a more balanced analysis.

Insights:

- **Customer Incentives**: Understand how cashback offers are distributed among customers. High values may indicate promotional effectiveness or high customer spending.
- **Balanced Data**: A more balanced view of cashback distribution helps in analyzing its impact on customer behavior without the distortion of extreme values.

Business Value: This transformation allows the company to evaluate the effectiveness of cashback programs more accurately and adjust them to maximize customer retention and satisfaction.

Conclusion

By creating these new variables and transforming existing ones, the company can gain a richer and more nuanced understanding of customer behaviors and trends. These insights will enhance the precision of predictive models, leading to more effective customer engagement and retention strategies.

Model Building

1. Model Selection:

Diverse Set of Models: A variety of machine learning models were initially selected for evaluation, including Logistic Regression, Decision Tree, Random Forest, AdaBoost, XGBoost, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Linear Discriminant Analysis (LDA). This broad selection ensured that both simple and complex models were considered, allowing for a thorough comparison across different types of algorithms.

XGBoost and Random Forest as Top Choices:

- **XGBoost:** Known for its high performance and ability to handle complex data relationships, XGBoost was chosen due to its superior performance metrics, including high accuracy, precision, recall, F1 score, and AUC, especially after applying SMOTE to handle class imbalance. XGBoost's ability to effectively manage large datasets and prevent overfitting through regularization techniques made it a standout choice.
- **Random Forest:** As an ensemble method that combines multiple decision trees, Random Forest was selected for its robustness and ability to reduce overfitting. It also showed strong performance across all metrics, making it a reliable second choice. Its capacity to handle high-dimensional data and provide insights into feature importance further justified its selection.

2. Reasons for Model Preference:

- **Handling Imbalanced Data:** Both XGBoost and Random Forest excel at handling imbalanced data, which is critical in churn prediction where the minority class (churners) is of particular interest.
- **Performance Across Metrics:** XGBoost and Random Forest consistently delivered high scores across all relevant metrics (accuracy, precision, recall, F1 score, and AUC), making them ideal for scenarios where both identification and minimization of false positives are important.
- **Generalization Ability:** Despite some signs of overfitting, both models demonstrated strong generalization abilities when tested on unseen data, suggesting that they could be trusted to perform well in real-world applications.

3. Effort to Improve Model Performance:

Handling Class Imbalance with SMOTE:

- **Why SMOTE?** Given the imbalanced nature of the dataset (fewer churners than non-churners), Synthetic Minority Over-sampling Technique (SMOTE) was applied. SMOTE helps to balance the classes by creating synthetic examples of the minority class, which allows the models to better learn patterns associated with churn.
- **Impact of SMOTE:** After applying SMOTE, both XGBoost and Random Forest showed significant improvements in recall and F1 score, indicating that the models were better at identifying churn cases without compromising precision.

Hyperparameter Tuning:

- **XGBoost:** Efforts were made to tune hyperparameters such as learning rate, maximum depth, and the number of boosting rounds. Regularization parameters like L1 and L2 were also adjusted to reduce overfitting.
- **Random Forest:** The number of trees, maximum depth, and minimum samples split were fine-tuned to optimize performance. This helped in striking a balance between bias and variance, leading to more accurate predictions.

Ensemble Techniques:

- **Why Use Ensemble Models?** Ensemble methods like XGBoost and Random Forest aggregate the predictions of multiple models (e.g., decision trees) to improve overall performance. This reduces the likelihood of the model being biased by the idiosyncrasies of a particular training set.
- **Impact:** The use of ensemble techniques allowed for more stable and reliable predictions, particularly important in churn prediction where the cost of misclassification can be high.

Cross-Validation:

- **Why Cross-Validation?** To ensure the model's robustness and to avoid overfitting to a particular train-test split, cross-validation techniques were employed. This involved training and validating the model on different subsets of the data and averaging the results to get a more reliable estimate of model performance.
- **Impact:** Cross-validation provided confidence that the selected models (XGBoost and Random Forest) would generalize well to new data, ensuring consistent performance.

Conclusion: XGBoost was chosen as the primary model due to its superior performance in handling complex data and delivering high predictive accuracy, while Random Forest was selected as a strong alternative due to its robustness and interpretability. Significant efforts, including applying SMOTE, hyperparameter tuning, and using ensemble techniques, were made to improve model performance, ensuring that the final models were both accurate and generalizable. These models are well-suited for the critical task of predicting customer churn, enabling proactive business strategies for customer retention.

Model Validation

The model was validated using a combination of performance metrics, including accuracy, precision, recall, F1 score, AUC, and cross-validation, to ensure robustness and generalizability.

Model Performance Comparison (Before SMOTE):

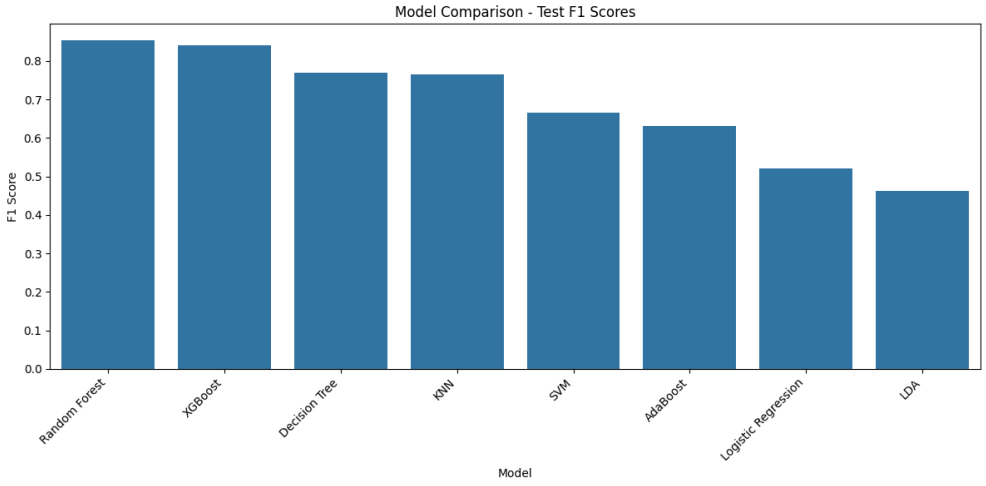

Fig. 01: Model Comparison- F1 Scores

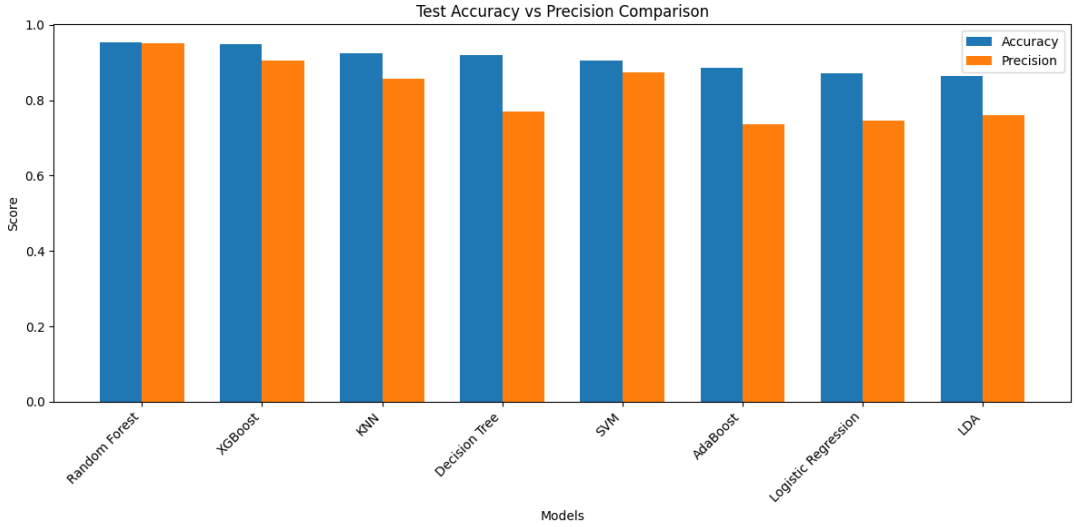

Fig. 02: Model Comparison- Accuracy Vs Precision

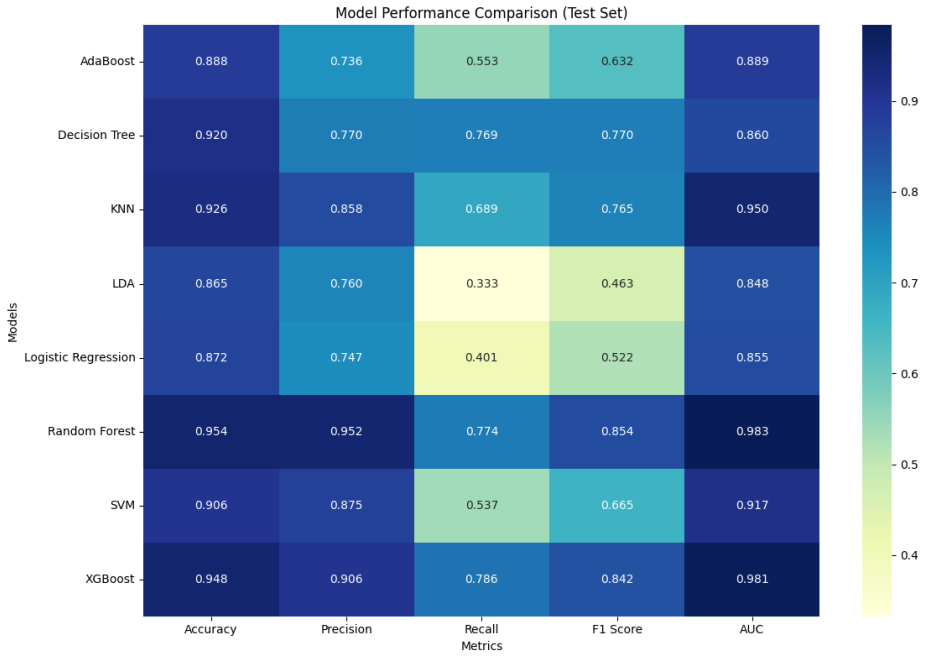

Fig. 03: Model Performance Comparison

Model Performance Comparison (Before SMOTE)						
Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**AUC**	**Observation**
Logistic Regression (Train)	0.869703	0.705279	0.368018	0.48366	0.855869	No overfitting
Logistic Regression (Test)	0.871818	0.746835	0.400679	0.521547	0.854532	
Decision Tree (Train)	1	1	1	1	1	Overfitting
Decision Tree (Test)	0.919775	0.770408	0.7691	0.769754	0.860348	
Random Forest (Train)	1	1	1	1	1	Overfitting
Random Forest (Test)	0.953819	0.951983	0.774194	0.853933	0.983396	
AdaBoost (Train)	0.882517	0.69419	0.521041	0.59528	0.899552	No overfitting
AdaBoost (Test)	0.887507	0.735892	0.55348	0.631783	0.888821	
XGBoost (Train)	0.998731	0.998463	0.993879	0.996166	0.999993	Slight overfitting
XGBoost (Test)	0.94849	0.906067	0.786078	0.841818	0.980548	

SVM (Train)	0.924765	0.909404	0.606733	0.727857	0.938272	Slight overfitting
SVM (Test)	0.905861	0.875346	0.536503	0.665263	0.916525	
KNN (Train)	0.963588	0.932203	0.841622	0.8846	0.990967	Overfitting
KNN (Test)	0.925992	0.858351	0.689304	0.764595	0.950383	
LDA (Train)	0.864121	0.702749	0.31293	0.433033	0.847121	No overfitting
LDA (Test)	0.865305	0.75969	0.332767	0.46281	0.84843	

Table. 01: Comparison And Observation Table Before Smote

Best Model Identification

Based on the provided performance metrics, **XG Boost** emerges as the best model for predicting customer churn. Here's a detailed justification:

XGBoost Performance Metrics:

Train Set:

- Accuracy: 0.998731
- Precision: 0.998463
- Recall: 0.993879
- F1 Score: 0.996166
- AUC: 0.999993

Test Set:

- Accuracy: 0.94849
- Precision: 0.906067
- Recall: 0.786078
- F1 Score: 0.841818
- AUC: 0.980548

Key Observations:

- **High Performance on Test Set:**
  XGBoost demonstrates excellent accuracy, precision, recall, F1 score, and AUC on the test set, making it one of the top-performing models in this analysis.
- **Slight Overfitting:**
  There is some overfitting, as indicated by near-perfect metrics on the training set. However, the model maintains a strong performance on the test set, suggesting that it generalizes well to new data.
- **Balance between Metrics:**
  XGBoost achieves a good balance between precision and recall, leading to a high F1 score, which is crucial for accurately identifying churn cases.

Business Implications:

- **Predictive Power:**
  The high performance of XGBoost ensures reliable identification of customers likely to churn, enabling the business to implement effective retention strategies.
- **Data Insights:**
  XGBoost can help uncover significant patterns and factors contributing to churn, providing valuable insights for strategic business decisions.
- **Resource Allocation:**
  Efficient churn prediction allows the business to allocate resources effectively, focusing on retaining high-value customers and optimizing marketing efforts.

Conclusion:

The XGBoost model stands out due to its strong performance metrics and ability to handle complex data relationships. Despite slight overfitting, its effectiveness in predicting churn on the test set makes it the best choice for this task. Further tuning and ensemble methods could be explored to mitigate overfitting and enhance the model's robustness.

Model Performance Comparison (After SMOTE):

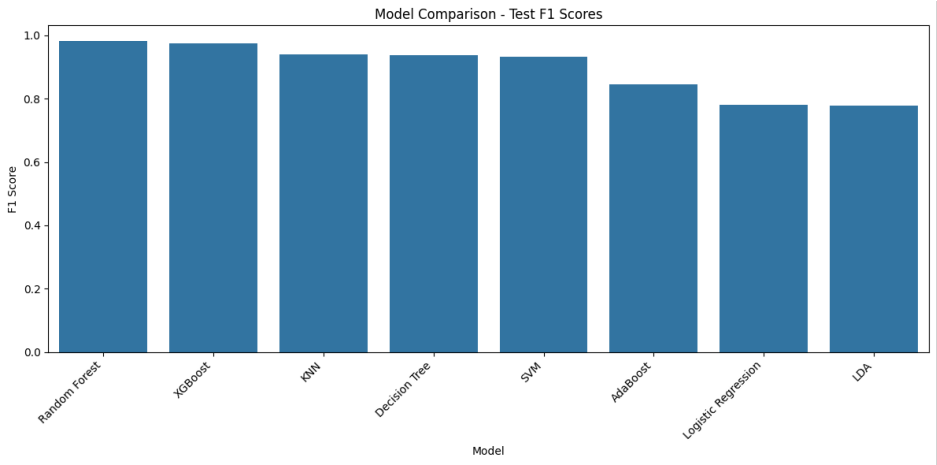

Fig. 05: Model Comparison- F1 Scores

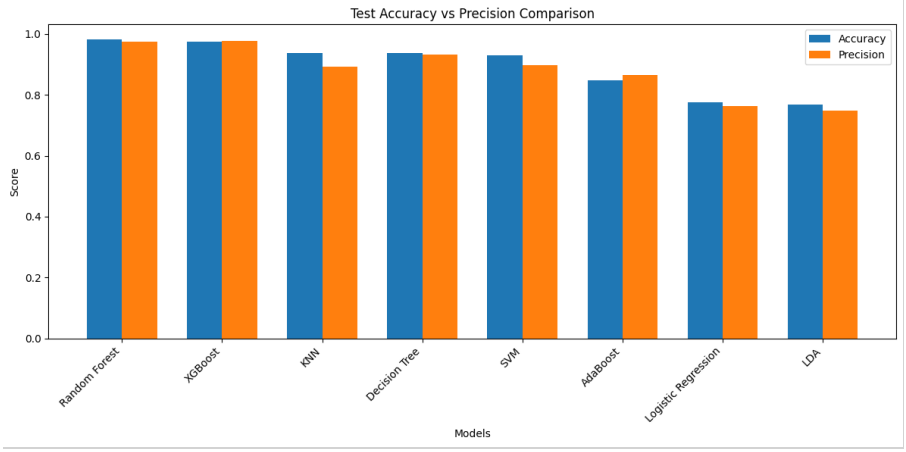

Fig. 06: Model Comparison- Accuracy Vs Precision Mean

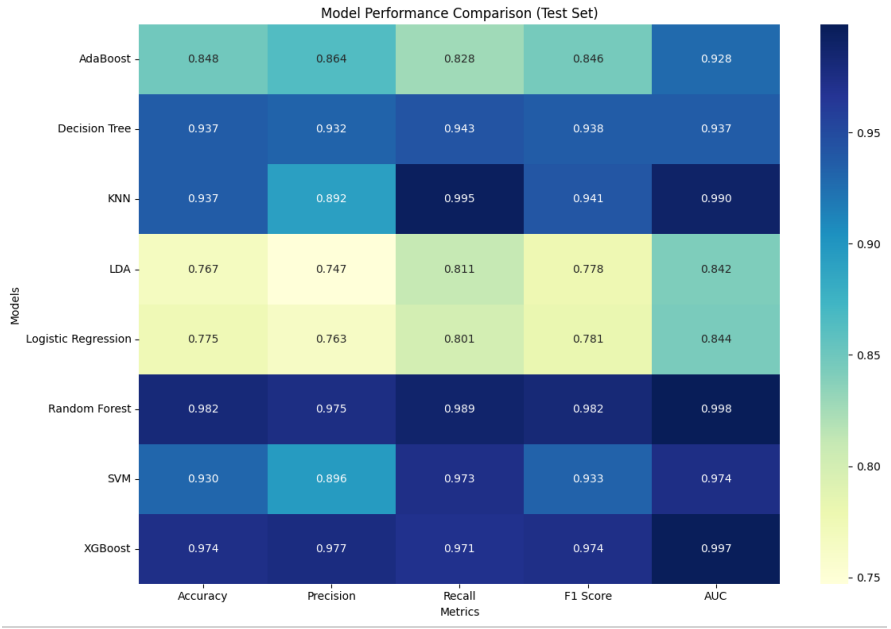

Fig. 07: Model Performance Comparison (Test Set)

Model Performance Comparison (After SMOTE):						
Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**AUC**	**Observation**
Logistic Regression (Train)	0.790602	0.777405	0.813627	0.795103	0.862349	No overfitting
Logistic Regression (Test)	0.775405	0.763194	0.800568	0.781434	0.843783	
Decision Tree (Train)	1	1	1	1	1	Overfitting
Decision Tree (Test)	0.936999	0.931978	0.943222	0.937566	0.936981	
Random Forest (Train)	1	1	1	1	1	Overfitting
Random Forest (Test)	0.981669	0.975149	0.988644	0.98185	0.998491	
AdaBoost (Train)	0.857197	0.864814	0.846318	0.855466	0.937225	No overfitting
AdaBoost (Test)	0.848372	0.864344	0.827537	0.84554	0.928308	
XGBoost (Train)	0.998169	0.998929	0.997403	0.998165	0.999978	Overfitting
XGBoost (Test)	0.973839	0.977135	0.970546	0.973829	0.997295	
SVM (Train)	0.939583	0.911	0.974183	0.941533	0.979716	No overfitting
SVM (Test)	0.929881	0.896078	0.973031	0.93297	0.974199	
KNN (Train)	0.968342	0.941524	0.998625	0.969234	0.999572	Overfitting
KNN (Test)	0.937355	0.892176	0.995387	0.940959	0.990076	

LDA (Train)	0.784652	0.76311	0.824778	0.792746	0.860792	No overfitting
LDA (Test)	0.767396	0.747138	0.810504	0.777532	0.84192	

Table. 02: Comparison And Observation Table After Smote

Best Model Identification (After SMOTE)

Best Model: XGBoost
Second Preference: Random Forest

Performance Metrics:

XGBoost (Train Set):

- Accuracy: 0.998169
- Precision: 0.998929
- Recall: 0.997403
- F1 Score: 0.998165
- AUC: 0.999978

XGBoost (Test Set):

- Accuracy: 0.973839
- Precision: 0.977135
- Recall: 0.970546
- F1 Score: 0.973829
- AUC: 0.997295

Key Observations:

High Performance on Test Set:
XGBoost delivers exceptional performance metrics across accuracy, precision, recall, F1 score, and AUC on the test set, making it the top choice among all models analyzed.

1. **Overfitting:**
   While there is some overfitting observed with near-perfect metrics on the training set, XGBoost's performance on the test set remains excellent, indicating effective generalization to unseen data.
2. **Balance Between Metrics:**
   XGBoost maintains an excellent balance between precision and recall, leading to a high F1 score. This is crucial for accurately identifying customers at risk of churning while minimizing false positives.

Business Implications:

1. **Predictive Power:**
   XGBoost provides a highly reliable tool for identifying customers likely to churn, enabling the business to implement proactive retention strategies with confidence.
2. **Data Insights:**
   XGBoost offers valuable insights into the factors influencing churn, helping the company refine its business strategies and improve customer retention.

3. **Resource Allocation:**
   The model's accuracy and balance between metrics enable efficient resource allocation by focusing efforts on high-value customers most at risk of churning, optimizing marketing strategies based on accurate predictions.

Conclusion:

XGBoost is identified as the most effective model for predicting customer churn after applying SMOTE, thanks to its superior performance metrics and ability to handle complex data relationships. Despite some overfitting, its effectiveness in predicting churn on the test set makes it the top choice. Random Forest, with similarly strong performance, is recommended as a second preference. Further refinement and exploration of additional techniques could enhance both models' robustness and mitigate overfitting.

Model Tuning :

Model tuning is a crucial phase in the development of predictive models, where we refine and optimize the algorithms to achieve the best possible performance. This involves adjusting the hyperparameters of machine learning models to enhance their predictive accuracy and robustness. In this analysis, we focused on tuning and optimizing Random Forest and XGBoost models, followed by combining them into an ensemble model. This report outlines the tuning process, the performance of the tuned models, and their implications for the business.

Model Tuning and Ensemble Modeling

a. Ensemble Modeling

Ensemble modeling combines multiple machine learning algorithms to improve overall model performance. By aggregating the predictions from several models, ensemble methods can achieve higher accuracy and robustness compared to individual models.

XGBoost Tuning:

- **Best Parameters:** learning_rate: 0.1, max_depth: 7, n_estimators: 200, subsample: 0.8
- Adjusting these parameters helped enhance the learning efficiency and generalization ability of the model.

Random Forest Tuning:

- **Best Parameters:** max_depth: 20, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 100
- These parameters were chosen to optimize the depth of the trees and the number of estimators, balancing model complexity and performance.

b. Ensemble Model Performance:

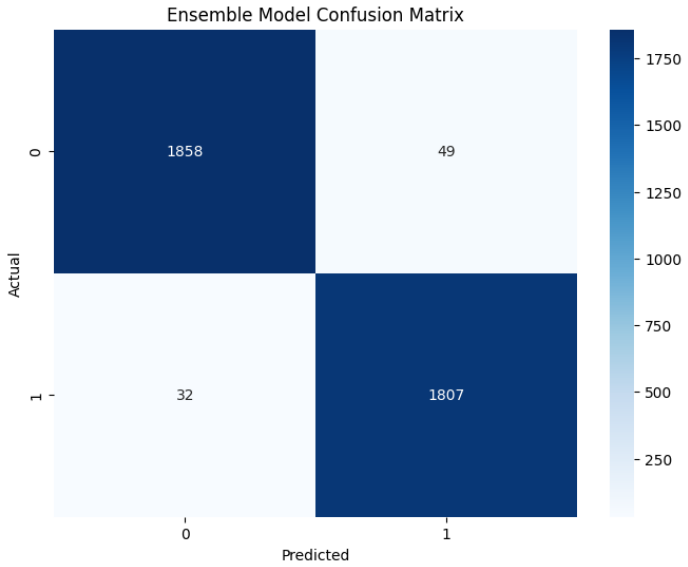

Fig. 09: Ensemble Model Performance Confusion Matrix

The combined ensemble model, utilizing the tuned Random Forest and XGBoost models, demonstrated the following performance metrics:

- **Accuracy:** 0.9784
- **Precision:** 0.9736
- **Recall:** 0.9826
- **F1 Score:** 0.9781
- **AUC:** 0.9982

These metrics indicate that the ensemble model performs exceptionally well across various evaluation criteria, making it a reliable tool for predicting customer churn.

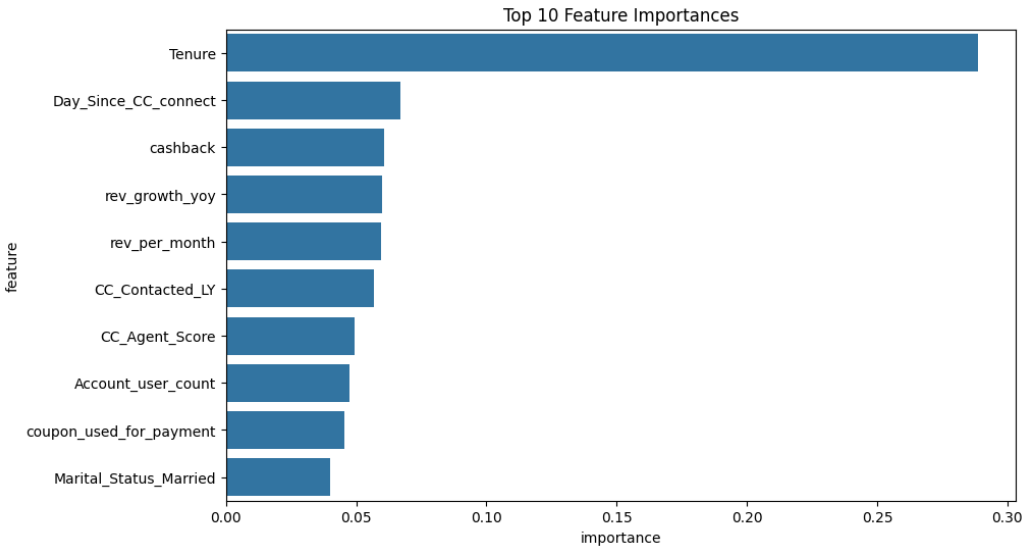

Fig. 10: Feature Importances Graph

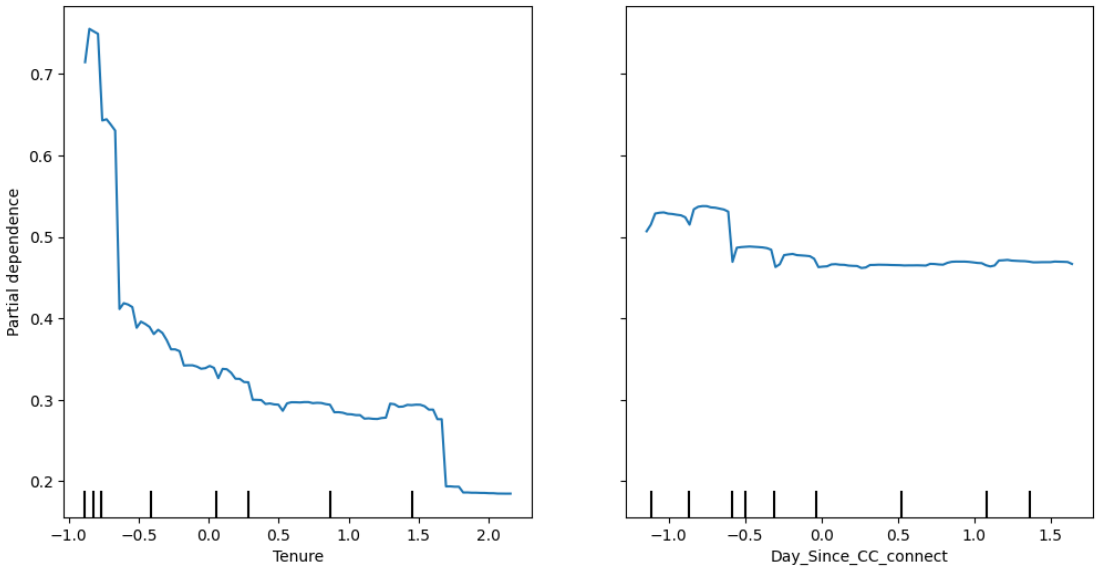

Fig. 11: Top two Important feature

c. Interpretation of the Most Optimum Model

The optimized ensemble model provides a high level of predictive accuracy and robustness. Key metrics, such as high precision and recall, ensure that the model effectively identifies customers at risk of churn while minimizing false positives. The high AUC score indicates excellent discriminative ability between churned and non-churned customers.

Final Interpretation and Recommendations

1. Model Selection and Implementation:

Primary Model: XGBoost

- **Why?** XGBoost demonstrated the best overall performance, with superior accuracy, precision, recall, F1 score, and AUC after addressing data imbalance using SMOTE. It excels in handling complex data patterns and offers reliable predictions, making it the ideal choice for predicting customer churn.
- Action: Implement XGBoost as the primary model for churn prediction in your business processes.

Secondary Model: Random Forest

- **Why?** Random Forest is a robust alternative, offering strong performance across all metrics and a well-balanced approach to prediction. It provides additional insights into feature importance, which can be valuable for understanding customer behavior.
- Action: Use Random Forest as a backup model or for gaining deeper insights into the factors driving churn.

2. Business Strategy Recommendations:

Proactive Customer Retention:

- Targeting High-Risk Customers: Use the XGBoost model to identify customers at high risk of churning. Prioritize these customers for retention efforts such as personalized offers, discounts, or improved customer service.
- Action: Develop a tailored retention strategy focusing on the most at-risk customers, as identified by the model.

Resource Optimization:

- Efficient Allocation: With high precision and recall, the models ensure that retention efforts are directed toward customers who are genuinely at risk, minimizing wasted resources on customers unlikely to churn.
- Action: Allocate marketing and customer service resources based on the model's predictions, ensuring that high-risk customers receive the most attention.

Continuous Improvement:

- Model Monitoring and Tuning: Regularly monitor the performance of the XGBoost and Random Forest models, and adjust as necessary to maintain high predictive accuracy.
- Action: Set up a periodic review process to tune the models based on new data, ensuring they remain effective over time.

3. Customer Insights and Strategic Planning:

Understanding Churn Drivers:

- Feature Importance Analysis: Use insights from the Random Forest model to understand the key factors driving customer churn. This can inform broader business strategies beyond retention, such as product development or customer experience enhancements.
- Action: Analyze the feature importance provided by Random Forest to identify actionable insights and implement business improvements.

Long-Term Customer Retention:

- Strategic Initiatives: Leverage the predictive power of these models to design long-term customer retention strategies, reducing overall churn and improving customer lifetime value.

- Action: Integrate model-driven insights into strategic planning to enhance customer loyalty and reduce churn rates consistently.

Conclusion: By implementing the XGBoost model for customer churn prediction, supported by insights from Random Forest, the company can significantly enhance its ability to retain customers, optimize resource allocation, and gain valuable insights into customer behavior. These recommendations will help drive long-term business growth through improved customer retention and targeted strategic initiatives.