MCA Semester – IV

Research Project – Interim Report

Name	Pavan Raj K G
Project	Churn Prediction Model for E-commerce Customer Retention
Group	Churn 2
Date of Submission	16-04-2024

A study on "Churn Prediction Model for E-commerce Customer Retention"

Research Project submitted to Jain Online (Deemed-to-be University)
In partial fulfillment of the requirements for the award of:
Master of Computer Application

Submitted by:

Pavan Raj K G

USN:

221VMTR02085

Under the guidance of:

Mr. Nimesh Marfatia

(Faculty-JAIN Online)

Jain Online (Deemed-to-be University)

Bangalore

2023-24

DECLARATION

I, *Pavan Raj K G,* hereby declare that the Research Project Report titled *"Churn Prediction Model for E-commerce Customer Retention" has been* prepared by me under the guidance of Mr. *Nimesh Marfatia.* I declare that this Project work is towards the partial fulfillment of the University Regulations for the award of the degree of Master of Computer Application by Jain University, Bengaluru. I have undergone a project for a period of Eight Weeks. I further declare that this Project is based on the original study undertaken by me and has not been submitted for the award of any degree/diploma from any other University / Institution.

Place: Bangalore **Pavan Raj K G**

Date: 15-04-2024 ***USN:* 221VMTR02085**

List of Tables

List of Graphs

1. Introduction

The e-commerce industry thrives on customer loyalty. Acquiring new customers is crucial for growth, but retaining existing ones is demonstrably more cost-effective. Studies show that a mere 5% increase in customer retention can lead to a 25-95% increase in net present value. However, customer churn, the rate at which customers cease business with a company, remains a significant challenge.

This project proposes a data-driven approach to tackle customer churn within an e-commerce company. We aim to develop a churn prediction model that analyzes customer data to identify individuals at high risk of leaving. This proactive approach allows us to implement targeted retention strategies, ultimately fostering customer loyalty and driving sustainable business growth.

2. Objectives of the Study

This study sets out to achieve the following key objectives:

- **Develop a High-Performance Churn Prediction Model**: Our primary objective is to construct a robust model that analyzes customer data and accurately identifies individuals with a high likelihood of churning. This proactive approach allows us to prioritize retention efforts and focus resources on those most susceptible to leaving.
- **Uncover Key Drivers of Customer Churn**: By analyzing the model's output and the underlying data, we aim to pinpoint the critical factors that contribute to customer churn. Understanding these pain points empowers us to address them directly, improve customer satisfaction, and enhance the overall customer experience.
- **Formulate Targeted Retention Strategies**: Leveraging the insights gleaned from the churn prediction model, we will develop targeted marketing campaigns and retention programs specifically designed to re-engage at-risk customers. These personalized interventions will incentivize them to continue doing business with E-commerce company, ultimately fostering customer loyalty and driving business growth.

Defining the Problem Statement

The problem we are addressing is customer churn, which represents a significant challenge for e-commerce companies. Customer acquisition is crucial for growth, but retaining existing customers is demonstrably more cost-effective. Despite this, a substantial portion of customers eventually cease business with a company.

This study aims to tackle this challenge by developing a data-driven approach to identify customers at risk of churning. By proactively intervening with targeted retention efforts, we can mitigate customer churn and foster long-term customer loyalty.

Need for the Study

Customer churn poses a significant threat to the long-term success of our E-commerce company. Studies have shown that retaining existing customers is significantly cheaper than acquiring new ones. Furthermore, loyal customers tend to make repeat purchases, spend more per transaction, and act as brand advocates, further promoting our business.

This project directly addresses this need by developing a churn prediction model. This model empowers us to proactively identify at-risk customers and tailor retention efforts to address their specific concerns. By mitigating customer churn and fostering loyalty, this study has the potential to significantly improve customer lifetime value, drive revenue growth, and enhance the overall profitability of our company.

3. Scope of the Study

This churn prediction model project resides within a well-defined scope that offers significant business and social opportunities for E-commerce company.

Business Opportunity

Reduced Customer Churn: The primary business opportunity lies in the model's ability to identify customers at risk of churning. By proactively intervening with targeted retention efforts, we can significantly reduce customer churn. This translates to a more stable customer base, leading to predictable revenue streams and improved financial performance.

Enhanced Customer Lifetime Value: Retained customers are more likely to make repeat purchases, spend more per transaction, and become brand advocates, attracting new customers through positive word-of-mouth. This fosters a loyal customer base with a higher lifetime value, ultimately driving sustainable business growth and profitability.

Data-Driven Decision Making: The churn prediction model empowers data-driven decision making. By analyzing customer behavior and churn drivers, we can gain valuable insights to optimize product offerings, personalize marketing campaigns, and enhance the overall customer experience.

Social Opportunity

Improved Customer Experience: By addressing the reasons behind customer churn, we can proactively improve the customer experience. This fosters customer satisfaction, loyalty, and trust in the brand.

Resource Optimization: The churn prediction model allows us to focus retention efforts on at-risk customers, optimizing the allocation of resources and maximizing the return on investment for customer retention programs.

Limitations of the Scope

While the project offers significant opportunities, it's important to acknowledge the limitations of the scope

Data Availability and Quality: The model's effectiveness relies heavily on the quality and comprehensiveness of customer data available. Limited or inaccurate data may hinder the model's accuracy.

External Factors: Customer churn can be influenced by external factors beyond our control, such as economic conditions or competitor offerings. The model may not capture these external factors entirely.

Model Maintenance: Customer behavior and churn patterns can evolve over time. The model will require ongoing monitoring and updates to maintain its effectiveness.

Overall, the scope of this churn prediction model project offers a well-defined approach with significant potential to improve customer retention, drive business growth, and enhance the overall customer experience. By acknowledging the limitations and maintaining responsible data practices, we can leverage this project to achieve both business and social benefits.

4. Data Collection Method

Understanding the Challenge:

The client, an E-commerce or DTH provider, faces high customer churn due to increased competition. They require a churn prediction model to identify accounts at risk and implement targeted retention strategies. Since an account can have multiple customers, losing an account result in losing multiple customers. The challenge lies in developing a model that balances customer retention with revenue assurance. Free or heavily subsidized offers might not be feasible.

Data Sources

To develop an effective churn prediction model, we will collect data from various internal sources.

The data consist of information as follows:

AccountID: Unique identifier for each account.

Churn: Flag indicating account churn within a defined timeframe (e.g., not active for X months).

Tenure: Length of time the account has been active.

City_Tier: Tier of the primary customer's city (based on population, development etc.).

Payment: Preferred payment mode of the customers associated with the account.

Account_user_count: Number of customers associated with the account.

account_segment: Account segmentation based on spending habits.

rev_per_month: Monthly average revenue generated by the account (past 12 months).

rev_growth_yoy: Revenue growth percentage of the account (past 12 months vs. previous 24 months).

coupon_used_l12m: Number of times coupons were used for payment in the last 12 months.

cashback_l12m: Monthly average cashback generated by the account (past 12 months).

Gender: Gender of the primary account holder (if provided with consent).

Marital_Status: Marital status of the primary account holder (if provided with consent).

Login_device: Preferred login device of the customers associated with the account.

CC_Contacted_L1_2m: Number of times any customer from the account contacted customer care in the last 12 months.

Day_Since_CC_connect: Number of days since the last customer service interaction for the account.

CC_Agent_Score: Average satisfaction score provided by customers for the received customer care service.

Complain_l12m: Flag indicating if the account raised any complaints in the last 12 months.

Data Collection Frequency

Customer service data will be collected continuously for real-time insights.

5. Data Analysis Tools

Exploratory Data Analysis

Following the data collection process outlined earlier, we can begin visually inspecting the data to gain initial insights and prepare it for model development.

About the Data:

Description	**Values**
Total number of rows	11260
Total number of columns	19
Number of Numerical columns	14
Number of object columns	5
Number of Numerical rows with missing values	785
Number of object rows with missing values	1891

Table 1: Overview on Data available

Rows: This represents the number of accounts in the dataset. A higher number of accounts allows for a more robust model.

Columns: These represent the various features used to predict churn. We'll have:

Description	**Columns**
Target Variable	Churn
Account-Level Features	▪ AccountID, ▪ Tenure, ▪ City_Tier,

	▪ Payment, ▪ Account_user_count, ▪ account_segment, ▪ rev_per_month, ▪ rev_growth_yoy, ▪ coupon_used_l12m, ▪ cashback_l12m
Customer-Level Features	▪ Gender, ▪ Marital_Status, ▪ Login_device
Customer Service Interaction Features	▪ CC_Contacted_L1_2m, ▪ Day_Since_CC_connect, ▪ CC_Agent_Score, ▪ Complain_l12m

Table 2: Overview on Columns in the Dataset

Data Cleaning

The data requires cleaning to address a significant number of missing values, totaling 2676 across 19 columns. This missing data needs to be addressed before proceeding with model development to ensure the model's accuracy and reliability.

<u>Identify Missing Value Patterns:</u>

- Analyze the distribution of missing values
- Investigate the reasons for missing values
- Understand the impact of missing values

Missing Value Treatment:

- Choose an Appropriate Missing Value Treatment Method
- Imputation: Estimate missing values based on available data. Techniques include:
  - Mean/Median Imputation: Replace missing values with the average or median value for the specific column. (Suitable for numerical data)
  - Mode Imputation: Replace missing values with the most frequent value in the column. (Suitable for categorical data)
  - Deletion: Remove rows or columns with a high percentage of missing values. This should be used cautiously to avoid losing valuable data.

Outlier Treatment: Identify outliers and decide on appropriate action like capping extreme values or removal based on their influence on the model.

Variable Transformation: Transformations like log scaling or normalization might be applied to improve model performance if necessary.

Data Visualisation & Insight

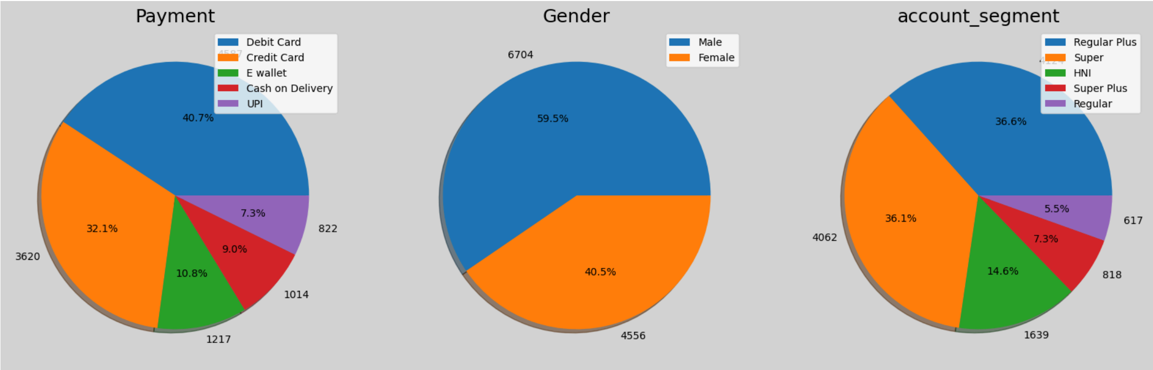

Fig 1 : Pie charts of Payment used for purchasing, Gender and Account segment

- **Payment Method Popularity:** Credit cards dominate as the payment method (40.7%), indicating a preference for credit over debit or e-wallets.

- **Account Segment Breakdown:** Regular Plus accounts are the most common (36.6%), potentially offering a good starting point for new customers.

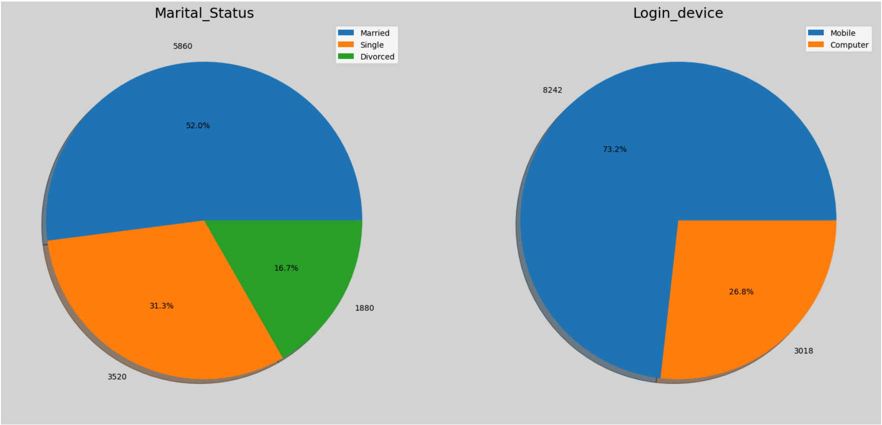

Fig 2: Pie chart of Marital Status of user and Login device used for ordering

Marital Status:

- Married is the most common marital status, making up over half (52.0%) of the people represented in the pie chart.
- Single is the second most common marital status at 31.3%, the remaining slices, Divorced (16.7%)

Login Device:

- Mobile is the most popular login device, used by over 73.2% of the people represented in the pie chart.
- Computer usage is significantly lower at 26.8%.

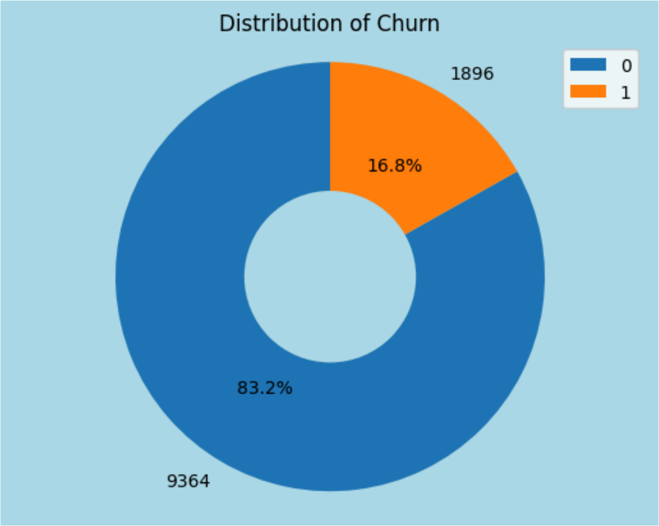

Fig 3 : Distribution of Churn

It shows the percentage of customers who churn, which presumably means they discontinue using a service or subscription.

The chart divides the data into two sections:

- No Churn (83.2%) This is the largest section of the pie chart, It indicates that the vast majority of customers (83.2%) do not churn within the measured timeframe.
- Churn (16.8%) This section is represents the percentage of customers who churned within the measured timeframe.
- Distribution of Churn: Customer churn is relatively low (16.8%), suggesting a healthy customer base.

Understanding Customer Churn: A Data-Driven Analysis

The charts below provide insight into where the churn rate is high. This can help you decide where to focus efforts to reduce churn and improve the quality of your services.

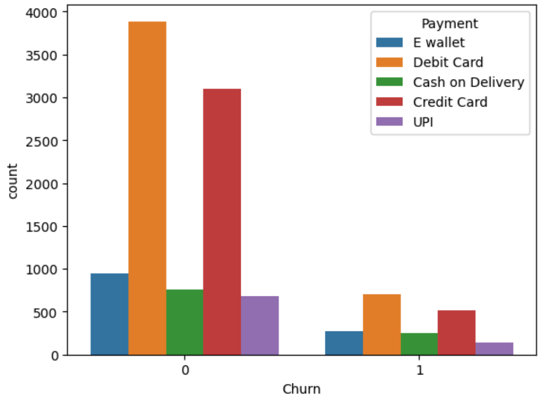

fig 4 : Bar chart of Payment v/s Churn

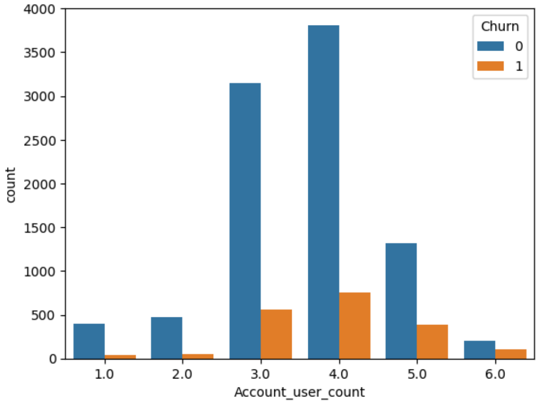

fig 5 : Bar chart of customers associated with the account v/s Churn

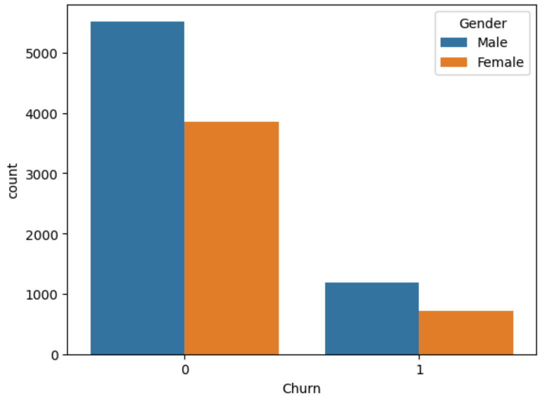

fig 6 : Bar chart of Gender v/s Churn

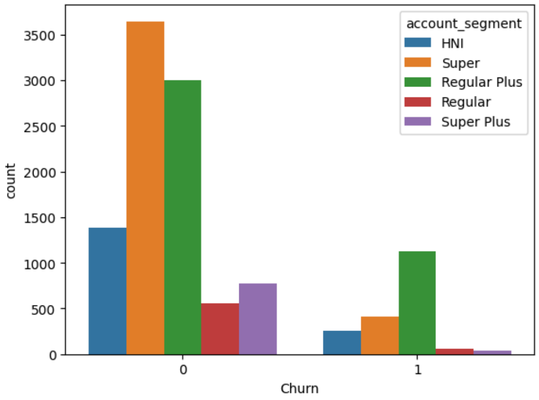

fig 7 : Bar chart of Account segment v/s Churn

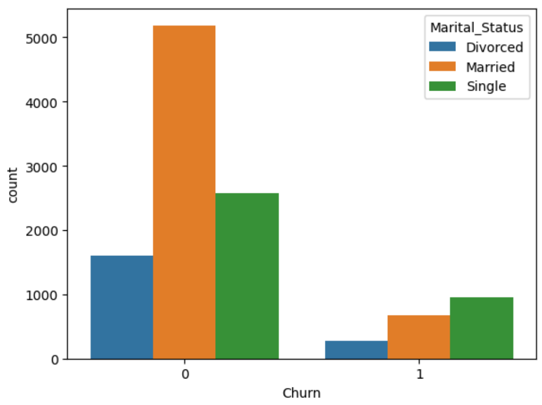

fig 8 : Bar chart of Marital Status v/s Churn

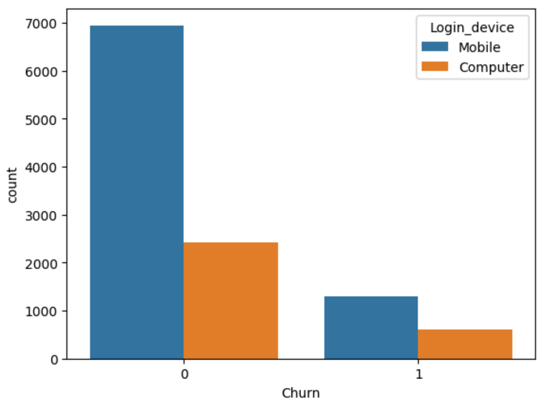

fig 9 : Bar chart of Device used v/s Churn

Business Insights from Exploratory Data Analysis (EDA)

This analysis combines the findings from the provided data snippets and the overall data exploration to offer business insights for churn prediction and targeted retention strategies.

Customer Lifetime Value and Churn Propensity:

- **High-Value Accounts:** A positive correlation between high rev_per_month and positive rev_growth_yoy suggests these accounts hold significant customer lifetime value. Retention efforts should prioritize these accounts to minimize revenue loss.

- **Service Satisfaction and Churn Risk:** The data snippet shows a potential link between lower Service_Score and CC_Agent_Score with churn. Focusing on improving customer service experience through staff training or feedback mechanisms can significantly reduce churn.

- **Customer Engagement and Churn:** Frequent customer service interactions (CC_Contacted_LY) could signal issues or dissatisfaction potentially leading to churn, while a long absence of contact (Day_Since_CC_connect) might indicate low engagement and indicates no issue with the service provided and leads to customer retension.

Segmentation and Retention Strategies:

- **Customer Segmentation:** Analyzing features like City_Tier, Account_user_count, and Login_device can reveal distinct customer segments with unique churn patterns. Tailored retention campaigns can then be developed to address the specific needs of each segment.

- **Targeted Retention Campaigns:** Based on the identified churn triggers, we can develop targeted retention campaigns that prioritize customer value and avoid excessive discounts:

    - **Tiered Loyalty Programs:** Offer progressively better rewards based on account value (rev_per_month) and tenure. This incentivizes continued engagement and spending without compromising revenue.

    - **Exclusive Content or Early Access:** Provide exclusive content, early access to new products/services, or priority customer service for high-value accounts. This fosters a sense of exclusivity and increases customer satisfaction.

    - **Personalized Recommendations and Upsells:** Leverage customer data (purchase history, login device) to recommend relevant add-on products or services. This caters to individual needs and potentially increases revenue without resorting to discounts.

    - **Gamification and Rewards:** Implement gamified elements like points, badges, and leaderboards for account activity or timely payments. This can incentivize engagement and on-time payments without direct financial discounts.

Additional Considerations:

- **Coupon Usage and Churn:** Analyzing how coupon_used_for_payment relates to churn. While it might indicate cost-consciousness, it could also suggest a customer looking for better deals elsewhere. Targeted promotions or loyalty programs can address this.
- **Complaint History and Churn:** Investigating the relationship between Complain_ly and churn. Accounts with a history of complaints might require more attention to address their concerns and prevent churn.

By leveraging these insights and conducting a thorough EDA, the E-commerce/DTH provider can develop data-driven churn prediction models and implement targeted retention campaigns that address churn triggers, improve customer satisfaction, and encourage continued business without compromising revenue. This approach will ensure long-term customer loyalty and sustainable business growth.

6. Model Building

Here are the steps involved in building the model:

1. **Data Preprocessing:**
   - **Cleaning:** Handle missing values through imputation or removal based on the variable and the proportion of missing data. Address outliers based on their influence on the model's performance.
   - **Feature Engineering:** Create new features by combining existing ones or deriving new metrics that might better capture churn signals.
   - **Feature Scaling:** If necessary, scale numerical features to a common range to ensure all features contribute equally to the model.
2. **Model Selection:** Choose an appropriate machine learning algorithm for churn prediction. Here are some common options:
   - **Logistic Regression:** A popular classification algorithm suitable for binary prediction tasks like churn (churn or not churn).
   - **Decision Trees:** Flexible and interpretable models that can handle both numerical and categorical features.
   - **Random Forest:** Ensemble method combining multiple decision trees for improved accuracy and reduced overfitting.
   - **Gradient Boosting:** Another ensemble technique that iteratively builds models to improve upon the previous ones.
3. **Model Training:** Split the data into training and testing sets. Train the chosen model(s) on the training set, allowing it to learn the patterns that differentiate churning customers from non-churning ones.

4. **Model Testing and Evaluation:**

   ○ The models were evaluated on a separate testing set using various performance metrics:

      i. **Accuracy:** Overall proportion of correct predictions (churned vs. not churned)

      ii. **F1 Score:** Harmonic mean of precision and recall

      iii. **Recall:** Proportion of actual churned customers who were correctly identified

      iv. **Precision:** Proportion of predicted churned customers who actually churned

Results

The table below summarizes the performance metrics for each model:

Model Name	**Accuracy**	**F1 Score**	**Recall Score**	**Precision Score**
Logistic Regression	0.8956	0.6010	0.4797	0.8045
Decision Tree	0.9347	0.8073	0.8347	0.7817
Random Forest	**0.9729**	**0.9127**	**0.8645**	**0.9667**
XGBoost Classifier	0.8832	0.4971	0.3523	0.8442

Table 3: Comparison of Models

Key Findings

- Random Forest emerged as the leading performer with the highest Accuracy (0.9729) and Precision Score (0.9667). This indicates its ability to accurately predict both churned and non-churned customers, minimizing false positives (unnecessary outreach).
- Decision Tree achieve-d a good balance between Recall and Precision, effectively identifying churned customers. However, its overall Accuracy was lower than Random Forest.
- Logistic Regression offered a lower F1 score compared to Decision Tree, suggesting it might miss a higher proportion of churned customers despite having a similar Accuracy.
- XGBoost Classifier exhibited the lowest performance among the evaluated models based on the provided metrics.

Next Steps

Based on these findings, the focus will shift towards:

- **In-depth Analysis of Random Forest:** We will delve deeper into the Random Forest model to understand
- **Model Refinement:** We will explore techniques like hyperparameter tuning to potentially optimize the Random Forest model's performance further.
- **Deployment Considerations:** We will evaluate the feasibility of deploying the chosen model in a production environment to enable real-time churn prediction and proactive retention strategies.

Conclusion

This interim report has successfully evaluated various machine learning models for customer churn prediction. Random Forest stands out as the most promising model based on its high Accuracy and Precision. Further analysis of this model and potential refinements will be conducted to establish a robust solution for minimizing customer churn and maximizing customer retention.