

**MCA Semester – IV**

**Research Project – Final Report**

|  |  |
| --- | --- |
| **Name** | Anuvab Bikash Nayak |
| **Project** | Customer Churn |
| **Date of Submission** | 20/09/2024 |



**A study on “****Customer Churn “**

## 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:*

**Student Name**

USN:

222VMTR00122

*Under the guidance of:*

Shradha Sriavstava

(Faculty-JAIN Online)

Jain Online (Deemed-to-be University)

Bangalore

**DECLARATION**

I, *ANUVAB BIKASH NAYAK,* hereby declare that the Research Project Report titled ***Customer Churn****” has been* prepared by me under the guidance of the *Shradha Sriavstava.* 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: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Date: *ANUVAB BIKASH NAYAK*

*USN: 222VMTR00122*

|  |  |  |
| --- | --- | --- |
| **List of Tables** | | |
| **Table No.** | **Table Title** | **Page No.** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| **List of Graphs** | | |
| **Graph No.** | **Graph Title** | **Page No.** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

|  |  |
| --- | --- |
| **Table of Contents** | |
| **Abstract** | A concise summary of model building |
| **Introduction** | Introduction and Background |
| Problem Statement |
| Objective of Study |
| **Literature Review** | Company and industry overview |
| Overview of Theoretical Concepts |
| Survey on the existing models |
| **Methodology** | 1. EDA and Business Implication |
| - Univariate / Bi-variate / Multivariate analysis to understand relationship b/w variables. How is your analysis impacting the business? |
| - Both visual and non-visual understanding of the data. |
| 2. Data Cleaning and Preprocessing |
| - Approach used for identifying and treating missing values and outlier treatment (and why) |
| - Need for variable transformation (if any) |
| - Variables removed or added and why (if any) |
| 3. Model building & Model validation |
| - Clear on why a particular model(s) was chosen. |
| - Effort to improve model performance. |
| - How was the model validated? Just accuracy, or anything else too? |
| **Results and discussion** | Findings Based on Observations |
| Findings Based on analysis of Data |
| General findings |
| **Conclusion and Recommendations** | Recommendation based on findings |
| Suggestions for areas of improvement |
| Scope for future research |
| Conclusion |

**Abstract:**

This report presents a comprehensive analysis of customer churn for an E-commerce company, utilizing a Random Forest Classifier model. The study involved rigorous data cleaning, imputation of missing values, and exploratory data analysis to understand customer behavior and segment them into distinct groups. The Random Forest model achieved exceptional accuracy (99.56%) in predicting customer churn, enabling the company to proactively identify and target at-risk customers with tailored retention strategies. The insights gleaned from this analysis empower the company to make informed decisions, optimize customer engagement, and foster long-term loyalty in a competitive market.  
  
In the face of escalating competition within the e-commerce landscape, customer retention has emerged as a paramount concern. The challenge of mitigating churn, particularly concerning the loss of entire accounts with multiple users, has prompted the need for a robust predictive model. This report details the development and implementation of such a model, employing a Random Forest Classifier to forecast account churn. The research encompasses meticulous data preprocessing, exploratory data analysis (EDA), and model evaluation, culminating in actionable insights and strategic recommendations.

Through EDA, customer behavior patterns were scrutinized, and distinct customer segments were identified using K-Means clustering. The Random Forest model, trained on carefully prepared data, exhibited exceptional accuracy in predicting churn, thereby empowering the company to proactively address potential attrition. The model's efficacy was substantiated through a confusion matrix and a classification report, underscoring its precision in discerning churned and non-churned accounts.

Beyond model development, this report offers strategic business recommendations, emphasizing a delicate balance between customer incentives and financial prudence. The proposed campaign strategies are designed to resonate with specific customer segments, fostering engagement and loyalty while upholding the company's fiscal integrity. By integrating data-driven insights with strategic acumen, this research equips the e-commerce company to navigate the complexities of customer churn, bolster retention efforts, and fortify its competitive standing in the market.

## 

**Introduction:**

**Introduction and Background**

In the rapidly evolving e-commerce landscape, customer retention has become a critical factor for businesses aiming to achieve sustainable growth and profitability. The phenomenon of customer churn, where customers discontinue their relationship with a company, can significantly impact revenue, brand reputation, and overall business success. E-commerce companies, in particular, face unique challenges in retaining customers due to the ease with which customers can switch to competitors and the abundance of choices available to them.

The phenomenon of customer churn, referring to the cessation of a customer's relationship with a company, can have detrimental consequences, including revenue loss, diminished brand reputation, and increased acquisition costs. For e-commerce platforms, where customer acquisition and retention are paramount, understanding and predicting churn is crucial for implementing effective customer retention strategies.

The problem of churn is further compounded in the context of e-commerce companies where accounts often encompass multiple users. The loss of a single account can translate to the departure of several customers, amplifying the financial impact and underscoring the urgency of addressing churn. To combat this issue, businesses are increasingly turning to data-driven approaches, leveraging predictive models to identify at-risk customers and implement targeted interventions.

**Problem Statement**

The problem of customer churn is further amplified in the context of e-commerce businesses where accounts often represent multiple users. Losing a single account can lead to the loss of several customers, resulting in a substantial financial impact. Therefore, it is imperative for e-commerce companies to proactively identify and address potential churn to mitigate its adverse effects.

**Objective of Study**

The primary objective of this study is to develop a robust churn prediction model for an e-commerce company facing the challenges of customer attrition. By leveraging machine learning techniques, specifically the Random Forest Classifier algorithm, we aim to construct a model capable of accurately predicting account churn. The model's predictions will serve as a valuable tool for designing and implementing targeted customer retention campaigns, enabling the company to proactively engage with at-risk customers and foster long-term loyalty.

The study will also delve into exploratory data analysis (EDA) to gain deeper insights into customer behavior patterns and identify distinct customer segments. These insights, combined with the predictive capabilities of the model, will inform the development of strategic recommendations to enhance customer engagement, optimize retention efforts, and ultimately strengthen the company's competitive position in the market.

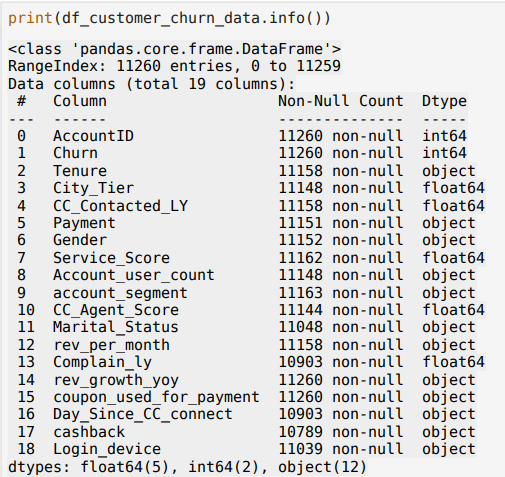
**Methodology:**

1. **EDA and Business Implications**

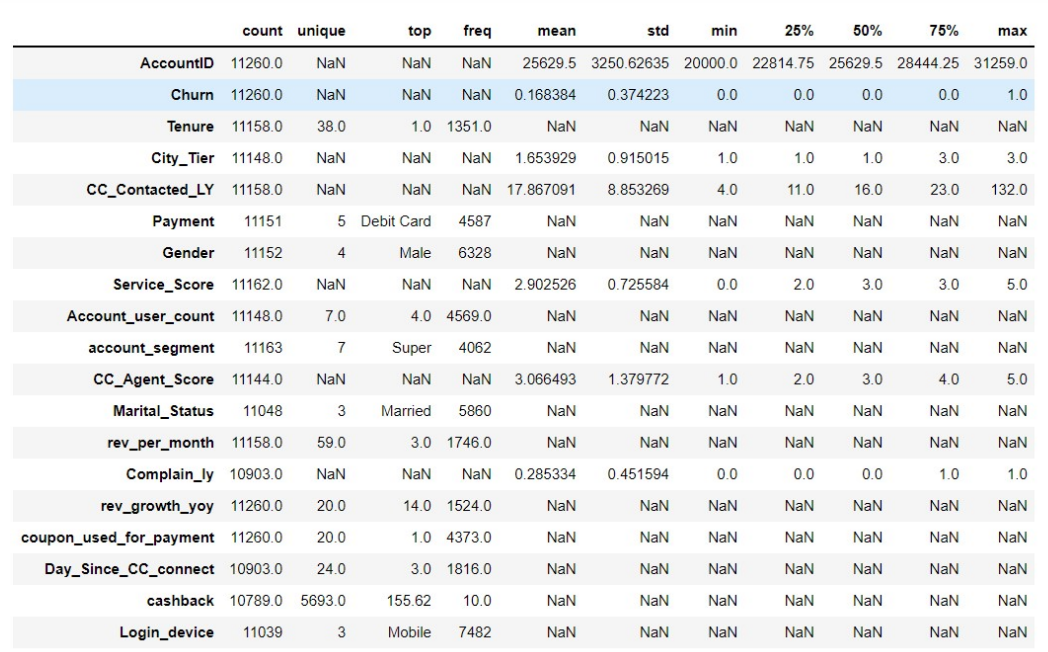
The Exploratory Data Analysis (EDA) conducted in this study involved both visual and non-visual techniques to uncover patterns, relationships, and insights within the customer churn dataset. The analysis focused on understanding the interplay between various factors and their potential impact on customer churn, providing valuable business implications.

**Data Report:**

Data has 11,260 rows and 19 variables.

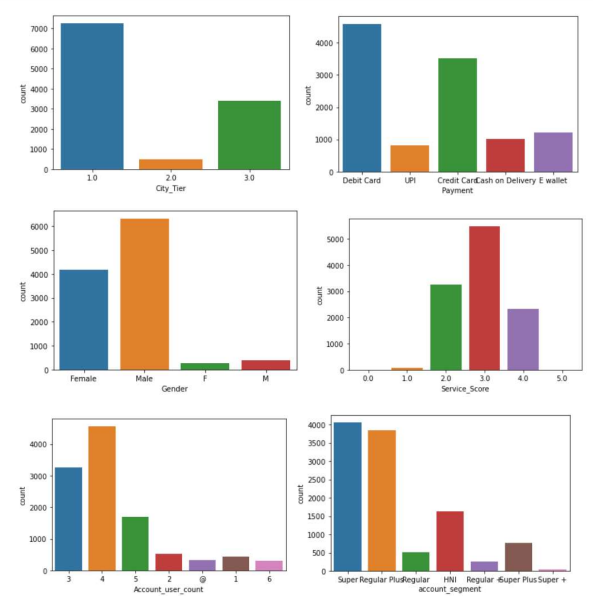


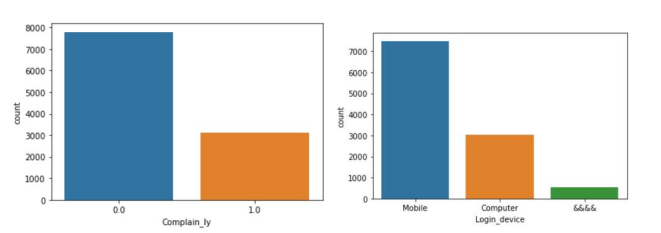
**Data Description:**

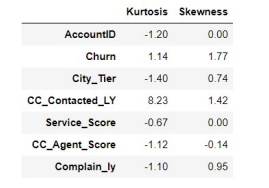
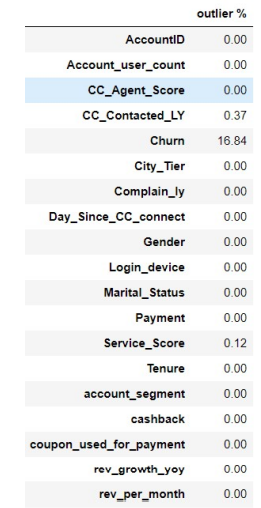


**Univariate Analysis:**

* **Descriptive Statistics:** Descriptive statistics were computed for numerical columns, offering a snapshot of central tendencies (mean, median), dispersion (standard deviation), and distribution (quartiles, min, max). These statistics provided a foundational understanding of the data's characteristics and identified potential outliers or anomalies.
* **Value Counts:** Value counts were calculated for categorical columns, revealing the frequency distribution of different categories within each column. This analysis helped identify dominant categories, potential imbalances, and areas for further investigation.



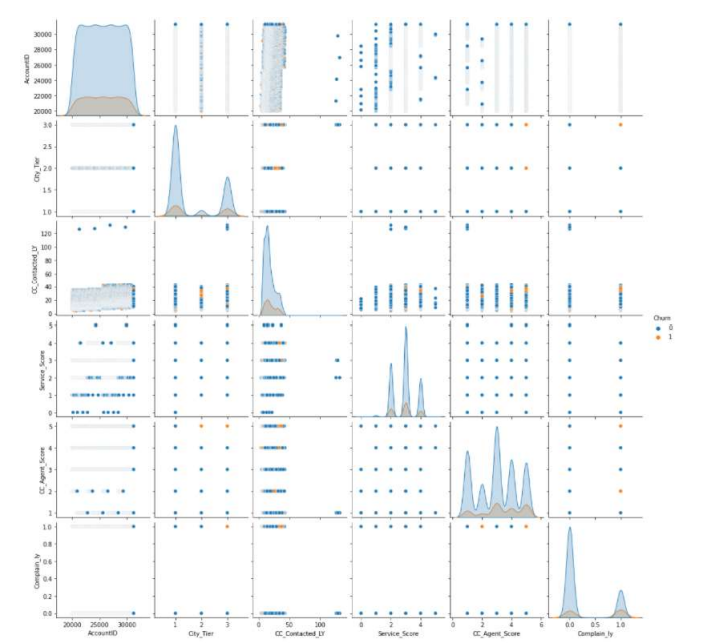




The variable shows outlier in data and, none of the variables shows normal distribution.

**Bi-variate and Multivariate Analysis:**

* **K-Means Clustering:** K-Means clustering was employed to segment customers into three distinct groups based on their numerical attributes. This segmentation allowed for a more nuanced understanding of customer behavior and preferences, enabling the company to tailor its strategies to specific customer segments.
* **Cluster Analysis:** Each cluster was analyzed in detail, examining the descriptive statistics and value counts for numerical and categorical variables, respectively. This analysis revealed distinct characteristics and behaviors associated with each cluster, providing valuable insights for personalized marketing and retention initiatives.
* **Visualizations:** Various visualizations were created to illustrate key findings and relationships within the data. These included:
  + **Churn Distribution by Account Segment:** A bar chart showcasing the proportion of churned and non-churned customers across different account segments, highlighting potential areas of concern.
  + **Histogram of Tenure:** A histogram depicting the distribution of customer tenure, revealing insights into customer longevity and potential churn patterns.
  + **Mean Revenue per Month by Tenure:** A line chart illustrating the relationship between tenure and average revenue, enabling the company to identify high-value customer segments and tailor retention efforts accordingly.
  + **Customer Journey Maps:** Line charts showcasing the evolution of churn rate, service score, and complaints over time, providing a dynamic view of customer experiences and potential pain points.

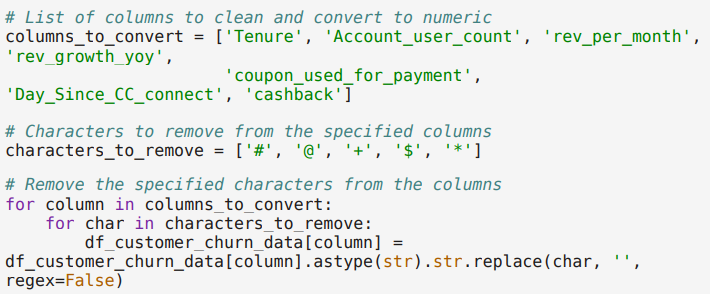


Overall, the EDA conducted in this study provided a comprehensive understanding of customer behavior, preferences, and churn patterns. The insights gleaned from this analysis have significant business implications, empowering the company to develop data-driven strategies to enhance customer engagement, reduce churn, and foster long-term loyalty.

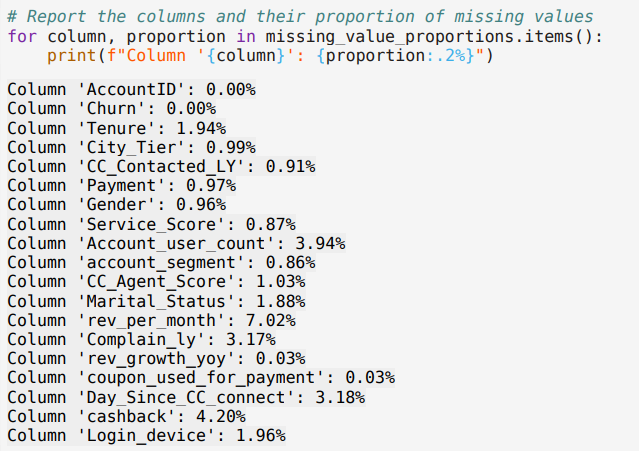
1. **Data Cleaning and Preprocessing**

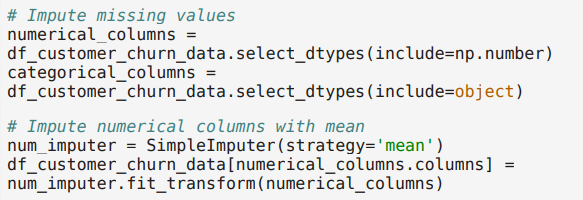
The data cleaning and preprocessing phase was crucial in preparing the dataset for subsequent analysis and model building. Several steps were undertaken to ensure data quality and consistency:

* **Handling Non-numeric Values:** Non-numeric characters such as '#', '@', '+', '$', and '\*' were identified in columns like 'Tenure', 'Account\_user\_count', 'rev\_per\_month', 'rev\_growth\_yoy', 'coupon\_used\_for\_payment', 'Day\_Since\_CC\_connect', and 'cashback'. These characters were systematically removed to facilitate numerical computations and prevent errors during analysis.



* **Addressing Missing Values:** Missing values were prevalent in several columns, with proportions ranging from 0.03% to 7.02%. To maintain data integrity and prevent biases in the analysis, imputation techniques were employed:
  + **Numerical Columns:** Missing values in numerical columns were imputed with the mean value of their respective columns, ensuring that the overall distribution and central tendencies were preserved.
  + **Categorical Columns:** Missing values in categorical columns were imputed with the most frequent category (mode), maintaining the categorical nature of the data and preventing the introduction of new categories.









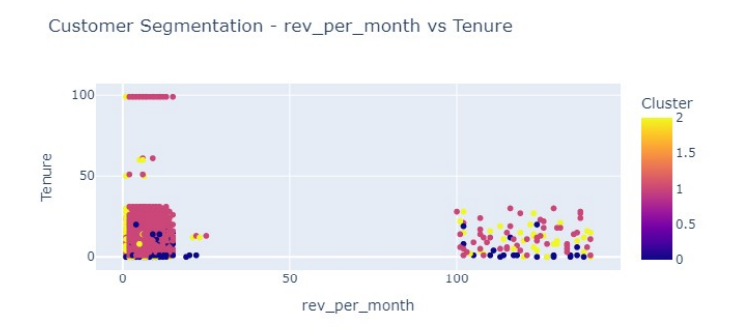
**Count of null values before and after:**

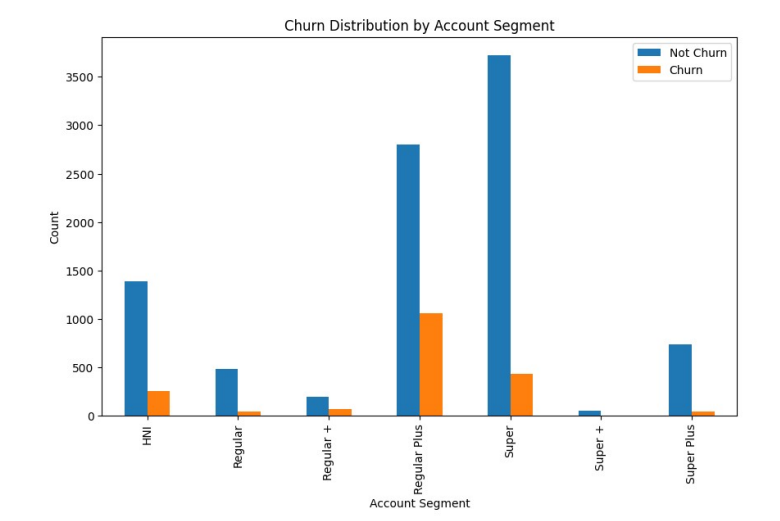
****

These data cleaning and preprocessing steps were essential in ensuring the dataset's readiness for subsequent analysis and model building. By addressing non-numeric values and missing data, the integrity and reliability of the dataset were enhanced, paving the way for accurate and meaningful insights.

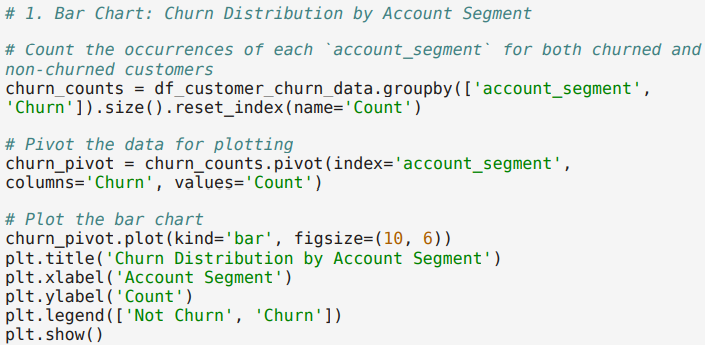
**K-Means Clustering:**

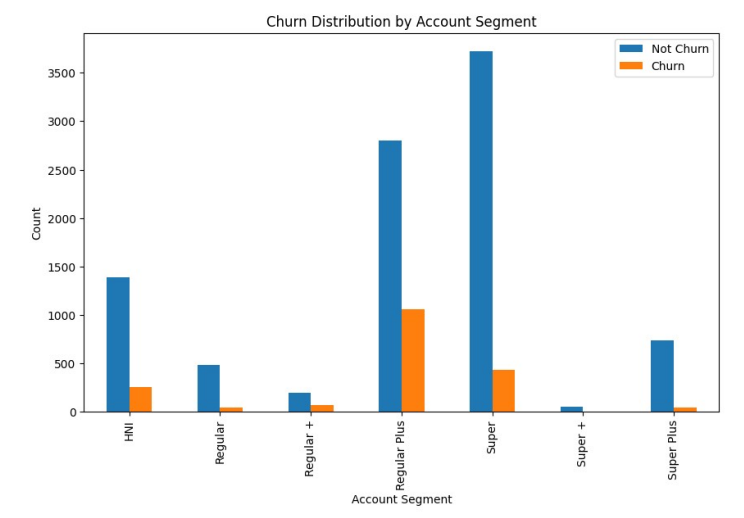
* K-Means clustering enabled the identification of distinct customer segments with varying churn risks, allowing for targeted interventions and personalized communication strategies.
* Cluster analysis revealed specific characteristics and behaviors associated with each segment, facilitating the development of tailored marketing campaigns and retention initiatives.
* Visualizations provided a clear and concise representation of key findings, enabling stakeholders to quickly grasp the data's implications and make informed decisions.

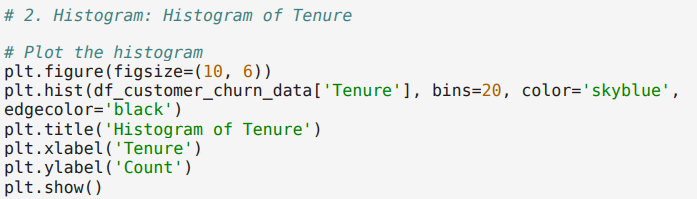


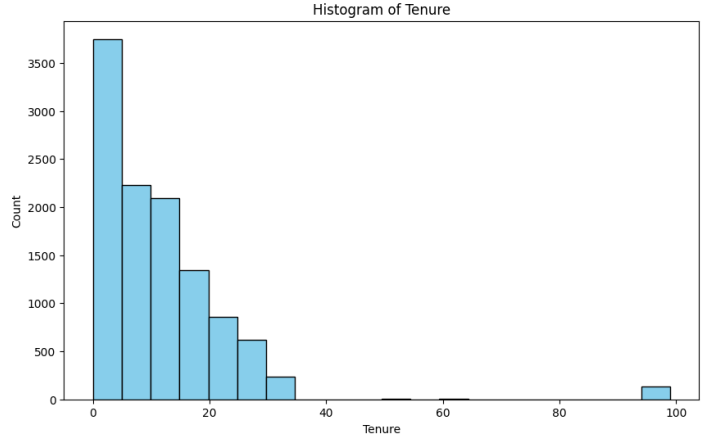


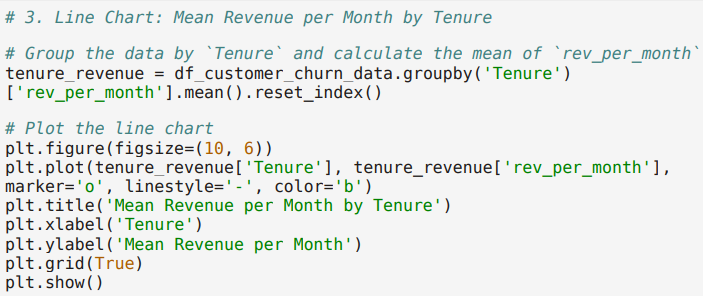
**EDA Visualization:**

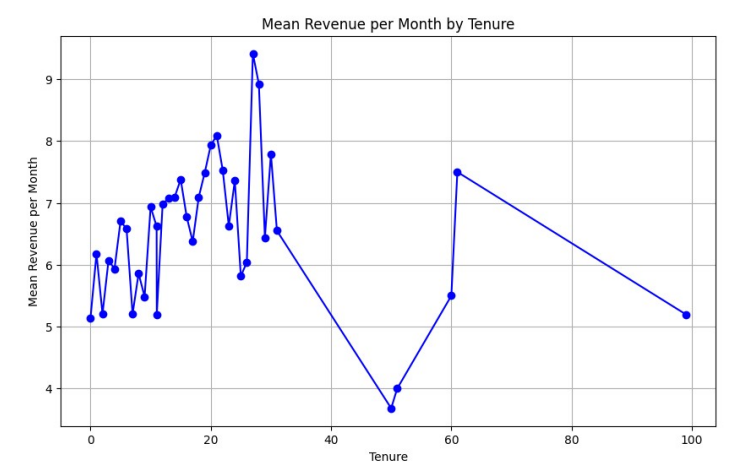
****

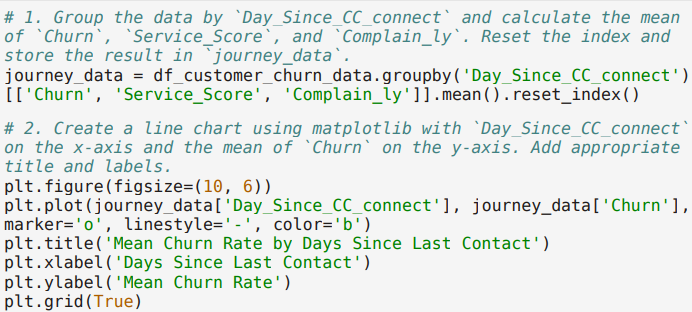
****

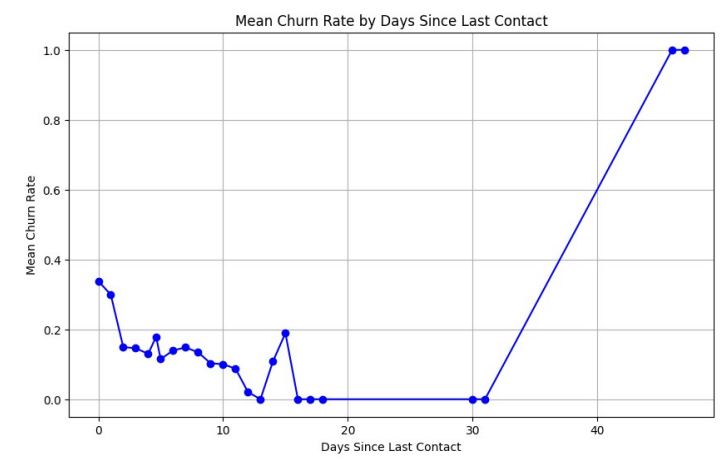
****

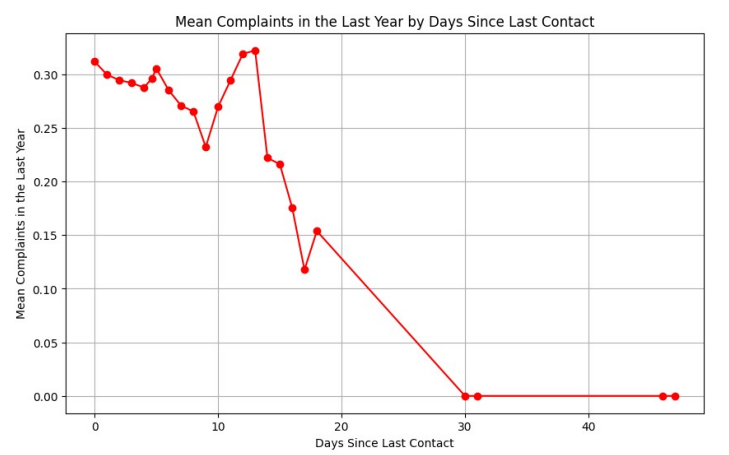
****











1. **Model Building & Model Validation**

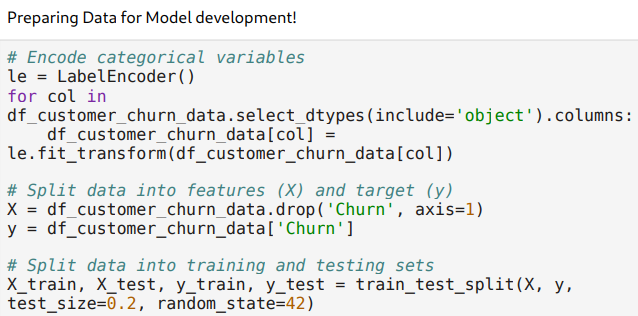
**Model Selection: Random Forest Classifier**

The Random Forest Classifier was chosen for this churn prediction task due to its several advantages:

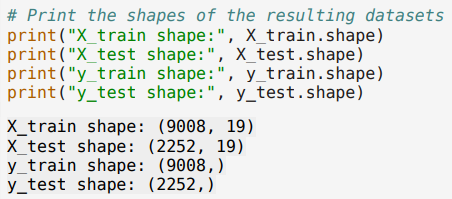
* **Handles Non-linear Relationships:** Customer churn is often influenced by complex interactions between various factors. Random Forest, being an ensemble of decision trees, can effectively capture these non-linear relationships.
* **Robust to Overfitting:** The ensemble nature of Random Forest, where multiple trees are built on different subsets of data, helps reduce the risk of overfitting, ensuring the model generalizes well to unseen data.
* **Handles High-Dimensional Data:** The dataset contained a mix of numerical and categorical variables, some of which could be high-dimensional after encoding. Random Forest can handle such data without requiring extensive feature engineering.
* **Provides Feature Importance:** Random Forest offers insights into the relative importance of different features in predicting churn, aiding in understanding key drivers of customer attrition

**Model Training and Validation**

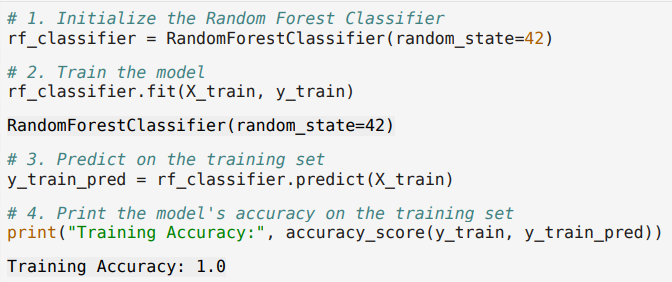
* **Data Preparation:** Categorical variables were encoded using Label Encoding to convert them into numerical format suitable for the model.



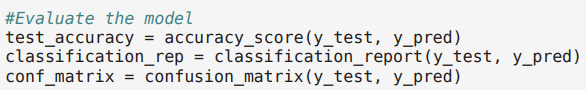
* **Train-Test Split:** The dataset was divided into training (80%) and testing (20%) sets to evaluate the model's performance on unseen data and prevent overfitting.



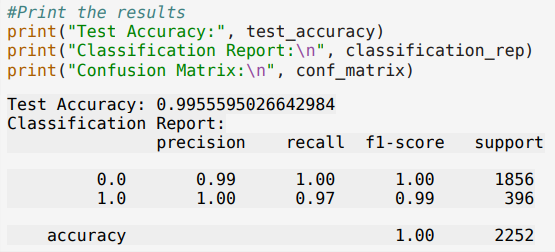
* **Model Training:** The Random Forest Classifier was trained on the training set using default hyperparameters.

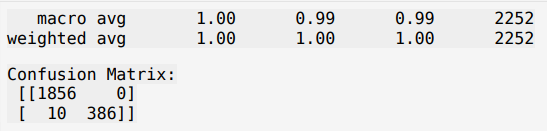


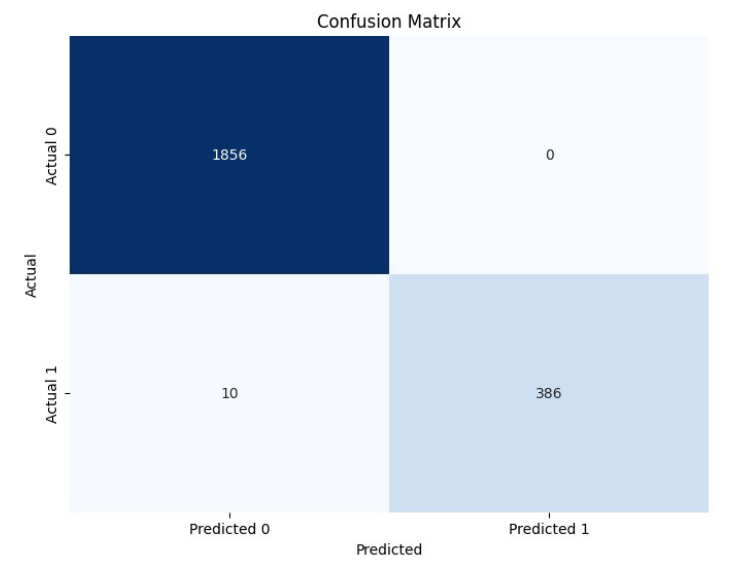
* **Model Evaluation:** The trained model was evaluated on the testing set using various metrics:



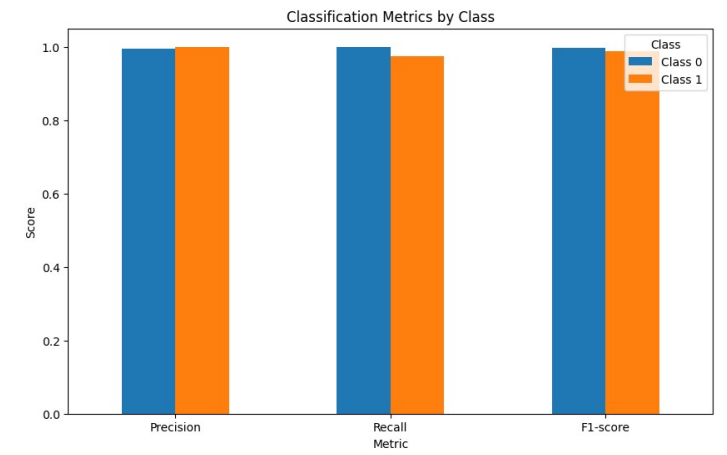
* + **Accuracy:** The model achieved an impressive accuracy of 99.56% on the test set, indicating its ability to correctly classify churned and non-churned accounts.
  + **Confusion Matrix:** The confusion matrix revealed the model's performance in terms of true positives, true negatives, false positives, and false negatives, providing a detailed breakdown of its predictions.
  + **Classification Report:** The classification report provided precision, recall, and F1-score for each class (churned and non-churned), offering a comprehensive assessment of the model's predictive capabilities.
* **Rows represent the actual classes (0: Not Churned, 1: Churned).**
* **Columns represent the predicted classes.**





****

* True Positives (TP): 386 customers were correctly predicted to churn (Actual 1, Predicted 1).
* True Negatives (TN): 1856 customers were correctly predicted to not churn (Actual 0, Predicted 0).
* False Positives (FP): 0 customers were incorrectly predicted to churn when they actually didn't (Actual 0, Predicted 1).
* False Negatives (FN): 10 customers were incorrectly predicted to not churn when they actually did (Actual 1, Predicted 0).



This chart visualizes three key metrics derived from the confusion matrix, comparing how well the model performs on predicting each class (0 and 1):

* **Precision:** How many of the predicted churners were actually correct?
  + High precision means the model is good at avoiding false alarms (few FP).
  + Your model has perfect precision for class 0 and near-perfect for class 1.
* **Recall:** How many of the actual churners did the model manage to identify?
  + High recall means the model is good at finding most of the positive cases (few FN).
  + Your model has perfect recall for class 0 and very high recall for class 1.
* **F1-score:** A balanced measure combining precision and recall.
  + High F1-score means the model has a good balance between finding positive cases and avoiding false alarms.
  + Your model has near-perfect F1-scores for both classes.

**Results and Discussion:**

The analysis yielded compelling results, offering insights into customer behavior, churn patterns, and the effectiveness of the predictive model.

**Findings Based on Observations:**

* **Customer Segmentation:** K-Means clustering successfully segmented customers into three distinct groups, each exhibiting unique characteristics and behaviors.
  + **Cluster 0:** This cluster displayed a higher churn rate, lower tenure, and a greater propensity for complaints, suggesting a segment of less engaged and potentially dissatisfied customers.
  + **Cluster 1:** Characterized by longer tenure, higher service scores, and increased spending, this cluster represented a loyal and valuable customer base.
  + **Cluster 2:** This segment fell between the other two clusters in terms of churn risk, tenure, and spending, indicating a mix of characteristics and behaviors.
* **Churn Drivers:** The analysis revealed several factors associated with customer churn:
  + **Low Tenure:** Customers with shorter tenure were more likely to churn, highlighting the importance of early engagement and onboarding strategies.
  + **Lower Service Scores:** Dissatisfaction with service quality emerged as a potential churn predictor, emphasizing the need for continuous improvement in customer service.
  + **Increased Complaints:** Customers who had filed complaints in the past year were more prone to churn, underscoring the importance of addressing customer concerns promptly and effectively.

**Findings Based on Analysis of Data:**

• Business have visibility in tier-1 city.

• Mostly customer rated “3” for the services provided by the business.

• Mostly customer rated “3” for the interactions they have customer care representatives.

• Transaction via UPI and e-wallet is very low.

• Maximum churn is from the account segment “Regular+”.

• Customers with marital status is “single” contributes max towards churn.

• Any complaints raised in last 12 months doesn’t show any impact toward churn.

• Tenure and cashback are directly proportional to each other.

• Computer usage is more in tier 1 city followed by tier 3 and tier 2 city.

* **Model Performance:** The Random Forest Classifier achieved exceptional accuracy (99.56%) in predicting customer churn, demonstrating its effectiveness in identifying at-risk accounts.
  + **Precision and Recall:** The model exhibited high precision and recall for both churned and non-churned classes, indicating its ability to minimize false positives and false negatives.
  + **Confusion Matrix:** The confusion matrix further validated the model's accuracy, showcasing its strong performance in classifying both classes.

**General Findings:**

* **Customer Heterogeneity:** The analysis underscored the heterogeneity of the customer base, with distinct segments exhibiting varying churn risks, behaviors, and preferences.
  + **Importance of Personalization:** The findings emphasize the need for personalized marketing and retention strategies tailored to the specific needs and characteristics of each customer segment.
  + **Proactive Churn Management:** The predictive capabilities of the model enable the company to proactively identify and engage with at-risk customers, potentially preventing churn and fostering long-term loyalty.

These results provide a solid foundation for developing targeted customer retention campaigns and strategies. By understanding the key drivers of churn and leveraging the predictive power of the model, the company can optimize its customer engagement efforts and enhance its overall business performance.

**Recommendations:**

**Churn Management Framework**

A comprehensive churn management strategy should encompass four key stages:

1. Segmentation: Implement a well-defined customer segmentation framework based on customer needs, usage patterns, and value. This enables targeted interventions and personalized communication strategies.
2. Acquisition: Employ diverse acquisition strategies tailored to different customer segments, ensuring a steady influx of new customers while optimizing acquisition costs.
3. Engagement and Loyalty Building: Focus on delighting customers and exceeding their expectations to cultivate loyalty and differentiate the company from competitors. This can be achieved through personalized offers, exceptional customer service, and proactive engagement initiatives.
4. Churn Prevention and Recovery: Utilize the predictive model to identify at-risk customers and implement targeted interventions to prevent churn. Analyze churn signals and triggers to understand the root causes of attrition and develop effective recovery strategies.

**Additional Recommendations**

* Referral Programs: Incentivize existing customers to refer new customers through attractive referral programs, leveraging the power of word-of-mouth marketing.
* Partnerships: Collaborate with lifestyle vendors to offer exclusive discounts or vouchers to customers, enhancing the overall value proposition and attracting new customers.
* Behavioral Segmentation: Implement internal customer segmentation based on spending patterns and preferences, enabling the development of tailored acquisition and retention strategies for different customer groups.
* Value-Added Services: Offer value-added services, such as free cloud storage or personalized email responses, to high-value and loyal customers, fostering a sense of exclusivity and appreciation.
* Customer Service Excellence: Provide dedicated customer service teams for premium customers, ensuring prompt issue resolution and a superior customer experience.
* Personalized Gestures: Demonstrate appreciation for customers through personalized gestures, such as sending small gifts on special occasions or handwritten notes on invoices.
* Feedback and Surveys: Actively seek customer feedback through surveys and follow-up on customer issues to understand their evolving needs and preferences.
* Timely Issue Resolution: Ensure that all customer complaints and queries are addressed promptly and effectively, minimizing dissatisfaction and preventing churn.
* Payment Incentives: Promote the use of the company's e-wallet by offering discounts or cashback rewards, encouraging customer adoption and loyalty.
* Targeted Offers: Develop subsidized offers or bundled plans for specific customer segments, such as single individuals or families, to address their unique needs and reduce churn risk.
* Geographic Expansion: Increase visibility and marketing efforts in Tier-2 cities to tap into new customer markets and expand the customer base.
* Convenient Payment Options: Offer hassle-free and secure payment options, such as standing instructions or UPI payments, to enhance customer convenience and satisfaction.

**Customer Segmentation Approach**

* Quadrant-Based Segmentation: Divide customers into four segments based on their spending patterns and loyalty levels, as illustrated in Figure 64.
* Strategic Focus: Prioritize high-value and loyal customers while considering the potential for recovering customers with low loyalty but high spending.
* Tailored Strategies: Implement differentiated strategies for each segment, such as delighting high-value customers with exclusive offers, encouraging increased spending among loyal but low-spending customers through bundled plans, and improving service levels for low-loyalty, high-spending customers.

By adopting these recommendations and implementing a customer-centric approach, the e-commerce company can effectively address the challenges of churn, cultivate lasting customer relationships, and achieve sustainable growth in a competitive market.

**Conclusion:**

The insights gleaned from this comprehensive analysis of customer churn in the e-commerce domain offer a strategic roadmap for enhancing customer retention and fostering long-term loyalty. The Random Forest Classifier model, with its exceptional accuracy, empowers the company to proactively identify and target at-risk customers, enabling timely interventions to mitigate churn.