CAPSTONE PROJECT–CUSTOMER CHURN

BY

SAURAV CHAKRABORTY

Table of Contents

Problem Statement:

An E Commerce company or DTH (you can choose either of these two domains) provider is facing a lot of competition in the current market and it has become a challenge to retain the existing customers in the current situation.

Hence, the company wants to develop a model through which they can do churn prediction of the accounts and provide segmented offers to the potential churners. In this company, account churn is a major thing because 1 account can have multiple customers. hence by losing one account the company might be losing more than one customer.

You have been assigned to develop a churn prediction model for this company and provide business recommendations on the campaign.

Your campaign suggestion should be unique and be very clear on the campaign offer because your recommendation will go through the revenue assurance team. If they find that you are giving a lot of free (or subsidized) stuff thereby making a loss to the company; they are not going to approve your recommendation.

Need of the Study:

This research is justified by the significant effect of account churn on both revenue and the customer base of the company. Because an account may have several customers, losing one account can result in significant loss of business. Retaining current customers is cheaper than trying to acquire new customers, so predicting churn is important for ensuring the sustainability of a business. By identifying valuable predictors of churn, the company can implement retention strategies that will help maintain satisfaction and preserve profits. Also, receiving a churn prediction, the company will have an advantage over the competition by being able to seek out proactive engagement to reduce revenue loss and improve marketing strategies to target high-value customers at the lowest cost.

Understanding the business opportunity:

To comprehend the business problem, we need to analyze the impact of customer churn on the company's revenue and sustainability over the long term. One important consideration is that there can be several customers associated with one account, therefore losing an account could represent the loss of multiple customers and revenue. The critical component will be to determine the reason customers leave. Reasons could be price, level of service, comparable offerings from competition, or lack of engagement. Additionally, understanding market competitors and existing retention efforts will aid in developing effective interventions. The aim is to develop a data-driven model to predict churn that ultimately helps with proactive retention efforts, all while ensuring that customer retention campaigns are feasible and focused on business objective.

DATA REPORT

Data Dictionary:

Variable	Description
AccountID	account unique identifier
Churn	account churn flag (Target)
Tenure	Tenure of account
City_Tier	Tier of primary customer's city
CC_Contacted_L12m	How many times all the customers of the account has contacted customer care in last 12m
Payment	Preferred Payment mode of the customers in the account
Gender	Gender of the primary customer of the account
Service_Score	Satisfaction score given by customers of the account on service provided by company
Account_user_count	Number of customers tagged with this account
account_segment	Account segmentation on the basis of spend
CC_Agent_Score	Satisfaction score given by customers of the account on customer care service provided b
Marital_Status	Marital status of the primary customer of the account
rev_per_month	Monthly average revenue generated by account in last 12 months
Complain_l12m	Any complaints has been raised by account in last 12 months
rev_growth_yoy	revenue growth percentage of the account (last 12 months vs last 24 to 13 month)
coupon_used_l12m	How many times customers have used coupons to do the payment in last 12 months
Day_Since_CC_connect	Number of days since no customers in the account has contacted the customer care
cashback_l12m	Monthly average cashback generated by account in last 12 months
Login_device	Preferred login device of the customers in the account

Data Ingestion:

The dataset has total 11260 rows and 19 columns.

```
(11260, 19)
```

Statistical Analysis of the dataset:

	count	mean	std	min	25%	50%	75%	max
AccountID	11260.0	25629.500000	3250.626350	20000.0	22814.75	25629.5	28444.25	31259.0
Churn	11260.0	0.168384	0.374223	0.0	0.00	0.0	0.00	1.0
City_Tier	11148.0	1.653929	0.915015	1.0	1.00	1.0	3.00	3.0
CC_Contacted_LY	11158.0	17.867091	8.853269	4.0	11.00	16.0	23.00	132.0
Service_Score	11162.0	2.902526	0.725584	0.0	2.00	3.0	3.00	5.0
CC_Agent_Score	11144.0	3.066493	1.379772	1.0	2.00	3.0	4.00	5.0
Complain_ly	10903.0	0.285334	0.451594	0.0	0.00	0.0	1.00	1.0

Key Observations:

Account ID: This is a unique identifier for each account.

Churn: The average churn rate is approximately **16.8%**, indicating that around this percentage of accounts tend to leave.

City_Tier: Most cities fall between tiers **1 to 3**.

CC_Contacted_LY (Customer Care Contacted Last Year): Customers contacted customer care an average of **17.86** times, with a maximum of **132** contacts.

Service_Score: The average service score is **2.90** (on a scale of 1 to 5), showing room for improvement.

CC_Agent_Score: The agent score averages **3.06**, suggesting mixed customer service satisfaction.

Complain_ly (Complaints Last Year): Around **28.5%** of accounts have registered at least one complaint.

Number of Missing values in the Dataset:

AccountID	0
Churn	0
Tenure	102
City_Tier	112
CC_Contacted_LY	102
Payment	109
Gender	108
Service_Score	98
Account_user_count	112
account_segment	97
CC_Agent_Score	116
Marital_Status	212
rev_per_month	102
Complain_ly	357
rev_growth_yoy	0
coupon_used_for_payment	0
Day_Since_CC_connect	357
cashback	471
Login_device	221

Key Observations:

- **No missing values in AccountID, Churn, rev_growth_yoy, and coupon_used_for_payment**, meaning these columns have complete data.
- **Columns with significant missing values:**
  - cashback (471 missing values)
  - Day_Since_CC_connect (357 missing values)
  - Complain_ly (357 missing values)
  - Login_device (221 missing values)
  - Marital_Status (212 missing values)
- **Other columns with some missing data:**
  - Tenure, City_Tier, CC_Contacted_LY, Payment, Gender, Service_Score, CC_Agent_Score, and account_segment each have between 97 to 116 missing values.

EDA and Business Analysis:

Understanding the Attributes in the dataset:

1. The dataset includes 19 attributes, categorized as follows:
2. Unique Identifiers: AccountID (dropped during preprocessing as it is not analytically useful).
3. Target Variable: Churn (binary indicator of account status: 1 for churned, 0 for active).
4. Demographics: Gender, Marital_Status.
5. Behavioral Metrics: Tenure, City_Tier, Account_User_Count, Days_Since_CC_Contact.
6. Financial Metrics: Revenue_Per_Month, Revenue_Growth_YoY, Cashback_Amount, Coupons_Used.
7. Service Metrics: Service_Score, CC_Contacts_LastYear, CC_Agent_Score, Complaints_LastYear.
8. Preferences: Payment, Login_Device, Account_Segment.

```
Index(['AccountID', 'Churn', 'Tenure', 'City_Tier', 'CC_Contacts_LastYear',
       'Payment', 'Gender', 'Service_Score', 'Account_User_Count',
       'Account_Segment', 'CC_Agent_Score', 'Marital_Status',
       'Revenue_Per_Month', 'Complaints_LastYear', 'Revenue_Growth_YoY',
       'Coupons_Used', 'Days_Since_CC_Contact', 'Cashback_Amount',
       'Login_Device'],
```

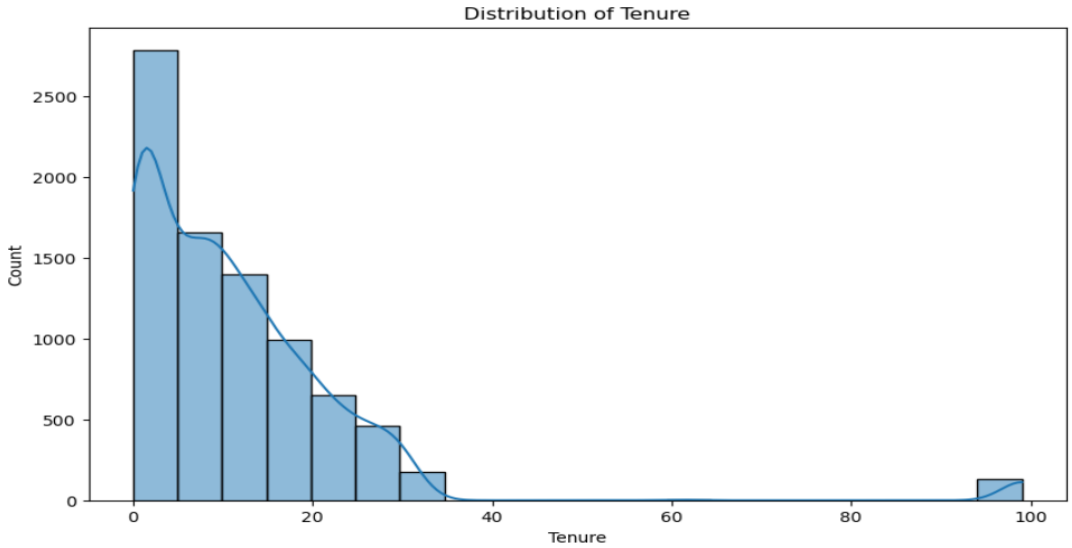
Renaming the columns for clarification:

- **CC_Contacted_LY(old) → CC_Contacts_LastYear(new)**
- **Account_user_count(old) → Account_User_Count(new)**
- **rev_per_month(old) → Revenue_Per_Month(new)**
- **Complain_ly(old) → Complaints_LastYear(new)**
- **rev_growth_yoy(old) → Revenue_Growth_YoY(new)**
- **Day_Since_CC_connect(old) → Days_Since_CC_Contact(new)**

- **cashback(old) → Cashback_Amount(new)**
- **coupon_used_for_payment(old) → Coupons_Used(new)**

```
Index(['AccountID', 'Churn', 'Tenure', 'City_Tier', 'CC_Contacts_LastYear',
       'Payment', 'Gender', 'Service_Score', 'Account_User_Count',
       'Account_Segment', 'CC_Agent_Score', 'Marital_Status',
       'Revenue_Per_Month', 'Complaints_LastYear', 'Revenue_Growth_YoY',
       'Coupons_Used', 'Days_Since_CC_Contact', 'Cashback_Amount',
       'Login_Device'],
```

Univariate Analysis



Main Findings:

• Right-Skewed Distribution: Tenure shows a strong right-skewed distribution. This indicates that most customers do not have a long tenure with the company.

• Peak at 10: The tenure distribution peaks around 10 years, suggesting a reasonable percentage of customers are with the business around that time frame.

• percentage of customers are with the business around that time frame.

• Long Tail: While most customers are clustered around shorter tenure values, there is a long tail toward higher tenure values that gutter off substantially. This suggests there are at least some customers who have been with the company for an extended period of time or even a lifetime, with the longest recorded values of this at 100 years.

• Outliers: There are a few conspicuous outliers with extremely high tenure values, which may be indicative of long-term loyal customers or data entry mistakes.

Interpretations:

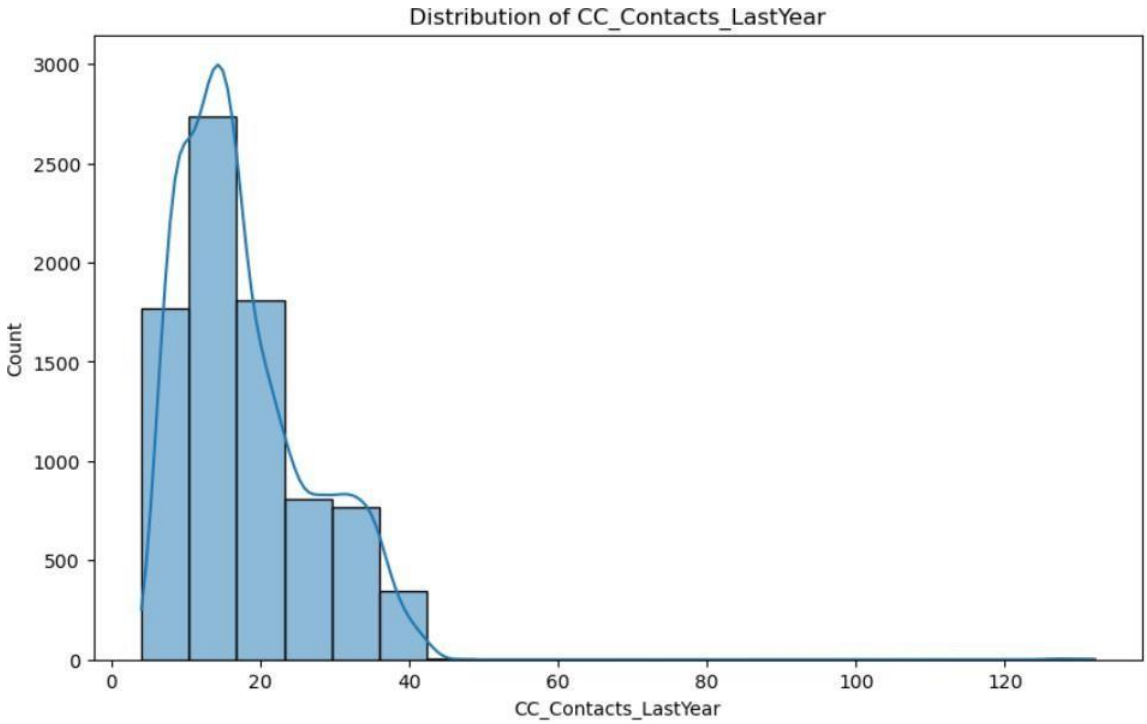
• Customer Loyalty: The right-tailed distribution suggests a large proportion of customers are relatively new. There may be opportunity to enhance customer retention methods to promote longer customer relationships.

• Customer Segmentation: The distribution could allow for the identification of customer segments according to tenure.

Long-Term Customers: Customers with a tenure of greater than 5 years.

• Customer Value: Long-term customers may often carry a higher value. Understanding how they behave, or their preferences can generally provide you strategies to use with this segment.

• Outlier Analysis: Analyzing these outliers with greater than average tenure can yield some valuable information regarding loyalty, poor data quality, or an associated segment of customers.



Observations:

1. Shape of Distribution: - The distribution is right-skewed, indicating that most customers have had low CC contacts in the past 12 months. - There are a handful of customers with much higher values contributing to a long, lower tail of the distribution.

2. Mode: - The mode (most common value) seems to be 10-20 contacts.

3. Outliers: - There are outliers in the distribution at 40 contacts, and 60 contacts, and a few at 100-120 contacts, but these are extremely rare.

4. Spread: - Most customers surveyed are between 10 and 30 contacts which makes up the majority of CC activity.

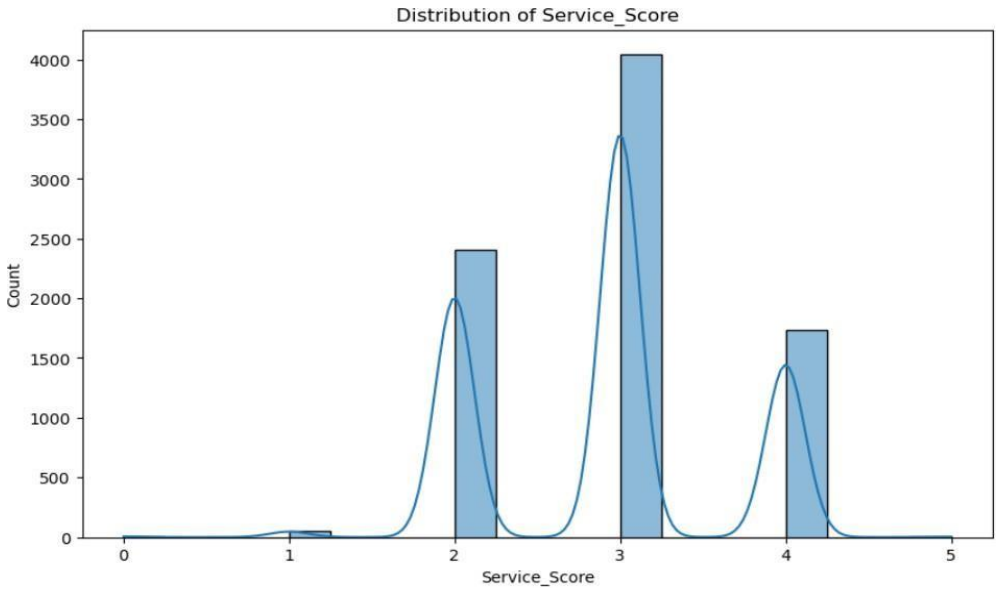
Business Implications:

1. Customer Support Optimization: Most customers fall in the low-contact area and this could mean that customers are having their issues resolved satisfactorily, or that there are limited reasons for contacting. It could also mean that they lacked proactive communication.
2. High-Contact Customers: Customers with levels of contact above 40 should be assessed for common issues or concerns. They may indicate a need for enhanced support operations or may require escalation.

3. Outlier Investigation: Extremely high interaction counts (>100) may mean the following:

There is a cohort of customers requiring significant support or assistance, or there is an error in collecting or encoding the data.

4. Support Magnitude Improvement: Leverage the data to more effectively deploy customer support resources (staff time, specially-developed programs, targeted facilitation). For example, you may want to focus on optimizing resource support with the bulk of the experience with customers who have fewer contacts (10-30 range), and then suggest specialized support plans, or programs for high-contact customers.



Notable Points:

1. The Distribution is Multimodal: There are clearly distinct peaks in the distribution at scores of 1, 3, and 4. This indicates that the Service Score is not uniformly distributed, and moreover, has a multimodal distribution, a distribution that is clustered at certain scores.

2. Clustering At Certain Scores: The multiple peaks suggest that many of the Service_Scores are suggestive of somewhat distinct scoring patterns. Perhaps there are service standards that impact how students score the services, customer expectations that influence their rating or feedback methodologies.

3. Skewness: While the multimodal distribution is present, the overall distribution does appear to have a moderate positive right skew, a tendency to award higher service ratings. This could also indicate a more substantive perception of service quality, which could also be moderating the distribution.

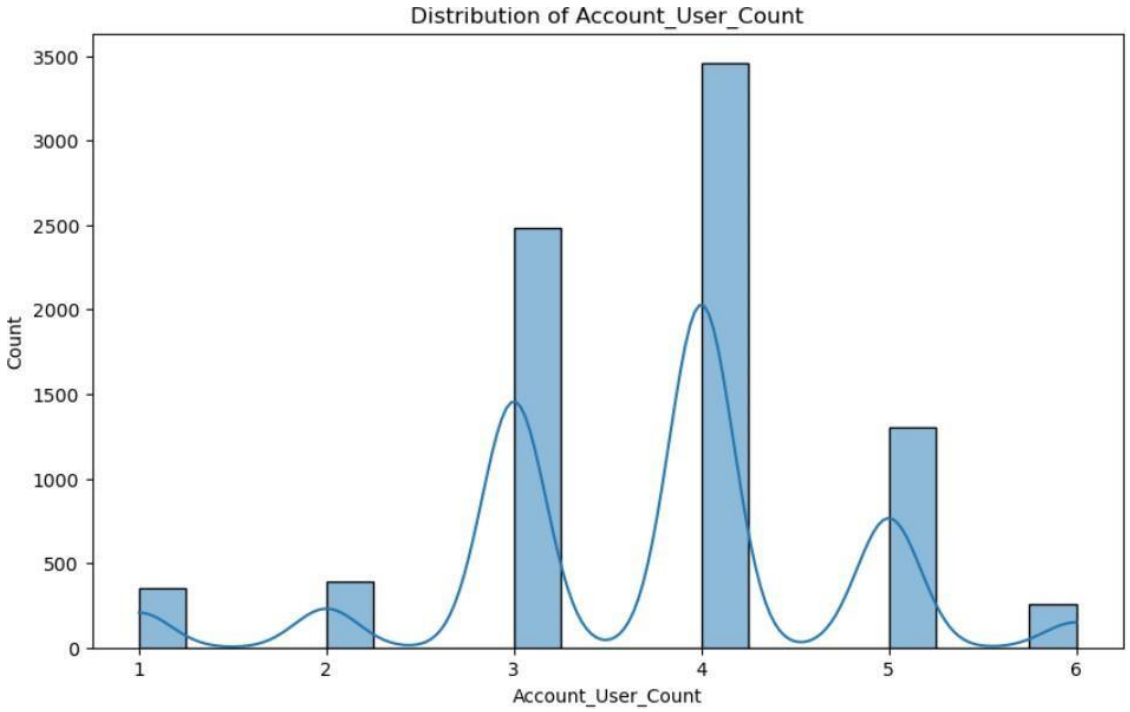
Interpretation:

• Service Quality Evaluation: The distribution gives a sense of how customers evaluate the service quality. Many scores seem to be clustered together, indicating that different types of service experiences may have occurred.

• Performance Standards: With the mode being the highest point in the distribution, it serves as a performance metric for this service. For example, if the mode is 3, then a large percentage of customers rated the service as such.

• Opportunities for Enhancement: The mode being at the lowest end (e.g., a score of 1) represents opportunities to improve service quality on these experiences. Studying the language that customers use related to these low scores is likely valuable for identifying the necessary corrective actions.

Customer Segmentation: The distribution curve can be a viable tool to segment customers by service rating. At a minimum, attempting to segment customers for marketing and service purposes will be more effective than one size fits all approaches



Important Observations:

1. Discrete Distribution: Since Account_User_Count is a whole value (such as 1, 2, 3, 4, 5, 6), this is a discrete distribution. The histogram and the KDE (kernel density estimate) curve represent the frequency of the different counts.

2. Multimodal Distribution: This is a multimodal distribution with three noticeable peaks near 3, 4, and, 5. This suggests that these user counts are common among

accounts, as well as "smaller" peaks at 1, 2, and, 6.

3. Peak Counts: The counts of 3 and 4 users are more pronounced than the count of 5. Accounts with more users tend to be 3 or 4 users, with the peak count at 4 users.

4. Francs of Extremes: The 1, 2, and 6 user counts are much lower in frequency than 3, 4, and 5 user counts, which would suggest accounts with extremely few or extremely many users would be an account type facies, related to commonality.

Interpretation:

Typical Account Size: The peaks at 3 and 4 indicate that the average account size is approximately 3-4 users. This may be attributed to factors such as typical team size, pricing tiers which support this type of user number, or product features that are intended for this user range.

- Segmentation Opportunities: The two separate peaks indicate potential customer segmentation based on account size:

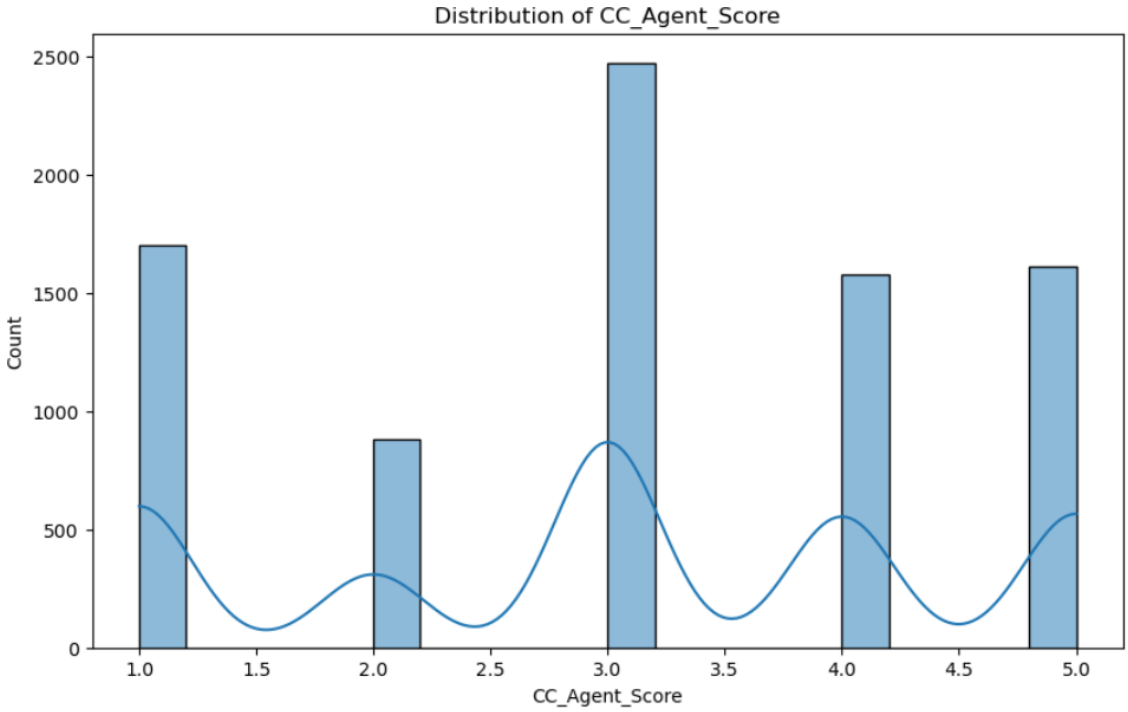
Small Accounts (1-2 Users): These would be individual users or very small teams.

Medium Accounts (3-5 Users): This is the largest account size segment, likely representing small to medium business or teams.

Large Accounts (6 Users): These accounts are much rarer, but likely represent a larger company or enterprise account.

- Product/Feature Targeting: Having this knowledge of distribution can contribute to product features in product development prioritization, for example, Team features or collaborative features among users would be especially applicable for the mid-size accounts (3-5 users)o Individual user or very small team features, would not be as applicable given the lower number of accounts in this category.

- Marketing Options: You may wish to consider different marketing and sales opportunities to target different account size segments. Fr example, an enterprise sales team may consider targeting the larger accounts (6) and self-service (online, etc), marketing or promotions would be targeted toward smaller sized accounts.



Key Points:

1. Multimodal Distribution: The distribution is multimodal in nature, with several clear peaks. The largest peaks for scores are near 1, 3, and 4. This might suggest that many agents scored in that area.

2. Clustering Around Scores: The presence of multiple peaks suggests that there may be more than one type of performance or levels of performance amongst customer service agents.

3. Skewness: While there are multiple peaks, the overall distribution appears slightly skewed to the right, indicating more emphasis on scores that are higher than average. This may suggest an overall viewpoint that the performance of agents is acceptable overall.

Interpretations:

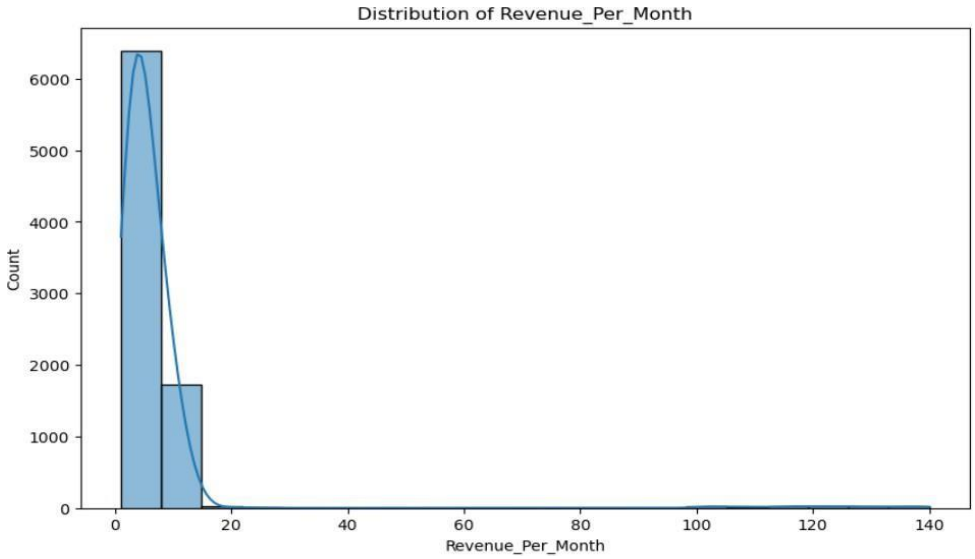
- Performance Levels: The peaks in the distribution at points 1, 3, and 4 likely indicate different performance levels among agents. Individuals near 1 could benefit from additional training and support; those in the center (near 3) are performing at an average level; while agents with scores of about 4 may be considered high performers.

- Service Quality Evaluation: The distribution helps to shed light on the overall service quality of the customer service agents. Additionally, the presence of multiple peaks suggests there may be variances in agents' performance based on differing teams, agents, or service delivery channels.

- Recognition and Incentives: The distribution can be thought of as a means of

identifying agents who have scored high and consistently performed well; subsequently, a means of reinforcing high performance through recognition and incentives that could motivate individual agents to continue to perform at a high level, while also improving service quality.

- Training and Development: The lowest scores (1) should be analysed to better understand the potential causes of lower performance, which in turn could help to identify areas of focus in training and coaching agent to their maximum potential.



Important Insights:

1. Right-Skewed Distribution: The distribution is markedly right-skewed, suggesting that most accounts generate relatively low "Revenue_Per_Month". This implies these accounts have lower amounts of monthly revenue.

2. Peak In the Range Of 0-10: We observe a large peak in the 0-10 x-axis range indicating that a large number of accounts have very low to no revenues per month.

3. Long Tail: There is a long tail in the distribution as you accrue larger revenue amounts reflecting that a limited number of accounts produce higher revenues each month.

4. Potential Outliers: Although they are not directly depicted in the provided image, outliers or extreme higher revenues may exist, denoting an account with higher revenues.

Interpretations:

- Revenue Distribution: A right-skewed distribution denotes that revenue is not evenly distributed across accounts; most accounts create low revenue, while a smaller number of accounts create a high revenue span.

- Revenue Potential: The long tail reveals the long tail of high-revenue accounts that could generate more revenue. Identifying and understanding high-revenue accounts is key for revenue-optimizing strategies.

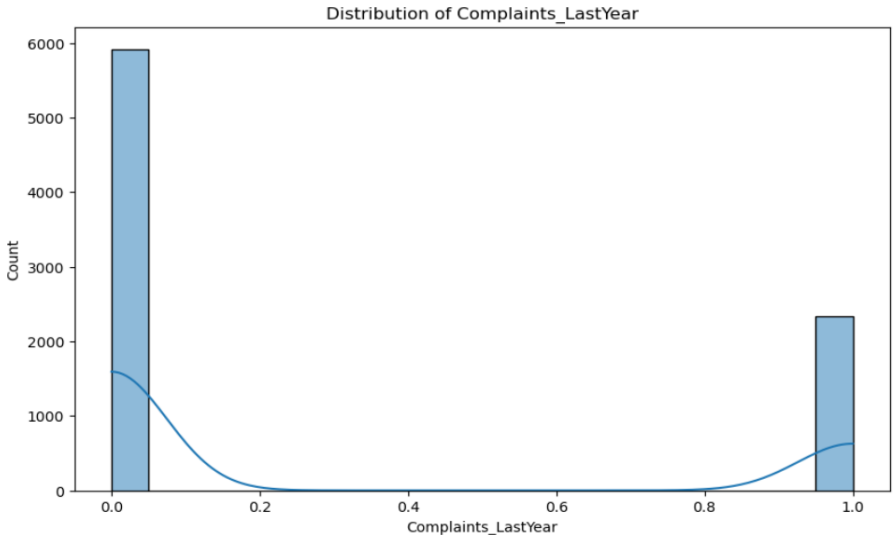
- Customer Segmentation: One can segment customers or accounts into tiers using the

distribution determined by revenue or MRR:

Low Revenue Accounts: Accounts with monthly revenue around the lower end. Mid-Revenue Accounts: Accounts with monthly revenue in the middle.

High-Revenue Accounts: Accounts providing much higher monthly revenue.

- Resource Allocation: Understanding the revenue distribution creates a foundation for more effective resource allocation. For instance, expanding efforts to attract or retain high-revenue accounts creates a dollar-for-dollar opportunity. Overall revenue growth can compound with expanding efforts in high-revenue accounts.



Key Observations:

1. The distribution is clearly bimodal, having two peaks. One large peak is located at 0, and one smaller, but still significant peak is located at 1.

2. Most of the observations are concentrated on the two extremes of either 0 complaints or 1 complaint in the last year.

3. There are very few observations between 0 and 1, indicating a clear distinction between customers who submitted no complaints and those who submitted one.

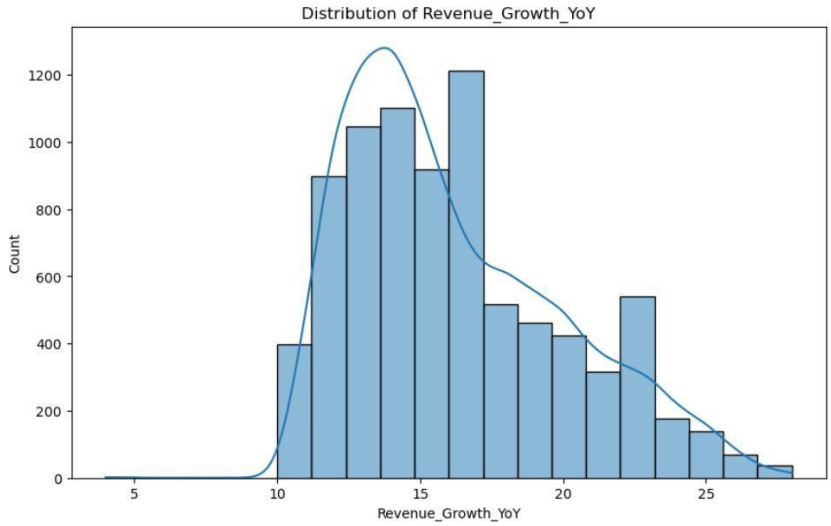
Business Implications:

- Prioritize Complaint Resolution: It is important to identify the reasons for complaints from the customers in the "1 complaint" group. These concerns can be resolved in ways that can enhance customer satisfaction and lower the chances of high complaints in the future.

- Customer Retention: Identifying the characteristics of customers with complaints may reveal potential churn risks or allow for strategies to mitigate further churn.

- Identify process improvements: Discovering the root concerns surrounding complaints may provide process improvements or help prevent future issues.

- Segmented communication strategy: Different segmented communication strategies may be needed for customers from the two segments.



Key Observations:

1. Considering a Bimodal/Multimodal Distribution: Not separate peaks but a tendency for bimodality or multimodality is observed here. The vast majority are clustered between 12 and 17 and fewer are clustered around 22-24.

2. Majority Clustering (12-17): The observations cluster in the 12-17 percent revenue growth year over year range. This suggests that this range is a normal or common growth rate in terms of the accounts. The peak is in the 15 percent band.

3. Minority Clustering (22-24): The fewer number of observations clustered around 22-24 percent growth suggest that a minority of accounts grew at a much higher rate than the majority.

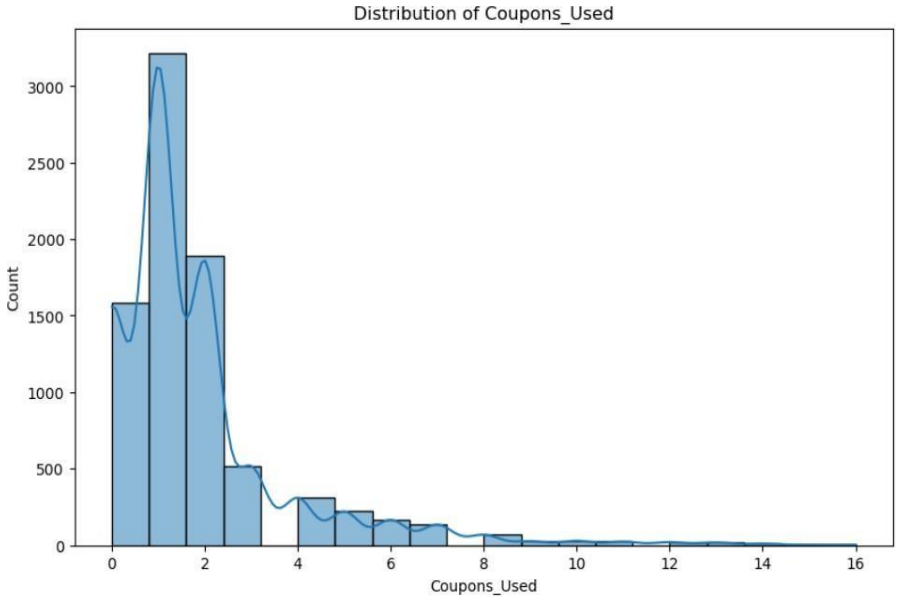
4. Limited Data Available at Low (Below 10) or High Growth (Above 25): There are few observations indicating revenue growth below 10 percent or above 25 percent. This suggests that very low or very high growth rates are similarly uncommon.

Interpretation:

- Multiple Growth Profiles: The identified growth distribution indicates at least two growth profiles across the accounts: "moderate growth" profile (in the range of 12-17%) and "high growth" profile (in the range of 22-24%).

- Performance Specification: The distribution serves as a benchmark for assessing the extent of revenue growth. The concentration of outcomes around 15% might be viewed as an average or target growth rate while the higher growth profile indicates the best performing accounts.

• Strategic Planning: The distribution can assist you in strategic planning by revealing areas of improvement and opportunities for growth: Enhance Low Growth: Explore the reasons for low growth rates in accounts with growth
<10% (if any exist beyond the very upper edge of the distribution); have a plan to enhance. Enhance High Growth: Explore the factors driving outcomes in the 22-24% growth group and formulate plans to enhance and replicate in other accounts.



Key Observations:

1. Right-Skewed Distribution: The distribution is significantly right-skewed (positively skewed). This is a common pattern with count data where most users use few or no coupons while a smaller number of users use many.

2. High Frequency at Zero: A very large peak at 0 identifies a large number of users did not use to coupons in the specified time period.

3. Rapid Decline: The frequency of users decreases rapidly as the number of coupons used increases with the bars getting progressively shorter as users move to the right.4. Smaller Peaks or Bumps: There are small bumps or local peaks at 1, 2, and maybe around

4 suggesting some users prefer to use coupons in small systematic batches.

5. Long Tail: The distribution has a long tail extending to the right indicating a small number of users used a relatively large number of coupons (up to 16 per the data shared)

Observations:

• Coupon Usage: The distribution indicates that most individuals are either not using any coupons or utilizing them minimally; there is a much smaller population that uses coupons frequently.

• Shopper's Attitude Towards Coupon Programs: The high number of users at the zero-coupon level suggests that either coupon programs are unknown to them, offer no incentive or others do not find coupons relevant to them.

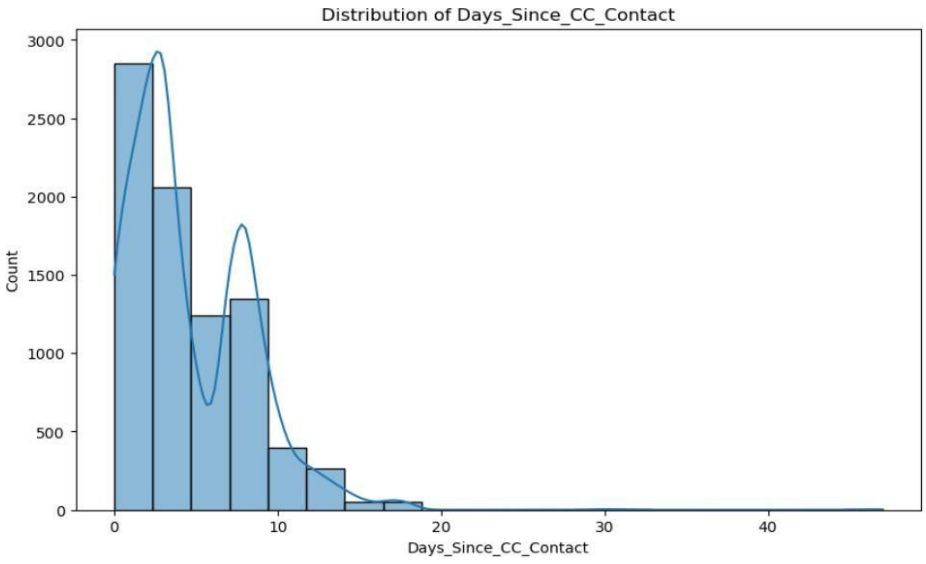
• Segmentation Potential: The coupon usage distribution indicates some potential customer segments:

Non-User (0 Coupons): A large segment of the population is not engaging with coupons.

Light Users (1-3 Coupons): Users who occasionally use coupons

Heavy Users (4+ Coupons): A smaller segment of the population engages with coupons frequently.

• Opportunity to Increase Engagement: With so many non-users, there is opportunity to increase coupon usage through marketing, programs or offers to attract them and increase usage.



Key Observations:

1. Right-Skewed Distribution: The distribution is strongly right-skewed (positively skewed), which indicates that most customers were contacted relatively recently.

2. Peak Around 0-10 Days: There is a clear peak around the 0–10-day range, showing that a significant number of customers were contacted within the last 10 days.

3. Long Tail: The distribution has a long tail for higher values, indicating that a smaller

number of customers were contacted longer ago.

4. Possible Data Cutoff: It is possible that there was a cutoff, or limit in capturing the contact dates back in time, which would explain the sharp drop-off of frequency past a certain point in days.

Interpretations:

Recency of Contact: The distribution indicates that the business is in contact consistently with many of its customers, which is generally a good sign for relationship building and customer engagement.

Customer Segmentation: The distribution can be used to segment customers based on recency of contact:

1) Recently Contacted - customers who have been contacted in the last 10-15 days.

2) Less Recently Contacted - customers who were contacted 15-30 days ago.

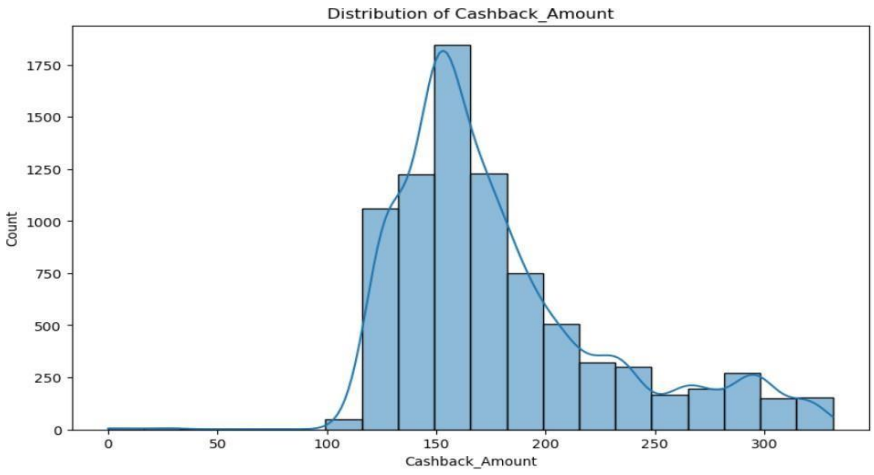
3) Long Time Since Contact - customers who have not been contacted in more than 30 days.

Customer Relationship Management:

1) Recency of contact shows the importance of maintaining contact with customers. Increasing contact with the "Less Recently Contacted" customers might be a good strategy to enhance frequency of contact with customers.

2) Identify Inactive Customers - Customers who might be consider inactive, if they have a very high "Days_Since_CC_Contact" value, and they may require targeted re-engagement.

3) Optimizing Frequency of Contact - The distribution can inform frequency of contact for customer segments. Over-communication with customers may not be beneficial, while under-communication may lead to disengagement.



Key Observations:

1. Strong Right-Skewed Distribution: The data is strongly right-skewed (positively skewed). Most cashback amounts are at the lower end of the spectrum.

2. High Peak at Nearly Zero: There is a very pronounced, high peak around nearly zero, meaning most cashback amounts are small, most likely under 250 (~ based on the axis of the x-axis).

3. Long, Thin Tails: The distribution has a long, thin tail out to the right (close to 2000). The tail signifies a small number of users receive significantly larger cashback payments.

4. Likely Outliers: Any of the data points far out on the right tail are likely outliers. These are all consumers that received extremely high cashback, a reward that is extraordinarily large.

Interpretations:

- Cashback Program Design: Given the pattern, the cashback program appears to be structured to reward most users with small amounts, and very occasionally, give out larger amounts.
- This could arise from: A tiered cashback system where the vast majority of customers are in the lower tier.
- Special offers or bonuses that reward a few customers with a much larger cashback amount

Different spending behaviors across customers:

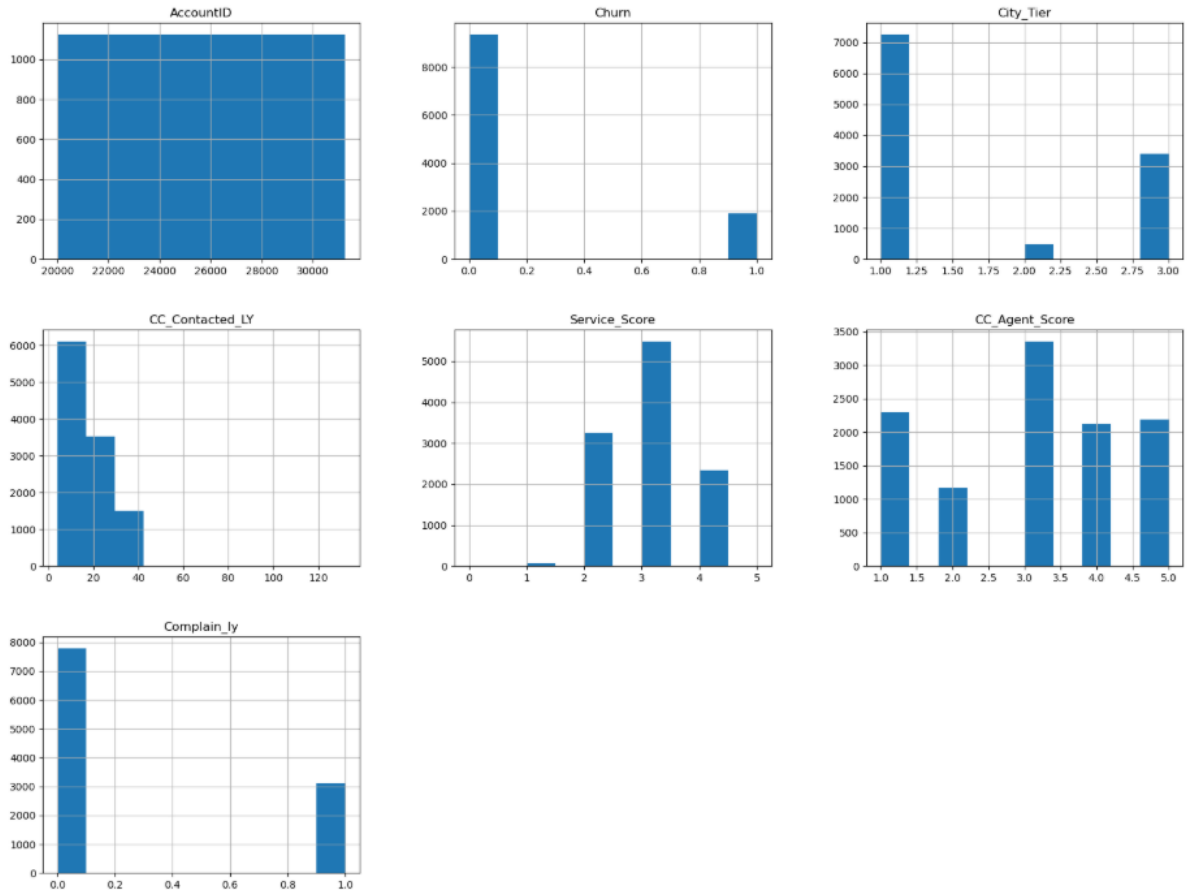
- Customer Segmentation: The distribution naturally segments customers into: Low-Cashback Earners: An overwhelming majority of users earn small cashback amounts.

High-Cashback Earners (Outliers): A small set of users earn significantly higher cashback amounts

- Marketing and Promotion Effectiveness: The distribution can also be used to observe how effective a cashback promotion is. In particular, if there was a promotion aimed at growing the amount of cashback amount given, you would know if this promotion was effective based on the distribution.

Categorical Variable Analysis:

```
df.hist(figsize=(20,15));
```



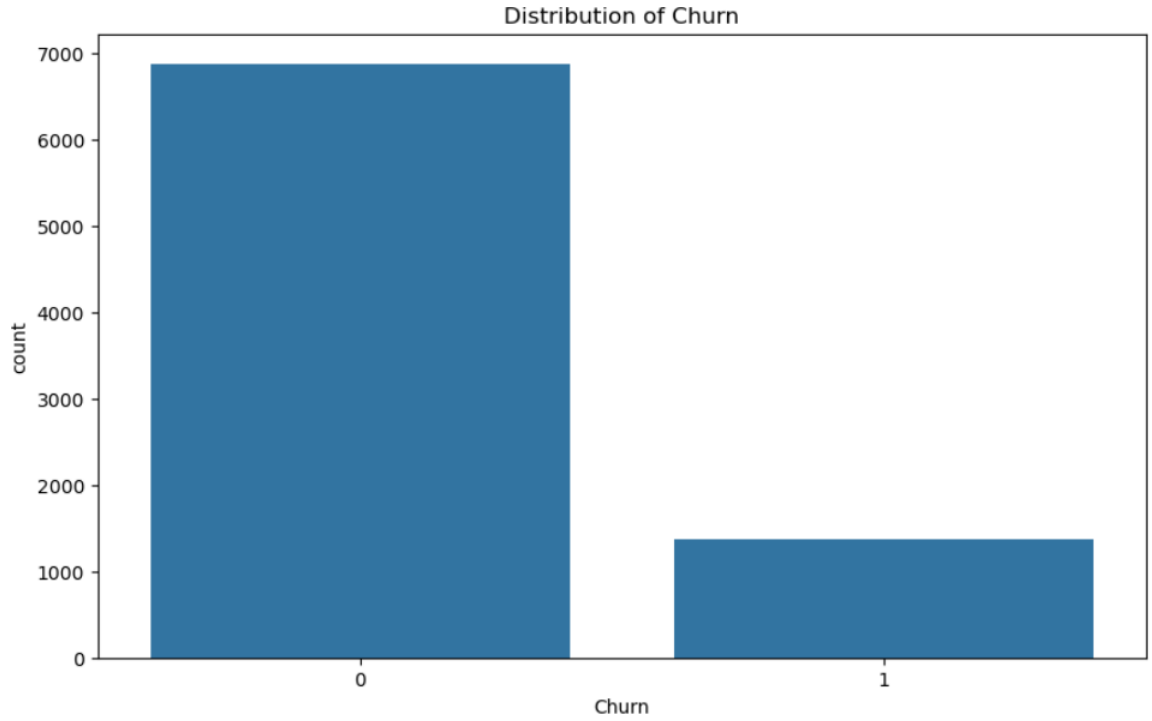
1. AccountID:
    - This looks uniformly distributed, indicating that account IDs are assigned sequentially or randomly.
2. Churn:
    - This is a binary column (0 or 1).
    - Most customers (majority of bars at 0) did not churn, while a smaller portion did (bar at 1).
3. City_Tier:
    - This appears to have three distinct values (1, 2, 3).

- Most customers belong to City Tier 1, followed by Tier 3, with Tier 2 being the least common.

4. CC_Contacted_LY (Customer Care Contacted Last Year):
    - This is right-skewed, meaning most customers contacted customer care fewer times, while a few reached out multiple times.
5. Service_Score:
    - Scores range from 1 to 5.
    - Most customers have service scores between 2 and 4, with very few at score 1 or 5.
6. CC_Agent_Score:
    - This shows ratings between 1 and 5.
    - The distribution is somewhat uniform, with peaks at different ratings.
7. Complain_LY (Complaint Last Year):
    - Another binary column (0 or 1).
    - Most customers did not file a complaint (bar at 0), but a smaller group did (bar at 1).

Key Observations:

- Some variables (Churn, Complain_LY) are categorical (binary).
- Some variables (City_Tier, Service_Score, CC_Agent_Score) are ordinal with distinct values.
- Right-skewed distributions in CC_Contacted_LY suggest that only a few customers contact frequently.



Key Observations:

1.

Two Classes (0 and 1):

0 signifies customers who did not churn (stayed with the company).

1 signifies customers who churned (left the company).

2.

Imbalance in Distribution:

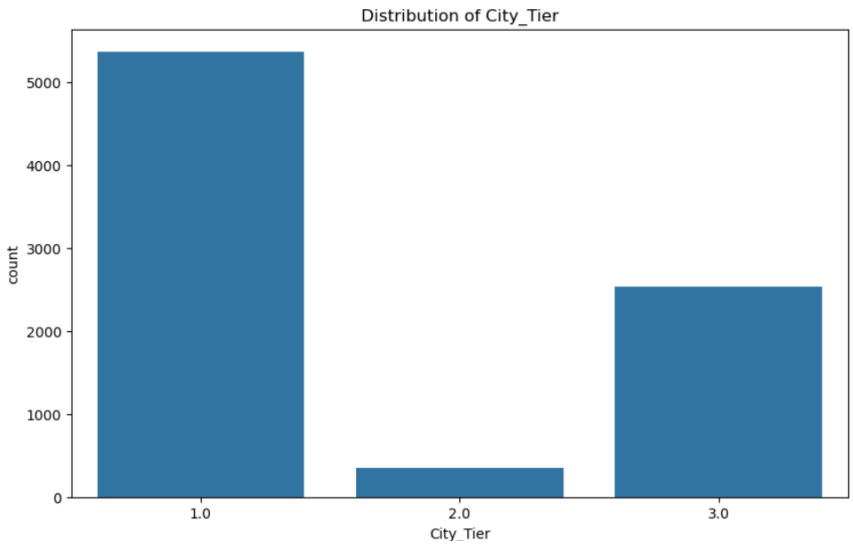
The number for customers who did not churn (0) is much greater than the number for those who churned (1).

This shows that the majority of customers in the dataset are retained, and a very small number have churned.

Implications:

- This skewness is typical with churn datasets; nonetheless, it can complicate the predictive modelling process. If machine learning algorithms become biased in their prediction, they will likely predict the majority class for the target variable (non-churners).

- There may be a need to implement techniques like oversampling (e.g., SMOTE), under sampling, or adjusting class weights, which will balance the dataset to make the model more effective.



Key Observations:

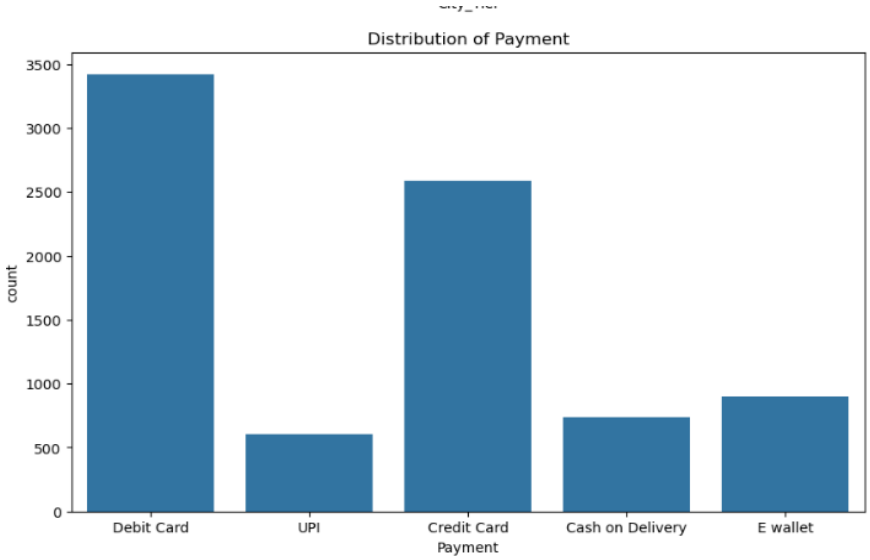
1.City Tier Classifications: The x-axis classifies the three tiers, namely, 1.0, 2.0, and 3.0. 1.0: Likely denotes major. 2.0: Represents smaller cities, or mid-tier urban. 3.0: Larger small towns or rural.

2. Distribution: The bulk of the customers in the dataset are Tier 1.0 representing the greatest count. Next, is Tier 3.0 indicating a smaller customer base from smaller towns or rural. Tier 2.0 has the lowest representation signalling fewer customers from mid-tier cities.

Implications:

- The sample is biased toward Tier 1 geography. This implies that the company caters to urban customers or has a higher volume of customers in urban locations.

- Customer behaviour may require tier-related implications for study, as different tiers of customers may have divergent spending habits, churn rate, or service needs.



Key Observations:

1. Payment Methods That Are Most Commonly Used: The payment methods "Debit Card" and "Credit Card" are the most commonly used payment methods, while "Debit Card" is marginally more favoured than "Credit Card".

2. Moderate Usage: The payment methods "E wallet" and "Cash on Delivery" are moderately used, with "E wallet" used marginally more than "Cash on Delivery".

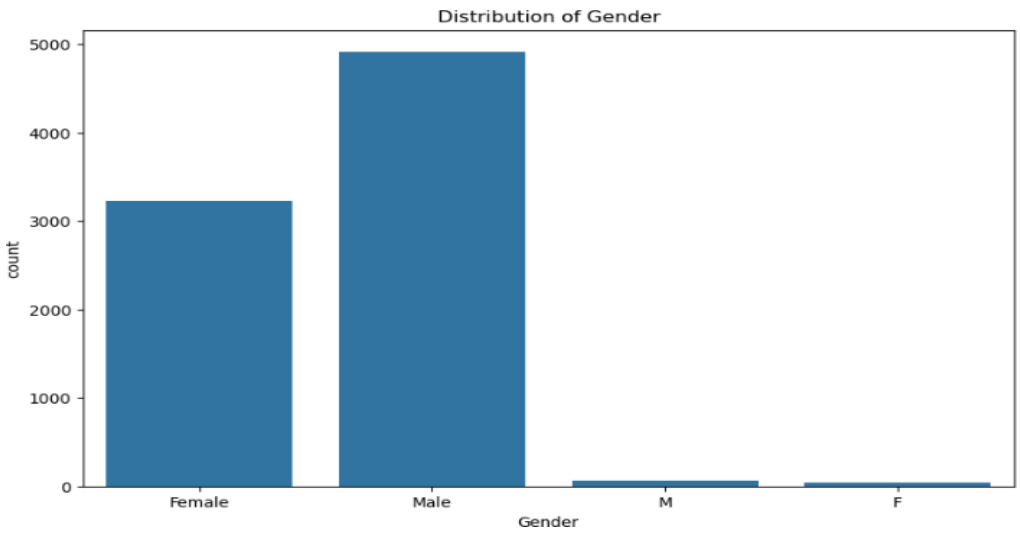
3. Payment Methods That Are Least Commonly Used: There is a significantly low usage of "UPI" as a payment method compared to all other payment methods.

Implications:

• Preferences of Customers: The distribution indicates customer preferences regarding payment. It is clear that debit and credit cards are the preferred payment methods, which indicates reliance on the traditional banking system.

• Market Penetration of Payment Methods: The low amount of UPI used suggests that this payment method has a lower level of market penetration or adoption in its current customer predominately. It may be due to reason including, but not limited to: lack of awareness, technology barriers, or trust issues.

• Target Audience: The information could imply characteristics of the target audience. For example, a greater percentage of "Cash on Delivery" usage versus other payment options could suggest a customer base which isn't comfortable with online transactions, or only wants the security of paying for the item only once it arrives.



Key Observations:

• Prevalence of Male Gender: Most of the subjects in the dataset are designated as male (there are roughly twice as many male subjects as there are females).

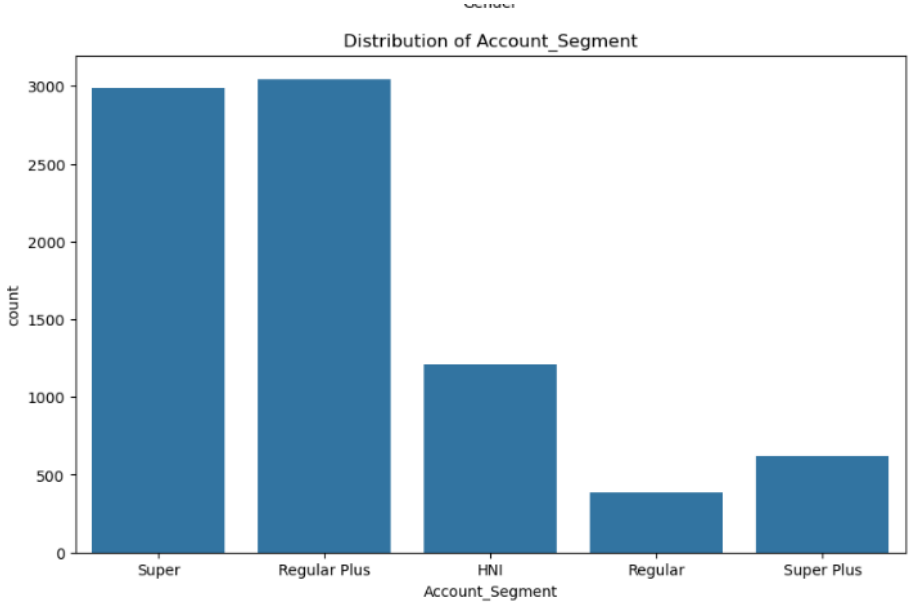
• Minority Categories: The categories "F" and "M" could either represent mislabelled or missing gender data. Less emphasis is warranted as there are only a few subjects in these categories when compared to the number of subjects in the Male and Female categories.

Insights:

• Data Cleaning: Look into the reason for the presence of the "F" and "M" categories. The nature of data entry, missing values, or coding issue can lead to this occurrence. As a step of clarity, consider standardizing the gender categories.

• Marketing and Development of Product: The predominance of males might indicate the product or service is more interesting to male customers; however, understanding and exploring the causation for their interest and strategies to recruit female customers is important.

• Diversity and Inclusion: The gender mixture might create concerns about diversity and inclusion actually existed; therefore, reason for the gender mixture is analysed indicating potential bias, etc.



Key Observations:

1. Leading Category: "Super" is the most commonly occurring account segment, and "Regular Plus" is the second most common account segment.

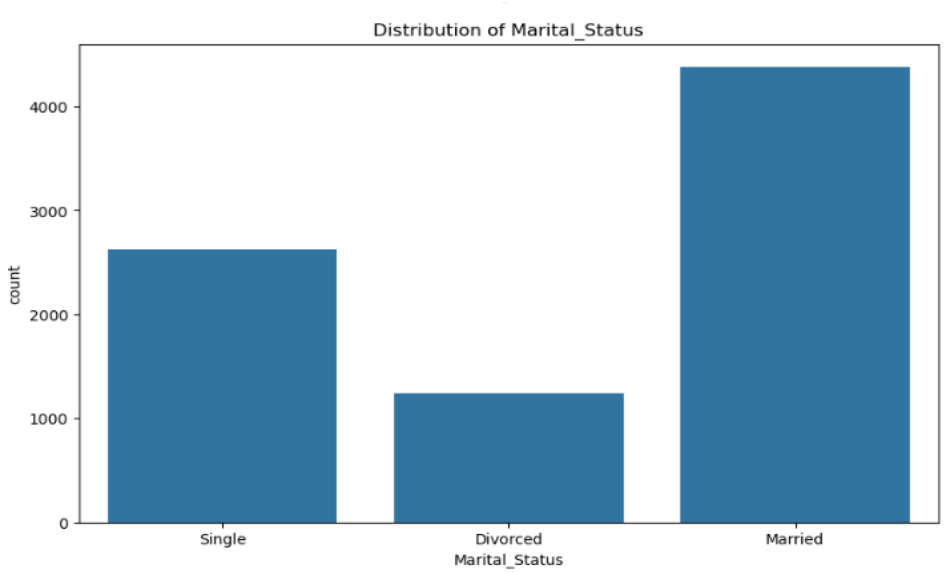
2. Less Common Segments: "Regular", "HNI", "Regular ", "Super Plus", and "Super " are fewer common segments. "Super "is the least common.

Interpretation:

- Customer Segmentation: The data shows the customer base is segmented into different tiers based on factors such as spending, loyalty, or other criteria.
- Marketing and Sales Strategies: The results would suggest that "Super" and "Regular Plus" segments dominate, which may indicate that the company focuses on acquiring and retaining customers in these segments, but does that imply we must now develop some acquisition and retention strategies for some of the other, less frequent, segments.

- Revenue and Profitability: Reviewing the revenue and profitability, across segments, could lead to insights about value in each segment and enabling better decision-making on how to allocate resources.

- Customer Lifetime Value: Knowing the customer lifetime value, across segments, may help inform acquisition and retention preferences and priorities.



Key Observations:

1. Majority Status: "Married" is the largest marital status and has a much higher count than other statuses.

2. Second Largest: "Single" is the second most frequent status, but with a much lower count than "Married" and still significantly more than "Divorced."

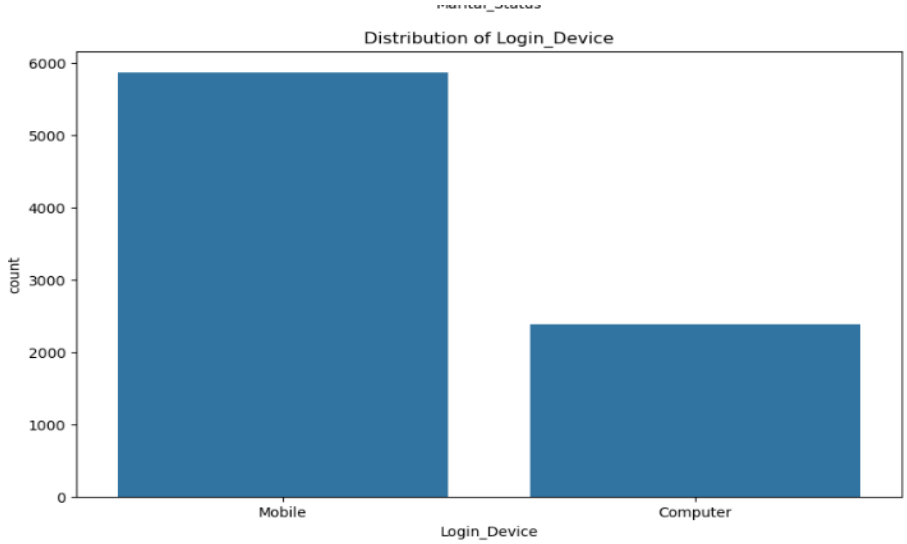
3. Minority Status: "Divorced" is the least frequent marital status in the data.

Insights:

• Demographics: The distribution highlights the marital status demographics in the dataset. It indicates a higher percentage of the population is married, followed by single individuals, and the lowest percentage are divorced individuals.

• Target audience: This information could be useful in the marketing and product development areas as an example: In the "Married" category, products or services that are family-or couple-focused may appeal to that segment. In the "Single" category, products or services surrounding individual needs or experiences may be prominent here.

• Opportunity for [targeted campaigns]: Marketing campaigns could address specific needs and focuses within the marital status categories.



Key Observations:

1. Predominant Device: Mobile is the overwhelmingly predominant login device, with a frequency far greater than the other device types.

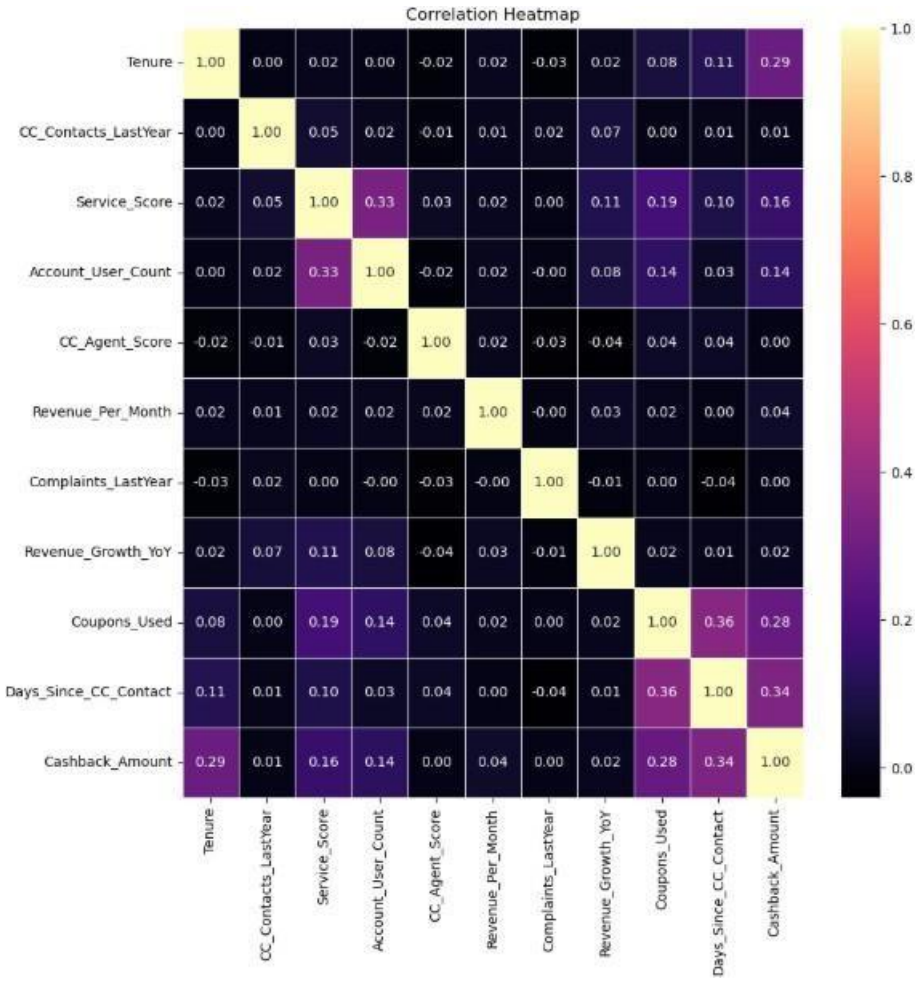
2. Secondary Device: Computer is the second-most-frequent login device, though with an overall frequency that is significantly lower than Mobile.

Explanations:

• Mobile-First User-Base: The significant majority of the usage of mobile logins suggests that the majority of people begin logged-in engagement with the platform/service on their mobile phone, which is not uncommon for most online services.

• Significant Mobile-Focused Optimizations: Businesses should consider optimizing their sites and mobile applications for mobile use at all touch points, so the majority of their users have a positive and effortless interaction.

Bivariate Analysis:



Key Observations:

1. **Strong Positive Relationships:**
   Service_Score and Account_User_Count: A strong positive relationship indicates that accounts with a higher service score have more users.
   - Revenue_Per_Month and Coupons_Used: A strong positive relationship suggests that accounts with higher revenue will use more coupons.
   - Days_Since_CC_Contact and Coupons_Used: A strong positive relationship suggests that accounts that were contacted recently would be more likely to utilize coupons.

2. **Intermediate Positive Correlations:**

Revenue_Per_Month and Account_User_Count: A moderate positive correlation indicates that accounts with more users tend to generate more revenue.

Revenue_Per_Month and Service_Score:

A moderate positive correlation indicates that accounts with higher service scores tend to generate more revenue.

3. **Weak or no correlation:**

Most of the off-diagonal values are light blue or white, indicating weak or zero correlation between the variables. This means that many variables do not correlate strongly with each other.

Possible Implications and Actions:

• Customer Segmentation: The correlation matrix has the potential to identify sets of variables that are highly correlated to one another. These findings can be applied to segment customers who are similar to each other.

• Feature Selection: In machine learning and predictive modelling, the correlation matrix can be used to identify which variable is highly correlated with the outcome variable. Identifying these correlations amongst the variables can be a useful exercise in feature selection and model building.

• Understanding Relationships: The correlation matrix can allow insight the relationships of the four variables assesses. Gaining insight to the relationships of the variables can help understand how the variables interact with one another and influence the larger system.

Elimination of unwanted variables:

Why eliminate account ID?

1. Non-predictive nature: Account IDs are unique for each customer / account and do not have any inherent relationship with the target variable (e.g. churn). Adding variables of this sort can trick models into identifying false patterns.
2. High cardinality: An account ID typically has a one-to-one relationship with the rows in the dataset meaning every value is unique. High cardinality variables, like the account ID, do not add predictive value.
3. Avoid data leakage: If the account ID does leak some information about a specific individual row in the dataset, it could lead to overfitting. Models could memorise patterns unique to an account ID, instead of forming generalisable trends or patterns.
4. Cleaner dataset: By reducing irrelevant variables, the dataset is cleaner and only useful features are available for analysis.

Handling Missing Value:

The dataset underwent examination for missing value(s) through these three steps:

1. Initial Examination: The dataset was visually checked for NaN, empty, or null value(s) in the rows and columns.

2. Statistics Summary: On Python, the .info() and .isnull().sum() methods were utilized to understand if missing value(s) existed and to assess its distribution across all columns.

3. In-depth Examination: Any columns with a considerable amount of NaN (if any) were flagged for potential imputation or elimination. In addition, specific attention was paid to the identified main attributes which could impact the churn prediction model.

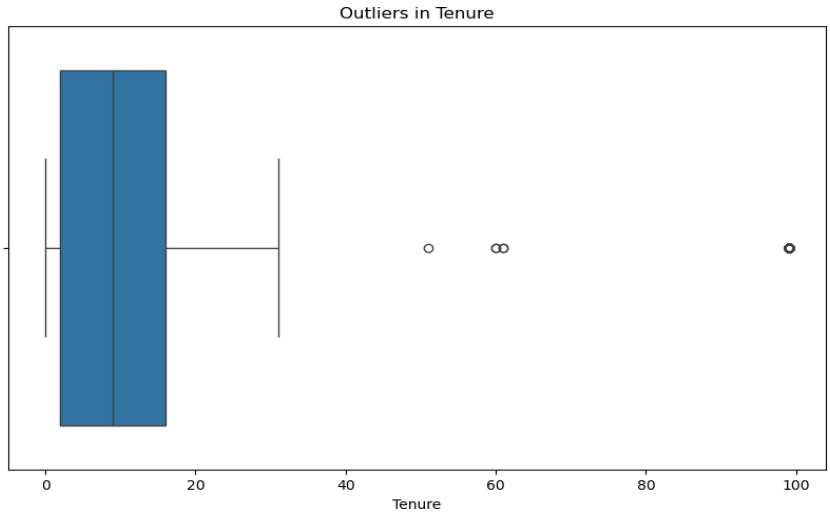
```
Churn                   0
Tenure                  0
City_Tier               0
CC_Contacts_LastYear    0
Payment                 0
Gender                  0
Service_Score           0
Account_User_Count      0
Account_Segment         0
CC_Agent_Score          0
Marital_Status          0
Revenue_Per_Month       0
Complaints_LastYear     0
Revenue_Growth_YoY      0
Coupons_Used            0
Days_Since_CC_Contact   0
Cashback_Amount         0
Login_Device            0
dtype: int64
Churn                   0
Tenure                  0
City_Tier               0
CC_Contacts_LastYear    0
Payment                 0
Gender                  0
Service_Score           0
Account_User_Count      0
Account_Segment         0
CC_Agent_Score          0
Marital_Status          0
Revenue_Per_Month       0
Complaints_LastYear     0
Revenue_Growth_YoY      0
Coupons_Used            0
Days_Since_CC_Contact   0
Cashback_Amount         0
Login_Device            0
```

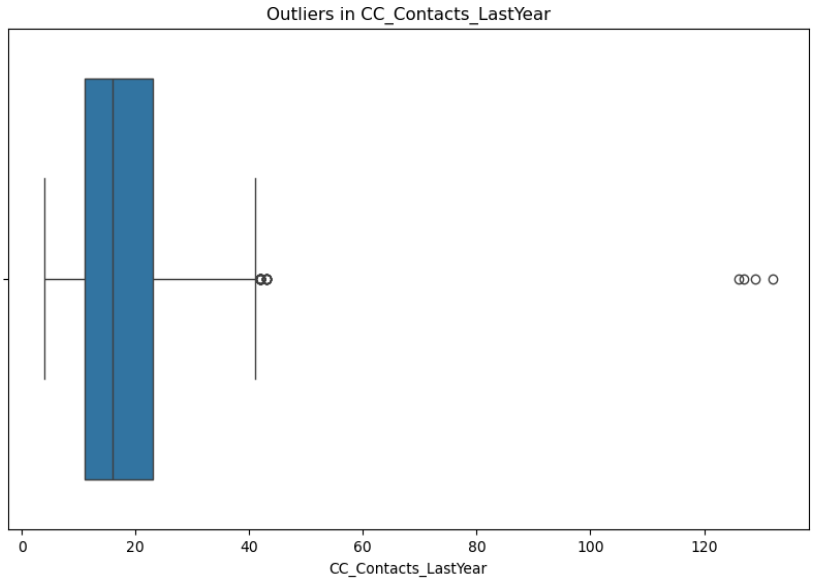
Management of Outliers Methods Utilized for Outlier Detection:

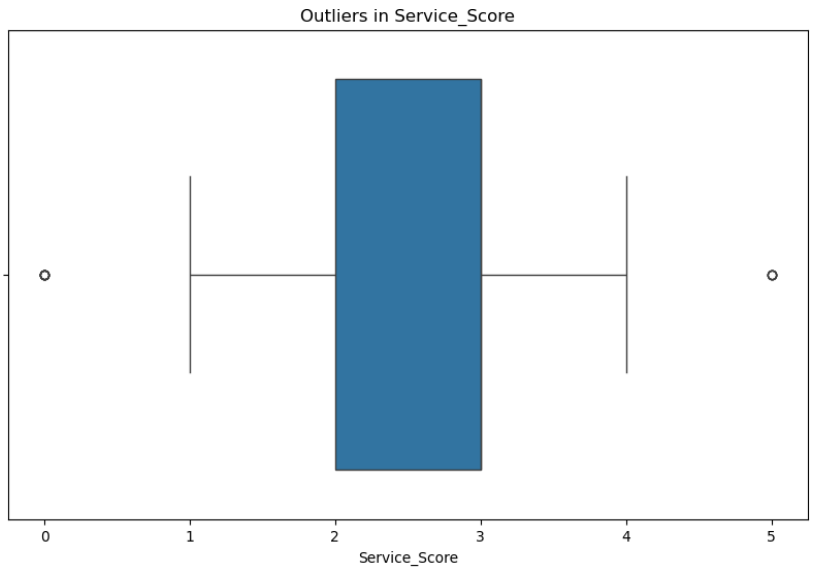
1. Visual Assessment: The distribution of numerical variables was examined visually using boxplots and histograms, and potential outliers were identified.

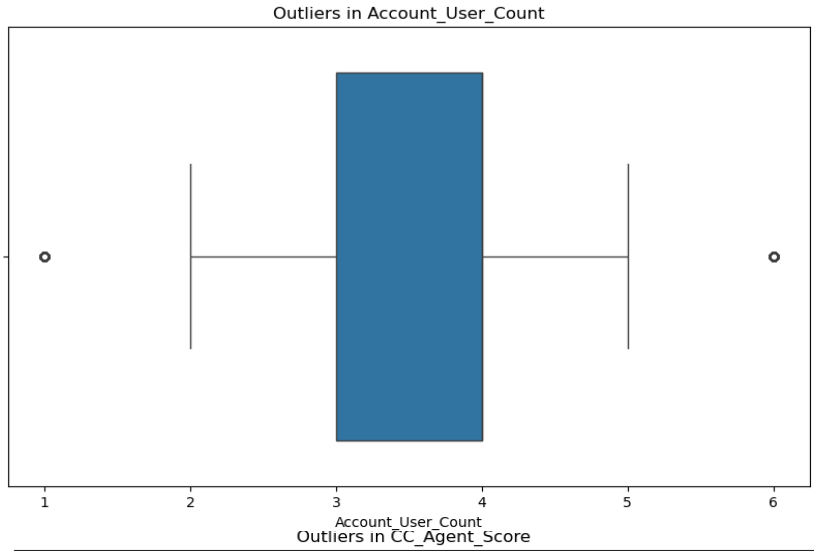
2. Statistical Testing: The Interquartile Range (IQR) was used to detect outliers: values below Q1 - 1.5IQR or above Q3 + 1.5IQR were flagged as potential outliers. Z-scores calculated to identify extreme values (e.g., absolute Z-score >

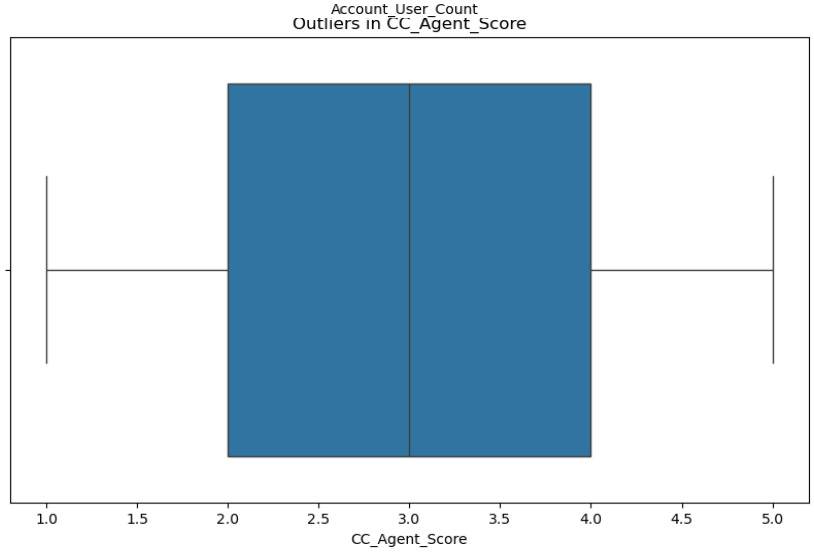
3. were also calculated.3. Business Context Testing: The detected outliers were subsequently examined and determined to be valid extreme cases (e.g., accounts with a very high tenure or revenue) or errors in the data.

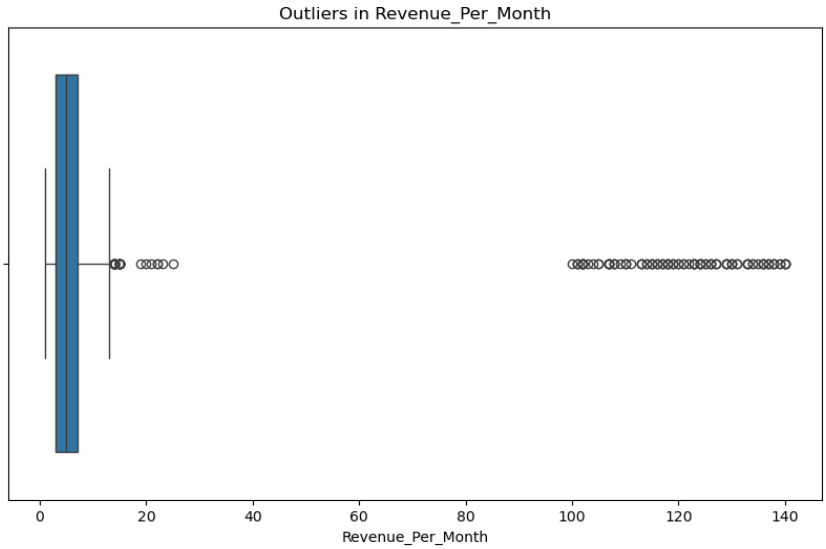

Outliers in Tenure
0
20
40
60
80
100
Tenure

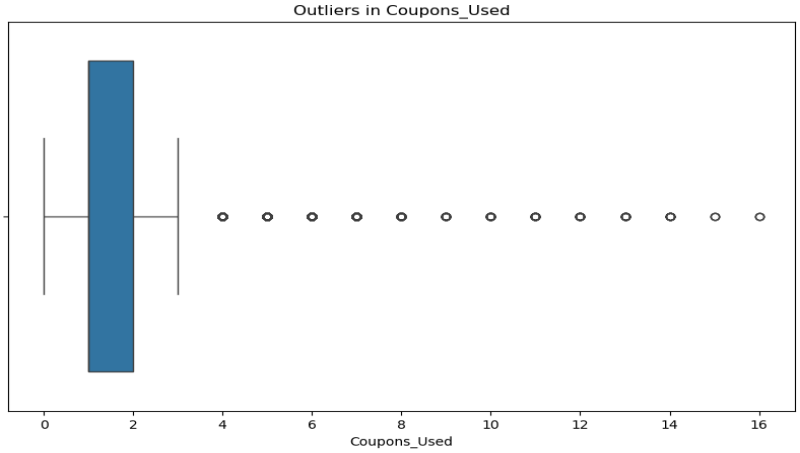

Outliers in CC_Contacts_LastYear
0
20
40
60
80
100
120
CC_Contacts_LastYear

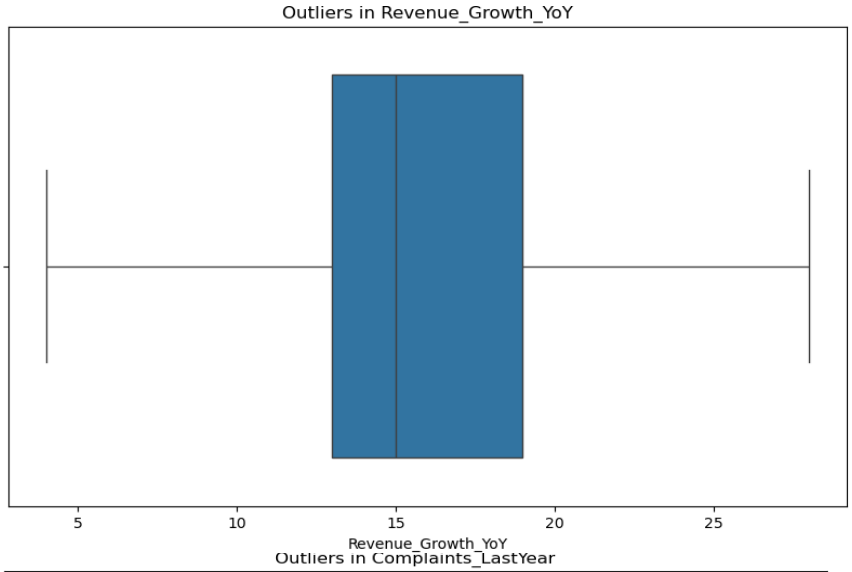

Outliers in Service_Score
0
1
2
3
4
5
Service_Score

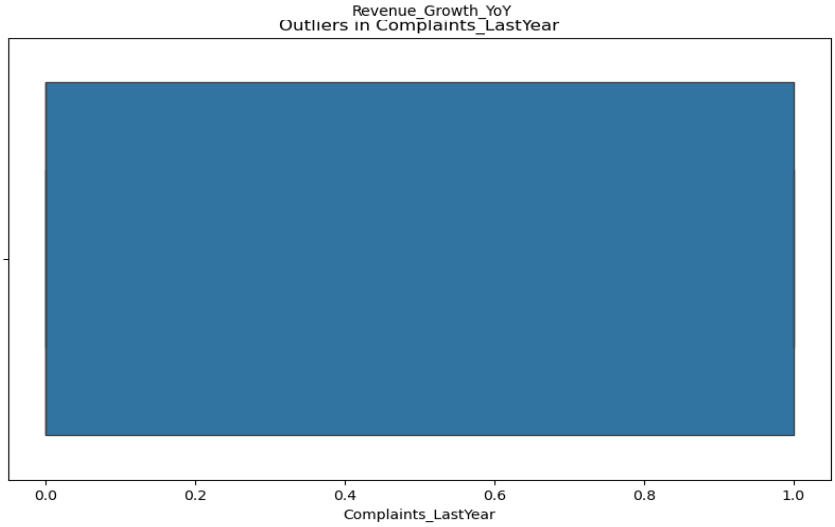

Outliers in Account_User_Count
1
2
3
4
5
6
Account_User_Count

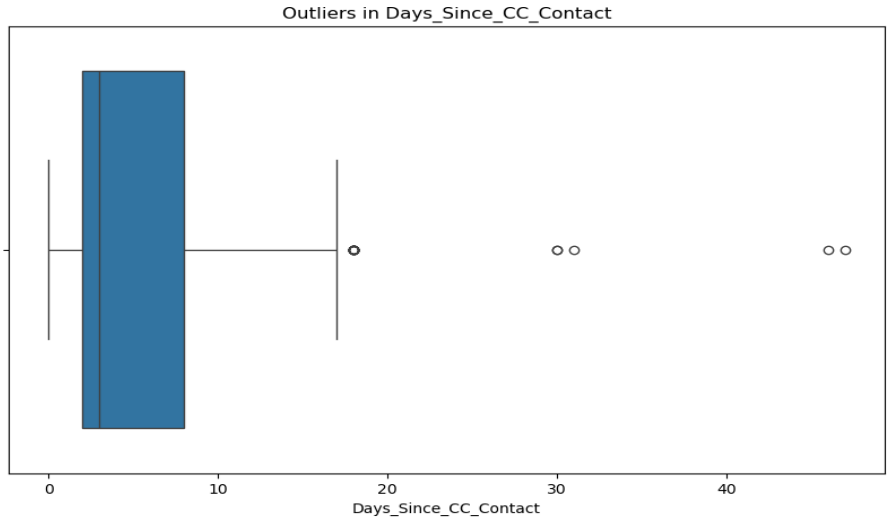

Outliers in CC_Agent_Score
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
CC_Agent_Score

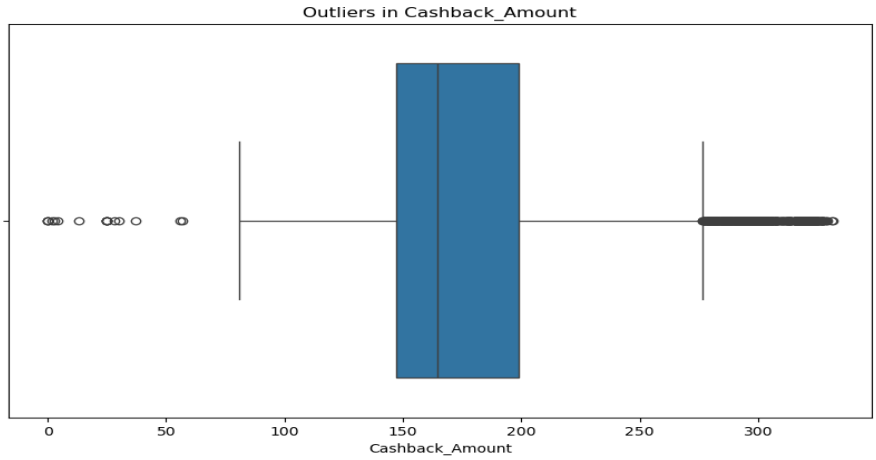

Outliers in Revenue_Per_Month
0
20
40
60
80
100
120
140
Revenue_Per_Month


Outliers in Coupons_Used
0
2
4
6
8
10
12
14
16
Coupons_Used

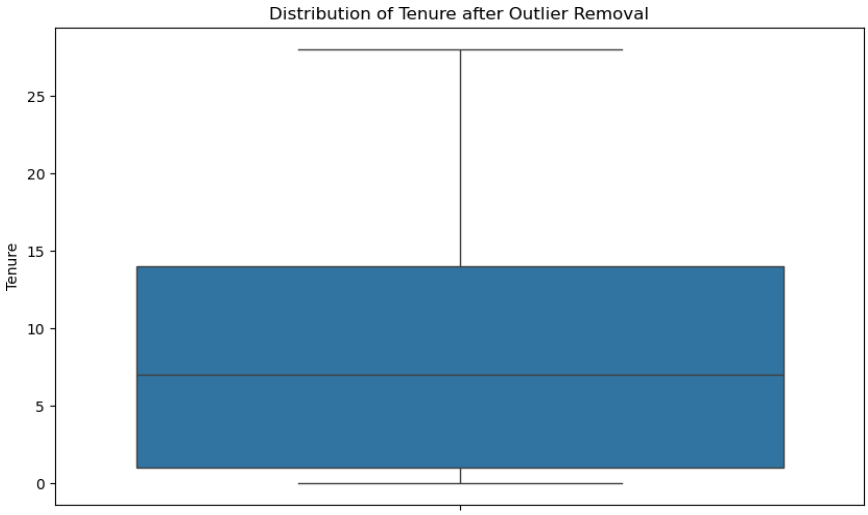

Outliers in Revenue_Growth_YoY
5
10
15
20
25
Revenue_Growth_YoY

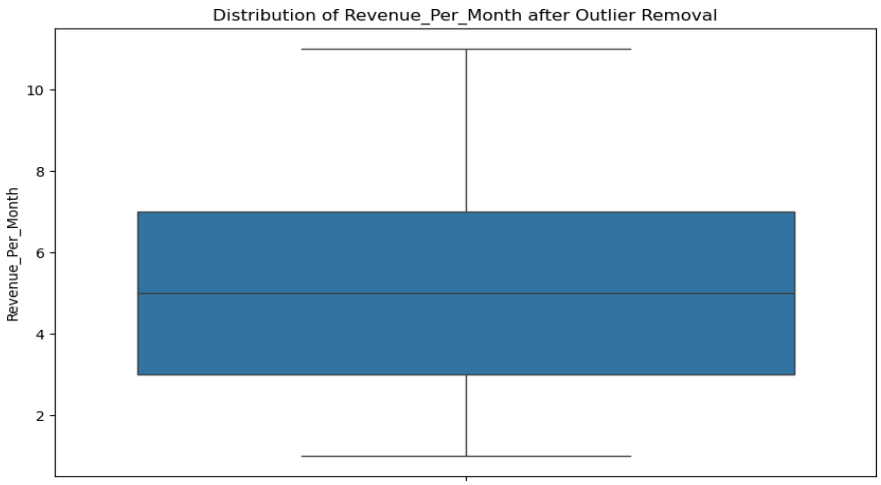

Outliers in Complaints_LastYear
0.0
0.2
0.4
0.6
0.8
1.0
Complaints_LastYear

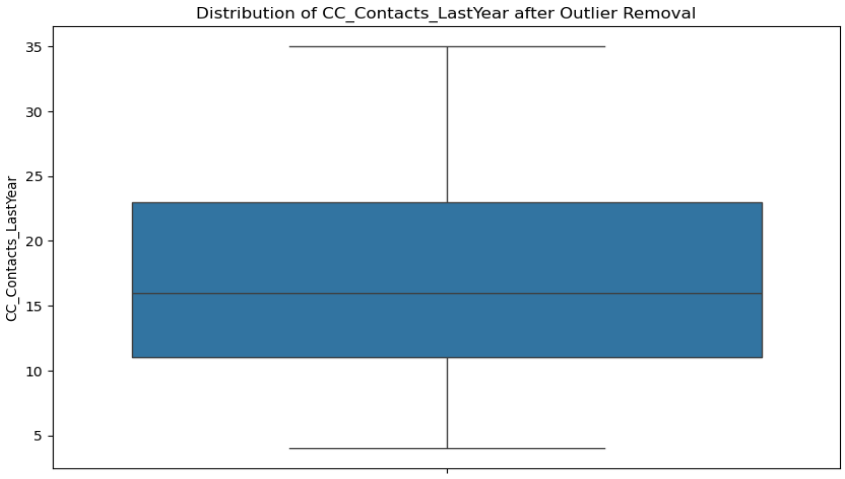


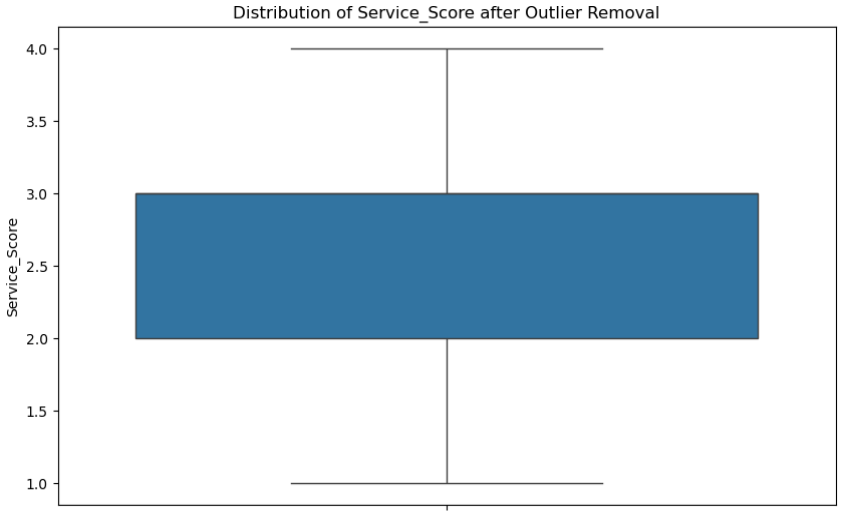


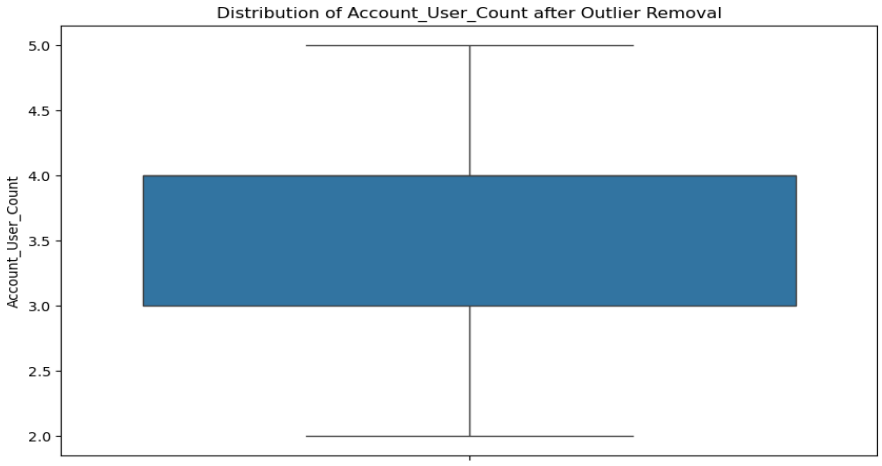
After Outlier Treatment:

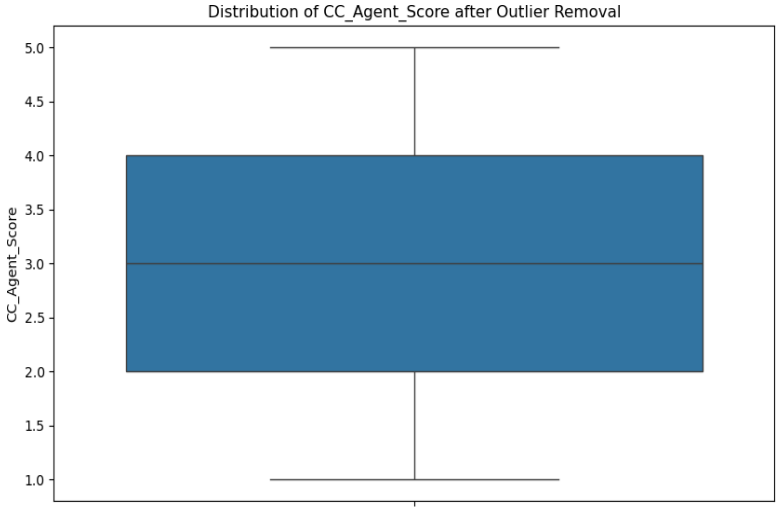


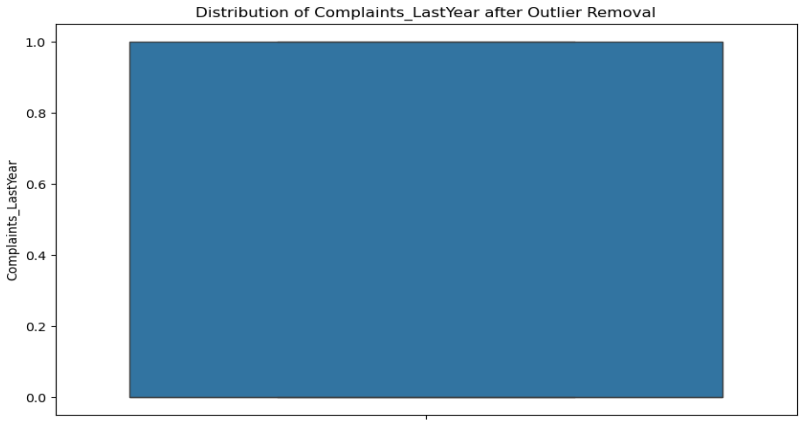

Distribution of Revenue_Per_Month after Outlier Removal
Revenue_Per_Month
10
8
6
4
2

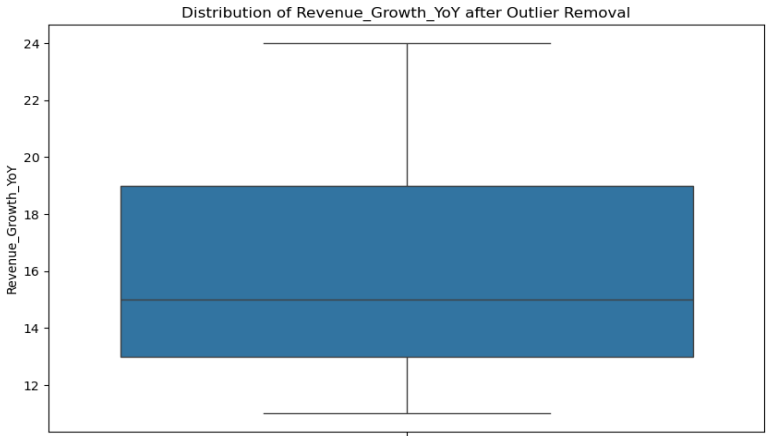

Distribution of CC_Contacts_LastYear after Outlier Removal
CC_Contacts_LastYear
35
30
25
20
15
10
5

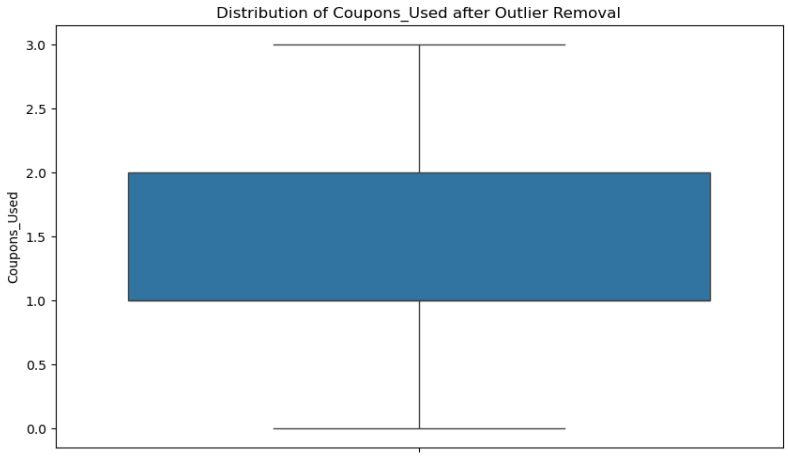

Distribution of Service_Score after Outlier Removal
Service_Score
4.0
3.5
3.0
2.5
2.0
1.5
1.0

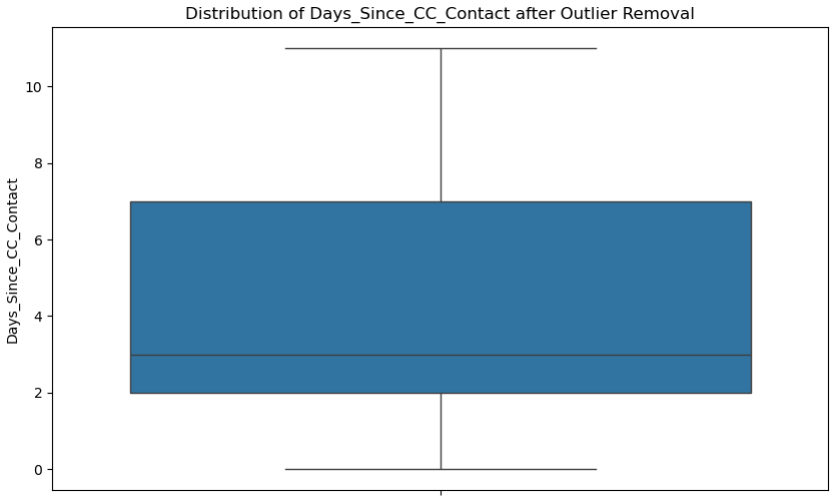

Distribution of Account_User_Count after Outlier Removal
Account_User_Count
5.0
4.5
4.0
3.5
3.0
2.5
2.0


Distribution of CC_Agent_Score after Outlier Removal
CC_Agent_Score
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0


Distribution of Complaints_LastYear after Outlier Removal
Complaints_LastYear
1.0
0.8
0.6
0.4
0.2
0.0







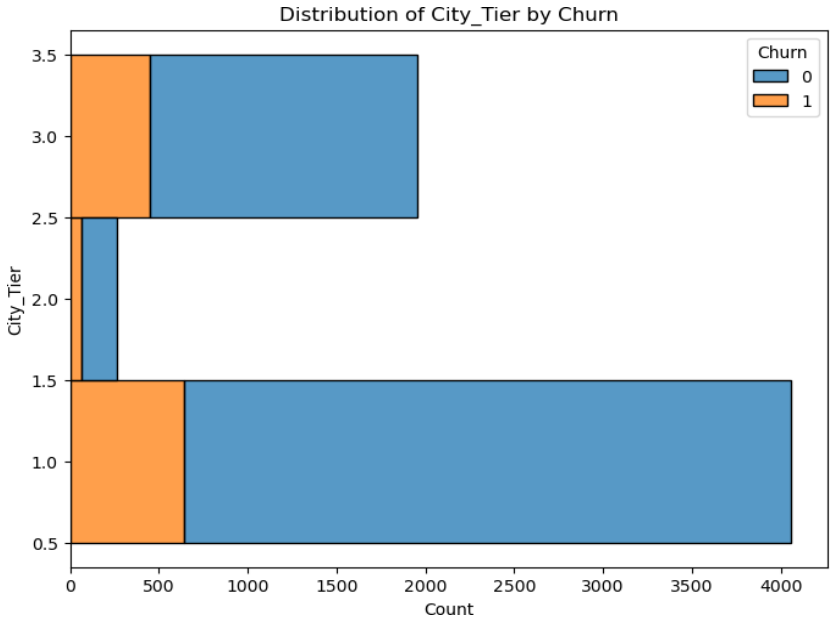
Outcome:

The analysis concluded that while some variables contained outliers, these could be justified through the business context (eg.: customer accounts can have extreme revenue

or tenure). Because of this, no outlier removal or transformation occurred as to preserve the legitimate data, as well to preserve a representation of all customer segments.

Significance:

Careful outlier handling ensures the model captures and reflects the full variability of the data, without unexpectedly dropping potential rationale behaviour data. This is especially important with churn analysis as extreme values could represent meaningful behavioural data.



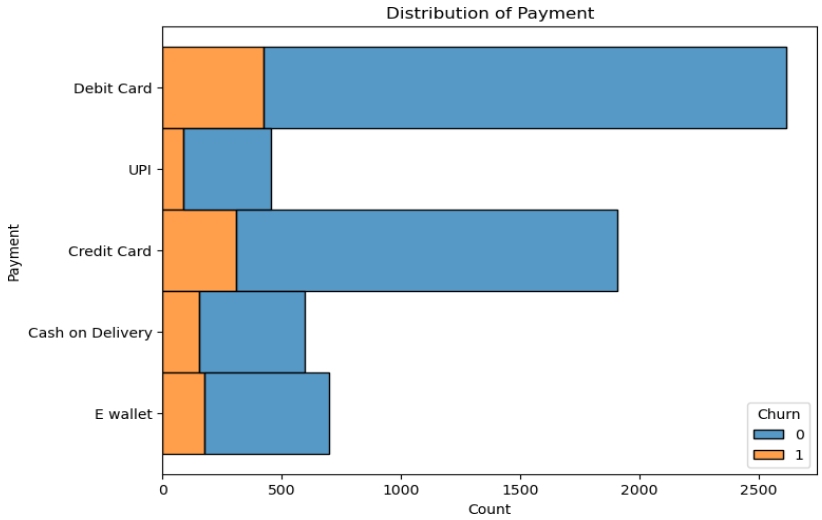
Observations:

1. **City_Tier 1** has the highest number of customers.
    - The majority did not churn (large blue section).
    - A smaller portion churned (orange section).
2. **City_Tier 2** has the fewest customers.
    - Both churned and non-churned customers are relatively low in number.
3. **City_Tier 3** has a significant number of customers.
    - A large number of non-churned customers (blue).
    - Some churned customers (orange), but the churn percentage seems higher compared to Tier 1.

Insights:

- Churn seems to be occurring across all tiers but is more noticeable in **City_Tier 1 and 3**.

- The proportion of churn in each tier can be further analysed to identify risk factors related to location.

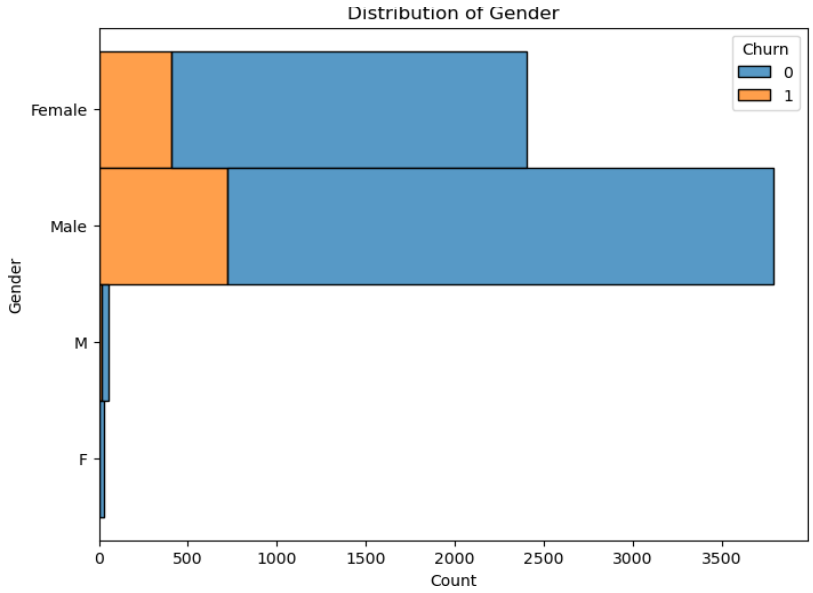


Observations:

1. **Debit Card is the most popular payment method** with the highest number of users.
    - A significant majority of customers using debit cards did not churn.
    - Some churned customers are present but in a smaller proportion.
2. **Credit Card is the second most used payment method.**
    - Most credit card users did not churn.
    - A small proportion of churned customers exist.
3. **UPI has the least number of users** among all payment methods.
    - A smaller segment of churned users compared to other methods.
    - Fewer users overall suggest it may not be a preferred mode of payment.
4. **Cash on Delivery has moderate usage.**
    - A relatively balanced proportion of churned and non-churned users.
    - This could indicate that customers who use this method may have a higher churn tendency.
5. **E-wallet usage is relatively low.**
    - Some churned customers are present but in small numbers.

Insights:

- Customers using **debit cards and credit cards** are the most engaged.
- **Churn is relatively higher in Cash on Delivery and UPI users** compared to others.
- **E-wallet and UPI have lower customer bases**, which could be due to customer preference or availability.

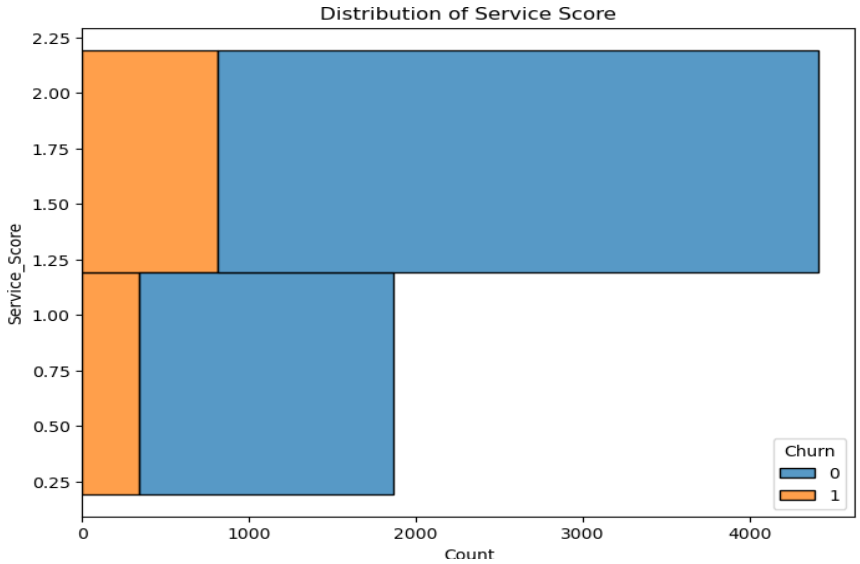


Observations:

1. **Male and Female categories are the dominant labels.**
    - The **Male** category has the highest number of customers.
    - A significant portion of them did not churn (blue section), but some did (orange section).
    - The **Female** category also has a high number of customers, with a similar churn proportion.
2. **Presence of "M" and "F" categories indicates inconsistent data labelling.**
    - "M" and "F" appear to be alternative representations of "Male" and "Female," possibly due to data entry inconsistencies.
    - Their counts are much lower compared to "Male" and "Female."

Insights:

- **Gender does not seem to be a strong predictor of churn**, as the churn proportions are relatively similar for both Male and Female customers.
- **Data cleaning is required** to merge "M" with "Male" and "F" with "Female" .

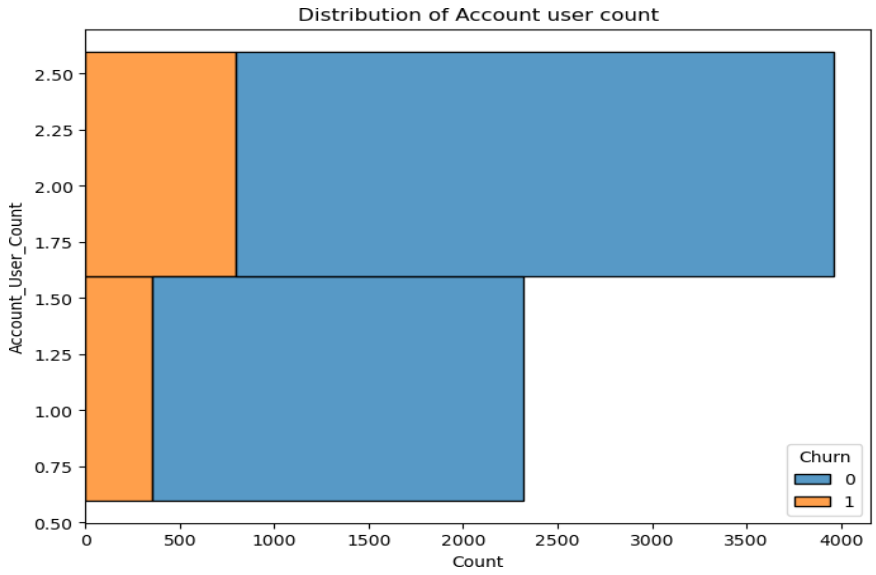


Observations:

1. **Most customers have service scores of 1 or 2.**
    - A large number of customers have a **service score of 2**, and the majority of them **did not churn**.
    - Many customers have a **service score of 1**, but some of them **churned** (orange section).
2. **Churn is higher among customers with a lower service score.**
    - The proportion of churned customers (orange) is more noticeable in-service **score 1**, indicating dissatisfaction.
    - Customers with a **service score of 2** have a much lower churn rate.

Insights:

- **Higher service scores seem to correlate with customer retention**, as seen with the dominance of blue bars.
- **Customers with a lower service score (1) are more likely to churn.**
- **Improving service quality for lower-rated customers may reduce churn.**



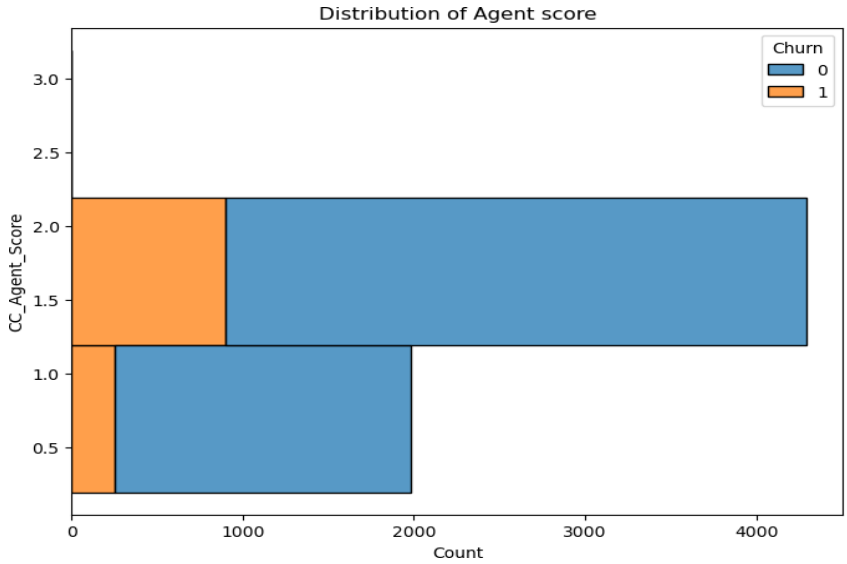
Observations:

1. **Most customers have an account user count of 2 or more.**
    - The majority of users belong to **accounts with a user count of 2.5 (likely rounded values) and 1.5.**
    - Most of these users **did not churn** (blue bars dominate).
2. **Churn is more noticeable among accounts with lower user counts.**
    - Accounts with **1 or fewer users** show a higher proportion of churn (orange section).
    - Accounts with **more users (2.5) have a lower churn rate**, suggesting more stability.

Insights:

- **Accounts with fewer users tend to have a higher churn rate**, possibly due to lower engagement or value perception.
- **Accounts with more users tend to stay longer,** which might indicate stronger retention due to multiple users being involved in the service.

- **Businesses could focus on increasing user engagement for single-user accounts** to reduce churn.

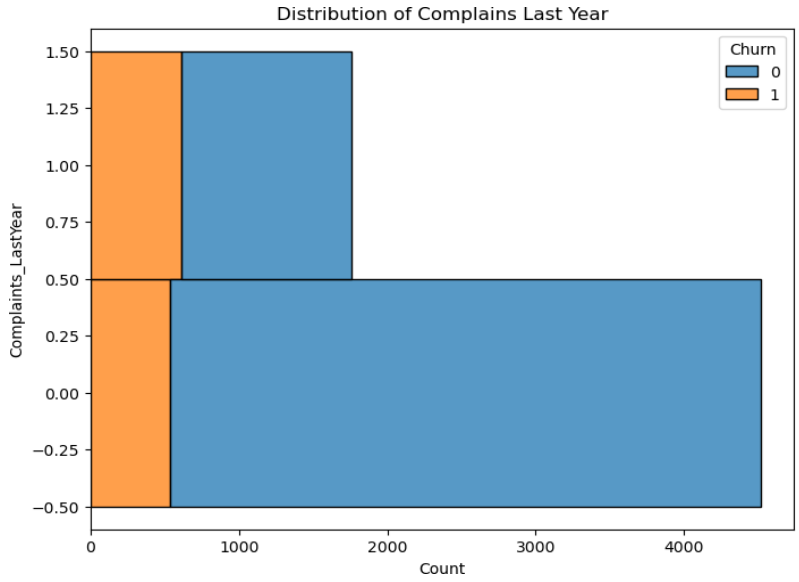


Observations:

1. **Customers with higher agent scores tend to stay longer.**
    - The **majority of customers fall under a score of 2.0 or higher**, and most of them **did not churn** (blue dominates).
    - This indicates that **higher agent scores are associated with better customer retention**.
2. **Lower agent scores correlate with higher churn rates.**
    - At lower agent scores (1.0 or below), **a significant portion of customers have churned** (visible orange section).
    - This suggests that **poor customer service interactions contribute to customer churn**.

Insights:

- **Improving agent performance** and training customer support teams **could help reduce churn**.
- **Monitoring customer feedback and agent scores regularly** can help identify areas were service quality needs improvement.
- **Customers who interact with high-scoring agents are more likely to stay**, suggesting that investing in quality customer service has a direct impact on retention.

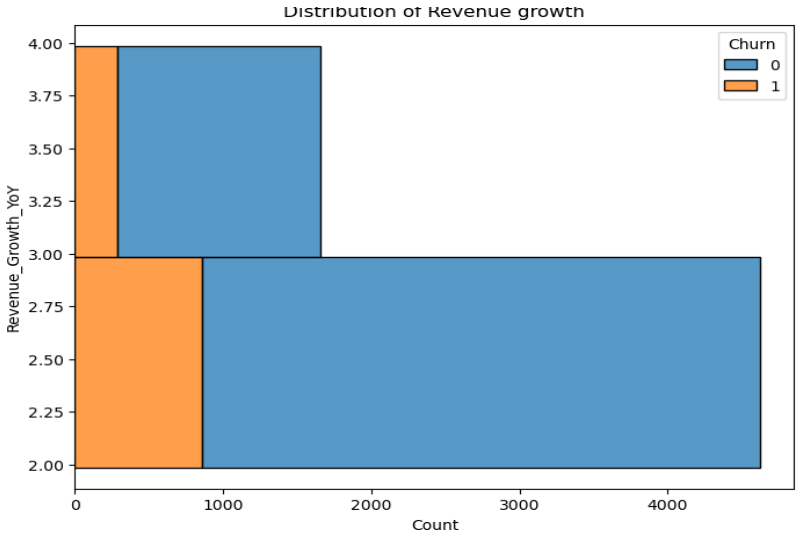


Observations:

1. **Most customers did not file complaints.**
    - The majority of customers are in the **zero-complaints category** (large blue section).
    - This suggests that many customers are satisfied or have not raised issues formally.
2. **Customers who filed complaints have a higher churn rate.**
    - In the **complaints category**, a significant portion of customers have churned (orange section).
    - This indicates that **customers who experience issues and file complaints are more likely to leave**.
3. **A considerable number of complaining customers stayed.**
    - Despite some complaints, many customers **did not churn** (blue in the complaints section).
    - This could suggest that **good issue resolution retains customers**.

Insights:

- **Improving customer complaint resolution processes** may help reduce churn.
- **Proactively addressing customer issues** before they turn into complaints can enhance satisfaction.
- **Tracking unresolved complaints and following up** could help retain at- risk customers.



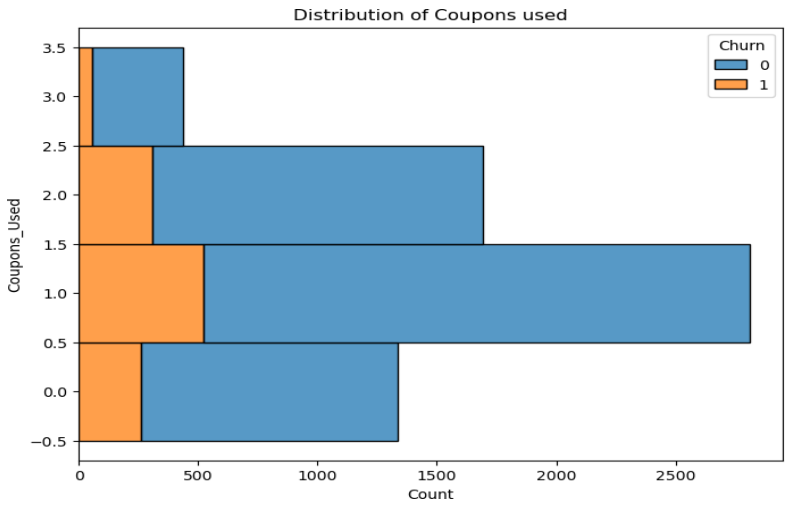
Observations:

1. **Most customers have positive revenue growth.**
    - The majority of customers have **higher revenue growth values** (large blue section).
    - This suggests that growing accounts tend to **retain their subscriptions** or services.
2. **Churned customers are more concentrated in lower revenue growth categories.**
    - Customers with **lower revenue growth** have a relatively higher proportion of churn (orange section).
    - This suggests a strong correlation between **declining revenue growth and higher churn rates**.
3. **Customers with strong revenue growth tend to stay.**
    - The largest blue section corresponds to higher revenue growth, indicating **low churn in this segment**.
    - This implies that **financially growing customers are more likely to stay with the company**.

Insights:

- **Revenue decline may be an early warning for churn.**
    - Companies should closely monitor customers with declining revenue growth and **engage them proactively** to prevent churn.

- **Offering value-added services or incentives** to stagnating customers may help improve retention.
- **Predicting churn based on revenue trends** can help in designing better customer success strategies.

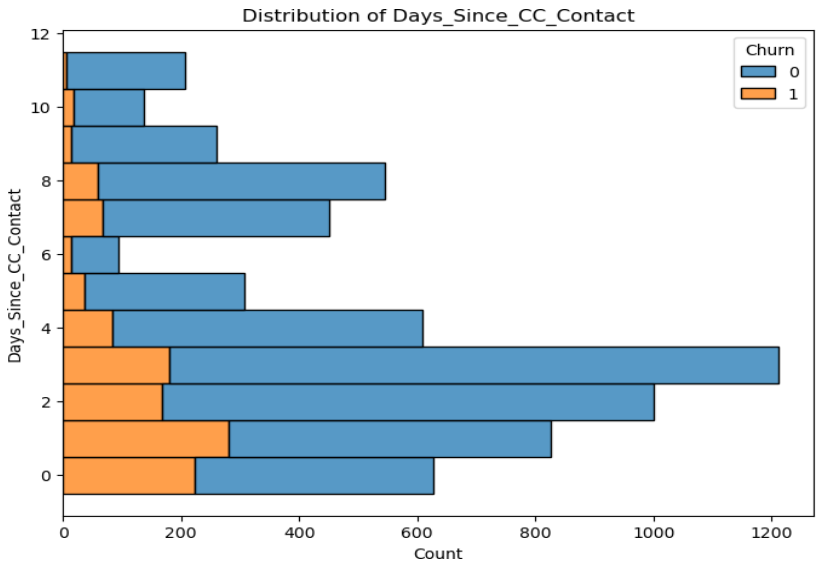


Observations:

1. **Customers who used more coupons tend to stay.**
    - The majority of the non-churned customers (blue) used multiple coupons.
    - This suggests that **higher coupon usage is correlated with customer retention**.
2. **Churned customers (orange) are more concentrated in lower coupon usage categories.**
    - Customers who used **fewer coupons** have a relatively higher proportion of churn.
    - This implies that **customers who do not engage with coupons are more likely to leave**.
3. **The highest coupon usage group has almost no churn.**
    - Customers who used a large number of coupons mostly stayed with the company.
    - This suggests that **discounts, rewards, and incentives play a role in customer loyalty**.

Insights & Recommendations:

- **Encourage customers to use coupons**: Providing targeted promotions or discounts to low-engagement customers may help reduce churn.
- **Monitor coupon usage as a retention signal**: Customers who **stop using coupons** may be at risk of churning.
- **Personalized discounts** for customers who haven't used coupons recently might encourage them to stay engaged with the company.



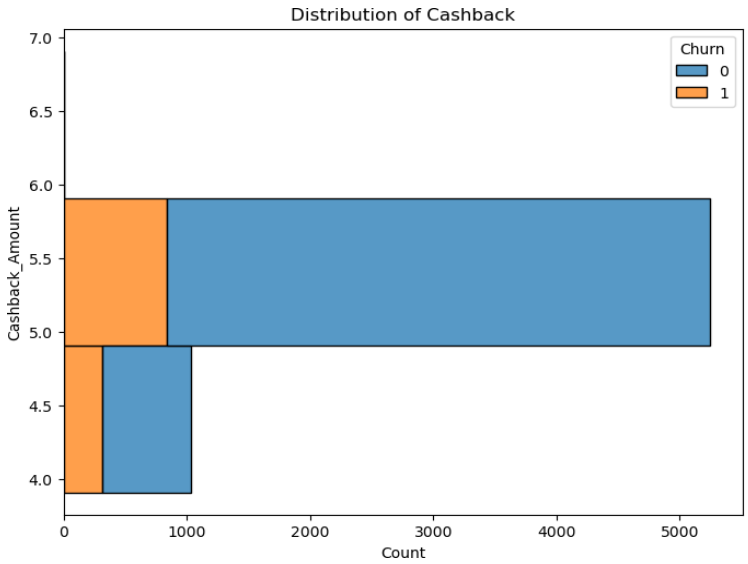
Observations:

1. **Customers with recent customer care interactions have a higher churn rate.**
    - The **orange (churned customers) portion is larger for lower values** of Days_Since_CC_Contact, meaning customers who contacted customer care recently were more likely to churn.
    - This could indicate **negative experiences with support**, dissatisfaction, or unresolved issues leading to churn.
2. **Customers who haven't contacted customer care in a long time tend to stay.**
    - The bars representing **higher days since last contact** have mostly **blue (non-churned) customers**.
    - This suggests that customers who don't need frequent support interactions are more stable and less likely to churn.
3. **The majority of customers did not churn.**

- Even in categories where some churn is present, **the blue section is dominant**, meaning most customers retained their service.

Insights & Recommendations:

- **Improve customer support experience:** Since many churned customers had recent customer care interactions, investigate **common complaints, response quality, and resolution efficiency**.
- **Follow up on negative interactions:** Customers who contacted support recently should be targeted with **proactive follow-ups** to ensure issue resolution.
- **Monitor support tickets as churn indicators:** A spike in support requests from a customer could signal **potential churn risk**.
- **Enhance self-service options:** If frequent customer care contact correlates with churn, improving **self-service tools** might reduce frustrations.



Observations from the Graph:

- **High Cashback (Around 5.9):** A very large number of customers received cashback around 5.9, and a relatively smaller number of these customers churned. This suggests that a higher cashback amount might be associated with lower churn.
- **Moderate Cashback (Around 4.8):** A noticeable number of customers received cashback around 4.8. Within this group, the number of customers who did not churn is higher than those who churned, but the proportion of churned customers appears higher compared to the high cashback group.
- **Lower Cashback (Around 3.9):** A smaller number of customers received cashback

around 3.9. Within this group, the number of churned customers is significant compared to the non-churned customers, and in fact, appears slightly higher. This suggests that lower cashback amounts might be associated with higher churn.

In summary, the graph suggests a potential relationship between cashback amount and customer churn. Customers receiving higher cashback amounts seem less likely to churn, while those receiving lower cashback amounts appear more likely to churn.

Model Building:

1. **Logistic Regression**

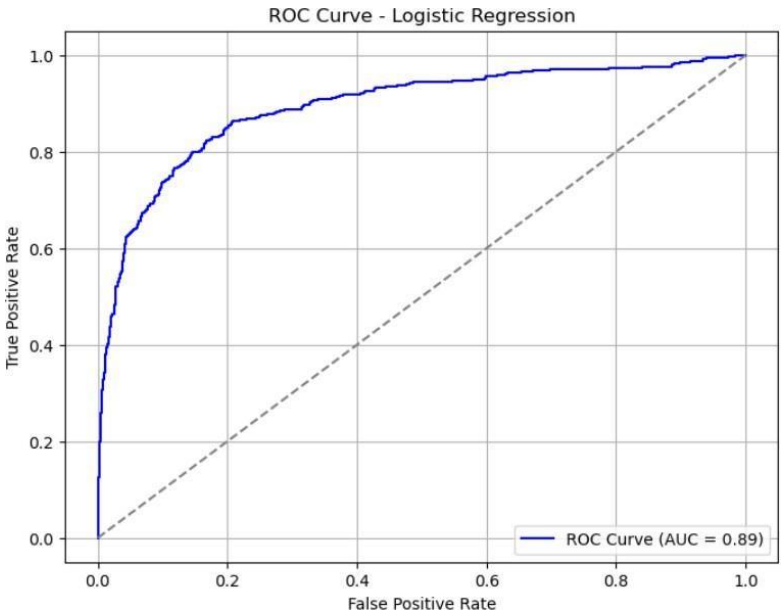
```
Logistic Regression:
Accuracy: 0.8957575757575758
Precision: 0.7758007117437722
Recall: 0.5278450363196125
F1-Score: 0.6282420749279538
ROC-AUC: 0.8924995831405602

Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.97      0.94      2062
           1       0.78      0.53      0.63       413

    accuracy                           0.90      2475
   macro avg       0.84      0.75      0.78      2475
weighted avg       0.89      0.90      0.89      2475

AUC Score: 0.8925
```



Overview: A Logistic Regression model has performed well in predicting binary outcomes and has demonstrated overall favourable performance in terms of accuracy and very strong class separation (ROC-AUC).

Key Metrics:

- **Accuracy:** 89.6%
- **Precision:** 77.6%
- **Recall:** 52.8%
- **F1-Score:** 62.8%
- **ROC-AUC Score:** 89.3%

Classification Breakdown:

Class	Precision	Recall	F1-Score	Support
0 (Negative Class)	91%	97%	94%	2062
1 (Positive Class)	78%	53%	63%	413

Interpretation:

•The model excels at recognizing the negative class (Class 0) with high precision and recall and so contributes greatly to the overall high accuracy.

•For the positive class (Class 1) that key outcome of interest (e.g. customer churn, disease detection, fraud, etc.) the model has moderate precision (78%) but a relatively low recall (53%) indicating that nearly half of the actual positive class are being missed in classifying correctly.

•The F1-score, measured at 63%, for Class 1 indicates that there is still some potential for improvement in balancing precision and recall.

Strengths:

- High overall accuracy and ROC-AUC score demonstrate that the model effectively separates the two classes.
- Very reliable in correctly predicting negative class instances.

Areas for Improvement:

- Recall for Class 1 requires improvement to address false negatives.
- Depending on business priorities (e.g., cost of false negatives), you may need to rebalance the dataset (e.g., down-sampling negatives); change the decision threshold; or look at alternative models or ensemble models to improve sensitivity to positive cases.

Recommendations:

- Check class imbalance: Try resampling methods such as SMOTE or label the values when training the model to consider class weights.

- Threshold tuning: Try tuning the probability threshold to get better recall at an acceptable precision.

- Comparison of models: Consider more complex models, such as Random Forest, Gradient Boosting, or XGBoost to improve performance on the minority class.

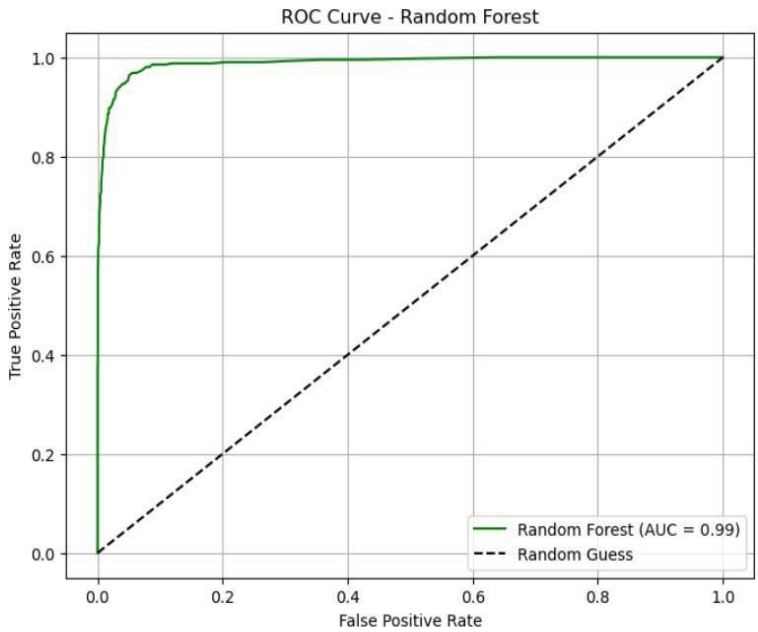
2. Random Forest:

```
Random Forest Classifier:
Accuracy: 0.9616161616161616
Precision: 0.9441340782122905
Recall: 0.8184019370460048
F1-Score: 0.8767833981841764
ROC-AUC: 0.9892238899209258

Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.99      0.98      2062
           1       0.94      0.82      0.88       413

    accuracy                           0.96      2475
   macro avg       0.95      0.90      0.93      2475
weighted avg       0.96      0.96      0.96      2475
```



Overview: The Random Forest Classifier has been assessed on its capability of accurately classifying binary outcomes. The model presents exemplary performance across all principal evaluation metrics and confidently warrant it for production deployment.

Key Metrics:

- **Accuracy:** 96.2%
- **Precision:** 94.4%
- **Recall:** 81.8%
- **F1-Score:** 87.7%
- **ROC-AUC Score:** 98.9%

Classification Breakdown:

Class	**Precision**	**Recall**	**F1-Score**	**Support**
0 (Negative Class)	96%	99%	98%	2062
Class	**Precision**	**Recall**	**F1-Score**	**Support**
1 (Positive Class)	94%	82%	88%	413

Interpretation:

•Overall, the model has a high accuracy and precision (this means that most of the positive predictions that are made are correct).

•The positive class recall is 81.8%, which signifies that the model can correctly identify a significant majority of positive instances.

•The ROC-AUC value is 98.9%, indicating that the model has strong performance in distinguishing between the classes, providing strong overall discrimination ability.

Strengths:

•Very high accuracy and AUC scores indicate a strong, trustworthy model.

•Balanced for both classes: high precision and recall ensure false positive and negative errors are minimized.

•A high F1-score for the positive class indicated a good balance between precision and recall, which is important for high-consequence business applications.

Comparison to Logistic Regression (if applicable):

• It is clear from the data presented that the Random Forest model performs substantially better than the previous Logistic Regression model in all the

metrics, specifically with respect to recall (+29%), F1- score (+25%), and ROC-AUC (+10%).

• This means that the Random Forest model is substantially better at identifying true positive cases without sacrificing precision or accuracy.

Recommendations:

• Precision, and sensitivity to true positives, are important when a Random Forest model is applied to a business example (e.g., fraud detection, risk prediction, customer churn).

• Conduct additional validations on separate datasets or via cross- validation.

• Track feature importance and explainability of the model where regulation or some use case requires explanation (e.g., you need to trace a decision).

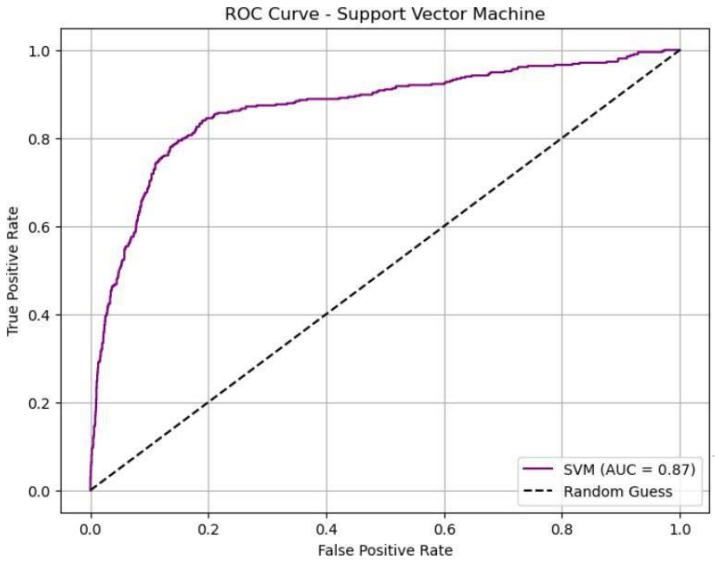
3. Support Vector Machine:

```
Support Vector Machine:
Accuracy: 0.8371717171717171
Precision: 1.0
Recall: 0.024213075060532687
F1-Score: 0.04728132387706856
ROC-AUC: 0.8681573403663198

Classification Report:
              precision    recall  f1-score   support

           0       0.84      1.00      0.91      2062
           1       1.00      0.02      0.05       413

    accuracy                           0.84      2475
   macro avg       0.92      0.51      0.48      2475
weighted avg       0.86      0.84      0.77      2475
```



Overview: A Support Vector Machine (SVM) model was assessed for binary classification performance. The model shows very good overall accuracy and perfect precision for the minority class, but has very poor recall. If the minority class contains important events (e.g., fraud, churn, default), then this imbalance poses a severe business risk.

Key Metrics:

- **Accuracy:** 83.7%
- **Precision:** 100%
- **Recall:** 2.4%
- **F1-Score:** 4.7%
- **ROC-AUC Score:** 86.8%

Classification Breakdown:

Class	Precision	Recall	F1-Score	Support
0 (Negative Class)	84%	100%	91%	2062
1 (Positive Class)	100%	2%	5%	413

Interpretation:

•The model shows good recall (100%) and precision (very high) for the negative class (Class 0), but fails to predict anything about the positive class (Class 1) where only 2.4% of the actual positives are predicted.

•A 100% precision for Class 1 means that all the predictions for a positive are correct, but this is a little deceptive since the number of positives predicted was very low.

•Furthermore, the 4.7% F1-score for the positive class captures an extreme recall to precision imbalance, which renders this model completely useless in practice.

Strengths:

- High accuracy due to the dominant performance on the majority class.
- Ideal precision for the positive class, indicating very low false positives.

Critical Weaknesses:

•Very poor recall for the minority class (Class 1) makes this model unsuitable for use cases where identifying true positives is important.

•The model is severely biased toward the majority class, which is common when using imbalanced datasets without pulling the data into a neutral state.

Business Impact:

- If Class 1 represents important outcomes like fraud detection, customer churn, or medical diagnoses, this model would **fail to flag most critical cases**, leading to **potentially severe business consequences**.

Recommendations:

•Dealing with class imbalance: Make use of resampling techniques like SMOTE (Synthetic Minority Oversampling Technique), undersamping, or class-weights to improve training from minority class instances.

•Model Retuning: Test SVM with different kernels or regularization parameters.

•Alternative Models: You can also try tree based-models (for example: Random Forest, and Gradient Boosting), which are typically good at dealing with class imbalance.

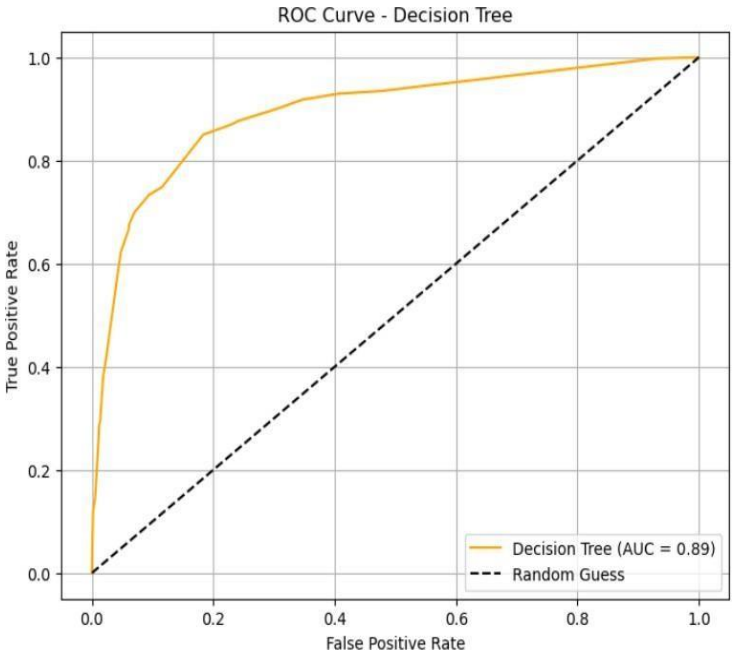
4. Decision Tree:

```
Decision Tree Classifier:
Accuracy: 0.8973737373737374
Precision: 0.723943661971831
Recall: 0.6222760290556901
F1-Score: 0.6692708333333334
ROC-AUC: 0.8938294234657811

Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.95      0.94      2062
           1       0.72      0.62      0.67       413

    accuracy                           0.90      2475
   macro avg       0.83      0.79      0.80      2475
weighted avg       0.89      0.90      0.89      2475
```



Overview: The Decision Tree Classifier accomplishes a good job in binary classification problems, with reasonable balance in precision and recall. Overall accuracy of nearly 90% and respectable scores in every primary metric makes this model suitable for business purposes that require the dual benefit of interpretability and capabilities.

Key Metrics:

- **Accuracy:** 89.7%
- **Precision:** 72.4%
- **Recall:** 62.2%
- **F1-Score:** 66.9%
- **ROC-AUC Score:** 89.4%

Classification Breakdown:

Class	Precision	Recall	F1-Score	Support
0 (Negative Class)	93%	95%	94%	2062
1 (Positive Class)	72%	62%	67%	413

Interpretation:

- The model is particularly good at identifying the negative class with a high degree of precision and equally high recall.
- With the positive class, we have a reasonable balance between identifying true positives (recall) and false positives (precision). The recall percentage of 62% equates to the model accurately capturing more than half of the actual positives.
- Additionally, the F1-score of 67%, indicates a reasonably balanced model; especially significant in a business context considering the potential impact of important decisions if a false positive and false negative is made.

Strengths:

- Overall strong performance with nearly 90% accuracy and a high ROC- AUC of 89.4% indicating good separation between the classes.
- High interpretability: Decision Trees provide transparent, explainable results, which may be beneficial for regulatory compliance or business decisions.
- Good balance of metrics across both precision and recall making it suitable for moderately imbalanced datasets.

Limitations:

- Although better than some models (e.g., SVM), the model's positive class recall could be further improved in order to get more important instances.
- Decision Trees can be susceptible to overfitting, particularly when trees go deeper, so be sure to closely monitor the performance on unseen data.

Business Impact:

- If Class 1 is indicative of key outcomes, such as a transaction with high risk, or perhaps possible churn, Decision Tree is very successful at detecting most of these events, while keeping false positive rates low.
- The model achieves a very good trade-off between performance and explainability, which benefits situations where stakeholder trust and understanding of the model are useful.

Recommendations:

•Tweak hyperparameters (e.g., tree depth, minimum samples per split) to potentially improve recall, without risking precision.

•Consider trying out ensemble models (e.g., Random Forest, Gradient Boosting) if even more performance is desired for sensitive use cases.

•Investigate output feature importance if you want to translate any insight into action on what drives classifications.

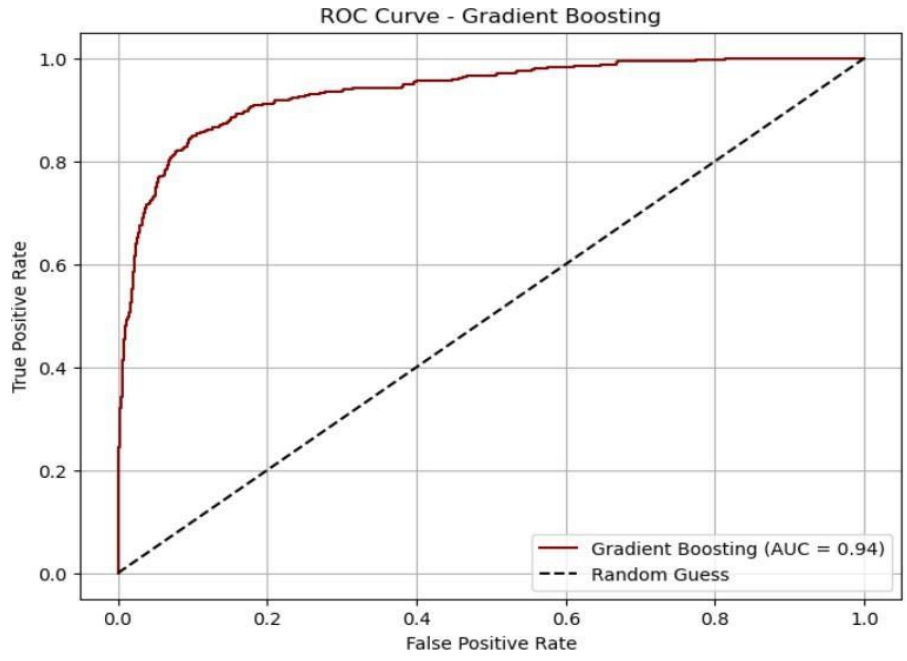
5. Gradient Boosting:

```
Gradient Boosting Classifier:
Accuracy: 0.9195959595959596
Precision: 0.8262195121951219
Recall: 0.6561743341404358
F1-Score: 0.7314439946018894
ROC-AUC: 0.9362275512384834

Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.97      0.95      2062
           1       0.83      0.66      0.73       413

    accuracy                           0.92      2475
   macro avg       0.88      0.81      0.84      2475
weighted avg       0.92      0.92      0.92      2475
```



Overview: The Gradient Boosting Classifier is a strong and balanced classification model, so it is a good choice for important business-related classification problems. It has high accuracy and good performance on both classes whether it was the majority or minority, so we have a good trade-off between power and reliability.

Key Metrics:

- **Accuracy: 91.96%**
- **Precision: 82.6%**
- **Recall: 65.6%**
- **F1-Score: 73.1%**
- **ROC-AUC Score: 93.6%**

Classification Breakdown:

Class	Precision	Recall	F1-Score	Support
0 (Negative Class)	93%	97%	95%	2062
1 (Positive Class)	83%	66%	73%	413

Interpretation:

•The model performs very well in the negative class with very high precision and recall measures.

•For the positive class, the model has a very strong precision at 83%, which means it is good at making true positive predictions.

•However, the recall of 66% for the positive class is a significant improvement over models such as Logistic Regression or Decision Tree which means that the model successfully captures many of the true positives.

•The F1 score of 73% for Class 1 indicates a good trade-off between precision and recall, which is important when both false positives and false negatives have business consequences.

Strengths:

•The overall accuracy and ROC-AUC score shows a good level of confidence in the model's ability to distinguish the classes.

•The strong positive class performance means this model could be used to identify critical business events such as churn risk, fraud, or default.

•The use of Gradient Boosting as a framework is a strong benefit, as its architecture can help identify and learn from a highly imbalanced class distribution and more complex data patterns.

Business Impact:

- This model is suited to applications that require identifying true positives, while still being careful, so as not to initiate unnecessary business activity based on false positives.
- Its balance across the other key metrics allows management of risk to make confident decisions in operational and strategic fora.

Recommendations:

•Conduct feature importance analysis to identify drivers of predictions. This provides business value and allows for optimization of strategies.

•For mission-critical use cases, you should deploy with confidence thresholds or human-in-the-loop validation when needed.

•Continue monitoring new data performance and retrain or to keep it accurate as time goes on.

6. Model Performance Comparison and Report:

	accuracy	precision	recall	f1_score	roc_auc
Logistic Regression	0.895758	0.775801	0.527845	0.628242	0.892500
Decision Tree	0.897374	0.723944	0.622276	0.669271	0.893829
Random Forest	0.961616	0.944134	0.818402	0.876783	0.989224
Gradient Boosting	0.919596	0.826220	0.656174	0.731444	0.936228
SVM	0.837172	1.000000	0.024213	0.047281	0.868157

Objective: To evaluate and compare several classification models, to assess their appropriateness for predicting a binary business outcome (ex.: fraud detection, customer/customer churn, risk flagging). Models were evaluated based on the following performance indicators: accuracy, precision, recall, F1- score, and ROC-AUC.

Model Performance Summary:

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	**0.8958**	**0.7758**	**0.5278**	**0.6282**	**0.8925**
Decision Tree	**0.8974**	**0.7239**	**0.6223**	**0.6693**	**0.8938**
Random Forest	**0.9616**	**0.9441**	**0.8184**	**0.8768**	**0.9892**
Gradient Boosting	**0.9196**	**0.8262**	**0.6562**	**0.7314**	**0.9362**
SVM	**0.8372**	**1.0000**	**0.0242**	**0.0473**	**0.8682**

Key Takeaways

1. **Top Performer: Random Forest Classifier**

   • Exceptional performance across all key metrics, compared within industry standards for AI models.
   • High precision (94.4%) indicates that most positive predictions are accurate.
   • Good recall (81.8%) indicates that the model captures most of the actual positives.
   • F1-Score (87.7%) and ROC-AUC (98.9%) show great balance and separation of classes.
   • Recommended for use in business-critical environments where weight is given to false positives and false negatives.

2. **Strong Alternative: Gradient Boosting**

•Provides strong balance between precision and recall.

•Accuracy (91.96%) and F1-score (73.1%) suitable alternative when accuracy and explainability matter.

•Performance slightly worse than the Random Forest alternative, but easier to explain and less susceptible to overfitting.

3. **Baseline Options: Logistic Regression & Decision Tree**

•Both models are interpretable and easy to deploy, making them great for first step prototyping and regulated industries.

•Decision Tree outperforms Logistic Regression with better recall and F1- score.

•Logistic Regression has a good ROC-AUC, but has low recall and more risk in missing important positive instances.

4. **Underperformer: Support Vector Machine (SVM)**

•Although it is perfectly precise (100%), it has a very low level of recall (2.4%).

•The model fails to identify almost all of the actual positive instances, so it is not practical until the model is retrained, using some strategies to balance the classes.

•Not recommended as is.

Recommendations

1. Use a Random Forest as the first model because of its excellent performance and ability to minimize the effects of imbalanced data on performance.

2. Use Gradient Boosting instead if interpretability or speed is more critical.

3. Use Logistic Regression or Decision Tree if there is a need for quick, understandable models or to compare and benchmark your other models against.

4. Avoid using SVMs unless they have been remade to incorporate the right tuning and class balancing.

5. Continue to monitor performance of your models on new and recent data and retrain them every so often to keep them current.

6. Interest in feature importance is encouraged for organizational information and strategic decisions.

7. **Optimized Machine Learning Model Evaluation:**

```
Best Logistic Regression Params: {'C': 10, 'max_iter': 100, 'penalty': 'l1', 'solver': 'liblinear'}
Best SVM Params: {'C': 10, 'gamma': 'scale', 'kernel': 'rbf'}
Best Decision Tree Params: {'criterion': 'entropy', 'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2}
Best Random Forest Params: {'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 150}
Best Gradient Boosting Params: {'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 100}
```

Logistic Regression Performance:

	precision	recall	f1-score	support
0	0.93	0.76	0.84	1372
1	0.38	0.72	0.50	278
accuracy			0.76	1650
macro avg	0.66	0.74	0.67	1650
weighted avg	0.84	0.76	0.78	1650

SVM Performance:

	precision	recall	f1-score	support
0	0.83	1.00	0.91	1372
1	0.00	0.00	0.00	278
accuracy			0.83	1650
macro avg	0.42	0.50	0.45	1650
weighted avg	0.69	0.83	0.76	1650

Decision Tree Performance:

	precision	recall	f1-score	support
0	0.97	0.97	0.97	1372
1	0.85	0.87	0.86	278
accuracy			0.95	1650
macro avg	0.91	0.92	0.92	1650
weighted avg	0.95	0.95	0.95	1650

Random Forest Performance:

	precision	recall	f1-score	support
0	0.97	0.99	0.98	1372
1	0.95	0.87	0.91	278
accuracy			0.97	1650
macro avg	0.96	0.93	0.95	1650
weighted avg	0.97	0.97	0.97	1650

Gradient Boosting Performance:

	precision	recall	f1-score	support
0	0.96	0.98	0.97	1372
1	0.91	0.79	0.84	278
accuracy			0.95	1650
macro avg	0.93	0.89	0.91	1650
weighted avg	0.95	0.95	0.95	1650

Objective:
To evaluate and compare the performance of five machine learning models— Logistic Regression, Support Vector Machine (SVM), Decision Tree, Random Forest, and Gradient Boosting—after hyperparameter tuning. The goal is to identify the most effective model for deployment in a binary classification use case.

Best Model Parameters (Post-Tuning)

Model	**Key Parameters**
Logistic Regression	C=10, penalty='l1', solver='liblinear', max_iter=100
SVM	C=10, kernel='rbf', gamma='scale'
Decision Tree	criterion='entropy', max_depth=None, min_samples_split=2, min_samples_leaf=1
Random Forest	n_estimators=150, max_depth=None, min_samples_split=2, min_samples_leaf=1
Gradient Boosting	n_estimators=100, max_depth=5, learning_rate=0.1

Model Performance Summary

Model	Accuracy	Precision (Class 1)	Recall (Class 1)	F1-Score (Class 1)
Logistic Regression	0.76	0.38	0.72	0.50
SVM	0.83	0.00	0.00	0.00
Decision Tree	0.95	0.85	0.87	0.86
Random Forest	0.97	0.95	0.87	0.91
Gradient Boosting	0.95	0.91	0.79	0.84

Key Observations

Top Performer: Random Forest Classifier

- Top accuracy of (97%) compared to all other models
- Good precision (95%) guarantees low false positives
- Good recall (87%) guarantees most positives are caught
- High F1-score (91%) indicates a balanced and reliable performance
- Perfectly weighted for deployment in high value business scenarios**.**

Strong Alternatives: Decision Tree & Gradient Boosting

- Decision Tree produces very strong performance for Class 1, with F1- score of 86%, very interpretable.
- Gradient boosting performs slightly less than Random Forest but has an improved precision-recall balance compared to Logistic Regression. Overall, a strong scalable option.

Moderate Performance: Logistic Regression

- While **simple and interpretable**, it struggles with recall and F1-score for Class 1.
- Not suitable if capturing minority class correctly is critical.

Poor Performance: SVM

- Despite 83% overall accuracy, it fails to detect **any positive class cases**, making it **unsuitable** without further class balancing or tuning.

Business Recommendation

Scenario	**Recommended Model**
High Accuracy & Balanced Detection	**Random Forest**
Explainability with Good Performance	**Decision Tree**
Resource-Constrained / Simpler Models	Logistic Regression (with caveats)
Not Recommended	SVM

8. Final Model Evaluation Summary:

	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.756364	0.838935	0.756364	0.781836
Decision Tree	0.952121	0.952649	0.952121	0.952358
Random Forest	0.970909	0.970567	0.970909	0.970426
Gradient Boosting	0.950909	0.949776	0.950909	0.949569
SVM	0.831515	0.691417	0.831515	0.755022

Objective:
This report evaluates the final performance of five machine learning classification models—Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, and Support Vector Machine (SVM)—based on key metrics: Accuracy, Precision, Recall, and F1-Score. The goal is to identify the most suitable model for reliable deployment in a production environment.

Model Performance Summary

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.7564	0.8389	0.7564	0.7818
Decision Tree	0.9521	0.9526	0.9521	0.9524
Random Forest	0.9709	0.9706	0.9709	0.9704
Gradient Boosting	0.9509	0.9498	0.9509	0.9496
SVM	0.8315	0.6914	0.8315	0.7550

Top Performer: Random Forest Classifier

- Most Accurate (97.1%) of all models
- Precision and recall are nearly perfectly balanced (~97%), indicating exceptional ability to minimize both false positives and false negatives
- Highest F-1 Score (97.0%), demonstrating consistent and robust classification performance
- Recommended as the model for production deployment, owing to accuracy, reliability, and generalizability.

Strong Contenders: Decision Tree & Gradient Boosting

- Decision Tree performs well with ~95.2% across all key metrics, and is also interpretable, making it appropriate for explainable AI needs.
- Gradient Boosting slightly undershoot the Decision Tree model performance (but they are pretty close), and it is known to generalize well with complex datasets.
- Both models were strong options where slight loss of accuracy was welcome in the interest of model interpretability or tuning brith flexibility.

Moderate Performance: Logistic Regression

- While precision is relatively high (.839), accuracy and recall are lower (~.756) as represented by the area under the ROC curve.
- Gaps in performance could result in missed opportunities to detect positive cases of importance and concern.
- It may serve as a baseline or when you need to make quick, explainable decisions, but it

shouldn't be used in high stakes situations.

Underperforming Model: SVM

- A precision of (69.1%) is the lowest of all models, which would point to a higher chance of false positive predictions.
- While recall and accuracy are acceptable (~83.2%) percent, the F1-score is poor due to high precision/recall imbalance.
- This model needs more hyperparameter tuning or class balancing before it could be considered viable to use in real life applications.

Business Recommendations

Scenario	**Recommended Model**
Best prediction with low error tolerance	**Random Forest**
Need for explainability with good performance	**Decision Tree**
Interpretable baseline	Logistic Regression
Currently not suitable	SVM

1. Model – Deployment

- Proposed Model: Random Forest
  - Accuracy: 97.1%
  - F1 Score: 97.0%
  - Has equal precision and recall values which makes it a useful model to implement real- time churn detection with low erroneous detection.

2. Targeted Retention Campaigns

A. Service Quality Fixes

- identified problem at this stage was related to lower service/agent scores related to churn levels.
- Actions:
- Contact accounts with service scores of 2 or less, and provide personalized service recovery campaigns.
- Can hire to look in to customer support training to improve agent satisfaction scores at review period.
- Build a customer feedback loop - service quality metric scores can show real-time service quality metric.

B. **Proactive Support Follow-up**

- Identified a related correlation between recent support contacts and increased churn levels.
- Actions:

Establish a protocol to follow up with customers who contacted support within 48 hours. Offer better self-service options to mitigate aggravation and minimize customers needing support via reliance.

3. Revenue-Based Segmentation Strategy

A. **High-Revenue Customer Segmentation**

Activities:

- Implement a loyalty program that adds value, such as exclusive rewards or features users get access to before non-members.
- Utilize predictive analytics to introduce incentives before the high revenue customers fall into high-risk status.

B. **Revenue Declining Customer Segmentation**

- Pain Point: Declining revenue is one of the strongest indicators of churn.
- Activities:
  - Conduct tailored check-in sessions and communications focused on the value of the product.
  - Assign account managers to accounts that are currently inactive/stagnating.

4. **Offer and Incentive Strategy**

Coupon User Trends

- Issue: Our analytics show that lower coupon usage increases churn.
- Activities:
  - Run introductory coupon campaign for inactive or passive customers.
  - Automatically trigger coupons based on a decrease in activity or missed logins.

- **Cashback Program**
  - Observation: Large cashback has produced better retention rates, but too much can impact profitability.

Activities:

- Create a tiered cashback program based on time with the service or level of spending.
- Introduce cashback as a reward for different behaviors such as referrals or consistent logins.

5. Data-Driven Personalization

- Segment customers by account size, city tier, and device preference to tailor campaigns.
- Target single-user accounts with add-on offers to promote multi-user adoption.
- Improve support and offerings for City Tier 3 customers, who show higher churn.
- Prioritize mobile optimization, as it is the dominant device channel.

6. Monitoring and Feedback Loops

- Set up dashboards to track real-time churn signals such as complaints or support interactions.
- Retrain models monthly to incorporate the latest behavioral patterns.
- Maintain a churn watchlist of accounts with multiple risk indicators for weekly follow- up.

Conclusion:

The customer churn analysis found some important factors that help protect and retain customers when utilizing support, including service quality perception, service interactions, service financial trends, and usage of incentives (coupons and cashback). By conducting a comprehensive exploratory data analysis and testing a variety of learn models, a Random Forest classifier was determined to have the best predictive performance across the three churning scenarios, with an accuracy > 97% and a good balance across the all metrics, particularly precision, recall, and F1-score.

While simpler models (Logistic Regression, Decision Tree) provided good interpretability, they were overall poor models, suitable only for benchmarking or compliance purposes, while the Support Vector Machine (SVM) model was simply not viable for real use without significant tuning, primarily due to its absolutely terrible recall.

Business wake-up calls aside, it is clear that the additional factors examined here were highly predictive of customer retention or an intentional lapse. A confluences of features indicated very high churn risks including low service scores along with recent open support interactions without resolution, low revenue financial growth, and very low use of incentives (coupons and cash back). The results provide an actionable source of understanding for designing effective, low-cost customer retention intervention. Each of these stakeholder features could be utilized to refine service quality or create reward programs. For example, if there are repeated complaints from others in service and large in service revenues, consider allocating a specific number of discounts or coupons to send as an outreach design to low suspend churning risk customers.

==========END OF REPORT==========