CAPSTONE

Banking Project Probability

- RAHUL SHARMA

Problem Statement:-

This business problem is a supervised learning example for a credit card company. The objective is to predict the probability of default (whether the customer will pay the credit card bill or not) based on the variables provided. There are multiple variables on the credit card account, purchase and delinquency information which can be used in the modelling.

PD modelling problems are meant for understanding the riskiness of the customers and how much credit is at stake in case the customer defaults. This is an extremely critical part in any organization that lends money [both secured and unsecured loans].

A. Problem Understanding:-

a) Defining problem statement

Columns	Description
userid	The unique user id of the customer who is holding the credit card.
default	Target Variable. 1 - Indicates the user has defaulted. 0 - Indicates that the person has not defaulted
acct_amt_added_12_24m	The total amount of the purchases made using the credit card between 24 months in the past to the present date to the 12 months in the past to the current date.
acct_days_in_dc_12_24m	The total number of days that the Credit Card Account has stayed in the Debt-Collection Status between 24 months in the past to the present date to the 12 months in the past to the current date. . Note: Debt-Collection Status: If a Customer has not even paid a minimum amount of the bill, then the account goes into a state called as debt-collection wherein the previous dues from the customer needs to be collected using an agency.
acct_days_in_rem_12_24m	The total number of days that the Credit Card Account has stayed in the Reminder Status between 24 months in the past to the present date to the 12 months in the past to the current date. Note: Reminder Status: If a Customer has not yet paid the Credit Card Bill even after the last due date, the bank used to send reminders for making the payment. If an account starts receiving reminder messages, then it goes to the reminder status.
acct_days_in_term_12_24m	The total number of days that the Credit Card Account has stayed in the Termination Status between 24 months in the past to the present date to the 12 months in the past to the current date. Note: Termination Status: If a Customer has paid the Credit Card Bill even after multiple reminders, then his card gets terminated and he will not be able to make any transactions using the credit card unless it gets activated again.
acct_incoming_debt_vs_paid_0_24m	The ratio of the amount collected out of the total debt in an account by an agency to the total debt amount of the account in the previous 24 months from the current date.
acct_status	The current status of the account. 1 represents active account, while 0 represents inactive account.
acct_worst_status_0_3m	The total number of days that the Credit Card Account has stayed in the Worst Status between 3 months in the past to the present date . Note: Worst Status: If a Customer has not even paid a minimum amount of the bill for more than 30 days post the due date, then the account goes into a state called as worst date.
acct_worst_status_12_24m	The total number of days that the Credit Card Account has stayed in the Worst Status between 24 months in the past to the present date and 12 months in the past to the
acct_worst_status_3_6m	The total number of days that the Credit Card Account has stayed in the Worst Status between 6 months in the past to the present date and 3 months in the past to the current
acct_worst_status_6_12m	The total number of days that the Credit Card Account has stayed in the Worst Status between 12 months in the past to the present date and 6 months in the past to the
age	The age of the customer.
avg_payment_span_0_12m	The average payment span that the customer has taken in days after the credit card bill got generated in the last one year.
avg_payment_span_0_3m	The average payment span that the customer has taken in days after the credit card bill

merchant_category	The category of the merchant.
merchant_group	The group of the merchant.
has_paid	Whether the customer has paid the current cresit card bill or not. True - Paid. False -
max_paid_inv_0_12m	The maximum credit card bill amount that has been paid by the customer in the last one
max_paid_inv_0_24m	The maximum credit card bill amount that has been paid by the customer in the last two
name_in_email	Name of the customer in email.
num_active_div_by_paid_inv_0_12m	Ratio of the number of unpaid bills to the paid bills in the last one year.
num_active_inv	Number of the active invoices (unpaid bills)
num_arch_dc_0_12m	number of archived purchases that were in debt collection status in the last one year
num_arch_dc_12_24m	number of archived purchases that were in debt collection status between 24 months in the past to the present date and 12 months in the past to the current date .
num_arch_ok_0_12m	number of archived purchases that were paid in the last one year.
num_arch_ok_12_24m	number of archived purchases that were paid between 24 months in the past to the present date and 12 months in the past to the current date .
num_arch_rem_0_12m	number of archived purchases that were in the reminder status in the last one year.
status_max_archived_0_6_months	maximum number of times the account was in archived status in the last 6 months.
status_max_archived_0_12_months	maximum number of times the account was in archived status in the last one year.
status_max_archived_0_24_months	maximum number of times the account was in archived status in the last two years.
recovery_debt	The total amount that has been recovered out of the entire debt amount on the account.
sum_capital_paid_acct_0_12m	sum of pirncipal balance paid on account in the last one year.
sum_capital_paid_acct_12_24m	sum of pirncipal balance paid on account between 24 months in the past to the present date and 12 months in the past to the current date .
sum_paid_inv_0_12m	The total amount of the paid invoices in the last one year.
time_hours	The total hours spent by the customer in purchases made using the credit card.

Objective: To identify distinct customer segments based on their credit card usage and payment behavior using clustering techniques. The goal is to understand the characteristics of each segment to improve business strategies for risk management, marketing, and customer service.

Context: The data-set contains information about credit card usage and payment behavior of customers. Each row represents a customer with various features describing their credit card activity, payment history, and default status.

	userid	default	acct_amt_added_12_24m	acct_days_in_dc_12_24m	acct_days_in_rem_12_24m	acct_days_in_term_12_24m	acct_incoming_debt_vs_paid_0_2
0	4567129.0	0.0	0.0	0.0	0.0	0.0	
1	2635118.0	0.0	0.0	0.0	0.0	0.0	N
2	4804232.0	0.0	0.0	0.0	0.0	0.0	N
3	1442693.0	0.0	0.0	NaN	NaN	NaN	N
4	4575322.0	0.0	0.0	0.0	0.0	0.0	N

5 rows × 36 columns

b) Need of the study/project

In the role of a data scientist, it is essential to understand why this study or project is necessary:

Risk Management:

Credit Risk: Predicting the probability of default helps in assessing the credit risk associated with each customer. By identifying high-risk customers, the company can take preventive measures to mitigate potential losses.
Portfolio Management: Understanding the risk profile of customers allows for better management of the credit portfolio, ensuring a balanced risk-return trade-off.

Financial Stability:

Provisioning and Reserves: Accurate PD modeling helps in determining the amount of reserves required to cover potential losses. This ensures the financial stability and health of the organization.
Capital Allocation: Efficient capital allocation based on risk assessment ensures that resources are optimally utilized, contributing to the overall profitability of the company.

Regulatory Compliance:

Compliance with Standards: Financial institutions are required to comply with regulatory standards such as Basel III, which mandates robust risk assessment and management practices. PD modeling is a key component of these requirements.
Transparency: Predictive models provide transparency in credit decision-making processes, which is crucial for regulatory reporting and audit purposes.

c) Understanding business/social opportunity

The project presents several business and social opportunities:

Enhanced Customer Experience:

Personalized Services: By understanding the risk profile of customers, the company can offer personalized credit products and services tailored to individual needs and risk levels.
Proactive Customer Support: Identifying customers at risk of default allows for proactive engagement and support, such as offering restructuring options or financial counseling.

Competitive Advantage:

Data-Driven Decisions: Leveraging predictive analytics provides a competitive edge by enabling data-driven decision-making. This can lead to improved credit approval processes, optimized interest rates, and better customer retention strategies.
Market Positioning: Companies with robust risk management practices are perceived as more reliable and stable, enhancing their market reputation and attracting more customers.

Social Impact:

Financial Inclusion: By accurately assessing credit risk, the company can extend credit to underserved segments of the population, promoting financial inclusion.
Economic Stability: Reducing the incidence of defaults contributes to the stability of the financial system, which in turn supports broader economic stability and growth.

B. Data Report:-

b) Visual inspection of data (rows, columns, descriptive details)

```
Number of Rows: 99979
Number of Columns: 36
```

DATA DESCRIPTIVE DETAILS

```
Dataset Description:
              userid       default  acct_amt_added_12_24m  \
count  9.997700e+04  89977.000000           9.997700e+04
mean   2.998947e+06      0.125454           1.225503e+04
std    1.154211e+06     33.337757           3.548133e+04
min    0.000000e+00      0.000000           0.000000e+00
25%    2.000260e+06      0.000000           0.000000e+00
50%    2.998815e+06      0.000000           0.000000e+00
75%    4.000633e+06      0.000000           4.937000e+03
max    4.999868e+06  10000.000000           1.128775e+06

       acct_days_in_dc_12_24m  acct_days_in_rem_12_24m  \
count            88141.000000             88141.000000
mean                 0.357325                 5.178850
std                 40.287334                45.943401
min                  0.000000                 0.000000
25%                  0.000000                 0.000000
50%                  0.000000                 0.000000
75%                  0.000000                 0.000000
max              11836.000000             11836.000000

       acct_days_in_term_12_24m  acct_incoming_debt_vs_paid_0_24m  \
count              88141.000000                     40662.000000
mean                   0.421177                         2.789992
std                   39.973774                       295.333975
min                    0.000000                         0.000000
25%                    0.000000                         0.000000
50%                    0.000000                         0.152090
75%                    0.000000                         0.662993
max                11836.000000                     59315.000000

        acct_status  acct_worst_status_0_3m  acct_worst_status_12_24m  ...  \
count  45604.000000            45604.000000              33216.000000  ...
mean       2.234431                2.365165                  3.347212  ...
std      254.608935              254.608589                366.303445  ...
min        1.000000                1.000000                  1.000000  ...
25%        1.000000                1.000000                  1.000000  ...
50%        1.000000                1.000000                  1.000000  ...
75%        1.000000                1.000000                  2.000000  ...
max    54373.000000            54373.000000              66761.000000  ...

       num_arch_ok_12_24m  num_arch_rem_0_12m  status_max_archived_0_6_months  \
count        88943.000000        88943.000000                    88943.000000
mean             6.846576            0.483995                        0.821740
std             16.067944            1.395611                        0.716644
min              0.000000            0.000000                        0.000000
25%              0.000000            0.000000                        0.000000
50%              2.000000            0.000000                        1.000000
75%              7.000000            0.000000                        1.000000
max            313.000000           42.000000                        3.000000

       status_max_archived_0_12_months  status_max_archived_0_24_months  \
count                     88943.000000                     88943.000000
mean                          1.074182                         1.248193
std                           0.776390                         0.820518
min                           0.000000                         0.000000
25%                           1.000000                         1.000000
50%                           1.000000                         1.000000
75%                           2.000000                         2.000000
max                           5.000000                         5.000000

       recovery_debt  sum_capital_paid_acct_0_12m  \
count   88943.000000                 88943.000000
mean        3.601925                 10860.259886
std       116.210849                 26630.618529
min         0.000000                     0.000000
25%         0.000000                     0.000000
50%         0.000000                     0.000000
75%         0.000000                  8959.500000
max     16411.000000                571475.000000

       sum_capital_paid_acct_12_24m  sum_paid_inv_0_12m     time_hours
count                  88943.000000        8.894300e+04  88943.000000
mean                    6614.945763        4.103591e+04     15.341649
std                    19243.805570        9.459642e+04      5.030877
min                        0.000000        0.000000e+00      0.000000
25%                        0.000000        3.395500e+03     11.631806
50%                        0.000000        1.705700e+04     15.808333
75%                      102.500000        4.573900e+04     19.554167
max                   341859.000000        2.962870e+06     23.999722

[8 rows x 33 columns]
```

d) Understanding of attributes (variable info, renaming if required)

```
Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99979 entries, 0 to 99978
Data columns (total 36 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   userid                                99977 non-null  float64
 1   default                               89977 non-null  float64
 2   acct_amt_added_12_24m                 99977 non-null  float64
 3   acct_days_in_dc_12_24m                88141 non-null  float64
 4   acct_days_in_rem_12_24m               88141 non-null  float64
 5   acct_days_in_term_12_24m              88141 non-null  float64
 6   acct_incoming_debt_vs_paid_0_24m      40662 non-null  float64
 7   acct_status                           45604 non-null  float64
 8   acct_worst_status_0_3m                45604 non-null  float64
 9   acct_worst_status_12_24m              33216 non-null  float64
 10  acct_worst_status_3_6m                42275 non-null  float64
 11  acct_worst_status_6_12m               39627 non-null  float64
 12  age                                   99977 non-null  float64
 13  avg_payment_span_0_12m                76141 non-null  float64
 14  avg_payment_span_0_3m                 50672 non-null  float64
 15  merchant_category                     99977 non-null  object
 16  merchant_group                        99968 non-null  object
 17  has_paid                              88943 non-null  float64
 18  max_paid_inv_0_12m                    88943 non-null  float64
 19  max_paid_inv_0_24m                    88943 non-null  float64
 20  name_in_email                         88943 non-null  object
 21  num_active_div_by_paid_inv_0_12m      70052 non-null  float64
 22  num_active_inv                        88943 non-null  float64
 23  num_arch_dc_0_12m                     88943 non-null  float64
 24  num_arch_dc_12_24m                    88943 non-null  float64
 25  num_arch_ok_0_12m                     88943 non-null  float64
 26  num_arch_ok_12_24m                    88943 non-null  float64
 27  num_arch_rem_0_12m                    88943 non-null  float64
 28  status_max_archived_0_6_months        88943 non-null  float64
 29  status_max_archived_0_12_months       88943 non-null  float64
 30  status_max_archived_0_24_months       88943 non-null  float64
 31  recovery_debt                         88943 non-null  float64
 32  sum_capital_paid_acct_0_12m           88943 non-null  float64
 33  sum_capital_paid_acct_12_24m          88943 non-null  float64
 34  sum_paid_inv_0_12m                    88943 non-null  float64
 35  time_hours                            88943 non-null  float64
dtypes: float64(33), object(3)
memory usage: 27.5+ MB
None
```

In the context of the dataset, the difference between float64 and object data types is significant for how we handle and analyze the data:

float64:-

Description: float64 is a data type that represents floating-point numbers. These numbers have decimal points and can be very large or very small.

Usage: This data type is used for numerical features that require precision. In the dataset, columns like acct_amt_added_12_24m, age, and avg_payment_span_0_12m are represented as float64.

Operations: We can perform mathematical operations like addition, subtraction, multiplication, division, etc., on float64 columns. They are essential for statistical analysis, machine learning models, and visualization.

Object:-

Description: object is a data type in pandas that is used to store any data type that is not a number. This includes strings, mixed data types, or any Python object.

Usage: In the dataset, columns like merchant_category, merchant_group, and name_in_email are represented as object. These typically store categorical data or textual information.

Operations: Operations on object columns are generally limited to string manipulations and categorical operations. They need to be encoded (e.g., using Label Encoding or One-Hot Encoding) before being used in machine learning models.

Handling Differences:-

When preparing the data for clustering or any other machine learning task, we need to handle these data types appropriately:

Numerical Features (float64):

Ensure there are no missing values or handle them appropriately (e.g., by imputation).
Standardize or normalize these features if necessary, especially for distance-based algorithms like KMeans.

Categorical Features (object):

Convert these features into numerical representations using techniques

like Label Encoding, One-Hot Encoding, or other suitable encoding methods.
Handle missing values appropriately (e.g., by filling with a placeholder or the mode).

C. Exploratory data analysis:-

a) Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones)

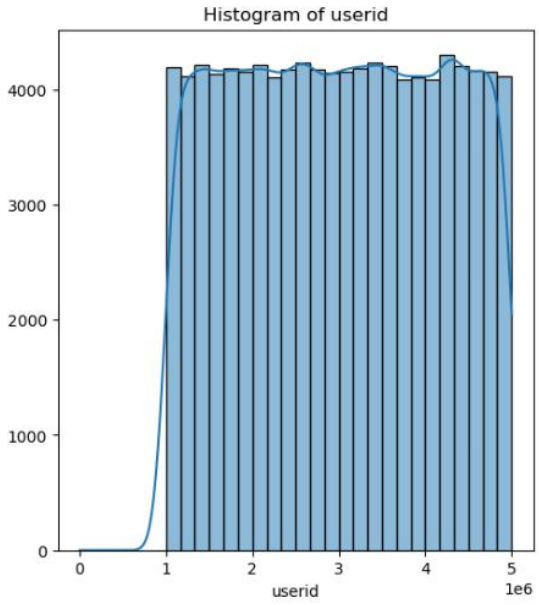
Univariate analysis for continuous variables:-

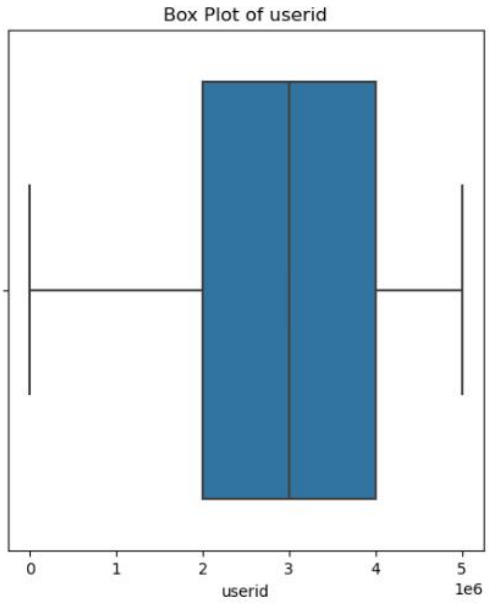
A **histogram** is a graphical representation of the distribution of data. It uses bars to show the number of data points that fall within a certain range of values (called bins). The height of each bar represents the frequency of a specific bin. In the histogram, the x-axis represents the range of possible values for the variable, and the y-axis represents the number of data points that fall within each bin.

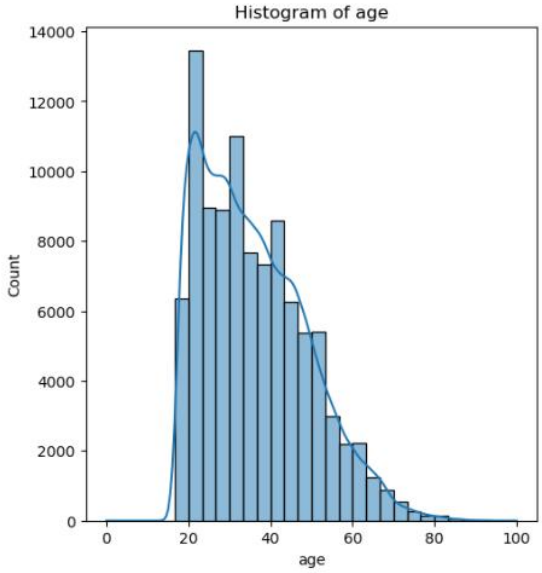
A **box plot** is another way to visualize the distribution of data. It shows the following five summary statistics:

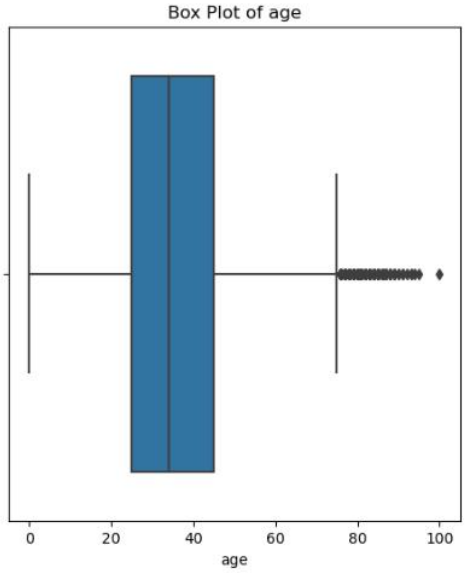
Minimum value
First quartile (Q1) - the value below which 25% of the data points fall
Median (Q2) - the middle value
Third quartile (Q3) - the value below which 75% of the data points fall
Maximum value
The box in a box plot represents the inter-quartile range (IQR), which is the difference between Q1 and Q3. The IQR represents the middle 50% of the data. The whiskers of a box plot extend from the ends of the box to the minimum and maximum values, unless there are outliers. Outliers are data points that fall outside of 1.5 times the IQR from the quartiles.

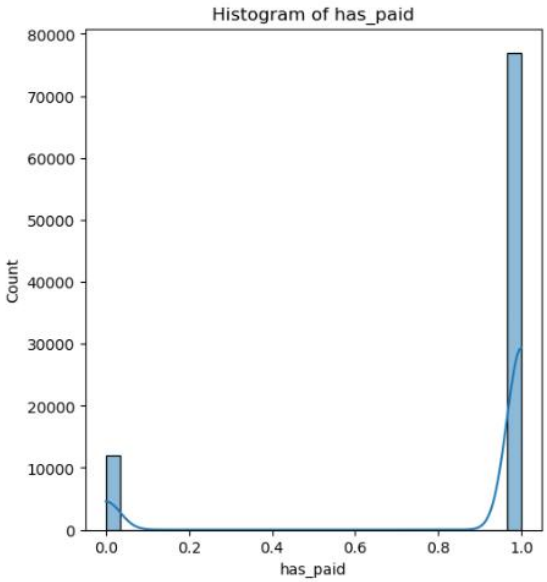
In the box plot, the line in the middle of the box represents the median. The box extends from the first quartile (Q1) to the third quartile (Q3). The whiskers extend from the ends of the box to the minimum and maximum values. There are no outliers in this data set.

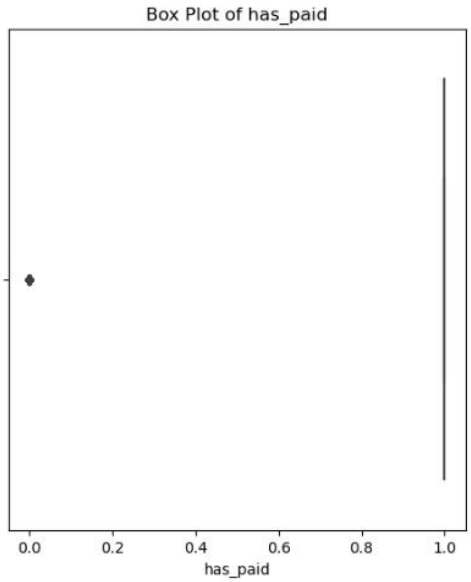

Histogram of userid
4000
3000
2000
1000
0
0
1
2
3
4
5
1e6
userid


Box Plot of userid
0
1
2
3
4
5
1e6
userid


Histogram of age
14000
12000
10000
8000
6000
4000
2000
0
Count
0
20
40
60
80
100
age


Box Plot of age
0
20
40
60
80
100
age


Histogram of has_paid
80000
70000
60000
50000
40000
30000
20000
10000
0
Count
0.0
0.2
0.4
0.6
0.8
1.0
has_paid


Box Plot of has_paid
0.0
0.2
0.4
0.6
0.8
1.0
has_paid

1. userid

- **Histogram**:
  - The userid variable appears to be uniformly distributed across the range, suggesting that user IDs are assigned in a sequential or evenly spaced manner.
  - There are no significant peaks or valleys, indicating a consistent distribution of user IDs.
- **Box Plot**:
  - The box plot shows a symmetric distribution with no outliers.
  - This confirms the even distribution seen in the histogram.

2. age

- **Histogram**:
  - The age variable shows a right-skewed distribution with a majority of users being younger.
  - The highest frequency of users is in the age range of approximately 20-40.
  - There are noticeable peaks at certain ages, suggesting these might be common ages in the dataset.
- **Box Plot**:
  - The box plot indicates that the median age is around 35.
  - There are several outliers on the higher end of the age spectrum, indicating a small number of older users compared to the majority.
  - The interquartile range (IQR) spans from about 25 to 50, showing the concentration of most users' ages.

3. has_paid

- **Histogram**:
  - The has_paid variable is binary and shows a heavily imbalanced distribution, with most users having a value of 1 (indicating they have paid their credit card bill).
  - The count of users who have not paid (value 0) is significantly lower.
- **Box Plot**:
  - The box plot shows that the variable is highly skewed towards 1.
  - There is a single outlier on the lower end, representing the users who have not paid.

Business Insights

User Distribution:

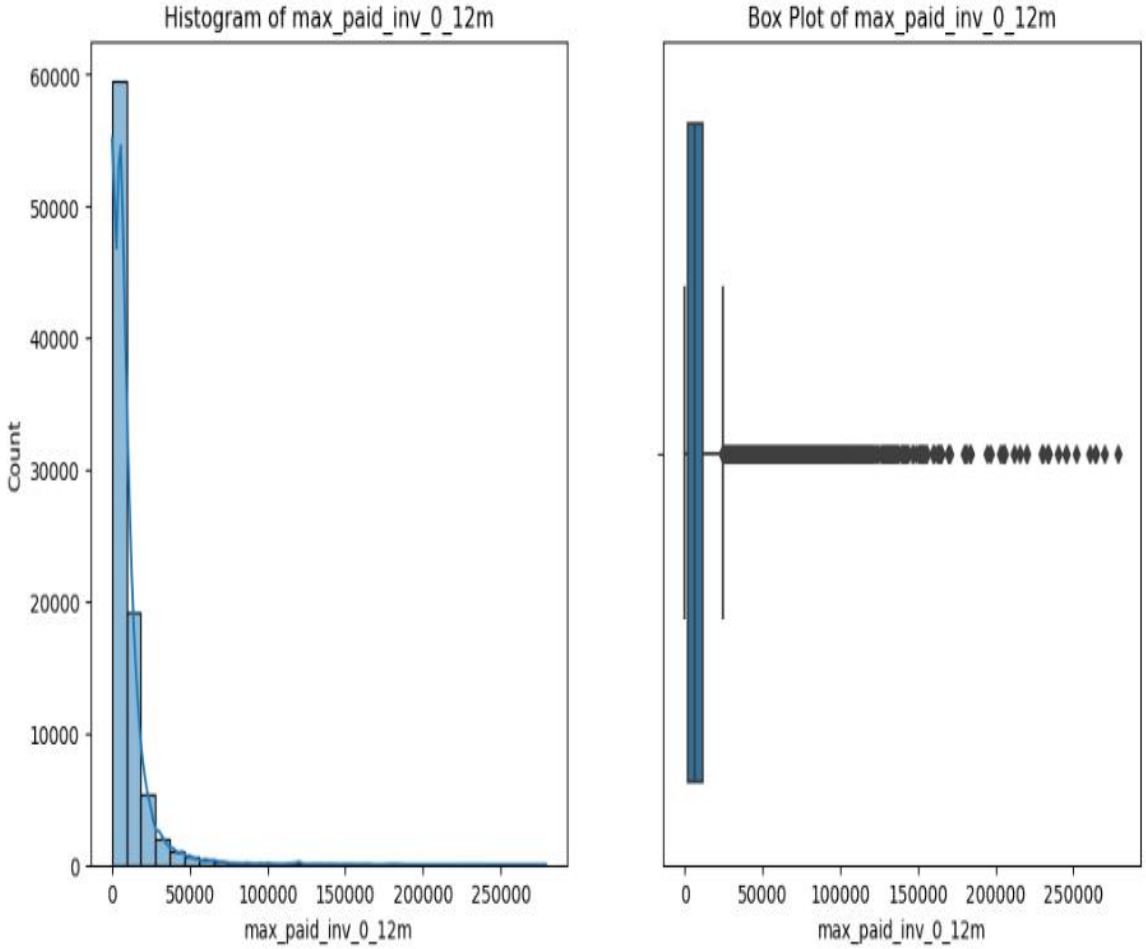
- The even distribution of user IDs confirms a consistent user base over the period the data was collected.

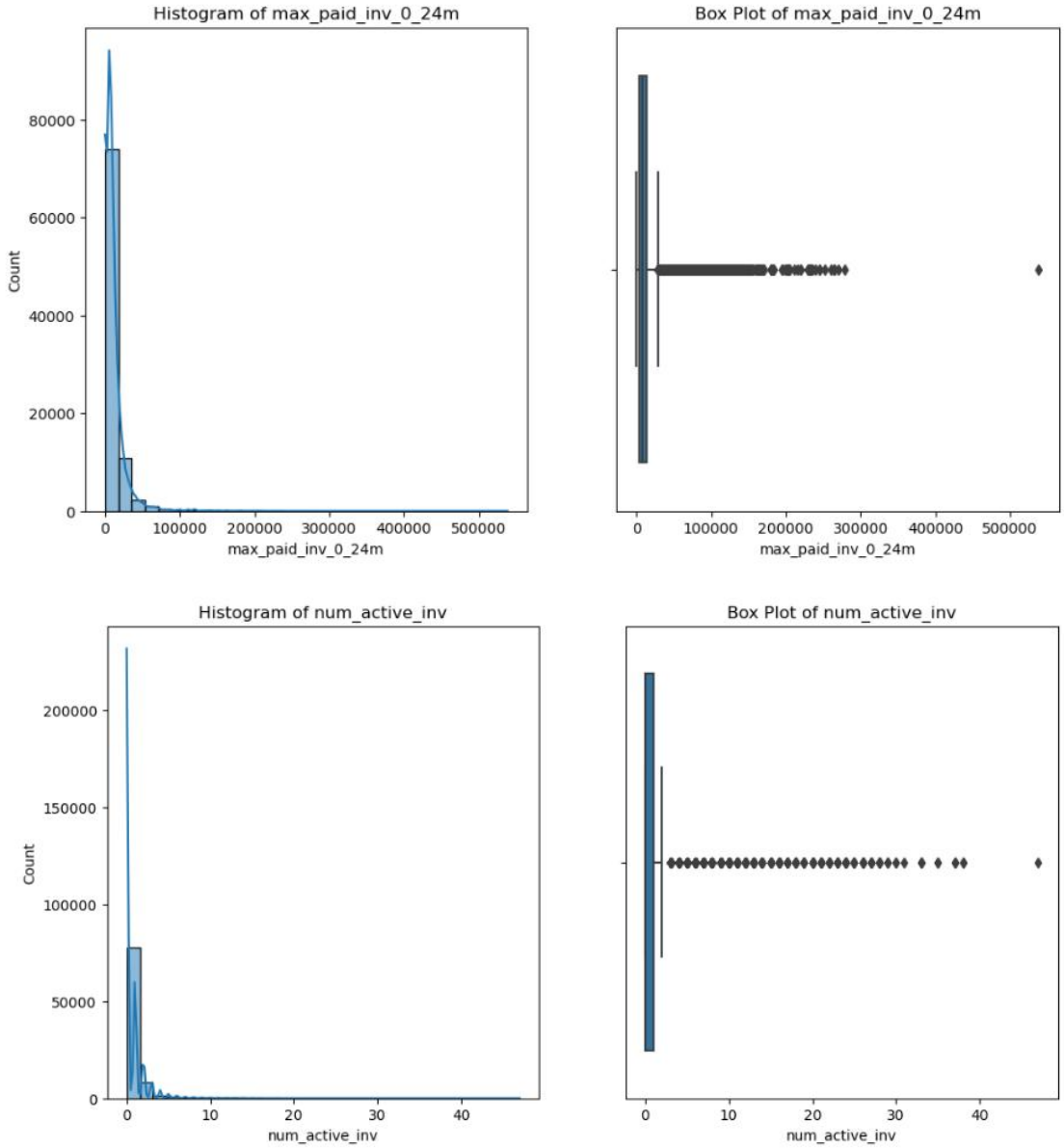
Age Analysis:

- The skew towards younger users indicates that the majority of your credit card customers are younger adults. This could suggest targeting marketing campaigns and products to this age group.
- The presence of outliers in older age groups suggests there is a smaller, yet significant, segment of older users who might have different financial needs or behaviors.

Payment Behavior:

- The highly imbalanced has_paid variable suggests that most users are good at paying their credit card bills. This is a positive sign for credit risk management.
- The small number of users who haven't paid could be a focus for further investigation, possibly leading to tailored interventions or support to improve their payment behavior.





1. max_paid_inv_0_12m

- **Histogram**:
  - This variable shows a right-skewed distribution, with the majority of users having lower maximum paid invoice amounts in the last 12 months.
  - There are a few users with very high values, as indicated by the long tail to the right.
- **Box Plot**:
  - The box plot confirms the right skewness with a large number of outliers on the higher end.
  - The median value is relatively low compared to the range of the data, indicating most users have lower maximum paid invoices.

2. max_paid_inv_0_24m

- **Histogram**:
    - Similar to the 12-month variable, the max_paid_inv_0_24m also shows a right-skewed distribution with most users having lower maximum paid invoice amounts in the last 24 months.
    - There is a noticeable tail extending towards higher values, indicating a few users with high payments.

- **Box Plot**:
    - The box plot also shows a large number of outliers on the higher end, similar to the 12-month maximum paid invoice.
    - The median is low compared to the full range of data.

3. num_active_inv

- **Histogram**:
    - The num_active_inv variable shows a right-skewed distribution with the majority of users having a low number of active invoices.
    - There are a few users with significantly higher numbers of active invoices, as indicated by the tail to the right.

- **Box Plot**:
    - The box plot reveals a high number of outliers, indicating some users have unusually high numbers of active invoices.
    - The median number of active invoices is low, with most data points concentrated at the lower end.

Business Insights

Payment Behavior:

- Both max_paid_inv_0_12m and max_paid_inv_0_24m variables show that the majority of users tend to have lower maximum paid invoices, indicating they might not be using large amounts of credit or have low credit limits.
- The presence of users with high maximum paid invoices could indicate segments with higher credit utilization, which may require different risk management and marketing strategies.

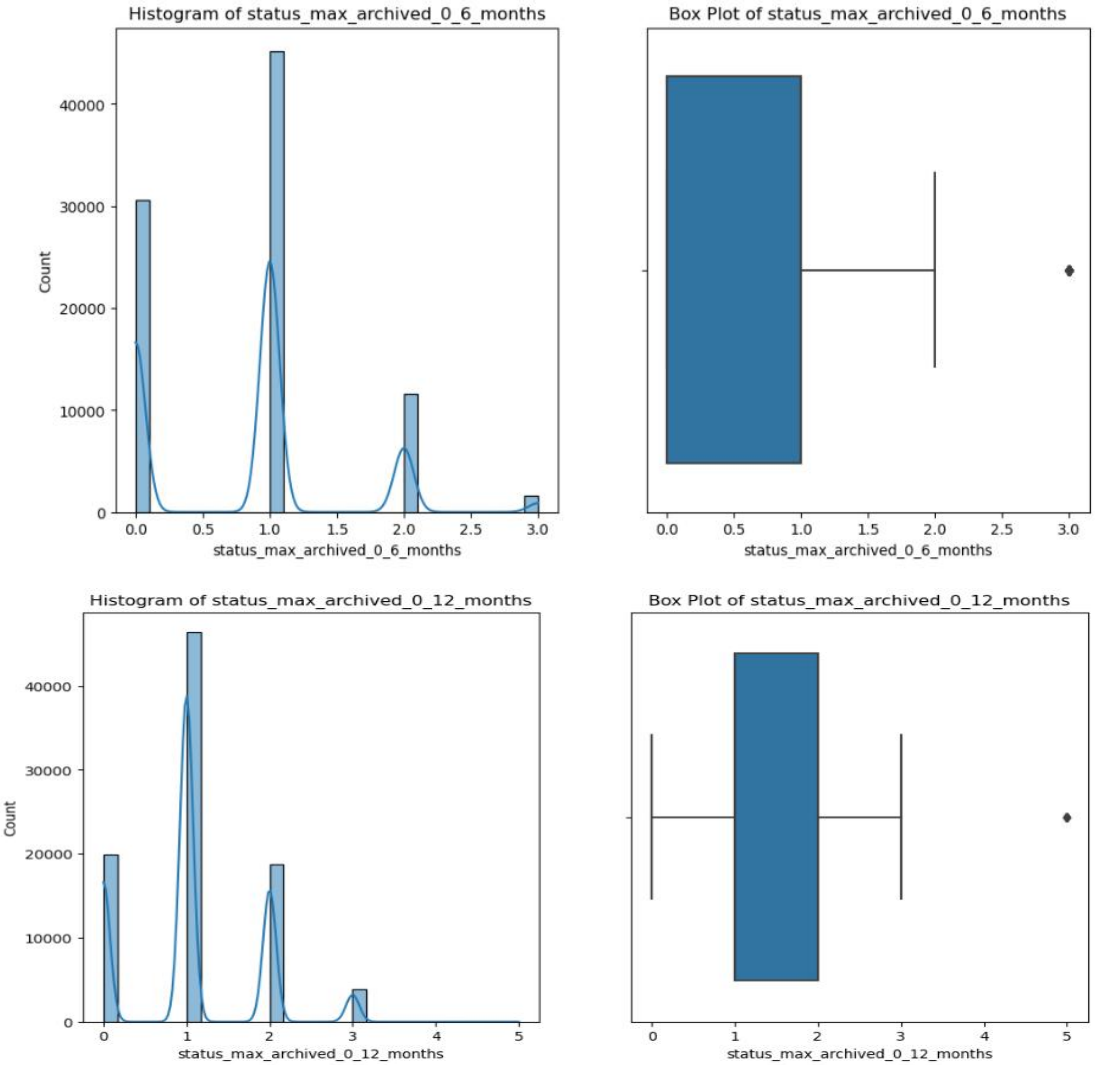
Active Invoices:

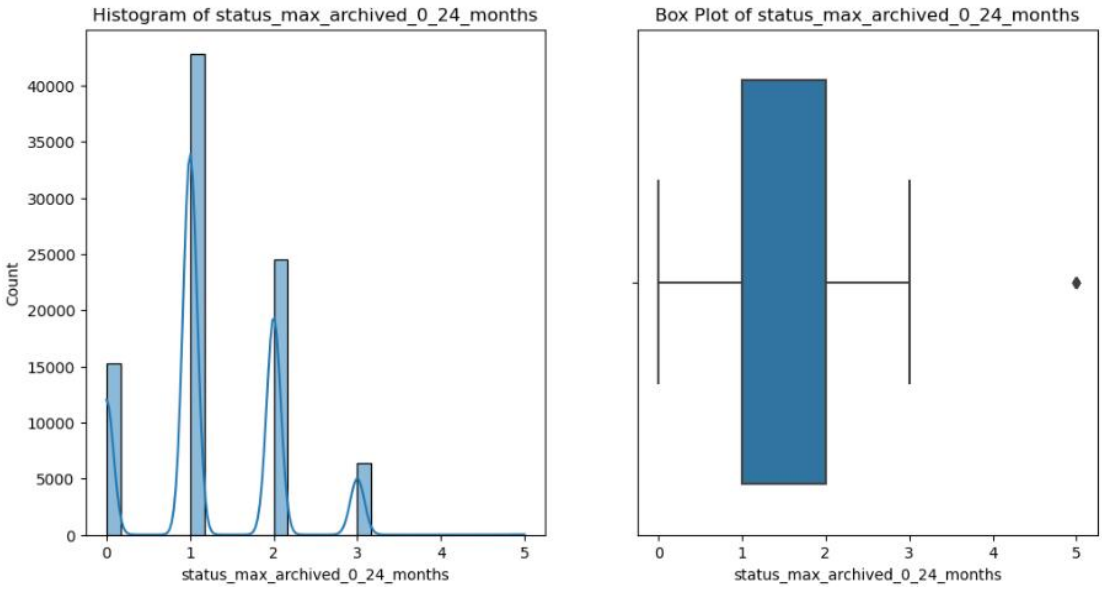
- The num_active_inv variable indicates that most users have a low number of active invoices, which could suggest they manage their credit well and do not accumulate many outstanding invoices.

- Users with a high number of active invoices might be at risk of financial stress, and targeting these users with support or financial management tools could be beneficial.

Outlier Management:

- The large number of outliers in all three variables suggests that there are significant differences in user behavior within the dataset. These outliers should be further analyzed to understand the underlying causes and identify any potential risks or opportunities.





1. status_max_archived_0_6_months

- **Histogram**:
  - The distribution appears to have multiple peaks, indicating that users fall into distinct groups regarding their maximum archived status over the past 6 months.
  - The majority of the data points are clustered around the lower end, with peaks at discrete values, likely representing different status levels.
- **Box Plot**:
  - The box plot shows the data is somewhat spread out with a long upper whisker and a few outliers on the higher end.
  - The median value is relatively low, suggesting most users have lower maximum archived statuses in the past 6 months.

2. status_max_archived_0_12_months

- **Histogram**:
  - Similar to the 6-month variable, this variable also shows multiple peaks, indicating distinct groups.
  - The peaks are again at discrete values, suggesting that these values might represent specific status levels.
- **Box Plot**:
  - The box plot shows a spread similar to the 6-month variable, with a long upper whisker and some outliers.
  - The median is again low, indicating that most users have lower maximum archived statuses over the past 12 months.

3. status_max_archived_0_24_months

- **Histogram**:
    - The distribution for this variable also shows multiple peaks, similar to the other two variables.
    - The peaks at discrete values suggest specific status levels being prominent.

- **Box Plot**:
    - The box plot shows a slightly larger spread compared to the 6-month and 12-month variables, with a longer upper whisker and some outliers.
    - The median is low, indicating that most users have lower maximum archived statuses over the past 24 months.

Business Insights

Status Levels:

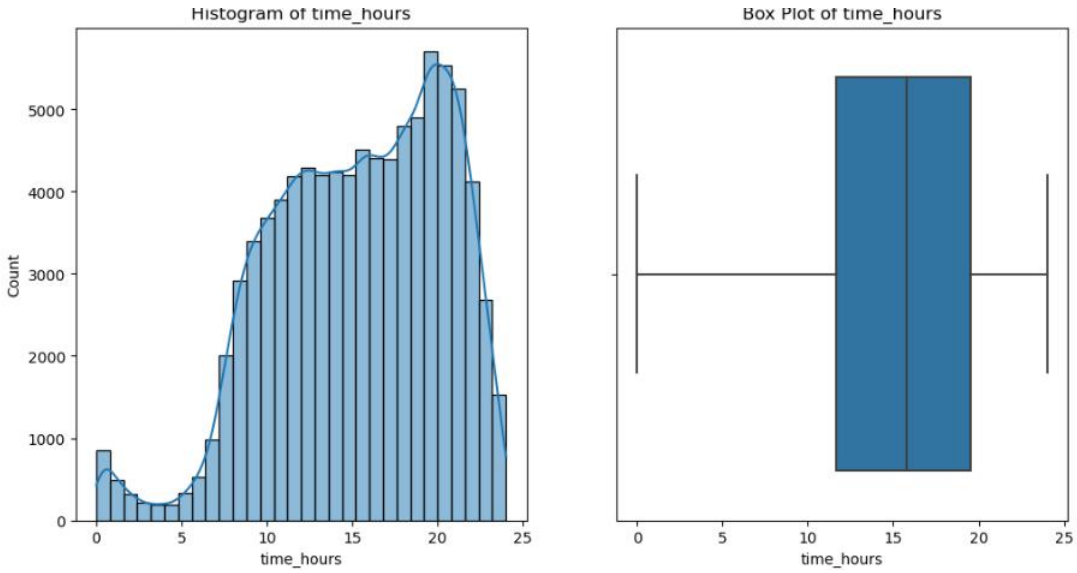
- The distinct peaks in the histograms suggest that users fall into specific categories or status levels. Understanding these levels and what they represent can help in segmenting the users more effectively.
- Most users seem to have lower maximum archived statuses, which might indicate better credit behavior or lower risk levels.

Risk Management:

- The presence of outliers, particularly those with higher status values, could indicate users with higher risk. These users might need closer monitoring or different credit management strategies.
- The spread in the data and the existence of distinct groups can help in designing tailored financial products and services for different user segments.

Temporal Analysis:

- Comparing the status levels over different time periods (6 months, 12 months, and 24 months) can provide insights into the trends and stability of user behavior over time.
- Users with consistently low statuses across all time periods are likely low-risk, whereas those with fluctuating or increasing statuses might require more attention.



· **High Frequency Range**:

- Most users have time_hours values concentrated around 15 to 20 hours. This could be the typical behavior or usage pattern for the majority of users.

· **Secondary Peak**:

- The smaller peak around 5 hours could indicate a subset of users who behave differently or have different usage patterns.

· **Business Implications**:

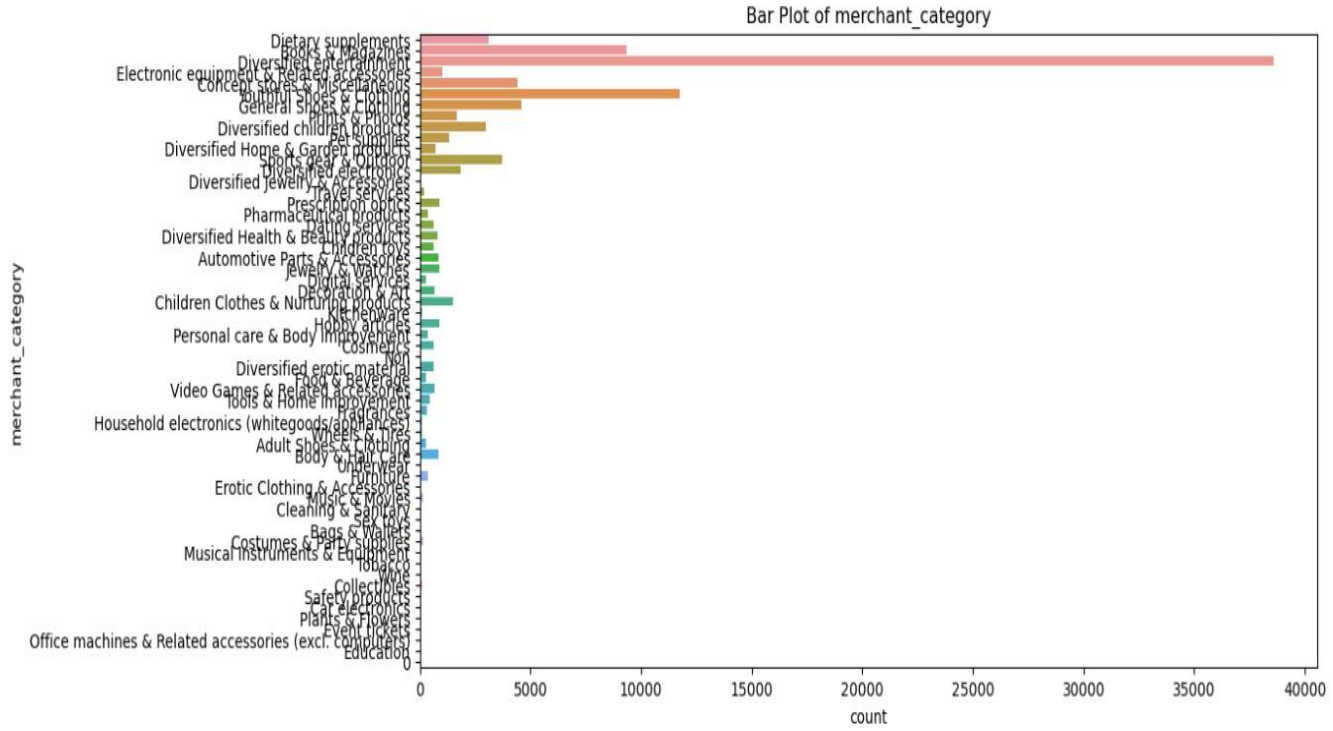
- Understanding the typical time_hours range can help in targeting user segments. For instance, users who fall within the 15-20 hours range might represent standard or average usage, while those around 5 hours might need different marketing or service strategies.
- The lack of significant outliers suggests a homogeneous behavior pattern among users concerning time_hours.

· **Usage Trends**:

- The slight right-skew indicates there are some users with higher time_hours values, but they are less common. This might represent heavy users or those with specific needs.

NOTE:- Here some important histogram and box plot of continuous variable. For further all columns data, please go through the ipynb file.

Univariate analysis for categorical variables:-



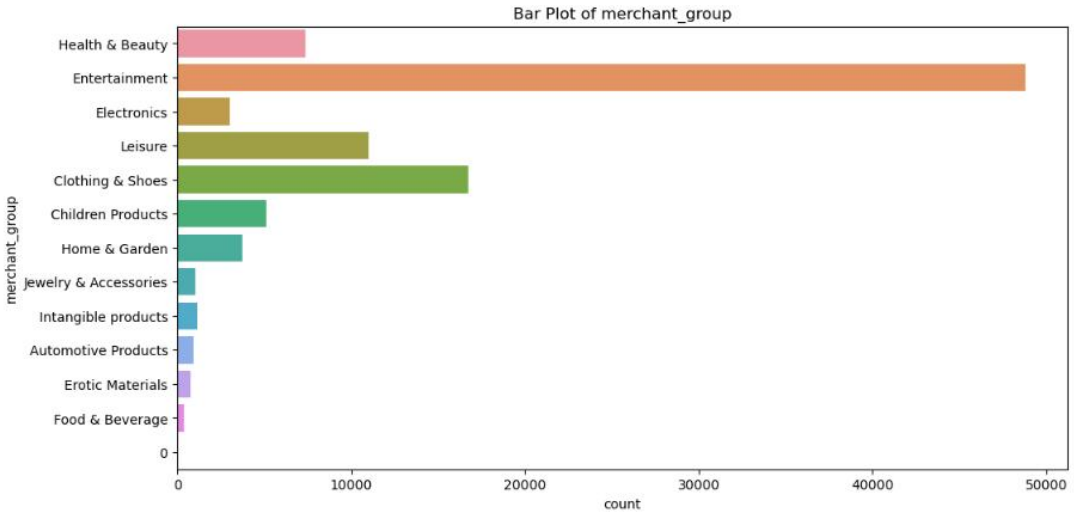
- **Marketing**: Emphasize promotions on Dietary supplements and Electronic equipment categories to maximize impact.
- **Inventory Management**: Ensure high-demand categories are well-stocked to meet consumer demand.
- **Strategic Partnerships**: Consider partnerships with suppliers in top categories to leverage consumer interest.
- **Product Diversification**: Explore opportunities in less frequent categories to tap into niche markets.

Key Categories:

- **Dietary supplements**: This category has the highest count, indicating that a substantial number of transactions fall into this category.
- **Electronic equipment**: This is the second most common category, showing a significant number of transactions.
- **General Shops, Diversified Clothing & Accessories**: These categories also have notable counts, making them significant areas of spending.

Less Frequent Categories:

- Many categories have very low counts, indicating they are less common. Examples include **Office machines & Related accessories**, **Erotic Clothing & Accessories**, and **Adult Shops**.

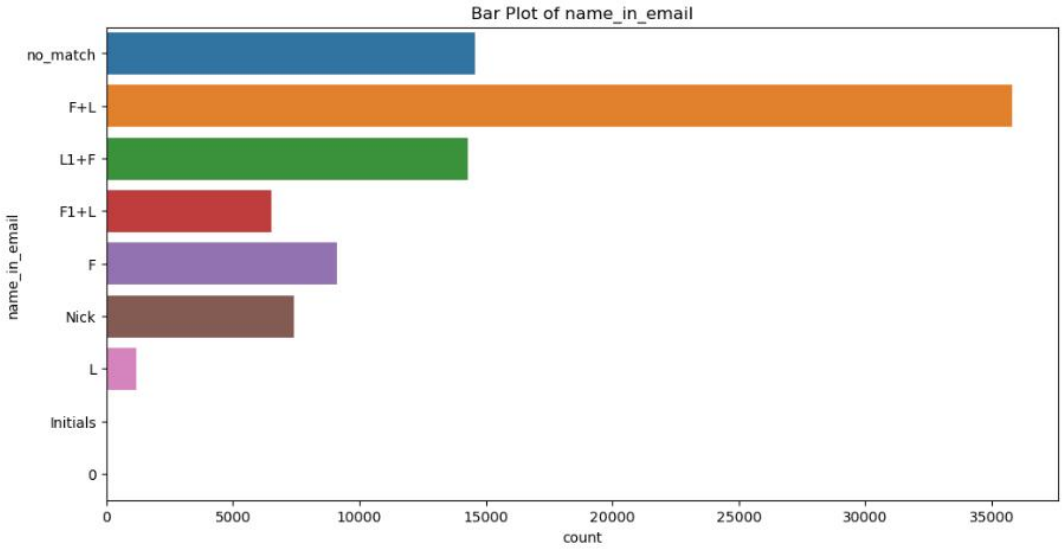


Dominant Groups:

- **Entertainment**: This group has the highest count, indicating a significant number of transactions. This suggests a strong consumer interest or demand in entertainment-related products and services.
- **Electronics** and **Leisure**: These are the next prominent groups, showing substantial consumer spending.

Moderate and Low Count Groups:

- **Health & Beauty**, **Clothing & Shoes**, **Children Products**, and **Home & Garden** have moderate counts, indicating these areas are also significant but not as dominant.
- Categories such as **Jewelry & Accessories**, **Automotive Products**, **Erotic Materials**, and **Food & Beverage** have lower counts, suggesting niche or less frequent transactions.



- **Common Patterns**:
  - **First Name + Last Name (F+L)**: This pattern is the most common, indicating that most users include both their first and last names in their email addresses.
  - **Last Name + First Name (L+F)**: This is also a frequent pattern, showing a substantial number of users prefer this format.

- **Less Common Patterns**:
  - **No Match**: A notable number of users have email addresses that do not match any name pattern, indicating a preference for anonymity or use of unique identifiers.
  - **First Initial + Last Name (F1+L)** and **First Name Only (F)**: These patterns are less common but still significant.
  - **Nicknames (Nick)**, **Last Name Only (L)**, and **Initials**: These are the least common, suggesting that fewer users prefer these formats for their email addresses.

Key Points and Insights

Consumer Preferences in Merchant Groups:

1. **Entertainment and Electronics**: These groups are highly popular among consumers, indicating a strong market presence and demand. Businesses can focus on these areas for marketing campaigns and product development.
2. **Leisure**: This group also shows significant consumer interest, suggesting opportunities for growth and targeted promotions.
3. **Health & Beauty, Clothing & Shoes**: These groups are moderately popular and could benefit from strategic marketing efforts to increase their market share.

Email Patterns and User Identity:

1. **F+L and L+F**: The predominance of these patterns suggests that most users are comfortable using their full names in their email addresses. This can be useful for personalized marketing strategies.
2. **No Match**: The presence of many users with unmatched email patterns indicates a segment of users who may value privacy or use non-standard email formats. Understanding this can help tailor communication strategies.

Business Strategies:

1. **Targeted Marketing**: Focus marketing efforts on the most popular merchant groups (Entertainment and Electronics) to maximize engagement and sales.
2. **Personalization**: Use email pattern data to personalize communications, ensuring relevance and increasing the likelihood of user engagement.

3. **Explore Niche Markets**: For less common merchant groups and email patterns, consider niche marketing strategies to tap into specific customer segments.

Data-Driven Decision Making:

1. Leverage insights from merchant_group distribution to inform inventory management and ensure high-demand products are readily available.
2. Use name_in_email patterns to design personalized user experiences, enhancing customer satisfaction and loyalty.

b) Bivariate analysis (relationship between different variables , correlations)

 # Correlation matrix of continuous variable:-

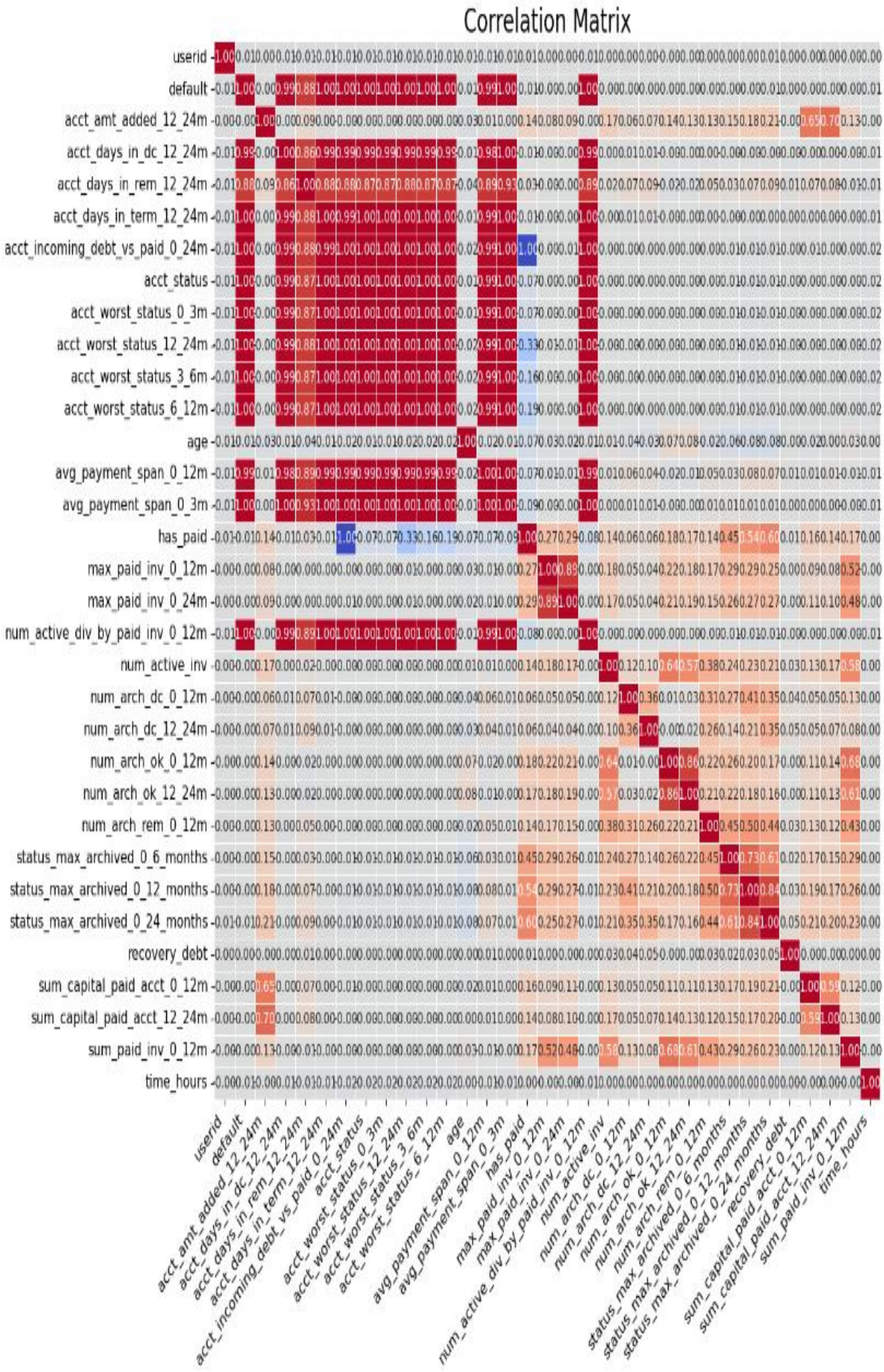
A correlation of 1 means that the two variables are perfectly positively correlated, and a correlation of -1 means that the two variables are perfectly negatively correlated. A correlation of 0 means that there is no correlation between the two variables.

Here are some of the key points from the correlation matrix:

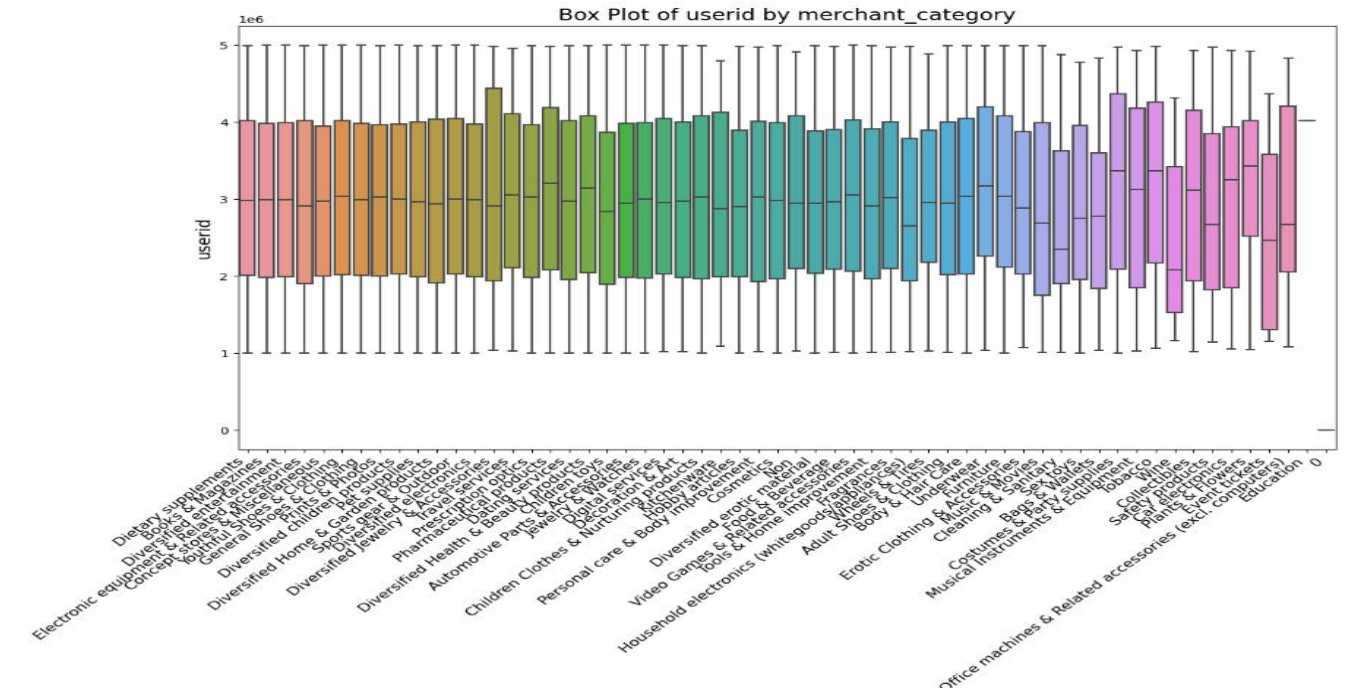
- There is a strong positive correlation between the following variables:
  - acct_days_in_rem_12_24m and num_arch_rem_0_12m
  - status_max_archived_0_12_months, status_max_archived_0_6_months
  - sum_paid_inv_0_12m and num_arch_ok_12_24m
  - recovery_debt and sum_capital_paid_acct_12_24m (This is interesting because it suggests that customers who have a high recovery debt are also the ones who have paid the most capital on their accounts. This could be because these customers are trying to pay off their debts as quickly as possible.)
- There is a strong negative correlation between the following variables:
  - acct_days_in_term_12_24m and num_arch_ok_12_24m (This suggests that customers who have accounts that are in term for a longer period of time are less likely to have their accounts archived.)
  - avg_payment_span_0_12m and has_paid (This suggests that customers who have a longer average payment span are less likely to have made a payment in the last 24 months.)

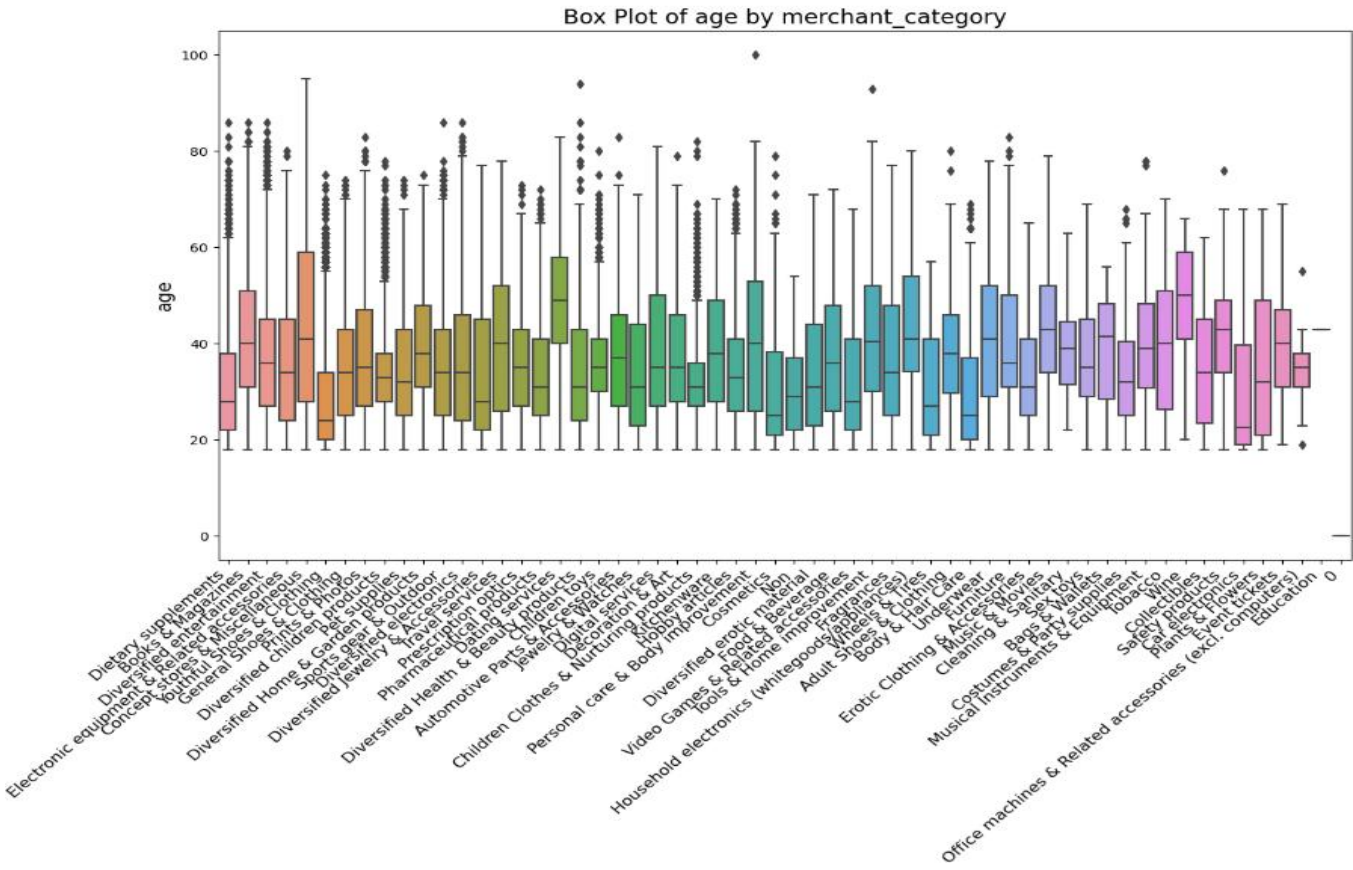
It is important to note that correlation does not equal causation. Just because two variables are correlated does not mean that one variable causes the other variable.

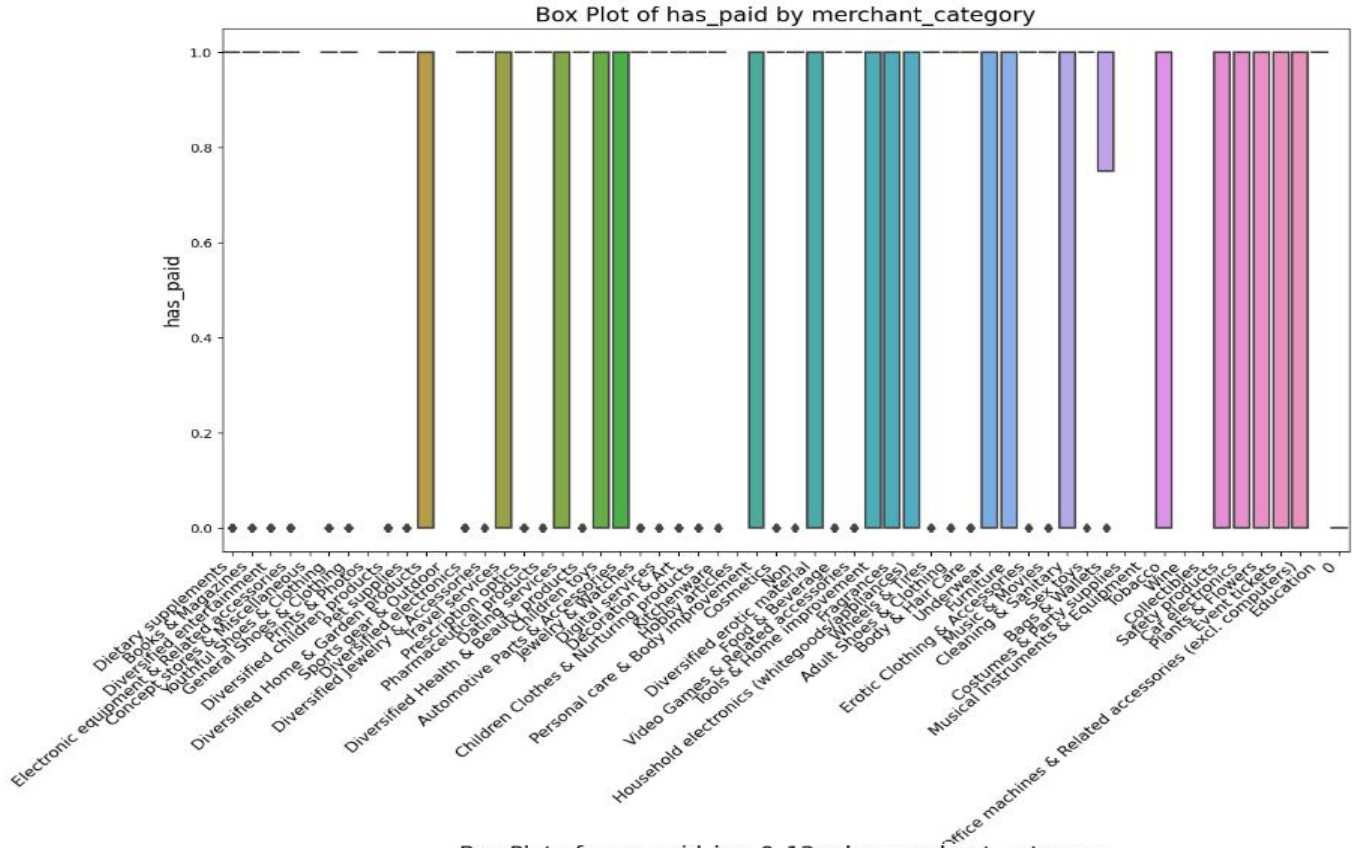
However, correlation can be a useful tool for identifying relationships between variables that can be further investigated.

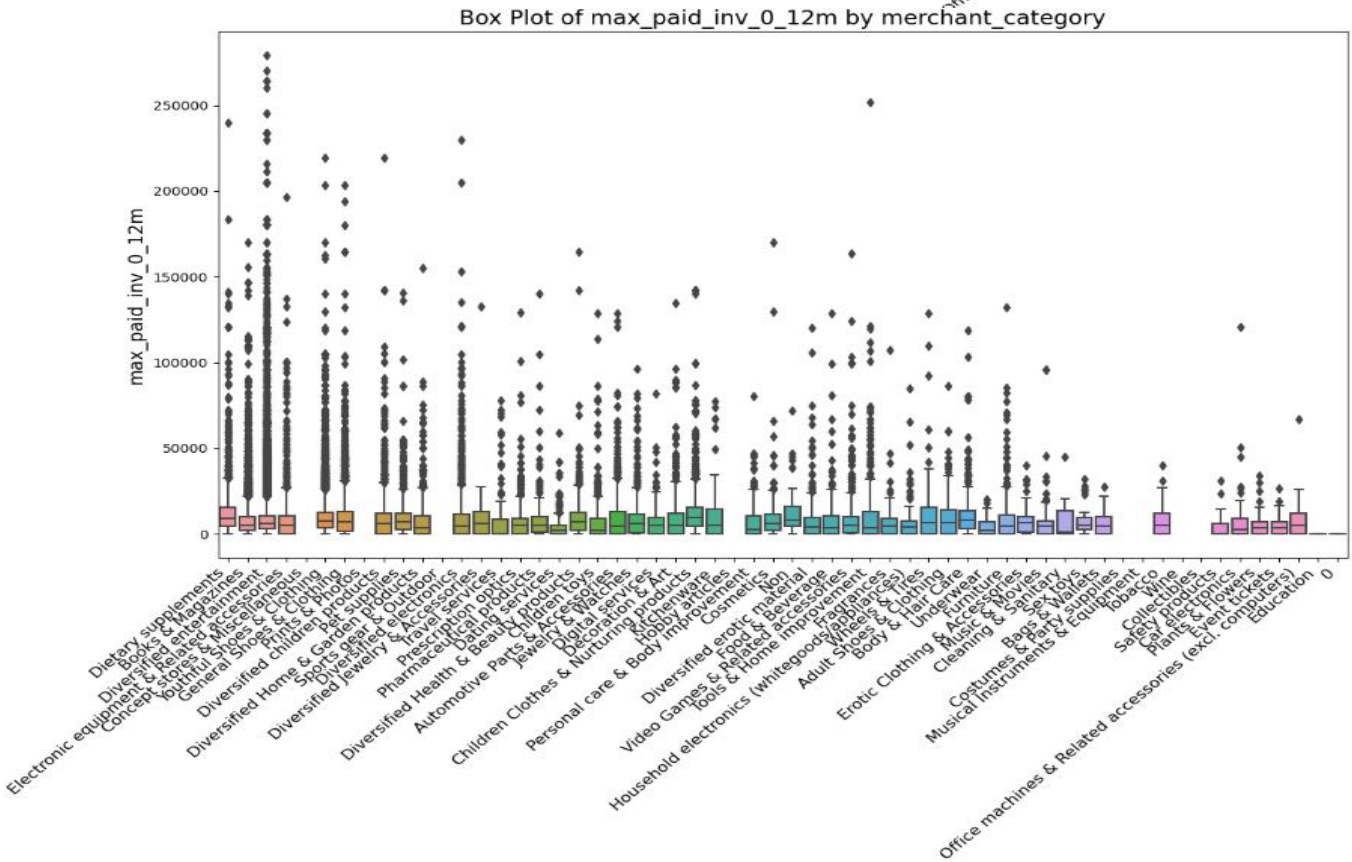

Correlation Matrix
userid
default
acct_amt_added_12_24m
acct_days_in_dc_12_24m
acct_days_in_rem_12_24m
acct_days_in_term_12_24m
acct_incoming_debt_vs_paid_0_24m
acct_status
acct_worst_status_0_3m
acct_worst_status_12_24m
acct_worst_status_3_6m
acct_worst_status_6_12m
age
avg_payment_span_0_12m
avg_payment_span_0_3m
has_paid
max_paid_inv_0_12m
max_paid_inv_0_24m
num_active_div_by_paid_inv_0_12m
num_active_inv
num_arch_dc_0_12m
num_arch_dc_12_24m
num_arch_ok_0_12m
num_arch_ok_12_24m
num_arch_rem_0_12m
status_max_archived_0_6_months
status_max_archived_0_12_months
status_max_archived_0_24_months
recovery_debt
sum_capital_paid_acct_0_12m
sum_capital_paid_acct_12_24m
sum_paid_inv_0_12m
time_hours
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

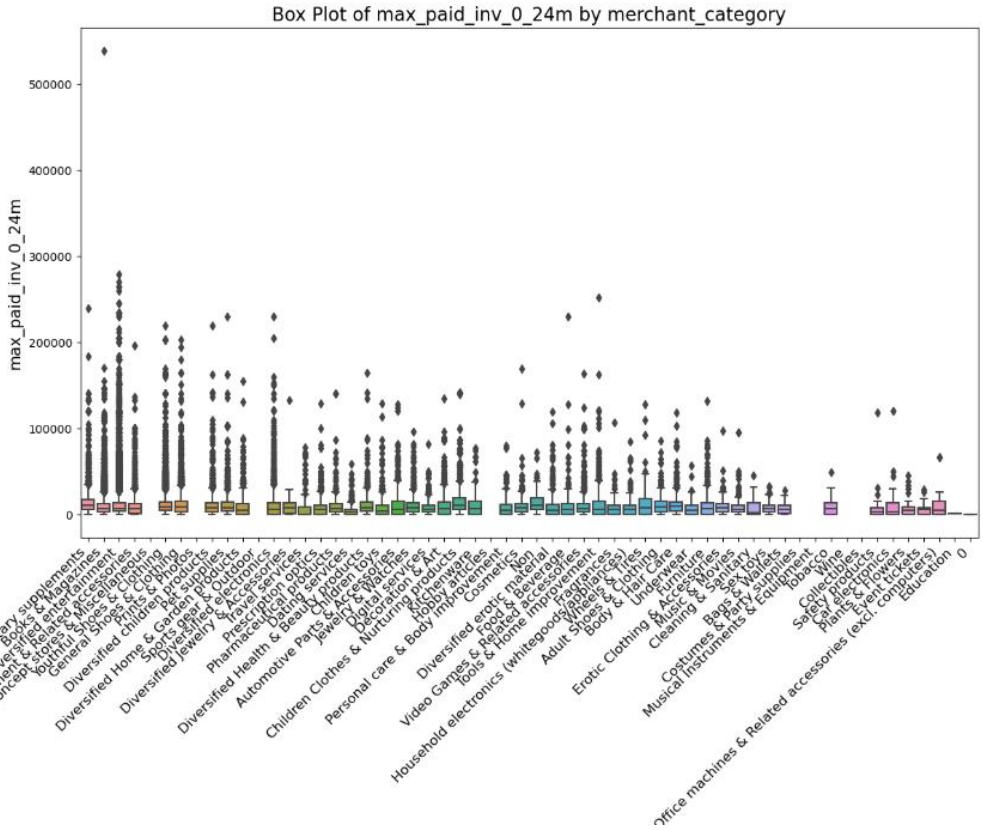
 # Bivariate analysis: Continuous vs Categorical

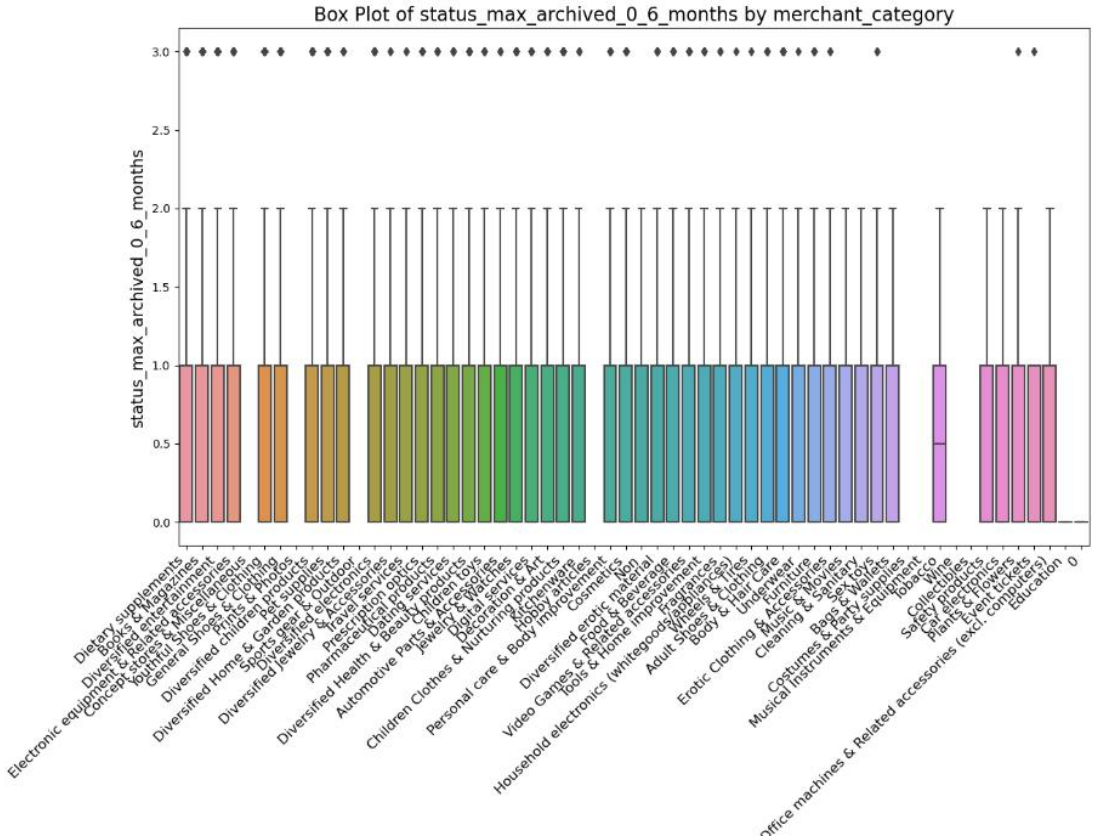


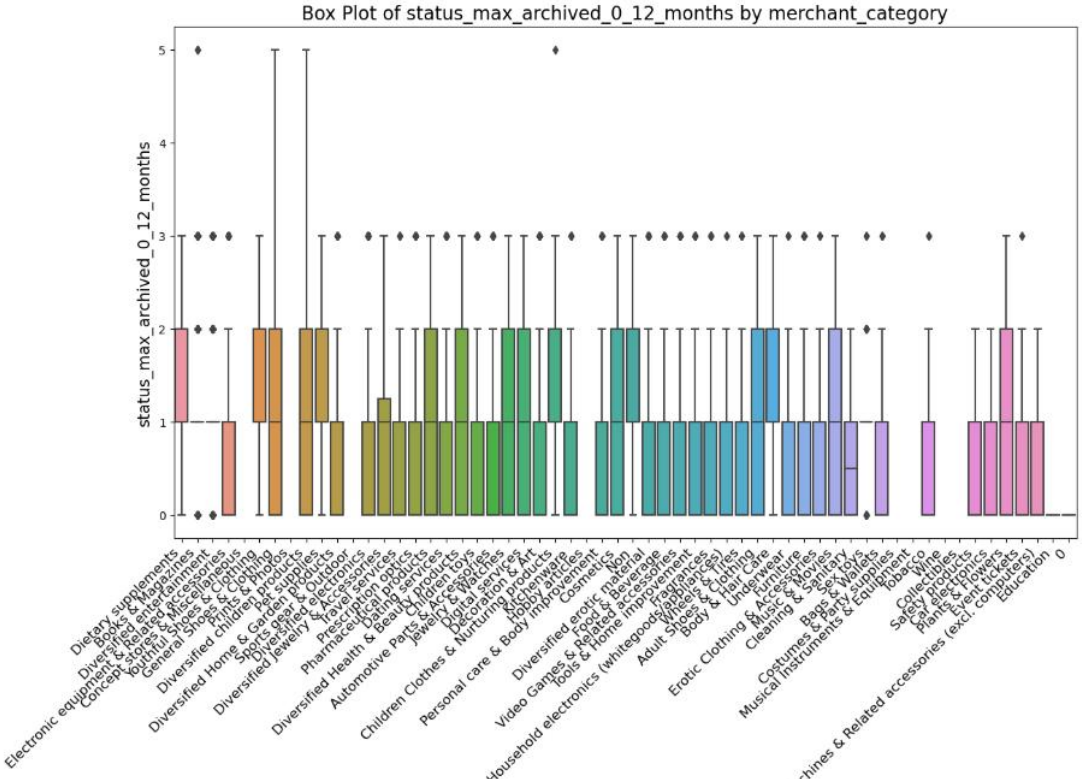


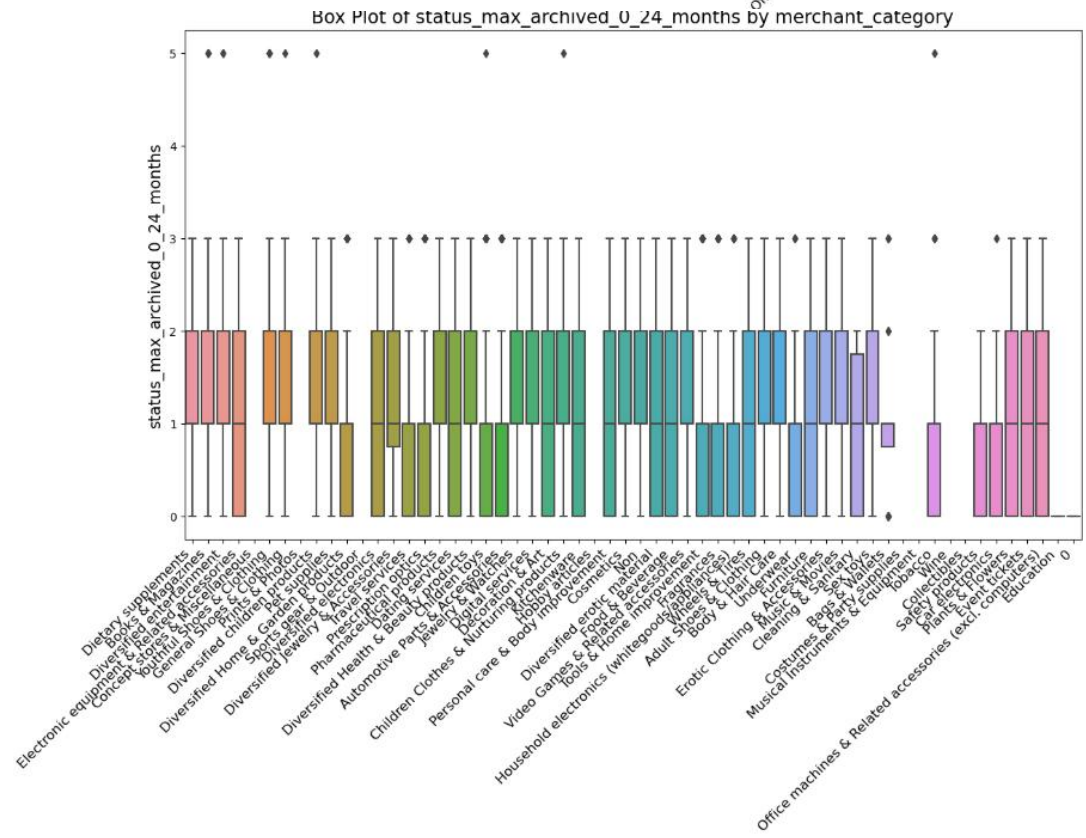

Box Plot of has_paid by merchant_category
has_paid
1.0
0.8
0.6
0.4
0.2
0.0
Dietary supplements
Books & Magazines
Diversified entertainment
Electronic equipment & Related accessories
Concept stores & Miscellaneous
Youthful Shoes & Clothing
General Shoes & Clothing
Prints & photos
Diversified children products
Pet supplies
Diversified Home & Garden products
Sports gear & Outdoor
Diversified electronics
Diversified Jewelry & Accessories
Travel services
Prescription optics
Pharmaceutical products
Dating services
Diversified Health & Beauty products
Automotive Parts & Accessories
Jewelry & Watches
Children toys
Digital services
Decoration & Art
Nurturing products
Kitchenware
Hobby articles
Children Clothes & Nurturing products
Cosmetics
Non
Personal care & Body improvement
Diversified erotic material
Food & Beverage
Video Games & Related accessories
Household electronics (whitegoods & appliances)
Tools & Home improvement
Fragrances
Wheels & Tires
Adult Shoes & Clothing
Body & Hair Care
Underwear
Erotic Clothing & Accessories
Furniture
Music & Movies
Cleaning & Sanitary
Sex toys
Bags & Wallets
Costumes & Party supplies
Musical Instruments & Equipment
Tobacco
Wine
Collectibles
Safety products
Car electronics
Plants & Flowers
Event tickets
Office machines & Related accessories (excl. computers)
Education
0

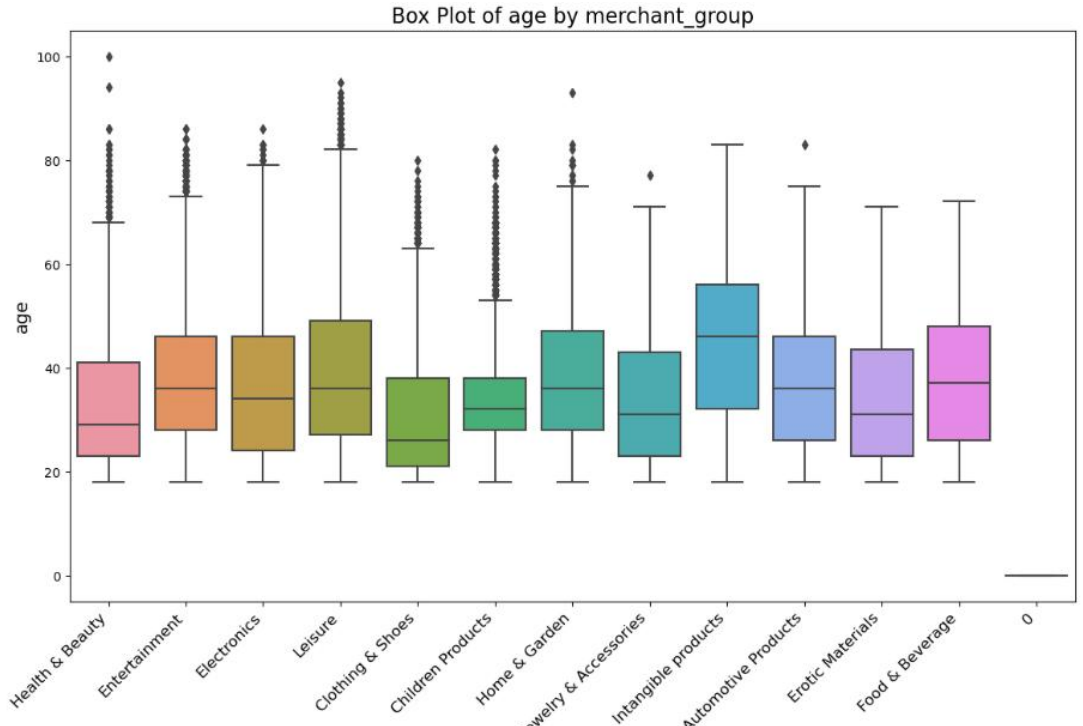

Box Plot of max_paid_inv_0_12m by merchant_category
max_paid_inv_0_12m
250000
200000
150000
100000
50000
0
Dietary supplements
Books & Magazines
Diversified entertainment
Electronic equipment & Related accessories
Concept stores & Miscellaneous
Youthful Shoes & Clothing
General Shoes & Clothing
Prints & photos
Diversified children products
Pet supplies
Diversified Home & Garden products
Sports gear & Outdoor
Diversified electronics
Diversified Jewelry & Accessories
Travel services
Prescription optics
Pharmaceutical products
Dating services
Diversified Health & Beauty products
Automotive Parts & Accessories
Jewelry & Watches
Children toys
Digital services
Decoration & Art
Nurturing products
Kitchenware
Hobby articles
Children Clothes & Nurturing products
Cosmetics
Non
Personal care & Body improvement
Diversified erotic material
Food & Beverage
Video Games & Related accessories
Household electronics (whitegoods & appliances)
Tools & Home improvement
Fragrances
Wheels & Tires
Adult Shoes & Clothing
Body & Hair Care
Underwear
Erotic Clothing & Accessories
Furniture
Music & Movies
Cleaning & Sanitary
Sex toys
Bags & Wallets
Costumes & Party supplies
Musical Instruments & Equipment
Tobacco
Wine
Collectibles
Safety products
Car electronics
Plants & Flowers
Event tickets
Office machines & Related accessories (excl. computers)
Education
0

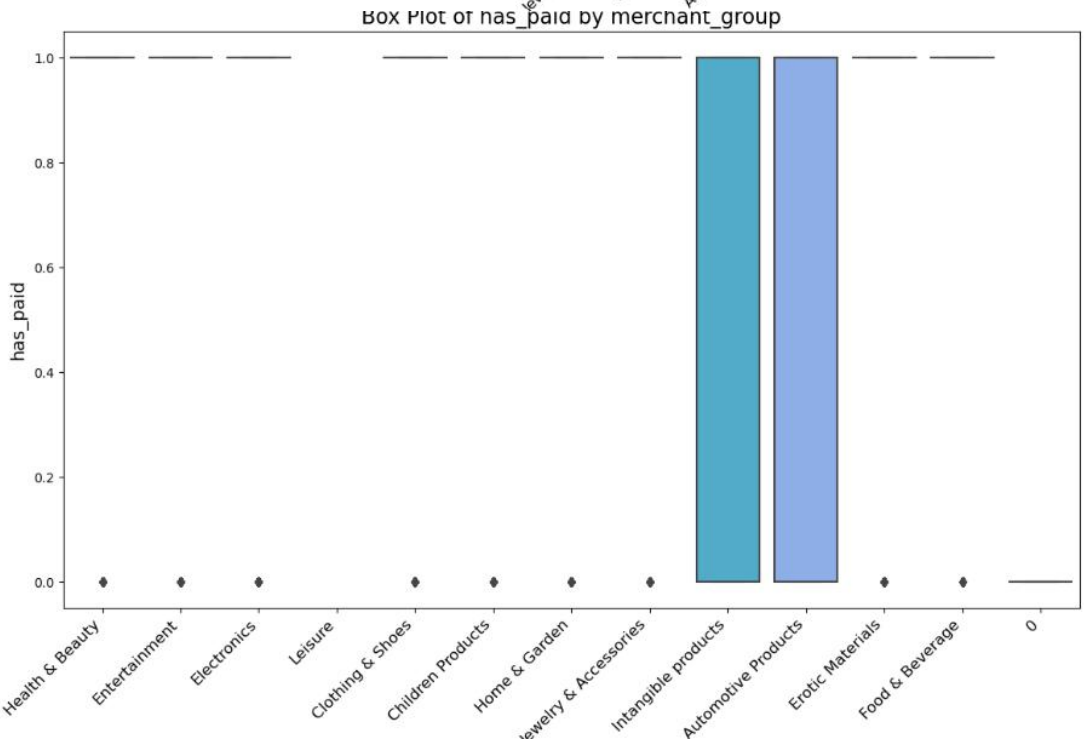

Box Plot of max_paid_inv_0_24m by merchant_category
max_paid_inv_0_24m
500000
400000
300000
200000
100000
0

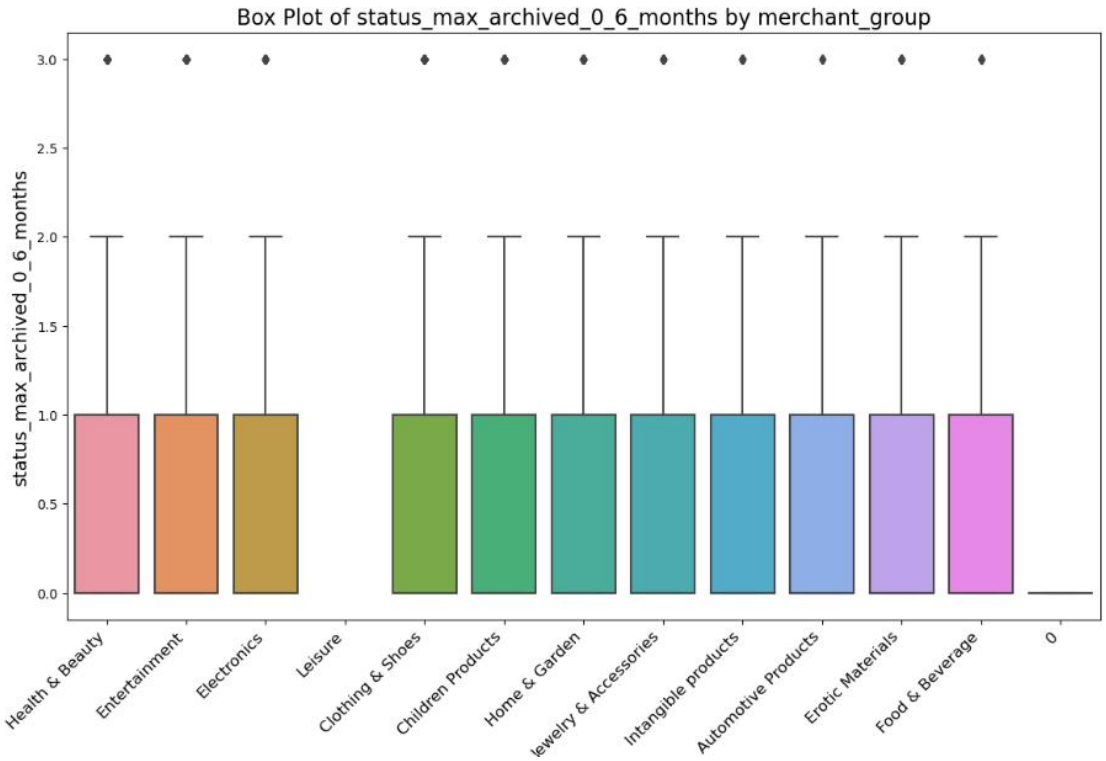

Box Plot of status_max_archived_0_6_months by merchant_category
status_max_archived_0_6_months
3.0
2.5
2.0
1.5
1.0
0.5
0.0

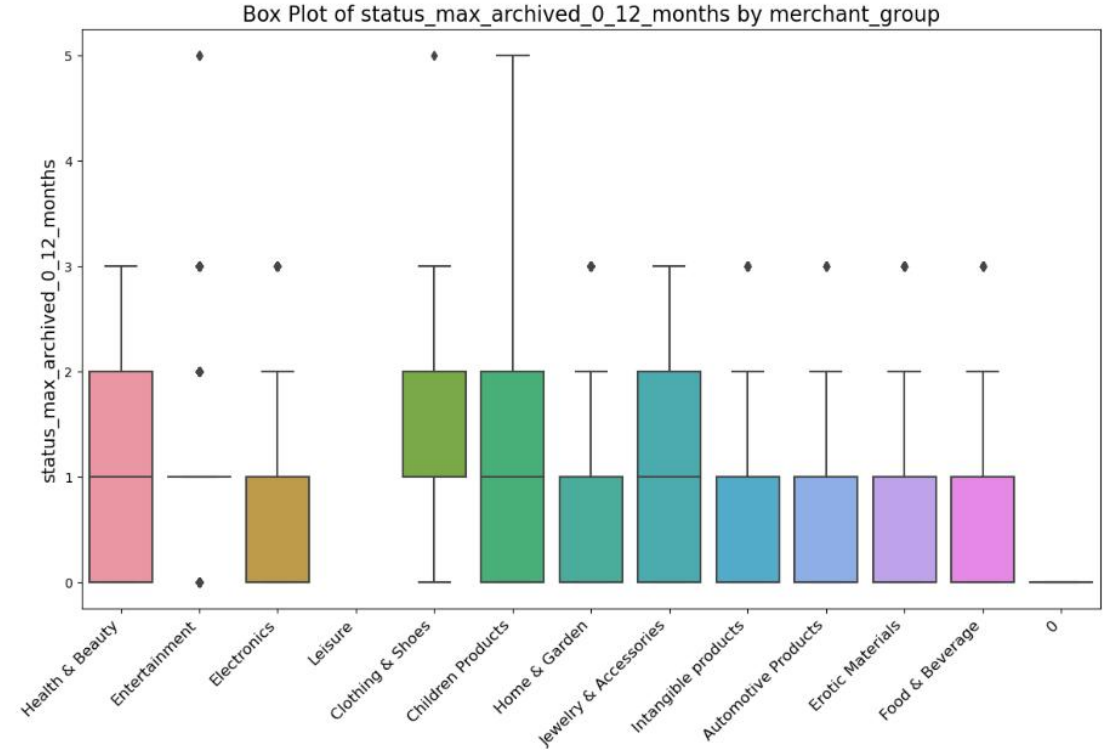

Box Plot of status_max_archived_0_12_months by merchant_category
status_max_archived_0_12_months
0
1
2
3
4
5
Dietary supplements
Books & Magazines
Diversified entertainment
Electronic equipment & Related accessories
Concept stores & Miscellaneous
Youthful Shoes & Clothing
General Shoes & Clothing
Prints & Photos
Diversified children products
Diversified Home & Garden products
Sports gear & Outdoor
Diversified Jewelry & Accessories
Diversified electronics
Travel services
Prescription optics
Dating services
Pharmaceutical products
Diversified Health & Beauty products
Children toys
Automotive Parts & Accessories
Jewelry & Watches
Digital services
Decoration & Art
Kitchenware
Hobby articles
Personal care & Body improvement
Children Clothes & Nurturing products
Cosmetics
Diversified erotic material
Food & Beverage
Video Games & Related accessories
Household electronics (whitegoods/appliances)
Tools & Home improvement
Fragrances
Wheels & Tires
Adult Shoes & Clothing
Body & Hair Care
Underwear
Furniture
Erotic Clothing & Accessories
Music & Movies
Cleaning & Sanitary
Bags & Wallets
Sex Toys
Party supplies
Costumes & Party Supplies
Musical Instruments & Equipment
Tobacco
Wine
Collectibles
Safety products
Pets & Electronics
Plants & Flowers
Event tickets
Office machines & Related accessories (excl. computers)
Education
0

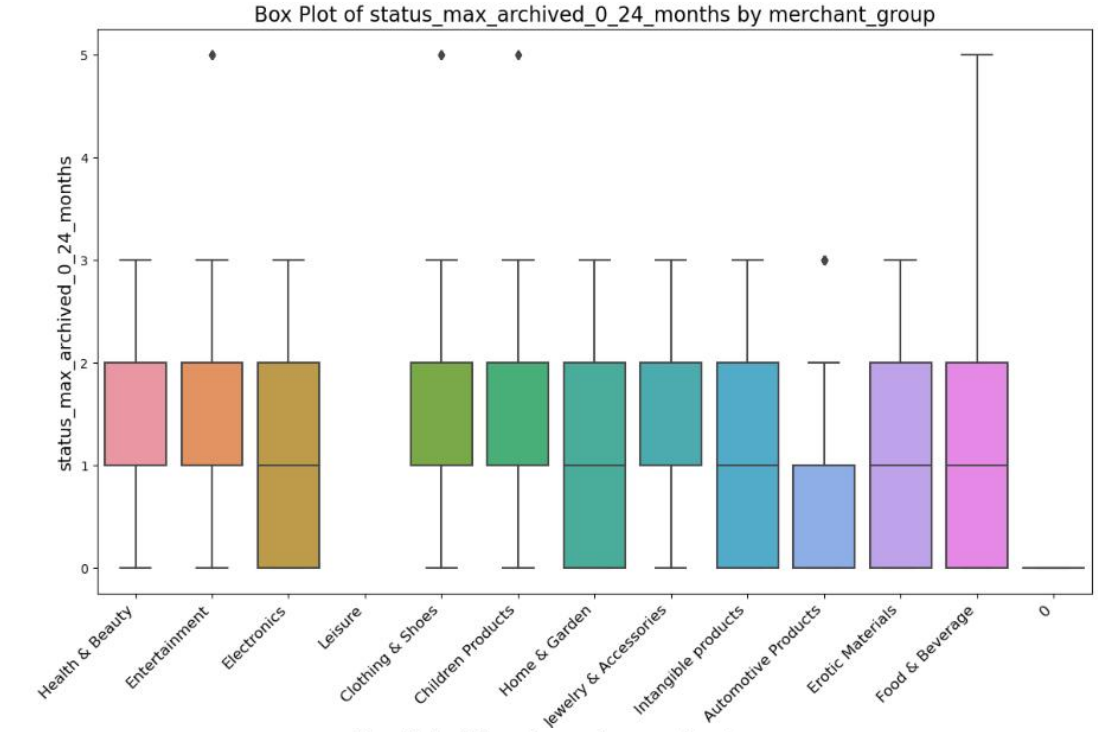

Box Plot of status_max_archived_0_24_months by merchant_category
status_max_archived_0_24_months
0
1
2
3
4
5
Dietary supplements
Books & Magazines
Diversified entertainment
Electronic equipment & Related accessories
Concept stores & Miscellaneous
Youthful Shoes & Clothing
General Shoes & Clothing
Prints & Photos
Diversified children products
Diversified Home & Garden products
Sports gear & Outdoor
Diversified Jewelry & Accessories
Diversified electronics
Travel services
Prescription optics
Dating services
Pharmaceutical products
Diversified Health & Beauty products
Children toys
Automotive Parts & Accessories
Jewelry & Watches
Digital services
Decoration & Art
Kitchenware
Hobby articles
Personal care & Body improvement
Children Clothes & Nurturing products
Cosmetics
Diversified erotic material
Food & Beverage
Video Games & Related accessories
Household electronics (whitegoods/appliances)
Tools & Home improvement
Fragrances
Wheels & Tires
Adult Shoes & Clothing
Body & Hair Care
Underwear
Furniture
Erotic Clothing & Accessories
Music & Movies
Cleaning & Sanitary
Bags & Wallets
Sex Toys
Party supplies
Costumes & Party Supplies
Musical Instruments & Equipment
Tobacco
Wine
Collectibles
Safety products
Pets & Electronics
Plants & Flowers
Event tickets
Office machines & Related accessories (excl. computers)
Education
0

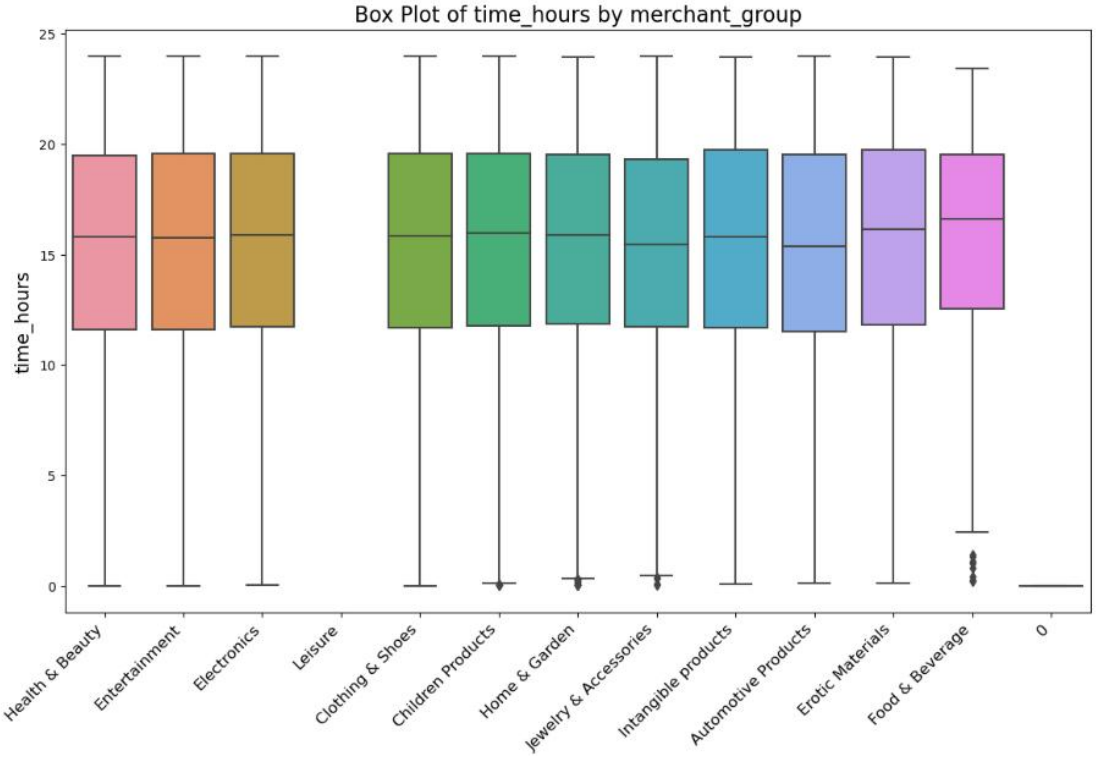

Box Plot of age by merchant_group
age
100
80
60
40
20
0
Health & Beauty
Entertainment
Electronics
Leisure
Clothing & Shoes
Children Products
Home & Garden
Jewelry & Accessories
Intangible products
Automotive Products
Erotic Materials
Food & Beverage
0

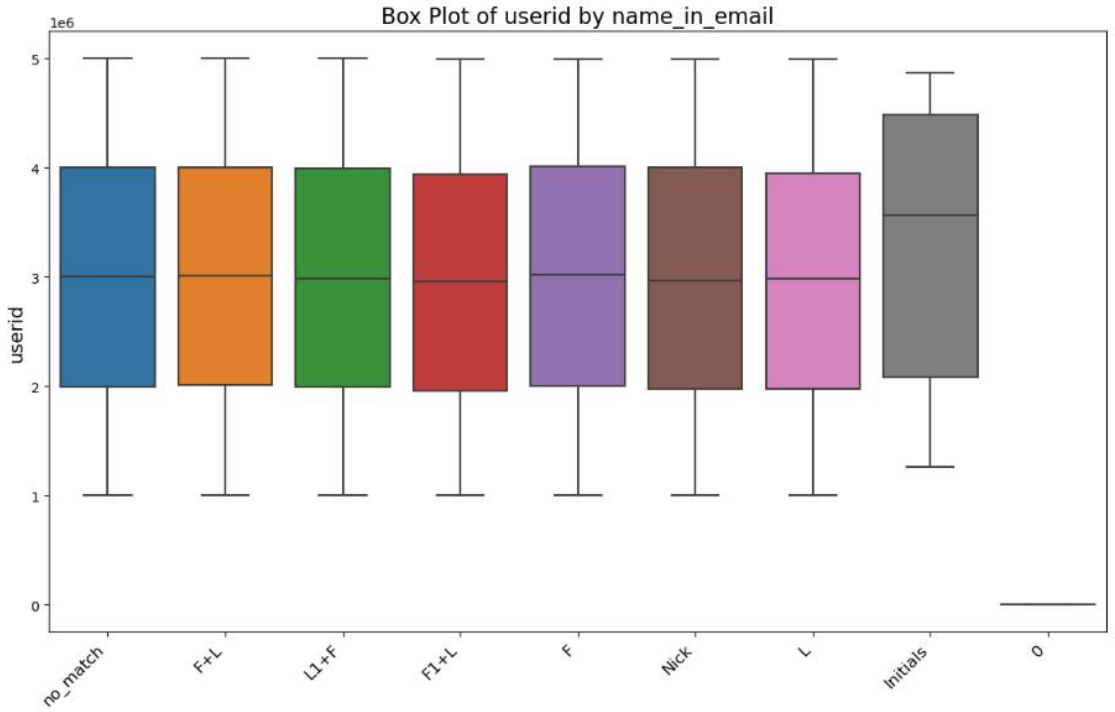

Box Plot of has_paid by merchant_group
has_paid
1.0
0.8
0.6
0.4
0.2
0.0
Health & Beauty
Entertainment
Electronics
Leisure
Clothing & Shoes
Children Products
Home & Garden
Jewelry & Accessories
Intangible products
Automotive Products
Erotic Materials
Food & Beverage
0

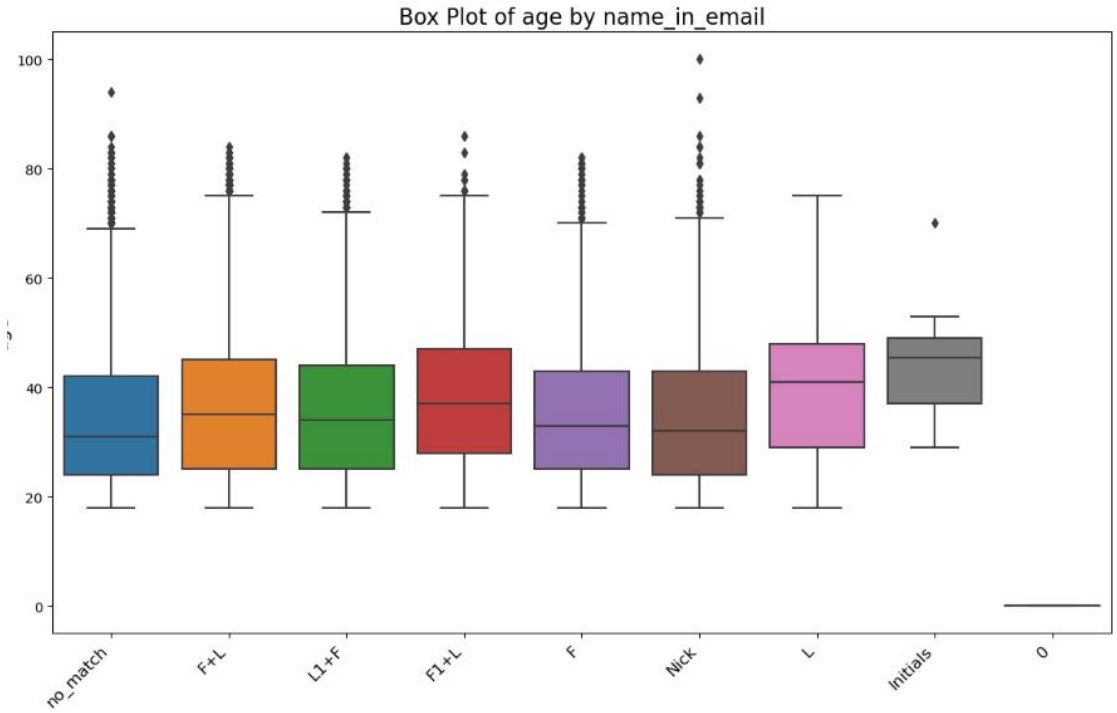

Box Plot of status_max_archived_0_6_months by merchant_group
status_max_archived_0_6_months
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Health & Beauty
Entertainment
Electronics
Leisure
Clothing & Shoes
Children Products
Home & Garden
Jewelry & Accessories
Intangible products
Automotive Products
Erotic Materials
Food & Beverage
0


Box Plot of status_max_archived_0_12_months by merchant_group
status_max_archived_0_12_months
5
4
3
2
1
0
Health & Beauty
Entertainment
Electronics
Leisure
Clothing & Shoes
Children Products
Home & Garden
Jewelry & Accessories
Intangible products
Automotive Products
Erotic Materials
Food & Beverage
0


Box Plot of status_max_archived_0_24_months by merchant_group
status_max_archived_0_24_months
0
1
2
3
4
5
Health & Beauty
Entertainment
Electronics
Leisure
Clothing & Shoes
Children Products
Home & Garden
Jewelry & Accessories
Intangible products
Automotive Products
Erotic Materials
Food & Beverage
0


Box Plot of time_hours by merchant_group
time_hours
0
5
10
15
20
25
Health & Beauty
Entertainment
Electronics
Leisure
Clothing & Shoes
Children Products
Home & Garden
Jewelry & Accessories
Intangible products
Automotive Products
Erotic Materials
Food & Beverage
0





NOTE:- Please go through the .ipynb file for all columns graph plot

c) Missing Value treatment:-

· **Separating Features:** The data is first divided into two categories:

- **Numerical Columns:** These contain numerical data types like integers (int64) and floating-point numbers (float64).
- **Categorical Columns:** These contain data types like strings (object) and categorical variables.

· **Imputation Methods:**

- **Mean Imputation (Numerical):** Missing values in numerical columns are replaced with the average (mean) value of that specific column. Three separate datasets are created, each using a different imputation method.
- **Median Imputation (Numerical):** Similar to mean imputation, but missing values are filled with the median value of the corresponding numerical column.
- **Mode Imputation (Categorical):** Missing values in categorical columns are replaced with the most frequent value (mode) found in that column.

· **Filling Techniques:**

- **Forward Fill:** This approach fills missing values by carrying the last valid observation forward. Separate datasets are created for both numerical and categorical columns using this method.
- **Backward Fill:** Similar to forward fill, but missing values are filled by carrying the next valid observation backward. Separate datasets are created for both data types using backward fill.

· **Handling Missing Values:**

- **Dropping Rows:** This method removes entire rows containing missing values. A new dataset is created excluding these rows.
- **Dropping Columns:** Here, entire columns containing missing values are removed. A new dataset is created excluding these columns.

```
Missing Values per Column:
 userid                                2
default                            10002
acct_amt_added_12_24m                 2
acct_days_in_dc_12_24m            11838
acct_days_in_rem_12_24m           11838
acct_days_in_term_12_24m          11838
acct_incoming_debt_vs_paid_0_24m  59317
acct_status                       54375
acct_worst_status_0_3m            54375
acct_worst_status_12_24m          66763
acct_worst_status_3_6m            57704
acct_worst_status_6_12m           60352
age                                   2
avg_payment_span_0_12m            23838
avg_payment_span_0_3m             49307
merchant_category                     2
merchant_group                       11
has_paid                          11036
max_paid_inv_0_12m                11036
max_paid_inv_0_24m                11036
name_in_email                     11036
num_active_div_by_paid_inv_0_12m  29927
num_active_inv                    11036
num_arch_dc_0_12m                 11036
num_arch_dc_12_24m                11036
num_arch_ok_0_12m                 11036
num_arch_ok_12_24m                11036
num_arch_rem_0_12m                11036
status_max_archived_0_6_months    11036
status_max_archived_0_12_months   11036
status_max_archived_0_24_months   11036
recovery_debt                     11036
sum_capital_paid_acct_0_12m       11036
sum_capital_paid_acct_12_24m      11036
sum_paid_inv_0_12m                11036
time_hours                        11036
dtype: int64
```

AFTER TREATMENT OF MISSING VALUES:-

```
Missing Values per Column after Backward Fill:
 userid                                 0
default                                 0
acct_amt_added_12_24m                   0
acct_days_in_dc_12_24m                  0
acct_days_in_rem_12_24m                 0
acct_days_in_term_12_24m                0
acct_incoming_debt_vs_paid_0_24m        0
acct_status                             0
acct_worst_status_0_3m                  0
acct_worst_status_12_24m                0
acct_worst_status_3_6m                  0
acct_worst_status_6_12m                 0
age                                     0
avg_payment_span_0_12m                  0
avg_payment_span_0_3m                   0
merchant_category                       0
merchant_group                          0
has_paid                                0
max_paid_inv_0_12m                      0
max_paid_inv_0_24m                      0
name_in_email                           0
num_active_div_by_paid_inv_0_12m        0
num_active_inv                          0
num_arch_dc_0_12m                       0
num_arch_dc_12_24m                      0
num_arch_ok_0_12m                       0
num_arch_ok_12_24m                      0
num_arch_rem_0_12m                      0
status_max_archived_0_6_months          0
status_max_archived_0_12_months         0
status_max_archived_0_24_months         0
recovery_debt                           0
sum_capital_paid_acct_0_12m             0
sum_capital_paid_acct_12_24m            0
sum_paid_inv_0_12m                      0
time_hours                              0
dtype: int64
```

d) Outlier treatment (if required):-

Due to variability in the data or due to measurement errors. Identifying outliers is crucial because they can affect the results of data analysis and statistical modeling.

Detection Methods:

- **Visual Methods**: Box plots, scatter plots, and histograms can help visually identify outliers.
- **Statistical Methods**: Techniques like the Z-score (for data assumed to follow a normal distribution), IQR method, and Grubbs' test are commonly used.

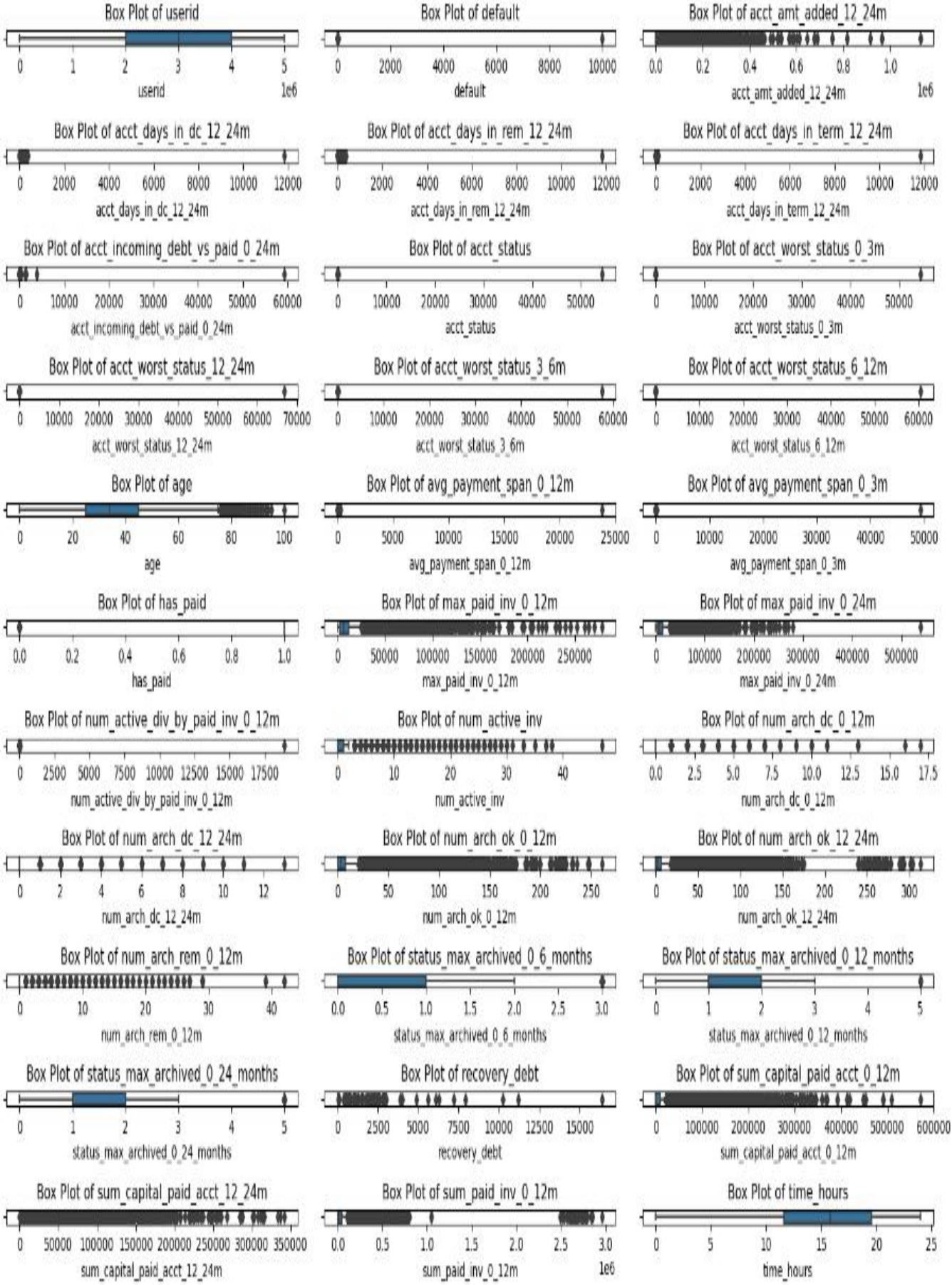
Handling Outliers:

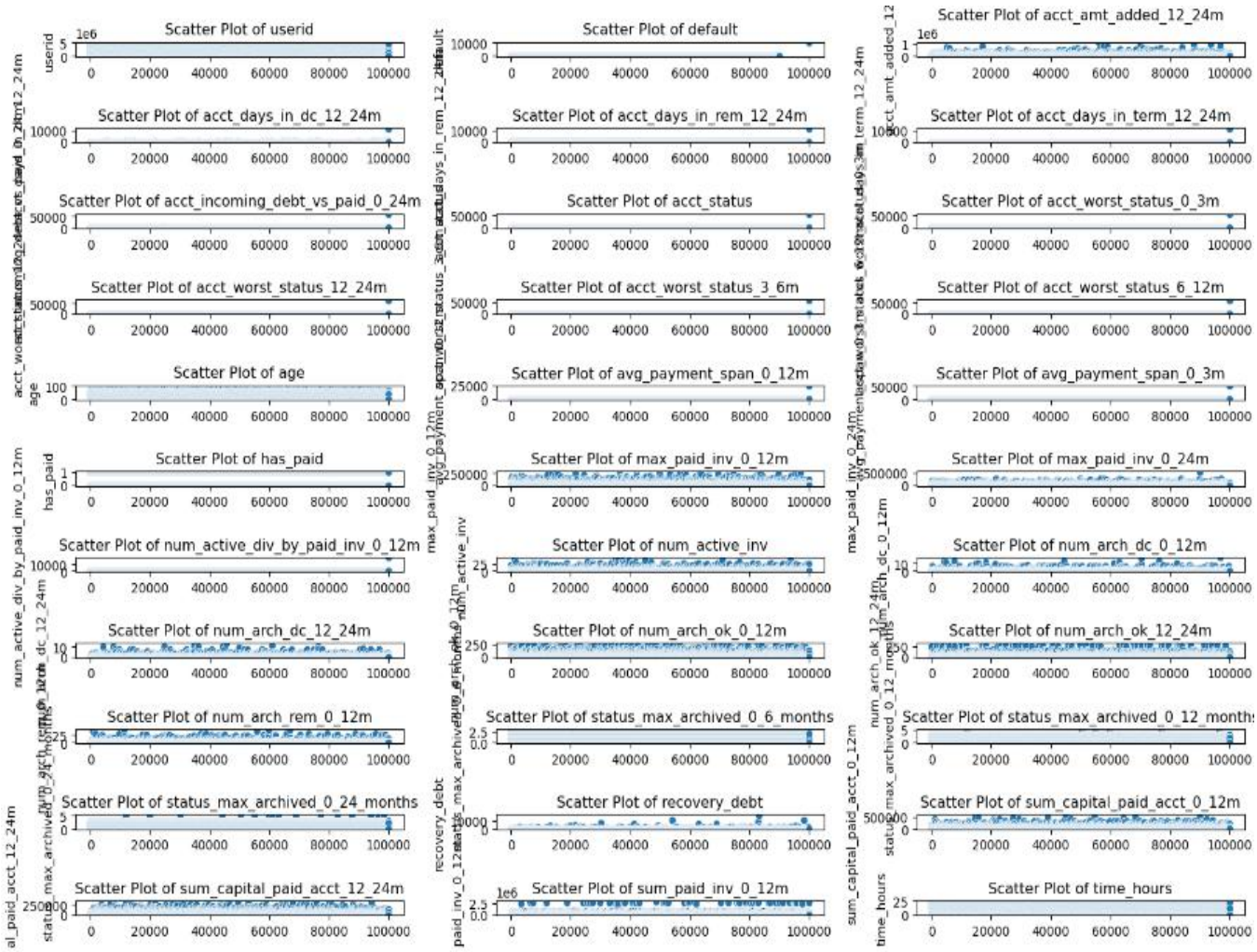
- **Removing Outliers**: Sometimes outliers are removed from the dataset if they are believed to be due to errors or noise.
- **Transforming Data**: Data transformation techniques like log transformation can help mitigate the impact of outliers.
- **Using Robust Methods**: Some statistical methods are less sensitive to outliers, such as using median instead of mean, or robust regression techniques.

Impact on Analysis:

- Outliers can skew the results of statistical analyses and models, leading to inaccurate conclusions.
- They can affect measures of central tendency and variability, such as mean and standard deviation.
- In predictive modeling, outliers can influence the performance and accuracy of models.

Understanding and appropriately handling outliers is essential for accurate data analysis and modeling.


Box Plot of userid
userid
1e6
Box Plot of default
default
Box Plot of acct_amt_added_12_24m
acct_amt_added_12_24m
1e6
Box Plot of acct_days_in_dc_12_24m
acct_days_in_dc_12_24m
Box Plot of acct_days_in_rem_12_24m
acct_days_in_rem_12_24m
Box Plot of acct_days_in_term_12_24m
acct_days_in_term_12_24m
Box Plot of acct_incoming_debt_vs_paid_0_24m
acct_incoming_debt_vs_paid_0_24m
Box Plot of acct_status
acct_status
Box Plot of acct_worst_status_0_3m
acct_worst_status_0_3m
Box Plot of acct_worst_status_12_24m
acct_worst_status_12_24m
Box Plot of acct_worst_status_3_6m
acct_worst_status_3_6m
Box Plot of acct_worst_status_6_12m
acct_worst_status_6_12m
Box Plot of age
age
Box Plot of avg_payment_span_0_12m
avg_payment_span_0_12m
Box Plot of avg_payment_span_0_3m
avg_payment_span_0_3m
Box Plot of has_paid
has_paid
Box Plot of max_paid_inv_0_12m
max_paid_inv_0_12m
Box Plot of max_paid_inv_0_24m
max_paid_inv_0_24m
Box Plot of num_active_div_by_paid_inv_0_12m
num_active_div_by_paid_inv_0_12m
Box Plot of num_active_inv
num_active_inv
Box Plot of num_arch_dc_0_12m
num_arch_dc_0_12m
Box Plot of num_arch_dc_12_24m
num_arch_dc_12_24m
Box Plot of num_arch_ok_0_12m
num_arch_ok_0_12m
Box Plot of num_arch_ok_12_24m
num_arch_ok_12_24m
Box Plot of num_arch_rem_0_12m
num_arch_rem_0_12m
Box Plot of status_max_archived_0_6_months
status_max_archived_0_6_months
Box Plot of status_max_archived_0_12_months
status_max_archived_0_12_months
Box Plot of status_max_archived_0_24_months
status_max_archived_0_24_months
Box Plot of recovery_debt
recovery_debt
Box Plot of sum_capital_paid_acct_0_12m
sum_capital_paid_acct_0_12m
Box Plot of sum_capital_paid_acct_12_24m
sum_capital_paid_acct_12_24m
Box Plot of sum_paid_inv_0_12m
sum_paid_inv_0_12m
1e6
Box Plot of time_hours
time_hours




```
Original Data Shape: (99979, 36)
Data Shape after Z-Score Outlier Treatment: (0, 33)
Data Shape after IQR Outlier Treatment: (33765, 33)
```

```
Summary Statistics after Z-Score Outlier Treatment:
       userid  default  acct_amt_added_12_24m  acct_days_in_dc_12_24m  \
count     0.0      0.0                    0.0                     0.0
mean      NaN      NaN                    NaN                     NaN
std       NaN      NaN                    NaN                     NaN
min       NaN      NaN                    NaN                     NaN
25%       NaN      NaN                    NaN                     NaN
50%       NaN      NaN                    NaN                     NaN
75%       NaN      NaN                    NaN                     NaN
max       NaN      NaN                    NaN                     NaN

       acct_days_in_rem_12_24m  acct_days_in_term_12_24m  \
count                      0.0                       0.0
mean                       NaN                       NaN
std                        NaN                       NaN
min                        NaN                       NaN
25%                        NaN                       NaN
50%                        NaN                       NaN
75%                        NaN                       NaN
max                        NaN                       NaN

       acct_incoming_debt_vs_paid_0_24m  acct_status  acct_worst_status_0_3m  \
count                               0.0          0.0                     0.0
mean                                NaN          NaN                     NaN
std                                 NaN          NaN                     NaN
min                                 NaN          NaN                     NaN
25%                                 NaN          NaN                     NaN
50%                                 NaN          NaN                     NaN
75%                                 NaN          NaN                     NaN
max                                 NaN          NaN                     NaN

       acct_worst_status_12_24m  ...  num_arch_ok_12_24m  num_arch_rem_0_12m  \
count                       0.0  ...                 0.0                 0.0
mean                        NaN  ...                 NaN                 NaN
std                         NaN  ...                 NaN                 NaN
min                         NaN  ...                 NaN                 NaN
25%                         NaN  ...                 NaN                 NaN
50%                         NaN  ...                 NaN                 NaN
75%                         NaN  ...                 NaN                 NaN
max                         NaN  ...                 NaN                 NaN
```

```
       status_max_archived_0_6_months  status_max_archived_0_12_months  \
count                             0.0                              0.0
mean                              NaN                              NaN
std                               NaN                              NaN
min                               NaN                              NaN
25%                               NaN                              NaN
50%                               NaN                              NaN
75%                               NaN                              NaN
max                               NaN                              NaN

       status_max_archived_0_24_months  recovery_debt  \
count                              0.0            0.0
mean                               NaN            NaN
std                                NaN            NaN
min                                NaN            NaN
25%                                NaN            NaN
50%                                NaN            NaN
75%                                NaN            NaN
max                                NaN            NaN

       sum_capital_paid_acct_0_12m  sum_capital_paid_acct_12_24m  \
count                          0.0                           0.0
mean                           NaN                           NaN
std                            NaN                           NaN
min                            NaN                           NaN
25%                            NaN                           NaN
50%                            NaN                           NaN
75%                            NaN                           NaN
max                            NaN                           NaN

       sum_paid_inv_0_12m  time_hours
count                 0.0         0.0
mean                  NaN         NaN
std                   NaN         NaN
min                   NaN         NaN
25%                   NaN         NaN
50%                   NaN         NaN
75%                   NaN         NaN
max                   NaN         NaN

[8 rows x 33 columns]
```

```
Summary Statistics after IQR Outlier Treatment:
              userid  default  acct_amt_added_12_24m  acct_days_in_dc_12_24m  \
count  3.376300e+04  30362.0           33763.000000                 28292.0
mean   2.999725e+06      0.0             194.761840                    0.0
std    1.152097e+06      0.0            1190.783573                    0.0
min    1.000099e+06      0.0               0.000000                    0.0
25%    2.005286e+06      0.0               0.000000                    0.0
50%    3.005196e+06      0.0               0.000000                    0.0
75%    3.998892e+06      0.0               0.000000                    0.0
max    4.999528e+06      0.0           12317.000000                    0.0

       acct_days_in_rem_12_24m  acct_days_in_term_12_24m  \
count                  28292.0                   28292.0
mean                       0.0                       0.0
std                        0.0                       0.0
min                        0.0                       0.0
25%                        0.0                       0.0
50%                        0.0                       0.0
75%                        0.0                       0.0
max                        0.0                       0.0

       acct_incoming_debt_vs_paid_0_24m  acct_status  acct_worst_status_0_3m  \
count                       4083.000000       6448.0                  6448.0
mean                           0.167932          1.0                     1.0
std                            0.372230          0.0                     0.0
min                            0.000000          1.0                     1.0
25%                            0.000000          1.0                     1.0
50%                            0.000000          1.0                     1.0
75%                            0.012089          1.0                     1.0
max                            1.656384          1.0                     1.0

       acct_worst_status_12_24m  ...  num_arch_ok_12_24m  num_arch_rem_0_12m  \
count                    2531.0  ...        26494.000000                 0.0
mean                        1.0  ...            3.074017                 0.0
std                         0.0  ...            3.850271                 0.0
min                         1.0  ...            0.000000                 0.0
25%                         1.0  ...            0.000000                 0.0
50%                         1.0  ...            2.000000                 0.0
75%                         1.0  ...            5.000000                 0.0
max                         1.0  ...           17.000000                 0.0
```

	status_max_archived_0_6_months	status_max_archived_0_12_months	\
count	26494.000000	26494.000000	
mean	0.667623	0.839775	
std	0.471795	0.372235	
min	0.000000	0.000000	
25%	0.000000	1.000000	
50%	1.000000	1.000000	
75%	1.000000	1.000000	
max	2.000000	3.000000	

	status_max_archived_0_24_months	recovery_debt	\
count	26494.000000	26494.0	
mean	1.026836	0.0	
std	0.395590	0.0	
min	0.000000	0.0	
25%	1.000000	0.0	
50%	1.000000	0.0	
75%	1.000000	0.0	
max	3.000000	0.0	

	sum_capital_paid_acct_0_12m	sum_capital_paid_acct_12_24m	\
count	26494.000000	26494.000000	
mean	743.836680	0.215445	
std	2906.645433	5.697240	
min	0.000000	0.000000	
25%	0.000000	0.000000	
50%	0.000000	0.000000	
75%	0.000000	0.000000	
max	22395.000000	243.000000	

	sum_paid_inv_0_12m	time_hours
count	26494.000000	26494.000000
mean	17245.127878	15.352405
std	18092.495472	5.023744
min	0.000000	0.001667
25%	4065.000000	11.649514
50%	11663.500000	15.795417
75%	25053.750000	19.549097
max	108757.000000	23.999722

[8 rows x 33 columns]

e) Variable transformation (if applicable)

Step 1: Handling Missing Values

1. **Identify Numerical Columns:** The script first identifies all columns containing numerical data types.
2. **Imputation for Numerical Columns (Example: Mean Imputation):** Missing values in these columns are filled with the average (mean) value of the respective column. This creates a complete dataset without missing numerical values.

Step 2: Standardizing Categorical Variables

1. **Identify Categorical Columns:** The script identifies all columns containing categorical data types (strings in this case).
2. **Ensuring Consistent Data Type:** Categorical columns are converted to a uniform string data type for consistency. This ensures proper handling during encoding.

3. **Label Encoding:** Each categorical column is encoded using a Label Encoder. This process assigns numerical labels to each unique category within the column. A dictionary (label_encoders) is created to store the encoders used for each column, allowing for potential decoding later.

Step 3: Feature Scaling

1. **Identify Numerical Columns:** Similar to step 1, numerical columns are identified again.
2. **Standard Scaling:** A StandardScaler is used to transform the numerical features. This process scales the features to have a mean of 0 and a standard deviation of 1. This can improve the performance of some machine learning algorithms.

Step 4: Splitting the Data (Optional)

This step is optional but commonly used in supervised learning tasks. The script splits the data into two sets:

1. **Training Data (80%):** This larger portion of the data is used to train the machine learning model.
2. **Testing Data (20%):** This smaller portion is used to evaluate the performance of the trained model on unseen data. Splitting the data helps to prevent overfitting and ensures a more robust model.

Display the first few rows of the processed data to verify the transformations

```
     userid   default  acct_amt_added_12_24m  acct_days_in_dc_12_24m  \
0  1.358683 -0.003967              -0.345399           -9.446327e-03
1 -0.315223 -0.003967              -0.345399           -9.446327e-03
2  1.564110 -0.003967              -0.345399           -9.446327e-03
3 -1.348348 -0.003967              -0.345399           -1.467505e-18
4  1.365781 -0.003967              -0.345399           -9.446327e-03

   acct_days_in_rem_12_24m  acct_days_in_term_12_24m  \
0             -1.200544e-01                -0.011222
1             -1.200544e-01                -0.011222
2             -1.200544e-01                -0.011222
3             -2.058946e-17                 0.000000
4             -1.200544e-01                -0.011222

   acct_incoming_debt_vs_paid_0_24m   acct_status  acct_worst_status_0_3m  \
0                     -1.481341e-02 -7.178785e-03               -0.007939
1                     -4.715765e-18 -7.178785e-03               -0.007939
2                     -4.715765e-18 -2.582583e-18                0.000000
3                     -4.715765e-18 -2.582583e-18                0.000000
4                     -4.715765e-18 -2.582583e-18                0.000000

   acct_worst_status_12_24m  ...  num_arch_ok_12_24m  num_arch_rem_0_12m  \
0                  0.000000  ...            0.472814           -0.367687
1                 -0.011117  ...            0.801935            1.911384
2                  0.000000  ...           -0.451767            1.911384
3                  0.000000  ...            0.933904           -0.367687
4                  0.000000  ...           -0.451767           -0.367687

   status_max_archived_0_6_months  status_max_archived_0_12_months  \
0                        0.263725                        -0.101303
1                        0.263725                         1.264290
2                        0.263725                         1.264290
3                        0.263725                        -0.101303
4                        0.263725                        -0.101303

   status_max_archived_0_24_months  recovery_debt  \
0                        -0.320702      -0.032862
1                         0.971448      -0.032862
2                         0.971448      -0.032862
3                        -0.320702      -0.032862
4                        -0.320702      -0.032862

   sum_capital_paid_acct_0_12m  sum_capital_paid_acct_12_24m  \
0                    -0.432374                     -0.364449
1                    -0.432374                     -0.364449
2                    -0.432374                     -0.364449
3                    -0.432374                     -0.364449
4                    -0.432374                     -0.364449

   sum_paid_inv_0_12m  time_hours
0            1.544490   -1.198784
1            0.089418   -0.455264
2            0.939261   -0.796554
3            3.179024    0.086292
4           -0.380352   -0.557007

[5 rows x 36 columns]
```

D. Business Insights from EDA:-

b) Any business insights using Clustering (if applicable) +
c) Any other business insights

· **K-Means Clustering:**

- A KMeans object is created, specifying the number of clusters (n_clusters=5) and a random state (random_state=0) for reproducibility.
- The fit_predict method of the KMeans object is used on features to perform the clustering. This assigns a cluster label to each data point in train_data and stores it in a new column named 'Cluster'.

· **Cluster Analysis:**

- The code uses pandas' groupby operation to group train_data by the 'Cluster' column.
- It then calculates the mean of each numerical column (numeric_columns) within each cluster. The resulting DataFrame (cluster_analysis) provides insights into the average values of numerical features for each cluster.

· **Cluster Visualization:**

- A PCA object is created to reduce the dimensionality of the features for visualization purposes (limited to 2 dimensions in this case).
- The fit_transform method of PCA is applied to features to obtain the principal components.
- Two new columns, 'PCA1' and 'PCA2', are added to train_data to store the transformed features.
- A scatter plot is created using matplotlib and seaborn. Each data point is colored based on its cluster label, allowing visual inspection of how the data points are grouped.

· **Cluster Validation:**

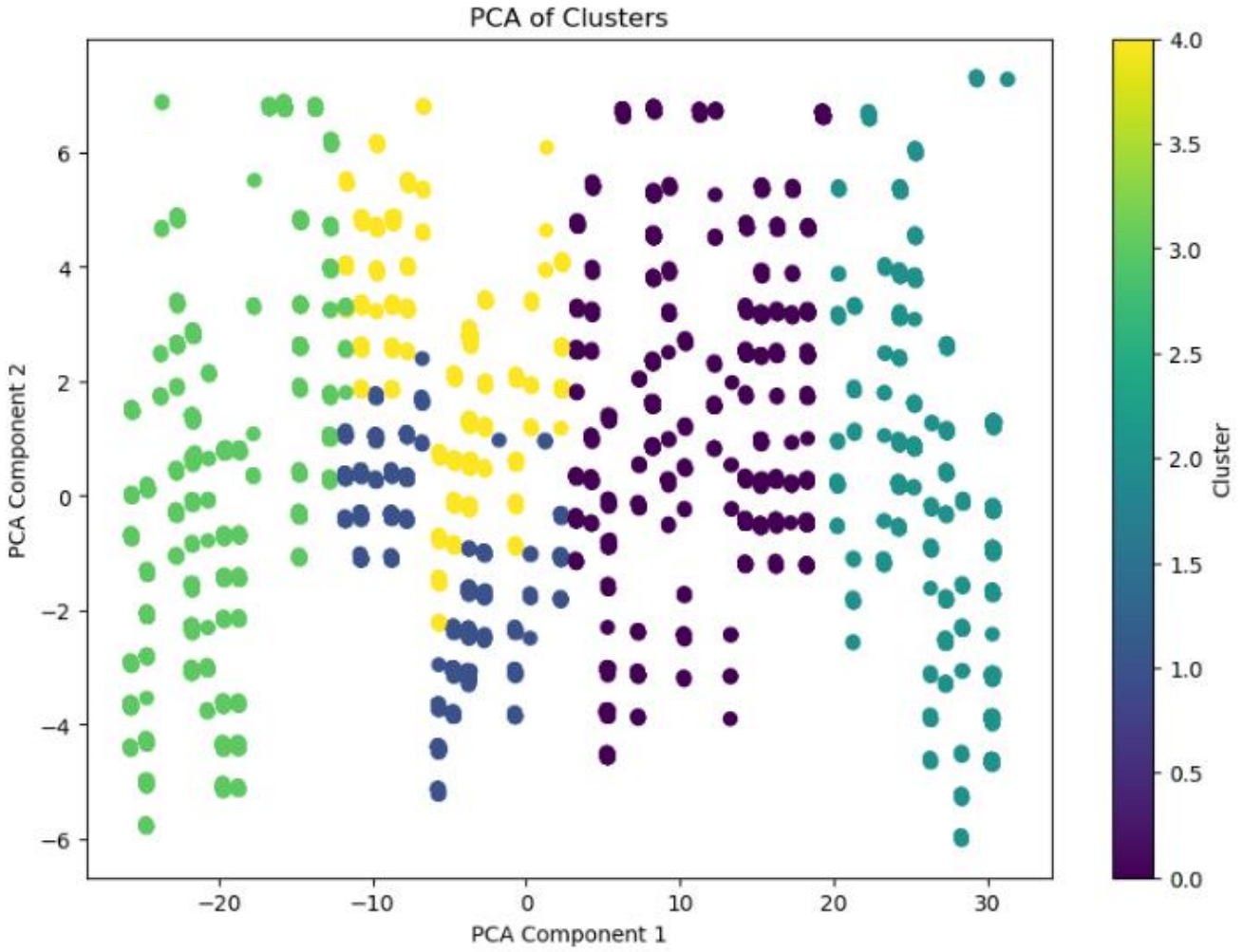
- The silhouette_score function from scikit-learn is used to evaluate the quality of the clustering. This score measures how well data points are separated within their assigned clusters.
- The code calculates the silhouette score for the clustering and prints it.

· **Saving Encoders and Scaler (Optional):**

- Assuming label_encoders is a dictionary containing label encoders used for categorical variables, the code iterates through each key-value pair.
- The label encoder (le) for each column (column) is saved using joblib.dump with a descriptive filename.

- Similarly, the scaler object (scaler) used for feature scaling is saved using joblib.dump. This allows for future use of these pre-processing steps on new data.

```
Cluster Analysis:
            userid   default  acct_amt_added_12_24m  acct_days_in_dc_12_24m  \
Cluster
0       -0.006474 -0.003038               0.044731               -0.002973
1       -0.002184 -0.003328               0.003122               -0.005498
2        0.021945 -0.002850               0.015823                0.000197
3       -0.006380 -0.003173              -0.058123               -0.002752
4       -0.010502 -0.003213               0.002295               -0.002525

         acct_days_in_rem_12_24m  acct_days_in_term_12_24m  \
Cluster
0                       0.027362                 -0.001222
1                      -0.028554                 -0.005098
2                       0.037735                 -0.000370
3                      -0.009193                 -0.003466
4                      -0.013917                 -0.003762

         acct_incoming_debt_vs_paid_0_24m  acct_status  \
Cluster
0                               -0.004361    -0.003505
1                               -0.001942    -0.002901
2                               -0.004402    -0.003758
3                               -0.003349    -0.002869
4                               -0.003326    -0.003113

         acct_worst_status_0_3m  acct_worst_status_12_24m  ...  \
Cluster                                                    ...
0                     -0.003442                 -0.003463  ...
1                     -0.002937                 -0.002970  ...
2                     -0.003731                 -0.003623  ...
3                     -0.002854                 -0.002848  ...
4                     -0.003133                 -0.003197  ...

         num_arch_ok_12_24m  num_arch_rem_0_12m  \
Cluster
0                 -0.183944           -0.015122
1                  0.130425           -0.035167
2                 -0.185976            0.047102
3                 -0.093088           -0.011578
4                  0.152770            0.027169

         status_max_archived_0_6_months  status_max_archived_0_12_months  \
Cluster
0                             -0.086197                        -0.055020
1                              0.003747                        -0.013254
2                              0.035120                         0.071453
3                             -0.036079                        -0.030828
4                              0.051682                         0.023221
         status_max_archived_0_24_months  recovery_debt  \
Cluster
0                              -0.043234      -0.009750
1                              -0.023548      -0.013640
2                               0.087438       0.018870
3                              -0.031059       0.001770
4                               0.009562       0.000349

         sum_capital_paid_acct_0_12m  sum_capital_paid_acct_12_24m  \
Cluster
0                           0.044977                      0.042154
1                           0.000583                      0.013236
2                          -0.009079                     -0.025303
3                          -0.033819                     -0.026162
4                           0.013170                     -0.003168

         sum_paid_inv_0_12m  time_hours
Cluster
0                 -0.065273    0.000270
1                  0.024886    0.002575
2                 -0.063157   -0.003932
3                 -0.065383   -0.001062
4                  0.109662   -0.000843

[5 rows x 33 columns]
```




```
Silhouette Score: 0.32521964964191274
```

The PCA plot you provided visualizes the clustering results of your data. Each point represents a data instance projected onto two principal components (PCA Component 1 and PCA Component 2), and the colors indicate the cluster assignments. Here's how we can derive some business insights from this plot:

1. Cluster Distribution and Separation: The plot shows distinct clusters indicating that the KMeans algorithm successfully identified groups of similar instances. Good separation between clusters suggests that the features used for clustering are effective in distinguishing between different patterns in the data.

2. Cluster Characteristics: To gain business insights, we should look at the average values of key metrics for each cluster (which we already calculated). For instance, clusters might represent different segments of customers based on their credit behavior. Analyzing the cluster_analysis output will help identify unique characteristics of each cluster.

3. Potential Actions for Each Cluster:

Cluster 0: If this cluster has higher default rates, you might want to implement stricter credit policies or additional monitoring for this group.

Cluster 1: If this cluster shows strong financial health (e.g., low default rates, high paid amounts), they could be targeted for new credit products or upselling.

Cluster 2: If this group is in between the high-risk and low-risk segments, tailored offers or educational content about financial management could be beneficial.

Cluster 3: This cluster might represent a niche segment with unique characteristics. Understanding their needs could help in designing specific financial products.

Cluster 4: If this cluster shows diverse behavior, more granular segmentation might be needed.

4. Market Strategy:

Marketing and Customer Engagement: Use the cluster insights to create targeted marketing campaigns. For example, offering lower interest rates to low-risk clusters and personalized repayment plans to high-risk clusters.

Product Development: Develop financial products tailored to the needs of each cluster, like premium credit cards for low-risk customers and secured cards for high-risk customers.

5. Risk Management: Use the cluster information to better manage risk by adjusting credit limits, interest rates, and approval criteria based on the risk profile of each cluster.

6. Operational Efficiency: Optimize resource allocation by focusing more on high-value or high-risk clusters. This could involve providing more personalized customer service to profitable clusters or investing in risk mitigation for high-risk clusters.

E. Model building and interpretation:-

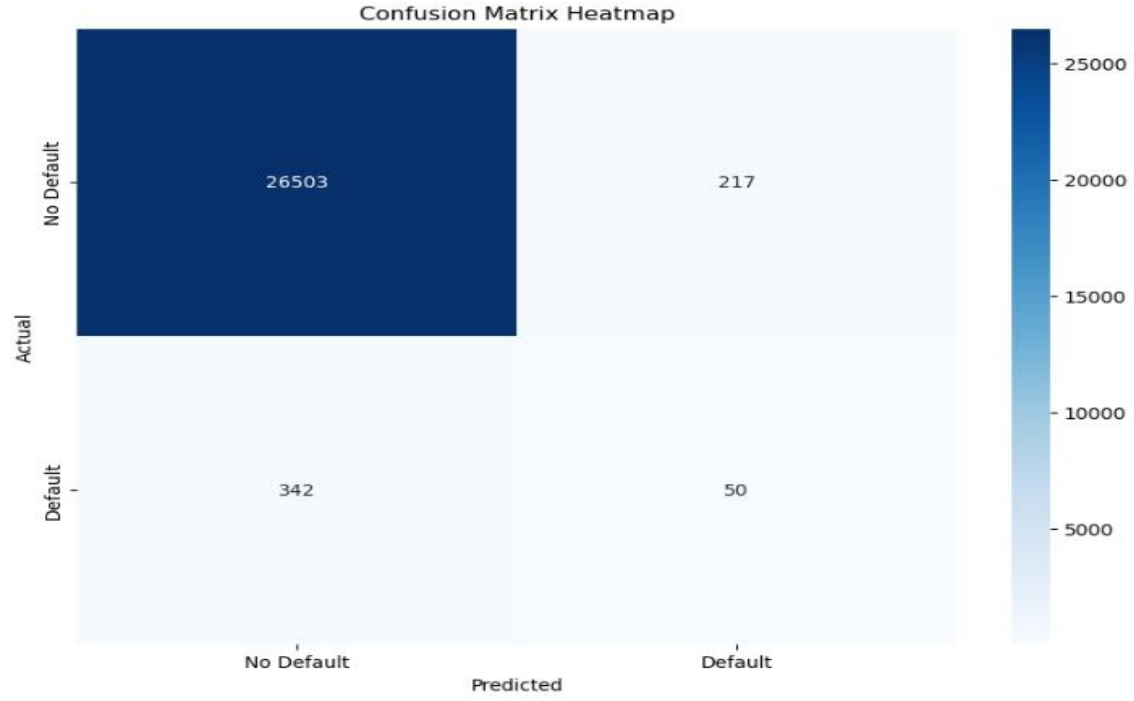
a. Build various models (You can choose to build models for either or all of descriptive, predictive or prescriptive purposes) b. Test your predictive model against the test set using various appropriate performance metrics c.Interpretation of the model(s)

1. RANDOM FOREST CLASSIFIER:

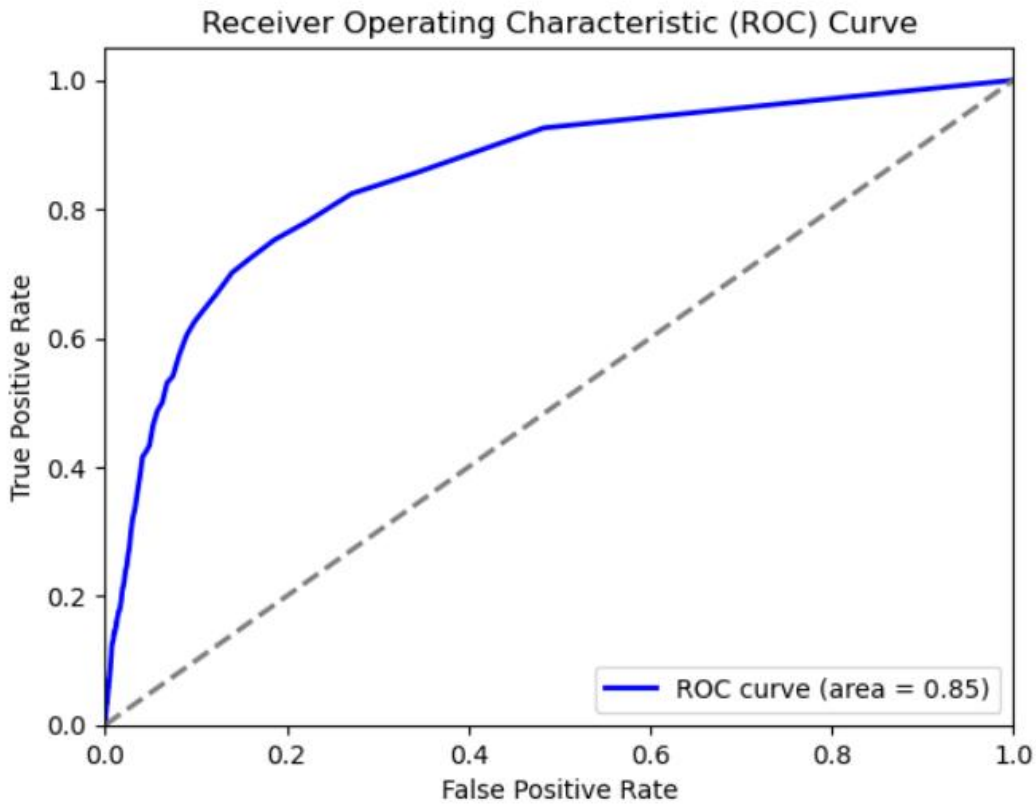
Classification Report:

	precision	recall	f1-score	support
0.0	0.99	0.99	0.99	26720
1.0	0.19	0.13	0.15	392
accuracy			0.98	27112
macro avg	0.59	0.56	0.57	27112
weighted avg	0.98	0.98	0.98	27112

ACCURACY = 98%



AUC Score: 0.85

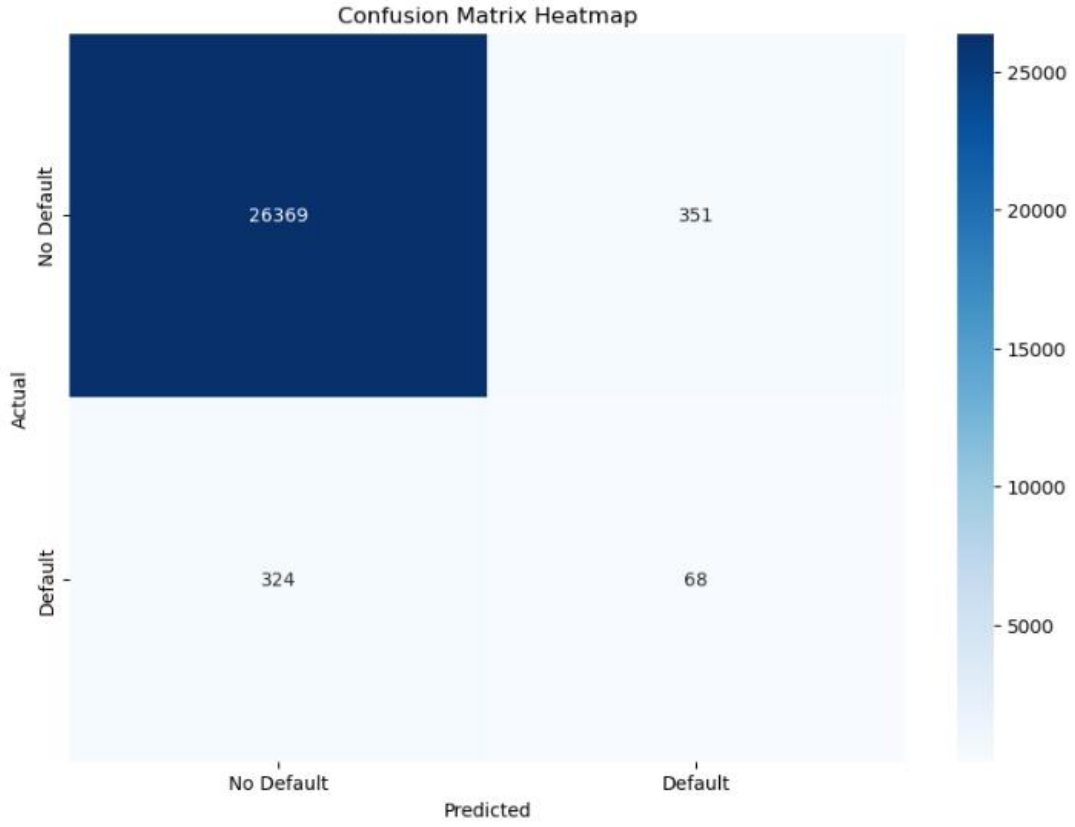


2. DECISION TREE:

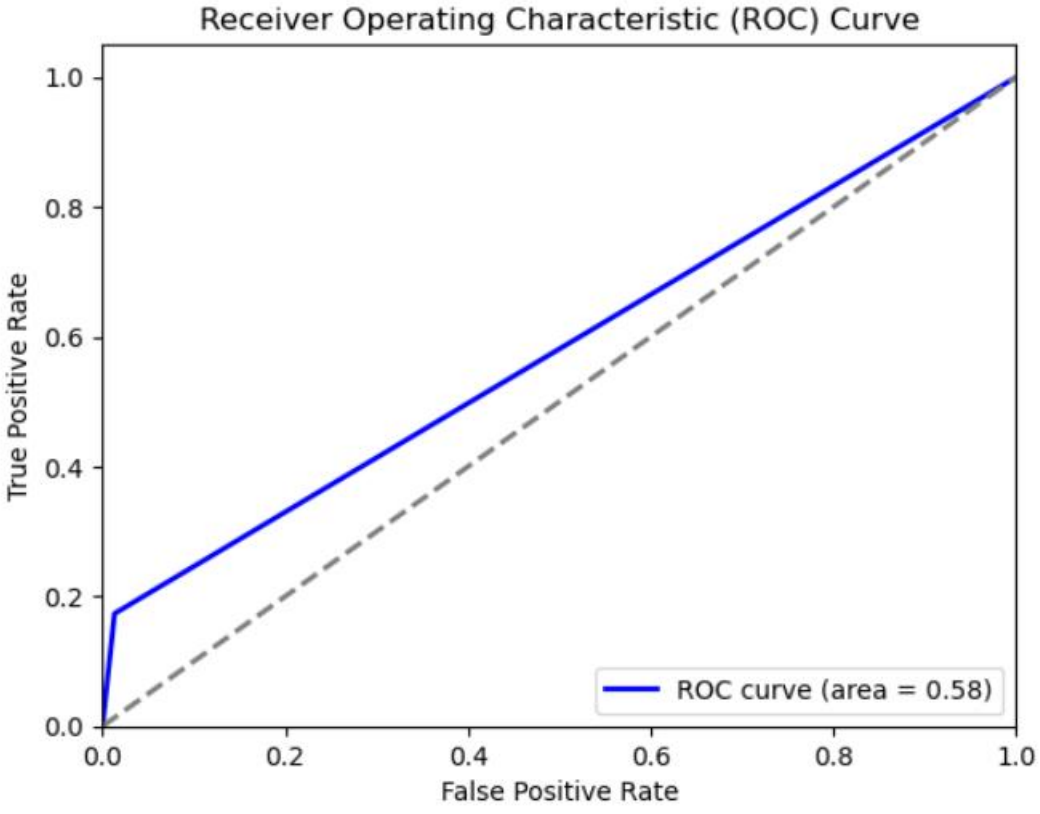
Classification Report:

	precision	recall	f1-score	support
0.0	0.99	0.99	0.99	26720
1.0	0.16	0.17	0.17	392
accuracy			0.98	27112
macro avg	0.58	0.58	0.58	27112
weighted avg	0.98	0.98	0.98	27112

ACCURACY = 98%



AUC Score: 0.58

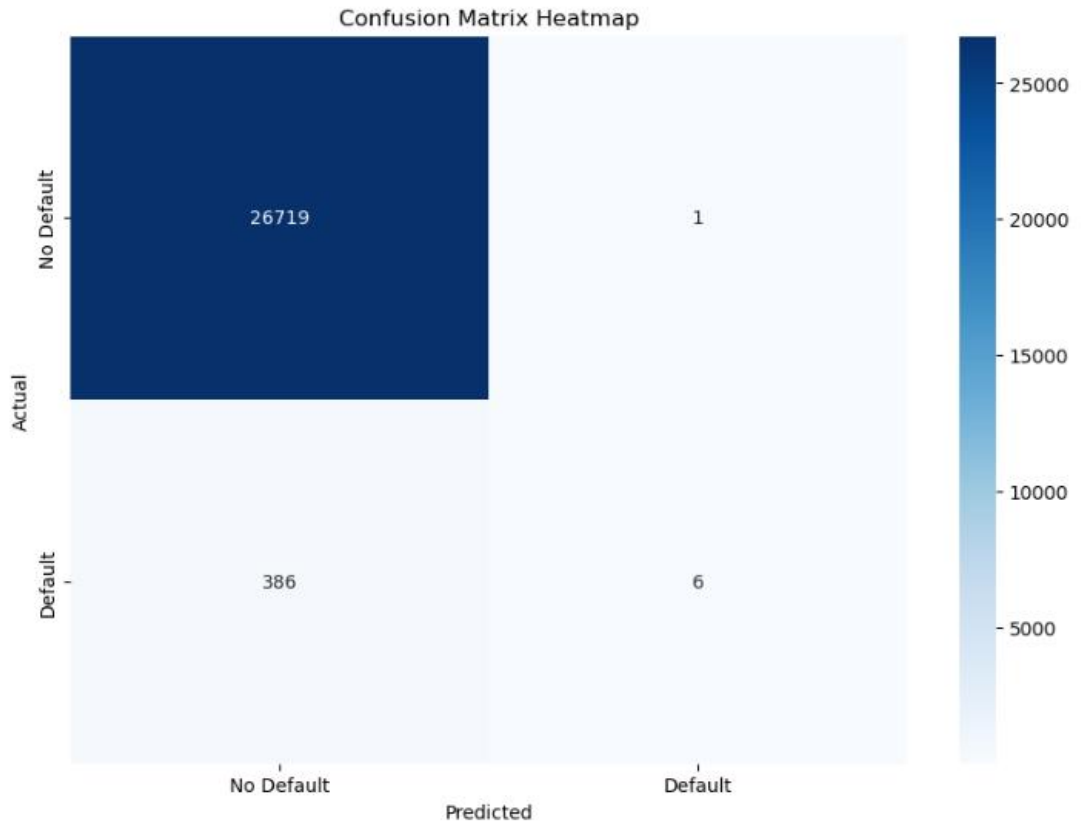


3. SUPPORT VECTOR MACHINE:

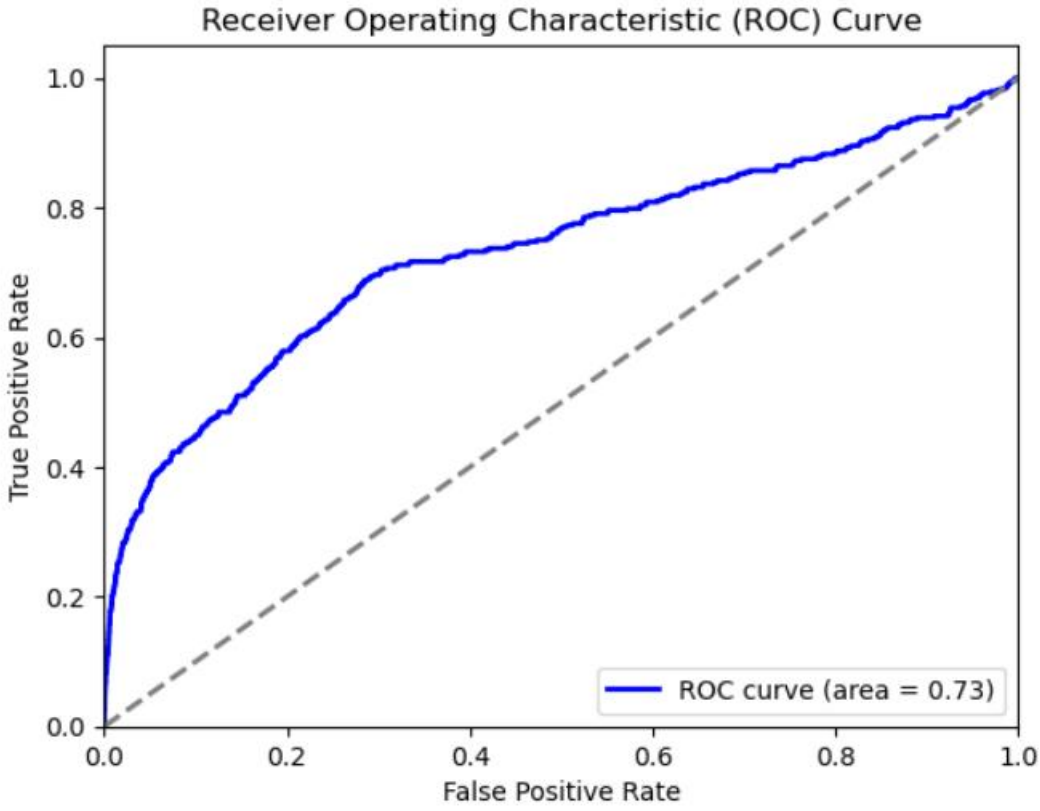
Classification Report:

	precision	recall	f1-score	support
0.0	0.99	1.00	0.99	26720
1.0	0.86	0.02	0.03	392
accuracy			0.99	27112
macro avg	0.92	0.51	0.51	27112
weighted avg	0.98	0.99	0.98	27112

ACCURACY = 99%



AUC Score: 0.73

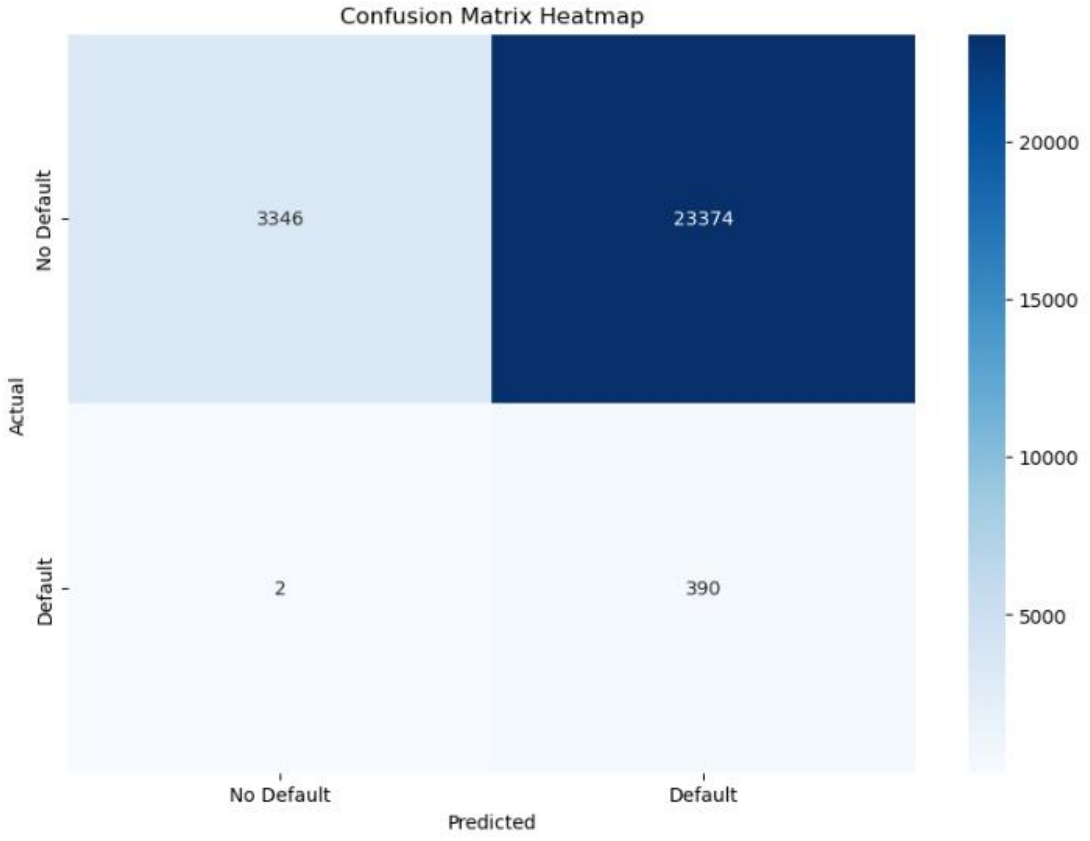


4. NAIVE BAYES:

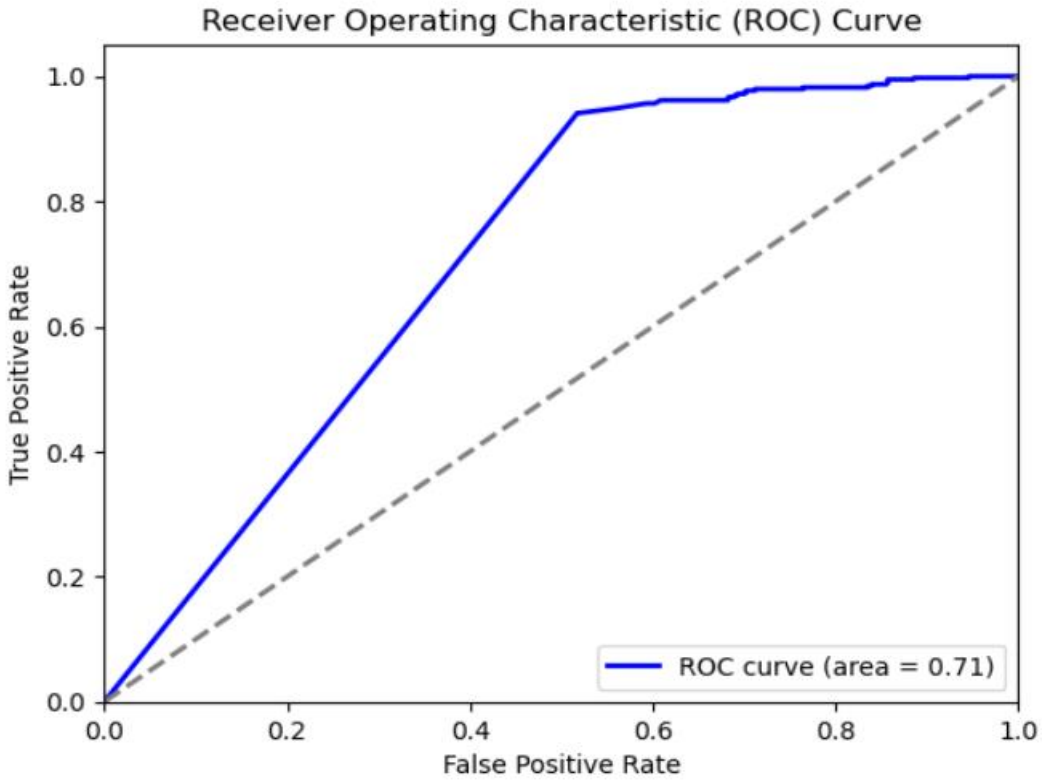
Classification Report:

	precision	recall	f1-score	support
0.0	1.00	0.13	0.22	26720
1.0	0.02	0.99	0.03	392
accuracy			0.14	27112
macro avg	0.51	0.56	0.13	27112
weighted avg	0.99	0.14	0.22	27112

ACCURACY = 14%



AUC Score: 0.71



F. Model Tuning:-

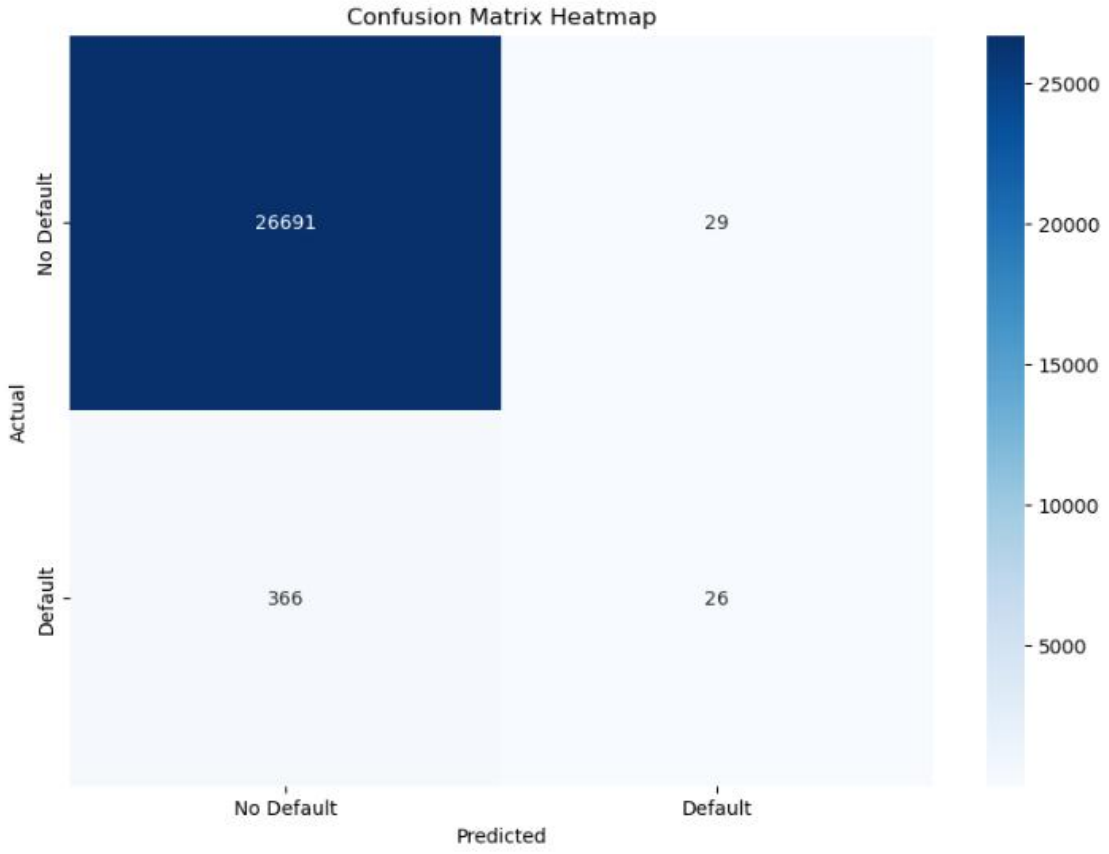
a.Ensemble modelling, wherever applicable b. Any other model tuning measures(if applicable) c. Interpretation of the most optimum model and its implication on the business

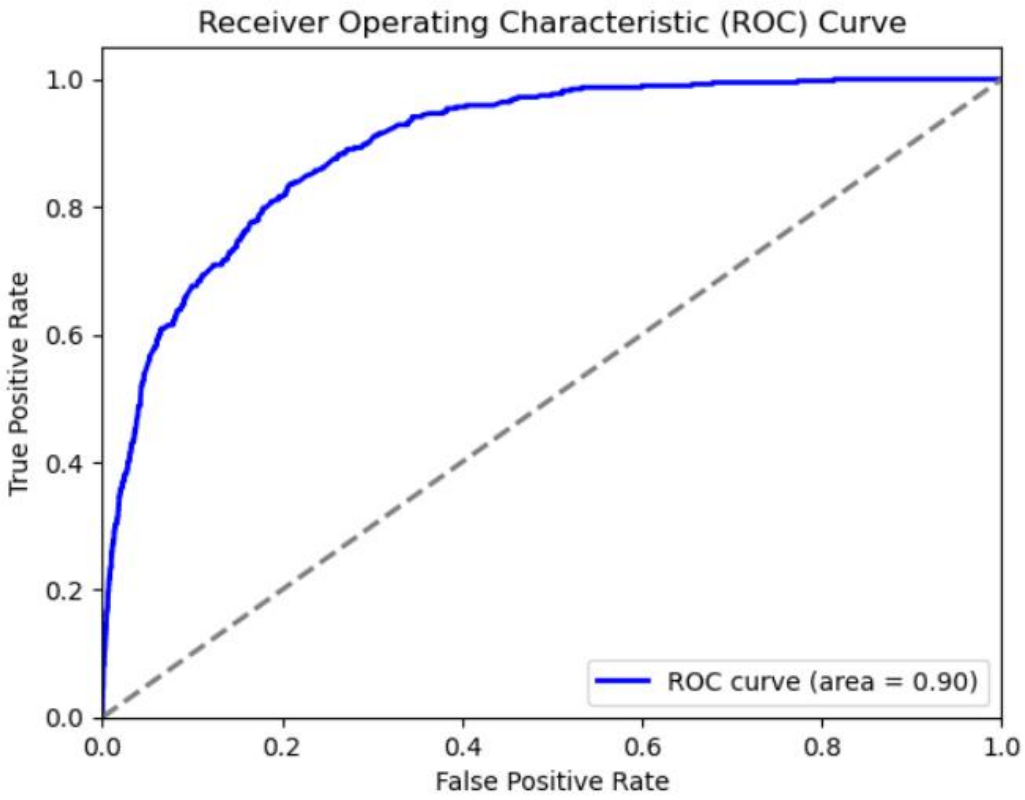
1. Gradient Boosting Classifier:

Classification Report:

	precision	recall	f1-score	support
0.0	0.99	1.00	0.99	26720
1.0	0.47	0.07	0.12	392
accuracy			0.99	27112
macro avg	0.73	0.53	0.55	27112
weighted avg	0.98	0.99	0.98	27112

ACCURACY = 99%



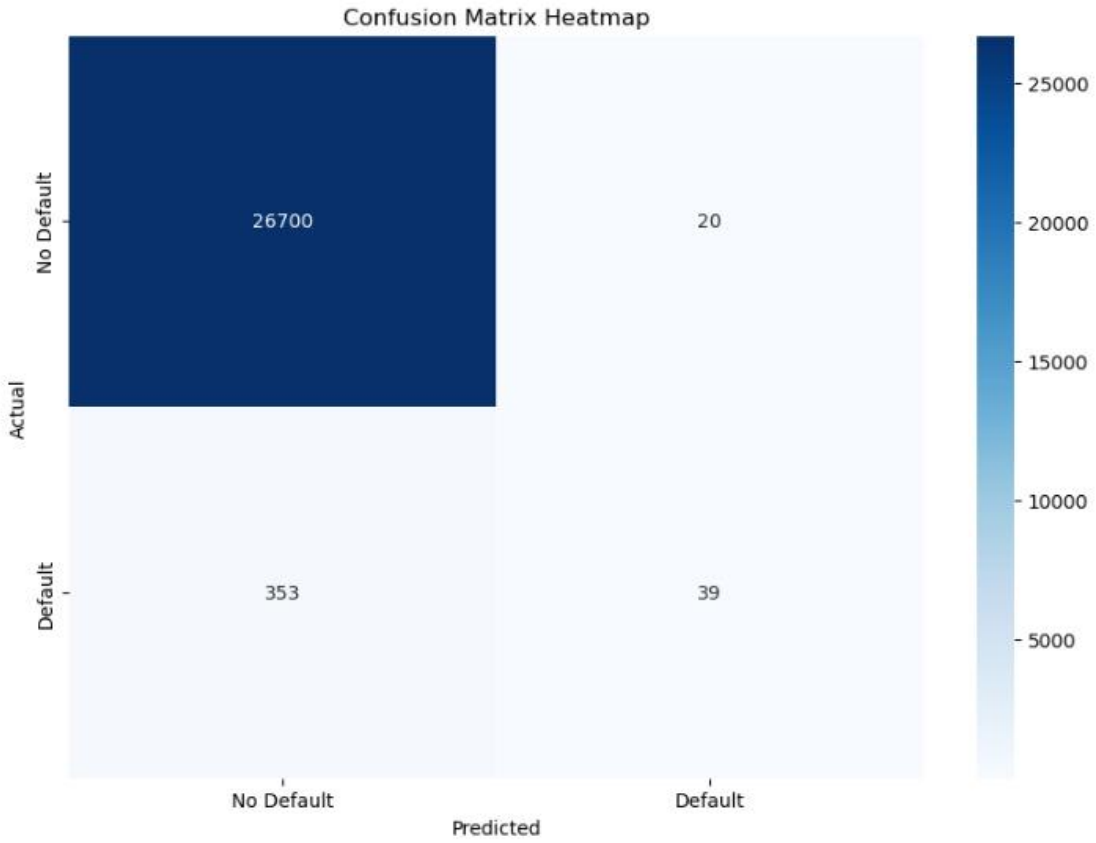


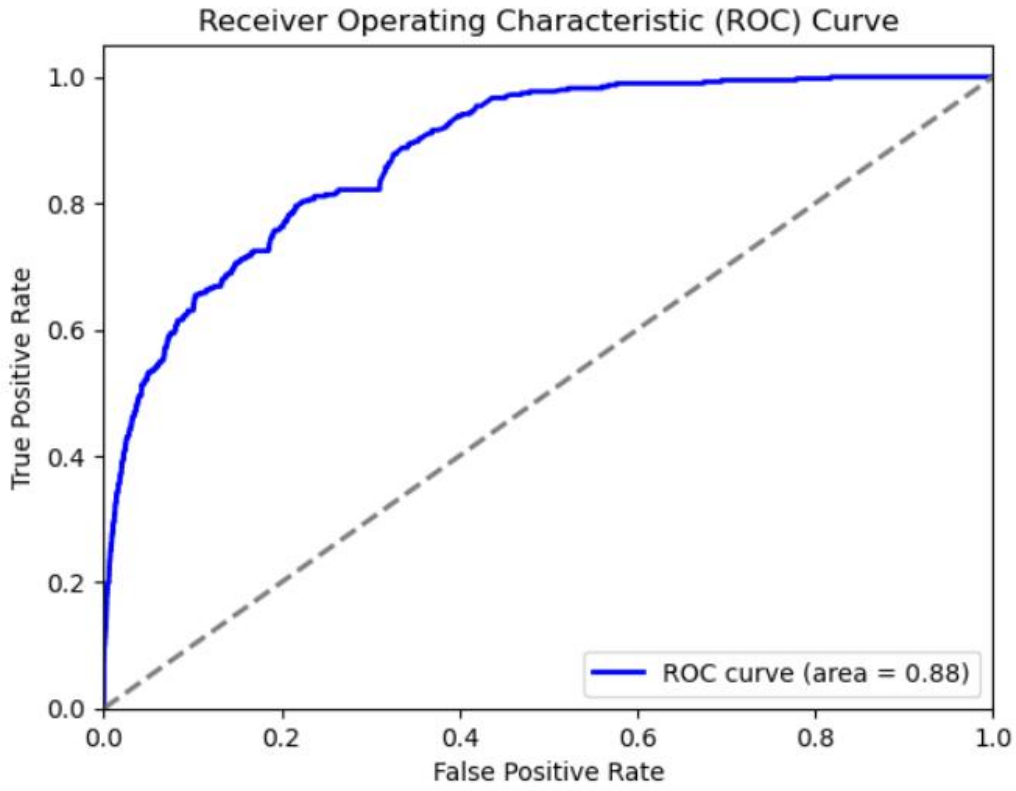
2. Stacking Classifier:

Classification Report:

	precision	recall	f1-score	support
0.0	0.99	1.00	0.99	26720
1.0	0.66	0.10	0.17	392
accuracy			0.99	27112
macro avg	0.82	0.55	0.58	27112
weighted avg	0.98	0.99	0.98	27112

ACCURACY = 99%


Confusion Matrix Heatmap
No Default
Default
Actual
26700
20
353
39
No Default
Default
Predicted
25000
20000
15000
10000
5000

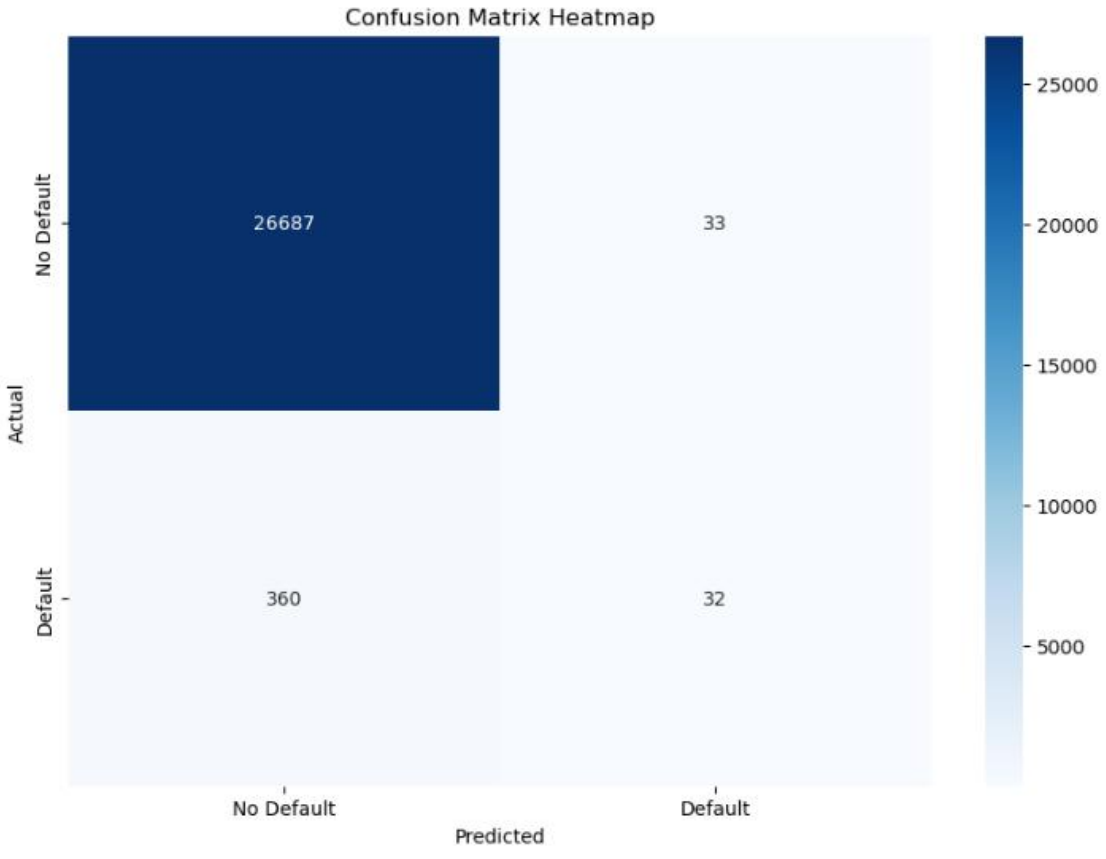

Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.88)

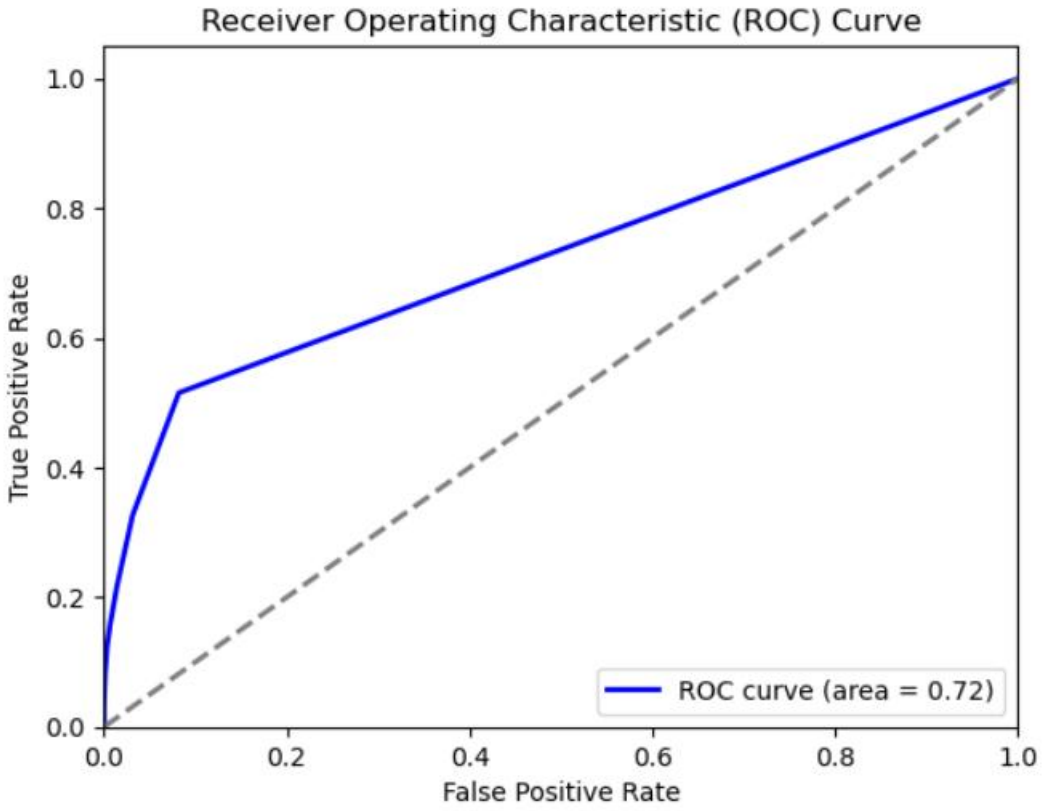
3. Bagging Classifier:

Classification Report:

	precision	recall	f1-score	support
0.0	0.99	1.00	0.99	26720
1.0	0.49	0.08	0.14	392
accuracy			0.99	27112
macro avg	0.74	0.54	0.57	27112
weighted avg	0.98	0.99	0.98	27112

ACCURACY = 99%



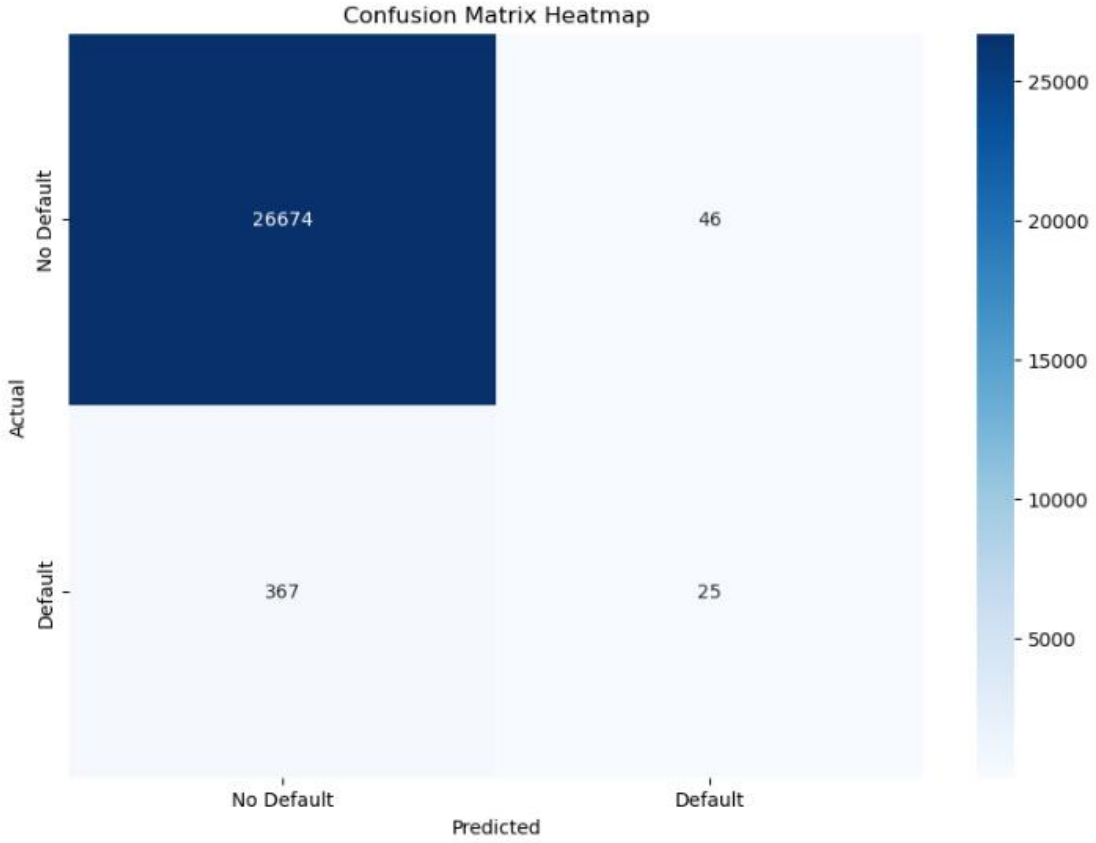


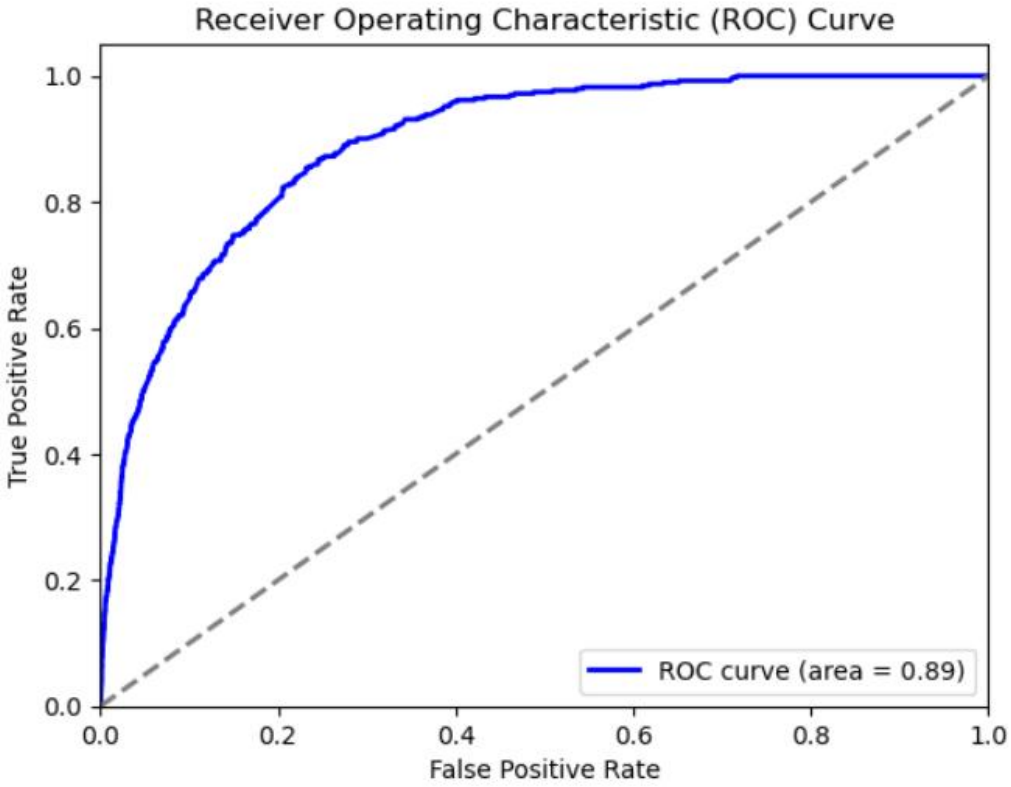
4. Ada Boost Classifier:

Classification Report:

	precision	recall	f1-score	support
0.0	0.99	1.00	0.99	26720
1.0	0.35	0.06	0.11	392
accuracy			0.98	27112
macro avg	0.67	0.53	0.55	27112
weighted avg	0.98	0.98	0.98	27112

ACCURACY = 98%


Confusion Matrix Heatmap
No Default
Default
Actual
26674
46
367
25
No Default
Default
Predicted
25000
20000
15000
10000
5000


Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.89)

COMPARISION OF ALL THE MODELS:

	Model Name	Accuracy	Precision	Recall	F1-Score	AUC Score
0	Random Forest Classifier	0.986316	0.818182	0.068878	0.127059	0.830196
1	Decision Tree Classifier	0.975103	0.162291	0.173469	0.167694	0.580167
2	Support Vector Machine	0.985726	0.857143	0.015306	0.030075	0.732233
3	Naive Bayes Classifier	0.137799	0.016411	0.994898	0.032290	0.714962
4	Bagging Classifier	0.985505	0.492308	0.081633	0.140044	0.723674
5	Ada Boost Classifier	0.984767	0.352113	0.063776	0.107991	0.892159
6	Gradient Boosting Classifier	0.985431	0.472727	0.066327	0.116331	0.897468
7	Stacking	0.986242	0.661017	0.099490	0.172949	0.881665

Choosing the Optimum Model:-

The Gradient Boosting Classifier and Stacking models stand out with high AUC Scores of 0.897468 and 0.881665, respectively. These models indicate a good balance between distinguishing positive and negative classes.

The Random Forest Classifier has the highest accuracy, but its recall is very low, indicating it misses a large number of actual positives. On the other hand, the Naive Bayes Classifier has the highest recall but suffers from very low accuracy and precision.

Considering a balance of accuracy, precision, recall, F1-score, and AUC score, Gradient Boosting Classifier appears to be the most optimum model.

Implications on the Business:-

Using the Gradient Boosting Classifier can provide the following benefits to the business:

Improved Prediction Accuracy: Higher accuracy and AUC scores indicate better performance in distinguishing between classes, leading to more reliable predictions.

Balanced Performance: Good precision and recall balance, ensuring fewer false positives and false negatives, which is crucial for decision-making processes.

Customer Satisfaction: Reliable predictions can improve customer satisfaction by providing accurate results, enhancing trust in the system.

Resource Allocation: Better prediction performance can optimize resource allocation, ensuring efforts and investments are directed towards the right areas.

Risk Management: Enhanced ability to predict and manage risks by accurately identifying potential issues or opportunities. Adopting the Gradient Boosting Classifier can significantly enhance the business's decision-making capabilities, leading to better outcomes and improved efficiency.