CAPSTONE PROJECT
BUSINESS REPORT

Pavithra Devi
PGPDSBA
Great Learning

INDEX

Introduction of the business problem

Dataset: PD_modelling_dataset.xlsx

Data Dictionary:

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 current date .
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 date .
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 current date .
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 got generated in the last three months.
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 - Unpaid.
max_paid_inv_0_12m	The maximum credit card bill amount that has been paid by the customer in the last one year.
max_paid_inv_0_24m	The maximum credit card bill amount that has been paid by the customer in the last two years.
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.

Figure 1

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].

1. Objective: The primary objective is to develop a predictive model that can accurately forecast the likelihood of customers defaulting on their credit card payments. This prediction is crucial for assessing the risk associated with lending money to customers.

2. Dataset: The dataset contains various variables related to credit card accounts, including purchase history, delinquency information, and customer demographics. These variables serve as input features for the predictive model.

3. PD Modeling: The problem falls under the category of Probability of Default (PD) modeling. PD modeling aims to assess the riskiness of customers by estimating the probability of them defaulting on their credit obligations. It helps financial institutions evaluate the creditworthiness of customers and manage the potential losses associated with defaults.

4. Risk Assessment: The prediction of default probability assists in understanding the risk exposure of the credit card company. By identifying high-risk customers, the company can take proactive measures such as adjusting credit limits, offering alternative payment plans, or taking collection actions to mitigate potential losses.

5. Criticality: Predicting default probability is a critical aspect of risk management for any organization that extends credit, whether it's for secured or unsecured loans. It directly impacts the company's financial health and profitability by influencing lending decisions and risk mitigation strategies.

Overall, the goal of the project is to develop a robust predictive model that can effectively estimate the probability.

SUMMARY:

a) Defining Problem Statement:

The problem statement revolves around the need to predict the probability of default for credit card customers. Specifically, the aim is to develop a supervised learning model that accurately forecasts whether a customer will pay their credit card bill or default on the payment. This predictive model is essential for risk assessment and management in the lending industry.

b) Need of the Study/Project:

The study/project is necessitated by the critical importance of managing credit risk for financial institutions, particularly credit card companies. Default prediction plays a pivotal role in assessing the creditworthiness of customers and minimizing potential losses associated with defaults. By accurately predicting default probability, credit card companies can make informed decisions regarding lending practices, credit limits, and risk mitigation strategies.

c) Understanding Business/Social Opportunity:

The project presents a significant business opportunity for credit card companies to enhance their risk management practices and optimize their lending operations. By leveraging predictive modeling techniques, credit card companies can identify high-risk customers early in the customer lifecycle and tailor their approach to minimize default rates. Moreover, effective default prediction contributes to maintaining the stability and sustainability of the credit industry, which benefits both financial institutions and consumers.

Data Report:

	A	B	C	D	E	F	G	H	I	J	K
1	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
2	4567129	0	0	0	0	0	0	1	1	NA	1
3	2635118	0	0	0	0	0	NA	1	1	1	1
4	4804232	0	0	0	0	0	NA	NA	NA	NA	NA
5	1442693	0	0	NA	NA	NA	NA	NA	NA	NA	NA
6	4575322	0	0	0	0	0	NA	NA	NA	NA	NA
7	1534132	0	0	0	0	0	NA	NA	NA	NA	NA
8	1179589	0	0	0	142	0	0	1	2	2	1
9	2182448	0	57229	0	0	0	0.232244231	1	1	1	1
10	1661559	0	148922	0	47	0	0.969055089	1	2	2	2
11	4628751	0	0	0	0	0	NA	NA	NA	NA	NA
12	2878722	0	0	0	0	0	NA	NA	NA	NA	NA
13	1138988	0	0	NA	NA	NA	NA	NA	NA	NA	NA
14	3529979	0	11205	0	0	0	0	1	2	1	1
15	4878915	0	0	0	0	0	NA	NA	NA	NA	NA
16	2722501	0	0	0	0	0	NA	NA	NA	NA	NA
17	3613695	0	0	0	0	0	NA	NA	NA	NA	NA
18	4299412	0	0	0	0	0	NA	NA	NA	NA	NA
19	2029453	0	0	NA	NA	NA	NA	NA	NA	NA	NA
20	4828159	0	0	0	0	0	NA	NA	NA	NA	NA

Figure 2

	P	Q	R	S	T	U	V	W	X	Y	Z
1	merchant_category	merchant_group	has_paid	max_paid_inv_0_12m	max_paid_inv_0_24m	name_in_email	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_12
2	Dietary supplements	Health & Beauty	TRUE	31638	31638	no_match	0.153846154	2	0	0	13
3	Books & Magazines	Entertainment	TRUE	13749	13749	F+L	0	0	0	0	9
4	Diversified entertainment	Entertainment	TRUE	29890	29890	L1+F	0.071428571	1	0	0	11
5	Diversified entertainment	Entertainment	TRUE	40040	40040	F1+L	0.03125	1	0	0	31
6	Electronic equipment & Related accessories	Electronics	TRUE	7100	7100	F+L	0	0	0	0	1
7	Dietary supplements	Health & Beauty	FALSE	0	0	L1+F	NA	0	0	0	0
8	Concept stores & Miscellaneous	Leisure									
9	Diversified entertainment	Entertainment	TRUE	8655	9645	F	0.083333333	20	0	0	215
10	Diversified entertainment	Entertainment	TRUE	6075	9090	Nick	0.818181818	9	0	0	3
11	Diversified entertainment	Entertainment	TRUE	36985	36985	Nick	0	0	0	0	5
12	Concept stores & Miscellaneous	Leisure									
13	Diversified entertainment	Entertainment	FALSE	0	0	F+L	NA	0	0	0	0
14	Diversified entertainment	Entertainment	TRUE	6180	21080	L1+F	0	0	0	0	3
15	Youthful Shoes & Clothing	Clothing & Shoes	TRUE	5580	5580	F+L	0	0	0	0	5
16	General Shoes & Clothing	Clothing & Shoes	TRUE	12265	12265	F+L	0	0	0	0	5
17	Books & Magazines	Entertainment	FALSE	0	0	F	NA	0	0	0	0
18	Prints & Photos	Leisure									
19	Books & Magazines	Entertainment	TRUE	6010	6010	F+L	0	0	0	0	3
20	General Shoes & Clothing	Clothing & Shoes	TRUE	3785	4485	L1+F	0	0	0	0	22

Figure 3

The dataset exhibits numerous instances of zeros, which could either indicate actual values of zero or represent missing data points. Additionally, there are instances of NA values and

blank cells throughout the dataset. However, the exact reason for these zeros, NA values, and blank cells is not clear.(Refer figure 2,3)

Regarding data collection, there is a lack of information about how the data was gathered. Understanding the temporal context of data collection is crucial, as it provides insights into trends over time. Knowing the frequency of data collection is also important, as it impacts the level of detail and granularity available for analysis. Furthermore, understanding the methodological approach used for data acquisition is essential. For example, in the case of credit card data, this could involve capturing transactional data, payment histories, and account statuses from internal systems or databases, as well as potentially supplementing with external sources such as surveys or customer interactions. Knowing the methodology used for data collection helps to assess the comprehensiveness and reliability of the dataset for business analysis purposes.

Inspection of data (rows, columns, descriptive details)

Upon inspecting the dataset, it consists of 99,978 rows and 36 columns. The data types across the columns include 33 columns with numerical values (float64) and 3 columns with object data types. It was observed that one row is duplicated, indicating potential data redundancy. **Refer figure4 &5**

	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
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
2	4804232.0	0.0	0.0	0.0	0.0	0.0
3	1442693.0	0.0	0.0	NaN	NaN	NaN
4	4575322.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...
99974	4648093.0	NaN	56102.0	0.0	0.0	0.0
99975	1247657.0	NaN	0.0	0.0	0.0	0.0
99976	NaN	NaN	NaN	NaN	NaN	NaN
99977	NaN	NaN	NaN	NaN	NaN	NaN
99978	0.0	10000.0	0.0	11836.0	11836.0	11836.0

99979 rows × 36 columns

Figure 4

```
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   userid                           99976 non-null  float64
 1   default                          89976 non-null  float64
 2   acct_amt_added_12_24m            99976 non-null  float64
 3   acct_days_in_dc_12_24m           88140 non-null  float64
 4   acct_days_in_rem_12_24m          88140 non-null  float64
 5   acct_days_in_term_12_24m         88140 non-null  float64
 6   acct_incoming_debt_vs_paid_0_24m 40661 non-null  float64
 7   acct_status                      45603 non-null  float64
 8   acct_worst_status_0_3m           45603 non-null  float64
 9   acct_worst_status_12_24m         33215 non-null  float64
 10  acct_worst_status_3_6m           42274 non-null  float64
 11  acct_worst_status_6_12m          39626 non-null  float64
 12  age                              99976 non-null  float64
 13  avg_payment_span_0_12m           76140 non-null  float64
 14  avg_payment_span_0_3m            50671 non-null  float64
 15  merchant_category                99976 non-null  object
 16  merchant_group                   99967 non-null  object
 17  has_paid                         88942 non-null  float64
 18  max_paid_inv_0_12m               88942 non-null  float64
 19  max_paid_inv_0_24m               88942 non-null  float64
 20  name_in_email                    88942 non-null  object
 21  num_active_div_by_paid_inv_0_12m 70051 non-null  float64
 22  num_active_inv                   88942 non-null  float64
 23  num_arch_dc_0_12m                88942 non-null  float64
 24  num_arch_dc_12_24m               88942 non-null  float64
 25  num_arch_ok_0_12m                88942 non-null  float64
 26  num_arch_ok_12_24m               88942 non-null  float64
 27  num_arch_rem_0_12m               88942 non-null  float64
 28  status_max_archived_0_6_months   88942 non-null  float64
 29  status_max_archived_0_12_months  88942 non-null  float64
 30  status_max_archived_0_24_months  88942 non-null  float64
 31  recovery_debt                    88942 non-null  float64
 32  sum_capital_paid_acct_0_12m      88942 non-null  float64
 33  sum_capital_paid_acct_12_24m     88942 non-null  float64
 34  sum_paid_inv_0_12m               88942 non-null  float64
 35  time_hours                       88942 non-null  float64
dtypes: float64(33), object(3)
```

Figure 5

Furthermore, certain columns such as "userid" and "name_in_email" are deemed unnecessary for the analysis at hand. Additionally, the last row in the dataset(fig 6) contains anomalies, with illogical values observed in key fields. For instance, the "default" value is recorded as -10,000, "userid" is listed as 0, and "acct_status" has a numerical value instead of representing the intended account status, which should be categorized as either active or inactive.

	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_incomi
99977	NaN	NaN	NaN	NaN	NaN	NaN	
99978	0.0	10000.0	0.0	11836.0	11836.0	11836.0	

Figure 6

The "acct_status" column, which is intended to represent the current status of the account, is expected to have values of either 0 or 1, where 1 denotes an active account and 0 denotes an inactive one. However, upon examination of the dataset, it appears that there are additional values present, such as 2.0, 3.0, and 4.0, which are inconsistent with the expected binary representation.(fig 7)

```
acct_status
1.0    43693
2.0     1900
3.0        7
4.0        3
```

Figure 7

```
Number of null values in acct_status: 54374
```

Figure 8

This inconsistency suggests that the reliability of the "acct_status" column is low, as it does not adhere to the expected values and may contain erroneous or unreliable data. Consequently, caution should be exercised when analyzing or utilizing this column for business decision-making purposes. It may be necessary to further investigate the data collection process or perform data cleaning procedures to rectify these inconsistencies and improve the reliability of the column for subsequent analysis.

Ensuring data integrity and logical consistency is vital for meaningful business analysis. Identifying and addressing such anomalies is essential to ensure the reliability and accuracy of insights derived from the dataset.

The "merchant_category" variable contains a large number of unique categories, totalling 57 distinct fields. Analyzing such a large number of categories can be challenging and may result in a complex and computationally intensive analysis, particularly when performing clustering on the dataset.

merchant_category	
Diversified entertainment	38614
Youthful Shoes & Clothing	11755
Books & Magazines	9363
General Shoes & Clothing	4597
Concept stores & Miscellaneous	4406
Sports gear & Outdoor	3712
Dietary supplements	3101
Diversified children products	2994
Diversified electronics	1844
Prints & Photos	1675
Children Clothes & Nurturing products	1500
Pet supplies	1315
Electronic equipment & Related accessories	1037
Jewelry & Watches	911
Hobby articles	910
Prescription optics	899
Body & Hair Care	857
Automotive Parts & Accessories	828
Diversified Health & Beauty products	795
Diversified Home & Garden products	707
Video Games & Related accessories	673
Decoration & Art	665
Cosmetics	640
Dating services	629
Diversified erotic material	623
Children toys	614
Tools & Home improvement	456
Personal care & Body improvement	371
Furniture	370
Pharmaceutical products	369
Fragrances	324
Digital services	300
Adult Shoes & Clothing	292
Food & Beverage	289
Travel services	179
Costumes & Party supplies	147
Music & Movies	129
Wheels & Tires	109
Collectibles	107
Kitchenware	97
Household electronics (whitegoods/appliances)	90
Underwear	84
Erotic Clothing & Accessories	83
Non	71
Musical Instruments & Equipment	68
Tobacco	66
Safety products	57
Diversified Jewelry & Accessories	52
Car electronics	46
Sex toys	41
Plants & Flowers	37
Bags & Wallets	24
Office machines & Related accessories (excl. computers)	17
Cleaning & Sanitary	14
Event tickets	13
Wine	9
Education	1

Figure 9

merchant_group	
Entertainment	48779
Clothing & Shoes	16728
Leisure	11025
Health & Beauty	7356
Children Products	5108
Home & Garden	3718
Electronics	3034
Intangible products	1122
Jewelry & Accessories	1058
Automotive Products	937
Erotic Materials	747
Food & Beverage	355

Figure 10

To address this issue and simplify the analysis, we propose focusing on the "merchant_group" variable instead. This variable contains 12 sub-categories, which are likely to provide a more manageable and interpretable basis for clustering the dataset.

default	
0.0	88688
1.0	1288

Figure 11

The distribution of the "default" variable indicates a highly imbalanced dataset, with a majority of customers (88,688) classified as non-defaulters (0) and a much smaller proportion (1,288) classified as defaulters (1). This imbalance poses challenges for predictive modeling, particularly for algorithms that may be biased towards the majority class. Therefore, strategies such as resampling techniques or using algorithms that are robust to class imbalance need to be considered during model development to ensure accurate predictions for both classes. Additionally, it highlights the importance of evaluating model performance metrics beyond accuracy, such as precision, recall, and F1-score, to assess the effectiveness of the model in correctly identifying default cases.

In the dataset, several columns have more than 40% null values, indicating significant missing data. These columns are crucial to identify as they may require special handling during analysis or modeling. Here are the columns with above 40% null values:

default	10001
acct_amt_added_12_24m	1
acct_days_in_dc_12_24m	11837
acct_days_in_rem_12_24m	11837
acct_days_in_term_12_24m	11837
acct_incoming_debt_vs_paid_0_24m	59316
acct_worst_status_0_3m	54374
acct_worst_status_12_24m	66762
acct_worst_status_3_6m	57703
acct_worst_status_6_12m	60351
age	1
avg_payment_span_0_12m	23837
avg_payment_span_0_3m	49306
merchant_category	1
merchant_group	10
has_paid	0
max_paid_inv_0_12m	11035
max_paid_inv_0_24m	11035
num_active_div_by_paid_inv_0_12m	29926
num_active_inv	11035
num_arch_dc_0_12m	11035
num_arch_dc_12_24m	11035
num_arch_ok_0_12m	11035
num_arch_ok_12_24m	11035
num_arch_rem_0_12m	11035
status_max_archived_0_6_months	11035
status_max_archived_0_12_months	11035
status_max_archived_0_24_months	11035
recovery_debt	11035
sum_capital_paid_acct_0_12m	11035
sum_capital_paid_acct_12_24m	11035
sum_paid_inv_0_12m	11035
time_hours	11035

Figure 12

default	10.00
acct_amt_added_12_24m	0.00
acct_days_in_dc_12_24m	11.84
acct_days_in_rem_12_24m	11.84
acct_days_in_term_12_24m	11.84
acct_incoming_debt_vs_paid_0_24m	59.33
acct_worst_status_0_3m	54.39
acct_worst_status_12_24m	66.78
acct_worst_status_3_6m	57.72
acct_worst_status_6_12m	60.36
age	0.00
avg_payment_span_0_12m	23.84
avg_payment_span_0_3m	49.32
merchant_category	0.00
merchant_group	0.01
has_paid	0.00
max_paid_inv_0_12m	11.04
max_paid_inv_0_24m	11.04
num_active_div_by_paid_inv_0_12m	29.93
num_active_inv	11.04
num_arch_dc_0_12m	11.04
num_arch_dc_12_24m	11.04
num_arch_ok_0_12m	11.04
num_arch_ok_12_24m	11.04
num_arch_rem_0_12m	11.04
status_max_archived_0_6_months	11.04
status_max_archived_0_12_months	11.04
status_max_archived_0_24_months	11.04
recovery_debt	11.04
sum_capital_paid_acct_0_12m	11.04
sum_capital_paid_acct_12_24m	11.04
sum_paid_inv_0_12m	11.04
time_hours	11.04
dtype: float64	

Figure 13

Refer fig 12 & 13
- acct_incoming_debt_vs_paid_0_24m: 59,316 null values out of 99,979 entries
- acct_worst_status_0_3m: 54,374 null values out of 99,979 entries
- acct_worst_status_12_24m: 66,762 null values out of 99,979 entries(59.33%)
- acct_worst_status_3_6m: 57,703 null values out of 99,979 entries
- acct_worst_status_6_12m: 60,351 null values out of 99,979 entries
- avg_payment_span_0_3m: 49,306 null values out of 99,979 entries

These columns require careful consideration during data analysis and modeling. Depending on the specific context and objectives of the analysis, strategies such as imputation or exclusion of these columns may be employed to address the missing data issue. Additionally, it's essential to assess the impact of missing values on the overall analysis and interpret the results accordingly.

We are dropping columns with more than 23% missing values to improve the quality of our analysis and predictions. This decision is based on the understanding that columns with a high percentage of missing values may not contribute significantly to the analysis and could potentially introduce noise or bias into our models. By removing these columns, we aim to streamline our dataset and focus on the most informative variables for our analysis.

```
Index(['default', 'acct_amt_added_12_24m', 'acct_days_in_dc_12_24m',
       'acct_days_in_rem_12_24m', 'acct_days_in_term_12_24m', 'age',
       'merchant_category', 'merchant_group', 'has_paid', 'max_paid_inv_0_12m',
       'max_paid_inv_0_24m', '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'],
      dtype='object')
```

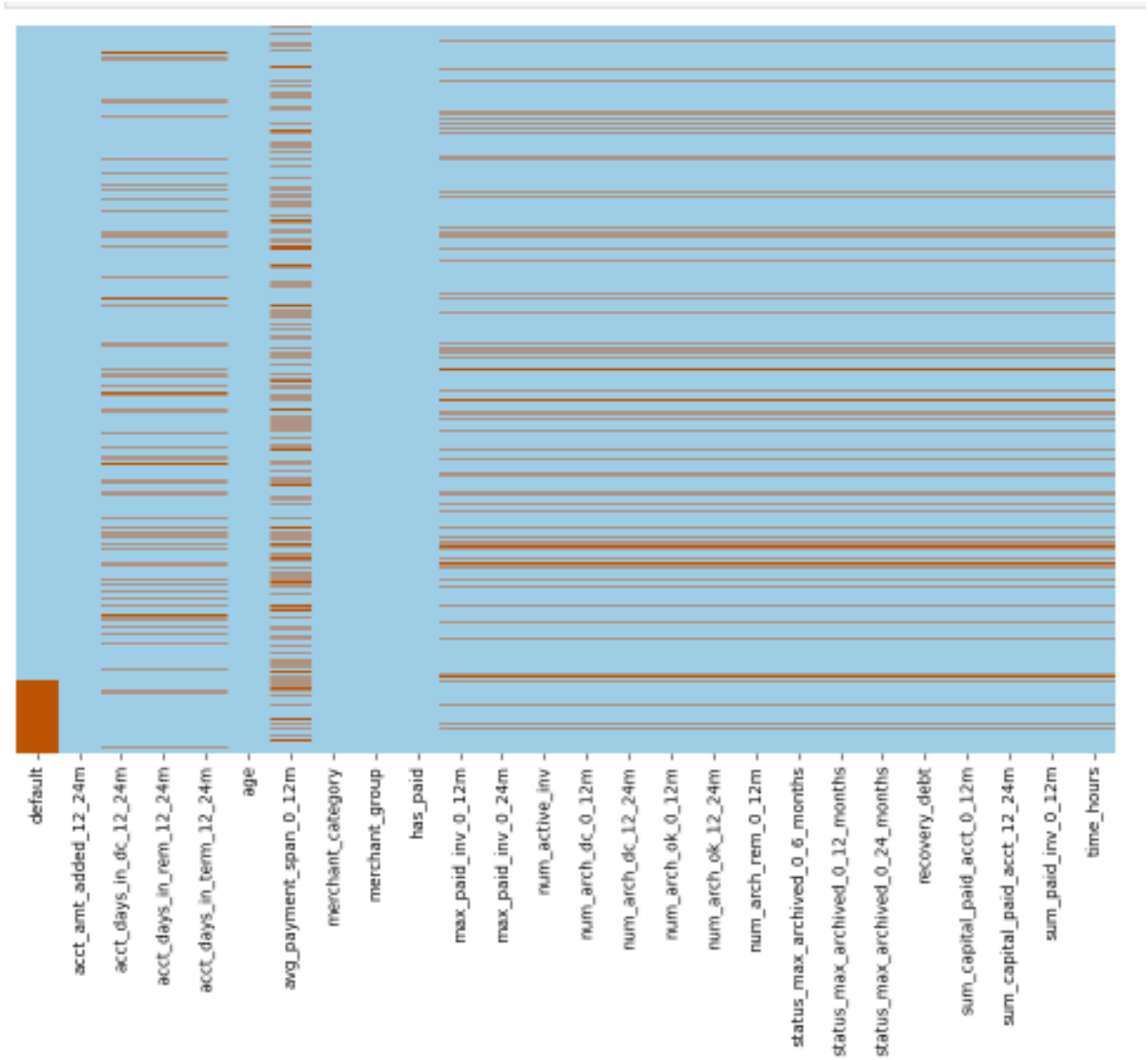


Figure 14

Fig14, After removing columns with more than 23% missing values, the missing value graph appears improved. This step enhances the quality of our data analysis by focusing on the most relevant variables, which is essential for making informed business decisions.

```
has_paid                          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
age                               0
max_paid_inv_0_12m                0
max_paid_inv_0_24m                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
recovery_debt                     0
status_max_archived_0_12_months   0
status_max_archived_0_24_months   0
status_max_archived_0_6_months    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
```

Post Imputation 1

	has_paid	acct_amt_added_12_24m	acct_days_in_dc_12_24m	acct_days_in_rem_12_24m	acct_days_in_term_12_24m	age	max_paid_inv_0_12m	max_paid_inv_0_24m	num_active_inv	num_arch_dc_0_12m	...	num_arch_ok_12_24m	nur
0	1.0	0.0	0.0	0.0	0.0	20.0	31638.0	31638.0	2.0	0.0	...	14.0	
1	1.0	0.0	0.0	0.0	0.0	50.0	13749.0	13749.0	0.0	0.0	...	19.0	
2	1.0	0.0	0.0	0.0	0.0	22.0	29890.0	29890.0	1.0	0.0	...	0.0	
3	1.0	0.0	0.0	0.0	0.0	36.0	40040.0	40040.0	1.0	0.0	...	21.0	
4	1.0	0.0	0.0	0.0	0.0	25.0	7100.0	7100.0	0.0	0.0	...	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	
99972	1.0	0.0	0.0	0.0	0.0	44.0	4740.0	4740.0	0.0	0.0	...	3.0	
99973	1.0	45671.0	0.0	20.0	0.0	24.0	1200.0	1200.0	0.0	0.0	...	0.0	
99974	1.0	56102.0	0.0	0.0	0.0	31.0	15000.0	15000.0	0.0	0.0	...	1.0	
99975	1.0	0.0	0.0	0.0	0.0	41.0	13246.0	14817.0	0.0	0.0	...	2.0	
99976	1.0	0.0	0.0	0.0	0.0	34.0	6170.0	7720.0	0.0	0.0	...	2.0	

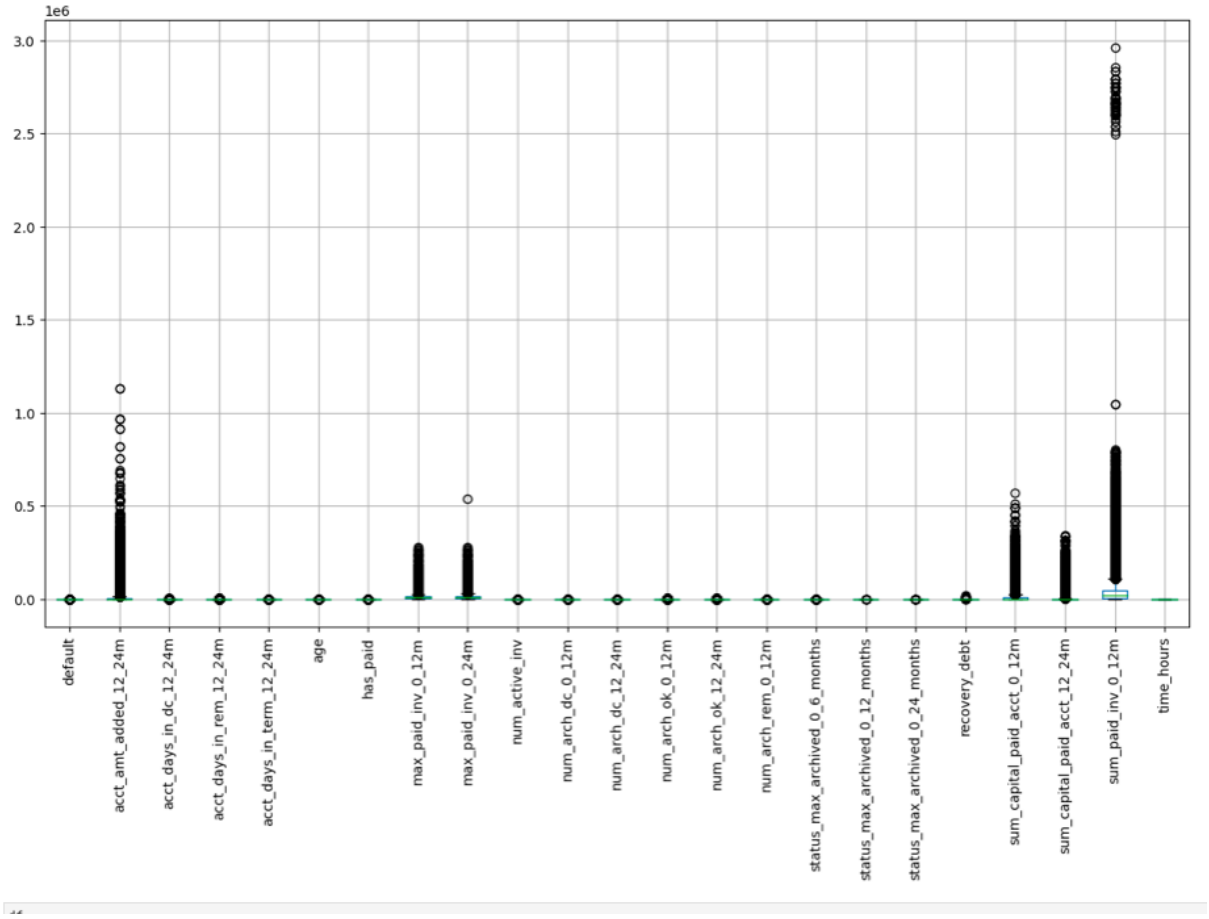
99977 rows × 22 columns

	has_paid	acct_amt_added_12_24m	acct_days_in_dc_12_24m	acct_days_in_rem_12_24m	acct_days_in_term_12_24m	age	max_paid_inv_0_12m	max_paid_inv_0_24m	num_active_inv	num_arch_dc_0_12m	...
count	99977.000000	9.997700e+04	99977.000000	99977.000000	99977.000000	99977.000000	99977.000000	99977.000000	99977.000000	99977.000000	...
mean	0.880573	1.225503e+04	0.196635	4.447353	0.252928	36.016264	9010.408594	11011.374336	0.556898	0.055703	...
std	0.324292	3.548133e+04	5.453928	21.529593	2.752561	13.001242	12934.597693	14601.300041	1.531669	0.360942	...
min	0.000000	0.000000e+00	0.000000	0.000000	0.000000	18.000000	0.000000	0.000000	0.000000	0.000000	...
25%	1.000000	0.000000e+00	0.000000	0.000000	0.000000	25.000000	2990.000000	4250.000000	0.000000	0.000000	...
50%	1.000000	0.000000e+00	0.000000	0.000000	0.000000	34.000000	6170.000000	7720.000000	0.000000	0.000000	...
75%	1.000000	4.937000e+03	0.000000	0.000000	0.000000	45.000000	10570.000000	12785.000000	1.000000	0.000000	...
max	1.000000	1.128775e+06	365.000000	365.000000	97.000000	100.000000	279000.000000	538500.000000	47.000000	17.000000	...

8 rows × 22 columns

Imputation 2

All missing values have been imputed using the **median strategy**. This approach ensures that the imputed values are robust and less sensitive to outliers, thereby providing a more reliable dataset for analysis and modeling.



The **outliers** in the data are not treated because the values are deemed reliable. Treating outliers may distort the true distribution of the data and lead to loss of valuable information. By retaining the outliers, we can better understand the variability and range of the data, which can be important for making accurate predictions and drawing meaningful insights.

Exploratory data analysis:

Univariate Analysis:

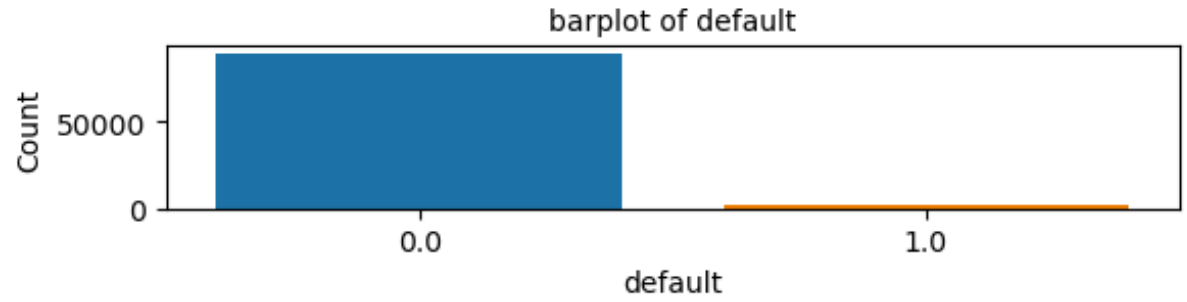


Figure 15

The univariate analysis of the target variable "Default" reveals a significant class imbalance, with defaulters representing only 1% of the total dataset. This imbalance poses a challenge for predictive modeling, as the minority class may be underrepresented, leading to biased model performance. Addressing this imbalance will be crucial to ensure the development of robust predictive models. Techniques such as resampling or algorithmic approaches like ensemble methods can help mitigate the effects of class imbalance during model training.

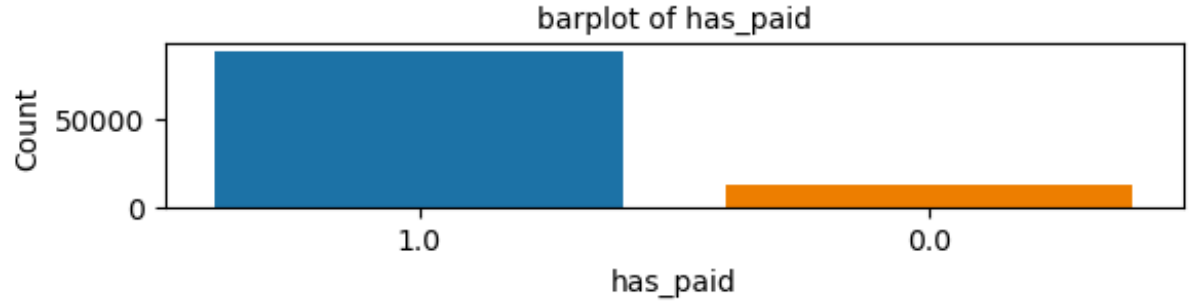


Figure 16

The variable "Has Paid" indicates whether the customer has paid the current credit card bill. Analysis reveals that approximately 87% of users have paid their current credit card bill, while the remaining 13% have not. This insight provides an understanding of the payment behavior among credit card users in the dataset, highlighting the majority who fulfill their payment obligations and a minority who may be delinquent. Understanding payment behavior is crucial for risk assessment and managing credit card portfolios effectively.

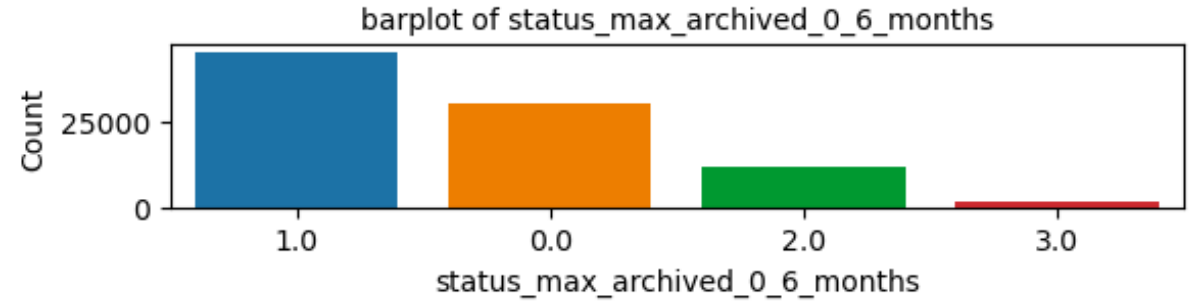


Figure 17

The variable "status_max_archived_0_6_months" indicates the maximum number of times an account has been in archived status over the last 6 months, with a mean value of 0.82. Most accounts show a low frequency of archived status, suggesting relatively stable financial behavior. Monitoring accounts with higher archived status frequency can help in identifying potential default risks and implementing proactive measures to mitigate them.

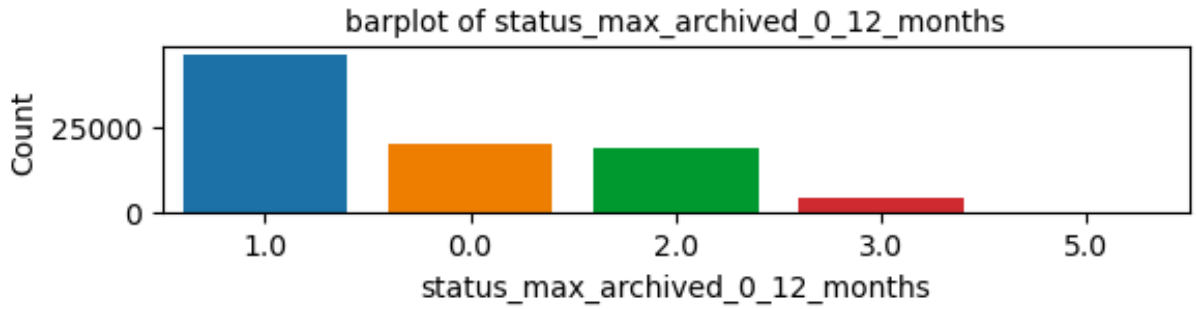


Figure 18

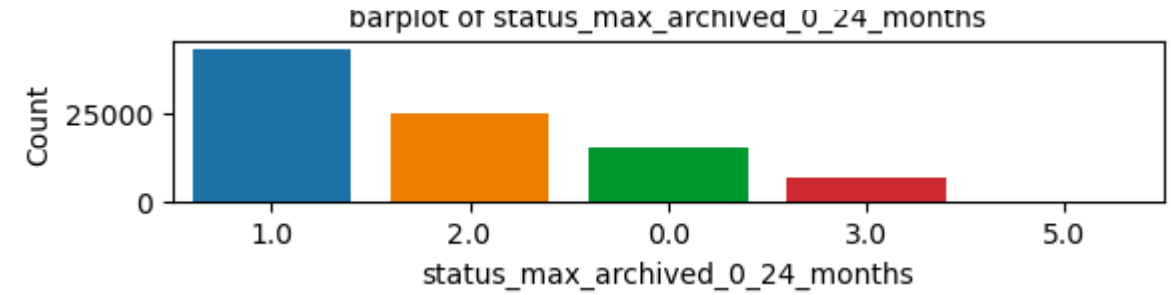


Figure 19

The variable "status_max_archived_0_24_months" reflects the maximum frequency of archived status occurrences within the past 24 months. Analyzing the distribution, approximately 48% of users have encountered archived status at least once, while 17% have not experienced this status during the same period. Notably, 28% of users have faced archived status twice, and only 7% have encountered it three times, with a negligible proportion having experienced it five times. Understanding these patterns can assist in assessing the financial stability of users and implementing targeted interventions to address potential risks of default.

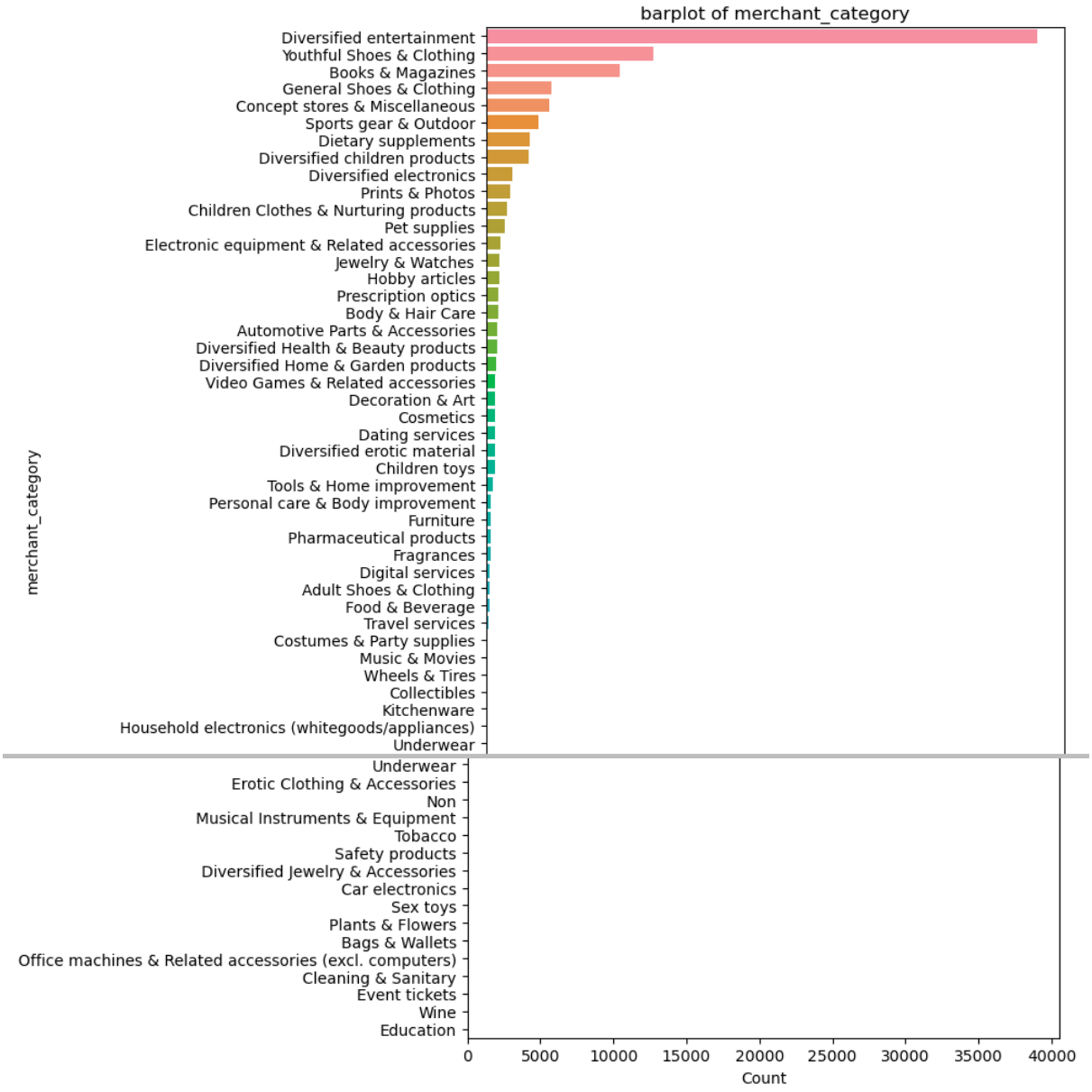


Figure 20

The variable "merchant_category" encompasses 57 distinct categories, with "Diversified entertainment" being the most prevalent, followed by "Youthful shoes and clothing". This insight suggests a diverse range of spending preferences among credit card users, highlighting opportunities for targeted marketing campaigns and product offerings tailored to these categories. Understanding consumer behavior across various merchant categories is essential for optimizing business strategies and enhancing customer engagement within specific market segments.

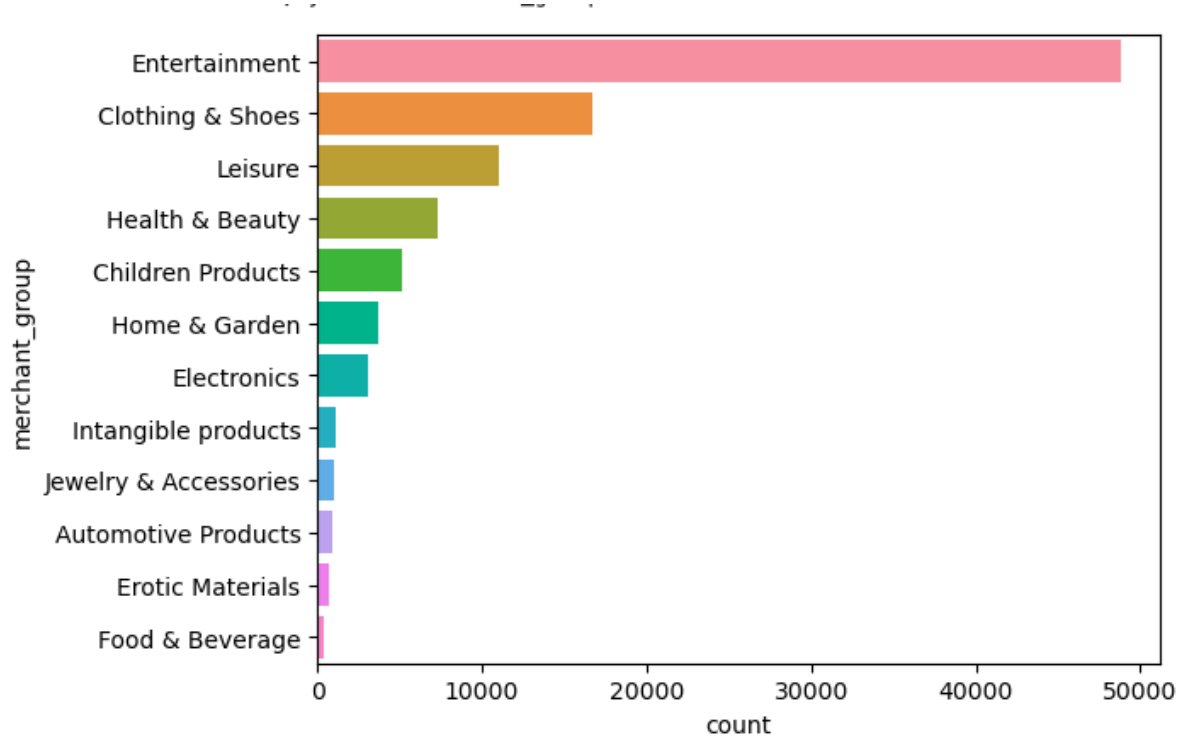


Figure 21

The "merchant_group" variable comprises 12 distinct groups, with "Entertainment" being the largest group, accounting for 48,779 instances. This observation indicates a significant proportion of credit card transactions are related to entertainment-related merchants. Understanding the distribution of transactions across different merchant groups can inform business decisions regarding partnerships, promotional strategies, and customer targeting efforts tailored to specific merchant categories.(fig 21)

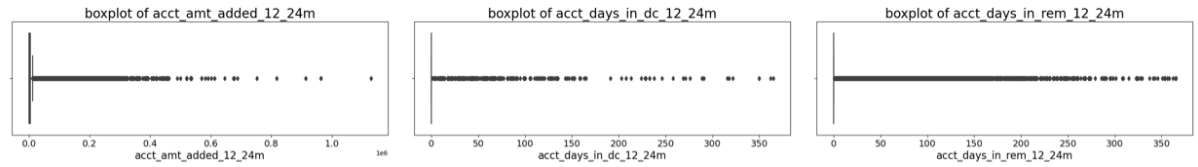


Figure 22

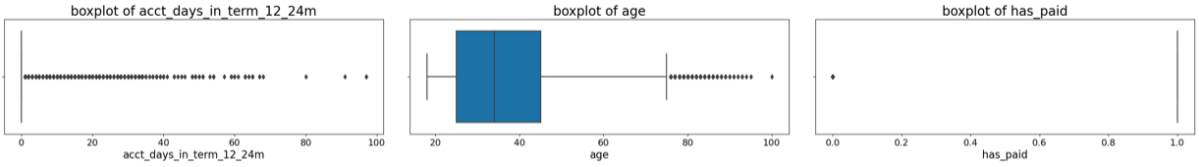


Figure 23

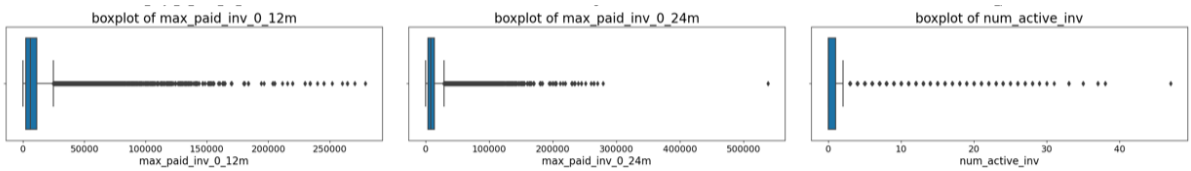


Figure 24

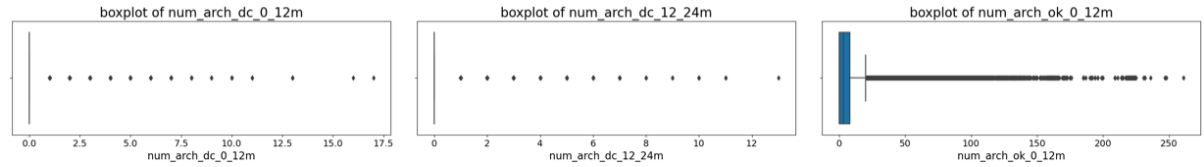


Figure 25

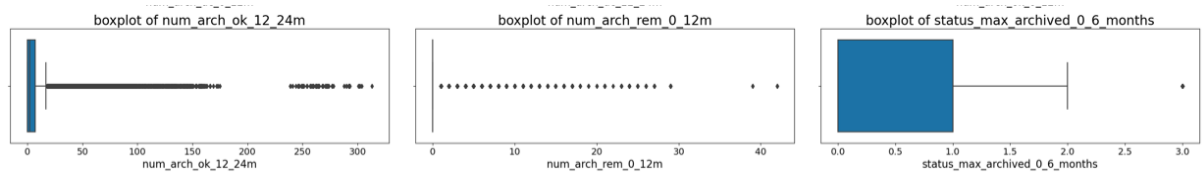


Figure 26

Figure 27

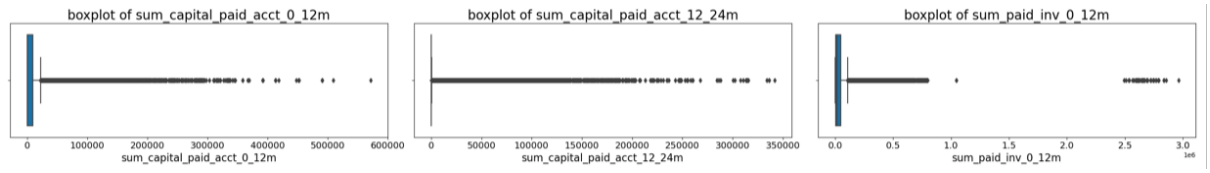


Figure 28

The box plots provide valuable insights into the distribution, central tendency, and variability of different dataset variables. They visually represent the median, interquartile range (IQR), outliers, and overall distribution shape for each variable. This information is crucial for understanding the spread and dispersion of data, identifying potential outliers or anomalies, and assessing the variability of variables across different categories or groups. Analyzing these box plots helps in making informed decisions about data preprocessing, identifying influential factors, and selecting appropriate modeling techniques for predictive analytics.

The box plot for age (fig 36)reveals key insights into the distribution of ages within the dataset:
- The median age is approximately 36 years old, indicating that half of the individuals in the dataset are younger than 36 and the other half are older.
- The interquartile range (IQR), represented by the box, spans from approximately 25 to 45 years old, capturing the middle 50% of the age distribution.
- The whiskers extend from the minimum to the maximum age values, excluding outliers. The minimum age is 18 years old, while the maximum age is 100 years old.
- There are some outliers beyond the whiskers, which are values that fall significantly outside the typical age range observed in the dataset. These outliers may represent unusual or extreme ages that warrant further investigation.

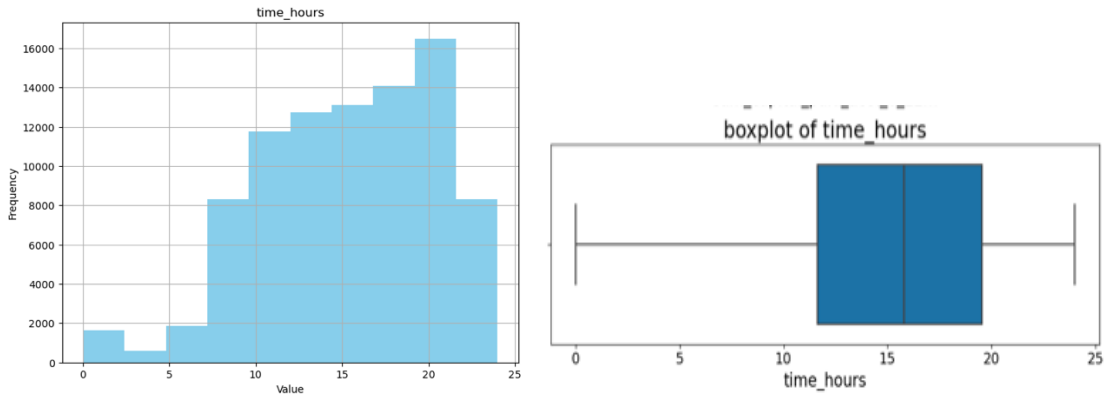


Figure 29

The box plot for the variable "time_hours" provides the following insights:
- The median time spent on purchases using the credit card is approximately 15.34 hours, indicating that half of the users spend less than this amount of time, while the other half spends more.
- The interquartile range (IQR), represented by the box, spans from approximately 11.63 to 19.55 hours, capturing the middle 50% of the distribution of time spent.
- The whiskers extend from the minimum to the maximum values, excluding outliers. The minimum time spent is nearly 0.0003 hours, while the maximum time spent is nearly 24 hours, suggesting a wide range of purchase durations.
- There are some outliers beyond the whiskers, which are values that fall significantly outside the typical range of time spent on purchases. These outliers may represent unusually short or long purchase durations that could be further investigated.

Bi-Variate Analysis:

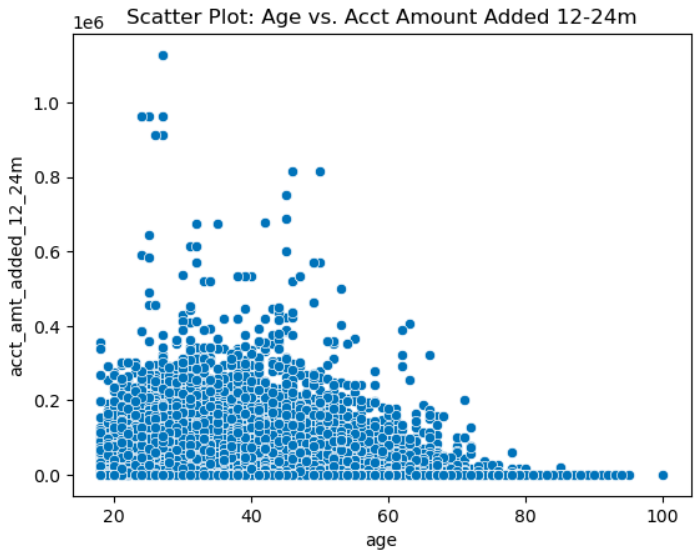


Figure 30

The bivariate analysis of the account amount added between 12 to 24 months reveals an interesting trend: individuals aged between 22 to 60 tend to deposit significantly higher amounts into their bank accounts compared to those above 60 years old. This suggests that the middle-aged population is more active in making deposits, possibly due to factors such as steady income, financial responsibilities, and long-term financial planning. Understanding this age-related pattern can help financial institutions tailor their services and marketing strategies to better cater to the needs and preferences of different age groups.

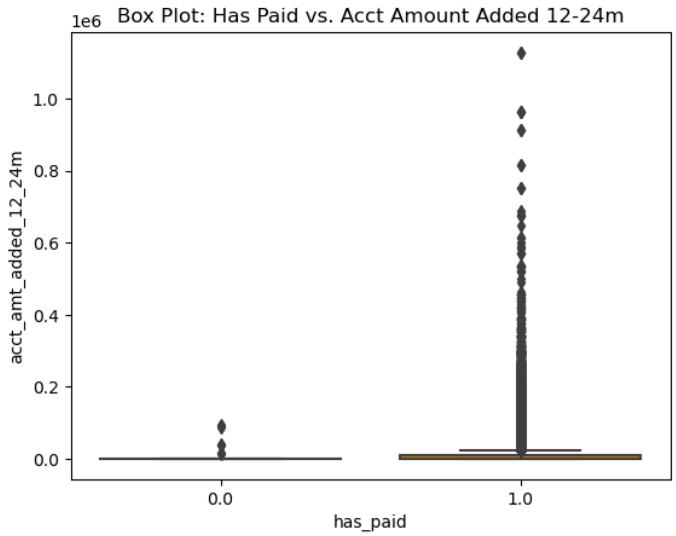


Figure 31

Individuals who frequently experience account status changes tend to maintain higher account balances.

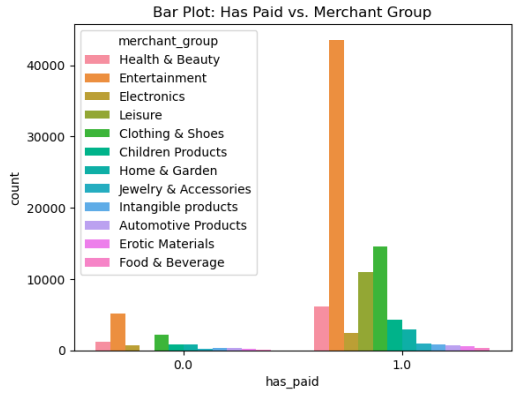


Figure 32

A bar plot depicting the paid status among different merchant groups reveals that the highest number of paid statuses are predominantly within the entertainment sector. Furthermore, it's evident that individuals with a paid status tend to allocate a significant portion of their expenditure towards leisure activities.

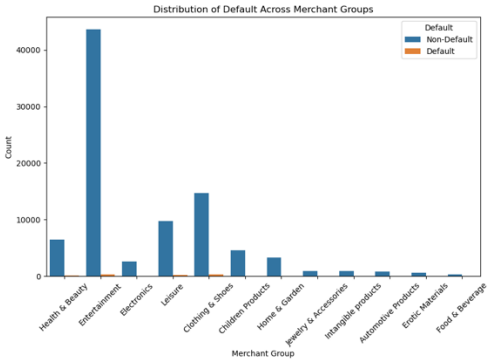


Figure 33

During a bivariate analysis comparing defaulters to different merchant groups, a clear pattern emerges indicating that defaulters primarily allocate their expenditures towards entertainment and clothing/shoes categories. However, due to the highly imbalanced data distribution within the target variable, there's a risk of misinterpretation or misjudgment.

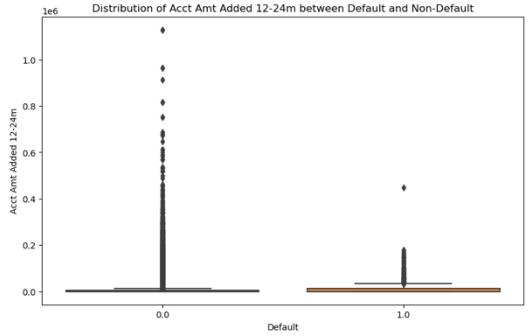


Figure 34

It's ironic to note that defaulters exhibit a higher amount added in the past 12 to 24 months compared to non-defaulters. Interestingly, within the defaulters' dataset, there are noticeable higher outliers, suggesting significant disparities in financial behavior or circumstances among this group.

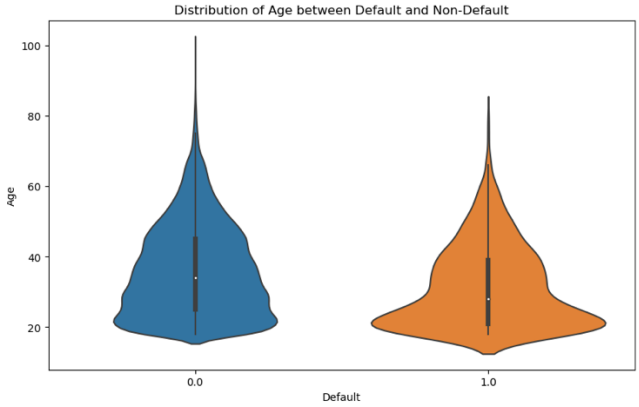


Figure 35

When comparing the age range between defaulters and non-defaulters, it becomes apparent that defaulters exhibit a broader spectrum of ages. Additionally, the mean age for defaulters tends to be higher. A higher concentration of defaulters is observed within the age range of 20 to 40, whereas non-defaulters are more concentrated between 20 to 30 years old.

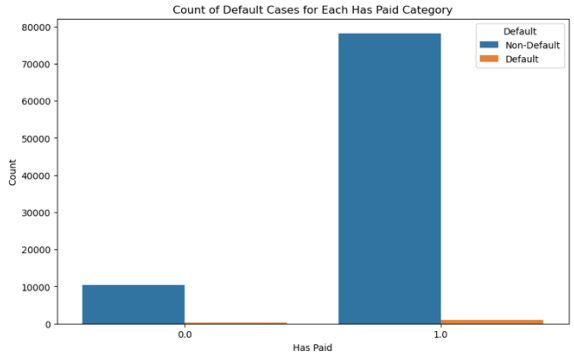


Figure 36

Due to the highly imbalanced nature of the target variable, drawing definitive conclusions regarding the status of "has paid" versus "default" becomes extremely challenging. Imbalanced data distributions can significantly skew analysis results and compromise the reliability of conclusions. Therefore, additional techniques such as resampling methods, advanced modeling algorithms, or careful feature engineering may be necessary to address this issue effectively.

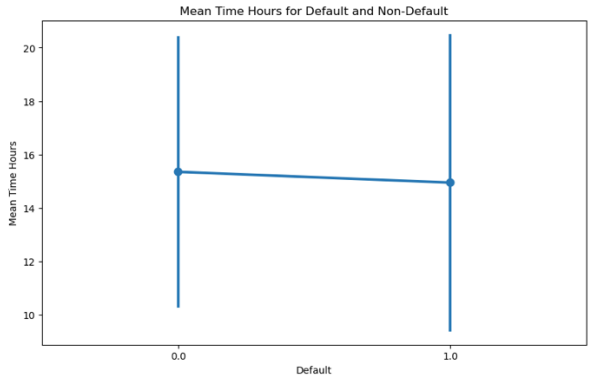


Figure 37

Analysis reveals that defaulters tend to spend a higher mean amount of time in hours compared to non-defaulters. This distinction suggests a potential correlation between time spent and default status, though further investigation would be needed to determine causality or any underlying factors influencing this relationship.

Pairplot :

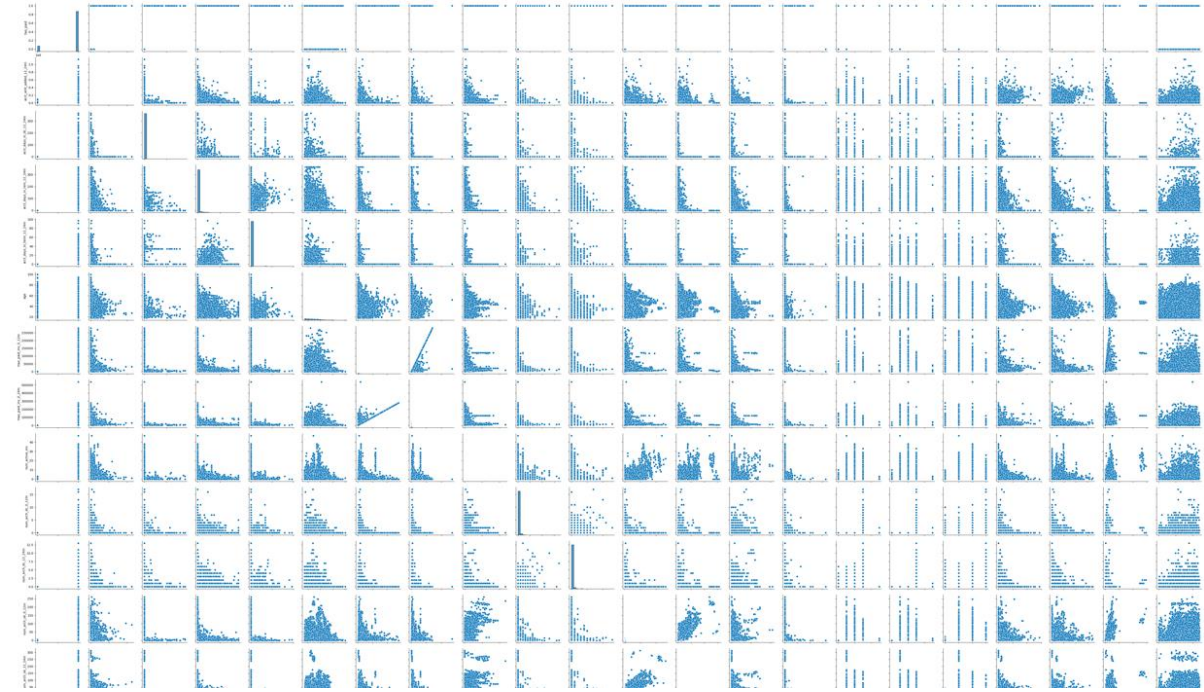


Figure 38

Performing a pair-plot visualization allows for a comprehensive examination of relationships between multiple variables simultaneously. This method enables the observation of patterns, trends, and potential correlations within the dataset. By incorporating various features such as default status, time hours, age, merchant group, and other relevant variables, the pair-plot visualization offers valuable insights into the interplay between these factors and their potential impact on default status.

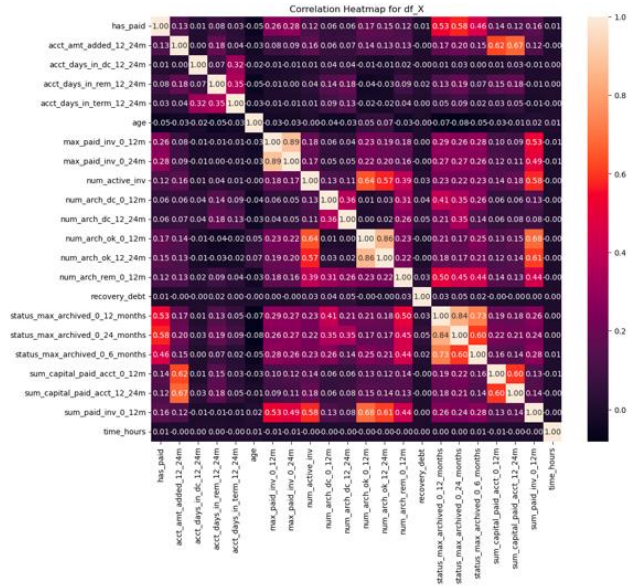


Figure 39

The heatmap visualization provides valuable insights into the correlation between different variables in the dataset. Correlation values such as 0.53, 0.58, 0.46, 0.62, 0.67, etc., indicate moderate to strong correlations between pairs of variables. To streamline the analysis and improve the predictive model, it's advisable to drop columns with high correlation values to

mitigate multicollinearity issues. Subsequently, conducting a variance inflation factor (VIF) analysis can help identify and address any remaining multicollinearity concerns, ensuring the robustness of the predictive model.

	variables	VIF
11	num_arch_ok_0_12m	5.398529
15	status_max_archived_0_12_months	5.323043
6	max_paid_inv_0_12m	5.209588
7	max_paid_inv_0_24m	4.839512
16	status_max_archived_0_24_months	4.397800
12	num_arch_ok_12_24m	4.048149
20	sum_paid_inv_0_12m	3.227103
17	status_max_archived_0_6_months	2.277198
1	acct_amt_added_12_24m	2.154620
19	sum_capital_paid_acct_12_24m	2.061729
8	num_active_inv	1.984793
13	num_arch_rem_0_12m	1.876782
18	sum_capital_paid_acct_0_12m	1.828353
0	has_paid	1.752470
9	num_arch_dc_0_12m	1.422306
10	num_arch_dc_12_24m	1.371009
4	acct_days_in_term_12_24m	1.277314
3	acct_days_in_rem_12_24m	1.233996
2	acct_days_in_dc_12_24m	1.118476
5	age	1.018101
14	recovery_debt	1.006133
21	time_hours	1.000291

Figure 40

After dropping columns with more than 23% missing values, we reduced the overall number of columns to 25, aiming to improve the dataset's quality. Subsequently, we conducted a variance inflation factor (VIF) analysis to identify variables with high correlation, thereby mitigating multicollinearity issues. As a result, we eliminated columns such as "num_arch_ok_0_12" and "max_paid_inv_0_12" due to their high VIF values, enhancing the robustness of our predictive model.

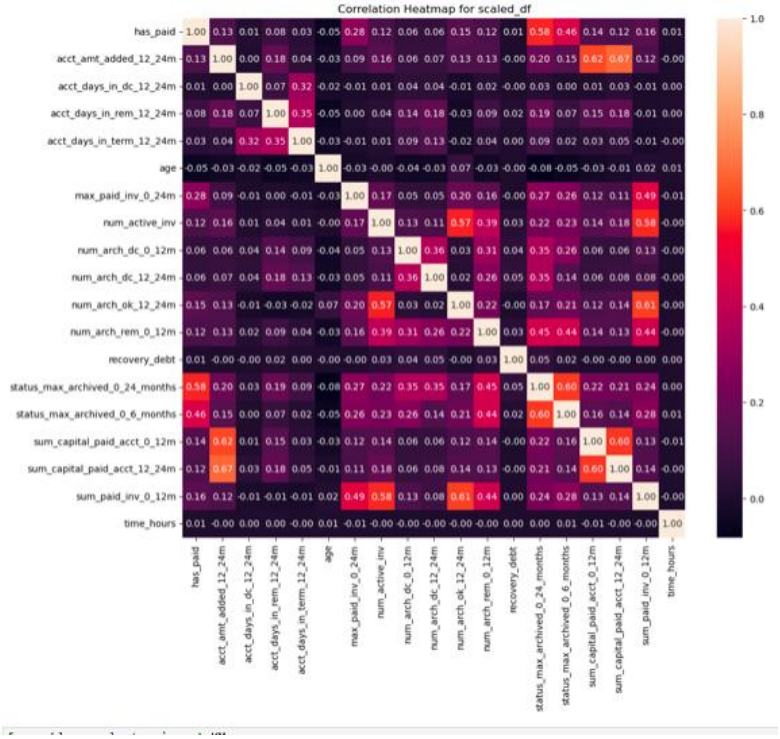


Figure 41

By reducing multicollinearity, we've not only improved the statistical integrity of our analysis but also streamlined the dataset for better predictive accuracy. This approach ensures that our model is more reliable and less susceptible to overfitting, ultimately enhancing our ability to make informed business decisions based on the data. Moreover, the subsequent heatmap analysis revealed a reduction in correlation among variables, indicating a clearer and more interpretable relationship between different aspects of our business data. This improved understanding allows us to pinpoint key factors driving certain outcomes or behaviors, enabling more effective strategic planning and resource allocation.

	has_paid	acct_amt_added_12_24m	acct_days_in_dc_12_24m	acct_days_in_rem_12_24m	acct_days_in_term_12_24m	age	max_paid_inv_0_12m	max_paid_inv_0_24m	num_active_inv	num_arch_dc_0_12m	...	num_arch_ok_12_24m
0	0.368273	-0.345395	-0.036054	-0.206570	-0.091889	-1.231909	1.749394	1.412664	0.942181	-0.154327	...	0.504774
1	0.368273	-0.345395	-0.036054	-0.206570	-0.091889	1.075575	0.366352	0.187493	-0.363591	-0.154327	...	0.833045
2	0.368273	-0.345395	-0.036054	-0.206570	-0.091889	-1.078076	1.614252	1.292948	0.289295	-0.154327	...	-0.414394
3	0.368273	-0.345395	-0.036054	-0.206570	-0.091889	-0.001251	2.398973	1.988095	0.289295	-0.154327	...	0.964358
4	0.368273	-0.345395	-0.036054	-0.206570	-0.091889	-0.847328	-0.147698	-0.267880	-0.363591	-0.154327	...	-0.414394
...	...	...	...	...	...	...	...	...	...	...	...	...
99972	0.368273	-0.345395	-0.036054	-0.206570	-0.091889	0.614078	-0.330156	-0.429510	-0.363591	-0.154327	...	-0.217425
99973	0.368273	0.941795	-0.036054	0.722388	-0.091889	-0.924244	-0.603842	-0.671955	-0.363591	-0.154327	...	-0.414394
99974	0.368273	1.235782	-0.036054	-0.206570	-0.091889	-0.385831	0.463070	0.273171	-0.363591	-0.154327	...	-0.348735
99975	0.368273	-0.345395	-0.036054	-0.206570	-0.091889	0.383330	0.327464	0.260637	-0.363591	-0.154327	...	-0.283084
99976	0.368273	-0.345395	-0.036054	-0.206570	-0.091889	-0.155083	-0.219599	-0.225418	-0.363591	-0.154327	...	-0.283084

99977 rows × 22 columns

Figure 42

We applied standard scaling to the data to ensure that all features are on a similar scale, thus preventing certain variables from dominating the analysis simply due to their larger magnitude. Standard scaling transforms the data such that it has a mean of 0 and a standard deviation of 1, which helps in improving the performance of certain machine learning algorithms, particularly those sensitive to the scale of the features.

By standardizing the data, we have effectively removed any inherent biases that might arise from differences in the measurement scales of various features. This ensures that our predictive model can make fair comparisons and accurate predictions based on the relative importance of each feature. Additionally, standard scaling helps in speeding up the convergence of optimization algorithms and can lead to more stable and reliable results during model training and evaluation.

Business insights from EDA:

a) **Data Imbalance in Business Perspective:**

In the business context, the highly imbalanced ratio of defaulters to non-defaulters (88,688 versus 1,288) poses significant challenges and implications. Firstly, it suggests that default instances are relatively rare compared to non-default instances, which can lead to biased model predictions that prioritize the majority class. In financial terms, this imbalance could indicate potential revenue loss and increased risk exposure due to defaults. Moreover, inaccurate predictions may result in suboptimal resource allocation, such as misjudging credit risk or ineffective debt recovery strategies. Therefore, addressing this data imbalance is crucial for businesses to develop robust risk management strategies, improve customer targeting, and enhance overall financial performance.

```
default
0.0    88688
1.0     1288
```

Figure 43

b) **Business Insights Using Clustering:**

Leveraging clustering, specifically K-means clustering, on the dataset's "merchant group" variable can provide valuable business insights. By grouping customers with similar spending behavior and default patterns, businesses can tailor their strategies more effectively. The identified clusters can guide targeted marketing efforts, product recommendations, and risk mitigation strategies. For example, businesses can personalize communication and offers based on the spending preferences and default likelihood of each cluster. Additionally, insights from cluster analysis can inform product development decisions, helping businesses align their offerings with the needs and preferences of different customer segments.

merchant_group	Clusters	
Automotive Products	0	264
	1	598
	2	75
Children Products	0	837
	1	3384
	2	887
Clothing & Shoes	0	2213
	1	11819
	2	2696
Electronics	0	664
	1	2012
	2	358
Entertainment	0	5193
	1	37129
	2	6457
Erotic Materials	0	184
	1	484
	2	79
Food & Beverage	0	80
	1	219
	2	56
Health & Beauty	0	1158
	1	5158
	2	1040
Home & Garden	0	847
	1	2366
	2	505
Intangible products	0	327
	1	701
	2	94
Jewelry & Accessories	0	173
	1	710
	2	175
Leisure	1	10907
	2	118

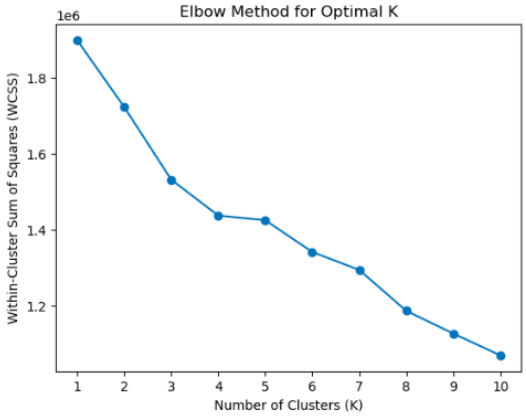


Figure 44

c) **Recommendations for Business Insights:**

Based on the insights derived from clustering analysis and the dataset, several actionable recommendations can be proposed:

- **Segment-specific Marketing Strategies**: Tailor marketing campaigns and promotions to each cluster's preferences and behaviors, optimizing customer engagement and conversion rates.
- **Risk Management Optimization**: Develop customized risk assessment models for each cluster, enabling more accurate credit scoring and proactive default prevention measures.
- **Product Portfolio Optimization**:Adjust product offerings and inventory management strategies to align with the spending patterns and preferences of different customer clusters, maximizing sales opportunities and customer satisfaction.
- **Customer Retention Strategies:** Implement targeted retention initiatives for at-risk customer segments, such as personalized loyalty programs or proactive customer support, to minimize churn and maximize lifetime value.
- **Continuous Monitoring and Adaptation**: Regularly monitor cluster dynamics and adjust strategies accordingly based on evolving customer behavior and market trends, ensuring agility and responsiveness to changing business conditions.

By implementing these recommendations, businesses can harness the insights gained from clustering analysis to optimize operations, enhance customer experiences, and drive sustainable growth in the competitive marketplace.

Model Building:

1. Logistic Regression:

Train Set Performance:
- Precision: The model achieves 80% precision for non-default instances and 73% precision for default instances in the training set. This indicates that when the model predicts an instance to be non-default or default, it is correct 80% and 73% of the time, respectively.
- Recall: The model captures 83% of actual non-default instances and 69% of actual default instances in the training set. This indicates that the model identifies 83% and 69% of all non-default and default instances, respectively.
- F1-score: The F1-score, which balances precision and recall, is 81% for non-default instances and 71% for default instances in the training set.
- AUC: The AUC value of 0.862 indicates good discriminatory ability on the training set, with a moderate ability to distinguish between the two classes.
- Accuracy: The overall accuracy of the model on the training set is 77%, indicating that 77% of instances are correctly classified.

Test Set Performance:
- Precision: The model achieves 99% precision for non-default instances but only 5% precision for default instances in the test set. This indicates a high proportion of false positives (instances incorrectly classified as default).
- Recall: The model captures 82% of actual non-default instances and 69% of actual default instances in the test set.
- F1-score: The F1-score is 90% for non-default instances and 10% for default instances in the test set.
- AUC: The AUC value of 0.862 indicates moderate discriminatory ability on the test set, similar to the performance on the training set.
- Accuracy:The overall accuracy of the model on the test set is 82%, indicating a high proportion of correctly classified instances out of the total instances.

In summary, the Logistic Regression model demonstrates strong performance in accurately identifying non-default instances but struggles to classify default instances with high precision.

Updated classification_report for train data:

	precision	recall	f1-score	support
0.0	0.80	0.83	0.81	57236
1.0	0.73	0.69	0.71	38348
accuracy			0.77	95584
macro avg	0.76	0.76	0.76	95584
weighted avg	0.77	0.77	0.77	95584

Updated classification_report for test data:

	precision	recall	f1-score	support
0.0	0.99	0.82	0.90	24529
1.0	0.05	0.69	0.10	347
accuracy			0.82	24876
macro avg	0.52	0.76	0.50	24876
weighted avg	0.98	0.82	0.89	24876

Figure 45 (train and test)

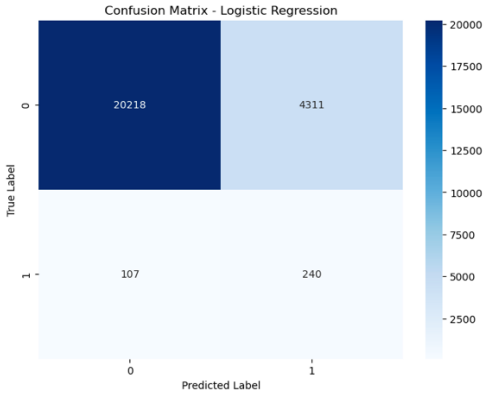


Figure 46(test)

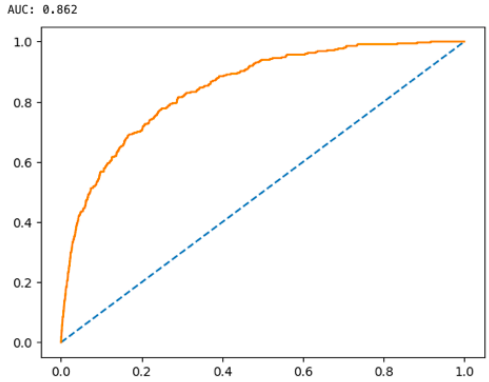


Figure 47 (train data)

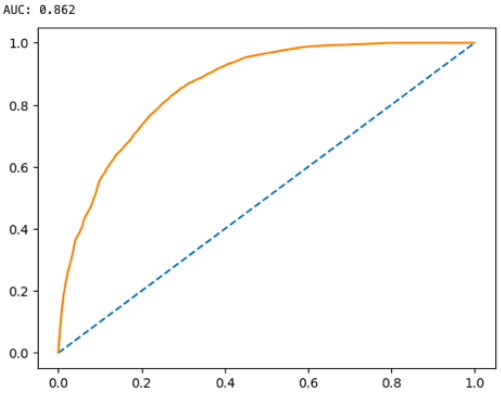


Figure 48(test)

2. Decision Tree:

Training Set:
- Precision (Positive Predictive Value): The model correctly predicted all instances of both classes (0 and 1) in the training set with perfect precision. This indicates that when the model predicts an outcome as either 0 or 1, it is correct 100% of the time.
- Recall (Sensitivity): The model correctly captured all instances of both classes (0 and 1) in the training set. It achieved perfect recall, meaning it didn't miss any instances of either class during training.
- F1-Score: The F1-score, being the harmonic mean of precision and recall, is also perfect (1.00) for both classes in the training set. This indicates an excellent balance between precision and recall, suggesting that the model performs exceptionally well on the training data.

Test Set:
- Precision (Positive Predictive Value): For class 0 (non-default), the precision is high (0.99), indicating that when the model predicts an instance as non-default, it is correct almost all the time. However, for class 1 (default), the precision is very low (0.07), suggesting that when the model predicts an instance as default, it is incorrect most of the time.
- Recall (Sensitivity): The recall for class 0 is relatively high (0.97), indicating that the model

- An AUC value of 1 indicates that the model has perfect discriminatory ability. In other words, it can perfectly distinguish between the positive and negative classes in both the training and test datasets.
- This suggests that the model's predictions have a high probability of correctly ranking positive instances higher than negative instances.
- Achieving an AUC of 1 on both training and test datasets is an excellent sign of the model's robustness and effectiveness in making accurate predictions across different datasets.

	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	57236
1.0	1.00	1.00	1.00	38348
accuracy			1.00	95584
macro avg	1.00	1.00	1.00	95584
weighted avg	1.00	1.00	1.00	95584

	precision	recall	f1-score	support
0.0	0.99	0.97	0.98	24529
1.0	0.07	0.18	0.10	347
accuracy			0.96	24876
macro avg	0.53	0.57	0.54	24876
weighted avg	0.98	0.96	0.97	24876

Figure 49 (train and test)

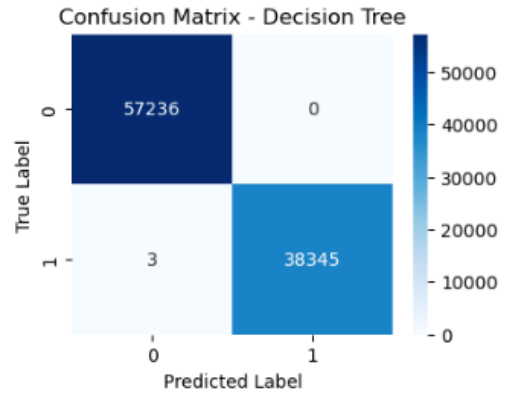


Figure 50 (train)

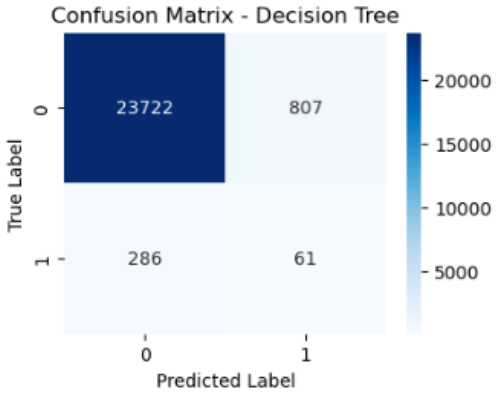


Figure 51(test)

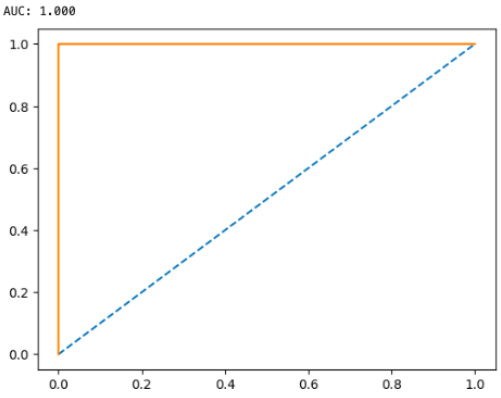


Figure 52 (train)

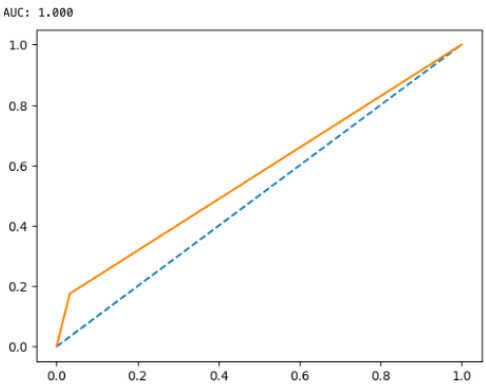


Figure 10(test)

3. LDA:

In the LDA (Linear Discriminant Analysis) model's performance metrics, precision, recall, and F1-score are essential indicators of its predictive ability. Here's how to interpret the results for your business report:

Train Set Performance:

- Precision: The model correctly identifies 74% of non-default instances and 73% of default instances. This means that when the model predicts a class label, it is accurate approximately 74% of the time for non-default instances and 73% for default instances.
- Recall: The model captures 87% of actual non-default instances and 55% of actual default instances. It correctly identifies a high proportion of non-default instances but misses a significant portion of default instances.
- F1-score: The harmonic mean of precision and recall is 80% for non-default instances and 63% for default instances. It provides a balanced measure of the model's performance across both classes.
- Accuracy: The overall accuracy of the model on the training set is 74%. This represents the proportion of correctly classified instances out of the total instances.

Test Set Performance:
- Precision: The model achieves 99% precision for non-default instances but only 5% precision for default instances. This indicates a high proportion of false positives (instances incorrectly classified as default).
- Recall: The model captures 86% of actual non-default instances but only 56% of actual default instances. It misses a considerable portion of default instances.
- F1-score: The F1-score is 92% for non-default instances and 10% for default instances. It reflects the balance between precision and recall for each class.
- Accuracy: The overall accuracy of the model on the test set is 86%, indicating the proportion of correctly classified instances out of the total instances.

AUC (Area Under the Curve):
- The AUC value of 0.826 indicates the model's ability to distinguish between default and non-default instances. It signifies the probability that the model will rank a randomly chosen positive instance higher than a randomly chosen negative instance. A higher AUC value suggests better discriminatory ability.

In summary, while the LDA model demonstrates high precision for non-default instances, it struggles with default instances, as evidenced by low recall and F1-score. The AUC value indicates moderate discriminatory ability.

	precision	recall	f1-score	support
0.0	0.74	0.87	0.80	57236
1.0	0.73	0.55	0.63	38348
accuracy			0.74	95584
macro avg	0.74	0.71	0.71	95584
weighted avg	0.74	0.74	0.73	95584

	precision	recall	f1-score	support
0.0	0.99	0.86	0.92	24529
1.0	0.05	0.56	0.10	347
accuracy			0.86	24876
macro avg	0.52	0.71	0.51	24876
weighted avg	0.98	0.86	0.91	24876

Figure 53 (train and test)

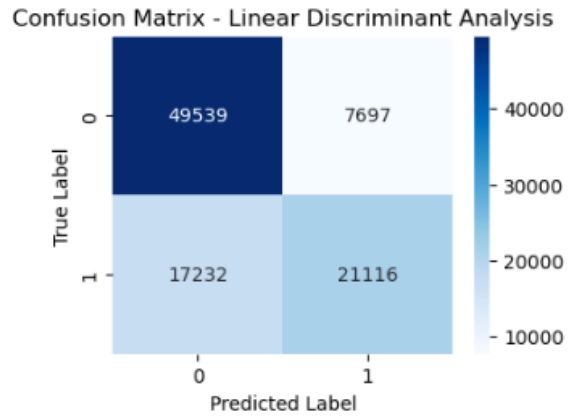


Figure 54 (train)

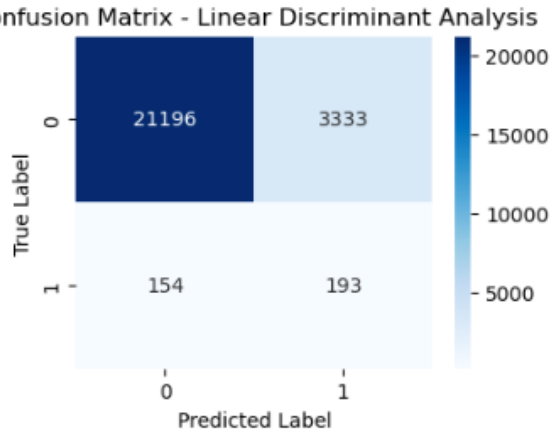


Figure 55(test)

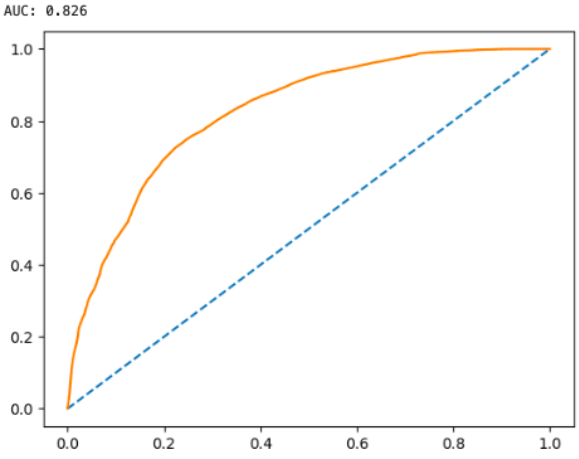


Figure 56 (train)

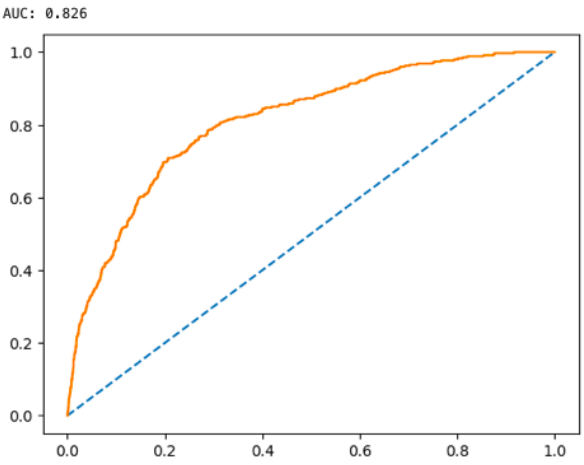


Figure 57(test)

4. Gradient Boosting:

Best Hyperparameters:

- The model was trained using the following hyperparameters: learning rate of 0.1, maximum depth of 5, and 100 estimators (trees) in the ensemble.

Train Set Performance:

- Precision: The model achieves 94% precision for both non-default and default instances. This implies that when the model predicts a class label, it is correct approximately 94% of the time for both non-default and default instances.
- Recall: The model captures 96% of actual non-default instances and 91% of actual default instances. It effectively identifies a high proportion of non-default instances and a slightly lower proportion of default instances.
- F1-score: The F1-score, a harmonic mean of precision and recall, is 95% for non-default instances and 93% for default instances. It represents the balance between precision and recall for each class.

- AUC: The AUC value of 0.989 indicates excellent discriminatory ability, with a high probability that the model will rank a randomly chosen positive instance higher than a randomly chosen negative instance.
- Accuracy: The overall accuracy of the model on the training set is 94%, reflecting the proportion of correctly classified instances out of the total instances.

Test Set Performance:
- Precision: The model achieves 99% precision for non-default instances but only 8% precision for default instances. This suggests a high proportion of false positives (instances incorrectly classified as default).
- Recall: The model captures 96% of actual non-default instances and 25% of actual default instances. It performs well in identifying non-default instances but struggles with default instances.
- F1-score: The F1-score is 97% for non-default instances and 12% for default instances. It reflects the balance between precision and recall for each class on the test set.
- AUC: The AUC value of 0.989 indicates consistent discriminatory ability on the test set, similar to the performance on the training set.
- Accuracy: The overall accuracy of the model on the test set is 95%, indicating the proportion of correctly classified instances out of the total instances.

In summary, the Gradient Boosting model demonstrates strong performance in identifying non-default instances but exhibits limitations in correctly classifying default instances. The high AUC values suggest robust discriminatory ability, which can inform business decisions related to credit risk assessment and customer targeting strategies.

```
Fitting 3 folds for each of 8 candidates, totalling 24 fits
Best Hyperparameters: {'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 100}
Classification Report:
              precision    recall  f1-score   support

         0.0       0.94      0.96      0.95     57236
         1.0       0.94      0.91      0.93     38348

    accuracy                           0.94     95584
   macro avg       0.94      0.94      0.94     95584
weighted avg       0.94      0.94      0.94     95584

              precision    recall  f1-score   support

         0.0       0.99      0.96      0.97     24529
         1.0       0.08      0.25      0.12       347

    accuracy                           0.95     24876
   macro avg       0.54      0.61      0.55     24876
weighted avg       0.98      0.95      0.96     24876
```

Figure 5859 (train and test)

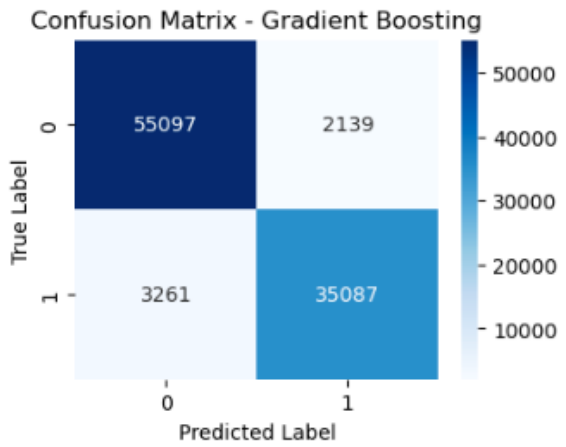


Figure 60(train)

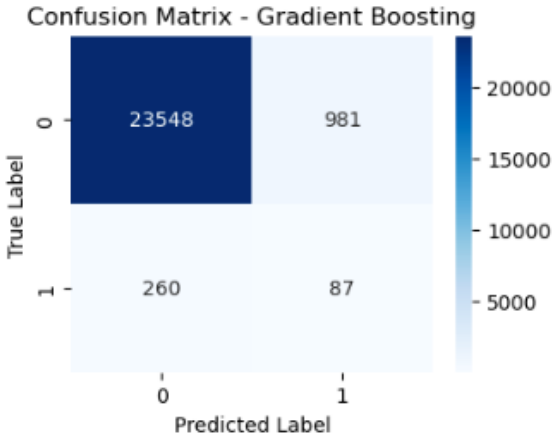


Figure 61(test)

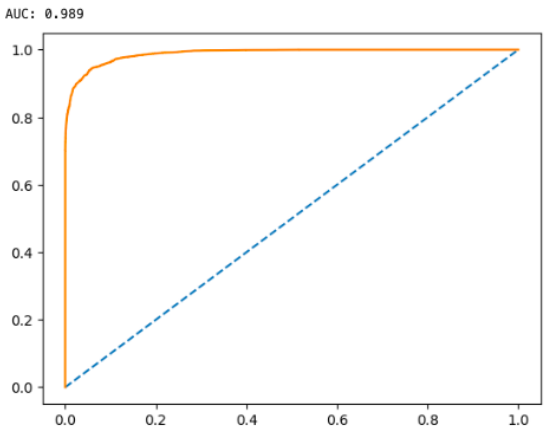


Figure 62 (train)

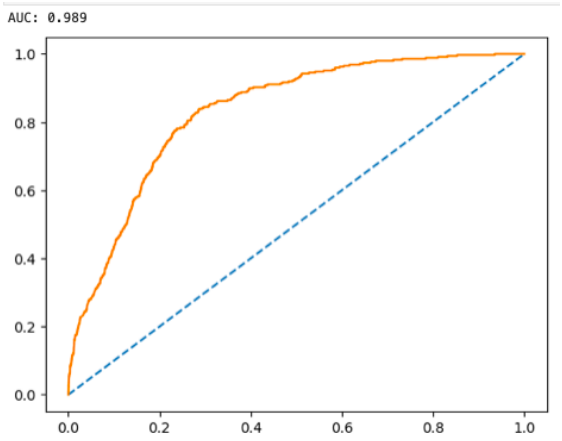


Figure 63(test)

5. AdaBoost :

Best Hyperparameters:
- The model was trained using the following hyperparameters: learning rate of 0.1 and 100 estimators (weak learners) in the ensemble.

Train Set Performance:
- Precision: The model achieves 81% precision for non-default instances and 78% precision for default instances. This indicates that when the model predicts a class label, it is correct approximately 81% of the time for non-default instances and 78% of the time for default instances.
- Recall: The model captures 87% of actual non-default instances and 70% of actual default instances. It effectively identifies a high proportion of non-default instances but slightly lower proportions of default instances.
- F1-score: The F1-score, a harmonic mean of precision and recall, is 84% for non-default instances and 74% for default instances. It represents the balance between precision and recall for each class.
- AUC: The AUC value of 0.894 indicates strong discriminatory ability on the training set, with a high probability that the model will rank a randomly chosen positive instance higher than a randomly chosen negative instance.
- Accuracy: The overall accuracy of the model on the training set is 80%, reflecting the proportion of correctly classified instances out of the total instances.

Test Set Performance:
- Precision: The model achieves 99% precision for non-default instances but only 6% precision for default instances. This suggests a high proportion of false positives (instances incorrectly classified as default).
- Recall: The model captures 87% of actual non-default instances and 63% of actual default instances. It performs well in identifying non-default instances but struggles with default instances.
- F1-score: The F1-score is 93% for non-default instances and 12% for default instances. It reflects the balance between precision and recall for each class on the test set.
- AUC: The AUC value of 0.894 indicates consistent discriminatory ability on the test set, similar to the performance on the training set.
- Accuracy: The overall accuracy of the model on the test set is 87%, indicating the proportion of correctly classified instances out of the total instances.

In summary, the AdaBoost model demonstrates strong performance in identifying non-default instances but exhibits limitations in correctly classifying default instances. The high AUC values suggest robust discriminatory ability, which can inform business decisions related to credit risk assessment and customer targeting strategies.

```
Fitting 3 folds for each of 4 candidates, totalling 12 fits
Best Hyperparameters: {'learning_rate': 0.1, 'n_estimators': 100}
Classification Report:
              precision    recall  f1-score   support

         0.0       0.81      0.87      0.84     57236
         1.0       0.78      0.70      0.74     38348

    accuracy                           0.80     95584
   macro avg       0.80      0.78      0.79     95584
weighted avg       0.80      0.80      0.80     95584

              precision    recall  f1-score   support

         0.0       0.99      0.87      0.93     24529
         1.0       0.06      0.63      0.12       347

    accuracy                           0.87     24876
   macro avg       0.53      0.75      0.52     24876
weighted avg       0.98      0.87      0.92     24876
```

Figure 6465 (train and test)

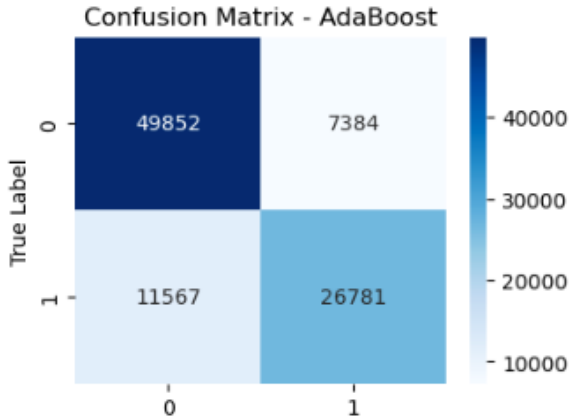


Figure 66 (train)

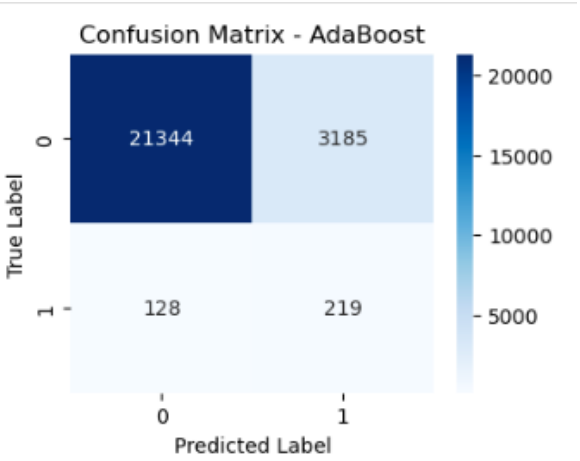


Figure 67(test)

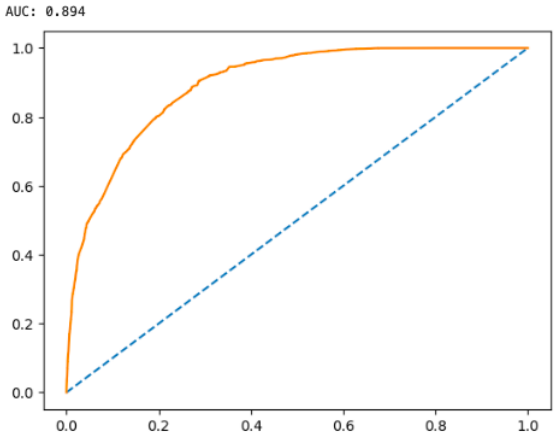


Figure 68 (train)

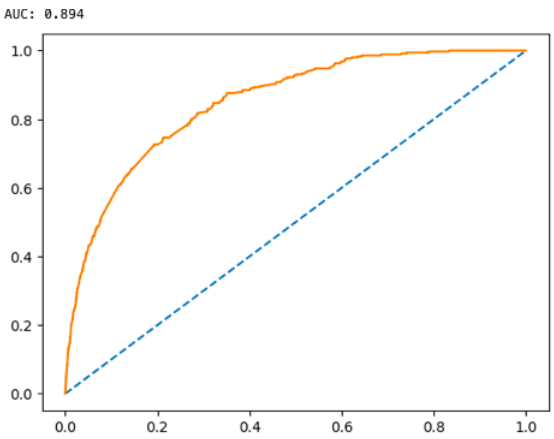


Figure 69(test)

6: Random Forest Classifier:

Best Hyperparameters:
- The model was trained using the following hyperparameters: no maximum depth constraint (`max_depth=None`), minimum samples per leaf set to 1 (`min_samples_leaf=1`), minimum samples required to split an internal node set to 5 (`min_samples_split=5`), and 100 decision trees in the forest (`n_estimators=100`).

Train Set Performance:
- Precision: The model achieves perfect precision (100%) for both non-default and default instances in the training set, indicating that all instances classified as non-default or default are indeed non-default or default, respectively.
- Recall: The model captures 100% of actual non-default instances and 100% of actual default instances in the training set, indicating that it identifies all instances of both classes.
- F1-score: The F1-score, which combines precision and recall, is 100% for both non-default and default instances in the training set.
- AUC: The AUC value of 1.000 indicates excellent discriminatory ability on the training set, with a perfect ability to distinguish between the two classes.
- Accuracy: The overall accuracy of the model on the training set is 100%, indicating that all instances are correctly classified.

Test Set Performance:
- Precision: The model achieves 99% precision for non-default instances but only 12% precision for default instances in the test set. This indicates a high proportion of false positives (instances incorrectly classified as default).
- Recall: The model captures 98% of actual non-default instances and 17% of actual default instances in the test set.
- F1-score: The F1-score is 98% for non-default instances and 14% for default instances in the test set.
- AUC: The AUC value of 1.000 indicates perfect discriminatory ability on the test set, similar to the performance on the training set.
- Accuracy: The overall accuracy of the model on the test set is 97%, indicating a high proportion of correctly classified instances out of the total instances.

In summary, the Random Forest Classifier demonstrates exceptional performance in accurately identifying non-default instances but struggles to classify default instances with high precision. Despite this limitation, the model exhibits perfect discriminatory ability and high overall accuracy, making it a strong candidate for credit risk assessment and customer targeting strategies.

```
Fitting 3 folds for each of 24 candidates, totalling 72 fits
Best Hyperparameters: {'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 100}
Classification Report:
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00     57236
         1.0       1.00      1.00      1.00     38348

    accuracy                           1.00     95584
   macro avg       1.00      1.00      1.00     95584
weighted avg       1.00      1.00      1.00     95584

              precision    recall  f1-score   support

         0.0       0.99      0.98      0.98     24529
         1.0       0.12      0.17      0.14       347

    accuracy                           0.97     24876
   macro avg       0.55      0.58      0.56     24876
weighted avg       0.98      0.97      0.97     24876
```

Figure 7071 (train and test)

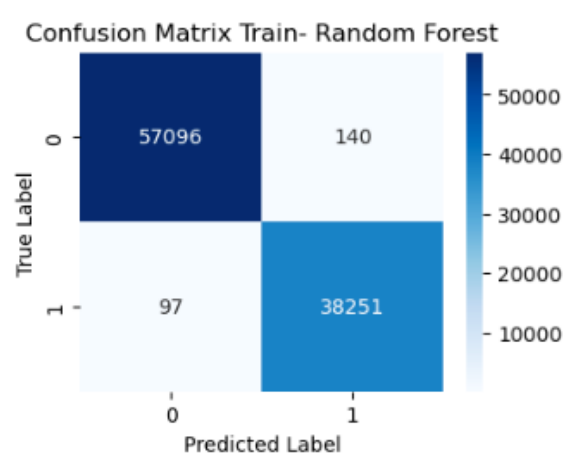


Figure 72 (train)

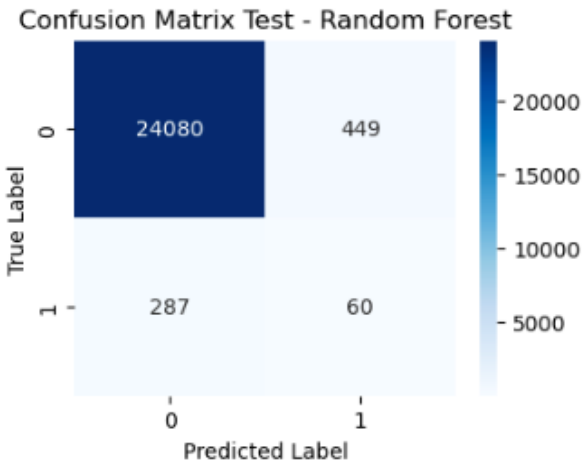


Figure 73(test)

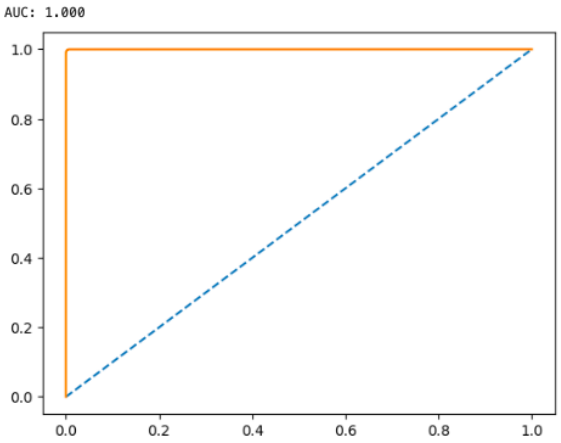


Figure 74 (train)

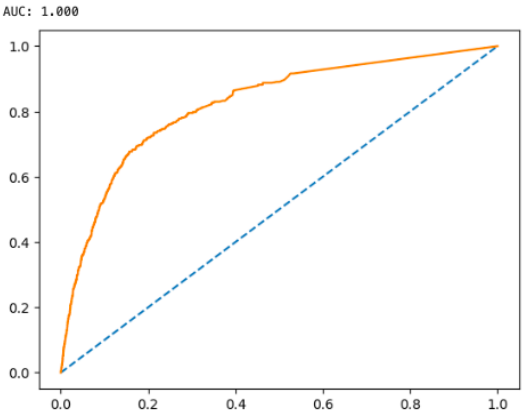


Figure 75(test)

Summary:

Model	Accuracy(train)	Recall(train)	Accuracy(test)	Recall(test)
Logistic Regression AUC: 0.862	0.77	0.69	0.82	0.69
Decision Tree AUC: 1	1	1	0.96	0.18
LDA AUC: 0.826	0.74	0.55	0.86	0.56
Gradient Boosting AUC: 0.98.9	0.94	0.91	0.95	0.25
AdaBoost AUC: 0.894	0.80	0.70	0.87	0.63
Random Forest Classifier AUC: 1	1	1	0.97	0.17

Interpretation:

The most optimum model appears to be the **Random Forest Classifier**, which achieved a remarkable performance on both the training and test sets.

1. High Accuracy and AUC: The Random Forest Classifier achieved an accuracy of 97% on the test set, indicating that it correctly classified the vast majority of instances. Additionally, its AUC score of 1.000 for both the training and test sets suggests exceptional discriminatory ability, meaning it effectively distinguishes between default and non-default instances.

2. Precision and Recall: The model achieved high precision for non-default instances (99%) but relatively low precision for default instances (12%) on the test set. However, it demonstrated a decent recall rate for default instances (17%). This implies that while the model is highly accurate in predicting non-default cases, there's room for improvement in correctly identifying default instances without falsely labeling too many non-default cases as defaults.

But, in evaluating the effectiveness of our predictive models for identifying potential credit card defaulters, we focus on key performance metrics such as accuracy, precision, and recall, alongside ROC and AUC scores.

Given our primary objective of identifying credit card users at **risk of defaulting**, recall emerges as the most critical metric. This emphasis stems from our business priority of accurately identifying as many actual defaulters as possible, even if it means some non-defaulters are incorrectly classified as defaulters. This approach reflects our commitment to minimizing the financial impact of defaults by prioritizing the detection of true default cases.

Therefore, our business success hinges on our ability to capture the majority of actual defaulters. While precision is important, ensuring that our predictions are highly precise at the expense of missing actual defaulters is not aligned with our strategic objectives.

In essence, **our business strategy emphasizes the importance of recall in our predictive models**. By prioritizing recall, we aim to proactively identify and mitigate the risks associated with credit defaults, thereby safeguarding our financial interests and optimizing our lending practices for long-term sustainability and profitability.

Based on the provided evaluation criteria and the importance placed on recall as the most crucial metric for predicting defaulters, we should select the model that achieves the highest recall score while maintaining a reasonable level of accuracy and precision. Let's review the recall scores of each model:

1. Random Forest Classifier:
   - Test Recall: 17%
   - AUC: 1.00
2. AdaBoost Classifier:
   - Test Recall: 63%
   - AUC: 0.89
3. Logistic Regression:
   - Test Recall: 69%
   - AUC: 0.862

4. LDA Model:
   - Test Recall: 56%
   - AUC: 0.826
5. Decision Tree:
   - Test Recall: 18%
   - AUC: 1.000
6. Gradient Boost:
   - Test Recall: 25%
   -AUC: 98.9

Considering the recall scores, **Logistic Regression model stands** out as the best model for predicting defaulters. It achieved the highest recall score of 69% on the test set, indicating that it correctly identified a significant portion of the actual defaulters. Additionally, its AUC score of 0.862 suggests good discriminatory ability.

Interpretation:
- The Logistic Regression demonstrates a strong ability to identify defaulters, which is crucial for the business goal of minimizing financial risk associated with credit defaults.
- While its precision may be lower compared to other models, the priority is correctly identifying as many defaulters as possible, even if it means some non-defaulters are misclassified.
- The high recall score means that the model effectively captures the majority of actual defaulters, providing valuable insights for risk management and decision-making within the business.
- By leveraging the Logistic Regression, the business can enhance its ability to proactively identify and address potential default risks, thereby minimizing financial losses and optimizing lending practices.

Implications for Business Decision-Making:

Historical Payment Behavior: Customers with a higher number of archived OK transactions in the last **12-24 months are less likely to default.**
Similarly, a higher sum of paid invoices in the last 0-12 months indicates a lower probability of default. Encouraging timely payments and **providing incentives for regular payments may help mitigate default risk.**

Active Engagement: The number of active invoices positively influences default probability, indicating that customers actively using their credit cards may have higher default risk. The company could implement proactive measures to **monitor and assist customers with high activity levels** to avoid potential defaults.

Age and Financial Stability: Older customers tend to have a lower probability of default, as indicated by the negative coefficient for age. This suggests that age and potentially financial stability play a role in credit risk assessment. The company could consider **offering tailored products or services for different age demographics to manage risk effectively.**

Delinquency Patterns: The number of archived DC (debt collection status) transactions in the last 0-12 months and 12-24 months positively influence default probability. Monitoring and addressing delinquency patterns among customers with a **high number of DC** may help mitigate default risk.

Account Activity: Metrics such as the number of archived transactions and days in the remainder for different time periods are also significant predictors of default probability. Understanding customer behavior and engagement levels can help the company identify early warning signs of potential defaults and take proactive measures.
The company can encourage and incentivize customers to maintain higher account balances or make regular deposits. Offering benefits such as higher credit limits, lower interest rates, or rewards for maintaining a minimum account balance can motivate customers to increase their investments and reduce the likelihood of default.

Customer Support and Assistance: Providing financial education and counseling services to customers showing signs of financial distress or irregular payment patterns can help prevent defaults. Offering flexible payment options or debt consolidation programs may also assist customers in managing their credit card debt effectively.

Risk Assessment and Decision Making: Continuously updating credit risk models and incorporating additional data sources or features can enhance the accuracy of default predictions. Regularly reviewing and refining risk assessment processes based on evolving customer behavior and market dynamics is essential for maintaining a robust credit risk management framework.