REVANTH JANAPRIYAN

**Probability of default**

Contents

[**INTRODUCTION** 2](#_Toc145846076)

[**Data Report** 4](#_Toc145846077)

[DATA DICTIONARY 5](#_Toc145846078)

[**Exploratory data analysis** 9](#_Toc145846079)

[**Univariate analysis** 9](#_Toc145846080)

[**Multivariate analysis** 17](#_Toc145846081)

[**Business insights from EDA** 22](#_Toc145846082)

[**MODEL BUILDING** 22](#_Toc145846083)

[**Encoding and scaling** 22](#_Toc145846084)

[**Feature selection** 23](#_Toc145846085)

[**Model building and interpretation** 23](#_Toc145846086)

[**Model Tuning and business implication** 27](#_Toc145846087)

# **INTRODUCTION**

1. *Defining problem statement*

This supervised learning scenario pertains to a credit card company aiming to predict the likelihood of defaulting on credit card payments. The goal is to determine the probability of customers failing to pay their credit card bills based on various variables encompassing account details, purchase history, and instances of delinquency. Through predictive modelling, the company seeks to evaluate the risk associated with individual customers and gauge the potential financial exposure if defaults occur. This analysis holds significant importance for lending institutions, encompassing those offering both secured and unsecured loans, as it aids in assessing customer risk profiles and managing credit-related uncertainties.

b) *Need of the study/project*

The need for the above study/project is to build a model that can predict the probability of a customer defaulting on their credit card bill. This is important for the credit card company because it allows them to understand the riskiness of their customers. This information can be used to make decisions about how much credit to extend to each customer, and to take steps to mitigate risk.

Here are some of the specific benefits of building this model:

* The company can avoid lending money to customers who are likely to default. This will help to protect the company's profits and reduce its risk of financial losses.
* The company can charge higher interest rates to customers who are at a higher risk of default. This will help to compensate the company for the risk it is taking on.
* The company can offer more tailored products and services to customers who are at a lower risk of default. This will help to attract and retain these customers.
* The company can improve its customer service by identifying customers who are at risk of default and providing them with early intervention services. This will help to prevent customers from defaulting and save the company money in the long run.

Overall, building a model to predict the probability of customer default is a valuable exercise for any credit card company. It can help the company to improve its risk management practices, attract and retain customers, and provide better customer service.

c) *Understanding business/social opportunity*

This study/project presents significant business and social opportunities that can positively impact both the credit card company and the larger society:

*Business Opportunities:*

1. **Enhanced Risk Management**: Accurate prediction of default probabilities empowers the credit card company to proactively manage and mitigate financial risks. This results in improved profitability and sustainability, as losses due to defaults are minimized.
2. **Optimized Operations**: By tailoring credit limits, interest rates, and credit approvals to individual customer risk profiles, the company can optimize its operational efficiency and resource allocation.
3. **Competitive Advantage**: Implementing sophisticated predictive models for default prediction can provide the credit card company with a competitive edge in the market. Offering responsible lending practices and personalized credit solutions can attract and retain customers.
4. **Data-Driven Decision-Making**: The study fosters a data-driven organizational culture, encouraging evidence-based decisions across various departments. This improves overall business decision-making and strategic planning.
5. **Product Development**: Insights derived from default prediction models can inform the development of innovative financial products and services that cater to different customer segments, thus expanding the company's product portfolio.

*Social Opportunities:*

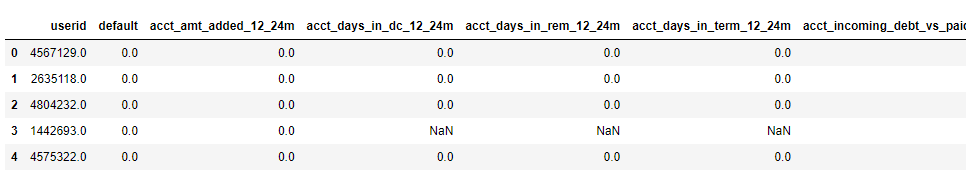
1. **Responsible Lending**: By accurately assessing default probabilities, the credit card company contributes to responsible lending practices. This ensures that customers are offered credit in alignment with their financial capabilities, reducing the risk of indebtedness.
2. **Financial Inclusion**: The project can lead to tailored credit solutions for customers with varying risk profiles. This fosters financial inclusion by providing credit opportunities to individuals who might otherwise be excluded from traditional lending processes.
3. **Education and Awareness**: As the project sheds light on factors influencing default probabilities, it can educate customers about the significance of responsible credit management and financial literacy.
4. **Reduced Default Burden:** Predictive models help customers make informed decisions, encouraging them to manage their credit responsibly and avoid defaults. This, in turn, alleviates the financial burden on individuals and families.
5. **Consumer Confidence:** Transparent and fair lending practices built upon accurate default prediction foster consumer confidence in the financial industry. This encourages healthier credit behaviour and cultivates trust.
6. **Economic Impact:** By minimizing credit-related risks, the credit card company contributes to the stability of the financial sector, which in turn positively impacts the broader economy.

In summary, this study/project presents a mutually beneficial opportunity that aligns business success with responsible lending practices and positive social outcomes. It showcases the potential for data-driven solutions to create value across sectors while promoting financial stability and individual well-being.

# **Data Report**

The Dataset used in the project/study is about the credit card usage of various customers present in the bank. The dataset contains 36 columns and 99979 rows (including null values).

Below is the figure provided with first few rows of certain columns present in the dataset.

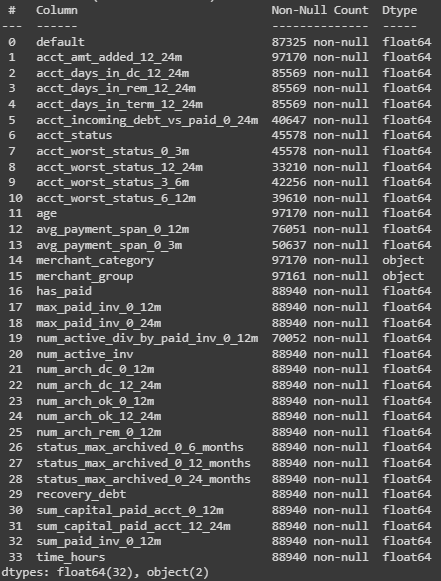


**Fig 1**

The variables ‘userid’ and 'name\_in\_email' present in the dataset have been removed as it provided no valuable information to align with the objective of this project/study. Thus, the number of columns in the dataset is further trimmed to 34.

## DATA DICTIONARY

* **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 credit 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 principal balance paid on account in the last one year.
* **sum\_capital\_paid\_acct\_12\_24m** -sum of principal 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.

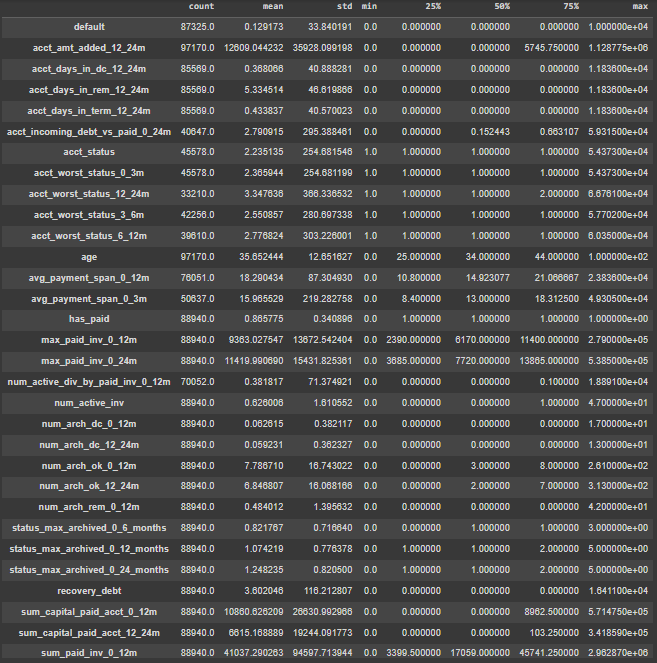


**Fig 2**

The figure 2 shows all the information about the dataset. There are 34 variables in the dataset after dropping 2 variables. Among the 34 variables there are 32 numeric variables and 2 object variables.

It can also be seen that non-null records in each variable are not same and it clearly confirms that there are null values present in the dataset. Also, datatype conversion is necessary for target variable along with few other variables.

The below figure shows the descriptive statistics of each variable in the dataset.



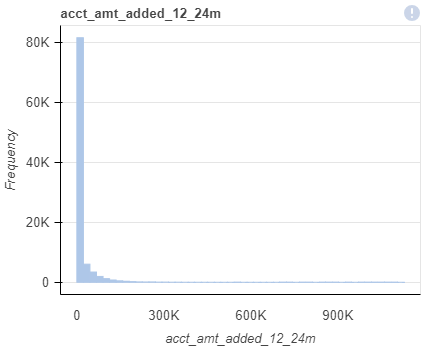
**Fig 3**

# **Exploratory data analysis**

## **Univariate analysis**

Univariate analysis is a statistical method that examines the distribution of a single variable. In this project, it used to describe the distribution of a variable, identify outliers, and assess the normality of the data.

**acct\_amt\_added\_12\_24m**

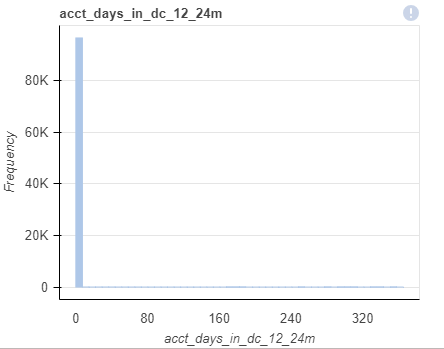


**Fig 4**

The acct\_amt\_added\_12\_24m variable is right skewed (refer fig 4). Almost 84% of the amount added falls into the bin between 0 and 22,575. On average 12,676 is added in the account during this 1 year. The maximum amount added is 1,128,800.

**acct\_days\_in\_dc\_12\_24m**

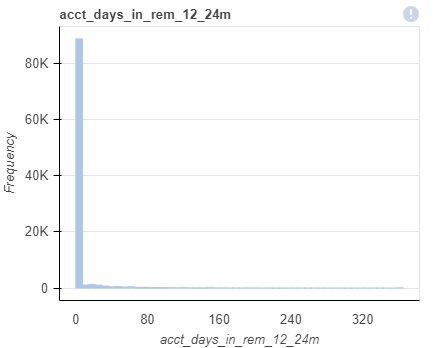
Like the previous variable, acct\_days\_in\_dc\_12\_24m is also right skewed. Approximately, 99% of customer’s account stayed in debt-collection for 0-7 days. The maximum number of days an account stayed in debt-collection status is 365 days. (Refer fig 5).



**Fig 5**

**acct\_days\_in\_rem\_12\_24m**

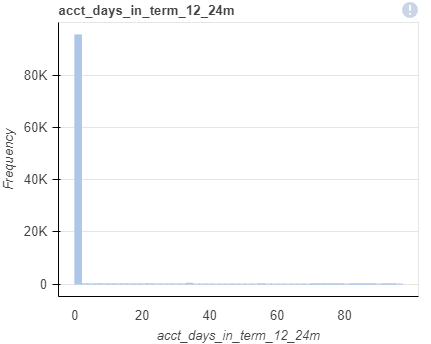
The acct\_days\_in\_rem\_12\_24m variable is right skewed. The maximum number of days an account stayed in reminder status in the past 24 months is 365 days, while the average number of days an account stayed in reminder status is 4.5 days.

****

**Fig 6**

**acct\_days\_in\_term\_12\_24m**

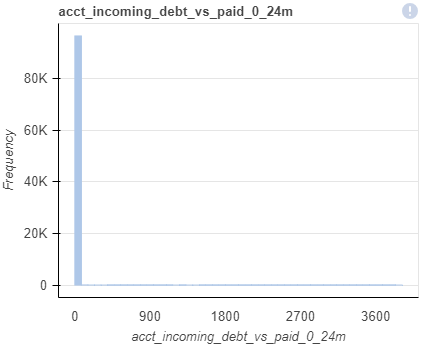
Like all the previous variables, acct\_days\_in\_term\_12\_24m is also right skewed. In the past 24 months the maximum number of days an account which stayed in termination status is 97 days.



**Fig 7**

**acct\_incoming\_debt\_vs\_paid\_0\_24m**

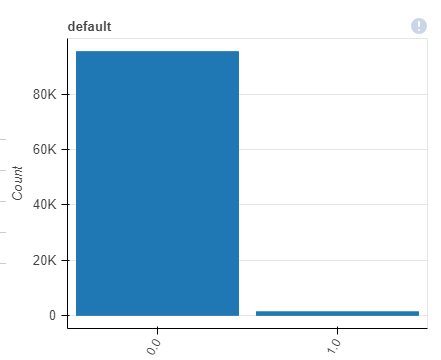
This variable is also a right skewed variable just like all the previous seen variables. Total amount collected in terms of ratio of debit amount in the account to the debit amount collected by the agency is 62,656.



**Fig 8**

**Default**

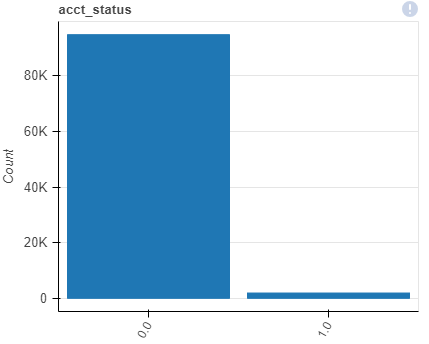
Default, is the target variable present in the dataset. The default variable has approximately 98% of non-defaulters and 2% of defaulters. The below bar graph shows it.



**Fig 9**

**acct\_status**

The acct\_status variable provides the information about number of active and inactive accounts. We, can clearly see in the figure 10, that almost 98% of the accounts is inactive and approximately 2% of the account is active.

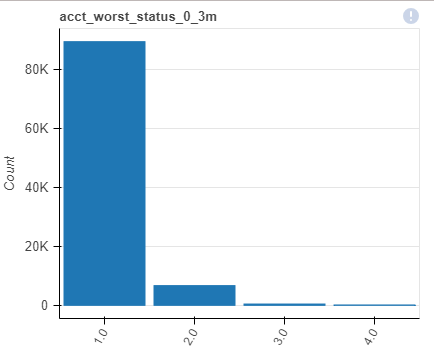


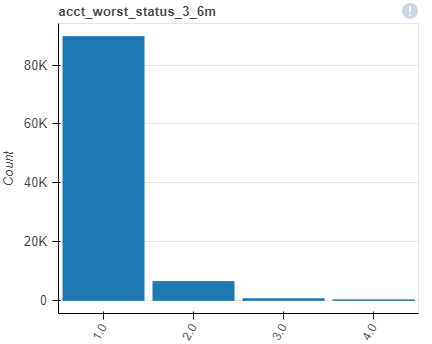
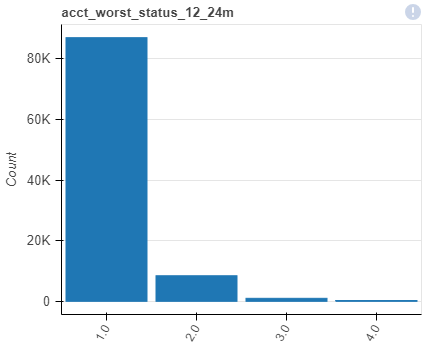
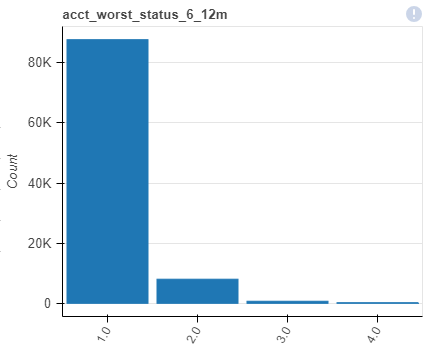
**Fig 10**

**acct\_worst\_status\_**

Below are count plot of three variable provides acct\_worst\_status\_0\_3m, acct\_worst\_status\_3\_6m, acct\_worst\_status\_6\_12m and acct\_worst\_status\_12\_24m. All four graphs provide the detail about accounts which were in worst status for the past 24 months. The maximum number of days an account was in worst status is 4 days and minimum is 1 day.

Almost in every term period, at least 98% of the worst status accounts stayed in worst status for 1 day.

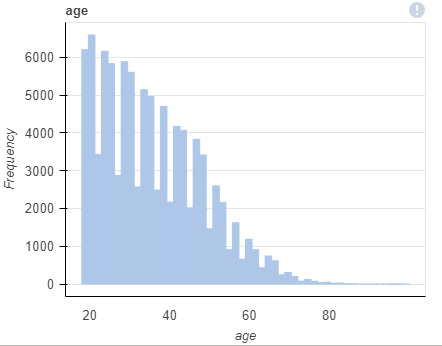
****

****

**Fig 11**

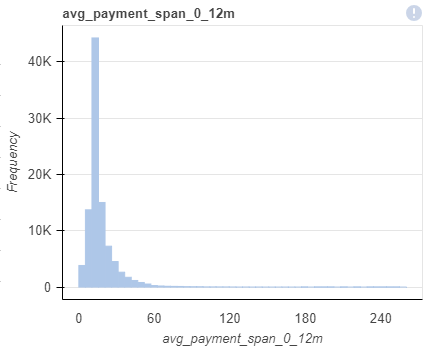
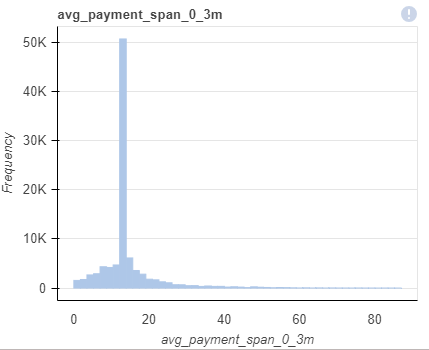
**Age**

The below histogram shows the distribution of age of the credit card users. The variable is heavily right skewed, with bin of age 19-21 having a greater number of credit card holders. The average of age of credit card holder is 35 years.



**Fig 12**

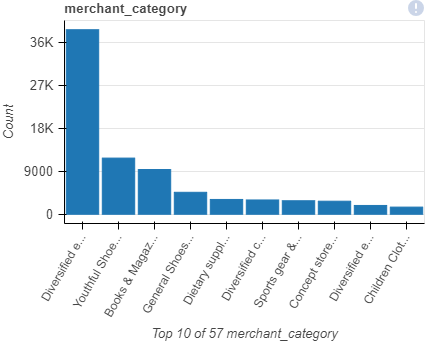
**average\_payment\_span**

  
**Fig 13**

The above figure shows the distribution of two variables avg\_payment\_span\_0\_3m and avg\_payment\_span\_0\_12m. From present day to past 3 months, on average customers pay credit bill mostly in 10-15 days, while referring for past 1 year the average number of days have declined to 12-14 days.

**merchant\_category**

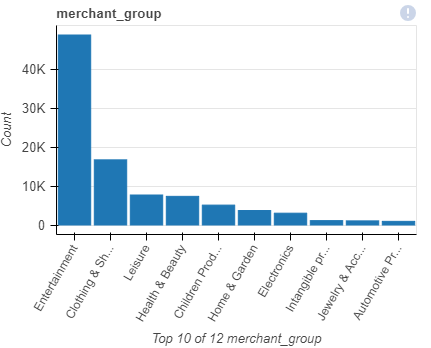
The merchant category variable has 57 unique categories. Among the 57 categories, diversity entertainment tops the list with the greatest number of merchants followed by youth shoes and clothing. Both contributing to 52% of the entire merchant category.



**Fig 14**

**Merchant\_group**

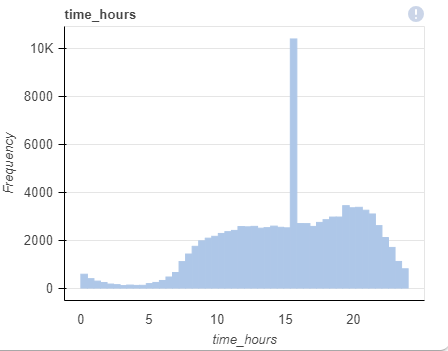
The merchant group variable contains 12 group of merchants. Similar to merchant category, Entertainment and clothing & Shoes topped the group contributing to 67% of merchant group.



**Fig 15**

**Time\_hours**

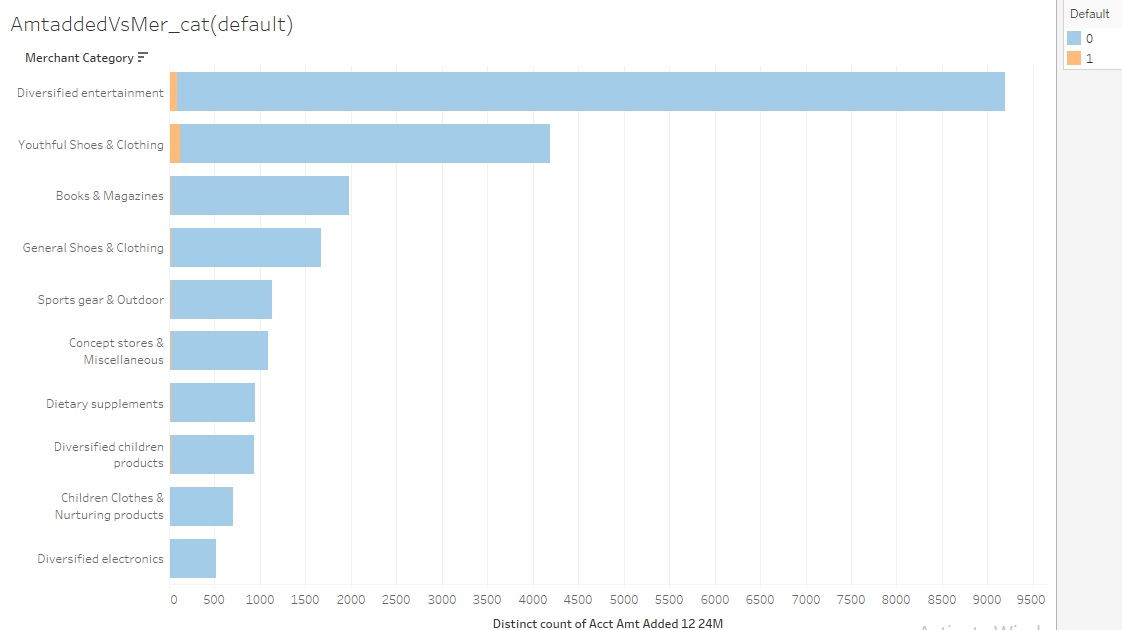
Below figure shows the time spend by customers to purchase using credit card. More than 10% of the customers spent around 15 hours in purchase using credit cards. The maximum time spent on purchase using credit card is 23.9 hours and mean is 15 hours.

****

**Fig 16**

## **Multivariate analysis**

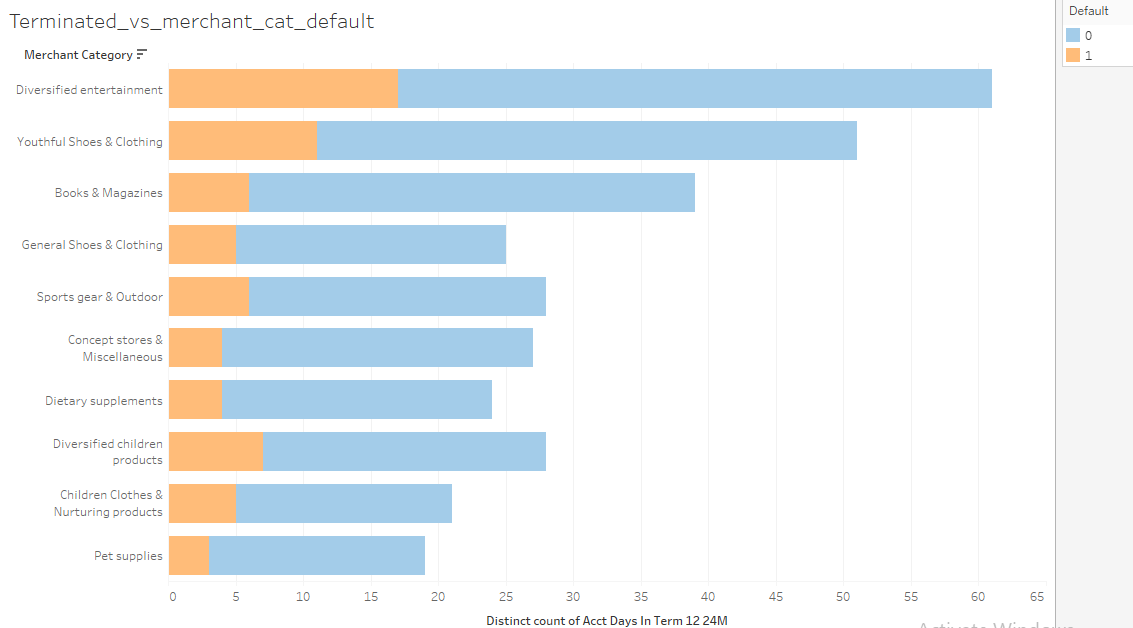
Now let’s explore multivariate analysis of the variables using tableau.



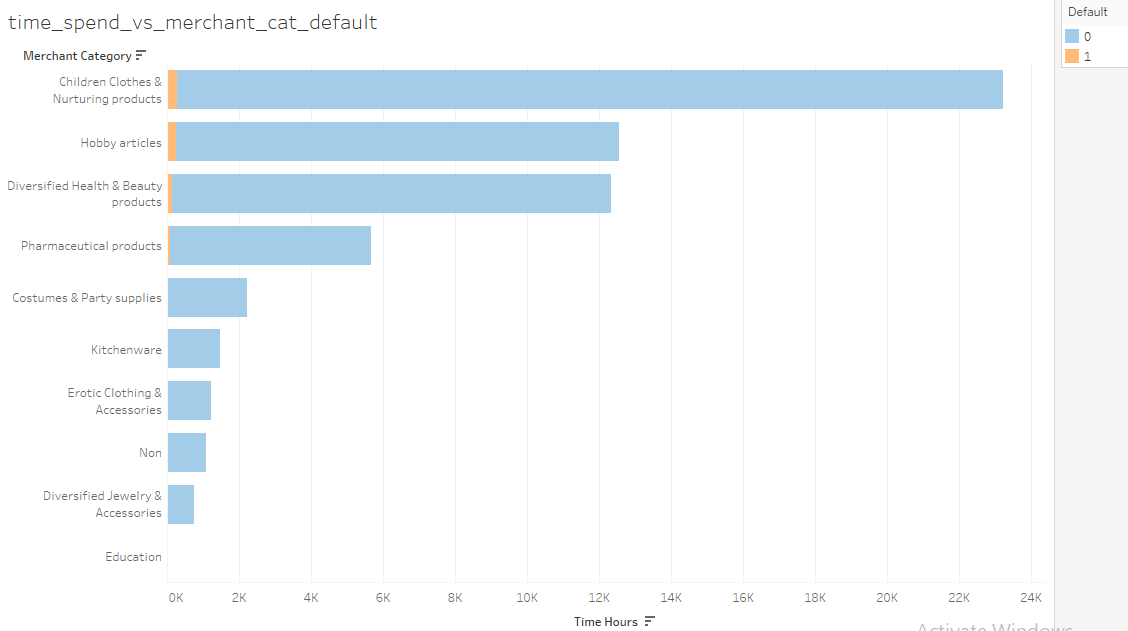
**Fig 17**

The above figure shows the top 10 merchant category which added amount in past 24 months along with the information whether there are any defaulters or not. It can be seen that Diversified entertainment and Youth shoes & clothing has few defaulters.

The figure 18, shows the number of inactive accounts among the merchant category along with the count of defaulters and non-defaulters in the merchant category. Diversified Entertainment has most defaulters as well non-defaulters along with most number of inactive accounts. (Refer fig 18)



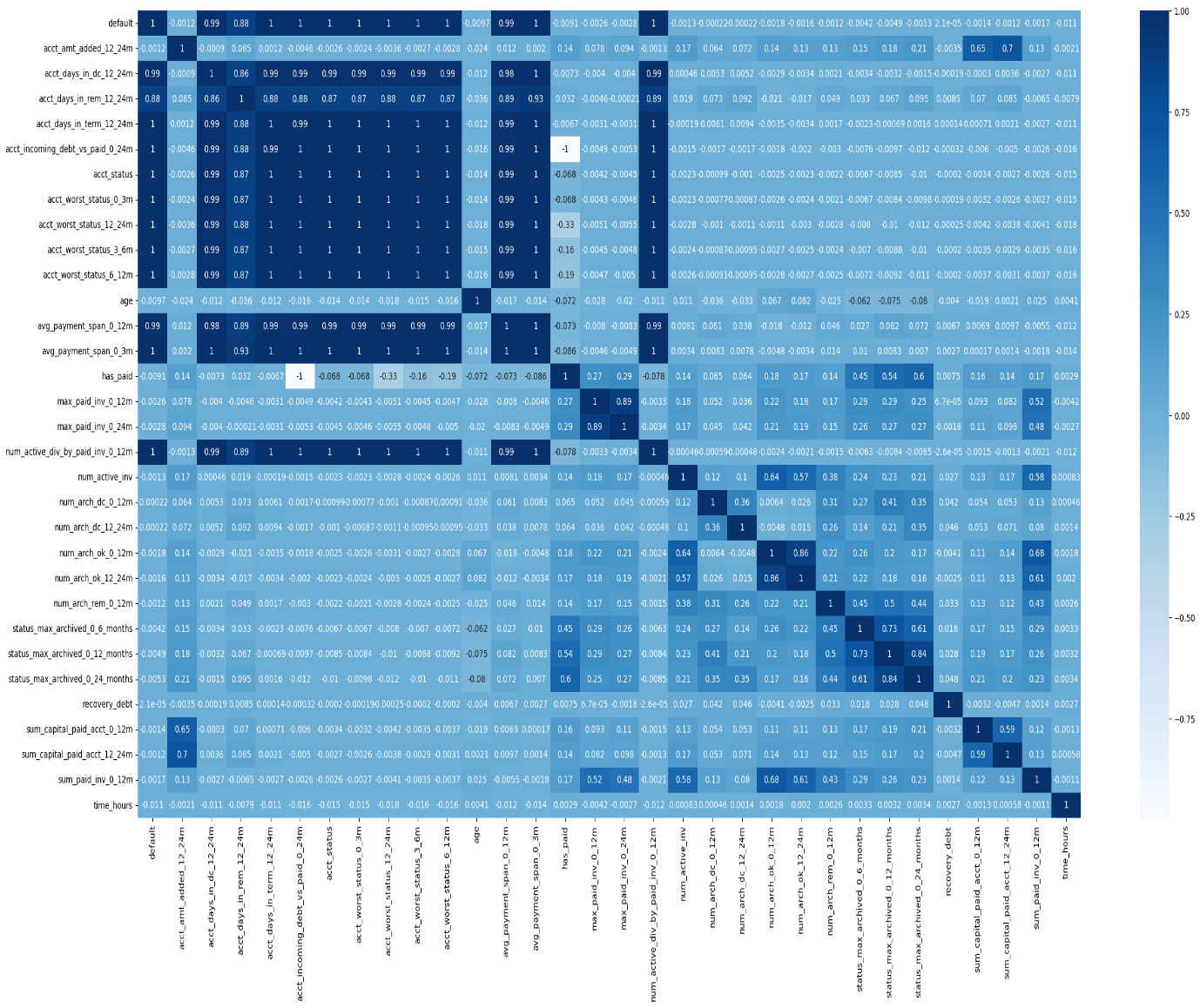
**Fig 18**



**Fig 19**

The figure 19 shows the top 10 merchant category which spent most time in purchasing using credit card along with defaulter and non-defaulter. Children product & Nurturing product tops the list and surprising education merchant category doesn’t spend any time on purchasing using credit card.

**HEATMAP**

****

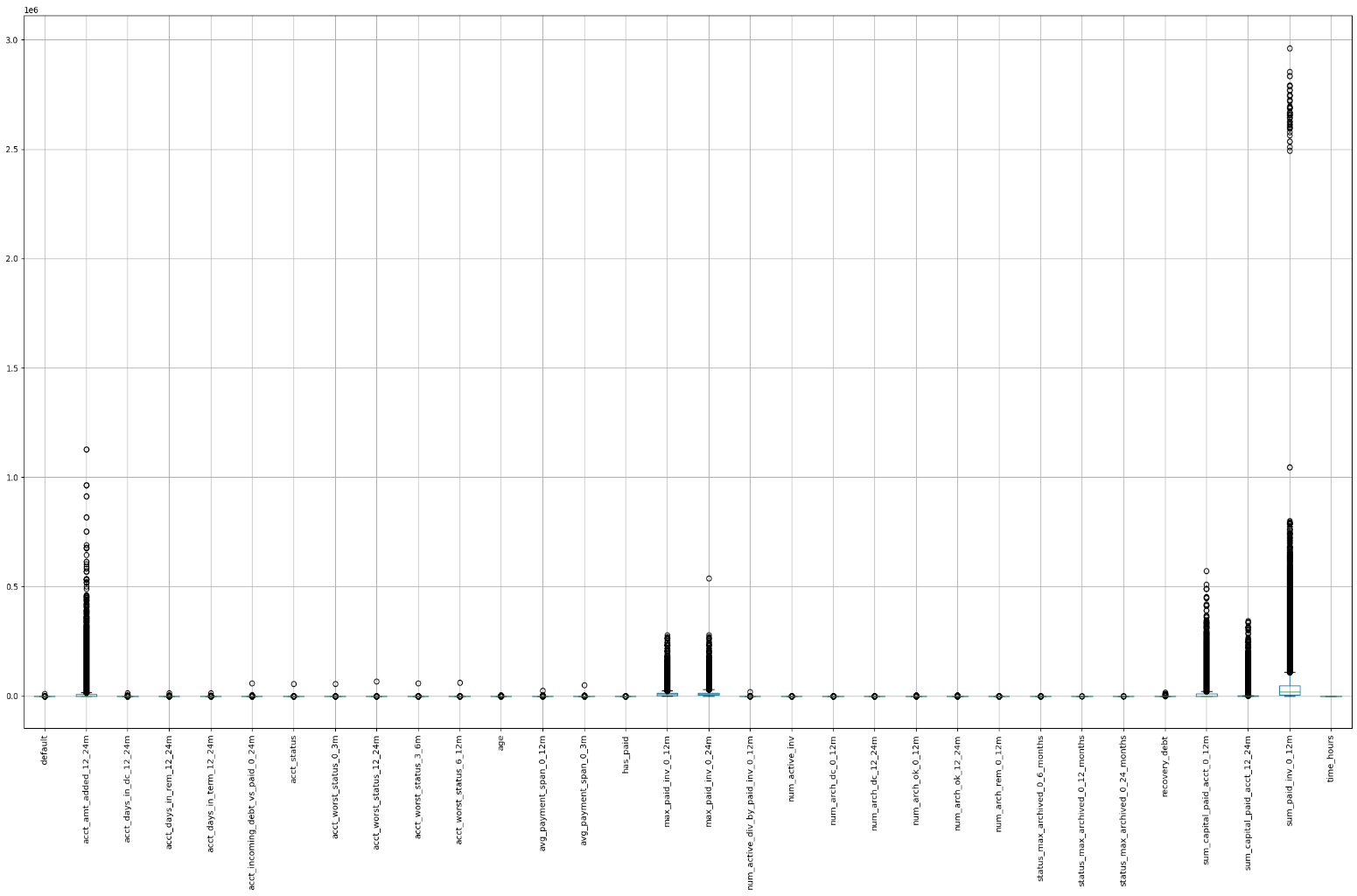
**Fig 20**

The figure 20 show the correlation among each variable present in the dataset using heatmap. Default variable and acct\_days\_in\_dc\_12\_24m has high correlation along with few other variables. While the least is between has\_paid and acct\_incoming\_debt\_vs\_paid\_0\_24m.

**Missing value**

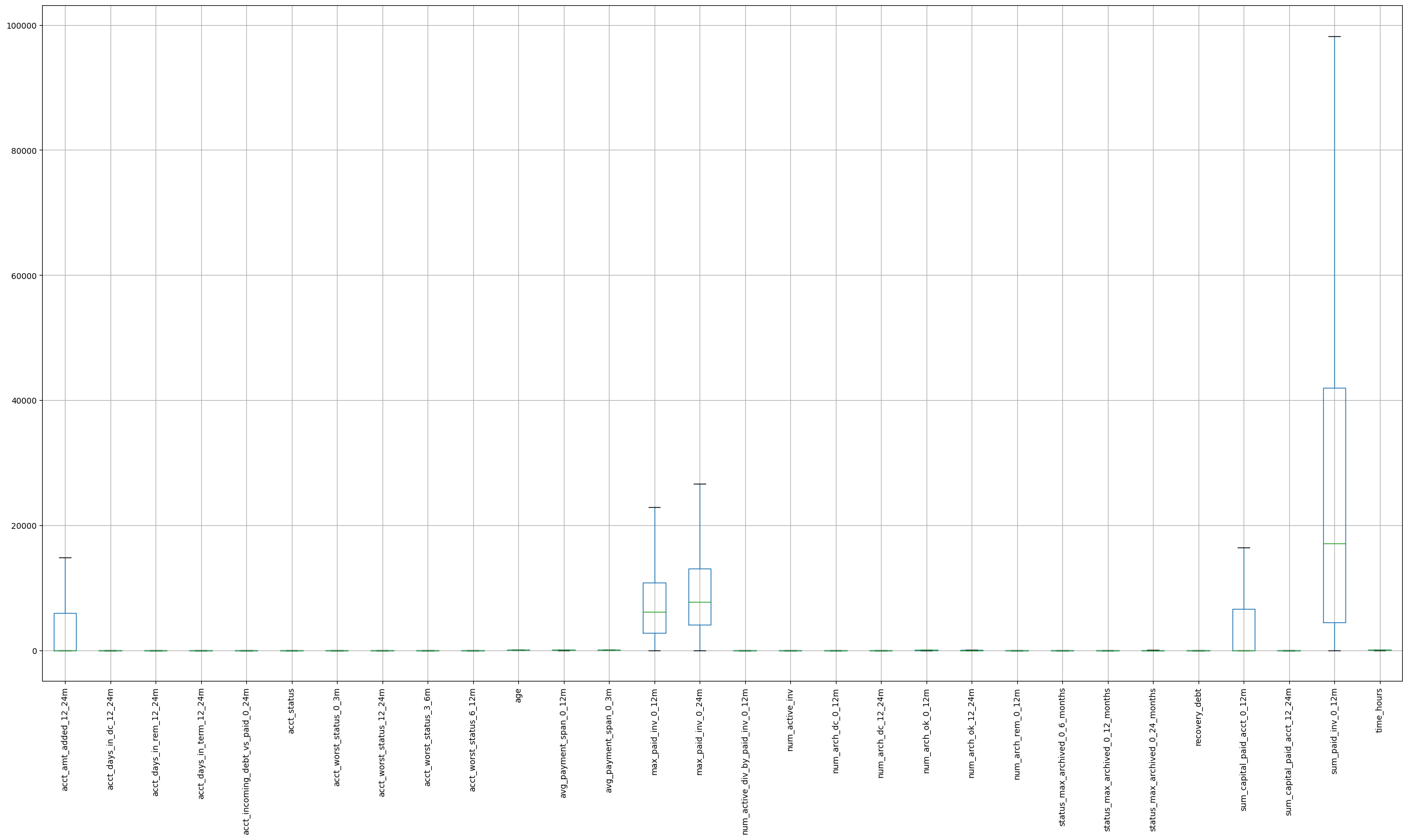
As, we saw previous in figure 2, there are missing values present in the dataset which need to be treated in such way that it does not impact the model performance during predictions.

The missing values in the numerical variables are imputed with median as it is the safety way to impute missing values in numerical variable and mode is used to impute the missing values in the categorical variables.

**Outliers  
**

**Fig 21**

The figure 21 show the outliers that are present in each variable in the dataset and the figure 22 shows the boxplot of the variables after the outliers have been treated using IQR.



**Fig 22**

# **Business insights from EDA**

The data is unbalanced. In the context of the credit card company, handling unbalanced data is crucial for accurate default prediction. If the number of default cases is much smaller, overlooking them might lead to significant financial risks. By applying the strategies such as Resampling Techniques, Weighted Loss Function, Synthetic Minority Over-sampling Technique (SMOTE), Ensemble Methods, etc.

# **MODEL BUILDING**

Before building model, there are few necessary steps need to be taken and carried out in order to make sure the predictions are unbiased.

Actions such as encoding categorical variables, scaling numeric variables and select best features for predictions.

## **Encoding and scaling**

The encoding technique used in this project is Label Encoding. Label encoding is a preprocessing technique used to convert categorical data into numerical form. It involves assigning a unique integer (label) to each category in a categorical variable. This technique is commonly used when dealing with machine learning algorithms that require numerical input data, such as decision trees and support vector machines.

The technique used to scale data(numeric) is StandardScaler. StandardScaler is a preprocessing technique in machine learning used for feature scaling or standardization. It's a method that transforms your data in a way that it has a mean of 0 and a standard deviation of 1. Standardization is important for certain machine learning algorithms that are sensitive to the scale of input features. StandardScaler is particularly useful when dealing with numerical features with varying units and scales.

## **Feature selection**

Feature selection is a critical preprocessing step in machine learning that helps improve model performance, reduce overfitting, enhance model interpretability, and make the modelling process more efficient. It allows you to focus on the most relevant and informative features, leading to better model generalization and more meaningful insights from the data.

In this project selectkbest and f\_regression is used. SelectKBest and f\_regression are feature selection techniques in machine learning that help us choose the most relevant features for the model. They are particularly useful when we have a large number of features, and we want to reduce the dimensionality of the dataset or improve the performance of the model by selecting the most informative features. Some of the unique features of selectkbest and f\_regression are:

SelectKBest:

* SelectKBest is a feature selection method provided by scikit-learn, a popular Python machine learning library.
* It selects the top K features with the highest scores based on a specified scoring function.
* The selection is based on statistical tests that assess the relationship between each feature and the target variable.

f\_regression:

* f\_regression is one of the scoring functions commonly used with SelectKBest.
* It is a statistical test used to assess the linear relationship between each numerical feature and the target variable.
* Specifically, it calculates the F-statistic and the p-value for each feature, and you can use these values to determine the importance of each feature.

Using the above two feature selection 15 best features have been selected for model building and predictions.

## **Model building and interpretation**

As, previously mentioned the dataset is imbalanced and it might lead to biased results. In order to eliminate that SMOTE (Synthetic Minority Over-sampling Technique) is used.

SMOTE is a technique used in machine learning to address the problem of class imbalance, particularly in classification tasks. Class imbalance occurs when one class (the minority class) has significantly fewer samples than another class (the majority class). SMOTE helps balance the class distribution by generating synthetic samples for the minority class.

**MODELS:**

* **Logistic Regression:**

Logistic Regression is a statistical and machine learning model used for binary classification and multi-class classification tasks. It is a fundamental model for classification tasks, known for its simplicity, interpretability, and effectiveness in a wide range of applications.

At first, the logistic regression model is built and predictions are made without tuning the model. The predictions of the model is evaluated using accuracy score, confusion matrix and classification report.

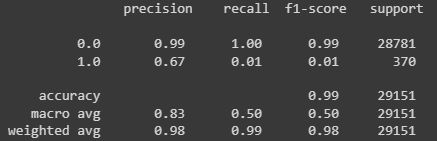
* The **accuracy score** of the untuned logistic regression model is 0.987 or 98.7%.
* The **confusion matrix** of the same model is:

[[28780 1]

[ 368 2]]

Here there is one false positive and 368 false negative.

* Finally, the classification report of the logistic regression model:



**Fig 23**

The recall of the model is 1 and which is good.

* **Naïve Bayes:**

Naive Bayes is a simple yet powerful probabilistic machine learning algorithm used primarily for classification and text categorization tasks. It is based on Bayes' theorem and makes a strong, naive assumption that all features are conditionally independent given the class label, which is why it's called Naïve.

Similar to the above model, the naïve bayes model is also built and used for prediction.

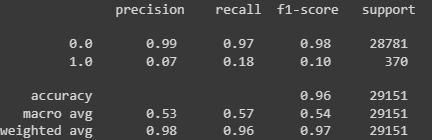
* The **accuracy score** of the untuned naïve bayes model is 0.960or 96%.
* The **confusion matrix** of the same model is:

[[27935 846]

[ 304 66]]

Here there is 846 false positive and 304 false negative, which is higher than the previous model

* Finally, the classification report of the logistic regression model:



**Fig 24**

The recall of the model is 0.97and also other parameters are lesser than that of the logistic regression model.

* **Decision Tree Classifier:**

Decision Tree Classifier is a popular supervised machine learning algorithm used for both classification and regression tasks. It is a tree-like model that makes decisions by recursively splitting the dataset into subsets based on the most significant attribute(s) at each node.

Below are the performance metrics of the untuned decision tree classifier model:

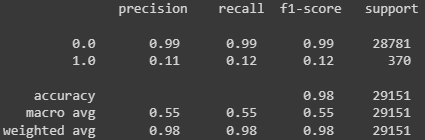
* The **accuracy score** of the untuned decision tree classifier model is 0.977 or 97.7%.
* The **confusion matrix** of the same model is:

[[28450 331]

[ 327 43]]

Here there is 331 false positive and 327 false negative, which is higher than the logistic regression model and lower than the naïve bayes model.

* Finally, the classification report of the logistic regression model:



**Fig 25**

Though the model’s metrics are good, it just fell short of logistic regression model.

* **Random Forest Classifier:**

A Random Forest Classifier is an ensemble machine learning model that builds a collection of decision trees during training and combines their predictions to make more accurate and robust classifications. It is an extension of the single Decision Tree model and is known for its high predictive accuracy and resilience to overfitting.

The performance metrics of the Random Forest Classifier model is given below:

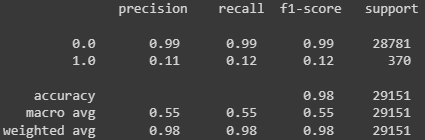
* The **accuracy score** of the untuned decision tree classifier model is 0.986 or 98.6%.
* The **confusion matrix** of the same model is:

[[28738 43]

[ 352 18]]

Here there is 43 false positive and 352 false negative, which is higher than the logistic regression model.

* Finally, the classification report of the logistic regression model:



**Fig 26**

Though the model’s accuracy is closer to that of Logistic regression model, the predictions in confusion matrix does not help it choose over the later model.

## **Model Tuning and business implication**

Though multiple machine Learning models has been built and used for predicting the defaulters, it is still not a great idea relay on its predictions. Out of all the machine Learning model Logistic Regression model and Random Forest Classifier model performed well. Yet, it is too early to conclude the best model for the given scenario, as it is yet to be fine-tuned.

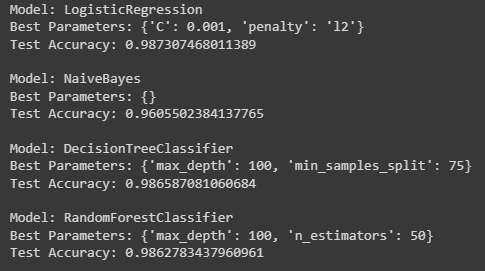
**GridSearchCV:**

GridSearchCV, or Grid Search Cross-Validation, is a technique used in machine learning to systematically search for the best combination of hyperparameters for a model. It's a method for hyperparameter tuning, which involves finding the set of hyperparameters that result in the best performance of a machine learning algorithm on a given dataset.

Some of the Importance of GridSearchCV are:

* GridSearchCV automates the process of hyperparameter tuning, saving time and effort compared to manual tuning.
* It systematically explores a range of hyperparameter values, ensuring to find the combination that works best for your specific problem.
* It helps prevent overfitting by using cross-validation to assess model performance.
* By using cross-validation, GridSearchCV provides a more reliable estimate of a model's performance than a single train-test split.

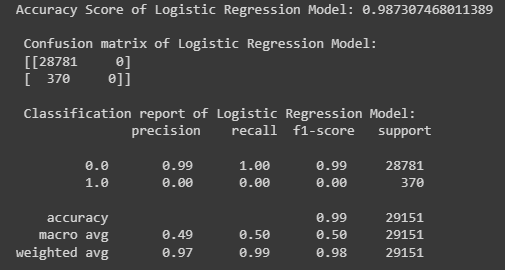
Using the GridSearchCV hyperparameters of each model is tuned and the best parameter for each model is found. Below Image shows all the best parameters of each model along with its accuracy score.



**Fig 27**

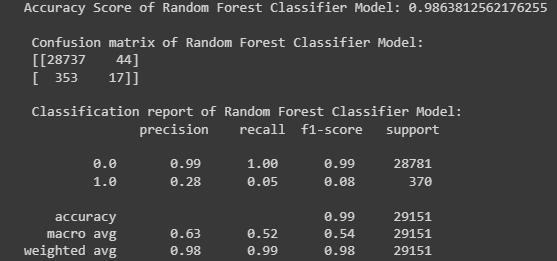
The figure 27 shows the best params for each machine learning model and its accuracy score. Just like a mirror image of the accuracy of the untuned models, the accuracy of the hyper tuned models is.

Even though, Accuracy score of the Logistic Regression model and Random Forest Classifier model is almost same, there is enough evidence to differentiate them in terms of other metrics. Below figures 28 and 29 confirms that.



**Fig 28**

Figure 28 shows all the performance metrics of the hyper tuned Logistic Regression model. Figure 29 shows all the performance metrics of the hyper tuned Random Forest classifier model.



**Fig 29**

Upon comparing both the models, it can be clearly seen that the logistic regression model out performs Random Forest Classifier in terms of over performance metrics. So, it can be concluded that the best model for the given scenario is Logistic Regression Model.

# **INSIGHTS AND RECOMMENDATIONS**

* Logistic Regression is the best model.
* Account status, merchant group, merchant category, age are some of the best features.
* Most time is spent by children clothing and nurturing products in merchant category.
* Diversity Entertainment is the High probability of default.
* Interest rates can be played around and see if there is any change in customer behaviour.
* Most of the credit card users are in 20-40 age group. Bank can create marketing strategies such as offers on most purchase made by them, etc.
* Cashbacks can be given for customers on day today usage such as grocery, fuel filling, etc.
* We can segment customers and give loyalty points to make them use more and at the same time we track which customers are less likely default.