DEFAULT OF CREDIT CARD CLIENTS ANALYSIS

Project by:

Shashank Gupta

Shivika Gupta

ABSTRACT:

1. **Introduction**

   The project is being prepared to understand the problems faced by the banks when a credit card is being issued to the customers to avoid the problem of default. Many customers tend to utilize their credit cards beyond their repaying capabilities which eventually results in high debt accumulation. In the last few years, credit card issuers have become one of the major consumer lending products. A credit card is a flexible tool by which you can use the bank's money for a short period of time. If you accept a credit card, you agree to pay your bills by the due date listed on your credit card statement. Otherwise, the credit card will have defaulted.

   There is much research on credit card lending, it is a widely researched subject. Many statistical methods have been applied to developing credit risk prediction, such as CatBoost Classifier, LightGBM Model, Logistic regression, K-nearest neighbour (KNN) classifiers, and probabilistic classifiers such as Bayes classifiers.

   A lot of exploratory data analysis has been done on the data set which helps us to understand which factors influence the credit card defaulters more and which are not that important. In all, there are 25 variables in the data set and it also includes the target variable i.e. whether the credit card client will default or not. Also, using the EDA it can be concluded that there are no missing values in the dataset which makes the cleaning process easier and a number of 6,636 out of 30,000 (22%) of clients will default the payment of credit card which has been inferred from the dataset.

   We can form different models using different combination of variables and check which model performs better in predicting the credit card default status like using the credit card history of a client for past 2 months or using the data for past 6 months and then checking the accuracy of both the combination of variables and then we can use the model with better accuracy for further analysis.

   In our project, we can further analyse the data using descriptive analysis, exploratory data analysis and then use the data to train the models like SVM, Random Forest classifier and Decision Tree classifier to predict the target variable and check the accuracy of the models using different combinations of variables like best subset selection or backward/forward selection and then select models with the best accuracy which can be calculated using validation sets and cross-validations so as to get better accuracy. Confusion matrix and ROC will be used to describe for each class the data that is correctly classified in a class.

2. **Problem Statement**

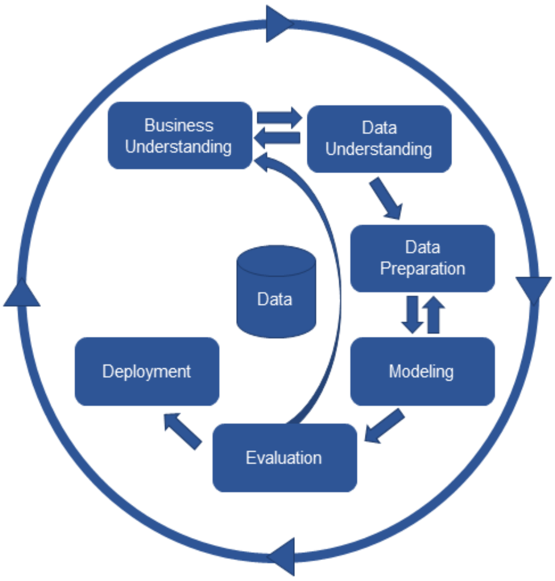
   The project is being prepared to understand whether the credit card customer will make payment default in the next month or not based on number of features.

3. CRISP-DM

In this project, default of credit client dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients.

There are 25 attributes out of which 23 are predictive attributes,1 non-predictive and 1 target variable i.e. default payment next month. Main aim of this project is to form structures and patterns to draw inferences about trends in the default of credit card by the client based on various factors such as the amount of given credit, marital status and education of each client, and 6 months bill history and payments, .

CRISP-DM, which stands for Cross-Industry Standard Process for Data Mining, is an industry-proven way to guide the data mining efforts. By following the methodology of the CRISP-DM process, every step of the project is defined. The following diagram explains it more:



To explain the process further, business understanding involves the study of the business domain. In this case, we study the financial industry where banking is one of the key players and to understand the problem at hand. Data Understanding deals with the study of the various independent variables in the dataset and the dependent variable. Data cleaning is followed by data understanding. Here, we deal with the missing values, outliers and NA values. As we want to understand each of the variables, the correlations among them and their relationship with the predictor variable, we proceed with Exploratory Data Analysis. As part of this, we do univariate, bivariate multivariate and correlation matrices. This also marks the final step in which we prepare the data for the modelling step.

Modelling is the step wherein the technique used for modelling is identified and stated along with any assumptions if required. In this case Random Forest Classifier, SVC, AdaBoost Classifier and Gradient Boosting Classifier algorithms are identified. This is also the stage where the dataset is divided into testing and training datasets. The model is built and assessed on the training dataset first. The parameters are well defined and revised according to the need after the assessment procedure.

Evaluation is the stage where data mining results are assessed, and approved models are finalized for implementation further. This is the stage where the entire process is summarized and those activities which were missed are highlighted for use in the future. All those activities that are approved to be used further are also stated to finish finalizing the exact modelling procedure. The decision on how to proceed is stated along with the rationale.

Deployment is the final step in the process wherein we summarize the strategy to deploy the project and how to perform the steps in order to implement in real-world.

METHODS AND ANALYSIS

1. Data Description

The data has been sourced from:
https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

The dataset being used for this project contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

There are 25 variables considered for the purpose of study, out of which 23 are predictive variables and 1 is non-predictive variable. The variable default payment next month is the target variable, which is a categorical variable where 0 is client does not make a default payment next month while 1 is client will make a default payment next month.

There are 25 variables that are the following:

- ID: ID of each client

- LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit

- SEX: Gender (1=male, 2=female)

- EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)

- MARRIAGE: Marital status (1=married, 2=single, 3=others)

- AGE: Age in years

- PAY_0: Repayment status in September 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, … 8=payment delay for eight months, 9=payment delay for nine months and above)

- PAY_2: Repayment status in August 2005 (scale same as above)

- PAY_3: Repayment status in July 2005 (scale same as above)

- PAY_4: Repayment status in June 2005 (scale same as above)

- PAY_5: Repayment status in May 2005 (scale same as above)

- PAY_6: Repayment status in April 2005 (scale same as above)

- BILL_AMT1: Amount of bill statement in September 2005 (NT dollar)

- BILL_AMT2: Amount of bill statement in August 2005 (NT dollar)

- BILL_AMT3: Amount of bill statement in July 2005 (NT dollar)

- BILL_AMT4: Amount of bill statement in June 2005 (NT dollar)

- BILL_AMT5: Amount of bill statement in May 2005 (NT dollar)

- BILL_AMT6: Amount of bill statement in April 2005 (NT dollar)

- PAY_AMT1: Amount of previous payment in September 2005 (NT dollar)

- PAY_AMT2: Amount of previous payment in August 2005 (NT dollar)

• PAY_AMT3: Amount of previous payment in July 2005 (NT dollar)

• PAY_AMT4: Amount of previous payment in June 2005 (NT dollar)

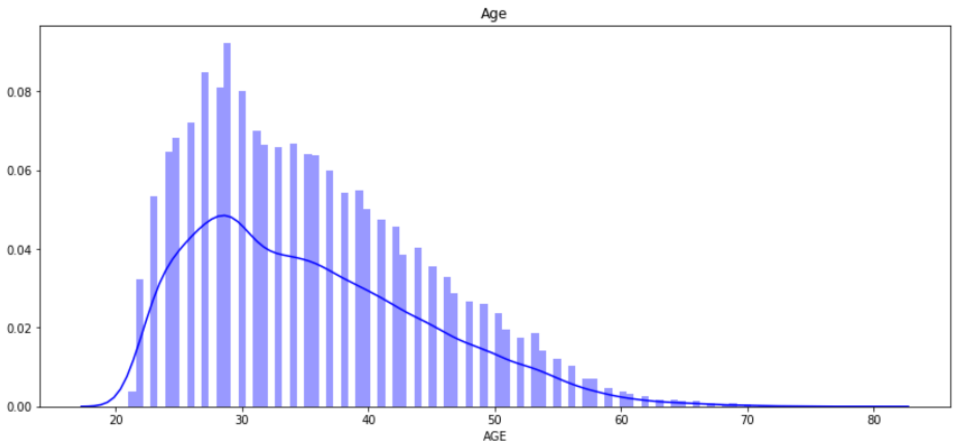
• PAY_AMT5: Amount of previous payment in May 2005 (NT dollar)

• PAY_AMT6: Amount of previous payment in April 2005 (NT dollar)

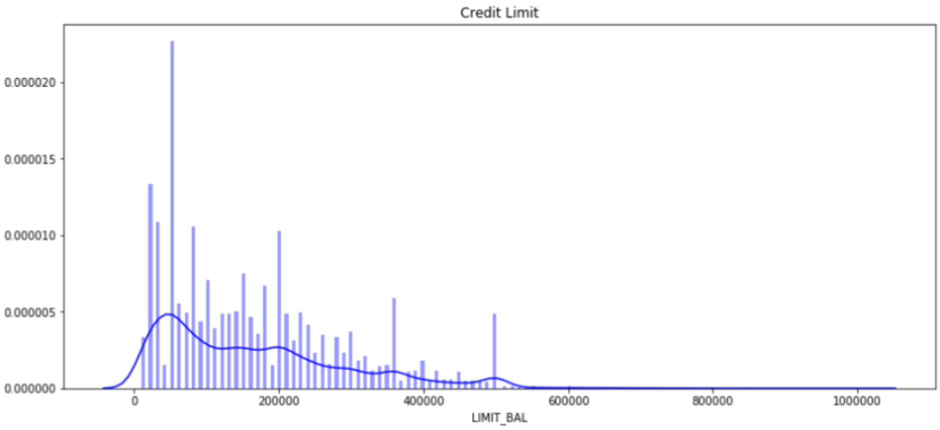
• default.payment.next.month: Default payment (1=yes, 0=no)

2. Univariate analysis

- **Analysis of Age**



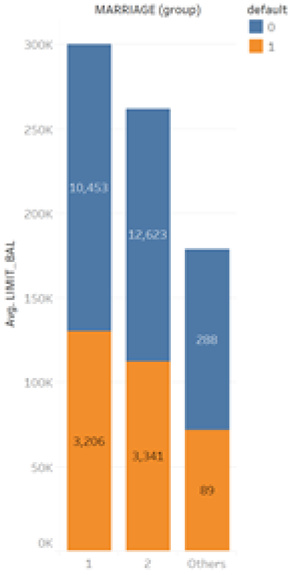
- **Analysis of Credit Limit**



3. Multivariate Analysis

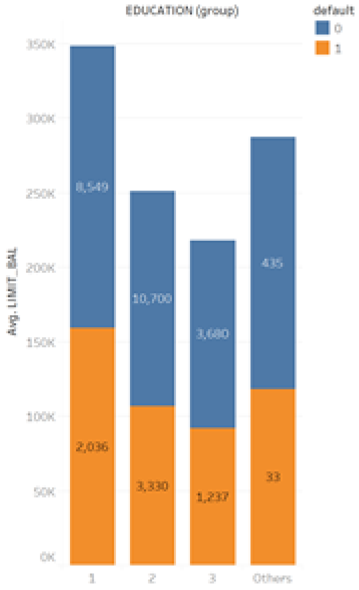
- **Analysis of Limit Balance Variable**

Comparision of Limit Balance, Marital Status and default payment next month status of the user



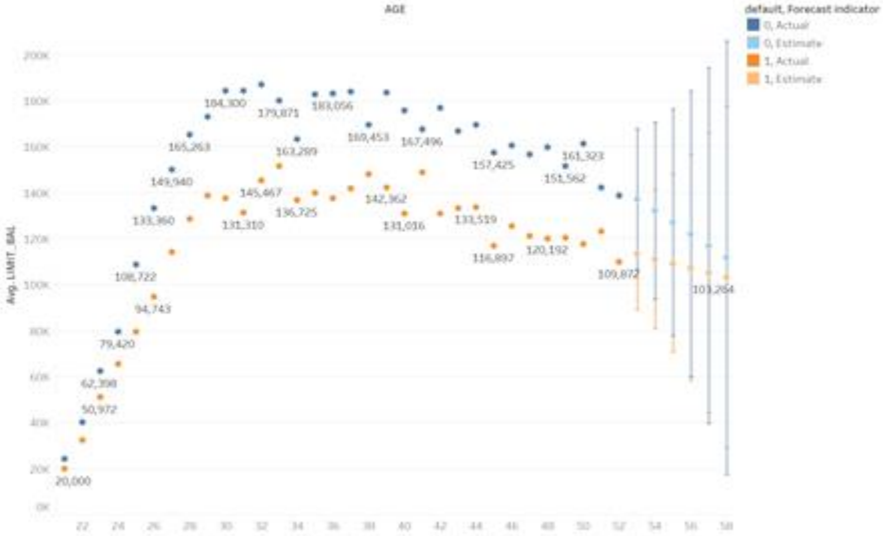
Average of LIMIT_BAL (actual & forecast) for each MARRIAGE (group). Colour shows details about default. The marks are labelled by count of MARRIAGE.

Comparision of Limit Balance, Education and default payment next month status of the user



Average of LIMIT_BAL for each EDUCATION (group). Colour shows details about default. The marks are labelled by count of EDUCATION.

Comparision of Limit Balance, Age and default payment next month status of the user

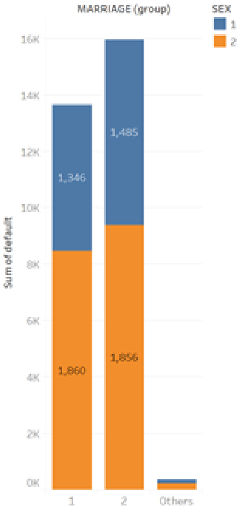


Average of LIMIT_BAL (actual & forecast) for each AGE. Colour shows details about default and forecast indicator. The marks are labelled by average of LIMIT_BAL (actual & forecast)..

In this first graphical analysis, we have kept limit balance as a constant while changing other variable to find the default payment. In the first graph, avg. limit balance has been compared with the marital status where 3341 clients in the 2nd category which represents that the client is single has the maximum default payment. In the second graph, avg. limit balance has been compared with the education of the client where 2036 clients in the first category which represents that the client has completed graduate school has the maximum default payment. In the third graph, avg. limit balance has been compared with the age of the client where the average limit balance below 146k tends to make a default payment.

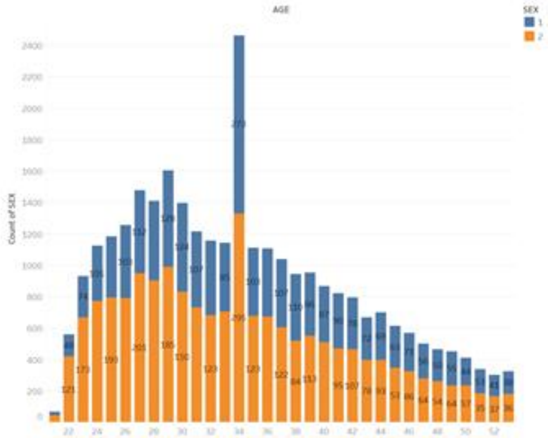
- **Analysis of Marital Status**

Comparision of Marital Status, Sex and default payment next month status of the user



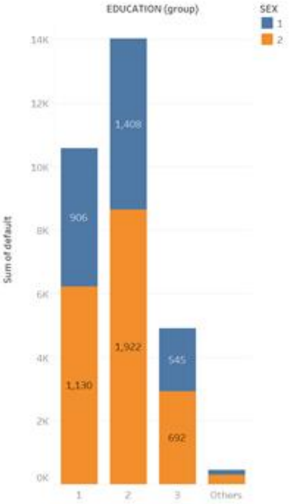
Count of default for each MARRIAGE (group). Colour shows details about SEX. The marks are labelled by sum of default.

Comparision of Sex, Age and default payment next month status of the user



Count of SEX for each AGE. Colour shows details about SEX. The marks are labelled by sum of default.

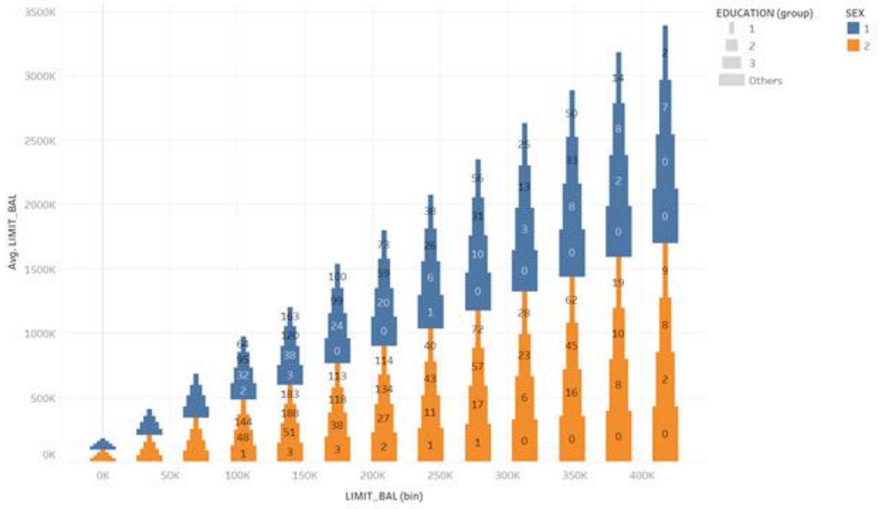
Comparision of Marital Status, Education and default payment next month status of the user



Count of default for each EDUCATION (group). Colour shows details about SEX. The marks are labelled by sum of default.

In this graphical analysis, marital status has been kept as a constant while changing other variable to find the default payment. In the first graph, marital status has been compared with the sex where 1860 client's in 2nd category of sex which represents female clients with graduate school education has the maximum default payment. In the second graph, age of the client has been compared with the sex of the client where 295 female clients of the age group of 34 years has the maximum default payment. In the third graph, marital status is compared with the sex and education of the client where 1922 female clients with university education tends to make the most default payments.

- **Visualisation of Education**



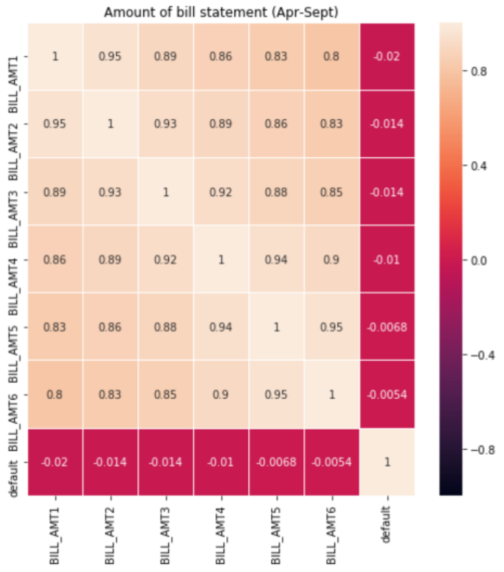
Comparision of Limit Balance, Sex, Education and default payment next month status of the user

The trend of average of LIMIT_BAL for LIMIT_BAL (bin). Colour shows details about SEX. Size shows details about EDUCATION (group). The marks are labelled by sum of default.

In this graph, the comparison between limit balance, sex and education of the client is analyzed to find the number of defaulters. From the graph we can see that 188 female university clients with the average limit balance of around 145000, tends to make the maximum default payments.

4. CORRELATION MATRICES

- **Correlation Matrix of Amount of Bill Statement**

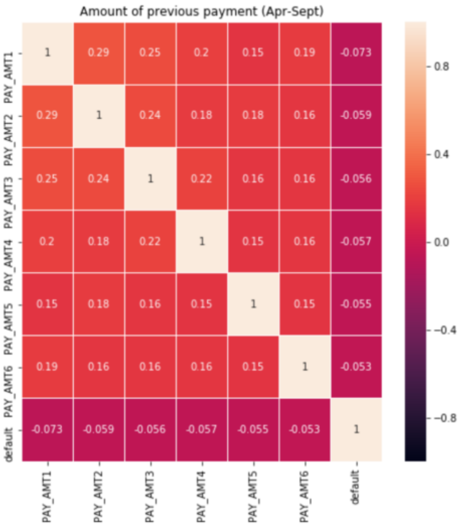


Amount of bill statement (Apr-Sept)

	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	default
BILL_AMT1	1	0.95	0.89	0.86	0.83	0.8	-0.02
BILL_AMT2	0.95	1	0.93	0.89	0.86	0.83	-0.014
BILL_AMT3	0.89	0.93	1	0.92	0.88	0.85	-0.014
BILL_AMT4	0.86	0.89	0.92	1	0.94	0.9	-0.01
BILL_AMT5	0.83	0.86	0.88	0.94	1	0.95	-0.0068
BILL_AMT6	0.8	0.83	0.85	0.9	0.95	1	-0.0054
default	-0.02	-0.014	-0.014	-0.01	-0.0068	-0.0054	1

The first correlation matrix explains the relationship between the bill amount of the consecutive months and gives the visibility whether the credit card client will make a default payment next month or not. Bill amount 1 column represents the bill amount in the first month. But as we see, its relationship with the following respective consecutive months, there is not much of a difference in the bill amount which means that the client still has that amount of spending power which gives -2% probability that the client will make a default payment and similar in the other cases.

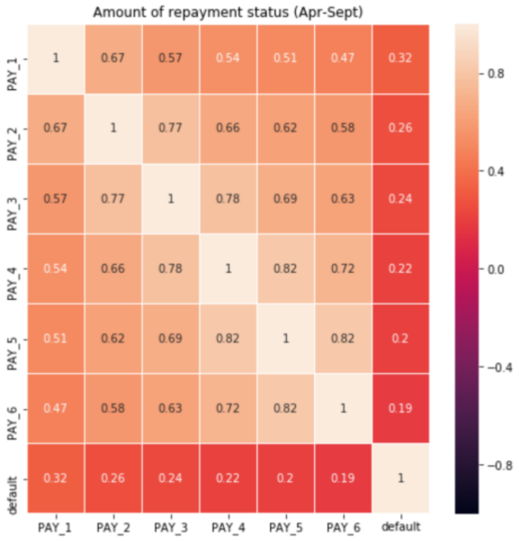
- **Correlation Matrix of Amount of Previous**


Amount of previous payment (Apr-Sept)

	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	default
PAY_AMT1	1	0.29	0.25	0.2	0.15	0.19	-0.073
PAY_AMT2	0.29	1	0.24	0.18	0.18	0.16	-0.059
PAY_AMT3	0.25	0.24	1	0.22	0.16	0.16	-0.056
PAY_AMT4	0.2	0.18	0.22	1	0.15	0.16	-0.057
PAY_AMT5	0.15	0.18	0.16	0.15	1	0.15	-0.055
PAY_AMT6	0.19	0.16	0.16	0.16	0.15	1	-0.053
default	-0.073	-0.059	-0.056	-0.057	-0.055	-0.053	1

This correlation matrix explains the relationship between the payment amount of the consecutive months and gives the visibility whether the credit card client will make a default payment next month or not. Pay amount 1 column represents the payment made by the client in the first month. But as we see, its relationship with the following consecutive months becomes less which means that the probability will be 29% that the client will pay the amount of first month in the second month, 25% person clients will pay the first month bill in the 3rd month and similarly in the other consecutive month. This results into -7.3% probability that the client will make a default payment and similarly in the other cases.

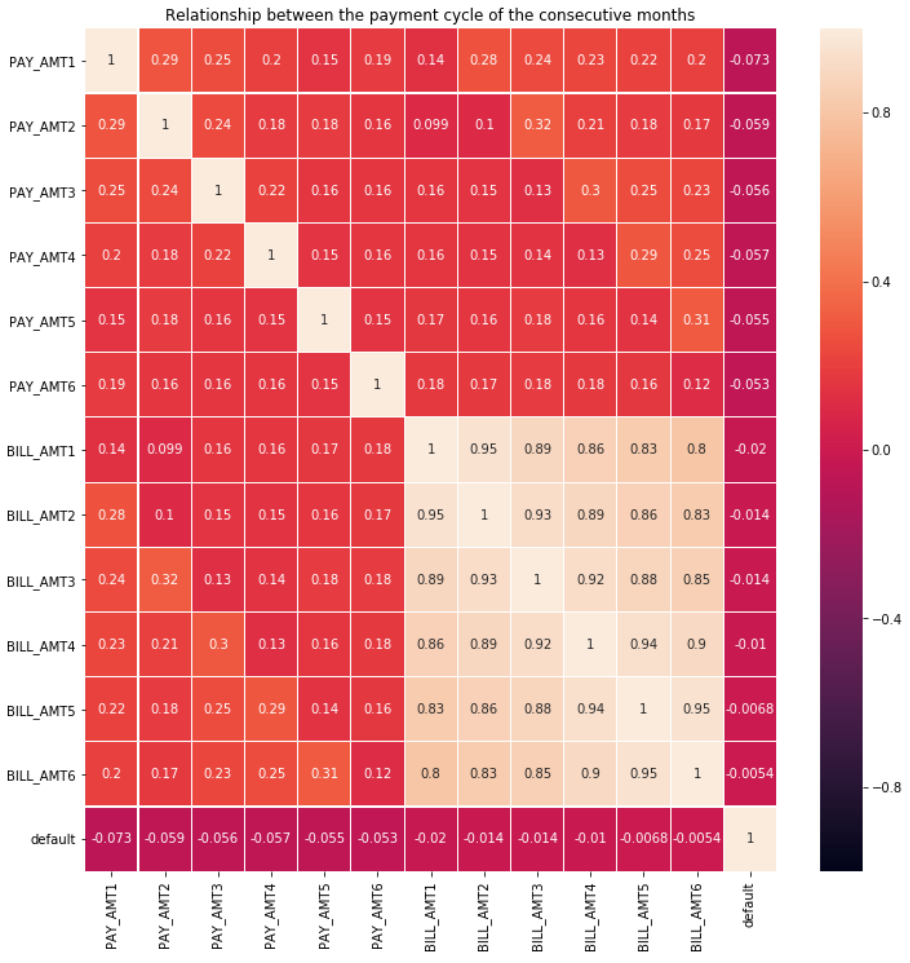
- **Correlation Matrix of Repayment Status**


Amount of repayment status (Apr-Sept)

	PAY_1	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	default
PAY_1	1	0.67	0.57	0.54	0.51	0.47	0.32
PAY_2	0.67	1	0.77	0.66	0.62	0.58	0.26
PAY_3	0.57	0.77	1	0.78	0.69	0.63	0.24
PAY_4	0.54	0.66	0.78	1	0.82	0.72	0.22
PAY_5	0.51	0.62	0.69	0.82	1	0.82	0.2
PAY_6	0.47	0.58	0.63	0.72	0.82	1	0.19
default	0.32	0.26	0.24	0.22	0.2	0.19	1

This correlation matrix explains the relationship between the repayment status of the consecutive months and gives the visibility whether the credit card client will make a default payment next month or not. So, a client in the month 1 has 32% chance that it will make a default payment and similar in other cases.

- **Correlation Matrix of Relationship Between the Payment Cycle of the Consecutive Months**



Relationship between the payment cycle of the consecutive months

	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	default
PAY_AMT1	1	0.29	0.25	0.2	0.15	0.19	0.14	0.28	0.24	0.23	0.22	0.2	-0.073
PAY_AMT2	0.29	1	0.24	0.18	0.18	0.16	0.099	0.1	0.32	0.21	0.18	0.17	-0.059
PAY_AMT3	0.25	0.24	1	0.22	0.16	0.16	0.16	0.15	0.13	0.3	0.25	0.23	-0.056
PAY_AMT4	0.2	0.18	0.22	1	0.15	0.16	0.16	0.15	0.14	0.13	0.29	0.25	-0.057
PAY_AMT5	0.15	0.18	0.16	0.15	1	0.15	0.17	0.16	0.18	0.16	0.14	0.31	-0.055
PAY_AMT6	0.19	0.16	0.16	0.16	0.15	1	0.18	0.17	0.18	0.18	0.16	0.12	-0.053
BILL_AMT1	0.14	0.099	0.16	0.16	0.17	0.18	1	0.95	0.89	0.86	0.83	0.8	-0.02
BILL_AMT2	0.28	0.1	0.15	0.15	0.16	0.17	0.95	1	0.93	0.89	0.86	0.83	-0.014
BILL_AMT3	0.24	0.32	0.13	0.14	0.18	0.18	0.89	0.93	1	0.92	0.88	0.85	-0.014
BILL_AMT4	0.23	0.21	0.3	0.13	0.16	0.18	0.86	0.89	0.92	1	0.94	0.9	-0.01
BILL_AMT5	0.22	0.18	0.25	0.29	0.14	0.16	0.83	0.86	0.88	0.94	1	0.95	-0.0068
BILL_AMT6	0.2	0.17	0.23	0.25	0.31	0.12	0.8	0.83	0.85	0.9	0.95	1	-0.0054
default	-0.073	-0.059	-0.056	-0.057	-0.055	-0.053	-0.02	-0.014	-0.014	-0.01	-0.0068	-0.0054	1

This correlation matrix gives a broader angle to understand the relationships between different payment cycle variables and their effect on the default payment. So, for example bill amount 1 has 14% chance that the pay amount 1 will be cleared and you can create a similar analysis for other cases as well.

5. Data Cleaning

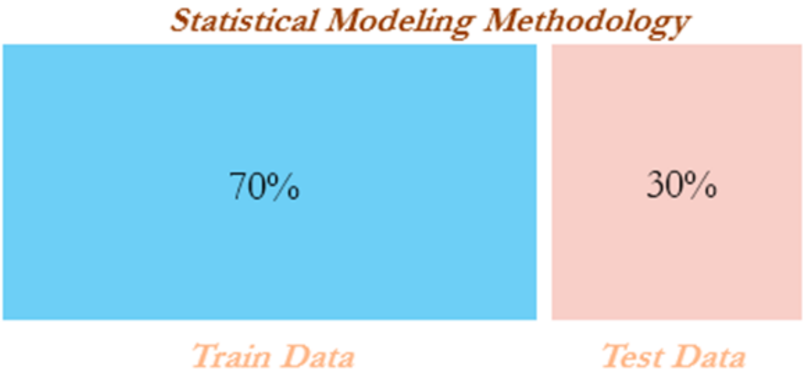
For data cleaning, analysis was done to find any null or missing values but there were some outliers in Limit balance and age and these outliers were replaced by the median of the respective columns. Also, ID variable is not significant hence we removed it.

RESULTS:

After completing the analysis, I summarised the findings so that I could apply algorithms like random forest classifier, ADA Boost Classifier, SVM, etc.

1. **Split data**

The dataset was divided into the training and test dataset in 7:3 ratio to find the accuracy of the model.



We applied four algorithms - Random Forest Classifier, SVC, ADA Boost Classifier and Gradient Boosting Classifying algorithms – that helped in building models with the train dataset.

1) **Random Forest Classifier:**
A random forest is a meta estimator that fits several decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

2) **ADA Boost Classifier:**
An AdaBoost classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.

3) **Support Vector Machine Classifier:**

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points.

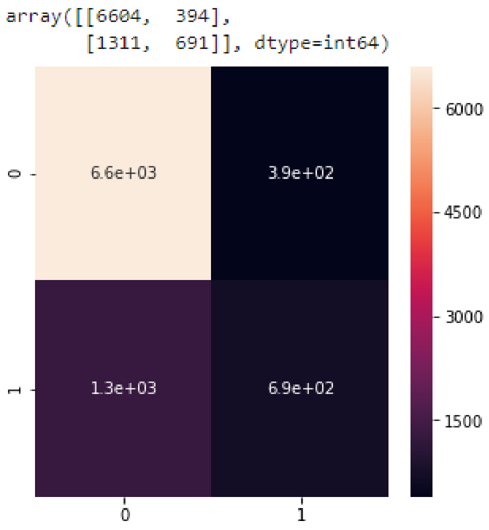
4) **Gradient Boosting Classifier:**
GB builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage n-class regression trees are fit on the negative gradient of the binomial or multinomial deviance loss function. Binary classification is a special case where only a single regression tree is induced.

Summary of Accuracy of the models:

Algorithm	**Accuracy**
Random Forest Classifier	81%
ADA Boost Classifier	82%
Support Vector Machine Classifier	77%
Gradient Boosting Classifier	79%

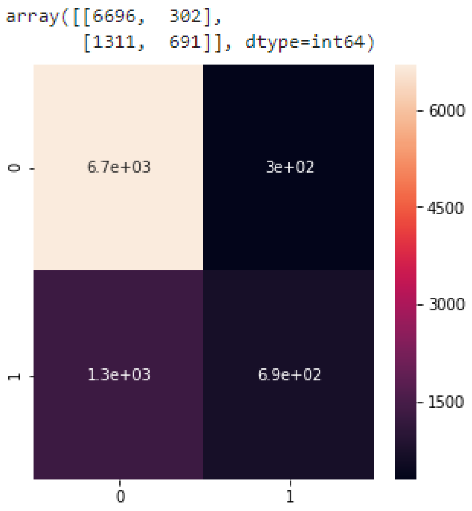
DISCUSSION AND CONCLUSION

1. **Conclusion (Confusion Matrix)**
ROC shows how better the model distinguishes between different classes i.e. default by credit card client or not. Prediction result or any point in the confusion matrix represents one space in ROC.



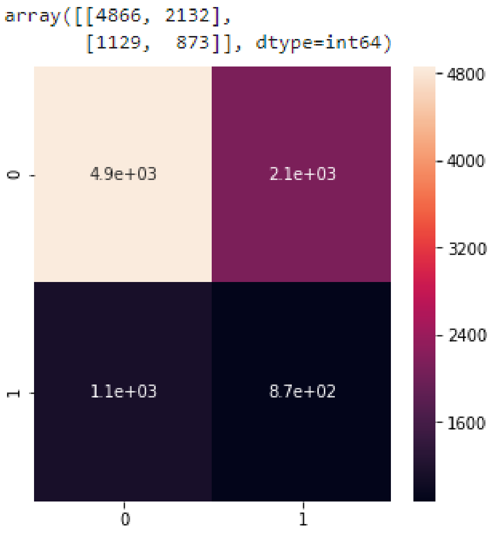
```
array([[6604,  394],
       [1311,  691]], dtype=int64)
```

Random Forest Classifier



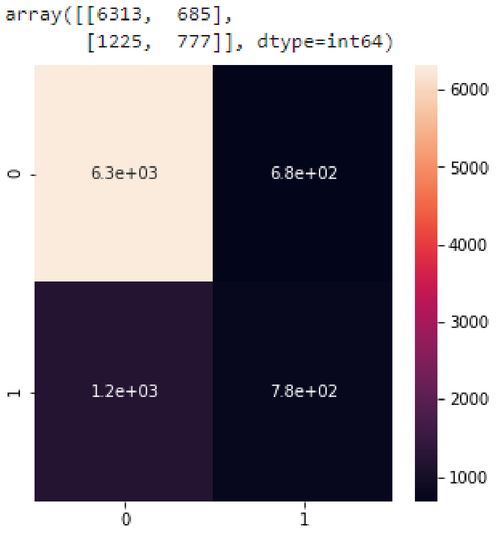
```
array([[6696,  302],
       [1311,  691]], dtype=int64)
```

Ada Boost Classifier



```
array([[4866, 2132],
       [1129,  873]], dtype=int64)
```

Support Vector Machine



```
array([[6313,  685],
       [1225,  777]], dtype=int64)
```

Gradient Boosting Classifier

After careful consideration we found the following:

- variables that affect the target variable i.e. default payment next month:

i. Limit Balance- Users with $145000 limit tends to make maximum defaults

ii. Marital Status- Clients those who are single tend to make maximum defaults

iii. Age- At the age of 34, clients tend to make maximum defaults.

iv. Bill Amount and Previous Payment History for 6 months tends to have direct impact on default payments.

- ADA Boost classifier gives the best accuracy i.e. 82%.

- Support Vector Machine Classifier is does not perform well on big datasets.

2. Limitations

1. The data was collected from only one geography so it might be biased for other geographies.
2. The range of Limit Balance and Age is very high as well as very low which resulted in outliers. This can be solved by building different clusters of different ranges and finding accuracy of different clusters.
3. Due to the computation power issue few algorithms couldn't be implemented to increase the accuracy. Computation power can be increased by using external GPUs like AWS, Azure, Google Cloud, etc..

3. Recommendation (Future Studies)

I. We can formulate more models using techniques like subset selections, forward selection and more to get better accuracy.

II. More relationships between independent variables can be analysed to overcome the problem of multicollinearity.

III. Increase the computational power issue to solve this problem with more complex algorithms.

IV. Clustering of data based on age of the users as one of the important deciding factors to implement the model and get better accuracy.

BIBLIOGRAPHY

1) Default Credit Card Clients Dataset, https://www.kaggle.com/uciml/default-ofcredit-card-clients-dataset/

2) Sharma, Sunakshi & Mehra, Vipul. (2018). Default Payment Analysis of Credit Card Clients. 10.13140/RG.2.2.31307.28967.

3) RandomForrestClassifier, 3.2.4.3.1. sklearn.ensemble.RandomForestClassifier — scikit-learn 0.22.2 documentation

4) SVM, https://scikit-learn.org/stable/modules/svm.htm

5) https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html

6) https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

7) https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html