Data Mining on FinTech related datasets

Amarnath Venkataramanan
MSc FinTech
National College of Ireland
Dublin, Ireland
x18105149@student.ncirl.ie

***Abstract*—In this project, the various data mining methods will be applied to three of fintech related datasets in order to discover the insights from the datasets. The methods will be applied to the datasets in R studio and corresponding outputs will be shown in this paper. Then the applied methods will be compared with each other to identify how well the method is performing on each of the datasets and then finally the better method will be chosen for each of the datasets. The methods were explained in detail along with their advantages with the help of the information from the relevant papers. Every method which has been applied to the datasets has been confirmed whether it follows the data mining methodologies. Data mining methodologies like CRISP-DM, KDD and SEMMA will also be explained in detail via this paper. The process flow of the data mining methodologies will be explained and also will be made sure that whether the process flow has been followed while applying each method on the datasets. The Data mining in now considered as a major factor of the risk management process of the financial institutions. Even though various data mining tools are existing in the market, this paper allows readers to understand how the algorithm works on the dataset and how to justify that whether algorithm's prediction.**

Keywords—Data mining, FinTech, CRISP-DM, KDD

I. INTRODUCTION

[1] states that FinTech is responsible for the all kind of data-oriented activities on financial services. Whereas FinTech adopts data relevant activities such as data analytics, data mining and data de-duplication in order to obtain a knowledge from the data that leads to enhancing financial services or introducing the entirely new services in financial sector. In general, data mining is the process of determining any valuable insights or patterns from the huge pool of datasets [2]. There are two variants of data mining techniques and they are, 1. Descriptive technique and 2. Predictive technique. Descriptive technique helps the analyst to fetch the characteristics of the data given as input. Predictive technique allows the analyst to predict the hidden information from the data given as input [3]. In FinTech, data mining plays a major role to improve the various existing services of the financial firms by discovering the insights from the financial data [4]. There are three process models exist for data mining and they are namely, 1. Knowledge Discover Databases (KDD), 2. Cross-Industry Standard Process for Data Mining (CRISP-DM) and 3. Sample, Explore, Modify, Model, Assess. The detailed description of these process model will be found below in this paper [5].

A. Objective and motivation of this project

In recent times one of the major problems faced by banking institution are follows,

a. Credit card fraud
b. Credit card default
c. Loan default

The ultimate aim of this project is to predict all of the above-mentioned fraudulent activities from the correspondingly related real time datasets by performing data mining process with the help of the certain algorithms under KDD data mining methodology to each of the dataset. And also, we will compare those algorithms applied to each of the dataset and we will be identifying the better algorithms with the highest accuracy of prediction for each of the dataset. All these data mining application are comes under the risk management process of the financial institutions.

Credit card fraudulent transactions. In order to detect those fraudulent transactions of credit card currently various predictive algorithms are currently used to perform the data mining to predict which of the transactions are legitimate and fraudulent. In this project we will be discussing by applying various data mining algorithms on the real time credit card data as per the KDD data mining methodology [12].

After the financial crisis in 2008, bank and other financial institutions are investing a huge amount in the risk management. Data analytics / mining shares a major portion of the potential risk management and investment on it. Even though with high level prediction algorithm and data mining tools available bank default is being frequent [13]. But, don't forget that "Something is better than nothing" even with high level data mining algorithm 100% detection of bank loan default is not possible atleast for some extent it would reduce bank loan default. Let us apply some of the high-level data mining algorithms to the real time bank loan default data and justify their predictions.

Another problem faced by bank is credit card default which is the situation where the user of the credit card failed to pay the debt amount within the due date. Once it happens the issuer (Bank) of the credit card will apply the penalty in addition to the due amount. There are two categories of the credit card default, they are follows,

a. Willingly credit card default
b. Unwillingly credit card default

Willingly credit card default is the situation where the users of the credit card intentionally will not pay the debt caused by credit card

Unwillingly credit card default is the situation where the users not able to pay their debts caused by credit card due to their financial or personal problems.

II. RELEVANT WORK

A. Datasets

Datasets used in this data mining process are follows,

1. **Credit card fraud:** Obtaining the dataset which consists of customer's transaction details are not easy due to the security policies and their organization's regulations. Presently, analysts are using the data from the data generators in order to test their newly created data mining tools [14]. The contributor of [15] have released the dataset for the public availability so that many analysts can use this data set for their analysis research purposes. This dataset consists of 31 variables. The feature "Time" refers to the seconds took for each transaction after first transaction in the dataset. The "Amount" term refers to the transaction amount. The "Class" feature is the response variable which is categorical with the values 0's and 1's whereas 0's represents normal transaction and 1's represents fraudulent transactions. The structure of the dataset denotes that this dataset is highly imbalanced whereas only 492 fraudulent transactions among 284807 transactions which makes the dataset highly skewed as well. Due to the data confidentiality reason some of the variables are not in detail and features have been transformed to principal components (V1, V2…, V28) via Principal Component Analysis. The identification of cardholders also made unavailable that makes each transaction are anonymous [15].

2. **Credit card default prediction (Taiwanese Dataset):** "Credit card default prediction as a classification problem" (Soui et al., 2018) is the data from Taiwan. With the help of predictive algorithm "Novel sorted smoothing" method, the probability of response variable "default payment next month" has been predicted. The response variable of this dataset is a binary categorical data which states customer credible or non-credible clients. As per [27] the prediction of the credit card default end up in the binary result as either good or bad. The prediction will help to reduce such delinquency of credit card default.

3. **P2P lending bank loan default: (Đurović, 2017)** states that the prediction of the bank loan default status is the highly dependent on the characteristics of the loan whereas the loan with short repayment duration is lesser risky than the loan with long repayment duration. [28] believes that the prediction of loan default in P2P lending usually takes duration of loan and type of loan as major prediction-affecting factors. The operational risk is considered as the crucial risk in the peer to peer lending system because of the failure to perform risk assessment on the right time. Such operational risks can be reduced to significant low level via certain standardized approaches. **[30]** believes that bidding credit is attached to lender credit in terms of P2P lending in China. Risk management is considering as a key factor in P2P lending system in order to avoid certain unfortunate circumstances like loan default. [23]

B. Methods

This section provides the brief details of various methods used in this data mining process and also discusses it advantages and disadvantages.

1. **Logistic regression:** As most of our datasets contain target variable as categorical so it is best to apply logistic regression rather than linear regression. [24] states that linear regression will not always return the prediction of the target variable between 0 and1, it might end up in beyond the range. It also expresses how the response variable depends on the predictive variables. The application of logistic regression is vast among various industries which can be used for the prediction of the certain activities they need based on the past data. [12] used logistic regression to predict the occurrences of the fraudulent credit card transactions and [16] used this model for the prediction of the credit card default. There are many more researchers and analysts have been logistic regression for the predicting purpose.

2. **K-Nearest Neighbors algorithm model:** Based on the process of clustering and similarity KNN predicts the behavior of the customer. [25] states that KNN model follows an assumption model by assuming if one entity belongs to a certain category then the entire entities of the same sample also falls under the same category. KNN model performs well with minimal error when the target dataset used is too large. The dependency of the model on the near neighbors varies with different problems. As it is mostly used in the big data the resources required by it will be way more. [25] The higher accuracy can be achieved with higher K value than 1NN. In recent times KNN being used by analyst for the various purposes of the data mining like [17] used KNN model to the automate the recording of the web usage details, [25] used KNN model in combination with Simulated Annealing model in order to predict the credit ratings and [18] used KNN model in an efficient way in order to predict the consumer credit risk.

3. **Random Forest model:** Random forest model can be used in the identification of useful features which will be ranked by the model based on the importance. It can be used for the classification and regression by generating the multitude of the decision trees. Significance of the variables are taken as deciding factor to identify the important variables. [29]. Layer construction and identify split variables among nodes are considered as the main challenges while using this model. Boosting and bagging methods are used to improve the performance of the random forest model. [31]. The P2P lending can be evaluated with the help of Random forest model and with the help of gentic algorithm optimal solution is obtained which have been detailly explained in [19].

4. **Support Vector Machine:** Support Vector Machine model is one of the supervised machine

learning algorith mostly used to perform the regrssion and classification. SVM model is the effective model if in case the target dataset is highly skewed due to the unbalanced nature. SVM is highly used by most of the financial institutions in order to differentiate the fraudulent financial activities among normal financial activities [20]. In the risk managent stages of the financial institution SVM model plays a major role in emprirical risk mitigation. (ping feng pai, 2011). [20] also discusses about how SVM model can be embedded into communal and spike detection as Hybrid SVM (HSCVM) in order to predict the fraudulent activities using credit card. SVM can also be efficiently used in [21] to evaluate credit report via data mining approach.

III. DATA MINING

Data mining is the process that includes both the machine learning and statistics in order to extract a pattern from the target dataset. The extracted pattern can be later used for the decision making in the financial institutions. The primary techniques of data mining can be categorized into two categories and they are,

1. Descriptive: It is a technique that exposes the quality of day by analyzing its properties.
2. Predictive: It is a technique that speculates the target data to predict the hidden pattern in the data.

There are three process models for the data mining and they are follows,

The **Knowledge Discovery Database (KDD)** is the step by step collaborative model [5], which consists of six steps to discover the knowledge from the input dataset [6]. [7] defines data mining as one of the steps in the KDD process.

Cross-Industry Standard Process for Data Mining (CRISP-DM) is an iterative model with total of six steps and was first introduced in 1996 [8].

Sample, Explore, Modify, Model, Assess (SEMMA) is the five steps process model for data mining developed by SAS institute [9].

A. Knowledge Discovery Database

KDD is the process of fetching the pattern or knowledge which is hidden in the input dataset. The relevant application of domain knowledge must be required by KDD. KDD is the step by step process with six stages that follows,

1. Select an appropriate dataset or a combination of certain sources of data.
2. Preprocessing involves the exploration of dataset to make it fit for further mining process.
3. Transformation of dataset will be carried out in order to convert the dataset into an appropriate format.
4. Build model (Data mining) includes the various statistical learning and machine learning algorithm in order to extract a hidden patter or knowledge from the target dataset.
5. Evaluation of model describes how good the model fits the dataset.
6. Extract the knowledge from the applied model.

The extraction of insights from the dataset via data mining enables decision making process [10].

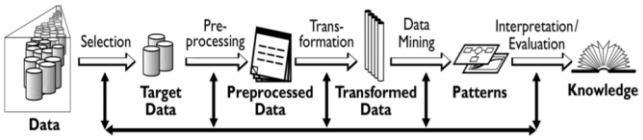


Figure 01: Various process stages of KDD.

B. CRoss-Industry Standard Process for Data Mining

Cross-Industry Standard Process for Data Mining [11] is the six-phase processing model for data mining that provides a regular framework and guidelines for data miners. The six stage of CRISP-DM have been explained below,

1. Business Understanding: Identifying the motivation and objectives of data mining.
2. Data Understanding: Collecting the data, exploring the data and identifying the quality of data.
3. Data Preparation: Transforming and cleaning the data.
4. Modelling: Application of the appropriate model for the target dataset.
5. Evaluation: Identifying the accuracy of the pattern or knowledge obtained by applying a suitable model to the target dataset.
6. Deployment: Deployment, update and improvement of models.

CRISP-DM and KDD are almost same in the functionality.

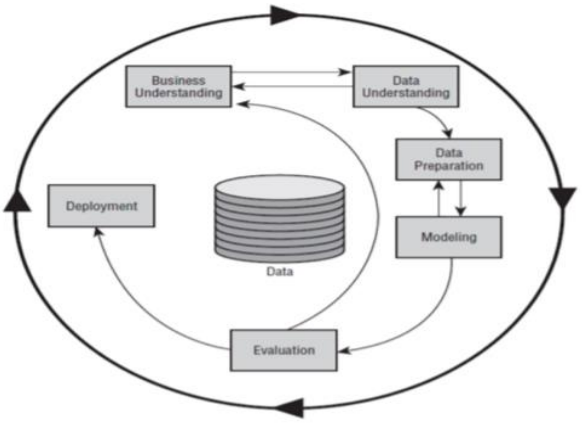


Figure 02: Process stages of CRISP-DM

C. SEMMA Process model

The SEMMA is the five-stage process model of data mining developed by SAS institute in order to provide better understanding, maintenance, development and organization for the data mining projects. SEMMA is in built with the SAS enterprise miner tool. The five stages of SEMMA process model follows,

1. Sample: It is the first stage of SEMMA process model in which a subset of large dataset that is more than enough to obtain the pattern and smaller enough to process faster.
2. Explore: Exploration of data, already discussed in the above-mentioned process model.
3. Modify: Transformation of data.
4. Model: Application of appropriate model.
5. Access: Evaluating the accuracy of the pattern or knowledge obtained from the target dataset.

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections A-D below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

D. Model Building on the datasets

Among different data mining methodologies let's adopt the Knowledge Discovery Database (KDD) and apply data mining algorithms on the dataset as per the KDD methodology.

Step01 - Selection of datasets:

There are three datasets which all are relevant to the FinTech domain. By applying four of the data mining algorithms namely, 1. Logistic regression, 2. SVM, 3. Random Forest and 4. KNN, we are intent to extract the pattern which will help the financial firm to get rid of the fraudulent financial activities. As per KDD let's start with the reading target datasets,

```
##########Step 01 - Selection of the target datasets#############
# Loading The Dataset ------------------------------------------------------
creditfraud <- read.csv("C:\Users\admin\Desktop\Data Analytics\Project/Creditfraud.csv", header=TRUE)
creditdefault <- read.csv("C:\Users\admin\Desktop\Data Analytics\Project/creditdefault.csv", header=TRUE)
loandefault <- read.csv("C:\Users\admin\Desktop\Data Analytics\Project/loandefault.csv", header=TRUE)
```

Step02 – Pre-processing of datasets

In this step let's explore the datasets with the help of their metadata. The metadata of all the three datasets can be extracted with the help of the characteristics of the features of the datasets.

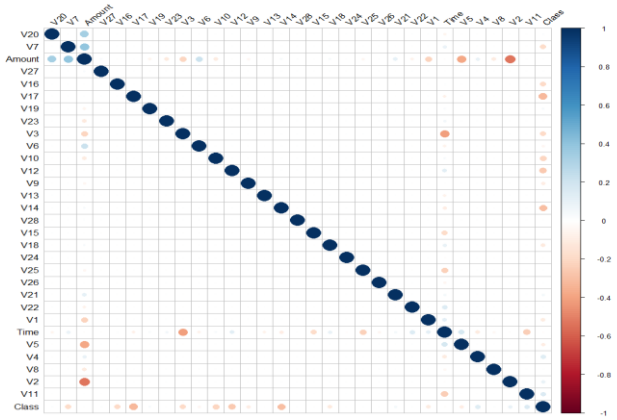
The most important three things we consider in data exploration are missing values, correlation and outliers which three can intervene the prediction so we should handle these but also, we must be careful while removing outliers because removal of certain values can seriously affect the nature of the datasets and so predictions.

1. **Missing Values:** All the three datasets used in this project doesn't have any missing values,

```
> ### Identifying missingvalues in each dataset #######
> sapply(fraud_df, FUN=function(x) {sum(is.na(x))})
  Time     V1     V2     V3     V4     V5     V6     V7     V8     V9    V10    V11    V12    V13    V14    V15    V16    V17    V18    V19
     0      0      0      0      0      0      0      0      0      0      0      0      0      0      0      0      0      0      0      0
   V20    V21    V22    V23    V24    V25    V26    V27    V28 Amount  Class
     0      0      0      0      0      0      0      0      0      0      0
> sapply(creddt_df, FUN=function(x) {sum(is.na(x))})
                        ID                  LIMIT_BAL                        SEX                  EDUCATION                   MARRIAGE
                         0                          0                          0                          0                          0
                       AGE                      PAY_0                      PAY_2                      PAY_3                      PAY_4
                         0                          0                          0                          0                          0
                     PAY_5                      PAY_6                  BILL_AMT1                  BILL_AMT2                  BILL_AMT3
                         0                          0                          0                          0                          0
                 BILL_AMT4                  BILL_AMT5                  BILL_AMT6                  PAY_AMT1                  PAY_AMT2
                         0                          0                          0                          0                          0
                 PAY_AMT3                  PAY_AMT4                  PAY_AMT5                  PAY_AMT6 default.payment.next.month
                         0                          0                          0                          0                          0
> sapply(loandt_df, FUN=function(x) {sum(is.na(x))})
                        id                       year                    issue_d                    final_d            emp_length_int
                         0                          0                          0                          0                          0
             home_ownership            home_ownership_cat           income_category                 annual_inc                 income_cat
                         0                          0                          0                          0                          0
                 loan_amount                       term                   term_cat           application_type application_type_cat
                         0                          0                          0                          0                          0
                    purpose                purpose_cat           interest_payments       interest_payment_cat             loan_condition
                         0                          0                          0                          0                          0
         loan_condition_cat                interest_rate                      grade                 grade_cat                        dti
                         0                          0                          0                          0                          0
                 total_pymnt            total_rec_prncp                 recoveries                installment                     region
                         0                          0                          0                          0                          0
>
```

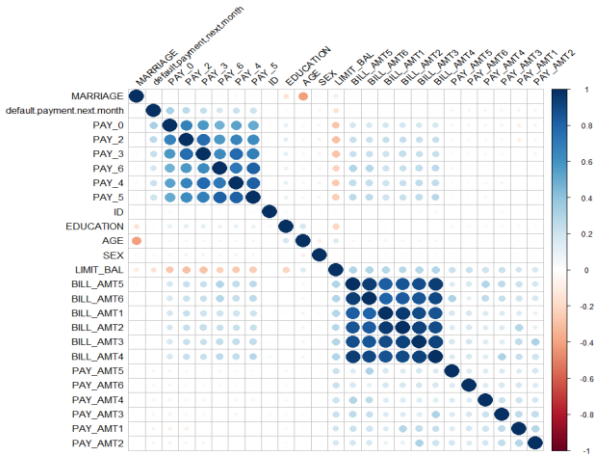
2. **Correlation:** The correlation among the variables of the datasets can be **visualized** with the help of "corrplot", following are the correlation plot of all three datasets,

Credit card fraud dataset:



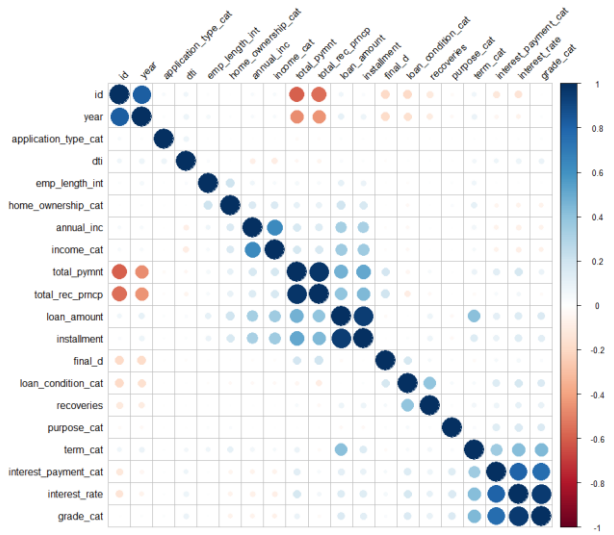
In the above plot, blue circle denotes positively correlated and red circle denotes negatively correlated. And the shade of circle denotes that how strongly variables are correlated. As previously mentioned, this dataset is highly skewed and unbalanced one so, the features are not much correlated with each other.

Credit Card Default:



Unlike the previous plot, correlation plot of the credit card default dataset has more correlation among the variables so it seems this dataset is highly balanced dataset.

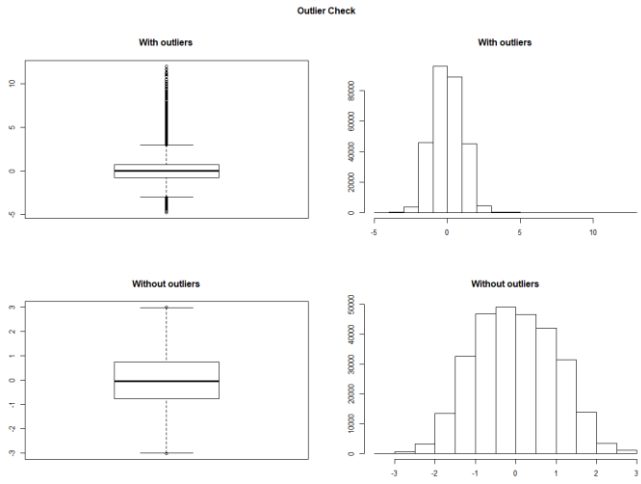
P2P Loan Default:



This dataset also more balanced than the first dataset. The negatively and un-correlated features with the target variable can be removed but in order to retain the characteristics of the dataset we could just ignore features “id” and “final_d” for the logistic regression.

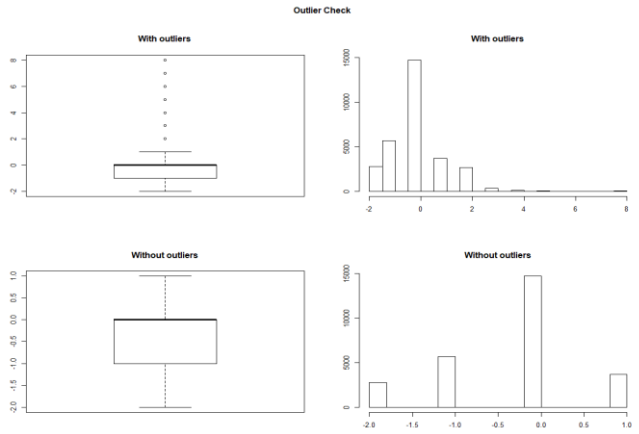
3. **Outliers:**

Credit Card Fraud: From the correlation plot we could see that "V11" feature is possess the highest correlation with target variable "Class" so, let’s find the outliers in the "recoveries" feature,



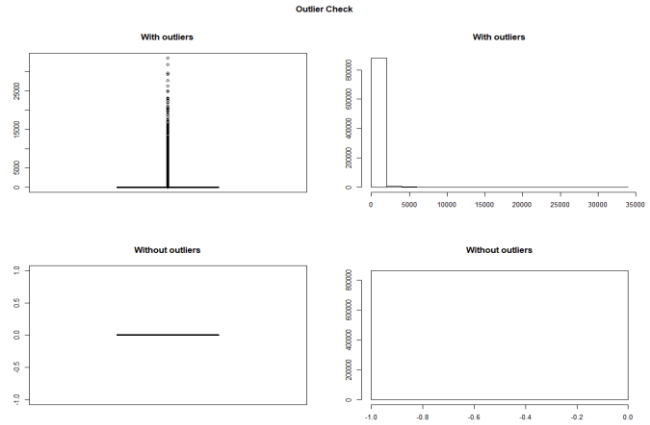
We could see that the V11 feature is normally distributed when there are no any outliers but we are not removing the outliers in order to withstand the characteristics of the dataset to apply prediction algorithm.

Credit Card Default: From the correlation plot we could see that "PAY_0" feature is possess the highest correlation with target variable "default.payment.next.month" so, let’s find the outliers in the "recoveries" feature,



We could see that the PAY_0 feature is normally distributed when there are no any outliers but we are not removing the outliers in order to withstand the characteristics of the dataset to apply prediction algorithm.

P2P loan Default: From the correlation plot we could see that "recoveries" feature is possess the highest correlation with target variable "loan_condition_cat" so, let’s find the outliers in the "recoveries" feature,



We could see that the “recoveries” feature is not looking good even when there are no any outliers so we are not removing the outliers in order to withstand the characteristics of the dataset to apply prediction algorithm.

Step03 – Transformation of datasets

Transformation of dataset mostly consists of following two activities,

1. Treating missing values
2. Treating outliers

As the three datasets used in this project doesn’t have any missing values so, there is no need for treating missing and as discussed above we are decided not to treat outliers in order to withstand the characteristics of the datasets in order to perform the prediction via applying prediction algorithms. Instead we are going to change the datatypes of certain features to the datatypes which they are supposed to be,

Credit card fraud:

In this dataset the response variable “Class” only has to be changed to categorical data type from integer,

```
> #########Step03 Transformation of datasets########
> #### Credit Card Fraud #######
> fraud_df$Class <- as.factor(fraud_df$Class)
> str(fraud_df$Class)
 Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
>
```

Credit card default:

The features like, "SEX", "Education", "Marriage", "PAY_0". "PAY_0", "PAY_0", "PAY_0", "PAY_0", "PAY_0" and "default.payment.next.month" are transformed to categorical variables which are they supposed to be,

```
> ####### Credit Card default ########
> creddt_df$SEX <- as.factor(creddt_df$SEX)
> str(creddt_df$SEX)
 Factor w/ 2 levels "1","2": 2 2 2 2 1 1 1 2 2 1 ...
> creddt_df$EDUCATION <- as.factor(creddt_df$EDUCATION)
> str(creddt_df$EDUCATION)
 Factor w/ 7 levels "0","1","2","3",..: 3 3 3 3 3 2 2 3 4 4 ...
> creddt_df$MARRIAGE <- as.factor(creddt_df$MARRIAGE)
> str(creddt_df$MARRIAGE)
 Factor w/ 4 levels "0","1","2","3": 2 3 3 2 2 3 3 3 2 3 ...
> creddt_df$PAY_0 <- as.factor(creddt_df$PAY_0)
> str(creddt_df$PAY_0)
 Factor w/ 11 levels "-2","-1","0",..: 5 2 3 3 2 3 3 3 3 1 ...
> creddt_df$PAY_2 <- as.factor(creddt_df$PAY_2)
> str(creddt_df$PAY_2)
 Factor w/ 11 levels "-2","-1","0",..: 5 5 3 3 3 3 3 2 3 1 ...
> creddt_df$PAY_3 <- as.factor(creddt_df$PAY_3)
> str(creddt_df$PAY_3)
 Factor w/ 11 levels "-2","-1","0",..: 2 3 3 3 2 3 3 2 5 1 ...
> creddt_df$PAY_4 <- as.factor(creddt_df$PAY_4)
> str(creddt_df$PAY_4)
 Factor w/ 11 levels "-2","-1","0",..: 2 3 3 3 3 3 3 3 3 1 ...
> creddt_df$PAY_5 <- as.factor(creddt_df$PAY_5)
> str(creddt_df$PAY_5)
 Factor w/ 10 levels "-2","-1","0",..: 1 3 3 3 3 3 3 3 3 2 ...
> creddt_df$PAY_6 <- as.factor(creddt_df$PAY_6)
> str(creddt_df$PAY_6)
 Factor w/ 10 levels "-2","-1","0",..: 1 4 3 3 3 3 3 2 3 2 ...
> creddt_df$default.payment.next.month <- as.factor(creddt_df$default.payment.next.month)
> str(creddt_df$default.payment.next.month)
 Factor w/ 2 levels "0","1": 2 2 1 1 1 1 1 1 1 1 ...
>
```

P2P loan default:

The features like, "home_ownership_cat", "income_cat", "term_cat", "application_type_cat", "purpose_cat", "interest_payment_cat", "loan_condition_cat" and "grade_cat" are transformed into categorical variables and it just repeats the meaning features like home_ownership, income_category, term, application type, purpose, interest payment, loan_condition and grade. All these features can be removed as it is of no use but still, we don't need to disturb the datasets to apply prediction algorithm.

```
> ###### P2P loan default ###########
> loandt_df$home_ownership_cat <- as.factor(loandt_df$home_ownership_cat)
> str(loandt_df$home_ownership_cat)
 Factor w/ 6 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 2 1 ...
> loandt_df$income_cat <- as.factor(loandt_df$income_cat)
> str(loandt_df$income_cat)
 Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
> loandt_df$term_cat <- as.factor(loandt_df$term_cat)
> str(loandt_df$term_cat)
 Factor w/ 2 levels "1","2": 1 2 1 1 2 1 2 1 2 2 ...
> loandt_df$application_type_cat <- as.factor(loandt_df$application_type_cat)
> str(loandt_df$application_type_cat)
 Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...
> loandt_df$purpose_cat <- as.factor(loandt_df$purpose_cat)
> str(loandt_df$purpose_cat)
 Factor w/ 14 levels "1","2","3","4",..: 1 2 3 4 4 5 6 2 3 4 ...
> loandt_df$interest_payment_cat <- as.factor(loandt_df$interest_payment_cat)
> str(loandt_df$interest_payment_cat)
 Factor w/ 2 levels "1","2": 1 2 2 2 1 1 2 2 2 1 ...
> loandt_df$loan_condition_cat <- as.factor(loandt_df$loan_condition_cat)
> str(loandt_df$loan_condition_cat)
 Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 2 2 ...
> loandt_df$grade_cat <- as.factor(loandt_df$grade_cat)
> str(loandt_df$grade_cat)
 Factor w/ 7 levels "1","2","3","4",..: 2 3 3 3 2 1 3 5 6 2 ...
> |
```

Numeric Data Quality Report:

Numeric data quality report defines each numeric feature of the datasets which ensures that dataset is ready to be predicted via application of the data mining algorithms. The following are the numeric data quality reports of each dataset,

Credit card fraud:

	Feature	Instances	Missing	Cardinality	Min	FirstQuartile	Median	ThirdQuartile	Max	Mean	Stdev
1	Time	284807	0	124592	0.000000000	54201.50000000000	84692.00000000000	139320.5000000000	172792.000000000	94813.8595750806562136858701706	47488.1459545662
2	V1	284807	0	275663	-56.407509631	-0.92037338439	0.018108799162	1.31564169388	2.454929991	0.0000000000000011665816574	1.9586958039
3	V2	284807	0	275663	-72.715727563	-0.59854991346	0.065485556396	0.80372387124	22.057728990	0.0000000000000003118990305	1.6513085795
4	V3	284807	0	275663	-48.325589362	-0.89036483816	0.179846343564	1.02719554247	9.382558433	-0.0000000000000013634475911	1.5162550052
5	V4	284807	0	275663	-5.683171198	-0.84864011633	-0.019846529481	0.74334128947	16.875344034	0.0000000000000021113342323	1.4158685749
6	V5	284807	0	275663	-113.743306711	-0.69159707089	-0.054335826736	0.61192643974	34.801665877	0.0000000000000009796989080	1.3802467340
7	V6	284807	0	275663	-26.160505936	-0.76829560846	-0.274187076507	0.39856489636	73.301625546	0.0000000000000015100375343	1.3322710898
8	V7	284807	0	275663	-43.557241571	-0.55407587904	0.040103082795	0.57043607288	120.589493945	-0.0000000000000005421869150	1.2370935982
9	V8	284807	0	275663	-73.216718455	-0.20862974404	0.022358036437	0.32734586192	20.007208365	0.0000000000000001026742902	1.1943529027
10	V9	284807	0	275663	-13.434066318	-0.64309757027	-0.051428731550	0.59713903028	15.594994607	-0.0000000000000024224601701	1.0986320892
11	V10	284807	0	275663	-24.588262437	-0.53542572649	-0.092917383596	0.45392344514	23.745136121	0.0000000000000022260734172	1.0888497654
12	V11	284807	0	275663	-4.797473465	-0.76249419551	-0.032757354499	0.73959340732	12.018913182	0.0000000000000017064050549	1.0207130277
13	V12	284807	0	275663	-18.683714633	-0.40557148544	0.140032588291	0.61823803295	7.848392076	-0.0000000000000012430515671	0.9992013895
14	V13	284807	0	275663	-5.791881206	-0.64853929911	-0.013568056785	0.66250495944	7.126882959	0.0000000000000008353085269	0.9952742301
15	V14	284807	0	275663	-19.214325490	-0.42557401245	0.050601319310	0.49314984922	10.526766052	0.0000000000000012286617343	0.9585956113
16	V15	284807	0	275663	-4.498944677	-0.58288427916	0.048071549763	0.64882080632	8.877741598	0.0000000000000048439271125	0.9153160116
17	V16	284807	0	275663	-14.129854517	-0.46803676671	0.066413320504	0.52329631248	17.315111518	0.0000000000000014336738623	0.8762528874
18	V17	284807	0	275663	-25.162799369	-0.48374831371	-0.065675753807	0.39967498265	9.253526250	-0.0000000000000003777280554	0.8493370637
19	V18	284807	0	275663	-9.498745921	-0.49884979867	-0.003636312355	0.50080674689	5.041069185	0.0000000000000009760659970	0.8381762095
20	V19	284807	0	275663	-7.213527430	-0.45629891874	0.003734822995	0.45894935576	5.591971427	0.0000000000000010384774122	0.8140405008
21	V20	284807	0	275663	-54.497720495	-0.21172136467	-0.062481092460	0.13304084099	39.420904248	0.0000000000000006406672131	0.7709250249
22	V21	284807	0	275663	-34.830382145	-0.22839494678	-0.029450167695	0.18637720338	27.202839157	0.0000000000000001686181118	0.7345240144
23	V22	284807	0	275663	-10.933143698	-0.54235037266	0.006781942528	0.52855363533	10.503090090	-0.0000000000000003380624490	0.7257015604
24	V23	284807	0	275663	-44.807735204	-0.16184634501	-0.011192930297	0.14764206386	22.528411690	0.0000000000000002673473153	0.6244602956
25	V24	284807	0	275663	-2.836626919	-0.35458613641	0.040976056057	0.43952660017	4.584549137	0.0000000000000044748702839	0.6056470678
26	V25	284807	0	275663	-10.295397075	-0.31714505407	0.016593501637	0.35071556287	7.519588679	0.0000000000000005105721961	0.5212780705
27	V26	284807	0	275663	-2.604550553	-0.32698392588	-0.052139108018	0.24095217371	3.517345612	0.0000000000000016846283090	0.4822270133
28	V27	284807	0	275663	-22.565679521	-0.07083952930	0.001342145979	0.09104511969	31.612198106	-0.0000000000000003673341355	0.4036324950
29	V28	284807	0	275663	-15.430083906	-0.05295979302	0.011243831565	0.07827995476	33.847807819	-0.0000000000000001230406267	0.3300832642
30	Amount	284807	0	32767	0.000000000	5.60000000000	22.00000000000	77.16500000000	25691.160000000	88.3496192509313402752013644	250.1201092402

Credit card default:

	Feature	Instances	Missing	Cardinality	Min	FirstQuartile	Median	ThirdQuartile	Max	Mean	Stdev
1	ID	30000	0	30000	1	7500.75	15000.5	22500.25	30000	15000.500000	8660.398374209
2	LIMIT_BAL	30000	0	81	10000	50000.00	140000.0	240000.00	1000000	167484.322667	129747.661567202
3	AGE	30000	0	56	21	28.00	34.0	41.00	79	35.485500	9.217904068
4	BILL_AMT1	30000	0	22723	-165580	3558.75	22381.5	67091.00	964511	51223.330900	73635.860575530
5	BILL_AMT2	30000	0	22346	-69777	2984.75	21200.0	64006.25	983931	49179.075167	71173.768782528
6	BILL_AMT3	30000	0	22026	-157264	2666.25	20088.5	60164.75	1664089	47013.154800	69349.387427037
7	BILL_AMT4	30000	0	21548	-170000	2326.75	19052.0	54506.00	891586	43262.948967	64332.856133916
8	BILL_AMT5	30000	0	21010	-81334	1763.00	18104.5	50190.50	927171	40311.400967	60797.155770265
9	BILL_AMT6	30000	0	20604	-339603	1256.00	17071.0	49198.25	961664	38871.760400	59554.107536746
10	PAY_AMT1	30000	0	7943	0	1000.00	2100.0	5006.00	873552	5663.580500	16563.280354026
11	PAY_AMT2	30000	0	7899	0	833.00	2009.0	5000.00	1684259	5921.163500	23040.870402057
12	PAY_AMT3	30000	0	7518	0	390.00	1800.0	4505.00	896040	5225.681500	17606.961469803
13	PAY_AMT4	30000	0	6937	0	296.00	1500.0	4013.25	621000	4826.076867	15666.159744032
14	PAY_AMT5	30000	0	6897	0	252.50	1500.0	4031.50	426529	4799.387633	15278.305679145
15	PAY_AMT6	30000	0	6939	0	117.75	1500.0	4000.00	528666	5215.502567	17777.465775435

P2P loan default:

	Feature	Instances	Missing	Cardinality	Min	FirstQuartile	Median	ThirdQuartile	Max	Mean	Stdev
1	id	887379	0	887379	54734.00	9206643.000	34433267.000000	54908135.00000	68617057.00000	32465133.055023842	22827341.721265946
2	year	887379	0	9	2007.00	2013.000	2014.000000	2015.00000	2015.00000	2014.021760713	1.261741262
3	final_d	887379	0	98	1012008.00	1012016.000	1012016.000000	1092015.00000	1122015.00000	1047089.324393523	45551.493573197
4	emp_length_int	887379	0	12	0.50	3.000	6.050000	10.00000	10.00000	6.050564359	3.507404696
5	annual_inc	887379	0	45784	0.00	45000.000	65000.000000	90000.00000	9500000.00000	75027.587860429	64698.154280715
6	loan_amount	887379	0	1372	500.00	8000.000	13000.000000	20000.00000	35000.00000	14755.264605090	8435.455601278
7	interest_rate	887379	0	542	5.32	9.990	12.990000	16.20000	28.99000	13.246739679	4.381867415
8	dti	887379	0	4086	0.00	11.910	17.650000	23.95000	9999.00000	18.157038740	17.190625688
9	total_pymnt	887379	0	505628	0.00	1914.590	4894.999117	10616.81423	57777.57987	7558.826683644	7871.243335697
10	total_rec_prncp	887379	0	260227	0.00	1200.570	3215.320000	8000.00000	35000.03000	5757.706423625	6625.441045703
11	recoveries	887379	0	23055	0.00	0.000	0.000000	0.00000	33520.27000	45.919243371	409.693873633
12	installment	887379	0	68711	15.67	260.705	382.550000	572.60000	1445.46000	436.717127360	244.186593476

Categorical Data Quality report:

Categorical data quality report defines each categorical feature of the datasets which ensures that dataset is ready to be predicted via application of the data mining algorithms. The following are the categorical data quality reports of each dataset,

Credit card fraud: Credit card fraud dataset consists of only one categorical feature so we could explore it via summary function in R, no need to draft the data quality report.

Credit card default:

	Feature	Inst	Miss	Card	FstMod	FstModFrq	Feature.1	FstModPnt	SndMod	SndModFrq	SndModPnt
1	SEX	30000	0	2	2	18112	SEX	60.37333333	1	11888	39.62666667
2	EDUCATION	30000	0	7	2	14030	EDUCATION	46.76666667	1	10585	35.28333333
3	MARRIAGE	30000	0	4	2	15964	MARRIAGE	53.21333333	1	13659	45.53000000
4	PAY_0	30000	0	11	0	14737	PAY_0	49.12333333	-1	5686	18.95333333
5	PAY_2	30000	0	11	0	15730	PAY_2	52.43333333	-1	6050	20.16666667
6	PAY_3	30000	0	11	0	15764	PAY_3	52.54666667	-1	5938	19.79333333
7	PAY_4	30000	0	11	0	16455	PAY_4	54.85000000	-1	5687	18.95666667
8	PAY_5	30000	0	10	0	16947	PAY_5	56.49000000	-1	5539	18.46333333
9	PAY_6	30000	0	10	0	16286	PAY_6	54.28666667	-1	5740	19.13333333
10	default.payment.next.month	30000	0	2	0	23364	default.payment.next.month	77.88000000	1	6636	22.12000000

P2P loan default:

	Feature	Inst	Miss	Card	FstMod	FstModFrq	Feature.1	FstModPnt	SndMod	SndModFrq	SndModPnt
1	issue_d	887379	0	103	01/10/2015	48631	issue_d	5.480296469	01/07/2015	45962	5.17952306737
2	home_ownership	887379	0	6	MORTGAGE	443557	home_ownership	49.985068387	RENT	356117	40.13133058141
3	home_ownership_cat	887379	0	6	3	443557	home_ownership_cat	49.985068387	1	356117	40.13133058141
4	income_category	887379	0	3	Low	729616	income_category	82.221463433	Medium	140977	15.88689838277
5	income_cat	887379	0	3	1	729616	income_cat	82.221463433	2	140977	15.88689838277
6	term	887379	0	2	36 months	621125	term	69.995458536	60 months	266254	30.00454146424
7	term_cat	887379	0	2	1	621125	term_cat	69.995458536	2	266254	30.00454146424
8	application_type	887379	0	2	INDIVIDUAL	886868	application_type	99.942414684	JOINT	511	0.05758531586
9	application_type_cat	887379	0	2	1	886868	application_type_cat	99.942414684	2	511	0.05758531586
10	purpose	887379	0	14	debt_consolidation	524215	purpose	59.074532979	credit_card	206182	23.23494245413
11	purpose_cat	887379	0	14	6	524215	purpose_cat	59.074532979	1	206182	23.23494245413
12	interest_payments	887379	0	2	Low	465316	interest_payments	52.437121005	High	422063	47.56287899533
13	interest_payment_cat	887379	0	2	1	465316	interest_payment_cat	52.437121005	2	422063	47.56287899533
14	loan_condition	887379	0	2	Good Loan	819950	loan_condition	92.401330210	Bad Loan	67429	7.59866979047
15	loan_condition_cat	887379	0	2	0	819950	loan_condition_cat	92.401330210	1	67429	7.59866979047
16	grade	887379	0	7	B	254535	grade	28.683910708	C	245860	27.70631263530
17	grade_cat	887379	0	7	2	254535	grade_cat	28.683910708	3	245860	27.70631263530
18	region	887379	0	5	leinster	214646	region	24.188762637	ulster	208731	23.52219288489

Step04 – Application of data mining algorithm on datasets:

Sampling of datasets: Each dataset set has to be split into test and training datasets, where as in training dataset data mining methods will be applied and predicted against the corresponding dataset.

In three datasets we have, from first dataset and third dataset (credit fraud and P2P loan default) we have taken only 20000 rows for the prediction purpose, as these are being big data it takes too long to process so we have taken limited number of rows to apply data mining techniques. From these 20000 rows, 70% of data are assigned to the training dataset to apply predictive methods and 30% to the testing dataset. Whereas we take entire second dataset (Credit default) into an account and the 70:30 partition has been made to this dataset also. Take a look at the snippet of code,

```
################ Sampling the Datasets ########################
########## Credit card fraud ############
DA_df1$Class <- as.factor(DA_df1$Class)
DA_df1 <- sample_n(DA_df1,20000) #Due to long run time we are considering only 20000 rows
index1 <- createDataPartition(DA_df1$Class, p=.70, list=FALSE, times=1)
train1 <- DA_df1[index1, ]
test1 <- DA_df1[-index1, ]
str

########## Credit card Default ############
DA_df2$default.payment.next.month <- as.factor(DA_df2$default.payment.next.month)
index2 <- createDataPartition(DA_df2$default.payment.next.month, p=.70, list=FALSE, times=1)
train2 <- DA_df2[index2, ]
test2 <- DA_df2[-index2, ]

########## P2P lending loan default ############
DA_df3$loan_condition_cat <- as.factor(DA_df3$loan_condition_cat)
DA_df3 <- sample_n(DA_df3,20000) #Due to long run time we are considering only 20000 rows
index3 <- createDataPartition(DA_df3$loan_condition_cat, p=.70, list=FALSE, times=1)
train3 <- DA_df3[index3, ]
test3 <- DA_df3[-index3, ]
```

IV. Evaluation

Applying methods to datasets: In this project, four data mining algorithms were chosen for the prediction purpose and they are follows,

1. Logistic Regression'
2. K – Nearest Neighbor algorithm
3. Random Forest algorithm
4. Support Vector Machine algorithm

Let's discuss about the advantages and disadvantages of the above-mentioned methods,

Methods	Advantages	Disadvantages
Logistic Regression	1. It performs well with linear data and also it is esay to be applied. It generates an easy predictive formula for classification. 2.It does well if that response variable to be categorical. Predictive variable can in any type while using this model.	1. It can be applied to only linear data and this model is not suitable for the non-linear data due to its linear decision surface. 2. It cannot predict continuous events. 3. There is a chance for overfitting when more variables were added to the dataset.
KNN	1. KNN is the algorithm which is non parmetric that makes assumption with the proximity. 2. The predictive power of this algorithm is high. 3. The processing time of this algortihm is comparitively less. 4. This method is suitable for the dataset which has no pior knowledge.	1. KNN requires high level of resources to process so this method is considered as the expensive method. 2. This method will not perform well with categorical datasets. 3. This is sensitive to outliers and missing values.
Random forest	1. Random forest algorithm performs well even with non transformed data which has outliers and missing values because random forest will take only relevant information for prediction. 2. Random forest can be applied to either linear dataset or non-linear dataset. 3. Random forest method comparitively overcome the overfitting in the better way. 4. It ranks the important variables in the excellent allow analysts to understand the importance of the variables in the dataset.	1. Manually assigning number of tree will be challenging and time consuming. 2. Prediction will be comparitiively slow. 3. Noisy classified dataset can't be handled properly by random forest, it may end up in overfitting. 4. Interpretation will be challenging in random forest method.
SVM	1. SVM can take up dataset without prior knowledge for prediction. 2. SVM can even performs on the unstructured and semi-structured data. 3. Overfitting is comparitively low in SVM. 4. Kernel function helps SVM to solve more complex problems.	1. Interpretation of the model is difficult. 2. Suitable kernerl version cannot be identified easily. 3. The processing time of this method will be slower when the target dataset is too large. 4. Since the discovery of pattern is being difficult, apply the business idea on the discovered pattern will be challenging.

Figure: Advantages and disadvantages of data mining methods

Logistic Regression:

Credit Card Fraud dataset (Dataset 01)	Credit Card default (Dataset 02)	Peer to Peer Loan default (Dataset 03)
Confusion Matrix and Statistics	Confusion Matrix and Statistics	Confusion Matrix and Statistics
Reference Prediction 0 1 0 5989 1 1 4 5	Reference Prediction 0 1 0 6810 199 1 1467 523	Reference Prediction 0 1 0 2 5531 1 186 280
Accuracy : 0.9991665 95% CI : (0.998056, 0.9997293) No Information Rate : 0.9989998 P-Value [Acc > NIR] : 0.4455993	Accuracy : 0.8148683 95% CI : (0.8066865, 0.8228458) No Information Rate : 0.9197689 P-Value [Acc > NIR] : 1	Accuracy : 0.0470078 95% CI : (0.04179, 0.0526717) No Information Rate : 0.9686614 P-Value [Acc > NIR] : 1
Kappa : 0.6662661 Mcnemar's Test P-Value : 0.3710934	Kappa : 0.3037097 Mcnemar's Test P-Value : <0.0000000000000002	Kappa : -0.0637847 Mcnemar's Test P-Value : <0.0000000000000002
Sensitivity : 0.8333333333 Specificity : 0.9993325546 Pos Pred Value : 0.5555555556 Neg Pred Value : 0.9998330551 Prevalence : 0.0010001667 Detection Rate : 0.0008334722 Detection Prevalence : 0.0015002500 Balanced Accuracy : 0.9163329440	Sensitivity : 0.72437673 Specificity : 0.82276187 Pos Pred Value : 0.26281407 Neg Pred Value : 0.97160793 Prevalence : 0.08023114 Detection Rate : 0.05811757 Detection Prevalence : 0.22113568 Balanced Accuracy : 0.77356930	Sensitivity : 0.0481844777 Specificity : 0.0106382979 Pos Pred Value : 0.6008583691 Neg Pred Value : 0.0003614676 Prevalence : 0.9686614436 Detection Rate : 0.0466744457 Detection Prevalence : 0.0776796133 Balanced Accuracy : 0.0294113878
'Positive' Class : 1	'Positive' Class : 1	'Positive' Class : 1

The above-mentioned table consists of confusion matrices of the logistic regression applied to three datasets. Let's discuss in detail about the performance of logistic regression on each dataset,

Credit card Fraud (Dataset 01): We could see logistic regression performed very well on the first dataset as the accuracy is 99.9%. The method managed to predict 5994 values correctly where as only 5 values were predicted incorrectly. Let us take the sensitivity and specificity into consideration which are comparatively most import method evaluation factors. Where, sensitivity is the true positive

which is nothing but the how accurately the method is predicted the positive (here, 1) and specificity is nothing but the measurement of how accurately the method predicted negativity. We could see for the dataset 01 the logistic regression is predicted with the 83.3% of sensitivity and 99.9% of specificity. With the obtained sensitivity and specificity, we could discover the pattern which can be further used to take a major business decision such as following, in this credit card default dataset we have obtained 83.3% sensitivity, which in turn to be predicted that the 83.3% of fraudulent transactions successfully so with this predicted pattern we could able to identify the similar fraudulent transaction and block it. And with the specificity of 99.9% percent we could identify authenticated honest customers to whom we could increase the credit limit for the either way benefits.

Credit Card Default (Dataset 02):
From the above table we could see logistic regression predicted on dataset 02 with the accuracy of about 81.4%, sensitivity and specificity of 72.4% and 82.2% respectively. With these obtained factors we could say that LG performed well on dataset 02 but not better than dataset 01. And the same business applications can be applied as above mentioned, with the discovered pattern we could able to predicts which customer could be credible or not.

Peer to Peer Loan default (Dataset 03):
With the obtained confusion matrices, it is proved that LG doesn't performed well with accuracy of 4.7% of accuracy, 4.8% and 1.06% of sensitivity and specificity respectively on the dataset 03 so whatever the pattern obtained from this model will not be useful to make any business-oriented decision due to low accuracy.

K-NEAREST NEIGHBOR ALGORITHM:

KNN Model

Credit Card Fraud dataset (Dataset 01)	Credit Card default (Dataset 02)	Peer to Peer Loan default (Dataset 03)
Confusion Matrix and Statistics	Confusion Matrix and Statistics	Confusion Matrix and Statistics
Reference	Reference	Reference
Prediction 0 1	Prediction 0 1	Prediction 0 1
0 5990 9	0 6428 1636	0 5371 401
1 0 0	1 581 354	1 162 65
Accuracy : 0.9984997	Accuracy : 0.7536393	Accuracy : 0.906151
95% CI : (0.997154, 0.9993138)	95% CI : (0.7445999, 0.7625143)	95% CI : (0.8984894, 0.913415)
No Information Rate : 0.9984997	No Information Rate : 0.7788643	No Information Rate : 0.9223204
P-Value [Acc > NIR] : 0.587408269	P-Value [Acc > NIR] : 1	P-Value [Acc > NIR] : 0.9999976
Kappa : 0	Kappa : 0.1172523	Kappa : 0.1440302
Mcnemar's Test P-Value : 0.007660761	Mcnemar's Test P-Value : <0.0000000000000002	Mcnemar's Test P-Value : < 0.00000000000000022
Sensitivity : 0.00000000	Sensitivity : 0.1778894	Sensitivity : 0.13948498
Specificity : 1.00000000	Specificity : 0.9171066	Specificity : 0.97072113
Pos Pred Value : NaN	Pos Pred Value : 0.3786096	Pos Pred Value : 0.28634361
Neg Pred Value : 0.99849975	Neg Pred Value : 0.7971230	Neg Pred Value : 0.93052668
Prevalence : 0.00150025	Prevalence : 0.2211357	Prevalence : 0.07767961
Detection Rate : 0.00000000	Detection Rate : 0.0393377	Detection Rate : 0.01083514
Detection Prevalence : 0.00000000	Detection Prevalence : 0.1039004	Detection Prevalence : 0.03783964
Balanced Accuracy : 0.50000000	Balanced Accuracy : 0.5474980	Balanced Accuracy : 0.55510305
'Positive' Class : 1	'Positive' Class : 1	'Positive' Class : 1

The above given table consists of confusion matrices of the KNN model for the three datasets with us.

Credit card Fraud (Dataset 01):
The confusion matrix of KNN on first dataset exposes that KNN performed neither good nor bad on the first dataset with accuracy of 99.8% but returned 0 for sensitivity so with the pattern predicted with this model we couldn't find any of the fraudulent transaction which ensure that this model failed on this dataset based on the balanced accuracy which is comparatively lowest.

Credit Card Default (Dataset 02):
From the confusion matrix obtained for the dataset 02 we could justify that KNN model performed better on dataset 02 than dataset 01. It predicted with the accuracy of 75.3%, 17.7% and 91.7 of sensitivity and specificity respectively. We could say KNN performed better on the dataset but we couldn't say that it performed well on the datasets.

Peer to Peer Loan default (Dataset 03):
The confusion matrix of KNN on dataset 03 clearly depicts that KNN performed better in dataset 03 than in both datasets. The accuracy level is high with 90% above but still lags in the sensitivity and specificity like all other datasets so we can ensure that KNN is not a very good model for the datasets we have.

RANDOM FOREST:

Random Forest Model

Credit Card Fraud dataset (Dataset 01)	Credit Card default (Dataset 02)	Peer to Peer Loan default (Dataset 03)
Confusion Matrix and Statistics	Confusion Matrix and Statistics	Confusion Matrix and Statistics
Reference	Reference	Reference
Prediction 0 1	Prediction 0 1	Prediction 0 1
0 5988 2	0 6642 367	0 5565 0
1 4 5	1 1228 762	1 0 435
Accuracy : 0.9989998	Accuracy : 0.8227581	Accuracy : 1
95% CI : (0.9978243, 0.9996329)	95% CI : (0.8147086, 0.8305981)	95% CI : (0.9993854, 1)
No Information Rate : 0.9988331	No Information Rate : 0.8745416	No Information Rate : 0.9275
P-Value [Acc > NIR] : 0.4496241	P-Value [Acc > NIR] : 1	P-Value [Acc > NIR] : < 0.00000000000000022204
Kappa : 0.6245071	Kappa : 0.391146	Kappa : 1
Mcnemar's Test P-Value : 0.6830914	Mcnemar's Test P-Value : <0.0000000000000002	Mcnemar's Test P-Value : NA
Sensitivity : 0.7142857143	Sensitivity : 0.67493357	Sensitivity : 1.0000
Specificity : 0.9993324433	Specificity : 0.84396442	Specificity : 1.0000
Pos Pred Value : 0.5555555556	Pos Pred Value : 0.38291457	Pos Pred Value : 1.0000
Neg Pred Value : 0.9996661102	Neg Pred Value : 0.94763875	Neg Pred Value : 1.0000
Prevalence : 0.0011668611	Prevalence : 0.12545838	Prevalence : 0.0725
Detection Rate : 0.0008334722	Detection Rate : 0.08467608	Detection Rate : 0.0725
Detection Prevalence : 0.0015002500	Detection Prevalence : 0.22113568	Detection Prevalence : 0.0725
Balanced Accuracy : 0.8568090788	Balanced Accuracy : 0.75944900	Balanced Accuracy : 1.0000
'Positive' Class : 1	'Positive' Class : 1	'Positive' Class : 1

The above given table consists of confusion matrices of the Random Forest model for the holding three datasets.

Credit card Fraud (Dataset 01):
While looking into the confusion matrix of random forest on dataset 01 and we could say it performed very well on the dataset and so the pattern obtained from this model could have a capability to make an efficient business decision which could reduce the fraudulent transactions and encourage the honest customers.

Credit Card Default (Dataset 02):
The confusion matrix of random forest on the dataset 02 depicts that random forest model performed better on dataset 02 than other two models with higher accuracy and balanced accuracy of 82.2% and 75.9% respectively so far.

Peer to Peer Loan default (Dataset 03):
The performance of the random forest model is the best on the dataset 03 with 100% accuracy, sensitivity, specificity and balanced accuracy. This is achieved because of the random forest characteristics which is not sensitive to the outliers and missing values and it process only with the relevant data.

SUPPORT VECTOR MACHINE ALGORITHM:

SVM Algorithm

Credit Card Fraud dataset (Dataset 01)	Credit Card default (Dataset 02)	Peer to Peer Loan default (Dataset 03)
Confusion Matrix and Statistics Reference Prediction 0 1 0 5988 7 1 2 2 Accuracy : 0.9984997 95% CI : (0.997154, 0.9993138) No Information Rate : 0.9984997 P-Value [Acc > NIR] : 0.5874083 Kappa : 0.3070526 Mcnemar's Test P-Value : 0.1824224 Sensitivity : 0.9996661 Specificity : 0.2222222 Pos Pred Value : 0.9988324 Neg Pred Value : 0.5000000 Prevalence : 0.9984997 Detection Rate : 0.9981664 Detection Prevalence : 0.9993332 Balanced Accuracy : 0.6109442 'Positive' Class : 0	Confusion Matrix and Statistics Reference Prediction 0 1 0 6721 1333 1 288 657 Accuracy : 0.8186689 95% CI : (0.8117702, 0.8277599) No Information Rate : 0.7788643 P-Value [Acc > NIR] : < 0.0000000000000022204 Kappa : 0.355993 Mcnemar's Test P-Value : < 0.0000000000000022204 Sensitivity : 0.9589100 Specificity : 0.3301508 Pos Pred Value : 0.8344922 Neg Pred Value : 0.6952381 Prevalence : 0.7788643 Detection Rate : 0.7468608 Detection Prevalence : 0.8949883 Balanced Accuracy : 0.6445304 'Positive' Class : 0	Confusion Matrix and Statistics Reference Prediction 0 1 0 5565 1 1 0 434 Accuracy : 0.9998333 95% CI : (0.9990717, 0.9999958) No Information Rate : 0.9275 P-Value [Acc > NIR] : <0.0000000000000002 Kappa : 0.9987594 Mcnemar's Test P-Value : 1 Sensitivity : 1.0000000 Specificity : 0.9977011 Pos Pred Value : 0.9998203 Neg Pred Value : 1.0000000 Prevalence : 0.9275000 Detection Rate : 0.9275000 Detection Prevalence : 0.9276667 Balanced Accuracy : 0.9988506 'Positive' Class : 0

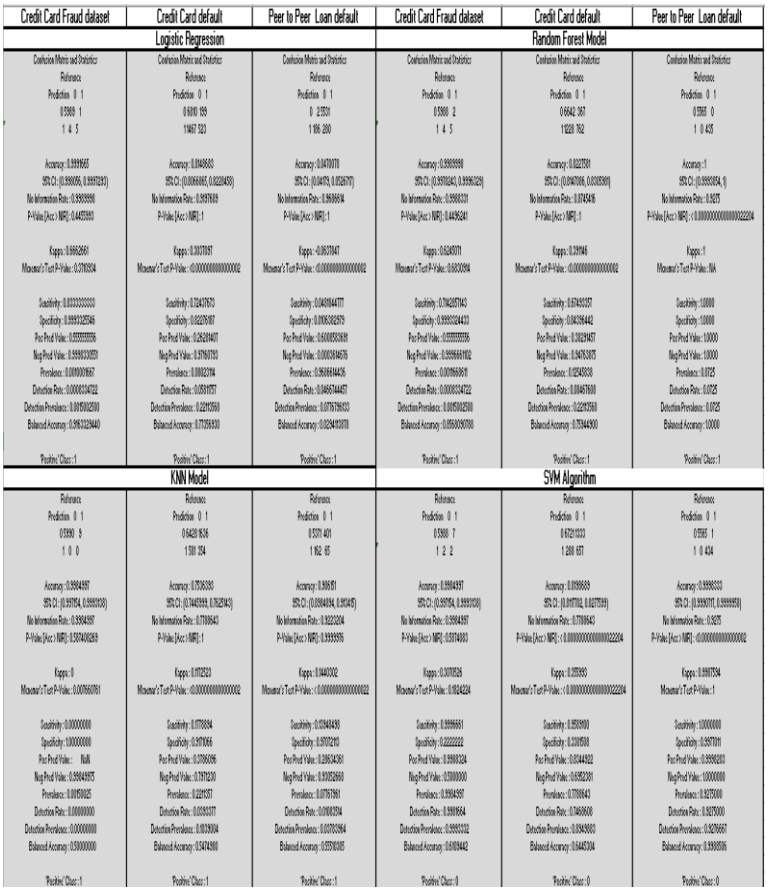
The above given table consists of confusion matrices of the SVM model for the three datasets with us.

Credit card Fraud (Dataset 01):
With the help of confusion matrix, we could so that SMV performed better than KNN on the dataset 01 but not fitting very well on the dataset as its balanced accuracy is just 61% even though its accuracy is 99.8%.

Credit Card Default (Dataset 02):
SVM outperforms the both KNN and Logistic regression model for the dataset 02 with the second highest accuracy of 81.9%. So that the pattern obtained from this model can be considered as the efficient enough to make any business decisions.

Peer to Peer Loan default (Dataset 03):
SVM again outperformed the both KNN and logistic regression model for the dataset 03 with again second highest accuracy of 99.98%, sensitivity of 100%, specificity of 99.7% and balanced accuracy of 99.88%. So, it also provides a worthy pattern which can be used to make a decision on the risk management.

Consolidated table and comparison:



By comparing all the applied methods with the help of above given consolidated confusion matrix let us choose the best and worst method for each of the datasets.

Credit card Fraud (Dataset 01):
For the dataset 01 we could say **logistic regression is the best model**. As per one of the advantages of the logistic regression, "it performs well when the response variable is the categorical variable and predictive variable can be any of type" the response variable of dataset 01 is the categorical variable. And, **KNN is the worst model** for the dataset 01, as KNN model will not perform well when the response variable is the categorical variable.

Credit Card Default (Dataset 02):
For the dataset 02 both logistic regression and random forest performed well but comparatively **random forest is the best** model due to higher accuracy. This is because random forest is not sensitive to the outliers and missing values and also this model will rank the important variable as well. And, **KNN is the worst model** for the dataset 01, as KNN model will not perform well when the response variable is the categorical variable.

Peer to Peer Loan default (Dataset 03):
No doubt, **random forest model outperformed all other model** with 100% accuracy on the dataset 03. **Logistic regression performed poorly** on the dataset 03 because it cannot predict continuous outcomes.

V. CONCLUSION

To conclude, though there are multiple and high-risk challenges of fraud and delinquencies in today's era compared to the earlier era for banking and financial institutions, timely decisions to prevent the customer likely to default augments profitability and loyal customers for the ban, which will ultimately increase the share-holders value. In this report, it has been proposed to identify and work on three datasets from reliable repositories and have applied four machine language algorithms using R function along with understanding of the datasets, pre-processing of datasets, infer the correlation and conducting the data quality report for numeric and categorical variable and assigning the binary operators to target variables for prediction purposes on all three3 datasets.
Further, this report also discussed about the comparison of the empirical result reflects in confusion matrix using the notions of performance namely the accuracy, sensitivity and specificity of four methods and 3 data sets. It has been identified that the "Random Forest" algorithm predicts better accuracy in all 3 methods. Also noted that in dataset 1 & 2 the Logistic regression is failed due to less accuracy predication and for dataset 3 the oversampling method is failed.

The purpose of this report is to provide the sense of behavioral scoring of delinquent borrowers to the banks and FI and they can use as reference for real world applications. However, there are limitations to this study where, this report included the raw data but not considered if customer have multiple debts, if any debt etc.

Additionally, as an extension to this study, it has also been thought that during the course of this study, the experiences learnt can be explored and applied to replicate the same in other industries like Healthcare, Stock markets, Insurances and other financial and non-financial domains.

VI. REFERENCES

[1] aK. Gai, M. Qiu, and X. Sun, "A survey on FinTech," Journal of Network and Computer Applications, vol. 103, pp. 262–273, Feb. 2018.

[2] D. J. Hand, "Principles of Data Mining:," Drug Safety, vol. 30, no. 7, pp. 621–622, 2007.

[3] S. Agarwal, "Data Mining: Data Mining Concepts and Techniques," in 2013 International Conference on Machine Intelligence and Research Advancement, 2013, pp. 203–207

[4] M. Kantardzic, Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons, 2011.

[5] Brachman, R. J. & Anand, T., "The process of knowledge discovery in databases.", AAAI Press / The MIT Press. 1996.

[6] U. Shafique and H. Qaiser, "A comparative study of data mining process models (KDD, CRISP-DM and SEMMA)," International Journal of Innovation and Scientific Research, vol. 12, no. 1, pp. 217–222, 2014.

[7] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," 1, vol. 17, no. 3, pp. 37–37, Mar. 1996.

[8] Brachman, R. J. & Anand, T., "The process of knowledge discovery in databases.", AAAI Press / The MIT Press. 1996.

[9] SAS Enterprise Miner – SEMMA. SAS Institute, 2014 [online] available: http://www.sas.com/technologies/analytics/datamining/miner/semma.html (September 2014.)

[10] J. D. Kelleher, B. M. Namee, and A. D'Arcy, "FUNDAMENTALS OF MACHINE LEARNING FOR PREDICTIVE DATA ANALYTICS," p. 31.

[11] P. Chapman, "CRISP-DM 1.0: Step-by-Step Data Mining Guide," 2000.

[12] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data mining for credit card fraud: A comparative study," Decision Support Systems, vol. 50, no. 3, pp. 602–613, Feb. 2011.

[13] L. Ying, "Research on bank credit default prediction based on data mining algorithm," 1, vol. 5, no. 6, pp. 4820–4823, Jun. 2018.

[14] Yufeng Kou, Chang-Tien Lu, Sirirat Sirwongwattana, Survey of Fraud Detection Techniques, Proc. IEEE : International Conference on Networking , Sensing & Control, pp. 749-754, 2004.

[15] A. D. Pozzolo, "Adaptive Machine Learning for Credit Card Fraud Detection," p. 199.

[16] I.-C. Yeh and C. Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," Expert Systems with Applications, vol. 36, no. 2, pp. 2473–2480, Mar. 2009.

[17] D. A. Adeniyi, Z. Wei, and Y. Yongquan, "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method," Applied Computing and Informatics, vol. 12, no. 1, pp. 90–108, Jan. 2016.

[18] J. Kruppa, A. Schwarz, G. Arminger, and A. Ziegler, "Consumer credit risk: Individual probability estimates using machine learning," Expert Systems with Applications, vol. 40, no. 13, pp. 5125–5131, 2013.

[19] X. Ye, L. Dong, and D. Ma, "Loan evaluation in P2P lending based on Random Forest optimized by genetic algorithm with profit score," Electronic Commerce Research and Applications, vol. 32, pp. 23–36, Nov. 2018.

[20] V. Mareeswari and G. Gunasekaran, "Prevention of credit card fraud detection based on HSVM," in 2016 International Conference on Information Communication and Embedded Systems (ICICES), 2016, pp. 1–4.

[21]C.-L. Huang, M.-C. Chen, and C.-J. Wang, "Credit scoring with a data mining approach based on support vector machines," Expert Systems with Applications, vol. 33, no. 4, pp. 847–856, Nov. 2007.

[22] Tao Guo and Gui-Yang Li, "Neural data mining for credit card fraud detection," in 2008 International Conference on Machine Learning and Cybernetics, Kunming, China, 2008, pp. 3630–3634.

[23] Pingfan Song, , Yunzhi Chen, , Zhixiang Zhou, & Huaqing Wu. (2018). Performance Analysis of Peer-to-Peer Online Lending Platforms in China. Sustainability -MDPI.

[24] S. K. S. P K Viswanathan, "Journal of Emerging Market Finance," Modelling Credit Default in Microfinance -An Indian case study , p. 3, 2017.

[25]Van Shi,Xin Fan and Wenjie Li, & lianfeng Liu and Hongbo Ma. (2016). Study on Key Technology of Power Users Credit Rating Evaluation Based on Big Data. IEEE.

[26]S. H. U. ,. S. P. H. D H Pandya, "Expert Systems with Applications," Fault Diagnosis of rolling element bearing with intrinsic mode function of acoustic emission data using APF - KNN, 2013.

[27]Yun-Huan Lee a , & Ya-Li Huang b. (2011). The expansion of the credit card market in Taiwan. Applied Economics Letters, 3.

[28]Yun-Huan Lee a , & Ya-Li Huang b. (2011). The expansion of the credit card market in Taiwan. Applied Economics Letters, 3.

[29]Yu Jina, & Yudan Zhua. (2015). A data-driven approach to predict default risk of loan for online Peer-to-Peer(P2P) lending. IEEE, 3.

[30]Zhang, Y. (2017). Influencing Factors of Online P2P Lending Sucess Rate in China. ScienceDirect -Procedia Computer Science 122 (2017) 896–901

[31]Milad Malekipirbazari, & Vural Aksakalli. (2015). Risk assessment in social lending via random forests. Expert Systems with Applications.