Data Mining Report

Lokesh Raj Duraisamy Nagarajan
Msc Fintech
National College of Ireland
Dublin,Ireland
x18142231@ncirl.ie

Abstract— The objective of this study is to examine all aspects of credit card defaulters by using data mining techniques which compares the predictive accuracy of the probability of defaulters among four data mining techniques (Logistic Regression, Oversampling, Random Forest, and Support vector machines). The findings of this review clearly show that data mining techniques have been applied most extensively to the detection of credit card defaulters, fraud and Loss assets, From the perspective of risk management. Peer loan lending is a financial process between individuals. Defaulters in peer to peer lending are damaging the credibility and development of the P2P lending platforms. Loan evaluation plays a major role to predict and reduce the loan defaults. The main goal of credit risk evaluation is to make practical decisions that guide lenders, which helps to gain more profits. The motivation is to anticipate the default on credit card, bad loan peer lending and credit card frauds that are anticipated to occur soon. Due to the sudden growth in the financial sector, banks are facing difficulties in knowing the customer behaviors. the result of predictive accuracy is to find the probability of defaulters. Therefore, it has been proposed to apply these 4 methods to evaluate the default occurrence accuracy and to minimize the risk of 'bad loans' by use of advanced machine learning algorithms. the motivation of this report is to gather financial data from multiple data sources and use data mining algorithms on this data to extract significant information and predict if a customer would be able to repay his loan or not. In other words, predict if the customer would be a defaulter or not.

Keywords— Logistic Regression, Oversampling, Random Forest, and Support vector machines, Data Mining, Financial Fraud.

I. INTRODUCTION

In recent years, Financial industries are facing major risk such as default Fraud activities which has become frequent and has become a major problem worldwide for financial industries, taking loans from financial institutions has turned into an extremely regular occasion.

Consistently countless make application for credits, for a variety of purposes. In any case, all these candidates are not solid, and everybody can't be approved. The decision making of providing a loan to the individual has become critical. So, the motivation of this project is to gather loan data from multiple data sources and implement data mining techniques to retrieve the essential data and predict if the customer would have the capacity to reimburse or repay his loan or not, the result is to foresee the customer.

The process involves the relationship between lenders and borrowers for transactions purposes without depending on any bank as a mediator major advantages of peer to peer lending is a loan in short periods and flexible trading.
Development of P2P Lending leads to the creation of scorecard and implement machine learning methods to design loan evaluation model these models help in predicting defaulters. The loan evaluation method consists of attributes such as profit score and actual profit where the actual profit considers the revenues and losses as non-defaulters. In the case of the lender when a non-defaulter is classified as a defaulter, the lender will face a potential loss and couldn't get any returns from the loans. So, from the loan evaluation technique if the defaulter is predicted it can help lender avoid losing principal.

Consequently, the latest profit score can help in examining the effects of loan evaluation model. Generally, credit card fraud is classified into online and offline fraud, fraud detection involves detecting the fraud as soon as possible these fraud detection methods are implemented to secure from delinquencies. These fraud occurrences are detected from the Anomalies in data and patterns [21].

Evaluating the risk, which is associated with a loan application, is a major standout issue amongst the most essential issues of the banks for survival in the focused market and for Profitability. Banks get a large number of loan applications from their clients and other individuals on regular basis. Effective Risk management appears to be essential for the long-term survival of banks and the stability of the whole financial system[16].

The greater part of the banks utilizes their very own credit scoring and risk assessment techniques so as to examine and evaluate to settle the decision on credit approval.

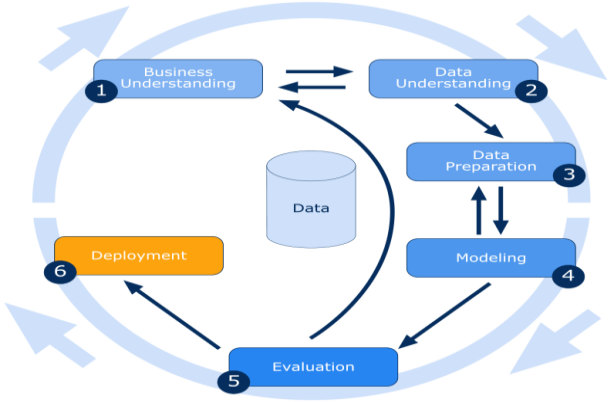
There are numerous cases happening each year, where individuals repay the loan amounts, or they become defaulters, because of this financial institution they suffer a huge amount of losses. Thus, these data mining methods will be used to examine the loan data and retrieve patterns which would help to identify the likely defaulters, which would help the financial institutions to settle on better decision making in future.

The significant motivation behind risk prediction is to utilize business budget summary, client transactions, and loan repayment records, etc. to predict customers credit risk and to reduce the damage and uncertainty.

In the next section, An introduction to data mining methodologies.

Data Mining:

The process of Data mining is about finding insights which extracting the information from large data sets to understand and infer the differences, designs, and correlations for predicting the outcomes. CRISP-DM methodology is used to perform the data mining process. CRISP-DM stands for the cross-industry process. There are 6 stages generally starting from (i)Determining the business objectives and plan of the project. (ii)Data understanding-describing and exploring the data and verifying its quality.(iii)Data preparation-Select your data which column and rows you need to analyze and integrate the data.(iv)Modelling-select the modelling technique which need to be implemented on the data , build it and asses the results of the models in this stage.(v)Evaluation-Evaluate the assessment results from the previous stage where the model is applied, then review the results and summarize the steps.(vi)Deployment-deployment off the application monitoring and maintained of the model .



There are four different types of machine learning algorithms used for the models to handle the data, termed as unsupervised learning, supervised learning, semi-supervised learning, and reinforcement learning

Data mining plays a major role in the Financial industry, it helps in understanding the detect any kind of hidden truths involved in large quantities of data. Data mining as a process FFD, as it is often applied to extract and uncover the hidden truths behind very large quantities of data. Data mining as a method is used for detecting unusual interesting patterns which are present in the database which helps the developers make critical decisions. Generally, a process that uses artificial intelligence and machine learning methods to extract and classify useful information and advance the facts gained from the large database [2].

Due to the sudden growth in the banking industry in recent years which primarily deals with the analysis and representation and transformation of data into useful meaning which is considered to be a task beyond human capability, in recent times various data mining techniques are Implemented in resolving business issues, associations and correlations, which are hidden in the business information stored in the database. These data mining techniques help in analyzing patterns, trends and predict how the customers will respond to on interest rates, which customers are more likely to accept recent offers and also analyzing which customers are likely to be in defaulters and how to engage customer relationship more gainful insights about their behavioral actions. Since they can prevent frauds [8].

The next section reviews the literature on Default credit card P2P lending and online and Credit card fraud detection.

II. RELATED WORKS

Dataset1: Default credit card (Taiwan)

This dataset conveys the goal of interpreting the information of defaulters who were compared and evaluated in various parameters such as Gender, Limit balance, education, marital status, and age. This data including the history of past payment gathered for the period of April to Sep. It contains 25 columns and 30000 rows and No missing values, default payment (Yes = 1, No = 0), as the response variable and (22 Numeric and 3 Categorical variables) Challenges in regard to Dataset -1, are that the columns of Sex, Education and Marital statuses are indicated in numerical wherein these values should be updated to categorical data.

From the viewpoint of risk control, evaluating the likelihood of default will be more important than classifying or segregating the customers in risky or non-risky. Regardless of whether the estimated probability of default produced from data mining methods can represent the genuine probability of default occurrence.

In an all-around created money related framework, emergency the board is on the downstream and hazard expectation is on the upstream. The real reason for hazard expectation is to utilize monetary data, for example, business budget summary, client exchange and reimbursement records, and so on., to foresee business execution or individual clients' credit chance and to decrease the harm and vulnerability.

As (FEI LI', 2004) [5], states that the indispensable assignments of a bank are to refresh the upgrade the procedure associated with the appraisal of planned to keep the peril of a credit misfortune, furthermore, to reduce the failure expenses over risk groups. requests are to fill a form by the candidates which will be reviewed by the bank thereby assign a reasonable score as indicated by a predefined scoring table. The requests is to fill a shape by the candidates which will be evaluated by the bank consequently allocate a sensible score as demonstrated by a predefined scoring table, based the score the bank will decide if the hopeful will be defaulter or not defaulter on different criteria parameter, for example, person's salary,

age, kind of occupation, the reason for loan, and etc., If the score is higher than the given an incentive than the application is acknowledged FICO assessment is conceded to the customer. In the event that it is higher than a given value, the application is accepted to provide loan otherwise refused.

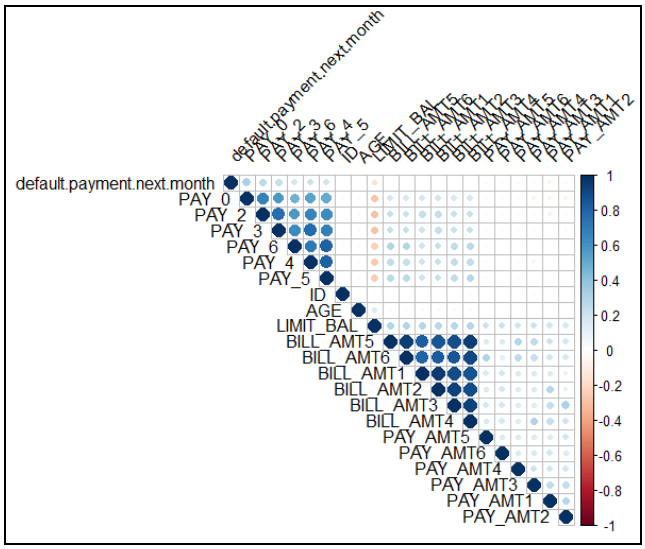
As referenced by (Singh, 2017) [17], Credit Scoring is the primary technique for assessing the Loan applicants and they are classified into two classes called Credible payers and defaulters. Credible payers are those who are trusted applicants who tend to pay the loan on the specified duration whereas the defaulters are those who fail to pay. This credit scoring procedure is utilized by banks and other cash moneylenders to construct a probabilistic prescient model, called a scorecard for evaluating estimating the probability of defaulters

DQR NUMERIC:

	Feature Na	Min	Median	Max	Mean	Stdev
1	ID	1	15000.5	30000	1.50E+04	8.66E+03
2	LIMIT_BAL	10000	140000	1000000	1.67E+05	1.30E+05
3	SEX	1	2	2	1.60E+00	4.89E-01
4	EDUCATION	0	2	6	1.85E+00	7.90E-01
5	MARRIAGE	0	2	3	1.55E+00	5.22E-01
6	AGE	21	34	79	3.55E+01	9.22E+00
7	PAY_0	-2	0	8	-1.67E-02	1.12E+00
8	PAY_2	-2	0	8	-1.34E-01	1.20E+00
9	PAY_3	-2	0	8	-1.66E-01	1.20E+00
10	PAY_4	-2	0	8	-2.21E-01	1.17E+00
11	PAY_5	-2	0	8	-2.66E-01	1.13E+00
12	PAY_6	-2	0	8	-2.91E-01	1.15E+00
13	BILL_AMT1	-165580	22381.5	964511	5.12E+04	7.36E+04
14	BILL_AMT2	-69777	21200	983931	4.92E+04	7.12E+04
15	BILL_AMT3	-157264	20088.5	1664089	4.70E+04	6.93E+04
16	BILL_AMT4	-170000	19052	891586	4.33E+04	6.43E+04
17	BILL_AMT5	-81334	18104.5	927171	4.03E+04	6.08E+04
18	BILL_AMT6	-339603	17071	961664	3.83E+04	5.36E+04
19	PAY_AMT1	0	2100	873552	5.66E+03	1.66E+04
20	PAY_AMT2	0	2009	1684259	5.92E+03	2.30E+04
21	PAY_AMT3	0	1800	896040	5.23E+03	1.76E+04
22	PAY_AMT4	0	1500	621000	4.83E+03	1.57E+04
23	PAY_AMT5	0	1500	426529	4.80E+03	1.53E+04
24	PAY_AMT6	0	1500	528666	5.22E+03	1.78E+04
25	nt.next.month	0	0	1	2.21E-01	4.15E-01

Correlation plot:

The Correlation plot states that the default payment attribute is related to age which is a critical factor in consideration for default assessment

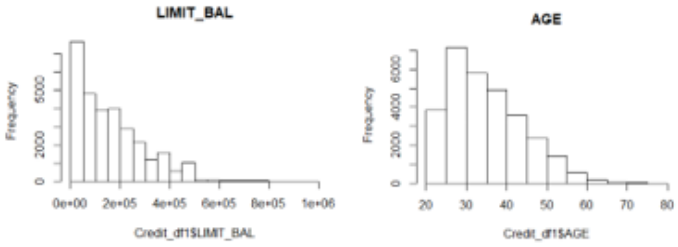


DQR CATEGORICAL:

No.	Instances	Missing	First Mode	Second Mode
1	30000	0	2	1
2	30000	0	2	1
3	30000	0	2	1

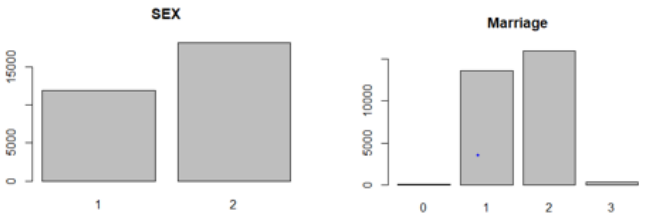
In this data set attributes such as_SEX, EDUCATION, MARRIAGE was in a numeric format which has been converted to categorical.

HISTOGRAM:



Based on the above histogram , It is observed that the LIMIT_BAL and default payment next month attributes are skewed and these data are not normally distributed.

BAR GRAPH:



From the above Bar Graph, it is stated that Loan frequency is high in Women gender and Single status.

Dataset 2: Peer to Peer Lending

In recent years there has been rapid growth peer-to-peer lending platform, P2P platform delivers marketplace in which customer can obtain loans directly without depending any financial institution such as bank. This process benefits the borrowers those have difficulty in obtaining loans from banks [7].

Currently, lenders are making possible to achieve a higher return than investing in financial products such as municipal bonds. As a result, this peer to peer lending has gained a huge number of clients all over the world since its beginning and it is considered to be a disruption in financial innovation sector. Moreover, it becomes volatile where the high return is associated with a high risk, due to this lender suffers a major loan default when the borrower fails to repay the lenders.

The fundamental highlights of P2P lending are adaptable exchanging, low access edges and short advance periods. The two lenders and borrowers can make more prominent benefits through P2P lending than conventional loan given by financial regulations. P2P lending has progressively sought out to be one of the significant types of credits for people and new businesses.

The difference between the benefit score and actual profit is depicted as. The actual profit considers the returns and losses of occasions that are named as non-defaulters and omits there turns and losses of the rest of the occurrences. While for the lender, when a non-defaulter has delegated a defaulter, the moneylender will lose the profits from the advance, which can be viewed as potential misfortunes [19].

So as to expand and correct the identification of good borrowers within the platform framework of social lending, study presents comparisons of different machine learning methods such as Random Forest, Support Vector Machines and Logistic Regression [11].

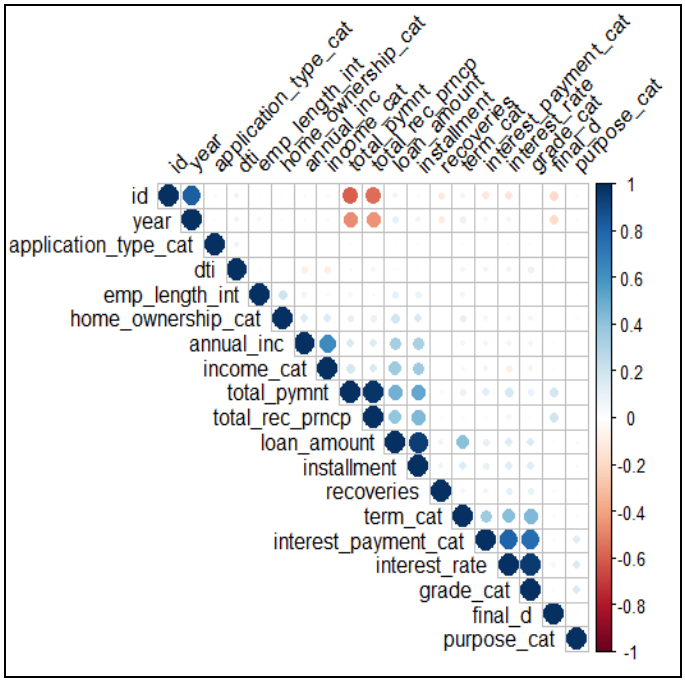
This data set used in this report which is retrieved from Kaggle for the period 2007 to 2015. Observed that this data included loans with the status of fully paid or defaulted only. The process of data manipulation and pre-processing were conducted with open-source statistical software R.

These Datasets are processed and help in identifying the defaulter in order to overcome the financial loss to face by lenders when loans are not repaid by the borrower and ensure the methodical functioning of P2P sectors, it is important to develop measures and methods to help detect possible fraudulent loan requests.

The Dataset contains 887379 rows and 30 columns and No missing values. Out of 30 variables, Let's consider the "Loan Condition" as the critical attribute and distinguish "**0" for Good loan and "1" for Bad loan** in binary number format. The remaining variables are discrete, categorical, and continuous response i.e 20 variables as numeric and 10 variables as categorical.

CORRPLOT:

The Correlation plot states that the blue spectrum indicates the positive correlation while red indicates the negative correlation. From the plot, it is detected that **total _pymnt** and **total_rec_prncp** attributes are highly negatively correlated.



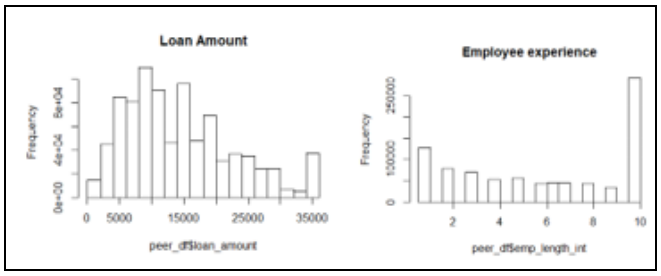
DQR NUMERIC:

	Feature	Cardinality	Min	Median	Max	Mean	Stdev
1	id	887379	54734	34433267	68617057	3.25E+07	2.28E+07
2	year	9	2007	2014	2015	2.01E+03	1.26E+00
3	final_d	98	1012008	1012016	1122015	1.05E+06	4.56E+04
4	emp_length_int	12	0.5	6.05	10	6.05E+00	3.51E+00
5	home_ownership_cat	6	1	3	6	2.10E+00	9.45E-01
6	annual_inc	45784	0	65000	9500000	7.50E+04	6.47E+04
7	income_cat	3	1	1	3	1.20E+00	4.43E-01
8	loan_amount	1372	500	13000	35000	1.48E+04	8.44E+03
9	term_cat	2	1	1	2	1.30E+00	4.58E-01
10	application_type_cat	2	1	1	2	1.00E+00	2.40E-02
11	purpose_cat	14	1	6	14	4.87E+00	2.38E+00
12	interest_payment_cat	2	1	1	2	1.48E+00	4.99E-01
13	loan_condition_cat	2	0	0	1	7.60E-02	2.65E-01
14	interest_rate	542	5.32	12.99	28.99	1.32E+01	4.38E+00
15	grade_cat	7	1	3	7	2.80E+00	1.31E+00
16	dti	4086	0	17.65	9999	1.82E+01	1.72E+01
17	total_pymnt	505628	0	4894.999	57777.58	7.56E+03	7.87E+03
18	total_rec_prncp	260227	0	3215.32	35000.03	5.76E+03	6.63E+03
19	recoveries	23055	0	0	33520.27	4.59E+01	4.10E+02
20	installment	68711	15.67	382.55	1445.46	4.37E+02	2.44E+02

DQR CATEGORICAL:

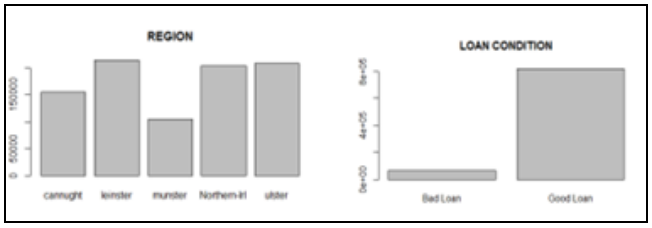
No.	Feature	Missing	Cardinality	FirstMode	SecondMode
1	issue_d	0	103	01-10-2015	01-07-2015
2	home_ownership	0	6	MORTGAGE	RENT
3	income_category	0	3	Low	Medium
4	term	0	2	36 months	60 months
5	application_type	0	2	INDIVIDUAL	JOINT
6	purpose	0	14	debt_consolidation	credit_card
7	interest_payments	0	2	Low	High
8	loan_condition	0	2	Good Loan	Bad Loan
9	grade	0	7	B	C
10	region	0	5	leinster	ulster

HISTOGRAM:



From the Histogram,It is observed that the frequency of loan amount 10000 is high proportional to the Employees having exp greater than 10 years .

BAR PLOT:



From the above bar plot, it states that Loan distribution frequency is high in Leinster which is inversely proportional to the rise of Good Loan condition.

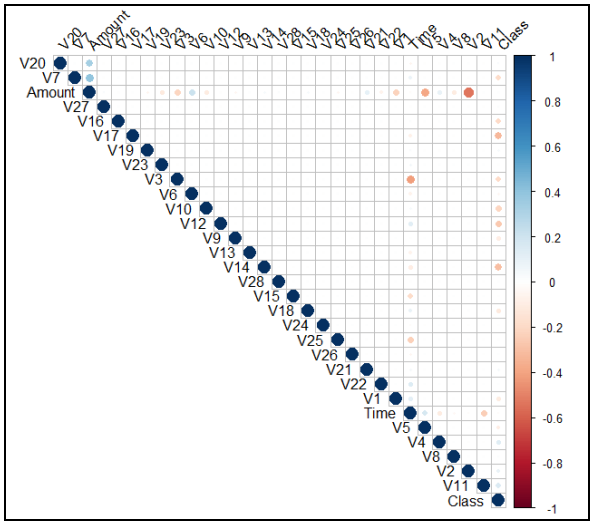
Dataset 3: Credit Card Fraud

According to (Andrea Dal Pozzolo, 2018) [1] — The credit card fraud and defaulter detection through credit card transactions or the social history of customers in unusual other difficulties for the computational insight calculations, Although, this issue can be settled by not of difficulties, for example, confirmation idleness and class unevenness. The majority of the calculations have been executed to depend on the fraud detection framework.

Credit card fraud detection solely depends on the transaction records which can be used for analysis, Transaction is performed in various ways such as Credit card identifier, transaction date, receive, amount of the transaction. Programmed systems are vital since it is not possible or easy for a human analyst to predict or identify the fraudulent patterns occurring in transactional datasets, often characterized by a large number of samples, many dimensions and online updates [14].

According to (E.W.T. Ngai, 2010)[3] Fraud activities in financial industries has become a major issue fraud occurs in various ways such as financial fraud, including credit card fraud, corporate fraud, and money laundering. Economically, financial fraud is becoming an increasingly serious problem.
Financial fraud detection (FFD) is crucial for the stoppage of the often damages, cost of financial fraud.
FFD comprises the classifying feature to segregate financial data from authentic data, thus this behaviors and client actions aiding the decision makers to implement productive methods to reduce the distinguishing fraudulent financial data from authentic data, thereby disclosing fraudulent behavior or activities and enabling decision makers to develop and reduce the effect in fraud.

CORRPLOT:



From the correlated plot, we can observe that column V2 is highly negatively correlated.

DQR NUMERIC:

	Feature	Min	Median	Max	Mean	Stdev
1	Time	0	8.47E+04	1.73E+05	9.48E+04	4.75E+04
2	V1	-56.40751	1.81E-02	2.45E+00	1.17E-15	1.96E+00
3	V2	-72.715728	6.55E-02	2.21E+01	3.12E-16	1.65E+00
4	V3	-48.325589	1.80E-01	9.38E+00	-1.36E-15	1.52E+00
5	V4	-5.683171	-1.36E-02	1.69E+01	2.11E-15	1.42E+00
6	V5	-113.74331	-5.43E-02	3.48E+01	9.80E-16	1.38E+00
7	V6	-26.160506	-2.74E-01	7.33E+01	1.51E-15	1.33E+00
8	V7	-43.557242	4.01E-02	1.21E+02	-5.42E-16	1.24E+00
9	V8	-73.216718	2.24E-02	2.00E+01	1.03E-16	1.19E+00
10	V9	-13.434066	-5.14E-02	1.56E+01	-2.42E-15	1.10E+00
11	V10	-24.588262	-9.29E-02	2.37E+01	2.23E-15	1.09E+00
12	V11	-4.797473	-3.28E-02	1.20E+01	1.71E-15	1.02E+00
13	V12	-18.683715	1.40E-01	7.85E+00	-1.24E-15	9.93E-01
14	V13	-5.791881	-1.36E-02	7.13E+00	8.35E-16	9.95E-01
15	V14	-19.214325	5.06E-02	1.05E+01	1.23E-15	9.59E-01
16	V15	-4.498945	4.81E-02	8.88E+00	4.84E-15	9.15E-01
17	V16	-14.129855	6.64E-02	1.73E+01	1.43E-15	8.76E-01
18	V17	-25.162799	-6.57E-02	9.25E+00	-3.78E-16	8.43E-01
19	V18	-9.498746	-3.64E-03	5.04E+00	9.76E-16	8.38E-01
20	V19	-7.213527	3.73E-03	5.59E+00	1.04E-15	8.14E-01
21	V20	-54.49772	-6.25E-02	3.94E+01	6.41E-16	7.71E-01
22	V21	-34.830382	-2.95E-02	2.72E+01	1.63E-16	7.35E-01
23	V22	-10.933144	6.78E-03	1.05E+01	-3.38E-16	7.26E-01
24	V23	-44.807735	-1.12E-02	2.25E+01	2.67E-16	6.24E-01
25	V24	-2.836627	4.10E-02	4.58E+00	4.47E-15	6.06E-01
26	V25	-10.295397	1.66E-02	7.52E+00	5.11E-16	5.21E-01
27	V26	-2.604551	-5.21E-02	3.52E+00	1.68E-15	4.82E-01
28	V27	-22.565679	1.34E-03	3.16E+01	-3.67E-16	4.04E-01
29	V28	-15.430084	1.12E-02	3.38E+01	-1.23E-16	3.30E-01
30	Amount	0	2.20E+01	2.57E+04	8.83E+01	2.50E+02

The next section reviews the methods which are applied to the datasets and the results across the three datasets.

III. METHODS

A. Logistic Regression

Logistic regression is implemented where the dependent variable is Binary. It describes that the dependent variable are likely have two values such as "Yes or No", "Default or No Default", "Living or Dead", "Responder or Non-Responder", "Yes or No" etc. the variables can be either categorical or numerical variables [12]. For example, we may ask the following: Will an individual making a purchase of an item in the near future?

Where Yi, {YES=1; NO =0}
In other words, "good" or "bad" customers are evaluated with the dependence on the values of explanatory variables of the applicant [2] (10).

ADVANTAGES:

- Logistic regression is robust as it does not require variables to be normally distributed.
- Logistic regression allows witnessing individual relationships between covariates and outcomes.
- LR model is sensitive to correlations amongst the independent variables.
- The key benefit of this approach is that it can produce a simple probabilistic formula of classification [6].

DISADVANTAGES:

- There are some possibilities of prediction bias if the dependent variables are not combined properly, also there is an increase in level correlation between the independent variables [4].
- Does not handle the missing value of continuous variables.
- Sensitive to extreme values of continuous variables.

Performance results of Logistic regression applied across three datasets:

Notions of Performance	CREDIT CARD DEFAULT	PEER TO PEER LENDING	CREDIT CARD FRAUD
	TEST	TEST	TEST
Accuracy	0.2097	0.949	0.9977
95% CI	(0.1953, 0.2248)	(0.9405, 0.9566)	(0.9952, 0.9991)
No Information Rate	0.9727	0.972	0.9967
P-Value [Acc > NIR]	1	1	0.2198
Kappa	-0.0391	0.5028	0.5323
Mcnemar's Test P-Value	<2e-16	<2e-16	0.1306
Sensitivity	0.209119	1	0.4
Specificity	0.231707	0.94751	0.999665
Pos Pred Value	0.906389	0.35443	0.8
Neg Pred Value	0.008169	1	0.997996
Prevalence	0.972658	0.02801	0.003334
Detection Rate	0.203401	0.02801	0.001334
Detection Prevalence	0.224408	0.07903	0.001667
Balanced Accuracy	0.220413	0.97376	0.699833
'Positive' Class	1	1	1

B. Random Forest

Random Forest Classifier is defined as a versatile method capable of performing both managed regression and classification algorithm. It also undertakes dimensional reduction methods, treats missing values, outlier values and other essential steps of data exploration Ensembled algorithms are those which combines more than one algorithm of a same or different kind for classifying objects.

Random forest classifier makes a lot of choice trees from a randomly chosen subset of preparing the set. It at that point totals the votes .from various decision trees to choose the last class of the test object (22). The random forest has the capacity of fitting various choice trees in different examples of datasets for greater improvement in forecast and decrease of over-fitting data.

The random forest can estimate the importance of a variable in a prediction task, and we could extract the relatively important features through importance ranking provided by the random forest model [20].

ADVANTAGES:

- This algorithm can solve both types of problems i.e. classification and regression.
- It has methods for balancing errors in data sets where classes are imbalanced.

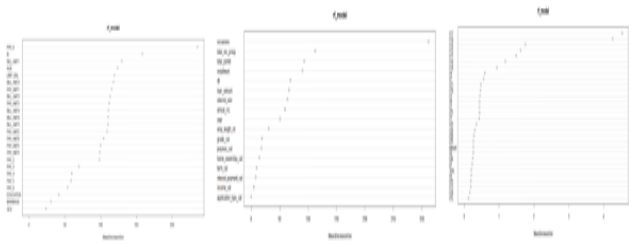
DISADVANTAGES:

- There will be no problems faced during classification however it is not the same case for regression since it doesn't give precise continuous nature prediction [19].
- Random forests act like a Black box approach since the user have little control over what the model does [20].

Performance results of Random Forest applied across three datasets:

Notions of Performance	CREDIT CARD DEFAULT	PEER TO PEER LENDING	CREDIT CARD FRAUD
	TEST	TEST	TEST
Accuracy	0.8183	0.955	0.9997
95% CI	(0.804, 0.8319)	(0.9469, 0.9621)	(0.9981, 1)
No Information Rate	0.8773	0.966	0.9987
P-Value [Acc > NIR]	1	0.9994	0.09143
Kappa	0.3777	0.5819	0.8887
Mcnemar's Test P-Value	<2e-16	<2e-16	1
Sensitivity	0.67391	1	1
Specificity	0.83846	0.9534	0.999666
Pos Pred Value	0.3685	0.43038	0.8
Neg Pred Value	0.94841	1	1
Prevalence	0.12271	0.03401	0.001334
Detection Rate	0.08269	0.03401	0.001334
Detection Prevalence	0.22441	0.07903	0.001667
Balanced Accuracy	0.75619	0.9767	0.999833
'Positive' Class	1	1	1

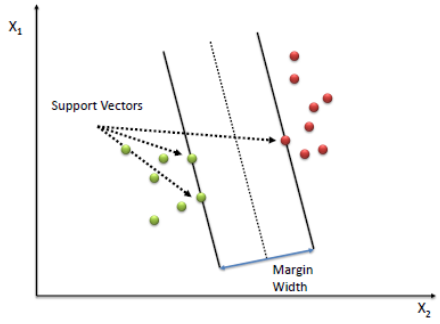
RF_ Graph for all 3 datasets:



- *Above RF graph states that the highest point is where the attribute which should be chosen for decision making.*

C. Support Vector Machines

Support Vector Machine" (SVM)" defined as a controlled machine learning algorithm which is suitable for classification tasks, but it can also be used in for classification or regression. SVM is capable to classify the large dataset [15]. Classification is implemented by detecting the hyperplane that distinguish the two classes very well. Support vector machine is the most used method for pattern recognition and classification [18].



ADVANTAGES:

- The result of SVM is optimal, exclusive and global since the training is done by solving a linearly constrained quadratic problem [13].
- In high dimensional spaces SVM is Effective
- It is highly recommendable in cases dimensions are greater than a number of samples.

DISADVANTAGES:

- The performance is low when they are unable to initialize dataset which is large with overlapping classes and includes more noise
- Performance is low when dataset has more noise

Performance results of Support Vector Machines applied across three datasets:

Notions of Performance	CREDIT CARD DEFAULT	PEER TO PEER LENDING	CREDIT CARD FRAUD
	TEST	TEST	TEST
Accuracy	0.8176	0.9547	0.9987
95% CI	(0.8033, 0.8313)	(0.9466, 0.9618)	(0.9966, 0.9996)
No Information Rate	0.7756	0.921	0.9983
P-Value [Acc > NIR]	9.397E-09	1.09E-13	0.4403
Kappa	0.3467	0.5799	0.333
Mcnemar's Test P-Value	< 2.2e-16	< 2.2e-16	0.1336
Sensitivity	0.963	0.9996	1
Specificity	0.315	0.4304	0.2
Pos Pred Value	0.8293	0.9534	0.9987
Neg Pred Value	0.7114	0.9903	1
Prevalence	0.7756	0.921	0.9983
Detection Rate	0.7469	0.9206	0.9983
Detection Prevalence	0.9006	0.9657	0.9997
Balanced Accuracy	0.639	0.715	0.6
'Positive' Class	0	0	0

D. Oversampling

Oversampling or under sampling are utilized for class irregularity issues, when you have low proportion for a specific class value for the dependent variable (class forecast demonstrate) in a given example , Random oversampling balances the information by oversampling the minority class[9].In spite of the fact that oversampling enhances the performance of some two-class classifier, the overall performance doesn't lead to best-performing classifiers [16].The goal of using oversampling is that to represent the learning classifier to be a balanced training set.

ADVANTAGES:

- An advantage of using this method is that it leads to no information loss

DISADVANTAGES:

- Over-sampling essentially includes repeated results of the original data eventually it adds up multiple observation of several types which leads to overfitting. Even though the accuracy of train data is high the accuracy of unseen data will be inferior.

Performance results of Over-Sampling applied across three datasets:

Notions of Performance	CREDIT CARD DEFAULT	PEER TO PEER LENDING	CREDIT CARD FRAUD
	TEST	TEST	TEST
Accuracy	0.1801	0.0984	0.0033
95% CI	(0.1665, 0.1943)	(0.0879, 0.1096)	(0.0016, 0.0061)
No Information Rate	0.855	0.8926	0.9963
P-Value [Acc > NIR]	1	1	1
Kappa	-0.1787	-0.0856	-0.002
Mcnemar's Test P-Value	<2e-16	<2e-16	<2e-16
Sensitivity	0.15172	0.03922	0.0006693
Specificity	0.34713	0.59006	0.7272727
Pos Pred Value	0.57801	0.44304	0.4
Neg Pred Value	0.06492	0.06879	0.002672
Prevalence	0.85495	0.89263	0.9963321
Detection Rate	0.12971	0.03501	0.0006669
Detection Prevalence	0.22441	0.07903	0.0016672
Balanced Accuracy	0.24942	0.31464	0.363971
Positive' Class	1	1	1

The next section reviews the Analysis of the methods which are applied to the datasets and evaluate the results.

IV. EVALUATION

A small brief about the Notion of performances:

A) **Cohens Kappa**: Cohens Kappa is a statistical measure i.e. to be more accurate a measure of agreement between the two individuals.

B) **RSME**: - The root-mean-squared error (RMSE) is a measure of how well your model performed. It does this by measuring the difference between predicted values and the actual values

C) **RSS**: - RSS is defined as Residual Sum of Squares i.e total of the squares of residuals between observed values and predicted values.

D) **Sensitivity**: -It is a measure of identified current positives.

E) **Specificity**: - It is a measure of identified current negatives

F) **F-Measure**: - defined as the measure the accuracy of the prediction models.

G) **MAPE**: - MAPE is defined as Mean Absolute Percentage Error is a method which is used for predicting and forecasting accuracy.

H) **Confusion Matrix**: - Confusion matrix also known as Error matrix which is used to evaluate the performance of a classification model. There are 4 cases in which a method is evaluated TP, TN, FP, and FN .

I) **Under sampling**: - Under sampling is defined as the small samples used in machine learning algorithms to predict accuracy.

	DATASET 1				DATASET 2				DATASET 3			
Notions of Performance	CREDIT CARD DEFAULT				PEER TO PEER LENDING				CREDIT CARD FRAUD			
	LR	Random Forest	Over Sampling	SVM	LR	Random Forest	OS	SVM	LR	RANDOM FOR	Over Sampling	SVM
Accuracy	0.2057	0.8183	0.1801	0.8176	0.949	0.955	0.0984	0.9547	0.9977	0.9997	0.0033	0.9987
95% CI	(0.1953, 0.2249)	(0.804, 0.8319)	(0.1665, 0.1943)	(0.8033, 0.8313)	(0.9405, 0.9566)	(0.9469, 0.9621)	(0.0879, 0.1096)	(0.9466, 0.9618)	(0.9952, 0.99	(0.9981, 1)	(0.0016, 0.00	(0.9966, 0.9996)
No Information Rate	0.9727	0.8773	0.855	0.7756	0.972	0.966	0.8926	0.921	0.9967	0.9987	0.9963	0.9983
P-Value [Acc > NIR]	1	1	1	9.40E-09	1	0.9994	1	1.09E-13	0.2198	0.09143	1	0.4403
Kappa	-0.0391	0.3777	-0.1787	0.3467	0.5028	0.5819	-0.0856	0.5799	0.5323	0.8887	-0.002	0.333
Mcnemar's Test P-Value	<2e-16	<2e-16	<2e-16	<2.2e-16	<2e-16	<2e-16	<2e-16	<2.2e-16	0.1306	1	<2e-16	0.1336
Sensitivity	0.209119	0.67391	0.15172	0.963	1	1	0.03922	0.9996	0.4	1	0.0006693	1
Specificity	0.231707	0.83846	0.34713	0.315	0.94751	0.9534	0.59006	0.4304	0.999665	0.999666	0.7272727	0.2
Pos Pred Value	0.906389	0.3685	0.57801	0.8293	0.35443	0.43038	0.44304	0.9534	0.8	0.8	0.4	0.9987
Neg Pred Value	0.008169	0.94841	0.06492	0.7114	1	1	0.06879	0.9903	0.997996	1	0.002672	1
Prevalence	0.972658	0.12271	0.85495	0.7756	0.02801	0.03401	0.89263	0.921	0.003334	0.001334	0.9963321	0.9983
Detection Rate	0.203401	0.08269	0.12971	0.7469	0.02801	0.03401	0.03501	0.9206	0.001334	0.001334	0.0006669	0.9983
Detection Prevalence	0.224408	0.22441	0.22441	0.9006	0.07903	0.07903	0.07903	0.9657	0.001667	0.001667	0.0016672	0.9997
Balanced Accuracy	0.220413	0.75619	0.24942	0.639	0.97376	0.9767	0.31464	0.715	0.699833	0.999833	0.363971	0.6
'Positive' Class	1	1	1	0	1	1	1	0	1	1	1	0

Figure 1:The consolidated table defines the performance values of methods applied to the dataset.

Analysis:

a) Dataset-I

- From the above report, it is observed that Random Forest has the highest accuracy rate(81.83%) which also out performs the other methods . support vector method(81.76%), which had the nearest accuracy to Random Forest both Logistic regression(20%) and Over Sampling is 18%. Hence Random Forest has been recommended for this dataset.

b) Dataset-II: -

- RF and SVM have similar prediction accuracy rates of RF 95.55% followed by SVM at 95.47% and LR at 94.09% and Over Sampling (0.09%). RF is Highly recommended.

c) Dataset-III: -

- In the third data set, RF seems to be having the high accuracy rate (99.97%) with the nearest method SVM (99.87%) accuracy followed by LR with an accuracy rate of (99.77%) and Over sampling with (0.3%). so, Oversampling is not recommended for this dataset since there is a high chance of over fitting. Hence RF is recommended for this dataset.

V. CONCLUSION

To conclude, on account of complex financial products, risk mitigation and credit scoring models play a vital role in reducing risks and thereby reducing loan losses. The datasets that have been identified here are with a purpose of predicting delinquencies and limiting loan losses. These datasets speak about credit card and P2P loans that are offered by banks and financial institutions. Hence, the application of algorithms is required for reducing loan losses though it cannot be prevented completely. The underlying shortlisting factor for methods that can be applied to the datasets is none other than the most basic utility of machine learning programs "Accuracy". The datasets which have been chosen to have similar business motive since all three datasets come under the financial risk management circle where the end goal is to predict the defaulter or to decide whether the customer eligible for a lending loan. these datasets define various issues such as loan default, credit card default and credit card fraud. So, these datasets are explored and applied to the 4 chosen models namely Logistic regression, Random Forest, Oversampling and Support vector machines. From the results its Random Forest more likely to be recommended model for the financial risk management division.

From the above work it is clearly inferred that Accuracy is the key aspect for any predicting function in this work through the credit card, Peer to Peer lending has been identified for delinquencies prediction purpose, but this scope is huge and are applicable to various financial domains and non-financial domains such as insurance, manufacturing, healthcare etc. Well researched blend of effective algorithms

combined with potent machine learning techniques can lead to predictive results.

VI. References

[1] Andrea Dal Pozzolo, G. B. (2018). Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy. *IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS*.

[2] CLIFTON PHUA, V. L. (2010). A Comprehensive Survey of Data Mining-based Fraud Detection Research. School of Business Systems, Faculty of Information Technology.

[3] E.W.T. Ngai, Y. H. (2010). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Institute of Business Intelligence and Knowledge Discovery.

[4] Emerland. (2017). Towards reliable prediction of academic performance of architecture students :using data mining techniques . Journal of Engineering, Design and Technology .

[5] FEI LI', J. X.-T.-L. (2004). DATA MINING-BASED CREDIT EVALUATION FOR USERS OF CREDIT. Third International Conference on Machine Learning and Cybernetics.

[6] I-Cheng Yeh, C.-h. L. (2007). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications.

[7] Jennifer Xu, M. C. (2016). Identifying Features for Detecting Fraudulent Loan Requests on P2P Platform. IEEE.

[8] Kalyani R. Rawate, P. P. (2017). REVIEW ON PREDICTION SYSTEM FOR BANK LOAN CREDIBILITY. Scientific Journal of Impact Factor (SJIF): 4.72.

[9] Kenneth Kennedy, B. M. (2013). Using Semi-supervised Classifiers for Credit. School of Computing.

[10] Lobna Abid, A. M.-G. (2016). THE CONSUMER LOAN'S PAYMENT DEFAULT PREDICTIVE MODEL: AN APPLICATION IN A TUNISIAN COMMERCIAL BANK . Asian Economic and Financial Review .

[11] Milad Malekipirbazari, V. A. (2015). Risk assessment in social lending via random forests . Elseiver.

[12] P. Ravisankar, V. R. (2009). Detection of financial statement fraud and feature selection using data. Institute for Development and Research in Banking Technology.

[13] Ping-Feng Paia, b. M.-F.-C. (2010). A support vector machine-based model for detecting top management fraud . Knowledge-Based Systems.

[14] Pozzolo, A. D. (2015). Adaptive Machine Learning for Credit Card Fraud Detection. Université Libre de Bruxelles.

[15] Rustam, F. Y. (2017). Application of Support Vector Machines for Reject Inference in Credit Scoring . Proceedings of the 3rd International Symposium on Current Progress in Mathematics and Sciences .

[16] Sihem Khemakhem, F. B. (2017). Credit risk assessment for unbalanced datasets based on data mining, artificial neural network and support vector machines . Journal of Modelling in Management .

[17] Singh, P. (2017). Comparative Study of Individual and Ensemble Methods of Classification for Credit Scoring. International Conference on Inventive Computing and Informatics (ICICI 2017).

[18] V.Mareeswari, D. G. (2016). Prevention of Credit Card Fraud Detection based on HSVM . International Conference On Information Communication And Embedded System(ICICES 2016) .

[19] XinYe, L.-a. (2018). Loan evaluation in P2P lending based on Random Forest optimized by genetic algorithm with profit score . Electronic Commerce Research and Applications.

[20] Yu Jina, Y. Z. (2015). A data-driven approach to predict default risk of loan for online Peer-to-Peer(P2P) lending-. Fifth International Conference on Communication Systems and Network Technologies.

[21] Yufeng Kou, C.-T. L.-P. (2004). Survey of Fraud Detection Techniques. International Conference on Networking, Sensing & Control.

[22] Zhao Wang, C. J. (2017). A novel behavioral scoring model for estimating probability of default over time in Peer-to-Peer lending. Electronic Commerce Research and Applications .