IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 9, September 2021

DOI: 10.17148/IJARCCE.2021.10925

Data Analytics for Credit Risk analysis in the Banking Sector: Linear Regression

Sumukh Mydur

Student, BE, Department of CSE, Dr. AIT, Bangalore, India

Abstract: Credit risk is the probability of a loss resulting from a creditor's failure to repay a loan or fulfil any other contractual obligations towards the investor. Traditionally, it relates to the hazard that a lender may not receive the owed head and premium, which follows a disruption of incomes and expanded expenses for collection. Unnecessary cash may be written to create additional income to cover for credit risk. Despite it is being impossible to know exactly who will default on commitments, satisfactorily surveying and overseeing credit risk can diminish the seriousness of a loss. The lender or investor earn a bonus for risking credit default and lending money in the form of interest from the borrower or issuer of a debt obligation. When lenders or banks provide mortgages, credit cards, visas or various types of credit or loans, there is a hazard that the borrower is probably not going to reimburse the loan. Likewise, if an organization provides credit to a client, there is a hazard that the client is not going to pay their solicitations. Credit risk additionally clarifies the risk that a guarantor may stall to make payment when asked or that an insurance company will be unable to pay a claim. Credit risks are determined based on the borrower's general ability to reimburse an advance as indicated by its unique terms. To assess credit risk on a consumer loan, loan specialists inspect the five Cs: credit history, capacity to repay, capital, the loan's conditions, and associated collateral. Banks have been the most important institutions of money lending and deposits. Primary functions include accepting deposits, offering loans, credit, overdraft, providing liquidity and discounting of bills. Secondary functions include providing safe custody of valuables, loans on valuables, corporate and consumer finances. Though the structure of banks has remained the same, the functionalities have been boosted. Automated tools, bots and computers have modernized the banking system. The dataset accumulated over a period of time is so huge that, automation tools and computer programs are the need of the day. In this paper we have tried to enhance the present bank credit-debit system by the use of Artificial Intelligence. Machine learning is a subset of AI and directly trains the machine by feeding the historic and runtime data collected during transactions. The machine which is trained is now capable of taking decisions, thereby making predictions. This would characterize the dataset as stored and predicted outcomes. Every business enthusiast would have keen interest to carefully study the performance of a financial institute for his/her benefit. In this assignment we have used both classification and regression algorithms to create a ML model of prediction. Linear regression model is designed from scratch using formula method. Classification algorithms like Support Vector Machine (SVM), Random Forest Classifier and KNN algorithms are effectively applied to fit to the dataset. Comparisons must be made during implementation to understand the pattern of predicted data. Regression algorithms like linear regression (developed from scratch) will be a boost to the accuracy of the assignment (categorical data excluded).

Keywords:accepting deposits, offering loans, credit, overdraft, providing liquidity and discounting of bills, Automated tools, bots and computers, Machine learning, Support Vector Machine (SVM), Random Forest Classifier and KNN algorithms, linear regression (developed from scratch) , historic and runtime data collected during transactions, AI, five Cs: credit history, capacity to repay, capital, the loan's conditions, and associated collateral.

I.INTRODUCTION

In this information era, huge amount of data is being stored, exchanged and conditioned. The volume of data that one has to deal with has exploded to unimaginable levels. Most of the data exists in its crude form and needs to be converted to useful format before analysis. This process of converting raw data into useful format is called data pre-processing. Real world data is [1]

- Incomplete: consists of missing attribute values or consists of only aggregate data.
- Noisy: containing errors or outliers.
- Inconsistent: containing discrepancies in code.

II.MOTIVATION

With the growth in financial services, the banks are facing a towering loss from inadvertent loans. In such case, there is a requirement for the bank to come up with their own credit risk evaluation framework. Nevertheless, a minority of the

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 10, Issue 9, September 2021
DOI: 10.17148/IJARCCE.2021.10925

banks have failed in developing software which precisely predicts customer's default and the said banks have undergone enormous loss. The loan breach still happens, usually in case of the commercial lender. From the Federal Reserve senior loan officer opinion survey report it was observed that, oil and gas companies defaulted on $39 billion in 2016, and the major yield bond default rate for the energy sector reached a peak at 18.8% during the year. In fact, the loan failure happened in every industry. Therefore, the commercial credit risk prediction is a vital research part that helps to uphold the economic environment. Software is very much required to differentiate good creditors from bad ones, which is a major decision for any credit giving organization (for example business banks and certain retailers). The accuracy of the software is delicate as the bank's durability is tied to taking appropriate risks; a non-risk-taking bank is as vulnerable as an overly-risk taking one. With the need of a credit risk assessment model with high accuracy, we develop a model using the statistical method, Linear Regression. Automated tools, bots and computers have modernized the banking system. The dataset accumulated over a period of time is so huge that, automation tools and computer programs are the need of the day. In this paper we have tried to enhance the present bank credit-debit system by the use of Artificial Intelligence.

III. ISSUES AND CHALLENGES

I. Inefficient Data Management

Credit risk management deliverables require the ability to securely store, categorize and search data based on a wide range of criteria. Any database needs to be revised in a real time to avoid potentially outdated information, as well as be keyword optimized to ensure easy location of information.

II. Limited Group-Wide Risk Modelling Infrastructure

Sometimes it's not enough to analyse the risk qualities posed by a single entity- a broad, comprehensive perspective of all risk measures as seen from above is key to understanding the risk posed by a new borrower to the group. Robust stress-testing capabilities and model management that spans the entire modelling lifecycle is the key to ensure accurate risk assessment.

III. Lacking Risk Tools

Identifying portfolio concentrations or re-grade portfolios is necessary to ensuring you're seeing the big picture. A comprehensive risk assessment scorecard should be able to quickly and clearly recognize positives and negatives associated with a loan.

IV. Less-than-intuitive Reporting and Visualization

Forget cumbersome spreadsheet-based processes- to glean the most valuable insights, data and analysis must be submitted in an intuitive, clean and clearly visualized way. Stripping away extraneous data that overburdens analysts and IT can help zero in on the most pertinent information.

IV. PROBLEM STATEMENT

In a bank huge dataset is produced with everyday transaction and with ever increasing deposits, loans, insurance policies, over drafts and other services. A bank with huge customers is considered and transaction data of 20 years has been recorded. The identity of the customer is morphed. Unique IDs to be presented to the same. Data needs to be examined for pattern recognition and data pre-processing needs to be carried out to –

- Fill the missing values or null values
- Remove redundant entries.
- Treat NaN values.
- Replace string values with their numerical counterparts.
- Create a sketch of post-assigned categorical values in each column defining a particular attribute.

The pre-processed data needs to be fed to the machine for training. The patterns would train the machine to make predictions in all possible situations. Classification algorithms like SVM, Random forest classifier, KNN and logistic regression need to be applied. Linear regression is modelled from scratch without using libraries for more accuracy and F1 score.

V. METHODOLOGY

i. Data acquisition – Data acquisition is carried out. Everyday transactions are recorded and stored in the database. Figure 1 shows the process of data acquisition. 20 years data transaction consists of about 1.5 million unique transaction IDs. About 140 parameters or attributes are part of this dataset. Few of them have been tabulated above. [1] [2]

ii. Data Inspection – The acquired data is inspected before data pre-processing. The data needs to be preprocessed before analytics or training. Figure 2 shows the process of data inspection. [3]

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 9, September 2021

DOI: 10.17148/IJARCCE.2021.10925

iii. Data Visualization – Graphical analysis of the dataset which is huge in nature is essential. Figure 3 shows a bar graph of verification_status v/s count. Figure 4 shows a count plot of loan purpose. Figure 5 shows a hue plot of home_ownership against loan_status. [4]

iv. Correlation is carried out and heat map plotted as shown in figure 6. Regions of strong and weak correlation is described by the color bar. Neutral values are ignored.

v. Linear regression is used to create a ML model for columns with non-categorical behavior. Figure 7 shows the code bit of the same using formula method (no libraries used). [5]

vi. Classification algorithms like SVM, KNN and Random forest classifier are applied to the model.

vii. The predicted values are well tabulated, accuracy measured and compared. [6]

viii. The predictions are sent back to be stored in the database for a closed loop execution in the coming years and will be continuously compared with the then run time values or transactions.

ix. The master table or the data table with raw or crude data is updated each time a new transaction takes place and the manipulated dataset is treated with time before execution.

```
In [1]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns
        from sklearn.linear_model import LogisticRegression
        from sklearn.model_selection import train_test_split

In [2]: df= pd.read_csv(r'C:\Users\nups0\Desktop\dataset.csv')

In [3]: df

Out[3]:
```

	id	loan_amnt	funded_amnt	funded_amnt_inv	term_months	int_rate	installment	emp_length	home_ownership	annual_inc	...	num_tl_90g_dpd_
0	1474286	30000	30000	30000	36	22.35	1151.16	5.0	1	100000.0	...	
1	1474287	40000	40000	40000	60	16.14	975.71	0.5	1	45000.0	...	
2	1474288	20000	20000	20000	36	7.56	622.68	10.0	1	100000.0	...	

Figure 1 shows the process of data acquisition.

```
In [6]: df.size

Out[6]: 614481

In [7]: df.shape

Out[7]: (7063, 87)

In [8]: df.dtypes

Out[8]: id                          int64
        loan_amnt                   int64
        funded_amnt                 int64
        funded_amnt_inv             int64
        term_months                 int64
                                    ...
        tot_hi_cred_lim             int64
        total_bal_ex_mort           int64
        total_bc_limit              int64
        total_il_high_credit_limit  int64
        disbursement_method         int64
        Length: 87, dtype: object
```

Figure 2 shows the process of data inspection.

IJARCCE

ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 9, September 2021

DOI: 10.17148/IJARCCE.2021.10925

```
In [14]: f= plt.subplots(figsize=(15,5))
         color_types=['#00FF00','#00FFFF','#FFAAEE']
         sns.countplot(x="verification_status",palette=color_types, data=df).set_title("Verification status graph")

Out[14]: Text(0.5, 1.0, 'Verification status graph')
```

Figure 3 shows a bar graph of verification_status v/s count.

```
In [17]: f= plt.subplots(figsize=(15,5))
         color_types=['#00FF00','#00FFFF','#FFAAEE']
         sns.countplot(x="purpose",palette=color_types, data=df).set_title("Loan purpose graph")

Out[17]: Text(0.5, 1.0, 'Loan purpose graph')
```

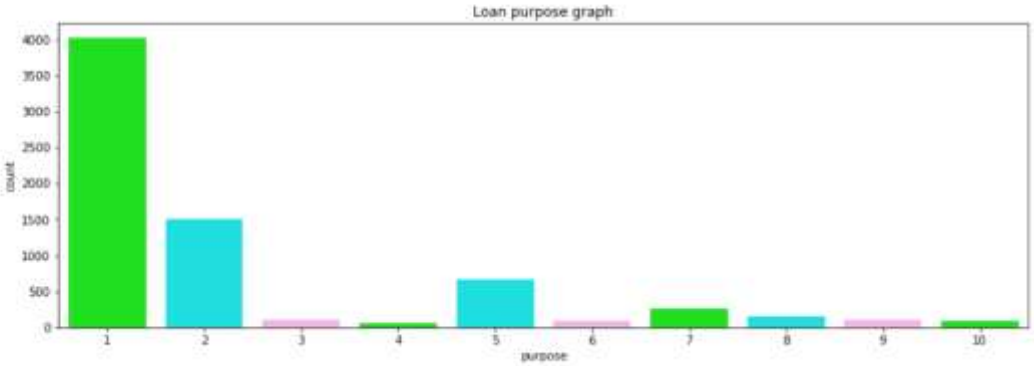

Figure 4 shows a count plot of loan purpose.

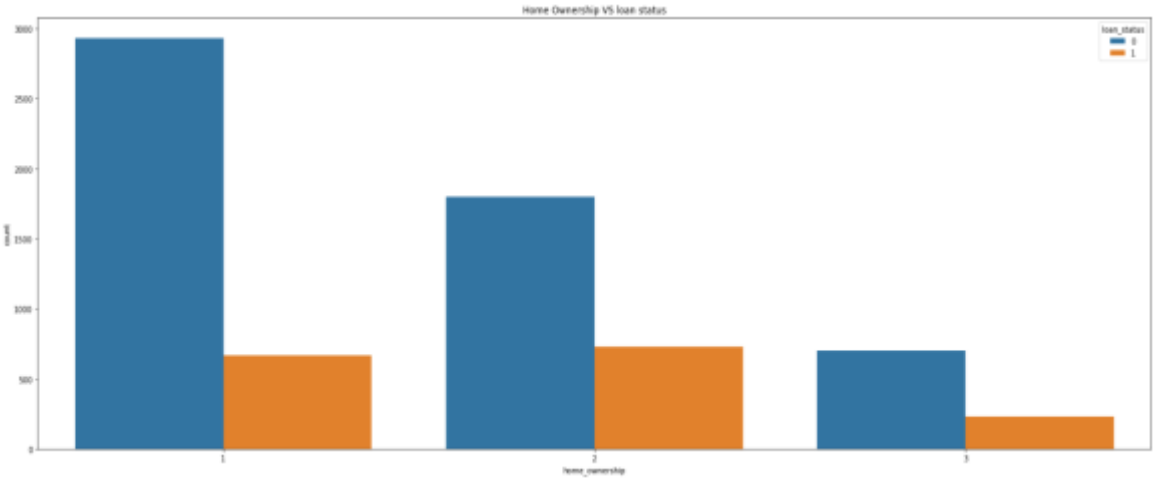

Figure 5 shows a hue plot of home_ownership against loan_status.

IJARCCE

ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 9, September 2021

DOI: 10.17148/IJARCCE.2021.10925

	id	loan_amnt	funded_amnt	funded_amnt_inv	term_months	int_rate	installment	emp_length	home_ownership	annual_inc	..
id	1.000000	0.027341	0.027341	0.027288	0.041211	0.130415	0.044465	-0.035195	0.057782	-0.055215	..
loan_amnt	0.027341	1.000000	1.000000	0.999995	0.390274	0.106911	0.951471	0.082418	-0.089565	0.324420	..
funded_amnt	0.027341	1.000000	1.000000	0.999995	0.390274	0.106911	0.951471	0.082418	-0.089565	0.324420	..
funded_amnt_inv	0.027288	0.999995	0.999995	1.000000	0.390464	0.107005	0.951404	0.082396	-0.089583	0.324391	..
term_months	0.041211	0.390274	0.390274	0.390464	1.000000	0.383300	0.167894	0.041518	-0.065941	0.041844	..
...	...	...	...	...	...	...	...	...	...	...	..
tot_hi_cred_lim	-0.096492	0.279343	0.279343	0.279400	0.079582	-0.112946	0.250768	0.139870	-0.368591	0.557846	..
total_bal_ex_mort	-0.032480	0.217649	0.217649	0.217665	0.084091	0.058044	0.213551	0.046978	-0.127184	0.399346	..
total_bc_limit	-0.109086	0.311789	0.311789	0.311762	0.045502	-0.234596	0.277310	0.066914	-0.062696	0.348812	..
total_il_high_credit_limit	-0.039759	0.159554	0.159554	0.159590	0.064376	0.032922	0.155136	0.048567	-0.113876	0.360716	..
disbursement_method	0.127683	-0.033862	-0.033862	-0.033713	-0.014596	0.116828	-0.011868	0.002189	0.013929	-0.005046	..

Figure 6 shows the heatmap coefficients.

```
In [12]: #mean of X and Y
         mean_x = np.mean(X)
         mean_y = np.mean(Y)
         #total number of values
         m = len(X)
         #formula to calculate b1 and b0
         numer= 0
         denom= 0
         for i in range(m):
             numer+= (X[i] - mean_x) * (Y[i] - mean_y)
             denom+= (X[i] - mean_x) ** 2
         b1 = numer/denom
         b0 = mean_y - (b1 * mean_x)
         #b1 and b0 are m and c respectively in y=mx+c
         print(b1,b0)

         0.08346747763810816 12009.934165736577
```

Figure 7 shows the code bit of the same using formula method (no libraries used).

VI. RESULTS

- Figure 8 shows the accuracy comparison of the classification algorithms. Random Forest Classifier model has an accuracy of 99.9%, Logistic regression model has an accuracy of 99.75%, KNN model has an accuracy of 80.89% at K=3 and 5 and SVM model has an accuracy of 75.87%. So, we Random Forest and Logistic Regression method show very good accuracies and are a very good fit to this assignment.
- Linear Regression model shows an accuracy of 91.155%. Figure 9 shows the r^2 value of the linear regression model.
- Figure 10 shows the predictions or the X_test values after linear regression.

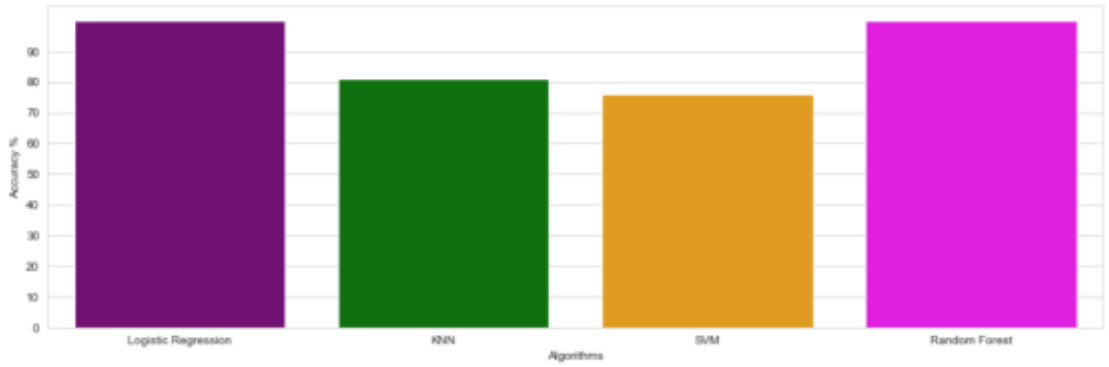

Figure 8 shows the accuracy comparison of the classification algorithms.

IJARCCE

ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 9, September 2021

DOI: 10.17148/IJARCCE.2021.10925

```
In [14]: ss_t = 0
         ss_r = 0
         for i in range(m):
             y_pred = b0 + b1 * X[i]
             ss_t += (Y[i] - mean_y) ** 2
             ss_r += (Y[i] - y_pred) ** 2
         r2 = 1 - (ss_r/ss_t)
         R2=r2*10
         print(R2*100)

         91.15564927258013
```

Figure 9 shows the r^2 value of the linear regression model.

```
In [25]: Y[250:500]

Out[25]: array([10000,  5000, 10000, 10000, 20000,  3500,  9500, 30000, 15000,
                15000,  5000, 25000, 10800, 40000,  5500, 10000, 10000, 16800,
                 3800, 10000,  7000,  6000,  8000, 30000, 12000, 15500, 25000,
                 1300, 25000,  5000, 10000,  6400, 11000, 28000, 12000, 12350,
                35000, 30000, 30000,  1000, 10000, 11000, 35000,  7000, 31400,
                16000,  3000, 10000,  1600, 17000,  5500,  8000,  2500, 15000,
                18000, 35000, 15000, 10000,  6500, 15000, 24000, 10000, 25000,
                29700, 37000, 16000, 19050,  9000, 24000, 35000, 10000, 10000,
                40000, 18000, 16000, 15000,  3200, 40000, 40000,  1000,  2500,
                 8000, 19000,  9000, 12000, 30000, 10000, 25000, 16000,  6000,
                12000, 30000,  6500, 27650, 24000, 20000, 25000, 25000, 10000,
                35500, 20000,  3000,  6000, 28000, 20000, 14500, 25000, 10000,
                 7500, 30000,  2850,  2000, 28000, 24000, 40000, 10000,  8800,
                19000,  6125, 20000, 40000, 19975, 21000, 23500,  3900, 16500,
                15000,  8500, 32425, 30000,  4000,  8000, 30000, 15000, 15000,
                 7200,  4800, 20000, 18000,  5000, 27000, 10000,  5000, 10000,
                 9500, 28800, 31200,  5000,  1400, 12000,  5000,  9000, 20550,
                 7000,  2000, 15600,  4500, 40000, 21000, 16000, 10000, 20000,
                 8000, 10000,  1500, 40000,  6600, 17000, 11000,  4800, 28000,
                12000,  4500,  2600,  5000,  6475,  4000,  2000,  3600, 10000,
                18000,  4000, 10000,  2000, 10000,  3000,  2000, 10000, 15000,
                30000,  4500,  3200,  7500, 35000, 10400,  6300,  8000, 34000,
                30000, 20000, 30000, 15000, 19200,  8000, 40000, 11200, 10000,
                 5500, 12000, 21000,  6000,  8500,  2000, 40000, 12000, 25000,
                13500, 10000, 40000,  3000, 12000, 23500, 10000, 11000, 22000,
                11775, 17000,  6000, 38000, 40000, 30000,  7200,  9600, 15000,
                10000,  5000,  6025,  9500,  9000,  8000, 15000,  3000,  5000,
                 3800,  8000, 12000, 16000, 14000,  4000,  4800], dtype=int64)
```

Figure 10 shows the predictions or the X_test values after linear regression.

VII. CONCLUSIONS

A Bank proactive in business in this 21st century world has many day to day transactions. Data analytics had to be carried out on the data –both historical and present trend to draw inference. The goal was to create or improve the ML model and carry out accuracy check comparison.

A python code was written and executed in the Jupyter platform to analyse and draw conclusions. Classification algorithms like Support Vector Machine (SVM), Random Forest Classifier and KNN algorithms are effectively applied to fit to the dataset. Comparisons must be made during implementation to understand the pattern of predicted data.Random Forest Classifier model has an accuracy of 99.9%, Logistic regression model has an accuracy of 99.75%, KNN model has an accuracy of 80.89% at K=3 and 5 and SVM model has an accuracy of 75.87%. We can conclude that Random Forest Classifier and Logistic Regression models are the best fit to this dataset. Since this data also behaves well for Linear regression algorithm, Linear regression is modelled from scratch without using libraries for more accuracy (91.155%) and F1 score.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 9, September 2021

DOI: 10.17148/IJARCCE.2021.10925

REFERENCES

[1] Principles of data mining- DJ Hand - Drug safety, 2007 - Springer
[2]The Python Standard Library — Python 3.7.1rc2 documentation-https://docs.python.org/3/library/
[3]Research on Data Preprocess in Data Mining and Its Application- J Zhi-gang, JIN Xu - Application Research of Computers, 2004 - en.cnki.com.cn
[4]Data Mining and Analytics: A Proactive Model - http://www.ijarcce.com/upload/2017/february-17/IJARCCE%20117.pdf
[5] A comparative analysis on linear regression and support vector regression-DOI: 10.1109/GET.2016.7916627
[6] Data Warehousing Architecture and Pre-Processing- Vishesh S, Manu Srinath, Akshatha C Kumar, Nandan A.S.- IJARCCE, vol 6, issue 5, May 2017.

OUR GUIDE

VISHESH S born on 13th June 1992, hails from Bangalore (Karnataka) and has completed B.E in Telecommunication Engineering from VTU, Belgaum, Karnataka in 2015. He also worked as an intern under Dr.Shivananju BN, former Research Scholar, Department of Instrumentation, IISc, Bangalore. His research interests include Embedded Systems, Wireless Communication, BAN and Medical Electronics. He is also the Founder and Managing Director of the corporate company Konigtronics Private Limited. He has guided over a hundred students/interns/professionals in their research work and projects. He is also the co-author of many International Research Papers. He is currently pursuing his MBA in e-Business and PG Diploma in International Business. Presently Konigtronics Private Limited has extended its services in the field of Software Engineering and Webpage Designing. Konigtronics also conducts technical and non-technical workshops on various topics.