International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-06, Issue-07, OCT 2020

Loan Fraud Detection – Using ML

[*]Angel Parmar, [#]Kelvin Parmar, [$]Premkumar Balani

[*,#]UG Student, [$]Assistant professor, G.H. Patel College of Engineering and Technology, V.V. Nagar, Gujarat, India. [*]angelparmar71@gmail.com, [#]kelvinparmar711@gmail.com, [$]prembalani@gcet.ac.in

Abstract- With the enhancement in the technology and availability of financial funds in banking sector, lots of people are applying for bank loans but the bank has its limited assets which it has to grant to limited people only, plus it needs to check in-depth about if the borrower will repay the loan in determined time-period properly. Thus, finding out to whom the loan can be granted and what would be a safer option for the bank is a typical process. So, in this paper we try to reduce this risk factor behind selecting the safe borrower to save lots of bank efforts and assets. This is done by mining the Big Data of the previous records of the people to whom the loan was granted and on the basis of these records/experiences the model is trained using the machine learning techniques which give the most accurate result. The main objective of this project is to form a prediction of whether assigning the loan to a particular person will be safe or not for the bank.

Keywords- Bank Frauds, Machine-learning, prediction, loan fraud detection

I. INTRODUCTION

This project forays into the sphere of loan sanctioning with an aim to expedite the conventional process of lending and borrowing deployed by the banking agencies as well as to reduce the red tape. A bank is a "financial institution that accepts deposits from the public and creates credit" which means that one of the two main responsibilities of a bank is to lend money to commercial and corporate clients. [1] Lenders' decision whether to issue credit is highly influenced by the borrower's credit score and credit report provided by Credit Information Bureau India Ltd (in India). The process of issuing credit has increased in complexity over the years due to the different possibilities, market demands and clients' circumstances. This makes the bank a highly regulated entity which is expected to act responsively when giving a loan. Together with the ever-increasing user demand for speed and personalization, banks and all credit issuers are turning to the power of machine learning algorithms. [2] This study explores a number of areas that reflect the rapidly changing socio-political and technical landscape in relation to the subject. Furthermore, also examines the implications of Machine Learning for management of the lending and the borrowing process. The primary objectives of the study are to:

1. Provide an introduction to Machine Learning technology and its core social value proposition.
2. Identify and engage with the key issues that are influencing policymakers and other key stakeholders in considering the use of Machine Learning as a value-added proposition within an education landscape.
3. Explain how banking institutions can use the technology as a transparent, trusted system for securing, analyzing, and predicting the pertinent data.
4. Identify a set of clear opportunities and challenges for the take-up of Machine Learning in banking institutions. The study also engages with issues relating to interoperability of technology and how the centralized nature of accreditation and the decentralized nature of the Machine Learning could be reconciled.

II. LITERATURE REVIEW

Paper 1: A machine learning approach for predicting bank credit worthiness

Abstract- This paper aspires to delve into the probe of determining the ability of machines to "learn" and do predictive analysis and its wide range of application areas. For instance, banks and financial institutions are sometimes faced with the challenge of what risk factors to consider when advancing credit/loans to customers. For several features/attributes of the customers are normally taken into consideration, but most of these features have little predictive effect on the credit worthiness or otherwise of the customer. Furthermore, a robust and effective automated bank credit risk score that can aid in the prediction of customer credit worthiness very accurately is still a major challenge facing many banks.

Summary- In this paper, author examines a real bank credit data and conduct several machine learning algorithms on the data for comparative analysis and to choose which algorithms are the best fit for learning bank credit data. The algorithms gave over 80% accuracy in prediction. Furthermore, the most important features that determine

 DOI : 10.35291/2454-9150.2020.0655

whether a customer will default or otherwise in paying his/her credit the next month are extracted from a total of 23 features. They, then applied these most important features on some selected machine learning algorithms and compare their predictive accuracy with the other algorithms that used all the 23 features. The results show no significant divergence, signifying that these features can accurately determine the credit worthiness of the customers. Finally, we formulate a predictive model using the most important features to predict the credit worthiness of a given customer. [3]

Paper 2: Credit risk analysis using machine and deep learning models

Abstract- In this work, we build binary classifiers based on machine and deep learning models on real data in predicting loan default probability. They then used in the modeling process to test the stability of binary classifiers by comparing their performance on separate data. The novelty of this paper lies in addressing some specific questions we encounter when we want to use Big Data and algorithms. We focus mainly on the questions related to the use of the algorithms to solve or attain an objective.

Summary- This exercise exhibits three key points: the necessity to use (i) several models, (ii) several sets of data (making them change to quantify their impact) and, lastly, (iii) several criteria. We observe that the choices of the features are determinant as are the criteria: they permit one to show the difficulty of getting a unique or "exact" answer and the necessity of analyzing all the results in detail or introducing new information before making any decision. They also cited the several performance measures to compare the performance of the models including AUC, Gini, RMSE and Akaike information criterion (AIC); in addition to different metrics like the F-score, the recall and the precision. In this paper, we will mainly present results on the AUC and RMSE criteria, although the results of the other metrics can be made available. [4]

Paper 3: Exploratory data analysis for loan prediction based on nature of the clients

Abstract- The main purpose of the paper was to classify and analyse the nature of the loan applicants. From a proper analysis of data set and constraints of the banking sector, seven different graphs were generated and visualized. From the graphs, many conclusions have been made and information were inferred such as short-term loan was preferred by majority of the loan applicants and the clients majorly apply loan for debt consolidation. This paperwork can be extended to higher level in future. Predictive model for loans that uses machine learning algorithms, where the results from each graph of the paper can be taken as individual criteria for the machine learning algorithm.

Summary- Whenever the bank makes decision to give loan to any customers then it automatically exposes itself to several financial risks. It is necessary for the bank to be aware of the clients applying for the loan. This problem motivates to do an EDA on the given dataset and thus analyzing the nature of the customer. The dataset that uses EDA undergoes the process of normalization, missing value treatment, choosing essential columns using filtering, deriving new columns, identifying the target variables and visualizing the data in the graphical format. Python is used for easy and efficient processing of data. This paper used the pandas library available in Python to process and extract information from the given dataset. The processed data is converted into appropriate graphs for better visualization of the results and for better understanding. For obtaining the graph Matplot library is used. [5]

Paper 4: A survey on ensemble model for loan prediction

Abstract- In this paper researchers analyze the data set using data mining technique. Data mining procedure provides a great vision in loan prediction systems, since this will promptly distinguish the customers who are able to repay the loan amount within a period. Algorithms like "J48 algorithm", "Bayes net", Naive Bayes" are used. On applying these algorithms to the datasets, it was shown that "J48 algorithm" has high accuracy (correct percent) of 78.3784% which provides the banker to decide whether the loan can be given to the costumer or not.

Summary- They used the "Tree model", "Random forest", "SVM model" and combined the above three models as Ensemble model. A prototype has been discussed in this paper so that the banking sectors can agree/reject the loan request from their customers. The main method used is real coded genetic algorithms. The combined algorithms from the ensemble model, loan prediction can be done in an easier way. It is found that tree algorithm provides high accuracy of 81.2%. [6]

Paper 5: Assessing loan risks: a data mining case study

Abstract- The paper used predictive model technique and descriptive model technique to predict the loan approval in banks. In predictive model technique, classification and regression were used and in descriptive model technique clustering and association were used.

Summary- Classifiers also implement several algorithms like naive Bayes, KNN algorithms of R language and regressors implements several algorithms like decision trees, neural networks, etc., To undergo this prediction analysis, out of all these algorithms, naive Bayes produces a most accurate classifier and the algorithms like decision tree, neural network, K-NN algorithms will be more accurate regressors. The main goal of the paper was to predict the loan classification based on the type of loan, loan applicant and the assets (property) that loan applicant holds. It was found that the decision tree algorithm gave an improved accuracy of almost 85% on doing the analysis. [7]

III. PROBLEM STATEMENT

Amidst the technological revolution and fast-paced life of 21st century, businesses and organizations are taking a leap

 DOI : 10.35291/2454-9150.2020.0655

in terms of development and marketing. Entrepreneurship and small businesses are attracting people's attention like never before. Thus, the demand for loan and credit has risen and banks and other financial organizations must catch up with this demand. But, with the help of latest technical facilities, faking of documents, fraudulent entries of finance, and bank frauds have touched a peak. Banks have to be cautious while granting loans and need to verify every single detail accurately in order to escape from such frauds. In most cases, lenders turn to established methods for determining credit worthiness, which are based on five C's of credit systems. These stand for, Character -a borrower's reputation or track record for repaying debts, credit history (record of consumer's ability to repay a loan). Capacity- comparing one's income against recurring debts and assessing the borrower's debt-to-income (DTI) ratio. Recurring debts are any payments required to be made on continuing basis (e.g. childcare support, loan). The debt-to-income ration determines whether enough income remains, after accounting for recurring debts, for the borrower to comfortably pay repay the loan. Capital-the amount of money a borrower puts towards potential investment. This practice is often applied to mortgages where some lenders require a down payment of 2%-3% to assure the seriousness of the borrower. Collateral-an asset the that borrow provides to secure the loan. This is often a car for car loans, home for mortgages or even bank saving deposits for personal loans. Conditions- this includes everything from the conditions of the loan (e.g. interested rate) to conditions, outside of borrowers control (e.g. the state of the economy). [8] These methods are somewhat reliable but at the same time they are highly time-consuming, require a lot of manual paper-work and are less accurate. To cope up with the never-ending demand of funds, better user experience and to expedite the loan granting procedure, the use of machine learning can be really beneficial for the organizations to lower the risk and improve the accuracy of the process. Tasks like managing customers' data and strategizing services according to their needs can easily be done with the help of data analysis techniques. [9]

IV. PROPOSED SYSTEM

The proposed machine-learning model is designed to be able to accurately predict whether the borrower would repay the credit amount on time or he would default it. In order to accurately predict this result, model was first trained on different classification techniques like Decision tree, Random forest etc. The training and testing datasets have detailed information about previous customers' credit details, personal information, and other essential data. Unnecessary columns of data have been removed from the datasets to improve the precision. The model is built on a tree-based classification method called as LightGBM, which provides higher accuracy than other traditional classifiers, has faster training and higher accuracy and it can also run on GPU. [10]

With the help of detailed information from the data dictionary(containing 41 columns and 233154 rows of training data), the machine learning model is built to predict result based on different important factors like bank performance, CNS score, disbursal amount, average acct. age, etc. Feature selection and data cleaning techniques are performed on training data first, to be able to use it in classification models efficiently i.e., null point removing, column dropping, feature scaling, etc. The proposed model is built on an open-source classification technique by Microsoft, which provides parallel computing and better accuracy. Also, the model is tested on popular classification techniques for accuracy and computing comparison, but the boosting method (LightGBM) has managed to provide better results without having any issues of overfitting. With the help of the predicted results, banks and lender organizations can improve their decision-making and mitigate risks in financial transactions.

V. METHODS

Loan fraud detection using ML model works on basic methodologies of classification and provides output prediction as a result. The main methods are as follows:

1. Data collection (Data dictionary)
2. Data pre-processing
3. Dividing data in train and test sets
4. Counting for the number of values
5. Resampling of datasets
6. Using lightGBM classifier to form predictions

Data collection:

The data dictionary consists of 40 features of related personal information of previous customers to the bank with more than 233154 rows of data i.e., date of birth, disbursed amount, employment type, etc.

Variable Name	Description	
UniqueID	Identifier for customers	
loan_default	Payment default in the first EMI on due date	
disbursed_amount	Amount of Loan disbursed	
asset_cost	Cost of the Asset	
ltv	Loan to Value of the asset	
branch_id	Branch where the loan was disbursed	
supplier_id	Vehicle Dealer where the loan was disbursed	
manufacturer_id	Vehicle manufacturer(Hero, Honda, TVS etc.)	
Current_pincode	Current pincode of the customer	
Date.of.Birth	Date of birth of the customer	
Employment.Type	Employment Type of the customer (Salaried/Self Employed)	
DisbursalDate	Date of disbursement	
State_ID	State of disbursement	
Employee_code_ID	Employee of the organization who logged the disbursement	
MobileNo_Avl_Flag	if Mobile no. was shared by the customer then flagged as 1	
Aadhar_flag	if aadhar was shared by the customer then flagged as 1	
PAN_flag	if pan was shared by the customer then flagged as 1	
VoterID_flag	if voter was shared by the customer then flagged as 1	
Driving_flag	if DL was shared by the customer then flagged as 1	
Passport_flag	if passport was shared by the customer then flagged as 1	
PERFORM_CNS.SCORE	Bureau Score	
PERFORM_CNS.SCORE.DESCRIPTION	Bureau score description	
PRI.NO.OF.ACCTS	count of total loans taken by the customer at the time of disbursement	
PRI.ACTIVE.ACCTS	count of active loans taken by the customer at the time of disbursement	
PRI.OVERDUE.ACCTS	count of default accounts at the time of disbursement	Primary accounts are those which the customer has taken for his personal use
PRI.CURRENT.BALANCE	total Principal outstanding amount of the active loans at the time of disbursement	
PRI.SANCTIONED.AMOUNT	total amount that was sanctioned for all the loans at the time of disbursement	
PRI.DISBURSED.AMOUNT	total amount that was disbursed for all the loans at the time of disbursement	
SEC.NO.OF.ACCTS	count of total loans taken by the customer at the time of disbursement	
SEC.ACTIVE.ACCTS	count of active loans taken by the customer at the time of disbursement	Secondary accounts are those which the customer act as a co-applicant or gaurantor
SEC.OVERDUE.ACCTS	count of default accounts at the time of disbursement	
SEC.CURRENT.BALANCE	total Principal outstanding amount of the active loans at the time of disbursement	
SEC.SANCTIONED.AMOUNT	total amount that was sanctioned for all the loans at the time of disbursement	
SEC.DISBURSED.AMOUNT	total amount that was disbursed for all the loans at the time of disbursement	
PRIMARY.INSTAL.AMT	EMI Amount of the primary loan	
SEC.INSTAL.AMT	EMI Amount of the secondary loan	
NEW.ACCTS.IN.LAST.SIX.MONTHS	New loans taken by the customer in last 6 months before the disbursment	
DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS	Loans defaulted in the last 6 months	
AVERAGE.ACCT.AGE	Average loan tenure	
CREDIT.HISTORY.LENGTH	Time since first loan	
NO.OF_INQUIRIES	Enquries done by the customer for loans	

Figure 1 Data dictionary

Data pre-processing:

The data provided in the data dictionary was raw and ambiguous, which needed to be cleaned first in order to be

 DOI : 10.35291/2454-9150.2020.0655

used as a training data set for the model. The employment type column was in categorical form which then was converted into numerical form using label encoding. Unnecessary columns like date of birth, unique_id and disbursal date were removed from the datasets. Null point values were given 'Not_provided' value and then converted into numeric form. Dates and years were converted completely into numeric form in order for model to work properly.

Dividing data in train and test-sets:

After pre-processing, the data is divided into two sets, 80% of the data used for training purpose of the model, while the remaining data was used for testing.

```
In [11]: train.shape, test.shape

Out[11]: ((233154, 41), (112392, 40))
```

Figure 2 Division of train and test datasets

Counting for the number of values:

Using the value.counts() method of pandas, the approx. number of which feature is seen the most in both training and testing datasets is visualized, which helps in deciding the most impactful parameters for the model.

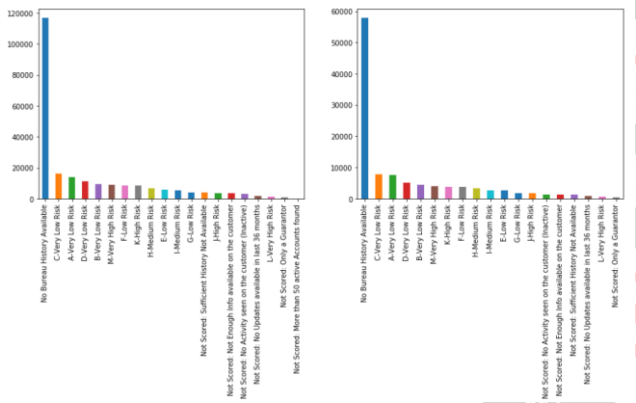

Figure 3 Visualization of column occurrence

Using LightGBM for forming predictions:

After resampling of data, the training datasets are used in five classifiers (all training on five different training datasets) and the method is run to form predictions. Furthermore, the formed predictions are compared to actual records from the test dataset, to calculate accuracy of these classifiers.

```
%%time
from lightgbm import LGBMClassifier
clf1 = LGBMClassifier(random_state=25)
clf1.fit(train, y)

# clf2 = LGBMClassifier(random_state=25, class_weight={0: 0.6386298893393229, 1: 1.5})
# clf2.fit(train, y)

# clf3 = LGBMClassifier(random_state=25)
# clf3.fit(train_smote, y_smote)

# clf4 = LGBMClassifier( random_state=25)
# clf4.fit(train_downsample, y_downsample)

# clf5 = LGBMClassifier( random_state=25)
# clf5.fit(train_upsample, y_upsample)

CPU times: user 11.1 s, sys: 452 ms, total: 11.5 s
Wall time: 3.42 s
```

Figure 4 creating classifiers

VI. RESULTS

Once the datasets are passed to classifiers and the model has produced outputs, the results of these classifiers are compared to check for the best accuracy. The classifier with the training dataset of down-sampled data showed highest accuracy of 80.29%, while the least accurate classifier was trained using the random dataset, showing only 32.45% of accuracy.

```
Best estimator:
LGBMClassifier(boosting_type='gbdt', class_weight=None, colsample_bytree=1.0,
        importance_type='split', learning_rate=0.1, max_depth=13,
        min_child_samples=20, min_child_weight=0.001, min_split_gain=0.0,
        n_estimators=400, n_jobs=-1, num_leaves=31, objective=None,
        random_state=51, reg_alpha=0.0, reg_lambda=0.0, silent=True,
        subsample=1.0, subsample_for_bin=200000, subsample_freq=0)

Best score:
0.8029181291907113

Best parameters:
{'learning_rate': 0.1, 'max_depth': 13, 'n_estimators': 400}
```

Figure 5 Finding out the most accurate classifier

VII. SYSTEM ARCHITECTURE

The proposed system consists of five main components.

1. User bank's details
2. Inserting borrower's details
3. The bank officials' user details
4. Passing the borrower's details into the model for prediction
5. Visualizing the result

The figures below show the flow of how the proposed model can be used in real-life banking systems and be useful for banks and financial lenders.

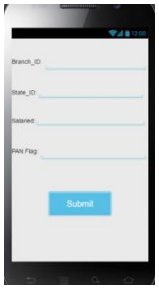

Figure 6 User bank's details

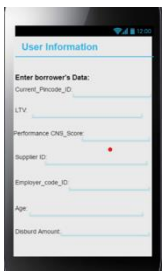

Figure 7 Inserting borrower's details

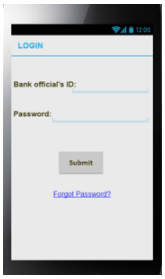

Figure 8 The bank official's user details

 DOI : 10.35291/2454-9150.2020.0655

Figure 9 Sending the borrower's details into model for prediction

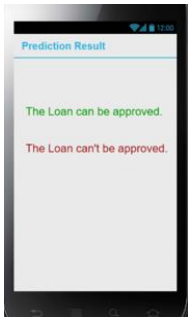

Figure 10 Visualizing the result

VIII. ADVANTAGES

1. With the help of improved accuracy and decision-making, customer satisfaction can be increased.
2. Due to less manual-work and paperwork, workload can be decreased.
3. Banks can achieve greater success in terms of data analysis and strategizing the services with the help of such advanced techniques.
4. More accurate results can prevent banks and lenders from falling into financial frauds and traps.
5. With the use of advanced technologies and inventions, banks and organizations can out-perform its competitors and gain more profits.
6. Use of the model in banks and other lender organization can promote the use of technology and innovation in financial sector, which can further help in boosting up a nation's economy.
7. The use of the proposed model can help in bringing transparency in financial transactions.

IX. LIMITATIONS AND FUTURE WORK

As of now, the model has the accuracy of approx. 80%, which is not highly reliable in huge financial transactions. Even though the model provides a solution for manual work in banking systems, the results need to be verified by higher authorities if there is any technical loophole.

In order to make the model more reliable and effective, we have determined to train the data on different latest classifiers and compare the accuracy.

We are also willing to make the interface of the mobile application more user-friendly and effective.

User reviews and feedback will be included in the application so that we know where to improve and what the system is lacking by real-world experience.

X. CONCLUSION

The technological revolution has changed and improved every possible field and the financial sector is no exception. In the last decade, the demand of automation in banking sector has grown rapidly and the area of machine learning and artificial intelligence has become broader. Machine learning is thought to have a significant impact on financial sector and bank loan procedures in upcoming future. Machine learning can also provide solutions for marketing, data analysis and business strategies with its advanced techniques and generate higher revenues for organizations. Use of machine learning and automation in banking sector is said to bring transparency in financial transactions and further, help boosting the economy of a nation. Machine learning has already established its roots in the financial sectors and is being warmly welcomed by the banks, other lender organizations and the borrowers as well.

XI. REFERENCES

[1] Manyizzy, "Economic functions," The university of Nairobi, Nairobi, 2014.

[2] D. Caicedo, "How AI and machine learning are improving the banking experience," 2019.

[3] E. Y. B. G. E. W. Regina Esi Turkson, "A machine learning approach for predicting bank credit worthiness," *Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR)*, vol. 3, 2016.

[4] D. G. B. H. Peter Martey Addo, "Credit Risk Analysis Using Machine and Deep Learning Models," 9 April 2018.

[5] V. J. S. S. X.Francis Jency, "An Exploratory Data Analysis for Loan Prediction," *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 7, no. 4S, 2018.

[6] R. K. Anchal Goyal, "A Survey on Ensemble Model For Loan Prediction," *International Journal of Advance Research and Innovative Ideas in Education*, 2016.

[7] R. Gerritsen, "Assessing loan studies," *IEEE*, 1999.

[8] T. Segal, "Understanding the Five Cs of Credit," 6 April 2019.

[9] R. Chunprina, "Machine Learning in Banking," 15 September 2020. [Online]. Available: https://spd.group/machine-learning/machine-learning-in-banking. [Accessed 30 October 2020].

[10] "Welcome to LightGBM's documentation," Github, 23 February 2020. [Online]. Available: https://lightgbm.readthedocs.io/en/latest/. [Accessed October 2020].

 DOI : 10.35291/2454-9150.2020.0655