

CREDIT FRAUD DETECTION

HOW MACHINE LEARNING SYSTEMS HELP REVEAL CREDIT SCAMS IN BANKS

**Kriti Palod- PGDM6/1927**

**Ashwani Gupta – PGDM6/1916**

**Shobhit Patni- PGDM6/1947**

**Ambuj Joshi- PGDM6**

**Kartik- PGDM6**

**Srishti Jain- CMBA2Y3/1944**

**ABSTRACT**

The finance and banking is very important sector in our present day generation, where almost every human has to deal with bank either physically or online. The productivity and profitability of both public and private sector has tremendously increased because of banking information system. Nowadays most of E-commerce application system transactions are done through credit card and online net banking. These systems are vulnerable with new attacks and techniques at alarming rate. Fraud detection in banking is one of the vital aspects as finance is major sector in our life. As data is increasing in terms of Peta Bytes (PB) and to improve the performance of analytical server in model building, we have interface analytical framework with python which can read data efficiently and give to analytical server for fraud prediction. In this paper we have discussed a Big data analytical framework to process large volume of data and implemented various machine learning algorithms for fraud detection and observed their performance on benchmark dataset to detect frauds on real time basis there by giving low risk and high customer satisfaction.

**KEYWORDS**

Hadoop, SAS, PB, Logistic Regression, Decision Tree, Big data analytics, Machine learning.

**INTRODUCTION**

Payments are the most digitalized part of the financial industry, which makes them particularly vulnerable to digital fraudulent activities. The rise of mobile payments and the competition for the best customer experience nudge banks to reduce the number of verification stages. This leads to lower efficiency of the rule-based approach. So, banks and payment companies switch to data analytics, machine learning, and AI-driven methods.

**BUSINESS PROBLEM**

Studies have shown that fraudsters, big and small, are able to take undue advantage of a number of well-documented weaknesses in the system. The central bank has an early warning signals system but, as had happened in the Nirav Modi case, PSBs do not always take advantage of it. Former RBI governor Urjit Patel made a presentation at Stanford University in June that showed most of the frauds are related to loans and occur due to poor operational risk management and ineffective internal audits at state-owned banks. These banks apply little risk analysis or due diligence.

**Domain of Business**

Bank fraud is the use of potentially illegal means to obtain money, assets, or other property owned or held by a financial institution, or to obtain money from depositors by fraudulently posing as a bank or other financial institution.

**Impact of credit fraud**

* Quantitative impacts

Public sector banks (PSBs) reported frauds of over ₹95,760 crore from April to September this year. According to the Reserve Bank of India’s latest annual report, all banks, including PSBs, reported frauds involving losses of ₹71,542.93 crore over the 12-month period of FY19. RBI data shows that the bulk of the frauds relate to loans and take place at PSBs. The incidence and cost of frauds is increasing year after year, posing threats to the financial stability and eroding the credibility of PSBs, auditors, credit rating agencies and the regulator RBI, as well as the trust of savers and depositors.

* Qualitative Impacts

The study found that big loan advance frauds happen as bank officials collude with borrowers and sometimes even with officials of third parties such as advocates and chartered accountants. Post loan sanction, the monitoring is weaker than at private banks due to lack of expertise and modern tech resources. Officers retire before they can be booked for fraud. Weak selection process and lower pay than at private banks are among key reasons. PSB staff are not offered appropriate incentives to prevent or detect frauds early.

**Positive effects of Solving the problem**

* **Better and speedier realization from Non-Performing Assets**: Provides a meeting ground to the lenders, borrowers and the prospective investors to speedily address the issue of Non-Performing Assets.
* **Better efficiencies in the management of NPAs**: While traditionally we used to avail the services of news media to create awareness about the Non Performing Asset, NPAsource.com brings into play the unlimited power of technology for the lender to undergo due-diligence.
* **Robust data collection for use by all lenders across the country**: Even lenders will find it very easy to upload data on the website and take huge advantage of the portal to release the value of their non-performing assets easily and effectively.
* **Attracting prospective investors / buyers from across the globe**: Geographical boundaries will not restrict the resolving of NPAs. Individuals from across the world will be able to access the portal and avail the benefits.
* **List of the NPA and their assets at a single location enabling lenders to take informed decision**: With maximum information available related to NPA, the prospective investor can take informed decision towards the NPA.
* **Revival of a sick unit/ industry by infusing required capital through prospective investors**: The portal will help greatly in unlocking huge values and once again bringing into action the units long lying sick.
* **Help from facilitators**: Investors, borrowers and lenders can take the service of facilitators such as Legal Advisors, Financial Advisors, Tax Advisors, Project Consultants & Valuation Experts thru the portal.

**Data identification and gathering**

**Data identification** refers to records that link two or more separately recorded pieces of information about the same individual or entity.

We identified various variables which were directly or indirectly connected to the contract type. We also made sure to protect the rights of customers and didn’t revealed name or any identification number.

**Data collection** is the process of gathering and measuring information on targeted variables in an established system, which then enables one to answer relevant questions and evaluate outcomes. The goal for all data collection is to capture quality evidence that allows analysis to lead to the formulation of convincing and credible answers to the questions that have been posed.

* **Gender**: Is collected and presented by sex as a primary and overall classification; Reflects gender issues; Is based on concepts and definitions that adequately reflect the diversity of women and men and capture all aspects of their lives; Is developed through collection methods that take into account stereotypes and social and cultural factors that may induce gender bias in the data.
* **Flag own car**: Motor vehicles generate a wealth of data. While most vehicle-generated data are of a technical nature and only used locally within your car, certain types of data can be used to provide services. Connected vehicles allow sharing of some of these data with third parties.
* **Flag own realty**: Realty is real estate; a piece of real property; land.
* **Family status**: This basically shows the personal status of each individual in relation to their marital status.
* **Children**: Basically, tells whether a family have Children or not. If they have Children, might be the reason for more loans with respect to their education and other expenses.
* **House Type**: This is a primary type of data which tells whether the person owns a house personally or is using a rental space.
* **Income**: Income is an important risk measure used by lenders for underwriting loans and a useful economic indicator of an area's standard of living. Income is money (or some equivalent value) that an individual or business receives, usually in exchange for providing a good or service or through investing capital.

**Reason for selection**

•**Gender**: In case of the variable gender, we cannot identify the reason as both the genders are equally capable of paying back the loan.

•**Flag own car**: This variable is used to identify the financial soundness of a person applying for loan if a person is having a car will definitely have a sound income and therefore, we can assume that he must be in a good position to repay the loan as compare to the person who doesn’t own a car. In general this is to check the financial soundness of the client.

•**Family status**: So, the reason for selection in case of family status will be single or married so we can say that if a person is married there are some chances that his/her income might get deviated in other expenses as well and for a married person to payoff loans becomes a hardship as compared to the person who is single.

•**Children**: So, in this variable set named children we can state the reason for selection is if a family consisting of children then its definite that the income of the family will be diverted towards those expenses so there might be a chance that a family with no children may have the high repaying capacity when compared to a family consisting of children.

•**House Type**: In the variable house type there are two factors, either the house is self-owned or its rented. So, if the house is rented, some proportion of the income of a particular household will go in an expense named “Rent” so, that the particular household will be left with low income after adjusting rent and there might be chances that that household will get difficulties in repaying the loan as compared to the families having self-owned houses.

•**Income**: It is the most simple and obvious variable where we can clearly say that there is indirect relationship between income of a person and the default risk as more the income there will be less chances of a person becoming defaulter & vice versa.

**Data type of every variable.**

So, there are only 2 types of data used in this particular type of data set which are numeric data and the second one is categorical data.

The following is the classification of variables in these 2 different types of data sets.

**Gender**: Categorical Data

**Flag own car:** Categorical Data

**Flag own reality**: Categorical Data

**Family status**: Categorical Data

**Children**: Numeric Data

**House Type**: Categorical Data

**Income**: Numeric Data

**SOLUTION**

There is little deterrence for fraudsters as conviction rates are low due to the lack of specialized financial sleuths. PSBs should set up an internal rating agency for stringent evaluation of big-ticket projects before sanctioning loans. Banks must evaluate projects on the basis of the business model and not get influenced by the brand name or creditworthiness of the parent firm. Strict punitive steps for bank staff and others who collude with fraudsters can help.

**ANALYSIS**

**Model Building**

**Q. Is there is a need for cleaning the dataset? Inputting the missing values?**

Ans. No, cleaning of dataset and inputting of missing values was not required.

**DATA VARIABLES-**

Following data variables are used in this dataset for building the model:-

**Independent Variables-** Gender, Candidate own car (Flag own car), candidate own property (Flag own realty), No. of children (Cnt\_children), relationship status (Name\_family\_status), house ownership type (House\_type), loan granted (Amt\_credit),

**Dependent Variable-** Loan default (Credit\_fraud)

**Training and Test Sets: Splitting Data**

1. **Training Set-** A subset to train a model. **70% of data is used for training**. This is used to build up our prediction algorithm and to adjust the weights on the neural network. Our algorithm tries to tune itself to the quirks of the training data sets.
2. **Test Set-** A subset to test the trained model. **30% of data is used for test**. It helps to apply our chosen prediction algorithm on our test set in order to see how it's going to perform to have an idea about our algorithm's performance on unseen data. This data set is used only for testing the final solution in order to confirm the actual predictive power of the network.

**MODEL SELECTION-**

1. **Logistic Regression-**Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability. The hypothesis of logistic regression tends it to limit the cost function between 0 and 1.

**Reason for selecting Logistic Regression-**

* Useful for binary data- As the data set was of binary nature i.e 0 and 1 so, logistic regression helps to explain the relationship between one dependent binary variable and one or more nominal, ordinal variables.
* **Overfitting-** logistic regression analysis, another **important consideration is the model fit**. Adding independent variables to a logistic regression model will always increase the amount of variance explained in the log odds
* **Reporting the R2-** As the dataset was binary. Numerous pseudo-R2 values have been developed for binary logistic regression.

1. **Decision Tree Classifier-** Decision tree is used when the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split. As dataset was of binary nature (0 and 1) so Decision tree classifier is used as it is useful when outcome is (0 and 1/ Yes and No).

**Reason for selecting Decision Tree Classifier-**

* Useful for binary data- As dataset was of binary nature ( 0 and 1) so Decision tree classifier is used as it is useful when outcome is (0 and 1/ Yes and No).
* Simple to understand- As it follows the same process which a human follow while making any decision in real-life.
* Solving critical problems- It can be very useful for solving decision-related problems.
* Outcomes for a problem- It helps to think about all the possible outcomes for a problem.
* Less data cleaning is required- There is less requirement of data cleaning compared to other algorithms.

**DEVELOP THE MODEL: CODE AND ASSOCIATED DIAGRAMS-**

**GoogleColabLink-https://colab.research.google.com/drive/1SQNXogVMymA\_8LssAqIefrBb36f5X66f?usp=sharing**

**Following images and diagrams of the model-**

1. **IMPORTING PANDAS AND INSERTING DATASET**

Table

Description automatically generated

1. **CONVERTING DATA INTO 0 AND 1 AS WELL IMPORTING SEABORN FOR VISUALIZATION**

Graphical user interface, application, table

Description automatically generated

Line chart

Description automatically generated

Graphical user interface, line chart

Description automatically generated

1. **IMPORTING VARIOUS PACKAGES TO PERFORM LOGISTIC REGRESSION, STATISTICS, CONFUSION MATRIX AND TRAINING AND TEST SET**

Graphical user interface, text, application, email

Description automatically generated

1. **IMPORTING CLASSIFICATION REPORT AND REGRESSION RESULTS LIKE – R-SQUARED**

Graphical user interface, text, application, email

Description automatically generated

A picture containing text, receipt, screenshot

Description automatically generated

1. **GETTING ACCURACY RESULT AND PLOTING HEAT MAP**

Graphical user interface

Description automatically generated

1. **IMPORTING SKLEARN METRICS AND ROC CURVE**

Graphical user interface, text, application

Description automatically generated

Chart, line chart

Description automatically generated

1. **IMPORTING DECISION CLASSIFIER**

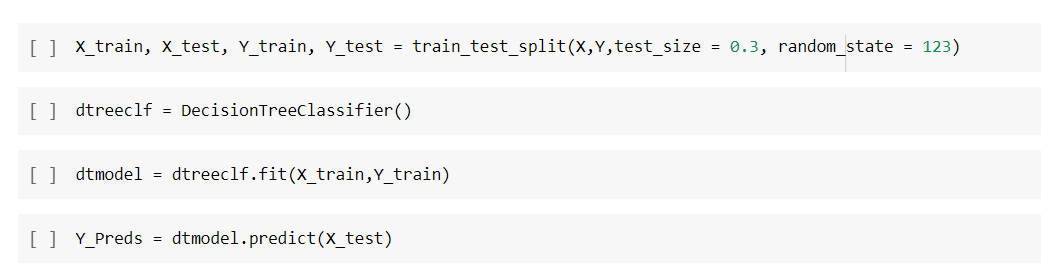
Graphical user interface, text, application

Description automatically generated

Table

Description automatically generated

1. **GETTING ACCURACY SCORE BY DECISION TREE CLASSIFIER- ACCURACY- 1.0**



Graphical user interface, text, application

Description automatically generated

1. **GETTING DECISION TREE CHART**

Timeline

Description automatically generated

**EVALUATION TECHNIQUES-**

1. **CONFUSION MATRIX-** It helps in evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. This gives us a holistic view of how well our classification model is performing and what kinds of errors it is making.

Confusion matrix gave **accuracy of model i.e 0.60**

b) **R-squared**- is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

A picture containing text

Description automatically generated

1. **Receiver Operating Characteristics (ROC) Curve-** Measuring the area under the ROC curve is also a very useful method for evaluating a model. ROC is the ratio of True Positive Rate (TPR) and False Positive Rate (FPR).

Graphical user interface, text, application

Description automatically generatedChart, line chart

Description automatically generated

**RECOMMENDATION**

* As discussed above our decision tree classifier had the accuracy of 1.0 and as for confusion matrix it was 0.60, which means it is highly recommended for banks to make use of this model.
* Applying machine learning to fraud detection enables financial firms to identify genuine transactions versus fraudulent transactions in real time, and with greater accuracy. AI can help reduce false positives that could occur with traditional fraud detection methods.
* While organizations may not be in a position to move to advanced data analytics immediately, they should begin examining their existing data, identifying data requirements, and developing the expertise necessary to begin as soon as possible.

**CONCLUSION**

The credit card fraud detection is becoming important topic of research, as different types of attacks are increasing at an alarming rate. In this paper we have proposed a robust framework to process large volume of data, the functionality of framework can be extended to extract real time data from different desperate sources. The extracted data is then used to build strong analytical model. To improve the analytical accuracy of fraud prediction, we have implemented three different analytical techniques. These analytical models are run on credit card dataset and accuracy of analytical model is evaluated with help of confusion matrix. Among the three models, random forest decision tree performs best in terms of accuracy, precision and recall. The only problem with random forest is over fitting of tree in memory as data increases. The future scope of this work is to remove over fitting problem of decision tree and to detect real time fraud transaction for high streaming real time data