IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Machine Learning Classification and Deep Learning based Credit Prediction and Risk analysis on Loans Data.

Ashutosh Kushwaha[1],
Department of Computer Engineering,
Wagholi, Pune

Aditya Kardile[1],
Department of Computer Engineering,
Wagholi, Pune

Mohd. Zaib Nawab[1]
Department of Computer Engineering
Wagholi, Pune

Dr. Vinod Wadne
Department of Computer Engineering
Wagholi, Pune

Abstract: - The improvement within the banking sector several individuals area unit applying for bank loans however the bank has its restricted assets that it's to grant to restricted individuals solely, thus sorting out to whom the loan are often granted which can able to be a safer choice for the bank is a typical method. Thus during this paper we tend to try and cut back this risk issue behind choosing the safe person thus on save several bank efforts and assets. This can be done by mining the large knowledge of the previous records of the individuals to whom the loan was granted before and on the idea of those records experiences the machine was trained mistreatment the machine learning model that offer the foremost correct result. the most objective of this paper is to predict whether or not assignment the loan to explicit person are going to be safe or not. This paper is split into four sections(i)Data assortment (ii) Comparison of machine learning models on collected knowledge (iii) coaching of system on most promising model(iv) Testing.

Keywords: **Loan Prediction, Big data, Machine Learning, Logistic Regression, SVM, Decision Tree, Naïve Bayes, etc.**

INTRODUCTION

Distribution of the loans is that the core business is a part of nearly every bank. most portion of the bank's assets are directly came from the profit attained from the loans distributed by the banks. The prime objective in the banking atmosphere is to invest their assets in safe hands wherever it's. these days several banks/financial corporations approve loans once a regress method of verification and validation however still there's no surety whether or not the chosen human is that the meriting right human out of all candidates. Through this method, we are able to predict whether or not that specific human is safe or not and also the whole method of validation of options is automatic by machine learning technique. The disadvantage of this model is that it emphasizes totally different weights to every issue however in real-world someday loan are often approved on the premise of single robust issue solely, that isn't doable through this method. Loan Prediction is extremely useful for workers of banks moreover as for the human conjointly. The aim of this paper is to produce fast, immediate, and straightforward thanks to opting for the meriting candidates. It will give special blessings to the bank. The Loan Prediction System will mechanically calculate the burden of every option participating in the loan process and on new check knowledge same options are processed with relevance to their associated weight. A point in time is often set for the human to visualize whether or not his loan is often sanctioned or not. Loan Prediction System permits jumping to specific applications so it often checks on a priority basis This Paper is completely for the managing authority of the Ban finance Company, the whole method of prediction is finished in private no stakeholders would be ready to alter the process. Result against specific Loan IDs is often sending to varied departments of banks so they'll take applicable action on the application. This helps all alternatives department to applied other formalities.

BACKGROUND

The most important background of machine learning algorithms their technique and mathematical formulation are outlined in this section. Analysing the Banking and credit data used these algorithms

1. Machine Learning

Machine learning algorithm can be group into two main categories, they include

1. **Supervised Learning**: supervised learning algorithm main feature is target variable and outcome variable to predict. Supervised learning technique is achieved using regression and classification problem.
2. **Unsupervised learning**: in unsupervised learning algorithm no target or outcome variable to predict. It is used for clustering entities into an different groups.

2. Classification Algorithms:

Classification algorithms work by predicting the simplest cluster to that a knowledge purpose belongs to by learning from labelled observations; it uses a group of input options for the educational method. Classification algorithms square measure sensible for grouping knowledge that square measure ne'er seen before into their numerous groupings and square measure thus extensively employed in machine learning tasks.

3. Evaluation Matrix:

1. Accuracy:

it is measured how many true positive and true negative cases is correct. Mathematically it is defined as

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

2. Sensitivity or Recall:

1. Recall: tells us how many of the actual positive cases we were able to predict correctly with our model. Mathematically it is defined as

$$\text{Recall} = \frac{TP}{TP+FN}$$

2. Specificity: tell us how many times classifier gets true negative correct value, mathematically is defined as

$$\text{Specificity} = \frac{TN}{TN+FP}$$

3. Precision:

Precision tells us how many of the correctly predicted cases actually turned out to be positive. Mathematically it is defined as,

$$\text{Precision} = \frac{TP}{TP+FP}$$

PROPOSED SYSTEM

In recent years, machine learning as become in style within the monetary sector. Machine learning is employed in credit risk analysis like increasing quantity of knowledge cc will provide higher insights compared to humans. Also, it's quicker than ancient approaches. many models of machine learning are tried on the given drawback, centering one bank, as well as linear, supply and multinomial regression by exploitation elastic internet approach. Random forest and gradient boosting algorithms have conjointly been with success tested . by experimentation the random forests work higher as a result of they're not forced to predict linear or continuous relationships. A machine learning model will yield far better insights from the info than a person's analyst. the chance context preponderantly determines that model to be used for the analysis. there's no prescriptive technique entirely tied to a group of algorithms. A report by McKinsey states that machine learning will scale back credit losses by 100 percent and credit call times by 20-25%. Also, the defrayal patterns of shoppers ar ever-changing and increasing. Machine learning helps the disposal establishments by decreasing idea. AI primarily based marking models combines' customers' credit history and also the power of massive knowledge to boost credit selections. exploitation prophetic models create it troublesome for establishments to clarify the scores to the shoppers. many varieties of risk models have a bigger level of transparency because the ancient strategies. Gradient Boosting Machines (GBM) ar prophetic models designed from sequence of many call tree sub-models. the character of GBM makes it easier than deep learning or neural network algorithms to clarify the logic behind the model's prophetic behavior. this can be as a result of GBMs ar pictured as sets of call trees which will be explained as opposition the neural networks that ar pictured as cryptic numbers that ar a lot of more durable to know. this technique predicts whether or not the loan is approved or reject.

This System refers the following things or ways.

- Data Collection
- Data Pre-processing (Data Cleaning)
- Model Selection
- Model Evaluation
- Classification Result (output)

PROPOSED ALGORITHM

The following shows the pseudo code for the proposed loan prediction method

1. Load the data
2. Determine the training and testing data
3. Data cleaning and pre-processing.

a) Fill the missing values with mean values regarding numerical values.

b) Fill the missing values with mode values regarding categorical variables.
c) Outlier treatment.
4. Apply the modelling for prediction
a) Removing the load identifier
b) Create the target variable (based on the requirement). In this approach, target variable is loan-status
c) Create a dummy variable for categorical variable (if required) and split the training and testing data for validation.
d) Apply the model: NB method, SVM method
5. Determine the accuracy followed by confusion Matrix.

IMPLEMENTATION

1. Modules:

1. Loan Dataset: Loan Dataset is incredibly helpful in our system for prediction of additional correct result. Victimisation the loan Dataset the system can mechanically predict that costumer's loan it ought to approve and that to reject. System can settle for application type as associate degree input. Even format of form ought to run as associate degree input to induce processed.
2. verify the coaching and testing data: usually , Here the system separate a dataset into a coaching set and testing set ,most of the info use for coaching ,and a smaller parts of knowledge is use for testing. when a system has been processed by victimisation the coaching set, it makes the prediction against the take a look at set.
3. Knowledge clean-up and processing: In knowledge clean-up the system observe and proper corrupt or inaccurate records from information and refers to distinctive incomplete, incorrect, inaccurate or immaterial components of the info then exchange, modifying or sleuthing the dirty or coarse knowledge. In processing the system converts knowledge from a given type to a far additional usable and desired type i.e. makes it additional meaning and informative.

2. Model Used:

SVM: during this approach, every knowledge item is aforethought in associate degree n dimensional area, wherever n represents the amount of options with every feature delineated during corresponding co- ordinates. A hyper plane is decided to tell apart the categories (possibly two) supported their options. Naïve Thomas (NB) Model: the idea for NB model is Bayes Theorem (BT), wherever events square measure reciprocally exclusive the same as rolling a die. Moreover, the BT presumes that the input options conjointly referred as predictor's square measure freelance in nature. Similarly, NB conjointly presumes that the input options square measure freelance in nature. But, this can be not possible within the realistic procedures. Since this assumption results in naïve, this rule is termed as Naïve mathematician rule. Thus, NB may be a probabilistic rule, wherever the probability is decided concerning the input options. On the opposite hand, throughout the dependent input options situation, probability is calculated double leading to improper results. Hence, for higher prediction results with relation to NB model, freelance input options square measure designated and processed. Dataset collected from Kaggle supply. The feature within the dataset embrace

MATHEMATICAL MODEL

Consider any decision problem, where for given number of inputs, decision oriented solution is available so our project is NP complete but some cases like not proper input format provided or if dataset not trained proper it's NP hard.
Let s be System:
S=I, P, O
S: is a System
I=I1, I2
P= DC, DP, DV, NBA, CL
O=RD
I1: Loan Dataset
I2: Trained Dataset.
DC: Data Cleaning D
DP: Data Processing
DV: Data Verification
NBA: Naïve Bayes Algorithm
CL: Classification
RD: Report Deliver Success
Condition: Proper features trained Dataset will give proper output Failure Condition No Trained Dataset.

BLOCK DIAGRAM

The proposed module can be divided in to different sections, machine learning, Flask, HTML, CSS, Anaconda-Jupiter notebook. Architecture used in proposed system are given below

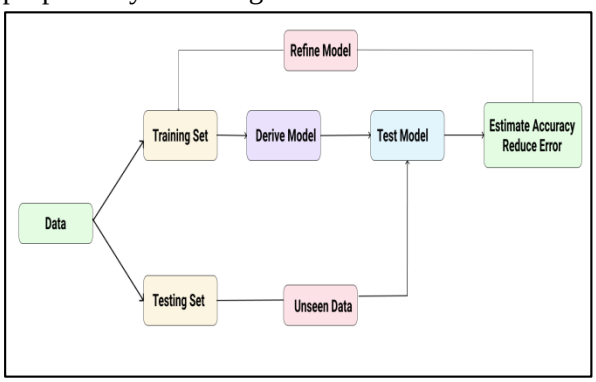

Figure: Architecture of System Design

RESULT

This section shows a comparative study of all the models that were built. These models are evaluated through accuracy, precision, and f1-score.

Below table is representing the values obtained for the various metrics from the different models. Since the f1-score and precision values of most of the models are similar excluding logistic regression model, we choose to measure the performance of the model using accuracy. It shows that the accuracy of Logistic Regression is less than other models. Also, the accuracy of Random Forest is comparatively low than Gradient Boosting, KNN Classifier and Naïve Bayes. Therefore, we can infer that Gradient Boosting and KNN Classifier are doing prediction well for our dataset.

PERFORMANCE EVALUATION OF MODELS

Sr. no	Model Used	Accuracy	Precision	F1 score
01	Logistic Regression	83.21	0.86	0.91
02	Naïve Bayes	90.46	0.94	0.95
03	Gradient Boosting	90.35	0.93	0.96
04	Random Forest	85.45	0.88	0.91
05	Knn Classifier	89.78	0.93	0.94

CONCLUSION

So here, it can be concluded with confidence that the Naïve Bayes, KNN, XGBoost, model is extremely efficient and gives a better result when compared to other models. It works correctly and fulfills all requirements of bankers. This system properly and accurately calculates the result. It predicts the loan is approved or reject to loan applicant or customer very accurately.

FUTURE WORK

Here in this paper, we have only considered loan prediction, a system could be made for predicting defaulters of other loans as well. Also, whether the non-defaulter would turn out to be a fraudster or not could be predicted.

BIBLIOGRAPHY

[1] Kacheria, A., Shivakumar, N., Sawkar, S. and Gupta, A. (2016). Loan Sanctioning Prediction System. [online] Ijsce.org.

[2]http://www.ijsce.org/wpcontent/uploads/papers/v6i4/D2904096416.pdf

[3] A. Gahlaut, Tushar, and P. K. Singh, “Prediction analysis of risky credit using Data mining classification models,” 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2017

[4] Hamid, Aboobyda & Ahmed, Tarig. (2016). “Developing Prediction Model of Loan Risk in Banks Using Data Mining”. Machine Learning and Applications: An International Journal. 3. 1-9. 10.5121/mlaij.2016.3101.

[5] Mrunal Surve, Pooja Thitme, Priya Shinde, Swati Sonawane, and Sandip Pandit. "Data mining techniques to analyze risk giving loan(bank)" Internation Journal Of Advance Research And Innovative Ideas In Education Volume 2 Issue 1 2016 Page 485-490

[6] P. Ravikumar and V. Ravi, "Bankruptcy Prediction in Banks by an Ensemble Classifier," 2006 IEEE International Conference on IndustrialTechnology, Mumbai, 2006, pp.

[7] S. Sathyadevan, D. M. S and S. G. S., "Crime analysis and prediction using data mining," 2014 First International Conference on Networks & Soft Computing (ICNSC2014), Guntur, 2014, pp.

[8] Lekha, K. and Prakasam, D. (2018). https://www.researchgate.net/publication/326147

[9] K. C. Lekha and S. Prakasam, "Data mining techniques in detecting and predicting cybercrimes in banking sector," 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS),Chennai, 2017, pp. 1639-1643.

[10] Yu Jin and Yudan Zhu, “ A data-driven approach to predict default risk of loan for online Peer-to-Peer(P2P) lending,” School of Information, Zhejiang University of Finance and Economics, 310018 Hangzhou, China.