Loan Prediction Using Machine Learning Techniques With Django Framework

[1]Mr. G. Vikram(Assistant Professor),
[2]Thanush, [3]Vinitha, [4]Sai.Harish, [5]WILSON VADLAPATI,
Department of CSE,
MALLA REDDY INSTITUTE OF TECHNOLOGY AND SCIENCE, Telangana, Hyderabad.

Abstract:

Banks get a lot of their money from loans. Loan approval procedures are highly prioritized by financial firms. Forecasting the probability of loan payback by consumers is becoming more complicated due to the increasing number of defaults and the increasing difficulties for banking authorities to appropriately analyze loan requests and manage the risks of persons defaulting on loans. A lot of research in the field of predictive algorithms for loan approval has been going on recently. Machine learning is most effective when used to forecast results from large datasets. To predict a customer's loan approval status, this research employs four algorithms: Random Forest, Decision Tree, Naive Bayes, and Logistic Regression. All four approaches will be evaluated on the same dataset to find the one that works best for deploying the model. Going future, we will construct a system that can anticipate which loan applicants will have their applications accepted by banks using machine learning techniques. The technology will thereafter be able to do the laborious tasks on our behalf.

I. INTRODUCTION

Loans are the principal source of revenue and risk transfer for the banking business. One important component of a bank's asset base is the interest it generates on its loans. The risk of the borrower not being able to repay the loan by the due date is one of the many dangers of making loans. The term "credit risk" applies to it. A numerical score called a "credit score" is used to determine whether an applicant is eligible for a loan or not. In order to assist banks reduce credit risk and choose who to lend to, this research intends to detail several Machine Learning techniques that successfully detect loan defaulters.

II. LITERATURE SURVEY

Title 1: Improving the Accuracy of Loan Approval Data to Promote Fair Pricing and Lending M. Cary Collins is the author.
Year: 2013 Material: Banks should do a better job of handling data associated with loan approval processes in terms of data quality and avoiding data errors, which would have a positive impact on areas
like fair pricing and lending. They quickly reviewed the typical methods used by numerous institutions to collect data before diving into the pricing and approval processes. Partially mandated by federal law are these data protocols. While discussing the collection and analysis of fair lending data, they brought attention to a number of critical first steps toward improving information quality for all stakeholders.
S. Sivasree and T. Rekha Sunny: A System for Predicting Loan Credibility Based on Decision Tree Algorithm
In 2015, data mining technologies are all the rage since there is so much publicly accessible data and we need to find ways to make sense of it. Scientists, retailers, researchers in the area of biology, those working in intrusion detection, those in the telecommunications industry, and countless more all make use of data mining techniques.
Data mining methods are also used by financial firms to gain a competitive edge. The authors of this study offered a prediction model to help financial institutions identify the most credible loan applicants. The Decision Tree Algorithm is being used for the purpose of forecasting aspects linked to believability. This article outlines a model prototype that might assist organizations in making educated decisions about client loan requests.
Third Title: Using Machine Learning to Predict Loan Approval Arun Kumar, Sanmeet Kaur, and Ishan Garg are the authors.
Year of Publication: 2016 Synopsis: Now that business is thriving, plenty of people are looking for loans from banks, but there's a limit to how many people any one bank can give money to because of resources. So, in order to figure out who may acquire a loan, banks usually follow a protocol. Thus, the bank

sought to reduce this risk in this paper by selecting the safe individual in order to save a significant amount of time and money. By analyzing the records of previous borrowers and training the computer with the most accurate machine learning model, we were able to base our choices on their information. The main purpose of this essay is to forecast the security of allocating a loan to a certain person. This paper is composed of the following four sections: first, acquire data; second, compare machine learning models using the collected data; and third, test the system after training it on the best performing model.

III. EXISTING SYSTEM

Anomaly detection relies on an individual's behavioral profile to identify out-of-the-ordinary occurrences. When it comes to detecting online banking fraud, there are three major problems. To begin, when there is little historical data, it could be difficult to profile a person's pattern of behavior. Two, the varied nature of transaction data could lead to inconsistent responses to various attribute values, which might hinder the development and future usage of the model. Thirdly, the transaction data is highly skewed, making it hard to utilize the label information meaningfully. Anomaly detection often has issues such as a high false alarm rate and a lack of generalizability. We hypothesize that this defect may be due to the very biased character of fraud data as well as the fact that individuals have little historical data available for behavior profiling. Although it's simple to collect data from individuals who are similar to you, similarity measurement is really challenging due to the variability of attribute values.

Chapter One: Cons

1) Instead of using machine learning approaches, a mathematical model was created.

2) Appropriate measures were not done once the problem of class imbalance had been sufficiently highlighted.

B. Proposed Structure

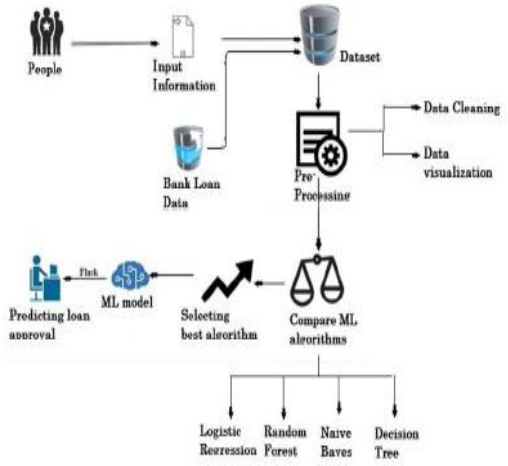
A generalized dataset, which is a collection of datasets from different sources, is used in our proposed approach. After that, four machine learning algorithms—Random Forest, Decision Trees, Naive Bayes, and Logistic Regression—are run on this dataset. The dataset that we used for prediction purposes has a training set and a test set that are divided 7:3. In order to determine which approach yields the best maximum test results, we apply the data model that was generated using machine learning algorithms to the training set. After that, we put the algorithm to work on the test set by making predictions there. Launching the model follows, with the help of the Flask Framework.

Benefits (Chapter C)

It is possible to evaluate and compare the algorithms' precision and efficiency.

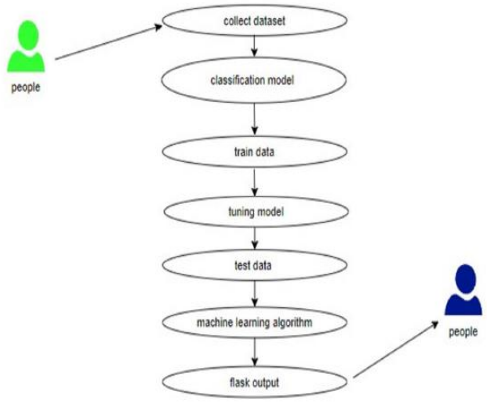
It is possible to address class imbalance using machine learning technologies.

IV. SYSTEM ARCHITECTURE



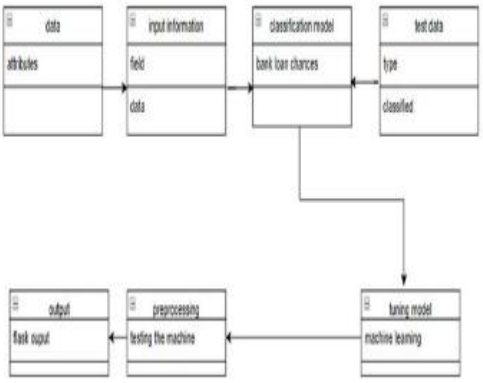
The process of creating a generalized dataset involves gathering and integrating many smaller datasets that fulfill certain requirements. Therefore, pre-processing, the cleaning up of the dataset, is necessary prior to data visualization. We next compare the outcomes of the four methods using the same pre-processed dataset to choose the winner. The next stage is to use the best method to train the model, and after that, to test its output prediction abilities. Then, we predicted whether a particular borrower would be approved for a bank loan using that model.

V. USE CASE DIAGRAM



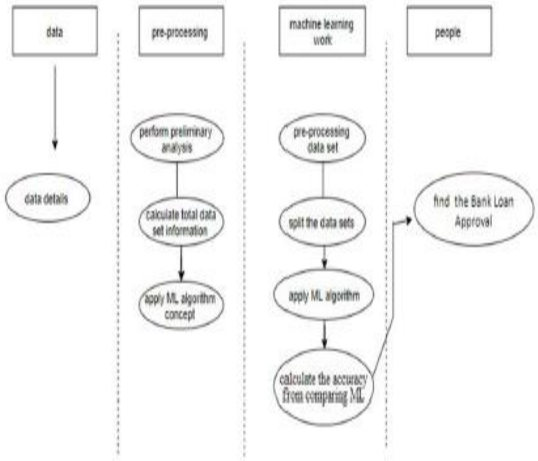
When doing a high-level examination of a system's requirements, use case diagrams are useful. Thus, use cases capture the functionality while analyzing a system's needs. In other words, use cases are just an orderly way to describe the system's functionality.

Section A: Iconography



In a class diagram, various components of the program are visually shown, providing a static overview of the system. The whole system will be shown by a set of class diagrams.

The Diagram Activity B



Not only can activity diagrams show how a system is always changing, but they are also useful for building executable systems via reverse and forward engineering. Although it is not, activity diagrams are often confused with flow charts.

VI. LIST OF MODULES

A. Data Pre-processing

B. Data Analysis of Visualization

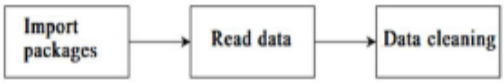
C. Comparing Algorithms

D. Deployment Using Flask

VII. DATA PRE-PROCESSING

Data pre-processing is a method for ensuring that data collections are clean and ready for analysis. Oftentimes, data collected from several sources is not processed and so useless for model analysis. For Machine Learning to be effective when dealing with data, a well-designed model is crucial. Some Machine Learning models have specific format requirements; for example, the Random Forest algorithm rejects null input. Thus, in order to proceed with the random forest approach, it is essential to handle null values in the original raw dataset. As part of the data pre-processing phase, we cleanse the data using the Python Pandas library. Data cleaning is the process of removing unnecessary, partial, or missing information. Methods and processes for cleaning data will vary from dataset to dataset. When integrating datasets, data mistakes, incompleteness, or duplicates might occur. The main objective of data cleaning is to identify and fix errors, since this improves the reliability of findings and

enhances the data value for analytics and decision making.



VIII. DATA ANALYSIS OF VISUALIZATION

Data visualization is a powerful tool for finding patterns, erroneous data, outliers, and more while learning about and playing with a dataset. You may use it to show and explain the deeper, more basic relationships between your charts and plots. Charts, graphs, and plots are visual representations of data that may help make it easier to understand. The capacity to quickly show data samples is vital in both applied machine learning and applied statistics. If you use it to remove outliers, you could get more reliable findings. In particular, it employs the Python program Matplotlib.

IX. COMPARING ALGORITHMS

We first construct a Machine Learning Model using the Scikit-Learn modules in Python before we compare algorithms. Preprocessing, linear model with logistic regression technique, cross validation with K-Fold method, ensemble with random forest method, and tree with decision tree classifier are all steps that are required in this library package. We also separate the data into a train set and a test set so that we may compare the two sets' accuracy and so make a prediction. We use the following performance measures to identify the top algorithm:

A. Matrix of Perplexity

One performance measure used to discover the model's correctness and accuracy is the confusion matrix. Here are the four parameters it uses:

B. False Positives (FP) When the real class is no and the anticipated class is yes, it is possible to mistake someone who will pay for a defaulter. Think about a scenario where a person's projected class states that they are married, but their actual class states that they are not married. A individual who is expected to pay gets mistakenly labeled as a defaulter when the actual class is yes and the projected class is no. This is known as a false negative (FN). This might happen, for instance, if a person's anticipated class states that they are single but their actual class value suggests that they are married. When both the actual and anticipated values of a class are yes, we say that the individual in question is likely to default on their payments (D. True Positives, or TP).

For instance, in the event that both the actual and anticipated class values reveal that the individual is married. If the value of the actual class is zero and the value of the anticipated class is zero as well, then the payer is expected to be a defaulter. If both the actual and anticipated classes indicate that a person is single, then the true positive rate (TPR) would be equal to the number of true positives divided by the sum of the two. The formula to calculate the false positive rate (FPR) is: FPR = (False Positives + True Negatives)/19.

A. Precision

The accuracy ratio, which measures the proportion of observations that are accurately predicted relative to the total number of observations, is the most crucial performance parameter. If we have symmetric datasets with almost same values for false positives and false negatives, then the model will generate correct results, and the accuracy will be higher.

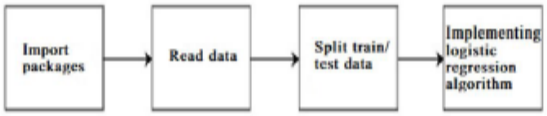
You may calculate the accuracy by dividing the sum of TP and TN by the sum of FP, FN, TN, and TP.

Section G. Accuracy

The accuracy of a prediction is defined as the proportion of expected positive observations that actually occur. Relatively low false positive rates are associated with high accuracy rates in the dataset. The 0.876 accuracy we achieved is commendable.

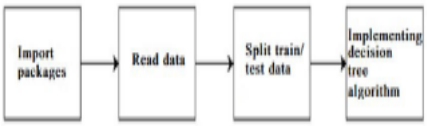
X. LOGISTIC REGRESSION

Logistic regression is a supervised method in machine learning that uses dependent variable classification with the aim of probability prediction. It is a method in statistics used to analyze datasets when the results are dictated by several independent variables. Since it is a dichotomous variable, there are only two possible outcomes. Logistic regression, which takes a dependent variable and a set of independent variables as inputs, aims to find the best model for this relationship. In logistic regression, there are only two possible values for the dependent variable: yes (representing success) or no (representing failure).



XI. DECISION TREE CLASSIFIER

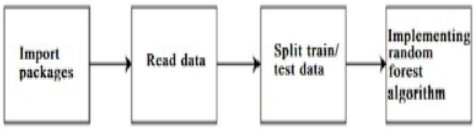
Classification models may be constructed using decision trees, which have the shape of a tree structure. It essentially builds an associated decision tree sequentially while breaking down a dataset into smaller and smaller sections. Each decision node contains at least two branches that terminate at either another decision node or a leaf node; the latter signifies the ultimate conclusion. The optimal predictor is represented by the root node, the highest decision node in a tree. You may use decision trees with numerical and categorical data. Dataset categorization is accomplished using an exhaustive and mutually exclusive collection of if-then rules. Each rule is being learned in turn from the training dataset. Each time a rule is learnt, the tuples that the rule applies to are eliminated. On the training set, this procedure is carried out until it encounters a terminating circumstance. Assuming your dataset only contains categorical characteristics, it is built using a top-down recursive divide-and-conquer approach. Alternatively, they need to be concealed beforehand. Using the information gain idea, we can determine which attributes are most important for categorization and place them at the top of the tree.



XII. RANDOM FOREST ALGORITHM

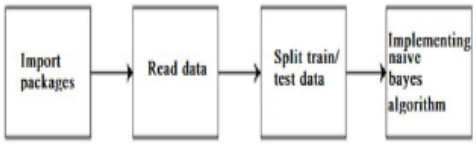
One supervised machine learning technique that relies on the ensemble learning approach is random forest. Ensemble learning allows us to combine many algorithms or iterate on the same strategy to create a robust prediction model. It does its job by training on a large number of decision trees and then producing the class, or the average of their classes. When decision trees tend to overfit their training set, random decision forests fix this problem. Each tree in the forest is used to forecast the category to which the new data belongs in classification difficulties.

At last, the category with the most votes will be the one to get the new record.



XIII. NAÏVE BAYES CLASSIFIER

Using Bayes' Theorem as its foundation, Naive Bayes is a supervised machine learning statistical classification method. It ranks among the simplest methods for supervised machine learning. Among other methods, the Naive Bayes classifier is the most dependable, accurate, and quick. On massive datasets, naive bayes classifiers perform well in terms of both accuracy and speed. The naive bayes classifier takes it as read that each given feature's impact on a dataset class is completely apart from any other features. An applicant's age, geography, income, and history of loans and transactions are some of the factors that determine whether or not they acquire a loan. The aforementioned characteristics are nevertheless thought of as distinct entities, regardless of how much they rely on one another. This assumption is seen as naïve since it simplifies calculation. Class conditional independence describes the aforementioned premise. The conditional probability is the likelihood of a class value given an attribute value. A data instance's likelihood of belonging to a certain class may be determined by multiplying the conditional probabilities for each attribute for a given class value in a dataset. In order to arrive at a forecast, we must first determine the likelihood of each class's occurrence and then choose the class value with the greatest probability.



XIV. DEPLOYMENT USING FLASK

Once we've determined which of the four algorithms performs the best based on performance indicators, we export the model as a PKL file and use Python's FLASK framework to provide an interface for deployment. Customers will see the outcome of their loan approval status as soon as they input their data into the user interface and hit the submit button.

XV. CONCLUSION

Data cleansing and missing value processing are the first steps in the analysis process. Next comes exploratory analysis, and lastly, model construction and assessment. A greater accuracy score, along with other performance parameters, will indicate optimal accuracy on the public test set. Predicting whether an applicant will be approved for a bank loan is possible with the aid of this study.

XVI. FUTURE WORKS

We can optimize the task to be implemented in an AI environment by making predictions about bank loan approvals and connecting them to the cloud.

REFERENCES

[1] Arun Kumar, Ishan Garg, and Sanmeer Kaur, "Loan Approval Prediction Using Machine Learning Approach," 2018.
[2] K. Hanumantha Rao, G. Srinivas, A. Damodhar, and M. Vikas Krishna at International Journal of Computer Science and Telecommunications published an article titled "Implementation of Anomaly Detection Technique Using Machine Learning Algorithms" (Volume2, Issue3, June 2011).
[3] G. Arutjothi and C. Senthamarai, "Prediction of loan status in commercial banks using machine learning classifier," International Conference on Intelligent Sustainable Systems (ICISS), 2017.
[4] "AzureML based analysis and prediction of loan applicants creditworthy," by Alshouiliy K, Alghamdi A, and Agrawal D P I n 2020, Third International conference on information and computer technologies.
[5] "Developing prediction model of loan risk in banks using data mining Machine Learning and Applications," Hamid A J and Ahmed T M, 2016.
[6] M. Li, A. Mickel, and S. Taylor "Should this loan be approved or denied?" published a paper in the Journal of Statistics Education in 2018.
[7] A. Vinayagamoorthy, M. Somasundaram, and C. Sankar, "Impact of Personal Loans Offered by Banks and Non-Banking Financial Companies in Coimbatore City," 2012.
[8] M. Cary Collins, Ph.D., and Frank M. Guess, Ph.D., MIT's Information Quality Conference, 2000, "Improving information quality in loan approval processes for fair lending and fair pricing."
[9] Arun Kumar, Ishan Garg, and Sanmeet Kaur, "Loan approval prediction based on machine learning approach," National Conference on Recent Trends in Computer Science and Information Technology, 2016.
[10] Sivasree M S and Rekha Sunny T, "Loan Credibility Prediction System Using Decision Tree Algorithm," International Journal of Engineering Research & Technology (IJERT), Vol. 4 Issue 09, September-2015.
[11] Jiří Doležal, Jiří Šnajdr, Jaroslav Belás, Zuzana Vincúrová, "Model of the loan process in the context of unrealized income and loss prevention", Journal of International Studies, Vol. 8, No 1, 2015, pp. 91-106. DOI: 10.14254/2071-8330.2015.