Topic – Lone prediction analysis

Name— Subham Nandi

Subject— Machine Learning Foundation

Course— B.Tech(CSE)

Institute—Lovely Professional University

**Abstract**

### With the improvement in the banking sector, lots of people apply for bank loans, but the bank has its limited slots that it only has to give and sell to limited people, so find out who can be given the loan that will be a safer option for the banks. So we're trying to reduce this risk factor behind choosing the safe individual in this paper to save lots of effort and resources from the bank. This paper proposes a loan approval system based on certain attributes to decide whether or not a loan should be granted to an individual. In this paper, the model we are proposing for the bankers would help them predict the trustworthy persons who have applied for a loan, thus increasing the chances of retaining their loans in time. This analysis is created using the algorithm of the Decision tree to estimate a loan's future. We will review the credit scoring for mortgage loans and the conditions that contribute to the rejection of the borrower. This will be that applicant's review and evaluate the percentage of applications that have been accepted but should have been rejected. The risk of mortgage loans demands a great detail of each applicant's review and walking the fine line of who should and should not be approved.

### Introduction

The term banking can be defined as receiving and protecting money that is deposited by the individual or the entities. This also includes lending money to the people which will be repaid within the given time. Banking sector is regulated in most of the countries as it is the important factor in determining the financial stability of the country. The provision of banking regulation act allows public to obtain loans.Loans are good sum of money borrowed for a period and expected to be paid back at given interest rate. The purpose of the loan can be anything based on the customer requirements. Loans are broadly divided as openended and close-ended loans. Open-ended loans are the loans for which the client has approval for a specific amount. Examples of open-end loans are credit cards and a home equity line of credit (HELOC). Close-ended loans decreases with each payment. In other words, it is a legal term that cannot be modified by the borrower. Personal loans, mortgages, auto payments, instalment loan and student loans are the most common examples of close-ended loans. Secured or collateral loan are those loans that are protected by an asset. Houses, Vehicles, Savings accounts are the personal properties used to secure the loan. Unsecured loans are also known as personal or signature loans. Here the lender believes that the borrower can repay the loan based on financial resources possessed by the borrower. Liquidity risk is the risk that arises from the lackof marketability of an investment that cannot be bought or sold quickly enough to prevent or minimize a loss. The interest rate risk is the risk in which the interest rates priced on loans will be too low to earn the bank money**.**

Literature Review :

We start our literature review with more general systematic literature reviews that focus on the application of machine learning in the general field of Banking Risk Management. Since the global financial crisis, risk management in banks has to take a major role in shaping decision-making for banks. A major portion of risk management is the approval of loans to promising candidates. But the black-box nature of Machine learning algorithms makes many loan providers vary the result. Martin Leo, Suneel Sharma and k. Maddulety's extensive report has explored where Machine Learning is being used in the fields of credit risk, market risk, operational risk, and liquidity risk only to conclude that the research falls short of extensive research is required in the field. We could not find any literature review for loan prediction for specific Machine learning algorithms to use which would be a possible starting point for our paper. Instead, since loan prediction is a classification problem, we went with popular classification algorithms used for a similar problem. Ashlesha Vaidya [2] used logistic regression as a probabilistic and predictive approach to loan approval prediction. The author pointed out how Artificial neural networks and Logistic regression are most used for loan prediction as they are easier comparatively develop and provide the most accurate predictive analysis. One of the reasoning behind this that that other Algorithms are generally bad at predicting from non-normalized data. But the nonlinear effect and power terms are easily handled by Logistic regression as there is no need for the independent variables on which the prediction takes place to be normally distributed.

**Existing System**

Bank employees check the details of applicant manually and give the loan to eligible applicant. Checking the details of all applicants takes lot of time. The artificial neural network model for predict the credit risk of a bank. The Feed- forward back propagation neural network is used to forecast the credit default. The method in which two or more classifiers are combined together to produce a ensemble model for the better prediction. They used the bagging and boosting techniques and then used random forest technique. The process of classifiers is to improve the performance of the data and it gives better efficiency. In this work, the authors describe various ensemble techniques for binary classification and also for multi class classification. The new technique that is described by the authors for ensemble is COB which gives effective performance of classification but it also compromised with noise and outlier data of classification. Finally they concluded that the ensemble based algorithm improves the results for training data set.

Proposed System

To deal with the problem, we developed automatic loan prediction using machine learning techniques. We will train the machine with previous dataset. so machine can analyse and understand the process . Then machine will check for eligible applicant and give us result. Advantages

• Time period for loan sanctioning will be reduced.

• Whole process will be automated , so human error will be avoided

• Eligible applicant will be sanctioned loan without any delay.

**Packages used :**

**NumPy:-** NumPy is a general-purpose array-processing package. It provides a high performance multidimensional array object, and tools for working with these arrays.It is the fundamental package for scientific computing with Python. As the whole project is based on whole complex stats ,we will use this fast calculations and provide results.

**Pandas: -** Pandas is the most popular python library that is used for data analysis. We will provide highly optimized performance with back-end source code with the use of Pandas.

**Matplotlib:-** Matplotlib tries to make easy things easy and hard things possible. We will generate plots, histograms, scatterplots, etc.,to make our project more appealing and easier to understand.

**Seaborn:-** We will use it for statistical data visualization as Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

**Scikit-learn:-** It is a Python library is associated with NumPy and SciPy. It is considered as one of the best libraries for working with complex data. There are a lot of changes being made in this library. We will use it for cross validation feature, providing the ability to use more than one metric. Lots of training methods like logistics regression will be used to provide some little improvements.

About The Dataset:

Dream Housing Finance company deals in all home loans. They have presence across all urban, semi urban and rural areas. Customer first apply for home loan after that company validates the customer eligibility for loan. Company wants to automate the loan eligibility process (real time) based on customer detail provided while filling online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, they have given a problem to identify the customers segments, those are eligible for loan amount so that they can specifically target these customers. This is a standard supervised classification task.A classification problem where we have to predict whether a loan would be approved or not. Below is the dataset attributes with description.

| **Variable** | **Description** |
| --- | --- |
| Loan\_ID | Unique Loan ID |
| Gender | Male/ Female |
| Married | Applicant married (Y/N) |
| Dependents | Number of dependents |
| Education | Applicant Education (Graduate/ Under Graduate) |
| Self-employed | Self employed (Y/N) |
| Applicant Income | Applicant income |
| CoapplicantIncome | Coapplicant income |
| Loan Amount | Loan amount in thousands |
| Loan\_Amount\_Term | Term of loan in months |
| Credit History | credit history meets guidelines |
| Property Area | Urban/ Semi Urban/ Rural |
| Loan Status | Loan approved (Y/N) |

ARCHITECTURE TECHNIQUES

Decision tree algorithm in machine erudition how’s which efficiently performs both family and retrogression tasks. It creates decision trees. Decision trees are universally used in the banking assiduousness due to their high exactitude and capableness to formulate a statistical model in plain language. In Decision tree each knot represents a criterion (diagnostic), each link (branch) represents a decision (rule) and each chip represents an outce (categorical or continues value).

Preprocessing

The collected data may contain missing values that may lead to inconsistency. To gain better results data need to be preprocessed and so it'll better the effectiveness of the algorithm. We should remove the outliers and we need to convert the variables. In order to flooring these issues we use chart function.

Train model on training data set:

Now we should train the model on the training dataset and make soothsayings for the test dataset. We can divide our train dataset into two tract train and testimony. We can train the model on this training part and using that make soothsayings for the testimony part. In this way, we can validate our soothsayings as we've the true soothsayings for the testimony part (which we don't have for the test dataset)

Correlating attributes:

Grounded on the correlation among attributes it was observed more likely to pay back their loans. The attributes that are individual and significant can include Property area, education, loan measure, and originally credit History, which is since by insight it's considered as important. The correlation among attributes can be associated using corplot and boxplot in Python platform.

Algorithm:

1. Import all the required python modules

2. Import the database for both TESTING and TRAINING.

3. Check any NULLVALUES are exists

4. If NULLVALUES exits ,fill the table with corresponding coding

5. Exploratory Data Analysis for all ATTRIBUTES from the table

6. Plot all graphs using MATPLOTLIB module

7. Build the DECISIONTREE MODEL for the coding

8. Send that output to CSV FILE

Future Scope:

This project is helpful for banking and finance companies because by using this this model they will understand that wheather they should give a lone to a person or not based upon their data.

CONCLUSION

From a proper analysis of positive points and constraints on the member, it can be safely concluded that the product is a considerably productive member. This use is working duly and meeting to all Banker requisites. This member can be freely plugged in numerous other systems. There have been mathematics cases of computer glitches, violations in content and most important weight of features is fixed in automated prophecy system, so in the near future the so – called software could be made more secure, trustworthy and dynamic weight conformation. In near future this module of prophecy can be integrated with the module of automated processing system. The system is trained on old training dataset in future software can be made resembling that new testing date should also take part in training data after some fix time.

For more information –

<https://github.com/subhamnandi12345/Lone-Prediction-Analysis>

THANK YOU