Loan Default Identification and its Effect
Gopal Choudhary¹, Yash Garud¹, Akshil Shetty¹, Rumit Kadakia¹, Sonali Borase²

¹Department of Computer Science and Engineering, MPSTME, NMIMS, Shirpur, District: Dhule, Maharashtra, India
²Assistant Professor, Department of Computer Science and Engineering, MPSTME, NMIMS, Shirpur, District: Dhule, Maharashtra, India

ABSTRACT

Now a days banking sector is on boom everyone is applying for loan but banks have limitation that they have limited assets so they can provide loan to limited loan applications but when they provide loan, they must assure that loan is being granted to only genuine customers. So, this paper focuses on we will try to lessen the uncertainty factor and assure the loan approval to genuine customers only and save the bank assets. That is performed by way of mining the massive data of the earlier data of the human beings to whom the loan become acknowledged earlier than and on the idea of those records/reviews the machine was skilled the use of the system mastering version which provide the maximum correct result. The main focus of the paper will be on the loan to be approved of those customers only who will be able to pay it back.

**Keywords**: Loan, Machine Learning, Training, Testing, Prediction.

I. INTRODUCTION

When a client takes loan from a bank and financers, he needs to pay it on time but he not paid the loan amount to that bank or financer and default that loan amount. Now that same customer willing is to take loan from more financers and banks and default that loan amount. So, these needs to be stopped somewhere.

Loan portfolio are the largest source of revenue of most of the banks and financers but it’s also leads to NPA (Non-Performing Assets) or end up with loan default. And this affect the financial performance of the financers and banks. The loan defaulters are trying to fool multiple financers and banks at a time. So, they needed to be stopped at right time.

The very obvious effect of loan default is that it causes the monitory growths of financers and banks. This result in the increase in the rate of interest on the loan, and this leads to overall failure to the economic growth.

II. METHODS AND MATERIAL

The implementation of the project can be divided into two parts.

1) Data analysis and data cleaning

Selecting relevant features

In machine learning and information, feature selection, additionally known as variable choice, attribute selection or variable subset choice, is the procedure of choosing a subset of applicable capabilities (variables, predictors) to be used in model creation. function choice techniques are used for four motives:

- simplification of models to cause them to simpler to interpret with the aid of researchers/customers
- reduces training time
- To avoid the curse of dimensionality
• enhanced generalization by reducing overfitting (formally, reduction of variance)

Null value imputation
Missing data can occur while no records is supplied for one or greater items or for a complete unit. missing records is a totally big problem in actual existence scenario. missing information can also talk to as NA (now not available) values. In Data Frame now and again many datasets sincerely arrive with lacking information, either as it exists and turned into no longer amassed or it in no way existed. for instance, suppose unique user being surveyed may also pick out now not to share their income, a few users may select no longer to proportion the cope with in this manner many datasets went lacking.

Handling Outliers
An outlier can be referred as a data point this is remote from different comparable factors. they’ll be due to variability inside the measurement or might also suggest experimental errors. If feasible, outliers must be excluded from the data set. but, detecting that anomalous times might be very difficult, and isn’t usually viable.

Different methods of dealing with outliers:
• Univariate method: This technique looks for data points with extreme values on one variable.
• Multivariate method: right here we search for unusual combinations on all the variables.
• Minkowski error: This approach reduces the contribution of ability outliers within the training method.

Training a machine learning model:
Logistic Regression
Rather than predicting precisely 0 or 1, logistic regression generates a possibility—a value between zero and 1, exclusive. for instance, don’t forget a logistic regression version for unsolicited mail detection. If the version infers a cost of 0.932 on a particular electronic mail message, it implies a 93.2% chance that the email message is junk mail. extra precisely, it way that inside the limit of infinite schooling examples, the set of examples for which the model predicts 0.932 will certainly be spam ninety 3.2% of the time and the last 6.8% will no longer.

Random Forest
Random forest is very easy and flexible, easy to apply ML algorithm that produces, even without hyper-parameter tuning, a top-notch result maximum of the time. it is also one of the maximum used algorithms, as its simplicity and the truth that it can be used for both category and regression duties. in this put up, you’ll analyze, how the random forest algorithm works and numerous different vital things approximately it.
Random Forest is a supervised studying set of rules. Like you can already see from its call, it creates a forest and creates it somehow random. The forest it builds, is an ensemble of decision tree, maximum of the time skilled with the “bagging” technique. the overall concept of the bagging technique is that a combination of learning models increases the overall result.

![Flow Diagram of the System](image)

**Fig 3.** Flow Diagram of the System

### III. EXPERIMENTAL RESULT

#### Data Set

The data which has been trained is being applied to machine learning model, whenever the new customer enters the detail in application form that act as test data set. So, after testing operation is performed, then afterwards the model will predict that will the new customer will be able to pay the loan on time or not on the process of loan approval on the basis of the training dataset.

<table>
<thead>
<tr>
<th>Attribute Name</th>
<th>Description</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cust_id</td>
<td>It's a Unique id</td>
<td>Integer</td>
</tr>
<tr>
<td>loan_amnt</td>
<td>The Loan amount that has been sanctioned</td>
<td>Integer</td>
</tr>
<tr>
<td>term</td>
<td>Duration time of Loan</td>
<td>Character</td>
</tr>
<tr>
<td>int_rate</td>
<td>The annual interest rate on the loan amount</td>
<td>Integer</td>
</tr>
<tr>
<td>installment</td>
<td>The emi that are monthly paid by the borrower</td>
<td>Integer</td>
</tr>
<tr>
<td>grade</td>
<td>LC assigned loan grade</td>
<td>Character</td>
</tr>
<tr>
<td>annual_inc</td>
<td>annual_inc: The self-claimed annual income given by the borrower during registration.</td>
<td>Integer</td>
</tr>
<tr>
<td>Credit_History</td>
<td>credit history meets guidelines</td>
<td>Integer</td>
</tr>
<tr>
<td>Property_Area</td>
<td>Type of area whether it is urban or rural</td>
<td>String</td>
</tr>
<tr>
<td>loan_status_coded</td>
<td>Loan Approved(Y/N)</td>
<td>Integer</td>
</tr>
</tbody>
</table>

#### Accuracy Measure

The model is able to predict 61% of defaulters this can be observed by ROI (Return on Investment). To calculate ROI here we have used fully paid, charged off and loan in grace period. So how to calculate ROI = 'total_payment' / 'funded_amount' ROI calculated without using this model = -4.57 and after using this model ROI calculated is 2.22.
Grade wise analysis of ROI

As you can see in the below table the ROI of C, D, E, G become positive using this model.

Table 2. ROI Index Table

IV. CONCLUSION

From a proper analysis of tremendous factors and constraints at the aspect, it is able to be appropriately concluded that the product is a distinctly green aspect. This utility is running nicely and meeting to all Banker necessities. This factor can be without difficulty plugged in lots of different systems. There have been numbers instances of computer system faults, mistakes in content and most essential weight of features is fixed in automatic prediction machine, so in the close to future the so-knew as software program may be made extra at ease, dependable and dynamic weight adjustment. In close to future this module of prediction can be combine with the module of automatic processing device, the device is trained on antique schooling dataset in destiny software program may be made such that new testing date have to additionally take part in education records after a few fix times.

V. REFERENCES


Cite this article as:

Journal URL : http://ijsrcseit.com/CSEIT1952198