**Term project**

**BSAN 734 – Data Mining Business Analytics**

**(Class Meeting Day and Time: Wednesday, 7:05 PM)**

**Fall 2022**

**Instructor: Dr. Ross Gruetzemacher**

**Project Title: Loan Approval Prediction**

**Team 3:**

Venkata Sai Nallapati

Bharath Muppidi

Akshita Reddy Nallamaddi

Yogananda Theeguru

1. **Introduction:**

One of the main services provided by banking and other financial institutions is loan lending. For banks and other financial organizations, insolvency or loan default is a significant issue, particularly for peer-to-peer (p2p) online lending platforms where a decision on whether to approve or reject a loan to a potential borrower must be made considerably more quickly than in a typical bank. Both individuals and businesses borrow money to fulfill a wide range of needs and desires. In this loan-giving system, the borrower's creditworthiness is a crucial factor. The customer's or company's creditworthiness has a big impact on whether a loan is approved or denied. Banks and other financial organizations rely on outside service providers, including credit bureaus and consultants to evaluate a potential borrower's creditworthiness. Numerous factors, like the customers' income, age, employment history, marital status, etc., have an impact on creditworthiness when loans are sanctioned to them. In this research, we suggest two classification models namely Binary Logistic Regression Model (LRM) and Random Forests which allow us to quickly forecast a potential borrower's creditworthiness using the historical data that is currently accessible.

1. **Problem Statement:**

Peer-to-peer lending is becoming more and more common in today's society as we move toward using digital methods for all kinds of needs and activities. When a borrower defaults, the lender loses their investment and loses interest, which makes them lose faith in the lending platform. Lack of trust in the platform will cause the lender to move their business to another reliable platform, costing the lending platform clients and income. As a result, there is a danger to our finances and potential losses for the platform. As a result, we provide a model that, given all the previously specified independent factors, may predict the risky customers. As a result, loan lending organizations can avoid those applicants and only lend loans to those eligible customers to profit from the interest payments they can make from them each month while avoiding late payments and repeated reminders that cost them labor and time. Additionally, they can give priority to those cities, occupations, and age groups who are creditworthy for loans and able to make timely payments.

**2.1 Drawbacks of the Current System:**

It takes a significant amount of time and effort to review all of the applicants' information. Due to thoroughly verifying every piece of information, there is a possibility of human error. There is a chance that a loan will be assigned to an ineligible applicant and the organization would have to waste their time and money trying to get the repayments. This would affect not just the organization from the borrower’s perspective but also from the lender’s side. Because when he doesn’t get his timely repayment, he would go for another platform.

**2.2 Proposed system:**

We created automatic loan prediction using machine learning approaches to solve the issue. With previously collected data, we will train the machine. In order for the machine to assess and comprehend the process. A dangerous applicant will thereafter be identified by a machine so that the lending organization can avoid him.

**Advantages:**

* The loan sanctioning period will be shortened.
* The entire procedure will be automated, preventing human error.
* A loan will be approved for a qualified applicant right away.

1. **Method:**

**3.1. Data understanding:**

The dataset contains 252,000 records containing 13 variables with 1 dependent variable and 12 independent variables. Following are the available variables with the last one being the dependent variable.

Table

Description automatically generated

Fig 1. All the columns of the dataset and the respective data types

To consider the data set as it is with 252,000 records and to split the data into 80:20 train and test data proportion. Considering the ‘Risk\_Flag’ as the dependent variable, we will be predicting the target with all other independent variables using classification models like Binary Logistic Regression and Random Forests. Evaluate both the classification models using various evaluation metrics like F1-score, accuracy, and ROC AUC. Choosing the best model which has a high evaluation score.

**3.2. Implemented methodology:**

**3.2.1. Data pre-processing:**

We initially thought of down sampling the dataset to 22,500 records as the raw dataset was huge. But the data would be insufficient to split into train and test data and work on it. So, instead we chose to continue with the raw dataset with 252,000 records. But we faced the problem of imbalance of the dependent variable. Following is a visual representation of the imbalance of dependent variable in the dataset.

Chart, pie chart

Description automatically generated

Fig 2. A pie chart showing the imbalance in the raw dataset.

**3.2.2. Data balancing:**

Our dependent variable is a binary variable with ‘0’s which represents non-risky candidates and ‘1’s representing the risky candidates. There is a high imbalance of 81% of 0s and 19% of 1s in this variable. Hence in order to solve this problem we did oversampling using ‘Random Over sampler’ balancing technique in python. This has balanced the 0s and 1s equally.

Graphical user interface, text, application, chat or text message

Description automatically generated

Fig 3. Using Random Over sampler to oversample the data to handle the imbalance.

Chart

Description automatically generated

Fig 4. A pie chart depicting the balanced dataset post oversampling.

**3.2.3. Data visualization:**

We have also plotted a heatmap to calculate the correlation among the variables. We found out that ‘House\_Ownership’ is the most important variable in predicting our dependent variable.

Graphical user interface

Description automatically generated

Fig 5. Heat map showing the correlation among the variables.

**3.2.4. Data engineering and splitting:**

In order to convert the unique categorical values of independent variables such as ‘Married/Single’, ‘House Ownership’, ‘Car Ownership’, ‘Profession’ and ‘State’ into numerical values, we used a label encoder.

Next, we split the balanced dataset of 442,008 rows into train and test data using ‘train\_test\_split’ package sklearn. We chose 80% as train data and 20% as test data.

**3.2.5. Modeling:**

**Logistic Regression:**

Once the data has been split, we performed Logistic Regression using the package of the same name. We predicted the f1-score and accuracy against the test data and got the following results.

Table

Description automatically generated

Fig 6. Results of logistic regression run with f1-score and accuracy.

As mentioned in the earlier parts of the report, detecting the risky candidates i.e., predicting 1s which are True positive is our aim. But the above model couldn’t predict the 1s at all, which doesn’t help us in taking a business decision. This was the case when we tried before oversampling. Although at this point it doesn’t seem necessary, we plotted ROC AUC in the process of learning. The result was ‘0.52’ which is not good.

**Random Forests:**

We thought there could be a mismatch between the dataset and the model and tried performing Random Forests. We chose estimators to be 350 trees and a maximum depth of 5.

Graphical user interface, text

Description automatically generated

Fig 7. Code and results of random forests run with f1-score and accuracy.

The results didn’t change. Not even by a percentage. The accuracy was good, but the sensitivity or true positive or recall rate was bad. Even random forests couldn’t predict properly. Just like earlier the evaluation metrics doesn’t seem to matter here also.

**Evaluation:**

We used the ROC curve and Area under the curve to evaluate the classification models for which Logistic regression got 0.50 and Random Forests got 0.60. Comparing both Random Forests was good but neither method has given us good true positive rates. When we cross-checked with other people’s work on this dataset from Kaggle the results were the same. Thus, we came to know that the dataset itself is bad and cannot be utilized in predicting a new individual’s creditworthiness and cannot make business decisions using this dataset.

Chart, line chart

Description automatically generated

Fig 8. ROC curve of Logistic Regression with AUC value.

Chart, line chart

Description automatically generated

Fig 9. ROC curve of Random Forests with AUC value.

**Conclusion:**

In a time where data analysis and business decisions taking from its results are shaping our economy, we think it is also important to understand which data cannot be used to take those critical decisions. Upon clear examination, we concluded that the dataset we chose is not suitable in making business decisions. Not all data can be used to make decisions. This would be our biggest learning from this project. If there are some more important independent variables such as credit score and past history of loans, then it would be more helpful to evaluate the results accurately and use it in decision-making.