International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 03 | Mar 2021 www.irjet.net p-ISSN: 2395-0072

LOAN RISK PREDICTION Using MACHINE LEARNING

Mrs.R.Radhika[1], Golla Venkata Lakshmi Dharani [2] , Gudipati Thanuja [3]

[1]Assistant Professor, Dept. Of CSE, SCSVMV (Deemed to be University), Kanchipuram, TamilNadu, India
[2]Student, Dept. Of CSE, SCSVMV (Deemed to be University), Kanchipuram, TamilNadu, India
[3]Student, Dept. Of CSE, SCSVMV (Deemed to be University), Kanchipuram, TamilNadu, India

---

Abstract - *In the recent years, the number of people applying for the loans gets increased for various reasons. Loans can be taken from many sources. One of them is by using the online platforms that connects the investors and borrowers. These platforms act as a connecting bridge for the people interested in investing their money to attain good amount of interest and the people who wants to borrow money with low-interest rates compared to other sources. As the investors involved in this platform are individual people not any organizations it is very important to make a wise decision in order to invest in a particular customer. The investor should be able to know whether is it safe to give loan to the borrower or not i.e., can the borrower be able to repay the loan.In this project, we are using Lending club data set to determine whether the loan is re-payed or charged-off., analyze the data using Exploratory Data Analysis and apply the machine learning algorithms like KNN Classifier, Random Forest Classifier, Decision Tree and Logistic Regression.*

Key Words: **P2P, KNN Classifier, Random Forest Classifier, Decision Tree and Logistic Regression, Smote, Nearmiss, Ensemble.**

1. INTRODUCTION

Peer-to-peer (P2P) lending, which is also known as Social lending, operates online trading platforms as a medium for lending money without the intrusion of traditional financial mediators, such as banks. Conducting business on peer platforms has recently become popular because it not only reduces financing costs but also has the potential for higher profitability for both investors and borrowers. Borrowers benefit from lower interest rates; investors receive a higher return than they would from a bank. However, evaluating the risk of investing is a common challenge in micro-financing, where loans are typically unsecured. Further, P2P lending usually occurs in settings with a high level of information asymmetry – that is, settings where the investors do not have complete information about the borrowers' credit history. Even though when the information is available, lenders might not know how to extract useful data from the given data. Therefore, predicting a borrower's creditworthiness on whether or not to fund particular loans has emerged as a critical problem for online lending platforms. Hence machine learning algorithms on used on the available data in order to generate a model that helps to predict the repayment of loan.

1.1 SCOPE OF THE PROJECT

Developing an application to analyze and predict the loan risk is a basic tool aimed at decreasing the risk for investors when investing in borrowers. The purpose of the system is mainly to use and concentrate on the data from the dataset that can be availed from the customer before giving the loan. This system will help to generate best suitable model for the data which helps in decreasing the risk of losing the money for an investor.

1.2 KNN Classifier

KNN collects a group of points that are labelled and uses the labelled points to learn how to label other points. A new point is labelled where, it checks all the labelled points closer to the point which is a new point (its nearest neighbors), and has nearest neighbors vote, so whichever label most of the neighbors have is that the label for the new point (the "k" is that the number of neighbors it checks).

1.3 Random Forest Classifier

Random forests/random decision forests is a multiple classifier systems for regression, classification and other tasks. They function by constructing a multiple decision trees at training time and outputting the category that's the Mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to training set.

1.4 Decision Tree

A Decision tree algorithm is a flow diagram like tree model, where each inner node represents a test on an attribute, every branch denotes an result of the test and each external node has a class label. Decision tree method analyzes examples and sorts them down the tree from the root to external node, with the external node providing the classification to the example. Every node in the tree works as a test case for some attribute, and every edge comes down from that node relates to one of the probable answers to the test case. The process is repeated for each subtree rooted at the new nodes and it's recursive in nature.

1.5 Logistic Regression

It is a statistical method for studying a data set where there are one or more independent variables that decides an outcome. The outcome is measured along with a dichotomous variable (where there are only two possible outcomes). The objective of logistic regression is to find out the best fitting model to represents the relationship between the set of independent (predictor or explanatory) variables and dichotomous characteristic of interest (dependent variable = outcome variable).

1.6 SMOTE

In this technique the minority class samples are randomly replicated to make it equal to the number of majority class samples.

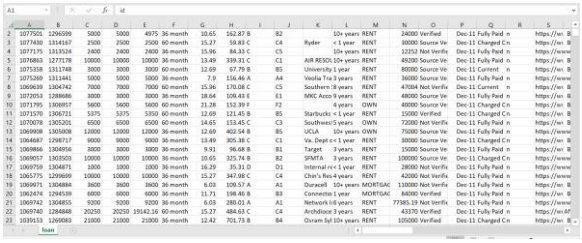
1.7 NearMiss

In this technique the majority class samples are randomly sampled to make it equal to the number of minority class samples

2. IMPLEMENTATION

2.1 Dataset:

The dataset is collected from a popular P2P lending platform "Lending Club". The dataset is obtained from the website Kaggle. The dataset consists of 74 columns and 887379 Records.



2.2 Data Preprocessing

Data preprocessing is the strategy used to convert the raw-data into comprehensible collection of data. As the data obtained from the dataset have different forms of data it should be converted into a format that can be understood by the machine to perform further actions. In our dataset, there are 74 columns and 887379 Records. All these columns are not used in further processes. Hence, the dataset is processed to convert it into suitable format. First import all the required libraries. The libraries imported for this are pandas and numpy Then the dataset is imported in to a dataframe for data processing of our dataframe the following methods are used:

info(),drop(),nunique(),unique(),replace(),isnull().sum(),dropna(),value_counts()

After performing the required operations using above methods the dataframe is converted into required format. The dataframe that obtained is used to create two dataframes. First dataframe consisting of input variables and second dataframe is target variable.

2.3 Feature Extraction

Feature Extraction may be a dimensionality reduction during which the info is reduced for better processing. Feature extraction gives the information on how each dimension/feature is influencing the target variable. Hence using feature extraction the dimensions that are highly influencing the target variable can be selected that helps in the further process. For feature extraction we used the ExtraTreesClassifier. A model for ExtraTreeClassifier is created and the data is fitted to using fit() method. Applying feature _importances_method for the model created gives the importance of each feature. The top most features are selected and remaining are dropped from the data. After performing this the data which is used to create model will be generated.

2.4 Splitting Data

For training the machine learning algorithms and to check for its performance the data must be divided into two parts. One part for training and other for testing the trained model. For splitting data into two parts train_test_split() method is used. This method takes the input-variables and output-variable data and divide into four parts. A train input-variables, test input-variables, train output variable and test output-variable. train_test_split() method takes test_size parameter to determine the ratio in which the splitting is to be performed. Test_size used in this scenario is 0.3

2.5 Evaluation Metrics

By using the Algorithms for data ,Evaluation metrics are obtained.Evaluation metrics are used to determine the performance of the model. These metrics are imported from sklearn.metrics

1.Confusion Matrix:

Confusion matrix is a n x n matrix where n determine the number of target classes.

2.accuracy_score:

accuracy_score gives the ratio of the number of instances that are correctly classified to the total number of instances.

3.precision_score:

precision_score gives the ratio of True Positives to the sum of True Positives and False Positives

4.Recall_score:

Recall_score is the number of true positives divided by the number of true positive plus the number of false negatives.

5.F1-score:

It is harmonic mean of precision and recall. Considered to get combined result of precision and recall

6.Classification_report:

Classification report gives the precision, recall, f1-score and support for both the classes. It gives the total performance of the model

2.6 Imbalance Data Problem

The data has imbalance problem i.e., the number of instances for both the target classes are not equal in the data. The number of instances for the target classes are 65591 and 842067. The percentage of majority class is 92.2% and minority class is 7.3% approximately. Whenever a new instance is given it is most likely be classified for majority class, therefore we can say we cannot rely on the models completely for new data. To solve this problem and increase efficiency of model we have to address this problem.

To deal with imbalanced data we are SMOTE and Near Miss Techniques.

2.7 Ensemble Model

Although we can select the best model based on its performance, there is the chance of over fitting of the data thus it is solved using ensemble technique. Here we are using Voting classifier method so we can create a more accurate model which will give the result based on outcomes generated by all the specified models. It will increase the chances of getting more desirable output.

To Create a combined model use VotingClassifier() of sklearn.ensemble

3. Testing

3.1 Black box Testing

In this testing, the testing is performed on inputs and its corresponding outputs.It cannot identify where is the exact fault occurred.

3.2 White box Testing

In this testing, it tests internal coding and infrastructure of a software specialise in checking of predefined inputs against expected and desired outputs. It can determine exactly where the faults occurred.

3.3 Unit Testing

In Unit testing individual units are teste to determine if they are fit for use and application is the smallest testable part. In procedural programming a unit could also be a private function or procedure. Unit tests are created occasionally by programmers and by white box testers.

3.4 Integration Testing:

Integration testing is that the activity of finding faults when testing the individual components together. Involving all components of the system in structural testion which is the cumulation of integration testing.

3.5 System Testing:

System testing tests all the components together, seen as one system to spot faults with reference to scenarios from the matter statement and therefore the requirements and style goals identified in the analysis and system design.

4. Visualization

4.1 Visualizing Feature Importance

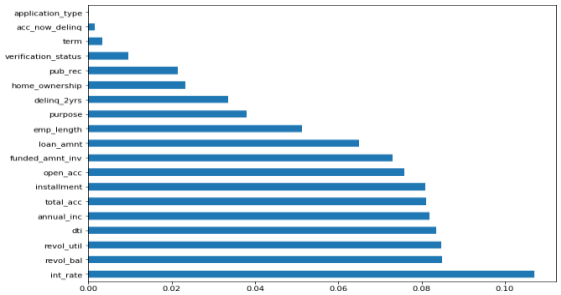


Chart -1: Feature Importance

4.2 Visualizing Some general Patterns

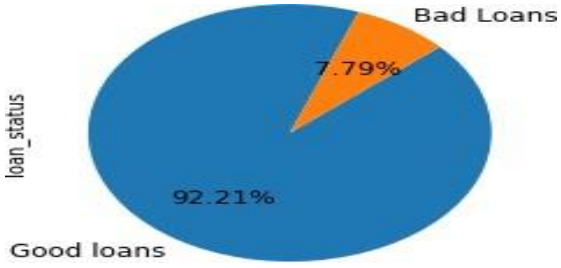


Chart -2: Loan Distribution

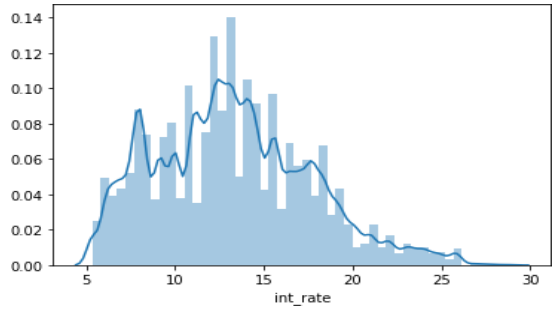


Chart -3: Distribution of Rate of interest

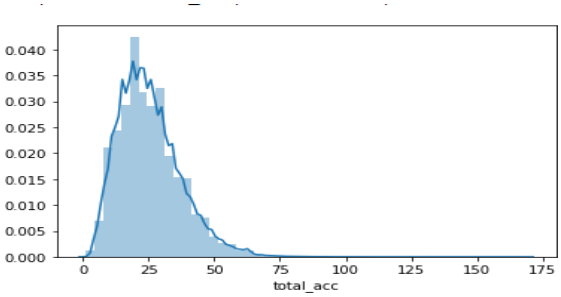


Chart -4: Distribution of Total Accounts

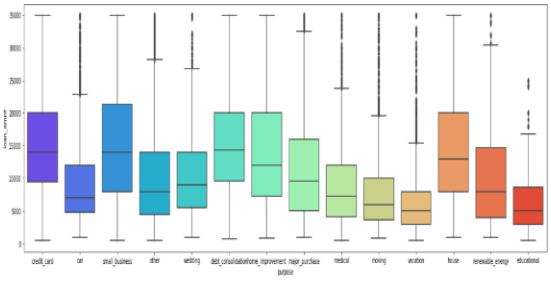


Chart -5: Loan Amount Vs Purpose

4.3 Heatmap

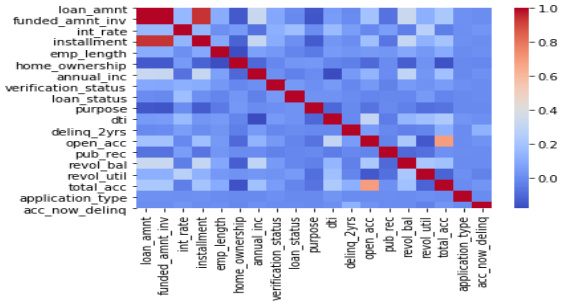


Chart -6: Heatmap

5. Results

1. Results obtained for Imbalanced Data

Evaluation Metrics Classification Model	Accuracy	F1-Score	False Negatives
Decision Tree classifier	0.85	0.92	16407
Random Forest classifier	0.91	0.95	18967
Logistic regression classifier	0.92	0.95	19580
K Nearest neighbour classifier	0.91	0.95	19361

2.Results obtained after balancing data using Smote Technique

Evaluation Metrics Classification Model	Accuracy	F1 - Score	False Negatives
Decision Tree classifier	0.91	0.91	18758
Random Forest classifier	0.92	0.92	17641
Logistic regression classifier	0.64	0.61	69394
K Nearest neighbour classifier	0.79	0.77	26273

3.Results obtained after balancing data using Nearmiss Technique

Evaluation Metrics Classification Model	Accuracy	F1 - Score	False Negatives
Decision Tree classifier	0.76	0.77	4464
Random Forest classifier	0.82	0.82	3327
Logistic regression classifier	0.64	0.63	6489
K Nearest neighbour classifier	0.84	0.85	5254

4.Results obtained after using Voting Classifier

->Smote

Evaluation Metric →	Accuracy	F1-score	False positives
Soft voting	0.93	0.93	13679
Hard voting	0.87	0.86	6082

->Nearmiss

Evaluation Metric →	Accuracy	F1-score	False positives
Soft voting	0.83	0.84	3914
Hard voting	0.82	0.81	1609

3. Final Result

Prediction for Sample Data

```
# Sample Data
sample_input=[[5000.0,4975.0,10.65,162.87,24000.0,27.65,3.0,13648.0,83.7,9.0]]
df_test=pd.DataFrame(sample_input, columns=['loan_amnt','funded_amnt_inv',
'int_rate','installment','annual_inc','dti','open_acc','revol_bal','revol_util',
'total_acc'])
predictions_test = voting_soft1.predict(df_test)
if predictions_test[0]==1:
    print('Risky')
else:
    print('No Risk')
```

```
No Risk
```

6. CONCLUSION

The data used while developing the model can be collected even from a new borrower easily. Hence the model developed can be used by any peer to peer lending platform. The Algorithms used to create models are KNN Classifier, Random Forest Classifier, Decision Tree Classifier, and Logistic Regression. Among these models Random Forest gives the best results. As the collected data is real-world data it has the problem of imbalance in target classes. Though a good amount of accuracy is generated while using the models, the models may not predict the minority class correctly because of the number of instances for minority class are low. Hence the models are resampled using SMOTE and NearMiss techniques to solve this problem. After applying the techniques, it is observed that Random Forest Classifier is performing well for the data. For solving problem of over fitting and for better performance and more accuracy all the models are combined into a single model using Voting Classifier which gives result considering all the outputs from all the models and predict the target class having maximum votes. For the voting classifier SMOTE model performance is higher because of the presence of large amount of data. Hence it is concluded that after balancing data using SMOTE technique it is suggested to use voting classifier for better performance. Among the hard and soft voting though false positives are less in hard voting this in turn implies that the false negatives are more that means rejecting the customers those who deserve loan thus soft voting is considered into account. Finally, the suggested model is voting classifier with Soft voting.

REFERENCES

[1] Liang, Junjie. "Predicting borrowers' chance of defaulting on credit loans."

[2] Pandey, Jitendra Nath. "Predicting Probability of Loan Default Stanford University, CS229 Project report Jitendra Nath Pandey, Maheshwaran Srinivasan."

[3] Tsai, Kevin, Sivagami Ramiah, and Sudhanshu Singh. "Peer Lending Risk Predictor."

[4] Caselli, Stefano, Stefano Gatti, and Francesca Querci. "The sensitivity of the loss given default rate to systematic risk: new empirical evidence on bank loans." Journal of Financial Services Research 34.1 (2008): 1-3

BIOGRAPHIES

Mrs.R.Radhika is Assistant Professor in Computer science and Engineering department in SCSVMV (Deemed to be University).

Golla Venkata Lakshmi Dharani is pursuing B. Eng. from SCSVMV (Deemed to be University).

Gudipati Thanuja is pursuing B. Eng. from SCSVMV (Deemed to be University).