Predicting Loan Repayment

Richa Sharma[1] and Nishita Malhotra[2]

Indira Gandhi Delhi Technical University for Women, Delhi, India
{richa.7327,malhotranishita}@gmail.com

Abstract. Lending and Borrowing is a mechanism that is completely based on the lender's trust and borrower's credibility. The problem arises when the borrower becomes risky and lender looses money. The aim is to solve this problem by analysing multiple factors (provided by Lending Club from 2007 to 2015) cumulatively. Finally, the lender will be able to lend money to the most eligible and reliable borrower.

Keywords: Loan · Borrower · Lender.

1 Introduction

Start writing the background of your problem here in introduction. Predicting the Loan Repayment is mostly dependent on the characteristics of the borrower. The less riskier the borrower is , the more cooperative is the lender.The previous records of the borrower are analysed to give assurance to the lender. Features like FICO Score,purpose of loan, annual income, revolving balance , inequiries about the borrower ,etc predicts the possibility of repayment by the borrower in a more efficient way. The appropriateness of the borrower is directly proportional to generosity of the lender providing better interest rates and leniency in installments. We believe that Machine Learning algorithms can help us in determining the loan defaulters and making this peer to peer lending smooth.

Problem Statement. *The objective is to build a machine learning model that is best suited for minimising losses of the lending organisation and provide loans to maximum number of loan applicants by analysing their credit history.*

2 Related Work

"Predicting Defaults in Lending Club Loans" [2] proposed a system which deals with an imbalanced data-set and applied algorithms such as logistic regression to compute defaulters correctly.

"Determinants of Default in P2P Lending"[1] uses the correlation between the attributes of the Lending Club data to make the peer to peer lending more relevant on both sides.

We plan to give improved results by studying different models for predictions and reduce the computed error by analysing dependencies of attributes on each other. Some factors have more impact on the predictions made, so our motive will be to give them prime importance in the model created.

3 Proposed Methodology

3.1 Dataset Description

We'll be using the publicly available loan_data on kaggle.com[1]that includes various attributes of borrower and if loan was fully paid or not. Table 1 describes the dimensions of the dataset used. Table 2 describes attributes of data. Here,

Details	**Count**
Number of Attributes	14
Total number of records	9578

Table 1: Details of the data-set.

not_fully_paid is considered to be the target label of data-set*labels*.

Data Attributes	**Type**	**Brief Explanation**
credit_policy	Categorical	1 If customer meets LendingClub.com criteria ,else 0.
purpose	Categorical	Purpose of the loan
int_rate	Numeric	Interest rate of loan
installment	Numeric	Monthly installments if loan is funded
log_annual_inc	Numeric	The natural log of the annual income of the borrower.
dti	Numeric	Debt-to-Income ratio of the borrower
fico	Numeric	FICO credit score of borrower
days_with_cr_line	Numeric	Number of days borrower had a credit line
revol_bal	Numeric	Borrower's revolving balance
revol_util	Numeric	Borrower's revolving line utilization rate
inq_last_6mths	Numeric	No of inquiries of a borrower in last 6months
delinq_2yrs	Numeric	30+ days due payment for last 2 years
pub_rec	Numeric	Borrower's number of derogatory public records
not_fully_paid	Categorical	Loan paid back fully or partially

Table 2: Details of Data Attributes.

4 Data Exploration

Data is available in one .csv file and does not contain any null values. Although it is quite imbalanced as the positive examples i.e. number of records where loans aren't paid back fully is only 19%. It is handled using suitable data balancing techniques.

[1] https://www.kaggle.com/braindeadcoder/lending-club-data#loan_data.csv

4.1 Visualisations

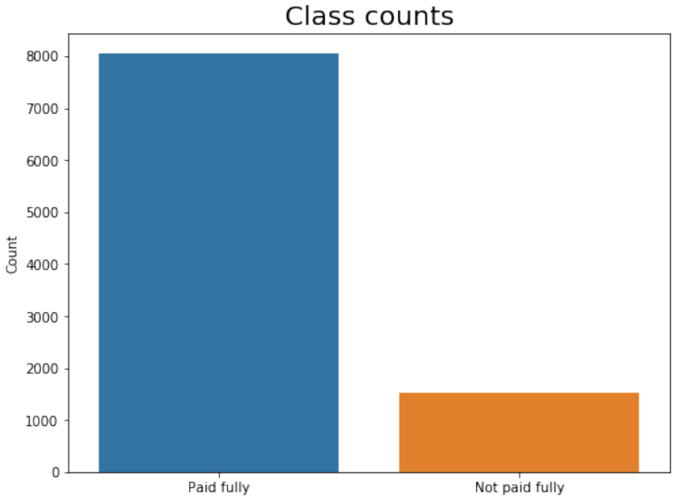


Fig. 1: Ratio of the people who paid loans fully to the ones who paid partially.

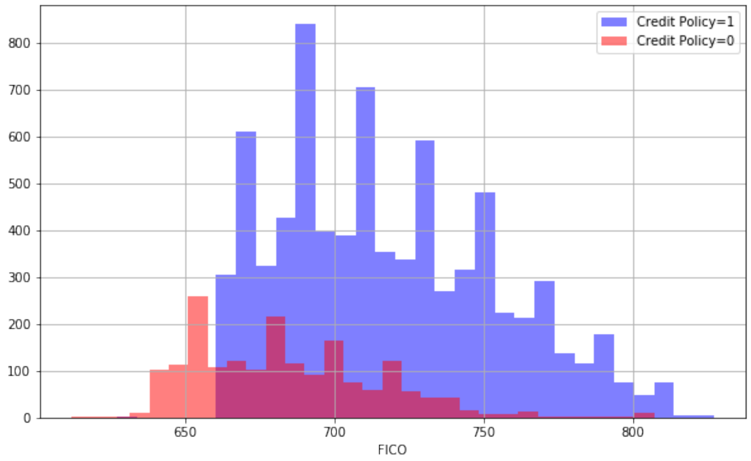


Fig. 2: Relationship between the FICO score of the borrower and Credit Policy. The histogram depicts that the borrower with a higher FICO Score, have credit policy as 1 because he/she tends to meet more criteria defined by the LendingClub.com whereas the ones with less FICO Score have credit policy as 0.

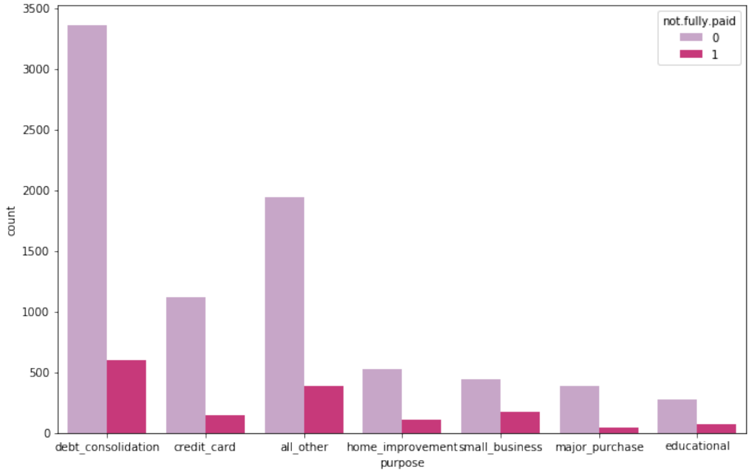


Fig. 3: Relationship between the Purpose of Loan and the Repayment. The Countplot depicts that the repayment of loans is independent of the purpose, since for each purpose ratio of fully paid and partially paid borrowers is almost same.

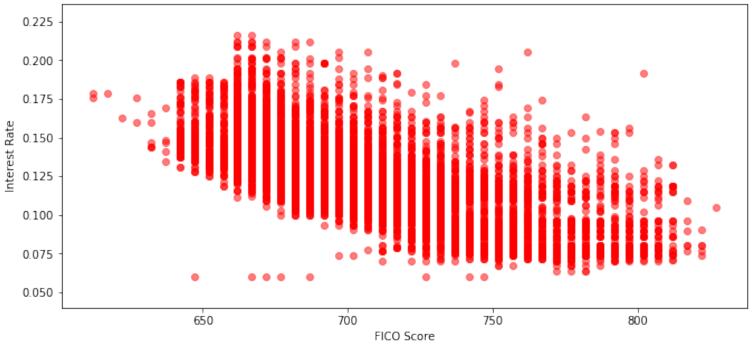


Fig. 4: Relationship between Interest rate and FICO Score. The Scatter Plot depicts that higher the FICO Score of borrower , lesser is the interest rate. It shows that a good FICO score gives lender a sense of trust on borrower.

4.2 Data Pre-processing

Data Cleaning : There were no rows containing null values. There were no duplicate rows. The data was sorted according to the Fico Score.

Direct Features : These features are already present in the data set as attributes so no computation is required as such.

Indirect Features : We have computed **purpose_encoded** using the direct feature purpose. Here each category of purpose is represented from 0 to 6.

5 Algorithms

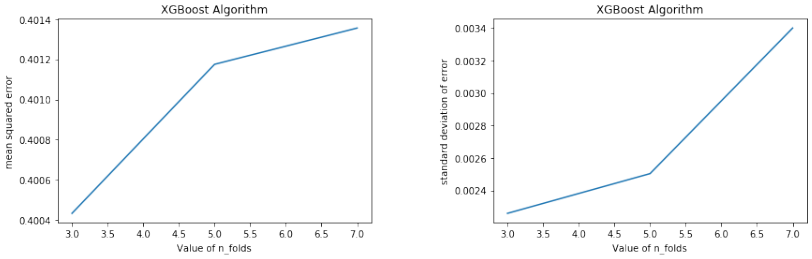
We have used **Ensemble Methods** which are a combination of various models referred to as Base Learners into a final model which is the Meta Learner.
It aims to combine diverse characteristic models into one so as to analyse wider aspects of our problem. Larger number of base learners help in maximising the accuracy and performance of our model and minimising the generalisation error. Base learners used are as follows:

- XGBoost Classifier
- Random Forest Classifier
- Gradient Boosting Regressor

5.1 XGBoost Classifier

XGBoost stands for e**X**treme **G**radient **B**oosting. It is a decision tree based ensemble algorithm which provides parellel processing and tree pruning making it fast and accurate.

It is a powerful algorithm that deals with missing values and regularisation which helps in avoiding overfitting. XGBoost helps in using the resources efficiently along with best computation time.



(a) Mean squared error

(b) Standard deviation of error

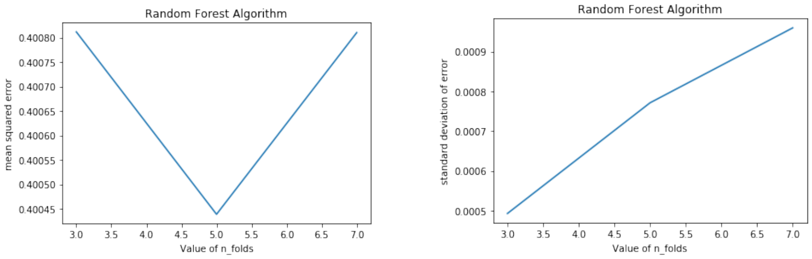
Fig. 5: Scores of XGBoost Algorithm for n_folds = 3, 5, 7

Conclusion : From the graphical representations which shows the mean squared error and standard deviation of error for different values of folds in cross-validation , XGBoost Classifier is giving least error for fold value 3 for both mean squared error and standard deviation of error .

5.2 Random Forest Classifier

Random Forest is a classification algorithm which is based on class predictions made by individual decision trees. Here, a large number of decision trees operate as an ensemble. Each decision tree splits up into class prediction and the class that receives maximum number of votes is the final predicted class.
The basic idea behind Random Forest is that if a group of many uncorrelated trees work together, they will produce more efficient results than any of the individual tree alone. Due to low correlations between the models, some trees are wrong and some are correct. Therefore, collectively they move in the required direction.

Conclusion : From the graphical representations which shows the mean squared error and standard deviation of error for different values of folds in cross-validation , Random Forest Classifier is giving least mean squared error for fold value 5 and least standard deviation of error for fold value 3.

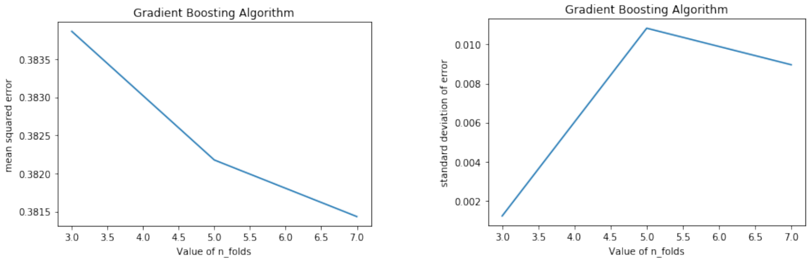


(a) Mean squared error (b) Standard deviation of error

Fig. 6: Scores of Random Forest Algorithm for n_folds = 3, 5, 7

5.3 Gradient Boosting Regressor

Gradient Boosting is a method of converting weak learners into strong learners. In boosting, each new tree is a fit on a modified version of the original data set. After evaluating the first tree, we identify the observations that are difficult to classify. The second tree is therefore grown to deal with the shortcoming of the first tree. Here, the idea is to improve upon the predictions of the first tree. Our new model is therefore Tree 1 + Tree 2. We then compute the classification error from this new 2-tree.
Here, we use the loss function to find the shortcomings of a particular decision tree. the loss function is defined as y = ax + b + e, where e is the error term. Our goal is to optimise the loss function in every subsequent tree made.



(a) Mean squared error (b) Standard deviation of error

Fig. 7: Scores of Gradient Boosting Algorithm for n_folds = 3, 5, 7

Conclusion From the graphical representations which shows the mean squared error and standard deviation of error for different values of folds in cross-validation,Random Forest Classifier is giving least mean squared error for fold value 7 and least standard deviation of error for fold value 3.

5.4 Stacking Models : Averaging Base Model

Averaging Base Model is a simple model in which the predictions of base models are averaged to yield better results than the results of a single model.
By averaging our base models : XGBoost Classifier, Random Forest Classifier and Gradient Boosting Regressor we got a significant decrease in the mean squared error and standard deviation of error for fold value 3. We chose fold value 3 because it showed least errors in our base models. The averaging base model approach also ensures that the results are not biased , since our data was imbalanced.

6 Future Scope

The binary classification of data can be extend to multiclassification where the machine can also predict numerous factors along with predicting the eligibility of the borrower.
Also more complex ensemble techniques can be used for predictions to deal with wider range of data.

References

1. Carlos Serrano-Cinca, Begoña Gutiérrez-Nieto, L.L.P.: Determinants of default in p2p lending (2015)
2. Shunpo Chang, Simon Dae-oong Kim, G.K.: Predicting default risk in lending club loans (2015)