ECE657A Report: The Prediction for the Bank Loans

Group 6
Xiyue Zhang 20601564 , Chang Liu 20632304, Su Wang 20625858

University of Waterloo

Abstract. This project focuses on binary classification problem. Given the financier basic information, we make prediction about the probability of the loan risk. Firstly we have detected the dataset feature and fixed missing or extreme value. Then we construct three classifiers (Logistic Regression, Decision Tree and Random Forest) and compare their classification performance by Python. Finally we choose the best classifier to make prediction. According to the experimental result, we will give some improvement for the further work.

Keywords: Binary classification, Classifiers, Python

1 Introduction

Nowadays, the accurate prediction of loans is very important for creditors or customers. The loan comes from different place like Commercial bank or Investing companies or online P2P lending platform where people can easily lend or borrow money. However it also increases the probability of loan risk. From investor perspective, they show concern about the possible of fraud. If the loan is started by single person, how they can know the financier repay money on time, which feature can greatly influence the loan risk, and which model is better to make accurate prediction. All these problems will be discussed in this paper.

We plan to detect the loan risk based on the financier information [9]. The dataset selected describes financier data by the features such as gender, education, employment, applicant income, co-applicant income, loan amount, loan term and credit history. The whole dataset is divided into training set and testing set. Different from testing set, the training set includes the label to show the loan is worthy to investing or not. The label is corresponded to Yes or No. Yes means we can invest on the trustful loan.

2 Literature Review

Some papers introduce data processing methods. For example, IG (information gain) can eliminate weak-related features [4] but it is only used for multi-class classification problem. Log transformation combines extreme value with other value within a range rather than deleting extreme value. Because some features like loan amount may contain large value. It should be consider as reasonable training data.

For classifier, the Bayesian classifier is based on the hypothesis of class independency that is hard to meet in reality [4]. The neural network is a black-box whose structure weight values are the hidden knowledge for classification, which is difficult for ordinary investors and finance majors to understand [4]. The SVM takes time to choose the right kernel function and parameter to train the best model. The Random Forest and Logistic Regression is better to classify financial data in terms of accuracy. However, Random Forest may lead to over-fitting problem that one of classification accuracy and cross-validation score is too high and the other one is low. The deep trees of Random Forest are susceptible to over-fitting. We can stop growing the tree when the data split is not statistically significant. Also we can use method called pruning, is to grow a full-sized deep tree and then to reduce the tree depth by recursively merging leaf nodes [6]. In this project we compare the accuracy with cross-validation score to decide proper parameters for Random Forest

For the imbalanced data, if we have 5% risky loan and 95% trustful loan, the prediction will be biased to trustful loan because the classifier is more sensitive to detecting the majority class and less sensitive to the minority class. So we need to preprocess imbalanced data before feeding them into a classifier. The first method is under-sampling the majority class. However the prediction may be poor because we throw away some useful information in the majority class. Actually the poor prediction may have nothing to do with under-sampling but due to few samples or bad features with poor discriminative power. So we need to detect the real reason of poor prediction to decide whether or not to use under-sampling. The second method is oversampling the minority class. We can just duplicate the sample or use SMOTE which creates new sample based on nearby sample. To avoid over-fitting, the oversampling should be inside the cross-validation loop and after the step to split validation set and training set. However the dataset we selected is not that imbalanced. There are 30% loan with label NO and 70% loan with label YES, so the trained classifier maybe slightly biased to label YES, but it will not influence much about the prediction accuracy.

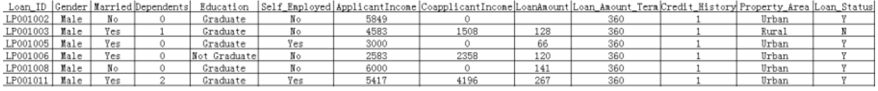
For the classification metrics, we can compare accuracy, recall or F-measure of each classifier to choose the best one. The recall is better than accuracy to analyze the financial problem because it gives a good indication about how the classifier performs on finding the positive instances. The classifier with higher recall is able to help investors avoid bad loans as many as possible [5], and the classifier with higher accuracy will help aggressive investors who are willing to take more risk to get every possible opportunity, so the proper classifier should be based on investors need

In addition, we can choose the right feature for training manually to reduce the cost of machine learning. The important feature to detect loan risk depends on loan type. For traditional bank loan, the credit history is important to predict loan risk. But for P2P lending loan, the loan interest rate is also important because the financier prefers to raise interest rate to attract investor on P2P platform, meanwhile it increase the loan risk [5]. The interest rate of traditional bank loan is always stable to avoid bankrupt.

3 Methods and Results

3.1 Dataset Introduction

The dataset for the project is the data of bank loan. Fig. 1 is the simple of the dataset.



Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
LP001002	Male	No	0	Graduate	No	5849	0		360	1	Urban	Y
LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
LP001006	Male	Yes	0	Not Graduate	No	2583	2358	120	360	1	Urban	Y
LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y
LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	Y

Fig. 1. Simple of Dataset

3.2 Data Preprocessing

Missing Values and Outliers Detection As we all know that in order to get the final correct results of the data analysis, the original dataset must be correct at first. So at first we check the dataset to see whether there are missing values and outliers.

Fig. 2 is the summary of numerical field of dataset. From the summary, we can get the count, mean, standard deviation, min, quartiles and max for each attribute.Based on these information of the summary, we can get the following conclusion:

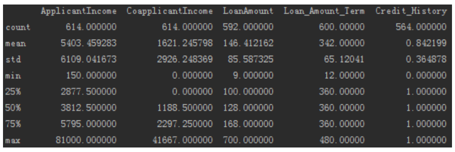
(1) There are (614 – 592) 22 missing values for LoanAmount.

(2) There are (614 – 600) 14 missing values for Loan_Amount_Term.

(3) There are (614 – 564) 50 missing values for Credit_History.

(4) About 84% applicants have a credit_history.

(5)ApplicatInclimie distribution and CoapplicatIncome seem to be in line with expectation.



```
      ApplicantIncome  CoapplicantIncome  LoanAmount  Loan_Amount_Term  Credit_History
count      614.000000         614.000000  592.000000         600.00000      564.000000
mean      5403.459283        1621.245798  146.412162         342.00000        0.842199
std       6109.041673        2926.248369   85.587325          65.12041        0.364878
min        150.000000           0.000000    9.000000          12.00000        0.000000
25%       2877.500000           0.000000  100.000000         360.00000        1.000000
50%       3812.500000        1188.500000  128.000000         360.00000        1.000000
75%       5795.000000        2297.250000  168.000000         360.00000        1.000000
max      81000.000000       41667.000000  700.000000         480.00000        1.000000
```

Fig. 2. The Summary of Numerical Field of Dataset

For the non-numerical values, like Property_Area, Credit_History etc, we can check its frequency distribution to see whether they make sense or not.

Fig. 3 is the histogram of ApplicationIncome (bins = 50). Fig. 4 is the box Plot of ApplicationInmcome. From these figures, we can observe that a lot of outliers/extreme values exist. Fig. 5 is the box plot of ApplicationIncome by Gender, Fig.6 is the box plot of ApplicationIncome by Education. From these two figures, we can see the gender and education have the effect on the ApplicationIncome.

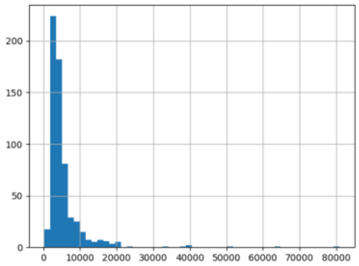


Fig. 3. The Histogram of ApplicationIncome

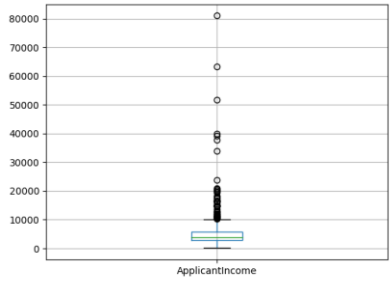


Fig. 4. The Box Plot of ApplicationInmcome

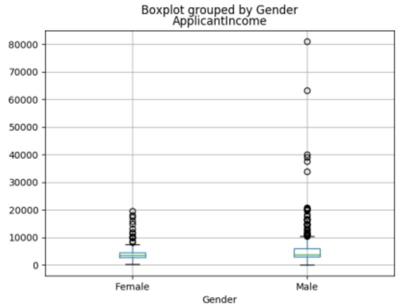


Fig.5. Box Plot of ApplicationIncome by Gen

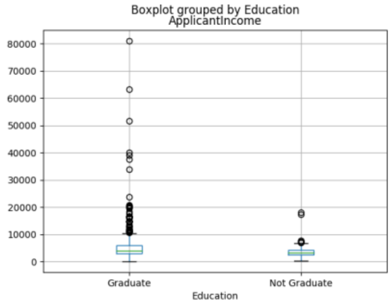


Fig.6. Box Plot of ApplicationIncome by Edu

Fig. 7 is the histogram of LoanAmount, Fig. 8 is the box plot of LoanAmout. From the two figures, we also can know there are some extreme values.

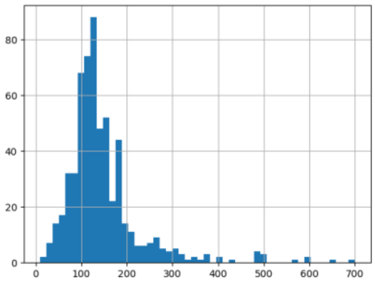


Fig. 7. The Histogram of LoanAmount

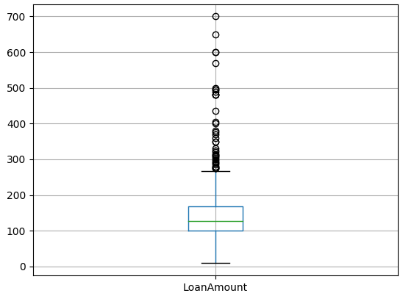
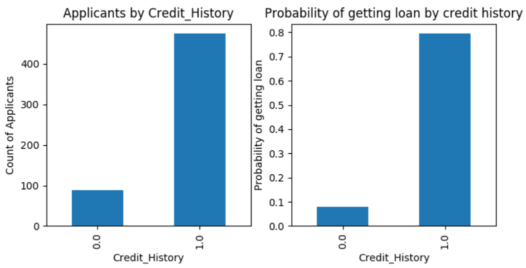


Fig. 8. The Box Plot of LoanAmount

Categorical Variable Analysis Fig. 9 is the loan status of applicants by credit history. From the figure, we can see the chances of getting a loan based on credit history. And the chances of getting a loan are eight-fold if the applicant has a valid credit history. Fig.10 is the loan status of applicants by gender and credit history. Based on this figure, we can know that the chances of getting a loan for the male are higher than the female. Meanwhile, we can also do the same analysis for the married and dependents etc.



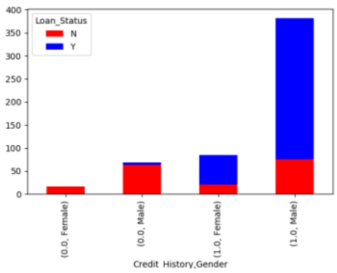


Fig. 9. Loan Status by Credit History **Fig.10.** Loan Status by Gender and Credit History

Missing Values and Extreme Values Process We know there a lot of missing values for some attributes and some extreme values for LoanAmount and ApplicationIncome. So before data mining, we should process these values.

Missing Values Process Because this is a loan dataset, many attributes do not change timely or uniformly, they are the ladder graph, we can use the mean value to replace the missing values. Here we used median value and most likely value to process the missing values based on the type of the attributes of the data.

Most Likely Value Process For the attributes, Credit History, Self Employed, Gender, Dependents, Married, they are not numeric, so we can use the most likely value to replace the missing values. Take Credit History for example. Table 1 is the count of credit History. We can see 475 values equal 1.0 and 89 values equal 0.0. The most likely value for credit history is 1.0, so use 1.0 to replace the missing values of credit history. Do the same process to the Self Employed, Gender, Dependents, Married.

Table 1. The Count of Credit History

value	count
1.0	475
0.0	89

Median Value Process For the LoanAmout and LoanTerm, we use their median value to replace the missing values separately.

For example, the LoanAmount process. We did not use the mean value to process the missing values simply. We wanted to find out whether combining some other attributes could give us some estimation about the loan amount. Fig. 11 is the box plot of LoanAmount by Education and Self_employed. Firstly, detect how Educated and Self_employed influence loan amount. Secondly, divide applications into 4 groups by educated and self_employed. We found different educated and self_employed indeed influence the loan mount. Thirdly, replace the missing values with the median value in each group.

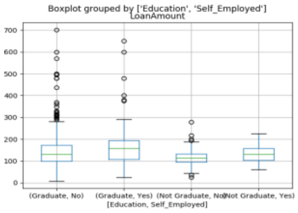


Fig. 11. The Box Pot of LoanAmount by Education and Self_employed

Extreme Values Process Usually, for the extreme values, we just treat them as the outliers and delete them directly. But sometimes these extreme values are the real data rather outliers, so we cannot just delete them. Here, we use Log Transformation to process the extreme values for the loan Amount and Application Income.

For loan amount, some applicants may apply for high loans due to specific need; for application income, some applicants have lower income but strong support Co-applicants, so combine both incomes as total income for Application Income. Finally, use Log Transformation to nullify their effect instead of treating them as the outliers.

For example of Loan Amount. Fig.12 is the histogram of Loan Amount before and after Log Transformation. We can see that Log transformation can help to spread the points more uniformly rather than tight cluster in the graph. The loan amount will be more interpretable within an range than in the point after Log Transformation.

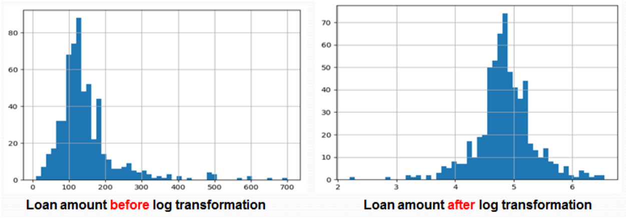


Fig. 12. Histogram of Loan Amount before and after Log Transformation

3.3 Classification

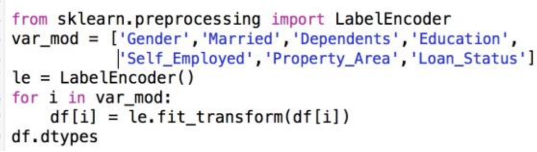
In this part, we will build three mathematic models, Logical Regression, Decision Tree, Random Forest , to classify the precessed data, and based on these models , to analysis the results(accuracy and cross-validation score)whether the bank will offer the loans to customers. In order to complete it, we will use sklearn as the machine learning library in Python. And I will introduce these three mathematic models briefly in the following.

For Logical Regression, it is a regression model where the dependent variable (DV) is categorical.The binary logistic model is used to estimate the probability of a binary response based on one or more predictor (or independent) variables (features). It allows one to say that the presence of a risk factor increases the probability of a given outcome by a specific percentage[1].

Secondly, Decision Tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm.Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning[2].

Finally, Random Forest is an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set[3].

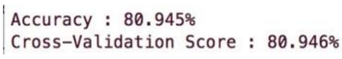
Data Encoding From the original data, we can see that some categorical variables, such as Loan_Status, are not numeric, just Yes or No. However, sklearn requires all inputs to be numeric. So we should convert all categorical variables into numeric by encoding the categories firstly.



```
from sklearn.preprocessing import LabelEncoder
var_mod = ['Gender','Married','Dependents','Education',
           'Self_Employed','Property_Area','Loan_Status']
le = LabelEncoder()
for i in var_mod:
    df[i] = le.fit_transform(df[i])
df.dtypes
```

Fig. 13. Transformation to be numeric

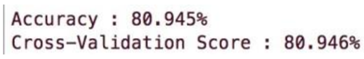
Logical Regression For this mathematic model, we will use three predictors as modules with different variables to analyze the influence of each variable for the final results. Among these three predictors, ‘Credit_History' is the common variable. And in the second and third predictors, we select other 4 variables for combination. We use the first predictor: predictor_var1 = [‘Credit_History’].The result is as following:



```
Accuracy : 80.945%
Cross-Validation Score : 80.946%
```

Fig. 14. Logical Regression predictor1

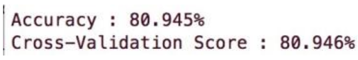
For the other two predictors:
predictor_var2
=[‘Credit_History’,'Education','Married','Self_Employed','Property_Area']



```
Accuracy : 80.945%
Cross-Validation Score : 80.946%
```

Fig. 15. Logical Regression predictor2

predictor_var3 =[‘Credit_History’,'Gender','Married','Education’,’Dependents’**]**

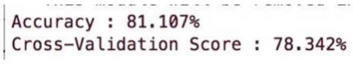


```
Accuracy : 80.945%
Cross-Validation Score : 80.946%
```

Fig. 16. Logical Regression predictor3

From the above results, it can be shown that all of three predictors have the same value of Accuracy and Cross-Validation Score. Therefore, a conclusion can be got that the variable “Credit_History” is dominating the model, and other variables almost have no influence on the model in this aspect.

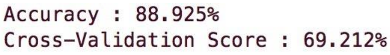
Decision Tree In this mathematic model, similar to the Logical Regression, we will use two predictors with different number variables for comparison. And “Credit_History” is still the common variable for two predictors.
predictor_var1 = [‘Credit_History','Gender','Married','Education','Dependents']



```
Accuracy : 81.107%
Cross-Validation Score : 78.342%
```

Fig. 17. Decision Tree predictor1

predictor_var2 = [‘Credit_History','Loan_Amount_Term','LoanAmount_log']

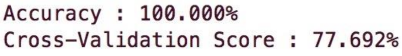


```
Accuracy : 88.925%
Cross-Validation Score : 69.212%
```

Fig. 18. Decision Tree predictor2

Firstly, we can see that the difference between “Accuracy” and “Cross-Validation Score” is large. So it results from the overfitting. Then comparing the results of predictor1 with that of Logical Regression, it can conclude the Decision Tree has a better performance than Logical Regression for the accuracy.

Random Forest In this model, we will use two predictors to analyze.
1> Frist Predictor

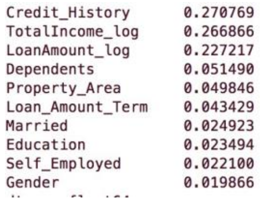


```
Accuracy : 100.000%
Cross-Validation Score : 77.692%
```

Fig. 19. Random Forest predictor1

In the first model, we will use all the features in the fixed data and the default parameters. Thus, we choose “n_estimators” as 100. It is clear that this predictor is overfitting due to 100% Accuracy. Therefore we need to use the second predictor.

2> Second Predictor In this predictor, we will select top 5 important features in the first models as the variables, and use the optimized parameters by plotting.



```
Credit_History      0.270769
TotalIncome_log     0.266866
LoanAmount_log      0.227217
Dependents          0.051490
Property_Area       0.049846
Loan_Amount_Term    0.043429
Married             0.024923
Education           0.023494
Self_Employed       0.022100
Gender              0.019866
```

Fig. 20. Top 5 features in the first model

From the above results,the values in the second column represents its importance to the results.Therefore,"Credit_History", "TotalIncome_log", "LoanAmount_log", "Dependents", "Property_Area" are selected as the predictor variables.

In order to get the optimized parameters, we plot four graphs for each parameters, "n_estimators", "min_samples_split", "max_depth", "max_features".

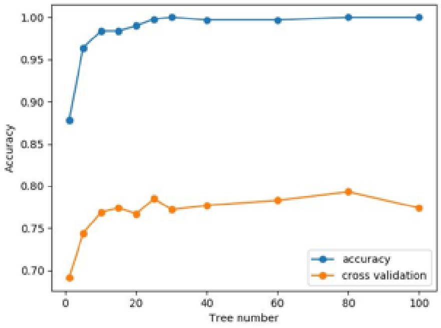


Fig. 21. Tree Number

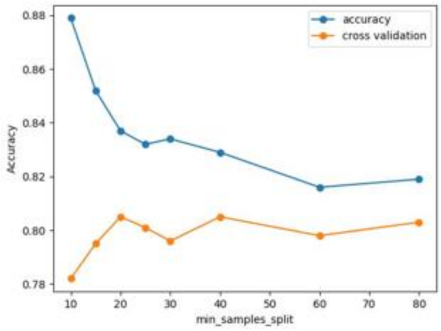


Fig.22. Min_Samples_Split

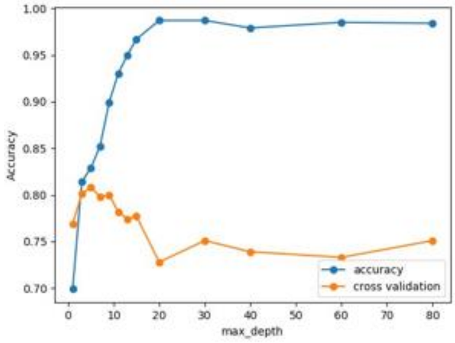


Fig. 23. Max_Depth

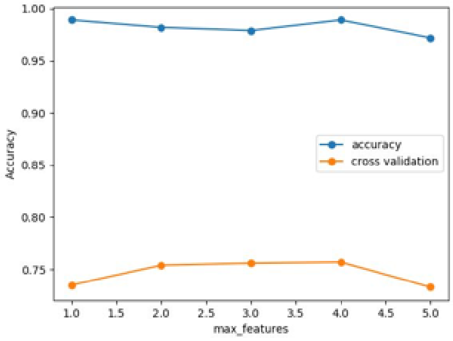
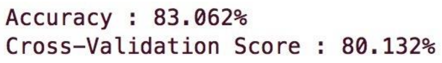


Fig.24. Max_Features

As shown in Figure 3.10-3.13, the blue line is accuracy and orange line is cross validation score for different parameter. To avoid overfitting, we should choose the model whose accuracy is similar to cross validation score. So we can confirm the optimized parameters, “n_estimators”=25, “min_samples_split”=20, “max_depth”=7, “max_features”=1.



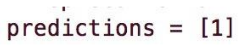
```
Accuracy : 83.062%
Cross-Validation Score : 80.132%
```

Fig. 25. Random Forest predictor2

Based on the accuracy of Random Forest, we can make a conclusion that among three mathematic models, Random Forest has the best performance for classification and predict the accuracy.

Results Validation In order to take one case for our model validation, we will select one case from the original data,(Loan_ID : LP001003 ApplicantIncome: 4583 Coapplicant Income: 1508 LoanAmount: 128 Credit_History: 1 Dependents: 1 Property_Area: Rural LoanStatus : Y).

Then the predictions will be made. During the process, we will use the best one the second model of Random Forest in this aspect.



```
predictions = [1]
```

Fig. 26. The predictions results

The value “1” just represents Yes, which means the bank will offer the loans to the customer, and is just equal to the real Loan_Status in the original data. So our classification and predictions models have a great performance.

4 Conclusion

This paper illustrates some approaches to detect loan risk. We define this problem into a binary classification problem, preprocess the data before the classification, and train three classifiers to predict the classification result. Finally we select the best classifier and best parameter to reach high accuracy and avoid over-fitting. So we can see that processing, classification and evaluation is different from other classification problem because it focuses on financial area. The imbalanced problem should be considered. The missing value should be replaced by proper value and extreme value should be combined rather than removed according to the feature. The important feature selected should be based on practical knowledge rather than reduce the dimension by PCA or LLE. Different metric like recall or accuracy can help different investor to predict the loan risk. So based on these knowledge, we can make conclusion that Random Forest with proper parameter works best to predict the dataset we selected.

References

1. Wikipedia for Logic Regression.https://en.wikipedia.org/wiki/Logistic_regression..
2. Wikipedia for Decision Tree. https://en.wikipedia.org/wiki/Decision_tree
3. Wikipedia for Random Forest. https://en.wikipedia.org/wiki/Random_forest
4. Sun, Jie, and Hui Li. "Data mining method for listed companies' financial distress prediction." Knowledge-Based Systems 21.1 (2008): 1-5.
5. Wu, Jiayu. "Loan Default Prediction Using Lending Club Data." (2014).
6. Pandey, Jitendra Nath. "Predicting Probability of Loan Default Stanford University, CS229 Project report Jitendra Nath Pandey, Maheshwaran Srinivasan."
7. Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.
8. API design for machine learning software: experiences from the scikit-learn project, Buitinck *et al.*, 2013.
9. Financier Data https://www.dropbox.com/s/hbsi3c9x7sf2j8a/train_loanspredict.csv?dl=0