Identifying Profitable LendingClub Loans using Decision Tree

Ngoc Phan
M.S. Artificial Intelligence
University of North Texas
Denton, TX
ngocphan@my.unt.edu

April 23, 2021

I. Introduction

LendingClub (LC) is an online peer-to-peer (P2P) platform that brings borrowers and lenders together. To apply for a loan, the borrower first submits a loan application to LC. Upon reviewing the application and borrower's credit history, LC may either approve or reject the application. Once the loan application has been approved, specific information related to the loan will be made available to the lenders. The lender then selects the loans that he or she wants to invest. The lender can also invest in either a fractional part of a loan or a whole loan.

Given a large number of loans available for investment, without careful consideration during the loan selection process, the lender may not maximize his or her return on investment (ROI) in two ways: (1) the borrowers pay off the loan too early which minimize the lender's ROI, and (2) the lender invests in a loan that has low ROI or a loan that would be defaulted in the near future. To enable profit maximization, the lenders would need to identify profitable loans and the length of the loans. This study focuses on building a Decision Tree classifier that would give the lenders the ability to classify profitable loans as well as unprofitable loans. Given the time constraint, an analysis has been performed to identify the relationship between loan's profitability and the length of loan.

II. Related Works

There have been many studies on credit risk prediction of LC loans. Lee, Jei Young [1] analyzed 2012 and 2013 LC's loan data and used a binary probit model with the Instrumental Variable (IV) method to identify a default loan. The most commonly used attributes were selected and broadly categorized into two groups: loan characteristics and borrower characteristics. Loan characteristics include attributes such as interest rate, loan amount, loan term, and dummies for loan purpose. Borrower characteristics include attributes such as income, home ownership, debt-to-income ratio (DTI ratio), revolving credit balance, number of past-due delinquencies, and borrower's overall grade assigned by LC. Lee, Jei Young [1] also created two derived attributes based on loan description by extracting sentiment score using sentiment analysis and counting the number of words for each description. Results of the study showed that borrowers having negative sentiment score on loan description are likely to get higher interest rate. The study also indicated that both loan characteristics and borrower characteristics were significantly associated with loan default rate.

Li, Peiqian & Han, Gao [2] studied LC's loan data from 2012 to 2015 and compared three classification methods (logistic regression, neural network, and random forest) for identifying credit risk loans using confusion matrix. Loans having status such as "current" and "late" were not used in the study because those loans are on-going and not yet final. For feature selection, columns that have missing values for most of the rows and columns that have same values across all rows were removed. Li, Peiqian & Han, Gao [2] also categorized the attributes with missing values into three cases: mean-set, zero-set, and max-set. For mean-set attributes, missing values were replaced with the mean. For max-set attributes, missing values were replaced with a constant factor multiplying the attribute's maximum value. For zero-set attributes, missing values were replaced with zeros. In addition to confusion matrix, the following metrics were also used to evaluate the result of classification methods: precision, recall, f1-score, support, and weighted average. Results of the study showed that the three classification methods have the same (or almost the same) weighted average, and the default loan classifier achieved 89% in terms of both weighted precision and recall metrics.

III. METHODOLOGIES

Some of the techniques mentioned in the previous works have been used in the project as follows:

- Li, Peiqian & Han, Gao [2]
  - Filter out loans with status such as "current" and "late" because those loans are not yet final.
  - Evaluate classification model using confusion matrix and the following metrics: precision, recall, and F-measure.
- Lee, Jei Young [1]
  - Include the most common attributes mentioned in the article in the study because results of previous study showed that those attributes were significantly associated with loan default rate.
  - Convert the following nominal attributes to symmetric binary attributes:
    - Loan term as a dummy where 0 indicates "36 months" and 1 indicates "60 months".
    - Home ownership as a dummy where 0 indicates rent, mortgage or other, and 1 indicates "own".
  - Convert ordinal attribute, Grade, to rank where A, B, C, D, E, F, G are assigned value 7, 6, 5, 4, 3, 2, 1, respectively.

In addition to the aforementioned techniques, the following techniques have also been applied in the project:

- Create the following derived attributes:
  - ROI = total payment – funded loan amount
  - Months in loan = last payment date - loan issued date
- Categorize the following continuous attributes: ROI, loan amount, interest rate, annual income, DTI ratio, revolving balance, and number of previous delinquencies.
- Implement classification model using Decision Tree induction to identify profitable loans.

IV. DATASET

LC's loan dataset is obtained on Kaggle. The dataset contains 2,260,701 accepted loans from 2007 to 2018 and 151 attributes (113 quantitative attributes, 9 date attributes, and 29 qualitative attributes). Among the accepted loans, there are 912,569 loans that are not yet final or still in-progress. There are 58 attributes that have at least 25% missing values, and 92 attributes that have at most 14% missing values. Table I shows the number of missing values for the attributes selected in the study after removing in-progress loans from the dataset.

TABLE I. MISSING VALUES

Attribute (s)	*Count*	*Percentage*
Loan amount, funded amount, interest rate, total payment, loan term, loan issued date, loan purpose, revolving balance, borrower's overall grade, home ownership	31	0.0023
Last payment date	2,356	0.1748
Loan description	1,222,175	90.6571
Annual income	35	0.0026
Debt-to-income ratio (DTI ratio)	405	0.03
Number of previous delinquencies	60	0.0045

Since loan description has about 91% of missing values, the attribute has been excluded from the study. All other attributes have a very low percentage of missing values, so all examples with those missing values have also been removed from the dataset. After removing irrelevant examples and examples that have missing values, there are remaining 1,345,375 examples to be included the study.

Table II shows summary statistics for attributes included in the study. The mean of funded loan amount is 0.001 lower than that of loan amount, which indicates that there are a small number of loans that had not been fully funded.

TABLE II. SUMMARY STATISTICS

Attributes	*Mean*	*Median*	*Std. Dev.*	*Min*	*Max*
ROI[c]	0.044	0.113	0.533	-3.997	2.83
Months in loan[d]	20.9	20	12.6	0	70
Interest rate	13.24	12.74	4.76	5.31	30.99
Loan amount[a]	1.441	1.2	0.871	0.05	4
Funded amount[a]	1.440	1.2	0.871	0.05	4
Loan term[e]	0.24	0	0.43	0	1
Home ownership[f]	0.11	0	0.31	0	1
Annual income[a]	7.626	6.5	6.994	0.0016	1099.92
DTI ratio	18.276	17.61	11.151	-1	999
Revolving balance[a]	1.628	1.114	2.247	0	290.484
Overall grade[b]	5.25	5	1.3	1	7
Num. of previous delinquencies	0.32	0	0.88	0	39

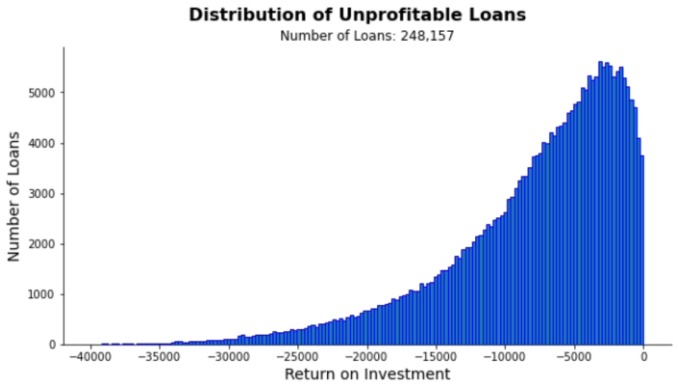
Note. [a]Values are in 10,000s of dollars. [b]Overall grade has seven values A, B, C, D, E, F, G which were assigned as 7, 6, 5, 4, 3, 2, 1, respectively. [c]ROI = total payment – funded loan amount. [d]Months in

loan = last payment date - loan issued date. [e]Loan term as a symmetry binary attribute, where 0 = 36 months, 1 = 60 months. [f]Home ownership as a symmetry binary attribute, where 1 = own, 0 = rent, mortgage, any, none or other.

V. Exploratory Data Analysis

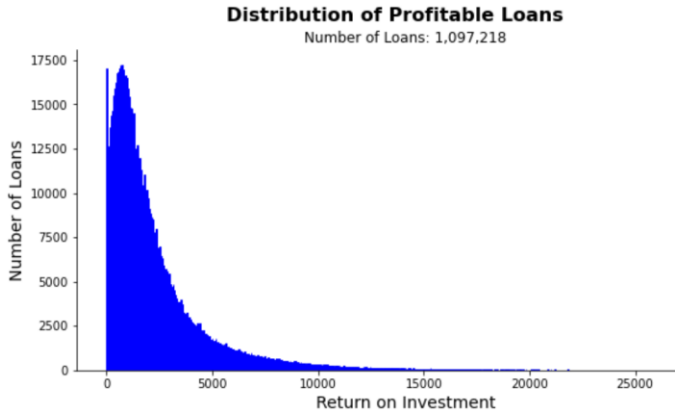
The distribution of unprofitable loans is left-skewed which indicates that the mean ROI is less than the median ROI, which is in turn less than the mode ROI (Figure I). There are a vast number of unprofitable loans that have ROI between negative $5,000 and $0. There are also a small number of unprofitable loans that have ROI less than negative $20,000. The total number of unprofitable loans is 248,157.

Figure I. Distribution of Unprofitable Loans



The distribution of profitable loans is right skewed which indicates that the mode ROI is greater than the median ROI, which is in turn greater than the mean ROI (Figure II). There are a vast number loans that have ROI between $0 and $3,000 while there are a small number loans that have ROI above $5,000. The total number of profitable loans is 1,097,218.

Figure II. Distribution of Profitable Loans



Since ROI is not normally distributed, ROI has been categorized based on two partitions: unprofitable loans and profitable loans. Unprofitable loans include loans that have ROI less than or equal to zero. Profitable loans are further divided into three partitions: Low (ROI > 0 and ROI <= 2,000), Medium (ROI > 2,000 and ROI <= 4,000), and High (ROI > 4,000). The categorized ROI is named as "loan return" and would be used as the target variable for developing Decision Tree classifier. Figure III shows the loan proportion in percentage per Loan Return.

Figure III. Loan Return

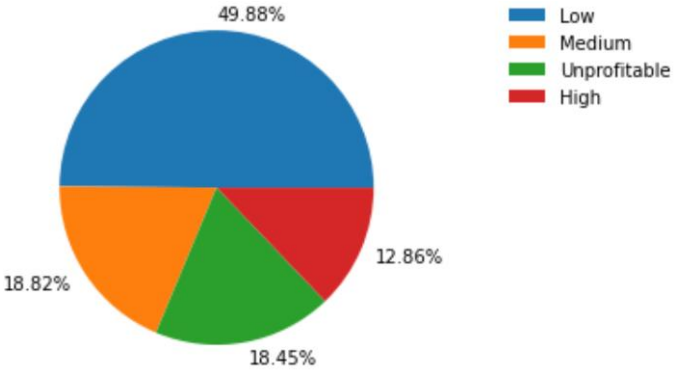


Table III shows six continuous attributes that have been categorized using the low value, high value, and step size. Low values are being categorized as "<= low", while high values are being categorized as "> high". The step values in between low and high are categorized as "(value_1 to value_2]" where value_1 is exclusive and value_2 is inclusive.

Table III. Categories of Continuous Attributes

Attributes	***Low***	***High***	***Step Size***
Loan Amount	10,000	30,000	10,000
Interest Rate	10	20	5
Annual Income	20,000	100,000	20,000
DTI Ratio	5	30	5
Revolving Balance	5,000	20,000	5,000
Num. of previous delinquencies	1	2	1

Figures IV to IX display the data distribution per category for the above six continuous attributes sorted by number of loans in descending order. Most borrowers applied for a loan amount of at least $20,000, and have an annual income greater than $20,000, a DTI ratio between 10 and 25, a number of past delinquencies less than 2, and a revolving balance greater than $20,000 or between $5,000 and $10,000. Many borrowers received an interest rate of at most twenty percent.

FIGURE IV. DISTRIBUTION OF LOAN AMOUNT

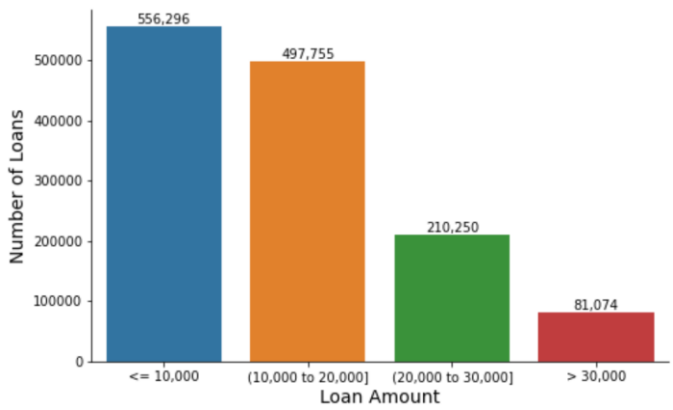


FIGURE V. DISTRIBUTION OF INTEREST RATE

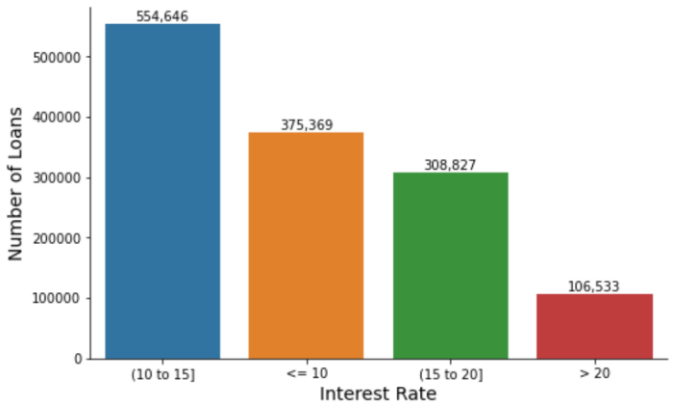


FIGURE VI. DISTRIBUTION OF ANNUAL INCOME

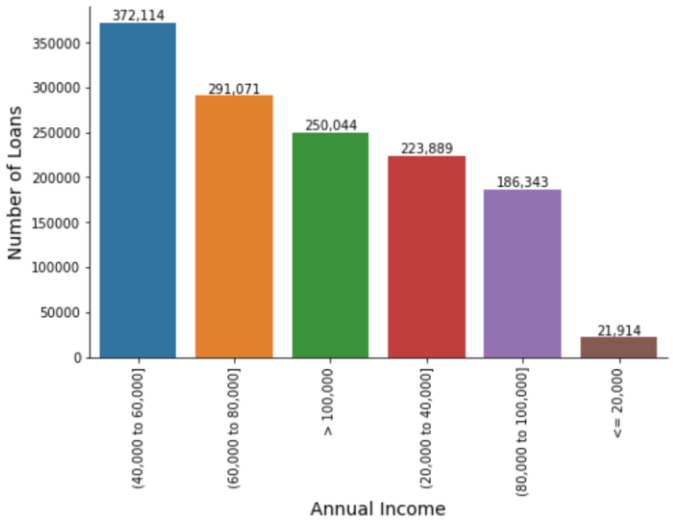


FIGURE VII. DISTRIBUTION OF DTI RATIO

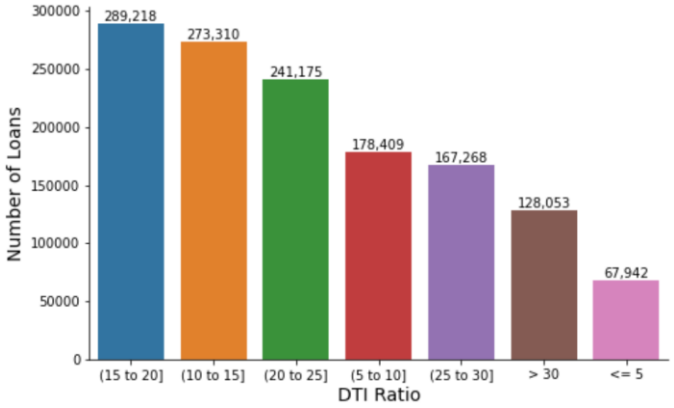


FIGURE VIII. DISTRIBUTION OF REVOLVING BALANCE

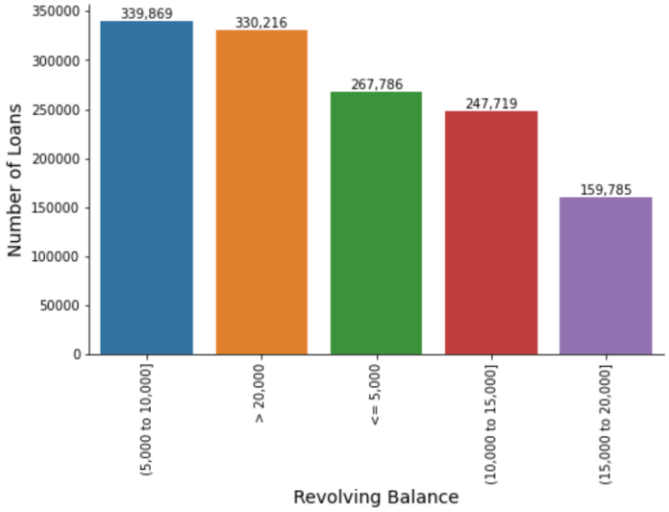
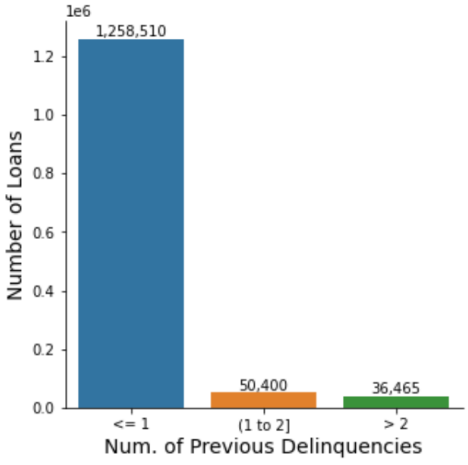


FIGURE IX. DISTRIBUTION OF NUM. OF PREV. DELINQUENCIES



As indicated by Figure X, the longer the borrower stays in a loan, the higher is the ROI. Most unprofitable loans have less than 50 months in loan. Loans that have ROI less than negative $20,000 have shorter length of loans which is less than 20 months. There are also some loan accounts

that had been closed immedidately in the same month the loans had been awarded where Months in Loan is zero.

FIGURE X. MONTHS IN LOANS VS. ROI

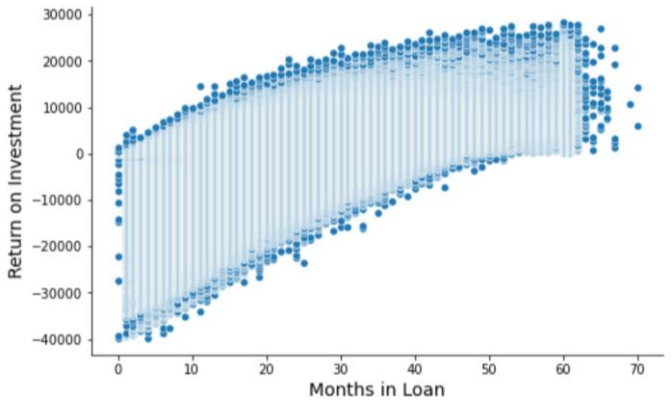


Figure XI show the correlation coefficient among attributes. The correlation score has been calculated using the following formula:

$$r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$$

where

- Covariance between A and B is computed as follow:

$$Cov(A,B) = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n}$$

- n is the number of tuples.
- $\bar{A}$ and $\bar{B}$ are the respective mean or expected values of A and B.
- σ_A and σ_B are the respective standard deviations of A and B.

FIGURE XI. PEARSON CORRELATION

Note. Attribute's name denoted as A-K, where A = loan amount, B = interest rate, C = loan term, D = annual Income, E = DTI Ratio, F = revolving balance, G = home ownership, H = number of delinquencies, I = ROI, J = months in loan, K = borrower's overall grade.

According to Figure XI, interest rate (B) has strong negative correlation (-0.95) with borrower's overall grade (K). This indicates that borrower who has low overall grade tends to receive high interest rate. Furthermore, the positive correlation (0.43) between ROI (I) and months in loan (J) also indicates that the longer the borrower stays in a loan, the higher the ROI the lender receives. Additionally, borrower who applies for longer loan term tends to receive higher interest rate because there is a positive correlation (0.42) between interest rate (B) and loan term (C).

VI. DATA SAMPLING

Stratified random sampling have been used to prepare data for model's training and testing. Seventy percent of number of examples in each Loan Return category has been selected for training the classification model, while the remaining thirty percent has been used for testing the model.

VII. CLASSIFICATION MODEL

Decision Tree induction has been used to develop Loan Return classifier. Figure XII shows the Decision Tree algorithm used in developing a classification model mentioned by Tan, Pang-Ning & Steinbach, Michael & Kumar, Vipin [3].

FIGURE XII. DECISION TREE ALGORITHM

```
TreeGrowth (E, F)
1: if stopping_cond(E,F) = true then
2:    leaf = createNode().
3:    leaf.label = Classify(E).
4:    return leaf.
5: else
6:    root = createNode().
7:    root.test_cond = find_best_split(E, F).
8:    let V = {v|v is a possible outcome of root.test_cond }.
9:    for each v ∈ V do
10:      E_v = {e | root.test_cond(e) = v and e ∈ E}.
11:      child = TreeGrowth(E_v, F).
12:      add child as descendent of root and label the edge (root → child) as v.
13:   end for
14: end if
15: return root.
```

The choice of the test condition (*test_cond*) has been determined using entropy as the impurity measure of a given node [3].

$$Entropy(t) = -\sum_{i=0}^{c-1} p(i|t)\, log_2 p(i|t)$$

where

- $p(i|t)$ denote the fraction of records belonging to class i at a given node t.
- c is the number of classes
- $0log_20 = 0$ in entropy calculation.

The stopping conditions to terminate the tree-growing process are given below:

- Testing whether all the records have either the same class label or the same attribute values.
- Testing whether the number of records has fallen below some minimum threshold based on the value of the hyper-parameters such as *max_depth*, and *min_samples_leaf*, etc.

Below is the specification of the Decision Tree model #1:

- Target variable: Loan Return
- Features (all categorical): loan amount, interest rate, loan term, loan purpose, annual income, DTI ratio, revolving balance, home ownership, and number of previous delinquencies.
- Multi-ways split is used for node splitting.
- Parameters:
  - Maximum depth of the tree: *max_depth = 7*.
  - Minimum number of samples required to be at a leaf node: *min_samples_leaf = 5*.
  - Set *max_leaf_nodes >= 100* to stop splitting when the total number of leaves equal or greater than 100.

After training the classifier, the built tree has a size of 183 and a total of 129 leaf nodes. The model's accuracy is 57%. The classification report and confusion matrix are shown in Table IV and V, respectively.

TABLE IV. CLASSIFICATION REPORT

	Precision	***Recall***	***F-Measure***
Unprofitable	0.42	0.02	0.04
Low	0.60	0.93	0.73
Medium	0.57	0.13	0.21
High	0.47	0.60	0.53
Weighted Average	0.52	0.42	0.38

TABLE V. CONFUSION MATRIX

	High	***Unprofitable***	***Medium***	***Low***
High	31,243	1,590	1,293	17,762
Unprofitable	15,912	1,716	1,273	55,546
Medium	11,353	361	9,516	54,744
Low	7,987	414	4,756	188,147

According to Table IV and V, class labeled Low has highest true positive, precision, recall, and F-measure while class labeled Unprofitable has lowest true positive, precision, recall and F-measure.

VIII. WEKA'S RESULT

The pre-processed data has been loaded into a data mining tool, Weka, to train a Decision Tree model #2 using the followings:

- Tree algorithm: REPTree
- Percentage split = 70%
- *max_depth = 7*
- *min_num = 5*

The result tree built using Weka has a size of 5,900 which is much larger than the tree's size of model #1. Table VI and Figure XIII show the classification report and confusion matrix output from Weka, respectively.

TABLE VI. CLASSIFICATION REPORT (WEKA)

	Precision	***Recall***	***F-Measure***
Unprofitable	0.381	0.190	0.254
Low	0.761	0.808	0.784
Medium	0.455	0.489	0.472
High	0.461	0.635	0.534
Weighted Average	0.595	0.611	0.595

FIGURE XIII. CONFUSION MATRIX (WEKA)

```
    a      b      c      d   <-- classified as
14243  29553  13748  17262 |      a = Unprofitable
 6428 162535  23298   8810 |      b = Low
 6369  20087  37085  12337 |      c = Medium
10315   1342   7296  32904 |      d = High
```

Model #2 has 61% accuracy which is 0.04% higher than that of model #1. Additionally, model #2 has higher recall for classes labeled Unprofitable, Medium, High and lower recall for class labeled Low. The weighted average for precision, recall, and F-measure are also higher for model #2 compared to that of model #1. Both models have highest precision, recall, and F-measure for class labeled Low, and lowest precision, recall and F-measure for class labeled Unprofitable.

IX. Conclusion

A Decision Tree algorithm has been used to train a classification model to identify profitable loans based on LendingClub Loan Data obtained from Kaggle. The result of the developed model (model #1) is then compared with that of a built model (model #2) using a data mining tool, Weka. Models #1 and #2 do not share the same result because each model has been trained on different train and test datasets. Furthermore, model #1 has a stopping condition when the total number of leaf nodes exceed 100 whereas model #2 could have an infinite number of leaf nodes. For model #1, post-pruning work is needed to decrease the number of leaf node to meet the maximum number of leaf nodes requirement (*max_leaf_node*). Additional parameters like *min_samples_split* (the minimum number of samples required to split an internal node), and *max_features* (the number of features to consider when looking for the best split) could be added to model #1. Furthermore, continuous variables could be used as the features instead of categorical variables, and two-ways split could be used instead of multi-ways split. Further data pre-processing is also needed to handle class imbalance issue because class labeled Low has almost twice of examples compared to that of other class labels.

References

[1] Lee, Jei Young (2020). Prediction of Default Risk in Peer-to-Peer Lending Using Structured and Unstructured Data. *Journal of Financial Counseling and Planning*, 31(1), 115-129. http://dx.doi.org/10.1891/JFCP-18-00073

[2] Li, Peiqian & Han, Gao (2020, March 20). *LendingClub Loan Default and Profitability Prediction*. Retrieved from http://cs229.stanford.edu/proj2018/poster/69.pdf

[3] Tan, Pang-Ning & Steinbach, Michael & Kumar, Vipin (2014). *Introduction to Data Mining* (1st ed.). London, UK: Pearson Education Limited.