DAO2702 Final Project Report

Session: A05

Team Name: Papa Python

David You Chen Ming (A0183268W)

Davin William (A0189019W)

Lee Lin Li Jasmine (A0191107M)

Stanley William Pudjihartono (A0188736L)

Steven Sebastian (A0187460Y)

Zheng Shu Min (A0172595U)

Contents Page

1. Introduction

In the United States (U.S.), peer-to-peer (P2P) lending platforms have grown consumer loans by more than USD$48 billion from 2006 to 2018 (Balyuk, 2019) as consumers increasingly turn to P2P lending as an alternative to traditional bank loans due to its higher accessibility. LendingClub is one of the leading P2P lending companies in the United States. With a rise in credit demand, there have been more instances of default on loans which will potentially squeeze LendingClub's profit margin. In anticipation, LendingClub would like to incorporate a robust model to better filter out borrowers who are likely to default on their loans.

To achieve that, we aim to improve LendingClub's recall rate which is the probability of predicting default given the borrowers will actually default. We thus recommend using machine learning to classify whether a potential borrower will default on his loan or not. In this report, we will analyse data on loans provided by LendingClub, and examine the relationships between loan status ('Fully Paid' or 'Charged Off') and variables such as loan amount and annual income to formulate a classification model to determine if a borrower will default on his or her loan based on a fixed set of predictors. Additionally, we will also provide recommendations based on our resultant model for LendingClub to complement the model to further minimise the risk of borrower defaults and consequently maximise its profits.

2. Methodology

2.1 Data Cleaning

The original dataset contained 42,537 rows of borrower/loan information and 52 columns of variables. Data cleaning was done to obtain a dataset containing only information that would be relevant and useful for our analysis.

Our target variable is 'loan_status'; specifically, we are interested in whether a loan is 'Fully Paid' or 'Charged Off'. Therefore, we first narrowed down the number of rows to include only those which were either 'Fully Paid' or 'Charged Off' under 'loan_status'. Additionally, unnecessary data such as duplicate rows or columns with non-applicable information were removed. We also reformatted columns with numeric variables to ensure the values are in numeric type.

2.2 Data Visualisation

We performed univariate data visualisation to analyse the distribution of each individual numerical variable as well as the distribution of our target variable (loan status), thereby providing us with a better understanding of the data. The main visualisation methods we have applied are Histogram plots and bar charts (Appendix B).

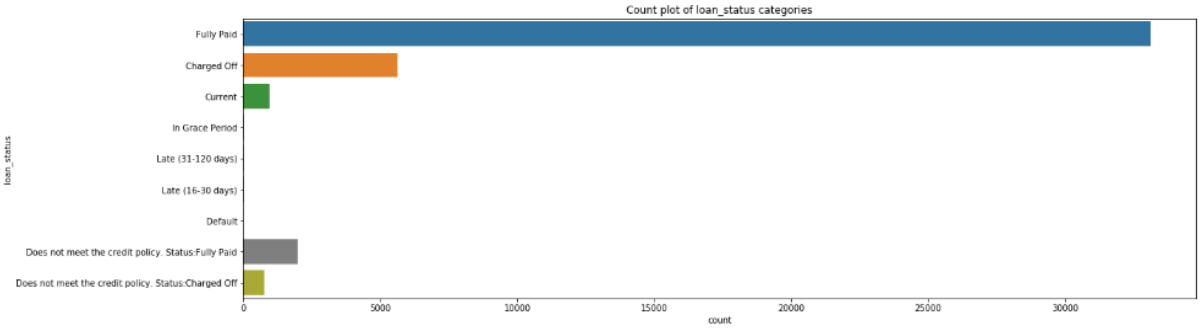


Fig 1. Count of loan_status across loan_status categories

By plotting the status of loans in a bar chart, we realised that the data was extremely skewed towards non-default. Therefore, we took necessary precautions in our data analysis to avoid undesirable results, by prioritizing recall and precision over accuracy.

2.3 K-Fold Cross Validation

Given our limited data sample, we used K-Fold Cross Validation (instead of single train-test split) to better estimate our model performance. In our approach, the training set was split into K equal non-intersecting, complementary subsets. In each iteration of training, the model could be trained on K-1 of the subsets and validated on the remaining one that was not used in training. The validation results from these K iterations were averaged out to give a more robust, less biased estimate of model performance than in a single train-test split. We selected K = 5 for cross-validation, and each Cross Validation fold was generated using a stratified split similar to the train-test split earlier. This was to ensure that the ratio of defaults to non-defaults remained the same in our training and validation sets.

2.4 Modelling Methods

2.4.1 Random Forest Classifier

Random Forest Classifier is an ensemble learning method which consists of many smaller individual decision trees that operate together to provide us with a better predictive performance. This model is applicable to our project as we ultimately want to predict which

group ('Fully Charged' or 'Charged off') the observation will belong to. It also helps us to predict whether a given loan will default or not by producing either a Positive or Negative result. Furthermore, Random Forest is also not sensitive to the scale of predictors, thus we did not have to standardize the predictors first. The fundamental idea of Random Forest Classifier lies in the random selection of data points and predictors in each iteration of constructing the individual decision trees. This avoids the problem of group bias of the decision trees and thus the end results of the prediction will be protected from the individual shortcomings and errors of individual trees.

2.4.2 Support Vector Machine (SVM) Modelling

SVM Modelling was used to represent our observations as points in a multi-dimensional space. SVM will draw linear boundaries in the multi-dimensional space which divides the space into sub-spaces that will define the classification for each observation point. The value of SVM comes from the fact that it finds the most optimum boundaries (furthest away from each observation point) to divide the space. This is done to minimize false observation (e.g. False Positive, False Negative). Once the boundary is set by SVM, the test observation points are simply mapped to the space and classified on whether they land on one or the other side of the boundary.

2.4.3 Logistic Regression Modelling

Logistic regression modelling uses linear regression to perform classification. Linear regression, by default, produces unbounded predicted y. Logistic regression modelling takes the predicted y produced by linear regression and uses it as an input for sigmoid function:

$$sig(t) = \frac{1}{1 + e^{-t}}$$

The sigmoid function makes the final outcome bounded between 0 to 1, making it suitable for classification.

3. Analysis of Results

3.1 Model Selection

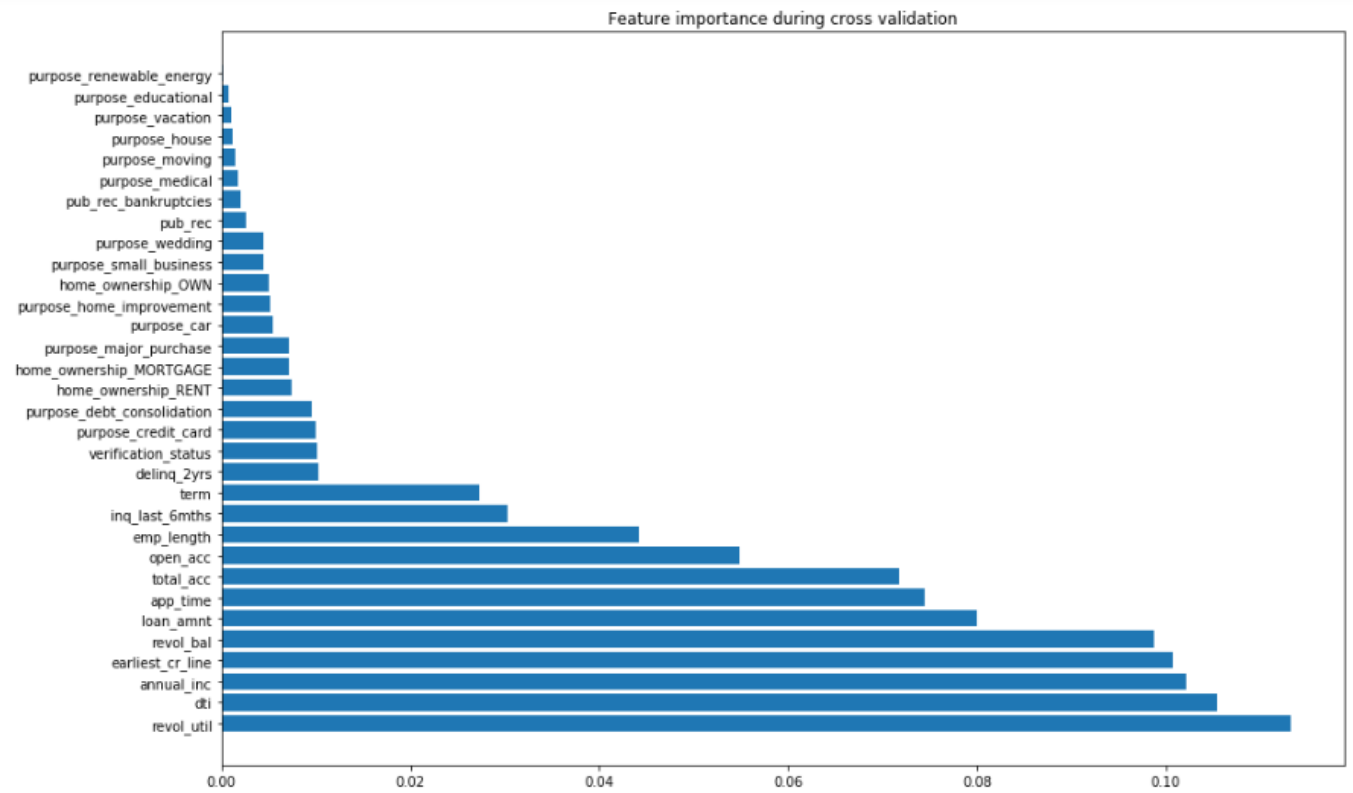
To evaluate our models, we used 3 different metrics: (1) Precision, (2) Recall, and (3) F5 score. Precision refers to the proportion of predicted defaults that turn out to be actual default. Recall refers to the proportion of actual defaults correctly predicted. F5 score is a composite metric

that takes into account both precision and recall, but weighs recall 5 times as much as precision. We gave more weight to recall as we recognized that in the context of default prediction, a false negative prediction (i.e wrongly predicting a default as non-default) is likely to be significantly more costly than a false positive one (which may only result in a premature sell-off and lost potential interests). On the other hand, we did not consider accuracy as from our data visualization we found that the data is already skewed towards non-default (86% vs 13%).

Comparing the candidate models' validation performance, we identified Random Forest Classifier as the best performing model, with slightly lower precision but much higher recall and F5 score than SVM and L1 Logistic Regression (Appendix B). As mentioned, we placed higher importance on recall over precision as false negative predictions are likely much more costly than false positive ones.

3.2 Variable Selection

Before training our selected model, Balanced Random Forest on the full training dataset, we performed variable selection to obtain a more compact model and reduce chances of overfitting.



We computed each variable's importance in each cross-validation fold and ranked the variables by their average feature importance. Least important features were removed.

Based on the chart, there appeared to be a significant drop off between term and delinq_2yrs. We thus selected the cut-off at 0.02 and removed all variables with mean importance less than the cut-off. As such, we identified 12 variables as useful in predicting whether a borrower is likely to default on a loan or not. They are namely: the credit utilization and credit balance, debt-to-income (DTI) ratio, annual income, earliest credit line, loan amount, application time, value of total account, open account, employment length, no. of inquiries in the last 6 months, and term.

3.3 Final Results

We obtained an excellent recall rate of about 70% based on our final model (Appendix B). However, we feel that the model may be too strict as it has classified several non-defaults as defaults, which may thus cause limitations as the precision is not very high at 19.90%. As mentioned, we place higher importance on recall over precision as false negative predictions are likely much more costly than false positive ones.

4. Discussion & Recommendations

Based on our modelling and analysis of the dataset, we have come up with several recommendations for LendingClub to help it minimise the risk of borrower default and maximise profits.

4.1 Usability of Model

As we have seen from our model, application of machine learning for data analysis in business setting has deep practical usability. Our model is able to classify data with recall percentage of 70%. This goes to show that the application of such technology can improve the profitability of business especially for business like our case, credit-lending, where an undetected default (False Negative) can be very costly for the profitability of business.

Furthermore, it is justified that the usage of such machine-learning technology to make credit-lending decisions would be more superior compared to using human discretion when classifying potential customers. It is less time-consuming, and less costly in the long run.

Utilizing the model would mean that LendingClub can hire less credit analysts to review loan applications, thus improving profit margin. Additionally, human beings will have a natural tendency to be subconsciously biased, and this may lead to less optimal decisions. With this knowledge, we would thus recommend that businesses consider using such machine-learning technology in their day-to-day decision making to obtain a superior and more consistent result.

4.2 Appeal Round for Rejected Loan Applicants

Another recommendation is for LendingClub to introduce an appeal round for borrowers who are initially classified by the model as likely to default. This is because as mentioned earlier, our model tends to be stricter than ideal, with precision rate of only 19.9% which means that more than 80% of predicted default do not translate to actual default. The false positive can be interpreted as a loss of potential revenue for LendingClub. Therefore, to mitigate the negative effect, LendingClub can introduce an appeal round to review the predicted default applications on a case-by-case basis. In this appeal round, LendingClub's staff should reassess the variables that have been determined to be significant in predicting if a borrower would default (i.e. revolving utilization, credit-line, DTI etc.) to make a more informed decision on whether to provide the loan. For example, a low revolving utilisation rate corresponds to a better credit score and is more likely to qualify a borrower for a loan (O'Brien, 2019). If a borrower has a low revolving utilisation rate, his loan application may therefore be reconsidered in the appeal round despite having been rejected by the predictive model. This would reduce machine error as our model may reject loan takers in the first round even though they qualify for loan taking, therefore ensuring that qualified loan takers are not being left out by the model and LendingClub is not losing out on potential revenue.

4.3 Ethical Considerations

However, we also recommend that LendingClub should exercise caution when utilising our predictive model. This is because the model is purely driven for the purposes of minimising risk of loan default. Consequently, certain variables such as the purpose of loan were deemed as unimportant and removed from the model (refer to section 3.2). However, this does not mean that LendingClub should stop collecting such information because it is not significant in determining if a borrower will default on the loan. This is because of ethical and liability concerns. For example, if a borrower did not have to declare the purpose of taking out a loan, he could use it for illegal purposes which could get LendingClub into legal trouble. Therefore,

the company should take care to continue to collect this sort of information and not simply rely on the model to streamline its lending processes.

5. Conclusion

Through our analysis of the different variables' relationship with loan status, we are able to produce a predictive model that identifies the most critical variables which LendingClub should consider when deciding on whether the loan takers will be less likely to default their loan. However, there are several limitations to our model which we have provided recommendations to further increase the model's precision. With the help of our model, LendingClub can better filter out the potential loan takers who may default their loan, thereby lowering their overall loan default rates.

6. References

O'Brien, S. (2019, August 20). *Why that 30% rule of thumb about credit card use could be costing you.* Retrieved from https://www.cnbc.com/2019/08/20/why-that-30percent-rule-of-thumb-about-credit-card-use-could-be-costing-you.html

Tetyana Balyuk. (2019, May 6). *Financial Innovation and Borrowers: Evidence from Peer-to-Peer Lending.* Retrieved from https://www.fdic.gov/bank/analytical/fintech/papers/balyuk-paper.pdf

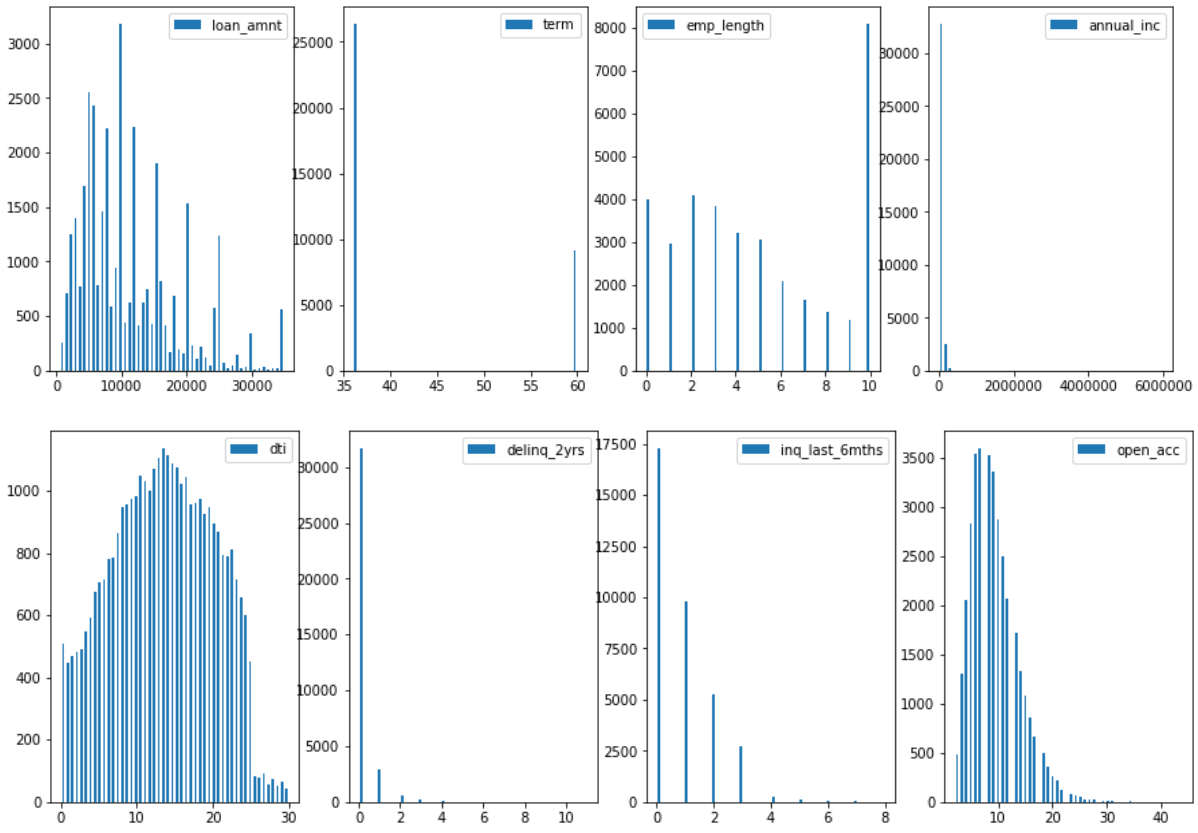
7. Appendix

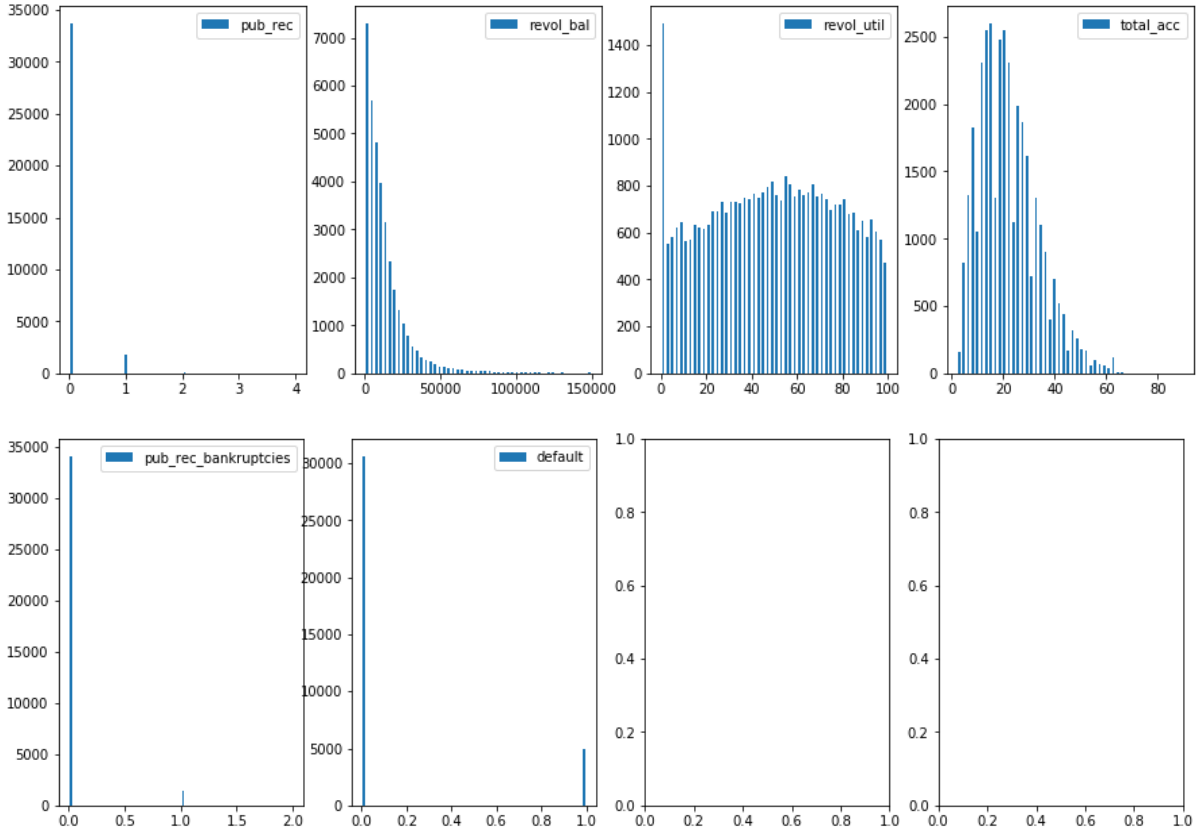
Appendix A: CSV File

Data set is retrieved from https://www.kaggle.com/samaxtech/lending-club-20072011-data.

Appendix B: Tables and Graphs

Section 2.1 - Histograms of Variable Distribution





Section 3.1 - Model Selection Retrieved from Jupyter notebook file.

	Fold	Precision	Recall	F5
0	1	0.199853	0.688131	0.629023
1	2	0.193136	0.667929	0.610231
2	3	0.202619	0.703283	0.642246
3	4	0.192322	0.664141	0.606878
4	5	0.198932	0.706700	0.643524
5	Average	0.197373	0.686037	0.626381

Section 3.3 - Final Results Retrieved from Jupyter notebook file.

	BRF	SVM	L1 Logit
Precision	0.197373	0.233319	0.229979
Recall	0.686037	0.593589	0.610261
F5	0.626381	0.560295	0.573755