Loan Default Prediction and Applicants Evaluation- Evidence in Lending Club

Section A, Team 13

Gamini Singh, Helena Gao, Max Liang, Cindy Yu, Maxwell Zhang

TABLE OF CONTENTS:-

BUSINESS UNDERSTANDING:

The Lending Club is a US peer-to-peer lending club (P2P) headquartered in San Francisco, it enables borrowers to obtain a loan and investors to purchase notes backed by payments made on loans, also makes traditional direct to consumer loans. The Lending Club needs to design and improve the algorithm for investors to find Notes they would like to purchase, using borrower and loan attributes such as the length of a loan term, interest rate, borrower credit score, home ownership status, and others. [1] To reduce default risk, Lending Club focuses on high-credit-worthy borrowers, declining approximately 90% of the loan applications it received as of 2012 and assigning higher interest rates to riskier borrowers within its credit criteria. [2]

During this process, investors face several inherent risks. For example, significant part of risks lies in the fact that some borrowers may not pay back what is owed. In such case, every dollar unable to collect become a dollar worth of potential loss. Therefore, it is critical for the Lending Club to recognise and efficiently manage the potential reputational risk, operational risk, and in particular credit risk associated with poor assessment of credit quality of its customers. On the other hand, Lending Club also need to enhance risk management by selecting high-credit borrowers based on their applicants analysis.

Thus, we aim to construct models with historical loan data that can be used to advise Lending Club of the default risk associated with its future borrowers and convey these information to investors to facilitate information transparency and mitigate their risks.

DATA UNDERSTANDING:

The Lending Club provides the dataset that our team uses to analyze the default behavior of loan applicants. In total, the dataset has 887,379 observations with 74 variables, which is rich in depth and breadth for a comprehensive data analysis. Each record comprises of information on the loan and its applicant such as the state, annual income, the purpose for the loan, home ownership, etc.. There are three data types - characters, numeric, and time. The target variable that we consider in the analysis is *loan_status*, which

is categorical consisting of *Fully Paid, Charged Off, Issued, Late*, etc.. In order for us to conduct analysis on the variable, we need to convert it into binary variable during the process of data preparation.

One potential bias is that the dataset only includes loans that the Lending Club approved but not those that were declined, so the data we have in hand are already those which the Lending Club consider has a low probability of default. So the sample size of default in the dataset is not large enough.

Other potential bias is that we cleaned the data for better analysis and future prediction by deleting the variables with large number of missing data and text data, but it is possible that these variables might have effects on predicting default as well.

DATA PREPARATION:

We divided data preparation into two stages. The first stage is data cleaning, which includes tasks such as removing redundant variables, dealing with null values，converting target variable to binary and constructing two datasets for analysis of different purposes. The second stage is data visualization, which provides valuable insights into the data and lay the foundation for further studies.

Stage 1: Data Cleaning:

First, we removed identifiers, columns of which large proportion of data is missing, variables that are unhelpful for our purpose such as *url*, and those that are difficult to deploy or interpret such as *emp_title*. Then, we substituted null values with the average value or zero. Also, we renamed the target variable as default and changed it from categorical to binary variable to facilitate analysis. Specifically, "*Default*" and "*Charged Off*" is grouped to be 1s and other categories become 0s. After that, we come up with the idea that the dataset can be used either to predict if the Lending Club should approve the loan to potential borrowers or to predict the likelihood of default of existing borrowers. So we ended up with two datasets - one contains only information about borrowers such as annual income while the other has information related to both borrowers and their current loans. As a result, we

can run models with two respective datasets and develop insights into the Lending Club's business from two perspectives.

Stage 2: Data Visualization

We used Tableau to create several diagrams to help us understand the dataset. In *Exhibit 2*, the time series plot shows that for loans issued right after financial crisis, the default rate surged possibly because financial status of many borrows was deeply affected. Then, it gradually comes down in the following years. In *Exhibit 3*, darker the color of a state, the higher the default rate is for borrowers from that state. It is obvious that Idaho, Iowa and Nevada are the three states with highest default rate, while North Dakota, Nebraska and Maine have lowest default rates. *Exhibit 4* describes the relationship between interest rate and default rate in different levels of income. It is obvious that default rate is positively correlated with interest rate and negatively correlated with income.

MODELING:

We build a framework to guide our workflow as the graph shown below. Generally, we used unsupervised models first for data exploration and then use supervised models for prediction. For better deployment and business application, two sets of models are constructed for its respective purposes.

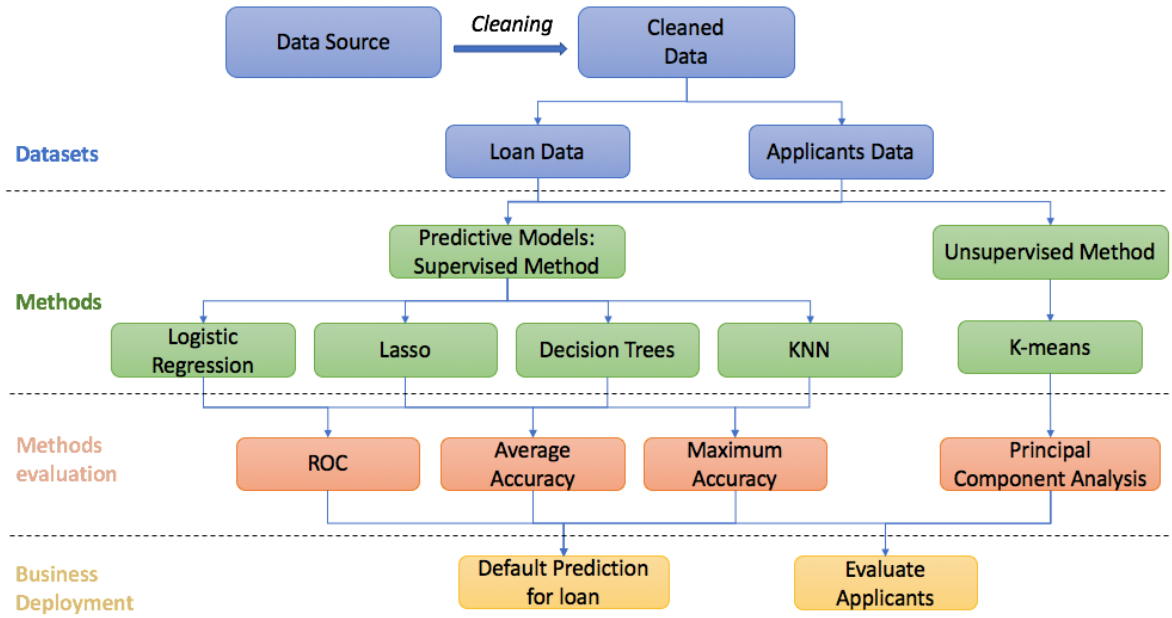


Figure 1. Framework

We built 4 supervised models with training data, made out of sample (OOS) predictions, and then evaluated performance using three metrics: areas under the ROC curve (AUC), average accuracy, and maximum accuracy before deciding the best model to use.

UNSUPERVISED MODEL:

1. *K-Means & Principle Component Analysis (PCA)*

For data exploration, we conducted k-means clustering method at the very beginning of our modeling process(See Exhibit 5). After getting the 5 clusters from running k-means method, we used PCA to find out and analyze the features that dominate each cluster. Out of the 5 clusters, only the second cluster consists of a large portion of default loans with 43668 defaults and 4 non-defaults as is shown in Exhibit 5. The Exhibit 6 illustrates the 6 attributes that dominate the second cluster such as total amount committed, implies that these 6 attributes are related to the default of loans and the latent feature under these factors can be summarized as payment schedule and payback ability. So, the Lending Club should be aware of how payment schedule is set in the loan agreement, which can affect borrowers' default likelihood.

SUPERVISED MODELS:

1. *Logistic Regression*

As default is binary, we used logistic regression. So, we ran the logistic regression with all variables from cleaned data. But only a few coefficients, such as loan amount, are statistically significant. In addition, with so many variables it is likely that the model overfits. Finally, it may be practically difficult to deploy given the amount of information needed.

2. *Lasso*

We used theory lambda for running the Lasso algorithm (Exhibit 7). The variables chosen by Lasso resemble what we see from the data visualization graphs and k-means method. Although using min lambda Lasso selects around 57 variables and has the lowest deviance, and using 1se lambda is only slightly worse than that, the theory lambda chooses an appropriate amount of variables which are more interpretable (Exhibit 8). The lasso is very convenient to use because the algorithms itself ignores unimportant variables and mitigates the problem of overfitting especially when combined with our manual variable selection process. The Lending Club only needs to input the 8 variables to predict probability of default for loans, which is sound and accurate.

3. *Classification Tree*

We used classification tree for prediction. Separating the data into comprehensive which contains all variables and applicant-specific, we built the tree using training data and predicted using testing data. The result can be seen in Exhibit 9. Classification tree is advantageous compared to other models in its interpretability as it shows clearly the decision making process at each node.

4. *K-Nearest Neighborhood (KNN)*

Using separate dataset for loan and applicants, we built K-NN model using training data and made prediction using testing data. Then, we compared the prediction with the actual data and assessed the prediction result (See Exhibit 10). The AUC for this method is rather unimpressive, because the number of non-default loans and applicants are much

higher than default. Thus, simply taking average of these neighborhoods would have no much higher prediction correctness than random guess. Since the Lending Club can use the data for similar loans and borrowers to evaluate default risk, it may want to group similar loans and borrowers into groups for more efficient comparison.

EVALUATION:

	Loan Data			Application Data		
	AUC	**AVG_ACC**	**MAX_ACC**	**AUC**	**AVG_ACC**	**MAX_ACC**
logistic	**0.525**	**0.975**	**0.995**	**0.553**	**0.889**	**0.948**
decision tree	**0.979**	**0.970**	**0.993**	**0.500**	**0.884**	**0.948**
knn	**0.513**	**0.903**	**0.949**	**0.500**	**0.929**	**0.948**
lasso	**0.984**	**0.951**	**0.995**	**0.650**	**0.888**	**0.948**

In total, we constructed 4 prediction models for assessing default risk. For each models, we first generated a ROC curve, which contains series of true positive rate and false positive rate under different thresholds (Exhibit 11) . Area under the ROC curve (AUC), our first performance metric shows lasso is the best model with AUC of 0.94. Classification tree is slightly worse in performance, while logistic and knn models have AUC of around 0.5. Moreover, we also considered the accuracy (ACC). We took into account the accuracy of models with all thresholds and generated the average accuracy of models and maximum accuracy of models accordingly. Although other three models have lower performance, both logistic regression and lasso achieve greater accuracy in both average and maximum accuracy. Consequently, lasso appears to be the best model in predicting default. Therefore, we decide to go ahead with it.

In practice, lenders want straightforward and accurate prediction of default based on loan and applicant information. One challenge commonly encountered is fast-growing and messy datasets. One way to improve is to perfect existing dataset while put consistent effort

into cleaning incoming data so that prediction performance can be accurately measured and compared. To evaluate the improvement after integrating our prediction models, the Lending Club can compare the percentage of default loans before and after.

BUSINESS DEPLOYMENT:

Credit risk exposure of a P2P platform varies with respect to its loan exposure, borrower characteristics and third-party collaborations. Before executing the proposed business deployment plan, it is critical to understand factors influencing credit risk exposure. These factors can stem from lenient credit standards, lax debt collection, poor portfolio risk management or insensitivity towards changing economic conditions. Borrower default serves as the common link between all these factors, either as the cause or the consequence. Therefore, it is imperative that P2P platforms develop a comprehensive understanding to proactively measure, monitor and predict borrower default.

In current project, we adopted similar strategy by deploying models for analysing the extensive loan data of Lending Club. We separated analysis into two parts. One part deals with only borrower details and other discusses different loan attributes. This method of analysis gives flexibility to break the analysis into two parts and predicts chances of default from business and applicants' points of view. Similar approach can be used for analysis of real time data generated by other P2P platforms. However, credit risk management strategies can change according to the sophistication of business activities investors are engaged in. Some attributes can contain missing values or can be irrelevant to the corresponding business setting. Since it is difficult to directly apply our models on data collected by other lending platforms, some customizations should be done. Moreover, the models used in our analysis use static data. This poses a limitation to predictions based on customer data like home ownership, and annual income that are prone to change. This dynamic aspect of data should be considered during deployment. Fundamentally, all model deployments should be targeted towards maximizing the risk adjusted return by keeping borrower default under check and avoiding lending to past defaulters.

While final models deployed will differ on case to case basis, the following issues still need to be addressed to develop a robust lending mechanism:

1. Following outdated credit risk strategies and lack of sound credit portfolio: This can be mitigated through periodic revision of lending policies. P2P platform's risk to reward threshold should be set according to profitability expectations and exposure limits on loans sanctioned. Arranging periodic external audits can also help (by checking fair value of outstanding debts or uncollectible etc.).
2. Poor credit administration: This can be mitigated through comprehensive assessment of borrower's or counterparty's risk profile and repayment capacity. Borrowing pool's credit default history can be used to revise details like maximum line of credit, interest rates and collection schedules.
3. Lack of scrutiny on fraudulent applications: This can be minimized through third party verification, external credit rating, utilizing public records etc. Addition of new borrowers should be done after thorough background checks on their reputation and legal capacity to repay.

Recent events of data breaches in the financial industry (Deloitte and Equifax breaches) have raised an alarm on the collection of personal details (annual income, delinquency etc.) by companies and the ethical issues related to sharing such confidential and sensitive information with other companies. However, collection, management and mining of data is an inevitable part of our analysis to effectively predict loan defaulters. It is critical that P2P companies work towards reducing their vulnerability to data breaches and actual or potential conflicts of interest by executing impenetrable confidentiality arrangements (a.k.a "Chinese walls"). Data collected, handled, stored and mined effectively facilitates seamless execution of prediction models that pave way for a superior customer experience.

APPENDIX:

Exhibit 1: Definitions of all the column names under study

Variable	Explanation
annual_inc	The self-reported annual income provided by the borrower during registration.
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
collection_recovery_fee	post charge off collection fee
collections_12_mths_ex_med	Number of collections in 12 months excluding medical collections
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
funded_amnt	The total amount committed to that loan at that point in time.
funded_amnt_inv	The total amount committed by investors for that loan at that point in time.
grade	LC assigned loan grade
home_ownership	The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.

initial_list_status	The initial listing status of the loan. Possible values are – W, F
inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
installment	The monthly payment owed by the borrower if the loan originates.
int_rate	Interest Rate on the loan
last_pymnt_amnt	Last total payment amount received
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
loan_status	Current status of the loan
open_acc	The number of open credit lines in the borrower's credit file.
out_prncp	Remaining outstanding principal for total amount funded
out_prncp_inv	Remaining outstanding principal for portion of total amount funded by investors
pub_rec	Number of derogatory public records
purpose	A category provided by the borrower for the loan request.
recoveries	post charge off gross recovery
revol_bal	Total credit revolving balance

term	The number of payments on the loan. Values are in months and can be either 36 or 60.
total_acc	The total number of credit lines currently in the borrower's credit file
total_pymnt	Payments received to date for total amount funded
total_pymnt_inv	Payments received to date for portion of total amount funded by investors
total_rec_int	Interest received to date
total_rec_late_fee	Late fees received to date
total_rec_prncp	Principal received to date
acc_now_delinq	The number of accounts on which the borrower is now delinquent.

Exhibit 2: Change in default rate with respect to time

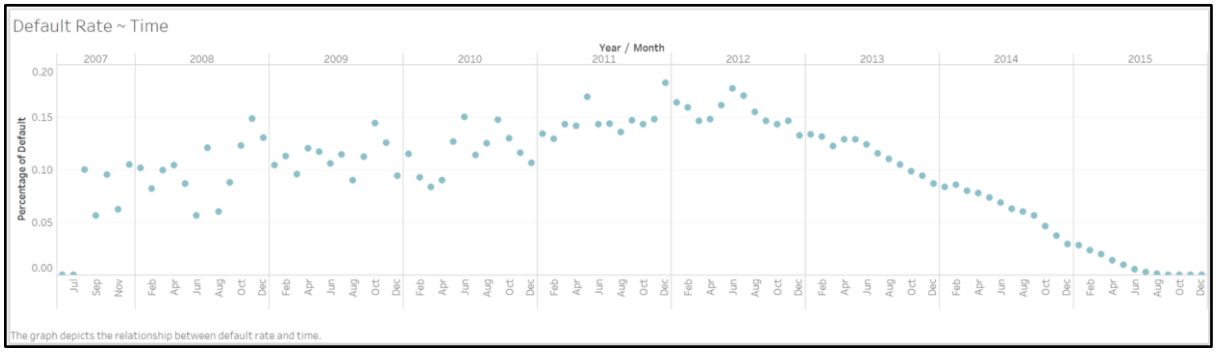


Exhibit 3: Spread of default rates across different U.S. states

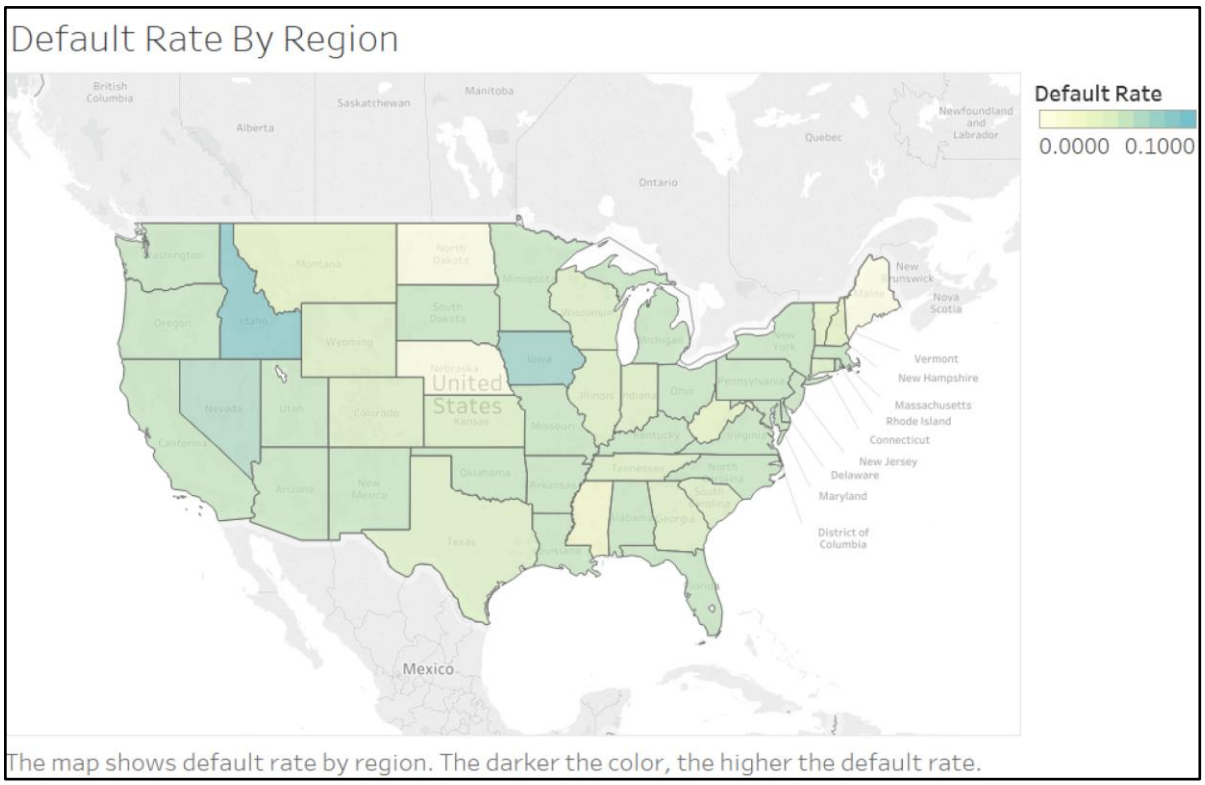


Exhibit 4: Change in default rate with respect to interest rate and borrower's annual income.

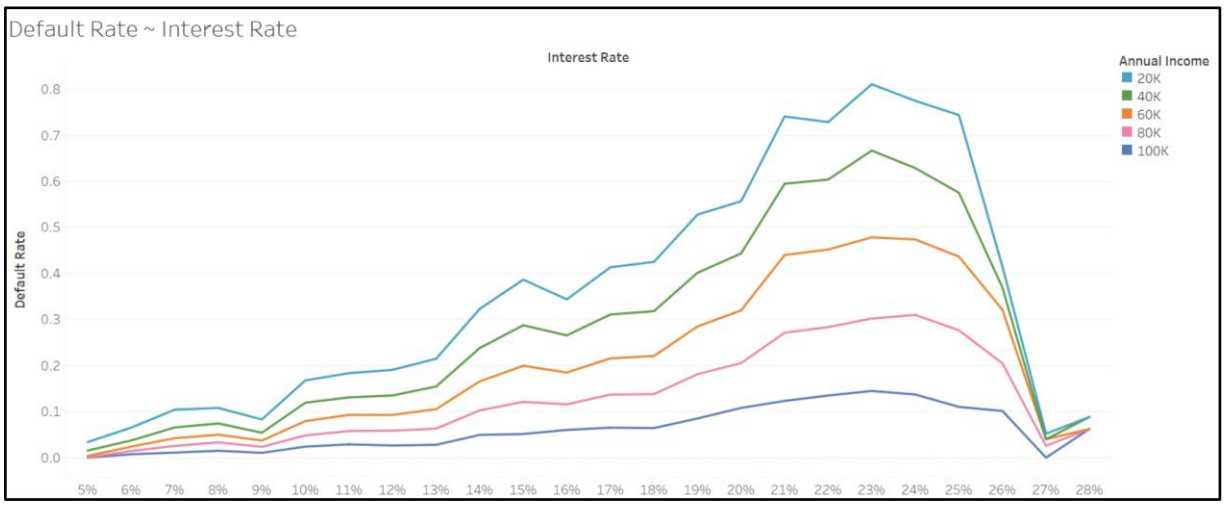


Exhibit 5: K-MEANS (K=5)

$`1`	**$`2`**		$`3`		$`4`	$`5`	
0	**0**	**1**	0	1	0	0	1
210774	**46**	**43668**	119627	184	300275	167932	19

Exhibit 6: PCA graph for Cluster 2

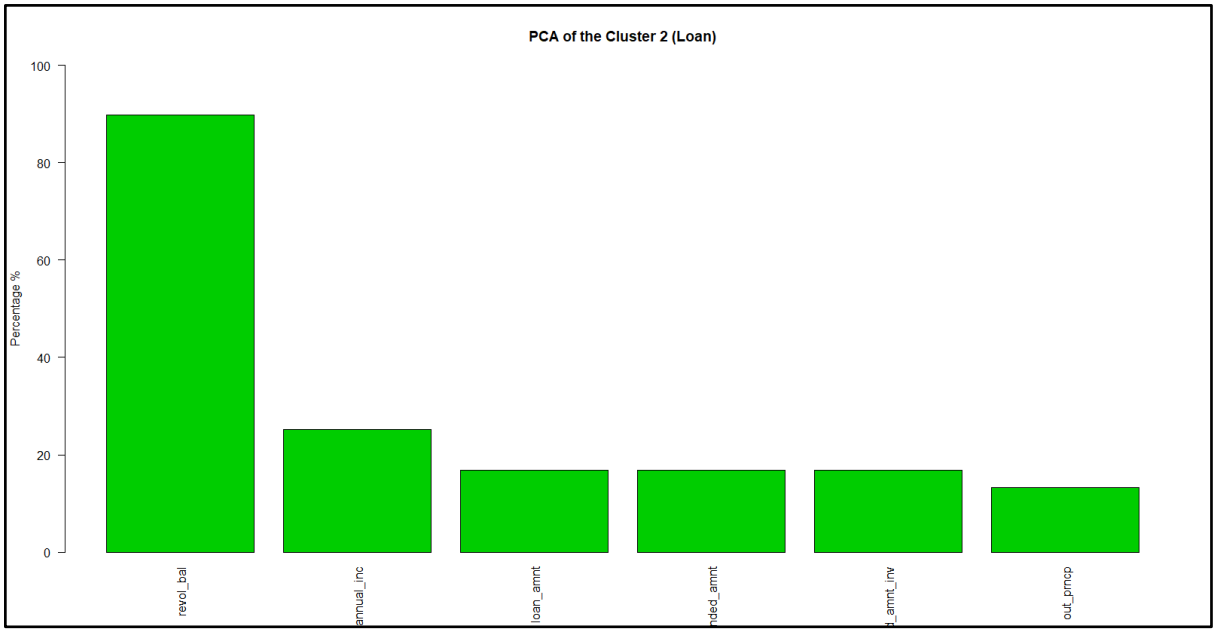


Exhibit 7: Lasso regression result

Regressor	Coefficient
"loan_amnt"	"0.000126400101282419"
"funded_amnt"	"9.17332020932901e-05"
"int_rate"	"0.0562922503691101"
"out_prncp"	"-0.000422966230100857"
total_rec_prncp"	"-0.000272867732524233"
"total_rec_late_fee"	"0.030556585017468"
"recoveries"	"0.00317055423699975"
"last_pymnt_amnt"	"-0.000136217232336222"

Exhibit 8: Fitting Graph for CV Lasso

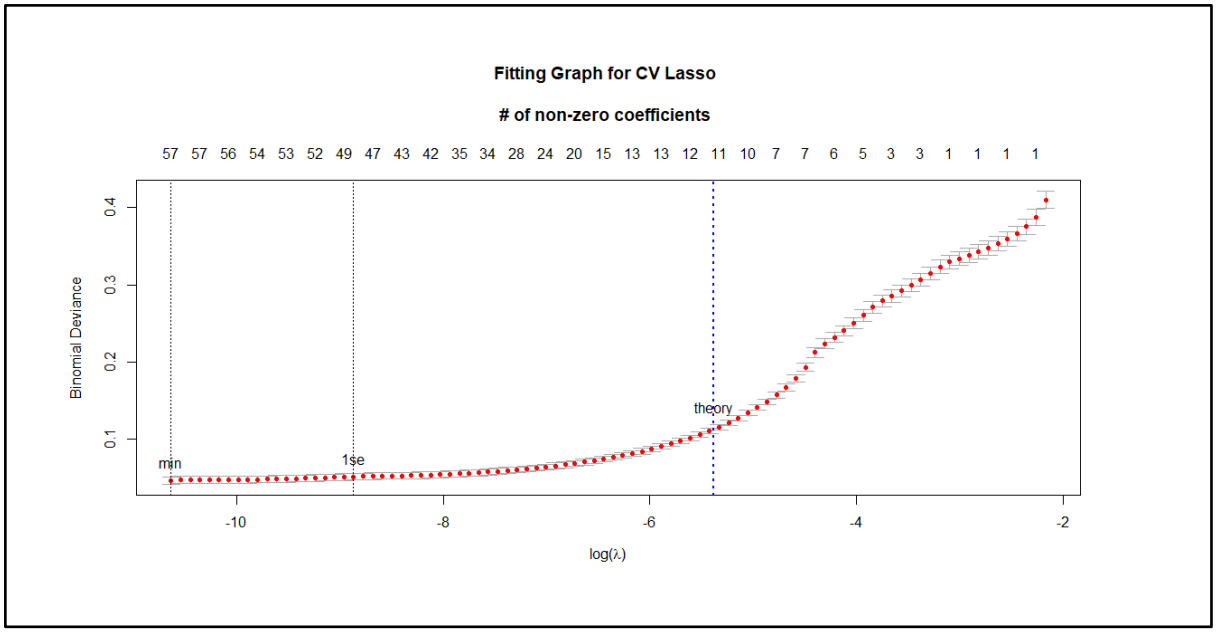


Exhibit 9：Classification Tree

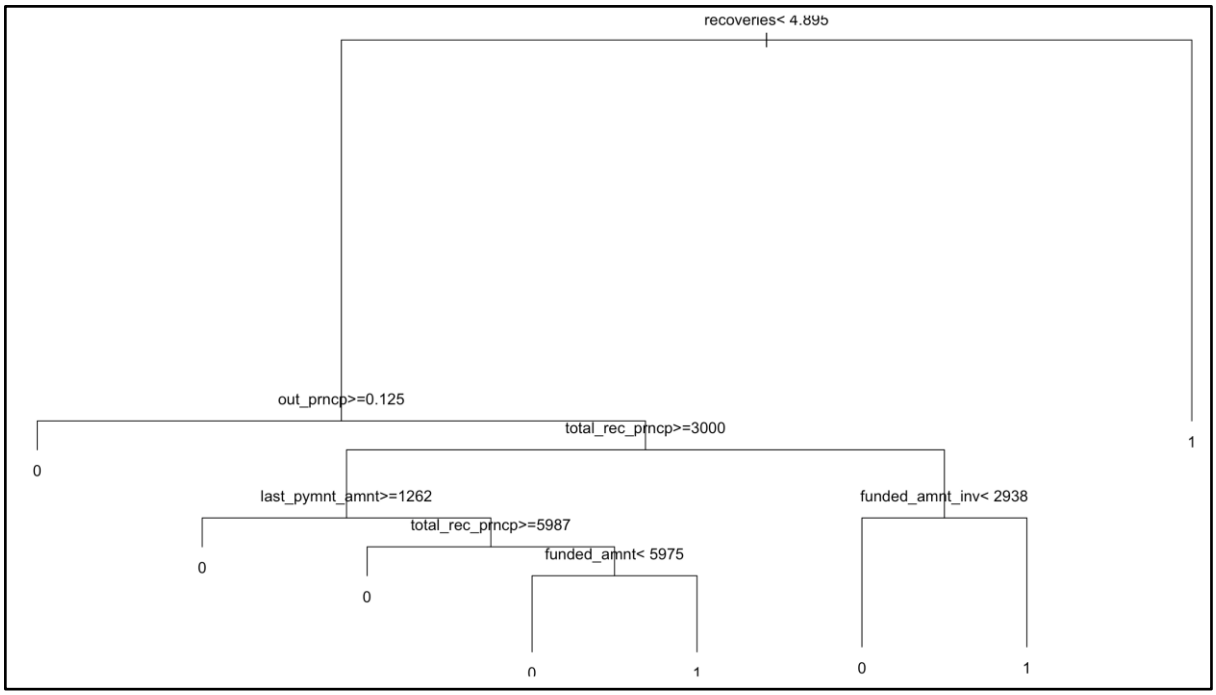
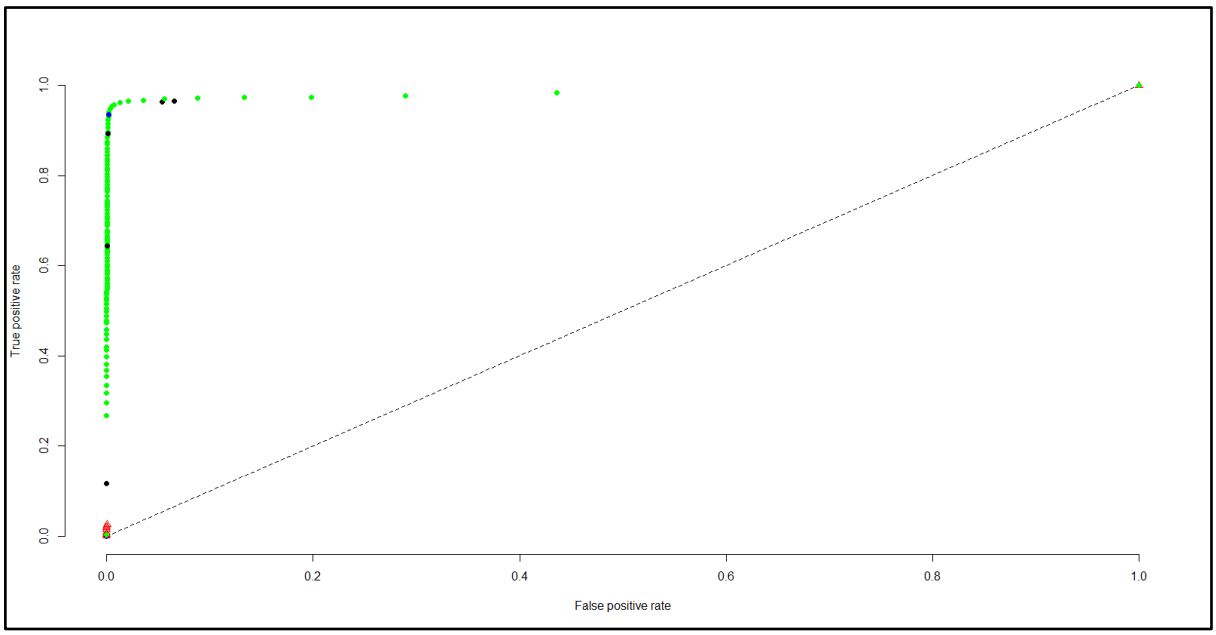


Exhibit 10: KNN Prediction Result

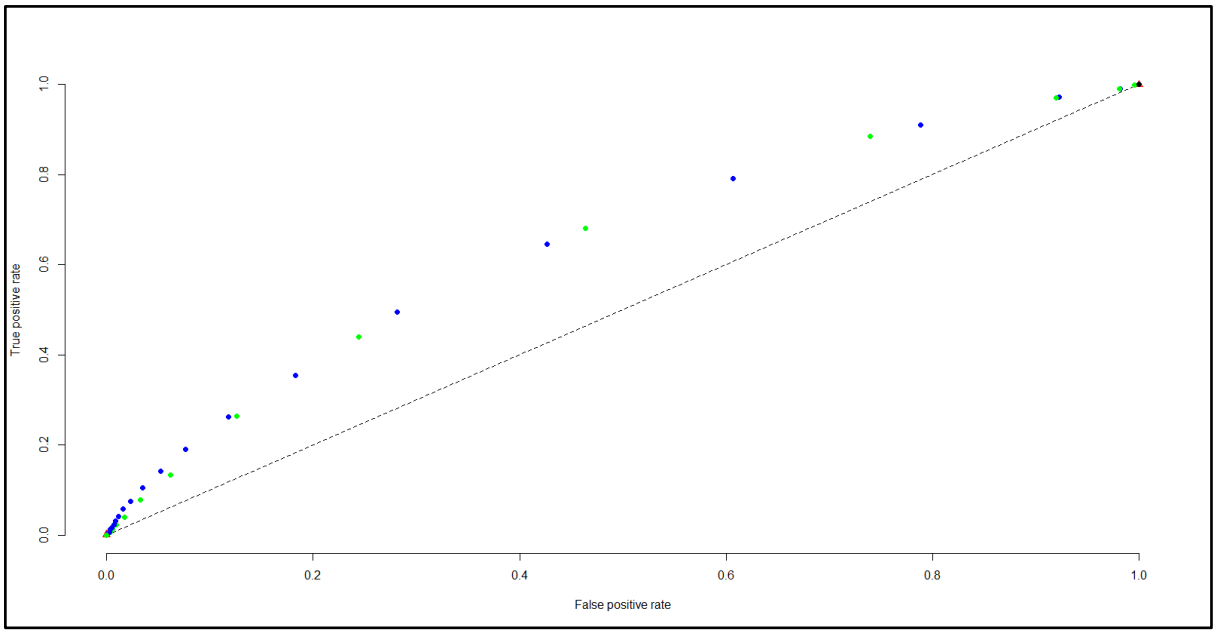
testg	0	1
0	0.90299184	0.04351768
1	0.04805077	0.00543971

Exhibit 11：ROC curve for four models



ROC Curve for Loan Data

(Blue-Logistic Regression, Green-LASSO, Black-Classification Tree, Red- KNN)



ROC Curve for Applicants Data

(Blue-Logistic Regression, Green-LASSO, Black-Classification Tree, Red- KNN)

Exhibit 12：Project Work Load Allotment

	Gamani Singh	Max Liang	Helena Gao	Cindy Yu	Maxwell Zhang
Business Understanding	*	*	*	*	
Data Understanding	*	*	*	*	*
Data Preparation		*	*	*	*
Modeling		*	*	*	*
Evaluation	*		*	*	*
Deployment	*	*			
R Script	*	*	*	*	*

REFERENCE:

1. "Lending Club Prospectus" (2010). Retrieved March 1, 2011.
2. Matthew Zietlin (2012). "Why Is Larry Summers Signing Up With Lending Club?". *The Daily Beast*. Retrieved May 28, 2013.
3. Federal Reserve Bank of Atlanta, Components of a Sound Credit Risk Management Program.
4. Jiménez G., Saurina J., (2002), Loan Characteristics and Credit Risk. Bank of Spain. *Directorate-General of Regulation.*
5. Basel. (1999). Principles for the Management of Credit Risk. *Basel Committee on Banking Supervision, Basel*