**Lending Club Issued Loans Analysis Protocol**

*Author(s): Feldman Nadav   
 Burkis Konstantin*

# Introduction

In this project we analyzed Landing Club[[1]](#footnote-1) Issued Loans data with the aim of estimating the credit risk of each loan, independent of LB's estimation.

Although there are several approaches to estimate credit risk, such as estimating the exposure at default (EAD) and the loss given default (LGD), in this work we studied only the probability of defaulted (PD) method, meaning estimating the probability to default of each given loan, by using classification algorithms to score each loan.

Credit risk problems have great importance for credit providers such as banks or credit cards companies and credit risk rating models are at the heart of their business model, evidence of this can also be seen in the constant preoccupation of the regulator with this subject.

Therefore, there is no wander that credit risk problems have been studied extensively, both in industry and academy and in particular, Lending Club data have been studied by many.

From the algorithmic point of view, there is a lot to studies from (Baesens et al., 2003)[[2]](#footnote-2) who set up a benchmarking state of the art classification algorithms for credit scoring by applied various state of the art classification algorithms on to eight real-life credit scoring data sets. they found that both the LS-SVM and neural network classifiers yield a very good performance, but also simple classifiers such as logistic regression and linear discriminant analysis perform very well for credit scoring.

Another great resource, and much up to date, are (Lessmann & e.g., 2015)[[3]](#footnote-3) whom updated the study of Baesens et al. and compared several novel classification algorithms to the state-of-the-art in credit scoring and found some advanced methods to perform extremely well on our credit scoring data sets, but never observe the most recent classifiers to excel.

From previous researches we can also study about important variables which are known to influence the probability to default, such as (Emekter, Tu, Jirasakuldech & Lu, 2015)[[4]](#footnote-4) , which analyst Lending Club data[[5]](#footnote-5) and found that credit grade (given by lending club), debt-to-income ratio, FICO score and revolving line utilization play an important role in loan defaults and that higher interest rates charged on the high-risk borrowers are not enough to compensate for higher probability of the loan default.

It should be noted that others also found loan purpose to be an important variable.

As mentioned, in this project we aimed to estimate the probability of defaulted independently to Lending Club estimation, hence we decided not to use any variable related to this estimation, including Lending Club credit grade.

In addition, while a proper definition of default is a case which a debtor has passed the payment deadline on a debt they were due to pay, in aim of making a better and more precise credit risk estimation and in order to decrease noises, we defined it as a case which the borrower failed to pay his debt. It should be noted that this definition is also consistent with definitions made in previous studies.

Finally, In order to create the best predicative model for probability of defaulted estimation, as defined above, we have set a number of baselines for comparison.

First, due to the unbalance nature of our problem and data (most of loans are paid), we would like to be more successful than if we had guessed that all the loans would be paid.

Second we would like to be more successful than Lending Club credit score[[6]](#footnote-6), and third, we would like to be more successful than previous works done on the same Lending Club data set.

Although on one hand this subject has been studied extensively in the academia, and on the other hand much work has been done on this issue in less formal frameworks, we hoped that we would be able to create a good model against the baselines that we have set.

This was because we operate according to academic standards, but unlike academic studies, our sole goal was to produce as successful predictive model as possible, based on our data, and we were willing to try an extensive toolkit to achieve this goal.

# Methodology (Project design) – Work on progress

## Data

In our project we are using a dataset which was posted in Kaggle, [Lending Club Loan Data.](https://www.kaggle.com/wendykan/lending-club-loan-data)

Also the data can be downloaded directly from Lending Club Webpage [Lending Club Statistics Page](https://www.lendingclub.com/info/download-data.action), but the data may differ. As stated in Lending Club Webpage the data is not full, full data can be downloaded only after registration to the site therefore our dataset have a lot of missing values and some features are missing at all.

The downloaded dataset from Kaggle comes with a form of one .csv file. The file contain loan data for all loans issued through 2007-2015, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information. Additional features include credit scores, number of finance inquiries, address including zip codes, and state, and collections among others. The file contains 887,382 observations and 75 variables. A data dictionary is provided by Lending Club in a separate file in appendix (lcdatadictionary.xlsx).

As a part of our project we divided the file into 6 tables which describe the loans details and the loaner's details. The final ER Diagram of the database that we created is shown below.

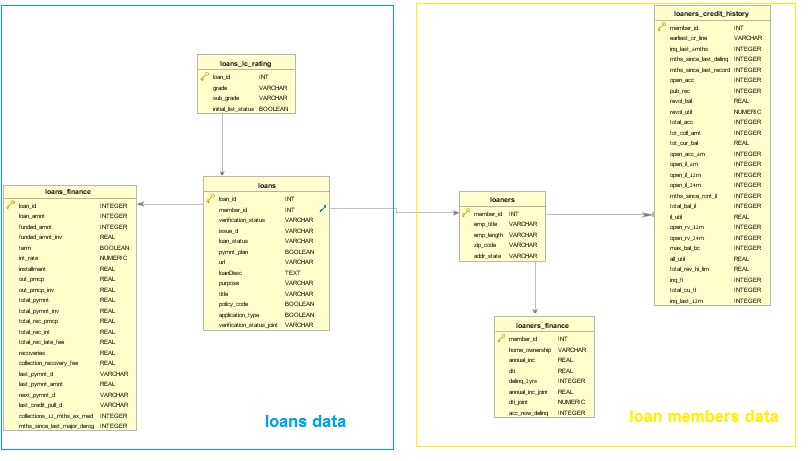


Figure 1 - ER Diagram of database

One of the possible external data that we thought and that may enrich our data considering that we have the state of each loan is the poverty rate of the state in that time.  
We downloaded poverty rates by states in the years 2007 – 2015 of loans from

<https://www.census.gov/cps/data/cpstablecreator.html>.

Our ultimate goal is the prediction of loan defaults. The variable loan status seems to be an indicator of the current state a particular loan is in.

The different loan statuses in our data and their spread shown in Table 1 - Loan statuses statistics and Figure 2 - Loans statistics.

| **loan\_status** | **count** | **rel\_count** |
| --- | --- | --- |
| Charged Off | 45248 | 0.0510198214 |
| Current | 601340 | 0.6780467509 |
| Default | 1219 | 0.0013744953 |
| Does not meet the credit policy. Status:Charged Off | 761 | 0.0008580729 |
| Does not meet the credit policy. Status:Fully Paid | 1988 | 0.0022415887 |
| Fully Paid | 207723 | 0.2342200839 |
| In Grace Period | 6250 | 0.0070472481 |
| Issued | 8396 | 0.0094669913 |
| Late (16-30 days) | 2357 | 0.0026576582 |
| Late (31-120 days) | 11589 | 0.0130672894 |

Table 1 - Loan statuses statistics

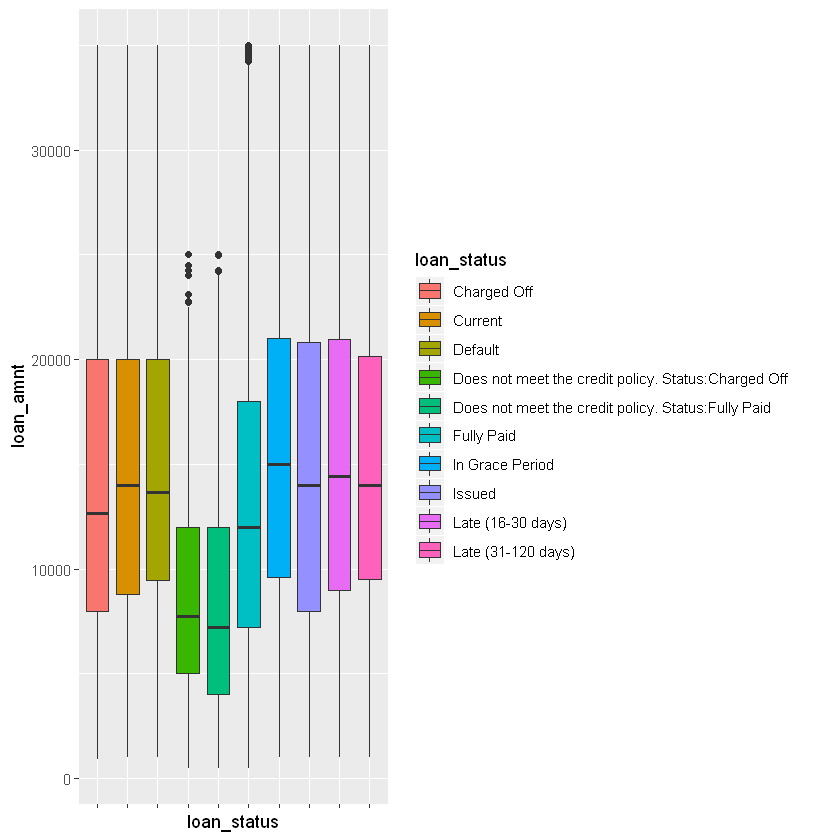


Figure 2 - Loans statistics

It is not immediately obvious what the different values stand for, so we refer to Lending Club’s documentation about “[What do the different Note statuses mean?](https://help.lendingclub.com/hc/en-us/articles/215488038-What-do-the-different-Note-statuses-mean-)”

* **Fully Paid:** Loan has been fully repaid, either at the expiration of the 3- or 5-year year term or as a result of a prepayment.
* **Current:** Loan is up to date on all outstanding payments.
* **Does not meet the credit policy. Status:Fully Paid:** No explanation but see “fully paid”.
* **Issued:** New loan that has passed all Lending Club reviews, received full funding, and has been issued.
* **Charged Off:** Loan for which there is no longer a reasonable expectation of further payments. Generally, Charge Off occurs no later than 30 days after the Default status is reached. Upon Charge Off, the remaining principal balance of the Note is deducted from the account balance. Learn more about the [difference between “default” and “charge off”](https://help.lendingclub.com/hc/en-us/articles/216127747).
* **Does not meet the credit policy. Status:Charged Off:** No explanation but see “Charged Off”
* **Late (31-120 days):** Loan has not been current for 31 to 120 days.
* **In Grace Period:** Loan is past due but within the 15-day grace period.
* **Late (16-30 days):** Loan has not been current for 16 to 30 days.
* **Default:** Loan has not been current for 121 days or more.

Given above information, we will define a default (**Outcome**) as follows:

**Defaulted**loans are in status:

1. Charged Off
2. Does not meet the credit policy. Status:Charged Off

**Fully Paid** loans are in status:

1. Fully Paid
2. Does not meet the credit policy. Status:Fully Paid

For the other possible statuses we don't have a definite outcome of the loan therefore we will exclude them from our dataset.

We have these variables in our dataset which are indications of the outcome - determined by Lending Club after the company evaluations of the credit risk.

* grade
* sub\_grade
* int\_rate

These are confounder variables that may affect the outcome so we drop them.

We also transform the variable ***installment*** as it was calculated with the interest rate that LC is calculated for the loan. New variable ***loan\_installment*** was calculated as the "clean" calculation of.

After definition of our ***outcome*** value and definition of our question we can remove all columns that are not defined at the beginning of the loan evaluation process:

* last\_pymnt\_d
* next\_pymnt\_d
* collection\_recovery\_fee
* last\_pymnt\_amnt
* out\_prncp
* out\_prncp\_inv
* recoveries
* total\_pymnt
* total\_pymnt\_inv
* total\_rec\_int
* total\_rec\_late\_fee
* total\_rec\_prncp

By their definitions below they all relate to later or current stages of the loan therefore are not relevant to the initiate state of the loan.

After that we continue our exploration of the data we checking NA and unique ratios.

TABLE 1 -- ??

####I don’t know if to show all the columns drops by NA or the data retrieval protocol is enough.

Which techniques will be applied to enrich the data? – Feature engineering?

Dealing With Outliers:

Dealing With NA:

We see that the variables ***verification\_status\_joint,annual\_inc\_joint,dti\_joint*** has a **99% NA**

**ratio**. This variables by their LC Definition indicate those are columns which describe a joint application loans.

| **variable** | **p\_zeros** | **p\_na** | **unique** |
| --- | --- | --- | --- |
| verification\_status\_joint | 0.00 | 99.94 | 3 |
| annual\_inc\_joint | 0.00 | 99.94 | 308 |
| dti\_joint | 0.00 | 99.94 | 449 |

Table - High NA Ratio

We will check for the ratio of joint applications in our dataset:

| **application\_type** | **N** | **app\_rat** |
| --- | --- | --- |
| 0 | 886871 | 0.9994241488 |
| 1 | 511 | 0.0005758512 |

Table - Application types ratio

As we can see by Table 2 - High NA Ratio and Table 3 - Application types ratio that exactly 99.4% of our loans are single type and that why these 3 columns have 99.4% NA.

This is as MNAR case and because the number of Joint loans is really small we will drop those observations.

* Add at the end of the protocol (appendix) the Data retrieval protocol

## 

## Models

Here you have to describe how do you plan to develop your models:

* How do you plan to divide your data
  + Training, validation, test - proportions, techniques
* Do you need to balance your data? How?
* Do you need to stratify/subsample your data? How?
* What techniques will you apply to model your outcome?
  + Unsupervised
  + Regression
  + Classification
* Will you use cross-validation and/or bootstrap?
* Which measures you will use to train and evaluate your models? Why?
* Do you plan to use ensembling or will use your best model?

## Deployment of your model

* Who will make the QA of the project?
  + Which units will be assessed
  + Write a QA protocol for each step of the project
* Who is the final user of the predictions?
* How the prediction will be presented to the final user?
* How will the final user be trained to use and interpret the prediction?
* On which platform the predictions will be deployed?
* How frequently the model will be updated?
* What will happen in cases where the model return a null prediction (eg. incomplete data)?
* Which models were used and which were selected for the final prediction.
* Which measurements were used to evaluate the prediction.
* Which results we got from those models.

# Results

Here you will present the main results of all the process. We will describe:

* The final amount of data used (total, train, test, etc)
* The amount of outliers and the way of treating them,
* The amount of missing values and the methods used for imputing them,
* The distribution of the data (timeframes)
* The methods used to transform the data and to generate new features.

# Conclusion

Here you will write about how the project began, which were the most important challenges you had when developing the project, and how did you get the final prediction. You have to discuss also the limitations of the model, when it can be used and when not.

# Appendix

1. Data Dictionary By Lending Club

|  |  |
| --- | --- |
| **LoanStatNew** | **Description** |
| acc\_now\_delinq | The number of accounts on which the borrower is now delinquent. |
| addr\_state | The state provided by the borrower in the loan application |
| all\_util | Balance to credit limit on all trades |
| annual\_inc | The self-reported annual income provided by the borrower during registration. |
| annual\_inc\_joint | The combined self-reported annual income provided by the co-borrowers 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 |
| desc | Loan description provided by the borrower |
| 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. |
| dti\_joint | A ratio calculated using the co-borrowers' total monthly payments on the total debt obligations, excluding mortgages and the requested LC loan, divided by the co-borrowers' combined self-reported monthly income |
| earliest\_cr\_line | The month the borrower's earliest reported credit line was opened |
| 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. |
| emp\_title | The job title supplied by the Borrower when applying for the loan.\* |
| 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. |
| id | A unique LC assigned ID for the loan listing. |
| il\_util | Ratio of total current balance to high credit/credit limit on all install acct |
| initial\_list\_status | The initial listing status of the loan. Possible values are – W, F |
| inq\_fi | Number of personal finance inquiries |
| inq\_last\_12m | Number of credit inquiries in past 12 months |
| 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 |
| is\_inc\_v | Indicates if income was verified by LC, not verified, or if the income source was verified |
| issue\_d | The month which the loan was funded |
| last\_credit\_pull\_d | The most recent month LC pulled credit for this loan |
| last\_pymnt\_amnt | Last total payment amount received |
| last\_pymnt\_d | Last month payment was 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 |
| max\_bal\_bc | Maximum current balance owed on all revolving accounts |
| member\_id | A unique LC assigned Id for the borrower member. |
| mths\_since\_last\_delinq | The number of months since the borrower's last delinquency. |
| mths\_since\_last\_major\_derog | Months since most recent 90-day or worse rating |
| mths\_since\_last\_record | The number of months since the last public record. |
| mths\_since\_rcnt\_il | Months since most recent installment accounts opened |
| next\_pymnt\_d | Next scheduled payment date |
| open\_acc | The number of open credit lines in the borrower's credit file. |
| open\_acc\_6m | Number of open trades in last 6 months |
| open\_il\_12m | Number of installment accounts opened in past 12 months |
| open\_il\_24m | Number of installment accounts opened in past 24 months |
| open\_il\_6m | Number of currently active installment trades |
| open\_rv\_12m | Number of revolving trades opened in past 12 months |
| open\_rv\_24m | Number of revolving trades opened in past 24 months |
| out\_prncp | Remaining outstanding principal for total amount funded |
| out\_prncp\_inv | Remaining outstanding principal for portion of total amount funded by investors |
| policy\_code | publicly available policy\_code=1 new products not publicly available policy\_code=2 |
| pub\_rec | Number of derogatory public records |
| purpose | A category provided by the borrower for the loan request. |
| pymnt\_plan | Indicates if a payment plan has been put in place for the loan |
| recoveries | post charge off gross recovery |
| revol\_bal | Total credit revolving balance |
| revol\_util | Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit. |
| sub\_grade | LC assigned loan subgrade |
| term | The number of payments on the loan. Values are in months and can be either 36 or 60. |
| title | The loan title provided by the borrower |
| tot\_coll\_amt | Total collection amounts ever owed |
| tot\_cur\_bal | Total current balance of all accounts |
| total\_acc | The total number of credit lines currently in the borrower's credit file |
| total\_bal\_il | Total current balance of all installment accounts |
| total\_cu\_tl | Number of finance trades |
| 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 |
| total\_rev\_hi\_lim | Total revolving high credit/credit limit |
| url | URL for the LC page with listing data. |
| verified\_status\_joint | Indicates if the co-borrowers' joint income was verified by LC, not verified, or if the income source was verified |
| zip\_code | The first 3 numbers of the zip code provided by the borrower in the loan application. |

1. Lending Club is a P2P lending company whom connected between people wishing to lend money to people who wish to borrow it, in the American market. Lending Club established at 2007 in California, USA ([Company website](https://www.lendingclub.com/company/about-us), [wiki page](https://en.wikipedia.org/wiki/Lending_Club) ). [↑](#footnote-ref-1)
2. Baesens, Bart & Van Gestel, Tony & Viaene, Stijn & STEPANOVA, M & Suykens, Johan & Vanthienen, Jan. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society. 54. 10.1057/palgrave.jors.2601545. [↑](#footnote-ref-2)
3. Lessmann, Stefan & Baesens, Bart & Seow, Hsin-Vonn & Thomas, Lyn. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research. (doi:10.1016/j.ejor.2015. [↑](#footnote-ref-3)
4. Riza Emekter, Yanbin Tu, Benjamas Jirasakuldech & Min Lu (2015) Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending, Applied Economics, 47:1, 54-70, DOI: [10.1080/00036846.2014.962222](https://doi.org/10.1080/00036846.2014.962222) [↑](#footnote-ref-4)
5. This data is not exactly the same data set which we make use of. [↑](#footnote-ref-5)
6. This is a difficult task considering that we do not have access to a number of very important variables used by Lending Club. [↑](#footnote-ref-6)