# UCSD Extension Bootcampspot Data Science Class Final Project

# Lending Club Loan Analysis.

# Can we predict the likelihood of a loan being paid back based on past loans?

Amir Vazquez

Narjes Taqvaei

Tolga Colgar

Kevin Mickey

## Background

Lending Club was founded as a mechanism to facilitate peer-to-peer loans. Lending Club acted as a middleman between borrowers who sought loans between $1,000 and $40,000 and investors who provided the capital that Lending Club would distribute. Lending Club made money by charging the borrowers an origination fee between 1 and 5% of the loan amount and investors a 1% service fee.[[1]](#footnote-1) Investors could see a list of loans that had been applied for, and direct their investment to specific loans that they felt were investment worthy, in amounts as small as $25.

In 2020 Lending Club bought a bank, changed their business model and stopped funding loans via individual investors. However, existing loans will continue to be serviced by Lending Club until they are paid off or declared in default.

## Data

There is a dataset of all Lending Club loans from 2007 through Q3 2020 available in CSV format. This file clocks in at approximately 1.77GB and contains 152 columns of data and 2,925,492 rows.

Not all rows have every column filled out, and we determined that for our purposes, over 100 of the columns were not of interest. After studying the data dictionary that accompanies the data set, we settled on 39 columns of data to import. This had several effects. It made loading the data set at least 3x faster (on my machine, other results may vary) and it made examining the Pandas frames much easier.

We also programmatically examined the data and determined which columns had more than 30% of the data missing. If they reached or exceeded this threshold, they were dropped. Fortunately, our manual list also encompassed all the columns in the programmatic list. When done, the data set was 870.5 MB.

Once the data was loaded, some of the fields had to be cleaned up. For instance, the percent field was a STRING value because every cell was formatted < xx.x%>, and the term (length) field was also a STRING in the format <XX months>. To be able to use these cells in calculations, the text needed to be stripped out, and then the data turned into floats. Similarly, the dates were formatted as STRINGS using MONTH-YEAR format. The data entry was consistent, so it was relatively easy to convert them to Python datetime objects. During the conversion, they all defaulted to having first of the month dates, so May-2020 became 01 May 2020. Since they all did this, and the loans are quoted in months, not days, it isn’t an issue.

One field that was disappointing was the Purpose field. It had over 65,000 unique reasons for a loan, and as such, no useful trends could be extracted.

The data dictionary is included in the github as a separate document, the fields we kept are attached as an appendix.

For any remaining cells with NAN values, we replaced them with the mean value from that column, otherwise the machine learning process breaks.

Now that we have a clean set of data, we can begin exploring if there are factors that point to a loan being successfully repaid or going to collections.

**The Rise and Fall of Lending Club as a Peer-to-Peer Network**

**Chart, bar chart

Description automatically generated**

2016-17 marked a period when Lending Club was caught cooking the books, and as a result the CEO and several other high level executives resigned.

**Loan Origination vs Borrower FICO ScoreChart, scatter chart

Description automatically generated**There is a significant dip in loans made between the C5 and D1 categories, after which the Paid (orange) vs Written-off (blue) categories rapidly converge, becoming indistinguishable in the F category.

Chart, bar chart

Description automatically generated

Not surprisingly, the higher the FICO score, the lower the chance of default on a loan. The less than 670 had the highest default rate at 25%.

**Loan Origination vs Income**

Chart, bar chart

Description automatically generated

Not surprisingly Lending Club doesn’t do much business at the low or high end of the income scale, focusing on the 25-150K range.

### Machine Learning

A screenshot of a computer

Description automatically generated with medium confidence

In this data set we have unbalanced data with 80% of the data belonging to Fully Paid and 20% belonging to Charged Off (defaulted) class.

We need to test which features are best for ML section.

Data for the 2 classes

Graphical user interface

Description automatically generated

Find and drop any columns where 80% or more data is missing, resulting in the following list.

Graphical user interface, text

Description automatically generated

Finding the correlation between variables

We will now look at the correlation between the variables selected above. This will tell us about any dependencies between variables and help us reduce the dimensionality a little bit more

Chart, scatter chart

Description automatically generated

It can be seen from the plot above that loan amount and installment have a very high correlation amongst each other. This is intuitive since a person who takes a large loan would require larger payments to repay it (Loans are either 36 or 60 months). Interest rate, sub grade and grade have a very high correlation between them. This is obvious since the interest rate is decided by grades/sub-grades

Choose only features that provide valuable info:



Text

Description automatically generated

First we convert object data types to numerical data types.

A screenshot of a computer

Description automatically generated with medium confidence

**Categorical variables**

Since the analysis and the machine learning algorithms can’t take categorical or string variables directly, we have to create dummy variables for them. Since many of these variables have multiple categories, using weights can cause discrepencies in the algorithm. Instead, we will hot encode these so that we have a 1 wherever that category turns up and 0 otherwise. This will also create separate columns for each item in the category. Finally, we'll be dropping one of the categories so that we have N-1 columns instead of N.

Text

Description automatically generated with medium confidence

**Final Data Types**

Text

Description automatically generated

Loan amount  
Interest Rate  
Employment Length  
DTI (Debt to Income)  
Loan Status  
Total Payment  
FICO Range High  
Home Ownership: Mortgage  
 None  
 Other  
 Own  
 Rent

**Using Random Forest Classifier**:

We decide to use Random Forest specifically for this inbalanced data set and check model performance.

**Imbalanced data**

Graphical user interface, text

Description automatically generated

**Performance metrics for imbalanced data sets:**

**Confusion Matrix**: a table showing correct predictions and types of incorrect predictions.

**Precision**: the number of true positives divided by all positive predictions. Precision is also called Positive Predictive Value. It is a measure of a classifier’s exactness. Low precision indicates a high number of false positives.

**Recall**: the number of true positives divided by the number of positive values in the test data. Recall is also called Sensitivity or the True Positive Rate. It is a measure of a classifier’s completeness. Low recall indicates a high number of false negatives.   
99.91%

**F1**: Score: the weighted average of precision and recall.  
98.64%

Chart, treemap chart

Description automatically generated

**Confusion Matrix**:

The diagonal elements represent the number of points for which the predicted label is equal to the true label, while off-diagonal elements are those that are mislabeled by the classifier. The higher the diagonal values of the confusion matrix the better, indicating many correct predictions.

**Balanced data**:

Used equal number of training data for both loan status 0 and 1.

Text

Description automatically generated

Results after using balanced data:

Graphical user interface

Description automatically generated

**Compare models:**

**Text

Description automatically generated**

**Linear Discriminant Analysis**

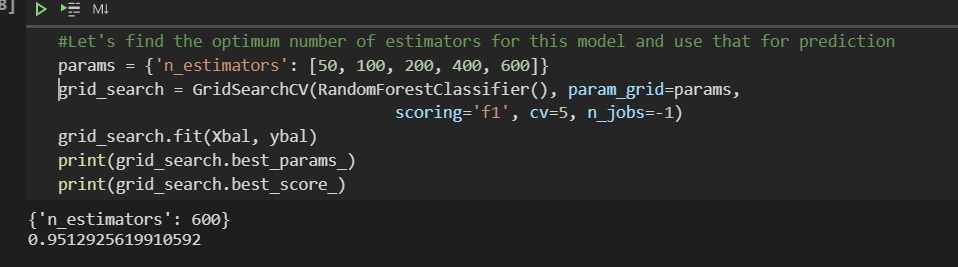
A classifier with a linear decision boundary, generated by fitting class conditional densities to the data and using Bayes’ rule.

The model fits a Gaussian density to each class, assuming that all classes share the same covariance matrix.

The fitted model can also be used to reduce the dimensionality of the input by projecting it to the most discriminative directions, using the transform method.

**Naive Bayes classifier for multinomial**:

The multinomial Naive Bayes classifier is suitable for classification with discrete features (e.g., word counts for text classification). The multinomial distribution normally requires integer feature counts. However, in practice, fractional counts such as tf-idf may also work.



95% Success rate

**Conclusion**

We explored different methods for dealing with imbalanced datasets:

Change the performance metric

Change the algorithm

Oversample minority class

Undersample majority class

It appears for this particular dataset random forest is among the best of the options we tried here.

## Appendix

Data Columns

loan\_amnt Loan Amount

term Term (in Months)

int\_rate Interest rate

installment Monthly Payment

grade Lending Club creditworthiness grade (A-G)

sub\_grade Lending Club creditworthiness subgrade (A1-A5 etc)

emp\_length Length at job (years)

home\_ownership Own, Mortgage, Rent, Any, None

annual\_inc Annual Income

issue\_d Loan Issue Date

loan\_status Current, paid or in collections

pymnt\_plan Appears to be a refinance/hardship repayment plan

addr\_state State of residence of borrower

dti Debt to income ratio

earliest\_cr\_line Earliest Credit Line

fico\_range\_low Fico Scores

fico\_range\_high

open\_acc How many open credit lines

pub\_rec How many derogatory public records

revol\_bal Total credit revolving balance

revol\_util % used vs % available revolving balance

total\_acc Total credit lines available

total\_pymnt Payments received on this loan

total\_rec\_late\_fee Total late fees on this loan

last\_pymnt\_d Last month a payment was recieved

application\_type Individual or joint

acc\_now\_delinq Number of accounts where borrower is delinquent

tot\_coll\_amt Total Collection amounts ever owed

tot\_cur\_bal Total current balance of ALL accounts

chargeoff\_within\_12\_mths Number of chargeoffs in last 12 months

mort\_acc Number of Mortgage accounts

num\_il\_tl Number of Installment Accounts

pub\_rec\_bankruptcies public record bankruptcies

tax\_liens Number of tax liens

tot\_hi\_cred\_lim Total high credit limit

total\_bal\_ex\_mort Total balance minus mortgage

total\_il\_high\_credit\_limit Total installment high credit limit

hardship\_flag Is borrower on hardship plan

debt\_settlement\_flag Loan has been charged off, is borrower working to fix

1. https://en.wikipedia.org/wiki/LendingClub [↑](#footnote-ref-1)