Rich Helle
August 13, 2018

Capstone Project

Project Overview

Lending Club is a peer-to-peer lending platform pairing individual investors with loan seekers. One of the draws of this platform for investors is the potential for higher returns because the traditional intermediary, a costly bank that shares little information, is replaced by a transparent online platform with minimal fees. Lending Club platform allows investors to individually select loans for their investment portfolio, providing certain data collected on each loan to aid the investor.

I invested with Lending Club using the automated investing tool, letting the platform choose the loans for me, this has not proved to be a good strategy. To date my adjusted net annualized return has been a dismal 0.07%. While the weighted average interest rate on my portfolio is above 10%, what's brought my return down to nearly zero is loan defaults which do not get fully repaid. My goal with this project is to use the data provided to determine how I might make a higher return.

Problem Statement

The problem I intend to solve with this project is to identify the feature values that best explain whether a loan will be repaid or default. Considering there are only two possible outcomes, this will be a binary classification problem.

Once the data is loaded I will explore the data visually, looking here and from other summary statistics for preprocessing that will need to be applied to the data. I anticipate applying s scaler to the data will help eliminate problems related to the scale of each feature. I also anticipate applying one hot encoding to any categorical features. The dataset will then be shuffled and split into training and test sets. These datasets will be separate and each model will be trained only on the training set.

With the data prepared, I'll apply a benchmark model, a K-nearest neighbors classifier. This algorithm has appeared in research on related topics and performs relatively well.

With a benchmark established, I will use methods applied elsewhere to similar problems with good results. Specifically, I will apply a support vector machine and ensemble methods, such as random forests, Adaboost and boosted gradients. Additionally, I will apply a naïve Bayes classifier, which I have not come across in other research but which I expect will perform at least as well.

Finally I'll apply a neural network to the problem. I expect this method will result in the optimal model, however significant parameter tuning may be necessary to achieve outstanding results.

Each model will be evaluated on the test set, with the paidoff precision metric used for evaluation.

Metrics

The most important metric will the precision of the model on the paidoff target variable. The reason precision is the most useful metric is that the investment return will depend on the repayment of the loans in which I invest. Precision here will measure the ratio of true positives, loans repaid, from the loans predicted to repay, true positives plus false positives.

Even if model only predicts a small percent of available loans will repay, this is unlikely to limit the model usefulness. There are a lot of loans available for purchase at any given time, as of this writing there are 164,596 loans available for investment and new loans arrive every day. My next investment of $1,000 can be invested in as few as one loan but no more than 40 loans, as the smallest investment in any loan is $25.

Precision: $\frac{True\ Positives}{True\ Positives+False\ Positives}$

Analysis

The Dataset

The dataset that will be used, linked provided in appendix, is provided by Lending Club and covers the loans issued on their platform from 2007 through 2011. This dataset was selected because each loan has terminated, either being fully repaid or charged off as defaulted. The original file consists of two parts, the loans made on the platform and loans which were declined. The declined loans were removed since these available to investors. The resultant data consists of 39,786 loans with 145 features. A copy of this data and the data dictionary in the zip file of this capstone proposal.

The features in the dataset can largely be broken down into three categories; borrower related features, loan related features and repayment related features.

Information related regarding the borrower, this includes information on their income, home ownership, residency, credit availability, credit usage. These are intended to help the investor adjudge the creditworthiness of the borrower and therefore their likelihood of repayment.

The loan features include the amount of the loan, the interest rate charged, loan term, and the monthly payment. These are amounts calculated as a result of the loan requested and the interest rate applied to the loan.

The repayment features relate to the performance of the loans during its life. This includes the amount pf principal repaid, the interest paid, and late payment information. Since this all occurs after the investment, these features will not be included in the analysis.

There is an imbalance in the target class between the number of defaulting and repaid loans. Defaults represent only 14.25%, so there are approximately 6 times as many repaid loans to unrepaid loans.
The data includes many fields, not all of which are observable at the time of the investment. Since the goal is to select loans based available information, the model input will be restricted to the data available at investment. Screenshots of the information available at investment and the filter options available for sorting loans have been included in the zip file of this capstone proposal.

Related Research

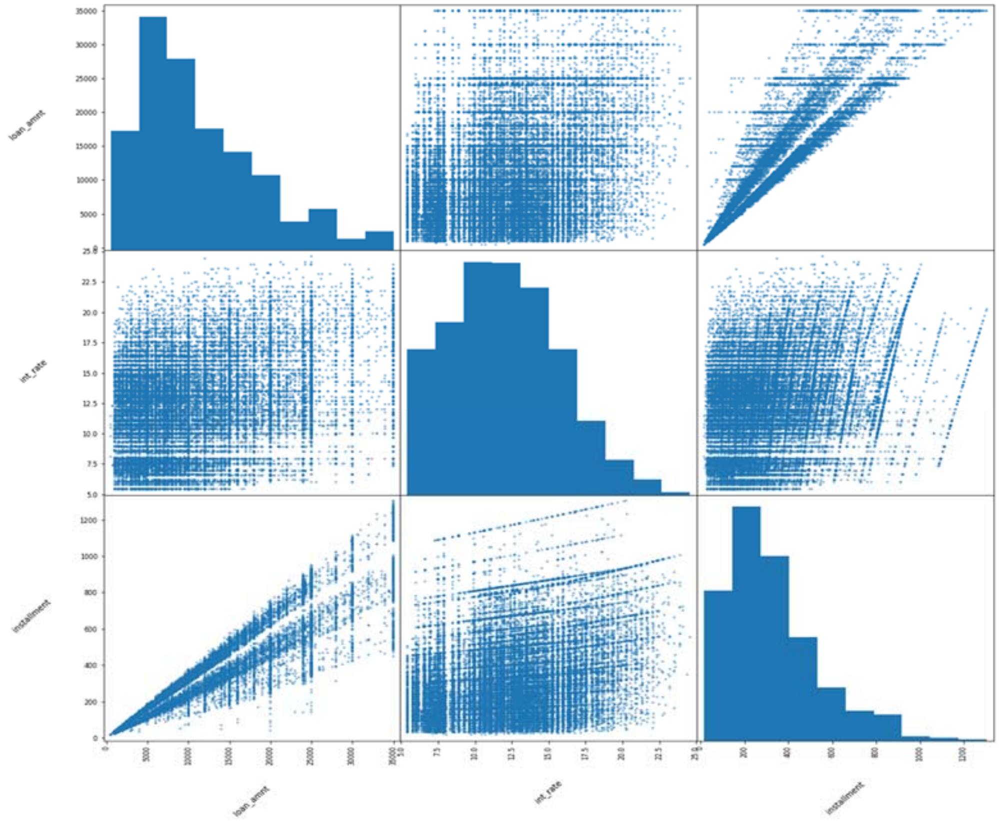
(Dabbura, 2018) undertakes a very similar project with a very similar dataset, ending in 2010 rather than 2011 and looking at a smaller subset of the features available. With the reduced feature set, he elects to drop observations with missing entries. He applies various ensemble methods finding they all preform about equally well. Notable he does not apply a naïve Bayes classifier, which may have enhanced his model performance.

(Steele, 2018) examines the repayment rates of mortgage loans using K-nearest neighbors and decision trees. These sort of loans may perform differently from the unsecured loans Lending Club provides to its borrowers. The dataset used consists of mortgages funded by Fannie Mae, for which the property would be the borrower's primary residence meaning it is their home, likely the borrower's largest investment and the equity in the home may be the borrower's most significant asset. Default on the mortgage would endanger these for the borrower and therefore they are less likely to default on the mortgage. In contrast the loans from Lending Club are unsecure and default may have less significant implications for borrowers. Steele finds that the two methods yield very similar results in the scoring metrics applied.

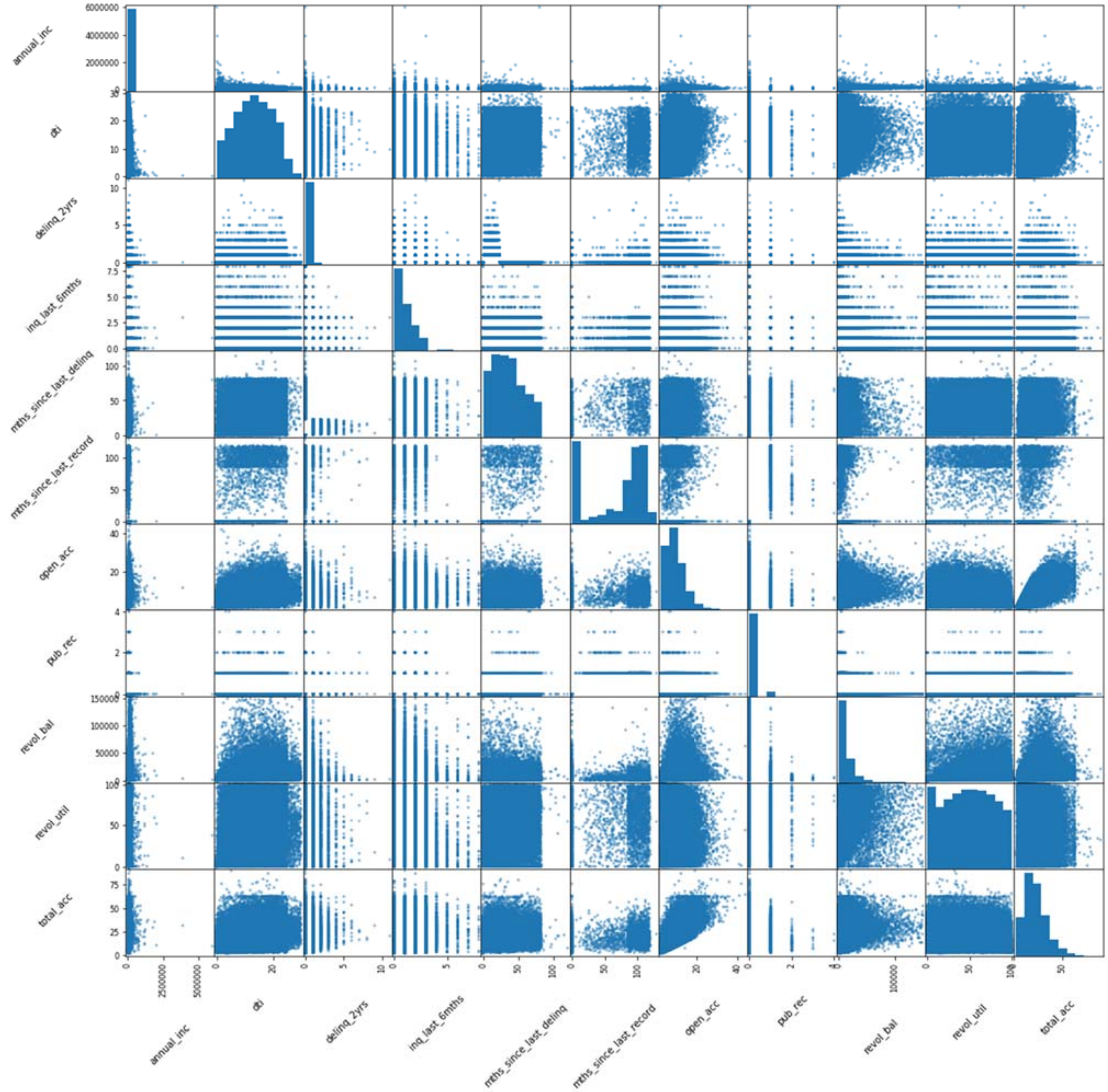
Data Exploration

Looking at the data, some of the fields convey no information because they contain only one value, either the same value or NaN. With these features removed, scatterplots of the loan features and the borrow features will give some insight.

Loan Features Scatterplot



Borrower Feature Scatterplot



Algorithms and Techniques

To identify loans that are likely to repay I'll be using a variety of supervised machine learning algorithms. The benchmark algorithms will be K-nearest neighbors. Its unclear how many neighbors to use for this technique, iterating over a range it seems there is little difference. The benchmark will use three neighbors since it performed best the paidoff precision. Related research, discussed above, indicated there was little difference in the performance of a variety of ensemble methods. K-nearest neighbors was chosen in part because of this and because it is an easily understandable method or readers to understand.

The algorithms I'll explore fall into three categories; support vector machines, naïve Bayes and ensemble methods. Comparing the performance on the training and test set, it is clear overfitting is not an issue.

Support Vector Machine

This algorithm generally performs well at classification tasks. I expect it will perform at least as well as the benchmark. My concern is that it that the radial basis function may be able to distinguish each point well with the default settings used.

Naïve Bayes

This probabilistic model will produce accurate parameter estimates for the features. The training set is not very large, and this is where the naïve Bayes tends to perform best.

Ensemble Methods

Other research on credit default has shown that random forests performed well in predicting defaults. I hope to repeat their success. The addition of boosted algorithms, in the form of Ada boost and Gradient boosted trees, will further the models efficacy.

Methodology

Before the dataset can be fed into the learning models some preprocessing is required. For this project, there are NaN values that will need to be addressed, numeric features will need to be explored for skew, a scaler adjustment will reduce the effect of feature scale from skewing the learning and the categorical features will need to be encoded.

NaN Values

Examining the data for NaN values, there are at least 36,995 loans containing at least one NaN, value. Considering there are only 39,786 loans in the dataset, excluding loans with NaN values is not an option.

Looking at the scatterplots for the features containing NaN values; mths_since_last_delinq, mths_since_last_record, and revol_util. The first two contain mainly zero entries, it also seems likely that a NaN here is because there were no instances of delinquencies or records. I believe replacing the NaN with zero will interrupt the dataset the least.

For the revol_util feature, the mode entry is zero. It also seems reasonable that the NaN value is an indication of a lack of a revolving credit facility, in which case filling in zero will interrupt the dataset the least.

Skewed Numeric Values

Reviewing the skew calculation of each numeric feature indicates there is a considerable amount of skew in the annual income feature. There is also some skewing in the revol_bal feature. To help condense the scale of these features, a logarithmic transformation will be applied.

The other features with a skew greater than 2 are features that are integer based and have few large values. For these reasons, I'll not be applying a transformation to these features.

Scaler

The varying scale of different features can impact their importance to the learning algorithm, to minimize this effect, a scaler will be applied to the numeric features. Specifically a sklearn's min max scaler which will scale each feature between zero and one.

One Hot Encoding

Finally one hot encoding will be applied to the categorical features, these features are difficult for the models to interpret without such an encoding.

I'm concerned about some features having so many unique values that the one hot encoding will drastically increase the scale of features and their sparsity will distort the learning algorithm. In our dataset there are three features with a large number of unique values. These will be removed from the dataset as part of preprocessing.

With one hot encoding applied to the reaming features there are now 142 features, where previously there are 22 before the encoding.

Implementation

With the data now preprocessed and ready for use in the learning algorithms, we break the data into training and test sets.

The test sets are used to train first the benchmark K-nearest neighbors model. Its unclear how many neighbors will produce the best results, iterating over the range there is little difference. The benchmark will use three neighbors since it performed best the paidoff precision.

The support vector classifier, naïve Bayes classifier, random forest classifier, Adaboost classifier and gradient booting classifier are in each trained on the data and evaluated on the test set for paidoff precision. The results clearly indicate the naïve Bayes performs best on the metric of concern.

Algorithm	**Paidoff Precision**
KNN (n=3) (Benchmark)	0.87
Support Vector Classifier	0.86
Naïve Bayes Classifier	0.95
Random Forest Classifier	0.86
Ada Boost Classifier	0.86
Gradient Boosting Classifier	0.86

With the naïve Bayes strongly outperforming the other classifiers, I've let go of the idea of turning the classifier.

Neural Network

With the naïve Bayes so clearly outperforming the other classifiers, I was curious if a neural network might be more suited to this task. For this I constructed a three layer dense neural network with 128 neurons per layer with relu activation functions feeding into an output later for classification.

Training the network for 100 epochs, the network performed incredibly well on the training set but performed no better than the benchmark on the test set.

Results

Model Evaluation and Validation

The paidoff precision is the metric by which the models will be judged. Reiterating the points made earlier, if only loans identified by the model as repaying are selected for investment, then the precision for this metric is all that will matter. The loans identified as not repaying will be avoided, rendering them unimportant for investment performance. There may be some concern that the model identifies very few loans as repaying, but as noted earlier the intended investment is small relative to the number loans available for investment at any given time.

The naïve Bayes model clearly outperforms the other models on the paidoff precision metric, and is therefore selected as the final model. The disparity between the naïve Bayes performance the other models was significant enough that tuning the other models seemed unlikely to yield equivalent results without significant exploration.

The final model provides a good solution for the problem. Had I applied this model to my own initial investment I would have experienced a default rate of only 5%. As it happens my performance was much worse, see the Lending Club Note Performance Summary in the appendix, suffering a default rate of 27%.

The final model fit with my expectation that a supervised machine learning algorithm could better identify loans likely to repay better than the Lending Club platforms automatic investment methodology. In part, my poor performance may have been due to the alleged unethical practices in its offerings to retail investors, see link in appendix.

This model could be used in a general setting for small investors. The importance given to the paidoff precision metric could cause strain for an investor seeking a large portfolio, however a more useful model for that case could be devised using the same algorithms.

Justification

The model outperformed the benchmark, on the paidoff precision. The benchmark scoring 0.87 and naïve Bayes scoring 0.95.

Conclusion

Reflection

I was surprised the neural network didn't perform better than even the benchmark, this is likely due to overfitting, but I expected the disparity between the neural network and naïve Bayes would have been less.

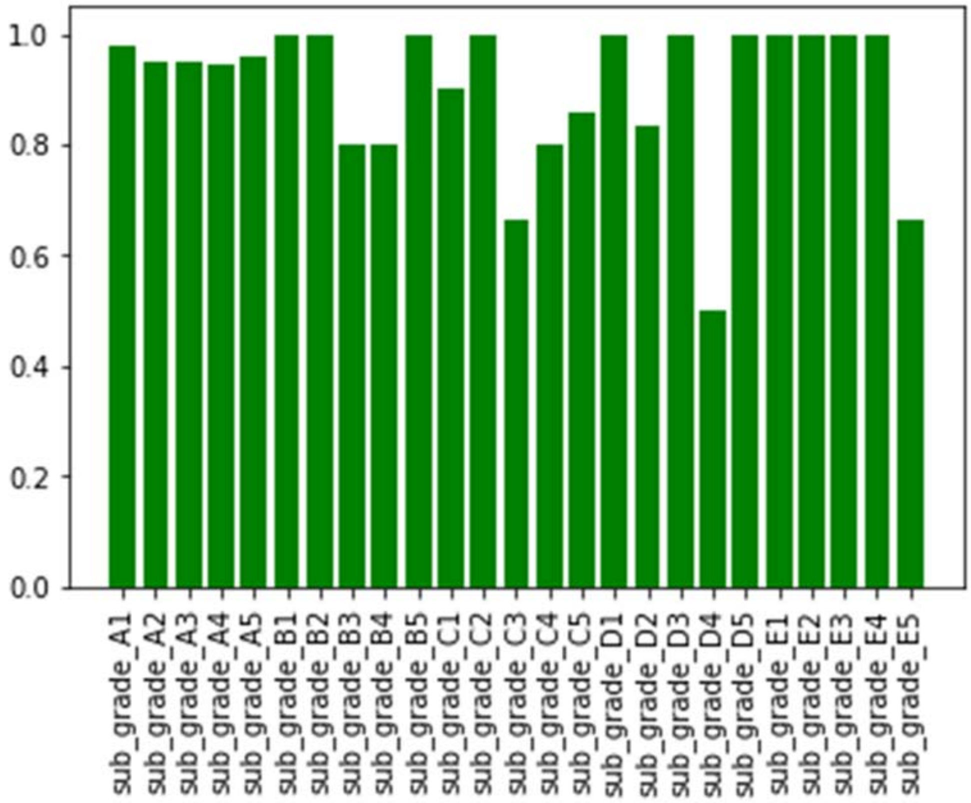
I was surprised how quickly the naïve Bayes classifier trained on the set given that its performance was well above the other models. The support vector machine and KNN models each took a significant amount of time to run in comparison.

I’m very pleased the model is able to provide some insight to which loans perform best. The task of picking loans manually is daunting given the amount of data available on each and the size of the portfolio available for investment.

Freeform Visualization

Some of the grades of loans are more likely to repay than others. Looking at the results of the naïve Bayes on the test set, the performance of the portfolio can be enhanced by looking taking this into consideration. Since the risk of default generally decreases with higher grades, A1 being the highest and E5 the lowest, it is surprising the performance of B1 and B2 is higher than all the A subgrades.

Naïve Bayes Precision Performance by loan Subgrade



Improvement

There are several areas that might yield improvement in the model and its usefulness.

From the existing dataset, the elimination of outliers in terms of income and revolving balance might could reduce the range of each variable significantly. This in turn would make the variability in remainder of the range more significant when the minmax scaler is applied.

The dataset could be increased by including completed loans from datasets in later years. This could have a couple harmful effects that would need to be considered. For example loans made in later years included both three and five year terms, data in the set used were all three years. More of the three year loans will have reached completion than the five year loans, possibly introducing a bias. Since some of the loans in the last three to five years have not reached term, it also means there would be a disproportionate number of defaulted loans since these nearly always end before the full term of the loan.

The neural network also presents significant opportunities to improve model performance. A custom loss function could apply more penalty to false positives. Network turning via the number of layers, training epochs, use of dropout layers and activation functions could enhance performance.

Works Cited

Dabbura, I. (2018). *Predicting Loan Repayment.* Towards Data Science. Retrieved from https://towardsdatascience.com/predicting-loan-repayment-5df4e0023e92

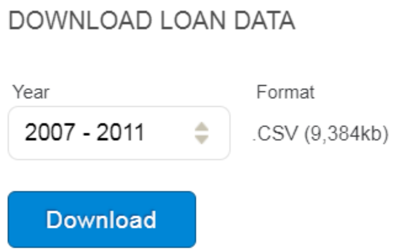
Steele, M. (2018). *Hands-On Machine Learning-Predicting Loan Delinquency.* Riskspan.com. Retrieved from https://blog.riskspan.com/hands-on-machine-learning-predicting-loan-delinquency

Appendix

Lending Club Dataset

https://www.lendingclub.com/info/download-data.action

With the options noted below.

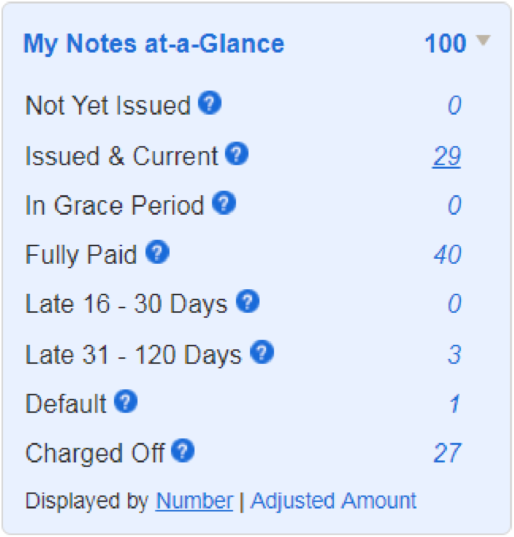


DOWNLOAD LOAN DATA

Year: 2007 - 2011

Format: .CSV (9,384kb)

Lending Club Note Performance Summary



My Notes at-a-Glance	100
Not Yet Issued	0
Issued & Current	29
In Grace Period	0
Fully Paid	40
Late 16 - 30 Days	0
Late 31 - 120 Days	3
Default	1
Charged Off	27

Displayed by Number | Adjusted Amount

Lending Club Scandal

https://bestcompany.com/loans/blog/a-running-timeline-of-the-rise-and-fall-of-lending-club