Macroeconomic Predictors of Credit Default

Code for this project, in addition to a slide presentation summarizing the process and results, is available in the following public GitHub repository:

https://github.com/mhardcastle0/Springboard/tree/master/Lending_Club_Default_Prediction

Part I - Introduction - Why is Credit Risk Important, and For Whom?

The Problem - Lending Risk and Exposure

Companies extending credit bear a significant risk that borrowers will not repay what is owed. Unsecured personal loans—closed-end loans not collateralized by houses, cars, or other assets—have grown significantly in popularity as online lending platforms have allowed companies to extend credit to borrowers with very short application and funding timelines.

Rapid online underwriting opens companies to instant nationwide exposure to fraudulent, low-credit, and other high-risk borrowers. Characteristics specific to the borrower have often been the only factors that determine whether the borrower will receive credit - characteristics such as borrower income, education level, credit score, or other credit attributes found in a credit report.

There are, however, several macroeconomic factors that influence borrowers' ability to repay loans, in aggregate - a contracting economy, all else equal, will typically expose lenders to greater risk, as borrowers are more likely to lose jobs or assets and to be unable to repay their loans. This project evaluates whether incorporating widely-available macroeconomic features into models can improve predictions of the likelihood that a borrower will default on his or her loan obligation.

The Client - Lenders of All Stripes

This information is valuable to any client originating loans - while macroeconomic impacts may be felt the hardest by companies that hold unsecured loans, it is likely that there is an association between macroeconomic factors at origination and loan performance for all lending segments. The clients most directly related to the project are those with data that resembles the unsecured personal loan data that is evaluated, such as Lending Club, Prosper, Avant, or Marcus by Goldman Sachs. Other unsecured products, which predominantly take the form of credit cards or retail installment loans (such as those offered by Affirm or Bread), would likely face similar exposure to macroeconomic factors as unsecured personal loans. Any results are likely to generalize to secured lending products, potentially to a lesser degree, such that a client could include any members of the, mortgage, HELOC, or other lending industries.

The Data - Loan Tapes and Macroeconomic Data

Lending Club is a Financial Technology company that offers unsecured personal loans ranging from $5,000 to $40,000 for 36- or 60-month terms. Applications are typically underwritten automatically, and lending decisions are reached in seconds. Several of Lending Club's loans are "peer-to-peer" - any individual can choose to purchase a loan that Lending Club has originated and not yet sold or securitized. Because of this peer-to-peer model, there is significant transparency over the company's loan portfolio; detailed data on each loan in the company's portfolio is publically available for would-be investors to analyze. This data, sourced from a flat file on Kaggle, was the basis of the analysis performed in this project, and provides the "outcome variable" - loan default - that is used. Data on monthly bankruptcy rates was sourced from Epiq Global, an organization that collects and publishes statistics on the number of monthly bankruptcy filings by region. Further macroeconomic data was downloaded from the public Federal Reserve Economic Data (FRED) published by the St. Louis Federal Reserve, which contains a large variety of macroeconomic data from which the seemingly most macroeconomically-relevant data was sourced.

The Project - Determining Who to Lend To

Keeping defaults low and setting interest rates at levels that are commensurate to a customer's credit risk are necessary to ensure that loans remain profitable. Lenders rely heavily on models predicting how individual borrowers' characteristics indicate their default risk—how a prospective borrower's income, years of job experience, credit score, or other factors in his or her credit report predict whether the loan will be repaid.

This project evaluates not only the impact of individual borrower characteristics, but how broader, non-individual factors may impact borrowers' default rates. A rise in nationwide bankruptcy rates, for example, may indicate a general deterioration of credit quality. Macroeconomic factors such as changes in stock market indexes, the real growth rate of the economy, or the unemployment rate might similarly predict changes in repayment rates. Proxies and measurements for each of these data points, and a large variety of others, were explored to evaluate whether adding macroeconomic factors improves a lender’s ability to identify unprofitable loans, before committing capital to them.

Part II - Sourcing and Cleaning the Data

The Primary Data - Lending Club Loan Tape

The primary datasource is the Lending Club dataset, which is hosted as a flat file on Kaggle. Preparing the data source required little modification: each row represents a loan with an origination month, term length, delinquency status, and several other credit- and performance-related metrics. The data contains several fields that may be relevant to analysis, but some of them would not be known at the time of origination and were discarded. Others, such as geographic features, can not be used to make lending decisions for legal compliance and disparate impact considerations and were also dropped. All of the fields that would have been known at the time of origination and which could be used for underwriting were maintained.

The dataset was evaluated for any missing or unexpected values that might impact analysis and was found to be very clean. Loan amounts at origination, origination dates, and loan statuses were evaluated for unexpected values and were all found to be in-line with expectation. Additionally, no missing values that would impact analysis were observed.

Only loans that would have completed their term lengths are relevant to this analysis. For this reason, all loans were removed from the dataset if they would not have seen through their term, plus one month to account for no payments being due in the first month, and three months for the possibility for loans to enter hardship and extend their term length by up to three months.

Macroeconomic Data

A variety of data sources were selected to act as proxies for macroeconomic health. They are generally sources that are known to be good predictors or indicators of macroeconomic well-being, and include:

1. The number of bankruptcies filed nationally during each month. Bankruptcies are an important predictor for loan default, as a larger number of bankruptcies very directly impacts lending companies’ expected number of borrower defaults. Additionally, the bankruptcy rate can serve as a proxy for the health of the economy - if more people are declaring bankruptcy, then it is expected that fewer people are economically well-off, all else equal.

2. The TED spread. This is the gap between the 3-month London Interbank Offer Rate and the three-month treasury bill interest rate, which is a measure of interbank lending risk. Higher TED spreads are an indication that the market believes that there is risk to lending to banks, which might indicate a weaker macroeconomy. This was a significant early-warning indicator during the 2008 recession, which differentially impacted banks and the finance sector as a whole.
3. The St. Louis Fed Financial Stress Index. This is an aggregation of 18 weekly data series that is designed to provide an aggregate measure of economic health - seven interest rates, six yield spreads, and five other indicators. The series is designed to average at 0, with values greater than 0 representing instability and values less than 0 representing stability.
4. The Civilian Labor Force Participation Rate. This measures the proportion of 25 to 54 year olds who are actively engaged in the workforce. There are several measures for unemployment other than the official Unemployment Rate; this measure is perhaps the broadest assessment of employment.
5. Total Vehicle Sales. This is the number of vehicles purchased in a given month. When people are concerned with job security, they are less likely to make large durable purchases like vehicles, so this may be a leading indicator of poor economic health.
6. Consumer Sentiment. This data is derived from a survey conducted by the University of Michigan each month. Questions are designed to solicit a broad overview of consumers' sentiment on the economy, including personal finance, spending, and the business climate. The results are aggregated into an index, where 100 represents consumer sentiment in 1966Q1.
7. The unemployment rate. This is defined as the proportion of people in the country who are both *looking for work* and who are not employed. Lower unemployment represents a higher likelihood of a borrower maintaining his or her job or getting another job in times of unemployment, both of which would predict higher repayment rates.
8. The change in the S&P 500 each month. Stock indices broadly track the health of the economy and rapidly respond when new information is available. These factors make this index a potential early indicator of economic performance.

Other than the Bankruptcy data sourced from Epiq Global, all macroeconomic data was extracted using the Federal Reserve Economic Data (FRED) API. As the Lending Club data is available only at the monthly level, several data series were modified to change daily or weekly datapoints into monthly datapoints. The S&P 500 data was sampled to include only the prices at the beginning and end of each month, in addition to the average monthly value. This data contained several missing values, as stocks are not traded on some weekend and holiday days; for any day that was missing stock S&P 500 data, the data from the previous available day was used. Several of the macroeconomic data sets are at weekly or daily granularity - for these, monthly values of each day's or week's data that fell during the month were averaged. The TED spread is a daily indicator - the average value of each day in each month was used. The St. Louis Federal Reserve Economic Stress Index is available weekly, so each value that was recorded during a month was averaged to reach that month's value.

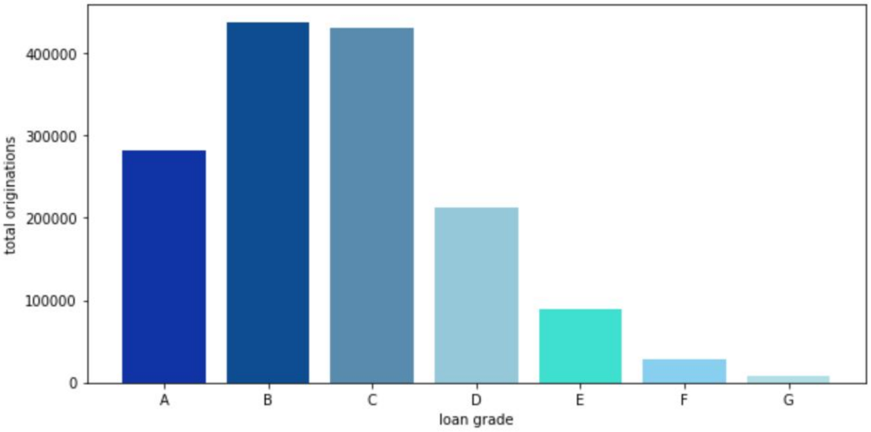
The bankruptcy data, sourced from an Excel file provided by Epiq Global, required the most modification. Several formatting issues were present, including merged cells containing the year portion of each period and summary and location-specific rows of data that needed to be dropped. The file required reading into a Pandas dataframe, transposing the data, removing superfluous columns, forward-filling the years to accommodate data missing due to merged cells, and merging the years and months into a single datetime column.

Finally, all of the data sources were merged together. The monthly bankruptcy quantities and all data sources from FRED were joined into the modified Lending Club dataset, where each row contains the data that is provided by Lending Club, as well as a column for each of the macroeconomic variables during the row corresponding to the month of that loan's origination.

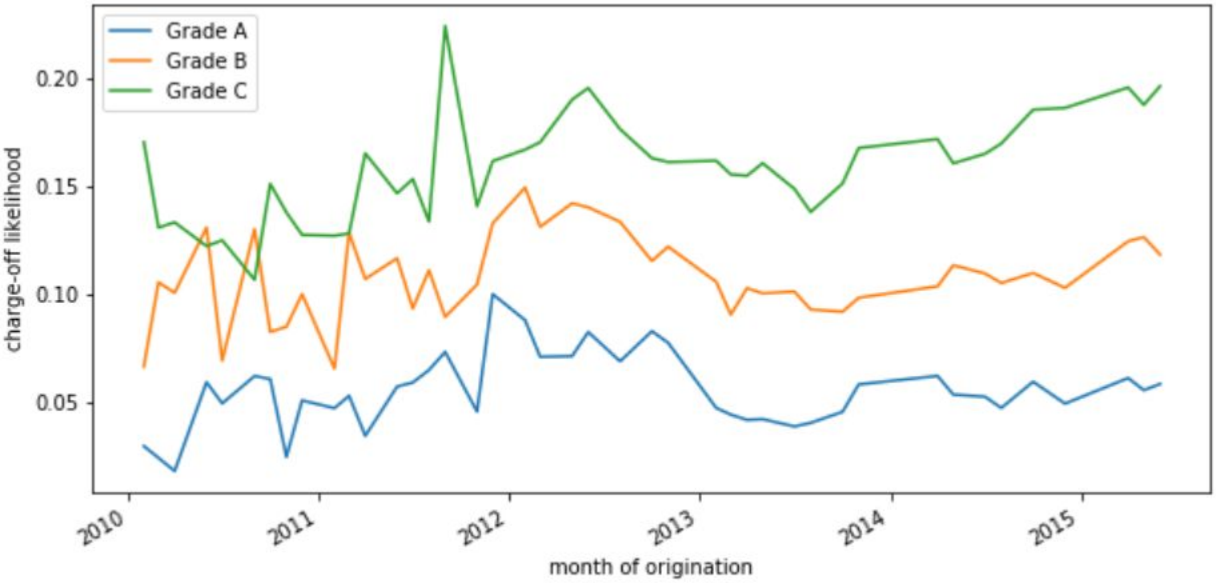
Part III - Exploring the Data

Initial Portfolio Exploration

To begin, an initial exploration of the loan portfolio was performed. The Lending Club portfolio is broken into grades A through G, with A representing the lowest-risk borrowers, and G representing the highest-risk borrowers. The total number of loans in each risk grade is shown in the graph to the right. Grades A through F have tremendous numbers of originations, from around 30,000 for grade F to over 400,000 for grades B and C, providing a sufficient sample for statistical testing.

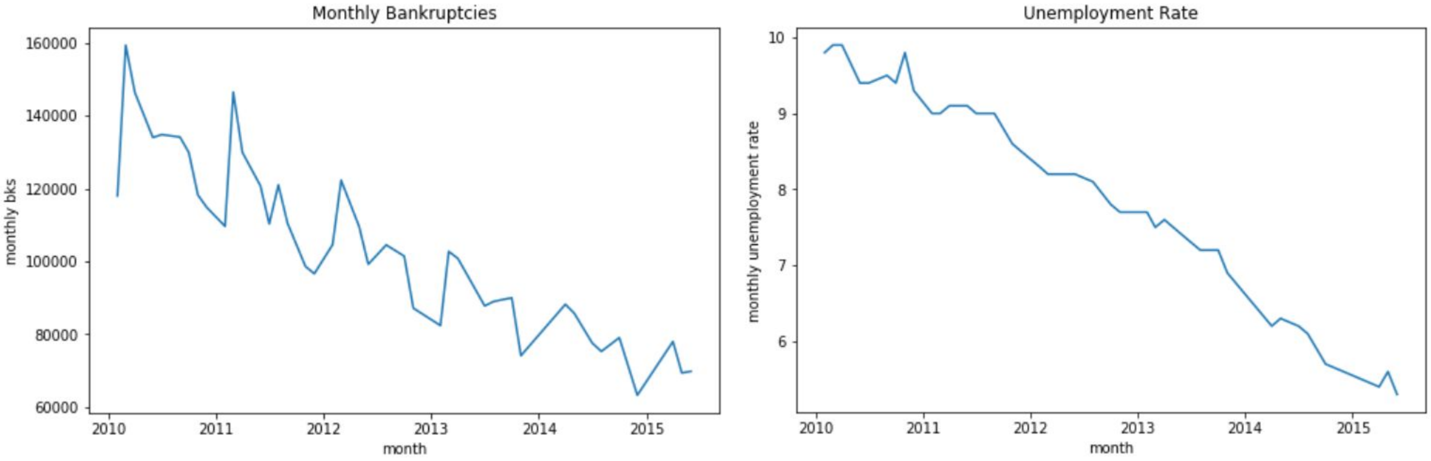


Next, default rates - and how they vary between the risk grades - is evaluated. The graph below shows the proportion of loans that were charged-off at any point, which is equivalent to the loans not being paid-in-full. Charge-off rates are higher for higher-grade loans, as expected.

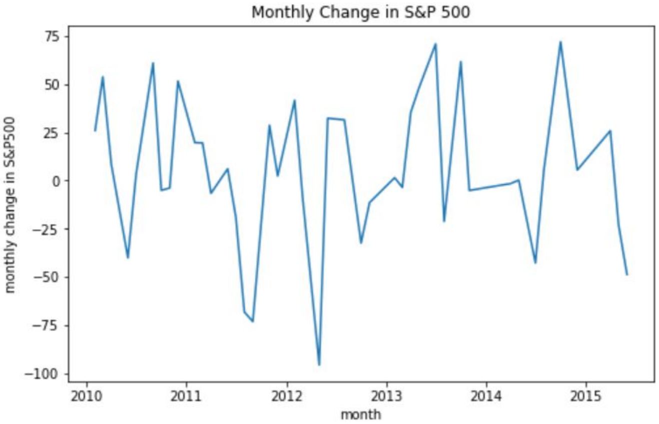


Macroeconomic Factors

Next, three macroeconomic variables of interest - the unemployment rate, the national monthly personal bankruptcy rate, and the monthly change in the S&P 500 - are explored. It is critical that there is variation in each variable that is tested, as a completely stable indicator would not be a usable predictor. The graphs below show that the rate of monthly bankruptcies and the unemployment rate have both fallen over the period for which loan data is available, with seasonality in the bankruptcy data.



The graph to the right shows a similar graph for the monthly change in the S&P 500 data, which evinces less of a trend and more of a "random walk." There is large variation for all variables, which should allow for relationships between each variable and the default rate to be evaluated.



The other macroeconomic variables were not explored in a similar fashion - the three tested are assumed to be roughly representative of the macroeconomic data.

Part IV - Basic Statistical Evaluation

Ordinary least squares linear regression was used to evaluate whether the hypothesized relationships exist between macroeconomic factors and the rate of loan defaults. Each of the three macroeconomic variables of interest shown above - the unemployment rate, the number of national monthly bankruptcy filings, and the monthly change is the S&P 500 - was correlated with the likelihood that a loan would default.

Each loan grade and term length was treated as a separate subgroup for the purpose of this analysis. For example, 36-month grade A loans, 36-month grade B loans, and 60-month grade A loans were all evaluated separately. This is done for two reasons: firstly, changes in portfolio composition over time might change portfolio-wide default rates; if there are more F-grade loans in later periods than earlier periods, then portfolio-wide default rates would likely increase over time regardless of macroeconomic impacts. Secondly, macroeconomic impacts might vary by risk tier—higher-risk borrowers may be more impacted by the unemployment rate than lower-risk borrowers, for example.

The table in the appendix shows the resultant confidence interval and sample size for each specification.

The large majority of results are not statistically significant. Of those that are significant, all results that correlate the unemployment rate with charge-off rates have the opposite sign than expected - a lower unemployment rate at the time of origination is correlated with a higher default likelihood. Similarly for the bankruptcy rate - lower monthly national bankruptcies at origination are associated with higher default rates.

There are two significant results correlating the monthly change in the S&P 500 with default rates, but one term is positive and the other negative, suggesting low generalizability of any results.

These unexpected results underscore the need for more rigorous evaluation - a simple comparison of macroeconomic variables to the outcome variables of interest provides little information, as the economy was improving for the entire range of the data being evaluated. Next, machine learning tools were used to provide a more targeted answer to the question of interest: can adding macroeconomic data to machine learning models of loan performance improve the models' accuracy?

Part V - Machine Learning Evaluation

The Setup

Next, the data was prepared for Machine Learning applications. The data was split into a matrix of features and a 1-dimensional matrix containing the outcome variable, the dummy variable indicating whether the loan defaulted. The data was split into 70% training and 30% test groups. The features were standardized so that larger-magnitude features did not have outsize impacts on the results.

Benchmarking With Lending Club Data Only

Next, three machine learning models were applied to the data: KNN, SVM, and Random Forest. Each was tested using 3-fold cross validation and using a large number of combinations of classifier parameters, with the best result used. The outcome variable that was evaluated was the "good-to-bad" ratio, a standard way of evaluating changes in loan underwriting that assesses the number of "good" non-defaulting borrowers who would be turned away for every one "bad" defaulting borrower kept out. Results for each test are shown in the table below. All models performed quite well: the worst by far, KNN, had a Good:Bad ratio of .9317, which would likely make implementation of the model profitable; a "bad" borrower can result in an entire loan balance being unpaid, while a "good" borrower contributes only interest payments and small other fees to profits, making a .9317 good:bad ratio very positive. The other two models are far better, with good:bad ratios of .5265 and .5509, respectively, for SVM and Random Forest.

Model Name	False Positives	True Positives	Good:Bad Ratio
KNN	4121	4423	.9317

SVM	2225	4226	.5265
Random Forest	2236	4059	.5509

Evaluating Macroeconomic Data

Next, a variety of economic data was incorporated into the dataset. Each of the macroeconomic variables described previously was incorporated into the lending club data.
Once these economic predictors were incorporated into the data, the data was prepared and the models ran exactly as before. The table below shows the good:bad ratio for the data which includes the newly incorporated features. All models have a more favorable good:bad ratio than the models that do not incorporate macroeconomic features.

Model Name	False Positives	True Positives	Good:Bad Ratio
KNN	3749	4079	.9191
SVM	2430	4698	.5172
Random Forest	2203	4487	.4910

The magnitude of the increase is quite large for the best-performing model, Random Forest: the number of false positives barely changed (from 2236 to 2203), but the number of true positives increased from 4059 to 4487. The reduction of 400 bad loans could have saved Lending Club several hundred thousand dollars, making incorporation of macroeconomic variables into lending models a potentially extremely profitable exercise.

Part VI - Conclusion

It is typical for lenders to use entirely borrower-specific data to evaluate whether an applicant should be extended credit. This project demonstrated that lending models could be improved dramatically by incorporating macroeconomic variables in decisions. The macroeconomic variables that were used represent a very small subset of the macroeconomic variables that are available - the FRED data source alone boasts over 500,000 economic datasets, and these could potentially be engineered and combined in any number of ways to extract more information from them. And potentially more powerful machine learning models, like boosting models, were not used - these could extract further signal out of the macroeconomic data. Overall, incorporation of macroeconomic variables is a potentially potent tool for evaluating whether an applicant should receive a loan.

Appendix – Confidence Intervals for Each Subset

The table below shows the confidence interval for the slope term for each variable of interest, broken out into each subgroup tested. Terms that are significant at the .05 level are in bold.

Unemployment Rate %

Term	Grade	95% CI Lower Bound	95% CI Upper Bound	Sample Size
36 months	A	-0.00135949	0.001641	60795
36 months	B	**-0.00468227**	**-0.00099431**	90624
36 months	C	**-0.01494143**	**-0.00967629**	68178
36 months	D	**-0.02550619**	**-0.01771471**	34891
36 months	E	**-0.04220207**	**-0.02550099**	9765
36 months	F	**-0.07110762**	**-0.02906948**	2113
60 months	A	-0.02670311	0.02487238	624
60 months	B	-0.00647652	0.02579703	4049
60 months	C	-0.00800087	0.02261887	6081
60 months	D	-0.02629441	0.012933	4112
60 months	E	**-0.04510762**	**-0.00650016**	4500
60 months	F	**-0.0741704**	**-0.01809013**	2380

National Monthly Bankruptcies (per 1,000)

Term	Grade	95% CI Lower Bound	95% CI Upper Bound	Sample Size
36 months	A	-7.85E-05	1.30E-04	60795
36 months	B	-2.40E-04	3.26E-05	90624
36 months	C	**-0.00095931**	**-0.00058006**	68178

36 months	D	**-0.00161385**	**-0.0010405**	34891
36 months	E	**-0.00260701**	**-0.00144237**	9765
36 months	F	**-0.00435358**	**-0.00160429**	2113
60 months	A	-0.00123861	0.00151179	624
60 months	B	-0.00044933	0.00099869	4049
60 months	C	-0.00084185	0.00061737	6081
60 months	D	-0.00137703	0.00043214	4112
60 months	E	**-0.00211259**	**-0.00030859**	4500
60 months	F	**-2.79E-03**	**-9.35E-05**	2380

Monthly Change in S&P 500 (EOM value minus BOM value)

Term	Grade	95% CI Lower Bound	95% CI Upper Bound	Sample Size
36 months	A	-8.49E-05	9.35E-06	60795
36 months	B	**-1.23E-04**	**-1.88E-05**	90624
36 months	C	-1.12E-04	3.08E-05	68178
36 months	D	-1.45E-04	7.54E-05	34891
36 months	E	-8.70E-05	3.80E-04	9765
36 months	F	-5.19E-04	0.00059952	2113
60 months	A	-4.22E-04	0.00090613	624
60 months	B	-3.11E-04	0.00028922	4049
60 months	C	-1.94E-05	5.79E-04	6081
60 months	D	-3.90E-04	0.000366	4112

60 months	E	-4.04E-04	0.00034348	4500
60 months	F	**4.41E-05**	**1.11E-03**	2380