How to Deal with Extreme Cases for Credit Risk Monitoring

A case study in a credit risk data science company.

Sebastião Cardoso Fernandes

Work Project

Supervisor: Professor Afonso Eça

Master's in Finance

January 2018

Dedicated to all my family, colleagues and professors
who have contributed to my progress.

Abstract

The Global Financial Crisis triggered a severe hold on credit lending due to the financial institutions' inability to assess credit applicants risk levels properly. Based on U.S. data from Lending Club, we conducted a study to evaluate the consequences of including macroeconomic risk factors in individual credit application observations. Through historical scenario stress testing, we find that this approach results in an increase in performance for credit scoring models developed in a stable economic cycle and applied to a recession. The inclusion of macroeconomic indicators reveals potential for credit institutions to better absorb shocks derived from economic downturns.

Keywords: Credit Scoring, Classification, Consumer Loans, Financial Stability, Stress Test, Macroeconomic Scenarios

Contents

1. INTRODUCTION 1
1.1 MOTIVATION 1
1.2 OBJECTIVES 3

2. LITERATURE REVIEW 4
2.1 MACROECONOMIC RISK FACTORS 4
2.2 EXTREME SCENARIOS AND STRESS TESTING 5

3. METHODOLOGY 6
3.1 MODELS 6
3.2 DEFINING THE DEVELOPMENT, VALIDATION AND TEST SETS 8
3.3 MODEL ASSESSMENT 8

4. DATA AND EMPIRICAL RESULTS 10
4.1 DATA DESCRIPTION 10
4.2 SETTING THE DEFAULT FEATURE 12
4.3 PERFORMANCE METRICS 13

5. CONCLUSIONS 15
5.1 RESULTS-BASED INFERENCE 15
5.2 LIMITATIONS 16

6. IMPACT IN THE BUSINESS WORLD AND FUTURE RESEARCH 17
6.1 CONSEQUENCES IN THE CREDIT SCORING MARKET 17
6.2 FUTURE WORK PROPOSITION 17

REFERENCES 18

APPENDIX 20

1. Introduction

1.1 Motivation

According to the U.S. Federal Reserve, consumer loans held by banks in late 2017 was $1,415bn, whereas commercial and industrial loans amounted to $2,125bn[1]. The retail credit business is therefore considered economically significant as it represents around 7.5% of the U.S. gross domestic product. Moreover, despite in 2016 the average delinquency rate on consumer loans being 2.16%, in 2009 it reached as high as 4.85%[2].

In terms of credit risk management, the goal of banks is to achieve a rather stable credit portfolio, whilst controlling for an acceptable credit risk level. Adjusting for this volatility is not only the duty of chief risk officers but also one of the greatest concerns of investors in every credit related security.

Considering the importance of the above stated features on the consumer credit market, it is crucial for financial institutions to have the appropriate tools for lending decisions. Being such a core activity for the bank's operations, the process and supervision of providing credit relies heavily on credit scoring models.

According to Louzada et al. [2016] credit scoring is "a numerical expression based on a level analysis of customer credit worthiness, a helpful tool for assessment and prevention of default risk, an important method in credit risk evaluation, and an active research area in financial risk management".

Credit scoring's most conventional form in today's financial markets can be described as a model whose goal is to categorize loan applicants according to their probability of default on credit payments. Thus, the result of this analysis is a binary classification where each applicant can be classified as credit worthy (good) or as someone who is very likely to default on its payments (bad). Although the output of these models is a probability of default, such result is then transformed into a binary variable according to a predefined threshold established by financial institutions to accommodate a specific risk level. Finally, as models are built with a broad variety of features, one of their roles in risk management is the identification of which features contribute positively or negatively to the increase of applicants' default risk.

There are, however, two types of credit scoring models worth distinguishing: application models which are designed to obtain a lending decision whenever a consumer applies for credit; and behavioural models that predict the delinquency rate of consumers on their current loans or of credit portfolios as a whole. The first carries severe impact since it rules the

[1] Data from the Federal Reserve as of November 1st (https://www.federalreserve.gov/releases/h8/current/)

[2] Data from the Federal Reserve Bank of St. Louis (https://fred.stlouisfed.org/series/DRCLACBS)

provision of credit for recent customers, whereas the latter is of extreme importance for financial institutions since it is a crucial input for capital requirements calculations according to the Basel III banking regulation.

The first reference to credit scoring analysis was made by Durand [1942] where he analysed a dataset containing good and bad loans and computed a credit rating formula, based on risk factors. 75 years later, research on such field has evolved exponentially and the development of state-of-the-art classifiers comprise machine learning algorithms and further enhanced technologies.

The systematic review performed by Louzada et al. covers the main classification methods in credit scoring, describing their development and usefulness for deployment within credit institutions. These methods can be neural networks, support vector machine, linear regression, decision trees, logistic regression, fuzzy logic, genetic programming, discriminant analysis, Bayesian networks, hybrid methods and ensemble methods. Although there exist various classification methods to build credit scoring models, for the purpose of this study, we will focus on logistic regressions and tree ensemble methods, given their explainability and wide adoption throughout banks and credit institutions.

Albeit much research, credit scoring literature has revealed itself to be insufficient to reflect recent advancements in state-of-the-art predictive learning models [Lessman et al., 2015]. More specifically, after the 2007 financial crisis, there is a gap in literature review in modifying models to calibrate for point-in-time probabilities of default. Thus, the quest for understanding financial stability has become the foundation of modern macroeconomic policy, especially after the recent global financial crisis, [Ali and Daly, 2010].

The 2007 financial crisis began with the subprime crisis in the mortgage market in the U.S. and was followed by a contagion to the global financial markets given the high exposure that financial institutions had in mortgage related securities. It is characterized by the worst financial crisis since the Great Depression and its consequences were a stagnated world economy and liquidity constraints across the majority of the financial markets.

To prevent further crisis like the one from 2007, financial regulatory agencies started implementing the Basel III accord in the world's major economies. This accord represents a regulatory framework comprising three major principles: capital requirements, leverage ratio and liquidity requirements.

To be compliant with the Basel III capital requirements, banks are obliged to perform a variety of analysis such as probabilities of default (PD), exposure at default (EAD) and loss given default (LGD). The execution of the previous analysis is not only required for credit granting but especially for risk management purposes as well.

The impact of such regulation on the financial institutions balance sheet has been immense, and the freedom to perform internal estimates to determine capital requirements is confined to those who meet the minimum conditions and receive supervisory approval to use the internal ratings-based (IRB) approach[3].

One problem arising from such tight control implemented by regulatory agencies is that it takes a long time for banks to deploy a credit scoring model in production. When considering that each day, banks could drop the oldest observations from the dataset used to build a model, and include the latest observations, we find that such procedure is not feasible given the regulatory agencies mechanisms. As so, in order to accommodate for newer information,

[3] Information from the Bank for International Settlements (https://www.bis.org/publ/bcbs128b.pdf)

credit scoring models need the ability to interpret the inputs and flag them as a signal of stability or not.

Some body of research advocates for the inclusion of macroeconomic variables to increase the predictability power of credit scoring models. As Malik and Thomas [2010] mentioned, PD's prediction at an individual level presents the limitation that it uses a snapshot of applicants who joined during certain time and does not allow macroeconomic changes to be included in the model.

Given the reliability of financial institutions and their operations on credit scoring models, having an accurate, robust and well calibrated model is fundamental to the financial health of these institutions. Our motivation arises then, as an attempt to further develop accuracy of credit scoring models for extreme scenarios, such as the 2007 financial crisis, and assess the applicability of models across several macroeconomic scenarios.

1.2 Objectives

The ultimate goal of this study is to increase the accuracy of credit scoring models developed in stable macroeconomic scenarios and tested in extreme scenarios such as one of economic recession. We aim to do so through the inclusion of macroeconomic risk factors to individual observations to provide a point-in-time calibration and the ability for the model to distinguish the macroeconomic scenario in which it is scoring.

Consequently, this paper contributes to the credit scoring literature as follows. It demonstrates the potential of including macroeconomic risk factors in consumer credit scoring models through a historical scenario stress testing methodology. We focus on improving the discriminatory power of models developed in relatively stable macroeconomic time spans and assess their performance for crisis scenarios. More specifically, we evaluate the business cycle adaptability of credit scoring models by allowing them to accommodate macroeconomic sensibility based on a point-in-time probability of default. Overall, the validation of these results is carried out through a misclassification cost analysis with statistical metrics such as the Gini coefficient and the ROC-AUC.

To briefly present our key results, the efficiency gains derived from the inclusion of macroeconomic risk factors is in the order of 2% for models developed in a stable macroeconomic scenario and tested in an extreme scenario such as the 2007 financial crisis. This discriminatory power improvement is given to the inclusion of specific features to each applicant individual features, such as GDP growth, unemployment rate and industrial production.

The paper is structured in the following manner. Chapter 2 features a literature review in macroeconomic variables and scenarios. Chapter 3 describes our methodology to achieve the desired goals and also proposes a valid methodology for further testing the results on future available information. Chapter 4 presents the dataset used in this study and the empirical results drawn out of the experiment. Chapter 5 concludes on the analysis performed throughout this study. Finally, Chapter 6 evaluates the impacts for the business world and proposes further research to complement the work developed herein.

2. Literature Review

2.1 Macroeconomic risk factors

There is a considerable body of research linking default risk with macroeconomic risk factors. Ali and Daly [2010] observe that with the beginning of globalisation, the concept of risk at both micro and macro levels changed completely, sparking the need for more standardized and innovative risk management tools. Their study provides a framework of a macroeconomic credit model to perform scenario analysis between two disparate economies, U.S and Australia. Their findings suggest that GDP is highly significant in explaining default risk and that when compared to Australia, the U.S. are much more sensitive to macroeconomic shocks. Fei et al. [2012] mention that credit risk measures are more realistic when derived from point-in-time methodologies that incorporate business cycles then through-the-cycle models that ease relevant economic fluctuations.

Through the use of Cox intensity models with macroeconomic risk factor and corporate specific rating features, Figlewski et al. [2012] verify that the inclusion of unemployment, inflation, real GDP growth and industrial production growth in the model leads to a statistically significant increase in explanatory power. Motivated by the lack of credit risk models, [Malik and Thomas, 2010] also incorporated consumer-specific ratings and macroeconomic factors in the framework of Cox Proportional Hazard models, revealing that default intensities of consumers are significantly influenced by macroeconomic factors and the inclusion of time of origination.

Even in more classical credit scoring literature, several attempts have been made to accommodate cyclicality. The use of rating transition matrices paired with a subdivision of economic regimes (normal, peak, through), according to GDP growth, revealed the strong dependence of default probabilities on the stage of the business cycle [Nickell et al., 2000]. Additionally, unemployment rates were also proved to significantly influence delinquency in the credit card market [Agarwal and Liu, 2003]. This is actually a very intuitive finding since as a result of negative macro shocks, when unemployment increases, people tend to support their loss of income by increasing credit card debt, which leads them to become delinquent on their monthly payments in case they do not find a job in the next few months.

According to the variety of models addressed in the Basel III accord requirements, we expanded the scope of our literature review to understand the impact of macroeconomic variables in models other than credit scoring ones. Bellotti and Crook [2012] work with Ordinary Least Squares models to incorporate macroeconomic variables to forecast LGD. This

approach was in line with Basel II requirements for LGD models to be able to forecast accurately in downturn conditions and enabling stress testing. Their findings were that interest rates and the unemployment level affected significantly the LGD forecast.

Finally, when considering the impact of doubtful and non-performing loans in banks, Mileris [2012] confirmed that macroeconomic changes in a country impacts significantly such loans. From this study, he concluded that, in association with GDP, macroeconomic risk factors such as inflation, interest rates, money supply, industrial production index and others influence directly the credit risk of debtors.

2.2 Extreme scenarios and stress testing

Considering our approach to evaluate the performance of credit scoring models, developed in stable macroeconomic scenarios, in extreme scenarios, we must lay down what defines an extreme scenario in the first place. Since the motivation of our work arises from extreme scenarios such as the global financial crisis, we resorted to the Business Cycle Dating Committee of the National Bureau of Economic Research (NBER) which provides a historical overview of business cycles turning points since 1978. According to NBER[4], "a recession is a significant decline in economic activity spread across the economy, lasting more than a few months, normally visible in real GDP, real income, employment, industrial production, and wholesale-retail sales". This definition corresponds perfectly to the motivation of this study regarding what an extreme scenario should be.

Having defined what an extreme scenario means, we then turn to the right methodology to apply in order to test our assumption. Stress testing encompasses several methodologies as it can be seen in Figure 1. It can initially be divided into two categories, sensitivity tests and scenario tests. Given that we want to test a specific scenario for the discriminatory power of our model, we are left with two possible approaches, historical scenarios or hypothetical scenarios. Considering that a hypothetical scenario would not only be difficult to determine, but also an attempt to preview the path of the economy in study, such approach would get the mandate of this analysis out of scope. Thus, the best fit for our purpose relies on stress testing with historical scenarios. Hence, historical scenario analysis is the methodology used to assess the impact of historical conditions in today's current models.

Stress Tests			
Sensitivity Tests		Scenario Tests	
Single Factor	Multi Factor	Historical Scenarios	Hypothetical Scenarios

Figure 1: Stress Testing Methodologies

[4] Information from the NBER (http://www.nber.org/cycles/jan2003.html)

3. Methodology

3.1 Models

From the broad range of types of credit scoring models, we have concluded that, when deciding which ones to use, controlling for explainability and benchmarking was fundamental. When considering explainability, models with linear parameters are easier to explain and understand since one can quickly grasp the impact of each parameter in the final result by assessing its coefficient. On the other hand, recent developments in machine learning algorithms and further integration with credit scoring has proved that ensemble methods outperform linear classifiers regularly [Lessman et al., 2015].

Considering the previous arguments, in this study we will use two types of models, logistic regression to perform the study and gradient boosting to benchmark the results obtained in the logistic regression. The logistic regression or the logit model consists in the estimation of a linear combination of $x = \{x_1, \dots, x_n\}$, the explanatory variables, and the logarithm of $y = \{y_1, y_2\}$, the output variable. Hence, considering y_1 as the target to predict, the model can be represented as:

$$\log\left(\frac{\pi}{1-\pi}\right) = x\theta\,,$$

where $\pi = P(Y = y_1)$ and θ is the vector representing the model factors. Since we are trying to predict how likely credit applicants are to become delinquent, the output should be a probability, thus, alternatively, the logit model can be represented as:

$$\pi_i = \frac{\exp\{x_i\theta\}}{1-\exp\{x_i\theta\}}\,,$$

where π_i is the probability of the ith applicant belong to the target y_1.

Ensemble methods are classifiers that build linear estimates over several other classifiers. In this specific case, the gradient boosting is a classifier which fits decision tree models by iteratively fitting sub-models to the residuals. Thus, the gradient boosting starts by fitting a rather simple model to the data:

$$F_1(x) = \hat{y}$$

This model can be characterized as a weak learner, which means that it performs slightly better than randomly picking the class. Here, the residuals can be represented by:

$$h_1(x) = y - F_1(x)$$

As a weak learner, the idea would be to modify it so that it can perform better. Hence, the gradient boosting creates a new model that integrates the residuals estimation with the aim to reduce overall error:

$$F_2(x) = F_1(x) + \hat{h}_1(x)$$

This process, however, is then repeated several times through an iteration that will keep improving the classifier:

$$F(x) = F_1(x) \rightarrow F_2(x) = F_1(x) + \hat{h}_1(x) \rightarrow \cdots \rightarrow F_m(x) = F_{m-1}(x) + \hat{h}_{m-1}(x)$$

It means that, through this optimization process, the gradient boosting model allows us to achieve a very accurate classifier that derives from a learning process of a really weak classifier.

To build the two types of models that we described above, we laid out the following mechanism. Each type of model will be built with and without the macroeconomic variables, to assess the impact of including this type of variable. Since gradient boosting models will serve the purpose of benchmarking, only two models will be built, where the only difference will be the inclusion of the macroeconomic risk factors. These models will, however, include all the variables available from the dataset since this algorithm has the ability to disregard statistically insignificant variables.

For the logit models, we will previously select features with two distinct methods that follow a statistical hypothesis approach. Firstly, we will build models where the features will be selected according to their importance using the scikit-learn[5], a machine learning library for scientific computation, module of feature selection. This module uses mean decrease impurity, which computes how much each feature decreases the weighted impurity in a random forest classifier by implementing a process of recursive feature elimination. Secondly, since this method is relatively biased towards variables with more categories, we will also select the included features through the computation of the Gini coefficient for individual variables. Hence, we will end up this analysis with four logit models, from which two use feature selection and two use individual Gini coefficients.

	No Feature Selection	Feature Selection	Individual Gini Coefficient
without Macroeconomic Variables	Gradient Boosting Model	Logit Model	Logit Model
with Macroeconomic Variables	Gradient Boosting Model	Logit Model	Logit Model

Table 1: Model Building Mechanism Output

[5] Documentation from scikit-learn (http://scikit-learn.org/stable/)

3.2 Defining the development, validation and test sets

As a result of our choice to perform an historical scenario stress testing approach, we need to define a methodology to test our assumption and assess the results. The main idea of our approach is to develop and validate the models in a stable macroeconomic scenario following an out-of-sample approach to make the partition of the data set. Afterwards, we will test this model in an extreme scenario, corresponding to a recession period as NBER defined, following an out-of-time approach. Figure 2 illustrates the mechanism of our methodology.

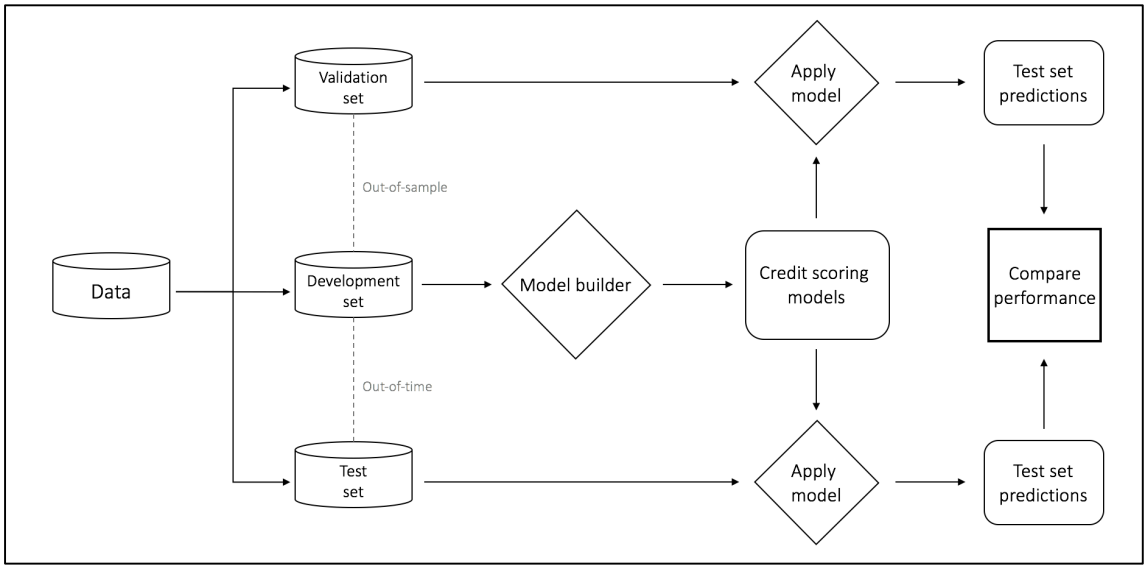

Figure 2: Methodology Mechanism

The descripted mechanism will be applied to both the data set that includes the macroeconomic risk factors and the one that does not. Such procedure, will not only enable us to take conclusions from the performance of the model in both scenarios, but also understand the impact of the inclusion of macroeconomic features in it.

3.3 Model assessment

In order to assess the predictive accuracy of a model, the standard metric among credit institutions is the Gini coefficient. This measure, however, is directly correlated to the receiver operating characteristic area under the curve (ROC-AUC) in the following manner:

$$Gini\ coefficient \approx 2 \times AUC - 1$$

Hence, it would be fair to assume that although the Gini coefficient is the standardized metric that we will be looking for, the basis of our assessment will be the ROC-AUC. The ROC-AUC, nevertheless, is the whole area beneath the ROC Curve. Its plot demonstrates the ability of a binary classifier throughout the variation of a discrimination threshold. Thus, the closer to the top left corner the curve is, the better the performance of the model. For further clarification Figure 3 illustrates an example of a ROC Curve.

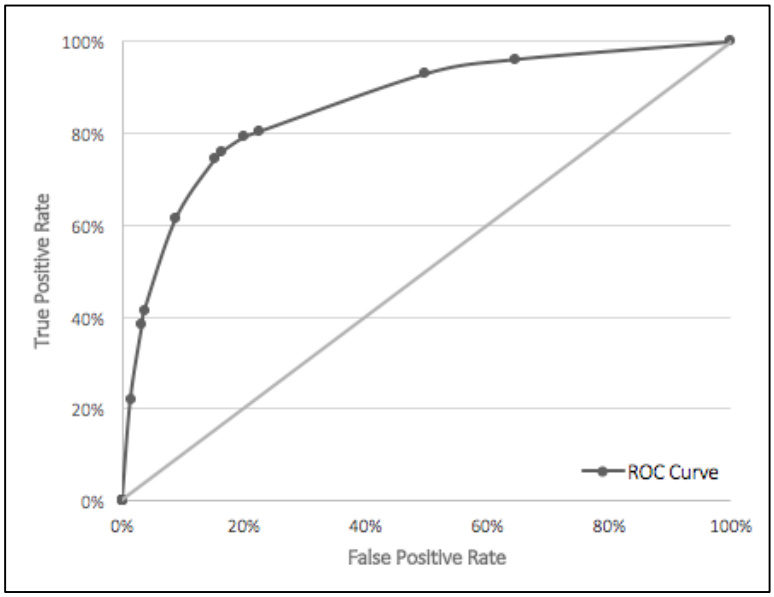

Figure 3: ROC Curve Example

In order to plot the ROC curve, one needs to first compute the true positive rate (TPR) and the false positive rate (FPR) for different thresholds of acceptance. The true positive rate indicates how much of the population is predicted to be delinquent correctly. On the other hand, the false positive rate, indicates how many of the applicants predicted to be delinquent on their loans were indeed good applicants. For the same model, the variation of these rates is only dependant on each credit institution's threshold. For instance, if the chief risk officer decides that the model's threshold is 30%, any individual scored with a higher probability of default will not be accepted, whereas the ones with probability of default lower than 30% will be conceded the loan. For the particular threshold of 30% there will be a specific TPR and FPR, which will be the coordinates of one of the ROC curve points.

In the credit risk industry, the Gini is preferred to the ROC-AUC since it takes values from 0 to 1 or (0% to 100%), making it easier to understand when assessing the discriminatory power of a model. For this study in particular, the Gini will be the metric taken into consideration since we can grasp the impact of our methodology with units that are coherent with the industry practices.

4. Data and Empirical Results

4.1 Data description

Our original data set contains information about 1,516,501 matured U.S. credit loans over the 9,5-year period, 01/06/2007 to 01/03/2017, from the Lending Club database. Lending Club[6] is a peer-to-peer lending company which makes his data available to the general public. Since the original sample had significant redundant data and missing information, we proceeded to some data cleaning. After doing so, we were left with 720,468 observations and 32 features, from which 23 are numerical, and 9 categorical. Figure 4 illustrates the distribution of loans issued and matured by the Lending Club throughout the sample period. We can notice an exponential increase of loans conceded between 2012 and 2014. The reason for the small number of loans between 2007 and 2011 is due to the fact that the Lending Club initiated its operations in 2007. Hence, this period relates to the initial activity of the peer-to-peer company, meaning that operations and reliability of the company were still gaining market confidence.

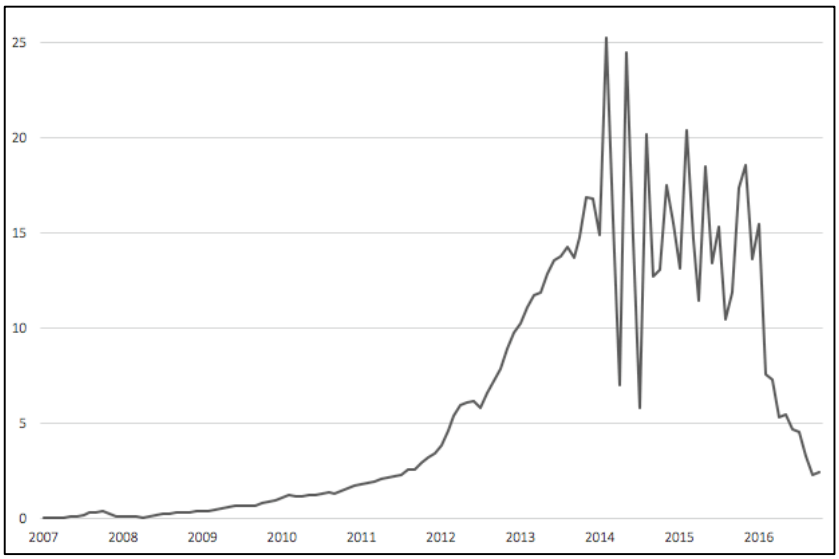

Figure 4: Evolution of Loans Issued and Matured in thousands ($)

[6] Data from the Lending Club (https://www.lendingclub.com/)

When considering the data set in which we will develop the credit scoring models, it is important to assess the stability of the populations from the different partitions that we will make to test our hypothesis. Hence, the distribution of observations is a limitation that we shall consider in our inferences. Given that Lending Club started its operations in 2007, the subsequent years maintained a low activity of granting loans. This translates in 4,386 completed loans from 2008 until mid 2009, 90,063 from mid 2009 until 2012 and 625,416 from 2013 until the first quarter of 2017. This evolution represents a limitation to our study since when comparing to the current activity of the credit institution, the values that resemble to the global financial crisis period are not much representative.

Each observation presents regularly the 32 features above mentioned, which are characteristics from the individuals to whom it was conceded credit. However, when some feature is not represented for a specific observation, for modelling purposes, we replaced the missing value by the median of the feature. A full description of each feature can be consulted in the data dictionary presented at Table A.1 of the Appendix. Additionally, it should be noted that another relevant limitation of this study was that due to their absence in the pre-2011 period, a lot of insightful features were dropped from the initial available features.

In the benchmarking state-of-the-art research update, Lessman et al. [2015] referenced that a common gap within credit scoring literature lies in the use of few or small data sets. However, as Louzada et al. [2016] correctly mentioned, as much as current information is widely available, given the modernization of the internet and the establishment of large data centres, availability of credit data is constrained given the confidential information of the applicants and the regulation that applies to it. Adding to this, the priority of our study to use a data set that covered the global financial crisis period further narrowed our alternatives. One particular aspect of this data set that should be taken into consideration is the large number of observations, a characteristic that contributes favourably for the stability of the models being built. Finally, it is very important to consider that credit data sets only have information available from the individuals to whom they have provided loans, thus, rejection inference plays a very important role in the credit risk landscape. Fortunately, through a rare sample that includes rejected applicants, Crook and Banasik [2004] were able to prove that rejection inference tends to leave regression coefficients unchanged, making this study valid to conclude from.

Since our study is based on the premise of adding macroeconomic risk factors to the original data set, we resorted to a credited U.S. database[7] to obtain our pre-selected features according to the literature review: quarterly real GDP growth rate, monthly civilian unemployment rate and monthly industrial production index. When adding these features to the original data set, one should bear in mind that the disposal of such figures happens at pre-defined dates after the reference period, which means that we would need to lag our features according to their release dates. To adjust for this aspect, the GDP growth[8] was lagged 6 months, whereas the unemployment rate[9] and industrial production[10] were only lagged by 2 months. Figure 5 illustrates the evolution of the macroeconomic risk factors used along the period of study.

One can easily observe how the variation of each macroeconomic indicator around the global financial crisis period clarifies the significant decline in economic activity as NBER has

[7] Data from the Federal Reserve Bank of St. Louis (A191RL1Q225SBEA; UNRATE and INDPRO respectively)

[8] Data from the Bureau of Economic Analysis (https://www.bea.gov/newsreleases/national/gdp/gdpnewsrelease.htm)

[9] Data from the Bureau of Labor Statistics (https://www.bls.gov/ces/ces_tabl.htm)

[10] Data from the Federal Reserve (https://www.federalreserve.gov/feeds/g17.html)

defined it. This observation allows us to elucidate the impact that these variables will have on positioning the model throughout the economic cycle.

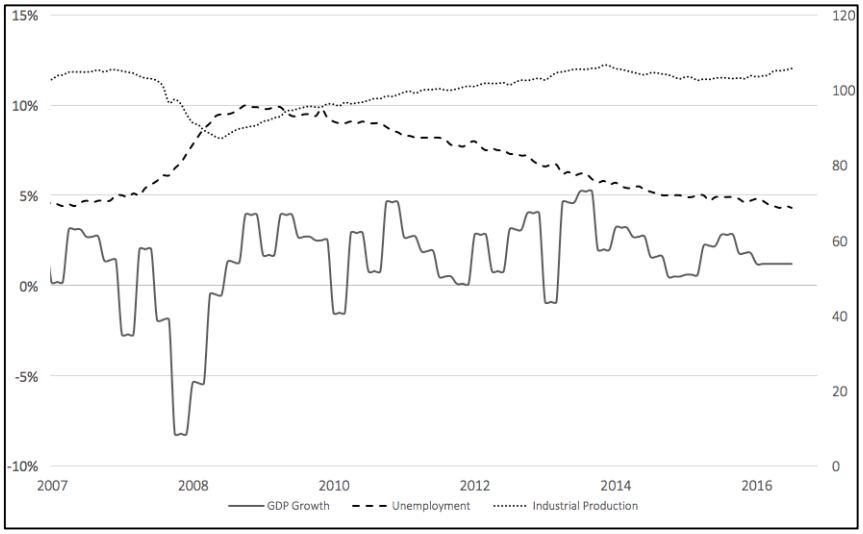

Figure 5: Macroeconomic Risk Factors

4.2 Setting the default feature

In order to build a binary classifier, it is crucial that the target feature, in this case an indicator of default on the loan, assumes one of two values. Thus, in order to guarantee delinquency predictability, we had to transform the loan status feature into a binary one and remove the observations with an inconclusive status. Table 2 illustrates the different values that loan status could take.

Non-default	*Default*	*Remove*
- Fully paid - Failed credit policy. Status: Fully paid	- Charged off - Failed credit policy. Status: Charged off - Default	- Late (16-30 days) - Late (31-120 days) - In Grace Period - Current

Table 2: Division of Loan Status Values into Default Feature

Our approach to define what would mean a default in this dataset came from a critical interpretation of Lending Club's data dictionary. We understood that both the categories "Fully paid" and "Failed credit policy. Status: Fully paid" meant that the loan has been paid, whilst the categories "Charged off", "Failed credit policy. Status: Charged off" and "Default" meant that the creditors did not meet their contractual obligations and have become delinquent on their loans. Finally, since the other categories did not allow us to correctly assess the situation of the credit, either because the loan was not concluded or indication of default was yet to be concluded, we have decided to remove those observations from our sample.

As consequence of our default engineering, our data set counts with 147,938 defaults, which translates in an overall default rate of 20.5% throughout mid-2007 until the first quarter of 2017. Figure 6 illustrates the evolution of default rate per loan term.

A conclusion that can be drawn from the observation of the graph is that although 60 month loans have started to be given in 2009 only, this category of loans carries much more risk for the population in study, thus increasing the probability of default.

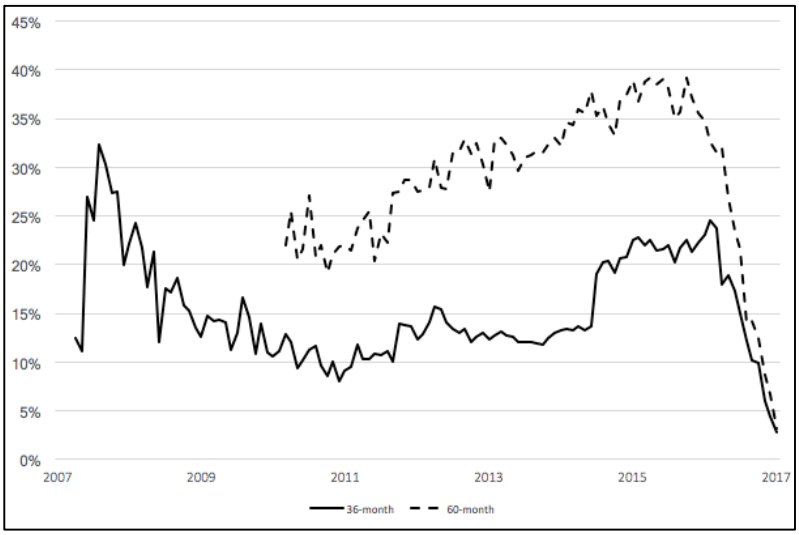

Figure 6: Average Monthly Default Rate

4.3 Performance metrics

In order to build different logistic models, we have conducted the feature selection and individual Gini analysis separately. The selection resulting from this process can be consulted in Table A.2 of the Appendix. An important result from this pre-modelling procedure is the importance that both analysis attribute to each macroeconomic risk factor. According to the feature selection analysis, the industrial production is the macroeconomic indicator with more impact in the model, whilst through the individual Gini analysis, the unemployment rate is the one with more importance. Nevertheless, it is critical to note that when ranking importance or explainability according to each analysis' scope, the three macroeconomic indicators included belong to the top 50% of all features.

To conduct our analysis, we have run several models, each one with 50 iterations and a three-fold cross validation for the development set, which contains 625,416 observations. The latter consists in an 80% out-of-sample partition of the observations from 2013 until the first quarter of 2017, a period that resembles macroeconomic stability in the U.S. Each model was then used to score the observations from the left 20% from this time frame, the validation set, to obtain our chosen performance metric, the Gini coefficient. Finally, we have run each model in the test set, which consists on the out-of-time partition of our initial sample corresponding to the global financial crisis period, from 2008 until mid-2009, containing 4,386 observations. Table 3 illustrates the results obtained throughout our study, plus the differential in performance of the models from the inclusion of the macroeconomic risk factors.

		Gradient Boosting (Benchmark)	Logit Model (Feature Selection)	Logit Model (Univariate Gini)
Raw Models	Validation set	45.14%	41.29%	41.52%
	Test set	25.91%	23.57%	25.76%
Macroeconomic Models	Validation set	47.61%	42.68%	43.11%
	Test set	28.10%	25.05%	27.96%
Differential (Macro - Raw)	*Validation set*	*+ 2.47%*	*+ 1.39%*	*+ 1.59%*
	Test set	*+ 2.19%*	*+ 1.48%*	*+ 2.20%*

Table 3: Gini Results and Comparison from Different Credit Scoring Models (Validation Set [2013-2017Q1] & Test Set [2008-2009Q2])

As previously noted, the gradient boosting is the type of credit scoring model that presents higher performance in the experiment. As such, it was important to include the ensemble method in our study to benchmark the results of the logistic models. For the validation set, the gradient boosting benchmark presents a Gini coefficient of 45.14% without macroeconomic risk factors, whilst when including them, it increases its Gini coefficient by 2.47%. On the other hand, for the test set, the period related to the global financial crisis, this model presents a Gini of 25.91% without macroeconomic risk factors and 28.1% with the inclusion of them. The increase in performance in both the validation and test sets, indicates that our approach not only increases the predictability power for eventual extreme scenarios but also increases the accuracy of a model built in its own current economic cycle. Although lower, the results for the approaches with the logistic model present a similar impact on the inclusion of macroeconomic indicators. The model built with the feature selection variables increases its performance in 1.39% for the validation set and 1.48% for the test set, whilst the model built with the univariate Gini analysis increases 1.59% and 2.20% respectively.

To validate our results and further develop on our conclusions, we have decided to apply the same methodology to the mild economic period that goes from the third quarter of 2009 until the end of 2012. Following such, we have then developed and validated the model in the indicated period and tested in the same extreme scenario of the global financial crisis. The results, which can be consulted in Table 4, follow the same conclusions from the initial analysis.

		Gradient Boosting (Benchmark)	Logit Model (Feature Selection)	Logit Model (Univariate Gini)
Raw Models	Validation set	39.20%	36.58%	38.03%
	Test set	30.18%	24.76%	29.02%
Macroeconomic Models	Validation set	39.50%	36.62%	38.19%
	Test set	31.83%	27.28%	30.77%
Differential (Macro - Raw)	*Validation set*	*+ 0.30%*	*+ 0.04%*	*+ 0.16%*
	Test set	*+ 1.65%*	*+ 2.52%*	*+ 1.75%*

Table 4: Gini Results and Comparison from Different Credit Scoring Models (Validation Set [2009Q3-2012] & Test Set [2008-2009Q2])

The performance metric increases for all models with the inclusion of the macroeconomic indicators, with the small caveat that for the validation sets, the increase is not as significant. Nonetheless, the increase in performance for the test set is the result we are concerned with in this study. Hence, this second analysis seems to reinforce our conclusion that by including macroeconomic indicators in the models, we observe an increase in the discriminatory power throughout economic cycles that resemble a recession.

5. Conclusions

5.1 Results-based inference

Our methodology, although a bit counter-intuitive, resorts to the application of a credit score model, developed in the time frame of 2013 until early 2017, to the time span of 2008 until mid-2009. One relevant aspect of our methodology, that might create a gap from a real-life situation, is the backwards application of a model. However, since availability of databases for credit portfolios is limited due to regulatory and confidentiality purposes, the ability to test such approach was restricted to the proposed methodology. Finally, the ambition to verify our hypothesis in the modern financial markets environment, left us with the global financial crisis as the one and most relevant extreme scenario to test upon.

The results obtained throughout our study are consistent and the conclusions to be drawn from it are threefold:

1) The increase in performance resulting from the inclusion of macroeconomic risk factors shows evidence that this approach not only perfections probability of default models to incorporate the economic cycle in which they are scoring but also improves the classification of applicants according to their risk level.
2) By increasing the validation set performance in the period of 2013 until early 2017, there is also evidence that for models built in their current economic cycle, few variations of the macroeconomic indicators can increase the sensitivity of the model to better score credit applicants.
3) Finally, among the three included macroeconomic indicators, the civilian unemployment rate and the industrial production index present more statistical significance than the GDP growth, as proven by the feature selection approach that applies recursive feature elimination, and by the individual Gini analysis. Nonetheless, although the three variables seem to carry explanatory power, it is important to infer about those who are more relevant in explaining the risk level of credit applicants.

As shown by the literature review, there is a considerable research panel advocating for the inclusion of macroeconomic risk factors in credit risk models in general. Since the major research coincides with credit portfolios rather than individual applications, we decided to test our hypothesis and fill the literature gap previously identified. Moreover, the decision to test this approach on an extreme scenario follows from the recent regulatory restrictions that

credit institutions have been facing and the willingness to further develop models that impact capital requirements. Ignoring economic cycles has a great impact on credit institutions balance sheets and the overwhelming process that goes from developing a model until having it approved and ready for deployment varies according to the specific business of the credit lender. For instance, whilst credit unions have more flexibility since they are not as tightly regulated as banks, their ability to deploy new models is much more flexible. On the other hand, banks face a lengthy process of approval and this type of approach aims to prevent banks from having obsolete credit scoring models in between the transitions of economic cycles.

An alarming result from our analysis is the huge decrease in performance when applying a credit scoring model, developed in a stable macroeconomic scenario, to a recession scenario. The best validation set performance presents a Gini coefficient of 47.61%, whilst by applying the same model to the period between 2008 and mid-2009, we only obtain a result of 28.10%. Still, although facing a huge decrease in explanatory power when transitioning to a recession, including macroeconomic indicators might smooth the burden of banks on either deploying updated models or being prepared to a sudden economic turn.

5.2 Limitations

There are some limitations that we have identified along our study worth developing on. The data set used has revealed some singularities that might impact the results obtained. Since Lending Club started its activity near the period of the global financial crisis, the majority of the variables available today, and the ones that are common to the credit risk market, were not available for the observation period in which we tested the model. Hence, the development of the model was made without these relevant variables since there would be no use to them under the described situation. Adding to this, our test set contains a small amount of observations when compared to the development set. Although there is no harm in testing a model under a restricted number of observations, in this case the fact that there were only so few observations for the testing period, might indicate a more cautious selection of the population to whom it was conceived credit. This might result in a through-the-door population slightly different than the one we have developed our models on. Such fact does not depend on the economic scenario under which the lending activity has been developed, but in the population's reach that the company initially had.

Other limitation that should be considered in this study, as Bellotti and Crook [2012] have noted, is the unavailability of other economic downturn or recession in our data set, such that we could include in the development of our model. Such feature would be of interest to consider, since we believe that an excellent credit scoring model should have training data across the entire economic cycle. We could however, have included part of the population from the global financial crisis period, but such approach would lead our analysis to a bias, in the sense that the specificity of the testing period would already be accounted for in our model. This would make our methodology invalid since we are isolating the extreme scenario as something that could occur in the future and not something that is already taking place.

6. Impact in the Business World and Future Research

6.1 Consequences in the credit scoring market

We expect our contribution to be valuable for the credit scoring market since it shows evidence of the benefits of including macroeconomic risk factors in the development of credit scoring models. It is possible to infer, that our analysis carries significant importance for the ongoing regulatory reforms. Through the adoption of more sophisticated credit scoring models that account for economic cycles, the financial system can improve resilience to economic shocks whilst solving for risk underwriting dilemma. This study aims to fill the gap in literature review on "How to Deal with Extreme Cases for Credit Risk Monitoring". It not only found a solution for credit scoring models' discriminatory power throughout different economic cycles, but it also achieved performance improvement for models built in their own economic cycles.

The findings of this work, should be further developed along with regulatory agencies to incorporate the benefits of our approach, but also along with credit institutions that seek to improve their current credit underwriting processes.

6.2 Future work proposition

The course of this study makes us advocate for further work on the inclusion of macroeconomic risk factors in credit scoring models. Firstly, further study to discover the impact of our approach to the financial institutions capital requirements is needed to assess the potential consequences that arise from it. Secondly, the uniqueness of our data set calls for additional testing of the used methodology among different data sets and recessions characterized by diverse economic conditions. Finally, although living in a rather globalized world, each country's reaction to a recession is different, and other macroeconomic indicators should be considered when assessing different populations' risk levels.

References

Agarwal, S., & Liu, C. (2003). Determinants of Credit Card Delinquency and Bankruptcy: Macroeconomic Factors. *Journal of Economics and Finance*, *27*(1), 75–84.

Ali, A., & Daly, K. (2010). Macroeconomic determinants of credit risk: Recent evidence from a cross country study. *International Review of Financial Analysis*, *19*(3), 165–171.

Banasik, J., Crook, J., & Thomas, L. (2003). Sample selection bias in credit scoring models. *Journal of the Operational Research Society*, *54*(8), 822–832.

Bellotti, T., & Crook, J. (2012). Loss given default models incorporating macroeconomic variables for credit cards. *International Journal of Forecasting*, *28*(1), 171–182.

Berge, T. J., & Jorda, O. (2011). Evaluating the classification of economic activity into recessions and expansions. *American Economic Journal: Macroeconomics*, *3*(2), 246–277.

Crook, J., & Banasik, J. (2004). Does reject inference really improve the performance of application scoring models? *Journal of Banking and Finance*, *28*(4), 857–874.

Desai, V. S., Crook, J. N., & Overstreet, G. A. (1996). A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research*, *95*(1), 24–37.

Durand, D. (1942). Risk Elements in Consumer Instalment Financing. *Journal of Marketing*, *6*(4), 407–408.

Fei, F., Fuertes, A.-M., & Kalotychou, E. (2012). Credit Rating Migration Risk and Business Cycles: CREDIT RATING MIGRATION RISK AND BUSINESS CYCLES. *Journal of Business Finance & Accounting*, *39*, 229–263.

Figlewski, S., Frydman, H., & Liang, W. (2012). Modeling the effect of macroeconomic factors on corporate default and credit rating transitions. *International Review of Economics and Finance*, *21*(1), 87–105.

Good, J. (1944). Application of the Logistic Function to Bio-Assay. *Journal of the American Statistical Association*, *77*(378), 342–344.

Hand, D. J., & Henley, W. E. (1997). Statistical Classification Methods in Consumer Credit Scoring: a Review. *Royal Statistical Society*, 523–541.

Huang, C.-L., Chen, M.-C., & Wang, C.-J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*, *33*(4), 847–856.

Hume, M., & Sentance, A. (2009). Journal of International Money The global credit boom : Challenges for macroeconomics and policy. *Journal of International Money and Finance*, *28*(8), 1426–1461.

Kosow, H., & Gassner, R. (2008). *Methods of Future and Scenario Analysis*.

Lan, Y., Janssens, D., Chen, G., & Wets, G. (2006). Improving associative classification by incorporating novel interestingness measures. *Expert Systems with Applications*, *31*, 184–192.

Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, *247*(1), 124–136.

Louzada, F., Ara, A., & Fernandes, G. B. (2016). Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science*, *21*(2), 117–134.

Louzada, F., Ferreira-Silva, P. H., & Diniz, C. A. R. (2012). On the impact of disproportional samples in credit scoring models: An application to a Brazilian bank data. *Expert Systems with Applications*, *39*(9), 8071–8078.

Malik, M., & Thomas, L. C. (2010). Modelling credit risk of portfolio of consumer loans. *Journal of the Operational Research Society*, *61*(3), 411–420.

Mileris, R. (2012). Macroeconomic Determinants of Loan Portfolio Credit Risk in Banks. *Inzinerine Ekonomika-Engineering Economics*, *23*(5), 496–504.

Nickell, P., Perraudin, W., & Varotto, S. (2000). Stability of rating transitions. *Journal of Banking & Finance*, *24*(1–2), 203–227.

Shi, Y. (2010). Multiple criteria optimization-based data mining methods and applications: A systematic survey. *Knowledge and Information Systems*, *24*(3), 369–391.

Stepanova, M., & Thomas, L. C. (2002). Survival Analysis Methods for Personal Loan. *Operations Research*, *50*(2), 277–289.

Taylor, P., Härdle, W., Mammen, E., Müller, M., Hardle, W., & Muller, M. (2012). Testing Parametric versus Semiparametric Modeling in Generalized Linear Models. *Journal of American Statistical Association*, (February 2013), 37–41.

Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, *16*(2), 149–172.

Vanek, T. (2016). Economic Adjustment of Default Probabilities. *European Journal of Business Science and Technology*, *2*(2), 122–130.

West, D. (2000). Neural network credit scoring models. *Computers and Operations Research*, *27*(11–12), 1131–1152.

Xu, X., Zhou, C., & Wang, Z. (2009). Credit scoring algorithm based on link analysis ranking with support vector machine. *Expert Systems with Applications*, *36*(2 PART 2), 2625–2632.

Ziari, H. A., Leatham, D. J., & Ellinger, P. N. (1997). Development of statistical discriminant mathematical programming model via resampling estimation techniques. *American Journal of Agricultural Economics*, *79*(4), 1352–1362.

Appendix

Table A.1: Data dictionary of features included in the final sample

#	Feature	Description
1	acc_now_delinq	The number of accounts on which the borrower is now delinquent.
2	addr_state	The state provided by the borrower in the loan application
3	annual_inc	The self-reported annual income provided by the borrower during registration.
4	application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
5	chargeoff_within_12_mths	Number of charge-offs within 12 months
6	collections_12_mths_ex_med	Number of collections in 12 months excluding medical collections
7	default	Flag indicating if the borrower defaulted on his loan or successfully completed the payment plan.
8	delinq_2yrs	The number of 30+ days past-due delinquencies in the borrower's credit file for the past 2 years
9	delinq_amnt	The past-due amount owed for the accounts on which the borrower is now delinquent.
10	dti	The borrower's debt-to-income ratio, excluding mortgage and the LC loan.
11	emp_length	Employment length in years. Possible values are between 0 and 10.
12	grade	LC assigned loan grade
13	home_ownership	The home ownership status. Our values are: RENT, OWN, MORTGAGE, OTHER
14	initial_list_status	The initial listing status of the loan. Possible values are – W, F
15	inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
16	installment	The monthly payment owed by the borrower if the loan originates.
17	issue_d	The month which the loan was funded
18	loan_amnt	The listed amount of the loan applied for by the borrower.
19	mths_since_earliest_cr_line	Months since the borrower's earliest reported credit line was opened
20	mths_since_last_delinq	The number of months since the borrower's last delinquency.
21	open_acc	The number of open credit lines in the borrower's credit file.
22	pub_rec	Number of derogatory public records
23	pub_rec_bankruptcies	Number of public record bankruptcies
24	purpose	A category provided by the borrower for the loan request.
25	pymnt_plan	Indicates if a payment plan has been put in place for the loan
26	revol_bal	Total credit revolving balance
27	revol_util	Revolving line utilization rate.
28	tax_liens	Number of tax liens
29	term_in_mths	The number of payments on the loan. Values are in months and can be either 36 or 60.
30	title	The loan title provided by the borrower
31	total_acc	The total number of credit lines currently in the borrower's credit file
32	verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified

Table A.2: List of selected features according to the method applied

Feature	Feature Selection Metric	Individual Gini
dti	0.076	0.189
revol_bal	0.068	*Lower than threshold. Not included*
revol_util	0.067	0.103
mths_since_earliest_cr_line	0.067	0.058
annual_inc	0.065	0.120
instalment	0.065	0.070
total_acc	0.055	*Lower than threshold. Not included*
loan_amnt	0.051	0.089
grade	0.051	0.361
addr_state	0.049	*Lower than threshold. Not included*
open_acc	0.046	*Lower than threshold. Not included*
mths_since_last_delinq	0.042	*Lower than threshold. Not included*
ind_pro*	0.041	0.056
unem_rate*	0.036	0.087
emp_length	0.032	*Lower than threshold. Not included*
title	0.027	*Lower than threshold. Not included*
gdp_gro*	0.027	0.042
term_in_mths	0.021	0.185
verification_status	*Lower than threshold. Not included*	0.113
home_ownership	*Lower than threshold. Not included*	0.103
inq_last_6mths	*Lower than threshold. Not included*	0.076
purpose	*Lower than threshold. Not included*	0.041

**Macroeconomic Risk Factors*