ESSAYS ON THE FINTECH LENDING MARKET

by

Eris Azizaj

A dissertation submitted to the Department of Economics,

College of Liberal Arts and Social Sciences

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in Economics

University of Houston

May 2021

ACKNOWLEDGMENTS

I would like to express that I am truly indebted to my advisor, Prof. Bent Sorensen, who has helped and supported me throughout the Ph.D. program. Prof. Bent has taught me how to be an empirical researcher, have the critical "eye" in the research we do, and target perfectionism. He has made the overwhelming process of the Ph.D. more fun and interesting with his anecdotes and smart jokes. I can not imagine how difficult these years would have been without his presence and support. Also, I am grateful to Prof. Aimee Chin and Prof. David Papell for their valuable contribution to this dissertation throughout their feedback. I would like to sincerely thank Prof. Rauli Susmel for being a member of my dissertation committee.

I am grateful to other faculty members: Kei-Mu Yi, Fan Wang, German Cubas, and participants of the Ph.D. Macro workshop at the University of Houston for their helpful comments and suggestions. I would like to thank my classmates Xavier Bautista, Priyam Verma, and Hamza Zahid for their friendship and enjoyable conversations about research, ideas, and life in general. A special thanks to Amber Pozo, who has always been very friendly in answering all my questions on various administrative issues.

I am truly blessed to have had my family by my side, especially my wife and my mom, who have made lots of sacrifices throughout this journey to help and support me complete my studies. They have been the true source of love, hope, inspiration, and have given me strength in every step of this process. To my beloved sons, who are the true faces of happiness and joy.

to my family

ABSTRACT

This dissertation consists of two essays on the FinTech lending market. I explore default in the FinTech lending market using a dataset of both extensive credit and soft information for borrowers from the largest FinTech lender in the United States. In the first essay, I study the default risk in this market over the business cycle. I find that both macro and regional economic conditions play a role in consumer default and should be taken into consideration when assessing credit risk. I show that lenders operating in this market increasingly focus on subprime borrowers, whose default rates are more sensitive to macro and regional economic conditions than those of prime borrowers. Based on estimates from a duration model, I provide counterfactual analyses of what default rates and the associated total losses would look like in different economic scenarios. In the case of a recession, the losses would be 37 percent higher than in the case of an expansion. For the same volume of loans in the recession, doubling the subprime share would lead to an additional 6.2 percent increase in losses.

In the second essay, I provide an overview of some of the most common machine learning methods used in modeling default risk and assess to what extent these methods are better than traditional approaches. Using the same datasets as in the first essay, I explore the determinants of default in the FinTech lending market. I apply different machine learning algorithms to predict out-of-sample default. I find that some of the machine learning algorithms, such as extreme gradient boosting and artificial neural networks, marginally outperform logistic regression. Annual income, loan purpose, revolving line utilization, and interest rate are the most important variables predicting default. Macro and regional variables are listed among the top 10 variables explaining consumer default behavior.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

1 FinTech Lending Default Risk over the Business Cycle

1.1 Introduction

The consumer lending market in the United States comprises one of the largest shares of the total U.S. credit market, with outstanding credit of $4 trillion (Reserve, 2021). Traditionally, this market has been dominated by banks, which have served as "middlemen" between investors and borrowers. However, recent developments in technology have created other matching options between parties through online platforms; known as marketplace lending (MPL) or FinTech lending.[1] Utilizing advancements in data processing and credit assessment tools—such as machine learning (ML) and artificial intelligence (AI)—lenders operating in the FinTech lending market claim to offer loans much faster and cheaper than traditional channels.

This market has seen enormous growth since its inception—an average of 84 percent growth per quarter—and many experts in the field think these platforms may revolutionize the banking system. According to TransUnion, FinTech lenging market's share of personal loan originations has grown from 1 percent in 2010 to nearly 40 percent in 2018.[2] Attracted by this growth, many banks have started investing in FinTech loans through securitization channels and over time have become the main source of capital for this market. As of today, about 68 percent of the investor

[1] There are some other terms used for the same type of loans such as crowd funding or peer-to-peer (P2P) lending, which was used initially when the structure of the market was different.

[2] Personal loans are unsecured installment loans that can be used for any purpose. They are typically short term, small loans, with higher APRs than other types of installment loans. More about the distribution of personal loans by lender category is in the Appendix. Retrieved November 10, 2020 from https://www.americanbanker.com/news/once-dismissive-of-fintechs-traditional-lenders-now-feeling-their-bite

base in Lending Club—the largest online lending platform in the United States—are banks, and the rest is composed of mutual and hedge funds, and other financial institutions.[3] This way FinTech lenders have been able to attract institutional capital to the market, which is expected to reach $450 billion or more by 2022 (S&P Global Market Intelligence, 2018).[4]

Consumer default has been extensively studied in the banking and finance literature, especially after the Great Recession. This paper seeks to shed light on the relationship between consumer default in the FinTech lending market and the business cycle, while the existing literature mainly focuses on determinants of default as a function of individual and loan characteristics. I consider the aggregate business cycle as well as regional fluctuations, because my sample does not contain many aggregate business cycles, so the use of regional cycles allows for more precise estimation. Studying the relationship between default and the business cycle is important, because if FinTech lenders are sensitive to macro and regional shocks and this market keeps growing at the present rate, significant losses may amplify the business cycle or even have the potential to disrupt financial stability.

In a report by Committee on the Global Financial System (CGFS) and Financial Stability Board (FSB), New York Fed President William Dudley and Chair of FSB Klaas Knot write “A bigger share of FinTech-facilitated credit in the financial system could have both financial stability benefits and risks in the future..." CGFS-FSB (2017). When these lightly regulated lenders asked for access to the payment system, settlement services, and other tools, St. Louis Fed President James Bullard told Reuters: “I am concerned that FinTech will be the source of the next crisis.”[5]

I find that over time FinTech lenders increasingly focus on a relatively risky segment of borrowers. The share of total loans issued to subprime borrowers has increased from 6 percent in 2009

[3] Retrieved September 07, 2020, from https://www.lendingclub.com/investing/institutional/banks.

[4] S&P Global Market Intelligence estimates that the market size has reached about $150 billion by 2018:Q2. Another source projects this market to reach $1 trillion by 2025 (PricewaterhouseCoopers, 2015).

[5] Retrieved February 15, 2018 from https://www.cnbc.com/2019/01/14/reuters-america-fintech-firms-want-to-shake-up-banking-and-that-worries-the-fed.html

to 20 percent in 2015. This has led to relatively higher default rates in this market—an average default rate of 14.2 percent—compared to the default rates on loans issued by traditional lenders. The subprime mortgage market in the U.S. accounted for only 4 percent of the total mortgage market in 2006, but many research studies (e.g., Demyanyk and Van Hemert (2008); Mian and Sufi (2009); Brunnermeier (2009)) consider the high default rates in this market and the associated total losses the trigger for the financial crisis. "Because of amplification effects, even small trigger events can lead to major financial crises and recessions" (Brunnermeier and Oehmke, 2013).

One important aspect when studying default behavior is the composition of loans by default maturity because losses are larger for earlier defaults and visa versa. Another aspect is whether the sensitivity of default to regional and macroeconomic shocks is different for subprime and prime borrowers. I conduct counterfactual analyses of what default rates and associated total losses would look like for different compositions of borrowers, economic scenarios, and changes in market size. I simulate the economy for an expansion and downturn, and calculate how the average probability of default varies across maturities. Using the default probabilities from a duration model combined with the composition of loans in each maturity group, I predict the total losses. I perform the same analysis for a scenario where the market size is tripled and the share of subprime borrowers is twice as large.

I find that aggregate inflation and GDP, and the regional unemployment rate, and house price growth (at the 3-digit ZIP code level) play a significant role in consumer default. Results from the counterfactual analysis show that in the case of a recession the losses from defaults would be 37 percent higher than in the case of an expansion.

I show that subprime borrowers are more vulnerable to macro and regional shocks than prime borrowers. The typical subprime borrower has 50 to 70 percent higher default probabilities than the typical prime borrower. Moreover, doubling the share of subprime borrowers leads to an additional 6.2 percent increase in losses.

I complement the analysis by briefly studying how individual and loan information affects the probability of default. In addition to the hard information provided, I use text analysis to extract soft information from the borrowers' loan description provided in the process of application. Searching for a list of keywords or common phrases that borrowers use to make positive s about their credit-worthiness, I create other variables that are significant in predicting default.[6]

The rest of this paper is organized in the following order: Section 2 gives some background information on how FinTech lending works and briefly summarizes the findings from the current literature. Section 3 provides descriptive statistics for FinTech lending data, and macro and regional economic indicators used in this study. Methodology and empirical results are in Section 4. Section 5 discusses the implications of my findings (i.e., the counterfactual analyses) with changes in market size, borrower composition, and different economic scenarios. Summary and some concluding remarks are in Section 6. The appendix provides further details on the data, complementary results, as well as additional figures and tables.

1.2 Background and Related Literature

1.2.1 Background on the FinTech Lending Market

The concept of peer-to-peer lending is as old as borrowing and lending to friends and family members. Having access to credit this way was a solution for individuals with high social capital, but not for others. Therefore, people started forming certain groups/circles to lend or borrow from each other. The idea first began in underdeveloped countries with "lending circles" of entrepreneurs helping other low-income entrepreneurs have access to capital (Conlin, 1999). Seeing its substantial success, the idea spread quickly around the world. Initially in the United States, P2P lending was implemented through 250 micro-lending programs to encourage economic activity in low-income communities (Conlin, 1999). However, with the innovation of Web 2.0 technology in the

[6] I follow Carmichael (2014) approach for the text analysis.

late 90s, this concept changed and was taken to another level. People started using online sources to borrow and lend to each other rather than just within small groups. This new way of lending is part of the fast-growing financial technology (FinTech) industry which has seen enormous growth since its inception. From a nationally representative survey, Adams et al. (2017) find that only 25 percent of the U.S. consumers are aware of the FinTech lending market and only 11.7 percent of them have applied for a marketplace loan, suggesting that this market is still quite unknown for the majority of U.S. consumers.

Lending Club (LC) is one of the online lending platforms in the United States, alongside Prosper, Sofi, UpStart, VirginMoney, Marcus by Goldman Sachs, and others. Founded in 2007, LC has grown remarkably and it has issued more than $56 billion in loans since its start until 2019:Q4 (Figure 1). This makes it the largest online consumer lending platform in the United States and a major role player in the market.

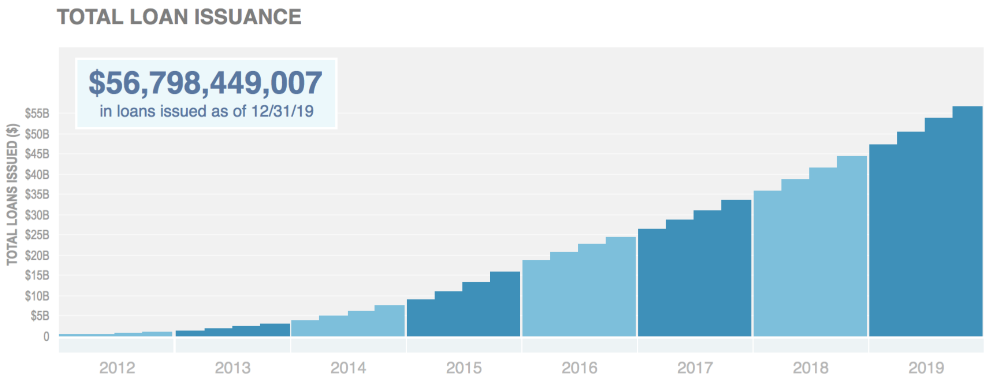

Figure 1: Growth of the origination loans for LC
Source: Lending Club website

Note: This figure shows the cumulative total loan amounts issued by LC—the largest FinTech lender in the United States—since its inception until 2019:Q4. Loan volumes have increased exponentially over time and the total amount is about $56.8 billion by the end of 2019.

The FinTech lenders seems to have several competitive advantages relative to traditional lending channels. First, taking a loan through these platforms is faster. Utilizing systems that allow them to screen borrowers in minutes after their application, saves consumers from the rigid and time-consuming loan application procedures. Second, utilizing the latest technological innovations in data processing, extracting new information (unconventional or alternative data), and using more sophisticated methods for credit assessment—such as machine learning (ML) and Artificial Intelligence (AI)—lenders operating in this market claim to arrive at more accurate credit scoring models for borrowers.[7] Third, FinTech lenders operate almost entirely online, with less to no physical spaces. According to Maudos and Guevara (2004), bank operating costs are one of the most important factors influencing interest rates. Online operation infrastructure removes the cost of the "middleman," potentially making these platforms able to offer loans at lower interest rates compared to any other type of lending (Adams, 2018; De Roure et al., 2016).

The three main players in this market are borrowers, lenders, and the intermediary firm, i.e., the online platform. All of them benefit in certain ways. For borrowers, the process of getting a loan is easier, less time consuming, and most of the times cheaper relative to getting a loan from traditional lending channels For lenders, instead of putting their savings into a bank account and getting a fixed interest rate—typically very low—investors can have higher returns by investing their savings into giving loans to others. Earnings for the online platform are based on origination and commission fees from monthly payments made on each loan.[8]

The online platform is the center point where borrowers and lenders interact with each other. Potential borrowers fill out an online application, where they provide personal information about their income, loan amount requested, the purpose of the loan, and more. Based on the information provided, the applicant's personal credit history, and other factors, the platform/firm accepts or rejects the request. In case of acceptance, the firm assigns a risk score (or grade) which determines the maximum loan amount this potential borrower can receive, the length of the loan, and the interest rate to be paid. After the borrower accepts one of the loan options offered, the firm asks

[7] In a forthcoming paper, I provide evidence that machine learning methods outperform traditional econometrics models used for borrower screening.

[8] LC charges one-time origination fees that vary from 2 percent to 6 percent and an extra 1 percent commission fee for each payment made to investors.

the partner bank (WebBank is the partner bank that LC uses to initiate the loans) to initiate the full amount of the loan to the borrower. Then the bank sells this loan back to the LC, which puts the loan on its website and investors start funding it partially or fully through acquisitions of tranches/notes. The platform is the place where the individual or institutional lenders have the opportunity to quickly and easily invest in a diversified portfolio of loans. They can choose to whom and how much they will lend based on their risk preferences. Furthermore, using the automated investing option, lenders can earn compounded returns on their investments.[9] The role of the online platform in this market is to do the match and screen out risky applicants. Utilizing sophisticated risk assessment models these platforms offer investors opportunities for returns on a pool of more trustworthy borrowers.

A major difference between a FinTech lender and a bank is that the former is not as aggressive as the latter in collecting the debt (Carmichael, 2014), which may result in an enormous quantity of defaulted loans.

1.2.2 Related Literature

This market has received recent attention from researchers. Many studies focus on the FinTech lending market, with some addressing the potential reasons for the market emergence, others focusing on the technology used by these platforms and the information asymmetry problem, and few examining the performance of loans issued through these platforms. Emekter et al. (2015) use a logit model to estimate the important factors of LC customer default. They find that credit grade, debt-to-income ratio, FICO score, and revolving line utilization have an important role in default. Carmichael (2014) uses more sophisticated techniques, where he combines both extensive credit information and soft data to perform a more “thorough” analysis of consumer credit. He finds that the FICO score, borrower-initiated credit inquiries, income, and loan purpose are the most

[9] https://www.lendingclub.com/investing/investor-education/automated-investing.

significant variables for explaining default. Serrano-Cinca, Gutierrez-Nieto and López-Palacios (2015) use univariate means tests and survival analysis to analyze the LC loan performance for loans issued during 2008−2014. They find that factors that best explain default are loan purpose, annual income, current housing situation, credit history, and indebtedness. Also by using a logistic regression model, they conclude that the LC assigned grade is the most predictive factor of default, but it can be improved by adding other information. De Roure et al. (2016) examine the interaction between bank lending and lending via FinTech online platforms in Germany. Their model predicts that transaction loans migrate to FinTech lending market and conclude that the decline in bank lending is correlated with the emergence of FinTech lending. Moreover, they find that risk-adjusted interest rates from FinTech lending are lower compared to those on bank loans. Similarly, Adams (2018) in his empirical study finds that FinTech lenders potentially offer lower interest rates compared to credit cards. Dietrich and Wernli (2016) analyze the determinants of interest rates in the FinTech lending market in Switzerland. Other than the significant effect of loan-specific and macro variables on the interest rates, they also find some discrimination by the lenders. Borrower-specific factors such as economic status have an important role in the lender's evaluation of borrowers' credit risk.

During the last few years, researchers from the Federal Reserve have been interested in this market. Jagtiani and Lemieux (2018) find that LC consumer lending activities have penetrated areas that are underserved by banks, and that the portion of the loans increases in areas where the local economy is not performing well. Jagtiani and Lemieux (2017) find that borrowers who took a loan from LC, on average are riskier than borrowers who choose to take a loan from banks. Also, they provide empirical evidence that for the same risk of default, consumers in FinTech lending market pay lower interest rates relative to traditional lending channels. Wang (2018) discusses how the advancements in information technology have affected the boundary of the firm in the retail lending market. She suggests that FinTech lenders' superior systems of information technology gives them the ability to collect and process information more effectively and efficiently, and this have played an important part in the rapid growth of the market. Hertzberg, Liberman and Paravisini (2018) find that borrowers self-select into different loans based on their private information. They find that borrowers who choose longer maturity loans are more likely to default. Mach, Carter and Slattery (2014) find that small businesses who took a loan in the FinTech lending market paid an interest rate two times higher relative to small business loans from traditional sources. Also, they find that loans taken for business were on average larger compared to other types of loans and had a higher probability of default. Dore and Mach (2019) find that borrowers who used the FinTech lending market, ex-post have higher credit scores and lower credit card utilization rates in the short term, but higher total debt levels in the long run relative to non-users of this market.

A number of studies from the banking and finance literature shows that macro and regional economic conditions play an important role in consumer default, even conditioning for different loan types. Gross and Souleles (2015) analyze credit card default estimating duration models. They find the unemployment rates and house price indices are among the significant variables in predicting default. Agarwal and Liu (2003) show that county unemployment rates significantly influence default in credit cards market. After the Great Recession a number of studies focuses on the determinants of default in non-performing loans (NPLs). Louzis, Vouldis and Metaxas (2012) look at the determinants of NPLs in the Greek banking sector. They show that GDP growth and the change in unemployment rate are among significant variables that explain NPLs and are positively and negatively correlated with NPLs, respectively. Ghosh (2015) studies the impact of macro and regional economic determinants on NPLs for all commercial banks and savings institutions in the United States. He finds that Real GDP and HPI reduces NPLs, while inflation and state unemployment rate are significantly positively correlated with NPLs.

1.3 Data and Descriptive Statistics

1.3.1 FinTech Lending Market Data

My source of loan-level data for the FinTech lending market is LC. In its beginning, the online platform offered only three-year term loans and in May 2010 started offering five-year term loans. The data used for this study includes all three-year matured loans issued between January 2009 to December 2015 with a total of 597,391 loans (Figure 2). I choose to use the universe of loans with clear payment status outcome—completed spell data with clear outcome status—to avoid data censoring issues.

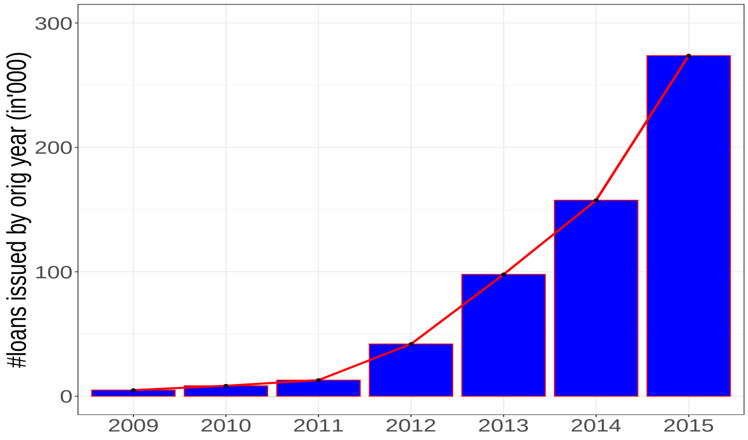

Figure 2: Three-year loan volumes growth by origination year

Note: This figure shows the universe of three-year term loans issued by LC between January 2009 to December 2015 used in this study. In total there are 597,391 loans issued, with an average annual growth of 103%. In a very short period, the number of loans has increased exponentially. More specifically, the total number of loans originated in 2015 is 56 times higher compared to loans originated in 2009.

Some of the loan and individual-specific variables with their respective description used for this study are shown in Table 1. A few other variables are generated based on the information given from the same dataset. For instance, to create the variable *claim creditworthiness*, I follow Carmichael (2014) approach and use text analysis to capture soft information about the borrower from the loan description provided in the process of application.

Table 1: Definitions for the main variables used in the statistical analysis

Variable	**Description**
loan amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
int rate	Interest Rate on the loan.
credit grade	LC assigned loan grade.
home ownership	The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are rent, own, mortgage, and other.
annual inc	The self-reported annual income provided by the borrower during registration.
issue date	The month in which the loan was funded.
last pymnt date	Last month payment was received.
months passed*	months passed since the loan is issued until the last payment date.
loan status	Current status of the loan: *charged-off* or *fully paid*.
default*	100 if default, 0 otherwise based on the loan status.
claim creditworthiness*	1 if the borrower uses some positive keywords in loan description and 0 otherwise.
purpose	The borrower's claim on the purpose s/he is using this loan.
desc	Loan description provided by the borrower.
zip code	The first 3 numbers of the zip code provided by the borrower in the loan application.
FICO	The borrower's FICO at loan origination.
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
delinq 2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years.
inq last 6mths	The number of inquiries in the past 6 months (excluding auto and mortgage inquiries).
open acc	The number of open credit lines in the borrower's credit file.
revol util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
acc now delinq	Number of accounts on which the borrower is now delinquent.
chargeoff within 12mths	Number of charge-offs within 12 months.

Note: This table shows the definitions of the variables used in this study. The explanation for most of the variables is taken from the data dictionary file. Variables with * are derived/created from other variables present in the data.

I search for a list of keywords that borrowers make to claim their creditworthiness and positive character attributes, such as "responsible," "trustworthy," "integrity," "always pay," and "never late," and words associated with job security such as "steady," "stable," and "stability." I also include some words that borrowers use in a variety of ways, such as "reliable" and "always been." *Claim creditworthiness* is a dummy variable indicating whether the borrower uses at least one of these keywords in his/her loan description.

The descriptive statistics for loans and individual characteristics are provided in Table 2. The median individual annual income is $60,000, which seems to be higher than expected compared to $32,632 as reported by U.S. Census Bureau in 2015.[10] The average size of issued loans is about $12,300 which shows that the majority of borrowers use LC for small amount of loans. The average monthly interest rate charged is 12.5 percent. On average, loans are paid after 24 months. This is low because some people paid the loan before the due date, while some others stopped paying. Out of 597,391 loans issued 85,027 of them were not fully paid back, making the average default rate about 14.2 percent. On average, borrowers have a FICO score of 698 with a minimum of 662, a maximum of 848, and a median of 692. The average debt to income ratio is 17 percent.[11] On average a borrower has 11 open credit lines and 24 total credit lines.

[10] Retreived August 12, 2019 https://fred.stlouisfed.org/series/MEPAINUSA672N.

[11] A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.

Table 2: Summary statistics of matured loans with 36 month term (originated Jan 2009 - Dec 2015)

Statistic	Mean	St. Dev.	Min	Median	Max	Skewness	Exc. Kurtosis
annual inc	$71,149	$58,268	$1,896	$60,000	$8,706,582	38.09	4291.2
loan amnt	$12,284	$7,560	$500	$10,000	$35,000	1.06	0.76
int rate (%)	12.53	3.871	5.42	12.50	26.06	0.37	-0.20
months passed	24.07	11.048	0	26	36	-0.42	-1.15
defaulted (%)	14.23	34.273	0	0	100	2.12	2.51
claim creditworthness	0.023	0.149	0	0	1	6.43	39.33
FICO	698.0	31.251	662	692	848	1.27	1.59
dti (%)	17.0	7.951	0	16.5	39.99	0.24	-0.52
delinq 2yrs	0.29	0.811	0	0	29	5.54	57.13
inq last 6mths	0.78	1.071	0	0	33	2.24	15.57
open acc	11.02	4.992	1	10	84	1.21	3.07
revol util (%)	54.78	23.834	0	55.8	892.30	-0.07	3.47
total acc	24.45	11.615	1	23	162	0.86	1.13
acc now delinq	0.004	0.068	0	0	5	20.54	560.74
chargeoff within 12 mths	0.008	0.100	0	0	7	16.78	410.48

Note: This table displays summary statistics for loan and individual attributes of three-year matured loans used in this study. Interest rate, defaulted, debt-to-income ratio, and revolving line utilization are in percentages; borrower self reported income and the loan amount issued are reported in U.S. dollars.

LC assigns a risk grade to each borrower based on the evaluation for credit riskiness to default. There are seven grade categories from A-G with each category having 5 different subgrades.[12] Therefore, in total there are 35 different subgrades from A1-G5 with A1 for borrowers with the lowest risk to default and G5 for borrowers with the highest default probability. These subscores decide the interest rate a borrower is expected to pay, the max amount they can borrow, and the term of the loan. Table 3 shows the distribution of loans by LC's grade and by the FICO score. The majority of borrowers have a grade of B, which makes about 34 percent of originated loans or issued amount. Borrowers with grades A and C have received 23 and 26 percent of the loans and about 25 and 24 percent share of the total amount invested, respectively. Besides, borrowers with very high credit risk, i.e., category F and G make less than 1 percent of the issued loans. Given

[12] As of 11/07/2017 LC has stopped issuing loans to borrowers that fall in the risk category of F and G.

the high default rate in this market (about 14.2 percent), this shows that the majority of defaulters are in other LC assigned grades rather than in the riskiest categories E, F, and G. This rises questions/concerns about the accuracy of the risk assessment models LC uses in pricing loans.[13].

Table 3: Distribution of loans by LC assigned grade and FICO score

Panel A: Grade	**#loans**	**%**	**amount ($ million)**	**%**
Grade-A	138,608	23.20	1,928.43	25.83
Grade-B	204,373	34.21	2,531.14	33.91
Grade-C	155,929	26.10	1,847.57	24.75
Grade-D	73,294	12.27	863.15	11.56
Grade-E	20,567	3.44	246.53	3.30
Grade-F	4,090	0.68	40.90	0.55
Grade-G	530	0.09	6.83	0.09
Panel B: FICO	**#loans**	**%**	**amount ($ million)**	**%**
Subprime (662 ≤ x ≤ 669)	108,311	18.13	1,150.67	15.42
Good (670 ≤ x ≤739)	428,813	71.78	5,476.68	73.37
Very Good (740 ≤ x ≤799)	54,347	9.10	756.16	10.13
Exceptional (800 ≤ x)	5,920	0.99	81.05	1.09
Total	**597,391**	**100.00**	**7,464.55**	**100.00**

Note: This table displays the share of loans issued to borrowers according to the LC assigned grade and traditional FICO score. In more detail, the rows in Panel A show the possible grades assigned by LC, while the columns represent the total number of loans issued to that particular grade, the share in terms of quantity, total amounts issued to that particular grade, and the respective share from total amounts issued. The rows in Panel B show the credit-worthiness categories of borrowers based on their FICO scores according to Equifax consumer credit reporting agency while columns represent the number of loans, quantity share, total amounts issued, and the respective share from total amounts issued for each particular category.

Expressed in FICO score terms, 72 percent of all loans are served to borrowers in the *Good* credit category, 18 percent to the *Subprime*, and the rest to borrowers in the *Very Good* or *Exceptional* group. In terms of monetary value, $5.5 billion has been issued to the *Good* borrowers. The categories *Very Good* and *Exceptional* have received a total of $837 million, and *Subprime* borrowers about $1.2 billion. Close to 90 percent of the loans issued in this market have gone to borrowers in the bottom distribution, as I will discuss this later in this section.

[13] This topic is beyond this paper and is discussed in Azizaj (2020b)

The distribution of FICO categories for each LC grade is shown in Figure 3. More than 55 percent of borrowers in each grade falls in the *Good* category. The share of *Subprime* borrowers is about 4 percent in grade A and it keeps increasing as we move to riskier grades—40 and 43 percent in grades F and G, respectively. Some borrowers in the *Subprime* category can take loans with very low-interest rates, while others with very high FICO scores pay higher prices relative to getting a loan from banks. This suggests that LC follows a very different risk assessment model compared to traditional lending channels where FICO scores play a significant role in default risk evaluation. These findings are consistent with those of Jagtiani and Lemieux (2017) and Adams (2018).

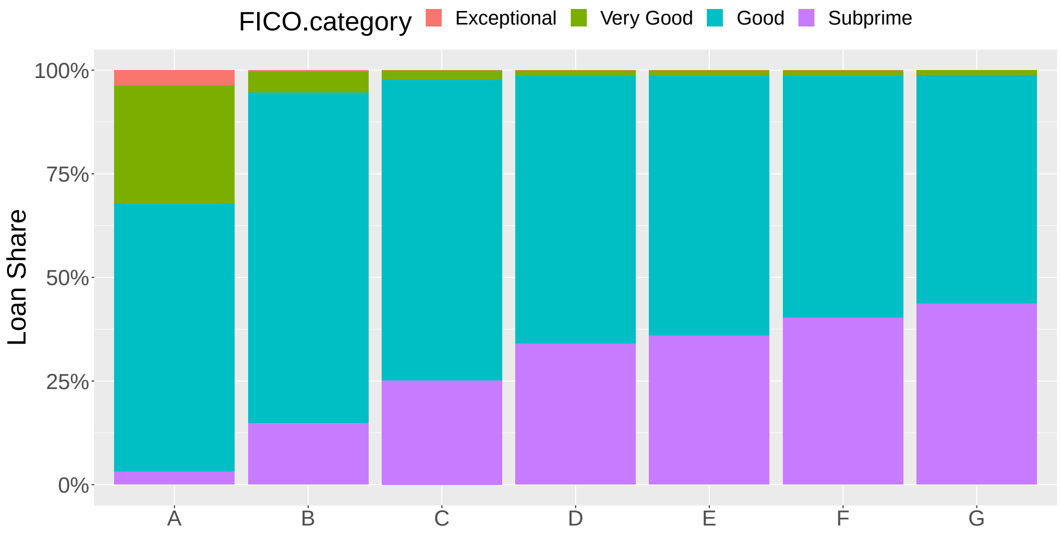

Figure 3: Share of FICO categories in each LC grade

Note: This figure shows the distribution of loans issued to borrowers with different FICO category (Subprime, Good, Very Good, and Exceptional) for each LC grade. Grade A represents borrowers with the lowest default risk and grade G borrowers with the highest default probability.

Figure 4 (left) shows the share of loans by origination year in the FinTech lending market issued to subprime borrowers between 2009–2015. During the first three years of our analysis (2009–2011), the share of subprime borrowers varies between 6 to 8 percent and increases significantly after 2011 until it triples in 2014. Figure 4 (right) shows the default rates in the FinTech lending market relative to traditional consumer loans and credit cards.Default rates for consumer loans offered through FinTech lending are higher relative to other consumer loans issued by traditional financial institutions. The default risk diverges after 2010 because lenders operating in the FinTech lending market increasingly focus on a relatively risky segment of borrowers, for which it appears that they have a comparative advantage relative to traditional lenders. A potential reason is that these lenders are less regulated compared with banks.

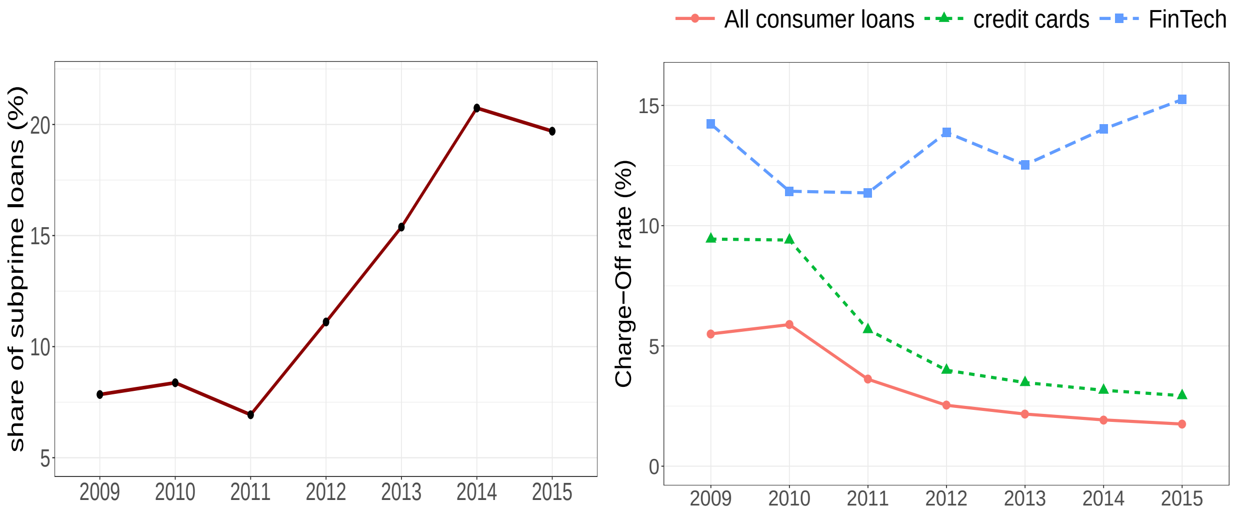

Figure 4: Share of subprime loans by origination year and charge-off rates

Source: Federal Reserve, LC, and author's calculations.

Note: The figure on the left shows the share of loans issued to subprime borrowers by each origination year. In the first years, the share varies between 6 to 8 percent which increases significantly after 2011 and triples in 2014. The figure in the right shows the default rates of loans issued in the FinTech lending market (the blue line) relative to default rates of consumer and credit cards issued by banks (red and green line, respectively).

Figure 5 shows the distribution of total loan amounts by state capita and their respective default rates.[14] States with the highest ratio of the loan amount per state capita are Nevada (NV), and the District of Colombia (DC) closely followed by New Jersey (NJ), and New York (NY). States with the lowest ratio of loan amounts per capita are Maine (ME), Nebraska (NE), and North Dakota (ND). States with the highest ratio of the loan per capita, NV, DC, NJ, and NY have default rates

[14] State population base year of 2010. Iowa (IA) and Idaho (ID) have very few loans during the timeline of our analysis, so I have excluded them from the map.

of 17.7 percent, 9.6 percent, 14.9 percent, and 15.6 percent, respectively. In this case, we can say that investors funding loans in Nevada (NV) are more exposed to risk because on average borrowers in NV have a higher probability of default compared to other borrowers living in DC, NJ, and NY. The states with the highest default rates are Mississippi (MS), NV, Arkansas (AR), and Oklahoma (OK) with associated rates of 19.3 percent, 17.7 percent, 17.5 percent, and 17.0 percent, respectively. The states with the lowest default rate are New Hampshire (NH), DC, Vermont (VT), and ME with respective default rates of 9.4 percent, 9.6 percent, 10.3 percent, and 10.7 percent.

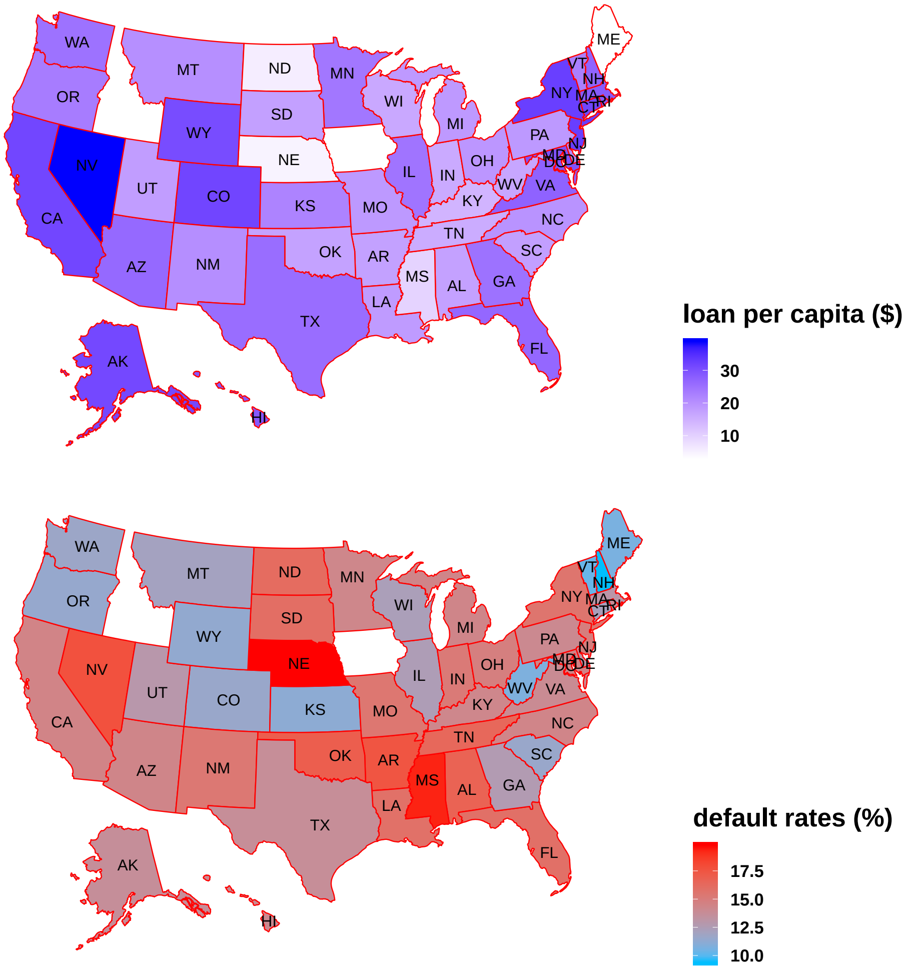

Figure 5: Distribution of loans by state capita and default rates across states

Note: The first figure shows the distribution of loan amounts per state capita and the second figure shows the respective default rates by each state.

1.3.2 Macro and Regional Data

Other than the loan or individual-specific variables, other factors push a borrower towards default. These exogenous factors might be regional or macro shocks. Referring to other similar studies from banking and finance literature, I use the macro and local variables that were found to be significant on default for non-performing loans. I base my initial selection of variables on theory and results from the empirical research (Agarwal and Liu, 2003; Louzis, Vouldis and Metaxas, 2012; Ghosh, 2015). The macroeconomics factors that I include in my model are *GDP growth* and *inflation*, which give information about how the economy is doing at a national level. Regional variables that serve as proxies for local economic conditions are the *unemployment rate*, and *house price index (HPI)* both at the 3-digit ZIP code level. All macro and regional variables expand between 2008–2018.

Table 4 shows the summary statistics for these variables. Macroeconomic indicators are at the national level and yearly frequency. Average GDP growth for the United States during the years of our analysis is 1.6 percent and the average inflation between 2008–2018 is 1.7 percent, with a minimum of –0.4 percent in 2009 and a max of 3.8 in 2008. Unemployment rate statistics are drawn at the county level and monthly frequency. The verage change in the unemployment rate is close to zero, with some counties experiencing a significant decrease in the unemployment rate (e.g., –4.4 percent change in the unemployment rate) and other counties with an increased unemployment rate (e.g., 7.8 percent increased unemployment). The unemployment rate is a lagging indicator and is expected to increase in case of a downturn when the jobs become more scarce. A borrower who loses his or her job has a higher probability of default on a loan compared to someone who is employed, all other variables equal. This is why we should take into consideration the location and the timing when issuing a loan and calculating the credit risk of a potential borrower in accordance. House price indices also are proxies for the regional economy. I present the summary statistics drawn at the 3-digit ZIP code level and quarterly frequency. The average change in HPI is 1.05

with a minimum of –32.7 percent and a maximum of 22.4 percent at some 3-digit ZIP code level and quarterly frequency.

Table 4: Summary statistics of macro and regional variables (2008–2018)

Statistic	**Mean**	**St. Dev.**	**Min**	**Median**	**Max**	**Skewness**	**Exc. Kurtosis**
GDPGR (%)	1.62	1.65	–2.75	2.03	2.88	–1.63	1.65
INF (%)	1.74	1.20	–0.40	1.60	3.80	–0.11	–0.77
UNEMPGR (%)	–0.06	1.37	–4.35	–0.41	7.84	1.75	3.59
HPIGR (%)	1.05	5.77	–32.68	1.26	22.44	–0.80	3.02

Note: This table shows the summary statics for national and regional variables used as proxies for macro and regional shocks in this study. GDP and inflation are at the national level and yearly frequency, unemployment, and house price indices are at 3-digit ZIP code level aggregated at yearly frequency. All variables are in growth terms and expand between 2008–2018.

Figure 6 shows growth rates for the U.S. wide GDP and inflation rate, and the 3-digit ZIP code regional variation for the growth rates of the unemployment rate, and HPI between 2008–2018. During this time frame, the U.S. economy has been in a continuous expansion and with small fluctuations on GDP growth and inflation. However, there is a lot of variation in growth rates of unemployment and HPI across 3-digit ZIP code regions.

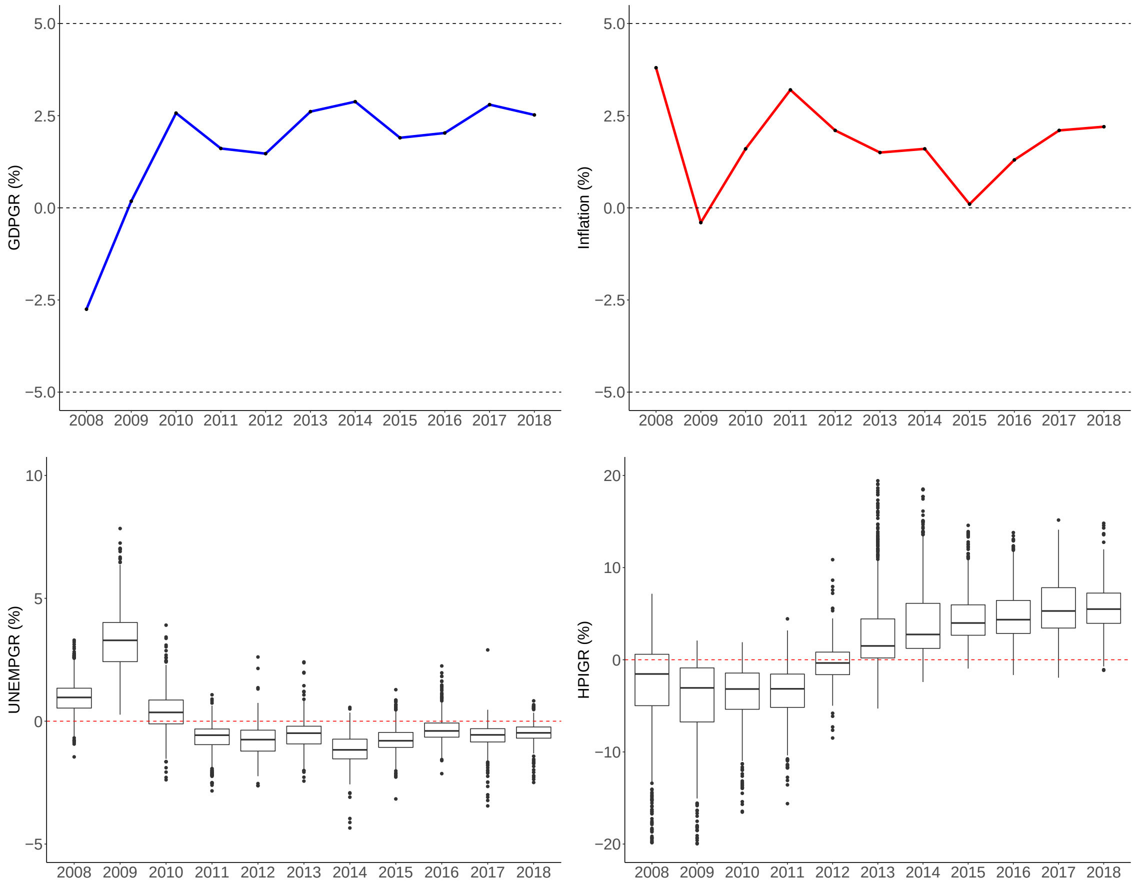

Figure 6: Macro and regional variables over time

Note: The figure in the top left corner shows U.S. wide GDP growth over time. The figure on the top right corner shows the time series of inflation rates. The figures in the bottom show the distribution of growth rates for unemployment and HPI across 3-digit ZIP-code-levels over the time frame of this study.

1.4 Methodology and Empirical Results

The main idea of this paper is to see whether regional and macroeconomic shocks play a role in the default probability of consumers in the FinTech lending market. Even though the economic conditions might be the same, some borrowers file for bankruptcy and others do not. Conditional on default, some people stop paying in the earlier stage of the loan while others stop paying towards the end of the maturity date. So, what are the drivers of default? Are certain categories of borrowers more sensitive to economic conditions than others? How does the probability of default change if the economic conditions had been different and how that changes the total losses in the economy?

In this section, I am going to show three types of analysis to answer the questions risen in this paper. First, I am going to analyze which factors play a significant role in default for the FinTech lending market (micro and macro variables). Second, I introduce the notion of timing in default and do the analysis year by year using a discrete hazard model. Third, I analyze the default behavior for different borrowers subgroups and see whether prime vs. subprime borrowers respond differently to macro and regional shocks.

1.4.1 Overall Loan Default

Individual and loan information might not be available to policy-makers but for completeness and comparison with the current literature, I will show the results that include loan information, individual characteristics, and macro and regional variables. The following linear probability model will give us the factors that play a key role in default for the FinTech lending market.[15]

$$\begin{aligned} \text{defaulted}_{irt} = \beta_0 + \beta_1 \cdot \text{GDPGR}_t + \beta_2 \text{INF}_t + \beta_3 \Delta\text{UNEMPR}_{r,t} + \beta_4 \%\Delta\text{HPI}_{r,t} \\ + \boldsymbol{\alpha} \cdot \mathbf{X}_i + \boldsymbol{\gamma} \cdot \mathbf{L}_i + \epsilon_{irt} \,. \end{aligned} \tag{1}$$

The dependent variable is a default dummy for the loan issued to individual *i*, at region *r*, at the origination year *t*. All macro and regional variables are in growth values at loan origination year t. Individual variables such as annual income, FICO, home ownership, debt to income ratio (dti), etc. are in the **X** vector of dimension 13x1 and all variables related to the loan such as interest rate, and loan amount are in the **L** vector with dimension 2x1. Instead of raw data, I use standardized values. This transformation will put independent variables on a common scale with unit variance and therefore it will be easier to compare the magnitude of coefficients.[16]

[15] I also perform the same exercise using a logit model. The results are similar and available upon request.

[16] Non-standardized regression results are shown in Appendix A.

Higher GDP growth is translated to a healthier economy, which reduces the probability of defaulting for consumers. So I expect a negative coefficient for *GDPGR*. High inflation may benefit borrowers because it lowers the value they are paying back to the lenders. So the higher the inflation rate the lower the probability of default. However, higher inflation also means that consumers are spending more for the same basket of goods and services if the wages are lagging behind the prices, which might leave borrowers with less money to pay the loan, and therefore, a higher probability of default. So both positive and negative coefficients for *INF* are intuitive and interpretable. In regions where the unemployment rate is increasing, the probability that a borrower will be unemployed next year is higher, which will increase the chances of him or her not making payments on an existing loan. I expect a positive coefficient for $\Delta UNEMPR$. The higher the percentage change in HPI in a particular region means houses are more valued. If the borrower is a home owner this means s/he possesses more wealth, which lowers the default probability because s/he can borrow against the house value. However, if a borrower rents, this is an indicator of higher rent prices in that area, therefore, the lower probability to pay the debt obligations which might increase the default probability. Both positive and negative signs for $\%\Delta HPI$ are intuitive and interpretable.

Table 5: Linear Probability Model of Default Behavior - Overall (with micro and macro variables)

	defaulted		
	(1)	(2)	(3)
constant	33.01***(1.10)	14.20***(0.05)	33.57***(1.16)
FICO	-1.50***(0.06)		-1.24***(0.06)
int_rate	5.14***(0.05)		5.50***(0.06)
loan_amnt (log)	2.02***(0.08)		2.06***(0.08)
home_ownership.rent	2.40***(0.09)		2.51***(0.10)
annual_inc (log)	-3.57***(0.11)		-3.65***(0.11)
verification_status.Verified	1.51***(0.10)		1.20***(0.10)
claim creditworthness	-1.92***(0.38)		-1.48***(0.41)
dti	1.53***(0.05)		1.35***(0.05)
delinq 2yrs	0.22***(0.05)		0.20***(0.05)
inq last 6mths	1.04***(0.05)		1.11***(0.05)
open_acc	1.03***(0.06)		0.93***(0.07)
revol_util	-0.63***(0.05)		-0.48***(0.06)
total_acc	-0.79***(0.06)		-0.65***(0.07)
acc_now_delinq	-0.06(0.04)		-0.08*(0.05)
chargeoff_within_12_mths	-0.05(0.04)		-0.07(0.05)
GDPGR_Y1		-0.10*(0.06)	-0.67***(0.06)
INF_Y1		-0.96***(0.06)	-1.27***(0.06)
ΔUNEMPR_Y1		-0.39***(0.05)	-0.39***(0.05)
%ΔHPI_Y1		-0.25***(0.06)	-0.48***(0.06)
Observations	597,391	550,146	550,146
Adjusted R^2	0.05	0.001	0.05
Residual Std. Error	34.11 (df = 597,375)	34.90 (df = 550,141)	34.04 (df = 550,126)
F Statistic	1,968.757*** (df = 15; 597,375)	109.037*** (df = 4; 550,141)	1,501.506*** (df = 19; 550,126)

Note: This table shows the results for the linear probability model of default behavior with individual characteristics, loan attributes, and macro and regional economic indicators at the time the loan is originated. I use the natural log of all variables that are in dollar amounts ($) such as *annual income* and *loan amount*. Categorical variables status is modified to fewer categories. For example, *verification status* categories of verified and source verified as *Verified* and the rest is kept unchanged. Home ownership categories of mortgage and own as *own*, and the rest of categories as *rent*. *Claim creditworthiness* and dependent variable *defaulted* are dummies. All other variables are in standardized values. *** (**) [*] indicate significance at the 1 (5) [10] percent level.

Table 5 shows the estimated results with and without macro and regional variables. The majority of loan and individual variables are significant and have the expected signs. The variable that best explains default is the interest rate. The higher the interest rates the higher the probability of default on a given loan. One standard deviation increase in interest rates increases the default probability by 5.1 percentage points. The FICO score has a negative impact on default, meaning the higher borrower's FICO score the lower the risk of defaulting. The higher loan amounts, debt-to-income ratio, delinquencies in the last two years, inquires in the last 6 months, the number of open accounts are, the higher the probability that borrower will default on a given loan. Also, the borrowers who rent have a higher risk of default than those who are homeowners. Borrowers with a high total number of credit lines currently in their credit file have a lower probability of default. This is because borrowers with more credit cards have the opportunity to manage debt by using the other cards to pay the minimum amounts required. The number of accounts the borrower is already delinquent does not play a role in the default probability. An interesting finding is the significance of the coefficient for the variable that borrowers make positive claims on their credit-worthiness, which is constructed based on text analysis of the loan description given during the application process. Therefore, borrowers who claim that they will pay back the loan because they have a stable job and/or they are reliable, have a lower probability of default.

All macro and regional variables are significant. High GDPGR in the year a loan is issued is an indicator of a healthy economy. Therefore, borrowers are more confident to pay back the loan and this lowers the risk of default. High inflation is a benefit for borrowers because the value of the loan paid back will be lower. Therefore, higher inflation rates incentivize borrowers to pay back the loan. The coefficient for the unemployment rate is significant, but with a negative sign. The unemployment rate is a lagging indicator, so the interpretation for the coefficient of change in the unemployment rate is hard to draw any conclusion. Higher the percentage change in HPI at the time a loan is issued lowers the chances the borrower will default. A one standard deviation increase in HPI inflation at the year the loan is issued decreases the probability of default by 0.48 percentage points. Overall, the significance and direction of the estimates for macro and regional variables are the same when we add individual and loan characteristics. This indicates that macro and regional variables are robust to the exclusion or inclusion of individual and loan attributes. Because the focus of this paper is on macro and regional variables, I will focus on the sensitivity of default to these variables from now on.

1.4.2 Importance of Timing in Default

Whether a borrower defaults in the first month after taking a loan or one month before the maturity date—when s/he has already paid almost the loan—the consequences are the same: borrower's credit score will plummet and the default remains in the credit report for 7–10 years. Therefore, it is going be hard for a borrower to take other loans from other places, and even if s/he can, the interest rates will increase significantly. This means that if a borrower knows that s/he is going to default on a loan, it is optimal to stop making payments from the beginning. This way s/he will have higher gains than defaulting at a later stage of the loan.

The existing literature has modeled the default for the FinTech lending market focusing on the overall default status of a loan only (i.e., whether a borrower defaults or not). However, we should not care only whether a borrower goes bankrupt or not, but also the time this event happens. Why should we care about the timing people default? Earlier a borrower stops making payments on a loan, higher the losses for the lenders, and vice versa. Imagine an extreme scenario where all people default in the third year (i.e., last year of the maturity). In this case, the losses would be small because borrowers have already paid 2/3 of the loan. However, if all borrowers default in the first year of the loan than the losses would be much higher since they are paying a very small portion of the loan. Therefore, the composition of default plays an important role in the destabilizing potential for the economy.

Table 6 shows the timing when borrowers choose to default and the respective losses for three-year matured loans used in this study. Conditional on default, 37 percent of borrowers have stopped making payments within 12 months after taking the loan, causing a total loss of \$337.3 million. The majority of defaulters file for bankruptcy in the 2nd year of the loan, 42 percent, therefore creating a total loss of \$229.3 million. Fewer people stop paying in the 3rd year, 21 percent with a total loss of $44.1 million. It is important to realize that even though the majority of borrowers have defaulted in the second year of the loan, when we focus on losses, the ones who have defaulted in the 2nd and 3rd year together do not add up the losses of the ones defaulted in the 1st year. Therefore, the composition of default-time in the FinTech lending market is important and should be taken into consideration.

Table 6: Default distribution by maturity and associated losses

	#loans	**%**	**Total share**	**Losses ($ million)**	**%**	**Total share**
defaulted_1stY	31,989	37.6%	5.35%	$337.3	55.23%	4.52%
defaulted_2ndY	35,441	41.7%	5.93%	$229.3	37.55%	3.07%
defaulted_3rdY	17,597	20.7%	2.95%	$44.1	7.22%	0.59%
Total defaults	**85,027**	**100%**	**14.23%**	**$ 610.7**	**100%**	**8.18%**

Note: This table shows the default distribution by duration and the associated losses. The dataset I use does not include intermediate payments but includes the total amount paid together with the total interest and installments. It updates periodically. Among all other variables, only the date of the last payment and loan status are updated over time.

I examine the impact of macro and regional shocks on default using a discrete hazard model. I split the timing of default in three different categories: the ones who defaulted within the first year after taking the loan, the ones who defaulted in the second year of duration and the ones who defaulted in the third year of duration (in the remainder of the paper I will call these as "1st year", "2nd year" and "3rd year", respectively). To capture which group of borrowers is more sensitive towards the business cycle, I run the following regressions:

$$\begin{aligned}\text{defaulted\_1}stY_{irt} = \alpha_0 + \alpha_1\text{subprime}_i + \alpha_2\text{GDPGR}_t + \alpha_3\text{INF}_t \\ + \alpha_4\Delta\text{UNEMPR}_{r,t} + \alpha_5\%\Delta\text{HPI}_{r,t} + \epsilon_{irt},\end{aligned} \tag{2}$$

$$\text{defaulted\_}2ndY_{irt} = \beta_0 + \beta_1\text{subprime}_i + \beta_2\text{GDPGR}_{t+1} + \beta_3\text{INF}_{t+1} + \beta_4\Delta\text{UNEMPR}_{r,t+1} + \beta_5\%\Delta\text{HPI}_{r,t+1} + \epsilon_{irt}\,, \tag{3}$$

$$\text{defaulted\_}3rdY_{irt} = \gamma_0 + \gamma_1\text{subprime}_i + \gamma_2\text{GDPGR}_{t+2} + \gamma_3\text{INF}_{t+2} + \gamma_4\Delta\text{UNEMPR}_{r,t+2} + \gamma_5\%\Delta\text{HPI}_{r,t+2} + \epsilon_{irt}\,. \tag{4}$$

For $x \in \{1,2,3\}$, the three different dependent variables for default are constructed as follows:

$$\text{defaulted\_}x = \begin{cases} 100, & \text{if defaulted within the } x^{th} \text{ year} \\ 0, & \text{otherwise} \end{cases}$$

The variable $subprime_i$ is a dummy variable created from the FICO score. If the borrower's FICO score is between 662 and 669 *subprime* equals one, and zero otherwise. Equation 2 includes all observations, while I am excluding the borrowers who did not survive the default event in the 1st year for equation 3, and excluding defaulters from the 1st or 2nd year in equation 4. Every macro and regional economic variable is at the year the default event has happened.

Table 7 shows the results from the discrete hazard model. The first group seems to be mostly sensitive to macro shocks. Significant coefficients have the expected signs. One standard deviation increase in GDPGR decreases the probability of default by 0.28 percentage points. One standard deviation increase in GDP growth at the year loan is issued reduces the default probability by 0.24p.p. One standard deviation increase in inflation decreases the average default probability by 0.24 percentage points. Regional variables seem to have little to no significant effect on the default probability for the first year of the loan. Among all macro and regional variables for the second year default, GDP growth and change in unemployment rate seems to be the most significant factors effecting default, and both have the expected sign. One standard deviation increase in GDP growth at the second year of the loan decreases default probability by 0.13 percentage points. Change in unemployment seems to have the highest impact on default among all macro and regional variables. While HPI inflation does not seem to play an important role on the default event for the first and second group, it is the only variable that is significant for the third group. One standard deviation increase in the HPI inflation decreases the default probability by 0.1 percentage points.

Table 7: Estimation Results from the Discrete Hazard Model

	defaulted_1stY	**defaulted_2ndY**	**defaulted_3rdY**
constant	4.883***(0.034)	5.699***(0.037)	3.018***(0.028)
subprime1	2.685***(0.079)	3.128***(0.088)	1.587***(0.068)
GDPGR_Y1	–0.282***(0.039)		
GDPGR_Y2		–0.113**(0.040)	
GDPGR_Y3			0.096*(0.040)
INF_Y1	–0.239***(0.038)		
INF_Y2		–0.084*(0.039)	
INF_Y3			0.070 (0.040)
ΔUNEMPR_Y1	–0.081*(0.035)		
ΔUNEMPR_Y2		0.218*** (0.038)	
ΔUNEMPR_Y3			0.014(0.028)
%ΔHPI_Y1	–0.043 (0.037)		
%ΔHPI_Y2		0.020(0.035)	
%ΔHPI_Y3			–0.104***(0.027)
Observations	550,146	520,579	488,053
Adjusted R^2	0.002	0.003	0.001
Residual Std. Error	22.523 (df = 550,140)	24.171 (df = 520,573)	17.828 (df = 488,047)
F Statistic	274.891*** (df = 5; 550,140)	274.575*** (df = 5; 520,573)	120.402*** (df = 5; 488,047)

Note: This table shows the estimated results from the discrete hazard model. *** (**) [*] indicate significance at the 1 (5) [10] percent level.

1.4.3 Business Cycle Default Sensitivity

Because the focus of this paper is default, it is important to see whether certain categories of borrowers are more likely to default in downturns. As the lenders in this market focus lending more and more to subprime borrowers, is this group of borrowers more sensitive to the business cycles than the prime borrowers? How do their probability of default change relative to the prime group? The following discrete hazard equations with the interaction terms would show us if this is the case.[17]

$$\begin{aligned} \text{defaulted\_1stY}_{irt} = {} & \alpha_0 + \alpha_1 \text{subprime}_i + \alpha_2 \text{GDPGR}_t + \alpha_3 \text{INF}_t + \alpha_4 \Delta\text{UNEMPR}_{r,t} \\ & + \alpha_5 \%\Delta\text{HPI}_{r,t} + \alpha_6 \text{GDPGR}_t \cdot \text{subprime}_i + \alpha_7 \text{INF}_t \cdot \text{subprime}_i \\ & + \alpha_8 \Delta\text{UNEMPR}_{r,t} \cdot \text{subprime}_i + \alpha_9 \%\Delta\text{HPI}_{r,t} \cdot \text{subprime}_i + \epsilon_{irt} \,, \end{aligned} \tag{5}$$

$$\begin{aligned} \text{defaulted\_2ndY}_{irt} = {} & \beta_0 + \beta_1 \text{subprime}_i + \beta_2 \text{GDPGR}_{t+1} + \beta_3 \text{INF}_{t+1} + \beta_4 \Delta\text{UNEMPR}_{r,t+1} \\ & + \beta_5 \%\Delta\text{HPI}_{r,t+1} + \beta_6 \text{GDPGR}_{t+1} \cdot \text{subprime}_i + \beta_7 \text{INF}_{t+1} \cdot \text{subprime}_i \\ & + \beta_8 \Delta\text{UNEMPR}_{r,t+1} \cdot \text{subprime}_i + \beta_9 \%\Delta\text{HPI}_{r,t+1} \cdot \text{subprime}_i + \epsilon_{irt} \,, \end{aligned} \tag{6}$$

$$\begin{aligned} \text{defaulted\_3rdY}_{irt} = {} & \gamma_0 + \gamma_1 \text{subprime}_i + \gamma_2 \text{GDPGR}_{t+2} + \gamma_3 \text{INF}_{t+2} + \gamma_4 \Delta\text{UNEMPR}_{r,t+2} \\ & + \gamma_5 \%\Delta\text{HPI}_{r,t+2} + \gamma_6 \text{GDPGR}_{t+2} \cdot \text{subprime}_i + \gamma_7 \text{INF}_{t+2} \cdot \text{subprime}_i \\ & + \gamma_8 \Delta\text{UNEMPR}_{r,t+2} \cdot \text{subprime}_i + \gamma_9 \%\Delta\text{HPI}_{r,t+2} \cdot \text{subprime}_i + \epsilon_{irt} \,. \end{aligned} \tag{7}$$

[17] I tried the Principal Component Analysis (PCA) to address these questions. PCA reduces the dimensionality of a dataset, possibly with many variables correlated to each other. Implemented on macro and regional variables, PCA gives independent latent variables. However, some of the regression coefficients on these variables are hard to interpret and draw intuitive conclusions. More on PCA and obtained results is found in Appendix B.

Table 8: Results from the Discrete Hazard Model with Interaction Terms

	defaulted_1stY	**defaulted_2ndY**	**defaulted_3rdY**
constant	4.885***(0.034)	5.699***(0.037)	3.018***(0.028)
subprime1	2.759***(0.081)	3.150***(0.088)	1.585***(0.068)
GDPGR_Y1	-0.216***(0.042)		
GDPGR_Y2		-0.045(0.043)	
GDPGR_Y3			0.103**(0.043)
INF_Y1	-0.242***(0.041)		
INF_Y2		-0.175***(0.042)	
INF_Y3			0.045(0.043)
ΔUNEMPR_Y1	-0.072*(0.038)		
ΔUNEMPR_Y2		0.190***(0.041)	
ΔUNEMPR_Y3			0.014(0.030)
%ΔHPI_Y1	-0.004(0.041)		
%ΔHPI_Y2		-0.021(0.038)	
%ΔHPI_Y3			-0.089***(0.029)
GDPGR_Y1*subprime1	-0.554***(0.119)		
GDPGR_Y2*subprime1		-0.580***(0.123)	
GDPGR_Y3*subprime1			-0.115(0.121)
INF_Y1*subprime1	0.235**(0.115)		
INF_Y2*subprime1		0.642***(0.112)	
INF_Y3*subprime1			0.238*(0.125)
ΔUNEMPR_Y1*subprime1	-0.010(0.099)		
ΔUNEMPR_Y2*subprime1		0.053(0.105)	
ΔUNEMPR_Y3*subprime1			-0.014(0.078)
%ΔHPI_Y1*subprime1	-0.207**(0.095)		
%ΔHPI_Y2*subprime1		0.190*(0.097)	
%ΔHPI_Y3*subprime1			-0.104(0.074)
Observations	550,146	520,579	488,053
Adjusted R^2	0.003	0.003	0.001
Residual Std. Error	22.522 (df = 550136)	24.170 (df = 520569)	17.828 (df = 488043)
F Statistic	157.584*** (df = 9; 550136)	157.810*** (df = 9; 520569)	67.708*** (df = 9; 488043)

Note: This table shows the estimated results from the discrete hazard model with the interaction terms. *** (**) [*] indicate significance at the 1 (5) [10] percent level.

Table 8 shows the estimation results of the above regressions with interaction terms. Now the interpretation of main variables coefficients will be different. The impact of GDP growth on the default probability of the first year depends not only on the main variable coefficient but also on the borrower's credit-worthiness (i.e., whether the borrower belongs to the subprime or prime category). The coefficients are estimated using the following equation:

$$\frac{\partial \text{defaulted\_1stY}}{\partial \text{GDPGR}_t} = \alpha_2 + \alpha_6 \text{subprime}_i\,. \tag{8}$$

Table 9 shows the marginal effects of macro and regional economic conditions on default for each borrower credit score category (i.e., prime vs. subprime) across maturities. The absolute values of coefficients for subprime borrowers are higher than the prime. This shows that subprime borrowers are more sensitive to the overall national and regional economic conditions than prime borrowers. Moreover, relative to macro variables' significance, regional economic conditions seem to have little to no significant impact on the default sensitivity of subprime borrowers.

Table 9: Marginal effects on defaults for *Prime* vs. *Subprime* borrowers

category	varible	defaulted_1stY	defaulted_2ndY	defaulted_3rdY
Prime	GDPGR_Y1	–0.216***(0.042)		
	GDPGR_Y2		–0.045(0.041)	
	GDPGR_Y3			0.103*(0.041)
	INF_Y1	–0.242***(0.041)		
	INF_Y2		–0.175***(0.040)	
	INF_Y3			0.045(0.041)
	ΔUNEMPR_Y1	–0.072*(0.038)		
	ΔUNEMPR_Y2		0.190***(0.039)	
	ΔUNEMPR_Y3			0.014(0.029)
	%ΔHPI_Y1	–0.004(0.041)		
	%ΔHPI_Y2		–0.021(0.037)	
	%ΔHPI_Y3			–0.089**(0.028)
Subprime	GDPGR_Y1	–0.770***(0.130)		
	GDPGR_Y2		–0.626***(0.135)	
	GDPGR_Y3			–0.012(0.133)
	INF_Y1	–0.007(0.127)		
	INF_Y2		0.467***(0.122)	
	INF_Y3			0.283*(0.138)
	ΔUNEMPR_Y1	–0.082(0.108)		
	ΔUNEMPR_Y2		0.243*(0.114)	
	ΔUNEMPR_Y3			0.0001(0.084)
	%ΔHPI_Y1	–0.210*(0.101)		
	%ΔHPI_Y2		0.169(0.105)	
	%ΔHPI_Y3			–0.194*(0.080)

Note: This table shows the the marginal effects on default for prime and subprime borrowers. *** (**) [*] indicate significance at the 1 (5) [10] percent level.

Prime: For those who defaulted within the first year after the loan was originated, only macro variables seem to be significant. One standard deviation increase in GDP growth at the year the loan is issued lowers the default probability for the prime by 0.22 percentage points. Inflation seems to be the most important variable with an estimated coefficient of -0.24. Higher inflation rates lower the real value of the outstanding loan and incentivize prime borrowers to repay the loan. For the second year, inflation and change in the unemployment rate are highly significant. One standard deviation increase of inflation in the second year of the loan lowers the default probability by 0.18

percentage points. One standard deviation increase in unemployment increases default probability by 0.19 percentage points. Only a small percentage of prime borrowers default in the third vintage. Therefore, estimated coefficients are not highly significant compared to other years. The percentage change in HPI seems to be the only variable that plays a significant role in default for this group.

Subprime: Borrowers in this category are more likely to default on a given loan. The model predicts that the main economic indicator that plays an important role in default for the first year is GDP growth. One standard deviation increase in GDP growth lowers the default probability by almost 0.8 percentage points, which is significantly higher from the estimated coefficient for prime borrowers. Most of the defaults take place in the second year of the loan, explaining the significance of macro and regional variables in the second year. Different from prime borrowers, this category of borrowers seem to be more sensitive to macro variables. Higher the GDP growth this year pushes borrowers towards paying their loans and decreases the default probability by 0.63 percentage points. Inflation plays a significant role in the default probability for the second year but has the opposite sign compared to inflation coefficients for prime borrowers. If subprime borrowers default more when prices go up, this may indicate that the increase in prices is not keeping up with wages for this category of borrowers. Therefore higher the inflation rate higher the default risk for subprime borrowers. In the third year, inflation and %ΔHPI seem to be significance at 10 percent level.

Main findings: Subprime borrowers are more sensitive to macro and regional economic conditions than prime borrowers. Macro variables play a higher role in default compared to regional variables for both prime and subprime borrowers. During economic scenarios with high inflation rates, prime borrowers are more likely to pay back the loan, while subprime borrowers are more likely to default.

1.5 Counterfactual Analyses

In this section, I use the estimated coefficients from the discrete hazard models and perform counterfactual analyses for different economic scenarios. I calculate the default probabilities across maturities in case of an unexpected recession or a boom and quantify the total losses coming from this market. Moreover, I predict the total losses for different growth rates and different borrower type composition for this market.

1.5.1 Different Economic Scenarios

One of the main roles of the Federal Reserve is to regulate and supervise credit markets and ensure financial stability. Supervising the FinTech lending market with exponential growth and high default rates would be part of this important role. They would be interested in knowing how much risk is in this market. Building counterfactual analyses for different economic scenarios and predicting the losses would give insights on the potential destabilizing effect for each economic scenario, and therefore insights on policy implications.

I simulate three different economic scenarios that from hereafter I will refer to as *actual*, *boom*, and *recession*. In the "actual" scenario, all macro and regional variables take the average values in the actual dataset. For a "boom" scenario, I increase the average GDP growth, inflation, and %ΔHPI by two standard deviations, and decrease the average ΔUNEMPR by the same magnitude.[18] For consistency preferences, I construct a "recession" scenario by decreasing the average GDP growth, inflation and %ΔHPI by two standard deviations and increase the average ΔUNEMPR by the same magnitude.

Using the estimated coefficients from the duration model, I calculate the default probability across maturities for each scenarios. Because the left hand side of the three main equations 5, 6, and 7 is a default dummy, the fitted values give the default probability for each borrower. The

[18] Remember that all variables are standardized.

mean of the fitted values for default gives us the average default probability across maturities.[19] The hazard functions for the prime borrowers in an actual scenario are as follows:

$$H_1^{prime.actual} = \hat{\alpha}_0 + \hat{\alpha}_2 \cdot \overline{\text{GDPGR}_t} + \hat{\alpha}_3 \cdot \overline{\text{INF}_t} + \hat{\alpha}_4 \cdot \overline{\Delta\text{UNEMP}_{r,t}} + \hat{\alpha}_5 \cdot \overline{\%\Delta\text{HPI}_{r,t}}\,,$$

$$H_2^{prime.actual} = \hat{\beta}_0 + \hat{\beta}_2 \cdot \overline{\text{GDPGR}_{t+1}} + \hat{\beta}_3 \cdot \overline{\text{INF}_{t+1}} + \hat{\beta}_4 \cdot \overline{\Delta\text{UNEMP}_{r,t+1}} + \hat{\beta}_5 \cdot \overline{\%\Delta\text{HPI}_{r,t+1}}\,,$$

$$H_3^{prime.actual} = \hat{\gamma}_0 + \hat{\gamma}_2 \cdot \overline{\text{GDPGR}_{t+2}} + \hat{\gamma}_3 \cdot \overline{\text{INF}_{t+2}} + \hat{\gamma}_4 \cdot \overline{\Delta\text{UNEMP}_{r,t+2}} + \hat{\gamma}_5 \cdot \overline{\%\Delta\text{HPI}_{r,t+2}}\,.$$

The hazard functions for subprime borrowers across maturities are calculated as follows:

$$\begin{aligned} H_1^{subprime.actual} &= (\hat{\alpha}_0 + \hat{\alpha}_1) + (\hat{\alpha}_2 + \hat{\alpha}_6) \cdot \overline{\text{GDPGR}_t} + (\hat{\alpha}_3 + \hat{\alpha}_7) \cdot \overline{\text{INF}_t} \\ &\quad + (\hat{\alpha}_4 + \hat{\alpha}_8) \cdot \overline{\Delta\text{UNEMP}_{r,t}} + (\hat{\alpha}_5 + \hat{\alpha}_9) \cdot \overline{\%\Delta\text{HPI}_{r,t}}\,, \end{aligned}$$

$$\begin{aligned} H_2^{subprime.actual} &= (\hat{\beta}_0 + \hat{\beta}_1) + (\hat{\beta}_2 + \hat{\beta}_6) \cdot \overline{\text{GDPGR}_{t+1}} + (\hat{\beta}_3 + \hat{\beta}_7) \cdot \overline{\text{INF}_{t+1}} \\ &\quad + (\hat{\beta}_4 + \hat{\beta}_8) \cdot \overline{\Delta\text{UNEMP}_{r,t+1}} + (\hat{\beta}_5 + \hat{\beta}_9) \cdot \overline{\%\Delta\text{HPI}_{r,t+1}}\,, \end{aligned}$$

$$\begin{aligned} H_3^{subprime.actual} &= (\hat{\gamma}_0 + \hat{\gamma}_1) + (\hat{\gamma}_2 + \hat{\gamma}_6) \cdot \overline{\text{GDPGR}_{t+2}} + (\hat{\gamma}_3 + \hat{\gamma}_7) \cdot \overline{\text{INF}_{t+2}} \\ &\quad + (\hat{\gamma}_4 + \hat{\gamma}_8) \cdot \overline{\Delta\text{UNEMP}_{r,t+2}} + (\hat{\gamma}_5 + \hat{\gamma}_9) \cdot \overline{\%\Delta\text{HPI}_{r,t+2}}\,. \end{aligned}$$

The hazard function for the first year default for prime borrowers if the economy experience a boom is as follows:

[19] Remember that prime and subprime borrowers have different coefficients.

$$
\begin{aligned}
H_1^{prime.boom} &= \hat{\alpha}_0 + \hat{\alpha}_2 \cdot (\overline{\text{GDPGR}_t} + 2) + \hat{\alpha}_3 \cdot (\overline{\text{INF}_t} + 2) + \hat{\alpha}_4 \cdot (\overline{\Delta\text{UNEMP}_{r,t}} - 2) + \hat{\alpha}_5 \cdot (\overline{\%\Delta\text{HPI}_{r,t}} + 2) \\
&= 2\hat{\alpha}_2 + 2\hat{\alpha}_3 - 2\hat{\alpha}_4 + 2\hat{\alpha}_5 + H_1^{prime.actual}.
\end{aligned}
$$

Applying the same logic, I construct all the other hazard functions for prime and subprime borrowers in case of a boom: $H_2^{prime.boom}$, $H_3^{prime.boom}$, $H_1^{subprime.boom}$, $H_2^{subprime.boom}$, $H_3^{subprime.boom}$. Similarly the hazard function for prime borrowers and in case of recession is as follows:

$$
\begin{aligned}
H_1^{prime.recession} &= \hat{\alpha}_0 + \hat{\alpha}_2 \cdot (\overline{\text{GDPGR}_t} - 2) + \hat{\alpha}_3 \cdot (\overline{\text{INF}_t} - 2) + \hat{\alpha}_4 \cdot (\overline{\Delta\text{UNEMP}_{r,t}} + 2) + \hat{\alpha}_5 \cdot (\overline{\%\Delta\text{HPI}_{r,t}} - 2) \\
&= -2\hat{\alpha}_2 - 2\hat{\alpha}_3 + 2\hat{\alpha}_4 - 2\hat{\alpha}_5 + H_1^{prime.actual}.
\end{aligned}
$$

Similarly, I construct all the other hazard functions for a recession: $H_2^{prime.recession}$, $H_3^{prime.recession}$, $H_1^{subprime.recession}$, $H_2^{subprime.recession}$, $H_3^{subprime.recession}$.

Figure 7 shows the predicted default probabilities for prime and subprime borrowers across maturities and for all three different economic scenarios simulated. The black line represents hazard values for the actual, while the blue and red lines for scenarios of an unexpected boom or a recession, respectively. The default hazard is highest in the second year of the loan for both categories of borrowers, which overall seem to have quite different hazard patterns. In each loan stage, the default probability for subprime borrowers is higher than the prime. For example, for the first and second year of the loan, the average default probabilities for a recession scenario for the subprime is 9.5 and 9.3 percent, while for a typical prime borrower the default probabilities are 5.7 and 6.6 percent, respectively. In this economic scenario, a typical subprime borrower who is in the first year of the loan has 66.7 percentage higher default probability than a typical prime borrower. Moreover, the hazard function for subprime borrowers in the first year is more spread out compared with other maturities and relative to prime borrowers'. The reason for this is that subprime borrowers are more sensitive to macro and regional economic conditions. The subprime hazard for the second year is not much spread because the inflation for this group of borrowers has a positive impact on default (as we show in Table 9. In case of a boom or a recession, the change in default probabilities for the subprime will be larger relative to a typical prime borrower.

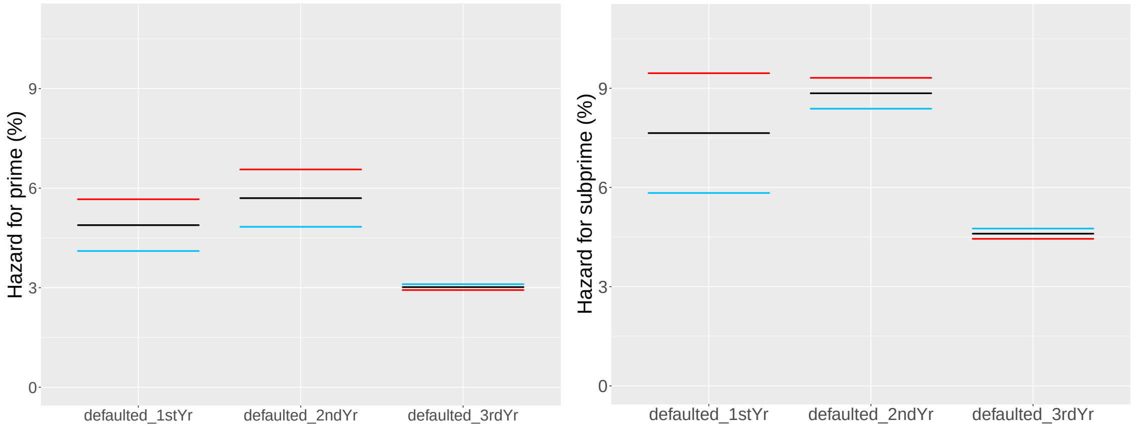

Figure 7: Hazard functions for prime and subprime borrowers on different economic scenarios

Note: This figure shows the predicted default probabilities for different economic scenarios: actual, boom, and recession (black, blue, and red, respectively). Hazard functions for subprime are higher than prime for each duration. Moreover, subprime borrowers default probabilities are more spread out because they are more sensitive to macro and regional shocks.

Assume it is the year 2015 and the Federal Reserve wants to quantify the total losses from the stock of outstanding loans in this market for different economic scenarios for 2015. For simplicity, assume that all loans are issued in the middle of the calendar year (i.e., July 1st). Because these are three-year loans, loans originated in 2013 are in the third year of their maturity in 2015; the ones originated in 2014 are in the second year and the ones originated in 2015 are in the first year. We know the composition of borrowers (prime and subprime) for each origination year. Using the hazard values from Figure 7 together with borrower compositions by loan origination year, we can predict the total losses in this market for all three economic scenarios of 2015. Figure 8 shows the flowchart of this counterfactual exercise.

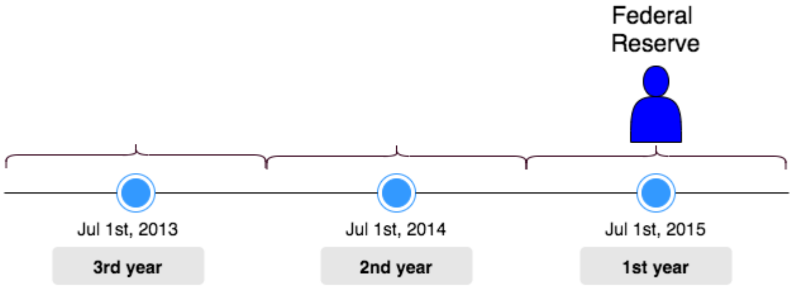

Figure 8: Snapshot of time

Note: This diagram illustrates the Federal Reserve standing in 2015 and want to quantify the total losses from defaults in this market for different economic scenarios of 2015. Knowing the default probabilities for each economic scenario, together with borrower composition in each origination year helps to predict the losses. The curly brackets show the year, the blue dot shows the middle of the year, and the straight line shows the time concept. Loans originated in 2013 are in the third year duration in 2015 and so on.

Losses associated with default from loans originated in a particular year equals the product of number of borrowers according to their type, the hazard for that particular type of borrower and at that maturity, and the average outstanding debt. An illustration of how I calculate the losses for loans issued in 2013 in case of a recession scenario in 2015 is as follows:

$$\text{Losses}^{2013}_{prime} = N^{prime}_3 \cdot H^{prime.recession}_3 \cdot \text{AVG.EXPECTED.AMNT}^{2013}_{prime} \cdot (1 - \frac{2}{3}),$$

$$\text{Losses}^{2013}_{subprime} = N^{subprime}_3 \cdot H^{subprime.recession}_3 \cdot \text{AVG.EXPECTED.AMNT}^{2013}_{subprime} \cdot (1 - \frac{2}{3}),$$

where $N^{subprime}_3$ is the number of loans issued to subprime borrowers who are in the third year of the loan (conditional on survival), $H^{subprime.recession}_3$ is the hazard for the subprime category in case of an expected recession in the following year, and $\text{AVG.EXPECTED.AMNT}^{2013}_{subprime}$ is the expected average amount (for those who are in the third year) to be paid by the subprime borrowers.[20] [21]

[20] The average outstanding debt for this group equals the average expected amount times $(1 - \frac{2}{3})$ since at this stage of the loan this group has already paid 2/3 of the loan.

[21] LC collects fixed monthly payments on the loan. Part of this payment goes for the repayment of the principal,

Following the same logic, I calculate the $\text{Losses}^{2014}_{prime}$, $\text{Losses}^{2014}_{subprime}$, $\text{Losses}^{2015}_{prime}$, $\text{Losses}^{2015}_{subprime}$.

Figure 9 shows the gross losses for different economic scenarios of 2015 coming from defaults (hereafter, for simplicity, I refer as "losses") for loans originated in 2013, 2014, and 2015 respectively. The losses coming from borrowers who are in the third year of the loan are small compared to losses from borrowers who are in the 2nd and 3rd year of the loan. Moreover, the difference in losses between different economic scenarios increases for loans issued more recently. The change in losses between the scenarios of 2015 reflects the change in profitability for three-year loans issued in this time frame due to macro and regional economic shocks. The losses for the actual, recession and boom economic scenarios look very different in our case. For a boom scenario in 2015, Federal Reserve would had observed a total loss of $271 million in this market, from which $177 million would have been only from the borrowers defaulting in the first year of the loan. If the economy in 2015 would have faced an unexpected downturn, the total losses would have been 37% higher (i.e., $371 million or $100 million higher in monetary term) then the losses in case of a boom.

and the rest is the interest rate to be paid. The monthly payment is calculated by the following formula:

$$mp = \frac{r \cdot P}{n[1 - (1 + \frac{r}{n})^{(-nt)}]},$$

where mp is the amount to be paid each month, P is the principal, r is the interest rate, n is the number of monthly payments within a year, and t is the loan term (3 years). I calculate the losses based on the year the loan is issued (i.e., whether the loan is in the 1st, 2nd, or 3rd year of the maturity condition on surviving the default event) and the borrower type.

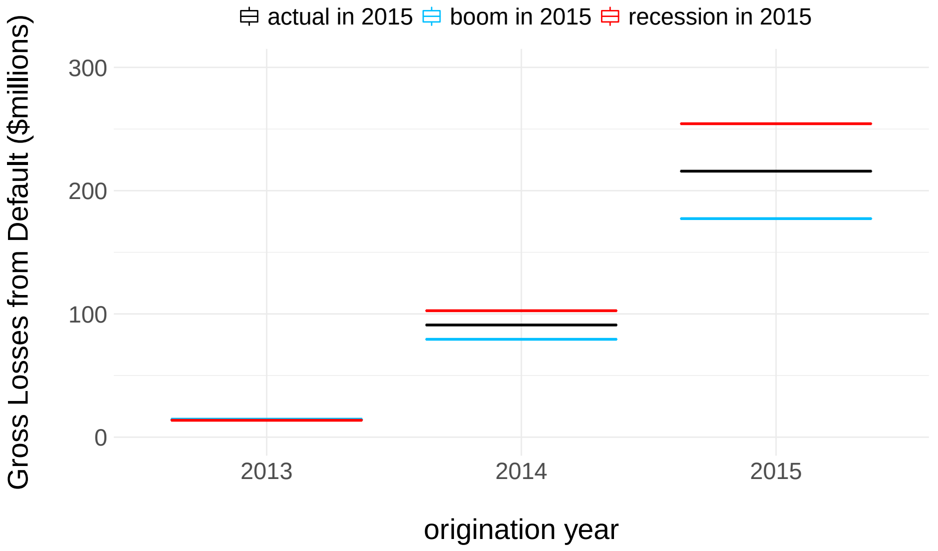

Figure 9: Gross losses for different economic scenarios of 2015 from loan defaults originated in 2013, 2014, and 2015 respectively

Note: This figure displays the default losses for different economic scenarios of 2015 for three-year loans originated in 2013, 2014, and 2015 respectively (e.g., the blue line in 2014 shows the default losses for loans originated in 2014 if there is a boom in 2015). Three different economic scenarios are constructed as follows: The black line shows losses for an economic scenario in 2015 in which macro and regional variables take the average values in the actual dataset. The blue line shows the losses for a "boom" scenario in 2015, which is constructed by increasing the average GDPGR, inflation, and percentage change in HPI by 2 standard deviations and decreasing the unemployment rate by the same magnitude in the actual data. The red line shows the losses for a "recession" scenario in 2015, which I construct by decreasing GDPGR, inflation, and percentage change in HPI by two standard deviations and increase the unemployment rate by the same magnitude.

1.5.2 Losses for Hypothetical FinTech Lending Market Size and Borrower Composition

It is important to realize that losses shown in Figure 9 are the total losses at a specific period in time—in this example losses observed in 2015—for three-year loans only offered by a single lender in this market. The total amount of three-year loans issued between 2013–2015 in my dataset is $6.7 billion and the actual amount issued from the start of the company to 2019:Q4 is $56.8 billion (as shown in Figure 1). So, these are the losses for only about 11.8% of the total amounts issued by this lender. Usually, lenders offer a large variety of products and with different maturities. Also, as the market has been growing it has captured the attention of many other new players, which has become part of the FinTech lending market later. Therefore, the total default losses in this market are much higher than what is reflected in Figure 9.

During the period of my analysis (2009–2015) the number of three-year loans issued by LC has an average annual growth of 103% in loan volumes and 118% in loan amounts, with 56 times more loans issued in 2015 than in 2009 or 72 times more in monetary terms (shown in Figure 2). According to S&P Global Market Intelligence report, the U.S. FinTech lending market size was about $150 billion by 2018:Q2 and is expected to grow to $450 or more billion by 2022. Avention, a venture capitalist, expects this market to reach $1 trillion by 2025 (PricewaterhouseCoopers, 2015). Assuming that all the online lenders operating in this market have the same lending standards as LC, what would the total losses from default be if in this market size is tripled in the future?

The outstanding balance on credit card debt in the United States for 2020 is $756 billion (Stolba, 2020). What if FinTech lending market gets as big as the credit card market, what would the total losses be? Table 10 shows the default losses if the value of outstanding loans in the FinTech lending market is 100 times more than the number of three-year loans issued by LC between 2013–2015 (i.e., if the market size is tripled, which is equivalent to $670 billion) and also if the subprime share doubles over time. In the case of a boom and a recession, the total losses are predicted to be about $27 billion and $37 billion, respectively. The $10 billion difference shows the change in losses due to macro and regional shocks. In addition to the increase in the number of loans originated in this period, I double the subprime share, as the lenders operating in this market increasingly focus on issuing more loans to subprime borrowers. Because subprime borrowers are more vulnerable to macro and regional economic conditions, losses for each economic scenario increases. The increase in the share of subprime borrowers increases the losses for every economic scenario of 2015. For a recession scenario in 2015, the change in borrower composition leads to a 6.2% increase in losses, namely from $37 billion to $39.3 billion. Moreover, due to the change in borrower composition the change in losses between a boom and a recession scenario in 2015 leads to a 10% increase in total losses, from $10 billion that was before to $11 billion.

Table 10: Default losses for hypothetical market volumes and borrower composition

	Total Losses ($ million)		
FinTech lending market size	**actual**	**recession**	**boom**
2013–2015 loan volume	321	371	271
2013–2015 loan volume x 100	32,087	37,055	27,118
2013–2015 loan volume x 100 and 2 x subprime share	33,915	39,349	28,482

Notes: This table shows the total losses if the outstanding loan volumes in the FinTech lending market are 100 times larger than three-year loan volumes originated between 2013–2015 by a single lender and also if the share of subprime borrowers doubles in addition to the increase in the size of the market. Columns under the “Losses ($million)” show the losses in this market for different economic scenarios of 2015. More details on how I construct the economic scenarios are found in Figure 9 and at the beginning of this section.

1.6 Conclusion

The FinTech lending market has seen enormous growth in the last decade and continues with the same momentum. This paper uses a dataset of borrower and loan information from the largest online lender in the United States, together with macro and regional conditions to analyze whether default in this market is sensitive to the business cycle. I find that lenders operating in the FinTech lending market increasingly focus on lending to subprime borrowers. Using a duration model, I estimate the default sensitivity to macro and regional shocks. I find that national-level changes in GDP growth, and inflation, and 3-digit ZIP-code-level changes in the unemployment rate, and percentage changes in HPI play a significant role in default and they should be taken into consideration when assessing credit risk. Also, I find that subprime borrowers are more vulnerable to the business cycle than prime borrowers.

Using the results from the duration model together with the composition of loans in this market I predict the total losses for different economic scenarios. I find that as this market grows, the total losses coming from default may become of macroeconomic importance.

The share of the FinTech lending market remains a small fraction of the consumer lending market, but given its high growth rates, it is only a matter of time for it to gain a significant share in the whole retail lending landscape. In a world where financial institutions are highly interconnected with each other, within and across the borders, it is important that institutional investors hedge against the high default risk in this market. I suggest that Federal Reserve and other institutions responsible for the national financial stability keep an eye on the total losses in this market.

1.7 Appendix A

Figure 10 shows the distribution of personal loans by banks, credit unions, finance companies, and the FinTech lenders. The share of personal loans issued by FinTech lenders increases significantly over time. In 2010 the share was about only 1 percent while in 2018 is about 40 percent.

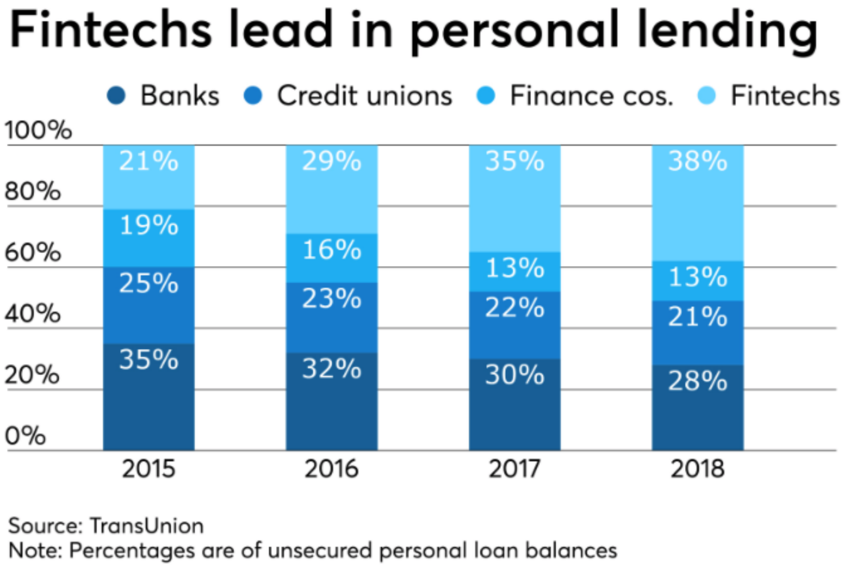

Figure 10: FinTech lenders lead in personal loan lending

Table 11 shows the distribution of loans by purpose based on borrowers' self-reported claims during the application process. About 81 percent of loans are used to help borrowers with credit card repayment and debt consolidation. Lower interest rates offered by these platforms help borrowers refinance their debt in credit cards and other loans. The total loan amounts originated for this purpose is close to $6.4 billion, which counts for about 85% of the total amount issued during our analysis. The rest of the loan purpose categories consumers choose LC to borrow money from are home improvement, small business loans, medical, education, etc. An interesting finding is that two of the most common loan types, car, and student loans, account for only 1.1 and 0.04 percent respectively, which are extremely low to the expectation. This might be because other online platforms, such as Sofi or LendKey and Lendingtree are "specialized" in student and auto loans, respectively; and potentially offer more convenient and better deals for this consumer type. Similarly, entrepreneurs also seem to use different borrowing sources for their business needs as they make only 1.2 percent of the share. Even though different online platforms offer similar products, each one of them focuses more on a specific target of loan type. For instance, if you need a loan for refinancing credit card debt, LC might be the best option. However, if you need a mortgage loan, Prosper might be the optimal online place to go. FinTech lenders keep increasing the variety of their products to make it a one-stop place for personal loans.

Table 11: Loan distribution by loan purpose, originated between Jan 2009 - Dec 2015

loan purpose	**#loans**	**%**	**amount ($ million)**	**%**
debt consolidation	339,258	56.79	4,432.68	59.38
credit card	145,041	24.28	1,944.73	26.05
home improvement	34,424	5.76	404.25	5.42
other	32,336	5.41	264.80	3.55
major purchase	12,399	2.08	116.21	1.56
small business	7,063	1.18	95.14	1.27
car	6,640	1.11	51.48	0.69
medical	6,635	1.11	48.66	0.65
house	2,507	0.42	31.41	0.42
moving	4,428	0.74	29.63	0.40
vacation	4,163	0.70	23.06	0.31
wedding	1,792	0.30	16.88	0.23
renewable energy	449	0.08	3.87	0.05
educational	256	0.04	1.76	0.02
Total	**597,391**	**100.00**	**7,464.55**	**100.00**

Figure 11 shows the default rates for different loan purposes issued by LC. Small business loans have the highest default rates, about 23 percent, closely followed by loans issued for renewable energy, and moving expenses. Default rate for refinancing purposes such as credit card and other debt obligations have a default rate of 12 percent and 15 percent, respectively.

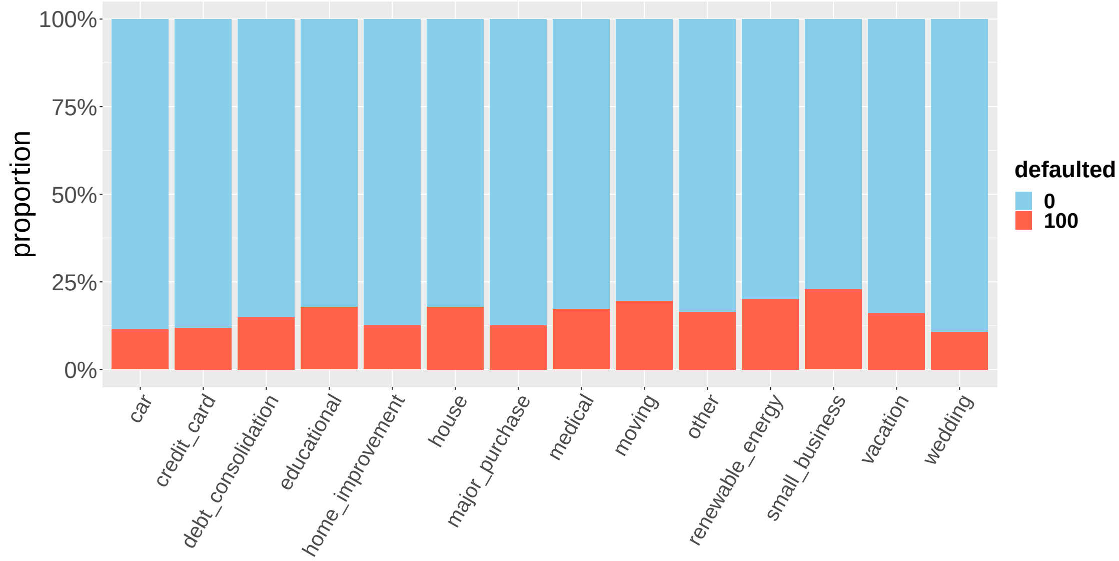

Figure 11: Default share by loan purpose

Table 12: Correlations for individual and loan attributes

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
(1)	default	1												
(2)	fico	-0.12	1											
(3)	int_rate	0.19	-0.49	1										
(4)	annual_inc	-0.05	0.09	-0.12	1									
(5)	loan_amnt	-0.02	0.14	-0.09	0.34	1								
(6)	dti	0.08	-0.09	0.14	-0.17	0	1							
(7)	delinq_2yrs	0.02	-0.17	0.06	0.05	0	0	1						
(8)	inq_last_6mths	0.07	-0.09	0.26	0.03	-0.03	-0.01	0.03	1					
(9)	open_acc	0.01	0.03	-0.07	0.14	0.19	0.30	0.06	0.12	1				
(10)	revol_util	0.05	-0.45	0.27	0.03	0.09	0.17	-0.01	-0.08	-0.15	1			
(11)	total_acc	-0.02	0.03	-0.11	0.19	0.21	0.21	0.13	0.14	0.69	-0.12	1		
(12)	acc_now_delinq	0	-0.04	0.03	0.01	0	0.01	0.13	0	0.02	-0.03	0.03	1	
(13)	chargeoff_within_12_mths	0	-0.05	0.01	0.01	0	0	0.14	0.01	0.01	-0.02	0.05	0.05	1

Table 13: Correlations for macro and regional variables (standardized values)

	GDPGR_Y1	INF_Y1	UNEMPR_Y1	HPI_Y1
GDPGR_Y1	1.00			
INF_Y1	0.47	1.00		
UNEMPR_Y1	-0.35	-0.12	1.00	
HPI_Y1	0.30	-0.19	-0.39	1.00

Table 14: Non-standardized Regression Results (overall default)

	defaulted		
	(1)	(2)	(3)
constant	63.50***(1.79)	15.44***(0.22)	61.32***(1.90)
FICO	-0.05***(0.002)		-0.04***(0.002)
int_rate	1.31***(0.01)		1.40***(0.02)
loan_amnt (log)	0.01***(0.00)		0.01***(0.00)
home_ownership.fixrent	2.40***(0.09)		2.51***(0.10)
annual_inc (log)	-3.45***(0.11)		-3.49***(0.11)
verification_status.fixVerified	1.56***(0.10)		1.26***(0.10)
claim_creditworthness	-1.89***(0.38)		-1.42***(0.41)
dti	0.19***(0.01)		0.16***(0.01)
delinq_2yrs	0.25***(0.05)		0.23***(0.06)
inq_last_6mths	1.06***(0.05)		1.13***(0.05)
open_acc	0.20***(0.01)		0.18***(0.01)
revol_util	-0.03***(0.002)		-0.02***(0.002)
total_acc	-0.07***(0.01)		-0.06***(0.01)
acc_now_delinq	-0.83(0.58)		-1.00*(0.61)
chargeoff_within_12_mths	-0.49(0.41)		-0.68(0.43)
GDPGR_Y1		-0.20*(0.12)	-1.25***(0.12)
INF_Y1		-1.13***(0.07)	-1.48***(0.07)
UNEMPR_Y1		-0.59***(0.08)	-0.58***(0.08)
HPI_Y1		-0.05***(0.01)	-0.10***(0.01)
Observations	597,391	550,146	550,146
Adjusted R^2	0.05	0.001	0.05
Residual Std. Error	34.11 (df = 597375)	34.90 (df = 550141)	34.05 (df = 550126)
F Statistic	1,948.60***(df = 15; 597375)	109.04***(df = 4; 550141)	1,484.83***(df = 19; 550126)

Notes: This table shows the non-standardized estimated results for the overall default linear probability model. Note:∗p<0.1; ∗∗p<0.05; ∗ ∗ ∗p<0.01

1.8 Appendix B

1.8.1 Principal Component Analysis and Interaction Terms

The main idea of PCA is to reduce the dimensionality of a dataset, possibly with many variables correlated to each other. If you apply PCA to X number of observed variables, it will give you exactly X independent latent variables, which are called principal components. The latent values in a principal component are called component scores or factor scores and the weights by which each standardized original variable should be multiplied to get the component score are called *loadings*. Moreover, PCA gives the amount of the variation captured by each principal component. PCA applies an orthogonal transformation to the observed variables and creates independent latent variables called principal components (PC). Based on how much of the variation in the data is captured by each PC, we determine whether to use them or not. The first PC on our time-frame of macro variables does not capture much of the variation from Real GDP growth and inflation. Therefore, I decided not to use it. However, the first PC (PC1) for regional variables, unemployment rate, and house price index, captures about 74% of the variation and would be a prime decision to include it in the regressions. The following three regression equations will show us whether subprime borrowers are more sensitive to the business cycle and by how much is this difference.

$$\begin{aligned} defaulted\_1stY_{irt} = {} & \beta_0 + \beta_1 subprime + \beta_3 \mathrm{GDPGR}_t + \beta_4 INF_t + \beta_5 regionPC1_t \\ & + \beta_6 \mathrm{GDPGR}_t \cdot \mathrm{subprime}_i + \beta_7 \mathrm{INF}_t \cdot \mathrm{subprime}_i \\ & + \beta_8 regionPC1_t \cdot \mathrm{subprime}_i + \epsilon_{irt} \end{aligned}$$

$$\begin{aligned} defaulted\_2ndY_{irt} = {} & \beta_0 + \beta_1 subprime + \beta_3 \mathrm{GDPGR}_{t+1} + \beta_4 INF_{t+1} + \beta_5 regionPC1_{t+1} \\ & + \beta_6 \mathrm{GDPGR}_{t+1} \cdot \mathrm{subprime}_i + \beta_7 \mathrm{INF}_{t+1} \cdot \mathrm{subprime}_i \\ & + \beta_8 regionPC1_{t+1} \cdot \mathrm{subprime}_i + \epsilon_{irt} \end{aligned}$$

$$\begin{aligned} defaulted\_3rdY_{irt} = {} & \beta_0 + \beta_1 subprime + \beta_3 \mathrm{GDPGR}_{t+2} + \beta_4 INF_{t+2} + \beta_5 regionPC1_{t+2} \\ & + \beta_6 \mathrm{GDPGR}_{t+2} \cdot \mathrm{subprime}_i + \beta_7 \mathrm{INF}_{t+2} \cdot \mathrm{subprime}_i \\ & + \beta_8 regionPC1_{t+2} \cdot \mathrm{subprime}_i + \epsilon_{irt} \end{aligned}$$

Estimation results for the discrete hazard model with regional PCA and interaction terms are shown in Table 15.

Table 15: Estimation results for the discrete hazard model with regional PCA and interaction terms

	defaulted_1stY	**defaulted_2ndY**	**defaulted_3rdY**
constant	4.83***(0.05)	5.87***(0.07)	3.16***(0.06)
subprime1	2.99***(0.13)	2.96***(0.17)	1.72***(0.14)
GDPGR_Y1	-0.22***(0.04)		
GDPGR_Y2		-0.07*(0.04)	
GDPGR_Y3			0.10**(0.04)
INF_Y1	-0.23***(0.04)		
INF_Y2		-0.17***(0.04)	
INF_Y3			0.04(0.04)
region.PC1_Y1	0.07(0.05)		
region.PC1_Y2		-0.21***(0.06)	
region.PC1_Y3			-0.17***(0.06)
GDPGR_Y1*subprime1	-0.55***(0.12)		
GDPGR_Y2 *subprime1		-0.66***(0.12)	
GDPGR_Y3 *subprime1			-0.10(0.12)
INF_Y1*subprime1	0.25**(0.11)		
INF_Y2 *subprime1		0.72***(0.11)	
INF_Y3 *subprime1			0.21*(0.12)
region.PC1_Y1*subprime1	-0.25**(0.12)		
region.PC1_Y2*subprime1		0.24(0.17)	
region.PC1_Y3*subprime1			-0.17(0.15)
Observations	550,146	520,579	488,053
Adjusted R^2	0.003	0.003	0.001
Residual Std. Error	22.52 (df = 550138)	24.17 (df = 520571)	17.83 (df = 488045)
F Statistic	201.85***(df = 7; 550138)	199.89***(df = 7; 520571)	86.61***(df = 7; 488045)

Note:$*$p<0.1; $**$p<0.05; $***$p<0.01

CHAPTER 2

2 Modeling Default Risk in the FinTech Lending Market: A Machine Learning Approach

2.1 Introduction

The consumer lending market in the United States has seen enormous growth in the last decades with a current outstanding credit of \$4.2 trillion (Reserve, 2021). Unsecured personal lending has experienced a greater growth rate than any other type of loans including mortgages, auto loans, credit cards, and student loans (Beiseitov, 2019). The growth in this segment of consumer lending has boomed mainly because of a new type of player known as FinTech, or marketplace lending.

Lenders operating in this market use technological advancements in data processing and credit risk assessment tools such as machine learning (ML) and artificial intelligence (AI) to evaluate the credit riskiness of prospective borrowers. Minutes after submitting an online application, borrowers get an offer specifying the interest rate, and in case of acceptance, they receive the loan within a few days. The FinTech lending market has seen exponential growth in the last decade and continues to do so. According to TransUnion, FinTech's share of personal loan originations has grown from 1 percent in 2010 to nearly 40 percent in 2018.[22]

Default rates are relatively high in this market compared to the default rates of loans issued by banks. Figure 12 shows the default rate across vintages for loans issued by FinTech vs. banks.

[22] Retrieved November 10, 2020 from https://www.americanbanker.com/news/once-dismissive-of-fintechs-traditional-lenders-now-feeling-their-bite.

FinTech lenders claim that they can better screen borrowers because of the complex credit screening algorithms and the alternative data they use when evaluating the creditworthiness of potential borrowers. Understanding the factors that play a role in capturing default becomes key to loan pricing and credit allocation.

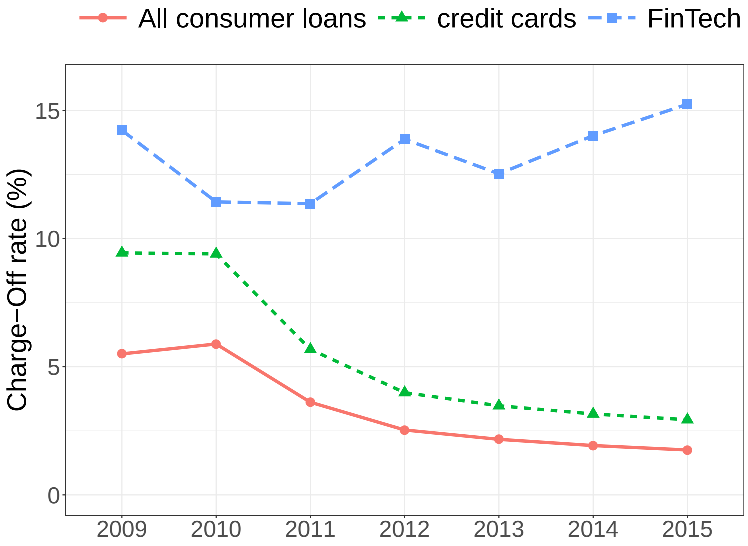

Figure 12: LendingClub vs bank consumer default rates

Source: Federal Reserve, LC, and author's calculations.

Note: This figure shows the default rates of loans issued in the FinTech lending market (the blue line) relative to default rates of consumer and credit cards issued by banks (red and green line, respectively).

This study aims to provide an overview of some of the most common ML methods used in modeling default risk and assesses to what extent these models are better than traditional econometrics approaches. Using a large dataset of three-year matured loans issued from the largest FinTech lender in the United States during the time interval of January 2009−December 2015, together with other macro and regional economic variables, this research sheds light on the determinant factors of default for this market. I compare the performance of four different ML methods to the traditional logistic regression results. ML methods used in this study are ideally suited for loan-level analysis and out-of-sample prediction due to the large sample size and for possible non-linear relationships among features that can not be captured by traditional approaches. Methods used to model and predict default are: Logistic Regression (LR), Decision Trees (DT), Random Forests (RF), eXtreme Gradient Boosting (XGB), and Artificial Neural Network (ANN). An oversampling technique known as Synthetic Minority Oversampling Technique (SMOTE) is implemented in order to balance the ratio between default and non-default classes. The imbalance between classes can yield low accuracy rate in the prediction results, especially in LR and tree-based models (King and Zeng, 2001; Cieslak and Chawla, 2008). The comparison between models forecast performance is done using Area Under the ROC Curve (AUC-ROC), which measures how well the model is capable to distinguish between default and non-default classes. This measure of fit is evaluated on a different sample (known as the testing set) of the one that the models are estimated on.

I find that some of the machine learning algorithms such as XGB and ANN outperform logistic regression, but the improvement is marginal. Annual income, loan purpose, revolving line utilization, and interest rate are the variables with the highest importance predicting default. Most of the macro and regional variables are listed among the top 10 variables in predicting default. Implementing SMOTE increased the model performance for some of the methods.

The rest of this paper is organized in the following order. Section 2 summarizes the current literature and their findings. Section 3 explains the theory behind the ML methods used in this study. Section 4 gives information on the performance metrics and why these metrics are important. Section 5 shows the empirical results and a comparison between methods. Section 6 offers some concluding remarks.

2.2 Literature Review

The usage of ML methods has started to get significant attention in the field of applied economics and finance.[23] Many research studies use these methods to examine the benefits and drawbacks

[23] For a broader discussion of the impact of ML on the field of economics see Athey (2018).

these methods might have when applied to different problems. Moreover, some studies have shown that these methods complement econometrics methods such as logit estimation and we can answer questions that we were not able to answer before.

For example, Björkegren and Grissen (2018) try to predict loan repayments for the unbanked in developing countries using ML methods. Because of no formal financial histories, traditional credit bureau models can not be implemented to evaluate the creditworthiness of potential borrowers. They show that by applying ML methods lenders can predict default for the unbanked using borrowers' behavioral signatures from mobile phone data. Albanesi and Vamossy (2019) use multiple ML models to predict consumer default, using the same information as traditional credit scoring models. They show that ML methods have much stronger performance than logistic regression. They find that the number of trades and the balance on outstanding loans are the most important factors predicting default, in addition to outstanding delinquencies and length of the credit history. Bagherpour (2017) uses different ML methods to model and forecast mortgage loan losses in the United States. He finds that machine learning models forecast performance are substantially better than traditional logistic regressions.

There are few papers that explore the performance of loans in the FinTech lending market, and they mainly use traditional econometrics approaches to explore the determinants of default. Emekter et al. (2015) uses a logit model and found that credit grade, debt-to-income ratio, FICO score, and revolving line utilization have an important role in loan defaults. Moreover, they state that the interest rates charged on the high-risk borrowers are not sufficiently large to compensate for the higher probability of default. So they suggest that LendingClub (LC) must find new ways to attract high FICO scores and high-income borrowers in order to sustain their businesses. Serrano-Cinca, Gutierrez-Nieto and López-Palacios (2015) use univariate means tests and survival analysis to analyze the LC data from $2008 - 2014$. They found that factors that best explain default are loan purpose, annual income, current housing situation, credit history, and indebtedness. Also by using logistic regression model, they found that the variable grade assigned by LC is the most predictive factor of default but it can be improved by adding other information. Carmichael (2014) uses a dynamic logistic regression where he combines both extensive credit information and soft data. He found that FICO score, borrower-initiated credit inquiries, income, and borrower self-stated loan purpose are the most significant variables for explaining default.

The closest research to this paper is the studies by Wang and Perkins (2019) and Turiel and Aste (2020). Turiel and Aste (2020) applies LR, SVM, and ANN to predicting default in the FinTech lending market for loans issued by LC between 2007 and 2017. They find that ANN outperforms the other methods but with marginal improvements on the LR. However, they do not incorporate the macro and regional variables, which are important variables that explain default behavior (Azizaj, 2020). Wang and Perkins (2019) use only RF and XGB (i.e., stochastic gradient boosting) to predict default in the FinTech lending market. They include unemployment rate and house price growth and found that especially the local unemployment rate is one of the most important factors predicting default. Different from the current literature, in this study I provide a theoretical overview of the ML methods suitable for modeling default risk. Moreover, in addition to loan and individual attributes, I use macro and regional economic variables that the literature have found important factors of default, and apply a larger variety of the ML methods to predict default behavior in the FinTech lending market.

2.3 Machine Learning Classification Models

ML methods have started being used widely in every industry, and the financial sector is not an exception. Predicting default is a very important task for lenders to help them with the loan pricing and mitigating potential losses. Lenders charge higher interest rates for riskier borrowers and lower rates for borrowers less likely to default.

In this paper, I will focus on some of the most common ML methods used in the lending industry. The problem in our case is a classification problem, where the output is a discreet binary variable with two classes: *default* and *non-default*. The objective for any ML model is to find a function that learns from the information available on a dataset (hereafter training set) and do a good job in forecasting out-of-sample outcomes (hereafter testing set).

The main difference between econometrics and ML methods is that in the latter it is impossible to directly quantify individual coefficients, i.e., there are no estimated coefficients as in regression models. However, to gain intuition on the importance of each variable used in predicting default, I use the variable importance measure provided by each ML method.[24]

2.3.1 Logistic regression

LR is the common method to deal with binary target variables. It is a parametric method, that considers linear relationships between inputs and the log of the odds ratio. This model returns probabilities of default for each loan issued. This research paper uses LR as the benchmark model to predict default in the FinTech lending market. Let Y be a binary variable that takes the value one if the loan has been classified as default and zero otherwise. If k explanatory variables used to predict default are $X = (x_1, x_2, x_3, ..., x_k)$ then the probability of defaulting for a loan given X is given by the logit function (a.k.a., sign function) and it is

$$p = E(Y = 1|X) = \frac{exp(\beta_0 + \sum_{i=1}^{k} \beta_i x_i)}{1 + exp(\beta_0 + \sum_{i=1}^{k} \beta_i x_i)} . \tag{9}$$

After taking the log of the odds-ratio we have

$$log(\frac{p}{1-p}) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i . \tag{10}$$

Equation 10 is a linear combination of independent variables. LR models are generally fit by

[24] I will provide more details on this matter later in the paper.

maximum likelihood, which gives us the estimated parameters $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k)$. Using these parameters estimated from the training set, I can make predictions on the testing set (i.e., out-of-sample forecast) which will give me probabilities of default for each sample. Then using a threshold we decide to which class each sample belongs. For example, using a threshold of $c = 0.5$, the decision about which class a certain loan belongs is as follows:

$$\hat{y}_i = \begin{cases} 1, & \text{if } P(Y_i = 1|X_i; \hat{\beta}) \geq 0.5 \\ 0, & \text{if } P(Y_i = 1|X_i; \hat{\beta}) < 0.5 \end{cases}, \tag{11}$$

where $\hat{y}_i$ is the forecast for borrower i in the testing set. However, choosing which threshold value to use to make the classifications will play an important role in the model performance because it allows us to trade precision for sensitivity (terms that I will define later in the paper).

2.3.2 Decision Trees and Random Forest

Decision trees are one of the simplest and easy to interpret machine learning algorithms in classification problems. The idea behind the algorithm is similar to an approach of asking a series of questions and each time we have an answer, a follow-up question is asked until we reach a conclusion about the classification problem. A very simple decision tree classifier example is shown in Figure 13, where we are trying to classify a potential borrower whether s/he will default on a loan or not. We have two classes: *default* and *non-default*; and for simplicity I have chosen only two variables: FICO and DTI. The algorithm finds cut-off points for each variable and based on these cut-offs makes the final decision. If the borrower has a $FICO < 670$ the classifier predicts that the borrower will default; If the borrower has a $FICO \geq 670$ and $DTI < 30$ the classifier assigns *non-default* and so on.

A decision tree has a hierarchical structure built top-down which consists of nodes (aka regions) and directed edges. Every tree has three types of nodes:

- A *root node* that has no incoming edges but one or more outgoing edges.
- *Internal nodes* which have one incoming edge and one or more outgoing edges.
- *Terminal nodes* or *leaf nodes* that have exactly one incoming edge and no outgoing edges.

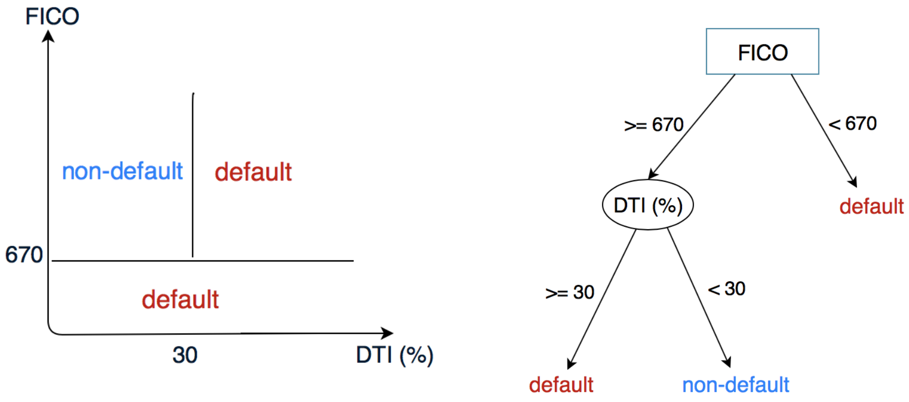

Figure 13: Classification tree example

The tree classifier's goal is to choose the variable that bests separate the classes and make it the root node (i.e., put it at the top of the tree) followed by other branches (i.e., internal nodes) until it reaches a terminal node with no further links. In decision tree classification problems every internal node represents a variable and every terminal node represents a class or output. Decision trees are built in a way that partition the data into subsets that contain observations with similar values. An important question would be how does the algorithm split the data and selects which variable to put at each node, and how does it decide on the test conditions at each node? Many measures have been developed to find ways how to best split the data into smaller subsets with more homogeneous classes, i.e., purer subset. These measures mainly are defined in terms of the classes for each observation before and after the split by considering the information gain after each split.[25] The variable with higher information gain (IG) splits the data into subsets with more homogeneous classes. Let $P_{k,m}$ be the proportion of observations belonging to class k at region (i.e., node) m and $I(\cdot)$ be the indicator function. Then mathematically we can write $P_{k,m}$ as:

$$P_{k,m} = \frac{1}{N_m} \sum_{i \in m} I(y_i = k)\,, \tag{12}$$

where N_m is the total number of observations at node m. The algorithm will classify the observations in node m to class k (which is the majority class in node m) because $k(m) = \arg\max_k P_{k,m}$. The measures developed for selecting the best split mainly consider the degree of impurity of the data at that particular node. Different measures of node impurity include:

$$\text{Entropy}(m) = -\sum_{k=1}^{K} P_{k,m} log_2 P_{k,m}\,, \tag{13}$$

$$\text{Gini Index}(m) = 1 - \sum_{k=1}^{K} P^2_{k,m}\,, \tag{14}$$

$$\text{Misclassification error}(m) = 1 - \underset{k}{max}(P_{k,m})\,. \tag{15}$$

For binary classification problems, as in our case, if p_0 is the proportion of observations that belong to the *non-default* class in a particular node, then $p_1 = 1 - p_0$ is the fraction of the observations that have defaulted. The values for the above impurity measures for this particular node are:

$$\text{Entropy} = -p_0 log_2 p_0 - p_1 log_2 p_1\,,$$

$$\text{Gini} = 1 - p_0^2 - p_1^2\,,$$

$$\text{Misclassification error} = 1 - max[p_0, p_1]\,.$$

Figure 14 shows the comparison among the impurity measures for binary classification models.

[25] I follow closely the book from Tan, Steinbach and Kumar (2016) and Hastie, Tibshirani and Friedman (2009) regarding the theory for machine learning methods used in this study.

Misclassification error is not sufficiently sensitive for tree growing, so entropy or Gini index measures are more preferable. Moreover, Entropy and Gini index functions are differentiable, therefore more amenable for numerical optimization (Hastie, Tibshirani and Friedman, 2009). This study uses the Gini index as the criteria for splitting. A value of $G = 0$ means the subset data is pure therefore no further splitting can be done since all the observations in this subset have the same class. If $G = 0.5$ then the subset data is at its highest level of impurity, meaning 50 percent of the observations have a *default* class and the other 50 percent *non-default* and therefore requires further splitting.

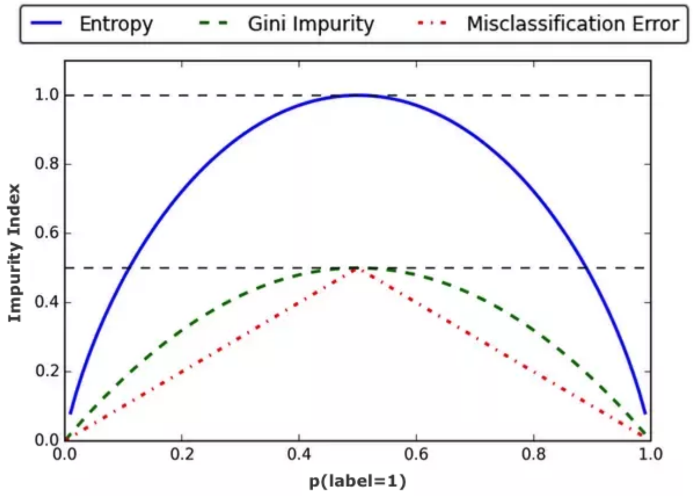

Figure 14: Impurity index for for different impurity measures.
Source: Medium website

Note: This figure shows the comparison among the impurity measures such as Entropy, Gini Impurity, and Misclassification Error for binary classification problems.

This research paper applies a classification and regression tree (CART) induction algorithm with only binary splits at each node, no matter the type of the variable. Consider an extremely simplified example of our dataset where we have 20 observations and three variables: income, home-ownership, and FICO. Based on the distribution of classes for each variable we want to calculate which variable is going to be the one that bests split the dataset into subsets with the lowest impurity. In each of the following tables in Figure 15, there are eight cases of default and twelve of non-default. After calculating the Gini index at each possible node, we calculate the average weighted Gini for each variable. The variable with the lowest weighted average Gini index (i.e., FICO in this case) is the root node as it best splits the data into purest subsets. In the same way, we find the variables in the internal nodes until we reach a node with no further possible splits which means it is a terminal node as all the observations in that node have the same class. But how does the decision tree classifier decide on the cut-off points? Based on the type of the variable the algorithm calculates the average weighted Gini indexes. An example for variables with two categories is shown in Figure 15 (home-ownership). If the categorical variable has more than two classes, then the algorithm will try all possible combinations of classes for binary splitting and will choose the one with the lowest average weighted Gini index.[26] For continuous variables, the test condition is expressed as a comparison test (FICO$\geq v$) or (FICO$< v$) where v is the cut-off point. FICO variable has a minimum value of 660 and a max value of 850 in this dataset. Computing all possible cut-off points is computationally infeasible. Therefore, the model uses only the unique values of FICO in the dataset and estimates the Gini indexes for average values between two adjacent values. Let FICO $\in \{660, 680, 726, 800\}$ in the example with twenty observations. Instead of considering each possible value between 660 and 800 when finding the best cut-off point, the algorithm will consider only the average values between the existing values in the dataset, that is $v \in \{670, 703, 763\}$. The algorithm will choose the cut-off point with the smallest average weighted Gini indexes.

[26] For ordinal variables the model preserves the order property of the values. If ordinal values are {Small, Medium, Large}, the possible combinations for binary splitting are: ({Small} and {Medium, Large}) or ({Small, Medium} and {Large}).

Income

	≥80K	<80K
default	3	5
non-default	8	4
Gini index	0.397	0.494

Gini = $(\frac{11}{20}) \cdot 0.397 + (\frac{9}{20}) \cdot 0.494 = 0.441$

Homeownership

	Rent	Own
default	6	2
non-default	3	9
Gini index	0.444	0.298

Gini = $(\frac{9}{20}) \cdot 0.444 + (\frac{11}{20}) \cdot 0.298 = 0.364$

FICO

	≤670	>670
default	7	1
non-default	0	12
Gini index	0	0.142

Gini = $(\frac{7}{20}) \cdot 0 + (\frac{13}{20}) \cdot 0.142 = 0.092$

Figure 15: Example of calculating Gini impurity.

Besides the advantage of easy interpretability, decision trees consider the non-linear connection between the target and input variables, a feature that is not in logistic regression. Moreover, this method considers interactions between explanatory variables. However, as in every other model decision trees have their cons. In general decision trees have low accuracy in prediction tasks and high variance. This is because they are very sensitive to the training data. In other words, a slightly different training set may result in a very different decision tree structure. Moreover, decision trees are prone to overfitting problems. The model performs well on the training set but has a poor prediction performance when it comes to unseen data. That being the case, decision tree classifiers have low bias and high variance. As a result, using a single tree to predict the outcome is not a good idea.

In order to avoid the high variance problem in tree classifiers, one can bootstrap the training set, find classifier predictions in each bootstrapped sample and take the average of the predictions.[27] This approach is known as *bagging* or bootstrap aggregation in the machine learning literature and is an ensemble learning method.[28] Even though the bagging method reduces variance and increases prediction, it considers all the variables when splitting. Therefore, it will put at the top of each grown tree the best predictor variables which tend to be the same if we consider all the variables, which will lead to correlated trees. RF method solves this problem by aggregating bootstrapped samples (exactly as bagging) but also uses different subsets of available features for each grown tree. In simple terms, RF is an ensemble learning method that grows a lot of trees and uses different subsets of variables for each grown tree. It assigns uniform probabilities to all variables and chooses only t of them where $t < T$ and T is the total number of variables.[29] Figure 16 shows an example of RF with three trees and $t = 2$. Higher the number of trees in the forests higher the robustness of the prediction, therefore higher accuracy.

[27] Bootstrapping is a random sampling with replacement technique; new simulated samples have the same size as the original dataset.

[28] Ensemble learning combines several ML models with the aim of enhancing the performance.

[29] The parameter t or *mtry* is a tuning specification and we play with different values of it to get the model that best predict the variable of interest. Moreover, you decide on how many trees you want to grow. More is better but at the computational cost.

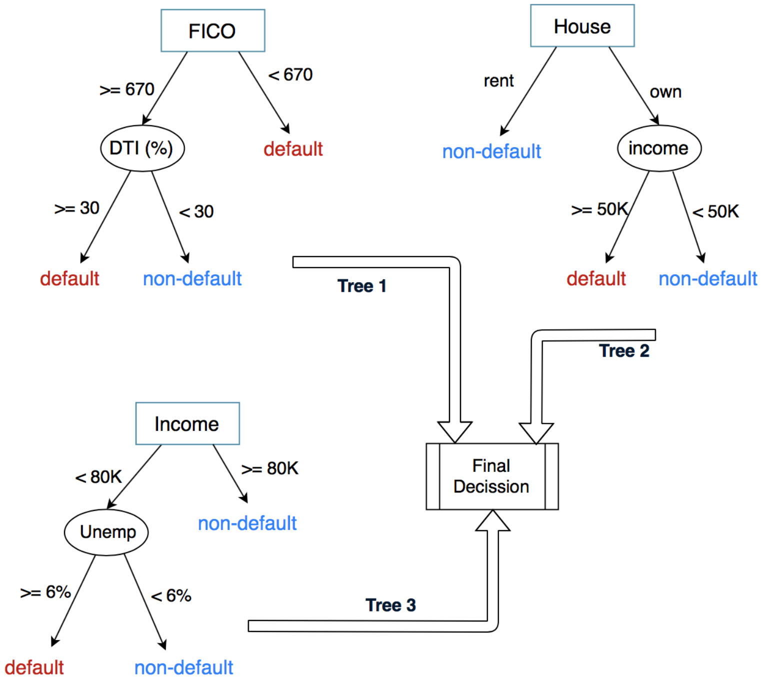

Figure 16: Random Forest example with $mtry = 2$

After the trees are constructed, the classification decision is based on votes coming from each grown tree. The majority votes decide on the final classification of the target variable. The advantages of using the RF method are that: it is very good in dealing with the overfitting problem, has high accuracy, runs efficiently on large datasets, can maintain high accuracy on datasets with a high proportion of missing data and it is computationally feasible. A major disadvantage of RF is that it is like a black box where there is very little control on the model, and also not easy for interpretation.

2.3.3 Boosting Methods

Similar to bagging or RF, boosting methods are ensemble learning methods that combine "weak" and shallow trees to form a more powerful one.[30] There are three main boosted methods: Adaptive

[30] A weak tree classifier is a tree that performs slightly better than random guessing. Most of the times a weak tree is a tree of depth one (i.e., one node and two leaves), also known as a stump.

Boosting or AdaBoost, Gradient Boosted Method (GBM), and XGB.

AdaBoost combines multiple weak trees into a strong one. In AdaBoost, all the trees are generated sequentially and have the same depth (generally stumps). The next tree is fitted by taking the previous tree's error into account. The idea is to create an iterative process $m = \{1, 2, 3, ..., M\}$ with M trees where each one is trained on a weighted version of the dataset. In each iteration, we modify the importance of each observation i in the data (i.e., training set) and give a weight parameter w_i. For each iteration we run the model, get the results, and update the weights of each observation. We give higher values to the observations which are misclassified and lower to those which were correctly classified. This way the classifier growing in the next iteration will focus on classifying correctly those observations which were misclassified in the previous iteration. This will lead to improvement in the overall model prediction. In contrast to RF where all the trees have equal weight on the final decision, in AdaBoost the final decision for the prediction will be based on a weighted majority vote, therefore some trees will get more say in the final decision than others.

GBM is similar to AdaBoost but with some minor differences. In GBM also the trees are generated sequentially but in a way that the next weak tree is more effective than the previous one (this was not the case in Ada Boost). Therefore, the overall model will improve sequentially with each iteration. The size of the tree in GBM is larger than in AdaBoost (generally between 8 and 32). Another difference to the Ada Boost is that the weights for each observation are not updated. GBM optimizes the loss function of the previous tree and tries to minimize the prediction errors by adding an adaptive model that adds weak learners. [31] The main idea is to reduce the errors in the previous learners' prediction. Also, a regularization strategy such as scaling the contribution of each tree (i.e., regularization by shrinkage) by a factor $0 < \nu < 1$ is used. This parameter is the learning rate of the boosting procedure. The generic algorithm is shown in Table 16.

[31] This study uses a cross entropy loss function $L(y_i, f(x_i)) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(f(x_i)) + (1 - y_1) \cdot log(1 - f(x_i))$.

The three main components of the GBM are loss function, weak learners, and an additive model which will regularize the loss function. While GBM often gives very good predictive results, one of the main disadvantages is that it is computationally very expensive (takes a long time to train the model).

Table 16: Generic GBM algorithm for classification problems

```
GBM Algorithm
1. Initiate the the optimal initial prediction $f_0(x) = \underset{\gamma}{argmin} \sum_{i=1}^{N} L(y_i, \gamma)$
2. For m = 1 to M:
   (a) For i = 1, 2, . . . , N compute residuals
       $$r_{im} = \left[ \frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f=f_{m-1}}$$
   (b) Fit a regression tree to the $r_{im}$ values and create terminal regions $R_{jm}$, for $j = 1, 2, ..., J_m$.
   (c) For $j = 1, 2, ..., J_m$ compute
       $$\gamma_{jm} = \underset{\gamma}{argmin} \sum_{x_i \in R_{jm}} L(y, \gamma)$$
   (d) Update $f_m(x) = f_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$
3. Output $\hat{f}(x) = f_M(x)$
```

XGB is an advanced version of GBM, where the focus is on enhancing the computational speed and the model performance. The main reason why this method was invented is that GBM is extremely slow. Different from the GBM, XGB creates weak trees in parallel, rather than sequentially as in GBM. It implements distributed computing methods, out-of-core computing, and cache optimizations in order to achieve faster and better results.[32]

2.3.4 Artificial Neural Network

ANN is another supervised learning algorithm that returns probabilities in binary classification problems. It is one of the most powerful statistical learning algorithms. This non-parametric model is closely related to how human brain neurons work. It learns to analyze and processes information and it has self-learning capabilities that produce better results as more data become

[32] For more detailed information on how this is achieved refer to Hastie, Tibshirani and Friedman (2009).

available. Training the model for ANN means to find the weights between neurons (i.e., input nodes) and other nodes in a hidden or output layer.

The simplest model of an ANN is a perceptron model (single layer) also known as the linear binary classifier. Figure 17 shows the structure of a perceptron model. In this simplified version we have five features and they form the input layer $L1$. The output node is a linear combination of input variables and their respective weight linkages. Higher weight values are an indication of strong relationship between the input and the target variable.

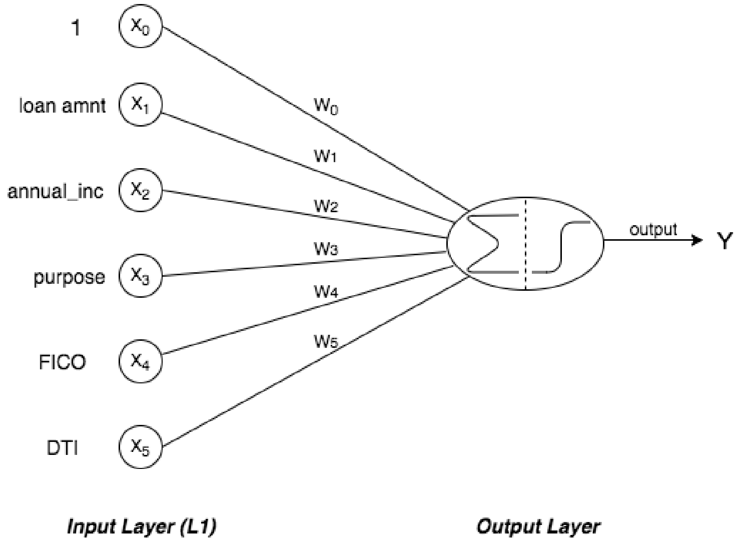

Figure 17: Perceptron structure

The perceptron computes the output value $\hat{y}$ by performing a weighted sum on its inputs and subtracting a bias factor $w_0 = -t$. If this value is greater than 0, then the model predicts a class of 1, and -1 otherwise.

$$\hat{y}_i = sign(\mathbf{w} \cdot \mathbf{x}_i - t) = \frac{1}{1 + e^{(\mathbf{w} \cdot \mathbf{x}_i - t)}} = \begin{cases} 1, & \text{if } \mathbf{w} \cdot \mathbf{x}_i - t > 0 \\ -1, & \text{if } \mathbf{w} \cdot \mathbf{x}_i - t < 0 \end{cases}, \tag{16}$$

where $\mathbf{w} = (w_1, w_2, w_3, ..., w_n)$ is a vector of weights, $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}, ..., x_{in})$ is the feature space for observation i and the *sign(.)* function or the logistic function acts as the activation function in the output node. By so far you must have realized a perceptron or an ANN with no hidden layer is a logistic regression. Training a perceptron means finding weights that reflect the input-output relationship of the underlying data. The key computation for this case is the weight update formula

$$w_j^{(k+1)} = w_j^{(k)} - \lambda(y_i - \hat{y}_i^{(k)})x_{ij}, \tag{17}$$

where $w_j^{(k)}$ is the weight parameter for variable j at the k^{th} iteration, λ is the learning rate parameter[33] with values between 0 and 1 and x_{ij} is the value of variable j for observation i. We put initial guest for the weight vector $\mathbf{w}$ and this will adjust the weights until the error term $(y_i - \hat{y}_i^{(k)})$ is less than the threshold or max number of iteration has been reached. For linearly separable classification problems, the algorithm guarantees an optimal solution conditioning in choosing small values for the learning rate. If the problem is not linearly separable, then the algorithm fails to converge and we should be looking for other possible solutions.

In cases of non-linearly separable problems, we use a more complex model than that of a perceptron. A model with extra layers (a.k.a., intermediate or hidden layers) in addition to the input and output ones is a model known as a multi-layer artificial neural network (MANN).[34] The goal of MANN is to find a weight vector $\mathbf{w}$ that minimizes the total sum of squared errors.

$$E(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \text{ for } i = 1, 2, 3, ..., N \quad , \tag{18}$$

where $\hat{y}_i = \mathbf{w} \cdot \boldsymbol{x}_i - t$ is the linear combination of weights and attributes for observation i. Notice that now the weight parameters in equation (18) are quadratic and therefore a global minimum exists. In cases we use other activation functions such as *sigmoid* or *tanh* then the algorithm applies a gradient descent method to find the global minima. The weight update formula for this case would

[33] The learning rate parameter controls the amount of adjustments made in each iteration.

[34] If the model has more than two hidden layers it is known as a Deep Learning (DL) model. However, for computational simplicity, I will use only one hidden layer in this paper.

be:

$$w_j^{(k+1)} = w_j^{(k)} - \lambda \frac{\partial E(\mathbf{w})}{\partial w_j}. \tag{19}$$

For non-linear functions, the gradient descent method does not always guarantee global minima. It is possible it can be trapped in a local minimum. The gradient descent is used to learn the output and hidden nodes of neural networks. However, this is a challenge since we can not compute the error term $\partial E(\mathbf{w})/\partial w_j$ without knowing the output value $\hat{y}_i$. Using a technique known as *back-propagation* helps to solve the issue. This technique allows us to use the errors for neurons at iteration $k + 1$ to calculate the errors for neurons at the k^{th} iteration.

When building a MANN there are many choices to make such as the number of hidden layers, the number of nodes in each layer, and the activation function. In this study, I use a feed-forward single-layer neural network, where the nodes in one layer are connected only to nodes in the next one, and each unit in one layer is connected to all units in the previous layer. Figure 18 shows the architecture structure of the best ANN model used in this study. The structure of the ANN consists of one input layer with the number of nodes dependent on the attribute type, one hidden layer with 9 nodes, and one output layer with one node. A popular activation function choice for non-linear problems is the sigmoid function. The choice of this function fits well with this study because it gives probabilities as an output, and also it is used in feed-forward single-layer ANN.

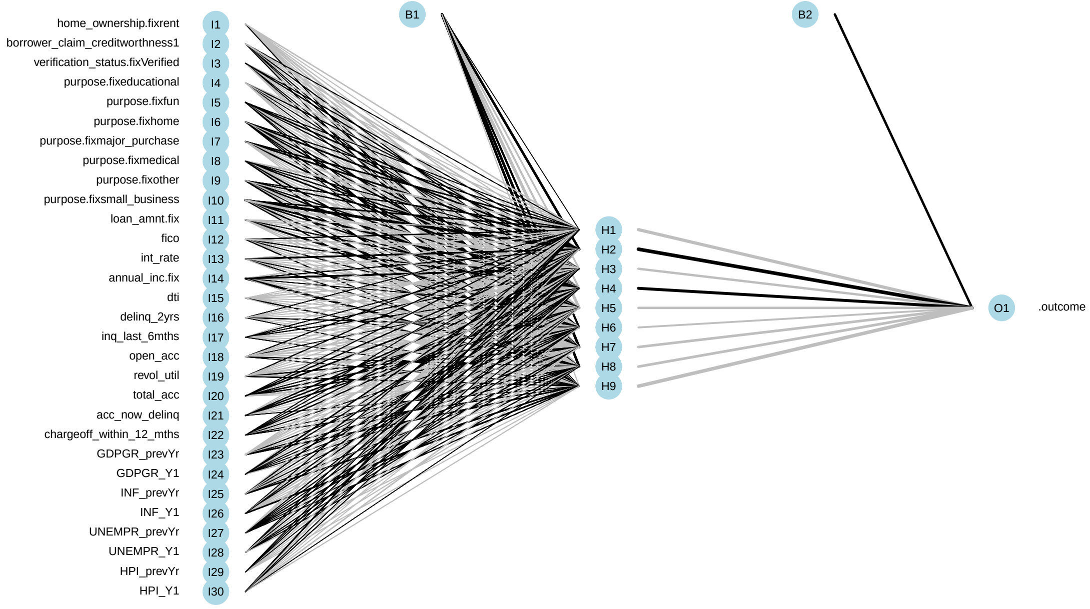

Figure 18: Architecture of the best feed forward single-layer ANN model

A critique for ANN is that sometimes it might get stuck at local minimum/maximum. Also, this method requires large amounts of data and it is computationally very expensive.

2.3.5 Support Vector Machine

SVM is another ML method related to statistical learning theory. It is useful when there are many input variables or when these input variables interact with the outcome or with each other in complicated (nonlinear) ways. SVMs make fewer assumptions about variable distribution than do many other ML methods, and this makes SVM especially useful when the training data is not completely representative of the way the data is distributed in production.

Consider an extremely simplified two-class linearly separable classification problem as shown in Figure 19 (left). Many decision boundaries will put the default class on one side and non-defaulters on the other side with zero misclassification error. However, which one of these possible boundaries will work best in an unseen dataset? A linear SVM tries to separate the data by maximizing the margin, which is the minimum distance between two data points of different classes that lie closest to the decision boundary. The data points or vectors are known as *support vectors*. Assume we have N observations in the training sample and each is denoted by a tuple $(\mathbf{x}_i, y_i)$ where $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}, ..., x_{id})$ is the feature space and for this particular method lets label the blue colored observations (i.e., non-default class) as $y_i = 1$ and red-colored observations (i.e., the default class) as $y_i = -1$.[35]

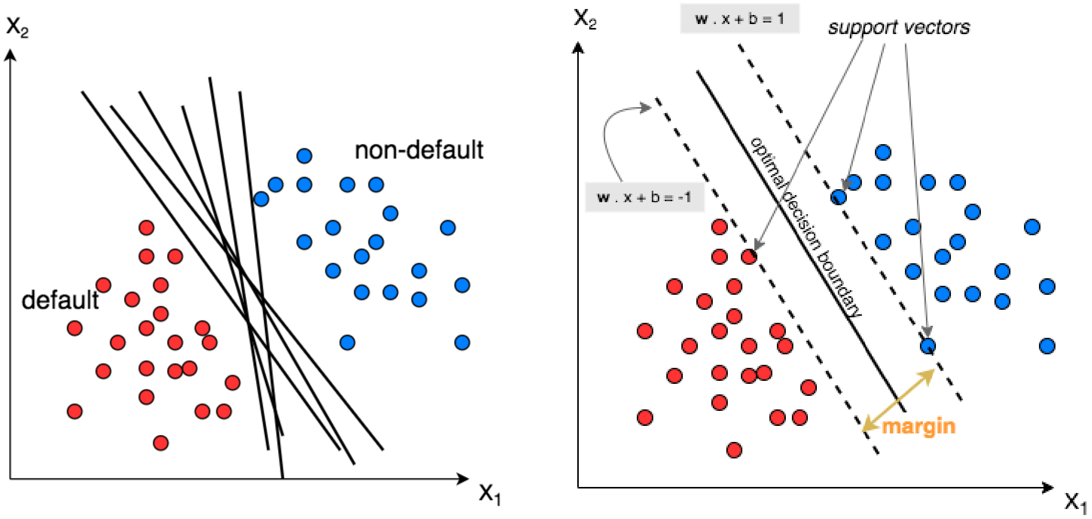

Figure 19: Linearly separable cases

The optimal decision boundary of a linear classifier can be written in the following form:

$$\boldsymbol{w} \cdot \mathbf{x}_i + b = 0\,, \tag{20}$$

where $(\boldsymbol{w}, b)$ are parameters of the model from the training data set, $\boldsymbol{w}$ is the weight vector and b

[35] I could have left the classes as 0/1 as in the previews methods but it is more convenient to label this way to avoid any confusion with the notation.

is a threshold. This linear classifier would classify any new observation from the unseen dataset according to the following equations:

$$\hat{y}_i = \begin{cases} 1, & \text{if } \boldsymbol{w} \cdot \mathbf{x}_i + b > 0 \\ -1, & \text{if } \boldsymbol{w} \cdot \mathbf{x}_i + b < 0 \end{cases} , \tag{21}$$

We can rescale the parameters $(\boldsymbol{w}, b)$ of the decision boundary so that the two parallel hyperplanes passing through the support vectors can be expressed as $\boldsymbol{\omega} \cdot \mathbf{x}_i + b = 1$ and $\boldsymbol{\omega} \cdot \mathbf{x}_i + b = -1$. The parameters (ω, b) should be chosen in such a way that all observations lying outside the hyperplanes satisfy the following conditions:

$$\boldsymbol{\omega} \cdot \mathbf{x}_i + b \geq 1 \text{ if } y_i = 1 , \tag{22}$$

$$\boldsymbol{\omega} \cdot \mathbf{x}_i + b \leq -1 \text{ if } y_i = -1 . \tag{23}$$

Equivalently we can write the above two inequality equations as:

$$y_i(\boldsymbol{\omega} \cdot \mathbf{x}_i + b) \geq 1 . \tag{24}$$

The distance between these two parallel hyperplanes, which is the length of the margin is $\frac{2}{||\omega||}$. The goal of the classifier is to find a decision boundary that maximizes this distance, which is equivalent to the minimization of $||\omega||$. Therefore our objective becomes an optimization problem with equation (24) as the constrain[36]:

$$\min_{\omega,b} \frac{||\boldsymbol{\omega}||^2}{2} \qquad \text{s.t}$$
$$y_i(\boldsymbol{\omega} \cdot \mathbf{x}_i + b) \geq 1, \qquad i = 1, 2, 3, ..., N .$$

We can solve this constrained optimization problem using the Lagrange multiplier method.

[36] Mathematically it is more convenient to minimize the $||\omega||^2$

$$L(\boldsymbol{\omega}, b, \lambda) = \frac{||\boldsymbol{\omega}||^2}{2} - \sum_{i=1}^{N} \lambda_i \Big[y_i (\boldsymbol{\omega} \cdot \mathbf{x}_i + b) - 1 \Big], \tag{25}$$

where λ_i's are the Lagrange multipliers. Taking derivatives w.r.t $\boldsymbol{\omega}$ and b we have:

$$\frac{\partial L}{\partial \boldsymbol{\omega}} = 0 \implies \boldsymbol{\omega} = \sum_{i=1}^{N} \lambda_i y_i \mathbf{x}_i, \tag{26}$$

$$\frac{\partial L}{\partial b} = 0 \implies \sum_{i=1}^{N} \lambda_i y_i = 0. \tag{27}$$

Equation (24) is an inequality constraint that forces us to apply Kuhn-Tucker conditions:

$$\lambda_i \geq 0, \tag{28}$$

$$\lambda_i \Big[y_i (\boldsymbol{\omega} \cdot \mathbf{x}_i + b) - 1 \Big] = 0. \tag{29}$$

Constraint (29) shows that most of the observations will have a Lagrange multiplier equal to zero as far as they reside on one side of the decision boundary and $\lambda_i > 0$ if observation i is a support vector. Moreover, equations (26) and (27) suggest that parameters ($\boldsymbol{\omega}$, b) are defined only by the support vectors and the other observations do not play a role in the decision boundary. This is another feature of SVM where only support vectors play a role in the classification decision, which is different from the logistic regression where all the observations are included as part of the decision problem. Substituting equations (26) and (27) into (25) we have

$$L(\lambda_i) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \mathbf{x}_i \mathbf{x}_j. \tag{30}$$

Equation (30) shows that now the optimization problem depends only on Lagrange multipliers and that the negative sign of the second term has changed it to a maximization problem due to the quadratic term. Once we have the λ_i's we can use equations (26) and (29) to solve for the optimal parameters ω and b, respectively.

In reality, very rarely we will have cases where we will have zero misclassification error. Most of the time the algorithm will follow a soft margin approach, which allows a limited fraction of the observations to be on the wrong side of the decision boundary. These misclassified examples are known as *slack variables*. The measure of the misclassification will depend on the trade-off between bias and variance. If we want to have a decision boundary with a small margin and a very low misclassification error, then we will have the overfitting problem because the classifier will perform very well on the training set but poorly on the unseen data. A decision boundary with a wider margin means more observations lie on the wrong side of the hyperplane and the classifier fits the data with lower variance and higher bias.

Figure 20 shows such a case where even though we can find a linear decision boundary to split the data with zero misclassification error we should prefer a decision boundary with some misclassification error because it will perform better in an unseen dataset. The SVM linear model uses a cost parameter C to decide on the trade-off between bias and variance. This parameter represents the penalty of misclassifying the training observations and it is determined by cross-validation.

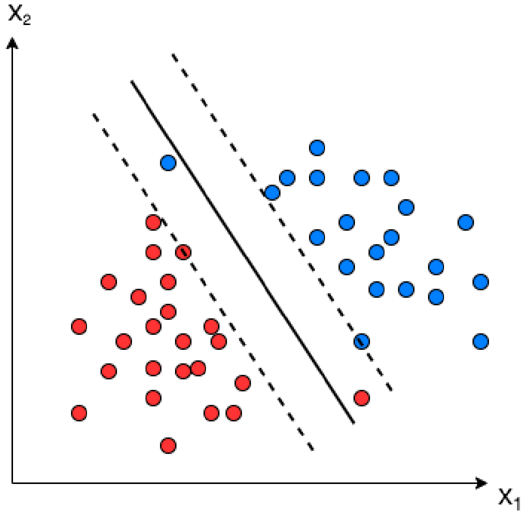

Figure 20: Linearly separable cases

For this scenario, we have to modify the constraint $y_i(\boldsymbol{\omega} \cdot \mathbf{x}_i + b) \geq 1$ because the decision boundary no longer satisfies this constrain. The inequality constraints must be relaxed to:

$$\boldsymbol{\omega} \cdot \mathbf{x}_i + b \geq 1 - \varepsilon_i \text{ if } y_i = 1\,, \tag{31}$$

$$\boldsymbol{\omega} \cdot \mathbf{x}_i + b \leq -1 + \varepsilon_i \text{ if } y_i = -1\,, \tag{32}$$

for $\varepsilon_i \geq 0 \quad \forall i$. Modifying the objective function to penalize decision boundaries with large values of slack variables, the constrained optimization problem becomes as follow:

$$\min_{\omega,b}\Big(\frac{||\boldsymbol{\omega}||^2}{2} + C\sum_{i=1}^{N}\varepsilon_i\Big) \qquad \text{s.t.,}$$
$$y_i(\boldsymbol{\omega} \cdot \mathbf{x}_i + b) \geq 1 - \varepsilon_i\,,$$
$$\varepsilon_i \geq 0 \quad \forall i\,.$$

Then the new Lagrange multiplier equation is:

$$L(\omega, b, \lambda_i, \varepsilon_i) = \frac{||\boldsymbol{\omega}||^2}{2} + C\sum_{i=1}^{N}\varepsilon_i - \sum_{i=1}^{N}\lambda_i\big[y_i(\boldsymbol{\omega} \cdot \mathbf{x}_i + b) - 1 + \varepsilon_i\big] - \sum_{i=1}^{N}\mu_i\varepsilon_i\,. \tag{33}$$

The KT conditions becomes:

$$\varepsilon_i \geq 0, \quad \lambda_i \geq 0, \quad \mu_i \geq 0\,, \tag{34}$$

$$\lambda_i\big[y_i(\boldsymbol{\omega} \cdot \mathbf{x}_i + b) - 1 + \varepsilon_i\big] = 0\,, \tag{35}$$

$$\mu_i\varepsilon_i = 0\,. \tag{36}$$

Taking derivatives w.r.t $\boldsymbol{\omega}$, b, and ε_i we have:

$$\frac{\partial L}{\partial \omega} = 0 \implies \omega = \sum_{i=1}^{N} \lambda_i y_i \mathbf{x}_i\,, \tag{37}$$

$$\frac{\partial L}{\partial b} = 0 \implies \sum_{i=1}^{N} \lambda_i y_i = 0\,, \tag{38}$$

$$\frac{\partial L}{\partial \varepsilon_i} = 0 \implies C = \lambda_i + \mu_i\,. \tag{39}$$

Substituting the above equations into the Lagrange equation (33) we have:

$$L(\lambda_i) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \mathbf{x}_i \mathbf{x}_j\,. \tag{40}$$

which is exactly the same as in the equation (30). Just for this example the constraints are different. Equation (39) together with the restriction (34) from KT conditions suggest that $0 \le \lambda_i \le C$.

In most of the cases the raw data will not be linearly separable, as shown in Figure 21. One way of dealing with a non-linearly separable case is by applying a data transformation from its original space $\mathbf{x}$ to a new space $\Phi(\mathbf{x})$ where we can use a linear decision boundary to separate the data. The objective function and the constraint have the same form as before in the new space:

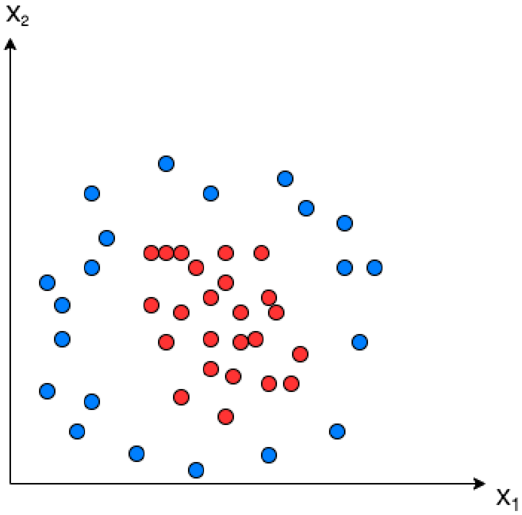

Figure 21: Non-Linearly separable cases

$$\min_{\omega,b} \frac{||\omega||^2}{2} \qquad \text{s.t.,}$$
$$y_i(\boldsymbol{\omega} \cdot \Phi(\mathbf{x}_i) + b) \geq 1, \qquad i = 1, 2, 3, ..., N,$$

and the last form of the Lagrange equation would be similar to equation (30) and (40) but the dot product of observation would be in the transformed space

$$L(\lambda_i) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \Phi(\mathbf{x}_i)\Phi(\mathbf{x}_j). \tag{41}$$

However, applying this transformation is computationally very expensive, especially if you are dealing with large dimensional spaces (i.e., a lot of variables) we might have the curse of dimensionality problem. Moreover, finding a function $\Phi(\mathbf{x})$ that will linearly separate the dataset is very cumbersome. Here we apply the so-called *Kernel trick*. The kernel trick is a method that finds the dot product of $\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$ without applying the transformation function. This means that we can find a functions **K(.)** such that

$$K(\mathbf{x}_j, \mathbf{x}_j) = \Phi(\mathbf{x}_i)\Phi(\mathbf{x}_j). \tag{42}$$

The kernel function $K(.)$ computes the dot product using the original space and is computationally cheaper. Moreover, since it operates in the original space, it avoids the curse of dimensionality problem. The real power of SVM relies on this trick. There are different types of kernel functions and the most applied is Gaussian or radial kernel function (RBF) which I use in this study. It has the following form:

$$K(\mathbf{x}_j, \mathbf{x}_j) = \exp(-\gamma ||\mathbf{x}_j - \mathbf{x}_j||^2), \tag{43}$$

where the γ coefficient is important to control the influence of function. Therefore, the Lagrange multiplier equations used in this study has the following form:

$$L(\lambda_i) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j K(\mathbf{x}_j, \mathbf{x}_j) . \tag{44}$$

While other ML methods apply gradient descent strategy on solving optimization problems, which does not guarantee that the solution is global minima or maxima. The setup of the SVM on the other side guarantees that the solution found is global minima or maxima.

2.4 Data, Pres-processing, Parameter Tuning and Performance Metrics

My dataset consists of 597,391 loan-level observations. After removing some of the observations with missing data—mainly some 3-digit zipcodes with missing values for unemployment rates and house price indices—the final version of my dataset consists of 550,146 observations and 25 variables. The variables used consists of borrower characteristics such as loan amount, interest rate, FICO score, annual income, debt-to-income ratio, delinquency in last 2 years, inquiries last six months, number of opened accounts, revolving line utilization rate, the total number of credit lines currently in the borrower's credit file, the number of accounts on which the borrower is now delinquent, number of charge-offs within 12 months; macroeconomic variables such as GDP growth, inflation at the origination year and one lag; regional economic indicators such as change in the unemployment rate and HPI inflation at the 3-digit zip code level aggregated yearly. Regional and macro variables are exogenous variables used as shocks. Macro and regional variables change over time and across geographic areas based on when the borrower has applied for the loan and the respective location.

All numerical variables are standardized with mean zero and unit variance before used in training the models.[37] This is done to avoid any scaling issues.

The data set is split into a training set (80 percent, or 440,118 observations) and a testing set (20 percent, or 110,028 observations). It is important to emphasize that the random split is performed in a way that the ratio between default and non-default classes is preserved in both training and testing sets. A different approach would lead to model under-performance for out-of-sample predictions. The training set is used to build the classification models, which are subsequently applied to the testing set for out-of-sample prediction and performance assessment.

In this paper I use a repeated 5 times 10-fold cross-validation to select which model among the same family performs best. I do this to validate the stability of the models on the unseen data and reduce the risk of overfitting the models. In other words, I use cross-validation to build a model on 9-folds (i.e., training fold) and test it on the complement 10th fold (i.e., test fold). This way we build 10 different models and make predictions in all of the data and estimate the accuracy of the models. For classification problems, stratified cross-validation is even better, which makes sure that each fold is a good representative of the whole data. For example, the default class is 14.2% of the data, so we need to make sure that this ratio between default and non-default class is preserved in each fold for better models.

I chose the hyper-parameters of the models via a grid-search algorithm. All the hyper-parameters were determined using 5 times repeated 10-fold cross-validation. This was done by specifying a range in the grid search or by setting a specific value for the hyper-parameter of interest.

[37] I use $z_i = \frac{x_i - \mu_x}{\sigma_x}$ values in training the models, where $\mathbf{x} = (x_1, x_2, ..., x_k)$ is the input vector, and z_i is the normalized data for i^{th} observation.

2.4.1 Model Performance

There are few measurements used to evaluate how accurately a predictive model will perform in practice, and all are derived from the confusion matrix.[38] Table 17 shows how the typical confusion matrix looks like. True Positive (TP) shows the number of loans that defaulted and are predicted from the model as default. True Negative (TN) shows the numbers of loans with non-default status and the model predicts them as non-default. False Negative (FN) shows the number of loans with the status of default but the model predicts them as non-default (also known as Type II error). False Positive (FP) shows the number of loans that are non-defaults but the model predicts them as default (also known as Type I error).

Table 17: Confusion Matrix

		Prediction	
		default (1)	non-default (0)
Actual	default (1)	TP	FN
	non-default (0)	FP	TN

The measurements used in this study are accuracy, sensitivity, specificity, positive and negative predicted values. Accuracy is the fraction of the sum of correctly predicted observations divided by the total number of observations in the data. It is important to emphasize that accuracy might be a misleading performance measure for a dataset with unbalanced classes of the target variable. Because of this problem, generally, credit risk models' performance is based on sensitivity, specificity, and precision.

[38] A confusion matrix summarizes the model's performance. It shows predictions against the actual outcomes. In other words, it is a table counting how often each combination of known outcomes (the truth) occurred in combination with each prediction type.

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} = \frac{TP + TN}{TP + FN + FP + TN}.$$

Sensitivity captures the ability of the model to correctly identify those who will default upon taking a loan. This measure is also known as true positive rate (TPR) or Recall.

$$Sensitivity/Recall/TPR = \frac{TP}{TP + FN}.$$

Specificity is the ability of the model to correctly identify those who will not default. Also, it is known as the true negative rate (TNR).

$$Specificity/TNR = \frac{TN}{TN + FP}.$$

False Positive Rate (FPR) is the rate between false predicted default and the actual total number of non-defaulters.

$$FPR = 1 - Specificity = \frac{FP}{TN + FP}.$$

The companion score to recall is precision or positive predicted values (PPV). PPV is the ratio of correctly identified defaulted loans by the model divided by the actual total number of defaults.

$$PPV = \frac{TP}{TP + FP}.$$

Negative Predicted Values (NPV) is the ratio of correctly identified non-default loans divided by the actual total number of non-defaults.

$$NPV = \frac{TN}{TN + FN}.$$

In the lending industry, predicting both the good and bad borrowers is important. Lenders do not want to lose money by giving loans to people who will default, but also do not want to reject good potential borrowers who will pay back the loan. So in this scenario, we want high values for both *Sensitivity* and *Specificity* because these are important model performance indicators for our problem.[39]

The comparison between models forecast performance is done using Area Under the Curve (AUC) which utilizes ROC Curve. ROC is a probability curve that plots the true positive rates against false positive rates. For classification problems, AUC-ROC is a metric that measures how well the model is capable to distinguish between default and non-default classes. Different ranges of AUC show the quality of the forecasts by the models. A value between 0.9 – 1 is considered an excellent forecast performance, 0.8 – 0.9 range is considered good, 0.7 – 0.8 fair, 0.6 – 0.7 poor and 0.5 – 0.6 fail. By adjusting the threshold, the ROC curve provides models with different discrimination capability.[40]

2.4.2 Balancing the Imbalanced Data with SMOTE

One common problem with predicting default is that the target variable is highly unbalanced: 85.8 percent of the loans are classified as *non-default* and 14.2 percent as *default*. An imbalanced dataset causes problems and poor performance of the prediction models. In a logistic regression

[39] These two measurements are important as we want to better price those who have higher chances of defaulting on a loan with higher interest rates and give low interest rates to those borrowers who have low default probabilities. Ideally we want the sensitivity and precision to be high and the false positive rate (i.e., non-defaulters that are classified as defaulters) low. Often for best results it will require some trade off between recall and precision.

[40] When AUC-ROC is close to 0.5, this means the model perform as random as throwing a coin with equal probabilities. This means the model has no discrimination capacity to identify between default and non-default classes.

framework, the conditional probability of minor classes will be underestimated (King and Zeng, 2001). Moreover, in tree-based models, the imbalance of the data can yield high specificity but low sensitivity (Cieslak and Chawla, 2008). These methods have a bias towards the majority class (non-default in our case), and they tend to ignore the minority class i.e., default. This will lead to misclassification of the minority class relative to the majority one, which will lead to low sensitivity. To improve models' performance different sampling techniques can be used, such as under-sampling, over-sampling and hybrid methods. Under-sampling or down-sampling technique tends to reduce the number of observations from the unbalanced training set until the ratio between classes is one. Even though theoretically this looks right in big datasets, it has the problem of losing valuable information about the majority class when throwing away observations. This way of balancing the data might lead to bias results. Oversampling or up-sampling technique replicates the data based on current observations of the minority group to balance the classes in the training set. However, this will lead to an overfitting problem, that means high accuracy in the training and low on the testing set. A hybrid method such as SMOTE deals with the above mentioned problems. SMOTE randomly increases the minority class of the target variable by replicating them. Different from the other oversampling techniques, instead of duplicating the current observations SMOTE creates new synthetic data points from the current observations with minority class (Chawla et al., 2002). It finds random points within nearest neighbors of each observation in the minority class and by boosting methods generates new observations with the minority class. Newly created synthetic data points generated by SMOTE are not duplicates of an existing observation, they are still based on original observations. This way overfitting problem in standard oversampling techniques will not be an issue anymore.

It is important to mention that this technique is implemented only on the training set and not on the testing set as the model performance will be evaluated on a real scenario case.

2.4.3 Implementation

Machine learning models' computations for such a dataset tend to be time and memory intensive, and can not be performed on a personal laptop/desktop. To run these models I use high performance computing resources at Hewlett Packard Enterprise Data Science Institute (HPEDSI) at the University of Houston. I parallelize the codes and train all the models in the cluster. Using cluster computing resources accelerated the learning rate and led to substantial computational speedup for each model.

All my analyses are done using R programming language and a Linux environment, and utilizing R packages such as *caret*, *dplyr*, *ggplot2*, *pROC*, *DMwR*, *parallel*, *doMC*, and *devtools*.

2.5 ML Models' Results

Table 18 reports the in- and out-of-sample performance of all models used in this paper, both with and without SMOTE implementation. Given the model simplicity, LR performs significantly well relative to the other machine learning models. Under no SMOTE implementation, the model gives an accuracy of 86 percent for both in- and out-of-sample prediction. for the training and testing set is 68.3 percent and 68.1 percent, respectively. Specificity is very high but the value for sensitivity is low, which is a common problem in unbalanced datasets. After SMOTE implementation we see a significant increase in the sensitivity score on both training and testing sets. This shows that balancing the dataset with SMOTE plays a significant role in predicting default with LR.

Table 18: Estimated results from all models

ML model	**Metric**	**No-SMOTE**		**SMOTE**	
		training	testing	training	testing
Logistic Regression	Accuracy	0.858	0.858	0.722	0.721
	Sensitivity	0.006	0.007	0.473	0.471
	Specificity	0.999	0.999	0.764	0.762
	Pos Pred Value	0.468	0.533	0.249	0.247
	Neg Pred Value	0.859	0.859	0.898	0.897
	AUC-ROC	0.683	**0.681**	0.682	0.680
Decision Tree	Accuracy	0.888	0.808	0.846	0.844
	Sensitivity	0.404	0.141	0.076	0.072
	Specificity	0.968	0.919	0.974	0.971
	Pos Pred Value	0.674	0.223	0.326	0.293
	Neg Pred Value	0.907	0.866	0.864	0.864
	AUC-ROC	0.825	0.608	0.665	0.662
Random Forest	Accuracy	1.000	0.858		
	Sensitivity	1.000	0.001		
	Specificity	1.000	0.999		
	Pos Pred Value	1.000	0.487		
	Neg Pred Value	1.000	0.858		
	AUC-ROC	1.000	0.674		
eXtreme Gradient Boost	Accuracy	0.858	0.858		
	Sensitivity	0.003	0.002		
	Specificity	0.999	0.999		
	Pos Pred Value	0.751	0.550		
	Neg Pred Value	0.858	0.858		
	AUC-ROC	0.710	**0.688**		
Artificial Neural Network	Accuracy	0.858	0.858		
	Sensitivity	0.002	0.002		
	Specificity	0.999	0.999		
	Pos Pred Value	0.581	0.578		
	Neg Pred Value	0.858	0.858		
	AUC-ROC	0.692	**0.687**		

DT model gives high accuracy and AUC-ROC values for the training set but lower AUC-ROC value on the testing set. This is because mainly DT models suffer from the overfitting problem. Different from the results in LR, SMOTE implementation seems to have a different effect in the DT model. After the SMOTE we see the values for sensitivity have decreased but there is a significant increase in the AUC-ROC value for the testing set which has increased from 60.8 percent to 66.2 percent.

Despite the usage of cross-validation, the RF model overfits the training data. However, when applied for the out-of-sample prediction, results are close to the ones from LR.

XGB outperforms LR and all the other ML methods. The model gives an AUC-ROC score of 71 percent on the training set and 68.8 percent on the testing set. The accuracy rate is 85.8 percent, which similar for both training and testing sets.

ANN performs very close to the XGB with an AUC-ROC of 69.2 percent on the training set and 68.7 percent on the testing set. Moreover, ANN has the highest positive predictive values (i.e., precision) among all the methods used in this study.

RF, XGB, and ANN models under SMOTE implementation run for 14 complete days on the cluster but this time was not enough to get results. I plan to do more on this in future work. In addition to the methods in Table 18, I used GBM and SVM with and without SMOTE implementation which also run for 14 complete days in the cluster and this time frame was not enough to train the models.

It is important to emphasize that results in Table 18 do not indicate that these are the best performance of each model. Making the models more complex or adding other variables will potentially lead to better out-of-sample predictions.

Variable importance is another feature in machine learning.[41] I use this feature to gain intuition and be able to interpret the model results. Figure 22 shows the importance of the variables in predicting default according to LR, DT, RF, and ANN. We can see that among the methods there are some significant differences in which variables best predict default behavior. For example, according to LR, the interest rate is the most important variable explaining default.[42] Higher the interest rate higher the burden on borrowers, therefore higher the probability of default. The second and third most important variables explaining default are borrowers' self-reported income and FICO score, respectively. Higher the income, lower the probability of default. FICO is a widely used measure by lenders to evaluate potential borrowers' creditworthiness. We can see that GDP

[41] Each method has a different way on how to measure the importance of the variables in predicting the outcome. For example for linear models, variable importance is based on the absolute values of the t-statistics.

[42] For LR estimated parameters with and without SMOTE implementation see Table 19 in the appendix.

growth is in the top 10 variables predicting default. Loans issued to small businesses are more likely to default than any other type of loan.

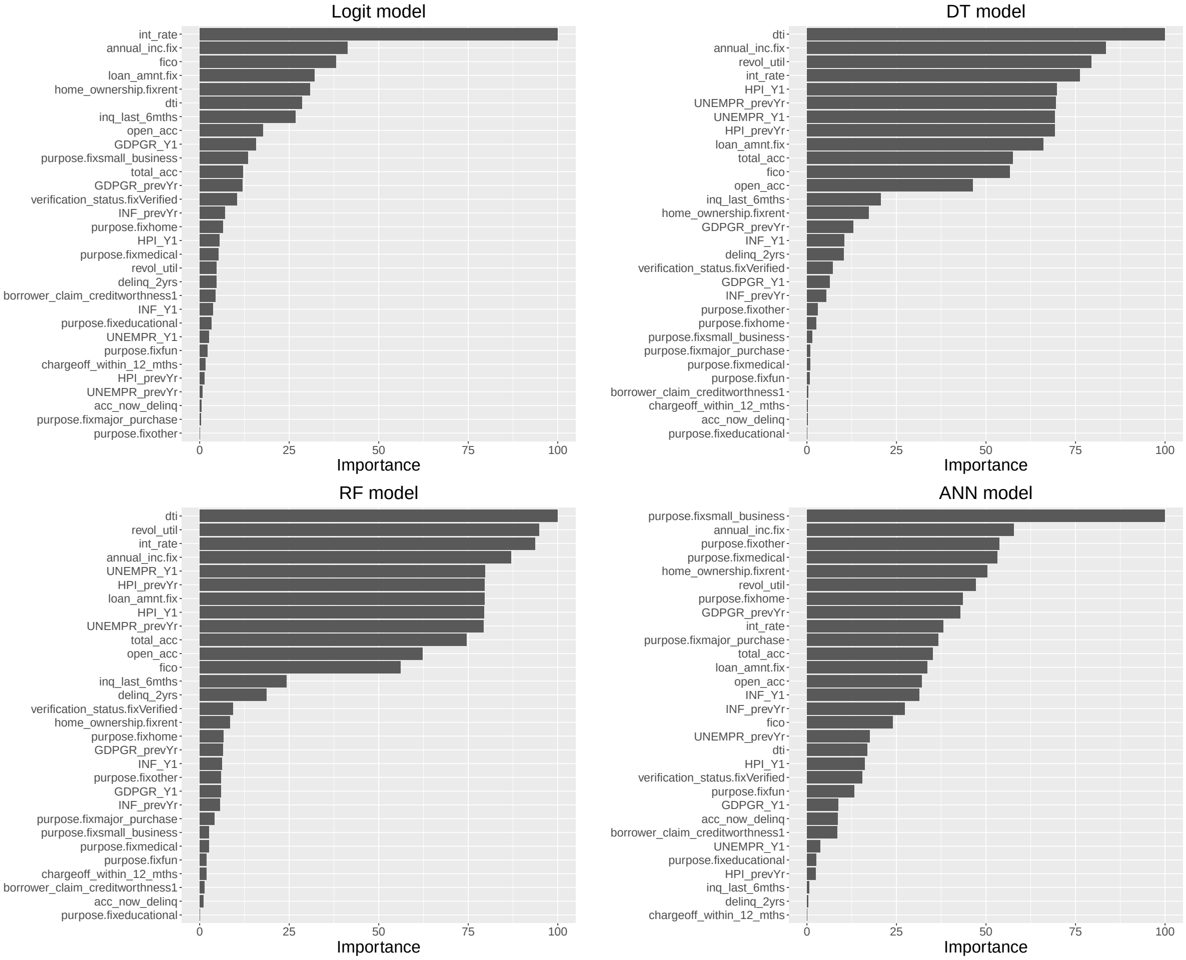

Figure 22: Importance of variables according to each model

DT and RF have very similar results. We see that for both methods debt-to-income ratio is the most important variable. the more a potential borrower is in debt, the higher the probability of default. For both these methods, we see that macro and regional variables make it in the top 10 important variables explaining default in the FinTech lending market. Borrowers' self-stated personal income, interest rate, and revolving line utilization are other important variables capturing default.

ANN which also has the best performance in predicting default lists borrower self-stated loan purpose as the variable that best explain default. Similar to finding from LR, loans issued to a small business are more likely to default than any other type of loan. Among the macro and regional variables GDP growth is one of the top 10 most important variables. Similarly, FICO does not seem to be among the most important variable. These results are consistent with the results by Jagtiani and Lemieux (2017) where they find that LC grades have a decreasing correlation with FICO scores and debt to income ratios over time, indicating that alternative data is being used.

2.6 Conclusion

This study provides an overview of some of the most common machine learning methods used in modeling default risk and assesses to what extent these methods are better than traditional approaches. I use a loan-level dataset from the largest FinTech lender in the United States, together with other macro and regional economic variables to evaluate the determinants of default in this market. I apply different machine learning algorithms to predict out-of-sample default. The ML methods used in this study consist of LR, DT, RF, XGB, ANN. An oversampling technique called SMOTE is implemented to balance the classes for the response variable.

I find that some of the machine learning algorithms such as XGB and ANN outperform logistic regression, but the improvement is marginal. The variables with the highest importance in predicting default are annual income, loan purpose, revolving line utilization, and the interest rate. Most of the macro and regional variables are listed among the top 10 variables predicting default.

I used high-performance computing resources (i.e., clusters) to train the models, which led to significant speedup for each method. It is important to mention that I tried also other methods such as GBM and SVM, which run for 14 days on the cluster and this time frame was not enough to get results. I plan to do more on this in future work.

The results in this paper do not indicate that these are the best results these methods can deliver. A possible improvement using more complex versions of ANN with multiple hidden layers and different activation functions such as hyperbolic tangent functions (tanh) or rectified linear unit (RELU) can potentially lead to better model performance. Adding more variables also can improve the out-of-sample predictions.

2.7 Appendix

The estimated coefficients from the logistic regression model with and without SMOTE implementation are shown in Table 19.

Table 19: Logistic Regression Results No-SMOTE vs. SMOTE

	defaulted	
	No-SMOTE	SMOTE
Constant	-2.12***(0.01)	-0.61***(0.01)
home_ownership.fixrent	0.22***(0.01)	0.20***(0.01)
borrower_claim_creditworthness1	-0.17***(0.04)	-0.20***(0.03)
verification_status.fixVerified	0.08***(0.01)	0.08***(0.01)
purpose.fixeducational	0.57***(0.20)	0.37**(0.15)
purpose.fixfun	-0.10**(0.05)	-0.07**(0.03)
purpose.fixhome	0.10***(0.02)	0.13***(0.01)
purpose.fixmajor_purchase	0.02(0.03)	0.07***(0.02)
purpose.fixmedical	0.18***(0.04)	0.21***(0.03)
purpose.fixother	-0.01(0.02)	0.01(0.01)
purpose.fixsmall_business	0.38***(0.04)	0.48***(0.03)
loan_amnt.fix	0.14***(0.01)	0.17***(0.004)
fico	-0.20***(0.01)	-0.25***(0.005)
int_rate	0.42***(0.01)	0.43***(0.004)
annual_inc.fix	-0.19***(0.01)	-0.22***(0.005)
dti	0.11***(0.01)	0.13***(0.004)
delinq_2yrs	0.02***(0.004)	-0.01***(0.003)
inq_last_6mths	0.09***(0.004)	0.08***(0.003)
open_acc	0.09***(0.01)	0.11***(0.005)
revol_util	-0.02***(0.01)	-0.03***(0.004)
total_acc	-0.06***(0.01)	-0.11***(0.01)
acc_now_delinq	-0.004(0.004)	-0.02***(0.004)
chargeoff_within_12_mths	-0.01*(0.004)	-0.02***(0.004)
GDPGR_prevYr	0.08***(0.01)	0.09***(0.01)
GDPGR_Y1	-0.10***(0.01)	-0.10***(0.01)
INF_prevYr	-0.05***(0.01)	-0.05***(0.01)
INF_Y1	-0.03***(0.01)	-0.02***(0.01)
UNEMPR_prevYr	0.01(0.01)	0.01*(0.01)
UNEMPR_Y1	0.02**(0.01)	0.03***(0.01)
HPI_prevYr	0.01(0.01)	0.004(0.01)
HPI_Y1	-0.04***(0.01)	-0.05***(0.01)
Observations	440,118	437,661
Log Likelihood	-168,479.90	-275,093.80
Akaike Inf. Crit.	337,021.70	550,249.70

Figure 23 shows the distribution of LC assigned grades by default status for those borrowers who own a house vs. those who rent. The share of borrowers who default with grade A and who are homeowners is lower relative to the share of borrowers with the same qualifications but who are renters. In addition, the difference in share between borrowers who are homeowners and renters keeps increasing for higher risk grades.

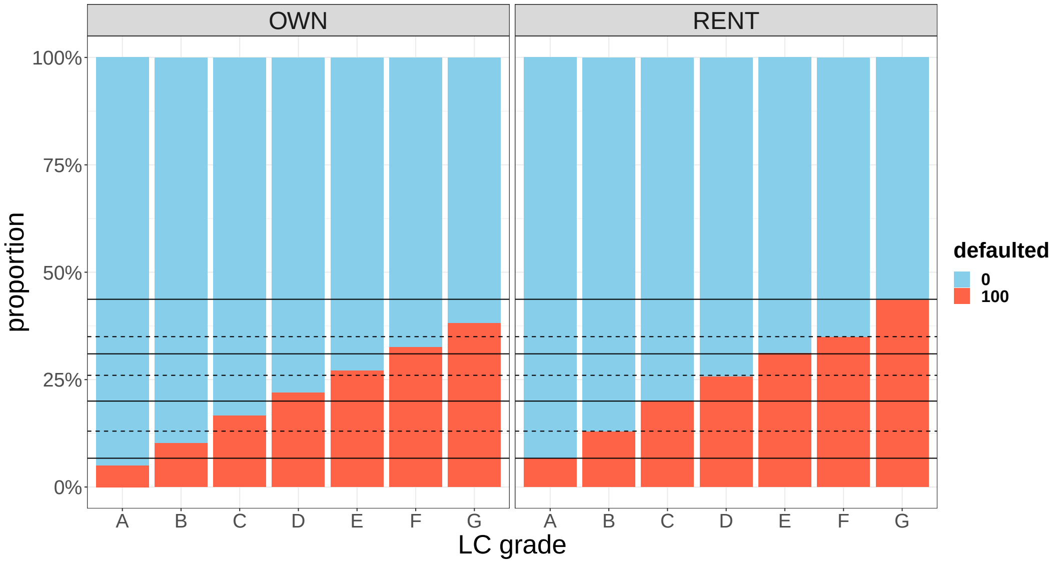

Figure 23: LC grade default share by home-ownership status

We know that subprime category have higher default risk, therefore higher interest rates should be charged. Theoretically, if FinTech lenders have better information and are able to measure risk more accurately, then the default rates between prime and subprime, condition on grade should not be different. Figure 24 shows the default share for prime and subprime borrowers. We can see that subprime borrowers have higher default probabilities than prime borrowers when conditioning on LC grade, which means their pricing model is not accurate when accounting for the risk.

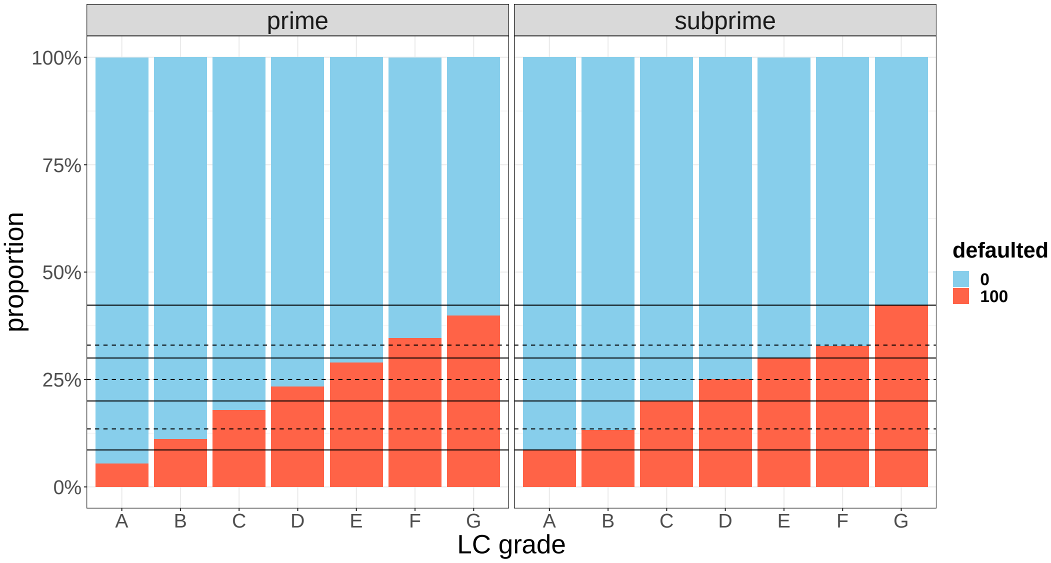

Figure 24: LC grade default share for prime vs. subprime borrowers

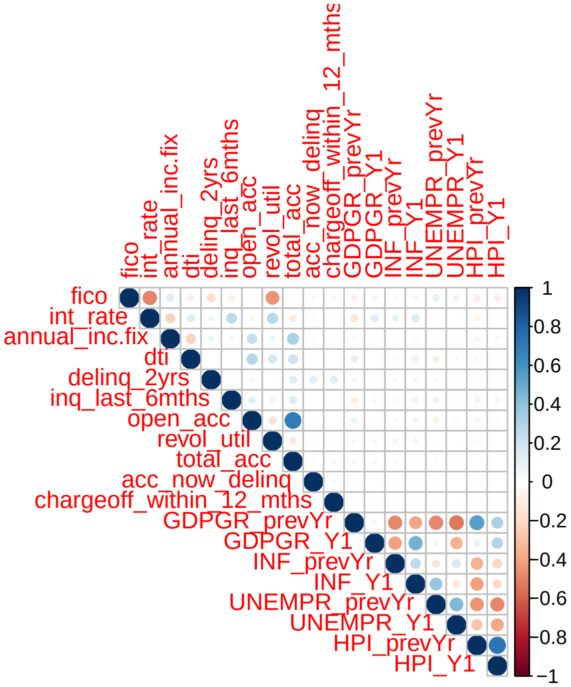

Figure 25: Correlation Plot

Bibliography

Adams, Robert. 2018. “Do Marketplace Lending Platforms Offer Lower Rates to Consumers?”

Adams, Robert, Tim Dore, Claire Greene, Traci Mach, and Jason Premo. 2017. “US Consumers’ Awareness and Use of Marketplace Lending.”

Agarwal, Sumit, and Chunlin Liu. 2003. “Determinants of Credit Card Delinquency and Bankruptcy: Macroeconomic Factors.” *Journal of Economics and Finance*, 27(1): 75–84.

Albanesi, Stefania, and Domonkos Vamossy. 2019. “Predicting Consumer Default: A Deep Learning Approach.” National Bureau of Economic Research.

Athey, Susan. 2018. “The impact of Machine Learning on Economics.” 507–547.

Azizaj, Eris. 2020. “FinTech Marketplace Lending, Default Risk, and the Business Cycle.” *SSRN Electronic Journal*.

Bagherpour, Ali. 2017. “Predicting Mortgage Loan Default with Machine Learning Methods.” *University of California/Riverside*.

Beiseitov, Eldar. 2019. “Unsecured Personal Loans Get a Boost from Fintech Lenders.”

Björkegren, Daniel, and Darrell Grissen. 2018. “Behavior Revealed in Mobile Phone Usage Predicts Loan Repayment.” *Available at SSRN 2611775*.

Brunnermeier, Markus K. 2009. “Deciphering the Liquidity and Credit Crunch 2007–2008.” *Journal of Economic Perspectives*, 23(1): 77–100.

Brunnermeier, Markus K., and Martin Oehmke. 2013. “Bubbles, Financial Crises, and Systemic Risk.” *Handbook of the Economics of Finance*, 1221–1288.

Carmichael, Don. 2014. “Modeling Default for Peer-to-Peer Loans.” *SSRN Electronic Journal*.

CGFS-FSB. 2017. “FinTech credit: Market Structure, Business Models and Financial Stability Implications.” Global Financial System and Financial Stability Board.

Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. “SMOTE: Synthetic Minority Over-Sampling Technique.” *Journal of artificial intelligence research*, 16: 321–357.

Cieslak, David A., and Nitesh V. Chawla. 2008. “Learning Decision Trees for Unbalanced Data.” *Lecture Notes in Computer Science*, 241–256.

Conlin, Michael. 1999. “Peer Group Micro-Lending Programs in Canada and the United States.” *Journal of Development Economics*, 60(1): 249–269.

Demyanyk, Yuliya S., and Otto Van Hemert. 2008. “Understanding the Subprime Mortgage Crisis.” *SSRN Electronic Journal*.

De Roure, Calebe, Loriana Pelizzon, Paolo Tasca, and Anjan Thakor. 2016. “How Does P2P Lending Fit into the Consumer Credit Market?” *SSRN Electronic Journal*.

Dietrich, Andreas, and Reto Wernli. 2016. “What Drives the Interest Rates in the P2P Consumer Lending Market? Empirical Evidence from Switzerland.”

Dore, Timothy, and Traci Mach. 2019. “Marketplace Lending and Consumer Credit Outcomes: Evidence from Prosper.”

Emekter, Riza, Yanbin Tu, Benjamas Jirasakuldech, and Min Lu. 2015. “Evaluating Credit Risk and Loan Performance in Online Peer-to-Peer (P2P) Lending.” *Applied Economics*, 47(1): 54–70.

Ghosh, Amit. 2015. “Banking-Industry Specific and Regional Economic Determinants of Non-Performing Loans: Evidence from US States.” *Journal of Financial Stability*, 20(C): 93–104.

Gross, David B., and Nicholas S. Souleles. 2015. "An Empirical Analysis of Personal Bankruptcy and Delinquency." *The Review of Financial Studies*, 15(1): 319–347.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Science & Business Media.

Hertzberg, Andrew, Andres Liberman, and Daniel Paravisini. 2018. "Screening on Loan Terms: Evidence from Maturity Choice in Consumer Credit." *The Review of Financial Studies*, 31(9): 3532–3567.

Jagtiani, Julapa, and Catharine Lemieux. 2017. "Fintech Lending: Financial Inclusion, Risk Pricing, and Alternative Information."

Jagtiani, Julapa, and Catharine Lemieux. 2018. "Do FinTech Lenders Penetrate Areas that are Underserved by Traditional Banks?" *Journal of Economics and Business*, 100: 43–54.

King, Gary, and Langche Zeng. 2001. "Logistic Regression in Rare Events Data." *Political Analysis*, 9(2): 137–163.

Louzis, Dimitrios P., Angelos T. Vouldis, and Vasilios L. Metaxas. 2012. "Macroeconomic and Bank-Specific Determinants of Non-Performing Loans in Greece: A Comparative Study of Mortgage, Business and Consumer Loan Portfolios." *Journal of Banking and Finance*, 36(4): 1012–1027.

Mach, Traci, Courtney Carter, and Cailin Slattery. 2014. "Peer-to-Peer Lending to Small Businesses."

Maudos, Joaquin, and Juan Fernandez De Guevara. 2004. "Factors Explaining the Interest Margin in the Banking Sectors of the European Union." *Journal of Banking and Finance*, 28(9): 2259–2281.

Mian, Atif, and Amir Sufi. 2009. “The Consequences of Mortgage Credit Expansion: Evidence from the U.S. Mortgage Default Crisis.” *The Quarterly Journal of Economics*, 124(4): 1449–1496.

PricewaterhouseCoopers. 2015. “Peer Presure. How Peer-to-Peer Lending Platforms are Transforming the Consumer Lending Industry.”

Reserve, Federal. 2021. Consumer Credit - G.19, Board of Governors of the Federal Reserve System.

Serrano-Cinca, Carlos, Begona Gutierrez-Nieto, and Luz López-Palacios. 2015. “Determinants of Default in P2P Lending.” *PloS One*, 10.

Stolba, Stefan L. 2020. “Credit Card Debt in 2020: Balances Drop for the First Time in Eight Years.” Experian.

Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. 2016. *Introduction to Data Mining.* Pearson Education India.

Turiel, J.D., and T. Aste. 2020. “Peer-to-Peer Loan Acceptance and Default Prediction with Artificial Intelligence.” *Royal Society Open Science*, 7(6): 191649.

Wang, J. Christina. 2018. “Technology, the Nature of Information, and FinTech Marketplace Lending.” *Federal Reserve Bank of Boston Research Paper Series Current Policy Perspectives Paper.*

Wang, J. Christina, and Charles B. Perkins. 2019. “How Magic a Bullet is Machine Learning for Credit Analysis? An Exploration with FinTech Lending Data.” Working Papers.