A Machine Learning Approach to Credit Allocation

by

Domonkos F. Vamossy

B.A. in Mathematical Economics and Mathematics, Whitworth University, 2015

M.A. in Economics, University of Pittsburgh, 2017

Submitted to the Graduate Faculty of

the Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2020

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Domonkos F. Vamossy

It was defended on

July 20, 2020

and approved by

Stefania Albanesi, Professor of Economics, University of Pittsburgh

Daniel Berkowitz, Professor of Economics, University of Pittsburgh

Douglas Hanley, Assistant Professor of Economics, University of Pittsburgh

Sera Linardi, Associate Professor of Economics, University of Pittsburgh

Dokyun Lee, Assistant Professor of Business Analytics, Carnegie Mellon University

Dissertation Director: Stefania Albanesi, Professor of Economics, University of Pittsburgh

A Machine Learning Approach to Credit Allocation

Domonkos F. Vamossy, PhD

University of Pittsburgh, 2020

This dissertation seeks to understand the shortcomings of contemporaneous credit allocation, with a specific focus on exploring how an improved statistical technology impacts the credit access of societally important groups. First, this dissertation investigates a variety of limitations of conventional credit scoring models, specifically their tendency to misclassify borrowers by default risk, especially for relatively risky, young, and low income borrowers. Second, this dissertation shows that an improved statistical technology need not to lead to worse outcomes for disadvantaged groups. In fact, the credit access for borrowers belonging to such groups can be improved, while providing more accurate credit risk assessment. Last, this dissertation documents modern-day disparities in debt collection judgments across white and black neighborhoods. Taken together, this dissertation provides valuable insights for the design of policies targeted at reducing consumer default and alleviating its burden on borrowers and lenders and across societally important groups, as well as macroprudential regulation.

Table of Contents

List of Tables

List of Figures

Preface

I am incredibly grateful to my advisor, Stefania Albanesi, for her constant guidance and encouragement. Her mentorship, collaboration, and the credit report data she provided, made this dissertation possible. I am also fortunate to have Dokyun Lee on my committee, whose class sparked my interest in exploring machine learning applications in economics, and whose help substantially improved my work. Sera Linardi is also someone I am extremely thankful for, her energy, enthusiasm and love of work set a great example for me. I would also like to thank Daniel Berkowitz and Douglas Hanley, along with the rest of the faculty members and administrators in the Economics Department at the University of Pittsburgh for their advice, support and assistance over the years. Special thanks to Osea Giuntella, Rania Gihleb, Georgia Spears, Brian Deutsch, Jessica LaVoice, Mallory Avery, Mark Azic, Lucy Wang, Lan Morrall, Max Myers, Tristan Cunha, and Viraj Mehta along with my amazing officemates and the rest of the graduate students in the Economics Department at the University of Pittsburgh for their friendship.

Finally, I would like to express my gratitude to my incredible family for their unconditional love and support. Special thanks to my parents, my great uncle, Julius Varallyay, and my grandmother, who set me up on this journey roughly eight years ago. Last, I dedicate this work to my late grandfather, Ferenc Vámossy, who I inherited my middle name from, and who would have been thrilled to read this dissertation.

1.0 Predicting Consumer Default: A Deep Learning Approach

(joint with Stefania Albanesi) We develop a model to predict consumer default based on deep learning. We show that the model consistently outperforms standard credit scoring models, even though it uses the same data. Our model provides favorable credit risk assessment to young borrowers and is better at capturing mortgage default relative to standard credit scoring models, while accurately tracking variations in systemic risk. We argue that these properties can provide valuable insights for the design of policies targeted at reducing consumer default and alleviating its burden on borrowers and lenders, as well as macroprudential regulation.

1.1 Introduction

The dramatic growth in household borrowing since the early 1980s has increased the macroeconomic impact of consumer default. Figure 1 displays total consumer credit balances in millions of 2018 USD and the delinquency rate on consumer loans starting in 1985. The delinquency rate mostly fluctuates between 3 and 4%, except at the height of the Great Recession when it reached a peak of over 5%, and in its aftermath when it dropped to a low of 2%. With the rise in consumer debt, variations in the delinquency rate have an ever larger impact on household and financial sector balances sheets. Understanding the determinants of consumer default and predicting its variation over time and across types of consumers can not only improve the allocation of credit, but also lead to important insights for the design of policies aimed at preventing consumer default or alleviating its effects on borrowers and lenders. They are also critical for macroprudential policies, as they can assist with the assessment of the impact of consumer credit on the fragility of the financial system.

This paper proposes a novel approach to predicting consumer default based on deep learning. We rely on deep learning as this methodology is specifically designed for prediction in environments with high dimensional data and complicated non-linear patterns of interaction among factors affecting the outcome of interest, for which standard regression approaches perform poorly. Our methodology uses the same information as standard credit scoring models, which are one of the most important factors in the allocation of consumer credit. We show that our model improves the accuracy of default predictions while increasing transparency and accountability. It is also able to track variations in systemic risk, and is able to identify the most important factors driving defaults and how they change over time. Finally, we show that adopting our model can accrue substantial savings to borrowers and lenders.

Credit scores constitute one of the most important factors in the allocation of consumer credit in the United States. They are proprietary measures designed to rank borrowers based on their probability of future default. Specifically, they target the probability of a 90 days past due delinquency in the next 24 months.[1] Despite their ubiquitous use in the financial industry, there is very little information on credit scores, and emerging evidence suggests that as currently formulated credit scores have severe limitations. For example, [5] show that during the 2007-2009 housing crisis there was a marked rise in mortgage delinquencies and foreclosures among high credit score borrowers, suggesting that credit scoring models at the time did not accurately reflect the probability of default for these borrowers. Additionally, it is well known that a substantial fraction of borrowers are unscored, which prevents them from accessing conventional forms of consumer credit.

The Fair Credit Reporting Act, a legislation passed in 1970, and the Equal Opportunity in Credit Access Act of 1974 regulate credit scores and in particular determine which information can be included and must be excluded in credit scoring models. Such models can incorporate information in a borrower's credit report, except age and location. These restrictions are intended to prevent discrimination by age and factors related to location, such as race.[2] The law also mandates that entities that provide credit scores make public the four most important factors affecting scores. In marketing information, these are reported to be

[1] The most commonly known is the FICO score, developed by the FICO corporation and launched in 1989. The three credit reporting companies or CRCs, Equifax, Experian and TransUnion have also partnered to produce VantageScore, an alternative score, which was launched in 2006. Credit scoring models are updated regularly. More information on credit scores is reported in Section 1.5 and Appendix A.9.

[2] Credit scoring models are also restricted by law from using information on race, color, gender, religion, marital status, salary, occupation, title, employer, employment history, nationality.

payment history, which is stated to explain about 35% of variation in credit scores, followed by amounts owed, length of credit history, new credit and credit mix, explaining 30%, 15%, 10% and 10% of the variation in credit scores respectively. Other than this, there is very little public information on credit scoring models, though several services are now available that allow consumers to simulate how various scenarios, such as paying off balances or taking out new loans, will affect their scores.

The purpose of our analysis is to propose a model to predict consumer default that uses the same data as conventional credit scoring models, improves on their performance, benefiting both lenders and borrowers, and provides more transparency and accountability. To do so, we resort to deep learning, a type of machine learning ideally suited to high dimensional data, such as that available in consumer credit reports.[3] Our model uses inputs as features, such as debt balances and number of trades, delinquency information, and attributes related to the length of a borrower's credit history, to produce an individualized estimate that can be interpreted as a probability of default. We target the same default outcome as conventional credit scoring models, namely a 90+ days delinquency in the subsequent 8 quarters. For most of the analysis, we train the model on data for one quarter and test it on data 8 quarters ahead, in keeping with the default outcome we are considering, so that our predictions are truly out of sample. We present a variety of performance metrics suggesting that our model has very strong predictive ability. Accuracy, that is percent of observations correctly classified, is above 86% for all periods in our sample, and the AUC-Score, a commonly used metric in machine learning, is always above 92%.

Our main comparison is between our model and a conventional credit score. By construction, credit scores only provide an ordinal ranking of consumers based on their default risk, and are not associated to a specific default probability. Yet, it is still possible to compare performance by assessing whether borrowers fall in different points of the distribution with the credit score compared to our model predictions. We find that our model performs significantly better than conventional credit scores. The rank correlation between realized default rates and the credit score is about 98%, where it is close to 1 for our model. Additionally, the Gini coefficient for the credit score, a measure of the ability to differentiate borrowers

[3] For excellent reviews of how machine learning can be applied in economics, see [69] and [9].

based on their credit score is approximately 81% and drops during the 2007-2009 crisis, while the Gini coefficient for our model is approximately 86% and stable over time. Perhaps most importantly, the credit score generates large disparities between the implied predicted probability of default and the realized default rate for large groups of customers, particularly at the low end of the credit score distribution. As an illustration, among Subprime borrowers, 17% display default behavior which is consistent with Near Prime borrowers and 15% display default behavior consistent with Deep Subprime. The default rates for Deep Subprime, Subprime and Near Prime borrowers are respectively 95%, 79% and 44%, so this misclassification is large, and it would imply large losses for lenders and borrowers in terms of missed revenues or higher interest rates. By contrast, the discrepancy between predicted and realized default rates for our model is never more than 4 percentage points for categories with at least a percent share of default risk.

Additional benefits of our approach when compared to conventional credit scoring models include better capturing mortgage default behavior, and providing more favorable risk assessment to young borrowers. In particular, we show that if we classify mortgage default into five categories (i.e., strategic, distressed, pay-down, cash-flow manager and no default), our risk assessment tracks actual default behavior closer in each of these groups. We also show that credit scores are indiscriminately low for young borrowers relative to our credit risk assessment, a feature that is particularly disadvantageous from a credit allocation perspective.

We also examine the ability of our model to capture the evolution of aggregate default risk. Since our data set is nationally representative and we can score all borrowers with a non-empty credit record, the average predicted probability of default in the population based on our model corresponds to an estimate of aggregate default risk. We find that our model tracks the behavior of aggregate default rates remarkably well. It is able to capture the sharp rise in aggregate default rates in the run up and during the 2007-2009 crisis and also captures the inversion point and the subsequent drastic reduction in this variable. With the growth in consumer credit, household balance sheets have become very important for macroeconomic performance. Having an accurate assessment of the financial fragility of the household sector, as captured by the predicted probability of default on consumer credit has become crucially important and can aid in macro prudential regulation, as well as for designing fiscal and monetary policy responses to adverse aggregate economic shocks. This is another advantage of our model compared to credit scores, since the latter only provides an ordinal ranking of consumers with respect to their probability of default. Our model can provide such a ranking but in addition also provides an individual prediction of the default rate which can be aggregated into a systemic measure of default risk for the household sector.

As a final application, we compute the value to borrowers and lenders of using our model. For consumers, the comparison is made relative to the credit score. Specifically, we compute the credit card interest rate savings of being classified according to our model relative to the credit score. Being placed in a higher default risk category substantially increases the interest rates charged on credit cards at origination and increasingly so as more time lapses since origination, whereas being placed in a lower risk category reduces interest rate costs. We choose credit cards as they are a very popular form of unsecured debt, with 74% of consumers holding at least one credit or bank card. In percentage of credit cards balances, average net interest rate expense savings are approximately 5% for low credit score borrowers. These values constitute lower bounds as they do not include the higher fees and more stringent restrictions associated with credit cards targeted to low credit score borrowers and the increased borrowing limits available to higher credit score borrowers. For lenders, we calculated the value added by using our model in comparison to not having a prediction of default risk or having a prediction based on logistic regression. We use logistic regression for this exercise as it is understood to be the main methodology for conventional credit scoring models. Over a loan with a three year amortization period, we find that the gains relative to no forecast are in the order of 60% with a 15% interest rate, while the gains for relative to a model based on logistic regression are approximately 3%. These results suggest that both borrowers and lenders would experience substantial gains from switching to our model.

Our analysis contributes to the literature on consumer default in a variety of ways. We are the first to develop a prediction model of consumer default using credit bureau data that complies with all of the restrictions mandated by U.S. legislation in this area, and we do so using a large and temporally extended panel of data. This enables us to evaluate model performance in a setting that is closer to the one prevailing in the industry and to train and test our model in a variety of different macroeconomic conditions. Previous contributions either focus on particular types of default or use transaction data that is not admissible in conventional credit scoring models.

The closest contributions to our work are [55], [23] and [78]. [55] apply a decision tree approach to forecast credit card delinquencies with data for 2005-2009. They estimate cost savings of cutting credit lines based on their forecasts and calculate implied time series patterns of estimated delinquency rates. [23] apply machine learning techniques to combined consumer trade line, credit bureau, and macroeconomic variables for 2009-2013 to predict delinquency. They find substantial heterogeneity in risk factors, sensitivities, and predictability of delinquency across lenders, implying that no single model applies to all institutions in their data. [78] examine over 120 million mortgages between 1995 to 2014 to develop prediction models of multiple states, such as probabilities of prepayment, foreclosure and various types of delinquency. They use loan level and zip code level aggregate information, and provide a review of the literature using machine learning and deep learning in financial economics. [59] predict mortgage default using use convolutional neural networks and emphasize the advantages of deep learning, but they do not evaluate their models out of sample the way we do. Finally, [60] reviews the recent literature on credit scoring, which is based on substantially smaller datasets than the one we have access to, and recommends random forests as a possible benchmark. However, we find that our hybrid model as well as our model components, a deep neural network and gradient boosted trees, improves substantially over random forests, possibly owing to recent methodological advances in deep learning, including the use of dropout, the introduction of new activation functions and the ability to train larger models.[4]

Our model is interpretable, which implies that we are able to assess the most important factors associated with default behavior and how they vary over time. This information is important for lenders, and can be used to comply with legislation that requires lenders and credit score providers to notify borrowers of the most important factors affecting their

[4] Other machine learning applications and reviews of default predictions include [68], [12], [20]. For deep learning applications see [81] and [78].

credit score. Additionally, it can be used to formulate economic models of consumer default. The literature on consumer default[5] suggests that the determinants of default are related to preferences, such as impatience which increases the propensity to borrow, or adverse expenditure of income shocks. Based on these theories, it is then possible to construct theoretical models of credit scoring, of which [26] is a leading example. We find that the number of trades and the balance on outstanding loans are the most important factors associated with an increase in the probability of default, in addition to outstanding delinquencies and length of the credit history. This information can be used to improve models of consumer default risk and enhance their ability to be used for policy analysis and design.

We also identify and quantify a variety of limitations of conventional credit scoring models, particularly their tendency to misclassify borrowers by default risk, especially for relatively risky borrowers. This implies that our default predictions could help improve the allocation of credit in a way that benefits both lenders, in the form of lower losses, and borrowers, in the form of lower interest rates. Our results also speak to the perils associated with using conventional credit scores outside on the consumer credit sphere. As it is well known, credit scores are used to screen job applicants, in insurance applications, and a variety of additional settings. Economic theory would suggest that this is helpful, as long as credit score provide information which is correlated with characteristics that are of interest for the party using the score ([34]). However, as we show, conventional credit scores misclassify borrowers by a very large degree based on their default risk, which implies that they may not be accurate and may not include appropriate information or use adequate methodologies. The broadening use of credit scores would amplify the impact of these limitations.

The paper is structured as follows. Section 1.2 describes our data. Section 1.3 discusses the patterns of consumer default that motivate our adoption of deep learning. Section 1.4 describes our prediction problem and our model. Section 1.5 compares our model to conventional credit scores. Section 1.6 illustrates our model's performance in predicting and quantifying aggregate default risk and calculates the value added of adopting our model over alternatives for lenders and borrowers.

[5] Some notable contributions include [27], [61], and [10].

1.2 Data

We use anonymized credit file data from the Experian credit bureau. The data is quarterly, it starts in 2004Q1 and ends in 2015Q4. The data comprises over 200 variables for an anonymized panel of 1 million households. The panel is nationally representative, constructed from a random draw for the universe of borrowers with an Experian credit report. The attributes available comprise information on credit cards, bank cards, other revolving credit, auto loans, installment loans, business loans, first and second mortgages, home equity lines of credit, student loans and collections. There is information on the number of trades for each type of loan, the outstanding balance and available credit, the monthly payment, and whether any of the accounts are delinquent, specifically 30, 60, 90, 180 days past due, derogatory or charged off. All balances are adjusted for joint accounts to avoid double counting. Additionally, we have the number of hard inquiries by type of product, and public record items, such as bankruptcy by chapter, foreclosure and liens and court judgments. For each quarter in the sample, we also have each borrowers's credit score. The data also includes an estimate of individual and household labor income based on IRS data. Because this is data drawn from credit reports, we do not know gender, marital status or any other demographic characteristic, though we do know a borrower's address at the zip code level. We also do not have any information on asset holdings.

Table 1 reports basic demographic information on our sample, including age, household income, credit score and incidence of default, which here is defined as the fraction of households who report a 90 or more days past due delinquency on any trade. This will be our baseline definition of default, as this is the outcome targeted by credit scoring models. Approximately 34% of consumers display such a delinquency.

1.3 Patterns in Consumer Default

We now illustrate the complexity of the relation between the various factors that are considered important drivers of consumer default. Our point of departure are standard credit scoring models. While these models are proprietary, the Fair Credit Reporting Act of 1970 and the Equal Opportunity in Credit Access Act of 1984 mandate that the 4 most important factors determining the credit scores be disclosed, together with their importance in determining variation in credit scores. These include credit utilization and number of hard inquiries, which are supposed to capture a consumer's demand for credit, the variety of debt products, which capture the consumer's experience in managing credit, and the number and severity of delinquencies. Each of these factors is stated to account for 10-35% of the variation in credit scores. The length of the credit history is also seen as a proxy on a consumer's experience in managing credit, and this is reported as accounting for 15% of the variation in credit scores.[6] The models used to determine credit scores as a function of these attributes are not disclosed, but they are widely believed to be based on linear and logistic regression as well as score cards. Additionally, available credit scoring algorithms typically do not score all borrowers.

Subsequently, we illustrate the properties of consumer default that suggest deep learning might be a good candidate for developing a prediction model. Specifically, we show that default is a relatively rare but very persistent outcome, there are substantial non-linearities in the relation between default and plausible covariates, as well as high order interactions between covariates and default outcomes.

1.3.1 Default Transitions

The default outcome we consider is a 90+ days delinquency, which occurs if the borrower has missed scheduled payments on any product for 90 days or more.[7] This is the default outcome targeted by the most widely used credit scoring models, which rank consumers based on their probability of becoming 90+ days delinquent in the subsequent 8 quarters. We refer to borrowers who are either current or up to 60 days delinquent on their payments as *current*.

[6] For an overview of the information available to borrowers about the determinants for their credit score, see `https://www.myfico.com/resources/credit-education/whats-in-your-credit-score`.

[7] For instance, if no payment has been made by the last day of the month within the past three months and the payment was due on the first day of the month three months ago. For credit cards, this occurs if the borrower does not make at least their minimum payment.

The transition matrix from current to 90+ days past due in the subsequent 8 quarters is given in Table 2. Clearly, the two states are both highly persistent, with a 77% of current customers remaining current in the next 8 quarters, and 93% of customers in default remaining in that state over the same time period. The probability of transition from current to default is 23%, while the probability of curing a delinquency with a transition from default to current is only 7%. These results suggest that default is a particularly persistent state, and predicting a transition into default is very valuable form the lender's perspective, since they are unlikely to be able to recuperate their losses. But it is also quite difficult, as the current state is also very persistent.

1.3.2 Non-linearities

Our model includes a relatively large list of features, which is presented in Table 27. The summary statistics for these features are reported in Table 28 in Section A.2.4. As is demonstrated in the table, there is a wide dispersion in the distribution of these variables. For example, the average balance on credit and bankcard trades is approximately $8,800, but the standard deviation, at $19,284, is more than twice as large. Similarly, average total debt balances are approximately $77,000, while the standard deviation is $170,000 and the 75th percentile $95,000, suggesting a high upper tail dispersion of this variable. Other features display similar patterns.

Figure 2 illustrates the highly non-linear relation between selected features and the incidence of default. In particular, it shows how the default rate, defined as the fraction of borrowers with a 90+ day past due delinquency in the subsequent 8 quarters, varies with total debt balances, credit utilization, the credit limit on credit cards, the number of open credit card trades, the number of months since the most recent 90+ day past due delinquency and the months since the oldest trade was opened. The figures show that while the relation between the features and the incidence of default is mostly monotone, it is highly nonlinear, with vary little variation in the incidence of default for most intermediate values of the variable and much higher or lower values at the extremes of the range of each covariate. The variables in the figure are just illustrative, a similar pattern holds for most plausible features.

1.3.3 High Order Interactions

Multidimensional interactions are another feature of the relation between default and plausible covariates, that is default behavior is simultaneously related with multiple variables. To see this, Figure 3 presents contour plots of the relation between the incidence of default and couples of covariates. The covariates reported here are chosen since they are important driving factors in default decisions, based on our model, as discussed in Section A.6.

Panels (a) and (b) explore the joint variation in the incidence of default with total debt balances, credit utilization (total debt balances to limits), and credit history. Blue values correspond to high delinquency rates while red values to low delinquency rates. As can be seen from both panels, higher credit utilization corresponds to higher delinquency rate, but for given credit utilization, an increase in total debt balances first decreases then increases the delinquency rate, where the switch in sign depends on the utilization rate. For given utilization rates, a longer credit history first increases then decreases the delinquency rate, provided the utilization rate is smaller than 1.[8] Panels (c) and (d) explore the relation between default and credit card borrowing. Default rates decline with the number of credit cards, though for a given number of credit card trades, they mostly increase with credit card balances. This relation, however varies with the level of both variables. An increase in the length of credit history is typically associated with lower default rates, however, if the number of open credit cards is low, this relation is non-monotone. The variables reported in the figures are illustrative of a general pattern in the joint relation between couples of covariates and default rates.

This pattern of multidimensional non-linear interactions across covariates is fairly difficult to model using standard econometric approaches. For this reason, we propose a deep learning approach to be explained below.

[8] Utilization rates above 1 can arise for a delinquent borrower if fees and other penalty add to their balances for given credit limits.

1.4 Model

Predicting consumer default maps well into a supervised learning framework, which is one of the most widely used techniques in the machine learning literature. In supervised learning, a learner takes in pairs of input/output data. The input data, which is typically a vector, represent pre-identified attributes, also known as features, that are used to determine the output value. Depending on the learning algorithm, the input data can contain continuous and/or discrete values with or without missing data. The supervised learning problem is referred to as a "regression problem" when the output is continuous, and as a "classification problem" when the output is discrete. Once the learner is presented with input/output data, its task is to find a function that maps the input vectors to the output values. A brute force way of solving this task is to memorize all previous values of input/output pairs. Though this perfectly maps the input data to the output values in the training data set, it is unlikely to succeed in forecasting the output values if (1) the input values are different from the ones in the training data set or (2) when the training data set contains noise. Consequently, the goal of supervised learning is to find a function that generalizes beyond the training set, so that it correctly forecasts out-of-sample outcomes. Adopting this machine-learning methodology, we build a model that predicts defaults for individual consumers. We define default as a 90+ days delinquency on any debt in the subsequent 8 quarters, which is the outcome targeted by conventional credit scoring models. Our model outputs a continuous variable between 0 and 1 that can be interpreted under certain conditions as an estimate of the probability of default for a particular borrower at a given point in time, given input variables from their credit reports.

We start by formalizing our prediction problem. We adopt a discrete-time formulation for periods 0,1,...,T, each corresponding to a quarter. We let the variable D_t^i prescribe the state at time t for individual i with $\mathrm{D} \subset \mathbb{N}$ denoting the set of states. We define $D_1^i = 1$ if a consumer is 90+ days past due on any trade and $D_1^i = 0$ otherwise. Consumers will transition between these two states over their lifetime. We allow the dynamics of the state process to be influenced by a vector of explanatory variables $X_{t-1}^i \in \mathbb{R}^{d_X}$, which includes the state D_{t-1}^i. In our empirical implementation, X_{t-1}^i represents the features in Table 27. Our target outcome is 90+ days past due in the subsequent 8 quarters, defined as:

$$Y_t^i = \begin{cases} 0 & \text{if } \sum_{n=t}^{t+7} D_n^i = 0 \\ 1 & \text{otherwise} \end{cases} \tag{1.1}$$

We fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and an information filtration $(\mathcal{F}_t)_{(t=0,1,\ldots,T)}$. Then, we specify a probability transition function $h_\theta : \mathbb{R}^{d_X} \to [0,1]$ satisfying

$$\mathbb{P}[Y_t^i = y | \mathcal{F}_{t-1}] = h_\theta(X_{t-1}^i), y \in D \tag{1.2}$$

where θ is a parameter to be estimated. Equation 1.2 gives the marginal conditional probability for the transition of individual i's debt from its state D_{t-1}^i at time $t-1$ to state y at time t given the explanatory variables X_{t-1}^i.[9] Let g denote the standard *softmax* function:

$$g(z) = \left(\frac{1}{1+e^{-z}} \right), z \in \mathbb{R}^K, \tag{1.3}$$

where $K = |D|$. The vector output of the function g is a probability distribution on D.

The marginal probability defined in equation 1.2 is the theoretical counterpart of the empirical transition matrix reported in Table 2. We propose to model the transition function h_θ with a hybrid deep neural network/gradient boosting model, which combines the predictions of a deep neural network and an extreme gradient boosting model. We explain each of the component models and their properties and the rationale for combining them below.

1.4.1 Deep Neural Network

One component of our model is based on deep learning, in the class used by [78]. We restrict attention to feed-forward neural networks, composed of an input layer, which corresponds to the data, one or more interacting hidden layers that non-linearly transform the data, and an output layer that aggregates the hidden layers into a prediction. Layers of the networks consist of neurons with each layer connected by synapses that transmit signals among neurons of subsequent layers. A neural network is in essence a sequence of nonlinear relationships. Each layer in the network takes the output from the previous layer and applies a linear transformation followed by an element-wise non-linear transformation.

[9] The state y encompasses realizations of the state between time t and $t+7$.

1.4.2 eXtreme Gradient Boosting (XGBoost)

The second component of our model is Extreme Gradient Boosting, which builds on decision tree models. Tree-based models split the data several times based on certain cut-off values in the explanatory variables.[10] Gradient Boosted Trees (GBT) are an ensemble learning method that corrects for tree-based models' tendency to overfit to training data by recursively combining the forecasts of many over-simplified trees. Though shallow trees are "weak learners" on their own with little predictive power, the theory behind boosting proposes that a collection of weak learners, as an ensemble, creates a single strong learner with improved stability over a single complex tree. For a more detailed description of our model components see Appendix A.3.

1.4.3 Hybrid DNN-GBT Model

We examined two techniques to create a hybrid DNN-GBT ensemble model. Ensemble models combine multiple learning algorithms to generate superior predictive performance than could be obtained from any of the constituent learning algorithms alone. The first method combines the two models by replacing the final layer of the neural network with a gradient boosted trees model. Examples of this approach are [28] and [73]. The second, uses both models separately and then averages out the final predicted probabilities of the two models. We found the latter to perform better on our dataset. This method is similar to [59], who combined a convolutional neural network with a random forest by averaging. Thus, our methodology relies on combining the output of the deep neural network with the output of a gradient boosted trees model. This is achieved in two steps:

1. For each observation, run DNN and GBT separately and obtain predicted probabilities for each of the models;

2. Take a weighted average of the predicted probabilities.[11]

[10] Splitting means that different subsets of the dataset are created, where each observation belongs to one subset.

[11] We have investigated several weighting schemes, and the results are reported in Table 31.

1.4.4 Implementation

Table 27 lists the features from the credit report data we use as inputs in the model. They include information on balances and credit limits for different types of consumer debt, severity and number of delinquencies, credit utilization by type of product, public record items such as bankruptcy filings, collection items, and length of the credit history. In order to be consistent with the restrictions of the Fair Credit Reporting Act on 1970 and the Equal Opportunity in Credit Access Act of 1984 we do not include information on age or zip code, and we do not include any information on income, to be consistent with current credit scoring models. Table 27 lists the full set of features used in our machine learning models. We describe the rationale behind our feature selection in Appendix A.4. Section A.5 provides a comprehensive performance assessment of our model, Section A.6 uses a variety of interpretability techniques to understand which factors are strongly associated with default behavior, while Appendix A.7 compares it to other approaches.

1.5 Comparison with Credit Score

In this section, we compare the performance of our hybrid model to a conventional credit score.[12] The credit score is a summary indicator intended to predict the risk of default by the borrower and it is widely used by the financial industry. For most unsecured debt, lenders typically verify a perspective borrower's credit score at the time of application and sometimes a short recent sample of their credit history. For larger unsecured debts, lenders also typically require some form of income verification, as they do for secured debts, such as mortgages and auto loans. Still, the credit score is often a key determinant of crucial terms of the borrowing contract, such as the interest rate, the downpayment or the credit limit. We have access to a widely used conventional credit score that uses information from the three credit bureaus.

[12] The hybrid model forecasts for each quarter are obtained using out-of-sample input data, as reported in Table 32.

1.5.1 Ranking

A common way to measure the accuracy of conventional credit scoring models is the Gini coefficient, which measures the dispersion of the credit score distribution and therefore its ability to separate borrowers by their default risk. The Gini coefficient is related to a key performance metric for machine learning algorithm, the AUC score, with $Gini = 2*AUC-1$, so we can compare the performance of the credit score to our model along this dimension. Figure 20 plots the Gini coefficient for the credit score and our predicted default probability by quarter. The Gini coefficient for our model is about 0.85 between 2006Q1 and 2008Q3, and then rises to 0.86. For the credit score, the Gini coefficient is close to 0.81 until 2012Q3 when it drops to approximately 0.79 until the end of the sample, suggesting a drop in performance of the credit score in the aftermath of the Great Recession.

Table 3 shows the relationship between credit score, predicted probability and realized default rate, where default is defined as usual as 90+ days delinquency in the subsequent 8 quarters. The calculation proceeds as follows. We first compute the number of unique credit scores in the data. We create the same number of bins of equal size in our predicted probability distribution, and calculate the realized frequency of 90+ days delinquencies in the subsequent 8 quarters for each of these bins. Since higher credit scores correspond to lower probability of default, we present the negative of the rank correlation with realized defaults for the credit score. The results indicate that even though credit score is successful in rank-ordering customers by their future default rates, with rank correlations between 0.980 and 0.994, our deep neural network performs better, with rank correlations always at 0.999. Figure 20 in Section A.9 plots the time series of these rank correlations by quarter for the entire sample period. The figure shows that the rank correlation for the predicted probability of default generated by our model is remarkably stable over time, while for the credit score it fluctuates from lows of around 0.975 before 2012 to a peak go 0.995 in 2013Q2 with notable quarter by quarter variation. This property of credit scores may be due to the fact that the credit score is an ordinal ranking and its distribution is designed to be stable over time, even if default risk at an individual or aggregate level may change substantially. Figure 19 in Section A.9 displays the histogram of credit score distributions in our sample for selected years, and show that these distributions are virtually identical over time.

Panel B of Table 3 reports the rank correlation with the realized default rate and the Gini coefficients by year for the credit score and the probability of default predicted by our model restricting attention to the current population, that is those borrowers who do not have any outstanding delinquencies in the quarter of interest. The rank correlation between the credit score and the realized default rate drops by 1-3 percentage points for these borrowers, whereas for our model it drops by less than a quarter of 1 percent. The Gini coefficient drops from 80-81% to 68-69% for the credit score and from 85-86% to 72-74% for our predicted probability. These results suggest that when measured on the population of current borrowers, the performance advantage of our model relative to a conventional credit score grows.[13]

Figure 4 plots a scatterplot of the realized default rate against the credit score (left panel) and our predicted probability (right panel) for all quarters in the year 2008. In addition to the raw data, we also plot second-order polynomial-fitted curves to approximate the relationship. The scatter plots of realized default rates against the predictions from our hybrid model lay mostly on the 45 degree line, consistent with the very high rank correlations reported in Table 3. By contrast, the relation between realized default rates and credit scores has an inverted S-shape, with the realized default rate equal to one for a large range of low credit scores and equal to zero for a large range of low credit scores, and a large variation only for intermediate credit scores.

Figure 5 plots second-order polynomial-fitted curves approximating the relation between realized default rates and those predicted by our model and the credit score for all years in which our model prediction is available, starting in 2006 until 2013, to examine how the relation between realized and predicted defaults varies with aggregate economic conditions. For the years at the height of the Great Recession, the default rate seems to be somewhat higher than our model prediction, but in all years the relation is very close to a 45 degree line. By contrast, there is virtually no change in the relation between the realized default

[13] Appendix A.9 also plots the time series of the rank correlation and the Gini coefficient for the credit score and our model for the current population. The credit score shows a large drop in these statistics for the credit score during the Great Recession whereas for our model they are both stable over time. This is consistent with the notion that the performance of the credit score dropped during the 2007-2009 period.

rate and the credit score. This is by construction, since the distribution of credit scores is designed to only provide a relative ranking of default risk across borrowers.[14] This property of the credit score implies that it is unable to forecast variations in aggregate default risk. In Section 1.6.1, we will show that our model is able to capture variations in aggregate default risk while retaining a consistent ability to separate borrowers by their individual default risk.

We next examine how the ranking of borrowers varies under credit score and our model to understand the differences in performance under the two approaches. To do so, we consider the industry classification of borrowers into five risk categories Deep Subprime, Subprime, Near Prime, Prime and Super Prime.[15] As shown in Table 4, these categories account for respectively, 6.5%, 21.2%, 14.1%, 33.3% and 24.9% of all borrowers. We then create 5 correspondingly sized bins in our predicted probability of default for each quarter separately with bin 1 corresponding to the 6.5% of borrowers with the highest predicted default risk and bin 5 to the 24.9% of all borrowers with the lowest predicted default risk. Finally, we calculate the fraction of borrowers in each credit score category that is in each of the 5 predicted default risk categories and their realized and predicted default rate. The results are displayed in Table 4. We also report the average realized and predicted default rate for each credit score category overall (columns 7 and 8) and for each predicted default risk category for all credit score (last 5 rows).

These results suggest that our model does well in predicting the default probability of borrowers in all categories, with a slight tendency to under-predict the probability of default by 1-5 percentage points for the Deep Subprime and Subprime borrowers. The majority of Deep Subprime borrowers fall in the two lowest categories of predicted default risk. For Subprime borrowers, 65% fall into the corresponding second category of default risk, while 15% fall in the first and 17% into the third. This corresponds to a sizable discrepancy as the average realized default probability for Subprime borrowers is 79%, whereas it is 95% for those in the first category and only 44% for those in the third. By contrast, the predicted default risk is very close to the realized default risk for Subprime borrowers in all categories

[14] Credit scores are specifically designed to provide a stable ranking by using multiple years of data.

[15] The threshold levels for these categories are: 1) Deep Subprime: up to 499 credit score; 2) Subprime: 500-600 credit score; 3) Near Prime: 601-660 credit score; 4) Prime: 661-780 credit score; 5) Super Prime: higher than 781 credit score.

of the predicted default risk distribution, with a discrepancy under 1 percentage point for predicted default risk category 1 and 2, and around 5 percentage points for category 3 and 4. Near Prime borrowers also display a wide dispersion across predicted default risk categories with only 43% falling into the corresponding third category, 24% falling in category 2 (higher default risk) and 31% falling into category 4 (lower default risk). Again, the realized default rates vary substantially for Near Prime borrowers by predicted default risk category, from 77% in category 2, to 41% and 20% in category 3 and 4, respectively, while the predicted default risk in much closer to the realized, with a maximum 3 percentage point discrepancy.

The discrepancy in classification for the credit score are lower for Prime and Super Prime borrowers. 13% of Prime borrowers fall into category 3 (higher default risk), 13% in category 5 (lower default risk) and 71% in the corresponding category 4. The realized default rates are 11% for Prime borrowers in category 4, and 34% and 3% respectively for Prime borrowers in category 3 and 5. Only 18% of Super Prime borrowers fall in category 4 of predicted default risk (higher risk) and 82% fall in the corresponding category 5. Moreover, the differences in realized default risk between these categories are minor, with a realized default rate of 6% and 2% for categories 4 and 5, respectively. These results suggest that credit scores misclassify borrowers across risk categories with very different realized default rates. By contrast, as shown in the bottom 5 rows of Table 4 and by columns (6) and (8), our model is very successful at predicting the default rate for borrowers irrespective of their credit score.

1.5.2 Feature Attribution

We next investigate feature attribution differences across credit scores and our hybrid model. We grouped our features into five categories to correspond to information we obtained from marketing resources for the credit score, and aggregated the absolute value of the SHAP values for each instance across each categories across our testing dataset for the pooled model. These categories are, payment history, amount owed, length of credit history, credit mix and new credit. Their contribution towards credit scores is reported in Table 5.

Contrary to credit scores, features relating to credit inquiries, debt products, and length of credit history account for 18% of the total variation in predicted probabilities for our hybrid model. The aggregate impact of these three factors is approximately half of the variation they explain of credit scores, which can partly be attributed to the low number of features we include in our models pertaining to these three groups. However, notice that even the per feature contribution is low for inquiries for our hybrid model.[16] Next, payment history accounts for 32% of the variation in predicted probabilities, being 3% short of its contribution towards credit scores. Perhaps most strikingly, accounts owed explain 50% of the variation for our hybrid model, while only 30% for credit scores. This exercise once again illustrates that features relating to debt balances are the most important determinants for our model's output, contrasted with credit scores, where payment history is registered as the most important predictor.

1.5.3 Vulnerable Populations

There has been a growing concern about the differential impacts of improved statistical technologies across categories such as age, race, gender, and income group. For instance, [43] found that though each group gains from improved predictive accuracy in creditworthiness (i.e., probability of mortgage default), Black and White Hispanic borrowers are predicted to lose, relative to White and Asian borrowers. They identified increased flexibility as the reason behind the unequal distribution of gains. Motivated by this finding, we compare the impacts of our technology relative to credit scores across age and income categories. To do so, we first rank our borrowers based on their percentile position in the predicted probability and credit score distribution respectively, and compute the difference.[17]. Then, to investigate which model provides better credit market access to the most vulnerable subgroup, the young with low income, we run regressions of the form:

$$Y_{ist} = \alpha + \beta_1 x_{1,ist} + \beta_2 x_{2,ist} + \beta_3 x_{1,ist} \times x_{2,ist} + \lambda_t + \delta_s + \delta_{st} + \epsilon_{ist} \tag{1.4}$$

where λ controls for any time varying changes common to all individuals (e.g., such as the Credit Card Act of 2009), while δ controls for any time invariant differences and for any

[16] The per feature contribution for credit inquiries is 0.006, contrasted with 0.01, 0.013, 0.014 and 0.016 for amounts owed, debt products, payment history, and length of credit respectively.

[17] Note that positive values imply lower credit risk by our model.

arbitrary trends across states (e.g., differential evolution of default rates across states over time). We include indicator variables for being under 30 years old (Young), for being in the bottom quintile of the income distribution (Income$_{p20}$), and for being in a county with above the median percentile unemployment shock in the quarter (e.g., measured by the difference between the unemployment rate in county c in quarter q and the average unemployment rate between 2000 and 2005 in county c). In most our specifications, we control for defaulting in the subsequent two years (Default) to ensure that differences are driven primarily by giving better access to credit to those who do not default on their loans.

We report descriptive statistics for this exercise in Table 6. We can see that young borrowers default on their loans at slightly higher rates, have lower debt and 90+ dpd debt balances, lower household income, and ranked lower in both the credit scores and our model's risk distribution. The cross-group differences are similar in direction but even larger in magnitude for the first income quantile vs. the rest comparison.

To complement Table 6, Panel (a) of Figure 6 plots the average difference across age and across default status, while Panel (b) displays these differences over time. We can see that the introduction of our hybrid model would mostly benefit young borrowers. For instance, the raw difference of 2 units for the youngest age group translates into improved credit ranking by 2 percentiles, which would improve their access to credit markets. Additionally, individuals who do not default would on average benefit across all age groups.

We report the regression results in Table 7. Columns (1) shows that precisely the most vulnerable subgroup, young individuals whose income is in the bottom quintile of the income distribution benefit the most from our credit risk model. In particular, our model would rank these individuals by 3.3 percentile higher on average in the credit risk distribution. We then control for default status, and while Column (2) shows that every borrower who do not end up defaulting would gain from our credit risk assessment, young and low income no-defaulters would be the largest beneficiaries, being ranked 5.2 percentile higher by our hybrid model. Columns (3-4) show that the benefits are similar in areas hard hit by unemployment shocks. This result holds during the Great Recession (i.e, 2007-2009), and the interpretation of this is that, relative to other groups, our model provides more favorable credit risk assessment for young and low income individuals, even in areas experiencing severe financial distress. Since credit scores do not provide an exact probability of default, we are unable to conclude whether our model would provide a larger number of borrowers access to credit (i.e., the extensive margin) across age and income groups in the borrower population. However, contrasting the relative ranking reiterates the notion that credit scores are indiscriminately low for young and low income borrowers.

1.5.4 Mortgage Default

We next compare our model with credit scores in predicting various categories of mortgage default. We restrict our sample to borrowers with mortgage debt and use the same percentile ranking as described in Section 1.5.3. We further restrict our sample to individuals who show new mortgage delinquency (i.e., current on all mortgage debt in the previous quarter), and who have debt outstanding aside from mortgage debt. We classify mortgage default into four categories: strategic, distressed, cash-flow, pay-down. We also include individuals who do not show any sign of mortgage delinquency throughout this period as our benchmark group.[18] The main caveat of this approach is our inability to distinguish between default due to income constraints and negative equity. For true strategic default, negative equity is a necessary condition, which we cannot control for. The magnitude of this bias was investigated by [19]. They show that between 2008 and 2011 the fraction of strategic defaulters accounted for 21% of all defaulters, of which 15.5% were truly strategic defaulters.

We report descriptive statistics for this exercise in Table 9. We can see that strategic defaulters have both the highest debt and 90 days late debt balances. One striking difference between our ranking and credit scores is the treatment of strategic defaulters. While credit scores rank them as the second least risky group, we rank them as the fourth.

We now estimate regressions as in Equation (1.4), with our interaction variables being default status and mortgage default type. Table 10 reports the results. In Column (1-3) our dependent variable is the quarter in which the borrower shows no delinquency

[18] These categories largely correspond to the work of Experian-Oliver Wyman (2009). Nonetheless, we are more stringent on classifying strategic default by requiring no delinquency on any other debt in the two consecutive quarters after showing mortgage delinquency, instead of requiring it only two quarters ahead. See `https://www.experian.com/assets/decision-analytics/reports/strategic-default-report-1-2009.pdf`.

on mortgage debt. Column (1-2) show that our model would rank strategic defaulters by approximately 2.5 percentiles lower on average in the quarter they yet to show a sign of mortgage delinquency. Given that our benchmark group is the "current" group, the constant in our regression illustrates that borrowers who stay current and do not default in the subsequent two years would be ranked 1 percentile higher on average. We next look at ranking differences in the quarter they become delinquent on their mortgage debt. Column (4) and (5) show that strategic defaulters would be ranked by approximately 5 percentile lower[19], and that all other groups who do not default would be ranked higher according to our model. Column (3) and (6) repeats the ranking exercise with lag and contemporaneous ranking for the Great Recession sample. The overall results are largely similar, however, we see an even larger gain for borrowers who do not default in the subsequent two years being ranked by approximately 2 percentile higher in the credit risk distribution on average. These results reaffirm the findings of [5], that is, credit scores did not accurately reflect the probability of default for a large group of mortgage owners during the 2007-2009 housing crisis. This misclasification is particularly severe for strategic defaulters.

1.6 Applications

In this section, we use our model in two applications. We first show that our model is able to accurately predict variations in aggregate default risk, and second, we illustrate the value added for lenders and borrowers from our hybrid model.

1.6.1 Predicting Systemic Risk

We first analyze the aggregate forecasting power of our hybrid model. We aggregate the deep-learning forecasts for individual accounts to generate macroeconomic forecasts of credit risk by taking the average of the predicted probabilities over a given forecast period. Since our sample of consumers in nationally representative in each quarter, this will provide

[19] There are no strategic defaulters who do not end up defaulting in the subsequent two years by definition.

an unbiased estimate of the aggregate default risk predicted by our model. We calculate the aggregate default probability for 2006Q1-2013Q4, and show that our model is able to predict the spike in delinquencies during the 2007-2009 financial crisis and also the reduction in delinquencies since then. This estimate of aggregate default risk could be used as a proxy of systemic risk in the household sector. The results are displayed in Figure 7.

Panel (a) plots the aggregate predicted default rate from our hybrid model and compares it to the aggregate realized default rate. While our predicted aggregate default rate is approximately 2 percentage points lower than the realized in 2006 and 2007, it rises at a similar speed as the realized default rate. It peaks in 2010Q2, approximately 2 quarters after the peak in the realized rate and then declines in the ensuing period, again reflecting the behavior of the realized rate, though it overestimates it by about 1 percentage point. Panel (b) shows a scatter plot of the predicted aggregate default rate against the realized for the different quarters in our sample period. The correlation between the predicted and realized aggregate default rate is 62%.

1.6.2 Value Added

We assess the economic salience of our hybrid DNN-GBT model by analyzing its value added for lenders and borrowers. For lenders, we examine the role our model can play in minimizing the losses from default. For borrowers, we calculate the interest savings for borrowers who are misclassified as having an excessively high probability of default based on the credit score compared to our model.

1.6.2.1 Lenders We follow the framework proposed by [55], which compares the value of having a prediction of default risk to having none, and we make the same simplifying assumptions with respect to the revenues and costs of the consumer lending business. Specifically, in absence of any forecasts, it is assumed a lender will take no action regarding credit risk, implying that customers who default will generate losses for the lender, and customers who are current on their payments will generate positive revenues from financing fees on their running balances. To simplify, we assume that all defaulting and non-defaulting customers have the same running balance, B_r, but defaulting customers increase their balance to B_d prior to default. We refer to the ratio between B_d and B_r as "run-up." It is assumed that with a model to predict default risk, a lender can avoid losses of defaulting customers by cutting their credit line and avoiding run-up. Then, the value added as proposed by [55] can be written as follows:

$$VA(r, N, TN, FN, FP) = \frac{TN - FN\big[1 - (1+r)^{-N}\big]\big[\frac{B_d}{B_r} - 1\big]^{-1}}{TN + FP} \tag{1.5}$$

where r refers to the interest rate, N the loan's amortization period, and TN, FN, FP refer to true negatives, false negatives and false positives respectively. Panel (a) of Figure 8 plots the Value Added (VA) as a function of interest rate and the ratio of run-up balance for our out-of-sample forecasts of 90+ days delinquencies over the subsequent 8 quarters for 2012Q4. These estimates imply cost savings of over 60% of total losses when compared to having no forecast model for a run-up of 1.2 at a 10% interest rate for an amortization period for 3 years.

We next compare the value added of our hybrid model with default predictions generated by a logistic regression. This exercise illustrates the gains from adopting a better technology for credit allocation. Panel (b) of Figure 8 shows more modest, but substantial cost savings in the range of 1-6% and approximately 2.5% for a 1.2 run-up at a 10% interest rate with a 3 year amortization period. Panel (c) calculates the cost savings associated to using our hybrid model in comparison to random forest. In this case the cost savings range from 0.1-0.7%. This exercise then confirms the advantages of using deep learning over other technologies in predicting default.

1.6.2.2 Borrowers

We now examine the potential cost savings for consumers who would be offered credit according to the predicted default probability implied by our model instead of a conventional credit score. Following our approach in Section 1.5, we create credit score categories based on common industry standards and corresponding predicted probability bins with the same number of observations for each quarter, and we place customers in these bins. The distribution of customers is summarized in Table 45. We then follow the information on interest rates by credit score category in Table 2 in [3] to obtain the cost of credit on credit card balances. [20]

To obtain the cost savings for consumers, we use the difference in interest rates by credit score category based on how they would be classified according to our model.[21] For customers who are placed in higher risk categories by the credit score compared to our predicted probability of default, interest rates on credit cards are higher than they would have been if they had been classified according to our model. Thus, using our model to score consumers rather than the credit score would generate the cost savings for them. For customers placed in risk categories by the credit score that are too low relative to the default risk predicted by our model, interest rates will be higher under our model. The calculation is made for each individual consumer. The average for each credit score category is then computed. The information on interest rates and balances, and the dollar value of cost savings for different credit card categories is reported in Table 11. We report this in current USD terms, since annual interest rate savings are symmetric by definition. The largest gains accrue to customers with Subprime and Near Prime credit scores. As we showed in Section 1.5, they are more likely to be attributed a probability of default by the credit score that is too low compared to our model predictions. Additionally, the biggest variation in credit card interest rates occurs across Subprime and Near Prime borrowers in comparison to Prime based on [3]. The cost savings for these borrowers average out to $1,104-1,401. Gains for Prime and Super Prime borrowers who are attributed a lower default probability by our model are very modest, as credit card interest rates vary little by credit score for Prime and Superprime borrowers. On the other hand, Prime and Superprime borrowers who are placed in group 1, corresponding to the highest predicted default probability based on our model face, substantial losses in the order of 4-5% of total credit card balances or $271-413. The cumulated interest rate cost savings across all consumers in our sample is $741,766,795, which amounts to $42 per capita.

This calculation provide us with a lower bound for the cost savings of being classified

[20] Credit card interest rates are notoriously invariant to overall changes in interest rates, so the calculations reported in this section apply irrespective of the time period. See [11] and [24].

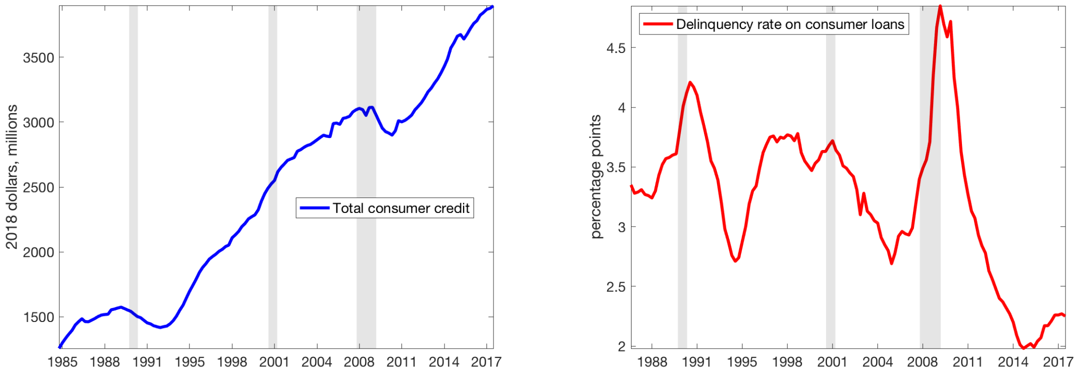
[21] We draw interest rates from a truncated normal distribution with mean and standard deviation as in [3].

according to our model in comparison to the credit score, as they do not take into account the higher credit limits and potential behavioral responses of customers faced with higher borrowing capacity and lower interest rates. As shown in [3], changes in the cost of funds for lenders mainly translate into changes in credit limits and exclusively for higher credit score borrowers. Therefore, being placed in a higher risk category for consumers also inhibits their ability to benefit from expansionary monetary policy. Additionally, we do not take into account the fact that more expensive credit in the form of higher interest rate costs makes it more likely that the consumer will incur missed payments in response to temporary changes in income. Fees for missed payments constitute a substantial component of credit card costs for consumers, and the ability to avoid these fees would contribute to substantial cost savings for consumers (see [2]).

1.7 Conclusion

We have proposed to use deep learning to develop a model to predict consumer default. Our model uses the same data used by conventional scoring models and abides with all legislative restrictions in the United States. We show that our model compares favorably to conventional credit scoring models in ranking individual consumers by their default risk, and is also able to capture variations in aggregate default risk. Our model is interpretable and allows to identify the factors that are most strongly associated with default. Whereas conventional credit scoring models emphasize utilization rates, our analysis suggests that the number and balances on open trades are the factors which associate more strongly to higher default probabilities. Our model is able to provide a default prediction for all consumers with a non-empty credit record. Additionally, we show that our hybrid DNN-GBT model performs better than standard machine learning models of default based on logistic regression and can accrue cost saving to lenders in the order of 1-6% compared to default predictions based on logistic regression, as well as interest rate cost savings for consumers of up to $1,401 per year.

1.8 Figures and Tables

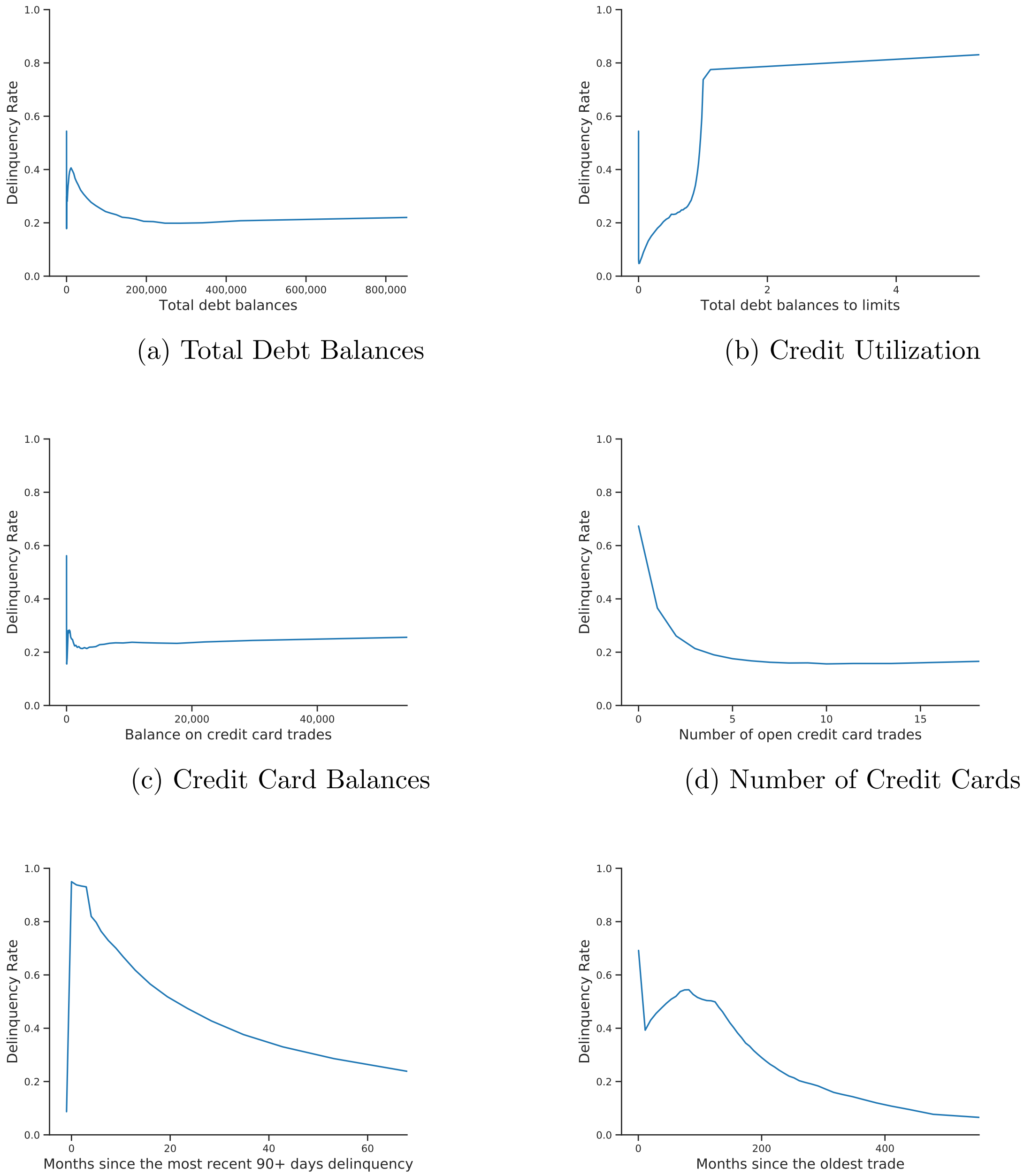


(a) Total Consumer Credit

(b) Delinquency Rate

Figure 1: Outstanding Consumer Credit and Delinquency over Time

Notes: Source: Author's calculations based on Federal Reserve Board data.



(a) Total Debt Balances

(b) Credit Utilization

(c) Credit Card Balances

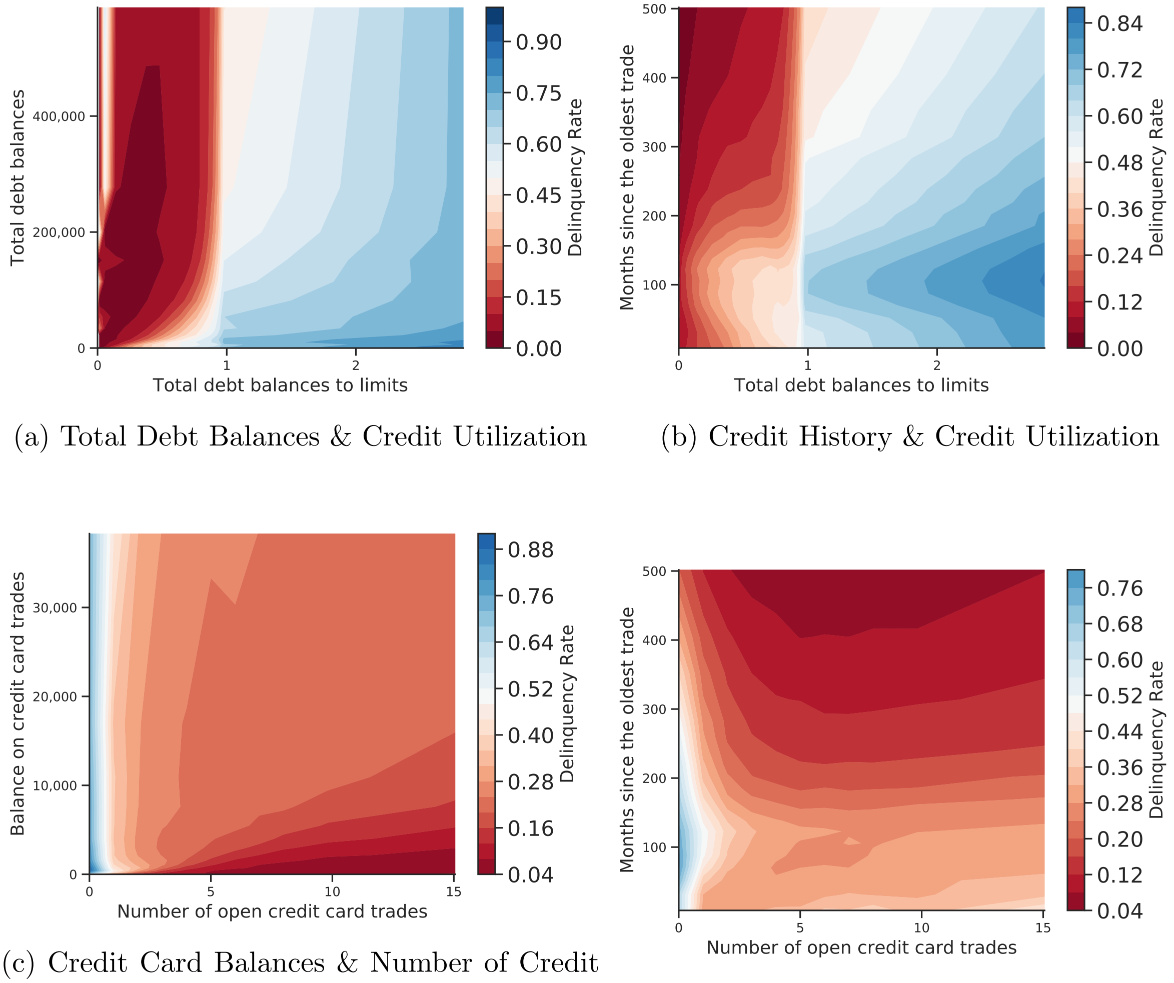
(d) Number of Credit Cards

(e) Proximity to Delinquency

(f) Length of Credit History

Figure 2: Nonlinear Relation Between Default and Covariates

Notes: Delinquency rate is the fraction with 90+ days past due trades in subsequent 8 quarters. In panel (e) and (f), -1 implies no past delinquency. Source: Authors' calculations based on Experian Data.



(a) Total Debt Balances & Credit Utilization

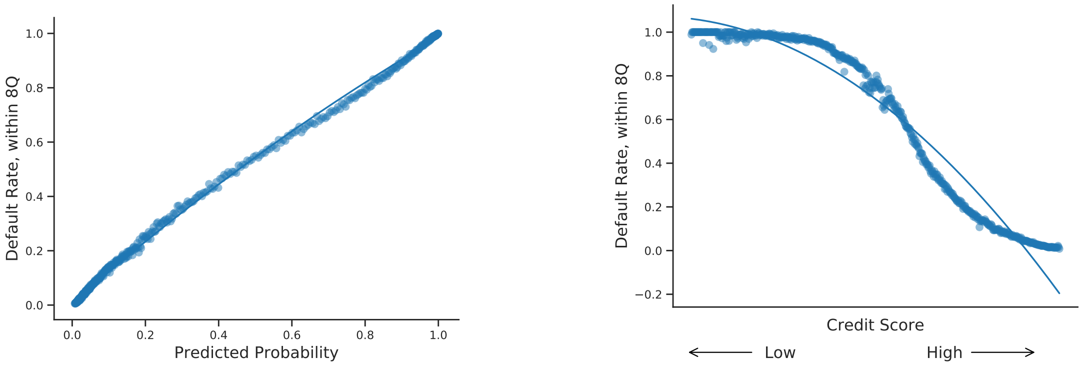
(b) Credit History & Credit Utilization

(c) Credit Card Balances & Number of Credit Cards

(d) Credit History & Number of Credit Cards

Figure 3: Multidimensional Relation Between Default and Covariates

Notes: Relationship between 90+ days past due delinquency rate and pairs of covariates. Source: Authors' calculations based on Experian Data.

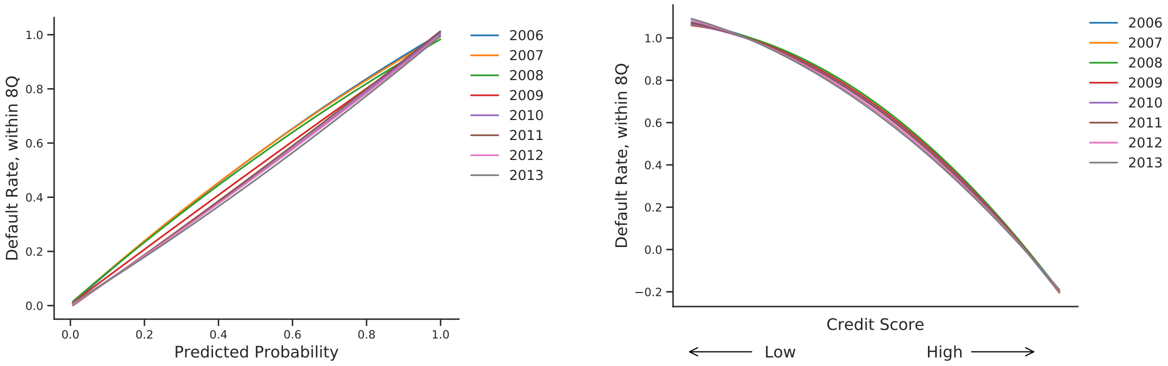


(a) Predicted Probability of Default

(b) Credit Score

Figure 4: Default Rates and Predicted Default Probability: Scatter Plot

Notes: Scatter plot of realized default rates against model predicted default probability (a) and the credit score (b), with associated second-order polynomial fitted approximations for the year 2008. Source: Authors' calculations based on Experian Data.

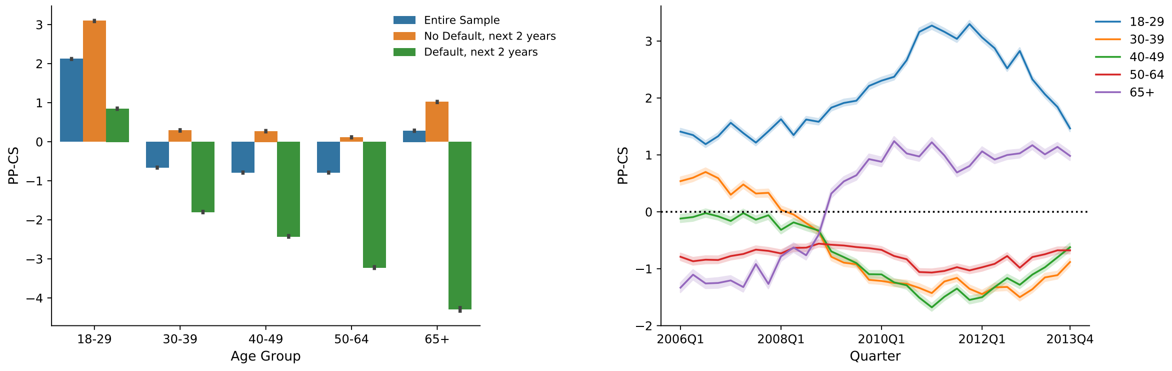


(a) Predicted Probability of Default

(b) Credit Score

Figure 5: Default Rates and Predicted Default Probability: Polynomial Approximation

Notes: Second-order polynomial approximation of the relationship between realized default rates against model predicted default probability (a) and the credit score (b) for selected years. Source: Authors' calculations based on Experian Data.

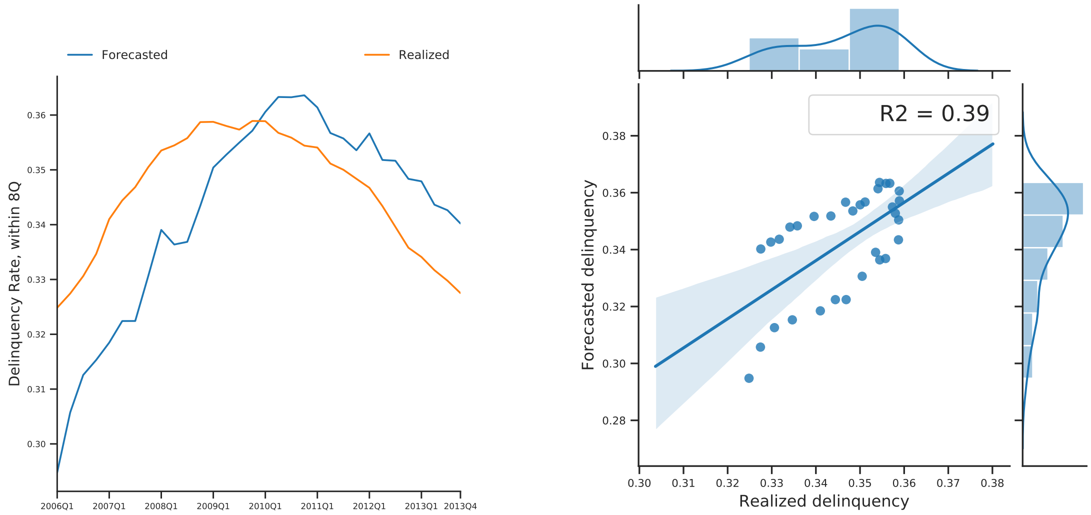


(a) Pooled

(b) Time Series

Figure 6: Differences in Creditworthiness by Age across and over Time

Notes: Source: Authors' calculations based on Experian Data.

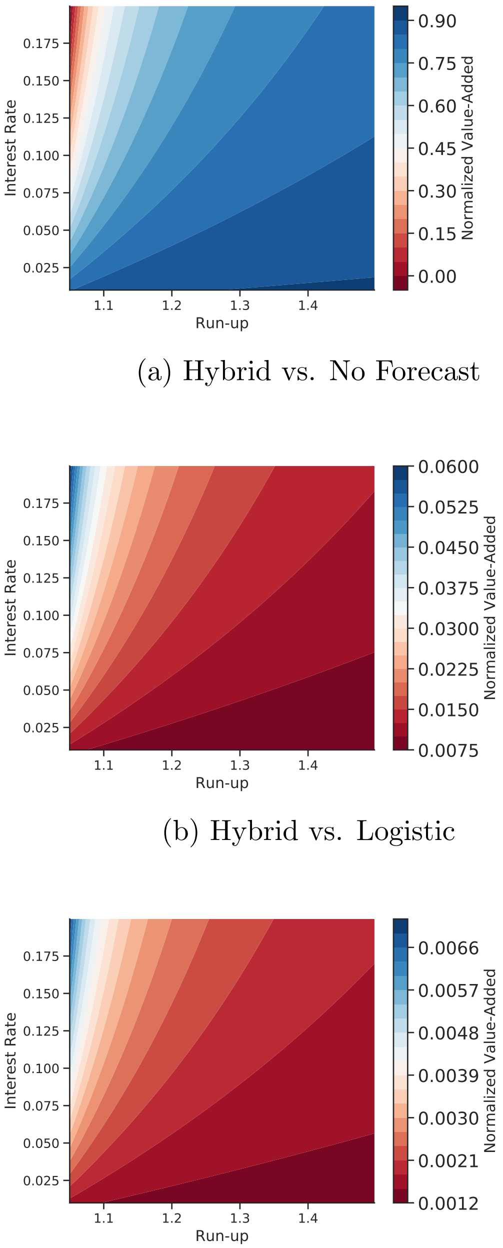


(a) Predicted and realized

(b) Correlation

Figure 7: Consumers with 90+ Days Delinquency: Predicted vs. Realized

Notes: Consumers with 90+ Days Delinquency within the Subsequent 8 Quarters. Aggregate default rates are obtained by averaging across all consumers in each period. Source: Authors' calculations based on Experian Data.



(a) Hybrid vs. No Forecast

(b) Hybrid vs. Logistic

(c) Hybrid vs. RF

Figure 8: Value-Added of Machine Learning Forecasts

Notes: Value-added of machine-learning forecasts of 90+ days delinquency over 8Q forecast horizons on data from 2012Q4. VA values are calculated with amortization period N = 3 years and a 50% classification threshold. Source: Authors' calculations based on Experian Data.

Table 1: Descriptive Statistics

Feature	Mean	Std. Dev.	Min	25%	50%	75%	Max
Age	45.8	16.3	18	32.2	45.1	57.8	83
Household Income	77.1	55.0	15	42	64	90	325
Credit Score	678.4	111.0	300	588	692	780	839
Default within 8Q	0.339	0.473	0	0	0	1	1

Notes: Credit score corresponds to Vantage Score 3. Household income is in USD thousands, trimmed at the 99th percentile. Source: Authors' calculations based on Experian Data.

Table 2: Default Transitions

Current/Next 8Q	No default	Default
No default	0.776	0.224
Default	0.073	0.927

Notes: Quarterly frequency of transition from current to default. Current corresponds to 0-89 day past due on any trade, Default corresponds to 90+ day past due on any trade in the subsequent 8 quarters. Source: Authors' calculations based on Experian Data.

Table 3: Borrower Rankings

Metric	2006	2007	2008	2009	2010	2011	2012	2013
PANEL A: Full Sample								
Rank Correlation								
Credit Score	0.9881	0.9804	0.9882	0.9816	0.9825	0.9861	0.9906	0.9944
Predicted Probability	0.9992	0.9994	0.9994	0.9993	0.9993	0.9993	0.9992	0.9992
GINI Coefficient								
Credit Score	0.8108	0.8142	0.8137	0.8143	0.8078	0.8008	0.7942	0.7898
Predicted Probability	0.8530	0.8529	0.8527	0.8598	0.8606	0.8592	0.8579	0.8563
PANEL B: Current Borrowers								
Rank Correlation								
Credit Score	0.9670	0.9613	0.9445	0.9653	0.9683	0.9585	0.9806	0.9559
Predicted Probability	0.9977	0.9983	0.9984	0.9978	0.9977	0.9977	0.9969	0.9973
GINI Coefficient								
Credit Score	0.6933	0.6935	0.6908	0.6807	0.6777	0.6833	0.6795	0.6810
Predicted Probability	0.7357	0.7242	0.7207	0.7178	0.7230	0.7324	0.7342	0.7373

Notes: Rank correlation between credit score, predicted probability of default according to our model and subsequent realized default frequency by year. Panel B only includes current borrowers, i.e., borrowers with no delinquencies. For the credit score, report the rank correlation between each unique value of the score and the default frequency. For predicted probability of default based on our hybrid DNN-GBT model, we first generate a number of bins equal to the number of unique credit score realizations in the data and then calculate the realized default frequency for each bin. Source: Authors' calculations based on Experian Data.

Table 4: Credit Risk Differences

Credit Score		Predicted Default Probability		Default Rate		Average Default Rate	
	Share		Share	Realized	Predicted	Realized	Predicted
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Deep Subprime	6.5%	1	49.5%	99.46%	99.20%	95.45%	94.40%
		2	49.5%	92.10%	90.47%		
		3	0.9%	63.69%	52.71%		
		4	0.0%	41.98%	13.79%		
		5	0.0%	37.14%	2.11%		
Subprime	21.2%	1	14.5%	99.21%	98.97%	78.64%	77.39%
		2	64.9%	84.34%	84.04%		
		3	16.7%	52.01%	46.78%		
		4	3.9%	21.58%	17.88%		
		5	0.0%	17.28%	2.46%		
Near Prime	14.1%	1	1.2%	98.97%	98.80%	43.71%	42.79%
		2	24.3%	76.60%	78.78%		
		3	43.0%	41.44%	40.59%		
		4	30.8%	19.69%	16.20%		
		5	0.7%	3.37%	2.93%		
Prime	33.3%	1	0.1%	98.82%	98.80%	14.31%	14.59%
		2	2.4%	75.15%	77.67%		
		3	13.1%	33.55%	36.68%		
		4	71.3%	10.67%	10.51%		
		5	13.1%	3.21%	2.67%		
Super Prime	24.9%	1	0.0%	99.04%	98.95%	2.56%	2.80%
		2	0.1%	81.19%	78.17%		
		3	0.2%	32.03%	34.54%		
		4	17.6%	5.56%	6.14%		
		5	82.1%	1.75%	1.92%		
All		1	6.5%	99.33%	99.08%		
		2	21.2%	83.92%	83.92%		
		3	14.1%	41.70%	40.96%		
		4	33.3%	11.45%	10.86%		
		5	24.9%	2.01%	2.05%		

Notes: Borrowers are classified into 5 categories of default risk standard in the industry named in column (1). The threshold levels for these categories are: 1) Deep Subprime: up to 499 credit score; 2) Subprime: 500-600 credit score; 3) Near Prime: 601-660 credit score; 4) Prime: 661-780 credit score; 5) Super Prime: higher than 781 credit score. The fraction of borrowers in each category is reported in column (2). Borrowers are also assigned to 5 categories of predicted default risk based on our hybrid model from the highest default risk (1) to the lowest (5), where the share of borrowers in each predicted default risk category is the same as for the credit score categories Deep Subprime to Super Prime. For each credit score risk category the share of borrowers in each predicted default risk category is reported in column (4). Columns (5) and (6) report the corresponding realized and predicted default probability for each credit score category interacted with predicted default risk category. Columns (7) and (8) report the average realized and predicted default probability for each credit score category. All rates, fractions and shares in percentage. Total # of observations: 17,732,772. Time period 2006Q1-2013Q4. Source: Authors' calculations based on Experian Data.

Table 5: Feature Attribution Differences

		Model	
Feature Group	# of Features	Hybrid	Credit Score
Payment History	23	0.32	0.35
Accounts Owed	50	0.50	0.3
Length of Credit	6	0.09	0.15
Debt Products	5	0.06	0.1
Inquiries	4	0.03	0.1

Notes: This table reports the Shapley values for five feature groups across four models. For each prediction window, we compute the Shapley value for each of the observations and for each feature. We then calculate the sum of the absolute value for each feature, aggregate it across the feature groups and report the results for the group. We normalized the results so that for each model the four groups sum up to 1. Source: Authors' calculations based on Experian data.

Table 6: Descriptive Statistics (Vulnerable Populations)

	Age: < 30	Age ≥ 30	t-test$_{Age}$	Income$_{p20}$	Income$_{>p20}$	t-test$_{Income}$
Default	0.43	0.32	-0.11***	0.65	0.27	-0.39***
	(0.50)	(0.47)		(0.48)	(0.44)	
Predicted Probability	39.26	53.49	14.23***	25.93	56.76	30.83***
	(22.76)	(29.57)		(18.49)	(27.75)	
Credit Score	37.14	54.05	16.91***	23.79	57.37	33.58***
	(22.94)	(29.24)		(17.79)	(27.19)	
Total debt ($)	25.38	93.15	67.77***	10.57	96.76	86.18***
	(68.87)	(192.75)		(40.29)	(192.98)	
90+ dpd debt ($)	1.51	4.11	2.60***	2.17	3.94	1.77***
	(16.23)	(39.93)		(16.27)	(39.88)	
Household Income ($)	61.11	99.33	38.22***	29.06	107.21	78.14***
	(423.20)	(315.58)		(6.43)	(380.03)	
Age	24.38	51.74	27.36***	32.22	49.62	17.40***
	(3.25)	(13.44)		(12.54)	(15.31)	
Unemployment Shock	2.15	2.21	0.06***	2.18	2.21	0.03***
	(2.37)	(2.39)		(2.43)	(2.37)	
N	3723837	14008935	17732772	3557963	14066498	17624461

Notes: *** denotes statistical significance at the 1% level. Income and debt variables in thousands.

Table 7: Vulnerable Populations

	(1)	(2)	(3)	(4)
Dependent Variable: Predicted Probability$_t$ - Credit Score$_t$				
Young	0.9633***	1.0077***	1.0273***	0.6923***
	(0.0350)	(0.0334)	(0.0449)	(0.0643)
Income$_{p20}$	1.5605***	3.2570***	3.2147***	2.5839***
	(0.0432)	(0.0473)	(0.0482)	(0.0624)
Young × Income$_{p20}$	1.4928***	0.7752***	0.9024***	1.0064***
	(0.0380)	(0.0346)	(0.0374)	(0.0526)
Default		-3.4441***	-3.4437***	-2.7325***
		(0.0440)	(0.0439)	(0.0601)
Unemployment Shock			0.1879***	0.4456***
			(0.0298)	(0.0448)
Young × Unemployment Shock			-0.0328	-0.1474**
			(0.0507)	(0.0730)
Income$_{p20}$ × Unemployment Shock			0.0955	-0.2791***
			(0.0611)	(0.0762)
Young × Income$_{p20}$ × Unemployment Shock			-0.2564***	-0.0522
			(0.0513)	(0.0679)
Constant	-0.7399***	0.1782***	0.0842***	-0.0560*
	(0.0122)	(0.0116)	(0.0186)	(0.0297)
State Fixed Effects	X	X	X	X
Quarter Fixed Effects	X	X	X	X
State X Quarter Fixed Effects	X	X	X	X
Sample: 2007-2009				X
Observations	17624453	17624453	17395749	6327506
R^2	0.0191	0.0373	0.0374	0.0276

Notes: This table reports regressions of the form specified by Equation (1.4). Model (1) does not control for default, Model (2) adds an indicator variable for default, while Model (3) and (4) include an indicator variable of being in a county with over the median unemployment shock as an additional dimension of interaction. We measure unemployment shock by the difference between unemployment rate in county c in quarter t and the average unemployment rate in county c between 2000 and 2005. Standard errors are clustered by state and quarter. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. To mitigate the influence of outliers, we winsorized are dependent variable at the at the 0.1 and 99.9 percentiles.

Table 8: Mortgage Default Categories

	Mortgage Debt		Other Debt
Default Type	Behavior$_t$	$\sum_{i=t+1}^{t+2}$ Behavior$_i$	$\sum_{i=t+1}^{t+2}$ Behavior$_i$
Strategic Default	30+ dpd	180+ dpd	current
Cash-flow Manager	30+ dpd	30-179 dpd	current
Distressed Default	30+ dpd	30-180+ dpd	30-180+ dpd
Pay-down	30+ dpd	current	
Current	current	current	

Table 9: Descriptive Statistics (Mortgage Default)

	Strategic	Cash-flow	Distressed	Pay-down	Current
Default	1.00	0.76	0.98	0.62	0.15
	(0.00)	(0.43)	(0.15)	(0.49)	(0.36)
Predicted Probability$_t$	21.79	23.26	13.09	22.03	65.51
	(9.35)	(9.33)	(8.65)	(12.50)	(23.11)
Credit Score$_t$	26.40	24.76	13.61	22.53	65.45
	(11.51)	(11.68)	(9.42)	(13.32)	(21.74)
Predicted Probability$_{t-1}$	44.17	36.49	21.81	32.78	65.42
	(17.67)	(16.66)	(13.72)	(19.58)	(23.07)
Credit Score$_{t-1}$	46.30	38.19	23.02	34.81	65.28
	(18.60)	(17.52)	(13.39)	(19.19)	(21.72)
Total debt	460.63	285.21	236.11	223.05	208.61
	(508.51)	(310.34)	(264.74)	(264.67)	(248.42)
90+ dpd debt	84.12	23.12	26.67	10.90	0.27
	(201.19)	(91.92)	(97.58)	(54.92)	(5.34)
Household Income	115.95	96.38	81.28	92.77	117.46
	(233.98)	(202.31)	(233.51)	(212.28)	(134.06)
Age	45.59	46.45	44.73	46.50	49.78
	(12.24)	(11.46)	(11.08)	(11.41)	(12.53)
N	4919	17459	46084	27291	4978377

Notes: Income and debt variables in thousands ($).

Table 10: Comparison across Mortgage Default Types

	(1)	(2)	(3)	(4)	(5)	(6)
	DV = PP_{t-1} - CS_{t-1}			DV = PP_t - CS_t		
Strategic	-2.3961***	1.5250***	1.4596***	-4.6024***	-0.1025	1.1758***
	(0.1808)	(0.1788)	(0.2657)	(0.2866)	(0.2695)	(0.3231)
Cash-flow	-1.5535***	-1.7307***	-1.4683***	-1.1805***	-0.1334	0.9873***
	(0.0846)	(0.1669)	(0.2714)	(0.1529)	(0.1713)	(0.2638)
Distressed	-1.2801***	0.5174*	-0.2224	-0.4365***	2.0476***	2.1810***
	(0.0601)	(0.3012)	(0.4567)	(0.0548)	(0.2839)	(0.4308)
Pay-down	-2.0488***	-1.6632***	-1.2766***	-0.4336***	-0.6702***	0.9276***
	(0.0688)	(0.1111)	(0.1700)	(0.0806)	(0.1474)	(0.1762)
Default		-4.6768***	-3.7876***		-5.3672***	-4.4260***
		(0.0781)	(0.1099)		(0.0790)	(0.1152)
Cash-flow × Default		3.9401***	3.3709***		2.8715***	2.5233***
		(0.2029)	(0.3249)		(0.2095)	(0.3216)
Distressed × Default		2.0404***	2.2633***		1.9118***	1.2084***
		(0.3106)	(0.4752)		(0.2935)	(0.4466)
Pay-down × Default		2.8285***	1.8904***		4.3412***	2.2444***
		(0.1521)	(0.2309)		(0.1764)	(0.2097)
Constant	0.1447***	0.8709***	1.9477***	0.0595***	0.8911***	1.8169***
	(0.0008)	(0.0126)	(0.0193)	(0.0013)	(0.0125)	(0.0196)
State Fixed Effects	X	X	X	X	X	X
Quarter Fixed Effects	X	X	X	X	X	X
State X Quarter Fixed Effects	X	X	X	X	X	X
Sample: 2007-2009			X			X
Observations	4895989	4895989	2078434	5074130	5074130	2079953
R^2	0.0319	0.0516	0.0337	0.0324	0.0583	0.0398

Notes: This table reports regressions of the form specified by Equation (1.4). PP and CS refer to percentile rank in predicted probability and credit score respectively. The dependent variable in Model (1-3) is difference in previous quarter ranking, while in Model (4-6) it is current quarter. Model (1) and (4) does not control for default, Model (2-3) and (5-6) add an indicator variable for default and interacts the mortgage default type with it. Standard errors are clustered by state and quarter. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. To mitigate the influence of outliers, we winsorized are dependent variable at the at the 0.1 and 99.9 percentiles.

Table 11: Cost of Credit Risk Misclassification

			Credit Score			
			Subprime, Near Prime	Prime Low	Prime Mid	Prime High, Superprime
			Annual Average Cost Saving ($)			
Predicted bin	Default	1	0	-413	-271	-299
		2	1104	0	84	15
		3	1401	-171	0	-56
		4	1268	-56	118	0

Notes: This table reports the average cost savings for consumers across credit score and predicted default probability bins for our sample. The cumulative savings for consumers on both credit card and bankcard debt adds up to $741,766,795. Time period 2006Q1-2013Q4. Source: Authors' calculations based on Experian Data.

2.0 Towards Fair Credit Allocation

Does a better statistical technology lead to worse outcomes for disadvantaged groups? Not necessarily. In this paper, we use constrained deep learning to predict consumer default, and show that by incorporating protected group status during training, the credit access for borrowers belonging to disadvantaged groups can be improved, while providing more accurate credit risk assessment. Additionally, we show that our constrained deep learning model performs better than standard machine learning models of default based on logistic regression and can accrue cost saving to lenders in the order of 0-2.5% compared to default predictions based on logistic regression. Taken together, our results suggest that introducing constraints can be effective at promoting better credit access to protected groups, while yielding substantial cost savings to lenders.

2.1 Introduction

New predictive statistical technologies and machine learning techniques are being eagerly implemented by profit maximizing businesses in a wide range of industries. The rapid of adoption of these technologies has led to concerns that society has not thoroughly evaluated the risks associated with their use. Indeed, intelligent technology can amplify the deepest divides and inequalities known to humans, but it can also promote inclusion and fairness. We need our technologies to produce inclusive and accessible outcomes, and in this paper, we study how to do so in the domain of household credit. As [33] put it:

> At the heart of it all, intelligent machines can help address the inequities in life. We need to resist the capitalistic impulse to try to maximize gains without paying attention to the negative consequences to society. In doing so, we can help shift our policy-makers from arguing over partisan issues to promoting our collective well-being. In the Intelligent Machine Age, free of mundane tasks; with fewer barriers to success for the disadvantaged, and less incentive for excessive personal gain; with more time, dedication, and purpose; in partnership with extraordinarily smart AI, we can choose to improve not just our own lives but the lives of many. . .

The key idea underlying our paper is that a more sophisticated statistical technology can be constrained to preserve certain favorable aspects of the more primitive technology, while improving on overall performance. Specifically, we enforce that the proportion of applicants wrongly rejected be lower than under the primitive technology across groups. In our context, the insight that this yields is that improvements in predictive technology does not necessarily worsen the credit access and need not to disproportionately impact already disadvantaged groups, when specifically instructed not to do so. This implies that even though there will always be some borrowers deemed less risky ("winners") by the better technology, while other borrowers will be considered riskier ("losers") relative to their position under the prevailing technology, at the group level, the new technology will not lead to worse outcomes. Hence, it is possible to train algorithms to more evenly distribute the gains from improved statistical technology. Nonetheless, the impacts before implementing such technologies need to be carefully considered across societally important categories such as race or age.

In this paper, we study how to increase fairness in the allocation of consumer credit. The work on fairness in credit scoring so far has focused only on whether credit scoring algorithms de facto learn whether a borrower belongs to a protected category and whether better prediction algorithms further penalize protected categories ([43]). We seek to address whether increased statistical accuracy necessarily benefits disadvantaged groups less. To do so, we will use recent methodological developments in constrained machine learning to ensure that the allocation of credit reflects the true probability of default and is not biased against protected groups. Doing this we obtain a more accurate, yet fairer prediction algorithm.

Credit scores constitute one of the most important factors in the allocation of consumer credit in the United States. They are proprietary measures designed to rank borrowers based on their probability of future default. Specifically, they target the probability of a 90 days past due delinquency in the next 24 months.[1] Despite their ubiquitous use in the financial industry, there is very little information on credit scores, and emerging evidence suggests that as currently formulated credit scores have severe limitations. For example, [5] show that

[1] The most commonly known is the FICO score, developed by the FICO corporation and launched in 1989. The three credit reporting companies or CRCs, Equifax, Experian and TransUnion have also partnered to produce VantageScore, an alternative score, which was launched in 2006. Credit scoring models are updated regularly.

during the 2007-2009 housing crisis there was a marked rise in mortgage delinquencies and foreclosures among high credit score borrowers, suggesting that credit scoring models at the time did not accurately reflect the probability of default for these borrowers. Additionally, it is well known that a substantial fraction of borrowers are unscored, which prevents them from accessing conventional forms of consumer credit.

The Fair Credit Reporting Act, a legislation passed in 1970, and the Equal Opportunity in Credit Access Act of 1974 regulate credit scores and in particular determine which information can be included and must be excluded in credit scoring models. Such models can incorporate information in a borrower's credit report, except age and location. These restrictions are intended to prevent discrimination by age and factors related to location, such as race.[2] Thus, from a legal standpoint, credit scoring models are restricted by law from using information on race, color, gender, religion, marital status, salary, occupation, title, employer, employment history, nationality. The reason behind this legislation was to promote fair credit allocation, but as we show, incorporating features pertaining to protected group status can improve their credit access.

The definition of "fair" is context dependent. Literature in computer science has recently devoted considerable attention to the notion of fair or ethical machine learning. One strand of this literature developed mathematical fairness criteria according to societal and ethical notions of fairness ([38], [47], [76], [58]), while another strand developed methods for building machine-learning models that satisfy such fairness criteria (see, e.g., [86] and [35]). As in this paper, mathematical fairness criteria are often group-based, where a target metric is enforced over subgroups in the data, also referred to as protected groups. For instance, [47] introduced the equality of opportunity criterion, which specifies that the true positive rates for a binary classifier are equalized across protected groups. Hence, according to computer science literature, it might be beneficial to use some information on protected groups status.

Algorithmic notions of fairness include: anti-classification, classification parity and cal-

[2] The law also mandates that entities that provide credit scores make public the four most important factors affecting scores. In marketing information, these are reported to be payment history, which is stated to explain about 35% of variation in credit scores, followed by amounts owed, length of credit history, new credit and credit mix, explaining 30%, 15%, 10% and 10% of the variation in credit scores respectively. Other than this, there is very little public information on credit scoring models, though several services are now available that allow consumers to simulate how various scenarios, such as paying off balances or taking out new loans, will affect their scores.

ibration. When applied to credit scoring models, anti-classification simply requires that protected attributes not be used in the model, and it is already embedded in US legislation. Classification parity requires that predicted performance of credit scoring models be equal across different groups, whereas calibration requires that outcomes that are determined as a function of the credit scoring models be independent of protected attributes conditional on risk estimates. Clearly, anti-classification does not imply classification parity or calibration, which is the most stringent of these requirements and involves both the risk assessment and the way lenders use it to make lending decisions. Nonetheless, little public information is available to assess classification parity for conventional credit scoring models. Some studies found, that if anything, blacks and individuals residing in lower-income tracts show higher incidences of poor performance than credit scores would predict ([74]). On the other hand, there is a large body of evidence that outcomes such as access to and price of credit are disparate across groups (see [32], [8], [25], [54], and [74]). For instance, [74] documents that blacks receive a larger inferred denial rate than their credit score would predict, and face higher interest rates than non-Hispanic whites. If possible, differences in credit scores are often used to justify such disparities, even if without classification parity, differences in credit scores across groups may not be meaningful. Even with classification parity, historically biased features may lead to low credit scores for some groups and there may be a potential for a self-confirming cycle of low credit scores, costly credit and high default rates. Our paper intends to make progress in improving fairness with respect to age and race. We investigate age, since our data contains this information, and young borrowers have indiscriminately low credit scores. We focus on race as it is the area where the potential for lack of fairness is likely to be greatest due to historical disadvantages and with our data we are able to construct a reliable proxy for race.

To investigate disparities in credit access, we first generate groups based on age and race. Our data contains information on the individual's age, and for age, we create two groups: (1) individuals who are younger than 30 years old, and (2) individuals older than 30 years old. Since we do not have exact racial information, we create a race proxy leveraging information on the zip code where an individual resides and their age to estimate their race. [46] apply a ZIP code proxy for race and were able to improve on fairness metrics on the true groups even with weakly related proxy groups. While having actual information on race for each consumer would be preferable, we believe adopting a race proxy approach can still provide valuable insights on the fairness question.

We contribute to our understanding of credit markets by introducing pioneering advances in constrained machine learning to develop predictions of consumer default that are fair, in the sense that they guarantee that the credit allocation is not too disparate for borrowers in protected groups. Our preliminary notion of fairness pertains to false positive rates, imposing that the better statistical technology have lower false positive rates than the worse one for each subgroups, so that the better prediction algorithm will not worsen the credit market access for already disadvantaged groups.

Our research has direct and important implications for policies and regulations pertaining to consumer debt. We show a variety of limitations of conventional credit scoring models, particularly their tendency to misclassify borrowers in our protected groups. These results can be used to improve credit scoring models and generate substantial savings for both borrowers and lenders on consumer credit markets, while ensuring that a better statistical technology do not harm the credit market access for disadvantaged groups. Finally, our analysis can be used to mitigate disparities in access to consumer credit and thus reduce inequality and improve social welfare.

2.2 Data

We use anonymized credit file data from the Experian credit bureau. The data is quarterly, it starts in 2004Q1 and ends in 2015Q4. The data comprises over 200 variables for an anonymized panel of 1 million households. The panel is nationally representative, constructed from a random draw for the universe of borrowers with an Experian credit report. The attributes available comprise information on credit cards, bank cards, other revolving credit, auto loans, installment loans, business loans, first and second mortgages, home equity lines of credit, student loans and collections. There is information on the number of trades for each type of loan, the outstanding balance and available credit, the monthly payment, and whether any of the accounts are delinquent, specifically 30, 60, 90, 180 days past due, derogatory or charged off. All balances are adjusted for joint accounts to avoid double counting. Additionally, we have the number of hard inquiries by type of product, and public record items, such as bankruptcy by chapter, foreclosure and liens and court judgments. For each quarter in the sample, we also have each borrowers's credit score. The data also includes an estimate of individual and household labor income based on IRS data. Because this is data drawn from credit reports, we do not know gender, marital status or any other demographic characteristic, though we do know a borrower's address at the zip code level. We also do not have any information on asset holdings.

2.2.1 Generating Groups

We generate subgroups along age and racial dimensions. First, we have information on each individuals age in our data. We create two subgroups: (1) individuals who are younger than 30 years old, and (2) individuals older than 30 years old. Second, our race proxy uses information on the zip code where an individual resides and their age to estimate their race. This method assigns a single proxy probability for race using Bayes' rule. Specifically, we first obtain information on the racial and ethnic composition of the U.S. population by zip code from Census 2010, which provides counts of enumerated individuals by race and ethnicity for various geographic area definitions across different age groups, including zip codes. Then, we select the racial composition of the age group that corresponds to the individual's age. We then create two subgroups: (1) individuals with $\geq 25\%$ probability of being black, and (2) individuals with $< 25\%$ probability of being black.

2.2.2 On Race Proxies

The absence of racial/demographic data has led researchers to develop techniques to estimate race/ethnicity indirectly from other sources. One of these methods is geocoding analysis. Geocoding connects a person's address to a census measure of their neighborhood's racial/ethnic composition and leverages that measure as the probabilities of an the person belonging to each of the major racial/ethnic groups. In the US, blacks are typically concentrated in certain neighborhoods ([41]); in such areas, geocoding alone can do a fairly decent job at distinguishing blacks from whites.[3] For instance, [39] find that geolocations alone predictions correlate with individual indicators at 0.57 for the black population. Incorporating age likely yields to marginal improvements.

[46] apply a similar proxy for race and were able to improve on fairness metrics on the true groups even with weakly related proxy groups. We acknowledge the shortcomings of using geocoding analysis, namely, that differences in credit scores across groups are smaller when compared to using actual racial information ([74]).[4] While having actual information on race for each consumer would be preferable, we believe adopting a race proxy approach can still provide valuable insights on the fairness question. For what follows, we use protected groups when we refer to young or black borrowers.

2.3 Motivating Evidence

There has been a growing concern about the differential impacts of improved statistical technologies across categories such as age, race, and gender. For instance, [43] found that though each group gains from improved predictive accuracy in creditworthiness (i.e., probability of mortgage default), Black and White Hispanic borrowers are predicted to lose, relative to White and Asian borrowers. They identified increased flexibility as the reason behind the unequal distribution of gains. Motivated by this finding, we compare the impacts of improved statistical technology across age and racial groups. To do so, we first predict the probability of default based on a logistic regression (worse technology), and then by a deep neural network (better technology). We then compute differences in predicted probabilities and false positives across the two models. The first (differences in predicted probabilities) tells us which model finds our protected groups riskier on average, while the second (differences in false positive rates) illustrates which model is more restrictive in credit access. Now,

[3] Unlike Blacks, Hispanics, Asians, and many Native Americans reside in significantly less segregated neighborhoods ([62]). For them, geocoding alone does poorly.

[4] When geocoding was used to estimate race, the mean difference in the credit scores between blacks and non-Hispanic whites decreased from 28.4 points to 15.1 points in [74]).

to investigate which model provides better credit market access to our protected groups, we run regressions of the form:

$$Y_{ist} = \alpha + \beta_1 x_{1,ist} + \beta_2 x_{2,ist} + \beta_3 x_{1,ist} \times x_{2,ist} + \lambda_t + \delta_s + \epsilon_{ist} \tag{2.1}$$

where λ controls for any time varying changes common to all individuals (e.g., such as the Credit Card Act of 2009), while δ_s controls for any time invariant differences across states (e.g., different base rates of default across states). We include indicator variables ($x_{1,ist}$) for being under 30 years old (Young) and for having an over 25% probability of being black (Black $\geq$ 25%). In some of our specifications, we control for defaulting ($x_{2,ist}$) in the subsequent two years (Default) to ensure that differences are driven primarily by giving better access to credit to those who do not default on their loans.

Table 12 presents the results. Panel A shows that our DNN assigns higher predicted probabilities of default for our protected groups, while Panel B shows that the false positive rates (individuals who are incorrectly rejected credit) are also higher under our deep neural network. For instance, the results in Column (1) of Panel A indicate that our DNN typically predicts lower credit risk on average (constant is negative), unless the individual is young, in which case it predicts 2.1% higher probability of default on average. Similarly, Panel B of Column (1) shows that the false positive rates are on average lower for the DNN, unless the individual is young, for this group the DNN has a 1% higher false positive rate than the logistic. Similarly, we find effects of the same direction for individuals who have a probability of over 25% of being black. We find that these differences cannot be explained by including credit report controls, and state and quarter fixed effects (Column (2), Column (6)), by controlling for realized default (Column (3), Column (7)), and by interacting protected group status with realized default (Column (4), Column (8)).

The results in Table 12 show that an improved technology may incorrectly reject applicants at a higher rate than the worse one. This finding is in line with [43], and motivates our paper to build a better statistical technology that preserves (or improves) the access to credit for our protected groups.

2.4 Constrained Machine Learning: A Conceptual Framework

We now provide a conceptual framework for incorporating fairness in a machine learning context. The approach was pioneered by Gupta et al. (2018), Cotter et al. (2018a) and Cotter et al. (2018b).

Consider a dataset of N examples $\mathcal{D} = \{(x_i, y_i, g_{i,1}, g_{i,2}, \ldots, g_{i,K})\}_{i=1,\ldots,N}$, where $x_i \in \mathbb{R}^D$ is the observed feature vector, $y_i \in 0, 1$ is the label, where 1 denotes default in the subsequent two years, and $g_{i,k}$ is the indicator of membership of individual i in group k. For what follows, suppose our individuals fall into one of the two mutually exclusive groups $g = G = \{A, B\}$. In absence of group membership, such as for race, we use proxy groups $\widetilde{g_k}$ as substitutes for true protected group status.

Further, let $r(x) : \mathbb{R}^D \to \mathbb{R}$ denote the true risk score as a function of attributes and $s_p(x) : \mathbb{R}^D \to \mathbb{R}$ an approximate risk score as predicted by the prevailing statistical technology p. Given statistical technology j, a decision maker determines the outcome y for a specific individual based on a decision rule $d(s_j) = 1 \to s_j(x) > t$ and 0 otherwise, where t denotes a threshold. For example, a lender may decide to reject a loan to a consumer with a credit risk above some threshold. We will assume that all observable attributes are unprotected from a legal standpoint, wheres the group membership is a protected attribute.

We will consider a logistic regression with parameters θ to construct the approximate score $s_p(x)$. Our goal is to show that it is possible to better this statistical technology while improving on a group-level fairness metric. We do this by placing constraints on decision rule outcomes:

$$J_k(\theta) = P(d(s_j(x)) = 1|g = k, y = 0) - P(d(s_p(x)) = 1|g = k, y = 0) \text{ for k = A, B} \quad (2.2)$$

This constraint ensures that given some threshold t, the new classifier does not wrongly reject a larger proportion of applicants from each subgroups. We then add this fairness constraint into the training optimization problem. Suppose a classifier is given by thresholding a classifier score $f(x_i; \theta)$, where $f \in \mathcal{F}$ for some function class $\mathcal{F}$ parameterized by $\theta \in \mathbb{R}^d$:

$$\min_{\theta} \sum_{i=1}^{N} L(f(x_i; \theta), y_i)$$

$$\text{s.t. } J_k(\theta) \leq 0 \text{ for k = A, B} \tag{2.3}$$

Then, we train a neural network with the same structure as in the previous chapter, with the only difference being the constraints implemented while training the network structure. This approach delivers a prediction model of consumer default that intrinsically embeds the notion of fairness that constrains it in the training phase. For what follows, we explore the trade-offs between model performance and fairness. Specifically, we will quantify how more stringent fairness requirements affect performance across different demographic groups, and how they affect costs for lenders. These results can constitute the basis for assessing the redistribution implications of developing fair credit scoring models.

2.5 Fairness Constraints

In this section, we investigate the ability of our constrained models to preserve the false positive rates of the worse statistical technology. Our first set of constraints relate to age, while our second set of constraints relate to race. In particular, we impose constraints during training to enforce that the false positive rates be the same for our two subgroups as they were under the logistic regression for each quarter. To illustrate this, suppose the false positive rates in quarter t with the logistic regression for borrowers aged 30 or less was X%, and for borrowers aged more than 30 was Y%. Then, for quarter t our constraints would be that the false positive rates for borrowers aged 30 or less be less than or equal to X%, and for borrowers aged more than 30 be less than or equal to Y%.

2.5.1 Descriptive Statistics

Table 13 reports descriptive statistics on our sample across protected groups, including household income, credit score, and incidence of default, which here is defined as the fraction of households who report a 90 or more days past due delinquency on any trade. This will be our baseline definition of default, as this is the outcome targeted by credit scoring models.

We also report the predicted probabilities, and false positive rates across models.

We can see that young or black borrowers default on their loans at higher rates, have lower debt and 90+ dpd debt balances, lower household income, and ranked lower in both the credit scores and our model's risk distribution. The false positive rates are also higher for our protected groups, however, the differences are smaller for the constrained models than the unconstrained model.

2.5.2 Age

We now estimate Equation (2.1) with the differences being between the age constrained model and the logistic regression. Contrasting Panel A of Table 14 with Table 12, we can see that the constrained model predicts probabilities that are closer to the logistic for the young population than it was with the unconstrained DNN. Looking at differences in false positive rates in Panel B of Table 14 shows that our constrained model is successful at reducing false positive rates for protected groups.

2.5.3 Race Proxy

Once again we estimate Equation (2.1) with the differences being between the race constrained model and the logistic regression.

Contrasting Panel A of Table 15 with Table 12, we can see that the constrained model predicts probabilities that are closer to the logistic for the black population than it was with the unconstrained DNN. Looking at differences in false positive rates in Panel B of Table 15 shows that our constrained model is successful at reducing false positive rates for individuals with over 25% probability of being black.

Taken together, these results imply that introducing fairness constraints during training is successful at preserving the false positive rates of the worse statistical technology. Concerns regarding disparate impacts with respect to credit access are therefore mitigated, while the benefits of adopting this improved technology we further investigate in Section 2.6 and Appendix B.2.

2.6 Applications

In this section, we use our models in three applications. We first show that under our constrained models credit applicants for our protected groups would be accepted at a higher rate. Second, we show that under our constrained models, borrowers belonging to our protected groups would be ranked as less risky. Last, we illustrate the value added for lenders from our constrained models.

2.6.1 Access to Credit

We assess the economic salience of our models by analyzing credit approval rates. We use four credit allocation rules that approves credit if expected default rate is smaller than the threshold T. Specifically, we set these thresholds to 10%, 30%, 50% and 70%. We report the approval rate and accuracy given the threshold in Table 16. Accuracy here is defined as approved applicant who does not default in the subsequent 2 years.

In Panel A of Table 16 we can see that young borrowers are accepted at the highest rate under the constrained model. The difference in acceptance rates is typically over 1%, while the accuracy is also improved for this subgroup compared to the logistic regression. Looking at Panel B of Table 16, we find comparable approval rates between the logistic regression and the constrained model, both of which are higher than the unconstrained model. We also find improved predictive accuracy from adopting the constrained model over the logistic regression.

2.6.2 Counterfactual Rankings

We now investigate counterfactual rankings of borrowers in our subgroups. To do so, for each quarter separately, we compute the percentile rank for individuals with a valid credit score based on their credit risk under various models (i.e., logistic, constrained, unconstrained, credit score). Specifically, each individual is ranked under each of the models for each quarter. Higher score implies lower credit risk.

Panel A of Table 17 shows counterfactual rankings for our two subgroups pertaining to age. When looking at young borrowers who do not default, we find that the constrained model ranks them over 1 percentile higher on average than the logistic or the unconstrained model, while 4.3 percentiles higher than the credit score. In contrast, we find smaller differences across models for borrowers older than 30 years. One striking difference in Panel A is the treatment of defaulters across age groups by the credit score: while defaulters are ranked similarly across our models by age groups, the credit score ranks defaulters older than 30 years by 5.4 percentile higher than young defaulters. Our findings here suggest that the credit score is indiscriminately low for young borrowers, and our fairness constrained model provides favorable credit risk assessment for young borrowers when compared to any of the alternatives we assessed.

Panel B of Table 17 compares counterfactual rankings for our two subgroups pertaining to race. We document smaller differences across models with our proxy fairness constraints, yet we still find that the constrained model ranks non-defaulting borrowers with a probability over 25% of being black higher than the alternative models considered. Our results here suggest that the credit score ranks this subgroup too low when compared with alternative machine learning models.

2.6.3 Value Added

We now assess the economic salience of our constrained models by analyzing its value added for lenders. In particular, we examine the role our model can play in minimizing the losses from default.

2.6.3.1 Lenders We follow the framework from the previous chapter, which compares the value of having a prediction of default risk to having one by a logistic regression, and we make the same simplifying assumptions with respect to the revenues and costs of the consumer lending business. Specifically, in absence of any forecasts, it is assumed a lender will take no action regarding credit risk, implying that customers who default will generate losses for the lender, and customers who are current on their payments will generate positive revenues from financing fees on their running balances. To simplify, we assume that all defaulting and non-defaulting customers have the same running balance, B_r, but defaulting customers increase their balance to B_d prior to default. We refer to the ratio between B_d and B_r as "run-up." It is assumed that with a model to predict default risk, a lender can avoid losses of defaulting customers by cutting their credit line and avoiding run-up. Then, the value added can be written as follows ([55]):

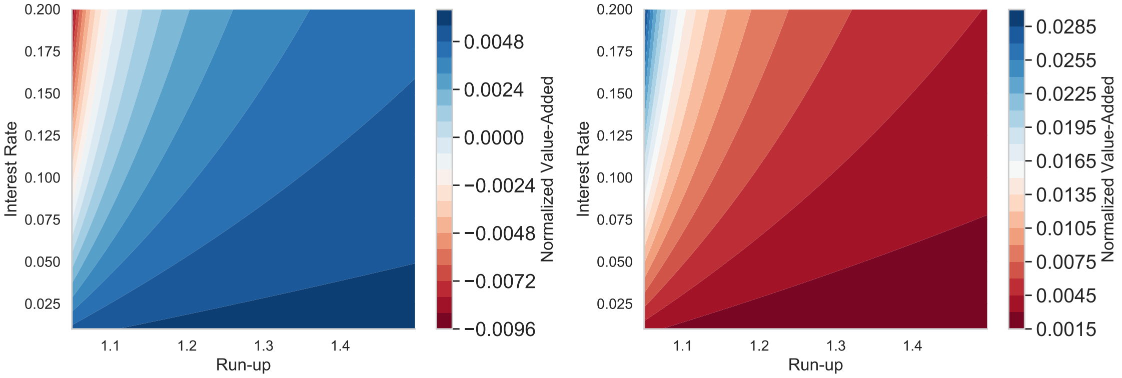
$$VA(r, N, TN, FN, FP) = \frac{TN - FN\big[1 - (1 + r)^{-N}\big]\big[\frac{B_d}{B_r} - 1\big]^{-1}}{TN + FP} \tag{2.4}$$

where r refers to the interest rate, N the loan's amortization period, and TN, FN, FP refer to true negatives, false negatives and false positives respectively. We compare the value added of our constrained models with default predictions generated by a logistic regression. This exercise illustrates the gains from adopting a better technology for credit allocation. Panel (a) of Figure 9 shows modest, but substantial cost savings in the range of 0-0.5% for our models with fairness constraints pertaining to age. Panel (b) calculates the cost savings associated to using our constrained model with fairness constraints pertaining to race. In this case the cost savings range from 0.1-2.5%. This exercise then confirms the advantages of using the better statistical technology.

2.7 Conclusion

We have proposed to use constrained deep learning to develop a model to predict consumer default. We show that by incorporating protected group status, the credit access for borrowers belonging to disadvantaged groups can be improved, while providing more accurate credit risk assessment. Additionally, we show that our constrained deep learning model performs better than standard machine learning models of default based on logistic regression and can accrue cost saving to lenders in the order of 0-2.5% compared to default predictions based on logistic regression. Taken together, our results suggest that introducing constraints can be effective at promoting better credit access to protected groups, while yielding substantial cost savings to lenders.

2.8 Figures and Tables



(a) Constrained^{age} vs. Logistic

(b) Constrained^{race} vs. Logistic

Figure 9: Value-Added of Constrained Machine-Learning Forecasts

Notes: Value-added of constrained machine-learning forecasts of 90+ days delinquency over 8Q forecast horizons. VA values are calculated with amortization period N = 3 years and a 50% classification threshold. Source: Authors' calculations based on Experian Data.

Table 12: The Unequal Effects of an Improved Statistical Technology

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Dependent Variable: PP$_{DNN}$ - PP$_{logistic}$								
Young	0.0210***	0.0234***	0.0207***	0.0194***				
	(32.52)	(59.94)	(56.23)	(31.97)				
Black $\geq$ 25%					0.0113***	0.0104***	0.0054***	0.0057***
					(12.73)	(12.68)	(8.20)	(7.21)
Default			0.0269***	0.0261***			0.0281***	0.0282***
			(27.02)	(24.66)			(29.42)	(27.42)
Young × Default				0.0030***				
				(3.19)				
Black $\geq$ 25% × Default								-0.0007
								(-0.94)
Constant	-0.0062***	-0.0087***	-0.0198***	-0.0195***	-0.0033***	-0.0060***	-0.0174***	-0.0174***
	(-12.95)	(-3.82)	(-8.63)	(-8.62)	(-6.59)	(-2.69)	(-7.75)	(-7.76)
adj. R^2	0.0133	0.0366	0.0610	0.0611	0.0027	0.0191	0.0456	0.0456
Panel B: Dependent Variable: FP$_{DNN}$ - FP$_{logistic}$								
Young	0.0100***	0.0075***	0.0076***	0.0155***				
	(17.84)	(20.33)	(22.07)	(19.42)				
Black $\geq$ 25%					0.0023***	0.0010***	0.0012***	0.0041***
					(6.82)	(3.52)	(3.94)	(6.94)
Default			-0.0014***	0.0155***			-0.0008*	0.0001
			(-3.14)	(5.72)			(-1.70)	(0.32)
Young × Default				-0.0191***				
				(-18.36)				
Black $\geq$ 25% × Default								-0.0059***
								(-8.66)
Constant	-0.0025***	-0.0223***	-0.0218***	-0.0236***	-0.0007***	-0.0216***	-0.0213***	-0.0216***
	(-8.08)	(-9.69)	(-9.14)	(-9.87)	(-2.86)	(-9.19)	(-8.73)	(-8.94)
adj. R^2	0.0009	0.0093	0.0093	0.0101	0.0000	0.0081	0.0082	0.0082
Observations	17860631	17860631	17860631	17860631	17860631	17860631	17860631	17860631
State FE		X	X	X		X	X	X
Quarter FE		X	X	X		X	X	X
Credit Report Controls		X	X	X		X	X	X

Notes: Robust standard errors clustered at the state level, t-statistic in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Young: indicator variable for being less than or equal to 30 years old. Black: indicator variable for having a probability over 25% of being black. We use 2010 Census data to estimate the probability of individual i being black in ZIP code z given age.

Table 13: Summary Statistics

	Age: $<$ 30	Age $\geq$ 30	t-test: Age	Black Share $\geq$ 25%	Black Share $<$ 25%	t-test: Race
	mean/sd	mean/sd	b	mean/sd	mean/sd	b
Household Income	91.400	103.645	12.246***	100.524	101.146	0.622
	(676.317)	(375.403)		(605.035)	(427.274)	
Age	24.363	51.709	27.345***	43.246	46.366	3.120***
	(3.259)	(13.440)		(15.518)	(16.506)	
Share of Black	0.133	0.110	-0.023***	0.497	0.054	-0.443***
	(0.186)	(0.178)		(0.210)	(0.060)	
Total Debt Balances	25160.564	92884.638	67724.074***	50023.481	83135.487	33112.006***
	(68584.997)	(192536.481)		(114408.128)	(183580.505)	
Mortgage Balances	15619.752	78096.554	62476.802***	38635.147	69090.228	30455.081***
	(62165.325)	(182562.298)		(105528.452)	(174003.573)	
Balance on Cards	3495.109	10140.684	6645.575***	6213.309	9139.744	2926.435***
	(8815.777)	(21102.405)		(15535.386)	(19879.183)	
90+ DPD Debt	1505.472	4124.973	2619.501***	4314.751	3454.076	-860.675***
	(16155.921)	(39957.633)		(34133.503)	(36602.924)	
Delinquent, next 8Q	0.436	0.322	-0.114***	0.543	0.315	-0.227***
	(0.496)	(0.467)		(0.498)	(0.465)	
Credit Score	621.013	686.626	65.612***	619.058	681.322	62.264***
	(120.767)	(122.286)		(130.120)	(121.852)	
PP$_{logistic}$	0.419	0.324	-0.095***	0.501	0.319	-0.182***
	(0.347)	(0.347)		(0.366)	(0.339)	
PP$_{DNN}$	0.434	0.318	-0.116***	0.509	0.316	-0.193***
	(0.356)	(0.354)		(0.373)	(0.348)	
PP$^{age}_{constrained}$	0.410	0.313	-0.097***	0.495	0.308	-0.187***
	(0.350)	(0.349)		(0.369)	(0.342)	
PP$^{black}_{constrained}$	0.423	0.314	-0.108***	0.501	0.311	-0.190***
	(0.354)	(0.351)		(0.372)	(0.345)	
FP$_{logistic}$	0.056	0.053	-0.003***	0.065	0.052	-0.014***
	(0.230)	(0.223)		(0.247)	(0.221)	
FP$_{DNN}$	0.064	0.050	-0.013***	0.067	0.051	-0.016***
	(0.244)	(0.219)		(0.250)	(0.220)	
FP$^{age}_{constrained}$	0.051	0.049	-0.002***	0.062	0.047	-0.015***
	(0.220)	(0.216)		(0.241)	(0.213)	
FP$^{black}_{constrained}$	0.060	0.050	-0.009***	0.065	0.050	-0.015***
	(0.237)	(0.219)		(0.247)	(0.219)	
N	3769240	14091391	17860631	2450540	15410091	17860631

Notes: Summary statistics across subgroups. PP denotes predicted probability of default, while FP denotes false positives.

Table 14: Constraining by Age

	(1)	(2)	(3)	(4)
Panel A: Dependent Variable: $\mathbf{PP}_{DNN}^{age}$ - $\mathbf{PP}_{logistic}$				
Young	0.0024***	0.0073***	0.0060***	0.0062***
	(5.17)	(27.59)	(25.57)	(14.28)
Default			0.0125***	0.0126***
			(15.44)	(13.54)
Young × Default				-0.0004
				(-0.58)
Constant	-0.0112***	0.0013	-0.0038*	-0.0039*
	(-25.66)	(0.62)	(-1.73)	(-1.76)
adj. R^2	0.0002	0.0263	0.0317	0.0317
Panel B: Dependent Variable: $\mathbf{FP}_{DNN}^{age}$ - $\mathbf{FP}_{logistic}$				
Young	-0.0013***	-0.0015***	-0.0021***	-0.0028***
	(-4.15)	(-7.71)	(-12.22)	(-7.64)
Default			0.0060***	0.0056***
			(9.66)	(7.68)
Young × Default				0.0016***
				(2.83)
Constant	-0.0037***	-0.0136***	-0.0160***	-0.0159***
	(-8.88)	(-6.05)	(-7.08)	(-6.91)
adj. R^2	0.0000	0.0052	0.0056	0.0056
Observations	17860631	17860631	17860631	17860631
State FE		X	X	X
Quarter FE		X	X	X
Credit Report Controls		X	X	X

Notes: Robust standard errors clustered at the state level, t-statistic in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Dependent variable is the difference between the constrained DNN's prediction and the logistic regressions. Young: indicator variable for being less than or equal to 30 years old. Default: indicator variable equal to 1 if the individual defaults in the subsequent 8 quarters.

Table 15: Constraining by Race

	(1)	(2)	(3)	(4)
Panel A: Dependent Variable: $\mathbf{PP}_{DNN}^{black}$ - $\mathbf{PP}_{logistic}$				
Black $\geq$ 25%	0.0081***	0.0075***	0.0027***	0.0038***
	(11.93)	(11.50)	(7.72)	(6.14)
Default			0.0127***	0.0187***
			(16.12)	(18.27)
Black $\geq$ 25% × Default				0.0005
				(0.75)
Constant	-0.0084***	-0.0078***	-0.0039*	-0.0153***
	(-21.82)	(-4.22)	(-1.78)	(-8.04)
adj. R^2	0.0013	0.0144	0.0308	0.0255
Panel B: Dependent Variable: $\mathbf{FP}_{DNN}^{black}$ - $\mathbf{FP}_{logistic}$				
Black $\geq$ 25%	0.0013***	0.0001	-0.0014***	0.0015***
	(4.49)	(0.26)	(-7.30)	(3.51)
Default			0.0059***	0.0013**
			(9.50)	(2.51)
Black $\geq$ 25% × Default				-0.0031***
				(-5.91)
Constant	-0.0013***	-0.0126***	-0.0160***	-0.0132***
	(-4.33)	(-6.18)	(-7.09)	(-6.27)
adj. R^2	0.0000	0.0052	0.0056	0.0053
State FE		X	X	X
Quarter FE		X	X	X
Credit Report Controls		X	X	X
Observations	17860631	17860631	17860631	17860631

Notes: Robust standard errors clustered at the state level, t-statistic in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Dependent variable is the difference between the constrained DNN's prediction and the logistic regressions. Black: indicator variable for having a probability over 25% of being black. We use 2010 Census data to estimate the probability of individual i being black in ZIP code z given age. Default: indicator variable equal to 1 if the individual defaults in the subsequent 8 quarters.

Table 16: Access to Credit across Protected Groups

Threshold →		10%		30%		50%		70%	
Model	Group	Approval Rate	Accuracy	Approval Rate	Accuracy	Approval Rate	Accuracy	Approval Rate	Accuracy
Panel A: Age									
	Age >30								
Logistic		0.451	0.737	0.636	0.852	0.708	0.865	0.775	0.85
Unconstrained		0.471	0.757	0.629	0.855	0.707	0.87	0.775	0.855
Constrained		0.467	0.754	0.646	0.857	0.713	0.867	0.782	0.851
	Age <= 30								
Logistic		0.205	0.614	0.546	0.824	0.621	0.831	0.697	0.808
Unconstrained		0.233	0.643	0.509	0.821	0.597	0.839	0.683	0.82
Constrained		0.24	0.646	0.561	0.832	0.627	0.834	0.708	0.807
Panel B: Race Proxy									
	Black <25%								
Logistic		0.428	0.708	0.649	0.849	0.72	0.862	0.784	0.849
Unconstrained		0.45	0.73	0.637	0.85	0.716	0.868	0.783	0.855
Constrained		0.451	0.731	0.652	0.852	0.719	0.865	0.787	0.851
	Black >= 25%								
Logistic		0.223	0.736	0.416	0.829	0.499	0.828	0.594	0.793
Unconstrained		0.236	0.75	0.397	0.832	0.485	0.838	0.585	0.805
Constrained		0.237	0.75	0.414	0.832	0.493	0.834	0.594	0.798

Notes: Classifier accuracy and approval rates under different classification thresholds. For instance, when the threshold is set at 10%, we classify individuals with over a 10% predicted probability as defaulters, while below this threshold as non-defaulters. Approval rate corresponds to the fraction of individuals classified as non-defaulters, while accuracy corresponds to the fraction of correctly tagged instances. In Panel A constrained refers to the age constrained model, while in Panel B it refers to the race constrained model.

Table 17: Counterfactual Rankings

Model	Protected Group	Default Status	Share	Mean	Std. Dev.	20%	40%	50%	60%	80%
Panel A: Age										
Constrained	Age >30	No default	53.7%	67.4	22.5	45	64	71	78	89
Logistic				67.4	22.7	45	64	71	78	89
Unconstrained				67.8	22.3	46	64	71	78	90
Credit Score				67.3	22.7	46	63	71	78	89
Constrained		Default	25.3%	23	18	8	15	19	24	35
Logistic				23.5	19	8	15	19	24	36
Unconstrained				23.1	18	8	15	19	24	35
Credit Score				25.4	19.2	9	17	21	26	39
Constrained	Age <= 30	No default	11.9%	54.3	16.9	41	50	54	58	69
Logistic				53	16.1	40	49	54	58	66
Unconstrained				53.2	16.5	39	50	54	58	68
Credit Score				50	18.4	34	46	52	57	67
Constrained		Default	9.1%	22.2	16.6	7	15	18	23	36
Logistic				22.6	16.7	8	15	19	23	36
Unconstrained				21.3	16.1	7	14	18	22	34
Credit Score				20	16.1	6	12	16	20	33
Panel B: Race Proxy										
Constrained	Black <25%	No default	59.3%	65.8	22	45	61	68	75	88
Logistic				65.6	22.1	45	60	67	74	88
Unconstrained				65.9	21.9	45	61	68	75	88
Credit Score				64.9	22.8	44	60	67	74	87
Constrained		Default	27.1%	23.6	18.1	8	16	20	24	37
Logistic				24	19	8	16	20	24	37
Unconstrained				23.4	18	8	15	20	24	36
Credit Score				24.9	19	8	16	21	26	39
Constrained	Black >= 25%	No Default	6.3%	57.8	22.8	36	50	57	64	80
Logistic				57.6	22.8	36	50	57	64	80
Unconstrained				57.8	22.7	36	50	57	64	80
Credit Score				57.2	24.1	35	50	58	65	81
Constrained		Default	7.4%	19.9	15.4	7	13	17	20	30
Logistic				20.5	15.9	7	14	17	21	31
Unconstrained				19.7	15.3	7	13	16	20	30
Credit Score				20.5	16.6	6	13	17	21	32

Notes: we compare the credit risk assessment of different models by normalizing the scores to a rank-order scale ranging from 1 to 100. This means that a score of 50 places an individual at the median of the distribution. We compute the rank-order separately for each quarter, and we require that the individual has a valid credit score. We report moments of the distributions across age and racial groups, and across default status. In Panel A constrained refers to the age constrained model, while in Panel B it refers to the race constrained model.

3.0 Racial Disparities in Debt Collection

(joint with Jessica LaVoice) A distinct set of disadvantages experienced by black Americans increases their likelihood of experiencing negative financial shocks, decreases their ability to mitigate the impact of such shocks, and ultimately results in debt collection cases being far more common in black neighborhoods than in non-black neighborhoods. In this paper, we create a novel dataset that links debt collection court cases with information from credit reports to document the disparity in debt collection judgments across black and non-black neighborhoods and to explore potential mechanisms that could be driving this judgment gap. We find that majority black neighborhoods experience approximately 40% more judgments than non-black neighborhoods, even after controlling for differences in median incomes, median credit scores, and default rates. The racial disparity in judgments cannot be explained by differences in debt characteristics across black and non-black neighborhoods, nor can it be explained by differences in attorney representation, the share of contested judgments, or differences in neighborhood lending institutions.

3.1 Introduction

The distinct disadvantages experienced by Black Americans increase their likelihood of experiencing negative financial shocks and limit their ability to mitigate the impact of such shocks. This ultimately makes them more likely to enter into default and have unpaid balances sent to collections. If a collection is brought to court and a guilty verdict is received, the defendant's wages can be garnished or their bank account can be seized. This fact, combined with compounding interest and various legal fees, could ultimately hinder the debtor's ability to accumulate wealth and overcome any initial economic disadvantage.[1] We document a disparity in debt collection judgments across Black and non-Black neighbor-

[1] ADP, the nation's largest payroll services provider, documented that more than one in 10 employees between the ages of 35 to 44 had their wages garnished in 2013.

hoods in this paper and explore potential mechanisms through which this racial disparity could be influencing the debt collection system.[2] Such mechanisms include differences in neighborhood-level income and credit score distributions, default rates, lending institutions, and debtors' likelihood to contest the debt in court.

To document this disparity, we construct a novel zip code-level panel dataset from 2004 to 2013 that links the number of debt collection judgments in each zip code in Missouri to data from Experian credit reports and the American Community Survey. Our main threat to identification is omitted variable bias. To mitigate this concern, we use a rich set of control variables from the Experian credit report data to track all aspects of a neighborhoods' financial liabilities, including the types of debt incurred with the number of accounts and balances as well as the neighborhoods' median credit score and detailed delinquency information. This helps to reduce omitted variable bias and provides insight into the specific mechanisms that may or may not be driving the racial disparity in debt collection judgments. We address concerns about differences in unobservable characteristics across Black and non-Black neighborhoods by limiting our sample to only neighborhoods with common support over observables and controlling for county and year fixed effects.

We find that the judgment rate is 85% higher in majority-Black neighborhoods than in majority non-Black neighborhoods. Over half of this baseline judgment gap can be explained by differences in incomes, credit scores, default rates, and housing values across these neighborhoods. Even after controlling for these differences, however, majority Black neighborhoods have approximately 40% more judgments than non-Black neighborhoods. Differences in total debt levels, debt composition, payment amounts, utilization ratios, and delinquency rates do not further mitigate the judgment gap, suggesting that credit scores are accurately capturing the relevant information from consumers' credit reports.

We also hypothesize that defendants from Black neighborhoods could be less likely to hire an attorney or to contest debt in court, making it less costly for debt collectors to obtain judgments in Black neighborhoods. We show, though, that there is no statistical difference

[2] This racial disparity was first stressed by Paul Keil and Anne Waldman in a ProPublica article. They found a disproportionate number of judgments in predominantly Black communities when compared to white ones, with the risk of judgment being twice as high in majority-Black neighborhoods than in majority-white neighborhoods with similar income levels.

in the share of contested judgments across Black and non-Black communities. Moreover, controlling for attorney representation has a limited effect on this judgment gap. Another potential theory regarding the racial gap in debt collection judgments is that differences in lending institutions across Black and non-Black neighborhoods cause the positive correlation between neighborhood racial composition and the judgment rate. The racial gap in judgments remains, though, after controlling for the number of banks and payday lenders in a given area. Lastly, we explore the extent to which certain plaintiff types (e.g. major bank, debt collector, high-cost lender) are driving our results. Judgment rates are higher in Black communities for every plaintiff type, but certain types of plaintiffs consistently show more racial imbalance in their lawsuits than others. Our results are adequately robust to offer an alternative measure to that of credit score or to using a gradient boosted trees machine learning estimation strategy. Furthermore, these results can be replicated by using the estimated racial composition of defendants as opposed to the racial composition of neighborhoods.

There are two potential explanations that we cannot explore using our current data: differences in wealth not driven by housing values and discrimination. Laws prohibit using race to make decisions regarding access to credit, and thus many creditors do not collect racial demographic information. Furthermore, juries are not typically used in debt collection cases and, if a case is heard in front of a judge, such cases are usually fairly algorithmic with a limited amount of subjectivity involved. It is more likely that the unexplained racial gap in debt collection judgments is the result of the broader disadvantages experienced by minority communities that have persisted in our modern society. For example, according to estimates provided by the United States Census Bureau in 2016, the typical Black household has a net worth of $12,920, while that of a typical white household is $114,700 - this is a $101,780 difference in wealth that could have important implications for a household's ability to mitigate negative financial shocks. About $35,000 of this wealth gap is not driven by home equity.

By translating this wealth gap into differences in annual income and using our estimates of the relationship between income and judgments, we calculate that a wealth gap of this size would explain almost all of the remaining judgment gap across Black and non-Black communities.[3]

There is a large literature that documents the important role that race plays in the labor market and the housing/mortgage markets (e.g. [14], [75], [17], and [80]). Blacks also face discrimination in the legal and criminal justice system; for example, they are more likely to be searched for contraband ([6]), to have biased bail hearings ([7]), and to be charged with a serious offense (Rehavi and Starr 2014). Furthermore, racial differences in wealth are large, and a growing literature explores how the racial wealth gap was propagated over time (e.g. [64] and [4]). While attention has been given to racial disparities in general, this is the first economic analysis to empirically document racial disparities in debt collection judgments.

Aside from the literature documenting racial disparities across many different dimensions, this paper also contributes to a growing literature about the debt collection industry.[4] We know that consumers who are sued by creditors and debt collectors are drawn predominantly from lower-income areas ([52]). Other more recent studies have investigated the role of information technology in the collection of consumer debts ([37]), documented the link between debt collection regulations and the supply of consumer credit ([40]), and determined if consumers are made better or worse off by settling their debt outside of court ([30]). Interpreted broadly, the impacts of the debt collection process on consumer outcomes have been well documented; however, racial disparities in debt collection have not been empirically explored.

The rest of our paper is outlined as follows: Section 2 discusses the typical debt collection litigation process in the United States, as well as the laws regulating access to credit and debt collection procedures, Section 3 describes our data, Section 4 outlines our empirical strategy, Section 5 documents the racial gap in debt collection judgments and discussions potential mechanisms driving the disparity, Section 6 presents various robustness checks,

[3] Our most conservative estimate of the judgment gap is 0.34 more judgments per 100 people in majority-Black neighborhoods compared to majority non-Black neighborhoods and is derived from [70]. We computed the difference in annual savings needed over a 40-year horizon to generate a wealth gap of $35,000. We found that an annual difference of $2,910 is sufficient to generate the wealth gap in net present value. For interest rate, we applied the historical return of the stock market, which between 1957 and 2018 was roughly 8%. Consistent with estimates from the U.S. Bureau of Economic Analysis, we assume an 8% personal savings rate. This translates into an annual income difference of $36,375. Increasing the median income of majority-Black neighborhoods by this amount would decrease the judgment rate by 0.25 judgments per 100 people.

[4] See [50] for an overview of the debt collection industry along with its institutional structure and regulatory environment.

and Section 7 concludes.

3.2 Background Information

The debt collection industry in the U.S. is large and the amount of debt being placed into collection continues to grow. According to a 2018 annual report by the Consumer Financial Protection Bureau, debt collection is a $10.9 billion dollar industry that employs nearly 120,000 people across approximately 8,000 collection agencies in the United States. In 2010 alone, U.S. businesses placed $150 billion in debt with collection agencies. When the debt is unsecured, the owner of the debt (i.e. the original creditor or the debt buyer) can either write off the debt, negotiate with the debtor to bring their debt to current, or file a debt collection lawsuit. In this section, we will summarize the key institutional details surrounding debt collection lawsuits and the laws regulating the debt collection industry.

3.2.1 Debt Collection Litigation Process

Debt collection litigation typically begins when a creditor files a “Summons and Complaint” in a state civil court.[5] This document names the parties involved and states the amount owed (including interest and, in some cases, attorney fees and court costs). The summons is served to the defendant to notify them that they are being sued. It provides additional information including the deadline for which the debtor must file a formal response, referred to as the “answer”, to the court. If this deadline is not met, the creditor will usually ask the court to enter a default judgment. Default judgments occur when the defendant has failed to perform a court-ordered action, and results in the court settling the legal dispute in favor of the plaintiff. The defendant is obligated to abide by the court’s ruling and is subject to the punishments requested by the court.

For most routine debt collection lawsuits, if the debtor files a formal response to the lawsuit a trial date will be requested and set by the court. In some courts, there will be

[5] These courts have many different names including municipal court, superior court, justice court, county court, etc.

a settlement conference before the trial date to try to settle the case before trial. Once a judgment is obtained by the creditor, the creditor might request a "debtor's examination," which would require the debtor to appear in court and answer questions about their finances. This process informs the creditor how it can collect the judgment. The most common methods for enforcing the judgments are to garnish wages or bank accounts.[6] If a dispute is settled before trial, the creditor gives up the ability to collect on the debt by garnishing the debtor's bank accounts or wages, and therefore often requires a one time lump sum payment to drop the suit.

3.2.2 Laws Regulating Debt Collection

Debtors are granted some protections throughout the debt collection process. The Fair Debt Collection Practices Act (FDCPA), which was enacted in 1977, is the primary federal law governing debt collection practices. The statute's stated purposes are as follows: to eliminate the abusive practices used to collect consumer debts such as calling the debtor at all hours of the night and showing up to their place of employment; to promote fair debt collection; and to provide consumers with an avenue for disputing and obtaining validation of debt information in order to ensure the information's accuracy.

Furthermore, the Consumer Credit Protection Act (CCPA) of 1968 restricts the amount of earnings that creditors can garnish from defendants' weekly disposable income to 25% or the amount by which disposable earnings are greater than 30 times the minimum wage. The share of wages protected from debt collection garnishments can be increased by state law. For example, a creditor in Missouri can garnish only 10% of after-tax wages if the debtor is the head of their household, though the burden to assert these protections is typically on the debtor and take-up is relatively low. There is no federal law limiting the amount of savings that can be seized from a debtor's bank accounts.

While not directly related to debt collection, other protections have been put in place to protect consumers in the credit market. For example, the Equal Credit Opportunity Act (ECOA) enacted in 1974 makes it illegal for creditors to discriminate against any applicant

[6] Courts can also seize and sell the debtor's personal property, though this is relatively uncommon.

on the basis of race, color, religion, national origin, sex, marital status, age, or participation in a public assistance program.[7] The law applies to everyone who regularly participates in a credit decision, including banks, retail and department stores, bankcard companies, finance companies, and credit unions. The ECOA applies both to the decision to grant credit as well as setting the terms of credit.

Furthermore, the Fair Credit Reporting Act of 1970 promotes the accuracy, fairness, and privacy of consumer information contained in the files of consumer reporting agencies. It was intended to protect consumers from the willful and/or negligent inclusion of inaccurate information in their credit reports. More recently, the Credit Card Accountability Responsibility and Disclosure (CARD) Act of 2009 established fair and transparent credit card practices. Key provisions include giving consumers enough time to pay their bills, prohibiting retroactive rate increases, making it easier to pay down debt, eliminating "fee harvester cards", and eliminating excessive marketing to young people.

Despite these protections, abusive debt collection practices still exist and, as we will show, minority neighborhoods are disproportionately impacted by debt collection judgments.

3.3 Data

We construct a zip code level panel dataset to document racial disparities in debt collection lawsuits across black and non-black neighborhoods. This panel dataset is constructed by combining multiple different data sources, the first of which documents debt collection court cases filed in Missouri from 2004 to 2013.[8] For each zip code in our sample, we know the number of debt collection lawsuits filed and the number of judgments arising from these lawsuits. We also know the number of cases that resulted in a default judgment (meaning the debtor did not show up to court), a consent judgment (meaning the debtor showed up

[7] This law is enforced by the Federal Trade Commission (FTC), the nation's consumer protection agency.

[8] This data was generously provided by Kiel and Waldman (2015). They acquired individual court case data from the state court administration. Their white paper focuses on three jurisdictions: Cook County, Illinois (composed of Chicago and surrounding suburbs), St. Louis City and St. Louis County, Missouri, and Essex county, New Jersey (composed of Newark and suburbs). The data included basic case information such as the plaintiff, the defendant, and the defendant's address. The race of the defendant is not reported.

to court and admitted to owing the debt), and the number of cases that were contested (meaning that some aspect of the debt was disputed). We know if the defendant was represented by an attorney and the plaintiff type (categorized into the following groups: auto, debt buyer, high-cost lender, major bank, medical, utility, and miscellaneous).

We also use Experian credit report data to control for credit scores, default rates, and other debt characteristics. This is an anonymous quarterly longitudinal panel of individuals who have an Experian credit report and spans from 2004 to 2013. The data contains many variables which allows us to track all aspects of individuals' financial liabilities, detailed delinquencies, various types of debt with the number of accounts and balances, as well as an individual's credit score. The data also contains each individual's zip code. We calculate the median credit score, the average number of delinquent accounts, and other credit measures for each zip code in our sample. We supplement these data with zip code tabulation data from the 2009-2013 American Community Survey to control for racial composition, median household income, the unemployment rate, and other socioeconomic variables of interest at the zip code level.[9]

We also document the number of lending institutions that are accessible to each neighborhood as an additional proxy for a neighborhood's financial well-being.[10] We use Census ZIP Code Business Patterns (ZCBP) data to get access to the number of banks and payday lenders in each zip code. The ZCBP data measure the number of establishments, number of employees and total payroll by ZIP and detailed industry code. Following [18] we use the following two North American Industrial Classification System (NAICS) codes to capture payday lending establishments: non-depository consumer lending (establishments primarily engaged in making unsecured cash loans to consumers) and other activities related to credit intermediation (establishments primarily engaged in facilitating credit intermediation, including check cashing services and money order issuance services).[11] For each zip code, we

[9] All dollar values are adjusted to be in terms of 2013 dollars. USPS ZIP Codes are not areal features used by the Census but a collection of mail delivery routes that identify the individual post office or metropolitan area delivery station associated with mailing addresses. ZIP Code Tabulation Areas (ZCTAs) are generalized areal representations of United States Postal Service (USPS) ZIP Code service areas.

[10] This measure can alternatively be thought of as a measure of access to credit markets. However, payday lenders could be the result of financial distress as opposed to the cause.

[11] [13] discuss how this proxy could likely overstate the number of payday lenders. We adjust these measures to correct for states that prohibit payday lending to help reduce this bias.

use arcGIS to create a weighted average (based on land area) of the number of banks and payday lenders that exist within a five mile radius from each zip code's centroid. Lastly, we use Fixed Broadband Deployment Data from Federal Communications Commission to document access to online credit markets for every zip code in our sample.

3.3.1 Race Proxies

In this paper, our primary focus is differences in the number of judgments per 100 people across majority black and majority non-black neighborhoods.[12] As such, we use neighborhood racial composition as our primary independent variable and classify a zip code as a majority black zip code if more than 50% of its residents are black. Information on the racial and ethnic composition of the U.S. population by geography comes from the Summary File 1 (SF1) from the 2010 Census, which provides counts of enumerated individuals by race and ethnicity for various geographic area definitions, including zip code tabulation areas.[13] However, one may wonder about the racial composition of the defendant pool specifically. In this section, we discuss various methods used by statisticians to estimate race when it is not available in administrative data. These methods generally use publicly available demographic information associated with an individual's surname and place of residence from the U.S. Census Bureau to construct proxies for race.

Our first proxy uses only surnames to predict the race of an individual, and thus the racial composition of the defendant pool. Information used to calculate the probability of belonging to a specific race given an individual's surname is based on data from the 2010 Census. This dataset provides each surname held by at least 100 enumerated individuals, along with a breakdown of the percentage of individuals with that name belonging to one of six race and ethnicity categories: Hispanic; non-Hispanic White; non-Hispanic Black or African American; non-Hispanic Asian/Pacific Islander; non-Hispanic American Indian and Alaska Native; and non-Hispanic Multiracial. In total, the surname list provides information on the

[12] The Equal Credit Opportunity Act (ECOA) generally prohibits a creditor from inquiring about the race, color, religion, national origin, or sex of an applicant, and as a result we lack information about race in both the Experian and debt collection datasets. One exception is applications for home mortgages covered under the Home Mortgage Disclosure Act (HMDA).

[13] Census block are the highest level of disaggregation (the smallest geography).

162,253 surnames covering approximately 90% of the population. While this proxy works well for Hispanic and Asian names, it is less accurate at predicting black-white differences since blacks and whites tend to have more similar surnames. We classify a defendant pool as being majority black if at least 50% of the defendants in the debt collection data were predicted to be black.

Our second proxy for the racial composition of defendants is constructed using Bayesian Improved Surname Geocoding (BISG) ([39]).[14] This method combines geography- and surname-based information into a single proxy probability for race using the Bayes updating rule. This method involves constructing a probability of assignment to race based on demographic information associated with surname and then updating this probability using the demographic characteristics of the zip code associated with place of residence. The updating is performed through the application of a Bayesian algorithm, which yields an integrated probability that can be used to proxy for an individual's race and ethnicity.[15] We once again classify the defendant pool as being majority black if at least 50% of the defendants in the debt collection data were predicted to be black. Research has found that this approach produces proxies that correlate highly with self-reported race and national origin and is more accurate than relying only on demographic information associated with a borrower's last name or place of residence alone ([22]).

3.3.2 Sample Selection

Our main specifications focus only on debt collection cases from Missouri. Aside from the fact that Missouri has a centralized database of cases tried in different circuit courts, previous research has documented that Missouri is a representative state in terms of collection ([72], [30]). More specifically, Missouri is representative in terms of percentage of consumers who are delinquent and the average amount of debt in collections ([72]). Missouri is also not particularly exceptional with regards to the law surrounding collections, its share of black residents, or its level of inequality ([30]). Finally, debt collectors in Missouri are obligated to

[14] Consumer Finance Protection Bureau's Office of Research (OR), the Division of Supervision, Enforcement, and Fair Lending (SEFL) rely on a Bayesian Improved Surname Geocoding (BISG) proxy method.

[15] Details of this algorithm are discussed in Section A2 of the Appendix.

file cases in the court associated with the borrower's address and their centralized database provides the defendants surname, both of which assists in our calculation of racial proxies of the defendant pool.

We have two additional sources of debt collection judgment data. The first is from all counties from New Jersey and Cook County, Illinois (composed of Chicago and surrounding suburbs). However, it is less detailed than the Missouri data. This data includes the number of cases filed in a five-year window from 2008-2012, but we don't have breakdowns about the type of judgment (default, consent, contested), we don't know the defendants' names, and we don't know if they were represented by an attorney. Our second additional source of judgment data comes from the Experian Credit Report data. This data once again lacks breakdowns by judgment and plaintiff type, as well as information about attorney representation. All of our main specifications use data from Missouri due to its representative nature, the high level of detail in the data, and because we want to keep our sample consistent across specifications. However, when comparable information is available, we use these additional data sources to test the robustness of our results.

3.4 Empirical Strategy

There are important differences in observable characteristics across majority black and non-majority black neighborhoods, which is documented in Figure 10. This figure shows kernel density estimates of various covariates that are used throughout this analysis. High share black neighborhoods tend to have lower median credit scores, lower median household incomes, lower median house values, higher unemployment rates and a higher share of divorced individuals. These disparities cause concern that there could also be important differences in unobservable characteristics that vary with racial composition of neighborhoods. We mitigate this concern by limiting our sample to only neighborhoods with common support over observables ([36]).

Specifically, we use a logistic regression to restrict our dataset to a common support. We estimate:

$$\Phi(M_{ict}) = \beta_0 + \theta X_{ic} + \epsilon_{ic} \tag{3.1}$$

where M_{ict} is an indicator variable equal to one if neighborhood i in county c in year t is predicted to be a majority black and X_{ict} is a vector of other controls for neighborhood i in county c in year t which includes quintiles of the income and credit score distributions, median income, median credit score, the gini index of income inequality, 90+ days-past-due debt balances, unemployment and divorce rates, population density, median house value, and education attainment levels such as fraction with at least a bachelors degree and fraction with less than a high school diploma. We plot the propensity score distribution in Figure 21, and restrict our sample to the intersection of the two curves. This drops very high income non-black neighborhoods and very low income black neighborhoods from our sample.

To further limit omitted variable bias, we use a rich set of control variables combined with both county and year fixed effects. County fixed effects will control for any time invariant unobservable characteristics of counties and year fixed effects control for any time varying changes that impact all of our neighborhoods.

3.4.1 Summary Statistics

Our common support sample consists of over 250 zip codes observed over 10 years. Table 18 presents summary statistics for our variables of interest across both majority black and majority non-black zip codes. Panel A shows our judgment data. On average, majority black neighborhoods had higher judgment rates than non-black neighborhoods (2.7 judgments per 100 people as opposed to 1.4). Black neighborhoods had a higher share of cases result in default judgments and a lower share of cases in which the defendant was represented by an attorney. Panel B summarizes differences in credit characteristics and lending institutions across black and non-black neighborhoods. On average, majority black neighborhoods have lower median credit scores and more debt balances that are 90 days past due. Lastly, Panel C summarizes other baseline control variables, including median income, median housing values, and educational attainment variables. In general, black neighborhoods have lower median household incomes, lower median house values, higher unemployment rates, and a lower share of college educated individuals. Most of these differences are significant at the 1% level.

3.4.2 Empirical Specification

We use a fixed effect framework with our common support sample to estimate the impact of racial composition on the debt collection judgment rate of neighborhoods. Our empirical specification is given by the following equation:

$$y_{ict} = \alpha + \beta M_{ic} + \theta X_{ict} + \gamma_c + \lambda_t + \epsilon_{ict} \tag{3.2}$$

where y_{ict} is the number of judgments per every 100 people in neighborhood i in county c in year t , M_{ict} is an indicator variable equal to one if neighborhood i in county c in year t has a black population greater than 50%, and X_{ict} is a vector of other controls for neighborhood i in county c in year t. Our main specification also includes county fixed effects to control for any time invariant differences across counties and year fixed effects to control for any time varying changes that impact all of our neighborhoods (like the Credit Card Accountability Responsibility and Disclosure Act of 2009). All regressions are weighted by population and standard errors are clustered at the county level.

To limit omitted variable bias and to better understand the mechanisms driving the racial disparity in debt collection judgments, we include a vast set of control variables. Aside from income and credit score, we add measures of debt balances by type of debt (credit card, medical, student loans, etc.), debt composition (type of debt as share of total debt balances), delinquent balances by length of delinquency (30 days, 60 days, 90 days), and bankruptcy/collection flags to X_{ict}. We also add controls for the number of banks and payday lenders within a five mile radius of each zip code and explore the results by plaintiff and judgment type.

3.5 Results

To establish a baseline judgment gap, we begin by documenting the racial disparity in judgments across black and non-black neighborhoods controlling only for county and year fixed effects. Figure 11 shows the relationship between the judgment rate and the percentage of blacks in a zip code. This figure classifies zip codes into one of a hundred bins based on their share of black residents and plots the average share of black of each bin against the average judgment rate of each bin. The size of the bubbles corresponds to the number of zip codes in each of the bins; as expected, there are many low share black neighborhoods and relatively less high share black observations. The regression line represents the fit between the average judgment rate in the bin and the share of black population weighted by the number of observations in the bin. We see that the judgment rate is positively correlated with the share of black residents residing in the zip code.

This relationship is formalized in Column (1) of Table 19. Column (1) shows that majority black neighborhoods have about 1.2 more judgments per every 100 people compared to non-black neighborhoods; this implies that the judgment rate in black neighborhoods is almost double that of non-black neighborhoods where the average judgment rate is 1.4 judgments per 100 people.

3.5.1 Income and Credit Score Distributions

One important difference in the observable characteristics between black and non-black neighborhoods is differences in income. Panel (a) of Figure 12 plots neighborhoods by their median income and the judgment rate per 100 people, with darker red circles representing neighborhoods with a higher share of black residents and darker blue circles representing neighborhoods with a higher share of white residents. This figure documents a negative relationship between the judgment rate and median income, with the judgment rate decreasing as median income increases. However, even looking at neighborhoods with similar income levels, we see higher judgment rates for majority black neighborhoods.

Column (2) of Table 19 adds controls for income quintiles and median income to the previous specification. After controlling for differences in the income distribution across black and non-black neighborhoods, we see that black neighborhoods are associated with 0.9 more judgments per 100 people. This implies that differences in the income distribution across black and non-black neighborhoods can explain 23% if the racial disparity in debt collection.

A second important difference in the observable characteristics between black and non-black neighborhoods is differences in credit scores. Panel (b) of Figure 12 plots neighborhoods by their median credit score and the judgment rate per 100 people, with darker red circles once again representing neighborhoods with a higher share of black residents and darker blue circles representing neighborhoods with a higher share of white residents. This figure documents a negative relationship between the judgment rate and median credit scores. Once again, racial bias is evident in this figure, with majority black neighborhood having higher judgment rates than non-black neighborhoods with similar median credit scores, suggesting that differences in credit scores may not be the primary mechanism driving the racial disparity in debt collection cases.

Column (3) in Table 19 adds credit score quintiles and median credit score to the baseline specification. We see that majority black zip codes are associated with 0.75 more judgments per 100 individuals, a 50% increase of the average judgment rate of non-black zip codes. This means that differences in the credit score distribution can explain 40% of the judgment gap between black and non-black neighborhoods. Column (4) adds income and credit score controls into the same specification and coefficient changes very little, suggesting that 60% of the judgment gap remains unexplained after controlling for differences in the income and credit score distributions across black and non-black communities.

Column (5) adds controls for total delinquent debt balances, unemployment rate, median house value, the fraction of the population with a college education, and population density.[16] In this specification, we see that majority black neighborhoods have a 40% higher judgment rate than non-black neighborhoods. It is primarily the inclusion of the unemployment rate

[16] According to estimates provided by the United States Census Bureau in 2016, the typical black household has a net worth of $12,920, while that of a typical white household is $114,700. This difference could play an important role in driving the racial gap in debt collection. Since a majority of wealth accumulated to middle or low income households is through home ownership, we use housing values to help control for differences in wealth levels across black and non-black neighborhood.

and median housing values that cause the decline in the estimated coefficient on our majority black indicator. This suggests that these variables can explain away an additional 10% of the baseline judgment disparity. Column (6) uses a one year lag of our baseline controls. Even after adding controls for these observable characteristics, 50% of the baseline judgment disparity still remains.

3.5.2 Debt Characteristics

We also explore whether differences in debt portfolios of black and non-black neighborhoods are driving the racial disparity in debt collection judgments independent of credit score. For example, it could be the case that majority black neighborhoods tend to acquire the type of debt that is more likely to be collected in court. To explore this hypothesis, we use the plethora of information from the Experian Credit Report data to control for differences in debt characteristics across black and non-black communities. Such controls include total debt levels, debt composition, payment amounts, utilization ratios, and delinquency rates.

Our results are presented in Table 20. Each specification includes the income, credit scores, and baseline controls discussed in Table 19, as well as county and year fixed effects. Column (1) adds additional controls for total debt levels, including breakdowns for the type of debt such as credit card debt, mortgage debt, and student loan debt. Column (2) includes controls for payment amounts and utilization rates. Column (3) includes debt composition controls, such as credit card debt as a share of total debt. Column (4) adds additional delinquency and collection controls, including the total debt balances that are 30 days, 60 days, or 90 days delinquent, as well as bankruptcy and collection flags. Lastly, Column (5) includes all of these controls together. In each specification we see that majority black neighborhoods have an additional 0.6 judgments per 100 individuals, a 40% higher judgment rate compared to non-black neighborhoods. These results indicate that after controlling for differences in credit scores, differences in debt characteristics cannot explain any additional share of the racial disparity in judgment rates.

3.5.3 Lending Institutions

We have shown that a racial disparity in debt collection judgments exists, even after controlling for differences in the income and credit score distributions across black and non-black neighborhoods. In this section, we explore another potential mechanism that could be driving the racial disparity in debt collection judgments - differences in lending institutions across black and non-black neighborhoods.[17]

To explore this potential explanation, we use arcGIS to create an index that measures the number of banks and payday lenders within 5 and 10 mile radii from each zip codes' centroid.[18] We also use broadband access as a proxy for access to online credit markets and the share of credit reports that are unscored as proxy for access to credit.[19] We add these variables as controls to our main specification.

The results are presented in Table 21. Column (1) shows our main result is present on this subsample of data. In Columns (2) and (3), we add controls for broadband access and the share of unscored accounts in a zip code respectively; both measures have no impact on our coefficient of interest. In Columns (4) and (5) we add our controls for banks and payday lenders with 5 and 10 mile radii. We see that the number of banks is negatively correlated with the judgment rate while the number of payday lenders is positively correlated with the number of judgments. Adding these controls decreases the coefficient on black majority by 15%, though, a large racial gap in the number of judgments issued across black and non-black communities remains.

3.5.4 Attorney Representation and Judgment Type

We next investigate whether debt collectors target neighborhoods where defendants are less likely to have an attorney or to contest the debt. It could be the case that debt collectors

[17] The presence of payday lenders is likely the result of financial distress as opposed to the cause. As such, we view these controls as an additional measure of the financial well-being of neighborhoods as opposed to a measure of credit access.

[18] This analysis only used data from 2008-2012 and thus our sample size is slightly smaller. We replicate the main results on this subsample of the data for context.

[19] An unscored credit report is one in which there is not enough information on a consumer's credit report to issue a formal credit score. This could serve as an access to credit if unscored reports are correlated with limited access to credit markets as opposed to a limited desire to obtain credit.

target their collection efforts in areas where defendants are less likely to show up to court, resulting in a default judgment, or in areas where defendants tend to acknowledge they owe the debt. In other words, debt collectors might avoid collecting in areas where defendants tend to argue some aspect of the debt owed, which could result in the plaintiff exerting more effort or spending money to collect the debt.

To explore the extent to which differences in attorney representation are driving our result, we document the disparity in attorney representation and show how this disparity impacts the judgment rate. These results are presented in Columns (1) and (2) of Table 22. Note that each specification in this table includes the income, credit score, and baseline controls discussed in Table 19, as well as county and year fixed effects. The outcome variable in Column (1) is the share of debt collection court cases where the defendant was represented by an attorney; this result shows that defendants in majority black neighborhoods are less likely to have an attorney represent them in a debt collection court case. However, as seen in Column (2) where our dependent variable is once again judgments per 100 people, controlling for the share of cases in which defendants are represented by an attorney cannot explain the racial disparity in debt collection cases.

We take this as evidence that attorney representation does not impact the number of debt collection judgments; this does not imply that attorney representation is not meaningful or important in debt collection court cases. Debt collection laws often place the burden to assert various legal protections, including the share of the debtor's wages that can be garnished as the result of a judgment, on the debtor. Attorney representation is likely important in protecting debtors' rights throughout the debt collection process, even if such cases ultimately end in judgments.

To explore the extent to which differences in the share of contested versus uncontested cases could be driving our result, we document the impact of neighborhood racial composition on the share of different types of judgments.[20] Our outcome variable in Column (3) is the

[20] We also document the share of cases that resulted in a judgment. These results are presented in Table 52 in the Appendix. While only marginally significant, we see that majority black neighborhoods are 2 percentage points more likely to have a case result in a judgment. This is primarily driven by a lower share of cases being settled before a case is tried. This translates to a 10% decrease from the non-black settlement rate. Since settling a case often requires a one time lump sum payment, defendants who settle tend to have worse subsequent credit outcomes ([30]). This suggests that a lower propensity to settle cases before trial could actually help defendants from majority black neighborhoods. This can also be seen as suggestive

share of cases in which the defendant admitted to owing the debt, our outcome variable in Column (4) is the share of cases that were contested, and our outcome variable in Column (5) is the share of cases resulting in default judgments. We see no racial differences along these dimensions. These results suggest that it is unlikely that debt collectors are targeting areas without attorney representation or areas where defendants are less likely to show up to court.

3.5.5 Non-linearities & Higher Order Interactions

In this section, we investigate if machine learning techniques that allow for high order interactions of observable characteristics can help inform what mechanisms are driving the remaining racial disparity in judgments. More specifically, we implement Gradient Boosted Trees (GBT) which is an ensemble learning method that recursively combines the forecasts of many shallow decision trees.[21] The theory behind boosting is that a collection of weak learners as a whole creates a single strong learner with improved stability over a single complex tree. There are pros and cons to using machine learning approaches to explore the racial disparity in debt collection. Two of the key benefits of applying GBT is that it is particularly well suited to capturing interactions between variables in the data, without ex-ante specifying what interactions to add and that it increases our predictive power.[22] The downside of this technique is that interpretability becomes more difficult and less precise.

3.5.5.1 Explanatory Power of Variables We use SHapley Additive exPlanations (SHAP), a unified framework for interpreting associations, to explain the output of our Gradient Boosted Trees (Nonlinear Model).[23] SHAP uses a game theoretical concept to assign each feature a local importance value for a given prediction. The SHAP value gives us individualized impacts for each predictor; positive SHAP values are associated with increased

evidence that defendants from majority non-black neighborhoods are better able to mitigate negative shocks. We see no statistical difference in the share of cases that are dismissed.

[21] For more information about the GBT model, see [42] and the Appendix.

[22] Table 39 contrasts the predictive power of the GBT model with the linear model. The RMSE is computed using a regression of baseline covariates on judgment rates, and the results show that the nonlinear model is better able to explain the variation in judgment rates. Note RMSE = $\sqrt{1-r^2} * \sigma_y$, and hence, a lower RMSE translates into higher predictive power.

[23] For more on SHAP, see [63].

judgment rates and negative SHAP values are associated with decreased judgment rates. Figure 13 plots the distribution of the impact each predictor, including first order interactions, has on the model output for the fifteen most important predictors. These distributions are shaded based on the value of the independent variable with blue dots representing lower values and red dots representing higher values. This figure orders our independent variables in order of their importance as a predictor of judgments in the GBT procedure. Neighborhood racial composition is the most important predictor of judgment rates. Thus, allowing for a nonlinear model with higher order interactions does not mitigate the impact of neighborhood racial composition on judgment rates; if anything, allowing for this more flexible model highlights the importance of race in predicting judgments.

Aside from neighborhood racial composition, high median house value and credit score decrease predicted judgment rate. The divorce rate is also a significant predictor; neighborhoods with higher divorce rates are associated with higher judgment rates and neighborhoods with lower divorce rates are associated with lower judgment rates. These results only point to correlations between the predictors and the judgment rate; they should not be interpreted causally. They are primarily used to understand the contribution each predictor on the final model output and to provide some comparative statics.

3.5.6 Differences in Plaintiff Type

Lastly, we explore if differences in judgment rates across black and non-black zip codes are driven by a specific plaintiff category. For each zip code, we know the number of judgments awarded to each of the following plaintiff types: auto, debt buyer, high-cost lender, major bank, medical, utility, and miscellaneous. Debt buyers account for 48% of plaintiffs in our sample. Medical lenders, major banks, and high-cost lenders are the next largest plaintiff categories accounting for 20%, 13%, and 6% of plaintiffs respectively. The other plaintiff categories are combined into the miscellaneous category.

Our results are presented in Table 23. Each specification includes the income, credit score, and baseline controls discussed in Table 19, as well as county and year fixed effects. Column (1) repeats the main analysis and includes judgments from all plaintiff types (this is the same result presented in Column (5) of Table 19). Column (2) limits the outcome variable to only judgments obtained by debt buyers, Column (3) to major banks, Column (4) to medical companies, Column (5) to high cost lenders, and Column (6) to any other lender. The racial gap in judgments is persistent across all plaintiff types.

The coefficients estimated across each specification should not be directly compared due to differences in the baseline judgment rates in non-black neighborhoods across these different plaintiff types. For example, the judgment rate in non-black neighborhoods was 0.58 judgments per 100 people for debt buyers, 0.46 judgments per 100 people for major banks, and 0.08 judgments per 100 people for high cost lenders. These baseline levels imply that majority black neighborhoods have a 33% higher judgment rate than non-black neighborhoods among debt buyers, a 9% higher judgment rate among major banks and a 128% higher judgment rate among high cost lenders. Thus, while the racial gap in debt collection judgments exists for every plaintiff type, high cost lenders consistently showed more of a racial imbalance in their lawsuits than others.

3.6 Robustness Checks

In this section, we provide various robustness checks including different measures of racial composition and exploring the impact of the racial composition of the defendant pool as opposed to the racial composition of the neighborhood. We also explore selection on unobservables, an alternative measure of credit score, and alternative judgment data sources.

3.6.1 Race Proxies

In this section we show that our results are robust to using the share of black residents in a neighborhood as opposed to a binary measure. Columns (1) and (2) of Table 24 present this result. Column (1) shows a positive and statistically significant coefficient on the share of black residents within a zip code and Column (2) shows our preferred specification from Table (2) which uses our binary measure for a black neighborhood. One potential concern with this analysis is that the racial composition of defendants within a neighborhood could be drastically different from the racial composition of the neighborhood itself. As such, we use our BISG measure of share black to estimate the racial composition of the defendant pool. Columns (3)-(4) present the results. Once again, the results are all positive and statistically significant. Column (5) uses only surname (and no information on zip code demographics) to predict the racial composition of defendants. The result is positive and statistically significant, although the effect size increases drastically.

Columns (6)-(7) of Table 24 present our results when the BISG method was used to estimate race but uses zip code fixed effects instead of county fixed effects. This is only possible with our proxies that utilize variation in defendants name because only these proxies give us variation in the racial composition of defendants over time. We once again get a positive and statistically significant coefficient, with black neighborhoods experiencing 0.6 more judgments per 100 people than comparable non-black neighborhoods. This is a 40% increase over the non-black neighborhood mean of 1.4 judgments per every 100 people.

3.6.1.1 Other Races One might wonder if this phenomenon is specific to the black population. In Table 25, we replicate Table 19 with additional controls for the share of Hispanic and Asian population within each zip code. While columns (1) and (2) of Table 25 show judgment gaps for both Asians and Hispanic neighborhoods (with share Asian being negatively related to judgments and share Hispanic being positively related to judgments), these disparities are completely explained away by differences in credit scores, income, and our other baseline controls; Columns (5) and (6) show no statistically significant coefficients for the share of Hispanic and Asian populations. The share of black residents remains positive and statistically significant in each of the specifications. These results indicate that there is something specific about black neighborhoods that is causing the gap in judgments.

3.6.2 Selection on Unobservables

We also investigated the impact of selection on unobservables on coefficient stability ([70]). In particular, we used Column (5) of Table 19 as our benchmark, and found that given a selection on unobservables that is half of the size of the selection on observables, our coefficient on black majority is reduced to 0.34 with a 95% confidence interval ranging from [0.13 to 0.54].[24] This suggests that 24% of our baseline judgment gap of 1.4 would remain after controlling for unobservable characteristics.[25] This finding suggests that a racial gap is unlikely to be zero, even after controlling for any unobservable characteristics.

3.6.3 Alternative Credit Score

In Table 53 in the Appendix, we add an alternative control for credit score, that was calculated using a deep learning algorithm. The model was shown to consistently outperform standard credit scoring models when predicting default rates (see Chapter 1). This alternative credit score has more predictive power then credit score in predicting default. However, it does not mitigate the racial bias we see in judgments across black and non-black communities. This provides additional support that differences in credit scores, which measure a borrowers likelihood of defaulting, is not the main factor driving the judgment gap between black and non-black communities.

3.6.4 Evolution of Disparity

Figure 14 plots the evolution of the racial disparity from 2004-2013. The racial disparity is present over our whole sample period, however it increases dramatically during the great recession. This could be taken as evidence that minority neighborhoods were disproportionately impacted by recession or that they had less wealth to help mitigate the negative shocks associated with the recession.

[24] We bootstrapped our treatment coefficient estimates 100 times, and assumed a maximum R^2 value of 0.9.

[25] We also examined the proportion of selection of unobservables to observables that would explain away our treatment effect. We found that a ratio of 1.08 with a 95% confidence interval ranging from [0.56, 1.6] is sufficient to explain away our findings.

3.6.5 Alternative Data Sources

Tables 18 and 19 are replicated using data from New Jersey and Cook County, Illinois. The results are presented in Table 54 and 55 in the appendix. Table 54 shows that judgments per 100 people are larger in majority black neighborhoods compared to majority non-black neighborhoods, while median income and median credit score tend to be lower. Table 55 confirms that the racial gap in debt collection judgments cannot be explained by differences in median income or median credit score. These results suggest that judgments are 30% higher in majority black neighborhoods compared to majority white ones. Differences in other observable characteristics, such as default rates, can explain some of this disparity, although even after controlling for these differences, judgments are still 22% higher in black neighborhoods compared to non-black neighborhoods.

Tables 18, 19, and 20 are replicated using merged Experian-ACS data. All of the credit variables, including judgments, are individual specific. Racial composition and other control variables from the census are imputed by zip code. The results are presented in Table 56, Table 57, and Table 58 in the appendix. About 75% of the racial disparity can be explained by differences in income and debt portfolios; being from a black neighborhood is associated with 0.02 more judgments, a 22% increase over the baseline rate of 0.07 judgments per 100 people.[26]

3.7 Conclusion

Our estimates suggests that there are 40% more debt collection judgments in majority black neighborhoods compared to non-black neighborhoods even after controlling for differences in incomes and credit scores. This racial disparity exists for different racial measures and cannot be fully explained away by the share of contested versus uncontested cases across black and non-black communities or by differences in debt characteristics. The racial gap in debt collection judgments cannot be explained by differences in lending institutions and

[26] Only 24% of the baseline judgment gap of 0.069 remains after the inclusion of income and credit controls. This is the same share of the judgment gap that remains unexplained from the Oster test presented above.

exists for every plaintiff type, however, certain types of plaintiffs consistently showed more of a racial imbalance in their lawsuits than others.

There are two potential explanations that we cannot explore using our current data: differences in wealth and discrimination. It is unclear where discrimination would occur during the legal process, as most cases are fairly algorithmic and heard by a judge with no jury necessary. Furthermore, Keil and Waldman (2015) quote Lance LeCombs, the Metropolitan St. Louis Sewer District's spokesman, who claims his company has no demographic data on its customers and treated them all the same. The racial disparity in its suits, he said, is the result of "broader ills in our community that are outside of our scope and exceed our abilities and authority to do anything about." According to estimates provided by the United States Census Bureau in 2016, one such broader ill is that the typical black household has a net worth of $12,920, while that of a typical white household is $114,700 - this is a $101,780 difference in wealth that could have important implications for a household's ability to mitigate negative income shocks. About 35,000 of this wealth gap is not driven by home equity. By translating this wealth gap into a difference in annual income and using our estimates of the relationship between income and judgements, we calculate that a wealth gap of this size would explain almost all of our most conservative estimate of the judgment gap across black and non-black communities.[27]

As the number of debt collection cases rise, identifying both the extent to which racial disparities exist and how they are entering the debt collection system are crucial. Future research should explore policies meant to provide more protections to consumers and how they impact the racial disparity in debt collection judgments. Such reforms could require debt buying companies to prove they own the debt before they can sue a debtor, preventing companies from winning judgments when the statute of limitations has expired on a debt[28],

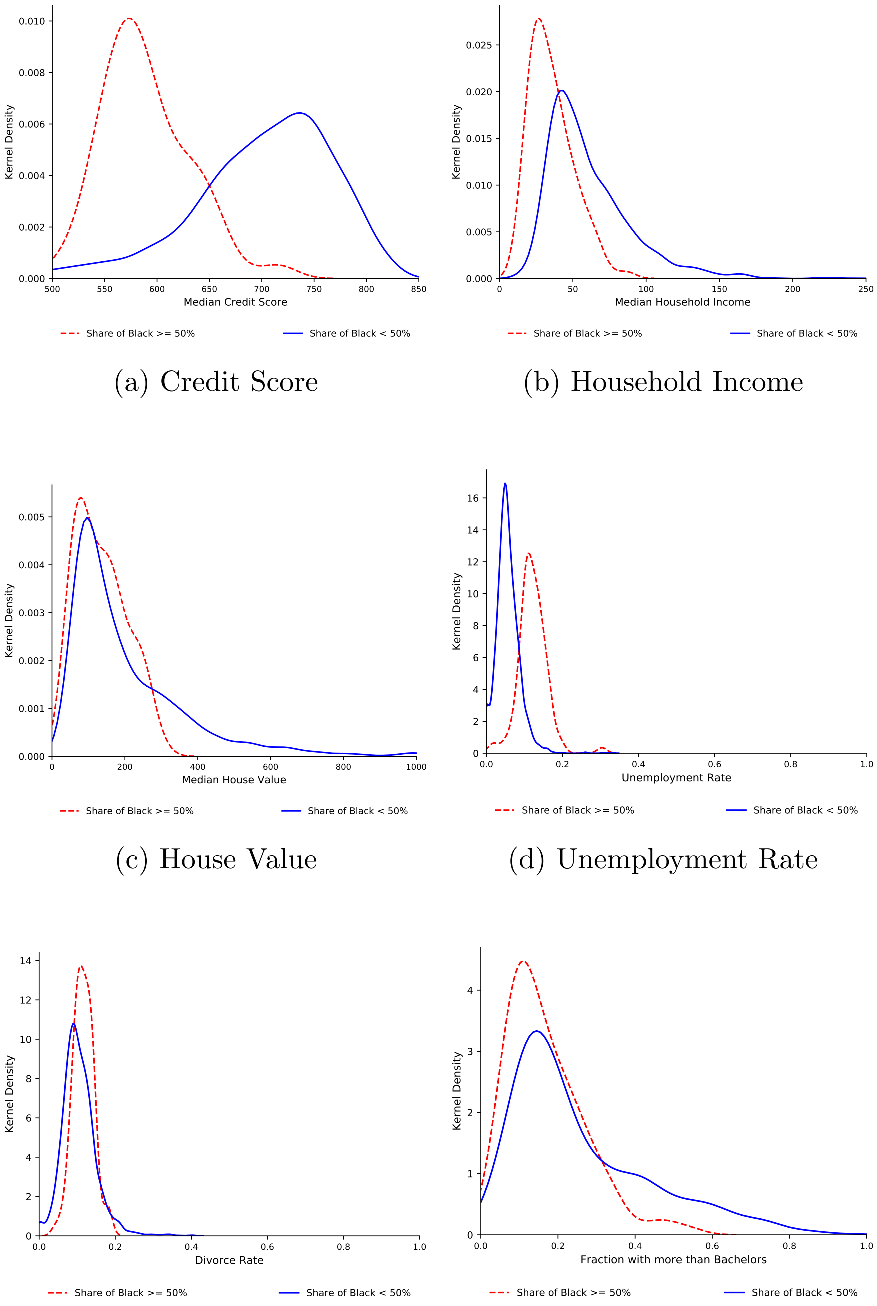
[27] Our most conservative estimate of the judgment gap is 0.34 more judgments per 100 people in majority black neighborhoods compared to majority white ones and is derived from [70]. We computed the difference in annual savings needed over a 40 year horizon to generate a wealth gap of $35,000. We found that an annual difference of $2,910 is sufficient to generate the wealth gap in net present value. For interest rate, we applied the historical return of the stock market, which between 1957 through 2018 is roughly 8%. Consistent with estimates from the U.S. Bureau of Economic Analysis, we assume an 8% personal savings rate. This translates into an annual income difference of $36,375. Increasing the median income of majority black neighborhoods by this amount would decrease the judgment rate by 0.25 judgments per 100 people.

[28] In most states, the law currently requires defendants to know that the statute of limitations has expired, and raise it as a defense in court.

or require collection attorneys to prove they have a legal right to collect attorney fees and provide an itemized list of their work on the case in order to win an attorney's fee through a default judgment[29]. When states do provide legal protections for debtors, such as allowing those with children to keep more of their pay under a head of family exemption, the burden is typically on the debtor to assert these protections. Another policy reform could require a clear notice that these are provided to debtors.

3.8 Figures and Tables

[29] Currently, when companies sue, they often request such fees, which are usually granted and passed on to the debtor as part of the judgment. For example, in Missouri, the fees are usually set at 15 percent of the debt owed, even though attorneys may spend only a few minutes on a suit.



(a) Credit Score

(b) Household Income

(c) House Value

(d) Unemployment Rate

(e) Divorce Rate

(f) >= Bachelor Degree

Figure 10: Kernel Density Estimates of Selected Covariates

Notes: This figure shows the kernel densities of select variables of interest broken down by the racial composition of neighborhoods.

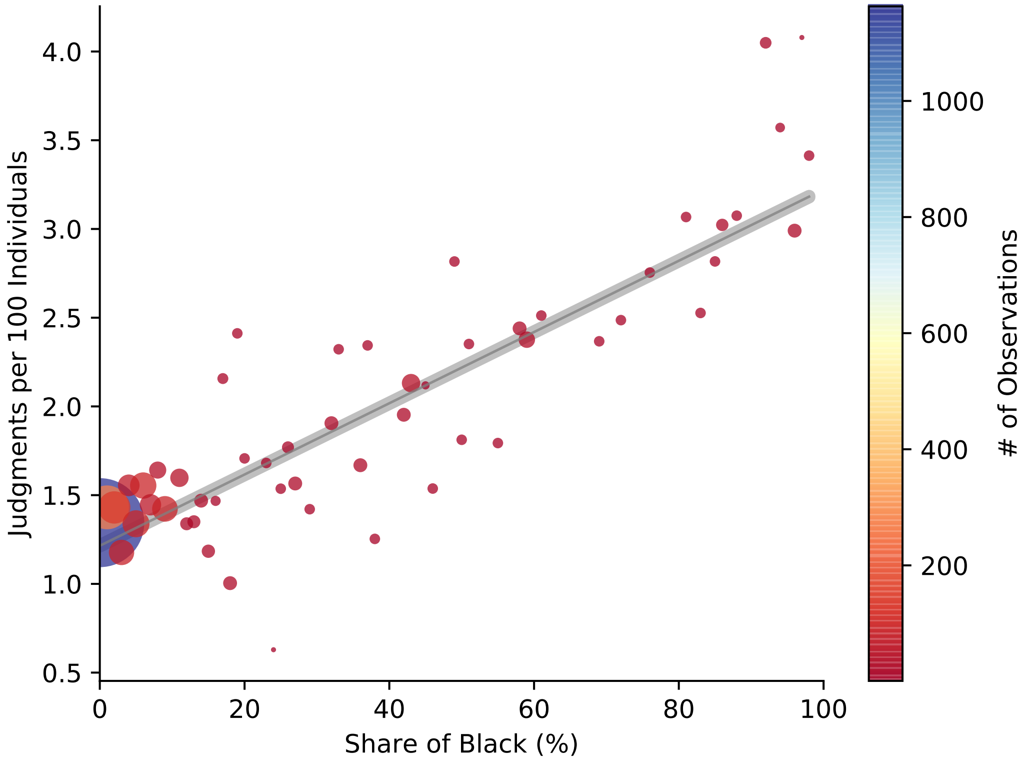
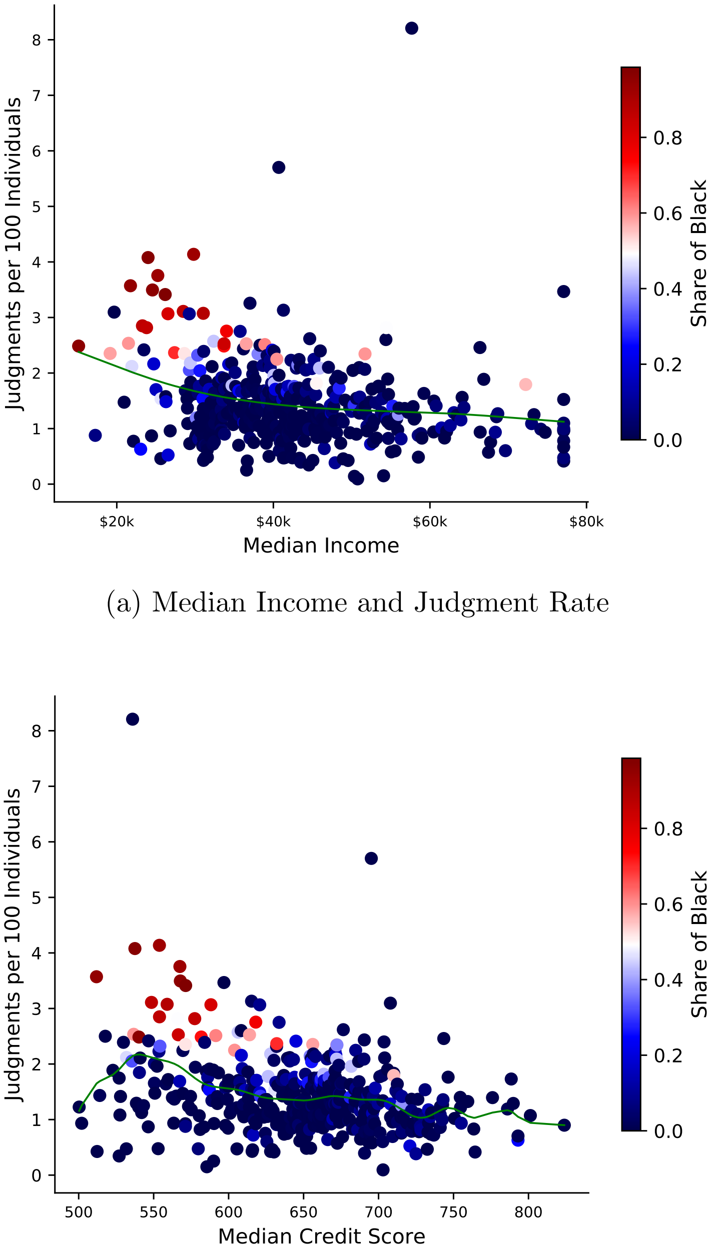


Figure 11: Judgments and Demographic Composition

Notes: Linear regression illustrates the relationship between share of black population and judgment rate. We categorized zip codes into one of a hundred bins based on their share of black residents and plotted the average share black of each bin against the average judgment rate of each bin. The size of the bubbles corresponds to the number of observations in each of the bins. The regression line represents the fit between the average judgment rate in the bin and the share of black population weighted by the number of observations in the bin.



(a) Median Income and Judgment Rate

(b) Median Credit Score and Judgment Rate

Figure 12: Income, Credit Scores, and Judgment Rate

Notes: The green line represents the non-parametric locally weighted regression line (LOESS) showing the smoothed fit curve of the data. Income is winsorized at the 98% level to mitigate the impact of outliers.

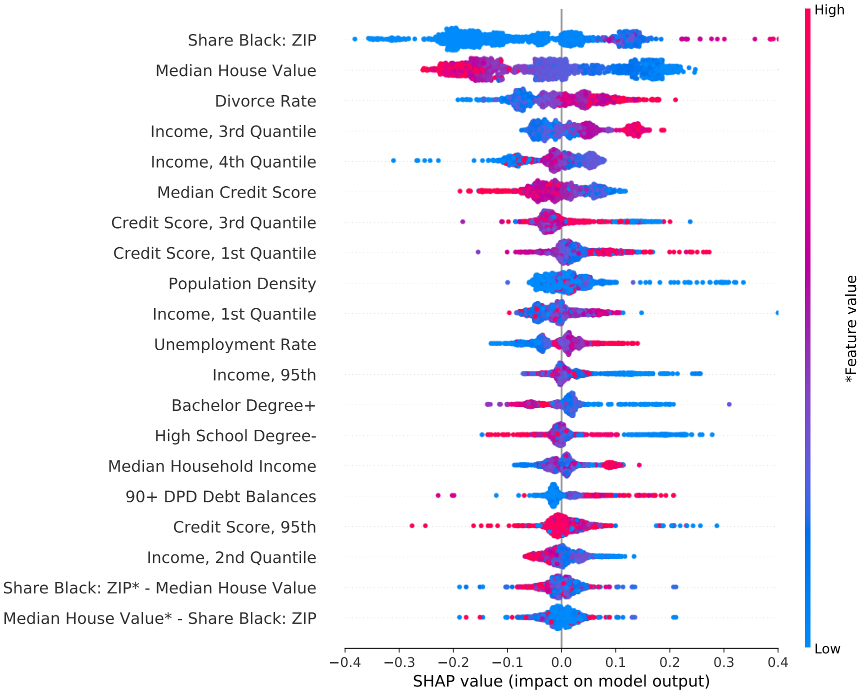


Figure 13: GBT Feature Explanations

Notes: This figure orders our independent variables in order of their importance in predicting judgments by our Gradient Boosted Trees. We use Shapley Additive exPlanations (SHAP) to explain the output of our Gradient Boosted Trees model. The SHAP value gives us individualized impacts for each predictor; positive SHAP values are associated with increased judgment rates and negative SHAP values are associated with decreased predicted judgment rates.

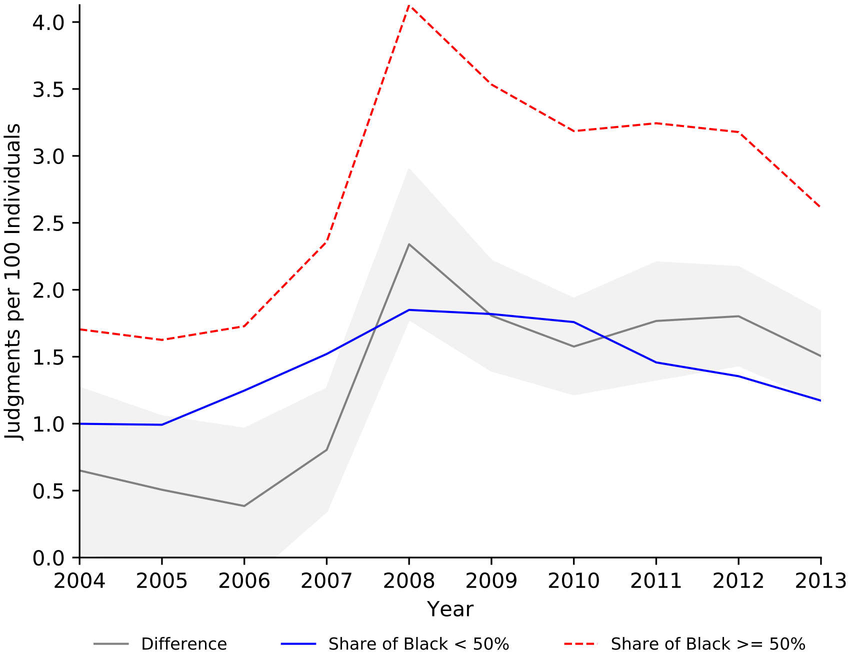


Figure 14: Disparity over Time

Notes: In this figure we graph the judgment rate for both majority black and majority non-black neighborhoods. We estimate the disparity in judgments by year and graph this disparity along with the 95% confidence interval.

Table 18: Summary Statistics

	Black	White	t-test
Panel A: Judgments			
Judgments per 100 People	2.73	1.43	-1.30***
	(1.31)	(0.91)	
Share of Default Judgments	0.45	0.38	-0.06***
	(0.07)	(0.12)	
Share of Consent Judgments	0.16	0.16	0.01
	(0.07)	(0.10)	
Share of Contested Judgments	0.06	0.05	-0.01*
	(0.04)	(0.06)	
Share w/ Attorney	0.04	0.10	0.06***
	(0.02)	(0.07)	
Panel B : Credit Variables			
Median Credit Score	606.21	647.30	41.09***
	(38.55)	(49.16)	
90+ DPD Debt Balances	3658.13	2540.28	-1117.85***
	(2911.64)	(6101.67)	
Banks (5 miles)	86.22	23.24	-62.99***
	(40.75)	(42.50)	
Payday Lenders (5 miles)	30.49	7.29	-23.20***
	(9.51)	(11.51)	
Panel C: Census Data			
Median Household Income (000s)	32.07	42.86	10.79***
	(12.04)	(12.07)	
GINI Index	0.46	0.42	-0.04***
	(0.06)	(0.05)	
Unemployment Rate	0.11	0.07	-0.04***
	(0.03)	(0.04)	
Divorce Rate	0.13	0.12	-0.01***
	(0.02)	(0.04)	
Median House Value (000s)	88.95	105.82	16.87***
	(35.83)	(41.42)	
Fraction with Bachelors Degree	0.17	0.19	0.02*
	(0.10)	(0.12)	
Fraction without High School Degree	0.19	0.15	-0.04***
	(0.06)	(0.07)	
Observations	227	2446	2673

Notes: Summary statistics for observations on the common support sample. Data is drawn from Missouri. Standard deviations are in parenthesis.

Table 19: Judgments, Income, and Credit Scores

	(1)	(2)	(3)	(4)	(5)	(6)
Black Majority: ZIP	1.2127***	0.9243***	0.7517***	0.7280***	0.5849***	0.6521***
	(0.0412)	(0.0877)	(0.0865)	(0.0690)	(0.1181)	(0.0555)
Median Household Income		0.0006		0.0137	0.0174	
		(0.0292)		(0.0180)	(0.0180)	
Median Credit Score			-0.0101***	-0.0085*	-0.0056	
			(0.0035)	(0.0044)	(0.0034)	
County Fixed Effects	X	X	X	X	X	X
Year Fixed Effects	X	X	X	X	X	X
Baseline Controls					X	
Income Quintiles		X		X	X	
Credit Quintiles			X	X	X	
Lagged Baseline Controls						X
Observations	2673	2673	2673	2673	2673	2407
R^2	0.5947	0.6355	0.6416	0.6527	0.6701	0.6823

Notes: Robust standard errors clustered at the county level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Dependent variable: Judgments per 100 individuals. All regressions are weighted by population and estimated on the common support sample.

Table 20: Judgments and Debt Portfolios

	(1)	(2)	(3)	(4)	(5)
Black Majority: ZIP	0.5853***	0.5873***	0.5957***	0.5614***	0.5671***
	(0.1193)	(0.1161)	(0.1123)	(0.1303)	(0.1339)
County Fixed Effects	X	X	X	X	X
Year Fixed Effects	X	X	X	X	X
Baseline Controls	X	X	X	X	X
Debt Levels	Yes				Yes
Monthly Payment and Utilization		Yes			Yes
Debt Composition			Yes		Yes
Delinquency/Bankruptcy/Collections				Yes	Yes
Observations	2673	2673	2673	2673	2673
R^2	0.6812	0.6733	0.6725	0.6745	0.6891

Notes: Robust standard errors clustered at the county level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Dependent variable: Judgments per 100 individuals. All regressions are weighted by population and estimated on the common support sample.

Table 21: Judgments and Lending Institutions

	(1)	(2)	(3)	(4)	(5)
Black Majority: ZIP	0.8265^{***}	0.8248^{***}	0.8263^{***}	0.7108^{***}	0.6965^{**}
	(0.2801)	(0.2786)	(0.2803)	(0.2121)	(0.2861)
Broadband		0.0930^{**}			
		(0.0431)			
Unscored			0.3830		
			(0.3923)		
Banks (5 miles)				-0.0068^{***}	
				(0.0020)	
Payday Lenders (5 miles)				0.0219^{***}	
				(0.0032)	
Banks (10 miles)					-0.0021^{***}
					(0.0001)
Payday Lenders (10 miles)					0.0099^{***}
					(0.0013)
County Fixed Effects	X	X	X	X	X
Year Fixed Effects	X	X	X	X	X
Baseline Controls	X	X	X	X	X
Observations	1703	1703	1703	1703	1703
R^2	0.8463	0.8474	0.8463	0.8597	0.8557

Notes: Robust standard errors clustered at the county level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Dependent variable: Judgments per 100 individuals. All regressions are weighted by population and estimated on the common support sample.

Table 22: Attorney Representation and Judgment Type

	(1) Attorney	(2) Judgments	(3) Consent	(4) Contested	(5) Default
Black Majority: ZIP	-0.013***	0.579***	0.006	0.002	0.010
	(0.004)	(0.121)	(0.004)	(0.006)	(0.008)
Attorney		-0.190	-0.015	0.058	-0.234***
		(0.691)	(0.059)	(0.050)	(0.057)
County Fixed Effects	X	X	X	X	X
Year Fixed Effects	X	X	X	X	X
Baseline Controls	X	X	X	X	X
Observations	2673	2673	2673	2673	2673
R^2	0.667	0.670	0.661	0.431	0.532

Notes: Robust standard errors clustered at the county level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. All regressions are weighted by population and estimated on the common support sample.

Table 23: Judgments by Plaintiff Type

	(1) Judgments	(2) Debt Buyer	(3) Major Bank	(4) Medical	(5) High-Cost	(6) Misc.
Black Majority: ZIP	0.582***	0.188***	0.041*	0.076*	0.105**	0.193***
	(0.118)	(0.045)	(0.022)	(0.040)	(0.047)	(0.049)
Mean	1.299	.58	.456	.288	.083	.161
Effect Size	44.8	32.5	8.9	26.4	125.7	119.7
County Fixed Effects	X	X	X	X	X	X
Year Fixed Effects	X	X	X	X	X	X
Baseline Controls	X	X	X	X	X	X
Observations	2673	2673	2673	2673	2673	2673
R^2	0.670	0.699	0.751	0.641	0.614	0.553

Notes: Robust standard errors clustered at the county level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. All regressions are weighted by population and estimated on the common support sample.

Table 24: Judgments and Other Measures of Racial Composition

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Share Black: ZIP	1.5253^{***}						
	(0.1532)						
Black Majority: ZIP		0.5819^{***}					
		(0.1175)					
Share Black: BISG			1.5163^{***}				7.1962^{***}
			(0.1389)				(1.5909)
Black Majority: BISG				0.5955^{***}		0.5553^{**}	
				(0.0905)		(0.2170)	
Share Black: Names					5.6546^{***}		
					(0.4711)		
County Fixed Effects	X	X	X	X	X		
ZIP Code Fixed Effects						X	X
Year Fixed Effects	X	X	X	X	X	X	X
Baseline Controls	X	X	X	X	X		
Income Quintiles	X	X	X	X	X		
Credit Score Quintiles	X	X	X	X	X	X	X
Observations	2673	2673	2671	2671	2671	2671	2671
R^2	0.6847	0.6704	0.6898	0.6769	0.6839	0.7296	0.7384

Notes: Robust standard errors clustered at the county level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Dependent variable: Judgments per 100 individuals. All regressions are weighted by population and estimated on the common support sample.

Table 25: Judgments and Other Demographic Groups

	(1)	(2)	(3)	(4)	(5)	(6)
Share Black: ZIP	2.1829***	1.8677***	1.6103***	1.6150***	1.4976***	1.6356***
	(0.0574)	(0.1594)	(0.1687)	(0.1021)	(0.1551)	(0.1052)
Share Asian: ZIP	-6.2263***	-3.2578***	-5.7555***	-3.3496***	-1.5249	-1.3548
	(1.2408)	(1.1516)	(1.0673)	(0.9776)	(1.3861)	(1.0703)
Share Hispanic: ZIP	1.3408***	0.1660	0.7119*	0.1395	-0.0104	0.1370
	(0.4118)	(0.3155)	(0.4169)	(0.3216)	(0.4319)	(0.3732)
Median Income		0.0193		0.0245**	0.0281**	
		(0.0170)		(0.0122)	(0.0129)	
Median Credit Score			-0.0061	-0.0051	-0.0033	
			(0.0037)	(0.0043)	(0.0039)	
County Fixed Effects	X	X	X	X	X	X
Year Fixed Effects	X	X	X	X	X	X
Baseline Controls					X	
Income Quintiles		X		X	X	
Credit Quintiles			X	X	X	
Lagged Baseline Controls						X
Observations	2673	2673	2673	2673	2673	2407
R^2	0.6521	0.6694	0.6689	0.6769	0.6850	0.6993

Notes: Robust standard errors clustered at the county level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Dependent variable: Judgments per 100 individuals. All regressions are weighted by population and estimated on the common support sample.

Appendix A - Predicting Consumer Default: A Deep Learning Approach

A.1 Performance Metrics

Suppose a binary classifier is given and applied to a sample of N observations. For each instance i, let y_i denote the true outcome. For each observation, the model generates a probability that an observation with feature vector x_i belongs to class 1. This predicted probability, $f(x_i)$ is then evaluated based on a threshold to classify observations into class 1 or 0. Given a threshold level (c), let True Positive (TP) denote the number of observations that are correctly classified as type 0, True Negative (TN) be the number of observations that are correctly classified as type 1, False Positive (FP) be the number of observations that are type 1 but incorrectly classified as type 0, and, finally, False Negative (FN) be the number of observations that are actually of type 0 but incorrectly classified as type 0. Based on these definitions, one can define the following metrics to assess the performance of the classifier:

$$\text{True Negative Rate (TNR)} \equiv \frac{\text{TN}}{\text{TN+FP}} \tag{A.1}$$

$$\text{False Positive Rate (FPR)} \equiv \frac{\text{FP}}{\text{FP+TN}} \tag{A.2}$$

$$\text{Precision} \equiv \frac{\text{TP}}{\text{TP + FP}} \tag{A.3}$$

$$\text{Recall} \equiv \frac{\text{TP}}{\text{TP + FN}} \tag{A.4}$$

$$\text{F-measure} \equiv \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Precision + Recall}} \tag{A.5}$$

$$\text{Accuracy} \equiv \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{A.6}$$

$$\text{Youden's J statistic} \equiv \frac{\text{TP}}{\text{TP + FN}} + \frac{\text{TN}}{\text{TN + FP}} - 1 \tag{A.7}$$

$$\text{ROC AUC} = \int_{\infty}^{-\infty} \text{TPR}(c)\text{FPR}'(c)\text{dc} \tag{A.8}$$

$$\text{Cross-entropy loss} = -\frac{1}{N}\sum_{i=1}^{N}(y_i \cdot \log(f(x_i)) + (1-y_i) \cdot \log(1-f(x_i)) \tag{A.9}$$

A.2 Data Pre-Processing

A.2.1 Sample Restrictions

Our original dataset contains 33,600,000 observations. We discard observations of individuals with missing birth information, deceased individuals and restrict our analysis to individuals aged between 18 and 85, residing in one of the 50 states or the District of Columbia, with 8 consecutive quarters of non-missing default behavior. This leaves us with 22,004,753 data points. Our itemized sample restrictions are summarized in Table 26 below.

A.2.2 Feature Scaling

We normalize all explanatory variables by their means and standard deviations:

$$z_i = \frac{x_i - \mu_x}{\sigma_x} \tag{A.10}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_k)$, and $\mathbf{z_i}$ is the i^{th} normalized data.

A.2.3 Train-Test Split

For most of our analysis we split the data to account for look-ahead bias, i.e., the training set consists of data 8Q prior to the testing data. Then, we scale the testing data by the mean and standard deviation of the training data.

In an alternative specification, we split our pooled data into three chunks: training set (60%), holdout set (20%), and testing set (20%). We report each specifications in Table 32 - Table 33. Except for parts of Section A.6, we used the predictions generated by our models on the temporal splits. In each specifications, we randomly shuffled the data to ensure that the mini-batch gradients are unbiased. If gradients are biased, training may not converge and accuracy may be lost.

A.2.4 Features

We summarize the list of features included in our model in Table 27, while Table 28 provides summary statistics for selected features.

A.2.5 Feature Groups

For the SHAP value analysis, we grouped features that had a correlation higher than 0.7. These groups are presented in Table 29.

A.3 Machine Learning Models

A.3.1 Deep Neural Network (DNN)

Figure 15 illustrates an example of a two layer neural network. This neural network has 3 input units (denoted x_1, x_2, x_3), 4 hidden units, and 1 output unit. Let n_l denote the number of layers in this network ($n_l = 2$). We label layer l as L_l, where layer L_0 is the input layer, and layer $L_{L=2}$ is the output layer. The layers between the input ($l = 0$) and the output layer ($l = L$) are called hidden layers. Given this notation, there are $L - 1$ hidden layers, 1 in this specific example. A neural network without any hidden layers ($L = 1$) is a logistic regression model.

There are two ways to increase the complexity a neural network: (1) increase the number of hidden layers and (2) increase the number of units in a given layer. Lower tier layers in the neural network learn simpler patterns, from which higher tier layers learn to produce more complex patterns. Given a sufficient number of neurons, neural networks can approximate continuous functions on compact sets arbitrarily well (see [49] and [48]). This includes approximating interactions (i.e., the product and division of features). There are two main advantages of adding more layers over increasing the number of units to existing layers; (1) later layers build on early layers to learn features of greater complexity and (2) deep neural networks– those with three or more hidden layers– need exponentially fewer neurons than shallow networks ([15] and [67]).

In the neural network represented in Figure 15, the parameters to be estimated are $(W, b) = (W^{(0)}, b^{(0)}, W^{(1)}, b^{(1)})$, where $W_{ij}^{(l)}$ denotes the weight associated with the connection between unit j in layer l and unit i in layer $l+1$, and $b_i^{(l)}$ is the bias associated with unit i in layer $l+1$. Thus, in this example $W^{(0)} \in \mathbb{R}^{3\times 4}, b^{(0)} \in \mathbb{R}^{4\times 1}$ and $W^{(1)} \in \mathbb{R}^{1\times 4}, b^{(1)} \in \mathbb{R}$. This implies that there are a total of 21 = (3+1)*4+5 parameters (four parameters to reach each neuron and five weights to aggregate the neurons into a single output). In general, the number of weight parameters in each hidden layer l is $N^{(l)}(1 + N^{(l-1)})$, plus $1 + N^{(L-1)}$ for the output layer, where $N^{(l)}$ denotes the number of neurons in each layer $l = 1, \ldots, L$.

Let $a_i^{(l)}$ denote the activation (e.g., output value) of unit i in layer l. Fix W and b, our neural network defines a hypothesis $h_{W,b}(x)$ that outputs a real number between 0 and 1.[1] Let $f(\cdot)$ denote the activation function that applies to vectors in an element-wise fashion. The computation this neural network represents, often referred to as forward propagation, can be written as:

$$z^{(1)} = W^{(0),T}x + b^{(0)}$$

$$a^{(1)} = f(z^{(1)})$$

$$z^{(2)} = W^{(1),T}a^{(1)} + b^{(1)}$$

$$h_{W,b}(x) = a^{(2)} = f(z^{(2)})$$

There are many choices to make when structuring a neural network, including the number of hidden layers, the number of neurons in each layer, and the activation functions. We built a number of network architectures having up to fifteen hidden layers.[2] All architectures are fully connected so each unit receives an input from all units in the previous layer.

[1] This is a property of the sigmoid activation function.

[2] The number of layers and the number of neurons in each layer, along with other hyperparameters of the model, are chosen by Tree-structured Parzen Estimator (TPE) approach. See Appendix A.4 for more details.

Neural networks tend to be low-bias, high-variance models, which imparts them a tendency to over-fit the data. We apply dropout to each of the layers to avoid over-fitting (see [79]). During training, neurons are randomly dropped (along with their connections) from the neural network with probability p (referred to as the dropout rate), which prevents complex co-adaptations on training data.

We apply the same activation function (rectified linear unit or RELU) at all nodes, which is obtained via hyperparameter optimization,[3] and defined as:

$$\text{RELU}(x) = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{A.11}$$

Let $N^{(l)}$ denote the number of neurons in each layer $l = 1,\ldots,\ L$. Define the output of neuron k in layer l as $z_k^{(l)}$. Then, define the vector of outputs (including the bias term $z_0^{(l)}$) for this layer as $z^{(l)} = (z_0^{(l)}, z_1^{(l)}, \ldots, z_{N^{(l)}}^{(l)})'$. For the input layer, define $z^{(0)} = (x_0^{(l)}, x_1^{(l)}, \ldots, x_{N^{(l)}}^{(l)})'$. Formally, the recursive output of the $l - th$ layer of the neural network is:

$$z^{(l)} = \text{RELU}(W^{(l-1),T} z^{(l-1)} + b^{(l-1)}), \tag{A.12}$$

with final output:

$$h_\theta(x) = g(W^{(L-1),T} z^{(L-1)} + b^{(L-1)}). \tag{A.13}$$

The parameter specifying the neural network is:

$$\theta = (W_0, b_0, \ldots, W_{L-1}, b_{L-1}) \tag{A.14}$$

A.3.2 Decision Tree Models

The second component of our model is Extreme Gradient Boosting, which builds on decision tree models. Tree-based models split the data several times based on certain cutoff values in the explanatory variables.[4] A number of such models have become quite prevalent in the literature, most notably random forests (see [21] and [23]) and Classification and Regression Trees, known as CART. We briefly review CART and then explain gradient boosting.

[3] There are many potential choices for the nonlinear activation function, including the sigmoid, relu, and tanh.

[4] Splitting means that different subsets of the dataset are created, where each observation belongs to one subset. For a review on decision trees, see [55].

A.3.2.1 CART There are a number of different decision tree-based algorithms. As an illustration of the approach, we describe Classification and Regression Trees or CART. CART models an outcome y_i for an instance i as follows:

$$\hat{y}_i = \hat{f}(x_i) = \sum_{m=1}^{M} c_m I\{x_i \in R_m\}, \tag{A.15}$$

where each observation x_i belongs to exactly one subset R_m. The identity function I returns 1 if x_i is in R_m and 0 otherwise. If x_i falls into R_l, the predicted outcome is $\hat{y} = c_l$, where c_l is the mean of all training observations in R_l.

The estimation procedure takes a feature and computes the cut-off point that minimizes the Gini index of the class distribution of $\mathbf{y}$, which makes the two resulting subsets as different as possible. Once this is done for each feature, the algorithm uses the best feature to split the data into two subsets. The algorithm is then repeated until a stopping criterium is reached.

Tree-based models have a number of advantages that make them popular in applications. They are invariant to monotonic feature transformations and can handle categorical and continuous data in the same model. Like deep neural networks, they are well suited to capturing interactions between variables in the data. Specifically, a tree of depth L can capture $(L-1)$ interactions. The interpretation is straightforward, and provides immediate counterfactuals: “If feature x_j had been bigger / smaller than the split point, the prediction would have been $\bar{y}_0$ instead of $\bar{y}_1$.” However, these models also have a number of limitations. They are poor at handling linear relationships, since tree algorithms rely on splitting the data using step functions, an intrinsically non-linear transformation. Trees also tend to be unstable, so that small changes in the training dataset might generate a different tree. They are also prone to overfitting to the training data. For more information on tree-based models see [66].

A.3.2.2 Gradient Boosted Trees (GBT) At each step m, $1 \leq m \leq M$, of gradient boosting, an estimator, h_m, is computed on the residuals from the previous models predictions. A critical part of gradient boosting method is regularization by shrinkage as proposed by [42]. This consists in modifying the update rule as follows:

$$F_m(x) = F_{m-1}(x) + \nu\gamma_m h_m(x), \tag{A.16}$$

where $h_m(x)$ represents a weak learner of fixed depth, γ_m is the step length and ν is the learning rate or shrinkage factor.

XGBoost is a fast implementation of Gradient Boosting, which has the advantages of fast speed and high accuracy. For classification, XGBoost combines the principles of decision trees and logistic regression, so that the output of our XGBoost model is a number between 0 and 1. For the remainder of the paper we refer to XGBoost as GBT.[5]

A.4 Model Estimation

Our estimation consists of seven steps. First, we specify the loss function. Second, we choose the optimization algorithm. Third, we optimize the hyperparameters of our GBT model. Fourth, we train the GBT model. Fifth, we restrict our feature set using the GBT model. Sixth, we optimize the hyperparameters (including the weighting parameter for our hybrid models), and seventh, we train our models.

A.4.1 Loss Function

Suppose $\mathbf{y}$ is the ground truth vector of default, and $\hat{\mathbf{y}}$ is the estimate obtained directly from the last layer given input vector $\mathbf{x} = (x_1, x_2, \ldots, x_k)$. By construction, $y_i = \{0, 1\}$ and $\hat{y}_i \in [0, 1]$. We minimize the categorical cross-entropy loss function[6] to estimate the parameter specified in (7). We do this by choosing θ that minimizes the distance between

[5] For more on XGBoost, see [28] and [73].

[6] Loss function measures the inconsistency between the predicted and the actual value. The performance of a model increases as the loss function decreases. There are several other types of loss functions, including mean squared error, hinge, and Poisson. The categorical cross-entropy is often used for classification problems.

the predicted $\hat{\mathbf{y}}$ and the actual $\mathbf{y}$ values. Given N training examples, the categorical cross-entropy loss can be written as:

$$L(\hat{y}, y) = -\frac{1}{N}\sum_{i=1}^{N}(y_i \cdot log(\hat{y}_i) + (1 - y_i) \cdot log(1 - \hat{y}_i) \qquad \text{(A.17)}$$

We apply an iterative optimization algorithm to find the minimum of the categorical cross-entropy loss function. We next describe this algorithm.

A.4.2 DNN Optimization Algorithm

Deep learning models are computationally demanding due to their high degree of non-linearity, non-convexity and rich parameterization. Given the size of the data, gradient descent is impractical. We follow the standard approach of using stochastic gradient descent (SGD) to train our deep learning models (see [44]). Stochastic gradient descent is an iterative algorithm that uses small random subsets of the data to calculate the gradient of the objective function. Specifically, a subset of the data, referred to as a mini-batch (the size of the mini-batch is called the batch size), is loaded into memory and the gradient is computed on this subset. The gradient is then updated, and the process is repeated until convergence.

We adopt the Adaptive Moment Estimation (Adam), a computationally efficient variant of the SGD introduced by (see [56]) to train our neural networks. The Adam optimization algorithm can be summarized as follows:

1. Fix the learning rate α, the exponential decay rates for the moment estimates: β_1,$\beta_2 \in [0, 1)$, and the objective function. Initialize the parameter vector θ_0, the first and second moment vector m_0 and v_0 respectively, and the timestep t.
2. While θ_t does not converge, do the following:
   a. Compute the gradients with respect to the objective function at timestep t:

$$g_t = \nabla_\theta f_t(\theta_{t-1}) \qquad \text{(A.18)}$$

b. Update the first and second moment estimates:

$$m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t \tag{A.19}$$

$$v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2 \tag{A.20}$$

c. Compute the bias-corrected first and second moment estimates:

$$\hat{m_t} = \frac{m_t}{1 - \beta_1^t} \tag{A.21}$$

$$\hat{v_t} = \frac{v_t}{1 - \beta_2^t} \tag{A.22}$$

d. Update the parameters:

$$\theta_t = \theta_{t-1} - \frac{\alpha \cdot \hat{m_t}}{\sqrt{\hat{v_t}} + \epsilon} \tag{A.23}$$

The hyperparameters have intuitive interpretations and typically require little tuning. We apply the default setting suggested by the authors of [56], these are $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-7}$.

A.4.3 GBT Algorithm

Fit a shallow tree (e.g., with depth L = 1). Using the prediction residuals from the first tree, fit a second tree with the same shallow depth L. Weight the predictions of the second tree by $\nu \in (0, 1)$ to prevent the model from overfitting the residuals, and then aggregate the forecasts of these two trees. Until a total of K trees is reached in the ensemble, at each step k, fit a shallow tree to the residuals from the model with k-1 trees, and add its prediction to the forecast of the ensemble with a shrinkage weight of ν.

A.4.4 Regularization

Neural networks are low-bias, high-variance models (i.e., they tend to overfit to their training data). We implement three routines to mitigate this. First, we apply dropout to each of the layers (see [79]). During training, neurons are randomly dropped (along with their connections) from the neural network with probability p (referred to as the dropout rate), which prevents complex co-adaptations on training data.

Second, we implement "early stopping", a general machine learning regularization tool. After each time the optimization algorithm passes through the training data (i.e., referred to as an epoch), the parameters are gradually updated to minimize the prediction errors in the training data, and predictions are generated for the validation sample. We terminate the optimization when the validation sample loss has not decreased in the past 50 epochs. Early stopping is a popular substitute to l2 regularization, since it achieves regularization at a substantially lower computational cost.

Last, we use batch normalization (see [53]), a technique for controlling the variability of features across different regions of the network and across different datasets. It is motivated by the internal covariate shift, a phenomenon in which inputs of hidden layers follow different distributions than their counterparts in the validation sample. This problem is frequently encountered when fitting deep neural networks that involve many parameters and rather complex structures. For each hidden unit in each training step, the algorithm cross-sectionally de-means and variance standardizes the batch inputs to restore the representation power of the unit.

A.4.5 Feature Selection

First we remove a subset of features which are classified at the discretion of the lender and may be inconsistent across lenders, trades, and borrowers. These includes any variables pertaining to charge-off, derogatory in isolation, Fannie Mae, Freddie Mac, presence of outstanding govt agency debt, utility trades, and judgements. Then, we exclude features having limited impact on the model. To do so, we use data from 2004Q1 with test set from 2006Q1, and train a GBT model. We extract each features' feature importance, and sort our features based on this metric. We then iteratively remove features whose feature importance score is the lowest among our features, and train a GBT model with the corresponding feature set. We do this until there is only one feature left. We keep the lowest possible number of features that have the same or better predictive power than the model with the full set of features.

We then compute pairwise correlations for our full set of features, and add variables whose pairwise correlation exceeds 0.7 (since high correlation impacts feature importance). Next, we remove features so that no variables have a pairwise correlation over 0.9[7]. Our feature selection leaves us with a total of 88 features, and we report summary statistics for the forty most influential features (we rank their influence based on the perturbation exercise) in Table 28.

A.4.6 Hyperparameter Selection

Deep learning models require a number of hyperparameters to be selected. We follow the standard approach by cross-validating the hyperparameters via a validation set. We fix a training and validation set, and then train neural networks with different hyperparameters on the training set and compare the loss function on the validation set. We cross-validate the number of layers, the number of units per layer, the dropout rate, the batch size, and the activation function (i.e., the type of non-linearity) via Tree-structured Parzen Estimator (TPE) approach (see [16]),[8] and select the hyperparameters with the lowest validation loss.

The training set for our out-of-sample hyperparameter optimization comes from 2004Q1, while the test set is from 2006Q1. Table 30 summarizes our machine learning model hyperparameters. For our neural network, we used 5 hidden layers, with 128-256-256-256-512 neurons per layer, RELU activation function, a batch size of 4096, a learning rate of 0.003, and a dropout rate of 50%. For our GBT, We found that a learning rate of 0.05, a max tree depth of 5, a max bin size of 256, with 1000 trees gave us good performance. All GBT

[7] When it comes to tie-breakers, we keep the more generic feature; e.g., total debt balances are kept over total mortgage balances.

[8] We use TPE since it outperformed random search (see [16]), which was shown to be both theoretically and empirically more efficient than standard techniques such as trials on a grid. Other widely used strategies are grid search and manual search.

models were run until their validation accuracy was non-improving for a hundred rounds and were trained on CPUs.

For the pooled sample prediction, we increased the batch size to 16,384 and the number of neurons per layers to 512,1024,2048,1024,512; while decreased the dropout rate to 20%, keeping the activation function, and the learning rate unchanged. We instituted early stopping with a patience of 1,000 for GBT, and trained a model of depth 6 with up to 10,000 trees and a learning rate of 0.3. We report the results of the best performing GBT.

A.4.7 Weighting

We use a grid search on our model trained on 2004Q1 data along with the 2006Q1 test data to find the optimal weights for our out-of-sample exercise, and we keep this weight constant going forward. For our pooled sample, we used the test data for finding the optimal weight[9].The results are reported in Table 31, and notice that the sample corresponds to Table 39. Based on this exercise, the optimal weight on the DNN is 0.2 for the out-of-sample exercise, and 0.7 for the pooled exercise.

A.4.8 Implementation

We include our listed features for each individual. Since we work with panel data, there is a sample for each quarter of data. We train roughly 20 million samples, which takes up around 20 gigabytes of data. Our deep learning models are made up of millions of free parameters. Since the estimation procedure relies on computing gradients via backpropagation, which tends to be time and memory intensive, using conventional computing resources (e.g., desktop) would be impractical (if not infeasible). We address the time and memory intensity with two methods. First, to save memory, we use single precision floating point operations, which halves the memory requirements and results in a substantial computational speedup. Second, to accelerate the learning, we parallelized our computations and trained all of our

[9] We only use the pooled sample for some interpretability and as a benchmark for our out-of-sample exercises, and such we are interested in finding the model with the best predictive power. We do not compare the pooled model's performance with the credit scores.

models on a GPU cluster[10]. In our setting, GPU computations were over 40X faster than CPU for our deep neural networks. For a discussion on the impact of GPUs in deep learning see [77].

We conduct our analysis using Python 3.6.3 (Python Software Foundation), building on the packages numpy ([82]), pandas ([65]) and matplotlib ([51]). We develop our deep neural networks with keras ([31]) running on top of Google TensorFlow, a powerful library for large-scale machine learning on heterogenous systems ([1]). We run our machine learning algorithms using sci-kit learn ([71]) and ([29]).

A.5 Classifier Performance

In this section, we describe the performance of our hybrid model under various training and testing windows. First, we evaluate our model on the pooled sample (2004Q1-2013Q4), where we apply a random 60%-20%-20% split to our training, validation, and testing sets. Then, to account for look-ahead bias, we train and test our models based on 8 quarter windows that were observable at the time of forecast. In particular, we require our training and testing sets to be separated by 8 quarters to avoid overlap. For instance, the second out-of-sample model was calibrated using input data from 2004Q2, from which the parameter estimates were applied to the input data in 2006Q2 to generate forecasts of delinquencies over the 8 quarter window from 2006Q3-2008Q2. This gives us a total of 32+1 calibration and testing periods reported in Table 32. The percentage of 90+ days past due accounts within 8 quarters varies from 32.5% to 35.9%.

The hybrid model outputs a continuous variable that, under certain circumstances, can be interpreted as an estimate of the probability of an account becoming 90+ days delinquent during the subsequent 8 quarters. One measure of the model's success is its ability to differentiate between accounts that did become delinquent and those that did not; if these two groups have the same forecasts, the model provides no value. Table 32 presents the average forecast for accounts that did and did not fall into the 90+ days delinquency category

[10] 1 node with 4 NVIDIA GeForceGTX1080 GPUs. The pooled model trains within 24 hours.

over the 32+1 evaluation periods. For instance, during the testing period for 2010Q4, the model's average prediction among the 35.44% of accounts that became 90+ days delinquent was 73.12%, while the average prediction among the 64.56% of accounts that did not was 16.18%. We should highlight that these are truly out-of-sample predictions, since the model is calibrated using input data from 2008Q4. This shows the forecasting power of our model in distinguishing between accounts that will and will not become delinquent within 8 quarters. Furthermore, this forecasting power seems to be stable over the 32+1 calibration and evaluation periods, partly driven by the frequent re-calibration of the model that captures some of the changing dynamics of consumer behavior.

We also look at accounts that are current as of the forecast date but become 90+ days delinquent within the subsequent 8 quarters. In particular, we contrast the model's average prediction among individuals who were current on their accounts but became 90+ days delinquent with the average prediction among customers who were current and did not become delinquent. Given the difficulty of predicting default among individuals that currently show no sign of delinquency, we anticipate the model's performance to be less impressive than the values reported in Table 32. Nonetheless, the values reported in Table 33 indicate that the model is able to distinguish between these two populations. For instance, using input data from 2008Q4, the average model prediction for individuals who were current on their debts and became 90+ days delinquent is 40.37%, contrasted with 10.53% for those who did not. As in Table 32, the model's ability to distinguish between these two classes is consistent across the 32+1 evaluation periods listed in Table 33.

Under certain conditions, the forecasts generated by our model can be converted to binary decisions by comparing the forecast to a specified threshold and classifying accounts with scores exceeding that threshold as high-risk. Setting the threshold level comes with a trade-off. A low level threshold leads to many accounts being classified as high risk, and even though this approach may accurately capture customers who are actually high-risk and about to default on their payments, it can also give rise to many low-risk accounts incorrectly classified as high-risk. By contrast, a high threshold can result in too many high-risk accounts being classified as low-risk.

This type of trade-off is inherent in any classification problem, and involves trading off

Type-I (false positives) and Type-II (false negatives) errors in a classical hypothesis testing context. In the credit risk management context, a cost/benefit analysis can be formulated contrasting false positives to false negatives to make this trade-off explicit, and applying the threshold that will optimize an objective function in which costs and benefits associated with false positives and false negatives are inputs.

A commonly used performance metric in the machine learning and statistics literature is a 2×2 contingency table, often referred to as the *confusion matrix*, that describes the statistical behavior of any classification algorithm. In our application, the two rows correspond to ex post realizations of the two types of accounts in our sample, *no default* and *default*. We define *no default* accounts as those who do not become 90+ days delinquent during the forecast period, and *default* accounts as those who do. The two columns correspond to ex ante classifications of the accounts into these categories. If a predictive model is applied to a set of accounts, each account falls into one of the four cells in the confusion matrix, thus the performance of the model can be assessed by the relative frequencies of the entries. In the Neymann-Pearson hypothesis-testing framework, the lower-left entry is defined as Type-I error and the upper right as Type-II error, while the objective of the researcher is to minimize Type-II error (i.e., maximize "power") subject to a fixed level of Type-I error (i.e., "size").

As an illustration, Figure 16 Panel (a) shows the confusion matrix for our hybrid DNN-GBT model calibrated using 2011Q4 data and evaluated on 2013Q4 data and a threshold of 50%. This means that accounts with estimated delinquency probabilities greater than 50% are classified as default and 50% or below as no default. For this quarter, the model classified 61.34% + 7.29% = 68.63% of the accounts as no default, of which 61.34% did indeed not default and 7.29% actually defaulted, that is, they were 90+ days delinquent in the subsequent 8 quarters. By the same token, of the 5.9% + 25.47% = 31.37% borrowers who defaulted, the model accurately classified 25.47%. Thus, the model's accuracy, defined as the percent of instances correctly classified, is the sum of the entries on the diagonal of the confusion matrix, that is, 61.34 % + 25.47% = 86.81%.

We can compute three additional performance metrics from the entries of the confusion matrix, which we describe heuristically here and define formally in the appendix. *Precision* measures the model's accuracy in instances that are classified as default. *Recall* refers to the number of accounts that defaulted as identified by the model divided by the actual number of defaulting accounts. Finally, the *F-measure* is simply the harmonic mean of precision and recall. In an ideal scenario, we would have very high precision and recall.

We can track the trade-off between true and false positives by varying the classification threshold of our model, and this trade-off is plotted in Figure 16 Panel (b). The blue line, called the Receiver Operating Characteristic (ROC) curve, is the pairwise plot of true and false positive rates for different classification thresholds (green line), and as the threshold decreases, the figure shows that the true positive rate increases, but so does the false positive rate. The ROC curve illustrates the non-linear nature of the trade-offs, implying that increase in true positive rates is not always proportionate with the increase in false positive rates. The optimal threshold then considers the cost of false positives with respect to the gain of true positives. If these are equal, the optimal threshold will correspond to the tangent point of the ROC curve with the 45 degree line.

The last performance metric we consider is the area under the ROC curve, known as AUC score, which is a widely used measure in the machine-learning literature for comparing models. It can be interpreted as the probability of the classifier assigning a higher probability of being in default to an account that is actually in default. The ROC area of our model ranges from 0.9238 to 0.9300, demonstrating that our machine-learning classifiers have strong predictive power in separating the two classes.

Table 34 reports the performance metrics widely used in the machine-learning literature for each of the 32+1 models discussed. Our models exhibit strong predictive power across the various performance metrics. For instance, the 85.71% precision implies that when our classifier predicts that someone is going to default, there is an 85.71% chance this person will actually default; while the 72.67% recall means that we accurately identified 72.67% of all the defaulters. Our approach of using only one quarter of data to train the model is rather restrictive. Using more quarters usually increases model performance, so since most credit scoring applications will use a training data that exceeds one quarter, performance metrics are likely to improve relative to what we report in our exercise.

Table 35 reports the same performance metrics for the population of borrowers who are current, that is, they do not have any delinquencies in the quarter they are assessed. As previously noted, this is a smaller population with a lower probability of default. Performance metrics drop marginally relative to those for the model applied to the population of all borrowers but they are still very strong. For example, the AUC score drops from 92-93% to 86-88%, accuracy and loss mostly remain in the same range.

A.6 Model Interpretation

We use our hybrid DNN-GBT model to uncover associations between the explanatory variables and default behavior. Since we do not identify causal relationships, our goal is simply to find covariates that have an important impact on default outcomes. Our findings can be used to better understand default behavior, further refine model specification and possibly aid in the formulation of theoretical models of consumer default. For this exercise, we mainly use the pooled model, which uses all available data. This allows us to assess factors that are critical in default behavior throughout the sample period with the best performing model. We also consider time variation in the factors influencing the default decision in subsets of our sample.

A.6.1 Explanatory Power of Variables

We start by examining the explanatory power of each of our features. We follow an approach similar to [78], which amounts to a perturbation analysis on the pooled sample using our hybrid model. First, we draw a random sample of 100,000 observations from the testing sample. Then, for each variable, we re-shuffle the feature, keeping the distribution intact and the model's loss function is evaluated with the changed covariate. We repeat this step 10 times, and report the average of the loss and accuracy. Then, the variable is replaced to its original values, and a perturbation test is performed on a new variable. Perturbing the variable of course reduces the accuracy of the model, and the test loss becomes larger. If a particular variable has strong explanatory power, the test loss will significantly increase. The test loss for the complete model when no variables are perturbed is the Baseline value. Features that have large explanatory power, and whose information is not contained in the other remaining variables will increase the loss significantly if they are altered. Table 36 reports the results. Features relating to credit history, debt balances, and the number and credit available on revolving trades dominate the list. Specifically, credit amount on open trades increases the loss by 24%, debt balances by 17%, the number of open credit and bankcards by 15%, while the total monthly payment on open trades, months since oldest trade, and months since the most recent 90+ days delinquency each increase the loss by 11%. These results suggest that debt balances, length of credit history and temporal proximity to a delinquency are all important factors in default behavior. Based on publicly available information, length of the credit history is also an important determinant of standard credit scoring models, though payment history rather than balances or number of trades is understood as the most critical. This approach to assessing the importance of different features for the predicted probability of default has two major shortcomings. First, when features are highly correlated, the interpretation of feature importance can be biased by unrealistic data instances. To illustrate this problem, consider two highly correlated features. As we perturb one of the features, we create instances that are unlikely or even impossible. For example, mortgage balances are highly correlated with and lower than total debt balances, yet this perturbation approach could create instances in which total debt balances are smaller than mortgage balances. Since many of the features are strongly correlated, care must be taken with interpretation of feature importance. We list the highly correlated features in Appendix A.2. An additional concern with this perturbation approach is that the distribution of some features are highly skewed, which implies that the probability of their value being different than where the mass of their distribution is concentrated is quite low. Moreover, skewness varies substantially across features, therefore the informativeness of the perturbation may differ across variables. In the next section, we examine a more robust approach that is less susceptible to these limitations.

A.6.2 Economic Significance of Variables

We now turn to analyzing the economic significance of our features for default behavior. We adopt SHapley Additive exPlanations (SHAP), a unified framework for interpreting predictions, to explain the output of our hybrid deep learning model (for a detailed description of the approach see [63]). SHAP uses a game theoretical concept to assign each feature a local importance value for a given prediction. Though Shapley values are local by design, they can be combined into global explanations by averaging the absolute Shapley values featurewise. Then, we can compare features based on their absolute average Shapley values, with higher values implying higher feature importance. Similarly to permutation feature importance, SHAP is a feature importance measure. The main difference between the two is that while permutation feature importance is based on the decrease in model performance, SHAP is based on the magnitude of feature attributions.

We first compute the Shapley values for the Deep Neural Network model and the Gradient Boosted Trees model separately, then simply average them for each individual and for each feature.[11] We use a random sample of 100,000 observations for explaining the model. By the Shapley efficient property, the SHAP values for an observation sum up to the difference between the predicted value of that observation and the expected value, computed using the background dataset:

$$f(x) = E_X[\widehat{f}(X)] + \sum_{j=1}^{M} \phi_j \tag{A.24}$$

where f is the model prediction, M is the number of features, and $\phi_j \in R$ is the feature attribution for feature j (i.e., the Shapley values). Thus, we can interpret the Shapley value as the contribution of a feature value to the difference between the model's prediction and the mean prediction, given the current set of feature values. As an illustration, a SHAP value of 0.1 implies that the feature's value for that particular instance contributed to an increase of 0.1 to the predicted probability compared to the mean prediction. Features that are highly correlated can decrease the importance of the associated feature by splitting the importance

[11] We implement Deep SHAP, a high-speed approximation algorithm for SHAP values in deep learning models to compute the Shapley values for our 5 hidden layer neural network. For GBT, we implement TreeExplainer, a high-speed exact algorithm for tree ensemble methods. Because our dataset is fairly large with many features, we pass a random sample of 100 observations, referred to as background observations, to compute the expected value for both models.

between both features. We account for the effect of feature correlation on interpretability by grouping features with a correlation larger than 0.7, and summing the SHAP values within each groups. We denote these groups with an asterisk for the rest of the analysis and report the composition of feature groups in Table 29 in the appendix.

Figure 17 sorts features by the sum of absolute SHAP value magnitudes, and plots the distribution of the impact each feature has on the model output for the twelve most important features or groups of correlated features. The color represents the feature value (red: high, blue: low), whereas the position on the horizontal axis denotes the contribution of the feature. The charts plot the distribution of SHAP values for individual instances in the 100,000 testing sample. The most important feature in terms of SHAP value magnitude is the worst status on any trades. High values of this variable tend to increase predicted default risk, whereas low values tend to decrease it, though the distribution of instances is dispersed. Features capturing credit history, such as length of credit history and recent delinquencies, also have high SHAP values, specifically, high values of these variables lower predicted default risk, with a much more dispersed distribution. Additionally, delinquent balances and outstanding collections are typically associated with an increase in predicted default probability. Higher total debt balances are also associated with a lower than expected predicted default risk, reiterating the notion that the borrowers with the most credit are also associated with lower predicted probability of default, which suggests that credit allocation decisions are made to minimize default probabilities. As in the perturbation exercise, we find that number of trades and balances seem to have the strongest association with variation in the predicted probability of default, whereas credit inquiries do not play a sizable role.

These results only point to correlations between the features and the predicted outcome and should not be interpreted causally. Yet, they can be used as a point of departure for a causal analysis of default and theoretical modeling. They are also important to comply with legal disclosure requirements. Both the Fair Credit Reporting Act ad the Equal Opportunity in Credit Access Act require lenders and developers of credit scoring models to reveal the most important factors leading to a denial of a credit application and for credit scores. The SHAP value provides an individualized assessment of such factors that can be used for making credit allocation decisions and communicating them to the borrower.

A.6.3 Temporal Determinants of Default

We next look at the changing dynamics of default behavior by comparing models that are trained in different periods of time. For this analysis, we use our hybrid model. Specifically, we target the following time periods: 2006Q1, 2008Q1 and 2011Q1 as time periods before, during and after the 2007-2009 crisis, and compute default predictions for them with data trained in the same quarter and two years prior, that is in 2004Q1, 2006Q1, 2007Q1 and 2009Q1, respectively. We then calculate Shapley values for the two models.[12] The first exercise provides an in-sample assessment for feature importance, while the second exercise can be used to assess feature importance out-of-sample. In both exercises, the model is the same, so comparing the results from the two exercises can help uncover which features are important for default prediction for a given period from an ex ante perspective and from an in-sample perspective. Table 37 reports the results.[13] For each period, it is interesting to compare the variation in SHAP values from an ex ante and contemporaneous perspective, and additionally we are interested in comparing variation in SHAP values for given features in the different time periods. In most testing windows, the temporal proximity to delinquency has the highest SHAP value.

Debt balances (i.e, total debt, revolving debt, auto debt), and utilization on installment loans are consistently among the five most influential features. Mortgage debt and credit card debt, are generally between the fifth and twenty-first most significant in terms of SHAP values. The SHAP value is quite stable over time for most features, but there are some variables for which it changes substantially. One example is number of open credit cards which ranks second and seventh for 2004Q1 and 2006Q1 but moves down to fourteenth in sample and nineteenth out-of sample for 2011Q1. The length of the credit history is never among the thirty most important features. Overall these results confirm our findings from the pooled model, suggesting the balances and number of trades, in addition to delinquency status, have a strong association with default risk according to our model.

[12] We do this for both the Deep Neural Network and the Gradient Boosted Trees and similarly to how we obtain the output, we simply take the average of the Shapley values. For both our models, we use a random sample of 100 observations of the testing data scaled by the mean and standard deviation of the corresponding training data for reference value.

[13] The features are sorted by the sum of absolute SHAP value magnitudes over the first period.

A.7 Model Comparisons

To better assess the validity of our approach, we compare our deep learning model to logistic regression and a number of other machine learning models. Deep learning models feature multiple hidden layers, designed to capture multi-dimensional feature interactions. By contrast, logistic regression can be interpreted as a neural network without any hidden layers.

A.7.1 Hidden Layers

To motivate our choice of deep learning, leading to our hybrid DNN-GBT model, we begin by illustrating the importance of hidden layers that enable us to capture multi-level interactions between features by comparing how neural networks of different depth perform on the pooled sample. For this exercise, we fix the number of neurons per layer at 512, and build neural networks up to 5 hidden layers.[14]

We benchmark our results against the logistic regression, which is a commonly used technique in credit scoring and can be interpreted as a neural network with no hidden layers. Table 38 reports the in- and out-of-sample behavior for neural networks with 0-5 hidden layers. The number of hidden layers measure the complexity of the network, and we found that the marginal improvements in performance beyond 3 layers are small. Table 38 also shows that applying dropout improves the out-of-sample fit for networks of higher depths. This demonstrates that dropout serves as an effective regularization tool and addresses overfitting for networks of greater depths.

The results in Table 38 suggest that there are complex non-linear relationships among the features used as inputs in the model. This is further supported by the fact that permitting non-linear relationships between default behavior and explanatory variables produces the largest model improvement. Going from a linear model (0 layers) to the simplest non-linear model (1 layer) generates the most sizable reduction in out-of-sample loss. To see this from another angle, we plot the ROC curves for our neural networks considered in Figure 18.

[14] The architecture reported in Table 38 was not optimized. We picked 512, as it is exactly the number of neurons in the first layer for our pooled model.

We can see that the logistic regression is dominated by all models that allow for non-linear relationships, while the improvements for deeper models are marginal.

A.7.2 Alternative Models

We next analyze a number of machine learning techniques that are possible alternatives to our hybrid model. These algorithms have been used in other credit scoring applications, and include decision trees (CART, see [55]), random forests (RF, see [23]), neural networks (see [83]), gradient boosting (GBT, see [84]) and logistic regression. We use the out of sample loss as our main comparison metric, with lower loss values corresponding to better model performance. We tune the hyper-parameters for each model and present the results in Table 39 for our baseline 1 quarter training/validation samples. Our hybrid model performs the best, with gradient boosting coming second.

It is important to emphasize that these results do not imply that there does not exist a random forest or CART model that cannot outperform our hybrid model. The best model will depend on the specific sample. The exercise is intended to illustrate that the complexity of the model is proportional to its accuracy to a certain degree, and that deep neural networks improve substantially on shallow models, such as logistic regression.

Empirically, ensembles perform better when there is a significant diversity among the models (see [57]). Table 40 shows the SHAP values for our hybrid DNN-GBT model in comparison to GBT and DNN models for the pooled sample. The results suggest there are significant differences between the DNN and GBT. For instance, monthly payment on mortgage trades is the third most important feature for GBT, while only twenty-fifth for DNN. Even more striking perhaps is that the number of credit cards is ranked most important for DNN, while only tenth for GBT.

To see this from another angle, we grouped features into debt categories, and computed the SHAP values for each groups. The results are presented in Table 41, and once again shows significant differences between our models. For instance, features pertaining to collections contribute twice as much to our DNN model's prediction than to our GBT model's prediction. The ensemble approach can thus be thought of as providing diversification, which can reduce the variance of any of the two models. If one of the models puts too high of a weight on a feature, the other model may mitigate this effect.

Overall, the results summarized in this section suggest that deep learning is necessary to capture the complexity associated with default behavior, since all deep models perform substantially better than logistic regression. The importance of feature interaction reflects the complexity associated with default behavior. Additionally, our optimized model combines a deep neural network and gradient boosting and outperforms other machine learning models, such as random forests and decision trees, as well as deep neural networks and gradient boosting in isolation. However, all approaches show much stronger performance than logistic regression, suggesting that the main advantage is the adoption of a deep framework.

A.7.3 Expanded Training Data

We next compare the performance our our DNN to GBT by keeping our models' architecture the same, but expanding the training data. Table 42 looks at performance differences when we allow only the most recent 4 quarters for training, while Table 43 uses observations up till the date specified by the training window. These two exercises illustrate that while the performance of GBT remains similar, DNN benefits from having more data to train on.

A.8 Monotonicity Constraints

We next look at the performance trade-offs in placing monotonicity constraints on some of our features for our GBT model. Placing constraints on certain features might be required by legislation and can be desirable from a fairness standpoint (e.g., individuals who are late on their debt should have higher default risk). Monotonicity constraints are easily implementable and enforceable for GBT: one just needs to specify the set features to constrain and the corresponding relationship between the feature and the model's output (i.e., increasing or decreasing).[15]

[15] As of now, imposing monotonicity constraints are not possible for standard DNN models.[85] proposes deep lattice networks that are monotonic with respect to a user-specified set of inputs as an alternative to

We investigated constraining two sets of features. Under Regime I (R I), we placed constraints on each of our “Worst status” features. In particular, under these constraints, we require that our GBT model’s output be increasing in these features. To illustrate this, suppose we have to individuals whose credit report file only differs in one of the ”Worst status” features. Then, under Regime I, the individual whose “Worst status” feature is higher will have a higher predicted probability of default.[16] Imposing Regime I only marginally affects the model’s performance, and results in a slightly lower average loss. We conjecture that our unconstrained model implicitly learns this relationship, but slightly overfits to the training data, which results in marginally worse model performance. This is also in line with our pooled sample SHAP value results, where “Worst status” features were positively associated with default risk.

Under Regime II, we place negative constraints on our credit limit variables, length of credit and temporal proximity to delinquency variables, while positive constraints on features pertaining to “Worst status”, number of delinquencies, bankruptcies, collections, utilization and fraction of delinquent balances. Notice, however that enforcing Regime II results in a sizeable performance drop, which could be due to the non-linear relationships we illustrated in Section 1.3. This finding suggests that one must be careful in selecting features to constrain, as it might result in significant performance drops.

A.9 Comparison with Credit Scores

The credit score is a summary indicator intended to predict the risk of default by the borrower and it is widely used by the financial industry. For most unsecured debt, lenders typically verify a perspective borrower’s credit score at the time of application and sometimes a short recent sample of their credit history. For larger unsecured debts, lenders also typically require some form of income verification, as they do for secured debts, such as mortgages and auto loans. Still, the credit score is often a key determinant of crucial terms of the

standard DNNs, while [45] implements a special loss function to achieve partial monotonicity for DNNs.

[16] “Worst status” features range from current → unrated → 30 days late → 60 days late → 90 days late → 120-180 days late → derogatory, with current having the lowest value.

borrowing contract, such as the interest rate, the downpayment or the credit limit.

The most widely known credit score is the FICO score, a measure generated by the Fair Isaac Corporation, which has been in existence in its current form since 1989. Each of the three major credit reporting bureaus– Equifax, Experian and TransUnion– also have their own proprietary credit scores. Credit scoring models are not public, though they are restricted by the law, mainly the Fair Credit Reporting Act of 1970 and the Consumer Credit Reporting Reform Act of 1996. The legislation mandates that consumers be made aware of the 4 main factors that may affect their credit score. Based on available descriptive materials from FICO and the credit bureaus, these are payment history and outstanding debt, which account for more than 65% of the variation in credit scores, followed by credit history, or the age of existing accounts, which explains 15% of the variation, followed by new accounts and types of credit used (10%) and new "hard" inquiries, that is credit report inquiries coming from prospective lenders after a borrower initiated credit application.

U.S. law prohibits credit scoring models from considering a borrower's race, color, religion, national origin, sex and marital status, age, address, as well as any receipt of public assistance, or the exercise of any consumer right under the Consumer Credit Protection Act. The credit score cannot be based on information not found in a borrower's credit report, such as salary, occupation, title, employer, date employed or employment history, or interest rates being charged on particular accounts. Finally, any items in the credit report reported as child/family support obligations are not permitted, as well as "soft" inquiries[17] and any information that is not proven to be predictive of future credit performance.

[17] These include "consumer-initiated" inquiries, such as requests to view one's own credit report, "promotional inquiries", requests made by lenders in order to make pre-approved credit offers, or "administrative inquiries", requests made by lenders to review open accounts. Requests that are marked as coming from employers are also not counted.

A.10 Additional Figures and Tables

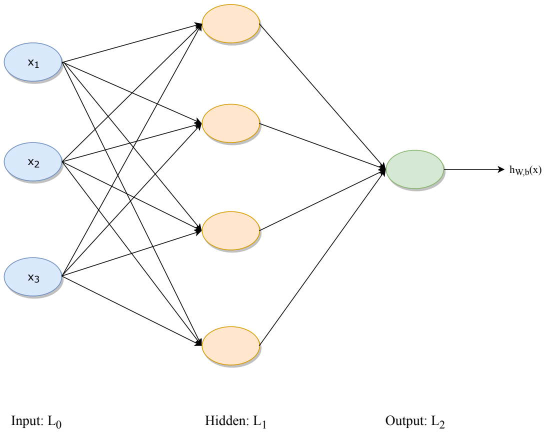
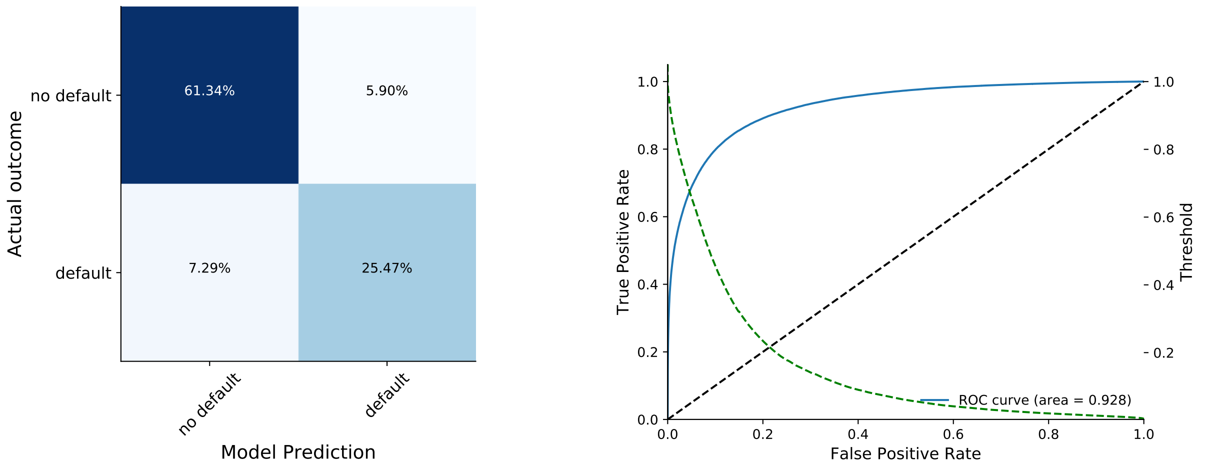


Figure 15: Two Layer Neural Network Example



(a) Confusion Matrix

(b) ROC Curve

Figure 16: Confusion Matrix and Receiver Operating Characteristic (ROC) Curve

Notes: Confusion matrix and Receiver Operating Characteristic (ROC) curve of out-of-sample forecasts of 90+ days delinquencies over the 8Q forecast horizon based on our model of default risk. In Panel (a), rows correspond to actual states, with default defined as 90+ days delinquent, no default otherwise. Classifier threshold: 50%. The numerical example is based on the model calibrated on 2011Q4 data and applied to 2013Q4 to generate out-of-sample predictions. Source: Authors' calculations based on Experian Data.

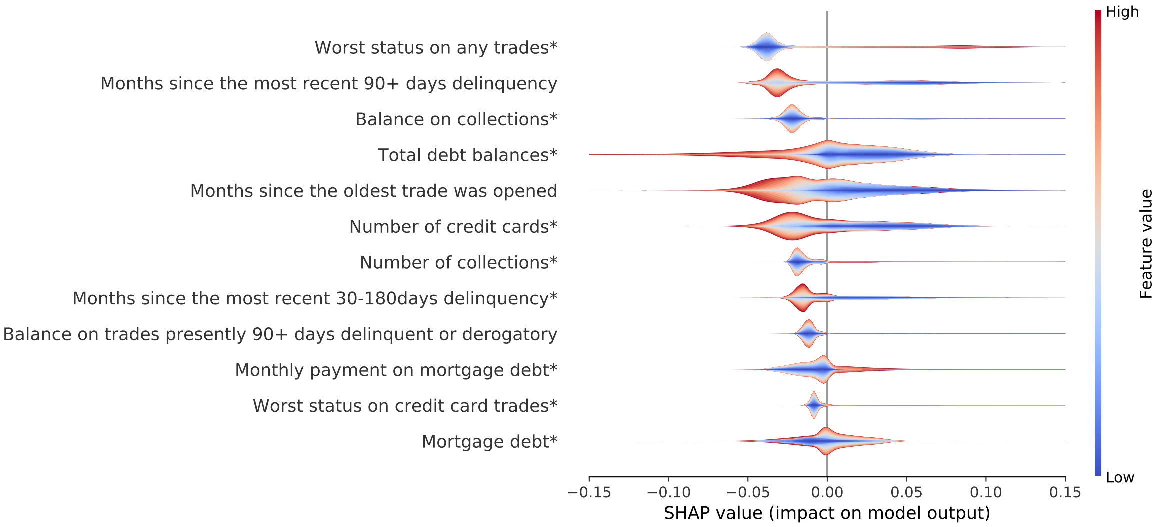


Figure 17: SHAP Applied to Model Output

Notes: SHAP applied to predicted 90+ days delinquency within 8Q. Source: Authors' calculations based on Experian Data.

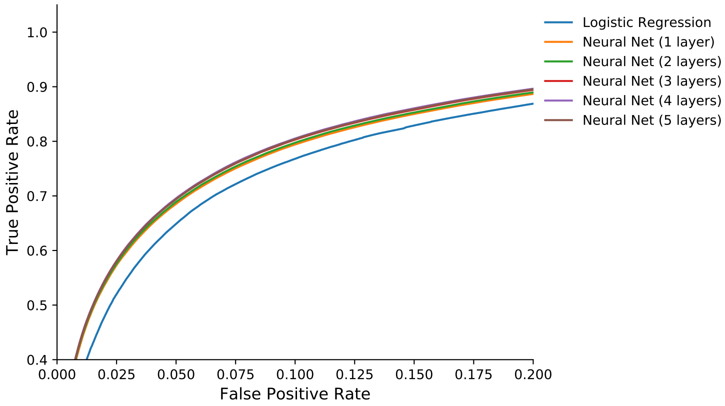


Figure 18: Out-of-Sample ROC Curves for Various Models with Dropout

Notes: Models are calibrated and evaluated on the pooled sample (2004Q1 - 2013Q4). Source: Authors' calculations based on Experian Data.

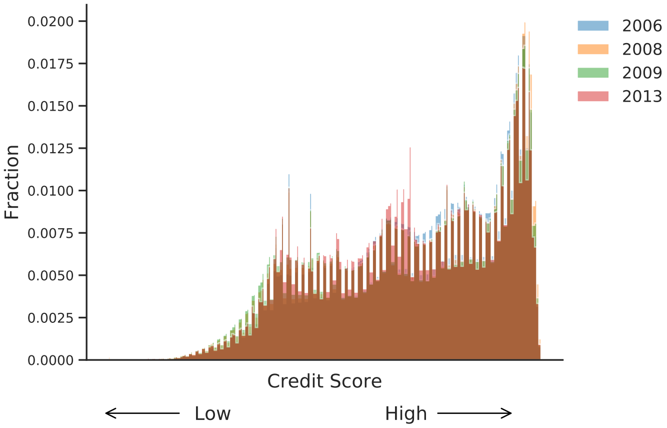
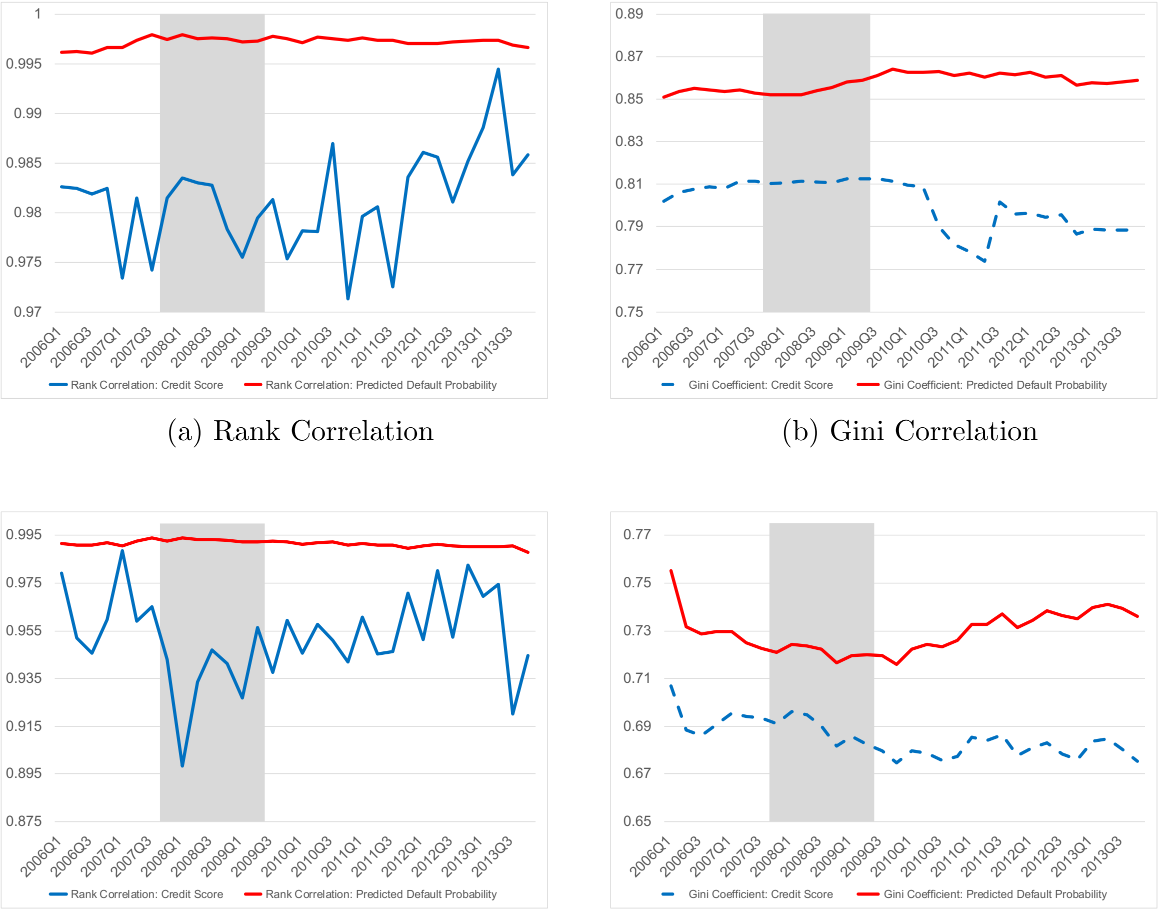


Figure 19: Credit Score Histogram by Years

Notes: Histogram of the credit score in our data by year for selected years. Source: Authors' calculations based on Experian Data.



(a) Rank Correlation

(b) Gini Correlation

(c) Rank Correlation: Current

(d) Gini Correlation: Current

Figure 20: Absolute Value of Rank Correlation with Realized Default Rate

Notes: Absolute value of rank correlation with realized default rate for the credit score and model predicted default probability for the full sample (a), for the current population (c), and Gini coefficients for the credit score and model predicted default probability by quarter for the full sample (b), and for the current population (d). Source: Authors' calculations based on Experian data.

Table 26: Itemized Sample Restrictions

	Observations
Credit Report Data	33,600,000
Remove	
Deceased	- 513,270
Age	- 4,718,804
Residence	- 953,215
Prediction Window	- 5,409,958
Prediction Sample	22,004,753

Table 27: Model Inputs

Amount past due on bankcard trades presently 30 dpd	Monthly payment on joint installment trades
Amount past due on credit card trades presently 90+ dpd	Monthly payment on joint mortgage type trades
Amount past due on installment trades presently 90+ dpd	Monthly payment on open first mortgage trades
Amount past due on joint mortgage type trades	Monthly payment on open non-deferred student trades
Amount past due on revolving trades presently 30 dpd	Monthly payment on second mortgage trades
Amount past due on revolving trades presently 90+ dpd	Months since the most recent 30-180 days delinquency on auto loan or lease trades
Amount past due on trades presently 30 dpd	Months since the most recent 30-180 days delinquency on credit card trades
Amount past due on trades presently 90+ dpd	Months since the most recent 30-180 days delinquency on trades
Balance on authorized user trades	Months since the most recent 90+ days delinquency
Balance on bankcard trades presently 90+ dpd	Months since the most recent foreclosure proceeding started on first mortgage trades
Balance on collections	Months since the most recently closed, transferred, or refinanced first mortgage trade
Balance on collections, last 24 months	Months since the most recently opened credit card trade
Balance on credit & bankcards	Months since the most recently opened first mortgage trade
Balance on home equity line of credit trades	Months since the most recently opened home equity line of credit trade
Balance on installment trades	Months since the oldest trade was opened
Balance on installment trades presently 90+ dpd	Mortgage to total debt
Balance on joint installment trades	Mortgage type inquiries made inthe last 3 months
Balance on joint revolving trades	Number of auto loan trades
Balance on open auto loan trades	Number of collections
Balance on revolving trades presently 90+ dpd	Number of credit & bankcards
Balance on second mortgage trades	Number of installment trades
Balance on trades presently 30 dpd	Number of 90 days delinquencies in the last 36 months
Balance on trades presently 60 dpd	Number of 90 days delinquencies in the last 6 months
Balance on trades presently 90+ days delinquent or derogatory	Number of open mortgage type trades
Bankcard inquiries made in the last 3 months	Open home equity line of credit trades
Credit amount on home equity line of credit trades	Public record bankruptcies
Credit amount on joint revolving trades	Public record discharged bankruptcies
Credit amount on joint trades	Public record dismissed bankruptcies
Credit amount on open credit card trades	Public records filed in the last 24 months
Credit amount on open deferred student trades	Ratio of inquiries (no deduplication) to trades opened in the last 6 months
Credit amount on open non-deferred student trades	Total debt balances
Credit amount on open trades	Trades legally paid in full for less than the full balance
Credit amount on revolving trades	Unsatisfied collections
Credit amount paid down on open first mortgage trades	Utilization ratio
Credit card utilization	Worst ever status on a credit card trade in the last 24 months
Fraction of 30 dpd debt	Worst ever status on a mortgage type trade in the last 24 months
Fraction of 60 dpd debt	Worst ever status on a trade in the last 24 months
Fraction of 90+ days delinquent debt	Worst ever status on an auto loan or lease trade in the last 24 months
Heloc utilization	Worst present status on a credit card trade
Inquiries made in the last 12 months (no deduplication)	Worst present status on a mortgage type trade
Installment utilization	Worst present status on a trade
Joint debt balances	Worst present status on a trade (excluding collections)
Monthly payment on credit card trades	Worst present status on an installment trade
Monthly payment on debt	Worst present status on an open trade

Notes: List of features included in our model.

Table 28: Summary Statistics

Feature	Mean	Std. Dev	25%	Median	75%
Open home equity line of credit trades	0.11	0.33	0	0	0
Installment utilization	0.28	0.38	0	0	0.66
Mortgage to total debt	0.3	0.42	0	0	0.82
Credit card utilization	0.32	8.19	0	0.05	0.35
Number of auto loan trades	0.34	0.6	0	0	1
Number of open mortgage type trades	0.5	0.83	0	0	1
Unsatisfied collections	0.73	2.12	0	0	0
Number of installment trades	0.74	1.3	0	0	1
Inquiries made in the last 12 months (no deduplication)	1.38	2.28	0	1	2
Number of credit & bankcards	5.73	6.19	0	4	9
Months since the most recent 90+ days delinquency	16.16	18.32	3	9	24
Months since the most recent 30-180 days delinquency on trades	22.65	23.26	2	14	39
Months since the most recent 30-180 days delinquency on credit card trades	26.45	23.91	4	20	45
Months since the most recent 30-180 days delinquency on auto loan or lease trades	28.08	24.36	5	22	48
Months since the most recently closed, transferred, or refinanced first mortgage trade	39.51	30.85	14	32	60
Worst ever status on a credit card trade in the last 24 months	45.26	116.61	1	1	2
Months since the most recently opened home equity line of credit trade	65.19	48.74	28	57	90
Worst present status on a trade (excluding collections)	68.34	145.61	1	1	2
Months since the most recently opened first mortgage trade	70.86	69.13	23	51	95
Monthly payment on credit card trades	121.22	366.75	0	31	125
Worst ever status on a trade in the last 24 months	126.6	180.08	1	1	400
Months since the oldest trade was opened	196.45	126.35	98	178	271
Monthly payment on open first mortgage trades	474.67	2163.29	0	0	694
Balance on collections, placed with the collector in the last 24 months	560.26	3277.47	0	0	0
Monthly payment on debt	907.95	11059.39	20	333	1232
Credit amount on open non-deferred student trades	2123.32	11983.58	0	0	0
Balance on trades presently 90+ days delinquent or derogatory	3125.34	33186.26	0	0	0
Balance on joint installment trades	4142.13	26920.71	0	0	0
Balance on open auto loan trades	4472.88	11608.48	0	0	3915
Credit amount paid down on open first mortgage trades	6109.98	164527.99	0	0	2255
Balance on installment trades	8754.12	32554.07	0	0	10583
Balance on credit & bankcards	8808.63	19284.41	0	1526	8624
Credit amount on home equity line of credit trades	9139.8	47969.14	0	0	0
Credit amount on joint revolving trades	10698.74	44129.16	0	0	2200
Credit amount on open credit card trades	21475.9	30662.29	0	8600	31947
Credit amount on revolving trades	30517.5	62226.97	0	9500	37311
Joint debt balances	50957.9	143808.27	0	0	29546
Credit amount on joint trades	64191.25	220910.44	0	0	55100
Total debt balances	77126.36	170742.97	318	11738	95808
Credit amount on open trades	108480.31	259094	3000	33146	146535

Table 29: Feature Groups

Total debt balances*
Joint debt balances
Credit amount on joint trades
Total debt balances
Credit amount on open trades

Balance on revolving debt*
Credit amount on home equity line of credit trades
Balance on home equity line of credit trades
Balance on joint revolving trades
Credit amount on joint revolving trades
Credit amount on revolving trades

Balance on auto loans*
Balance on open auto loan trades
Number of auto loan trades

Balance on collections*
Balance on collections
Balance on collections, placed in the last 24 months

Balance on installment loans*
Balance on installment trades
Balance on joint installment trades

Number of bankruptcies*
Public record bankruptcies
Public record discharged bankruptcies

Number of collections*
Number of collections
Unsatisfied collections

Number of credit cards*
Credit amount on open credit card trades
Number of credit & bankcards

Monthly payment on mortgage debt*
Monthly payment on joint mortgage type trades
Monthly payment on first mortgage trades

Monthly payment on debt*
Monthly payment on debt
Monthly payment on joint installment trades

Balance on 90-180 days late installment loans*
Amount past due on installment trades presently 90+ dpd
Balance on installment trades presently 90+ dpd

Fraction of 90+ days delinquent debt*
Fraction of 90+ days delinquent debt
Worst present status on an open trade

90 days late credit card debt*
Amount past due on credit card trades presently 90+ dpd
Balance on bankcard revolving and charge trades presently 90+ dpd

Months since the most recent 30-180days delinquency*
Months since the most recent 30-180 days delinquency on credit card trades
Months since the most recent 30-180 days delinquency

Worst status on credit card trades*
Worst present status on a trade (excluding collections)
Worst present status on a credit card trade

Worst status on any trades*
Worst present status on a trade
Worst ever status on a trade in the last 24 months

Mortgage debt*
Mortgage to total debt
Number of open mortgage type trades
Months since the most recently opened first mortgage trade
Months since the most recently closed, transferred or refinanced first mortgage trade

Table 30: Hyperparameters for Machine Learning Models: Out-of-sample Exercise

Model	Tree Depth	# of Trees
CART	7	
RF	20	800
GBT	5	1000

Table 31: Weighting Schemes and Loss

Weight on DNN	Out-of-Sample Loss	Pooled Loss
0.2	0.3230	0.2778
0.3	0.3230	0.2745
0.1	0.3231	0.2823
0.4	0.3232	0.2721
0	0.3236	0.2890
0.5	0.3237	0.2705
0.6	0.3244	0.2695
0.7	0.3252	0.2693
0.8	0.3263	0.2700
0.9	0.3277	0.2717

Notes: Performance comparison of our hybrid DNN-GBT model under different weighting schemes. The results of predicted probabilities versus actual outcomes over the following 8Q (testing period) are used to calculate the loss metric for 90+ days delinquencies within 8Q. DNN refers to deep neural network, Source: Authors' calculations based on Experian Data.

Table 32: 1 Quarter Ahead Predictions, Full Sample– Hybrid DNN-GBT

Training Window	Testing Window	Data	Predicted	Delinquents	Non-Delinquents
2004Q1-2013Q4	2004Q1-2013Q4	0.3396	0.3359	0.7475	0.1242
2004Q1	2006Q1	0.3248	0.2948	0.6543	0.1218
2004Q2	2006Q2	0.3274	0.3057	0.6748	0.1260
2004Q3	2006Q3	0.3306	0.3126	0.6851	0.1286
2004Q4	2006Q4	0.3347	0.3153	0.6850	0.1293
2005Q1	2007Q1	0.3410	0.3185	0.6867	0.1279
2005Q2	2007Q2	0.3444	0.3224	0.6897	0.1295
2005Q3	2007Q3	0.3469	0.3224	0.6872	0.1287
2005Q4	2007Q4	0.3505	0.3306	0.6975	0.1327
2006Q1	2008Q1	0.3535	0.3390	0.7093	0.1366
2006Q2	2008Q2	0.3545	0.3364	0.7022	0.1355
2006Q3	2008Q3	0.3558	0.3369	0.7046	0.1338
2006Q4	2008Q4	0.3587	0.3434	0.7109	0.1379
2007Q1	2009Q1	0.3588	0.3504	0.7221	0.1425
2007Q2	2009Q2	0.3580	0.3528	0.7250	0.1452
2007Q3	2009Q3	0.3573	0.3550	0.7269	0.1482
2007Q4	2009Q4	0.3589	0.3571	0.7286	0.1492
2008Q1	2010Q1	0.3589	0.3606	0.7319	0.1527
2008Q2	2010Q2	0.3568	0.3633	0.7352	0.1570
2008Q3	2010Q3	0.3559	0.3632	0.7336	0.1586
2008Q4	2010Q4	0.3544	0.3636	0.7312	0.1618
2009Q1	2011Q1	0.3541	0.3614	0.7296	0.1595
2009Q2	2011Q2	0.3511	0.3567	0.7221	0.1590
2009Q3	2011Q3	0.3500	0.3557	0.7214	0.1588
2009Q4	2011Q4	0.3484	0.3536	0.7221	0.1565
2010Q1	2012Q1	0.3467	0.3567	0.7300	0.1585
2010Q2	2012Q2	0.3434	0.3518	0.7257	0.1563
2010Q3	2012Q3	0.3396	0.3517	0.7307	0.1568
2010Q4	2012Q4	0.3358	0.3484	0.7285	0.1562
2011Q1	2013Q1	0.3341	0.3479	0.7318	0.1553
2011Q2	2013Q2	0.3317	0.3436	0.7266	0.1536
2011Q3	2013Q3	0.3298	0.3426	0.7286	0.1527
2011Q4	2013Q4	0.3275	0.3402	0.7289	0.1509

Notes: Performance metrics for our model of default risk over 32+1 testing windows. For each testing window, the model is calibrated on data over the period specified in the training window, and predictions are based on the data available as of the data in the training window. For example, the fourth row reports the performance of the model calibrated using input data available in 2004Q3, and applied to 2006Q3 data to generate forecasts of delinquencies for within 8 quarter delinquencies. Average model forecasts over all customers, and customers that (ex-post) did and did not become 90+ days delinquent over the testing window are also reported. Source: Authors' calculations based on Experian Data.

Table 33: 1 Quarter Ahead Predictions, Current– Hybrid DNN-GBT

Training Window	Testing Window	Data	Predicted	Delinquent	Non-delinquent
2004Q1-2013Q4	2004Q1-2013Q4	0.1676	0.1629	0.5304	0.0889
2004Q1	2006Q1	0.1844	0.1568	0.4292	0.0952
2004Q2	2006Q2	0.1702	0.1475	0.3972	0.0963
2004Q3	2006Q3	0.1695	0.1499	0.4007	0.0987
2004Q4	2006Q4	0.1727	0.1506	0.4010	0.0983
2005Q1	2007Q1	0.1805	0.1542	0.4047	0.0990
2005Q2	2007Q2	0.1813	0.1545	0.3994	0.1003
2005Q3	2007Q3	0.1831	0.1527	0.3929	0.0988
2005Q4	2007Q4	0.1847	0.1566	0.4032	0.1007
2006Q1	2008Q1	0.1890	0.1650	0.4195	0.1057
2006Q2	2008Q2	0.1896	0.1626	0.4098	0.1048
2006Q3	2008Q3	0.1872	0.1593	0.4043	0.1028
2006Q4	2008Q4	0.1817	0.1595	0.4037	0.1053
2007Q1	2009Q1	0.1781	0.1650	0.4205	0.1097
2007Q2	2009Q2	0.1752	0.1668	0.4240	0.1122
2007Q3	2009Q3	0.1713	0.1689	0.4302	0.1149
2007Q4	2009Q4	0.1661	0.1669	0.4230	0.1160
2008Q1	2010Q1	0.1683	0.1722	0.4372	0.1186
2008Q2	2010Q2	0.1668	0.1778	0.4508	0.1231
2008Q3	2010Q3	0.1661	0.1795	0.4559	0.1244
2008Q4	2010Q4	0.1644	0.1787	0.4509	0.1252
2009Q1	2011Q1	0.1674	0.1812	0.4616	0.1248
2009Q2	2011Q2	0.1668	0.1768	0.4514	0.1218
2009Q3	2011Q3	0.1669	0.1769	0.4520	0.1218
2009Q4	2011Q4	0.1597	0.1699	0.4380	0.1189
2010Q1	2012Q1	0.1604	0.1724	0.4468	0.1200
2010Q2	2012Q2	0.1622	0.1705	0.4477	0.1168
2010Q3	2012Q3	0.1598	0.1676	0.4434	0.1152
2010Q4	2012Q4	0.1575	0.1668	0.4432	0.1152
2011Q1	2013Q1	0.1606	0.1710	0.4576	0.1162
2011Q2	2013Q2	0.1603	0.1692	0.4541	0.1149
2011Q3	2013Q3	0.1578	0.1660	0.4496	0.1128
2011Q4	2013Q4	0.1548	0.1623	0.4430	0.1109

Notes: Performance metrics for our model of default risk over 32+1 testing windows for customers who are current as of the forecast date but become 90+ days delinquent in the following 8 quarters. For each testing window, the model is calibrated on data over the period specified in the training window columns, and predictions are based on the data available as of the data in the training window. For example, the fourth row reports the performance of the model calibrated using input data available in 2004Q3, and applied to 2006Q3 data to generate forecasts of delinquencies for within 8 quarter delinquencies. Average model forecasts over all current customers, and all current customers that did and did not become 90+ days delinquent over the testing window are also reported. Source: Authors' calculations based on Experian Data.

Table 34: Performance Metrics using Hybrid DNN-GBT, Full Sample

Training Window	Testing Window	AUC score	Precision	Recall	F-measure	Accuracy	Loss
2004Q1-2013Q4	2004Q1-2013Q4	0.9494	0.8546	0.8104	0.8319	0.8888	0.2693
2004Q1	2006Q1	0.9243	0.8508	0.7056	0.7714	0.8642	0.3230
2004Q2	2006Q2	0.9250	0.8486	0.7170	0.7773	0.8654	0.3187
2004Q3	2006Q3	0.9259	0.8492	0.7261	0.7828	0.8668	0.3165
2004Q4	2006Q4	0.9250	0.8517	0.7232	0.7822	0.8652	0.3200
2005Q1	2007Q1	0.9257	0.8561	0.7235	0.7842	0.8642	0.3208
2005Q2	2007Q2	0.9257	0.8571	0.7267	0.7865	0.8641	0.3217
2005Q3	2007Q3	0.9251	0.8592	0.7224	0.7849	0.8626	0.3247
2005Q4	2007Q4	0.9238	0.8541	0.7290	0.7866	0.8614	0.3278
2006Q1	2008Q1	0.9246	0.8505	0.7390	0.7908	0.8618	0.3265
2006Q2	2008Q2	0.9245	0.8542	0.7319	0.7883	0.8607	0.3275
2006Q3	2008Q3	0.9255	0.8556	0.7342	0.7902	0.8613	0.3259
2006Q4	2008Q4	0.9257	0.8529	0.7402	0.7926	0.8610	0.3259
2007Q1	2009Q1	0.9277	0.8487	0.7540	0.7986	0.8635	0.3210
2007Q2	2009Q2	0.9279	0.8441	0.7614	0.8006	0.8642	0.3198
2007Q3	2009Q3	0.9286	0.8422	0.7673	0.8030	0.8655	0.3177
2007Q4	2009Q4	0.9300	0.8454	0.7723	0.8072	0.8676	0.3143
2008Q1	2010Q1	0.9299	0.8415	0.7787	0.8089	0.8679	0.3153
2008Q2	2010Q2	0.9296	0.8334	0.7858	0.8089	0.8675	0.3159
2008Q3	2010Q3	0.9292	0.8323	0.7864	0.8087	0.8676	0.3161
2008Q4	2010Q4	0.9290	0.8297	0.7850	0.8067	0.8667	0.3171
2009Q1	2011Q1	0.9296	0.8341	0.7842	0.8084	0.8683	0.3154
2009Q2	2011Q2	0.9282	0.8341	0.7748	0.8033	0.8668	0.3179
2009Q3	2011Q3	0.9284	0.8379	0.7721	0.8036	0.8679	0.3168
2009Q4	2011Q4	0.9288	0.8378	0.7702	0.8026	0.8680	0.3154
2010Q1	2012Q1	0.9293	0.8323	0.7770	0.8037	0.8684	0.3142
2010Q2	2012Q2	0.9280	0.8290	0.7731	0.8001	0.8673	0.3162
2010Q3	2012Q3	0.9277	0.8248	0.7746	0.7989	0.8676	0.3151
2010Q4	2012Q4	0.9271	0.8172	0.7769	0.7965	0.8667	0.3167
2011Q1	2013Q1	0.9280	0.8160	0.7790	0.7971	0.8675	0.3141
2011Q2	2013Q2	0.9276	0.8158	0.7754	0.7951	0.8674	0.3139
2011Q3	2013Q3	0.9281	0.8127	0.7792	0.7956	0.8680	0.3123
2011Q4	2013Q4	0.9284	0.8118	0.7776	0.7943	0.8681	0.3104

Notes: Performance metrics for our model of default risk. The model calibrations are specified by the training and testing windows. The results of classifications versus actual outcomes over the following 8Q are used to calculate these performance metrics for 90+ days delinquencies within 8Q. Source: Authors' calculations based on Experian Data.

Table 35: Performance Metrics using Hybrid DNN-GBT, Current

Training Window	Testing Window	AUC score	Precision	Recall	F-measure	Accuracy	Loss
2004Q1-2013Q4	2004Q1-2013Q4	0.9225	0.781	0.5632	0.6545	0.9003	0.2458
2004Q1	2006Q1	0.8773	0.7686	0.4059	0.5313	0.8679	0.3165
2004Q2	2006Q2	0.8653	0.7298	0.3569	0.4794	0.8681	0.3159
2004Q3	2006Q3	0.8638	0.7227	0.3606	0.4811	0.8681	0.3162
2004Q4	2006Q4	0.8642	0.7307	0.3609	0.4832	0.8666	0.3196
2005Q1	2007Q1	0.8649	0.7384	0.3658	0.4892	0.8621	0.3275
2005Q2	2007Q2	0.8621	0.7302	0.3591	0.4814	0.8597	0.3318
2005Q3	2007Q3	0.8616	0.7382	0.3454	0.4706	0.8577	0.3353
2005Q4	2007Q4	0.8599	0.7261	0.3594	0.4808	0.8566	0.3382
2006Q1	2008Q1	0.8612	0.7192	0.3818	0.4988	0.8550	0.3402
2006Q2	2008Q2	0.8615	0.7236	0.3656	0.4858	0.8532	0.3414
2006Q3	2008Q3	0.8609	0.7219	0.3590	0.4795	0.8541	0.3403
2006Q4	2008Q4	0.8578	0.7061	0.3552	0.4727	0.8560	0.3366
2007Q1	2009Q1	0.8599	0.6993	0.3818	0.4939	0.8607	0.3285
2007Q2	2009Q2	0.8596	0.6870	0.3903	0.4978	0.8621	0.3251
2007Q3	2009Q3	0.8594	0.6802	0.3990	0.5030	0.8649	0.3207
2007Q4	2009Q4	0.8573	0.6747	0.3898	0.4941	0.8674	0.3170
2008Q1	2010Q1	0.8607	0.6822	0.4156	0.5165	0.8690	0.3152
2008Q2	2010Q2	0.8617	0.6645	0.4385	0.5284	0.8694	0.3136
2008Q3	2010Q3	0.8619	0.6608	0.4480	0.5340	0.8701	0.3128
2008Q4	2010Q4	0.8632	0.6653	0.4388	0.5288	0.8715	0.3098
2009Q1	2011Q1	0.8664	0.6769	0.4524	0.5423	0.8721	0.3092
2009Q2	2011Q2	0.8666	0.6833	0.4386	0.5342	0.8724	0.3085
2009Q3	2011Q3	0.8678	0.6891	0.4373	0.5351	0.8732	0.3076
2009Q4	2011Q4	0.8651	0.6866	0.4123	0.5152	0.8760	0.3029
2010Q1	2012Q1	0.8664	0.6781	0.4255	0.5229	0.8754	0.3029
2010Q2	2012Q2	0.8684	0.6851	0.4265	0.5257	0.8752	0.3025
2010Q3	2012Q3	0.8673	0.6854	0.4178	0.5192	0.8763	0.3012
2010Q4	2012Q4	0.8664	0.6768	0.4195	0.5180	0.8770	0.2998
2011Q1	2013Q1	0.8691	0.6744	0.4420	0.5340	0.8761	0.3005
2011Q2	2013Q2	0.8694	0.6781	0.4372	0.5317	0.8765	0.2997
2011Q3	2013Q3	0.8690	0.6766	0.4336	0.5285	0.8779	0.2975
2011Q4	2013Q4	0.8671	0.6706	0.4270	0.5218	0.8789	0.2958

Notes: Performance metrics for our model of default risk for the current population. Borrowers who are current do not have any delinquencies. The model calibrations are specified by the training and testing windows. The results of classifications versus actual outcomes over the following 8Q are used to calculate these performance metrics for 90+ days delinquencies within 8Q. Source: Authors' calculations based on Experian Data.

Table 36: Explanatory Power of Variables

Feature	Accuracy	Loss
Credit amount on open trades	0.8587	0.3333
Total debt balances	0.8627	0.3245
Number of credit & bankcards	0.8677	0.3156
Credit amount on open credit card trades	0.8714	0.3064
Monthly payment on open trades	0.8751	0.3039
Credit amount on open revolving trades	0.8751	0.3036
Months since the most recent 90 or more days delinquency	0.8757	0.3034
Months since the oldest trade was opened	0.8722	0.3026
Balance on credit & bankcards	0.8751	0.2974
Number of open mortgage type trades	0.8758	0.2953
Monthly payment on credit card trades	0.8777	0.2946
Credit amount on open joint revolving trades	0.8811	0.2930
Monthly payment on open first mortgage trades	0.8817	0.2876
Number of installment trades	0.8790	0.2874
Months since the most recent 30-180days delinquency on trades	0.8806	0.2861
Worst ever status on a trade in the last 24 months	0.8797	0.2859
Balance on installment trades	0.8802	0.2859
Joint debt balances	0.8818	0.2856
Installment utilization	0.8804	0.2853
Mortgage to total debt	0.8817	0.2846
Number of open auto loan trades	0.8803	0.2843
Months since the most recently closed, transferred, or refinanced first mortgage trade	0.8809	0.2835
Months since the most recently opened home equity line of credit trade	0.8816	0.2835
Months since the most recent 30-180days delinquency on credit card trades	0.8800	0.2835
Balance on open auto loan trades	0.8812	0.2826
Inquiries made in the last 12 months	0.8818	0.2826
Months since the most recently opened first mortgage trade	0.8819	0.2822
Unsatisfied collections	0.8832	0.2814
⋮	⋮	⋮
Mortgage type inquiries made in the last 3 months	0.8879	0.2694
Bankcard revolving and charge inquiries made in the last 3 months	0.8881	0.2693
Baseline	0.8881	0.2691

Notes: This table reports a perturbation analysis on the pooled sample using our hybrid model. For each variable, we re-shuffle the feature, keeping the distribution intact in the test dataset and the model's loss function is evaluated on the test dataset with the changed covariate. We repeat this step 10 times, and report the average of the loss and accuracy. Then, the variable is replaced to its original values, and a perturbation test is performed on a new variable. Perturbing the variable of course reduces the accuracy of the model, and the test loss becomes larger. If a particular variable has strong explanatory power, the test loss will significantly increase. The test loss for the complete model when no variables are perturbed is the Baseline value.

Table 37: SHAP Values over Time

	Prediction Date					
	2006Q1		2008Q1		2011Q1	
	Model					
	2004Q1	2006Q1	2006Q1	2008Q1	2009Q1	2011Q1
Features						
Ratio of inquiries to trades opened in the last 6 months	0.035 (1)	0.029 (3)	0.026 (3)	0.027 (4)	0.029 (5)	0.023 (4)
Number of credit cards*	0.032 (2)	0.014 (7)	0.013 (7)	0.007 (17)	0.007 (19)	0.013 (14)
Total debt balances*	0.03 (3)	0.022 (5)	0.021 (5)	0.014 (8)	0.016 (10)	0.027 (3)
Balance on revolving debt*	0.029 (4)	0.041 (2)	0.034 (2)	0.031 (3)	0.032 (4)	0.013 (15)
Balance on auto loans*	0.028 (5)	0.02 (6)	0.018 (6)	0.015 (7)	0.028 (6)	0.015 (8)
Installment utilization	0.027 (6)	0.025 (4)	0.024 (4)	0.044 (2)	0.049 (2)	0.055 (1)
Worst status on any trades*	0.02 (7)	0.008 (10)	0.007 (8)	0.017 (5)	0.011 (14)	0.015 (9)
Months since the most recently opened home equity line of credit trade	0.018 (8)	0.004 (24)	0.004 (19)	0.003 (33)	0.003 (37)	0.005 (20)
Months since the most recent 90+ days delinquency	0.015 (9)	0.052 (1)	0.043 (1)	0.044 (1)	0.07 (1)	0.049 (2)
Monthly payment on credit card trades	0.014 (10)	0.002 (35)	0.002 (33)	0.002 (39)	0.004 (29)	0.01 (18)
Credit card utilization	0.014 (11)	0.008 (8)	0.007 (9)	0.01 (11)	0.019 (8)	0.021 (5)
Mortgage debt*	0.012 (12)	0.005 (16)	0.004 (20)	0.005 (21)	0.015 (12)	0.017 (7)
Heloc utilization	0.011 (13)	0.0 (64)	0.0 (64)	0.0 (64)	0.0 (64)	0.0 (64)
Number of bankcard revolving and charge inquiries	0.009 (14)	0.002 (36)	0.002 (38)	0.016 (6)	0.01 (15)	0.004 (23)
Worst status on credit card trades*	0.009 (15)	0.003 (28)	0.003 (31)	0.002 (37)	0.004 (27)	0.007 (19)
Balance on credit & bankcards	0.009 (16)	0.004 (19)	0.005 (15)	0.013 (10)	0.006 (24)	0.018 (6)
Number of open installment trades	0.005 (17)	0.008 (9)	0.007 (10)	0.006 (19)	0.004 (30)	0.005 (21)
Number of mortgage type inquiries made in the last 3 months	0.005 (18)	0.002 (41)	0.002 (37)	0.001 (43)	0.002 (41)	0.0 (56)
Balance on installment loans*	0.005 (19)	0.005 (15)	0.006 (11)	0.009 (14)	0.041 (3)	0.012 (16)
Balance on trades presently 90+ days delinquent	0.005 (20)	0.002 (33)	0.002 (36)	0.004 (25)	0.003 (32)	0.001 (43)

Notes: This table reports the Shapley values for a selected 20 features for three out-of-sample models. For each prediction window, we compute the Shapley value for each of the observations and for each feature. We then calculate the average of the absolute value for each feature, and report the results for the selected features. Finally, we rank the results based on the feature's relative rank in the given prediction window in parentheses. Source: Authors' calculations based on Experian data.

Table 38: Neural Networks Comparison: Loss & Accuracy

Model	In-sample Loss		Out-of-sample Loss	
	w/o Dropout	Dropout	w/o Dropout	Dropout
Logistic Regression	0.3454	0.3454	0.3452	0.3452
1 layer	0.3167	0.3182	0.3179	0.3187
2 layers	0.3079	0.3139	0.3163	0.3160
3 layers	0.3018	0.3057	0.3149	0.3117
4 layers	0.2978	0.3017	0.3140	0.3101
5 layers	0.2955	0.3038	0.3138	0.3108

Model	In-sample Accuracy		Out-of-sample Accuracy	
	w/o Dropout	Dropout	w/o Dropout	Dropout
Logistic Regression	0.8564	0.8564	0.8564	0.8564
1 layer	0.8666	0.8660	0.8661	0.8658
2 layers	0.8707	0.8679	0.8670	0.8669
3 layers	0.8737	0.8718	0.8680	0.8689
4 layers	0.8756	0.8733	0.8685	0.8695
5 layers	0.8767	0.8726	0.8688	0.8691

Notes: In-sample and out-of-sample loss (categorical cross-entropy) and accuracy for neural networks of different depth and for logistic regression. Models are calibrated and evaluated on the pooled sample (2004Q1 - 2013Q4). Source: Authors' calculations based on Experian Data.

Table 39: Model Comparison: Out-of-Sample Loss

Training Window	Testing Window	Combined	GBT	RF	DNN	CART	Logistic
2004Q1	2006Q1	0.3230	0.3236	0.3266	0.3294	0.3416	0.3486
2004Q2	2006Q2	0.3187	0.3189	0.3224	0.3276	0.3372	0.3469
2004Q3	2006Q3	0.3165	0.3167	0.3206	0.3258	0.3349	0.3463
2004Q4	2006Q4	0.3200	0.3203	0.3240	0.3278	0.3396	0.3475
2005Q1	2007Q1	0.3208	0.3212	0.3253	0.3281	0.3426	0.3497
2005Q2	2007Q2	0.3217	0.3217	0.3261	0.3302	0.3424	0.3514
2005Q3	2007Q3	0.3247	0.3248	0.3292	0.3326	0.3454	0.3551
2005Q4	2007Q4	0.3278	0.3283	0.3323	0.3339	0.3490	0.3576
2006Q1	2008Q1	0.3265	0.3269	0.3312	0.3339	0.3476	0.3569
2006Q2	2008Q2	0.3275	0.3280	0.3317	0.3338	0.3473	0.3587
2006Q3	2008Q3	0.3259	0.3262	0.3298	0.3339	0.3453	0.3563
2006Q4	2008Q4	0.3259	0.3264	0.3291	0.3331	0.3457	0.3552
2007Q1	2009Q1	0.3210	0.3219	0.3244	0.3289	0.3407	0.3513
2007Q2	2009Q2	0.3198	0.3204	0.3238	0.3279	0.3397	0.3501
2007Q3	2009Q3	0.3177	0.3179	0.3211	0.3262	0.3386	0.3481
2007Q4	2009Q4	0.3143	0.3145	0.3180	0.3232	0.3352	0.3442
2008Q1	2010Q1	0.3153	0.3157	0.3190	0.3248	0.3347	0.3461
2008Q2	2010Q2	0.3159	0.3161	0.3197	0.3253	0.3366	0.3486
2008Q3	2010Q3	0.3161	0.3164	0.3195	0.3254	0.3356	0.3493
2008Q4	2010Q4	0.3171	0.3177	0.3208	0.3254	0.3365	0.3466
2009Q1	2011Q1	0.3154	0.3154	0.3181	0.3257	0.3323	0.3459
2009Q2	2011Q2	0.3179	0.3182	0.3209	0.3266	0.3350	0.3470
2009Q3	2011Q3	0.3168	0.3170	0.3191	0.3263	0.3347	0.3455
2009Q4	2011Q4	0.3154	0.3158	0.3182	0.3242	0.3316	0.3440
2010Q1	2012Q1	0.3142	0.3143	0.3172	0.3239	0.3312	0.3427
2010Q2	2012Q2	0.3162	0.3164	0.3195	0.3255	0.3335	0.3441
2010Q3	2012Q3	0.3151	0.3152	0.3186	0.3248	0.3321	0.3440
2010Q4	2012Q4	0.3167	0.3167	0.3204	0.3258	0.3353	0.3446
2011Q1	2013Q1	0.3141	0.3143	0.3172	0.3236	0.3324	0.3424
2011Q2	2013Q2	0.3139	0.3140	0.3174	0.3231	0.3322	0.3415
2011Q3	2013Q3	0.3123	0.3123	0.3159	0.3219	0.3301	0.3403
2011Q4	2013Q4	0.3104	0.3104	0.3143	0.3207	0.3274	0.3393
Average		0.3186	0.3189	0.3222	0.3272	0.3376	0.3480

Notes: Performance comparison of machine learning classification models of consumer default risk. The model calibrations are specified by the training and testing windows. The results of predicted probabilities versus actual outcomes over the following 8Q testing period are used to calculate the loss metric for 90+ days delinquencies within 8Q. Combined refers to the hybrid DNN-GBT model, DNN refers to deep neural network, RF refers to random forest, GBT refers to gradient boosted trees, while CART refers to decision tree. Source: Authors' calculations based on Experian Data.

Table 40: Shap Values across Models

Feature	Hybrid	Logistic	DNN	GBT
Worst status on any trades*	0.048 (1)	0.051 (2)	0.031 (3)	0.089 (2)
Months since the most recent 90+ days delinquency	0.037 (2)	0.048 (3)	0.015 (8)	0.091 (1)
Total debt balances*	0.031 (3)	0.028 (4)	0.023 (5)	0.062 (5)
Balance on collections*	0.031 (4)	0.002 (27)	0.017 (7)	0.065 (4)
Months since the oldest trade was opened	0.029 (5)	0.002 (30)	0.025 (4)	0.047 (8)
Number of credit cards*	0.028 (6)	0.027 (5)	0.037 (1)	0.033 (10)
Number of collections*	0.025 (7)	0.059 (1)	0.037 (2)	0.008 (30)
Balance on trades presently 90+ days delinquent	0.021 (8)	0.002 (29)	0.022 (6)	0.018 (19)
Months since the most recent 30-180days delinquency*	0.021 (9)	0.025 (6)	0.011 (11)	0.047 (7)
Monthly payment on mortgage debt*	0.017 (10)	0.0 (49)	0.005 (25)	0.066 (3)

Notes: This table reports the Shapley values for four selected machine learning classification models of consumer default risk. We sorted the features based on the feature's relative rank (in parentheses) using the hybrid model. Source: Authors' calculations based on Experian Data.

Table 41: Shap Values across Debt Categories

	Model			
	Hybrid	Logistic	DNN	GBT
Feature Group				
Total Debt	0.4896	0.6241	0.4791	0.4607
Revolving Debt	0.2228	0.1816	0.2348	0.2374
Installment Debt	0.1895	0.1030	0.1594	0.2364
Collections	0.0981	0.0914	0.1266	0.0655

Notes: This table reports the aggregate absolute Shapley values for four selected machine learning classification models of consumer default risk. We grouped our features into debt categories, and computed the sum of the absolute SHAP values. Installment debt includes auto loans, mortgage debt, and student debt; while revolving debt includes credit and bankcard debt, and home equity line of credit trades. For ease of interpretability, we normalized our feature groups to 1 for each of our models. Source: Authors' calculations based on Experian Data.

Table 42: Model Comparison: DNN vs. GBT, 1 Year

Training Window Start	Training Window End	Testing Window	AUC-score		Loss	
			DNN	GBT	DNN	GBT
2004Q1	2004Q1	2006Q1	0.9219	0.9239	0.3294	0.3236
2004Q1	2004Q2	2006Q2	0.9222	0.9249	0.3279	0.3193
2004Q1	2004Q3	2006Q3	0.9236	0.9262	0.3245	0.3165
2004Q1	2004Q4	2006Q4	0.9232	0.9252	0.3260	0.3198
2004Q2	2005Q1	2007Q1	0.9234	0.9257	0.3289	0.3208
2004Q3	2005Q2	2007Q2	0.9237	0.9256	0.3281	0.3222
2004Q4	2005Q3	2007Q3	0.9224	0.9250	0.3334	0.3247
2005Q1	2005Q4	2007Q4	0.9220	0.9240	0.3342	0.3276
2005Q2	2006Q1	2008Q1	0.9221	0.9246	0.3364	0.3276
2005Q3	2006Q2	2008Q2	0.9223	0.9240	0.3349	0.3291
2005Q4	2006Q3	2008Q3	0.9232	0.9250	0.3338	0.3269
2006Q1	2006Q4	2008Q4	0.9236	0.9254	0.3319	0.3266
2006Q2	2007Q1	2009Q1	0.9248	0.9272	0.3290	0.3224
2006Q3	2007Q2	2009Q2	0.9254	0.9274	0.3268	0.3214
2006Q4	2007Q3	2009Q3	0.9263	0.9285	0.3246	0.3178
2007Q1	2007Q4	2009Q4	0.9279	0.9301	0.3207	0.3138
2007Q2	2008Q1	2010Q1	0.9274	0.9299	0.3223	0.3150
2007Q3	2008Q2	2010Q2	0.9267	0.9297	0.3236	0.3149
2007Q4	2008Q3	2010Q3	0.9266	0.9294	0.3234	0.3149
2008Q1	2008Q4	2010Q4	0.9267	0.9291	0.3238	0.3165
2008Q2	2009Q1	2011Q1	0.9271	0.9299	0.3221	0.3142
2008Q3	2009Q2	2011Q2	0.9254	0.9283	0.3257	0.3172
2008Q4	2009Q3	2011Q3	0.9254	0.9282	0.3256	0.3168
2009Q1	2009Q4	2011Q4	0.9257	0.9285	0.3242	0.3159
2009Q2	2010Q1	2012Q1	0.9261	0.9292	0.3222	0.3137
2009Q3	2010Q2	2012Q2	0.9250	0.9278	0.3244	0.3160
2009Q4	2010Q3	2012Q3	0.9249	0.9277	0.3230	0.3146
2010Q1	2010Q4	2012Q4	0.9245	0.9273	0.3243	0.3158
2010Q2	2011Q1	2013Q1	0.9254	0.9280	0.3216	0.3134
2010Q3	2011Q2	2013Q2	0.9248	0.9277	0.3220	0.3132
2010Q4	2011Q3	2013Q3	0.9256	0.9283	0.3189	0.3114
2011Q1	2011Q4	2013Q4	0.9255	0.9284	0.3192	0.3103
Average			0.9247	0.9272	0.3262	0.3186

Notes: Performance comparison of the two best performing machine learning classification models of consumer default risk. The model calibrations are specified by the training and testing windows. The results of predicted probabilities versus actual outcomes over the following 8Q (testing period) are used to calculate the loss metric and the AUC-score for 90+ days delinquencies within 8Q. DNN refers to deep neural network, GBT refers to gradient boosted trees. Source: Authors' calculations based on Experian Data.

Table 43: Model Comparison: DNN vs. GBT, Full

Training Window*	Testing Window	AUC-score		Loss	
		DNN	GBT	DNN	GBT
2004Q1	2006Q1	0.9219	0.9239	0.3280	0.3236
2004Q2	2006Q2	0.9226	0.9249	0.3264	0.3193
2004Q3	2006Q3	0.9238	0.9262	0.3238	0.3165
2004Q4	2006Q4	0.9234	0.9252	0.3264	0.3198
2005Q1	2007Q1	0.9234	0.9257	0.3281	0.3212
2005Q2	2007Q2	0.9230	0.9255	0.3309	0.3228
2005Q3	2007Q3	0.9223	0.9249	0.3342	0.3255
2005Q4	2007Q4	0.9217	0.9238	0.3366	0.3288
2006Q1	2008Q1	0.9226	0.9245	0.3362	0.3292
2006Q2	2008Q2	0.9223	0.9240	0.3363	0.3308
2006Q3	2008Q3	0.9229	0.9250	0.3354	0.3285
2006Q4	2008Q4	0.9229	0.9252	0.3354	0.3284
2007Q1	2009Q1	0.9246	0.9271	0.3314	0.3237
2007Q2	2009Q2	0.9249	0.9273	0.3292	0.3225
2007Q3	2009Q3	0.9260	0.9284	0.3257	0.3188
2007Q4	2009Q4	0.9278	0.9300	0.3213	0.3146
2008Q1	2010Q1	0.9277	0.9300	0.3224	0.3155
2008Q2	2010Q2	0.9277	0.9300	0.3215	0.3147
2008Q3	2010Q3	0.9276	0.9299	0.3211	0.3140
2008Q4	2010Q4	0.9278	0.9297	0.3205	0.3147
2009Q1	2011Q1	0.9285	0.9308	0.3188	0.3119
2009Q2	2011Q2	0.9265	0.9291	0.3224	0.3152
2009Q3	2011Q3	0.9267	0.9292	0.3216	0.3142
2009Q4	2011Q4	0.9269	0.9291	0.3206	0.3141
2010Q1	2012Q1	0.9268	0.9296	0.3206	0.3125
2010Q2	2012Q2	0.9255	0.9282	0.3226	0.3151
2010Q3	2012Q3	0.9252	0.9278	0.3214	0.3140
2010Q4	2012Q4	0.9246	0.9274	0.3236	0.3150
2011Q1	2013Q1	0.9256	0.9281	0.3198	0.3131
2011Q2	2013Q2	0.9251	0.9277	0.3202	0.3130
2011Q3	2013Q3	0.9257	0.9282	0.3192	0.3112
2011Q4	2013Q4	0.9256	0.9283	0.3181	0.3101
Average		0.9250	0.9273	0.3256	0.3185

Notes: Performance comparison of the two best performing machine learning classification models of consumer default risk. The model calibrations are specified by the training and testing windows. * implies that all data was used up to the quarter specified. The results of predicted probabilities versus actual outcomes over the following 8Q (testing period) are used to calculate the loss metric and the AUC-score for 90+ days delinquencies within 8Q. DNN refers to deep neural network, GBT refers to gradient boosted trees. Source: Authors' calculations based on Experian Data.

Table 44: Constraining Model Behavior

Training Window	Testing Window	AUC-score			Loss		
		UN	R I	R II	UN	R I	R II
2004Q1	2006Q1	0.9239	0.9239	0.9222	0.3236	0.3237	0.3262
2004Q2	2006Q2	0.9247	0.9247	0.9230	0.3189	0.3189	0.3219
2004Q3	2006Q3	0.9257	0.9257	0.9239	0.3167	0.3167	0.3199
2004Q4	2006Q4	0.9246	0.9247	0.9229	0.3203	0.3202	0.3235
2005Q1	2007Q1	0.9254	0.9254	0.9237	0.3212	0.3212	0.3243
2005Q2	2007Q2	0.9255	0.9255	0.9238	0.3217	0.3217	0.3249
2005Q3	2007Q3	0.9249	0.9249	0.9231	0.3248	0.3248	0.3281
2005Q4	2007Q4	0.9235	0.9235	0.9218	0.3283	0.3284	0.3313
2006Q1	2008Q1	0.9245	0.9246	0.9229	0.3269	0.3267	0.3298
2006Q2	2008Q2	0.9242	0.9244	0.9227	0.3280	0.3277	0.3308
2006Q3	2008Q3	0.9253	0.9254	0.9237	0.3262	0.3258	0.3291
2006Q4	2008Q4	0.9255	0.9257	0.9240	0.3264	0.3260	0.3292
2007Q1	2009Q1	0.9274	0.9276	0.9259	0.3219	0.3213	0.3246
2007Q2	2009Q2	0.9277	0.9278	0.9261	0.3204	0.3201	0.3237
2007Q3	2009Q3	0.9284	0.9285	0.9269	0.3179	0.3178	0.3210
2007Q4	2009Q4	0.9297	0.9299	0.9284	0.3145	0.3142	0.3176
2008Q1	2010Q1	0.9296	0.9296	0.9281	0.3157	0.3157	0.3190
2008Q2	2010Q2	0.9293	0.9294	0.9277	0.3161	0.3161	0.3200
2008Q3	2010Q3	0.9289	0.9290	0.9273	0.3164	0.3165	0.3202
2008Q4	2010Q4	0.9286	0.9287	0.9269	0.3177	0.3178	0.3216
2009Q1	2011Q1	0.9293	0.9294	0.9278	0.3154	0.3155	0.3190
2009Q2	2011Q2	0.9279	0.9279	0.9261	0.3182	0.3184	0.3222
2009Q3	2011Q3	0.9281	0.9281	0.9264	0.3170	0.3170	0.3208
2009Q4	2011Q4	0.9285	0.9285	0.9267	0.3158	0.3157	0.3195
2010Q1	2012Q1	0.9290	0.9291	0.9273	0.3143	0.3144	0.3181
2010Q2	2012Q2	0.9277	0.9277	0.9260	0.3164	0.3165	0.3202
2010Q3	2012Q3	0.9275	0.9276	0.9257	0.3152	0.3153	0.3192
2010Q4	2012Q4	0.9269	0.9269	0.9250	0.3167	0.3168	0.3209
2011Q1	2013Q1	0.9278	0.9278	0.9257	0.3143	0.3143	0.3186
2011Q2	2013Q2	0.9274	0.9274	0.9255	0.3140	0.3141	0.3180
2011Q3	2013Q3	0.9280	0.9280	0.9260	0.3123	0.3124	0.3164
2011Q4	2013Q4	0.9283	0.9282	0.9262	0.3104	0.3106	0.3148
Average		0.9270	0.9270	0.9253	0.3189	0.3188	0.3223

Notes: Performance comparison of GBT models of consumer default risk under various monotonicity constraint regimes. UN denotes the unconstrained model, while R I and R II are the models under Regime I and Regime II respectively. The model calibrations are specified by the training and testing windows. The results of predicted probabilities versus actual outcomes over the following 8Q testing period are used to calculate the loss metric for 90+ days delinquencies within 8Q. Source: Authors' calculations based on Experian Data.

Table 45: Distribution of Customers by Credit Score and Predicted Default

		Credit Score			
		Subprime, Near Prime	Prime Low	Prime Mid	Prime High, Superprime
Predicted Default bin	1	36.24%	3.60%	1.24%	0.42%
	2	3.81%	3.50%	2.40%	1.07%
	3	1.12%	2.70%	3.96%	3.32%
	4	0.32%	0.97%	3.51%	31.81%

Notes: This table reports the share of customers in each predicted credit score categories and corresponding predicted default probability bins. Customers are classified based on credit scores and predicted default probabilities at account origination for each of their credit cards included in the balances. Source: Authors' calculations based on Experian data.

Appendix B - Towards Fair Credit Allocation

B.1 Features

The model inputs and feature groups are as in the first chapter.

B.2 Classifier Performance

In this section, we describe the performance of our models under various training and testing windows. To account for look-ahead bias, we train and test our models based on 8 quarter windows that were observable at the time of forecast. In particular, we require our training and testing sets to be separated by 8 quarters to avoid overlap. For instance, the second out-of-sample model was calibrated using input data from 2004Q2, from which the parameter estimates were applied to the input data in 2006Q2 to generate forecasts of delinquencies over the 8 quarter window from 2006Q3-2008Q2. This gives us a total of 32 calibration and testing periods.

We compute seven metrics, which we have provided earlier, but describe heuristically here. The area under the ROC curve, known as *AUC score*, can be interpreted as the probability of the classifier assigning a higher probability of being in default to an account that is actually in default. *Precision* measures the model's accuracy in instances that are classified as default. *Recall* refers to the number of accounts that defaulted as identified by the model divided by the actual number of defaulting accounts. The *F-measure* is simply the harmonic mean of precision and recall. In an ideal scenario, we would have very high precision and recall. *Accuracy* refers to the fraction of accounts classified correctly, and *Loss* incorporates the magnitude of the mistakes made.

B.2.1 Unconstrained Models

We first look at the performance of our unconstrained models. Table 46 presents the results for the logistic regression, while Table 47 presents it for the unconstrained DNN. We can see that the DNN improves on the performance of the logistic regression substantially. For instance, for our testing window of 2006Q1, our logistic regression accurately classified 85.6% of instances, while our DNN accurately classified 86.3% of instances. For instance, the 84.44% precision implies that when our classifier predicts that someone is going to default, there is an 84.44% chance this person will actually default; while the 70.78% recall means that we accurately identified 70.78% of all the defaulters, both of these are higher than for the logistic regression (83.78% and 68.94% respectively).

B.2.2 Constrained Models

Table 48 presents the results for our age constrained DNN, while Table 49 presents it for the race constrained DNN. We can see that these models outperform the logistic regression, but in a sense of standard machine learning metrics, they underperform the DNN.

B.3 Model Interpretation

We now turn to analyzing the economic significance of our features for default behavior across our models. We adopt SHapley Additive exPlanations (SHAP), a unified framework for interpreting predictions, to explain the output of our hybrid deep learning model (for a detailed description of the approach see [63]). SHAP uses a game theoretical concept to assign each feature a local importance value for a given prediction. Though Shapley values are local by design, they can be combined into global explanations by averaging the absolute Shapley values featurewise. Then, we can compare features based on their absolute average Shapley values, with higher values implying higher feature importance.[1]

[1] We implement Deep SHAP, a high-speed approximation algorithm for SHAP values in deep learning models to compute the Shapley values for our 5 hidden layer neural network. We use a random sample of 100,000 observations for explaining the model.

By the Shapley efficient property, the SHAP values for an observation sum up to the difference between the predicted value of that observation and the expected value, computed using the background dataset:

$$f(x) = E_X[\widehat{f}(X)] + \sum_{j=1}^{M} \phi_j \tag{B.1}$$

where f is the model prediction, M is the number of features, and $\phi_j \in R$ is the feature attribution for feature j (i.e., the Shapley values). Thus, we can interpret the Shapley value as the contribution of a feature value to the difference between the model's prediction and the mean prediction, given the current set of feature values. As an illustration, a SHAP value of 0.1 implies that the feature's value for that particular instance contributed to an increase of 0.1 to the model's output compared to the mean prediction. SHAP values are in model output units before it is passed to the final softmax layer.

Features that are highly correlated can decrease the importance of the associated feature by splitting the importance between both features. We account for the effect of feature correlation on interpretability by grouping features with a correlation larger than 0.7, and summing the SHAP values within each groups. We denote these groups with an asterisk for the rest of the analysis and these groups are as in the first chapter.

Table 50 presents the feature attributions for selected quarters for our unconstrained models. Though the seven most important features are similar, the logistic regression places significantly higher weight on collections. Mortgage debt, card balances, and inquiries do not make the top 10 for the unconstrained deep neural network, or for any of the constrained deep neural networks reported in Table 51. On the other hand, total debt balances and monthly payment on credit cards are among the ten most important features for the deep neural networks, but not for the logistic regression. A key difference between the constrained and the unconstrained DNNs is credit card utilization, which is picked up as one of the ten most important features for the unconstrained model, while for the constrained model it is replaced by balances on revolving debt.

These results only point to correlations between the features and the predicted outcome and should not be interpreted causally. Yet, they can be used to compare feature attributions across models. They are also important to comply with legal disclosure requirements.

Both the Fair Credit Reporting Act ad the Equal Opportunity in Credit Access Act require lenders and developers of credit scoring models to reveal the most important factors leading to a denial of a credit application and for credit scores. The SHAP value provides an individualized assessment of such factors that can be used for making credit allocation decisions and communicating them to the borrower.

B.4 Additional Tables

Table 46: Performance Metrics using Logistic Regression

Training Window	Testing Window	AUC score	Precision	Recall	F-measure	Accuracy	Loss
2004Q1	2006Q1	0.9123	0.8378	0.6894	0.7564	0.8558	0.3469
2004Q2	2006Q2	0.9126	0.8338	0.7037	0.7633	0.8570	0.3449
2004Q3	2006Q3	0.9138	0.8367	0.7111	0.7688	0.8585	0.3463
2004Q4	2006Q4	0.9132	0.8386	0.7127	0.7705	0.8577	0.3453
2005Q1	2007Q1	0.9120	0.8453	0.7077	0.7704	0.8561	0.3484
2005Q2	2007Q2	0.9128	0.8487	0.7080	0.7720	0.8558	0.3491
2005Q3	2007Q3	0.9111	0.8516	0.7035	0.7705	0.8545	0.3529
2005Q4	2007Q4	0.9106	0.8507	0.7080	0.7728	0.8538	0.3555
2006Q1	2008Q1	0.9111	0.8468	0.7139	0.7747	0.8531	0.3557
2006Q2	2008Q2	0.9110	0.8482	0.7123	0.7744	0.8525	0.3564
2006Q3	2008Q3	0.9119	0.8507	0.7125	0.7755	0.8530	0.3552
2006Q4	2008Q4	0.9136	0.8485	0.7210	0.7796	0.8538	0.3532
2007Q1	2009Q1	0.9157	0.8414	0.7373	0.7859	0.8561	0.3502
2007Q2	2009Q2	0.9165	0.8350	0.7471	0.7886	0.8567	0.3481
2007Q3	2009Q3	0.9182	0.8335	0.7541	0.7918	0.8582	0.3452
2007Q4	2009Q4	0.9195	0.8368	0.7579	0.7954	0.8600	0.3420
2008Q1	2010Q1	0.9186	0.8325	0.7610	0.7951	0.8594	0.3446
2008Q2	2010Q2	0.9186	0.8239	0.7673	0.7946	0.8585	0.3464
2008Q3	2010Q3	0.9188	0.8259	0.7643	0.7939	0.8584	0.3452
2008Q4	2010Q4	0.9185	0.8216	0.7666	0.7931	0.8583	0.3446
2009Q1	2011Q1	0.9185	0.8229	0.7665	0.7937	0.8589	0.3447
2009Q2	2011Q2	0.9179	0.8230	0.7595	0.7899	0.8580	0.3445
2009Q3	2011Q3	0.9187	0.8255	0.7577	0.7901	0.8585	0.3426
2009Q4	2011Q4	0.9186	0.8224	0.7594	0.7897	0.8590	0.3418
2010Q1	2012Q1	0.9186	0.8192	0.7643	0.7908	0.8597	0.3409
2010Q2	2012Q2	0.9173	0.8163	0.7619	0.7882	0.8591	0.3425
2010Q3	2012Q3	0.9172	0.8092	0.7684	0.7882	0.8592	0.3418
2010Q4	2012Q4	0.9152	0.8024	0.7643	0.7829	0.8576	0.3446
2011Q1	2013Q1	0.9161	0.8037	0.7637	0.7832	0.8587	0.3422
2011Q2	2013Q2	0.9158	0.8011	0.7628	0.7815	0.8585	0.3420
2011Q3	2013Q3	0.9162	0.8009	0.7629	0.7815	0.8592	0.3405
2011Q4	2013Q4	0.9163	0.7970	0.7661	0.7813	0.8594	0.3393

Notes: Performance metrics for our model of default risk for the logistic regression. The model calibrations are specified by the training and testing windows. The results of classifications versus actual outcomes over the following 8Q are used to calculate these performance metrics for 90+ days delinquencies within 8Q. Source: Authors' calculations based on Experian Data.

Table 47: Performance Metrics using Unconstrained DNN

Training Window	Testing Window	AUC score	Precision	Recall	F-measure	Accuracy	Loss
2004Q1	2006Q1	0.9229	0.8444	0.7078	0.7701	0.8627	0.3276
2004Q2	2006Q2	0.9236	0.8406	0.7208	0.7761	0.8638	0.3244
2004Q3	2006Q3	0.9241	0.8449	0.7228	0.7791	0.8644	0.3239
2004Q4	2006Q4	0.9244	0.8447	0.7275	0.7818	0.8639	0.3241
2005Q1	2007Q1	0.9230	0.8582	0.7093	0.7767	0.8609	0.3314
2005Q2	2007Q2	0.9244	0.8598	0.7136	0.7799	0.8611	0.3311
2005Q3	2007Q3	0.9231	0.8511	0.7248	0.7829	0.8604	0.3321
2005Q4	2007Q4	0.9230	0.8470	0.7342	0.7866	0.8601	0.3322
2006Q1	2008Q1	0.9225	0.8504	0.7303	0.7858	0.8591	0.3342
2006Q2	2008Q2	0.9231	0.8450	0.7392	0.7886	0.8592	0.3328
2006Q3	2008Q3	0.9240	0.8453	0.7407	0.7895	0.8593	0.3317
2006Q4	2008Q4	0.9251	0.8469	0.7450	0.7927	0.8602	0.3291
2007Q1	2009Q1	0.9257	0.8385	0.7627	0.7988	0.8623	0.3277
2007Q2	2009Q2	0.9258	0.8434	0.7564	0.7976	0.8626	0.3262
2007Q3	2009Q3	0.9271	0.8450	0.7591	0.7998	0.8641	0.3238
2007Q4	2009Q4	0.9283	0.8470	0.7665	0.8047	0.8665	0.3203
2008Q1	2010Q1	0.9276	0.8277	0.7910	0.8089	0.8660	0.3225
2008Q2	2010Q2	0.9273	0.8293	0.7858	0.8070	0.8659	0.3230
2008Q3	2010Q3	0.9274	0.8189	0.7985	0.8086	0.8651	0.3229
2008Q4	2010Q4	0.9272	0.8206	0.7936	0.8069	0.8654	0.3234
2009Q1	2011Q1	0.9272	0.8218	0.7929	0.8071	0.8657	0.3226
2009Q2	2011Q2	0.9262	0.8290	0.7765	0.8019	0.8651	0.3244
2009Q3	2011Q3	0.9268	0.8322	0.7750	0.8026	0.8660	0.3254
2009Q4	2011Q4	0.9266	0.8275	0.7793	0.8027	0.8664	0.3222
2010Q1	2012Q1	0.9266	0.8214	0.7850	0.8028	0.8662	0.3215
2010Q2	2012Q2	0.9254	0.8399	0.7497	0.7922	0.8647	0.3241
2010Q3	2012Q3	0.9257	0.8152	0.7823	0.7984	0.8653	0.3225
2010Q4	2012Q4	0.9236	0.8038	0.7826	0.7931	0.8628	0.3273
2011Q1	2013Q1	0.9245	0.8158	0.7673	0.7908	0.8643	0.3239
2011Q2	2013Q2	0.9240	0.7925	0.7932	0.7929	0.8625	0.3269
2011Q3	2013Q3	0.9249	0.8131	0.7673	0.7895	0.8651	0.3212
2011Q4	2013Q4	0.9249	0.8045	0.7764	0.7902	0.8649	0.3204

Notes: Performance metrics for our model of default risk for the unconstrained DNN model. The model calibrations are specified by the training and testing windows. The results of classifications versus actual outcomes over the following 8Q are used to calculate these performance metrics for 90+ days delinquencies within 8Q. Source: Authors' calculations based on Experian Data.

Table 48: Performance Metrics using Constrained DNN by Age

Training Window	Testing Window	AUC score	Precision	Recall	F-measure	Accuracy	Loss
2004Q1	2006Q1	0.9238	0.8357	0.7602	0.7961	0.8621	0.3309
2004Q2	2006Q2	0.9215	0.8664	0.6910	0.7688	0.8567	0.3367
2004Q3	2006Q3	0.9229	0.8622	0.7069	0.7769	0.8553	0.3390
2004Q4	2006Q4	0.9218	0.8510	0.7002	0.7683	0.8602	0.3337
2005Q1	2007Q1	0.9232	0.8077	0.7810	0.7941	0.8619	0.3289
2005Q2	2007Q2	0.9215	0.8607	0.7059	0.7756	0.8549	0.3420
2005Q3	2007Q3	0.9225	0.8283	0.7582	0.7917	0.8597	0.3337
2005Q4	2007Q4	0.9215	0.8650	0.6849	0.7645	0.8561	0.3358
2006Q1	2008Q1	0.9243	0.8529	0.7350	0.7896	0.8597	0.3321
2006Q2	2008Q2	0.9226	0.8114	0.7670	0.7886	0.8626	0.3272
2006Q3	2008Q3	0.9257	0.8406	0.7643	0.8006	0.8635	0.3269
2006Q4	2008Q4	0.9211	0.8549	0.6830	0.7593	0.8582	0.3351
2007Q1	2009Q1	0.9208	0.8596	0.7088	0.7769	0.8560	0.3407
2007Q2	2009Q2	0.9202	0.8695	0.6907	0.7699	0.8549	0.3401
2007Q3	2009Q3	0.9195	0.8688	0.6823	0.7643	0.8539	0.3505
2007Q4	2009Q4	0.9237	0.8328	0.7599	0.7947	0.8631	0.3295
2008Q1	2010Q1	0.9250	0.8390	0.7605	0.7978	0.8621	0.3293
2008Q2	2010Q2	0.9230	0.8122	0.7608	0.7857	0.8631	0.3251
2008Q3	2010Q3	0.9245	0.8247	0.7781	0.8007	0.8627	0.3290
2008Q4	2010Q4	0.9245	0.8219	0.7849	0.8030	0.8626	0.3297
2009Q1	2011Q1	0.9262	0.8412	0.7647	0.8011	0.8637	0.3258
2009Q2	2011Q2	0.9228	0.8310	0.7517	0.7893	0.8620	0.3298
2009Q3	2011Q3	0.9242	0.8310	0.7668	0.7976	0.8611	0.3305
2009Q4	2011Q4	0.9201	0.8504	0.6802	0.7559	0.8573	0.3335
2010Q1	2012Q1	0.9229	0.8160	0.7590	0.7864	0.8632	0.3259
2010Q2	2012Q2	0.9234	0.8211	0.7730	0.7963	0.8628	0.3298
2010Q3	2012Q3	0.9218	0.8612	0.6909	0.7667	0.8591	0.3333
2010Q4	2012Q4	0.9231	0.7929	0.7849	0.7889	0.8623	0.3255
2011Q1	2013Q1	0.9238	0.8435	0.7518	0.7950	0.8613	0.3301
2011Q2	2013Q2	0.9245	0.8451	0.7431	0.7908	0.8618	0.3296
2011Q3	2013Q3	0.9215	0.8185	0.7535	0.7846	0.8611	0.3305
2011Q4	2013Q4	0.9235	0.8619	0.7139	0.7809	0.8563	0.3371

Notes: Performance metrics for our model of default risk for the age constrained DNN model. The model calibrations are specified by the training and testing windows. The results of classifications versus actual outcomes over the following 8Q are used to calculate these performance metrics for 90+ days delinquencies within 8Q. Source: Authors' calculations based on Experian Data.

Table 49: Performance Metrics using Constrained DNN by Race

Training Window	Testing Window	AUC score	Precision	Recall	F-measure	Accuracy	Loss
2004Q1	2006Q1	0.9245	0.8259	0.7789	0.8017	0.8636	0.3289
2004Q2	2006Q2	0.9224	0.8672	0.6915	0.7695	0.8571	0.3346
2004Q3	2006Q3	0.9229	0.8619	0.7086	0.7777	0.8557	0.3399
2004Q4	2006Q4	0.9220	0.8501	0.7055	0.7711	0.8614	0.3326
2005Q1	2007Q1	0.9221	0.8064	0.7754	0.7906	0.8599	0.3310
2005Q2	2007Q2	0.9213	0.8566	0.7111	0.7771	0.8551	0.3400
2005Q3	2007Q3	0.9236	0.8244	0.7715	0.7971	0.8619	0.3305
2005Q4	2007Q4	0.9217	0.8615	0.6949	0.7693	0.8578	0.3349
2006Q1	2008Q1	0.9245	0.8450	0.7470	0.7930	0.8603	0.3305
2006Q2	2008Q2	0.9227	0.8064	0.7746	0.7902	0.8625	0.3273
2006Q3	2008Q3	0.9263	0.8306	0.7827	0.8059	0.8648	0.3259
2006Q4	2008Q4	0.9216	0.8539	0.6869	0.7614	0.8589	0.3331
2007Q1	2009Q1	0.9215	0.8623	0.7052	0.7759	0.8559	0.3402
2007Q2	2009Q2	0.9204	0.8685	0.6917	0.7701	0.8549	0.3391
2007Q3	2009Q3	0.9195	0.8677	0.6852	0.7657	0.8544	0.3504
2007Q4	2009Q4	0.9239	0.8207	0.7791	0.7993	0.8636	0.3285
2008Q1	2010Q1	0.9253	0.8331	0.7695	0.8001	0.8625	0.3283
2008Q2	2010Q2	0.9231	0.8002	0.7773	0.7886	0.8625	0.3255
2008Q3	2010Q3	0.9249	0.8169	0.7907	0.8036	0.8630	0.3280
2008Q4	2010Q4	0.9255	0.8145	0.7962	0.8053	0.8626	0.3276
2009Q1	2011Q1	0.9264	0.8323	0.7788	0.8047	0.8643	0.3252
2009Q2	2011Q2	0.9234	0.8205	0.7705	0.7947	0.8631	0.3284
2009Q3	2011Q3	0.9249	0.8280	0.7766	0.8015	0.8627	0.3285
2009Q4	2011Q4	0.9202	0.8540	0.6766	0.7550	0.8574	0.3363
2010Q1	2012Q1	0.9229	0.8064	0.7717	0.7887	0.8628	0.3261
2010Q2	2012Q2	0.9243	0.8143	0.7862	0.8000	0.8636	0.3281
2010Q3	2012Q3	0.9220	0.8599	0.6941	0.7681	0.8596	0.3318
2010Q4	2012Q4	0.9232	0.7888	0.7910	0.7899	0.8621	0.3261
2011Q1	2013Q1	0.9235	0.8338	0.7634	0.7971	0.8610	0.3307
2011Q2	2013Q2	0.9239	0.8248	0.7730	0.7981	0.8625	0.3290
2011Q3	2013Q3	0.9219	0.8097	0.7677	0.7881	0.8613	0.3301
2011Q4	2013Q4	0.9231	0.8559	0.7229	0.7838	0.8570	0.3373

Notes: Performance metrics for our model of default risk for the race constrained DNN model. The model calibrations are specified by the training and testing windows. The results of classifications versus actual outcomes over the following 8Q are used to calculate these performance metrics for 90+ days delinquencies within 8Q. Source: Authors' calculations based on Experian Data.

Table 50: SHAP Values: Unconstrained Models

Feature	2006Q1	2008Q1	2010Q1	2012Q1
Panel A: Logistic Regression				
Number of collections*	0.601 (1.0)	0.595 (1.0)	0.532 (1.0)	0.697 (1.0)
Number of credit cards*	0.388 (2.0)	0.331 (4.0)	0.351 (4.0)	0.344 (4.0)
Worst status on any trades*	0.379 (3.0)	0.389 (2.0)	0.414 (2.0)	0.483 (2.0)
Months since the most recent 90+ days delinquency	0.356 (4.0)	0.352 (3.0)	0.356 (3.0)	0.432 (3.0)
Months since the most recent 30-180days delinquency*	0.224 (5.0)	0.222 (6.0)	0.214 (7.0)	0.204 (5.0)
Worst status on credit card trades*	0.216 (6.0)	0.23 (5.0)	0.222 (6.0)	0.193 (6.0)
Mortgage debt*	0.208 (7.0)	0.176 (8.0)	0.115 (11.0)	0.12 (11.0)
Fraction of 90+ days delinquent debt*	0.206 (8.0)	0.215 (7.0)	0.225 (5.0)	0.136 (8.0)
Balance on credit & bankcards	0.151 (9.0)	0.159 (10.0)	0.212 (8.0)	0.076 (14.0)
Inquiries made in the last 12 months (no deduplication)	0.131 (10.0)	0.162 (9.0)	0.142 (10.0)	0.126 (9.0)
Panel B: DNN				
Worst status on any trades*	0.365 (1.0)	0.389 (2.0)	0.395 (3.0)	0.5 (2.0)
Number of collections*	0.355 (2.0)	0.465 (1.0)	0.416 (1.0)	0.54 (1.0)
Number of credit cards*	0.31 (3.0)	0.369 (3.0)	0.325 (4.0)	0.402 (3.0)
Months since the most recent 90+ days delinquency	0.225 (4.0)	0.21 (5.0)	0.225 (5.0)	0.252 (4.0)
Worst status on credit card trades*	0.191 (5.0)	0.224 (4.0)	0.175 (6.0)	0.146 (7.0)
Total debt balances*	0.179 (6.0)	0.176 (6.0)	0.078 (18.0)	0.086 (14.0)
Months since the most recent 30-180days delinquency*	0.145 (7.0)	0.152 (10.0)	0.146 (9.0)	0.138 (9.0)
Fraction of 90+ days delinquent debt*	0.143 (8.0)	0.145 (11.0)	0.162 (7.0)	0.125 (10.0)
Monthly payment on credit card trades	0.118 (9.0)	0.155 (7.0)	0.397 (2.0)	0.181 (6.0)
Credit card utilization	0.109 (10.0)	0.089 (15.0)	0.152 (8.0)	0.082 (15.0)

Notes: This table reports the Shapley values for a selected 10 features for four out-of-sample models. For each prediction window, we compute the Shapley value for 100,000 observations and for each feature. We then calculate the average of the absolute value for each feature, and report the results for the selected features. Finally, we rank the results based on the feature's relative rank in the given prediction window in parentheses. Source: Authors' calculations based on Experian data.

Table 51: SHAP Values: Constrained Models

Feature	2006Q1	2008Q1	2010Q1	2012Q1
Panel A: Race Constrained				
Number of collections*	0.441 (1.0)	0.379 (2.0)	0.514 (1.0)	0.546 (1.0)
Worst status on any trades*	0.411 (2.0)	0.46 (1.0)	0.425 (2.0)	0.488 (2.0)
Number of credit cards*	0.31 (3.0)	0.336 (3.0)	0.31 (4.0)	0.443 (3.0)
Months since the most recent 90+ days delinquency	0.211 (4.0)	0.188 (5.0)	0.222 (5.0)	0.304 (4.0)
Worst status on credit card trades*	0.193 (5.0)	0.225 (4.0)	0.211 (6.0)	0.142 (7.0)
Months since the most recent 30-180days delinquency*	0.181 (6.0)	0.164 (6.0)	0.179 (9.0)	0.181 (5.0)
Balance on revolving debt*	0.174 (7.0)	0.102 (13.0)	0.094 (16.0)	0.063 (18.0)
Fraction of 90+ days delinquent debt*	0.161 (8.0)	0.16 (7.0)	0.187 (8.0)	0.121 (10.0)
Total debt balances*	0.157 (9.0)	0.145 (10.0)	0.071 (21.0)	0.048 (23.0)
Monthly payment on credit card trades	0.146 (10.0)	0.154 (9.0)	0.407 (3.0)	0.172 (6.0)
Panel B: Age Constrained				
Number of collections*	0.416 (1.0)	0.395 (2.0)	0.513 (1.0)	0.631 (1.0)
Worst status on any trades*	0.407 (2.0)	0.446 (1.0)	0.424 (2.0)	0.529 (2.0)
Number of credit cards*	0.29 (3.0)	0.37 (3.0)	0.302 (4.0)	0.429 (3.0)
Months since the most recent 90+ days delinquency	0.216 (4.0)	0.224 (5.0)	0.226 (5.0)	0.325 (4.0)
Worst status on credit card trades*	0.205 (5.0)	0.228 (4.0)	0.206 (6.0)	0.165 (6.0)
Months since the most recent 30-180days delinquency*	0.17 (6.0)	0.168 (7.0)	0.176 (9.0)	0.197 (5.0)
Total debt balances*	0.168 (7.0)	0.155 (9.0)	0.068 (21.0)	0.048 (21.0)
Balance on revolving debt*	0.164 (8.0)	0.087 (13.0)	0.101 (14.0)	0.075 (14.0)
Fraction of 90+ days delinquent debt*	0.153 (9.0)	0.179 (6.0)	0.19 (7.0)	0.154 (8.0)
Monthly payment on credit card trades	0.142 (10.0)	0.164 (8.0)	0.358 (3.0)	0.155 (7.0)

Notes: This table reports the Shapley values for a selected 10 features for four out-of-sample models. For each prediction window, we compute the Shapley value for 100,000 observations and for each feature. We then calculate the average of the absolute value for each feature, and report the results for the selected features. Finally, we rank the results based on the feature's relative rank in the given prediction window in parentheses. Source: Authors' calculations based on Experian data.

Appendix C - Racial Disparities in Debt Collection

C.1 Judgment Data

We obtained our judgment data from Paul Kiel and Annie Waldman at ProPublica. This data included all debt collection judgments in New Jersey, Missouri, and Cook County Illinois from 2008 to 2012.[1] Both Missouri and New Jersey have state-wide databases. The Missouri dataset was provided by Missouri's Office of the State Courts Administrator (OSCA) and included all debt collection cases filed in Associate Circuit Court for which OSCA has an electronic record through early 2014.[2] For each case in Missouri, the data contained the following information: court (judicial circuit), county, case ID, filing Date, case type, disposition, plaintiff, plaintiff attorney, defendant, defendant date of birth, defendant address, defendant attorney, judgment amount, date of judgment satisfaction, date of first garnishment attempt.[3] Kiel and Waldman added two fields: a standard name for each plaintiff and a plaintiff type. St. Louis County joined the state's online system in 2007 and St. Louis City has been online since 2000. Missouri's court system has some variation among the judicial circuits in how case types are categorized, so, in consultation with OSCA employees, Keil and Waldman selected a range of case types that could be reasonably construed as debt collection cases. For St. Louis City and County courts, these were: Breach of Contract, Promissory Note, Suit on Account, Contract /Account (Bulk), Misc Associate Civil-Other, Small Claims under $100, Small Claims over $100.[4] They limited the dataset to cases that had resulted in a judgment.

[1] In their ProPublica articles, Paul and Annie focus on Essex County, St. Louis City, St. Louis County, and Cook County due to the cities high segregation indexes. Due to a peculiarity of the court system database, the Essex County window is slightly different: July 1, 2007, through June 30, 2012. Futhermore, various circuits in Missouri came online at different times, but all circuits were online by 2008.

[2] The max amount sought in associate circuit courts in Missouri is $25K.

[3] The judgment amount was determined to be unreliable and is not used throughout this analysis.

[4] Together, the small claims and "misc associate" cases comprised less than four percent of cases.

C.2 BISG Algorithm

Vectors of six racial/ethnic probabilities for each listed surname (corrected for suppression and for low-frequency surnames) are used as the first input into the BISG algorithm. This information is used to calculate a prior probability of an individual's race/ethnicity. The algorithm updates these prior probabilities with geocoded ZCTA proportions for these groups from the 2010 Census SF1 files to generate posterior probabilities. Let J equal the number of names on the enhanced surname list plus one to account for names not on the list and let K equal the number of ZCTA in the 2010 census with any population. We define the prior probability of a person's race on the basis of surname, so that for a person with surname j = 1, ..., J on the list, the prior probability for race, i = 1,...,6, is p(i—j) = proportion of all people with surname j who report being of race i in the enhanced surname file (the probability of a selected race given surname). This probability is updated on the basis of ZCTA residence. For ZCTA k = 1,...,K, r(k—i) = proportion of all people in redistributed SF1 file who self report being race i who reside in ZCTA k (the probability of a selected ZCTA of residence given race/ethnicity). Let $u(i, j, k) = p(i|j) * r(k|i)$. According to Bayes' Theorem and the assumption that the probability of residing in a given ZCTA given a person's race does not vary by surname, the updated (posterior) probability of being of race/ethnicity i given surname j and ZCTA of residence k can be calculated as follows:

$$q(i|j,k) = \frac{u(i,j,k)}{u(1,j,k)+u(2,j,k)+u(3,j,k)+u(4,j,k)+u(5,j,k)+u(6,j,k)} \tag{C.1}$$

Note that all parameters needed for BISG posterior probabilities are derived only from Census 2010 data, and that none are derived from administrative sources.

C.2.1 Gradient Boosted Trees (GBT)

Gradient Boosted Trees (GBT) is an ensemble learning approach that mitigates the tendency of tree-based models' to overfit to training data. This is accomplished by recursively combining the forecasts of many over-simplified trees. The theory behind boosting proposes that a collection of weak learners as an ensemble create a single strong learner with improved stability over a single complex tree.

At each step m, $1 \leq m \leq M$, of gradient boosting, an estimator, h_m, is computed on the residuals from the previous models predictions. A critical part of gradient boosting method is regularization by shrinkage as proposed by [42]. This consists in modifying the update rule as follows:

$$F_m(x) = F_{m-1}(x) + \nu\gamma_m h_m(x), \tag{C.2}$$

where $h_m(x)$ represents a weak learner of fixed depth, γ_m is the step length and ν is the learning rate or shrinkage factor.

The estimation procedure begins with fitting a shallow tree (e.g., with depth L = 1). Using the prediction residuals from the first tree, you then fit a second tree with the same shallow depth L. Weight the predictions of the second tree by $\nu \in (0, 1)$ to prevent the model from overfitting the residuals, and then aggregate the forecasts of these two trees. At each step k, fit a shallow tree to the residuals from the model with k-1 trees, and add its prediction to the forecast of the ensemble with a shrinkage weight of ν. Do this until a total of K trees is reached in the ensemble. For our GBT model, we split the data into three chunks: training set (60%), holdout set (20%), and testing set (20%). We relied on XGBoost for the implementation of our GBT model ([29]).

C.3 Expanded Sample

For the interested reader, we relax the common support assumption and replicate Table 18, Table 19, Table 24, and Table 25 on the entire MO sample. Table 59, Table 60, Table 61, and Table 62 report the results. We find very similar results to our main specification which uses only the common support sample, suggesting that omitted variables that are correlated with observable neighborhood characteristics are not biasing our results in any particular direction.

C.4 Additional Figures and Tables

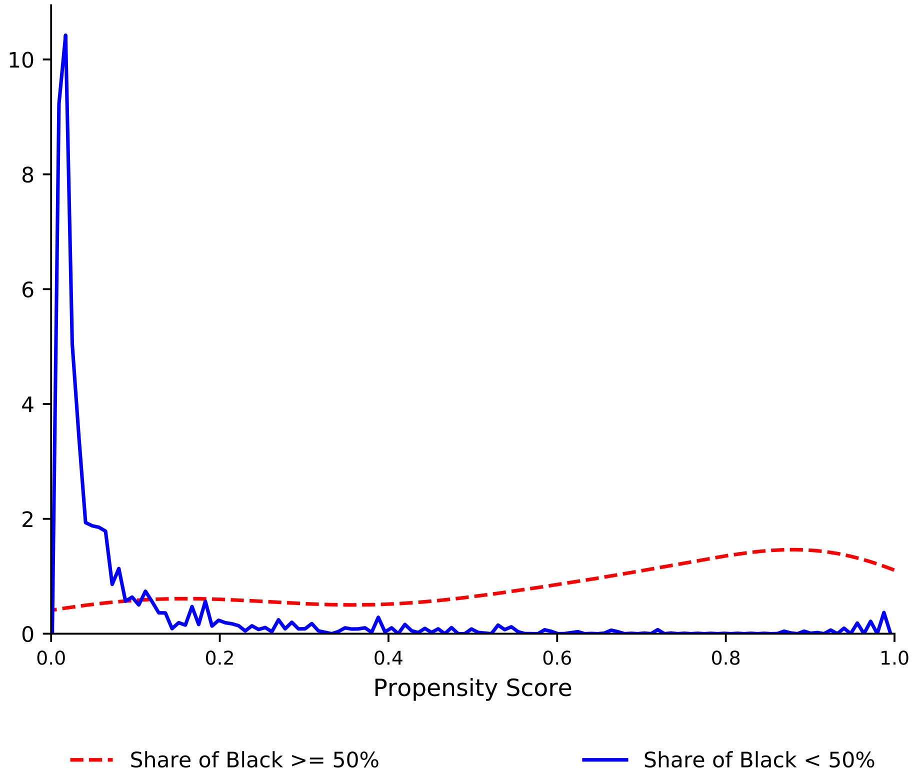


Figure 21: Propensity Score Distributions

Notes: The common support for the propensity score distributions: [0.0013, 0.9750]

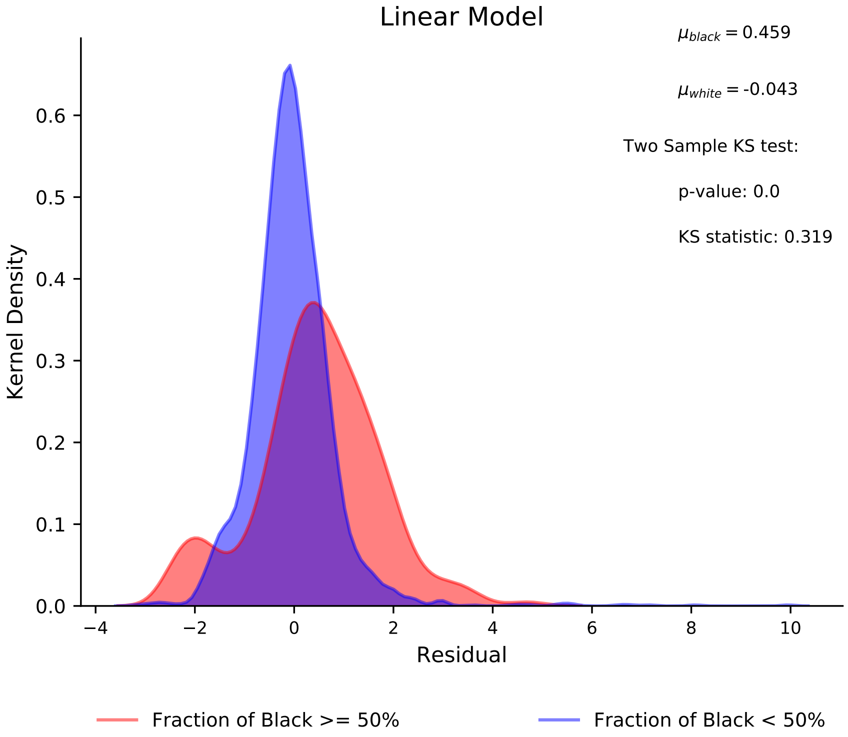


Figure 22: Residual Distributions by Race

Notes: We fit a linear model using our baseline controls and compute corresponding fitted values. We then obtain and plot the residuals separately for black majority and non-black majority ZIP codes. We test whether the distribution of black majority ZIP codes is to the right of the non-black majority distribution, and report the results of the KS-test in the upper right corner.

Table 52: Judgment Rates

	(1) Judgment Rate	(2) Dismissed	(3) Settle
Black Majority: ZIP	0.017*	0.007	-0.017***
	(0.009)	(0.011)	(0.003)
Median Household Income	-0.002	-0.001	0.002**
	(0.001)	(0.001)	(0.001)
Median Credit Score	-0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
Attorney	-0.192***	0.036	0.204***
	(0.068)	(0.067)	(0.075)
County Fixed Effects	X	X	X
Year Fixed Effects	X	X	X
Baseline Controls	X	X	X
Observations	2673	2673	2673
R^2	0.490	0.737	0.686

Notes: Robust standard errors clustered at the county level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. All regressions are weighted by population and estimated on the common support sample.

Table 53: Judgments and an Alternative Credit Score

	(1)	(2)	(3)	(4)	(5)	(6)
Black Majority: ZIP	1.3307***	0.8139***	0.8310***	0.8139***	0.6317***	0.7410***
	(0.1444)	(0.1383)	(0.1317)	(0.1383)	(0.1511)	(0.0519)
Median Household Income		-0.0040		-0.0040	0.0029	
		(0.0259)		(0.0259)	(0.0203)	
Predicted Probability		1.6121**	2.3538***	1.6121**	1.1066***	
		(0.6432)	(0.5456)	(0.6432)	(0.3345)	
Median Credit Score					-0.0026	
					(0.0039)	
County Fixed Effects	X	X	X	X	X	X
Year Fixed Effects	X	X	X	X	X	X
Baseline Controls					X	
Income Quintiles		X		X	X	
Lagged Baseline Controls						X
Observations	2204	2204	2204	2204	2204	1949
R^2	0.6175	0.6821	0.6687	0.6821	0.7090	0.7707

Notes: Robust standard errors clustered at the county level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Dependent variable: Judgments per 100 individuals. All regressions are weighted by population and estimated on the common support sample.

Table 54: Summary Statistics (NJ and IL Sample)

	Black	White	t-test
	mean/sd	mean/sd	b
Judgments	2.81	2.10	-0.70***
	(2.05)	(2.24)	
Median Household Income	41.37	51.64	10.26***
	(15.31)	(16.85)	
Median Credit Score	605.52	644.63	39.10***
	(25.99)	(39.71)	
GINI Index	0.46	0.44	-0.01***
	(0.05)	(0.06)	
90+ DPD Debt Balances	7803.14	6752.62	-1050.52
	(7445.88)	(12317.03)	
Unemployment Rate	0.12	0.09	-0.03***
	(0.02)	(0.03)	
Median House Value (000s)	161.47	201.05	39.58***
	(65.88)	(107.93)	
Fraction with Bachelors Degree	0.19	0.25	0.07***
	(0.10)	(0.17)	
Fraction without High School Degree	0.18	0.14	-0.04***
	(0.08)	(0.08)	
Observations	224	596	820

Notes: Summary statistics for observations on the common support sample. Data is drawn from NJ & IL.

Table 55: Judgments, Income, and Credit Scores (NJ and IL Data)

	(1)	(2)	(3)	(4)	(5)	(6)
Black Majority: ZIP	0.8525^{***}	0.6745^{***}	0.6049^{***}	0.6338^{***}	0.4700^{***}	0.4453^{***}
	(0.1294)	(0.1269)	(0.1601)	(0.1532)	(0.1372)	(0.1436)
Median Household Income		0.0625^{*}		0.0609	0.0331	
		(0.0370)		(0.0397)	(0.0372)	
Median Credit Score			0.0056	0.0086	0.0066	
			(0.0072)	(0.0072)	(0.0068)	
County Fixed Effects	X	X	X	X	X	X
Year Fixed Effects	X	X	X	X	X	X
Baseline Controls	X	X	X	X	X	
Lagged Baseline Controls						X
Observations	820	820	820	820	820	579
R^2	0.9268	0.9406	0.9329	0.9412	0.9472	0.9539

Notes: Robust standard errors clustered at the county level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Observations on the common support sample from NJ and IL; estimated by a logistic regression, ran separately for each year. Dependent variable: Judgments per 100 individuals. All regressions are weighted by population and estimated on the common support sample.

Table 56: Summary Statistics (Individual Judgment Data)

	Black	White	t-test
	mean/sd	mean/sd	b
Judgments	0.14	0.07	-0.07^{***}
	(0.47)	(0.36)	
Household Income	60.28	83.75	23.46^{***}
	(41.12)	(57.67)	
Credit Score	611.93	683.13	71.20^{***}
	(107.69)	(107.84)	
Observations	399723	6722540	7122263

Notes: Summary statistics for observations on the common support sample. Data is a 1% representative sample of the U.S. for individuals with a credit report.

Table 57: Judgments, Income, and Credit Scores (Individual Judgment Data)

	(1) Judgments	(2) Judgments	(3) Judgments	(4) Judgments	(5) Judgments
Black Majority	0.0689**	0.0559***	0.0148*	0.0158*	0.0162*
	(0.0012)	(0.0006)	(0.0014)	(0.0015)	(0.0013)
Household Income		-0.0006***		0.0002***	0.0002***
		(0.0000)		(0.0000)	(0.0000)
Credit Score			-0.0008***	-0.0008***	-0.0008***
			(0.0000)	(0.0000)	(0.0000)
County Fixed Effects	X	X	X	X	X
Quarter Fixed Effects	X	X	X	X	X
Baseline Controls					X
Observations	7122263	7122263	7122263	7122263	7105641
R^2	0.0026	0.0102	0.0530	0.0541	0.0567

Notes: Robust standard errors clustered at the county level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$ Observations on the common support sample. Data is a 1% representative sample of the U.S. for individuals with a credit report. Dependent variable: current judgments.

Table 58: Judgments and Debt Portfolios (Individual Judgment Data)

	(1) Judgments	(2) Judgments	(3) Judgments	(4) Judgments	(5) Judgments
Black Majority	0.0125*	0.0136*	0.0157*	0.0158*	0.0126*
	(0.0015)	(0.0020)	(0.0014)	(0.0014)	(0.0016)
County Fixed Effects	X	X	X	X	X
Quarter Fixed Effects	X	X	X	X	X
Debt Levels	Yes				Yes
Payment and Utilization		Yes			Yes
Debt Composition			Yes		Yes
Delinquency				Yes	Yes
Observations	7122263	7122263	7122263	7122263	7122263
R^2	0.0601	0.0580	0.0541	0.0629	0.0691

Notes: Robust standard errors clustered at the county level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$ Observations on the common support sample. Data is a 1% representative sample of the U.S. for individuals with a credit report. Dependent variable: current judgments.

Table 59: Summary Statistics (Full Sample)

	Black	White	t-test
	mean/sd	mean/sd	b
Panel A: Judgments			
Judgments per 100 People	2.76	1.25	-1.51***
	(1.33)	(0.82)	
Share of Default Judgments	0.45	0.37	-0.08***
	(0.07)	(0.14)	
Share of Consent Judgments	0.16	0.17	0.01
	(0.07)	(0.12)	
Share of Contested Judgments	0.06	0.05	-0.01*
	(0.04)	(0.08)	
Share w/ Attorney	0.04	0.10	0.07***
	(0.02)	(0.09)	
Panel B: Credit Variables			
Median Credit Score	603.63	680.50	76.86***
	(37.89)	(66.59)	
90+ DPD Debt Balances	3527.08	1556.61	-1970.47***
	(2850.97)	(4629.70)	
Banks (5 miles)	87.91	13.74	-74.17***
	(39.65)	(33.24)	
Payday Lenders (5 miles)	30.74	3.72	-27.01***
	(9.27)	(8.27)	
Panel C: Census Data			
Median Household Income (000s)	31.41	45.91	14.50***
	(11.73)	(17.10)	
GINI Index	0.46	0.40	-0.05***
	(0.05)	(0.06)	
Unemployment Rate	0.12	0.05	-0.06***
	(0.03)	(0.03)	
Divorce Rate	0.13	0.12	-0.01***
	(0.02)	(0.05)	
Median House Value	86.47	116.08	29.61***
	(35.28)	(67.95)	
Fraction with Bachelors Degree	0.16	0.18	0.01*
	(0.10)	(0.13)	
Fraction without High School Degree	0.19	0.15	-0.04***
	(0.06)	(0.08)	
Observations	248	7865	8113

Notes: Summary statistics for observations for the entire Missouri sample.

Table 60: Judgments, Income, and Credit Scores (Full Sample)

	(1)	(2)	(3)	(4)	(5)	(6)
Black Majority: ZIP	1.5063***	1.0203***	1.0202***	0.8521***	0.7632***	0.8334***
	(0.1460)	(0.1088)	(0.1016)	(0.0757)	(0.1082)	(0.0324)
Median Household Income		-0.0101		-0.0046	-0.0012	
		(0.0091)		(0.0069)	(0.0071)	
Median Credit Score			-0.0002	-0.0005	0.0003	
			(0.0018)	(0.0019)	(0.0020)	
County Fixed Effects	X	X	X	X	X	X
Year Fixed Effects	X	X	X	X	X	X
Baseline Controls					X	
Income Quintiles		X		X	X	
Credit Quintiles			X	X	X	
Lagged Baseline Controls						X
Observations	8113	8113	8113	8113	8113	7197
R^2	0.5884	0.6545	0.6367	0.6660	0.6761	0.6835

Notes: Robust standard errors clustered at the county level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Observations for the entire Missouri sample. Dependent variable: Judgments per 100 individuals. All regressions are weighted by population.

Table 61: Judgments and Other Measures of Racial Composition (Full Sample)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Share Black: ZIP	1.6431*** (0.1037)						
Black Majority: ZIP		0.7632*** (0.1082)					
Share Black: BISG			1.5541*** (0.1189)				5.1912*** (1.3026)
Black Majority: BISG				0.7144*** (0.0950)		0.4766** (0.2131)	
Share Black: Names					4.5415*** (0.5285)		
Median Household Income	0.0012 (0.0057)	-0.0012 (0.0071)	0.0037 (0.0056)	0.0037 (0.0059)	0.0048 (0.0060)		
Median Credit Score	0.0010 (0.0018)	0.0003 (0.0020)	0.0009 (0.0018)	-0.0004 (0.0019)	-0.0003 (0.0017)	-0.0001 (0.0022)	0.0002 (0.0022)
County Fixed Effects	X	X	X	X	X		
ZIP Code Fixed Effects						X	X
Year Fixed Effects	X	X	X	X	X	X	X
Credit Quintiles	X	X	X	X	X	X	X
Baseline Controls	X	X	X	X	X		
Observations	8113	8113	8106	8106	8106	8106	8106
R^2	0.6926	0.6761	0.6951	0.6802	0.6805	0.7491	0.7530

Notes: Robust standard errors clustered at the county level are in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Observations for the entire Missouri sample. Dependent variable: Judgments per 100 individuals. All regressions are weighted by population.

Table 62: Judgments and Other Demographic Groups (Full Sample)

	(1)	(2)	(3)	(4)	(5)	(6)
Share Black: ZIP	2.4034***	1.8654***	1.9266***	1.6624***	1.6533***	1.7849***
	(0.1141)	(0.1480)	(0.0909)	(0.1015)	(0.1003)	(0.0545)
Share Asian: ZIP	-4.6405***	-2.2187**	-3.9005***	-2.1501**	-0.6058	-0.6912*
	(0.9563)	(0.8761)	(1.0624)	(0.9451)	(0.7972)	(0.4098)
Share Hispanic: ZIP	1.5121***	0.2103	0.9982***	0.1223	0.2984	0.3032*
	(0.3017)	(0.2648)	(0.2462)	(0.2330)	(0.2363)	(0.1805)
Median Household Income		-0.0038		-0.0008	0.0013	
		(0.0062)		(0.0056)	(0.0056)	
Median Credit Score			0.0007	0.0002	0.0010	
			(0.0016)	(0.0017)	(0.0018)	
County Fixed Effects	X	X	X	X	X	X
Year Fixed Effects	X	X	X	X	X	X
Baseline Controls					X	
Income Quintiles		X		X	X	
Credit Quintiles			X	X	X	
Lagged Baseline Controls						X
Observations	8113	8113	8113	8113	8113	7197
R^2	0.6577	0.6814	0.6717	0.6860	0.6927	0.7029

Notes: Robust standard errors clustered at the county level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Observations for the entire Missouri sample. Dependent variable: Judgments per 100 individuals. All regressions are weighted by population.

Table 63: Model Comparison

Model	RMSE: All	RMSE: Black	RMSE: White
Linear Model	0.8079	1.2704	0.7809
Non-linear Model	0.6294	0.9556	0.6106

Notes: We split our sample into training and test dataset according to an 80-20 split. Then, we fit both of our models on the training dataset, and using the parameter estimates obtained from the training dataset, we compute fitted values on the test dataset. Then we compute the residual and obtain the corresponding Root Mean Square Error (RMSE). We repeat this procedure 10,000 times with distinct train-test splits and the statistics reported are the averages obtained from this exercise.

Bibliography

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[2] Sumit Agarwal, Souphala Chomsisengphet, Neale Mahoney, and Johannes Stroebel. Regulating consumer financial products: Evidence from credit cards. *The Quarterly Journal of Economics*, 130(1):111–164, 2014.

[3] Sumit Agarwal, Souphala Chomsisengphet, Neale Mahoney, and Johannes Stroebel. Do banks pass through credit expansions to consumers who want to borrow? *The Quarterly Journal of Economics*, 133(1):129–190, 2018.

[4] Prottoy A Akbar, Sijie Li, Allison Shertzer, and Randall P Walsh. Racial segregation in housing markets and the erosion of black wealth. Technical report, National Bureau of Economic Research, 2019.

[5] Stefania Albanesi, Giacomo De Giorgi, and Jaromir Nosal. Credit growth and the financial crisis: A new narrative. Working Paper 23740, National Bureau of Economic Research, August 2017.

[6] Kate Antonovics and Brian Knight. A new look at racial profiling: evidence from the boston police department. *Review of Economics and Statistics*, 91:163–177, 2009.

[7] David Arnold, Will Dobbie, and Crystal S. Yang. Racial bias in bail decisions. *Quarterly Journal of Economics*, pages 1885–1932, 2018.

[8] Elizabeth Asiedu, James A Freeman, and Akwasi Nti-Addae. Access to credit by small businesses: How relevant are race, ethnicity, and gender? *American Economic Review*, 102(3):532–37, 2012.

[9] Susan Athey and Guido W Imbens. Machine learning methods that economists should know about. *Annual Review of Economics*, 11, 2019.

[10] Kartik Athreya, Xuan S Tam, and Eric R Young. A quantitative theory of information and unsecured credit. *American Economic Journal: Macroeconomics*, 4(3):153–83, 2012.

[11] Lawrence M Ausubel. The failure of competition in the credit card market. *The American Economic Review*, pages 50–81, 1991.

[12] Luca Barbaglia, Sebastiano Manzan, and Elisa Tosetti. Forecasting loan default in europe with machine learning. *Available at SSRN 3605449*, 2020.

[13] James R. Barth, Jitka Hilliard, and Yanfei Sun. Do state regulations affect payday lender concentration? *Journal of Economics and Business*, 84:14–29, 2016.

[14] Robert Bartlett, Adair Morse, Richard Stanton, and Nancy Wallace. Consumer-lending discrimination in the fintech era. Technical report, National Bureau of Economic Research, 2019.

[15] Yoshua Bengio, Yann LeCun, et al. Scaling learning algorithms towards ai. *Large-scale kernel machines*, 34(5):1–41, 2007.

[16] James S Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In *Advances in neural information processing systems*, pages 2546–2554, 2011.

[17] Marianne Bertrand and Sendhil Mullainathan. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review*, 94(4):991–1013, 2004.

[18] Neil Bhutta. Payday loans and consumer financial health. *Journal of Banking and Finance*, 47:230–242, 2014.

[19] Michael G Bradley, Amy Crews Cutts, and Wei Liu. Strategic mortgage default: The effect of neighborhood factors. *Real Estate Economics*, 43(2):271–299, 2015.

[20] Nicola Branzoli and Ilaria Supino. Fintech credit: A critical review of empirical research literature. *Bank of Italy Occasional Paper*, (549), 2020.

[21] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[22] Consumer Financial Protection Bureau. Using publicly available information to proxy for unidentified race and ethnicity. 2014.

[23] Florentin Butaru, Qingqing Chen, Brian Clark, Sanmay Das, Andrew W Lo, and Akhtar Siddique. Risk and risk management in the credit card industry. *Journal of Banking & Finance*, 72:218–239, 2016.

[24] Paul S Calem and Loretta J Mester. Consumer behavior and the stickiness of credit-card interest rates. *The American Economic Review*, 85(5):1327–1336, 1995.

[25] Colleen Casey, Davita Silfen Glasberg, and Angie Beeman. Racial disparities in access to mortgage credit: Does governance matter? *Social Science Quarterly*, 92(3):782–806, 2011.

[26] Satyajit Chatterjee, Dean Corbae, Kyle Dempsey, and Jose-Victor Rios-Rull. A theory of credit scoring and competitive pricing of default risk. *Unpublished paper, University of Wisconsin*, 31, 2016.

[27] Satyajit Chatterjee, Dean Corbae, Makoto Nakajima, and Jose-Victor Rios-Rull. A quantitative theory of unsecured consumer credit with risk of default. *Econometrica*, 75(6):1525–1589, 2007.

[28] Hugh Chen, Scott Lundberg, and Su-In Lee. Hybrid gradient boosting trees and neural networks for forecasting operating room data. *arXiv preprint arXiv:1801.07384*, 2018.

[29] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[30] Ing-Haw Cheng, Felipe Severino, and Richard R. Townsend. How do consumers fare when dealing with debt collectors? evidence from out-of-court settlements. *Working Paper*, 2019.

[31] François Chollet et al. Keras: Deep learning library for theano and tensorflow. *URL: https://keras. io/k*, 7(8), 2015.

[32] Ethan Cohen-Cole. Credit card redlining. *Review of Economics and Statistics*, 93(2):700–713, 2011.

[33] Flynn Coleman. *A Human Algorithm: How Artificial Intelligence is Redefining who We are*. Counterpoint Press, 2019.

[34] Dean Corbae and Andrew Glover. Employer credit checks: Poverty traps versus matching efficiency. Working Paper 25005, National Bureau of Economic Research, September 2018.

[35] Andrew Cotter, Heinrich Jiang, Maya R Gupta, Serena Wang, Taman Narayan, Seungil You, and Karthik Sridharan. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *Journal of Machine Learning Research*, 20(172):1–59, 2019.

[36] Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009.

[37] Lukasz A. Drozd and Serrano-Padial. Modeling the revolving revolution: The debt collection channel. *American Economic Review*, 107(3):897–930, 2017.

[38] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[39] Marc N Elliott, Peter A Morrison, Allen Fremont, Daniel F McCaffrey, Philip Pantoja, and Nicole Lurie. Using the census bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology*, 9(2):69, 2009.

[40] Viktar Fedaseyeu. Debt collection agencies and the supply of consumer credit. *Federal Reserve Bank of Philadelphia Working Paper*, 15-23, 2015.

[41] AM Fremont, AS Bierman, SL Wickstrom, CE Bird, MM Shah, JJ Escarce, and TS Rector. Use of indirect measures of race/ethnicity and socioeconomic status in managed care settings to identify disparities in cardiovascular and diabetes care quality. *Health Aff*, 24(2):516–526, 2005.

[42] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[43] Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. Predictably unequal? the effects of machine learning on credit markets. *The Effects of Machine Learning on Credit Markets (November 6, 2018)*, 2018.

[44] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[45] Akhil Gupta, Naman Shukla, LLC Deepair, Lavanya Marla, Arinbjörn Kolbeinsson, and Kartik Yellepeddi. How to incorporate monotonicity in deep networks while preserving flexibility? 2020.

[46] Maya R. Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. Proxy fairness. *ArXiv*, abs/1806.11212, 2018.

[47] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[48] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251 – 257, 1991.

[49] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359 – 366, 1989.

[50] Robert Hunt. Collecting consumer debt in america. *Business Review*, Q2:11–24, 2007.

[51] John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(3):90–95, 2007.

[52] R. M. Hynes. Broke but not bankrupt: Consumer debt collection in state courts. *Florida Law Review*, 60(1), 2008.

[53] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[54] Maurice Jourdain-Earl. The foreclosure crisis and racial disparities in access to mortgage credit 2004-2009. *Compliance Tech, Arlington VA*, 2011.

[55] Amir E Khandani, Adlar J Kim, and Andrew W Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787, 2010.

[56] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[57] Ludmila I Kuncheva and Christopher J Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 51(2):181–207, 2003.

[58] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in neural information processing systems*, pages 4066–4076, 2017.

[59] Håvard Kvamme, Nikolai Sellereite, Kjersti Aas, and Steffen Sjursen. Predicting mortgage default using convolutional neural networks. *Expert Systems with Applications*, 102:207–217, 2018.

[60] Stefan Lessmann, Bart Baesens, Hsin-Vonn Seow, and Lyn C Thomas. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124–136, 2015.

[61] Igor Livshits, James MacGee, and Michele Tertilt. Consumer bankruptcy: A fresh start. *American Economic Review*, 97(1):402–418, 2007.

[62] John R Logan. Ethnic diversity grows, neighborhood integration lags. *Redefining urban and suburban America: Evidence from Census*, 1:235–255, 2000.

[63] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[64] Signe-Mary McKernan, Caroline Ratcliffe, Margaret Simms, and Sisi Zhang. Do racial disparities in private transfers help explain the racial wealth gap? new evidence from longitudinal data. *Demography*, 51(3):949–974, 2014.

[65] Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX, 2010.

[66] Christoph Molnar. *Interpretable Machine Learning*. 2019. `https://christophm.github.io/interpretable-ml-book/`.

[67] Guido Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 2924–2932, Cambridge, MA, USA, 2014. MIT Press.

[68] Mirko Moscatelli, Fabio Parlapiano, Simone Narizzano, and Gianluca Viggiano. Corporate default forecasting with machine learning. *Expert Systems with Applications*, page 113567, 2020.

[69] Sendhil Mullainathan and Jann Spiess. Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, 2017.

[70] Emily Oster. Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204, 2019.

[71] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[72] Caroline Ratcliffe, Signe-Mary McKernan, Brett Theodos, Emma Kalish, John Chalekian, Peifang Guo, and Christopher Trepel. Delinquent debt in america. *Urban Institute Opportunity and Ownership Initiative Brief*, 2014.

[73] Xudie Ren, Haonan Guo, Shenghong Li, Shilin Wang, and Jianhua Li. A novel image classification method with cnn-xgboost model. In *International Workshop on Digital Watermarking*, pages 378–390. Springer, 2017.

[74] US Federal Reserve. Report to the congress on credit scoring and its effects on the availability and affordability of credit. *Board of Governors of the Federal Reserve System*, 2007.

[75] Joseph A Ritter and Lowell J Taylor. Racial disparity in unemployment. *The Review of Economics and Statistics*, 93(1):30–42, 2011.

[76] Chris Russell, Matt J Kusner, Joshua Loftus, and Ricardo Silva. When worlds collide: integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems*, pages 6414–6423, 2017.

[77] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[78] Justin Sirignano, Apaar Sadhwani, and Kay Giesecke. Deep learning for mortgage risk. *Available at SSRN*, 2018.

[79] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[80] Margery Austin Turner, Stephen L Ross, George C Galster, John Yinger, et al. Discrimination in metropolitan housing markets: National results from phase i hds 2000. *Washington, DC: The Urban Institute*, 2002.

[81] Domonkos F Vamossy. Investor emotions and earnings announcements. *Available at SSRN 3626025*, 2020.

[82] Stéfan van der Walt, S Chris Colbert, and Gael Varoquaux. The numpy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22–30, 2011.

[83] David West. Neural network credit scoring models. *Computers & Operations Research*, 27(11-12):1131–1152, 2000.

[84] Yufei Xia, Chuanzhe Liu, YuYing Li, and Nana Liu. A boosted decision tree approach using bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications*, 78:225–241, 2017.

[85] Seungil You, David Ding, Kevin Canini, Jan Pfeifer, and Maya Gupta. Deep lattice networks and partial monotonic functions. In *Advances in neural information processing systems*, pages 2981–2989, 2017.

[86] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970, 2017.