NBER WORKING PAPER SERIES

MACHINE LEARNING FOR SOLAR ACCESSIBILITY: IMPLICATIONS FOR LOW-INCOME SOLAR EXPANSION AND PROFITABILITY

Sruthi Davuluri
René García Franceschini
Christopher R. Knittel
Chikara Onda
Kelly Roache

Working Paper 26178
http://www.nber.org/papers/w26178

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2019

As is customary in economics, the authors are listed alphabetically. This paper has benefited from conversations with the Solstice Initiative. Financial support from the Department of Energy is gratefully acknowledged. Onda's work was supported by the Kimmelman Family E-IPER Fellowship and the Satre Family Fellowship during his doctoral work at Stanford University. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Machine Learning for Solar Accessibility: Implications for Low-Income Solar Expansion and Profitability

Sruthi Davuluri, René García Franceschini, Christopher R. Knittel, Chikara Onda, and Kelly Roache

NBER Working Paper No. 26178
September 2019
JEL No. C53,L11,L94,Q2

ABSTRACT

The solar industry in the US typically uses a credit score such as the FICO score as an indicator of consumer utility payment performance and credit worthiness to approve customers for new solar installations. Using data on over 800,000 utility payment performance and over 5,000 demographic variables, we compare machine learning and econometric models to predict the probability of default to credit-score cutoffs. We compare these models across a variety of measures, including how they affect consumers of different socio-economic backgrounds and profitability. We find that a traditional regression analysis using a small number of variables specific to utility repayment performance greatly increases accuracy and LMI inclusivity relative to FICO score, and that using machine learning techniques further enhances model performance. Relative to FICO, the machine learning model increases the number of low-to-moderate income consumers approved for community solar by 1.1% to 4.2% depending on the stringency used for evaluating potential customers, while decreasing the default rate by 1.4 to 1.9 percentage points. Using electricity utility repayment as a proxy for solar installation repayment, shifting from a FICO score cutoff to the machine learning model increases profits by 34% to 1882% depending on the stringency used for evaluating potential customers.

Sruthi Davuluri
Center for Energy and Environmental Policy Research
77 Massachusetts Avenue
Cambridge, MA 02139
sruthi@mit.edu

René García Franceschini
77 Massachusetts Ave Cambridge,
MA 02139 ragarcia@mit.edu

Christopher R. Knittel
MIT Sloan School of Management
100 Main Street, E62-513
Cambridge, MA 02142
and NBER
knittel@mit.edu

Chikara Onda
Stanford University
Stanford, CA 94305
conda@stanford.edu

Kelly Roache
Energy and Policy Institute
kellyroache@gmail.com

1 Introduction

Most solar companies currently use credit scores to determine whom to approve for solar installations. Despite their widespread use, credit scores consider many aspects of a consumer's credit history that are not directly related to utility payment; therefore, the FICO score is an imperfect proxy for predicting utility payment performance. Furthermore, approximately 5 million low-income consumers are credit invisible or have unscored records, representing 45% of consumers in low-income neighborhoods [5]. This implies that traditional credit score cutoffs exclude people with low credit scores and those with insufficient credit history. Simultaneously, Low-to-Moderate Income (LMI) households bear a disproportionate energy burden, paying on average three times as much for energy as wealthier households [6]. Thus, by depending on credit scores as the sole indicator of consumer payment performance, the community solar market reproduces existing inequalities and limits its own potential for growth by excluding potential consumers.

The goal of this research is: (1) to develop an alternative prediction model of default based on machine learning algorithms, specifically LASSO, SVM, and random forests; and (2) to compare its overall forecasting performance, as well as its implications for LMI consumers, to traditional credit metrics. We do so by developing a model that predicts the probability of non-delinquency of utility bill payments using a large data set of utility repayment and other financial data obtained from a credit reporting agency (CRA). We find that a traditional regression analysis using a small number of variables specific to utility repayment performance greatly increases accuracy and LMI inclusivity relative to FICO score, and that using machine learning techniques further enhances model performance. Our preferred model increases the number of LMI applicants approved by 1.1% to 4.2% depending on the stringency used in evaluating potential customers, while decreasing the default rate by 1.4 to 1.9 percentage points. Our analysis shows that it is possible to extend solar to a larger number of qualified applicants with lower or no credit scores, while at the same time decreasing default risk, thus opening access to an untapped, low-risk market segment.

The paper proceeds as follows. Section 2 provides a broad review of the community shared solar (CSS), its current qualifying mechanism, and of the use of alternative credit qualifying scores across various industries. Section 3 describes the data set and data processing. Section 4 outlines the models underlying the prediction models. In particular, we use traditional regression methods and machine learning techniques on account-level payment performance, financial, and demographic data to predict the probability of delinquency. Section 5 assesses the models developed in Section 4 by comparing to FICO based on accuracy, default rates, and LMI inclusion. Section 6 explains the analysis of profitiability, which is followed by the Conclusion in Section 8.

2 Literature Review

2.1 Community Shared Solar

Community shared solar provides a solution to expand solar access to consumers currently locked out of the rooftop solar market. In a community shared solar project, individuals subscribe to an off-site solar farm from which they receive credits on their electricity bill. This model is particularly attractive for those who have explored rooftop solar but are not eligible. Approximately 80% of Americans are currently locked out of the solar market. This includes renters, households with unsuitable roofs, and those not able to afford the high cost of installing rooftop panels. Through community solar, customers have access to renewable energy

and savings without needing to invest in rooftop solar. According to industry estimates, the community solar market is expected to boom in the coming years, with community solar capacity in the US expected to reach as much as 11,000 MW by 2020, compared to 250 MW in early 2017, and just 1 MW of production in 2014 [4, 7, 9, 10].

2.2 Credit Score Requirements for Community Shared Solar

Community solar has the potential to expand renewable energy access to a much wider demographic than rooftop solar. However, the community solar market is still developing and thus subject to considerable uncertainty within the financial community. In many cases, financiers require that solar developers vet customer credit scores in order to mitigate perceived subscription payment risk. Oftentimes, developers set a minimum score of 700 on the FICO scale. These high credit requirements exclude a significant portion of the population, yet there is little evidence that FICO scores can accurately predict defaults on community solar payments in the same way as they predict defaults on loans.

Additionally, the direct correlation between credit scores and income results in the disproportionate exclusion of LMI households from the community solar market [8]. Credit scores are developed for consumers actively participating in the banking and credit system, which naturally favors higher-income consumers. While the exact formulas for calculating credit scores are industry secrets, the score is determined based on five categories of information: 1) payment history, 2) utilization ratio (the amount owed vs. the individual's maximum credit limit), 3) length of credit history, 4) recent activity, and 5) how much debt remains unpaid [26]. Data for each of these categories is collected from a variety of types of credit, including mortgages, credit cards, auto loans, student loans, etc. A lack of ability to engage with these systems leads to credit scores that are often inadequate to participate in community solar. This affects the 56% of American consumers who have subprime credit scores [3].

While some are excluded from mainstream credit because their credit score is too low, many are denied access because their credit score is nonexistent. In order to investigate groups excluded from mainstream credit, the Consumer Financial Protection Bureau has defined the terms *credit invisibles* and *credit unscorables*; credit invisibles include individuals without any records with national credit rating agencies, and unscorables include those with thin credit files or stale records [5]. The same organization estimates that in 2010, 26 million Americans were credit invisible while an additional 19.4 million were unscorable [5].

LMI households are also disproportionately more likely to be unscored than their wealthier counterparts. Nearly 50% of low-income consumers and 30% of moderate income consumers are unscored, compared to only 10% of upper income consumers [5]. Lenders generally consider consumers without credit scores to be high risk [8]. This means that many customers without credit scores are subject to predatory lenders who charge both high interest rates and high penalty fees.

The exclusion of LMI households from community solar is even more impactful as these households stand to benefit the most from subscription-model community solar. A 2016 report cited that the median energy burden for households with less than 80% of their area median income was 7.2%, while non low-income households had a median energy burden of 2.3% [6]. In other words, the energy burden among LMI households is disproportionately higher than the total population. Worse yet, there is evidence that LMI households could be included in community-solar projects without additional risk to the investors of these developments.

2.3 Utility Bills as Proxies for Community Shared Solar Payments

This paper hypothesizes that FICO scores and other traditional credit score indicators are an imperfect predictor of community solar payments, and that utility payment history can better predict the risk of community solar payment default. This study has chosen to use utility payment history rather than community solar subscription payment history for several reasons. Since the market is still developing, there is both limited historical data on community solar payments and inherent selection bias in the existing data. The selection bias stems from the existing high FICO requirements that make it impossible to assess repayment rates of households with lower FICO scores. Therefore, this analysis draws from the assumption that utility payment history can serve as a proxy for community solar subscription payments.

Since community solar payments and utility payments are generally similar in amounts, we hypothesize utility and community solar payments will be adequate proxies for one another. Energy spending is a necessary good for most consumers, meaning that it tends be one of the first household expenses to be paid. Status quo bias suggests that customer prioritization of electric utility bills will extend to community solar energy bills as well.

Another argument that supports the use of utility payment history as a proxy for community solar is the potential for bill consolidation. A few states with emerging community solar markets are considering legislation that would consolidate utility and community solar subscription bills. If community solar charges appear on a customer's utility bill, then consumers will treat utility bills and community solar bills exactly the same. Therefore, for the purposes of this study, we assume a close proxy relationship between community solar subscription payments and electric utility bill payments.

2.4 Alternative Credit Metrics in Industry

Alternative credit scoring mechanisms would provide value in other industries as well, such as student loans, vehicle purchases, mortgage applications, credit card applications, and a number of other industries which rely on the existing FICO credit score. Incorporating alternative data can generate credit scores for those currently without scores. For example, LexisNexis has developed the RiskView Score, an alternative credit metric, which scored nearly 10% of the sample that did not have a score previously [25]. Another alternative credit metric, Link2Credit, created scores for 19 million previously unscored records. In 2012, the Policy and Economic Research Council (PERC) conducted a study on the impact of alternative data on credit scores using both non-financial tradeline data and utility data. The study found that 74% of sampled customers that were previously unscorable could be scored using alternative data [28]. Alternative data can therefore create a creditworthiness metric to extend credit to those without scores.

In addition to creating scores for the unscored, alternative data increases the efficacy and precision of traditional credit scoring. The RiskView Score improved the segmentation of consumers within credit ranges, allowing for expanded and more precise lending. LexisNexis used a cross section of traditional credit scores and the RiskView Score to determine which consumers within each range of credit scores were higher risk borrowers than others [8, 25]. Alternative data improves the precision with which credit rating agencies can measure creditworthiness, which in turn can extend credit beyond traditional scoring boundaries without negatively impacting bill payment rates. A 2015 PERC study found that non-financial utility and telecom delinquencies were predictive of future mortgage, bank card, and public record delinquencies [27]. Empirically, alternative data has successfully predicted financial default.

Several national credit agencies have created products using alternative data. The aforementioned Link2Credit score uses phone payment history and other public record metrics, while Fair Isaac developed a FICO expansion score including debit data, utility data, and public record attributes [25]. Equifax marketed their Advanced Energy Plus score to use energy payment data to augment thin file consumers' credit history. Alternative credit metrics are more useful if they are widely trusted and usable in the finance community. While the array of alternative credit products does not signify widespread use, it certainly signals market interest and credibility of such products.

2.5 Alternative Credit Metrics in Academia

In addition to the industry-led initiatives, there has been other research in academia exploring alternative credit scoring mechanisms for various other purposes beyond the solar industry. There has been literature which uses regression discontinuity to display the moral hazard effect induced when private lenders employ strict FICO Score cutoffs [12, 14, 16, 18, 24]. In other words, private lenders are more likely to offer services to customers with a FICO score just above a certain threshold than customers below the same threshold. In an analysis of subprime mortgage loan contracts in the United States, Keys et. al show that such securitization practices adversely affects the incentives for lenders to carefully screen borrowers [15].

To address the issue, a number of researchers and academics have used statistics and machine learning to provide an alternative credit scoring mechanism. Nikravesh uses fuzzy query and ranking as a method of predicting the default risk associated with lending to a new customer, and to serve as an alternative to the FICO score [22]. Yu et. al propose a multistage neural network ensemble learning model to predict credit risk [20]. Huang et. al investigate a data mining approach with support vector machines as a credit scoring model, which required a long training time [11]. Wang et. al experiment with fuzzy SVMs and traditional SVMs for predicting credit risk to show that the fuzzy SVM achieves better generalizability by being less sensitive to outliers than alternative machine learning methods [32]. Antonakis et. al analyzes the predictive ability of several machine learning approaches, including Naïve Bayes Rule, k-Nearest Neighbors, classification trees, and neural networks, for screening credit applicants [2]. Khandani et. al use generalized classification and regression trees to classify the rates of credit-card holder delinquencies and defaults, and use their results to study nonlinear relationships that are not captured by traditional credit scores [17]. Wang et. al demonstrate the feasibility of using bagging and random subspace, together with Support Vector Machines, as an alternative method to predict credit risk assessment [29, 30]. Wang et. al also compared the predictive ability of logistic regression analysis (LRA), linear discriminant analysis (LDA), multi-layer perceptron (MLP), and radial basis function network (RBFN), with decision trees with and without bagging as alternative methods of credit scoring [31]. Their decision trees model demonstrated some of the lowest performance ratings due to noise, while this paper found a decision trees model to be the most accurate. Finally, Kruppa et. al also demonstrated the accuracy of random forest, k-Nearest Neighbors, and bagged k-Nearest Neighbors to predict consumer credit risks [19]. Kruppa et. al found results consistent with those presented in this paper, demonstrating that the random forest algorithm showed higher accuracy rates than the alternative methods tested. In addition to those listed, there has been other research in using machine learning techniques to assess credit risk. However, the other literature was focusing on credit risk from a general perspective, and did not identify the impacts for lower-income customers or on the solar industry specifically.

	Testing sample	US average
Income (median)	\$55,000-\$59,999	\$55,322
College	19.3%	30.3%
Female	26.6%	50.8%
Black	10.5%	13.3%
Hispanic	8.4%	17.8%

Table 1: Descriptive statistics: demographic variables

3 Data

3.1 Significance of Data

This study uses account-level credit score and monthly payment performance between December 2009 and November 2016, obtained from a credit reporting agency (CRA) along with other financial and demographic data. Because we are interested in using the data to predict payment performance in the last 12 months of the data, we use records with at least 24 months of consecutive utility payment performance data in the period (December 2014 to November 2016).

The full universe of data from the CRA include 8.3 million records, of which we procured the 10.6% (872,382) with 24 consecutive months of payment history for an individual utility account. Of those individuals with a full history, 61.1% (535,931) have no negative record and 38.9% (341,372) have at least one negative record. Here, we define a negative record as any delinquency of at least 30 days. It is important to note that utilities are more likely to report a delinquent account, and therefore that such accounts may be over-represented in the set of accounts with 24 months of consecutive payment data.[1]

3.2 Descriptive Statistics

In addition to payment history, we use demographic data, including features such as home ownership, length of residence, level of education, and age. In order to see whether the sample differs from the population of U.S. utility account holders, we compare to national averages from the Census in Table 1. We see that the sample is more or less representative in terms of annual income, but it under-represents women and minorities. However, it is important to note that the utility account holders across the United States will most likely differ from the entire U.S. population.

Looking across geographies in Table 2, a few observations bear mentioning. First, we see that urban, suburban, and rural households are all well-represented in the sample. Looking across regions, however, we see that the majority of observations come from the East North Central region (82.6%). Most of these observations are from Wisconsin (74.1%), although this is of the 64.3% of the sample that report the state of residence. Though this may lead to some concern over the external validity of this study, this would only affect the accuracy of the alternative scoring mechanisms if Wisconsinites systematically differ from the rest of the country in terms of the relationship between past and future payment performance. Ther eis no intuitive explanation as to why this should be the case, and we confirm this further below in Table 4

[1]Since we do not have access to the full universe of Experian data, we have also constructed a sample requiring 36 months of data to illustrate that the effect of restricting the data in this manner, assuming that moving from an unrestricted sample to the sample requiring 24 months of data has a similar effect to moving from this sample to an even more restrictive sample requiring 36 months of data.

Population density		
	%	N
Rural areas	26.9	234,181
Smaller suburbs and towns	38.5	335,960
City and surrounds	34.6	302,047
Census division		
	%	N
New England	2.6	14,710
Middle Atlantic	2.6	14,309
East North Central	82.6	463,04
West North Central	0.9	4,992
South Atlantic	8.3	46,367
East South Central	0.7	3,814
West South Central	0.3	1,626
Mountain	0.6	3,557
Pacific	1.5	8,314

State					
	%	N		%	N
Alabama	0.2	915	*Montana*	0.0	47
Alaska	0.0	35	*Nebraska*	0.1	292
Arizona	0.2	918	*Nevada*	0.1	405
Arkansas	0.0	117	*New Hampshire*	0.0	90
California	0.3	1,640	*New Jersey*	0.1	560
Colorado	0.1	435	*New Mexico*	0.1	738
Connecticut	2.5	13,942	*New York*	1.7	9,505
Delaware	0.0	99	*North Carolina*	0.9	5,201
District of Columbia	0.0	66	*North Dakota*	0.0	23
Florida	0.8	4,201	*Ohio*	1.3	7,266
Georgia	0.7	4,175	*Oklahoma*	0.0	126
Hawaii	0.0	49	*Oregon*	0.4	2,331
Idaho	0.2	878	*Pennsylvania*	0.8	4,244
Illinois	2.2	12,463	*Rhode Island*	0.0	112
Indiana	2.0	11,162	*South Carolina*	4.1	23,197
Iowa	0.1	725	*South Dakota*	0.0	42
Kansas	0.0	78	*Tennessee*	0.4	2,255
Kentucky	0.1	315	*Texas*	0.2	1,209
Louisiana	0.0	174	*Utah*	0.0	109
Maine	0.0	73	*Vermont*	0.0	52
Maryland	0.1	650	*Virginia*	0.2	909
Massachusetts	0.1	441	*Washington*	0.8	4,259
Michigan	3.0	16,680	*West Virginia*	1.4	7,869
Minnesota	0.6	3,536	*Wisconsin*	74.1	415,473
Mississippi	0.1	329	*Wyoming*	0.0	27
Missouri	0.1	296			

Table 2: Descriptive statistics: geography

by showing that the accuracy of the alternative scoring mechanisms marginally increases when running the analysis on a sample excluding Wisconsin.

3.3 Data Processing

Our data set has 872,382 data points and 5,022 variables. We use the entire data set to improve the accuracy of our model and to classify all the records in our data set. This involves three main processing steps. First, we randomize the order of the examples in the data set by shuffling the rows. Second, since the machine learning algorithms require there to be no missing values, we replace each missing value with a zero and generate a corresponding indicator variable for each variable, taking on the value 1 if the value is missing. We also include variables with existing numeric missing-value codes in this process (e.g. FICO score).

We then divide the full data set into a training data set, a validation data set, and a testing data set, comprising of 60%, 20%, and 20% of the data, respectively. We use the same data sets across all of the models in order to appropriately compare the accuracy rates between them. For the traditional regression analysis, we combine the training and validation data sets to estimate the models.

4 Developing the Alternative Scoring Mechanism

We now turn to developing the preferred alternative to traditional credit scores created specifically to evaluate customers for community solar participation, which leverages the rich data set on utility repayment history. We develop a number of alternative models that use the 12 months of data prior to December 2015 to predict the likelihood of being delinquent at least once in the following 12 month period (December 2015 to November 2016). The models, which we test in section 4, vary on two dimensions. First, we develop alternatives using a traditional regression model as well as machine learning techniques. Second, we vary the definition of a delinquency to be used as the dependent variable between a delinquency of greater than 30 days and greater than 90 days.

4.1 Traditional Regression Analysis

We start by estimating a set of models with a small number of variables, which we deem to be the most relevant for the probability of being delinquent in a given 12-month period. Using this regression method may present an improvement over using FICO alone. Unlike FICO, model would specifically predict the probability of delinquency in utility payments, rather than general financial habits, which may include many other categories not directly relevant for utility payment performance such as credit card debt and installment loans.

In particular, we estimate linear probability and probit models of the form:

$$Pr(D_{it}) = \alpha + \gamma_1 D^{30}_{it-1} + \gamma_2 D^{60}_{it-1} + \gamma_3 D^{90}_{it-1} + \gamma_4 FICO_{it-1} + \gamma_5 noFICO_{it-1} + \mathbf{X}'_{it}\beta, \tag{1}$$

where $Pr(D_{it})$ is the probability of at least one delinquency for individual i in the 12-month period using the 30-day or 90-day definition depending on the specification. The various D^{j}_{it-1} variables are indicator variables for at least one delinquency of more than j days for the individual in the previous 12-month

	NotCurrent		*>90DaysPastDue*	
	LPM	Probit	LPM	Probit
Regressor	(1)	(2)	(3)	(4)
FICO	-0.00105***	-0.00279***	-0.00139***	-0.00828***
	(5.49e-06)	(2.51e-05)	(4.27e-06)	(4.00e-05)
FICOBlank	-0.574***	-1.316***	-0.723***	-4.237***
	(0.00338)	(0.0213)	(0.00291)	(0.0274)
30DaysPastDue	0.147***	1.483***	0.479***	2.209***
	(0.00122)	(0.0172)	(0.00196)	(0.0121)
60DaysPastDue	-0.00688***	0.104***	-0.110***	-0.0498***
	(0.00124)	(0.0130)	(0.00204)	(0.0107)
>90DaysPastDue	0.0886***	0.891***	0.421***	1.664***
	(0.00101)	(0.00941)	(0.00128)	(0.00831)
NewMover	0.0394***	0.129***	-0.00324	-0.0176
	(0.00427)	(0.0156)	(0.00246)	(0.0257)
HomeOwner	-0.0479***	-0.290***	-0.0358***	-0.205***
	(0.00102)	(0.00624)	(0.000912)	(0.00694)
Multifamily	0.0809***	0.362***	-0.0171***	-0.0872***
	(0.00130)	(0.00702)	(0.000992)	(0.00823)
Constant	1.415***	2.544***	1.138***	4.154***
	(0.00384)	(0.0190)	(0.00330)	(0.0264)
N	697,762	697,762	697,762	697,762
R^2	0.204	0.233	0.758	0.754

Robust standard errors in parentheses (*** p<0.01, ** p<0.05, * p<0.1). All specifications also include indicator variables for missing demographic variables (*NewMover, HomeOwner, Multifamily*).

Table 3: Regression models of probability of delinquency with limited variables and varying delinquency definitions

period. $FICO$ and $noFICO$ represent the individual's FICO score and an indicator variable equal to one if that individual does not have a FICO score. The matrix $\mathbf{X}$ contains various demographic and housing characteristics, which in our preferred specification presented in Table 9 includes a binary variables for new movers (within the past 12 months), home ownership, and residence in a multifamily building. We estimated a number of other specifications, including those with the following demographic characteristics in addition to the aforementioned: a binary variable for college education and a categorical variable for income in $10K increments up to $120K+. These only marginally increase the R-squared of the model while adding in variables that are at odds with the LMI inclusion goal. The full set of specifications are presented in the appendix.

The regression models are presented in Table 9 using a 30 day definition for delinquency in columns (1) and (2), and a 90 day definition for columns (3) and (4), using both linear probability model and probit specifications. For the probit models, we report the marginal effects at the means of continuous variables and for binary variables, the average effect of moving from 0 to 1. The variables for days past due indicate

whether, at any point in the past 12 months, the account was 30, 60, or >90 past due (these variables are not mutually exclusive).

Immediately, we see that the coefficients on FICO score are negative and highly significant across specifications as expected. Looking at the linear probability models, all else held constant, a 10-point decrease in a FICO score would increase the probability of an account being 30 and over 90 days past due by 1.1 and 1.4%. Interestingly, however, having no FICO score seems to have a negative effect on the likelihood of being delinquent (a positive effect on payment performance). One plausible explanation is that those with poor payment performance and with no credit score already have their risk captured by the three delinquency variables, which are, for the most part negative and highly significant.

Interestingly, using the less strict 90 day definition for a delinquency as the dependent variable captures a much greater share of the variation than the 30 day definition (20.4% vs 75.8%). This is likely due to the fact that 30 day delinquencies are a much noisier measure of financial habits than delinquencies of greater than 90 days. For instance, a 30 day delinquency could be due to a one-time error such as a misplaced envelope, whereas a 90 day delinquency is more likely to be an indicator of being a risky consumer. This intuition is consistent with the fact that being a new mover only has a statistically significant effect in the 30 day models.

4.2 Machine Learning Techniques

We classify records with several different machine learning techniques in order to compare the performances of each one and select the one with the highest accuracy rates. First, we use different algorithms using a smaller data set, as described below. Since there are both continuous and categorical variables, it is important to normalize the data. We try three different normalization techniques in order to find the one that gives the best fit. We then perform dimensional reduction on our entire data set to prioritize the important features and create a condensed data set. We test several different machine learning techniques, such as LASSO, support vector machines, and a random forest algorithm using the condensed data set.

4.2.1 Creating Architectures on a Subset

Instead of creating all of our models using the entire data set, we use samples from the whole data. In particular, we take a subset of 10,000 samples and 13 features to build our algorithms. This enables us to conduct tests more rapidly. We incorporate the entire data set after we perfect our methodology.

In order to obtain a high level of accuracy, we normalize the data by rescaling using the following equation:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}. \tag{2}$$

This yields better accuracy, as it gives only non-negative values, and performed with the highest accuracy on the linear regressions, relative to the following alternatives:

$$x' = \frac{x - \mu}{\sigma}, \tag{3}$$

$$x' = \frac{x - \mu}{x_{max} - x_{min}}. \tag{4}$$

4.2.2 Dimensional Reduction Using LASSO

In order to reduce computing times, we perform dimensional reduction on the large data set to identify the important features and use the most significant variables. This agrees with the economic intuition of the data set. While some of the demographic variables hold economic significance (e.g. home market value, income code, and number of cars owned), other variables seemed extraneous and unnecessary for our analysis (e.g. whether the individual was a movie collector, type of preferred vacation, and women's suit size). Removing these parameters speeds up the computing time, maintains relevance of the parameters, and increases accuracy of the model. This also decreases the data requirements for the alternative scoring method. We use LASSO for feature selection and as a shrinkage method, to reduce the size of the data set we would use to train the model, identify the most important features, and use them to conduct the rest of the analysis [21]. We perform LASSO on the entire data set using remote computing, with a $\lambda = 0.05$, yielding 20 important features. The most important features were the deliquency in the previous time period, values from the payment grid, and the amount past due. The top 5 variables and their respective weights are displayed in Table 4.2.2.

Var Name	Meaning	Absolute Value of Weight
CURR KEYCD 24	Current on Utility Payments in previous year	0.3577
PAYMENT GRID81	Payment history grid	0.0359
ACCT PAST DUE AMT	Amount Past Due	0.0292
DELQ DT 1 BLANK	Most recent delinquency date unavailable	0.024654
DELQT DT 2 BLANK	Second most recent delinquency date unavailable	0.008477

Weights for 5 Most Important Values

All of these variables make intuitive economic sense, and should be useful in calculating the probability of delinquency for an individual. It is interesting to note that some of the most important features were as expected, such as delinquency in the last year and FICO score, while some other significant financial data we had not initially predicted would be so important. After looking through the top 20 features, we noticed that all of the key features described financial payment history, and none of them were demographic variables.

4.2.3 Support Vector Machines

Support Vector Machines (SVMs) are a method of supervised machine learning which uses labeled training data to formulate the optimal hyperplane that can classify new data points [23]. SVMs create decision boundaries between different labels (in our case, delinquent and not delinquent) in high dimensional spaces [1]. This means that if there is no clear decision boundary in a two-dimensional place, SVMs can extrapolate to higher dimensions to create a hyperplane that can be used to classify various data points. The dual form of linear SVMs is specified below, where x_i represents the input parameters, y_i is the decision variable, and α_i is the dual variable and related to the weight vector.

$$\max_{\alpha} -\frac{1}{2}||\sum_i \alpha_i y_i x_i||^2 + \sum_i \alpha_i, \tag{5}$$

$$\sum_i y_i \alpha_i = 0, \tag{6}$$

	NotCurrent		*>90DaysPastDue*	
	Training Accuracy	Testing Accuracy	Training Accuracy	Testing Accuracy
ML Method	(1)	(2)	(3)	(4)
LASSO	91.14 %	90.37%	96.26%	96.05%
SVM	94.44%	89.71%	99.02%	87.82%
Random Forest Algorithm	100%	97.49%	100%	98.99%
Random Forest without WI	100%	97.85%	100%	99.05%

Table 4: Accuracy Rates for Machine Learning Algorithms with Different Definitions of Delinquency

$$0 \leq \alpha \leq C. \tag{7}$$

The algorithm will perform certain transformations on the data points, known as kernels, to translate it into higher dimensions. Kernels are useful tools to express complicated feature functions in a simple way. Beyond the linear kernel, the Gaussian radial basis function (rbf) kernel is a popular kernel function. The Gaussian RBF kernel has special properties which allows it to classify correctly almost all of the time. However, one must be wary of overfitting when using the Gaussian RBF Kernel, which is clearly stated in Equation 8.

$$K(x, x') = exp(-\gamma||x - x'||^2). \tag{8}$$

A regularization term, C, is added to prevent overfitting and accommodate cases when the data is linearly inseparable. The regularization parameter represents the importance of the training errors. As the regularization parameter, C, increases, the margin width becomes smaller, and therefore there are less margin violations. An increase in the regularization term correlates with a greater emphasis on margin violations, and the margin becomes tighter around the decision boundary. Thus, the number of support vectors, and violations, decreases as C increases. However, it is imperative to consider the tradeoff between accuracy and robustness, as it important to prevent the algorithm from overfitting to the training data.

The γ term reflects a certain margin of error surrounding the decision boundary. A small gamma corresponds to a decision boundary that underfits the data. In parallel, a larger gamma value tends to overfit the data.

We tune the hyperparameters, C, γ, and the kernel type, on the validation data set. The SVM we use has the following specifications: C = 10, γ= 0.1, and it utilizes a radial basis function (rbf) kernel. While this method displays high accuracy rates, it is very time consuming, which is an important factor when comparing it to other methods.

4.2.4 Random Forest Algorithm

We use the random forest algorithm, another type of supervised machine learning, which essentially separates the data into multiple smaller datasets, or *bags*, and forms decision trees with the smaller data sets, and uses the many decision trees to classify the input parameters, as further described below.

Decision trees are particularly appropriate for our data due to the fact that it includes many features of varying importance, on different scales. Decision trees are useful for finding the appropriate feature to split

on, and the value of that feature in order to minimize the cost function [13]. We use a greedy heuristic model, which locally minimizes the cost in order to find the global optimum.

Since we have a large amount of data, we find that bagging, or the bootstrap algorithm, is the best way to improve our accuracy rates while preventing overfitting the data set. Bagging essentially means that the algorithm is taking random samples, creating several different classifiers, and uses the errors from one classifier to 'learn' from its mistakes and create future classifiers. The random forest algorithm creates many random samples (many decision trees) and essentially averages the outcome over all of the decision trees to come up with one final answer. We used *SkLearn*'s learning implementation of random forests in order to label our records using this technique, and to predict the probabilities of delinquency and non-delinquency. The depth of each tree is limited to 150 levels and the seed of the forests is predetermined to 27. The results are given in Table 4.

Figure 4.2.4 shows a visual representation of the random forest algorithm. While the entire random forest is very large, has many branches and nodes, and can be complicated to follow, we displayed one of the decision trees in order to visualize how the architecture works. This visualization enables us to understand how some variables specifically affect the labelling, which could be useful for further applications.

Not only are decision trees accurate, they have a relatively short running time. Not all of the features have the same level of importance for our model; the decision tree can distinguish which features are important, and rank them in order of importance. While other models mainly consider linear or non-linear combinations of the features, the decision tree algorithm is able to solve the best splitting criteria: this may be a binary split, a specific threshold, a quadratic term, or another non-linear representation of a feature. It is particularly efficient here as, on the one hand, it accounts for highly non-linear combination and gives interpretability, and on the other hand, it does not require dimensionality reduction, which is a time-consuming process.

4.2.5 Summary

Among the variety of models that we explored, the random forest algorithm is clearly superior in terms of accuracy. Moreover, the random forest algorithm not only has better accuracy, but it also requires less data pre-processing. Finally, it is easier to interpret and runs more quickly. These are the three main reasons as to why we decide to use a random forest architecture as a preferred scoring mechanism, rather than the FICO score or other techniques tested.

5 Results

We now turn to comparing the alternative scoring methods developed with traditional regression analysis and machine learning techniques to standard FICO cutoffs, in terms of accuracy, default rate, and LMI inclusion.

Figure 2 displays the probabilities of non-delinquency using the random forest algorithm against the individual's FICO Score. There are many individuals who have a high probability of non-delinquency with the random forest algorithm, but do not have a very high FICO score, which demonstrates the amount of people that would have been rejected with the FICO cutoff, but accepted according to the random forest algorithm ("false negatives"). Additionally, there are quite a few data points with high FICO scores but do not have a very high probability with the random forest algorithm, who would be erroneously accepted ("false positives"). Figure 2 suggests that there are a high numbers of false negatives and false positives under

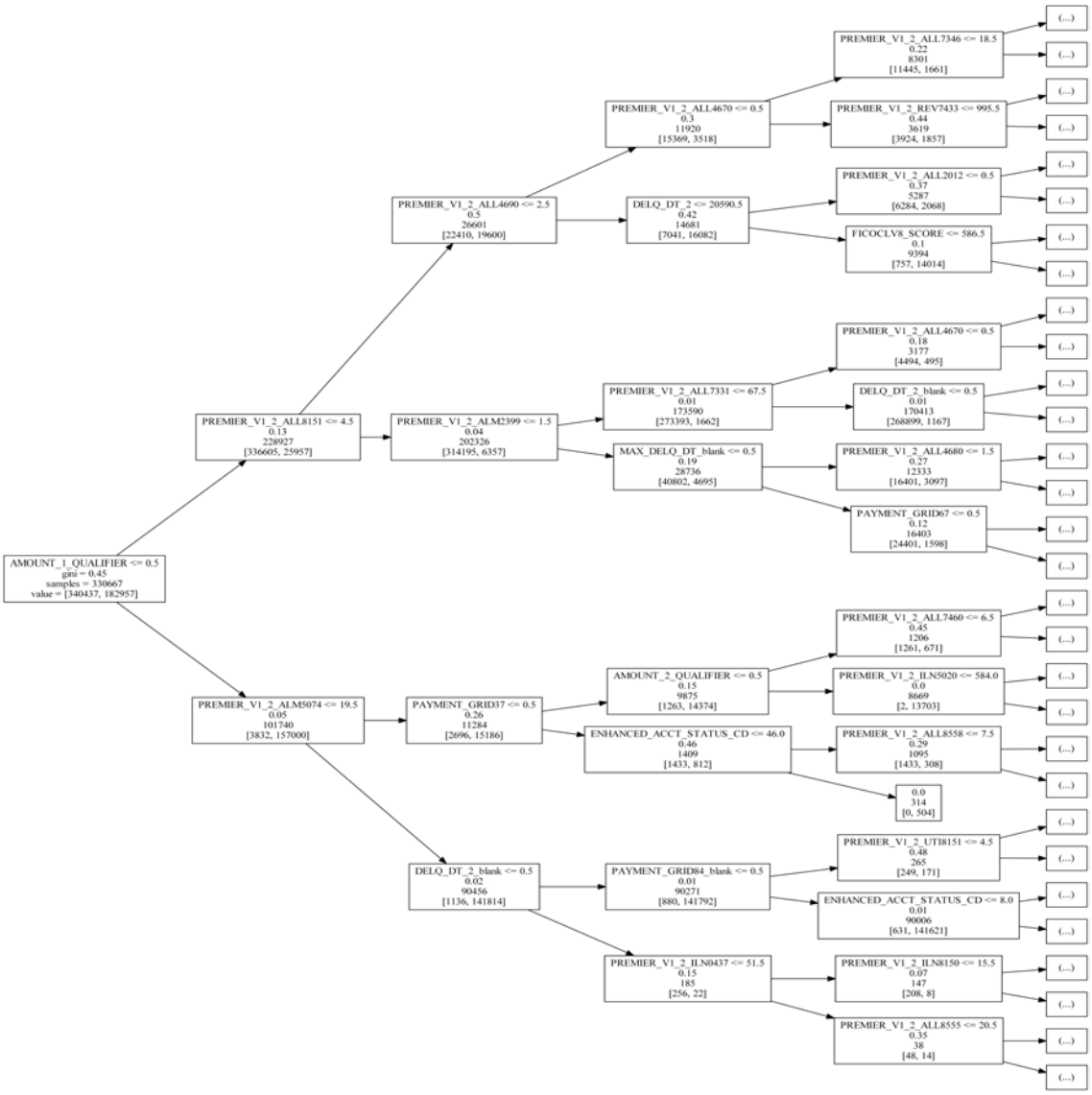

Figure 1: Visual representation of Random Forest Algorithm

traditional FICO scoring. Though the FICO Score is one variable used by the random forest algorithm, there are many other variables as well. In order to further render the random forest algorithm comparable to the FICO score, we compute the share of the sample approved under all possible FICO cutoffs and compare FICO to an equivalently selective random forest algorithm.

Figure 3 shows the accuracy of the random forest algorithm relative to FICO. The false positive rate on the graphs in the first row indicate the percentage of those accepted that are ultimately delinquent on their payments, and the false negative rate on the graphs in the bottom row is the inverse: those rejected that would have been current on their payments. The graphs in the left-hand column are those using models that predict delinquencies of 30 days or more, and the graphs on the right-hand column are those using delinquencies of 90 days or greater.

There are a number of interesting trends identified from these results. First, as discussed above, because monthly utility payment performance histories are incomplete, those that are delinquent on their payments

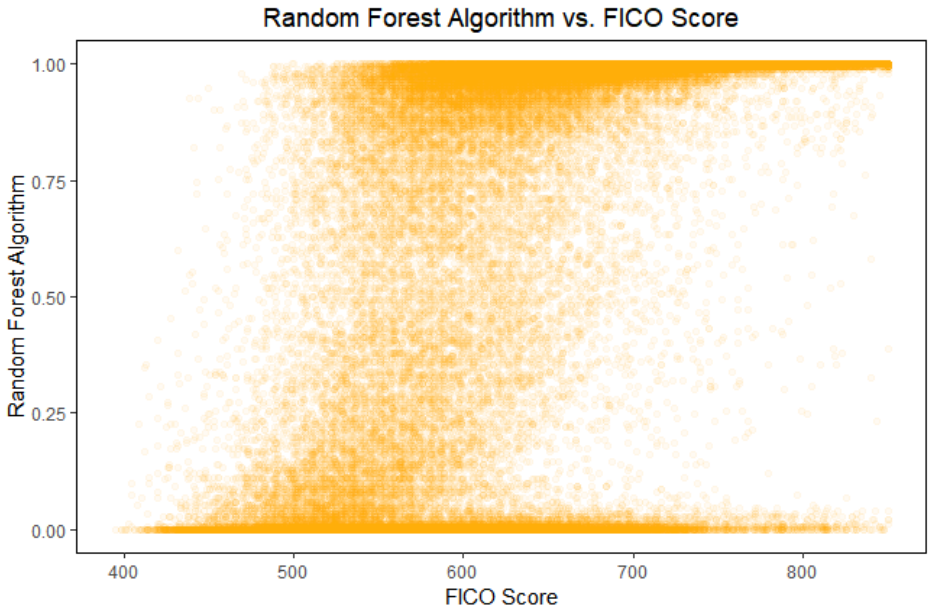

Figure 2: Random Forest Algorithm and FICO Score

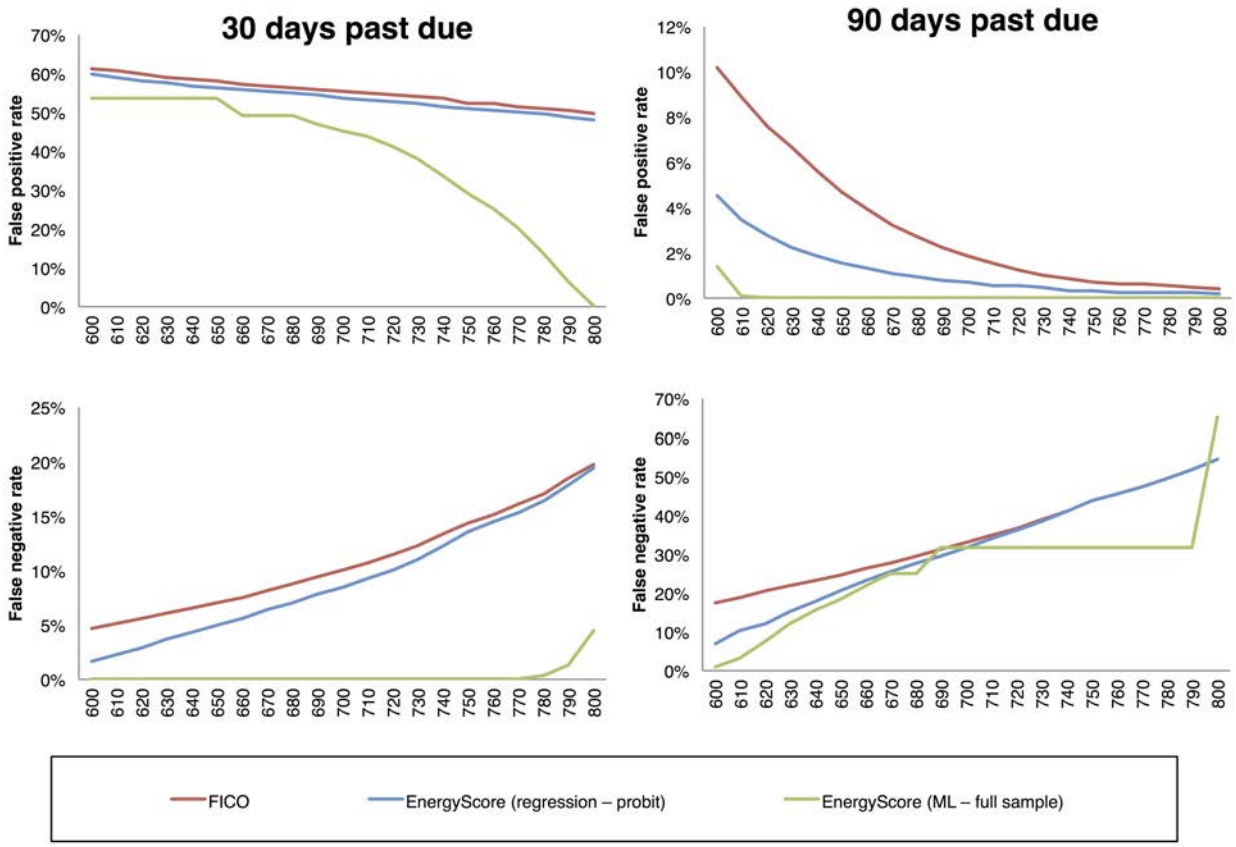

Figure 3: Accuracy rates for models using 30 day and 90 day definitions of delinquency over equivalent FICO scores

tend to be over-represented in our restricted sample of accounts with 24 months of consecutive data. Second are the overall trends; the false positives are all downward-sloping and the false negatives are all upward-sloping. This agrees with our intuition. Higher FICO-equivalent cutoffs imply higher selectivity, leading to lower delinquency rates and hence lower false positive rates. More stringent cutoffs also imply that more qualified applicants are being rejected, driving up the share of rejected applicants that end up paying on time, and by extension, the false negative rate.

Third is the fact that the machine learning curve has sections that are flat. This is because the random forest algorithm optimizes the best splitting criterion for each branch of the decision tree in order to calculate the probability of delinquency. In other words, it assigns probabilities of delinquency by putting data points into categories according to the independent variables. This could consist of a binary variable, or it could split based on a specific threshold of a continuous variable. For example, it may calculate the probability of delinquency based on whether the income code is below the $110K to $120K category. Therefore, accounts with the same values for certain categories will have the same probability of delinquency, as opposed to a regression in which different values for the covariates necessarily leads to differences in the dependent variable. The random forest algorithm based on a 90 day definition of delinquency therefore assigns roughly 28% of the sample the same minimum probability of delinquency and another 1% the next lowest probability of delinquency. Since the accuracy rates are computed such that those below a particular cutoff are rejected, the accuracy curves move in a stepwise manner in the relevant ranges. For false negatives, since a high cutoff means most applicants are rejected, the false negative rate tends to the sample non-delinquency rate above an 800 FICO equivalent cutoff.

Comparing between models, we can immediately see that the random forest algorithm yields great gains in accuracy over a strict FICO cutoff. This is especially true using the machine learning methods, when compared to the regression techniques. For instance, when comparing to a FICO cutoff of 680, the random forest algorithm developed using a 30-day delinquency definition decreases the false positive rate by 7.0 percentage points (56.4% to 49.4%) and the false negative rate by 8.7 percentage points (8.7% to 0.0%). Similar gains are observed using a 90-day definition of delinquency. Here, the false positive rate (i.e. delinquencies among the approved pool) falls 2.7 percentage points (2.7% to 0.0%), while the false negative rate (i.e. rejected applicants being non-delinquent) falls 4.2 percentage points (29.5% to 25.3%).

The higher accuracy of the 90 day definition also agrees with our intuition. If we are using delinquency as a measure of creditworthiness, delinquency using a 30 day definition could be noisier (i.e., due to an error such as a misplaced bill) than a 90 day delinquency, which would be more indicative of financial tendencies. This is consistent with the much higher explanatory power of the regressions using 90 day delinquency as the dependent variable.

The stringency of the FICO score cutoff affects the accuracy comparisons with the alternative scoring methods, as seen in Figure 4. Here, we define a default as an account that has at any point in the 12-month period either been transferred to a collections agency or charged off. Compared to a FICO score cutoff of 680, the default rate decreases by 1.4 percentage points (1.9% to 0.5%) using a 30 day delinquency definition and by 1.9 percentage points (1.9% to 0.0%) using a 90 day delinquency definition.

Importantly, the random forest algorithm, when tested with both 30 and 90 day definitions of delinquency, increase the number of LMI applicants approved, as seen in Table 5. The random forest algorithm using a 30 day definition increases the number of LMI accounts approved by 11.4% to 14.0% depending on the stringency, while that using a 90 day definition increases LMI customers by 1.1% to 4.2%. However, worth

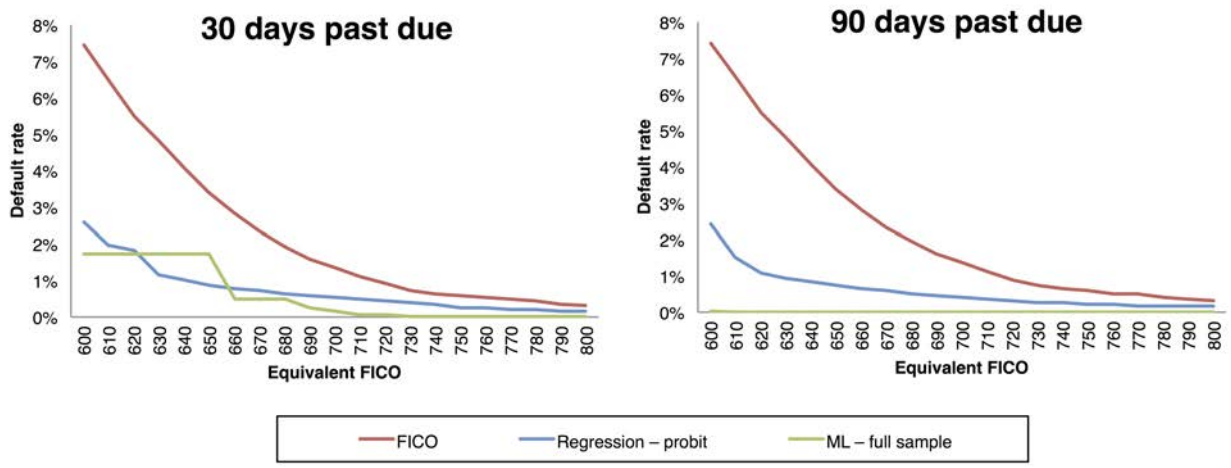

Figure 4: Default rate using 30 day and 90 day delinquency definitions over equivalent FICO scores

FICO equiv.	Regression	Machine Learning
30 days past due		
650	-3.8%	13.4%
680	-7.3%	14.0%
700	-8.9%	11.4%
90 days past due		
650	-1.0%	2.4%
680	-1.2%	4.2%
700	-1.8%	1.1%

Table 5: Change in number of LMI customers approved relative to a FICO cutoff

noting is that this is not the case with the random forest algorithm using traditional regression techniques on a smaller set of variables, which see slight decreases in the LMI population approved. This could be due to the limited number of variables used in the regressions, which are highly correlated with income, whereas the random forest algorithm uses the full data set.

6 Implications for Profitability

In this section we develop a profit model to predict the expected profits of the firm when using the random forest algorithm. For now, we assume that if a customer is offered the product, they purchase it. In this case, expected profits depend on the rule that dictates whether the product is offered to the customer and customers' default rates. Let the rule dictating whether the product is offered be denoted as, $I(X)$, where X is a set of variables the firm uses to generate the "offering rule." Similarly let $I(X_i)$ represent the indicator variable for whether consumer i is offered the product.

Consumers may default on paying for the service. For simplicity, imagine that the consumer defaults right away and all costs are up front, so that the firm never collects any revenues and incurs all of the costs. Let $\Pr(D_i = 1|X_i)$ represent the probability consumer of type X_i defaults.

First consider the profits of a firm that are not conditional on any information in X. Profits can be

written as:

$$E\left[\pi_i(P, MC)\right] = \sum_i \left[P \cdot (1 - \Pr(D_i)) - MC(X_i)\right], \tag{9}$$

where P is the price and MC is the marginal cost. If this expression is positive, the firm offers the product to everyone; if it is negative, the firm exits. If the firm offers the product then it must be the case that the average repayment rate is greater than the ratio of marginal cost to price, given by:

$$1 - \overline{\Pr(D_i)} > \frac{MC}{P}. \tag{10}$$

We can also rearrange this expression to yield a condition on how the average default rate relates to the Lerner index:

$$\overline{\Pr(D_i)} < \frac{P - MC}{P}. \tag{11}$$

Better scoring technology allows the firm to increase profits through eliminating customers that have negative expected profits. In particular, the firm's first order condition will imply probability of default for the marginal customer equals the Lerner index, given by:

$$\overline{\Pr(D_i|I(X_i))} = \frac{P - MC}{P} \tag{12}$$

6.1 Empirical Implementation

Though we lack data on prices and costs, we can use the fact that the industry appears to utilize decision rules based on a customer's FICO score to bound the ratio of marginal cost to price. For example, an often-cited decision rule is to offer customers the product if their FICO score is above 650. We can use this rule, and similar rules based on FICO scores in two ways.

The first way uses the FICO cut off as a way to estimate the ratio of marginal cost to price. Notice that if decision rule is optimal, it implies that:

$$\overline{\Pr(D_i|FICO = 650)} = \frac{P - MC}{P}. \tag{13}$$

Therefore, given an estimate of the expected default rate of customers with FICO scores of 650, we can estimate the Lerner index. One estimate of the left hand side of this equation is the average default rate for customers with FICO scores of 650. To empirically implement this we take the empirical average default rate of customers with FICO scores of $650 \pm X$ where we will vary X to gauge robustness. This defines the Lerner index which we will use to gauge the benefits of improvements in credit scoring. To estimate the change in profits from different scoring rules, we normalize price to be 1 implying marginal cost is $\Pr(D_i|FICO = 650)$. We calculate the profit for the random forest algorithm and the industry standard, using Equation 14.

$$E\left[\pi_i(P, MC, I(X_i))\right] = \sum_i P \cdot I(X_i) \cdot (1 - \Pr(D_i)) - MC \cdot I(X_i). \tag{14}$$

The results are displayed in Table 6.

Regardless of the FICO score cutoff, the random forest algorithm leads to an increase in profits for the firm, which is a very significant result from our study. The random forest algorithm both benefits the

FICO equiv.	Industry Standard	Random Forest Algorithm	Total Percent Increase
650	$ 20,337.37	$ 27,287.22	34%
680	$ 5,393.77	$ 9,428.54	75%
700	$ 127.65	$ 2,529.99	1882%

Table 6: Profit estimates for industry standard and random forest algorithm at three different FICO cutoffs

FICO equiv	π from New Customers	π from Less Delinquents	Total π Increase
650	$ 8,232.05	$ 4,216.79	$ 6,949.84
680	$ 5,932.67	$ 3,943.36	$4,034.77
700	$ 4,057.96	$ 3,618.99	$2,402.34

Table 7: Profit increase between Random Forest Algorithm and FICO Score attributed to New Customers and by Preventing Delinquent Customers

customers, by accepting more LMI customers, and benefits the firms, by increasing profits. As the FICO score becomes more stringent, the firm's profits decrease drastically using a FICO score cutoff, while the decrease is much more modest using comparatively stringent cutoffs using the random forest model. Hence, the percentage increase in profits by moving from the industry standard to the random forest algorithm increases with stringency. However, as shown in Table 7, the dollar value of the increase in profits from the random forest algorithm relative to a FICO score cutoff decreases as the FICO score cutoff becomes more stringent, because the firm is accepting less customers overall. We can decompose the increase in profits from the random forest algorithm to two sources. First is the increase in profits due to accepting new customers who would have been denied under the FICO score cutoff, or a decrease in false negatives ("π from New Customers). Second is a reduction in losses from rejecting those who are accepted under the FICO Score cutoff but whom the random forest algorithm identifies as high-risk, or a decrease in false positives ("π from Less Delinquents"). Note that these two columns do not sum up to the value in Total π Increase because the firm that uses the random forest algorithm could still lose profits by denying access to a customer that would have brought them profits, or by accepting some delinquents, who would have been correctly classified under a FICO score cutoff. Overall, however, the random forest algorithm leads to an increase in profits when compared to the FICO score cutoff, regardless of the stringency of the industry standard, due to the overall decrease in false positives and false negatives.

7 Conclusion and Future Implications

In this paper, we develop an alternative score based on a more accurate model of utility bill payment performance that would be more inclusive of LMI households than a traditional credit score. We do so using a variety of traditional regression approaches, as well as machine learning techniques, on a large data set from a credit reporting agency (CRA) to develop a model that predicts the probability of non-delinquency. The final alternative scoring mechanism is based on a random forest algorithm, because it has the highest accuracy rates, a reasonable computation time, and a more comprehensive interpretability. We find that even a traditional regression analysis using a small number of variables specific to utility repayment performance greatly increases accuracy and LMI inclusivity relative to FICO, and that using machine learning techniques enhances this further. Our preferred random forest algorithm increases the number of LMI

applicants approved by 1.1% to 4.2%, while decreasing the default rate by 1.4 to 1.9 percentage points depending on the stringency of the cutoff. Our analysis shows that it is possible to extend solar to a larger number of qualified applicants with lower or no credit scores while decreasing default risk, thus representing an untapped, low-risk market segment.

References

[1] 1.4. support vector machines. Available at scikit-learn.org/stable/modules/svm.html.

[2] A. C. Antonakis and M. E. Sfakianakis. Assessing naïve bayes as a method for screening credit applicants. *Journal of Applied Statistics*, 36(5):537–545, 2009.

[3] Jennifer Brooks, Kasey Wiedrich, Lebaron Sims Jr., and Solana Rice. *Excluded from the Financial Mainstream: How the Economic Recovery is Bypassing Millions of Americans*. CED, Washington, D.C, 5 edition, 2015.

[4] Brian Carey, Debi Gerstel, and Sean Jang. Community solar: Share the sun rooflessly. Technical report, London UK, 2017.

[5] CFPB. Who are the credit invisibles? how to help people with limited credit histories. Technical report, Washington DC, 2016.

[6] Ariel Drehobl and Lauren Ross. Lifting the high energy burden in america's largest cities: How energy efficiency can improve low-income and under-served communities. Technical report, Washington DC, 2016.

[7] EPA. Community solar: An opportunity to enhance sustainable development on landfills and other contaminated sites. Technical report, Washington DC, 2016.

[8] Jeffrey Feinstein. Alternative data and fair lending. Technical report, New York NY, 2013.

[9] David Feldman, Anna M. Brockway, Elaine Ulrich, and Robert Margolis. Shared solar: Current landscape, market potential, and the impact of federal securities regulation. Technical report, Golden, CO, 2015.

[10] Cory Honeyman. U.s. community solar outlook. Available at http://www.blowinglotsofweirdstuffup.com/guide.html, 2 2017.

[11] Cheng-Lung Huang, Mu-Chen Chen, and Chieh-Jen Wang. Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*, 33(4):847–856, 2007.

[12] Wei Jiang, Ashlyn Aiko Nelson, and Edward Vytlacil. Securitization and loan performance: Ex ante and ex post relations in the mortgage market. *The Review of Financial Studies*, 27(2):454–483, 2013.

[13] Zachary M. Jones and Fridolin Linder. Exploratory data analysis using random forests. Available at https://cran.r-project.org/web/packages/edarf/vignettes/edarf.html.

[14] Benjamin J. Keys, Tanmoy Mukherjee, Amit Seru, and Vikrant Vig. Financial regulation and securitization: Evidence from subprime loans. *Journal of Monetary Economics*, 56(5):700 – 720, 2009.

[15] Benjamin J. Keys, Tanmoy Mukherjee, Amit Seru, and Vikrant Vig. Did securitization lead to lax screening? evidence from subprime loans. *The Quarterly Journal of Economics*, 125(1):307–362, 2010.

[16] Benjamin J. Keys, Amit Seru, and Vikrant Vig. Lender screening and the role of securitization: evidence from prime and subprime mortgage markets. *The Review of Financial Studies*, 25(7):2071–2108, 2012.

[17] 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.

[18] John Krainer and Elizabeth Laderman. Mortgage loan securitization and relative loan performance. *Journal of Financial Services Research*, 45(1):39–66, 2014.

[19] Jochen Kruppa, Alexandra Schwarz, Gerhard Arminger, and Andreas Ziegler. Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications*, 40(13):5125–5131, 2013.

[20] Kin Keung Lai Lean Yu, Shouyang Wang. Credit risk assessment with a multistage neural network ensemble learning approach. *Expert Systems with Applications*, 34(2):1434–1444, 2008.

[21] Kin Keung Lai Lean Yu, Shouyang Wang. Credit risk assessment with a multistage neural network ensemble learning approach. *Expert Systems with Applications*, 34(2):1434–1444, 2008.

[22] Masoud Nikravesh. Credit scoring for billions of financing decisions. In *Proceedings Joint 9th IFSA World Congress and 20th NAFIPS International Conference*, volume 1, pages 191–196, July 2001.

[23] Savan Patel. Chapter 2 : Svm (support vector machine) - theory. Available at medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72. (2017/05/03).

[24] Uday Rajan, Amit Seru, and Vikrant Vig. The failure of models that predict failure: Distance. incentives and defaults. *Journal of Financial Economics*, 115(2):237–260, 2015.

[25] Rachel Schneider and Arjan Schütt. The predictive value of alternative credit scores. Available at hePredictiveValueofAlternativeCreditScores, November 2007.

[26] Stacy Smith. How is your credit score determined? Available at https://www.experian.com/blogs/ask-experian/how-is-your-credit-score-determined/, August 2016.

[27] Michael A. Turner and Patrick Walker. Predicting financial account delinquencies with utility and telecom payment data. Technical report, Durham, NC, 2015.

[28] Michael A. Turner, Patrick D. Walker, Sukanya Chaudhuri, and Robin Varghese. A new pathway to financial inclusion: Alternative data, credit building, and responsible lending in the wake of the great recession. Technical report, Durham, NC, 2012.

[29] Gang Wang, Jinxing Hao, Jian Ma, and Hongbing Jiang. A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications*, 28(1):223–230, 2011.

[30] Gang Wang and Jian Ma. A hybrid ensemble approach for enterprise credit risk assessment based on support vector machine. *Expert Systems with Applications*, 39(5):5325–5331, 2012.

[31] Gang Wang, Jian Ma, Lihua Huang, and Kaiquan Xu. Two credit scoring models based on dual strategy ensemble trees. *Knowledge-Based Systems*, 26:61–68, 2012.

[32] Yongqiao Wang, Shouyang Wang, and K. K. Lai. A new fuzzy support vector machine to evaluate credit risk. *IEEE Transactions on Fuzzy Systems*, 13(6):820–831, 2005.

8 Appendix

Regresssor	LPM (1)	Probit (2)	LPM (3)	Probit (4)	LPM (5)	Probit (6)	LPM (7)	Probit (8)
30DaysPastDue	0.252***	1.852***	0.146***	1.474***	0.147***	1.483***	0.142***	1.472***
	(0.00122)	(0.0166)	(0.00120)	(0.0173)	(0.00122)	(0.0172)	(0.00122)	(0.0172)
60DaysPastDue	-0.0352***	0.243***	-0.0160***	0.0457***	-0.00688***	0.104***	-0.00388***	0.120***
	(0.00139)	(0.0136)	(0.00122)	(0.0129)	(0.00124)	(0.0130)	(0.00125)	(0.0131)
90DaysPastDue	0.274***	1.532***	0.0939***	0.908***	0.0886***	0.891***	0.0867***	0.888***
	(0.000766)	(0.00770)	(0.00101)	(0.00919)	(0.00101)	(0.00941)	(0.00102)	(0.00947)
FICO			-0.00115***	-0.00328***	-0.00105***	-0.00279***	-0.000951***	-0.00244***
			(5.29e-06)	(2.42e-05)	(5.49e-06)	(2.51e-05)	(5.64e-06)	(2.56e-05)
FICOBlank			-0.631***	-1.653***	-0.574***	-1.316***	-0.525***	-1.107***
			(0.00327)	(0.0207)	(0.00338)	(0.0213)	(0.00344)	(0.0217)
NewMover					0.0394***	0.129***	0.0434***	0.141***
					(0.00427)	(0.0156)	(0.00425)	(0.0157)
HomeOwner					-0.0479***	-0.290***	-0.0288***	-0.217***
					(0.00102)	(0.00624)	(0.00105)	(0.00636)
Multifamily					0.0809***	0.362***	0.0771***	0.343***
					(0.00130)	(0.00702)	(0.00130)	(0.00706)
Income							-0.0134***	-0.0526***
							(0.000225)	(0.000930)
College							-0.0388***	-0.124***
							(0.00114)	(0.00388)
Constant	0.617***	0.249***	1.459***	2.699***	1.415***	2.544***	1.422***	2.584***
	(0.000689)	(0.00189)	(0.00378)	(0.0184)	(0.00384)	(0.0190)	(0.00385)	(0.0192)
N	697,762	697,762	697,762	697,762	697,762	697,762	697,762	697,762
R^2	0.147	0.193	0.197	0.223	0.204	0.233	0.211	0.239

Robust standard errors in parentheses (*** p¡0.01, ** p¡0.05, *p¡0.1). All specifications also include indicator variables for missing demographic variables.

Table 8: Full regression specifications for probability of deliquency using 30-day definition

Regresssor	LPM (1)	Probit (2)	LPM (3)	Probit (4)	LPM (5)	Probit (6)	LPM (7)	Probit (8)
30DaysPastDue	0.621***	2.808***	0.480***	2.213***	0.479***	2.209***	0.477***	2.205***
	(0.00205)	(0.0118)	(0.00196)	(0.0122)	(0.00196)	(0.0121)	(0.00196)	(0.0122)
60DaysPastDue	-0.138***	0.0485***	-0.113***	-0.0662***	-0.110***	-0.0498***	-0.109***	-0.0420***
	(0.00237)	(0.0117)	(0.00203)	(0.0106)	(0.00204)	(0.0107)	(0.00203)	(0.0108)
90DaysPastDue	0.649***	2.720***	0.425***	1.676***	0.421***	1.664***	0.421***	1.662***
	(0.00108)	(0.00703)	(0.00128)	(0.00829)	(0.00128)	(0.00831)	(0.00128)	(0.00833)
FICO			-0.00143***	-0.00852***	-0.00139***	-0.00828***	-0.00136***	-0.00805***
			(4.13e-06)	(3.92e-05)	(4.27e-06)	(4.00e-05)	(4.33e-06)	(4.03e-05)
FICOBlank			-0.745***	-4.392***	-0.723***	-4.237***	-0.709***	-4.110***
			(0.00285)	(0.0267)	(0.00291)	(0.0274)	(0.00293)	(0.0276)
NewMover					-0.00324	-0.0176	-0.00199	-0.00153
					(0.00246)	(0.0257)	(0.00246)	(0.0257)
HomeOwner					-0.0358***	-0.205***	-0.0301***	-0.144***
					(0.000912)	(0.00694)	(0.000922)	(0.00713)
Multifamily					-0.0171***	-0.0872***	-0.0188***	-0.0958***
					(0.000992)	(0.00823)	(0.000994)	(0.00826)
Income							-0.00514***	-0.0467***
							(0.000143)	(0.00139)
College							-0.00354***	-0.0711***
							(0.000599)	(0.00689)
Constant	0.0882***	-1.623***	1.133***	4.149***	1.138***	4.154***	1.143***	4.220***
	(0.000326)	(0.00321)	(0.00327)	(0.0260)	(0.00330)	(0.0264)	(0.00331)	(0.0266)
N	697,762	697,762	697,762	697,762	697,762	697,762	697,762	697,762
R^2	0.687	0.667	0.757	0.753	0.758	0.754	0.759	0.756

Robust standard errors in parentheses (*** p¡0.01, ** p¡0.05, *p¡0.1). All specifications also include indicator variables for missing demographic variables.

Table 9: Full regression specifications for probability of deliquency using 90-day definition