Risk, Return, and Matching in Online Lending

ECON 21150 Final Project

Daniel Chavez and Ariel Goldszmidt

June 2018

Abstract

We study the problem faced by an online lending company in assigning interest rates to borrowers. The lender assigns borrowers to interest rate grades in order to optimize a combination of return and risk, but must make this assignment decision before observing direct information on the borrower's potential returns. In this paper, we present a simple model in which the lender makes ex ante predictions of the mean and variance of a potential borrower's returns under each grade to which they may be assigned. We model grade assignment as a discrete choice problem over the loans to identify the risk-reward preferences of the lender. Using Lending Club data, we find evidence that the lender behaves as a highly risk averse investor in assigning borrowers to grades. However, we also find that the observed distribution of grade assignments cannot be fully explained by simple mean-variance optimization.

1 Introduction

As a substitute for traditional brick-and-mortar payday lending storefronts, online peer to peer (P2P) lenders such as Lending Club, Avant, and Lending Tree have quickly captured much of the short-term loan industry and have expanded upon the traditional business model of payday lenders. For example, Lending Club, one of the largest and oldest P2P lenders not only lends money to millions of borrowers, but also packages those loans into assets which investors can buy shares in and fund. The loans comes with different terms, sizes, and interest rates, and are bucketed into loan grades A through G. In this paper, we study the process through which heterogeneous borrowers are assigned interest rates and placed into loan grades, based on the limited borrower information which an online lender can observe.

P2P lenders operate in a manner broadly reminiscent of an investor managing a portfolio of assets, which leads us to study the process of interest rate assignment through the lens of portfolio optimization. The lender faces a trade-off in assigning interest rates: assigning a higher interest rate can potentially give a higher return, but may also increase the probability that the borrower will default. Can the principles of risk and return optimization explain the observed matching of borrowers to rates?

Studying this process is important for several reasons. Most importantly, it may give insight into the methods private lenders use to price credit, which are largely black-box. Additionally, the analysis will show whether the common sense economics of risk and reward can adequately explain lender behavior. Furthermore, an analytic framework for examining this process could be extended and applied to various other markets in which one side must assign applicants into grades with limited information.

To study our research question, we first model econometrically how the lender can use their limited information about borrowers to predict the important risk and reward features of a loan in each grade. We then use these econometric models to construct counterfactual measures of expected return and risk for each loan in the data by using borrower characteristics. Once these risk and return measures are in place, we study the lender's decision problem as a discrete choice problem where the lender must decide for each borrower (and hence for each risk and return profile) in which loan grade to place them in order to maximize overall utility. Estimating the parameters of this discrete choice model which will inform us of the relative importance of risk and return in the lender's decision, and show whether the assignment process can really be fully explained by mean-variance optimization.

2 Literature Review

The subject matter of this paper intersects many bodies of literature, including those of personal lending (both traditional and online), credit and risk scoring, and portfolio optimization.

We both build on and diverge from the traditional theory of consumer credit pricing, a general overview of which is offered by Özer and Phillips (2012). Before the 1980's, lending institutions usually offered only one interest rate for consumer loans, which was set to hit a target return on capital, adjusting for costs and expected losses. Later, lenders adopted risk-based pricing models, which offer different rates to different types of consumers based on some measure of riskiness. While the adoption of risk-based pricing expanded consumer credit channels, it also made lenders' interest rate decisions less transparent. More recently, many lending institutions have moved to profit-based pricing, which sets interest rates for different consumer segments to maximize expected profit.

Phillips (2013) gives a theoretical treatment of risk- and profit-based pricing, modelling a lender which sets interest rates for each segment to maximize expected net present value of interest income. The author presents conditions under which this maximization problem has a unique solution. He also discusses the phenomenon of price-dependent risk—through which borrowers given higher interest rates become riskier—and underscores the importance of accounting for this fact in the optimization problem.

We also reference theoretical work done to extend portfolio optimization theory to the context of personal loans. Mencía (2012) develops a model for the returns on such loans, and considers the mean-variance efficient frontier of portfolios consisting of such loans. The author finds evidence in Spanish banking data that interest rates are generally consistent with those produced

by mean-variance optimization with a value-at-risk constraint.

The rapid growth of online lending over the past decade has also produced many research papers focusing specifically on P2P loan markets. Wei and Lin (2017) compare two mechanisms for setting personal loan interest rates, one in which an auction is used to set rates and another in which the platform itself sets rates. Using an auction-theoretic model and data from Prosper.com, they find evidence that when platforms set rates manually, interest rates are higher, more loans are funded, and more loans default. Klafft (2008), also using Prosper.com data, shows that return on investment has been generally poor in online lending, and even negative for loans to highly risky borrowers. Based on this evidence, the author argues that online lending markets are too easy in issuing credit (or at least were before the financial crisis).

3 Data

We use publicly available Lending Club loan data containing information on each borrower and the terms of their loan. The data can be downloaded at `https://www.lendingclub.com/info/download-data.action`. Borrower features may be divided into those which are observed before a loan is given—employment length, annual income, debt-to-income ratio, credit history, and so on—and those which are only observed after a loan is given—whether a borrower pays on time or defaults, for example.

The data records observations of 145 features for 1,765,451 loans issued between 2007 and 2017. Many features are missing for large numbers of loans. We first remove all current loans from the data, as we have not yet observed whether these loans will default. We then remove all features for which observations are missing for more than 200 loans, and remove any categorical variables which take on too many categories (zip code and employment title, for example). We then remove any loans for which we are still missing some features, and encode the remaining categorical variables as dummies. Our final data set consists of 958,132 loans, and includes 100 features which are observed before a loan is given and a few observed after the loan is given (interest rate, default indicator, recovery rate, etc.).

Table 1 gives summary statistics for all the non-categorical variables in our final data set. The data set also includes dummy variables for home ownership type, loan purpose, home state, year issued, verification status, disbursement method, and application type.

In Figure 1, we plot normalized histograms of the distribution of interest rates in each loan grade. Note that in each grade, interest rates seem to cluster around a grade mean. The variance of interest rates between grades also appears to exceed the variance within grades. These observations motivate a modeling approach that discretizes the lender's interest rate assignment decision; the details of this approach are provided in Section 4.2 below.

For our analysis, we randomly split our data into two halves, which we refer to as the training set and the testing set. The training set will be used to fit predictive models of returns, which will then be applied to the test set to predict the mean and variance each loan would have in each grade. These predictions on the test set are then used to estimate our discrete choice

	Mean	St. Dev.	Min.	Max.
loan_amnt	14,393.814	8,577.206	500.000	40,000.000
annual_inc	75,523.486	65,616.948	100.000	9,550,000.000
dti	18.043	9.424	-1.000	999.000
delinq_2yrs	0.316	0.873	0.000	39.000
inq_last_6mths	0.714	1.005	0.000	33.000
open_acc	11.560	5.381	0.000	90.000
pub_rec	0.212	0.594	0.000	86.000
revol_bal	16,153.393	22,204.127	0.000	2,904,836.000
total_acc	25.310	11.993	1.000	176.000
collections_12_mths_ex_med	0.016	0.141	0.000	20.000
acc_now_delinq	0.005	0.078	0.000	14.000
chargeoff_within_12_mths	0.009	0.108	0.000	10.000
delinq_amnt	14.401	751.569	0.000	94,521.000
tax_liens	0.049	0.387	0.000	85.000
int_rate_num	0.135	0.047	0.053	0.310
emp_length_num	5.721	3.709	0.000	10.000
term_years	3.491	0.861	3.000	5.000

Table 1: Summary statistics of non-categorical variables. Descriptions of each variable can be found in Lending Club's data dictionary at `https://www.lendingclub.com/info/download-data.action`.

model. The purpose of this split is to avoid overestimating the accuracy with which the lender can forecast mean and variance of returns on a loan, and to better model the lender's own assignment decision, which must always be made "out-of-sample" for new borrowers.

4 A Model of Returns

In this section, we present a simple model of returns from loans, and use it to derive a methodology for predicting mean and variance for new loan applicants, for each possible loan grade.

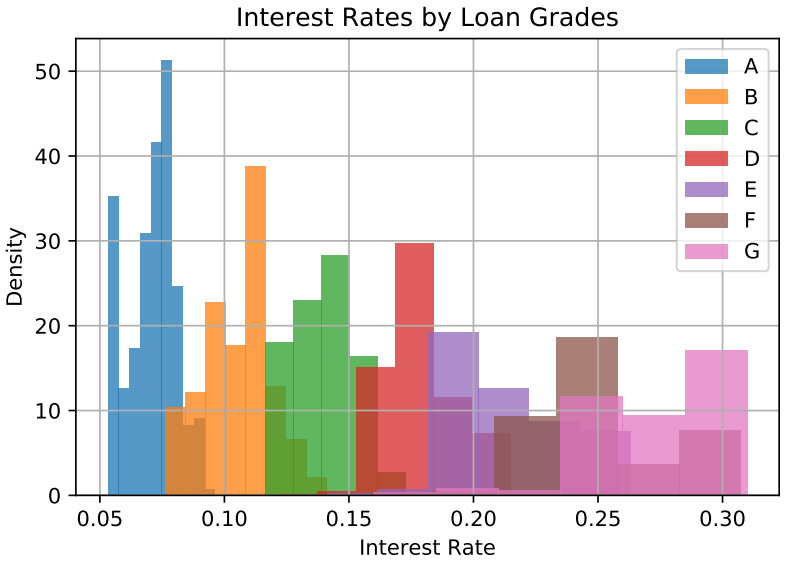
4.1 Setup

Our setup loosely follows that presented by Mencía (2012). Let the subscript i index borrowers, D_i be the indicator that i defaults,[1] ϕ_i be the annualized gross recovery rate[2] received from i if they default, and r_i be the annualized interest rate on i's loan. Then the lender's annualized

[1] We count both defaults and charge-offs in the data as defaults. Lending Club defines as defaults loans for which payments are 120+ days late, and defines as charge-offs loans which are 150+ days late and are not expected to make any future payments.

[2] The gross recovery rate is the ratio of the total amount received on a defaulted loan to the total amount given by the lender, up to the present date. As such, it includes both amounts obtained through debt collection after default and payments made by the borrower before default. To annualize the recovery rate, we subtract 1 from it, divide by the term length of the loan, and add 1.

Figure 1: Distribution of interest rates within each loan grade.



gross rate of return from lending to borrower i at interest rate r_i is

$$R_i = (1 - D_i)(1 + r_i) + D_i\phi_i. \tag{1}$$

In the above equation, D_i and ϕ_i are both random variables, jointly following some unknown distribution F which may depend on a vector of borrower i's personal characteristics, X_i, and on the interest rate given to i:

$$(D_i, \phi_i) \sim F(X_i, r_i). \tag{2}$$

Using equations (1) and (2), one could derive the distribution of returns R_i for each interest rate r_i if one knew F. The lender, if they knew F, could select for each borrower the interest rate which would give the most attractive return distribution, depending on the lender's investment preferences. Deriving the exact distribution F would require a highly sophisticated microeconomic model to fully account for how a borrower's age, education, employment, debt level, income, and other individual characteristics determine the distribution of the stream of payments they can make to the lender at every interest rate. Given the wide variety of borrower characteristics observed by lenders, as well as the large amount of parametric assumptions such a microeconomic model would require, this approach is largely infeasible for our purposes. We as econometricians cannot know the full distribution F, and suspect that the lender itself can at best model it imperfectly.

Fortunately, portfolio theory dating back to Markowitz (1952) suggests that investors need not know the entire distribution of returns to make investment decisions; it should suffice to know the mean and variance of returns. Consequently, we will assume that the expected utility the lender receives from lending to borrower i at rate r_i is

$$EU_i(r_i) = \mathrm{E}\left[R_i | X_i, r_i\right] - \frac{\lambda}{2}\mathrm{Var}\left[R_i | X_i, r_i\right], \tag{3}$$

where λ is the coefficient of absolute risk aversion.[3] The lender's problem is to choose r_i for each i to maximize the above quantity.

The above expression may be derived by taking a second-order Taylor expansion of the lender's utility function; see Appendix A.1 for details.

Those who have studied portfolio theory may immediately raise a few important objections to the form of equation (3); we address these now.

4.1.1 Where is the target mean return?

The mean-variance utility optimization problem

$$\max_w \left\{ w'\mu - \frac{\lambda}{2} w'\Sigma w \right\} \text{ s.t. } w'\mathbf{1} = 1$$

and the more traditional Markowitz (1952) variance minimization problem

$$\min_w \left\{ w'\Sigma w \right\} \text{ s.t. } w'\mu = \mu_0, w'\mathbf{1} = 1$$

are equivalent, in the sense that both produce solutions on the same mean-variance efficient frontier (Bodnar et al. (2013)). Varying the risk parameter λ or the target return μ_0 traces out portfolios on this frontier. The form of the expected utility given in equation (3) is based on the mean-variance utility optimization problem. We choose such a form rather than one based on the classical variance minimization problem because the objective of former problem is easier to interpret as the lender's expected utility, and because such an objective allows for econometric identification of the lender's risk preferences.

4.1.2 Where are the covariances?

Given that we have nearly one million individual loans in our final data set, we would need to estimate nearly half a trillion covariances to model a complete portolio optimization problem. Moreover, to model the lender's decision in assigning interest rates, we would need to consider the covariance between each set of two borrowers for every possible assignment of interest rates to the two, further multiplying the number of parameters to be estimated.

It is well known that return covariance estimates are heavily error prone, even when years of data are available. Moreover, when the number of assets under consideration grows large, the covariance estimates with the largest errors tend to grossly distort the optimal portfolio weights, leading to highly suboptimal portfolios (Michaud (1989)). In our context, estimating half a trillion (or some multiple of that) covariances based solely on individual borrower characteristics would almost certainly introduce extreme amounts of estimation error in some cases which would come to dominate the results of our analysis.

[3] Because we are dealing with preferences over the gross return from each borrower (rather than the actual payout from a borrower), we have effectively normalized the lender's initial wealth to 1, so that the notions of absolute and relative risk aversion will coincide.

In addition to this estimation error problem, there are also theoretical grounds for omitting consideration of variances. The lender's aggregate portfolio, consisting of hundreds of thousands of individual loans, is highly "granular." Mencía (2012) has shown that as the granularity of a loan portfolio increases to infinity, the diversifiable risk in that portfolio goes to zero. In a portfolio consisting of many individual loans, the non-diversifiable risk dominates the diversifiable risk, meaning that estimating covariances to perfectly diversify the portfolio becomes less important.

For these reasons, as well as for computational tractability, we assume that the utility the lender receives from a given vector of interest rate assignments for each borrower is separable across borrowers. This means that the lender can obtain their most preferred vector of interest rate assignments by individually choosing the interest rate for each borrower to maximize the expected utility from that borrower, following equation (3).

That said, it is possible to consider correlations on aggregate levels of loans, for example the correlations between portfolios consisting of all loans of a particular grade. Such consideration is useful in analyzing the problem faced by Lending Club's end investors, who allocate their money across such buckets of loans. This problem has been considered by Polák (2017) using the same Lending Club data as this paper.

4.1.3 Where are the weights?

Portfolio weights do not appear in the lender's utility function because Lending Club itself does not decide how much to lend to any particular borrower; it simply decides whether or not to lend the amount requested by the borrower (which is usually in the \$5,000–\$40,000 range). Moreover, our data on declined loans is severely limited, so that we focus only on the lender's decision in assigning interest rates to loans it has already decided to fund. This modelling decision leaves the lender with no control over the exact weight of any particular loan in the portfolio.

Because we have also assumed that the lender's optimization problem is separable across borrowers, we imagine that the lender individually tries to attain the best risk-return characteristics on a number of single asset portfolios, each consisting of one loan with weight $w = 1$. Consequently, weights do not appear in the utility function in (3).

This assumption is perhaps made more reasonable by remembering that Lending Club is not the final holder of the loans. Lending Club acts as an intermediary, selling these loans to retail investors, and as such cares about providing individually attractive loans to its investors, rather than crafting a full optimized portfolio of loans.

4.1.4 What about higher moments?

The derivation of equation (3) given in Appendix A.1 is based on a second order Taylor expansion, but there is *a priori* no reason to stop at the second order term. Taking a third-, fourth-, or higher-order Taylor expansion would result in the third-, fourth-, and higher-order central moments of the return R_i appearing in the utility function. Such moments correspond to the

skewness, kurtosis, or higher-order non-normalities of the return distribution. The problem of portfolio optimization with higher moments has been considered frequently in the literature (see, for example, Jondeau and Rockinger (2006) or Harvey et al. (2010)). Moreover, it is reasonable to imagine that returns on personal loans have highly non-normal distributions, making higher moments meaningful for our study.

It is both theoretically and practically straightforward to extend our analysis to include higher moments. Following the methodology outlined below in Section 4.2, we could build predictive models of skewness and kurtosis in addition to those for mean and variance, and use those predictive models to study the lender's preferences over these additional features of the return distribution. However, when we formed such models, we found that predicted skewnesses and kurtoses were highly correlated with predicted means and variances. The sample correlation of variance and skewness predictions was -0.861, while the correlation of mean and kurtosis predictions was -0.863. (For comparison, the sample correlation of mean and variance predictions was only 0.047.) Due to these high correlations, including skewness and kurtosis terms in the lender's utility leads to strong multicollinearity problems in estimating our discrete choice model.

We interpret these high correlations to mean either that we cannot estimate skewness and kurtosis well enough to properly identify these features in the return distribution, or that mean and variance alone provide enough information about the distribution of returns. Either way, we consider only mean and variance in our analysis of the lender's problem.

4.2 Predicting Mean and Variance of Returns

Using the model setup given in the previous section, we now turn to the problem of forecasting mean and variance of the returns from a borrower assigned to a particular grade. From equation (1), it is straightforward to derive

$$\mathrm{E}\left[R_i|X_i,r_i\right] = (1+r_i)(1-\mathrm{E}\left[D_i|X_i,r_i\right]) + \mathrm{E}\left[\phi_i|D_i=1,X_i,r_i\right]\mathrm{E}\left[D_i|X_i,r_i\right], \tag{4}$$

$$\begin{aligned}\mathrm{Var}\left[R_i|X_i,r_i\right] = {} & (1+r_i)^2(1-\mathrm{E}\left[D_i|X_i,r_i\right])\mathrm{E}\left[D_i|X_i,r_i\right] \\ & + \mathrm{E}\left[\phi_i^2|D_i=1,X_i,r_i\right]\mathrm{E}\left[D_i|X_i,r_i\right] - \mathrm{E}\left[\phi_i|D_i=1,X_i,r_i\right]^2\mathrm{E}\left[D_i|X_i,r_i\right]^2 \\ & - 2(1+r_i)\mathrm{E}\left[\phi_i|D_i=1,X_i,r_i\right]\mathrm{E}\left[D_i|X_i,r_i\right](1-\mathrm{E}\left[D_i|X_i,r_i\right]);\end{aligned} \tag{5}$$

see Appendix (A.2) for details. In the above, the "$|r_i$" notation does not denote conditioning on r_i, which is a nonrandom constant chosen by the lender; rather, it denotes that the expectations are being taken for a fixed value of r_i, which is a parameter of the distribution of (D_i, ϕ_i) following equation (2).

Notice that equations (4) and (5) depend only on the conditional expectations $\mathrm{E}\left[D_i|X_i,r_i\right]$, $\mathrm{E}\left[\phi_i|D_i=1,X_i,r_i\right]$, and $\mathrm{E}\left[\phi_i^2|D_i=1,X_i,r_i\right]$. Thus, if one has predictive models of these conditional expectations, one can plug them into equations (4) and (5) to predict mean and variance of returns for a given borrower under a given interest rate.

In all of the above, r_i is a continuous variable chosen by the lender for each borrower.[4] But studying the lender's decision problem across a range of continuous variables is inherently challenging. One might hope to derive first order conditions for the problem of maximizing the expected utility given in (3), but doing so would require knowing exactly how $\mathrm{E}\left[R_i|X_i, r_i\right]$ and $\mathrm{Var}\left[R_i|X_i, r_i\right]$ depend on r_i. But knowing the form of this dependence requires knowing the exact distribution F, which is unfeasible.

The continuity of r_i also presents problems in forming predictive models of $\mathrm{E}\left[D_i|X_i, r_i\right]$, $\mathrm{E}\left[\phi_i|D_i = 1, X_i, r_i\right]$, and $\mathrm{E}\left[\phi_i^2|D_i = 1, X_i, r_i\right]$. We would need to make some assumption about the form in which r_i enters these conditional expectations functions. The problem of maximizing (3) would be very sensitive to the form in which r_i enters these functions, and we cannot justify any particular parametric assumption about how r_i should enter.

Motivated by the observation that interest rates cluster around grade averages (see Figure 1), we choose to confront these difficulties by discretizing the interest rates as follows. For each grade $g \in \{A, \ldots, G\}$, we let r_g denote the mean interest rate given to borrowers in grade g. We approximate the lender's continuous interest rate decision as the discrete decision of assigning a borrower to a grade A through G and charging them the average interest rate of that grade. The lender's utility maximization problem is now the discrete problem

$$\max_{g \in \{A,\ldots,G\}} \left\{ \mathrm{E}\left[R_i|X_i, r_g\right] - \frac{\lambda}{2}\mathrm{Var}\left[R_i|X_i, r_g\right] \right\} \tag{6}$$

for each borrower i. By discretizing the interest rate in this way, we lose the ability to account for interest rate variation within each loan grade, but we also greatly simplify the problem. We now only need to model the dependence of the conditional expectation functions $\mathrm{E}\left[D_i|X_i, r_i\right]$, $\mathrm{E}\left[\phi_i|D_i = 1, X_i, r_i\right]$, and $\mathrm{E}\left[\phi_i^2|D_i = 1, X_i, r_i\right]$ on X_i and seven discrete values of r_i. Additionally, once we have developed such functions, we will be able to use a straightforward discrete choice framework for estimating λ.

For simplicity, we assume linear forms for the conditional expectation functions of ϕ and ϕ^2, and a logistic form for those of D:

$$\mathrm{E}\left[\phi_{i,g}|D_{i,g} = 1, X_i, r_g\right] = X_i'\beta_g^{(\phi)}, \tag{7}$$

$$\mathrm{E}\left[\phi_{i,g}^2|D_{i,g} = 1, X_i, r_g\right] = X_i'\beta_g^{(\phi^2)}, \tag{8}$$

$$\mathrm{E}\left[D_{i,g}|X_i, r_g\right] = \sigma(X_i'\beta_g^{(D)}), \tag{9}$$

where σ is the sigmoid functions $\sigma(z) = \frac{1}{1+e^{-z}}$. The regression equations (7), (8), and (9) are *not* intended to be causal, nor are they thought to be completely representative of how Lending Club

[4]Lending Club in fact advertises its interest rates as discrete, with one interest rate per loan subgrade (finer classifications into which the loan grades A through G are divided). While this is true at any fixed time, Lending Club frequently shifted these discrete rates by varying amounts and in varying directions over the time period of our data. Consequently, there are borrowers who obtained loans at similar times and were placed into the same subgrade, but were given slightly different interest rates. In light of this fact, we think of Lending Club's interest rates over time as being continuously rather than discretely selected. Archival snapshots of Lending Club's rates at various points in time can be found at `https://web.archive.org/web/*/https://www.lendingclub.com/public/rates-and-fees.action`.

or any other online lender actually makes its predictions. The true causal relationships are likely much more complex than the above functional forms, and large online lenders almost certainly use more sophisticated models than linear or logistic regressions in making predictions. We choose these simple forms because they are computationally easy to estimate and have limited risk of overfitting or data, leading to better extrapolation in predicting counterfactual means and variances. The possibility of using more sophisticated forms for the conditional expectation functions will be considered in Section 5.3.

In order to estimate equations (7), (8), and (9), we assume

Assumption 1. *Such* $\beta_g^{(\phi)}$, $\beta_g^{(\phi^2)}$, *and* $\beta_g^{(D)}$ *exist for all* $g \in \{A, \ldots, G\}$, *and do not depend on* X_i.

We illustrate this assumption with an example. Suppose all borrowers in grade G have annual incomes under $80,000, while all borrowers in grade A have annual incomes above $80,000. Assumption 1 implies that if some borrower i in grade A had instead been put in grade G, on average ϕ_i, ϕ_i^2, and D_i would depend on i's income in the same way as it does for borrowers with incomes under $80,000. Essentially, Assumption 1 says that we can extrapolate the relationship between income and default and recovery rates that *any* potential borrower would have in grade G just by looking at the relationship observed for individuals who were actually assigned to grade G.

Assumption 1 is rather strict and difficult to verify empirically as it deals entirely with counterfactuals.[5] We partially justify it by arguing that though the true data generating process may violate the assumption, the lender itself would need to make some similar assumption to predict counterfactual mean and variance of returns. This is because the only data that the lender has about returns for individuals given, say, G grade loans comes from individuals who were actually given G grade loans. If the lender wanted to forecast the mean and variance of returns a new borrower would have on a G grade loan, the lender would have little choice but to base this forecast on past data of borrowers who have received G grade loans. Consequently, the predictions we will produce using Assumption 1 may be biased, but should at least be biased in the same ways as the predictions made by the lender itself would be. Because we are ultimately interested in how lenders make grade assignments based on their own (potentially biased) predictions of mean and variance, making Assumption 1 should not invalidate our results, even if the assumption is incorrect.

The main value of Assumption 1 is that it allows us to estimate all the β's by running regressions on subsets of our data. Specifically, we can estimate equations (7) and (8) by running a linear regression of recovery rate and its square on personal characteristics for individuals in

[5]Of course, there exists a wide range of econometric tools for dealing with counterfactuals. A regression discontinuity method could be useful in attempting to empirically verify Assumption 1. One could consider borrowers on the edge between two grades, and test the hypothesis that their β's are close to those of other borrowers in their grades. Rejecting this hypothesis would be tantamount to rejecting Assumption 1. We have not pursued this analysis, though it could in itself be an interesting topic for future research.

grade g who have defaulted, and we can estimate equation (9) by running a logistic regression of a default indicator on personal characteristics for all individuals in grade g. Our large data set makes such regressions on subsets feasible: even the smallest subset on which we run a regression still contains over $1,700$ observations, allowing for reasonably precise parameter estimates.[6]

The results of the linear regressions of ϕ and ϕ^2 and logistic and linear regressions of D for each loan grade are given in Tables 3, 4, 5, and 6 in Appendix A.3. Though we cannot interpret any of the estimated parameters as causally meaningful, studying the tables allows for some important observations about how the prediction models are working; these are discussed in Section 4.4 below.

4.3 Omitted Variables

There is reason to believe that the regressions described in Section 4.2 will suffer from omitted variable problems. We know, for example, that Lending Club observes FICO credit scores of borrowers, but does not include these in their publicly available data. There may be other such features observed by the lender but not the econometrician. In addition, there are features reported by Lending Club which do not enter into our regressions, such as the text description each borrower writes about why they are applying for a loan. We as econometricians are not able to include such data in our regressions, but the lender may have algorithms or human readers which can extract information from such variables.

We therefore consider in this section the effect of omitted variables on our prediction models. Suppose we as econometricians only observe personal characteristics X_i, while the lender observes both X_i and additional characteristics Z_i. The regression equation the lender uses to predict, say, ϕ_i would take the form

$$\phi_i = X_i'\beta + Z_i'\gamma + \varepsilon_i.$$

If we regress ϕ_i on X_i alone, we obtain the biased and inconsistent estimate

$$\hat{\beta}^{\text{OV}} = \beta + \left(\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} X_i Z_i'\right)\gamma + \left(\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} X_i \varepsilon_i\right).$$

Suppose we also write a vector regression equation of Z_i on X_i:

$$Z_i' = X_i'\delta + \eta_i',$$

where δ is a dimension$(X) \times$ dimension(Z) matrix and η_i is a dimension$(Z) \times 1$ error. The above equation is defined so that $X_i'\delta$ is the best linear of Z_i, so η_i is necessarily uncorrelated with X_i. OLS regression of Z_i on X_i will therefore produce a consistent estimate of δ:

$$\left(\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} X_i Z_i'\right) \xrightarrow{p} \delta.$$

[6]This size consideration is also the main reason we do not model interest rate assignments as a discrete choice problem over all loan *sub*grades: doing so would require regressions on much smaller subsets, leading to imprecisely estimated β's and unreliable prediction models.

If X_i and ε_i are assumed to be uncorrelated, then the above equation for $\hat{\beta}^{\text{OV}}$ shows that

$$\hat{\beta}^{\text{OV}} \xrightarrow{p} \beta^{\text{OV}} = \beta + \delta\gamma.$$

Now suppose that we observe personal characteristics X_j for some new borrower j and wish to predict ϕ_j. Using our biased estimation of β, we would in the probability limit predict

$$\begin{aligned}
\phi_j^{\text{predicted}} &= X_j'\beta^{\text{OV}} \\
&= X_j'(\beta + \delta\gamma) \\
&= X_j'\beta + (Z_j' - \eta_j')\gamma \\
&= X_j'\beta + Z_j'\gamma - \eta_j'\gamma.
\end{aligned}$$

Note that $X_j'\beta + Z_j'\gamma$ is, in the probability limit, exactly the prediction of ϕ_j the lender would make from observing both X_j and Z_j. Our prediction is the same as the lender's prediction, except with an added mean zero error term $-\eta_j'\gamma$. In other words, our prediction is an unbiased noisy proxy for the prediction made by the lender, even though our estimate of β is biased and the lender observes variables we do not.

The degree of the noise in our predictions depends on the magnitude of η_i, which may be thought of as the "orthogonal information" possessed by the lender and not the econometrician: it is precisely the component of the unobserved features Z_i which cannot be explained by the observed features X_i. While we cannot estimate the variance of η_i, we have reason to believe that this orthogonal information is typically small, or equivalently that the unobserved features are predictable from the observed features. For instance, though we do not observe FICO credit scores, the features we do observe (income, employment duration, debt-to-income ratio, various credit history factors, etc.) are likely strong predictors of FICO scores.

Note that the argument of this section depends entirely on the algebraic properties of *linear* regression estimates, and does not necessarily carry over to the *logistic* regressions used to predict default rates. Consequently, the predictions generated by our baseline model may not be unbiased proxies for the predictions made by the lender. To overcome this hurdle, we will also consider an alternative model in which linear regression is used to predict default rates.

4.4 Results from Estimation of Predictive Models

In this section, we give the results of the regressions outline in Section 4.2. We begin with the estimation of equation (7), the results of which are shown in Table 3. Each column in the table corresponds to a separate regression for each loan grade of borrower and loan characteristics on the recovery rate ϕ. For completeness, we also include regressions on all loan grades in the last column. We observe that the signs and magnitudes of estimates for each parameter are similar across all loan grades, suggesting that these features have comparable effects on recovery rates across grades. R^2 values, especially for lower grades, are all fairly high, and the differences between adjusted and non-adjusted R^2s suggest that the models are not badly over fitting the data. Both R^2 and adjusted R^2 values are higher for the lower grade loans, suggesting that

the recovery rates of lower grade borrowers are more predictable from the data than are rates for higher grade borrowers. Sample sizes are large across grades, allowing for precise parameter estimates.

Table 4 presents results from the estimation of equation (8). Similar observations regarding the magnitudes and signs of parameters and R^2 and adjusted R^2 values apply to these results as well.

Finally, Tables 5 and 6 present results from regressing defaults on personal characteristics, using both logistic and linear regressions. The logistic regressions were estimated with a small L_2 regularization term to deal with occasional numerical errors.[7] Again, magnitudes and signs of parameters are generally close across grades for both specification forms. We also see R^2, adjusted R^2, and McFadden pseudo-R^2 are higher for the lower grade loans.[8] However, the R^2 values in these regressions are all much lower than those obtained in the recovery rate regressions, suggesting that default rates are generally harder to predict than recovery rates.

As mentioned previously, we should not interpret the results of these models as causal. These regressions serve as elements of our predictive models of the economically relevant features of the loans: their return's estimated mean and variance. These tables and descriptions of results are given primarily to assess how well the regressions appear to be working.

As a sanity check of our prediction models, we plot loans from the test set in estimated mean-variance space in Figure 2 below. The color of each point indicates the grade to which the loan was actually assigned, and the estimated mean and variance are the estimates produced by our models for that grade. For visibility, we plot 4,000 loans from each grade. We observe that loans in the riskier grades are predicted to have both higher mean returns and higher variance, as one would expect. Additionally, we notice that almost all loans are predicted to have positive variance, even though the structure of our prediction models does not guarantee this property. These observations provide reassurance that our prediction models are reasonable.

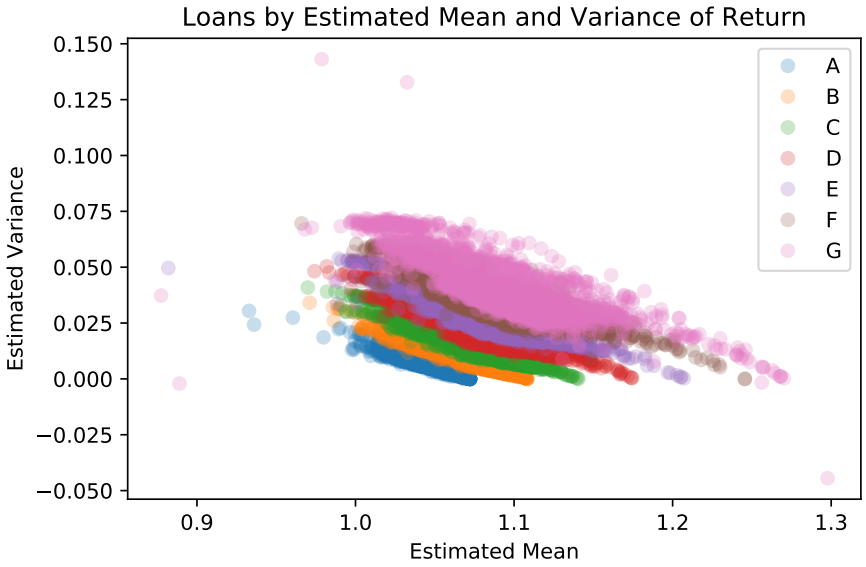
5 Discrete Choice

To analyze how predicted mean and variance affect the lender's grade decisions, we introduce a latent-utility discrete choice model. Following equation (3), we model the lender's utility from assigning borrower j to grade g as a linear function of the predicted mean and variance of the return from borrower j if they are assigned to grade g.

[7] Because we must regress on multiple subsets of data, and our data includes multiple dummy variables, we occasionally encounter perfect separability issues. For example, in the training set no borrowers from Iowa defaulted. When we try to fit a logistic regression to this data, the optimizer tries to push the coefficient on the Iowa dummy variable to $-\infty$, causing a failure to converge. Adding an L_2 regularization penalty stops this problem.

[8] The McFadden pseudo-R^2 is defined as $1 - \frac{\text{log likelihood of null model}}{\text{maximized log likelihood of full model}}$.

Figure 2: Loan grades in estimated mean-variance space.



5.1 Model

When the lender encounters a new borrower, they do not know the exact distribution of that borrower's potential returns for every interest rate. As shown in the previous sections, however, the lender can predict return mean and variance from borrower and loan characteristics, using predictive models developed by observing other borrowers. Formally, the lender can construct $M_{j,g}$ and $V_{j,g}$ for each borrower j and grade g. The lender then wishes to assign the loan to the grade which will maximize their expected utility, by solving

$$\max_{g\in\{A,\ldots,G\}} \left\{ \mathrm{E}\left[R_j|r_g\right] - \frac{\lambda}{2}\mathrm{Var}\left[R_j|r_g\right] \right\}. \tag{10}$$

The lender does not know the true $\mathrm{E}\left[R_j|r_g\right]$ and $\mathrm{Var}\left[R_j|r_g\right]$, but has predictions of these values, and can solve

$$\max_{g\in\{A,\ldots,G\}} \left\{ M_{j,g} - \frac{\lambda}{2}V_{j,g} + e_{j,g} \right\}, \tag{11}$$

where $e_{j,g}$ is an unobserved idiosyncratic preference shock. If we assume that the $e_{j,g}$ are independent and identically distributed from a Gumbel distribution,[9] we obtain the conditional logit model

$$\Pr\left\{j \text{ assigned to } g\right\} = \frac{\exp(X'_{j,g}\beta)}{\sum_{g'\in\{A,\ldots,G\}}\exp(X'_{j,g'}\beta)} \tag{12}$$

[9]There is no theoretical reason to assume the errors follow a Gumbel distribution; we make this assumption solely to obtain a convenient analytic form for the problem.

where $X_{j,g} = (M_{j,g}, V_{j,g})'$ and $\beta = (\beta_1, \beta_2)'$. Note that we can recover the risk aversion parameter λ from β as $\lambda = -2\frac{\beta_2}{\beta_1}$. This model has log likelihood function

$$l(X, \beta) = \sum_{i=1}^{N} \left(\sum_{g \in \{A,\ldots,G\}} 1\{i \text{ assigned to } g\} X'_{i,g}\beta - \log \sum_{g \in \{A,\ldots,G\}} \exp(X'_{i,j}\beta) \right), \qquad (13)$$

and numerical optimization of the log likelihood function allows us to estimate β and λ.

5.2 Estimation Procedure

We use maximum likelihood to estimate the parameters of our model. Since our analysis involves two stages (estimating predictive models and then using those predictions to estimate a discrete choice model), analytic expressions for the standard errors of our estimates are difficult to derive. Instead, we use a bootstrap procedure to estimate standard errors. We bootstrap both the estimation of the predictive models and the estimation of the discrete choice model simultaneously in each iteration. We used 100 bootstrap iterations.

Because $\hat{\beta}_1$ and $\hat{\beta}_2$ are maximum likelihood estimates, they will be asymptotically normally distributed, so computing the sample standard deviation of these estimates across bootstrap samples will provide a good estimate of the true standard errors. In contrast, $\hat{\lambda} = -2\frac{\hat{\beta}_2}{\hat{\beta}_1}$ is asymptotically the ratio of two normal random variables. The distribution of such a quantity can be thought of as a generalization of the Cauchy distribution, and the moments of such a distribution will generally be undefined.[10] Sample standard deviations will thus not accurately describe the variability in $\hat{\lambda}$. Instead, we report the range from the 25th to 75th percentile of $\hat{\lambda}$ across bootstrap samples.

We consider two specifications. The main specification is based on equations (7), (8), and (9). The alternative specification uses a linear regression to predict default D, and is based on the discussion given in Section 4.3.

5.3 Results from Estimation of Discrete Choice Model

The results of running maximum likelihood estimation on our baseline model—in which default rate is predicted with a logistic regression and recovery rate and squared recovery rate are predicted with linear regressions—are given in column (1) of Table 2.

We find a slightly significant positive coefficient on mean return M and a significant negative coefficient on return variance V. Because the features we construct are noisy proxies for the features constructed by the lender itself, we suspect that these parameters are more significant then the bootstrapping suggests. The signs of the coefficients match our intuition that the lender likes higher returns and lower variance. The McFadden pseudo-R^2 of 0.110 suggests that the model provides a reasonably good improvement over the empty null model (at least in terms of log likelihood). Both the magnitudes and the significance levels of the estimated coefficients on

[10] For a more complete account of such distributions, see Marsaglia (1965).

Table 2: Results of discrete choice model estimation.

	(1)	(2)
M	4.990^{*}	-8.821^{**}
	(2.176)	(2.064)
V	-52.108^{**}	-35.041
	(2.488)	(13.857)
λ	20.887^{*}	-7.944
	(398.788)	(3.789)
	IQR: $(23.154, 47.283)$	IQR: $(-8.406, -2.465)$
$-2\log(\text{LR})$	205546	165674
R^2_{McFadden}	0.110	0.089
% Correct	16.7%	16.2%

$p < 0.05^{*}, p < 0.01^{**}, p < 0.005^{***}$.

Bootstrapped standard errors in parentheses.

M and V suggest that the lender shows more evidence of variance minimization than of return maximization.

Indeed, our coefficients lead to a slightly significant point estimate of $\hat{\lambda} = 20.887$. This estimate is substantially larger than usual estimates of the risk aversion coefficient, which range from 1 to 10 (Ang (2014)).[11] Our maximum likelihood estimates therefore suggest that the lender acts as a highly risk averse agent in assigning borrowers to loan grades.

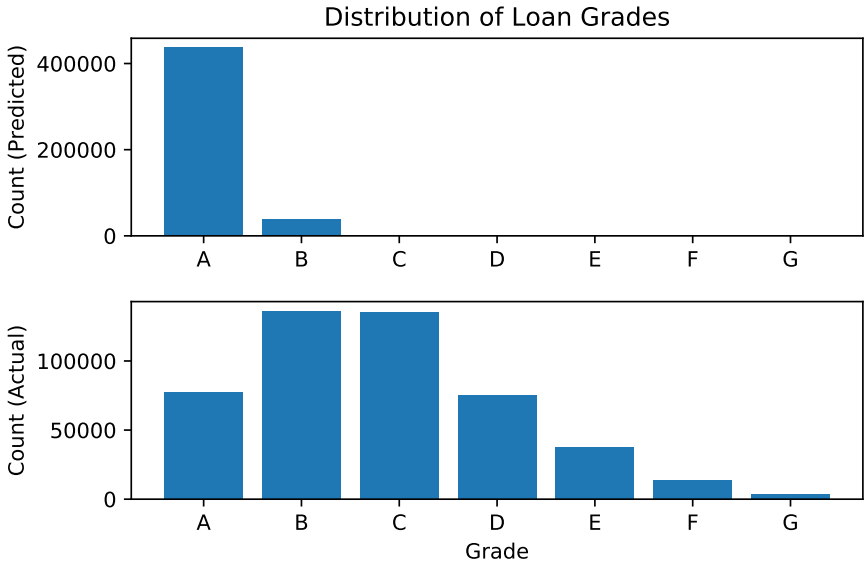
Though the McFadden pseudo-R^2 metric suggests a good fit, the percent of loans whose grades are correctly predicted by the model is disappointingly low. Upon closer inspection, we find that the discrete choice model predicts that the vast majority of loans should go to grade A. We plot in Figure 3 the distribution of loan grades predicted by our model against the actual distribution of grades in the data. The fact that so many loans are predicted to go to the safest grade is consistent with our estimate that the lender is highly risk-averse, but clearly does not fit the observed distribution, which centers around grades B and C.

There are many possible explanations for why the model would overpredict the number of borrowers that will be matched to grade A. From an economic perspective, it may be that the lender's optimization problem has constraints on the quantities of high-grade loans that can be given. From an empirical perspective, the conditional logit model's log likelihood function may be the wrong criterion to maximize if one's goal is to recreate the observed distribution of loan grades. Optimizing a different cost function could provide a better fit to the loan grade distribution, but the results of such an optimization would not tell us anything about the lender's risk preferences.

This failure of our model to account for the large quantity of low-grade loans given suggests that the lender's grade assignment cannot be fully explained by mean-variance optimization of

[11] Janěcek (2004) finds that risk aversion coefficients may be on the order of 30 or more, so estimates of this magnitude are not unheard of.

Figure 3: Predicted vs. actual loan grade distributions in the test data.



each loan. Some factors other than mean and variance of loan returns must be important in the lender's decision, leading to the assignment of many more low-grade loans than mean-variance optimization would imply.

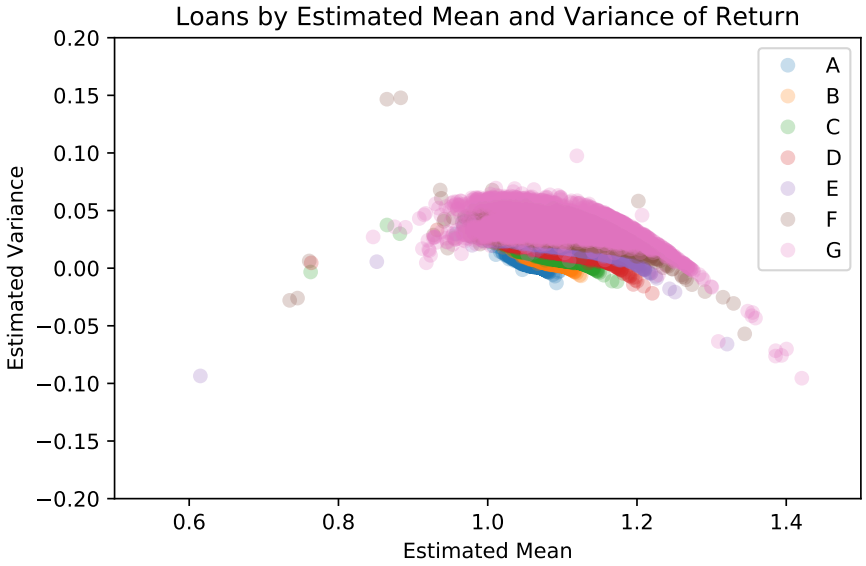
To see how the model performs on data from different loan years, we attempted to replicate the loan grade decision process made by the lender. For each loan in the data, we first use our predictive models from Section 4.2 to calculate for each borrower their predicted mean return and variance for each possible grade. Using these values and the parameters estimated from the discrete choice model, we assign utilities for each loan being assigned to each grade. We then run a greedy (myopic) algorithm which goes through each row in the data and assigns a loan to the grade which yields the highest utility. The algorithm repeats this process until each grade has hit its capacity (in this case, the actual number of loans assigned to that grade in the data). The results of this procedure are similar to prior results in that we overestimate the amount of loans that are assigned to higher grades and are unable to accurately predict which loans should go to lower grades. There is perhaps some room for improving our predictions by using an algorithm which is not myopic but instead looks ahead to find the global optimum of the system, but such a procedure would be computationally expensive for the number of loans in our data set.

We next turn to the results of the alternative specification in column (2) of Table 2, which uses a linear regression to predict defaults. While we again estimate a large negative coefficient on V, the coefficient on M and the estimate of λ have changed from positive to negative. These estimate makes little economic sense, suggesting that the lender is risk-loving and dislikes higher returns. The likelihood ratio statistic, McFadden pseudo-R^2, and percent of loans correctly predicted all suggest that model (2) fits the data slightly worse than model (1). Model (2) exhibits the same problem of overpredicting the number of A grade loans, and suffers from it

even worse than model (1), predicting that 99.7% of loans should go to grade A. (Model (1) predicts that 91.2% of loans will go to grade A.)

To understand the worse performance of model (2), we recreate Figure 2 using model (2). We see an increase in the number of loans which are outliers in terms of mean and variance, as well as a breakdown of the clean separations between the grades. These observations suggest that model (2) does not generate reasonable predictions of loan return and variance, and consequently we do not place much faith in the results of estimating this model.

Figure 4: Loan grades in estimated mean-variance space, using model (2).



In inspecting equations (4) and (5), we see that if a predicted default rate is not between 0 and 1, the signs of terms in the equations can change, resulting in unreasonable predictions of mean and variance of returns. This may partially explain the poor performance of model (2)'s mean and variance predictions, as using a linear regression to predict D can result in predictions above 1 or below 0.

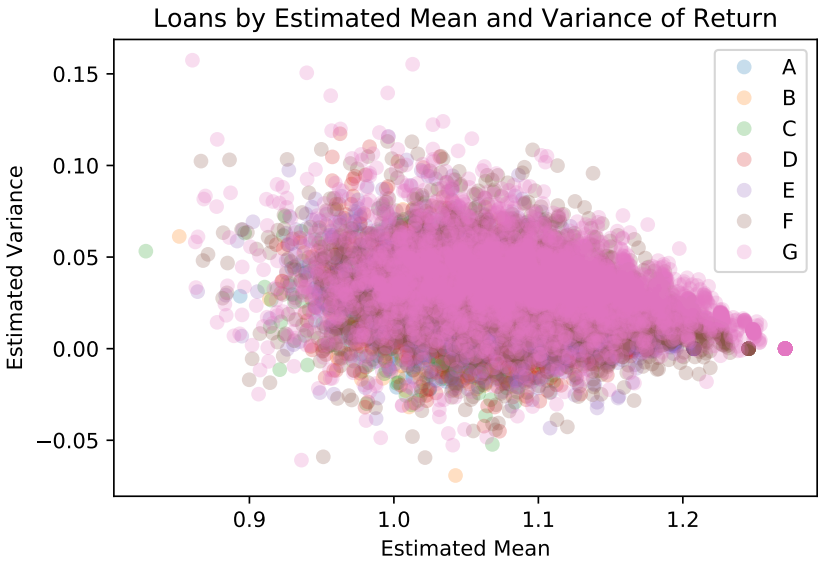
Though we do not give much weight to the results of model (2), the fact that the estimates of λ differ so much between the two models suggests that the results of our baseline model may be highly sensitive to misspecification. The estimates of the coefficient on V, however, are comparable between the two specifications, suggesting that this estimate is at least somewhat robust to model misspecification.

To further explore the robustness of our results against model misspecification, we considered alternative predictive models for ϕ, ϕ^2, and D. Specifically, we fit predictive models using random forest regressions to estimate predictive models of these three quantities for each loan grade. We choose random forest regression for three main reasons: first, it is a highly flexible nonparametric model, contrasting the more rigid logistic and linear regressions used in our main specification; second, it is computationally easier to fit than many other machine learning models; and third, bagging methods have been found to be effective in consumer credit scoring

applications (Costangioara (2011)).

Unfortunately, the random forest-based model produced the mean-variance plot shown below in Figure 5. The plot appears to be totally random, suggesting that the random forest model performs very poorly at forecasting mean and variance of returns out of sample. These results leads us to believe that fitting more sophisticated models for predicting ϕ, ϕ^2, and D will in general require much care and calibration, and that simpler linear and logistic regressions may function more reasonably for out of sample prediction.

Figure 5: Loan grades in estimated mean-variance space, using a random-forest based model.



Despite the apparent limitations in the random forest-based prediction model, we estimated our discrete choice model using forecasts generated with random forest regressions. The point estimates found were $\hat{\beta} = (-5.027, -26.730)'$, $\hat{\lambda} = -10.634$, $R^2_{\text{McFadden}} = 0.065$, the signs and magnitudes of which are comparable to those found using model (2). Unfortunately, computational limitations prevented us from bootstrapping standard errors for these estimates.

6 Conclusions

The estimates from our baseline model—which, of the three models considered, appears to give the most reasonable forecasts of out-of-sample mean and variance—show significant evidence that the lender acts as a highly risk-averse agent in assigning borrowers to loan grades. These results support the hypothesis that the lender's behavior in assigning grades is broadly consistent with the mean-variance optimization framework of portfolio theory. However, explorations of alternative models for predicting mean and variance suggest that our results may be highly sensitive to model misspecification.

Additionally, comparison of the actual and model-predicted loan grade distributions shows that the lender assigns many more borrowers to low-grade loans than can be explained by

mean-variance optimization alone. This result suggests that the lender may be subject to some constraint limiting the number of high-grade loans which can be assigned, or that the lender cares about some aspect of each loan that is not reflected in its individual risk and return. Alternatively, we may simply be observing a breakdown of our Assumption 1, which may lead us to exaggerate how safe borrower's debts will be if given high-grade loans.

In developing our predictive models, we also find that recovery rates are much more predictable than default rates, and that both recovery and default rates are more predictable for borrowers in low grades than borrowers in high grades. This evidence points to the counterintuitive result that, in a sense, risky borrowers are more predictable. The default and recovery rates of the riskiest borrowers are the most easily forecasted from personal data. Additionally, it is generally hard to tell whether a borrower will default, but given that they will default, it is much easier to tell how much will be recovered from them.

Many extensions to this paper are possible. It would be useful to study still more forms for the predictive models of ϕ, ϕ^2, and D, to better model how a sophisticated lender can use all the information in the data. One potentially promising form is linear and logistic regressions with polynomial and interaction terms. Unfortunately, due to the number of variables observed, including even second-order polynomial and interaction terms greatly increases the computational cost of fitting the models.

It would also be useful to account for borrower prepayment, which is an important risk in consumer lending. If a borrower's interest rate affects their propensity to prepay, then prepayment risk would enter into our model as another component of returns which the borrower can influence in making grade assignments. A prepayment risk term could enter equation (1), and predictive models could be developed for prepayment rates just as they were for default and recovery rates. The main impediment to this extension is that our data set does not give full time series of borrower payments, so it is difficult to determine which borrowers prepaid and by how much.

Accounting for rejected loans would also improve our analysis. As it stands, our model may suffer from selection bias, in that we do not include borrowers who were declined by the lender. While Lending Club does publish data on declined applicants, it reports many fewer features than are reported about accepted applicants, and some of the features it reports for declined applicants are not reported for accepted applicants. It is therefore quite difficult to compare declined and accepted borrowers using the available data. However, it may at least be possible to do more theoretical work toward understanding how the omission of declined borrowers biases our results.

References

Ang, Andrew (2014) *Asset Management: A Systematic Approach to Factor Investing*, Chap. 2: Oxford University Press, URL: https://EconPapers.repec.org/RePEc:oxp:obooks:9780199959327.

Bodnar, Taras, Nestor Parolya, and Wolfgang Schmid (2013) "On the equivalence of quadratic optimization problems commonly used in portfolio theory," *European Journal of Operational Research*, Vol. 229, pp. 637 – 644, URL: http://www.sciencedirect.com/science/article/pii/S0377221713002105, DOI: http://dx.doi.org/https://doi.org/10.1016/j.ejor.2013.03.002.

Costangioara, Alexandru (2011) "Consumer Credit Scoring," *Journal for Economic Forecasting*, Vol. 0, pp. 162–177, URL: https://ideas.repec.org/a/rjr/romjef/vy2011i3p162-177.html.

Harvey, Campbell R., John C. Liechty, Merrill W. Liechty, and Peter Müller (2010) "Portfolio selection with higher moments," *Quantitative Finance*, Vol. 10, pp. 469–485, URL: https://doi.org/10.1080/14697681003756877, DOI: http://dx.doi.org/10.1080/14697681003756877.

Janěcek, Karel (2004) "What is a realistic aversion to risk for real-world individual investors?", URL: https://www.semanticscholar.org/paper/What-is-a-realistic-aversion-to-risk-for-real-world-Janecek/03d84ee02071f197cd3fcb60330cd5c4df766933, working paper.

Jondeau, Eric and Michael Rockinger (2006) "Optimal Portfolio Allocation under Higher Moments," *European Financial Management*, Vol. 12, pp. 29–55, URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1354-7798.2006.00309.x, DOI: http://dx.doi.org/10.1111/j.1354-7798.2006.00309.x.

Klafft, Michael (2008) "Online Peer-to-Peer Lending: A Lenders' Perspective," *Proceedings of the International Conference on E-Learning, E-Business, Enterprise Information Systems, and E-Government*, pp. 371–375.

Markowitz, Harry (1952) "Portfolio Selection," *Journal of Finance*, Vol. 7, pp. 77–91, URL: https://EconPapers.repec.org/RePEc:bla:jfinan:v:7:y:1952:i:1:p:77-91.

Marsaglia, George (1965) "Ratios of Normal Variables and Ratios of Sums of Uniform Variables," *Journal of the American Statistical Association*, Vol. 60, pp. 193–204, URL: http://www.jstor.org/stable/2283145.

Mencía, Javier (2012) "Assessing the risk-return trade-off in loan portfolios," *Journal of Banking & Finance*, Vol. 36, pp. 1665 – 1677, URL: http://www.sciencedirect.com/science/

article/pii/S0378426612000222, DOI: http://dx.doi.org/https://doi.org/10.1016/j.jbankfin.2012.01.007.

Michaud, Richard O. (1989) "The Markowitz Optimization Enigma: Is 'Optimized' Optimal?" *Financial Analysts Journal*, Vol. 45, pp. 31–42, URL: http://www.jstor.org/stable/4479185.

ΩÖzer and Phillips

Özer, Özalp and Robert Phillips (2012) *The Oxford Handbook of Pricing Management*: Oxford University Press, pp.138-153.

Phillips, Robert (2013) "Optimizing prices for consumer credit," *Journal of Revenue and Pricing Management*, Vol. 12, pp. 360–377, URL: https://doi.org/10.1057/rpm.2013.9, DOI: http://dx.doi.org/10.1057/rpm.2013.9.

Polák, Petr (2017) "Portfolio diversification on P2P loan markets," Master's thesis, Charles University.

Wei, Zaiyan and Mingfeng Lin (2017) "Market Mechanisms in Online Peer-to-Peer Lending," *Management Science*, Vol. 63, pp. 4236–4257, URL: https://doi.org/10.1287/mnsc.2016.2531, DOI: http://dx.doi.org/10.1287/mnsc.2016.2531.

A Appendix

A.1 Derivation of Lender's Expected Utility

Let $u : \mathbb{R} \to \mathbb{R}$ denote the utility function of the lender, and assume u is increasing and twice continuously differentiable. Let W be a real-valued random variable representing the outcome of some risky prospect. By taking a second-order Taylor expansion of u about $\mathrm{E}[W]$, we obtain

$$\begin{aligned} u(W) &\approx u(\mathrm{E}[W]) + u'(\mathrm{E}[W])(W - \mathrm{E}[W]) + \frac{u''(\mathrm{E}[W])}{2}(W - \mathrm{E}[W])^2 \\ &= u(\mathrm{E}[W]) - \mathrm{E}[W]\, u'(\mathrm{E}[W]) + u'(\mathrm{E}[W]) \left[W + \frac{u''(\mathrm{E}[W])}{2u'(\mathrm{E}[W])}(W - \mathrm{E}[W])^2 \right] \\ &= u(\mathrm{E}[W]) - \mathrm{E}[W]\, u'(\mathrm{E}[W]) + u'(\mathrm{E}[W]) \left[W - \frac{\lambda(\mathrm{E}[W])}{2}(W - \mathrm{E}[W])^2 \right], \end{aligned}$$

where $\lambda(\cdot)$ is the Arrow-Pratt index of absolute risk aversion,

$$\lambda(c) = -\frac{u''(c)}{u'(c)}.$$

Note that $u(\mathrm{E}[W]) - \mathrm{E}[W]\, u'(\mathrm{E}[W])$ is a constant and $u'(\mathrm{E}[W])$ is a positive constant. Since preferences are invariant under monotonic increasing transformations, the same preferences represented by the above equation for $u(W)$ are also represented by the function

$$U(W) \approx W - \frac{\lambda(\mathrm{E}[W])}{2}(W - \mathrm{E}[W])^2.$$

Taking the expectation of both sides, we obtain

$$\mathrm{E}[U(W)] \approx \mathrm{E}[W] - \frac{\lambda(\mathrm{E}[W])}{2}\mathrm{Var}[W].$$

If W is the return R_i from lending to borrower i at rate r_i, then we recover the expression for lender preferences given in equation (3).

If u shows constant absolute risk aversion, i.e. if $u(c) = 1 - e^{-ac}$ for some positive constant a, then λ will be a constant. Otherwise, λ will not be constant, and we treat the problem of estimating the λ in equation (3) as estimating a "typical" value of λ for the lender.

Note that in the above derivation, we could equally well consider a third- or fourth-order Taylor expansion of u about $\mathrm{E}[W]$. Doing so would result in a similar expression for the lender's expected utility, but which would also include the skewness and kurtosis of returns.

A.2 Derivation of Expectation and Variance of Return

Applying the expectation operator to both sides of equation (1), recalling that r_i is nonrandom, and using linearity gives

$$\mathrm{E}[R_i | X_i, r_i] = (1 + r_i)(1 - \mathrm{E}[D_i | X_i, r_i]) + \mathrm{E}[\phi_i D_i | X_i, r_i].$$

We next condition on D_i to break up $\mathrm{E}\left[\phi_i D_i | X_i, r_i\right]$ as

$$\begin{aligned}
\mathrm{E}\left[\phi_i D_i | X_i, r_i\right] &= \mathrm{E}\left[\phi_i D_i | D_i = 1, X_i, r_i\right] \Pr\left\{D_i = 1 | X_i, r_i\right\} \\
&\quad + \mathrm{E}\left[\phi_i D_i | D_i = 0, X_i, r_i\right] \Pr\left\{D_i = 0 | X_i, r_i\right\} \\
&= \mathrm{E}\left[\phi_i | D_i = 1, X_i, r_i\right] \Pr\left\{D_i = 1 | X_i, r_i\right\} \\
&= \mathrm{E}\left[\phi_i | D_i = 1, X_i, r_i\right] \mathrm{E}\left[D_i = 1 | X_i, r_i\right].
\end{aligned}$$

The last equality follows from the fact that D_i is an indicator variable. Plugging this expression for $\mathrm{E}\left[\phi_i D_i | X_i, r_i\right]$ into the above expression for $\mathrm{E}\left[R_i | X_i, r_i\right]$ gives equation (4).

To derive equation (5), we start by squaring both sides of equation (1) to obtain

$$\begin{aligned}
R_i^2 &= (1 + r_i)^2 (1 - D_i)^2 + 2(1 + r_i) D_i (1 - D_i) \phi_i + \phi_i^2 D_i^2 \\
&= (1 + r_i)^2 (1 - D_i) + \phi_i^2 D_i.
\end{aligned}$$

The second equality follows since $D_i \in \{0, 1\}$, so $D_i^2 = D_i$, $(1 - D_i)^2 = (1 - D_i)$, and $D_i(1 - D_i) = 0$. Taking the expectation of both sides of the above equation gives

$$\mathrm{E}\left[R_i^2 | X_i, r_i\right] = (1 + r_i)^2 (1 - \mathrm{E}\left[D_i | X_i, r_i\right]) + \mathrm{E}\left[\phi_i^2 D_i | X_i, r_i\right].$$

An argument identical to the one given above for $\mathrm{E}\left[\phi_i D_i | X_i, r_i\right]$ shows that $\mathrm{E}\left[\phi_i^2 D_i | X_i, r_i\right] = \mathrm{E}\left[\phi_i^2 | D_i = 1, X_i, r_i\right] \mathrm{E}\left[D_i | X_i, r_i\right]$, so

$$\mathrm{E}\left[R_i^2 | X_i, r_i\right] = (1 + r_i)^2 (1 - \mathrm{E}\left[D_i | X_i, r_i\right]) + \mathrm{E}\left[\phi_i^2 | D_i = 1, X_i, r_i\right] \mathrm{E}\left[D_i | X_i, r_i\right].$$

Finally, we compute

$$\begin{aligned}
\mathrm{Var}\left[R_i^2 | X_i, r_i\right] &= \mathrm{E}\left[R_i^2 | X_i, r_i\right] - \left(\mathrm{E}\left[R_i | X_i, r_i\right]\right)^2 \\
&= (1 + r_i)^2 (1 - \mathrm{E}\left[D_i | X_i, r_i\right]) + \mathrm{E}\left[\phi_i^2 | D_i = 1, X_i, r_i\right] \mathrm{E}\left[D_i | X_i, r_i\right] \\
&\quad - \left((1 + r_i)(1 - \mathrm{E}\left[D_i | X_i, r_i\right]) + \mathrm{E}\left[\phi_i | D_i = 1, X_i, r_i\right] \mathrm{E}\left[D_i | X_i, r_i\right]\right)^2 \\
&= (1 + r_i)^2 (1 - \mathrm{E}\left[D_i | X_i, r_i\right]) \left(1 - (1 - \mathrm{E}\left[D_i | X_i, r_i\right])\right) \\
&\quad + \mathrm{E}\left[\phi_i^2 | D_i = 1, X_i, r_i\right] \mathrm{E}\left[D_i | X_i, r_i\right] - \mathrm{E}\left[\phi_i | D_i = 1, X_i, r_i\right]^2 \mathrm{E}\left[D_i | X_i, r_i\right]^2 \\
&\quad - 2(1 + r_i) \mathrm{E}\left[\phi_i | D_i = 1, X_i, r_i\right] \mathrm{E}\left[D_i | X_i, r_i\right] \left(1 - \mathrm{E}\left[D_i | X_i, r_i\right]\right) \\
&= (1 + r_i)^2 (1 - \mathrm{E}\left[D_i | X_i, r_i\right]) \mathrm{E}\left[D_i | X_i, r_i\right] \\
&\quad + \mathrm{E}\left[\phi_i^2 | D_i = 1, X_i, r_i\right] \mathrm{E}\left[D_i | X_i, r_i\right] - \mathrm{E}\left[\phi_i | D_i = 1, X_i, r_i\right]^2 \mathrm{E}\left[D_i | X_i, r_i\right]^2 \\
&\quad - 2(1 + r_i) \mathrm{E}\left[\phi_i | D_i = 1, X_i, r_i\right] \mathrm{E}\left[D_i | X_i, r_i\right] \left(1 - \mathrm{E}\left[D_i | X_i, r_i\right]\right),
\end{aligned}$$

which is exactly equation (5).

A.3 Results from Estimation of Prediction Models

Table 3: Linear prediction of recovery rate ϕ.

	Grade							
	A	B	C	D	E	F	G	All
const	76.128^{***}	81.296^{***}	82.292^{***}	66.777^{***}	82.929^{***}	76.505^{***}	66.517^{***}	82.276^{***}
	(1.538)	(8.282)	(2.465)	(2.282)	(3.406)	(3.876)	(5.626)	(1.641)
loan_amnt	0.000	0.000	0.000	−0.000	-0.000^{*}	-0.000^{**}	−0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
annual_inc	0.000	−0.000	0.000	0.000	0.000	0.000^{*}	−0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
dti	−0.016	-0.025^{***}	-0.014^{***}	0.002	−0.009	0.006	0.005	-0.009^{***}
	(0.016)	(0.007)	(0.006)	(0.007)	(0.005)	(0.011)	(0.018)	(0.003)
delinq_2yrs	−0.044	0.032	0.094^{*}	0.173^{***}	−0.022	0.170	0.134	0.102^{***}
	(0.143)	(0.060)	(0.044)	(0.052)	(0.054)	(0.078)	(0.122)	(0.024)
inq_last_6mths	-0.530^{***}	-0.305^{***}	-0.193^{***}	-0.269^{***}	-0.210^{***}	-0.216^{***}	0.113	-0.200^{***}
	(0.152)	(0.066)	(0.044)	(0.046)	(0.057)	(0.073)	(0.132)	(0.023)
open_acc	0.071^{*}	0.068^{***}	0.056^{***}	0.053^{***}	0.038^{***}	0.073^{***}	0.056	0.056^{***}
	(0.029)	(0.014)	(0.010)	(0.012)	(0.013)	(0.019)	(0.035)	(0.006)
pub_rec	1.079^{***}	0.266^{*}	0.333^{***}	0.335^{***}	0.195	−0.006	−0.322	0.330^{***}
	(0.317)	(0.131)	(0.091)	(0.097)	(0.114)	(0.165)	(0.275)	(0.049)
revol_bal	−0.000	−0.000	-0.000^{***}	-0.000^{*}	−0.000	-0.000^{***}	−0.000	-0.000^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
total_acc	-0.044^{***}	-0.014^{*}	−0.008	−0.009	0.000	-0.018^{*}	−0.001	-0.011^{***}
	(0.014)	(0.007)	(0.005)	(0.006)	(0.006)	(0.008)	(0.017)	(0.003)
collections_12_mths_ex_med	0.755	0.054	0.561^{*}	0.218	0.065	−0.178	0.635	0.235
	(0.820)	(0.272)	(0.240)	(0.264)	(0.300)	(0.451)	(1.188)	(0.128)
acc_now_delinq	0.626	−0.842	0.121	−1.061	−0.108	−1.508	-3.168^{*}	-0.561^{*}
	(2.064)	(0.673)	(0.437)	(0.557)	(0.699)	(0.893)	(1.260)	(0.261)
chargeoff_within_12_mths	−1.367	−0.445	-0.745^{*}	−0.646	−0.319	0.527	−0.434	-0.508^{*}
	(1.340)	(0.443)	(0.358)	(0.424)	(0.433)	(0.505)	(1.380)	(0.192)
delinq_amnt	−0.033	−0.000	−0.000	−0.000	−0.000	0.001	0.002	−0.000
	(0.019)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.002)	(0.000)
tax_liens	-1.000^{*}	−0.180	−0.171	−0.313	−0.085	0.369	0.443	-0.176^{*}
	(0.441)	(0.180)	(0.135)	(0.153)	(0.170)	(0.239)	(0.473)	(0.082)
emp_length_num	0.048	0.055^{***}	0.048^{***}	0.038^{***}	0.053^{***}	0.072^{***}	0.020	0.049^{***}
	(0.029)	(0.015)	(0.011)	(0.013)	(0.015)	(0.022)	(0.039)	(0.006)
term_years	2.029^{***}	1.943^{***}	2.188^{***}	2.302^{***}	2.523^{***}	2.742^{***}	3.548^{***}	2.339^{***}
	(0.177)	(0.063)	(0.043)	(0.051)	(0.076)	(0.147)	(0.300)	(0.024)
R^2	0.233	0.250	0.306	0.291	0.316	0.307	0.401	0.292
Adj. R^2	0.217	0.246	0.303	0.288	0.311	0.295	0.366	0.291
N	4822	17919	29010	21913	13743	5765	1709	94881

$p < 0.05^{*}, p < 0.01^{**}, p < 0.005^{***}$. Heteroskedasticity-robust standard errors in parentheses. Categorical dummy variables in the regressions (home ownership status, loan purpose, state of residence, loan issuance year, borrower verification status, disbursement method, and application type) omitted for space. Dependent variable measured in percentage points. Variable descriptions available at `https://www.lendingclub.com/info/download-data.action`.

Table 4: Linear prediction of squared recovery rate ϕ^2.

	Grade							
	A	B	C	D	E	F	G	All
const	7401.060^{***}	6742.197^{***}	6913.682^{***}	5489.611^{***}	7026.468^{***}	5991.706^{***}	5774.744^{***}	6894.227^{***}
	(280.565)	(1377.209)	(430.804)	(405.623)	(600.380)	(708.469)	(1018.517)	(280.767)
loan_amnt	−0.000	0.001	0.001	−0.002	-0.003^{*}	-0.005^{*}	−0.001	0.000
	(0.003)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.004)	(0.001)
annual_inc	0.000	−0.000	0.000	0.000	0.000	0.001^{*}	−0.001	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)
dti	−2.496	-4.361^{***}	-2.375^{***}	0.266	−1.614	1.080	0.305	-1.471^{***}
	(2.677)	(1.290)	(1.001)	(1.179)	(0.825)	(1.919)	(3.279)	(0.449)
delinq_2yrs	−10.760	4.245	16.421^{*}	29.136^{***}	−3.870	27.908	21.353	17.447^{***}
	(23.803)	(10.218)	(7.657)	(9.072)	(9.520)	(14.103)	(22.040)	(4.213)
inq_last_6mths	-88.429^{***}	-52.504^{***}	-34.684^{***}	-47.772^{***}	-36.736^{***}	-38.790^{***}	19.240	-33.987^{***}
	(25.939)	(11.396)	(7.659)	(8.084)	(10.020)	(13.224)	(23.928)	(4.029)
open_acc	11.504^{*}	11.231^{***}	9.310^{***}	8.858^{***}	6.132^{**}	12.764^{***}	9.749	9.351^{***}
	(4.897)	(2.479)	(1.823)	(2.108)	(2.308)	(3.381)	(6.297)	(0.990)
pub_rec	182.190^{***}	45.017	56.340^{***}	56.205^{***}	31.112	−3.383	−50.653	55.610^{***}
	(54.504)	(22.550)	(15.847)	(16.853)	(19.908)	(29.414)	(49.425)	(8.564)
revol_bal	−0.000	−0.001	-0.001^{***}	-0.001^{*}	−0.001	-0.003^{***}	−0.003	-0.001^{***}
	(0.001)	(0.001)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.000)
total_acc	-7.456^{***}	-2.401^{*}	−1.350	−1.483	0.161	-3.310^{*}	−0.254	-1.960^{***}
	(2.378)	(1.160)	(0.873)	(0.970)	(1.135)	(1.530)	(3.086)	(0.467)
collections_12_mths_ex_med	127.147	3.188	93.993^{*}	35.328	9.066	−29.054	113.174	38.081
	(136.816)	(45.738)	(41.540)	(45.827)	(52.242)	(77.446)	(215.607)	(22.040)
acc_now_delinq	118.681	−132.736	21.778	−186.979	−13.620	−269.595	-580.339^{*}	-93.525^{*}
	(354.442)	(112.907)	(76.315)	(97.461)	(124.913)	(156.739)	(221.787)	(45.510)
chargeoff_within_12_mths	−238.449	−77.093	-127.773^{*}	−113.483	−52.542	90.777	−52.414	-87.584^{*}
	(225.426)	(75.649)	(60.846)	(71.771)	(75.937)	(91.955)	(250.187)	(32.820)
delinq_amnt	−5.754	−0.004	−0.009	−0.007	−0.007	0.207	0.378	−0.007
	(3.187)	(0.010)	(0.005)	(0.006)	(0.008)	(0.270)	(0.278)	(0.004)
tax_liens	-169.042^{*}	−28.250	−28.784	−52.830	−11.742	72.597	70.567	-27.709^{*}
	(75.998)	(31.099)	(23.400)	(26.657)	(29.413)	(43.980)	(85.045)	(14.782)
emp_length_num	8.201	9.290^{***}	8.411^{***}	6.467^{***}	9.322^{***}	12.748^{***}	3.485	8.408^{***}
	(4.943)	(2.535)	(1.938)	(2.260)	(2.622)	(3.958)	(6.987)	(1.069)
term_years	325.766^{***}	313.655^{***}	353.031^{***}	370.759^{***}	405.854^{***}	438.853^{***}	560.647^{***}	380.687^{***}
	(31.030)	(11.065)	(7.506)	(9.049)	(13.299)	(25.617)	(51.941)	(4.167)
R^2	0.228	0.243	0.296	0.278	0.300	0.292	0.378	0.281
Adj. R^2	0.213	0.239	0.293	0.274	0.295	0.280	0.342	0.280
N	4822	17919	29010	21913	13743	5765	1709	94881

$p < 0.05^{*}, p < 0.01^{**}, p < 0.005^{***}$. Heteroskedasticity-robust standard errors in parentheses. Categorical dummy variables in the regressions (home ownership status, loan purpose, state of residence, loan issuance year, borrower verification status, disbursement method, and application type) omitted for space. Dependent variable measured in squared percentage points. Variable descriptions available at `https://www.lendingclub.com/info/download-data.action`.

Table 5: Logistic prediction of default D.

	Grade							
	A	B	C	D	E	F	G	All
const	-295.796^{*}	-213.917^{***}	-164.120^{***}	-125.906	-77.961	-71.397	-39.533	-193.092^{***}
	(133.744)	(74.310)	(55.214)	(73.972)	(85.051)	(135.856)	(162.283)	(28.984)
loan_amnt	0.000	0.001^{***}	0.001^{***}	0.001^{***}	0.001^{***}	0.002^{***}	0.001^{*}	0.001^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
annual_inc	-0.000^{***}	-0.000^{***}	-0.000^{***}	-0.000^{***}	-0.000^{***}	-0.000^{***}	-0.000^{***}	-0.000^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
dti	2.659^{***}	1.619^{***}	1.850^{***}	1.555^{***}	2.494^{***}	1.531^{***}	1.490^{***}	2.593^{***}
	(0.221)	(0.112)	(0.090)	(0.106)	(0.144)	(0.235)	(0.448)	(0.050)
delinq_2yrs	14.912^{***}	5.888^{***}	4.505^{***}	2.710^{***}	3.098^{*}	4.431^{*}	1.815	7.730^{***}
	(1.951)	(0.898)	(0.733)	(0.886)	(1.238)	(2.079)	(3.480)	(0.417)
inq_last_6mths	11.799^{***}	8.084^{***}	6.827^{***}	7.145^{***}	6.490^{***}	2.295	7.590^{**}	15.903^{***}
	(2.003)	(0.948)	(0.701)	(0.766)	(0.985)	(1.512)	(2.723)	(0.370)
open_acc	2.569^{***}	2.356^{***}	1.928^{***}	1.809^{***}	0.962^{***}	1.785^{***}	1.575	2.130^{***}
	(0.405)	(0.218)	(0.179)	(0.216)	(0.283)	(0.452)	(0.886)	(0.100)
pub_rec	33.296^{***}	7.723^{***}	3.865^{*}	3.127	1.490	-1.935	2.348	10.823^{***}
	(4.452)	(1.937)	(1.506)	(1.774)	(2.471)	(4.034)	(7.428)	(0.860)
revol_bal	0.000	-0.000	-0.000	-0.000^{***}	-0.001^{***}	-0.000^{**}	-0.001^{*}	-0.000^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
total_acc	-1.251^{***}	-1.031^{***}	-1.064^{***}	-0.976^{***}	-0.786^{***}	-0.922^{***}	-0.771	-1.577^{***}
	(0.196)	(0.103)	(0.085)	(0.101)	(0.132)	(0.211)	(0.418)	(0.047)
collections_12_mths_ex_med	25.494	21.818^{***}	11.170^{*}	14.469^{**}	8.031	10.310	12.560	17.366^{***}
	(14.173)	(5.096)	(4.433)	(5.505)	(6.585)	(12.479)	(24.001)	(2.537)
acc_now_delinq	16.382	3.465	4.637	1.404	-10.945	13.848	52.590	8.914
	(37.866)	(11.410)	(8.092)	(9.222)	(9.934)	(21.279)	(32.301)	(4.617)
chargeoff_within_12_mths	5.902	2.315	-0.439	-3.230	-0.629	-19.093	-27.293	0.578
	(18.397)	(7.112)	(5.988)	(7.612)	(9.091)	(16.726)	(35.537)	(3.419)
delinq_amnt	-0.444	0.002	0.001	0.000	0.002	-0.022	-0.008	0.001^{*}
	(0.472)	(0.001)	(0.001)	(0.001)	(0.001)	(0.017)	(0.010)	(0.000)
tax_liens	-24.874^{***}	-4.373	0.715	-5.760^{*}	0.538	4.260	-0.658	-6.950^{***}
	(6.297)	(2.681)	(2.229)	(2.769)	(3.921)	(5.572)	(11.870)	(1.270)
emp_length_num	-1.003^{*}	-1.382^{***}	-1.263^{***}	-1.297^{***}	0.006	-0.442	0.311	-1.064^{***}
	(0.416)	(0.225)	(0.187)	(0.228)	(0.307)	(0.506)	(0.973)	(0.104)
term_years	14.358^{***}	20.246^{***}	22.476^{***}	20.506^{***}	13.962^{***}	11.800^{***}	7.013	39.471^{***}
	(4.402)	(1.282)	(0.825)	(0.950)	(1.333)	(2.567)	(5.478)	(0.447)
R^2_{McFadden}	0.028	0.026	0.031	0.036	0.040	0.045	0.078	0.070
$-2\log(\text{LR})$	1027.760	2709.515	4357.236	3273.755	1974.265	830.946	407.385	33333.387
N	77142	136732	135007	75326	37428	13630	3801	479066

$p < 0.05^{*}, p < 0.01^{**}, p < 0.005^{***}$. Standard errors in parentheses. Categorical dummy variables in the regressions (home ownership status, loan purpose, state of residence, loan issuance year, borrower verification status, disbursement method, and application type) omitted for space. Parameter estimates and standard errors multiplied by 100, for ease of reading and interpretation. Variable descriptions available at `https://www.lendingclub.com/info/download-data.action`.

Table 6: Linear prediction of default D.

	Grade							
	A	B	C	D	E	F	G	All
const	-4.856	-2.726	1.343	-17.713	29.412	100.817^{*}	-11.401	-11.360^{*}
	(5.554)	(10.062)	(8.677)	(8.357)	(18.715)	(15.717)	(25.238)	(5.089)
loan_amnt	-0.000	0.000^{**}	0.000^{***}	0.000^{***}	0.000^{***}	0.000^{***}	0.000	0.000^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
annual_inc	-0.000^{***}	-0.000^{***}	-0.000^{***}	-0.000^{***}	-0.000^{***}	-0.000^{***}	-0.000^{***}	-0.000^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
dti	0.180^{***}	0.177^{***}	0.284^{***}	0.273^{***}	0.440^{***}	0.199^{***}	0.282^{**}	0.371^{***}
	(0.014)	(0.029)	(0.036)	(0.053)	(0.054)	(0.130)	(0.101)	(0.026)
delinq_2yrs	1.078^{***}	0.649^{***}	0.679^{***}	0.497^{***}	0.699^{*}	0.857	0.344	1.155^{***}
	(0.170)	(0.113)	(0.127)	(0.181)	(0.277)	(0.474)	(0.758)	(0.070)
inq_last_6mths	0.756^{***}	0.975^{***}	0.992^{***}	1.280^{***}	1.288^{***}	0.371	1.940^{***}	2.473^{***}
	(0.129)	(0.113)	(0.119)	(0.156)	(0.219)	(0.351)	(0.591)	(0.062)
open_acc	0.116^{***}	0.230^{***}	0.338^{***}	0.381^{***}	0.343^{***}	0.471^{***}	0.309	0.313^{***}
	(0.024)	(0.027)	(0.032)	(0.046)	(0.065)	(0.107)	(0.196)	(0.018)
pub_rec	2.157^{***}	0.649^{***}	0.158	0.281	0.292	-0.636	0.378	1.452^{***}
	(0.354)	(0.227)	(0.251)	(0.364)	(0.549)	(0.910)	(1.627)	(0.141)
revol_bal	-0.000	-0.000^{**}	-0.000^{***}	-0.000^{***}	-0.000^{***}	-0.000^{*}	-0.000	-0.000^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
total_acc	-0.059^{***}	-0.087^{***}	-0.172^{***}	-0.176^{***}	-0.185^{***}	-0.166^{***}	-0.154	-0.221^{***}
	(0.011)	(0.011)	(0.014)	(0.020)	(0.030)	(0.049)	(0.094)	(0.007)
collections_12_mths_ex_med	1.966^{*}	3.073^{***}	1.779^{*}	3.049^{**}	1.829	2.714	2.870	2.925^{***}
	(1.197)	(0.761)	(0.787)	(1.171)	(1.612)	(2.917)	(5.790)	(0.478)
acc_now_delinq	-0.921	0.291	0.463	0.225	-2.532	2.907	12.082	1.270
	(2.160)	(1.409)	(1.444)	(1.764)	(1.949)	(4.433)	(7.250)	(0.876)
chargeoff_within_12_mths	0.338	0.202	-0.107	-0.755	0.023	-4.552	-6.086	-0.031
	(1.357)	(0.868)	(0.969)	(1.563)	(1.993)	(3.631)	(8.170)	(0.546)
delinq_amnt	-0.000	0.000^{*}	0.000	0.000	0.000	-0.003	-0.001	0.000^{*}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
tax_liens	-1.598^{***}	-0.200	0.575	-0.787	0.144	1.219	0.192	-0.842^{***}
	(0.498)	(0.345)	(0.411)	(0.519)	(0.872)	(1.199)	(2.642)	(0.218)
emp_length_num	-0.052^{*}	-0.157^{***}	-0.226^{***}	-0.281^{***}	-0.088	0.002	0.182	-0.174^{***}
	(0.025)	(0.026)	(0.031)	(0.046)	(0.068)	(0.116)	(0.217)	(0.016)
term_years	2.229^{***}	2.631^{***}	3.857^{***}	4.378^{***}	4.309^{***}	3.834^{***}	3.732^{***}	7.258^{***}
	(0.334)	(0.168)	(0.144)	(0.194)	(0.291)	(0.580)	(1.157)	(0.081)
R^2	0.014	0.020	0.031	0.042	0.050	0.057	0.101	0.068
Adj. R^2	0.012	0.019	0.030	0.040	0.048	0.051	0.078	0.068
N	77142	136732	135007	75326	37428	13630	3801	479066

$p < 0.05^{*}, p < 0.01^{**}, p < 0.005^{***}$. Standard errors in parentheses. Categorical dummy variables in the regressions (home ownership status, loan purpose, state of residence, loan issuance year, borrower verification status, disbursement method, and application type) omitted for space. Parameter estimates and standard errors multiplied by 100, for ease of reading and interpretation. Variable descriptions available at `https://www.lendingclub.com/info/download-data.action`.