How Does A Firm Adapt in A Changing World? The Case of Prosper Marketplace

Xinlong Li and Andrew T. Ching*

First draft: May 31st, 2019
This draft: August 2nd, 2021

Abstract

In a rapidly changing world, older data may not be as informative as the most recent data. This is known as the *concept drift* problem in statistics and machine learning. How does a firm adapt in such an environment? To address this research question, we propose a *generalized revealed preference approach*. We argue that by observing a firm's choices, we can uncover how it uses past data to adapt when making business decisions. We apply this approach to study Prosper Marketplace, which is an online peer-to-peer (P2P) lending platform. More specifically, we develop a structural model, where Prosper uses a data selection algorithm to continuously update its training sample and re-estimate the borrower side and lender side models, and use these updated model to do risk assessment for loan applications over time. To infer which data selection algorithm Prosper may use, we consider a set of algorithms motivated by the machine learning literature. For each data selection algorithm, we assume Prosper takes the algorithm specific borrower side and lender side model parameter estimates as given, and use Prosper's loan risk rating classification decisions to estimate its objective function's structural parameters. By comparing the goodness-of-fit of these algorithm specific models, we find that the ensemble recession probability method is the most plausible adaptive learning algorithm used by Prosper. We also use our model to simualte what would happen if Prosper does not adaptively learn, and switch to other adaptive learning methods.

Keywords: Peer-to-peer Lending, Two-sided Market, Concept Drift, Machine Learning, Structural Model, Fintech, Generalized Revealed Preference

*Xinlong Li (xinlong.li@ntu.edu.sg) is Assistant Professor of Marketing at Nanyang Business School, Nanyang Technological University. Andrew T. Ching (andrew.ching@jhu.edu) is Professor of Business, Economics and Public Health at Carey Business School, Johns Hopkins University.

1 Introduction

What is the hottest item one day later may become just a fad. Individual's preferences and behavior may change rapidly these days, due to constant technological innovations and an ever-changing economic environment. Therefore, treating all historical data to be equally informative when building an analytical model to guide business decisions could be quite misleading. More formally, the problem can be framed as follows. Suppose the independent variables are denoted by X_t (e.g., consumer characteristics) and the dependent variable is denoted by Y_t (e.g., consumer choice), where t indexes time. The relationship between X_t and Y_t can be formulated as $Y_t = F(X_t, \epsilon_t, \theta_t)$, where ϵ_t is the stochastic element and θ_t is the parameter vector at time t. Note that θ_t may change over time. If the change in θ_t is unaccounted for, the estimated θ can be very far from the true θ_t, and that could lead to very poor predictive power of the estimated model. Researchers in statistics and machine learning call this the *concept drift* problem.

The concept drift problem has received more attention lately because the recent technology allows many companies to collect customer data that arrive in a stream; hereafter, we refer to this type of data as *streaming data*. Some common examples of streaming data include online reviews, website clicks, mobile phone apps, credit card transactions, E-commerce purchases, and social networks. Businesses are harnessing streaming data to develop deeper evidence-based insights about their customers to radically change the way they run their businesses. However, streaming data is likely subject to the pitfall of concept drifts. Traditional estimation approaches usually pool most data together and treat them equally when conducting estimation. But the predictive performance of models estimated using older data could quickly deteriorate over time. Therefore, conclusions drawn from streaming data analysis will be questionable if the concept drift problem is not accounted for. Acknowledging the concept drift problem, the machine learning literature has taken advantage of the very large number of observations from streaming data and introduced ways to handle the concept drift problem (Tsymbal, 2004).

How does a firm account for the concept drift problem when using streaming data? Recognizing the potential of streaming data, tech companies have recruited a large number of researchers in Data Science, Statistics and Computer Science to help them analyze their data and develop models to predict the behavior of their customers. It is conceivable that tech companies are using some methods from the machine learning literature to address the concept drift problem. There are a number of ways that a firm can use. Some examples of these methods range from selecting the most recent data based on a moving window of observations, only using "relevant" past observations based on some economic environment metrics, etc. From the market intelligence viewpoint, if a firm can learn how its competitors use their data to make business decisions (e.g., setting prices), it can gain valuable competitive advantages. In this paper, we propose extending the revealed preference approach to infer how a firm uses its available data (i.e., its adaptive learning/data selection algorithm) from its choices. In particular, we use a peer-to-peer (P2P) lending platform to illustrate our approach.

The platform we study is Prosper Marketplace Inc. (hereafter Prosper), which is one of the largest online P2P lending platforms in the U.S.. Prosper connects people who need to borrow money with people who have savings to invest. As a platform, one of the key services Prosper provides is to evaluate each loan application's risk level and assign it to one of the seven ratings: AA, A, B, C, D, E and HR, where AA means the lowest risk and HR means the highest risk. Each rating corresponds to a certain interest rate. In other words, the interest rate is rating specific. Prosper also posts the expected loss rate for each loan application to capture its downside risk in case of default. This *posted loss rate* is also rating specific.

On the one hand, Prosper's rating should reflect a loan's true risk accurately in order to maintain Prosper's long-term reputation. On the other hand, Prosper might take into account that the interest rate is also a tool to increase the likelihood of funding a loan because Prosper only earns revenue when a loan is successfully funded. If the interest rate is set too high, borrowers will be discouraged to borrow from this platform; if it is set too low, it may fail to attract enough investors to finance the loan.

Our key insight is that Prosper's decision on classifying risk categories will not only reveal the parameter values of its objective function, but also the way Prosper selects the past data to adaptively learn and update the parameter values of its borrower and lender predictive models. Specifically, we develop a structural model to capture Prosper's decision process. We assume that in each period, as more data come in, Prosper selectively uses the past data available to update and re-estimate its borrower and lender models, and then uses them to analyze the risks of the incoming loan applications and assign each loan to a risk rating to maximize its objective function. Because the estimates of borrower and lender model parameters are a function of the way that Prosper selects and weighs the past data, Prosper's loan rating choice policy is also a function of its adaptive learning method. Hence, by observing Prosper's choices, it should be possible to infer how Prosper uses the past data (i.e., its adaptive learning/data selection algorithm). We call this the *generalized revealed preference approach*. It is important to highlight that we do not assume firms know how the true data generating process evolves over time because the environment may be changing in a highly unpredictable way. Instead, we assume that they know concept drifts exist, and hence their data scientists would use data selection algorithms from machine learning to try to address it. We rely on revealed preference to infer which data selection algorithm is most likely used by Prosper.

To implement our approach, we face one major challenge. There are infinitely many non-nested methods that a firm could use to weight the past data. It is impossible for us to consider all of them. Hence, we restrict our attention to a set of machine learning methods in handling the concept drift problem. For each adaptive learning algorithm, we re-estimate our structural model. Then we compare their goodness-of-fit. The adaptive learning algorithm which leads to the best model fit will be regarded as the most plausible algorithm adopted by Prosper among our consideration set.

Our data set consists of 31,807 unsecured personal loan applications from Dec 2010 to Dec 2012 provided by Prosper. Using our generalized revealed preference approach, we find that among the set of data selection

algorithms that we consider, an ensemble recession probability method best describes its data selection process. Note that because we do not assume Prosper knows exactly how the data generating process changes over time, it may not be using the best adaptive learning algorithm to update its borrower side and lender side models.

The rest of the paper is organized as follows. Section 2 gives an overview of related literature. In section 3, we discuss the P2P lending industry and the data in detail, and provide some reduced form evidence that concept drifts exist in our context. The model framework is summarized in section 4. We introduce different algorithms of using historical data in section 5. We discuss the identification issues in section 6. Section 7 presents the estimation results. Section 8 is the conclusion.

2 Literature

Our paper is related to an emerging stream of literature on P2P lending and more generally, crowdfunding. Wei and Lin (2017) and Liu, Wei and Xiao (2020) studied how the change in pricing mechanism (from auction to posted price) at Prosper affected its participants' behavior. Fu, Huang, and Singh (2021) studied whether machine learning algorithms can beat the crowds of investors in predicting loan defaults and lead to greater welfare for both investors and borrowers. Lin and Viswanathan (2016), Lin et al. (2013), and Iyer et al. (2016) investigated if soft information, such as home bias, online friendships of borrowers, pictures and text descriptions acts as signals of credit quality.[1] Zhang and Liu (2012) studied rational herding decisions among lenders. Freedman and Jin (2011) investigated the role of learning-by-doing in alleviating the information asymmetry about borrowers. Kawai, Onishi and Uetake (2020) examined how signalling can mitigate welfare losses that result from adverse selection. All of these papers, except Wei and Lin (2017) and Liu, Wei and Xiao (2020), study the early period when Prosper use the auction mechanism to fund loans (similar to eBay auction).

None of the papers above consider the possibility that firms may face a concept drift problem. However, in the machine learning and statistics literature, many studies have shown the prevalence of concept drift and how it may bias analysis results when ignoring it. Schlimmer and Granger (1986) first point out concept drift may affect model's prediction performance. Kelly, Hand and Adams (1999) show that the concept drift problem exists in credit card default detections. Crespo and Weber (2005) show that adaptive data mining methods that update the model continuously outperform static models in customer segmentation analysis. In the famous Netflix Prize Competition, one of the lessons learned by the winning team is that taking temporal dynamics into account substantially contributes to building accurate models. They allow

[1] Another relevant paper is Ni and Xin (2020). They study the existence of local bias in one of the largest online crowdfunding marketplaces in China and quantify the importance of information asymmetry and preference toward local projects on inducing local biases using a structural model.

users' average rating to change over time to capture their drifting preference. Hoens et al. (2012) provided a review of other related works on concept drift.

Our paper is also related to the economics literature on firms' adaptive learning behavior. The idea of adaptive learning posits that agents proceed like an econometrician and use the available data to estimate a model of the economy and update their beliefs about the model parameter values as new data arrives. The pioneer works include Sargent (1993), Evans and Honkapohja (2001, 2013), and more recent works include Doraszelski, Lewis and Pakes (2018), Kozlowski, Veldkamp, Venkateswaran (2020). Aguirregabiria and Jeon (2020) provide a comprehensive review about how firms learn about demand, costs, or the strategic behavior of other firms in the changing market. Ching (2010) and Hitsch (2006) model firms to learn about their demand conditions in a Bayesian manner.[2]

Similar to these studies in economics, we also model how a firm revises its belief in an adaptive manner. But the economics literature tends to assume that the data generating process is stable at least for an extended period of time, and the timing of any structural breaks are known to agents in the model. Within each stable period, this literature typically assumes an adaptive learner will simply pool all data available up to the current time stamp to re-estimate and update the parameter values of the model that he/she faces.[3] In contrast to this literature, we assume agents do not know when concept drifts (corresponds to structual breaks) happen, and even allow for the possibility that concept drifts happen continuously over time. Instead, we hypothesize that agents in the model know concepts drifts may have happened, and hence they try to use the data available selectively to address such this problem. This is the underlying assumption that data scientists have when they propose ways to address the concept drifts problem. Because firms (managers) need to make real time decisions, and existing tests to detect concept drifts can only detect it after a concept drift had happened for some time (because it relies on its cumulative effect on the data), we believe it is reasonable to assume that firms do not try to uncover how the data generating process changes over time. Because firms do not know the exact nature of concept drifts, they may just rely on the knowledge of their data scientists and choose one adaptive learning method to handle this problem, even though the method may not be the most optimal. In this research, we focus on uncovering the adaptive learning algorithm that a firm uses.[4]

[2]Hitsch (2006) models firms learn about the paramter values of the reduced form demand curves of their products. Ching (2010) models both firms and consumers are uncertain about product quality, and social learning allows both sides to learn its true quality over time.

[3]For instance, Kozlowski et al. (2020) assumes consumers use data available to update their estimate of the distribution of macroeconomic shocks to the economy.

[4]The marketing literature has also found evidence that consumer preferences evolves over time, e.g., Sriram, Chintagunta, and Neelamegham (2006), Liechty, Fong, and DeSarbo (2005), Dew, Ansari and Li (2020), etc. Preference evolution can be one source of concept drift. More generally, concept drift can also be due to change in the macro environment, competitor's policies, etc. Another source of concept drift could be cross-sided network externality, which

3 Institutional Background and Data

3.1 Institutional Background

Prosper is the first P2P lending platform in the US. It facilitated over $18 billion of unsecured loans by 2020. Over time, more and more people have started to accept P2P lending as one of the main alternative finance markets in the U.S. and view peers as having an equal level of credibility as experts. The value of global P2P lending is expected to rise to one trillion U.S. dollars by 2050.[5]

It should be highlighted that in the early years of Prosper, it used an eBay style auction system that allowed lenders and borrowers to determine the interest rates. This business model earned Prosper the name "eBay of Loans." However, Prosper was not very successful during this early phase. The credit score of the borrowers were relatively low, and the average investor returns were negative.[6] On December 20, 2010, the platform switched to a post-price model with pre-set interest rates for each loan application (hereafter listing). In this study, we focus on listings initiated during the two-year window, December 20, 2010 to Dec 31, 2012, after the regime change. The post-price model marked a new era for Prosper. Since the launch of this new model, Prosper has been growing rapidly.

To register as a borrower on Prosper, an applicant needs to provide some basic identity information including name, Social Security number, address, telephone number, etc. If the information provided is consistent with information in the anti-fraud and identity verification databases, the registration will be approved. Borrowers can request anywhere from $2,000 to $25,000 [7] per loan on Prosper and choose to repay over a 36- or 60-month amortization periods. In our analysis, we only consider loans with a fixed loan length of 36 months because this is what most borrowers choose, and the interest rate for 60-month loans is determined quite differently.

Before a borrower's loan application is approved, Prosper pulls the borrower's credit history from its credit reporting partner Experian. Some key credit history variables include the borrower's credit score, number of delinquencies in the last seven years, total number of inquiries, bankcard utilization rate, etc., which help Prosper to assess loan risk by mitigating asymmetric information, and determine eligibility (Chan et al. 2020). Table 3 provides a full list of the variables in the data set. Loan applications will not be approved unless the borrower's credit score needs to be above a certain threshold, which is 600 in our sample period. If a loan application is approved, Prosper then uses its risk assessment algorithm to assess its risk. Prosper states that their risk assessment algorithm is trained based on the historical performance of Prosper loans with

we do not explicitly model. Some of our adaptive learning algorithms should be able to address it because the parameter estimates at each time stamp will be specific to the data being selected for estimation.

[5] https://www.statista.com/statistics/325902/global-p2p-lending/

[6] https:www.lendacademy.comprosper-review, accessed on June 4th, 2021.

[7] The amount limit only applies to our sample period. Nowadays borrowers can request as much as $40,000 per loan.

similar characteristics, without providing any details. Prosper assigns each listing a rating using a system with seven grades: AA, A, B, C, D, E and HR, where AA is the lowest risk and HR stands for the highest risk. Each rating corresponds to an interest rate, and this mapping changes very occasionally over time. Once the loan application is approved, it is open for investment for 14 days. Moreover, for each listing, Prosper posts Loss Rate (hereafter, we refer it to "posted loss rate"), which is a proxy for expected loss rate. It is important to highlight that posted loss rate only depends on its risk rating. So, similar to loan interest rates, posted loss rates are not listing specific, instead, they are rating specific.

Prosper charges lenders a 1% annual loan service fee based on the current outstanding loan principal lenders hold. Prosper charges borrowers an origination fee on each completed loan. The origination fee is a percentage of the funded loan amount and varies across Prosper's rating. We should highlight that this is the only revenue source for Prosper. In other words, if a loan is not funded, neither borrowers nor lenders need to pay any fees to Prosper. Table 1 shows the average interest rate, average annual return, origination fee rate and posted loss rate for each rating level. It is worth noting that, on average, lower rated loans have higher annual returns, and E loans have the highest average annual return. This is consistent with the general finance principle that high risk investment (i.e., high variance) carries higher expected return.

We should highlight that both interest rate and posted loss rate of a listing depend only on what Prosper rating it receives. This could simply reflect that risk assessment is intrinsically a difficult problem because of the unobserved heterogeneity of borrowers, and hence Prosper may not feel comfortable with assigning different interest rates and posted loss rates to loans within one rating class, knowing that there are noises in what their model can predict.

Borrowers make repayments of equal amount on a monthly basis. If a borrower misses four repayments in a row, the loan is marked as "Defaulted."[8] Lenders lose their outstanding principal in defaulted loans. Note that defaulted loans hurt Prosper in three ways: (i) Prosper loses the annual service fee on the unpaid principal of defaulted loans; (ii) Prosper needs to incur the operating cost of loan collections; (iii) the default rate is one of the main factors that affects lenders' investment decisions, and it is widely discussed online.[9] Hence, the credit screening and risk assessment processes are very important to Prosper, because risk rating should accurately inform how much default risks investors face.

Lenders are allowed to contribute as little as $25 to a loan and as much as the full amount requested by the borrower. Lenders have access to all the financial history information of borrowers when they make their investment decisions. Prosper provides full historical loan performance data on its website.

[8] For instance, if borrower A did not make repayment for four months in a row starting month 2, then he/she will be coded as default in month 2.

[9] http://www.lendingmemo.com/lending-club-prosper-default-rates/

3.2 Data Description

Our data set spans from Feb 2007 to Dec 2012. In this paper, we are interested in studying Prosper's problem after the regime change. Thus, we only describe the data on loans initiated after Prosper has switched to the posted price business model since Dec 19, 2010. We use data prior to Dec 2010 to construct initial conditions for our models.

Our data set consists of 31,807 listings. Among them, 22,277 (70.04%) loan applications were funded, 5,825 (18.31%) were withdrawn by the borrowers, and 3,705 (11.65%) were expired (i.e., they were not funded). Table 2 provides summary statistics of listings' outcome based on their Prosper ratings. Listings that are rated D or below D account for 59.44% of the total listings. The completed percentage (i.e., percentage of listings that are successfully funded) of HR listings is significantly lower than listings in all other rating levels. It indicates that the interest rate of HR listings were not high enough to offset their risks, and so lenders were less interested in investing in those listings. The distribution of ratings varies over time. Figure 1 presents the percentage distribution among seven ratings over time. The number of C and HR loans grows gradually over time. The variations in rating percentage could be partly driven by changes in borrower's characteristics over time. Another possibility is that Prosper may adaptively learn from more recent or selected data and change the rating assignment accordingly to improve its profits.

Among those 22,277 originated loans, 3,529 (15.84%) were defaulted. Figure 2 shows how the conversion rate (i.e., the rate at which listing become loans) and default rate change over time. Large variation can be observed in the figure. For instance, the platform experiences high default rate and low conversion rate in August, 2011. While at the start of 2012, the conversion rate is relatively high and default rate is low. These variations can help us to identify lenders' and borrowers' changing investing and borrowing behaviors.

For each loan application, we observe borrower's credit history variables (e.g., the range of FICO score, credit lines, current delinquencies, etc.), Prosper rating, interest rate, monthly loan payment amount, and whether this loan application is completed, withdrawn, or expired. We also observe the borrower's full repayment history, including the percentage of principal loss in case of default. Table 3 provides a full list of the variables we can observe in the data. Table 4 presents their summary statistics. The average amount requested by each borrower is $6,829 and the average interest rate a borrower needs to pay is 23.34%. The average interest rate is relatively high, and this may reflect that it is harder for some people to borrow from traditional channels (e.g., banks). The high interest rate should also attract potential lenders to invest in this platform.

On average Prosper charges borrowers origination fee 4.36% of the loan amount. The average lower bound of FICO score across all loan applications is 697.47, which is much higher than the minimum eligibility requirement of 600 on Prosper. The mean annual income of borrowers was $68,004, which was more than

two times higher than the per capita annual income of $29,173 in the U.S in 2011.[10] Borrowers' average monthly debt is $873, which is 15.40% of their average monthly income. Borrowers average bankcard utilization rate is 51.50%, which is much higher than the average 30% credit utilization rate.[11] About half of the borrowers own a home and the average mortgage balance is $107,599. For comparison, the average home ownership rate in 2011 is 66.1% in the U.S. Hence, a typical borrower in Prosper is relatively young and middle class. Prosper makes key borrower credit data available in their website. We can observe many borrower characteristics, including key credit related characteristics such as credit score, bank card utilization rate, number of delinquencies in the past, number of credit lines, etc. These observed characteristics allow us to capture heterogeneity in borrowers and loans, how the relationship between them and the outcome variables changes over time.

3.3 Reduced Form Evidence of Concept Drifts

This section provides evidence of the concept drift problem in the environment that we study. We follow Gama et al. (2004) to test for the existence of concept drifts. The intuition of this test relies on detecting a significant deterioration in model prediction measured by its prediction error rate using data points just beyond the training set. The test makes use of a growing sample of observations to keep revising the prediction error rates of the model. If there is no concept drift, we do not expect to see any significant change of the prediction error rate. But if a major concept drift happened, it would lead to a significant increase in the prediction error rate. The basic mechanism of the test is as follows: It starts with a base window of sample period, and use it to estimate a model and compute its prediction error in the period just beyond the base window. If the prediction error rate is below a threshold, we will expand the base window, and repeat this procedure until the prediction error rate is above the threshold. This is when we detect at least one concept drift has happened (in a sense this is conservative test because it relies on the "cumulative" effects of possibly multiple concept drifts). As soon as we detect a positive test outcome, we will start a new data window by discarding the old data.

We will use the following example to illustrate how this test works. Let us consider how to test the existence of concept drift in lender's investing behavior. Consider a set of listings, in the form of pairs (x_i, y_i), where x_i represents borrower i's characteristics (e.g., listing's interest rate, the borrower's credit score, monthly income, etc. The full set of x_i's are listed in Table 3.), and y_i takes value 1 if the listing is funded and 0 otherwise. Suppose that we use a logistic regression model to predict each listing's probability of getting funded. Let $\hat{y}_l$ denotes the predicted outcome for listing l, where subscript l denotes observations in the period just beyond the base window. The event of false prediction, $\hat{y}_l \neq y_l$, is a random variable from Bernoulli trials.[12] For a sequence of n_t observations, the number of false prediction events follows a Binomial

[10] https://www.deptofnumbers.com/income/us/

[11] https://www.creditcards.com/credit-card-news/credit-card-use-availability-statistics-1276.php

[12] We set $\hat{y}_l = 1$ if our logistic model predicts listing l's probability of getting funded is larger than 0.5, and $\hat{y}_l = 0$

distribution. Let p_t denote the prediction error rate (i.e., the percentage of false predictions) in period t, the corresponding standard deviation is $s_t = \sqrt{p_t(1-p_t)/n_t}$. If lender's investment behavior is stationary, the prediction error rate (p_t) of our model should remain stable over time. A significant change in the error rate suggests a change in lender's investment behavior. We describe the details of the test in Appendix A, including how we define the detection threshold and update the data window. In Figure 3(a), we show the prediction error rate on whether listings are funded in each time stamp. The red dots represent when concept draft are detected by the test. For lenders' investment behavior, the test is able to detect concept drift 17 times during our sample period. It is important to note that because the test relies on the "cumulative" effects of concept drift on the prediction error, when the test result is positive, one or more concept drifts may have already happened for some time. Hence, the number of times detected should be interpreted as the lower bound for the number of concept drifts happened in the sample period.

In addition to observing whether listings are funded or not, Prosper also observe whether borrowers decided to withdraw their applications, and defaulted their loans. We test concept drift for these two outcomes, using a logit model with the same set of explanatory variables as in the loan funding model, to construct prediction error rates for borrower's withdrawal and default decisions over time. Figures 3(b) and 3(c) show the test results. The test is able to detect concept drifts 7 and 5 times for the withdrawal and default decisions, respectively.

In light of this evidence for concept drift, we consider Prosper uses five different ways to handle the historical data when it estimates the demand (lenders) and supply (borrowers) side parameters. In the benchmark method (EWM), we assume Prosper ignores the concept drift problem and simply use all historical data equally. In all other methods, we assume Prosper recognizes the concept drift problem, and weights the past data depending on different conditions. For each version, we embed the estimated demand and supply sides in Prosper's objective function, which drives its risk assessment decisions for each loan application. By comparing the goodness-of-fit of Prosper's risk assessment choices predicted by each model, we can infer which adaptive learning algorithm is closest to what Prosper actually uses.

Remark: It is important to note that concept drift in our context could be due to several sources:

- Changes in the distribution of unobserved characteristics over time – they can affect the stochastic component of data generating process, or change how observed independent variables affect the dependent variable via their interactions with observed independent variables;
- Changes in the fundamental impact of x on y, because of (a) the changes in competitive or macro environments, (b) the fundamental differences between new vs. old borrowers/lenders over time, (c) borrowers/lenders learn and adapt over time.

otherwise. Note that we model x_i's to enter the model linearly.

We should stress that our goal (and the machine learning literature of concept drift) is not to separately identify these sources, and measure their relative importance (which could also change over time). Instead, knowing that concept drift is in the data generating process, the machine learning literature proposes different ways to handle it. We hypothesize that Prosper uses one of the adaptive learning algorithms to handle this problem. Our goal is to infer which adaptive learning algorithm Prosper uses. Our intuition is that disentangling the exact sources of concept drift is a very challenging problem that firms may not be able to solve either. But even if Prosper does not know what the exact sources are or their relative importance, it can still try to come up with an adaptive learning algorithm to address the concept drift. Following this logic, Prosper's loan classification choices is a function of its adaptive learning algorithm, and hence by observing their choices, we should able to uncover it. This is the underlying intuition behind our generalized revealed preference approach.

4 Model

We model Prosper as an adaptive decision maker. At the end of period t, we assume Prosper updates and re-estimates its borrower side and lender side models by incorporating the new information it receives in that period, and then uses the newly updated borrower side and lender side models to assign rating to loan applications in period $t+1$.

First and foremost, assigning loan risk rating is a risk assessment exercise. The key measures of a loan's risk is its default probability, and expected (principal) loss given default. If the sole purpose of Prosper is to provide an accurate underlying risk of each loan application, Prosper needs to estimate the default rate, and expected loss given default so that when setting a rating, its corresponding posted loss rate (which is a function of Prosper's rating) matches with actual expected loss rate as closely as possible. This is why it is important for us to model how Prosper estimates the default probability and expected loss given default (i.e., its risk assessment model).

But we should also highlight that Prosper only earns revenue when a loan application is funded. Hence, it is possible that Prosper may take into account how the loan rating might influence borrowers' and lenders' behavior. Recall that interest rate is a function Prosper's rating (again, 1-1 mapping). On the one hand, the lower the rating (i.e., the higher the interest rate), the more likely a borrower may withdraw the listing. On the other hand, the higher the interest rate, the more investors the loan can attract. Hence, loan ratings, via these two conflicting forces, can affect the expected revenue of a loan listing. In our model, we allow Prosper to take both accuracy of risk and expected revenue of a listing into account. However, we do not impose the relative utility weights on these two components, and will let the data reveal them as the output of our structural estimation.

Our model has four parts: (i) the borrower side model; (ii) the lender side model; (iii) the risk assessment

model; (iv) Prosper's utility function, which consists of the expected revenue from each loan application, and the penalty on assigning a risk level deviated from its true risk. We consider five different data selection/adaptive learning algorithms that Prosper may utilize to estimate parts (i)-(iii) continuously. We investigate which adaptive learning algorithm and investigate which one best describes Prosper's decision process For each adaptive learning algorithm, we assume Prosper takes the series of estimates of parts (i)-(iii) as given, and assigns loan rating to each listing to maximize its utility; we use Prosper's utility function and rating choices to estimate its utility structural parameters for each adaptive learning algorithm.

The adaptive learning algorithms are motivated by some common approaches proposed in the machine learning literature (e.g., gradual forgetting, moving window, etc.). We wait until section 5 to discuss the details of these adaptive learning algorithms.

4.1 Borrower Side

We assume that borrowers arrive exogeneously, and submit their loan applications and find out their Prosper rating. They then decide whether to withdraw their loan applications; it is this withdrawal decision that we model.[13]

Let i denote the ith listing that arrives in period t. X_i is the observed characteristics of the borrower, which consists of the borrower's financial history information, such as credit score, monthly income, number of delinquencies, number of prior Prosper loans, etc.[14] Let l denote the loan's risk rating assigned by Prosper, and Z_{il} denote Prosper-determined variables which include interest rate, posted loss rate, origination fee rate and monthly repayment. We should highlight that the interest rate, posted loss rate and origination fee rate have 1-1 relationship with Prosper's risk rating.

In each period t, borrower i's utility of staying with Prosper depends on the rating, l, assigned by Prosper:

$$\begin{aligned} U_{ilt}(X_i, Z_{il}; \boldsymbol{\gamma_t}) &= \gamma_{lt} + \gamma_{1t} \cdot O_l \cdot M_i + \gamma_{2t} \cdot MP_{il} + \gamma_{3t} \cdot M_i + \epsilon_{i1t}, \\ &= f_1(X_i, Z_{il}; \boldsymbol{\gamma_t}) + \epsilon_{i1t}, \end{aligned} \tag{1}$$

where $\boldsymbol{\gamma_t}$ represents the vector of borrower's side coefficients in period t; O_l is the origination fee rate for a rating l loan; M_i indicates the amount of loan i; MP_{il} is borrower i's monthly payment. Note that MP_{il} is a function of the interest rate associated with rating l and M_i. This is why borrower i's utility is a function of

[13] We should note that the arrival process may also change over time for various reasons. This can be one of the sources for the concept drift problem.

[14] Each borrower can only have one submitted loan application at any point of time. So we use listing i and borrower i interchangeably. A borrower can submit another loan application once his previous loan application is funded, expired or withdrawn. A repeated borrower will receive a new index value, and the prior Prosper loans variable will capture whether he/she is a repeated borrower.

rating l. We add subscript t to the utility function here to show that if the same listing arrives in a different period, the utility can be different because Prosper may have changed its belief and therefore may assign it with a different rating. For the same reason, we have subscript t in all the lender side model.

For a 36-month loan, monthly payment is calculated as follows:

$$MP_{il} = R^*_{il} \cdot M_i / [1 - (1 + R^*_{il})^{-36}],$$

where $R^*_{il} = R_l/12$ and R_l represents the annual interest rate a borrower needs to pay for a listing with rating l; M_i is the amount requested by borrower i.

Note that if the interest rate set by Prosper is too high, a borrower may find it more attractive to take out a loan somewhere else and withdraw her loan application. Moreover, the macro environment can also affect an individual's likelihood to participate in Prosper. Therefore, we model a borrower's outside option value as a function of her characteristics and the macro environment,

$$U_{i0t} = f_0(X_i, E_t; \boldsymbol{\gamma_t}) + \epsilon_{i0t}, \tag{2}$$

where E_t represents macroeconomic variables (e.g., S&P 500 closing quotes, the TED spread, the U.S. 30-year mortgage rate, etc.). Assuming type-I extreme value distribution for the idiosyncratic errors ϵ_{i1} and ϵ_{i0}, the withdrawal probability of listing i with rating l is:

$$W_{ilt} = Pr(\text{Withdraw} = 1 | X_i, Z_{il}, E_t; \boldsymbol{\gamma_t}) = \frac{1}{1 + \exp(f_1(X_i, Z_{il}; \boldsymbol{\gamma_t}) - f_0(X_i, E_t; \boldsymbol{\gamma_t}))}. \tag{3}$$

The full set of variables included in X_i, Z_{il} and E_t are listed in Table 3. We should highlight that the parameter vector, γ_t, depends on t. This is because we allow Prosper to use new information available in each period to update its model and adapt to an non-stationary environment with concept drifts.

In each period, Prosper observe which listings are funded, withdrawn or defaulted, and use the new information to update its models accordingly. Suppose that some listings were withdrawn in period t. At the end of period t, Prosper labels these listings as withdrawn and updates the model that predicts a listing's withdrawal probability by taking these newly withdrawn listings into account.[15] For all the loan applications arrive in period $t + 1$, Prosper uses the updated models to predict their withdrawal probabilities. Prosper updates its lender side and risk assessment model in the same way.

4.2 Lender Side

Because of the main focus of our research is to infer Prosper's adaptive learning algorithm, we abstract away modeling the lender's dynamic investment decisions. Instead, we focus on modeling the likelihood that a loan is funded (hereafter, we refer it to the loan's funding probability). Predicting a loan's funding probability

[15] Prosper does not use listings that are still open for investment to update its model.

(as a function of loan rating along with other characteristics) should be of the first order importance to Prosper because it plays a significant role in determining its expected revenue.[16] In each period t, the probability that listing i getting funded (hereafter we refer it to "funding probability") is modeled as follows:

$$F_{ilt} = Pr(\text{Funded} = 1 | X_i, Z_{il}, E_t; \boldsymbol{\beta_t}) = \frac{\exp(\beta_{lt} + X_i\beta_{1t} + Z_{il}\beta_{2t} + E_t\beta_{3t})}{1 + \exp(\beta_{lt} + X_i\beta_{1t} + Z_{il}\beta_{2t} + E_t\beta_{3t})}, \tag{4}$$

where the definitions of X_i, Z_{il} and E_t are the same as those in the borrower side model; $\boldsymbol{\beta_t}$ represents the parameter vector on lender's side. Again, the parameter values is a function of t, because we assume Prosper updates its lender side model in each period using its adaptive learning algorithm.[17]

Although our models of borrowers and lenders are relatively simple, we believe it reflects what a company like Prosper needs to do. In practice Prosper needs to re-calibrate these models very quickly in each period, in order to apply them to make their loan rating assignment decisions. By contrast, a more complex structural model of borrowers or lenders behavior could take up to weeks to estimate. By the time Prosper obtains the results, such a calibrated model may already be outdated.

4.3 Risk Assessment Model: Naive Bayes Classifier

In addition to updating the borrower side and lender side models, Prosper also needs to re-evaluate its risk assessment model in each period. In this section, we explain the model for evaluating each loan's default probability, and its loss given default (hereafter LGD), which is the share principal lost if a borrower defaults.

We will use Naive Bayes classifier in our study to assess the risk of each loan application. In practice, a naive Bayes classifier performs very well compared with other machine learning algorithms in predicting financial risk. Viaene et al. (2002) compared different classification methods in the context of expert automobile insurance claim fraud detection and find that the naive Bayes method outperformed other methods. Wang et al. (2003) showed that Naive Bayes outperformed decision trees in dealing with the concept drift problem. Humpherys et al. (2011), Domingos and Pazzani (1997), Friedman et al. (1997) also provided evidence to support the superior performance of naive Bayes classifier in other contexts. In addition, Naive Bayes model can be estimated very quickly. This is important here because Prosper needs to evaluate the risk for a large number of loan applications in each period, after incorporating new data points.

Let us explain the Naive Bayes model. Suppose that we have a labeled data set $\{(X_i, y_i), i = 1, 2, ..., n\}$,

[16] Note that Prosper only released lender's investment decisions during their auction period, but did not release this level of micro data since switching to the posted price business model. Zhang and Liu (2012) study the dynamic investment decisions of lenders in Prosper during the auction period.

[17] At the end of period t, Prosper runs its adaptive learning algorithm by taking into account the the listings that are funded or expired in that period, and re-estimates the model parameters by using the newly selected past data. It then uses this updated model to predict the funding probability of the loan applications arrived in period $t + 1$.

where X_i is a vector of characteristics for listing i; and $y_i \in \{1, 2, ..., K\}$ represents listing i's label (e.g., the label can indicate whether loan i defaults or not). Assume each listing has m characteristics, which can be represented by $X_i = (X_{i1}, X_{i2}, ..., X_{im})$. Let $p(y_i = k)$ be the prior belief about the label. According to the Bayes's rule, the posterior belief is:

$$p(y_i = k|X_i) = \frac{p(y_i = k) \cdot p(X_i|y_i = k)}{p(X_i)} = \frac{\omega_k \cdot p(X_i|y_i = k)}{p(X_i)}, \tag{5}$$

where $\omega_k = p(y_i = k)$. To approximate the initial prior belief on a loan's default probability, we use the empirical default rate for all loans prior to the period under consideration. To apply equation (5), one challenge is to obtain $p(X_i|y_i = k)$. Estimating this joint conditional distribution is very data-demanding when X_i consists of many variables.[18] The Naive Bayes estimator uses several simplifying assumptions to by-pass this hurdle.

First, given the category y_i, it assumes that listing characteristic X_{ij} are conditionally independent of each other. That is,

$$p(X_i|y_i = k) = \prod_{j=1}^{m} p(X_{ij}|y_i = k),$$

Therefore, we have,

$$\begin{aligned} p(y_i = k|X_i) &\propto \omega_k \cdot p(X_i|y_i = k) \\ &\propto \omega_k \cdot \prod_{j=1}^{m} p(X_{ij}|y_i = k). \end{aligned}$$

Second, it assumes the conditional distributions $p(X_{ij}|y_i = k)$ follows a multinomial distribution with the number of trials equal to 1. For example, suppose that the jth listing characteristic is income and it takes four levels: 1, 2, 3 and 4. If borrower i has income level 2, then $X_{ij} = (X_{ij1}, X_{ij2}, X_{ij3}, , X_{ij4}) = (0, 1, 0, 0)$.

Let the vector $P_{kj} = (p_{kj,1}, ..., p_{kj,n_j})$ denote the probability distribution of characteristic j given $y_i = k$, where n_j represents the number of values or levels that characteristic j can take. We have

$$p(X_{ij}|y_i = k) \propto \prod_{h=1}^{n_j} p_{kj,h}^{X_{ijh}}.$$

Moreover, we have

$$P(X_i|y_i = k) = \prod_{j=1}^{m} p(X_{ij}|y_i = k) = \prod_{j=1}^{m} \prod_{h=1}^{n_j} p_{kj,h}^{X_{ijh}}. \tag{6}$$

To obtain an estimate for P_{kj}, we use the crude frequency estimator. That is, for each characteristic j, we use the empirical frequency of its values conditional on the data with label k, up until the current period. Note that in contrast to estimating the conditional joint distribution, it is far much less data-demanding to

[18] If X_i consists of 10 variables, and each variable takes 5 values, there will be 9.7 million cells. The true distribution could be very sparse for many cells, and that will require a very large sample size to obtain a precise estimate.

estimate the conditional marginal distribution of each characteristic precisely because it only takes a handful number of values. Hence, even a sample size with just a few hundred observations can give us a fairly precise estimate.

In the actual implementation, we keep updating ω_k and P_k over time. We use $\boldsymbol{\Omega_t} = \{\omega_{1t}, \omega_{2t}, ..., \omega_{Kt}\}$ to represent prior belief at time t and $\boldsymbol{P}_{kt} = (P_{k1t}, P_{k2t}, ..., P_{kmt})$ to represent the joint distribution of all m characteristics in class k at time period t. The details on how we update $\boldsymbol{\Omega_t}$ and $\boldsymbol{P}_{kt}$ can be found in Appendix B.

The naive Bayes classifier assumes every characteristic is independent conditional on category. Although this assumption seems quite strong, the naive Bayes classifier performs very well in practice (e.g., Viaene et al., 2002; Wang et al., 2003). Hand and Yu (2001) pointed out that because the naive Bayes model is less susceptible to the sparsity of data in the multidimensional space (i.e., the curse of dimensionality problem) compared with other models that assume dependence between characteristics, it usually result in lower variance for the estimates of classification probabilities. Often times, the reduction in variance resulting from the relatively few parameters involved in the independent model can compensate for any increase in bias due to the naive independence assumption.

We apply the naive Bayes classifier to estimate each listing's expected default probability (ED) and expected LGD ($ELGD$). In particular, in period t, we use ED_{ilt} and $ELGD_{ilt}$ to represent listing i's expected default probability and loss given default, respectively. Notice both ED_{ilt} and $ELGD_{ilt}$ are functions of rating. The reason is different ratings correspond to different interest rates and therefore a borrower needs to pay different monthly payment to the lenders, which in turn will affect the borrower's default probability and loss given default. When estimating a listing's LGD, we discretize the LGD into four groups using the 25%, 50% and 75% quantiles, and apply the standard naive Bayes classifier to the discretized data to get each listing's probabilities of belonging to each group. At the same time, we compute the average LGD, which is denoted as $\overline{ELGD}_{kt}$, in each loss group k at time period t, $k = 1, 2, 3, 4$. Then, the predicted LGD for listing i with rating l, $ELGD_{ilt}$, is defined as $ELGD_{ilt} = \sum_{k=1}^{4} p_{ilk} \cdot \overline{ELGD_{kt}},\ k = 1, 2, 3, 4$, where p_{ilk} is the probability that listing i belongs to LGD group k if it is assigned with rating l. For default prediction, We use $\overline{\boldsymbol{\Omega}}_{\boldsymbol{dt}}$ and $\overline{\boldsymbol{P}}_{\boldsymbol{dt}}$ to denote all the priors and posteriors beliefs, respectively. For LGD, we use $\overline{\boldsymbol{\Omega}}_{\boldsymbol{lt}}$ and $\overline{\boldsymbol{P}}_{\boldsymbol{lt}}$ to denote all the priors and posteriors beliefs, respectively.

We should point out that at any given period, some loans will still be "on-going," and we do not know whether they will eventually default or not; if so, what their LGDs are. Therefore, we follow Prosper and define a loan as defaulted if a borrower misses 4 repayments in a row.

4.4 Firm's Objective Function (Indirect Utility Function)

Prosper considers each loan application i separately, and assigns it with a risk rating l to maximize the expected utility from this loan application. The decision is based on the state vector $S_t = (\boldsymbol{\gamma_t}, \boldsymbol{\beta_t}, \overline{\boldsymbol{\Omega}}_t, \overline{\boldsymbol{P}}_t)$, where $\boldsymbol{\gamma_t}$ and $\boldsymbol{\beta_t}$ are borrower and lender side parameters explained in sections 4.1 and 4.2, respectively; $\overline{\boldsymbol{\Omega}}_t = (\overline{\boldsymbol{\Omega}}_{dt}, \overline{\boldsymbol{\Omega}}_{lt})$, and $\overline{\boldsymbol{P}}_t = (\overline{\boldsymbol{P}}_{dt}, \overline{\boldsymbol{P}}_{lt})$ denote the the prior and posterior beliefs parameters for the default and lost given default prediction models explained in section 4.3. Notice that interest rates and posted loss rates at each rating level are determined outside of our model. In our sample period, Prosper seldom changes the interest rate and posted loss rate associated with each risk category. This study will focus on modeling Prosper's rating assignment decisions and take its mapping to interest rates as given. Prosper's utility of assigning rating l to listing i posted in t is:

$$U_{ilt}(X_i, Z_{il}|S_t, AL, \alpha, \delta_1, \delta_2) = \alpha_l + \delta_1 \cdot (O_l + L_{ilt}) \cdot M_i \cdot F_{ilt} \cdot (1 - W_{ilt}) + \delta_2 \cdot |PLR_l - ED_{ilt} \cdot ELGD_{ilt}| + \epsilon_{ilt}, \tag{7}$$

where O_l is the origination fee rate for a rating l listing; L_{ilt} is the expected annual loan service fee rate that Prosper charges the lenders;[19] M_i is the loan amount requested; F_{ilt} and W_{ilt} are the predicted funded and withdrawal probability, respectively; PLR_l is the posted loss rate; ED_{ilt} and $ELGD_{ilt}$ are the predicted default probability and loss given default, respectively, and $ED_{ilt} \cdot ELGD_{ilt}$ is the expected loss rate for listing i given rating l; δ_1 is the utility weight on expected revenue; δ_2 captures the cost of misreporting the listing's risk (e.g., it may damage Prosper's long-term reputation, and affects its future profits); α_l's are the alternative specific intercepts; AL denote a specific adaptive learning algorithm used in step 1 estimation to obtain $F_{ilt}, W_{ilt}, ED_{ilt}$ and $ELGD_{ilt}$; ϵ_{ilt} is the random utility term which follows the type I extreme value distribution.

Note that O_l and PLR_l do not depend on i because they are rating specific. Note also that it is almost surely $|PLR_l - ED_{ilt} \cdot ELGD_{ilt}| \neq 0$ because PLR_l is rating specific, and can only take one of seven values, but $ED_{ilt} \cdot ELGD_{ilt}$ is continuous. Our intuition suggests that $\delta_1 > 0$ and $\delta_2 < 0$. As we report our results in the

[19] Notice that the annual service fee is charged monthly based on the current outstanding loan principal a lender has. So if a borrower defaults on his loan, lenders who invested in this particular loan will lose part of their principal, and Prosper cannot collect a service fee from lost principal. Therefore, the expected service fee rate for listing i is a function of the annual service fee rate and when the loan will default. The later the loan defaults, the more service fee Prosper can charge from the lenders. In our data sample, the service fee rate is fixed at 1%. We approximate the number of repayments borrower i has made as $PT_{ilt} = \textbf{round}((\text{1-ELGD}_{ilt})/\frac{1}{36}) = \textbf{round}((\text{1-ELGD}_{ilt}) \cdot 36)$. The monthly service fee rate is $r_s = 0.01/12$. Since the service fee is charged monthly on the current outstanding loan principal, L_{ilt} is calculated as

$$L_{ilt} = \begin{cases} 0, & PT_{ilt} = 0; \\ ED_{ilt} \cdot r_s \cdot \sum_{j=1}^{PT_{ilt}} (37 - j)/36 & PT_{ilt} = 1, 2, ..., 36. \end{cases}$$

Hence, L_{ilt} is a function of ED_{ilt} and $ELGD_{ilt}$.

next section, the estimated signs of these two structural parameters confirm our intuition. Like standard discrete choice models, $\frac{\delta_1}{\delta_2}$ will determine how Prosper does the trade-off between expected profits and reporting the truth risks of a listing, when it set the loan rating. If the estimated $\frac{\delta_1}{\delta_2}$ is very close to zero, then the data reveals that Prosper is practically only interested in using the rating to reflect a listing's true risk.

We assume Prosper takes a two-step approach to determine loan ratings. First, it uses an adaptive learning algorithm to estimate the predictive models in sections 4.1 and 4.2 and obtain $F_{ilt}, W_{ilt}, ED_{ilt}$ and $ELGD_{ilt}$. Second, it then takes the predicted $F_{ilt}, W_{ilt}, ED_{ilt}, ELGD_{ilt}$ as given, and choose a loan rating to maximize its objective function for each loan listing.

$$l^* = arg\max_{l} \left[U_{ilt}(X_i, Z_{il}|S_t, AL, \theta)\right]. \tag{8}$$

It follows that the likelihood function We can then write the likelihood function in the following closed form:

$$L(X, Z|\mathbf{S}, AL, \theta) = \prod_{t=1}^{T}\prod_{i=1}^{I_t}\prod_{l=1}^{7}\left(\frac{\exp(U_{ilt}(X_i, Z_{il}|S_t, AL, \theta))}{\sum_{j=1}^{7}\exp(U_{ij}(X_i, Z_{ij}|S_t, AL, \theta))}\right)^{\mathbb{1}\{\text{Rating}_i^o = l\}}, \tag{9}$$

where $\mathbf{S}$ denote the state variables in all periods; T represents the total number of time periods; I_t represents the number of listings in period t; Rating_i^o is listing i's observed Prosper rating. We estimate the structural parameters of Prosper's objective function $\theta = (\alpha_1, ..., \alpha_7, \delta_1, \delta_2)$ using maximum likelihood.[20] It is important to highlight that the likelihood function is conditioning on an adaptive learning algorithm, AL, and so as Prosper's policy function. Hence, when we repeat two-step estimation exercise for each adaptive learning algorithm that we consider, we can compare the goodness-of-fit of this set of policy functions and find out which explains the data the best. This is essentially our inference strategy.[21] Figure 4 illustrates our model framework. We now turn to discuss the set of adaptive learning algorithms that we consider.

5 Adaptive Learning Algorithms

According to Tsymbal (2004), there are three general approaches of handling the concept drift problem: (1) observation selection; (2) observation weighting; and (3) ensemble learning. The goal of the observation selection approach is to select observations relevant to the current situation to estimate the model. Observation weighting is about assigning a weight to each observation depending on its relevance to the current situation. Ensemble learning method maintains multiple individual models and use them to make a combined prediction. Each individual model relies on its own observation selection and weighting

[20] Note that one of the utility intercept needs to normalized for identification reason. We normalize $\alpha_7 = 0$, where alternative 7 corresponds to the HR rating. Similarly, alternative 1 corresponds to A rating.

[21] Similar two-step estimation approach has been used in the structural estimation literature (e.g., Bajrari et al. 2007; Benkard, 2004; Ching, 2010; Diermeier et al. 2005; Erdem et al. 2003; Hendel and Nevo, 2006; Yao et al. 2012).

approach. We consider five different adaptive learning algorithms, which represent these three approaches well. In particular, algorithms (1)-(4) rely on observation selection and weighting approaches, and algorithm (5) makes use of all three approaches.

Of course, there are many more adaptive learning algorithms that Prosper can use. Unfortunately, it is impossible to examine every algorithm. Our approach follows McCullagh and Nelder (1983), who argue that "all models are wrong; some, though, are better than others and we can search for the better ones." As a proof-of-concept study, our goal is to demonstrate that our framework allows us to compare different adaptive learning algorithms and see which one is closest to what Prosper actually uses. If researchers are interested in other adaptive learning algorithms, they can apply our framework to evaluate them as well.

We emphasize that the model framework remains the same for all algorithms discussed below. Each algorithm leads to its own set of parameter values in the borrowers and lenders models (the withdrawal, funding, default probabilities and loss given default), which go into the firm's objective function as we explained in section 4.

5.1 Equal Weight Method

The first algorithm that we consider is the Equal Weight Method (EWM). This is by far the most commonly used method when companies or researchers face limited sample size. It simply makes use of all the data available and weight them equally. Its underlying assumption is that there is no concept drift and the data generating process remains unchanged over time.[22] We treat this algorithm as the benchmark case. For EWM, the length of a period is a day. That is, Prosper updates its model every day.

5.2 Moving Window Method (MWM)

The second method we consider is the Moving Window Method (MWM), which belongs to the observation selection approach. MWM is one of the most widely used observation selection approaches to deal with the concept drift problem (Pechenizkiy et al. 2010, Forman 2006). At each time step, the algorithm re-estimates the parameters of the model using the data from the new training window. The model is updated in the following two processes: a learning process (update the model based on the new window) and a forgetting process (discard older data). We consider the most recent N periods as a fixed window size. As the window moves forward, new arrived observations are added to the window and oldest observations are discarded. We use grid search to determine the optimal window size. To selection is done by using the receiver operating characteristic (ROC) curve as the measurement metric for funded, withdrawal and default probabilities

[22] Kozlowski et al. (2020) model consumers as adaptive learners and assume that they use the equal weight method to update their belief about economic shock distribution over time.

predictions, and mean square error (MSE) for LGD prediction. We provide more details about ROC curve in Appendix D.

The optimal window size is 7 months for the withdrawal probability model; and the optimal window sizes are 3, 36 and 48 months for funding probability, default probability and LGD models, respectively. Notice the optimal window sizes for the default prediction and LGD prediction are much larger. This is because only 70.04% of the listings get funded, and among the funded listings, only 15.84% of them defaulted. Hence, the number of defaulted loans is much smaller than the number of listings. The default prediction and LGD prediction models need larger window sizes to have enough observations to reach better prediction performance. We use data between Feb 2007 and Dec 2010 as our initial data set. Detailed prediction performance can be found in Table 6.

5.3 Recession Probability Method (RPM)

Our third adaptive learning algorithm is Recession Probability Method (RPM), which is a moving window based method with varying window size. More generally, it is an observation selection approach. This method selects observations based on the closeness of economic conditions. The intuition is that borrowers and lenders behavior should depend on the economic environment.

To define similar economic environments, we make use of the Smoothed U.S. Recession Probabilities released by the Federal Reserve Bank. U.S. Recession Probabilities are the smoothed probabilities of a recession in the U.S., which are calculated from a dynamic-factor Markov-switching model on non-farm payroll employment, industrial production index, real personal income, real manufacturing, and trade sales. This model was originally developed by Chauvet (1998) and has been used to study stock market volatility business cycle turning points in many studies including Sornette (2017), Kim and Nelson (1998), Hamilton and Lin (1996), etc.

We need to find the optimal similar economic environments, or in other word, the optimal windows, for borrower side, lender side, default prediction and LGD prediction models. Take the borrower side model for example, the optimal similar economic environment should enable the model to make best predictions for borrower's withdrawal decisions. Assume month t's recession probability is RP_t. We find that a model estimated using data generated from months whose recession probability is within [RP_t-0.4%, RP_t+0.4%] can best predict borrower's withdrawal decisions. Hence, the optimal marginal recession probability for the borrower side model is 0.4%. As RP_t changes over time, the observations within the range [RP_t-0.4%, RP_t+0.4%] change as well. That is to say, the size of the training window is varying over time. The corresponding optimal marginal recession probabilities for lender side model, default prediction model and LGD prediction model are 0.3%, 3.1% and 1.1%, respectively. As in the previous section, the optimal marginal recession probabilities are selected by examining each model's out of sample predictions. We use the data from February 01, 2007 to December 19, 2010 as our initial data set. That is, we select similar

economic environment from February 01, 2007 to December 19, 2010 when we begin our model estimation. Therefore, we use 71 months data in total for our estimation. For RPM, since the recession probability is a monthly measure, the length of a period is a month. Even though RPM is updated each month, we still use the model to make daily decisions. For instance, we use the model updated on June 30th, 2012 to make rating assignment decisions for every day in July, 2012.

Unlike MWM, RPM does not discard older information in a mechanical way. Instead, it utilizes older information in a more complicated way. The rationale behind RPM is that consumer behavior is highly correlated with the economic environment. Therefore, taking advantage of historical consumer behavior data that is generated from similar economic environments as today should be beneficial for the current period's prediction problems.

5.4 Gradual Forgetting Method (GFM)

The fourth algorithm that we consider is the Gradual Forgetting Method (GFM), which is an observation weighting approach.

As Koren (2009) pointed out, one disadvantage of moving window method is that it gives the same weight to all observations within the window being considered, while completely discarding all observations outside the window. This may be reasonable when the drift is abrupt, but less so when the drift is gradual. Thus, a refinement is observation weighting, depending on its relevance to the current situation. Many studies find that observation weighting approach is able to improve the model's adaptability to drifting concepts (Koren, 2009; Klinkenberg, 2004). To implement GFM, we follow Klinkenberg (2004) and weight observations according to their age using an exponential aging function. To be specific, let $\lambda \in (0,1)$ denote the forgetting factor. Then all the observations in period t will be assigned with weight λ^t. The older an observation is, the less weight it carries. To estimate the logistic regression models for the borrower side and lender side, we employ the estimation method proposed by Balakrishnan and Madigan (2008). This method is based on a quadratic Taylor approximation to the log-likelihood. A forgetting factor can be easily incorporated into this estimation scheme and the model can be recursively estimated, which significantly reduces the computational burden. We provide the estimation details of this method in Appendix C.1. With respect to the prediction about default and LGD, we incorporate a forgetting factor to introduce temporal adaptivity in the naive Bayes model as well, and we explain the details in Appendix C.2. For GFM, the length of a period is a day. That is, Prosper updates its model every day.

For each of the four models (funding probability, withdrawal probability, default probability, and LGD models), we employ a grid search over the forgetting factor. As in section 5.2, we compare the their out of sample prediction performance with different forgetting factors. We find that the best forgetting factor in terms of prediction performance for the funding probability, withdrawal probability, default probability, and LGD models are 0.997, 0.990, 0,995 and 0.997, respectively. Note that these are daily forgetting factors. Take

the default prediction model as an example. An observation which is one year "old" only carries weight of $0.995^{365} = 0.16$. The full prediction performance of the four models can be found in Table 6.

5.5 Ensemble Recession Probability Method (E-RPM)

The fifth algoirthm that we consider is Ensemble Recession Probability Method (E-RPM), which makes use of both observation weighting and ensemble learning approaches.[23]

In all the previous exercises, we try to create one model to deal with concept drift. However, in most cases, it is hard to make good predictions by using a single model. In the machine learning literature, researchers find that instead of relying on a single model, the simple average of multiple individual models can be a powerful heuristic that captures "the wisdom of the crowd." Even if each individual model might be weak, the aggregated model can perform very well in prediction. The literature refers this approach to an *ensemble* method.

In E-RPM, we applies this ensemble approach to RPM. To train a model which specializes in a specific economic environment, we first use the recession probability index to divide the whole data set into different groups. For instance, data generated from recession period is assigned to the recession group, while data generated from economic booms is assigned to the booming group. We then estimate a model using each data group separately. We take a weighted average of each individual model to get our ensemble model. Note that each sub-model has the same model framework, as shown in Figure 4. The only difference among the sub-models is that they are estimated using data from different economic environments. As in RPM, we use the data from February 01, 2007 to December 19, 2010 as our initial data set. That is, we select similar economic environment from February 01, 2007 to December 19, 2010 when we begin our model estimation. Therefore, we have 71 months data in total for our estimation. For E-RPM, the length of a period is a month. Similar with RPM, even though E-RPM is updated each month, we still use the model to make daily decisions. As an example, we will show how to use E-RPM to train each individual model and determine weights to predict each listing's funding probability. The cases for predicting withdrawal probability, default probability and LGD are similar. If readers are not interested in the exact details, they can skip the following discussion and jump to section 6.

(i) Initial individual models: We use the data from February 01, 2007 to December 19, 2010 (i.e., the pre-posted price period) as our initial data set. We first divide the data into four sub-samples, Q_0^1, ..., Q_0^4, using the first, second, and third quantiles of the data's recession probability. Here, the subscript 0 represents the initial period. Then we train four sub-sample specific logistic regression models, C_0^1,...,C_0^4. Notice that the specifications of the four individual logistics regression models are the same. But their parameters are different because they are trained using different sub-samples of the data set. We can think of C_0^1,...,C_0^4 as

[23] Cogley and Sargent (2005) also use an ensemble modeling approach to model adaptive learning.

four experts who specialize in different economic environments.

(ii) Initial weights: Initialize the weights we put on each individual model to be $H = (0.25, 0.25, 0.25, 0.25)$.

(iii) Predictions: At time t, we denote the four individual models as C_t^1,...,C_t^4. Those four individual models are trained using data sets Q_t^1, ..., Q_t^4, respectively. Q_t^1, ..., Q_t^4 represent the four sub-samples divided by the first, second, and third quantiles of the data's recession probability in period t. We use C_t^1,...,C_t^4 to make predictions for listings arriving in period $t+1$. For listing i with characteristics X_i, Z_{li} (rating specific characteristics) and macro environment index E_{t+1} at time $t+1$, we use each individual model to predict a corresponding funding probability. Let $F_{jit} = C_t^j(X_i, Z_{li}, E_t)$ represent individual model j's predicted funding probability for listing i. We take the weighted average of F_{1it}, F_{2it}, F_{3it}, and F_{4it} as the predicted funding probability for listing i. That is:

$$F_{it} = \sum_{j=1}^{4} h_{jt} F_{jit} = \sum_{j=1}^{4} h_{jt} C_t^j(X_i, Z_{li}, E_t), \tag{10}$$

where h_{jt} is the weight we put on individual model j.

(iv) Update weights: We now explain how we update the weight for each individual model for $t > 1$. Following Wang et al. (2003), the weight for each individual model in time t is a function of the inverse of its prediction mean squared error (MSE) in time $t-1$. Let n_{t-1} be the number of listings which borrowers decided to proceed (i.e., they decide not to withdraw after knowing its Prosper rating) in time $t-1$. The MSE for individual model j is given by

$$B_{jt} = \frac{1}{n_{t-1}} \sum_{i=1}^{n_{t-1}} (y_i - F_{jit})^2, t = 2, 3, ..., T \tag{11}$$

where y_i takes value 1 if listing i is indeed funded and 0 otherwise; F_{jit} represents individual model j's predicted funding probability for listing i. That is to say, we measure an individual model's prediction performance using the model's prediction accuracy in the last period. Wang et al. (2003) proposes setting weight for poorly performed model to zero. We experimented this approach, and find that if we only assign positive weights to the top 2 best models in each period (and setting the weights for the other two inferior models to zero), we get better prediction results. Hence, we put zero weights on the individual models with the largest and second largest MSEs. For the other two individual models, we put a weight proportional to the inverse of their MSEs. Let

$$h'_{jt} = \begin{cases} 0 & \text{if individual model } j \text{ has the largest or second largest MSE in period t.} \\ [B_{jt}]^{-1} & \text{otherwise.} \end{cases}$$

Then we set the weight $h_{jt} = \frac{h'_{jt}}{\sum_j h'_{jt}}$.

(v) Update individual models: Let Q_t denote the whole data set we have at the end-of-period t. We divide the data into four sub-samples, Q_t^1, ..., Q_t^4, using the first, second, and third quantiles of Q_t's recession probability. We train individual model C_t^j using data Q_t^j.

(vi) Repeat steps (iii) to (v) until the last period of the data.

We apply the same procedures to estimate each listing's withdrawal probability, default probability and LGD.

6 Identification

In this section, we provide some intuitions about our model identification.

Recall that we take a two-step approach to estimate different parts of our model. The first step is to obtain the lender and borrower side parameters, and it requires us to re-estimate the funding probability and withdrawal probability models, Navie Bayes models for default probability, as well as LGD in each period as new data come in. Hence, all these parameters obtained in step 1 are time specific. In each period, the funding probability and withdrawal probability models can be identified from the variations in loan characteristics, lenders' investment decisions and borrowers' withdrawal decisions. Similarly, parameters in the risk assessment model can be identified from variations in borrowers' default behavior, conditional on loan characteristics. Clearly, the parameter estimates change with adaptive learning algorithms because each of them has its own way to select the available data to be used for estimation. It should be highlighted that we assume both researchers and Prosper have the same information, and we conduct the same estimation exercise. Hence, we are able to use the same predicted values of these objects in Prosper's objective function.

Step 2 is to estimate the structural parameters in Prosper's objective function $(\alpha_1, ..., \alpha_6, \delta_1, \delta_2)$, which reveals how Prosper may make the trade-off between reporting the true risk and increasing the chance of having a loan be funded. Unlike the predictive models for funding probability, withdrawal probability, default probability and LGD, which we assume Prosper needs to adaptively learn about their parameter values, we assume Prosper's objective function remains unchanged during our sample period. The intercept terms $(\alpha_1, ..., \alpha_6)$ are pinned down by the average shares of Prosper rating in the sample period. To see the intuition about how the data identify the relative values of δ_1 and δ_2, let's consider the following example. Suppose that Prosper only cares about reporting LGD truthfully. Then the variation of the expected revenue term conditioning on rating will have no influence on Prosper's rating decisions, and hence we will obtain $\delta_1 = 0$ and $\delta_2 < 0$. Now suppose that Prosper receives a loan application with relatively high risk. Suppose further that Prosper only cares about the short-term profits of earning the commission for successfully funding this application, it would assign a rating better than what it actually deserves, so as to increase the chance of getting it funded. Such choices would reveal that $\delta_1 > 0$ and δ_2 is close to zero. Basically, the identification of the structural parameters in Prosper's objective function is straightforward once we obtain $W, F, ED, ELGD$ from step 1.

If $|\frac{\delta_1}{\delta_2}|$ is close to zero, it maybe tempting to interpret that Prosper does not care about profits. But we note

that such estimates are still consistent with the hypothesis that Prosper focuses on long-term profits (e.g., total discounted profits), which is captured by the truthful reporting of loan risk term. If borrowers and lenders find that Prosper risk assessment provides additional values to standard credit scores such as FICO score, this could help attract more participants to use its platform in the future.

We need to identify how Prosper uses historical data to make decisions. We will use a simplified and hypothetical example to show the intuition. Suppose the average default rate for borrowers from State A is 50% from Jan 2011 to Dec 2012, while only 6% of borrowers from State A default from Jul 2012 to Dec 2012. Suppose further that the only observable characteristic of borrowers is which state they reside (hence there is no changes in the distribution of the observable characteristic over time). The default rate changes are shown in Figure 5. At the very beginning of 2013, Prosper needs to determine what rating to assign to a borrower from State A. If Prosper used the data from Jan 2011 to Dec 2012 to inform its decision making, it would likely expect that a borrower from State A is very risky and assign a low rating to such a borrower, given the other borrower characteristics are the same. However, if Prosper only uses data from Jul 2012 to Dec 2012 to learn about the market, it may believe that borrowers from State A are quite credible and assign a relatively good rating to such a borrower, given the other borrower characteristics are the same. From this example, we can see that different ways of using historical data will help Prosper to form very different beliefs about the market, and Prosper will make different decisions, i.e., rating assignments, under different beliefs. Therefore, by analyzing Prosper's rating assignment decisions, we can identify how Prosper uses historical data in its practice.

One potential concern is that the changes in the competitive or macro environment may lead to changes in the composition of borrowers and lenders who come to Prosper over time. Hence, the change in the observed distribution of the dependent variables (withdrawal, funded and default probabilities and LGD) may reflect shifts in the covariates of the independent variables instead of concept drift. We do not think this is a problem. First, we have controlled for covariate shift explicitly over time (see Table 3). Those covariates are included in the lender side, borrower side and risk assessment models. Some key covariates include: FICO score range, income, debt to income ratio, number of delinquencies, etc. Second, it is possible that we do not observe all relevant characteristics of borrowers, and the distribution of unobserved characteristics may change over time. But if this happens, it will change the data generating process over time, which is what concept drift captures. Recall that we use $Y_t = F(X_t, \epsilon_t, \theta_t)$ to describe the data generating process at the beginning of the Introduction (section 1). The shift in the distribution of unobserved characteristics can affect the relationship between Y_t and X_t if unobserved characteristics interact with X_t, and that would reflect in changes in θ_t over time.

7 Results

7.1 Estimates of Prosper's Objective Function

For each data selection/adaptive learning algorithm, we take the estimated model parts (i)-(iii) described in section 4 as given, and then estimate the structural parameters of Prosper's objective function using the likelihood function stated earlier. We will use the model fit w.r.t. Prosper's choice on loans' risk ratings to infer which data selection/adaptive learning algorithm is closest to the one used by Prosper. The estimates of α's, δ_1, δ_2, log-likelihood and BICs are shown in Table 5. The parameter estimates are all statistically significant. It is worth highlighting that even though we did not impose any restrictions on the sign of the estimates, the results show $\delta_1 > 0$ and $\delta_2 < 0$ in all adaptive learning algorithms, as the theory suggests.[24]

The structural parameters are all statistically significant in all five data selection algorithms. Comparing the goodness-of-fit, it is clear that E-RPM gives the best log-likelihood and BIC. Hence, we infer that among the five data selection methods, E-RPM is closest to what Prosper uses. Under E-RPM, the estimates of δ_1 and δ_2 are 4.26 and -13.98, respectively. Notice these two parameters represent the weights Prosper puts on the short-term profit part and the long-term reputation part. Both of them are significant, which suggests that Prosper takes both short-term profits and long-term reputation into account. But long-term reputation is a lot more important. To better illustrate the relative importance of the short term profit and the long-term reputation, let us consider an average listing (a loan application with average loan characteristics) in the data. If Prosper assigns rating A to this listing, the estimated loss rate is 3.10%. Notice a listing's estimated loss rate is a function of the rating because a borrower needs to pay different amount of interest depending on the rating he got, which in turn affects his default probability and LGD. Since the posted loss rate at rating A is 3.03%, the long term reputation term equals $-13.98 \cdot |3.03\% - 3.10\%| = -0.01$. At the same time, if Prosper assigns rating B to this listing, the corresponding estimated loss rate is 7.17%. Given the posted loss rate at rating A is 5.56%, the long term reputation term equals $-13.98 \cdot |5.56\% - 7.17\%| = -0.22$. On the other hand, the profit terms are 1.08 and 1.10 at ratings A and B, respectively. In this example, the reputation term dominates. Even though assigning rating B yields a slightly larger utility from the profit term, the utility decrease from the reputation term is way larger. In this case, Prosper is more likely to assign rating B instead of rating A. The profit terms plays a more important role when switching the rating assignment can lead to large increase of profit but relatively small decrease of reputation. This happens more frequently between ratings AA and A, as well as B and C. This is because there is a origination fee rate jump from AA to A and B to C (Table 1).

[24] Note that there is no reason for us to expect that the structural parameter estimates of Prosper's objective function to be close across different adaptive learning algorithms because the distribution of (F, W, D, LGD) should be very different across algorithms.

In the second example, let us consider the median listing (a loan application with median loan characteristics) in our sample. If Prosper assigns rating B to this listing, the estimated loss rate would be 5.93%. Given the posted loss rate for rating B listings is 5.56%, the long term reputation term equals $-13.98 \cdot |5.93\% - 5.56\%| = -0.05$. While if Prosper assigns rating C to this listing, the estimated loss rate and corresponding long term reputation term are 6.78% and $-13.98 \cdot |6.78\% - 7.94\%| = -0.16$. If Prosper only cares about truthfully reporting the risk, it should choose to assign rating B to this listing. However, the utility from profit terms are 0.89 and 1.03 for rating B and C, respectively. Taking both the short-term profit and long-term reputation into account, Prosper is more likely to assign rating C to the median listing.

7.2 A Closer Look at E-RPM

The estimation results in the previous section suggest that among the set of adaptive learning algorithms being considered, Prosper is more likely to use an E-RPM (ensemble recession probability model). We now take a closer look at how well E-RPM model predicts customer behavior compared with other algorithms and how Prosper adjusts the weights on each individual model over time. We first compare E-RPM's prediction performance on listings' funding, withdrawal, default, and LGD outcomes against EWM, MWM, RPM and GFM. We use the receiver operating characteristic (ROC) curve as the measurement metric for funding, withdrawal and default predictions (since they are all binary prediction problems) and use MSE for LGD prediction.

Roughly speaking, the larger the area under ROC (AUC), the better the model's overall prediction performance is. Table 6 summarizes the comparison results. For funding probabilities, E-RPM's AUC is 0.854 and it outperforms the other four models; MWM with AUC of 0.847 is the runner-up. For default probabilities, E-RPM's AUC is 0.617 and it also performs the best; however, GFM's AUC is 0.615, which comes very close to E-RPM; it is worth pointing out that for withdrawal prediction, MWM with AUC equals 0.618 performs the best; E-RPM with AUC equals 0.595 is the runner-up. In addition, E-RPM gives the smallest MSE for the LGD prediction. This piece of evidence shows that the way Prosper uses historical data (E-RPM) can help Prosper better understand customer behavior compared with using all historical data equally (EWM), mechanically discounting (GFM) or cutting data into fixed-length windows (MWM). Figures A1, A2, A3 and A4 visualize the ROC comparisons. It is worth noting that RPM and E-RPM are quite close w.r.t. these three ROC curves.

Next, we examine how Prosper adaptively adjusts the weights it puts on each individual model in E-RPM. Figure 6 displays the weights on the individual models that predict listing's funding outcomes. In 2011, the weights on different models change frequently and significantly, which suggests that the market was changing fast and the model needs to experiment with different weights to better learn about the market. During 2012, adjustments of weights are less frequent and smaller than in the previous year. For instance, the weight on individual model 4 is 0 in most times. The findings are consistent with the fact that Prosper

just changed its business model in Dec 2010. In 2011, both lenders and borrowers need to learn how to participate in this platform and their behavior may change more dramatically in this early stage. At the same time, Prosper tends to actively experiment with different policies to make sure their new business model could succeed. Hence, 2011 is characterized with fast changing lender and borrower behavior, as well as experimental platform policies. In contrast, 2012 is a relative stable phase. Lenders and borrowers know better about how the new business model works and Prosper also has more information on how to satisfy its customers. Changes in weights become less frequent and less dramatic.

Changes in weights on withdrawal prediction models exhibit similar pattern. While for default prediction models, the weights remain unstable by the end of the second year. The reason is default is less often to happen. E-RPM needs to actively experiment with different weights to better learn about borrower's default behavior over this two-year period. Figures A5 and A6 visualize the patterns.

7.3 Counterfactual Experiments

Our research finds evidence that Prosper uses a relatively sophisticated ensemble adaptive learning algorithm to deal with a general concept drift problem. To shed some light on the importance of using the E-RPM method to adapt, we conduct a series of counterfactual experiments by assuming (i) Prosper does not care about expected short-term profits, (ii) Prosper does not adaptively learn at all, (iii)-(vi) Prosper uses one of the four other methods to adapt. In experiment (i), we set $\delta_1 = 0$, and keep α's and δ_2 remains unchanged in Prosper's objective function; we also keep their beliefs about each listing's funding, withdrawal, default and LGD at each rating l $(F_{ilt}, W_{ilt}, ED_{ilt}, ELGD_{ilt})$ unchanged as in the actual scenario. In all other experiments, we assume the structural parameters (α's, δ_1, δ_2) remain unchanged, but their beliefs about $(F_{ilt}, W_{ilt}, ED_{ilt}, ELGD_{ilt})$ changes according to counterfactual adaptive learning method.

Experiment (i) assumes that Prosper does not take the short-term profits into account. We set $\delta_1 = 0$. The counterfactual rating distribution is reported in Figure 7. Interestingly, the rating distribution now has thicker tails on both ends – Prosper assigns more AA and HR loans. It is sensible to assign more AA loans because they are actually less profitable due to the lower commission fee. It also makes sense that we see more HR loans because they typically have high withdrawal probability and low funding probability, and it is less likely that they get funded and generate revenue. In other words, both AA and HR are the least profitable loans. Because Prosper cares about short-term profits, it has assigned fewer AA and HR loans in the actual scenario (about 900 less for AA loans, and 300 less for HR loans).

In experiment (ii) (no adaptive learning), we assume Prosper only used data in the initial period to estimate the models to predict $(F_{ilt}, W_{ilt}, ED_{ilt}, ELGD_{ilt})$, and then we simulate Prosper rating choices using the structural parameter estimated in step 2. The counterfactual rating distribution is reported in Figure 8. We can see that Prosper would assign much more loans to AA, A and B rating. It doubled the number of AA loans, and categorized around 1,500 more A loans. Prosper would also assign fewer loans to D, E and HR

ratings. In particular, the number of HR loans drops by around 2,000.[25] The results show that if Prosper always holds her beliefs calibrated by the data during the auction period, she will tend to assign higher ratings to loan listings. The main reason is in the auction period, funding probability is quite low in general, especially for listings with poorer ratings. Holding the belief calibrated in this counterfactual experiment, Prosper tends to assign listings with better ratings to increase their probability of getting funded, in an attempt to obtain higher expected revenue.

In experiment (iii), we assume Prosper use EWM method to update its belief. The counterfactual rating distribution is shown in Figure 9. The results here are qualitatively similar to experiment (ii). It still predicts that Prosper would assign more loans to AA, A and B ratings, and less loans to D, E and HR ratings. But the differences are not as large as experiment (ii).

In experiment (iv), we assume Prosper used MWM to adaptively learn. Figure 10 shows the simulated ratings under this experiment has a thicker right tail in E and HR loans, but they are not too different from E-RPM.

Experiment (v) assumes Prosper used RPM instead of E-RPM. Figure 11 shows the simulated ratings under this experiment are also not too different from E-RPM, but are slightly thicker on the left tail in AA, A, B loans.

Experiment (vi) assumes Prosper uses GFM. Figure 12 shows the simulated ratings under this experiment are much heavier on the left tail in AA, A, B loans compared with E-RPM.

We further analyze how the above counterfactual experiments may affect Prosper's revenue. Before we discuss the results, it is important to stress that it is Prosper's beliefs about listing i's funding, withdrawal, default probabilities and LGD with rating l (hereafer $F^P_{ilt}, W^P_{ilt}, ED^P_{ilt}, ELGD^P_{ilt}$) that go in her objective function. These beliefs are not supposed to be unbiased estimates of their true values (hereafter $F^T_{ilt}, W^T_{ilt}, ED^T_{ilt}, ELGD^T_{ilt}$). In order to map Prosper's choices to expected revenue, we need to estimate $(F^T_{ilt}, W^T_{ilt}, ED^T_{ilt}, ELGD^T_{ilt})$. How do we achieve this? From the econometrician's viewpoint, we use data of moving windows of [t - 14 days, t + 14 days] to estimate funding and withdrawal probability models, and [t - 30 days, t + 30 days] for default probability and loss given default models. Let's highlight the differences between the econometrician's estimate of their true values vs. Prosper's adaptive learning of their values: at any given t, as an adaptive learner, Prosper can only use data available at the beginning of t, i.e., she can only use data before t; in contrast, an econometrician can use all data available for research, i.e., he can use data before and after t. Because an econometrician can use data from both sides of the neighborhood of t, we should be able to more accurately uncover the true values of $(F^T_{ilt}, W^T_{ilt}, ED^T_{ilt}, ELGD^T_{ilt})$. Specifically, given an objective

[25] Recall that the initial parameter values for $(F_{ilt}, W_{ilt}, ED_{ilt}, ELGD_{ilt})$ are calibrated based on the data prior to Dec 2010, which is the auction period. We expect that the data during the auction period can still provide Prosper with some decent information about the borrowers' and lenders' behavior.

function or adaptive learning method (AL), we calculate Prosper's estimates of $(F^P_{ilt}, W^P_{ilt}, ED^P_{ilt}, ELGD^P_{ilt})$. Then, we use the structural parameters $(\alpha's, \delta_1, \delta_2)$ estimated from the E-RPM method to obtain Prosper's rating choice probabilities for each listing i, $P_{ilt}(AL)$. To map $P_{ilt}(AL)$ to expected revenue, we need to use $(F^T_{ilt}, W^T_{ilt}, ED^T_{ilt}, ELGD^T_{ilt})$: The expected revenue of listing i is:

$$E[Revenue_{it}|AL] = \sum_l (O_l + L^T_{ilt}) \cdot M_i \cdot F^T_{ilt} \cdot (1 - W^T_{ilt}) \cdot P_{ilt}(AL).$$

Definitions of the variables can be found in section 4.4. Notice L^T_{ilt} has superscript T because it is a function of ED^T_{ilt} and $ELGD^T_{ilt}$ (see footnote 19). Then we sum up all the expected revenue across listings to obtain the total expected revenue in the two year period. Tables 7 reports the counterfactual expected revenue and their percentage differences across different counterfactual scenarios. In addition to the counterfactual discussed above, we also add one that assumes Prosper knows the true data generating process, $(F^T_{ilt}, W^T_{ilt}, ED^T_{ilt}, ELGD^T_{ilt})$.

Column (2) of Table 7 shows the result if Prosper ignores the short-term profits (setting $\delta_1 = 0$), the expected revenue drops by about 4.5%. This is consistent with our intuition. As our estimates suggests, Prosper does not only rely on listings' estimated risk to assign ratings, but also its expected revenue. But this "profit" factor plays a relatively small role in the decision process. The biases caused by the profit term mainly affect AA and HR rating groups, as shown in Figure 7.

Column (3) of Table 7 shows that if Prosper does not adaptively learn at all, its expected revenue drops by about 3.06%. Hence, actively adapting to the changing market can indeed help a firm increase its revenue.

Columns (4) to (7) show how the expected revenue changes if Prosper uses a different adaptive learning method to learn. Under EWM and GFM, Prosper's expected revenue will drop by 3.11% and 3.7%, respectively. In contrast, if Prosper uses MWM and RPM, the counterfactual revenue only drops by 0.36% and 0.28%. The results are consistent with the overall predictive performance of different methods: recall that ERPM outperforms EWM and GFM, but EPRM, MWM and RPM are all quite close. It is possible that MWM and RPM outperform E-RPM for some listings that are more profitable. Taking results from Columns (4) to (7) together, we should recommend firms to carefully examine the adaptive learning methods before they implement them to deal with the concept drift problem. Well suited learning methods (E-RPM, MWM and RPM in our case) can indeed help the firm to better learn the market, while other adaptive learning methods (EWM and GFM in our case) may lead to even worse results compared with no adaptive learning at all.

Finally, when assuming Prosper knew the true data generating process, it does achieve the highest expected revenue (column 8 of Table 7); it gives 1.74% higher expected revenue compared with the benchmark E-RPM model. The increase is relatively small, suggesting that Prosper is doing a reasonably good job when using

ERPM to adapt.

8 Conclusion

This is the first structural econometric modeling paper that proposes a framework to infer which adaptive learning algorithm that a firm is most likely to use in handling concept drifts. We refer to this approach as *generalized revealed preference*. To illustrate our approach, we apply it to Prosper's adaptive learning behavior. We first provide evidence that consumers' borrowing and lending behavior are changing over time. We then show that by analyzing Prosper's choices and the lens of a structural model, we can uncover not only the parameters in its objective function, but also the way Prosper selects data to make decisions. Among the five adaptive learning algorithms we consider, we find that an ensemble method that relies on macroeconomic conditions of the data (E-RPM) is the most plausible method adopted by Prosper. In one counterfactual experiment where we assume Prosper does not care about short-term profits, we demonstrate that Prosper likely assigns too few AA and HR ratings, because these two categories are the least profitable loans. In other counterfactual experiments where we assume Prosper uses other adaptive learning algorithms, they demonstrate how different ways to learn can lead Prosper to make different decisions.

To conclude, we reiterate that the concept drifts could happen often, and it can be very difficult for a firm to figure out how the underlying data generating process changes over time, in particular, when it needs to make real time decisions. Hence, we do not assume that Prosper uses an optimal adaptive learning method. An optimal adaptive learning method would require Prosper to know exactly how and when concept drifts happen. Such an assumption does not seem very plausible here, especially because we are studying the period when Prosper just started to use posted price business model. Hence, we hypothesize that during our sample period, Prosper uses one adaptive learning algorithm to address concept drifts, even though it may not be a perfect solution. Our research does not address how Prosper settled with this algorithm. Also, it is certainly possible that Prosper may change its adaptive learning algorithm from time to time. Studying this problem is beyond the scope of this paper. We will leave these questions for future research.

References

1. Aguirregabiria, Victor, and Jihye Jeon. "Firms' beliefs and learning: Models, identification, and empirical evidence." *Review of Industrial Organization* 56, no. 2 (2020): 203-235.
2. Balakrishnan, Suhrid, and David Madigan. "Algorithms for sparse linear classifiers in the massive data setting." Journal of Machine Learning Research 9, no. Feb (2008): 313-337.
3. Bajari, Patrick, C. Lanier Benkard, and Jonathan Levin. "Estimating dynamic models of imperfect competition." *Econometrica* 75, no. 5 (2007): 1331-1370.
4. Benkard, C. Lanier. "A dynamic analysis of the market for wide-bodied commercial aircraft." *The Review of Economic Studies* 71, no. 3 (2004): 581-611.
5. Chan, Tat, Naser Hamdi, Xiang Hui, and Zhenling Jiang. "Digital Verification and Inclusive Access to Credit: An Empirical Investigation." Available at SSRN (2020).
6. Chauvet, Marcelle. "An econometric characterization of business cycle dynamics with factor structure and regime switching." *International Economic Review* (1998): 969-996.
7. Ching, Andrew T. "A dynamic oligopoly structural model for the prescription drug market after patent expiration." *International Economic Review* 51, no. 4 (2010): 1175-1207.
8. Cogley, Timothy, and Thomas J. Sargent. "The conquest of US inflation: learning and robustness to model uncertainty." *Review of Economic dynamics* 8.2 (2005): 528-563.
9. Crespo, Fernando, and Richard Weber. "A methodology for dynamic data mining based on fuzzy clustering." *Fuzzy Sets and Systems* 150.2 (2005): 267-284.
10. Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society: Series B (Methodological)* 39, no. 1 (1977): 1-22.
11. Dew, Ryan, Asim Ansari, and Yang Li. "Modeling dynamic heterogeneity using Gaussian processes." Journal of Marketing Research 57, no. 1 (2020): 55-77.
12. Diermeier, Daniel, Michael Keane, and Antonio Merlo. "A political economy model of congressional careers." *American Economic Review* 95, no. 1 (2005): 347-373.
13. Domingos, Pedro, and Michael Pazzani. "On the optimality of the simple Bayesian classifier under zero-one loss." *Machine learning* 29.2 (1997): 103-130.
14. Doraszelski, Ulrich, Gregory Lewis, and Ariel Pakes. "Just starting out: Learning and equilibrium in a new market." American Economic Review 108, no. 3 (2018): 565-615.
15. Erdem, Tülin, Susumu Imai, and Michael P. Keane. "Brand and quantity choice dynamics under price uncertainty." *Quantitative Marketing and Economics* 1, no. 1 (2003): 5-64.
16. Evans, George W., and Seppo Honkapohja. 2001. Learning and Expectations in Macroeconomics. Princeton, NJ: Princeton University Press.

17. Evans, George W., and Seppo Honkapohja. 2013. "Learning as a Rational Foundation for Macroeconomics and Finance." In Rethinking Expectations: The Way Forward for Macroeconomics, edited by Roman Frydman and Edmund S. Phelps, 68–111. Princeton, NJ: Princeton University Press.

18. Forman, George. "Tackling concept drift by temporal inductive transfer." In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 252-259. 2006.

19. Freedman, Seth M., and Ginger Zhe Jin. *Learning by doing with asymmetric information: Evidence from Prosper.com*. No. w16855. National Bureau of Economic Research, 2011.

20. Friedman, Nir, Dan Geiger, and Moises Goldszmidt. "Bayesian network classifiers." *Machine learning* 29, no. 2-3 (1997): 131-163.

21. Fu, Runshan, Yan Huang, and Param Vir Singh. "Crowds, Lending, Machine, and Bias." Information Systems Research (2021).

22. Gama, Joao, Pedro Medas, Gladys Castillo, and Pedro Rodrigues. "Learning with drift detection." In *Brazilian symposium on artificial intelligence*, pp. 286-295. Springer, Berlin, Heidelberg, 2004.

23. Gordon, Brett R., Avi Goldfarb, and Yang Li. "Does price elasticity vary with economic growth? A cross-category analysis." Journal of Marketing Research 50, no. 1 (2013): 4-23.

24. Hamilton, James D., and Gang Lin. "Stock market volatility and the business cycle." *Journal of Applied Econometrics* 11, no. 5 (1996): 573-593.

25. Hand, David J., and Keming Yu. "Idiot's Bayes-not so stupid after all?." *International statistical review* 69.3 (2001): 385-398.

26. Hendel, Igal, and Aviv Nevo. "Measuring the implications of sales and consumer inventory behavior." *Econometrica* 74, no. 6 (2006): 1637-1673.

27. Hoens, T. Ryan, Robi Polikar, and Nitesh V. Chawla. "Learning from streaming data with concept drift and imbalance: an overview." *Progress in Artificial Intelligence* 1, no. 1 (2012): 89-101.

28. Humpherys, Sean L., Kevin C. Moffitt, Mary B. Burns, Judee K. Burgoon, and William F. Felix. "Identification of fraudulent financial statements using linguistic credibility analysis." *Decision Support Systems* 50, no. 3 (2011): 585-594.

29. Iyer, Rajkamal, Asim Ijaz Khwaja, Erzo FP Luttmer, and Kelly Shue. "Screening peers softly: Inferring the quality of small borrowers." *Management Science* 62, no. 6 (2016): 1554-1577.

30. John, George H., and Pat Langley. "Estimating continuous distributions in Bayesian classifiers." *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1995.

31. Kawai, Kei, Ken Onishi, and Kosuke Uetake. "Signaling in online credit markets." Available at SSRN 2188693 (2020).

32. Kelly, Mark G., David J. Hand, and Niall M. Adams. "The impact of changing populations on classifier performance." *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1999.

33. Klinkenberg, Ralf. "Learning drifting concepts: Example selection vs. example weighting." *Intelligent data analysis* 8, no. 3 (2004): 281-300.

34. Kim, Chang-Jin, and Charles R. Nelson. "Business cycle turning points, a new coincident index, and tests of duration dependence based on a dynamic factor model with regime switching." *Review of Economics and Statistics* 80.2 (1998): 188-201.

35. Koren, Yehuda. "Collaborative filtering with temporal dynamics." In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 447-456. 2009.

36. Kozlowski, Julian, Laura Veldkamp, and Venky Venkateswaran. "The tail that wags the economy: Beliefs and persistent stagnation." Journal of Political Economy 128, no. 8 (2020): 2839-2879.

37. Liechty, John C., Duncan KH Fong, and Wayne S. DeSarbo. "Dynamic models incorporating individual heterogeneity: Utility evolution in conjoint analysis." Marketing Science 24, no. 2 (2005): 285-293.

38. Lin, Mingfeng, Nagpurnanand R. Prabhala, and Siva Viswanathan. "Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending." *Management Science* 59.1 (2013): 17-35.

39. Lin, Mingfeng, and Siva Viswanathan. "Home bias in online investments: An empirical study of an online crowdfunding market." *Management Science* 62.5 (2016): 1393-1414.

40. Liu, Xinyuan, Zaiyan Wei, and Mo Xiao. "Platform mispricing and lender learning in peer-to-peer lending." Review of Industrial Organization 56, no. 2 (2020): 281-314.

41. McCullagh, P.; Nelder, J. A. (1983), *Generalized Linear Models*, Chapman and Hall, §1.1.4.

42. Ni, Jian, and Yi Xin. "Financing Micro-entrepreneurship in Online Crowdfunding Markets: Local Preference versus Information Frictions." Available at SSRN 3580585 (2020).

43. Pechenizkiy, Mykola, Jorn Bakker, I. Žliobaitė, Andriy Ivannikov, and Tommi Kärkkäinen. "Online mass flow prediction in CFB boilers with explicit detection of sudden concept drift." *ACM SIGKDD Explorations Newsletter* 11, no. 2 (2010): 109-116.

44. Phua, Clifton, Vincent Lee, Kate Smith, and Ross Gayler. "A comprehensive survey of data mining-based fraud detection research." *arXiv preprint arXiv*:1009.6119 (2010).

45. Schlimmer, Jeffrey C., and Richard H. Granger. "Incremental learning from noisy data." *Machine learning* 1.3 (1986): 317-354.

46. Sornette, Didier. *Why stock markets crash: critical events in complex financial systems*. Vol. 49. Princeton University Press, 2017.

47. Sriram, Srinivasaraghavan, Pradeep K. Chintagunta, and Ramya Neelamegham. "Effects of brand preference, product attributes, and marketing mix variables in technology product markets." Marketing Science 25, no. 5 (2006): 440-456.

48. Tsymbal, Alexey. "The problem of concept drift: definitions and related work." Working paper, Computer Science Department, Trinity College Dublin. Available at: https://www-ai.cs.tu-

dortmund.de/LEHRE/FACHPROJEKT/SS12/paper/concept-drift/tsymbal2004.pdf [1013 citations according to Google Scholar, accessed on Dec 19, 2020]

49. Viaene, Stijn, Richard A. Derrig, Bart Baesens, and Guido Dedene. "A Comparison of State-of-the-Art Classification Techniques for Expert Automobile Insurance Claim Fraud Detection." *Journal of Risk and Insurance* 69, no. 3 (2002): 373-421.

50. Wang, Haixun, Wei Fan, Philip S. Yu, and Jiawei Han. "Mining concept-drifting data streams using ensemble classifiers." In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 226-235. AcM, 2003.

51. Wei, Zaiyan, and Mingfeng Lin. "Market mechanisms in online peer-to-peer lending." *Management Science* 63.12 (2017): 4236-4257.

52. Yao, Song, Carl F. Mela, Jeongwen Chiang, and Yuxin Chen. "Determining consumers' discount rates with field studies." *Journal of Marketing Research* 49, no. 6 (2012): 822-841.

53. Zhang, Juanjuan, and Peng Liu. "Rational herding in microloan markets." *Management science* 58.5 (2012): 892-912.

Table 1: Interest Rates and Service Fee Rates

	AA	A	B	C	D	E	HR
Average Interest Rate	7.74%	10.80%	15.88%	19.92%	24.85%	30.43%	31.78%
Average Annual Return	7.03%	8.58%	10.11%	10.99%	12.15%	13.13%	12.17%
Origination Fee Rate	0.5%	3%	3%	4.5%	4.5%	4.5%	4.5%
Posted Loss Rate	1.42%	3.03%	5.56%	7.94%	10.83%	14.41%	17.08%

Notes: The interest rate of each rating is set by Prosper and changes very occasionally over time. The interest rates are calculated by taking the average across all listings in the data. Annual return is calculated by subtracting real loss rate from interest rate. The real loss rate is calculated from the data since we can observe all the loans repayment outcomes. If a loan defaults, we can observe the principal loss as well. The posted loss rate in this Table is reported by Prosper. For loan applications with the same rating, Prosper posts the same loss rate. The posted loss rate is not necessary to coincide with the real loss rate. This is why the the average annual return and the posted loss rate do not add up to the average interest rate.

Table 2: Listing Status by Prosper Ratings

Prosper Rating	Completed	Expired	Withdrawn	Total
AA	1,030 (69.59%)	165 (11.15%)	285 (19.26%)	1,480
A	2,714 (72.55%)	408 (10.91%)	619 (16.55%)	3,741
B	2,965 (72.99%)	351 (8.64%)	746 (18.37%)	4,062
C	2,729(75.49%)	310 (8.58%)	576 (15.93%)	3,615
D	4,675 (71.23%)	573(8.73%)	1,315 (20.04%)	6,563
E	3,402 (76.43%)	263(5.91%)	786 (17.66%)	4,451
HR	4,762 (60.32%)	1,635 (20.71%)	1,498 (18.97%)	7,895
Total	22,277(70.04%)	3,705 (11.65%)	5,825 (18.31%)	31,807

Notes: Percentages are calculated rowwise. For instance, 1,030 completed AA listings account for 69.59% of the 1,480 total AA listings.

Table 3: List of Variables

Variable Name	Description
	Listing Features
Listing Number	The unique identifier for a listing.
Amount Requested*	Amount requested by borrower in a listing.
Amount Funded	Amount funded by lenders.
Prosper Rating	A series of dummy variables assigned by Prosper to indicate borrower's credit grade (AA, A, B, C, D, E, HR). AA indicates the best credit and HR indicates the worst credit.
Origination Fee Rate*	Origination fees are a percentage of the amount borrowed varying by Prosper Rating. It ranges from 0.5% to 4.5% according to the listing's rating.
Interest Rate*	The interest rate a borrower needs to pay for this loan.
Listing Status	Withdrawn - The listing was withdrawn by customer request. Expired - The listing failed to fund in time. Completed - The listing ran to completion and funded. Cancelled - The listing was canceled by Prosper.
	Borrower Characteristics
Bankcard Utilization*	The percentage of available revolving credit that is utilized at the time the credit profile was pulled.
Home Owner*	A Borrower will be classified as a homowner if they have a mortgage on their credit profile or provide documentation confirming they are a homeowner.
Posted Loss Rate*	The posted principal loss percentage on default. Prosper reports the same Posted Loss Rate for listings with the same rating.
Credit Score Range Lower*	The lower value of the range of the borrower's credit score provided by the consumer credit rating agency
Current Credit Lines*	Number of current credit lines at the time the credit profile was pulled.
Credit Lines Last 7 Years*	Number of credit lines in the past seven years at the time the credit profile was pulled.
Current Delinquencies*	Number of accounts delinquent at the time the credit profile was pulled.
Delinquencies Last 7 Years*	Number of delinquencies in the past 7 years at the time the credit profile was pulled.
Monthly Income*	The monthly income the borrower stated at the time the listing was created.
Income Verifiable*	Prosper will request documents such as recent paystubs, tax returns, or bank statements to verify a borrower's income.
Inquiries Last 6 Months*	Number of inquiries in the past five months at the time the credit profile was pulled.
Total Inquiries*	Total number of inquiries at the time the credit profile was pulled.
Open revolving accounts*	Number of open revolving accounts at the time the listing was created.
Revolving Credit Balance*	The monetary amount of revolving credit balance at the time the listing was created.
Prior Prosper Loans*	Number of Prosper loans the borrower has borrowed at the time they created this listing.

Table 3: List of Variables (Continued)

Prior Prosper Loans Active*	Number of ongoing Prosper loans the borrower has at the time they created this listing.
Monthly Debt*	The amount of debt the borrower needs to pay each month.
Month*	The month in which the listing was submitted.
Group*	Equals 1 if the borrower belongs to a group. A Group is a collection of Members who share a common interest or affiliation.
Real Estate Balance*	The mortgage balance.
Market Level Variables	
Mortgage Rate*	30-Year Fixed Rate Mortgage Average in the United States.
TED Spread*	The difference between the interest rate on short-term US government debt and the interest rate on interbank loans.
Adjusted Closing Price*	S&P 500 daily closing price.

Notes: Variables with * enter our lender side, borrower side and risk access model to help us calculate each listing's funding, withdrawal and default probabilities. We also use these variables in the logistic regressions in section 3.3 to show how to detect concept drift.

Table 4: Variables: Summary Statistics

Variable	Mean	SD	Max	Min
Panel A: Listing Features				
Amount Requested ($1000)	6.83	0.095	25	2
Amount Funded ($1000)	4.98	4.58	25	0
Funded Percentage (%)	75.48	39.04	100	0
Interest Rate (%)	23.34	8.14	32.20	5.99
Origination Fee (%)	4.36	0.92	4.95	0.5
Panel B: Borrower Credit Variables				
Bank Card Utilization (%)	51.50	33.12	223	0
Home Owner (0/1)	0.49	0.50	1	0
Posted Loss Rate (%)	10.53	5.29	20.3	0.49
Credit Score Range Lower	697.47	44.55	778	600
Current Credit Lines	9.11	5.39	56	0
Credit Lines Last 7 Years	25.76	13.93	120	2
Current Delinquencies	0.48	1.32	27	0
Delinquencies Last 7 Years	0.48	1.32	27	0
Monthly Income ($1000)	5.67	13.84	1,750	0
Income Verifiable (0/1)	0.86	0.35	1	0
Inquiries Last 6 Months	1.28	1.74	27	0
Total Inquiries	4.33	4.07	73	0
Prior Prosper Loans	0.33	0.69	7	0
Prior Prosper Loans Active	0.15	0.35	2	0
Monthly Debt ($1000)	0.87	1.35	100.28	0
Real Estate Balance ($1000)	107.60	162.81	3,830.08	0
Open revolving accounts	6.19	4.31	47	0
Revolving Credit Balance ($1000)	19.04	37.15	1,066.76	0
Group (0/1)	0.033	0.18	1	0
Panel C: Market Level Variables				
Mortgage Rate (%)	4.38	0.51	5.33	3.58
TED Spread (%)	0.33	0.11	0.57	0.14
Adjusted Closing Price	1,322	78.30	1,466	1,099

Notes: Summary statistics in this table are calculated based on listings submitted between Dec 20, 2010 and Dec 31, 2012.

Table 5: Estimation Results of Prosper's Objective Function

	EWM	MWM	RPM	GFM	E-RPM
Log-Likelihood	-51999.34	-53343.86	-51121.85	-52219.56	-49314.87
BIC	104081.62	106770.66	102326.64	104522.06	98712.68
$\hat{\delta}_1$	2.21 (0.378)	1.58 (0.132)	4.41 (0.169)	2.09 (0.401)	4.26 (0.185)
$\hat{\delta}_2$	-11.0 (0.161)	-6.41 (0.120)	-10.44 (0.145)	-10.94 (0.165)	-13.98 (0.154)
$\hat{\alpha}_1$	-2.33 (0.035)	-1.44 (0.033)	-1.70 (0.037)	-2.38 (0.035)	-1.41 (0.037)
$\hat{\alpha}_2$	-1.41 (0.024)	-0.69 (0.021)	-1.27 (0.023)	-1.44 (0.024)	-0.99 (0.023)
$\hat{\alpha}_3$	-1.30 (0.023)	-0.69 (0.021)	-1.18 (0.022)	-1.30 (0.023)	-0.96 (0.023)
$\hat{\alpha}_4$	-1.30 (0.024)	-0.84 (0.022)	-1.27 (0.023)	-1.30 (0.024)	-1.06 (0.023)
$\hat{\alpha}_5$	-0.61 (0.020)	-0.32 (0.018)	-0.57 (0.019)	-0.59 (0.020)	-0.42 (0.019)
$\hat{\alpha}_6$	-0.74 (0.020)	-0.67 (0.020)	-0.75 (0.020)	-0.74 (0.020)	-0.68 (0.020)

Notes: For GFM, the forgetting factor takes value 0.997, 0.99, 0.995 and 0.997 for the borrower side, lender side, default prediction and LGD prediction models, respectively. In MWM, the window size equals to 3, 7, 36 and 48 months for the borrower side, lender side, default prediction and LGD prediction models, respectively. The model comparison is done by evaluating their performance on one-period ahead out of sample prediction. For instance, when predicting loan applications' rating assignment in period $t+1$, we use data until period t to calibrate our model and then make predictions. After we have our model's prediction accuracy for period $t+1$, we move one period forward to period $t+2$ and re-estimate our model by taking data from period $t+1$ into account and make predictions for period $t+2$. We repeat this process until the last period of our data. The length of period for different adaptive learning models can be found in sections 5.1-5.5. BICs are calculated based on one-period ahead out of sample log-likelihood. Standard errors are in brackets. $^{***}p<0.001$.

Table 6: Comparison of Prediction Results

	EWM	MWM	RPM	GFM	E-RPM
Funding AUC	0.712	0.847	0.796	0.638	0.854
Withdrawal AUC	0.534	0.618	0.587	0.535	0.595
Default AUC	0.597	0.609	0.603	0.615	0.617
Loss Given Default MSE	0.045	0.092	0.110	0.095	0.035

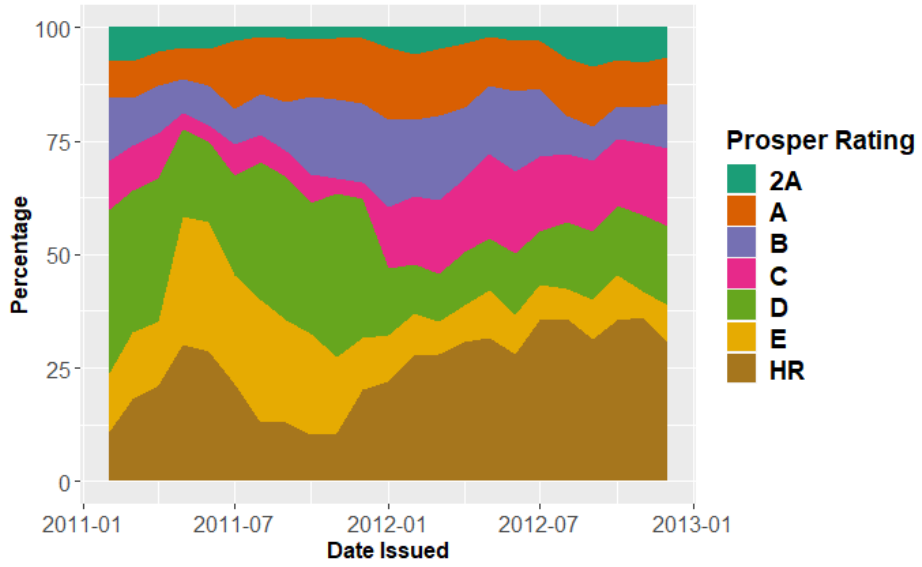
Notes: AUC represents area under ROC. The larger the AUC is, the better prediction performance the corresponding method has.

Table 7: Counterfactual Revenue (E-RPM as Baseline)

	E-RPM	$\delta_1 = 0$	No AL	EWM	GFM	MWM	RPM	True DGP
Revenue ($)	7,380,725	7,066,938	7,154,574	7,150,842	7,107,806	7,3542,71	7,360,216	7,509,171
ΔRevenue ($)	0	-313,787	-226,151	-229,883	-272,919	-26,454	-20,509	128,446
ΔRevenue (%)	0	-4.25%	-3.06%	-3.11%	-3.70%	-0.36%	-0.28%	1.74%

Notes: This table shows the counterfactual expected revenue in different scenarios. We use E-RPM as the baseline.

Figure 1: Rating Distribution of Loans Over Time



Notes: Each color represents the percentage of a certain rating category over time.

Figure 2: Conversion Rate and Default Rate

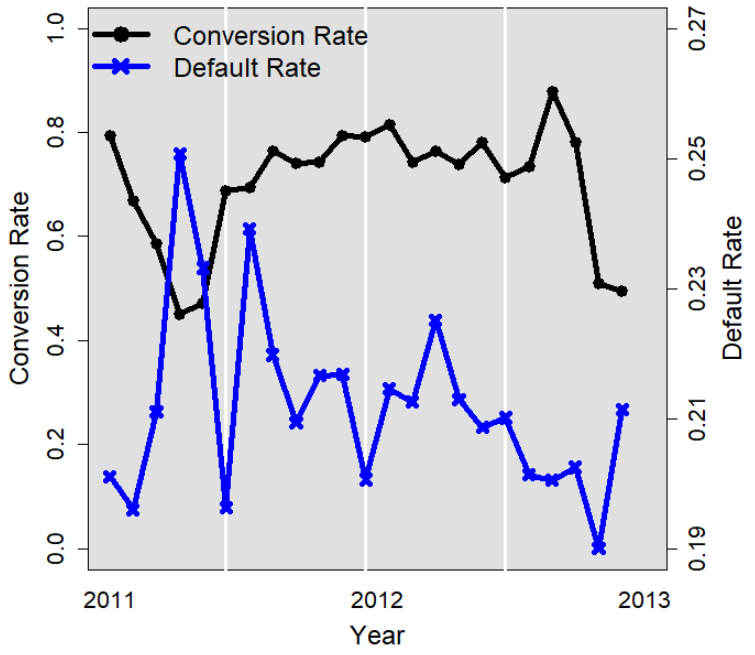
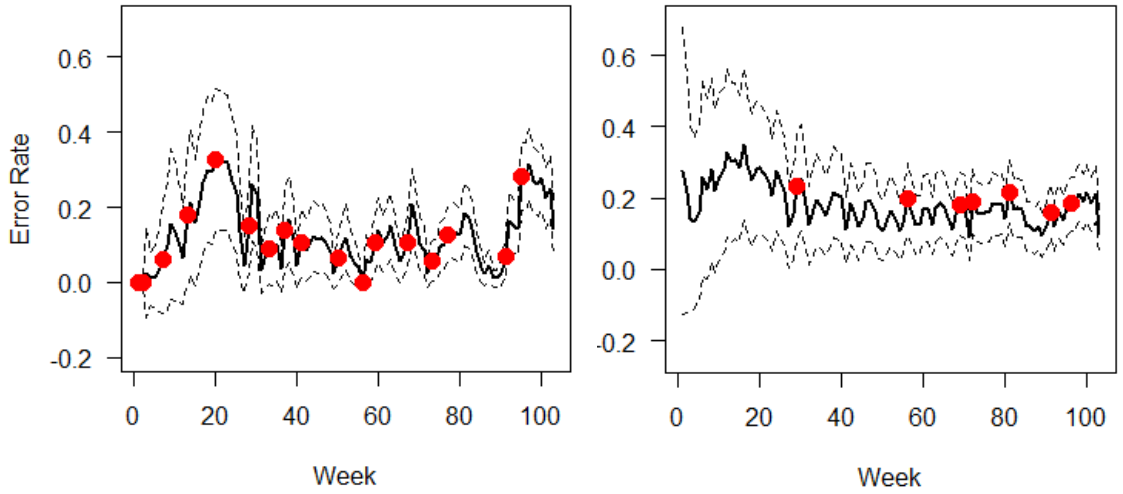
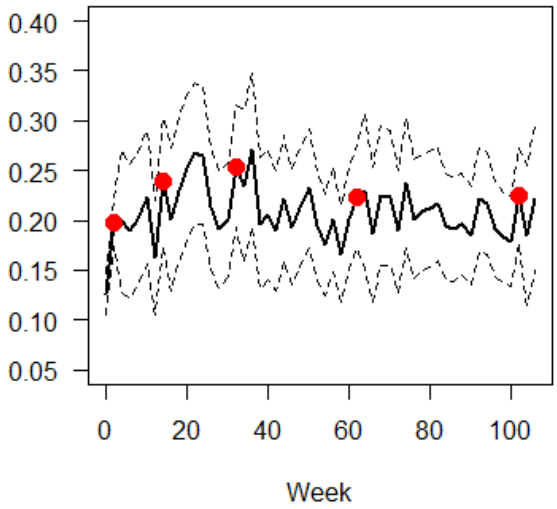


Figure 3: Concept Drift Detection



(a) Prediction Error Rate in Loan Completion

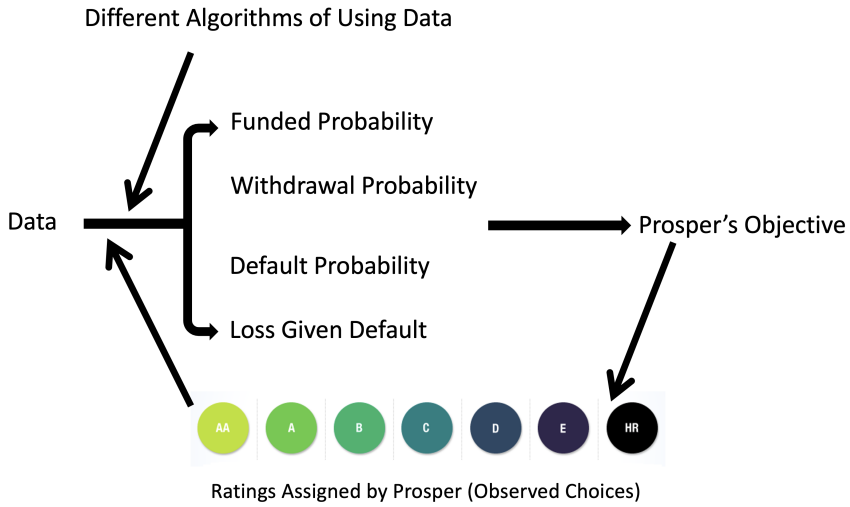
(b) Prediction Error Rate in Withdrawal Decisions



(c) Prediction Error Rate in Default Outcomes

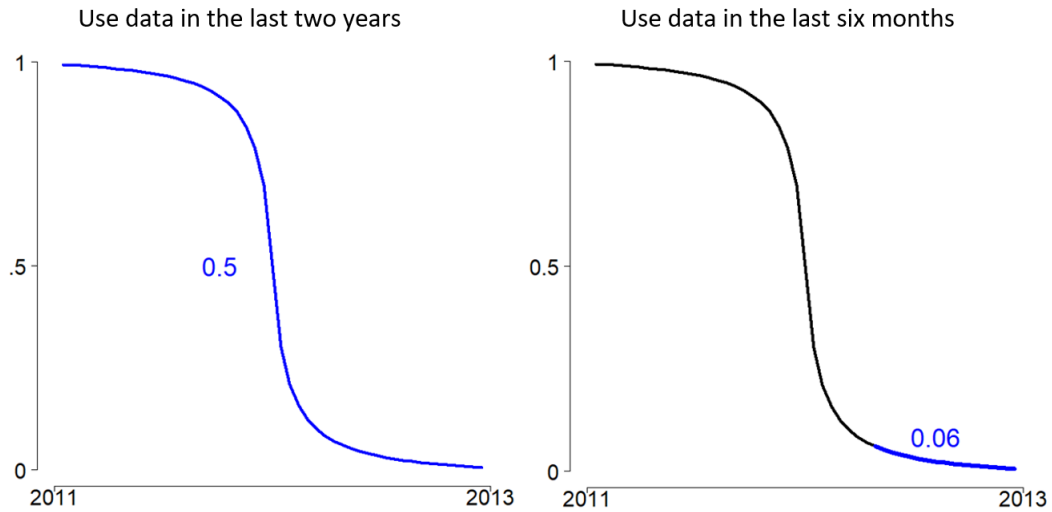
Notes: The red dots indicate when concept drifts are detected by Gama et al. (2004) Test. Concept drifts in lenders' investment behavior, borrowers' withdrawal and default behavior were detected 17, 7 and 5 times, respectively. Dotted lines represent the 95% confidence interval.

Figure 4: Model Framework



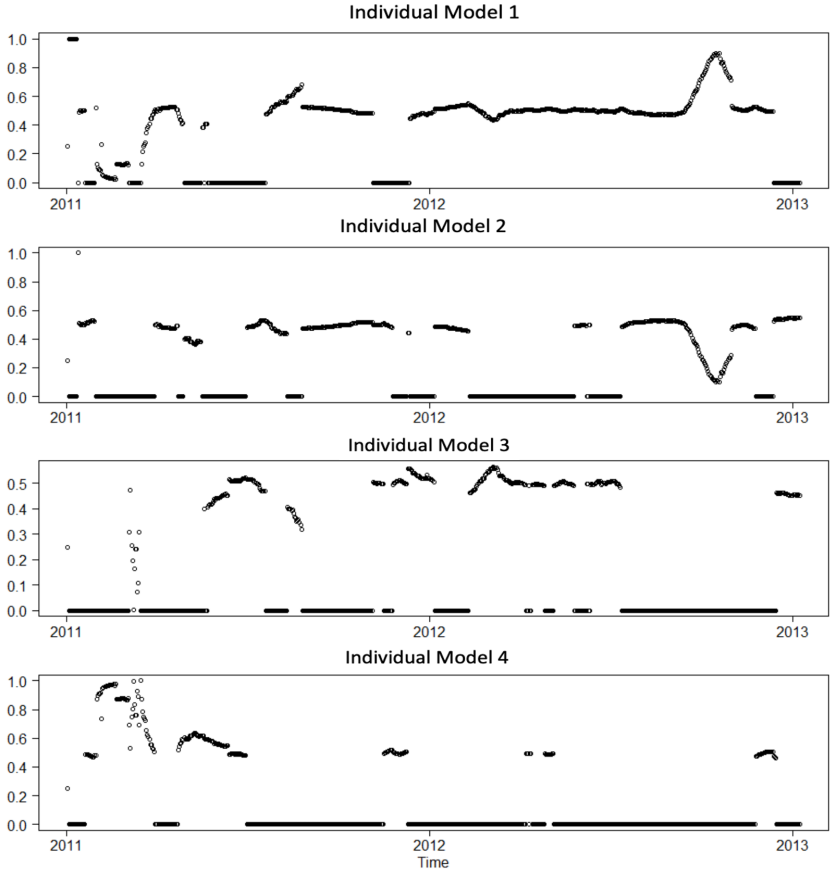
Notes: We use lender side model and borrower side mdoel to compute each listing's funding probability and withdrawal probability, respectively. Default probability and LGD are computed using Naive Bayes classifier. Prosper then puts these four variables into its objective function and chooses one of the seven ratings to maximize its objective. What we are interested in is to infer the way Prosper uses historical data to compute funding, withdrawal, default probabilities and LGD. Because these calibrated components are a function of the way Prosper weights past data, Prosper's choice in classifying loans should reveal their data weighting mechanism.

Figure 5: Average Default Rate of Borrowers from State *A*



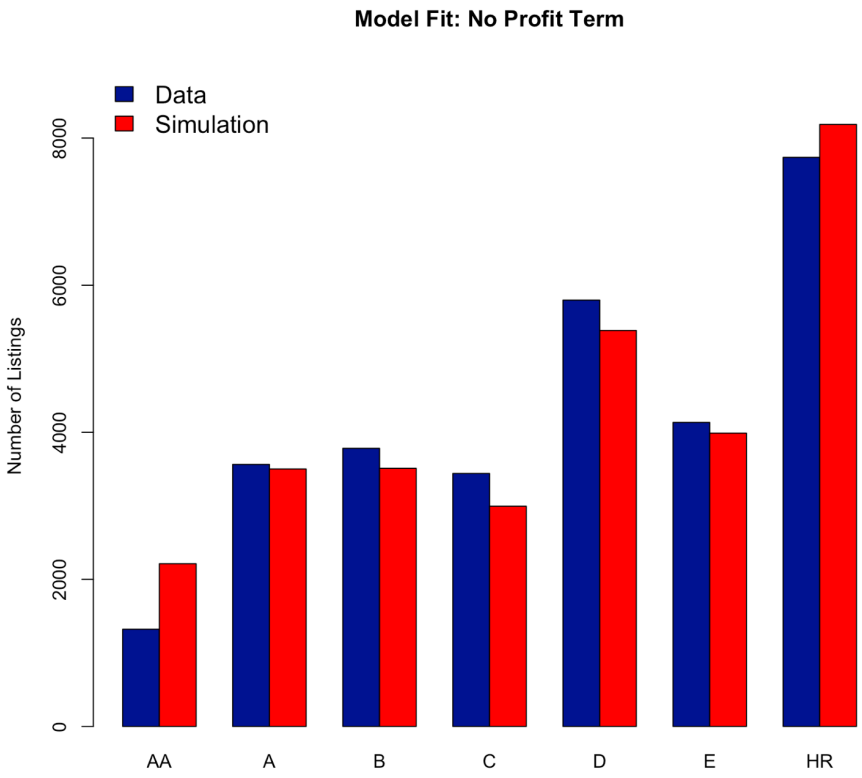
Notes: This is a purely hypothetical example. In this example, the average default rate for borrowers from State *A* is 50% from Jan 2011 to Dec 2012, while only 6% of borrowers from State *A* default from Jul 2012 to Dec 2012.

Figure 6: Weights on Individual Funding Model in E-RPM



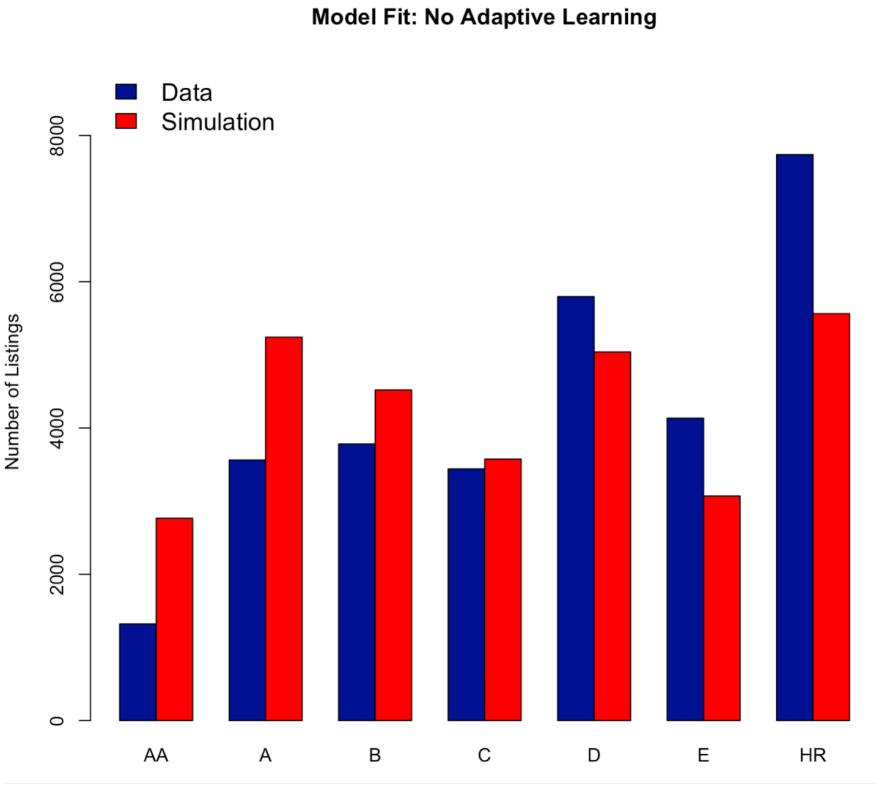
Notes: This figure shows the weights E-RPM puts on each individual model on funding predictions. The weights are updated every day.

Figure 7: Rating Distribution: No profit Term



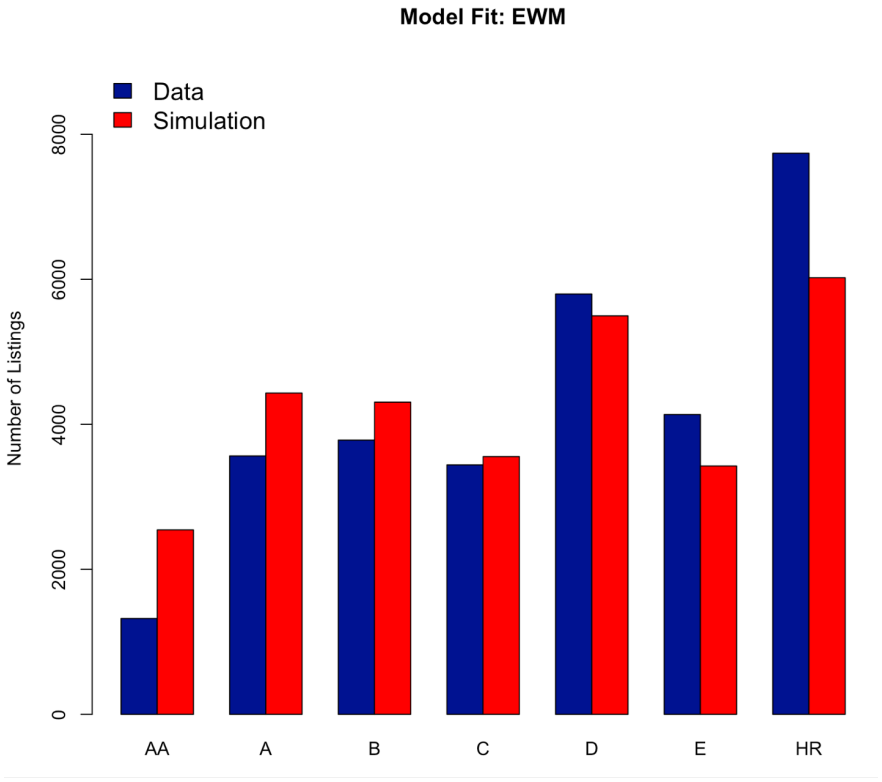
Notes: Red bars represent the expected rating distribution if Prosper does not care about profit when making rating assignments.

Figure 8: Rating Distribution: No Adaptive Learning



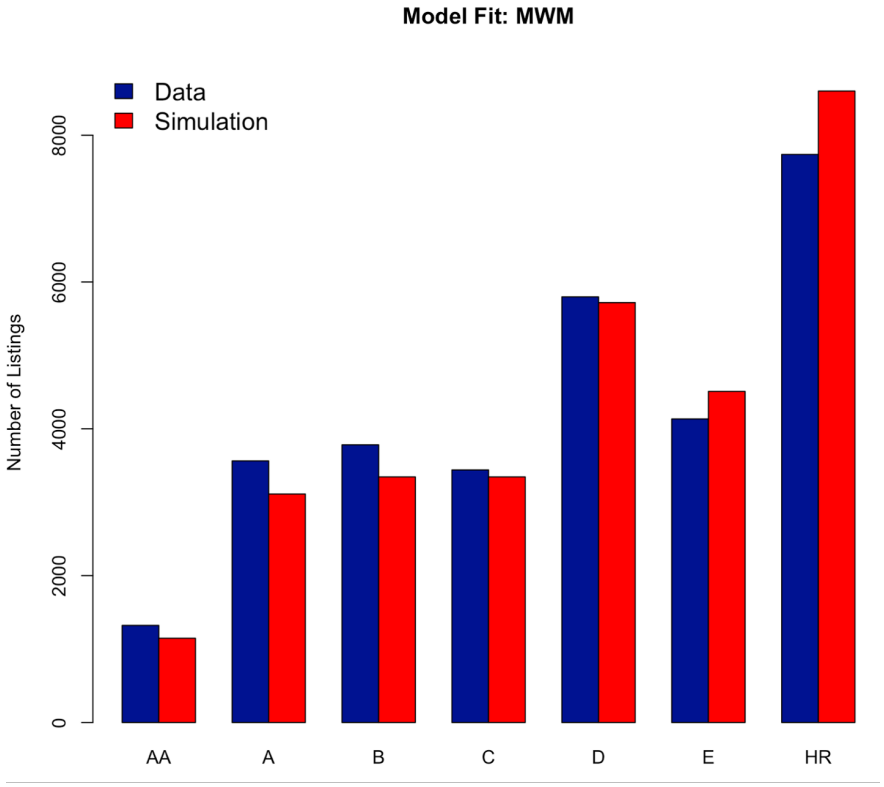
Notes: Red bars represent the expected rating distribution if Prosper did not adaptively update its model but kept using the model calibrated using data from the auction period.

Figure 9: Rating Distribution: EWM



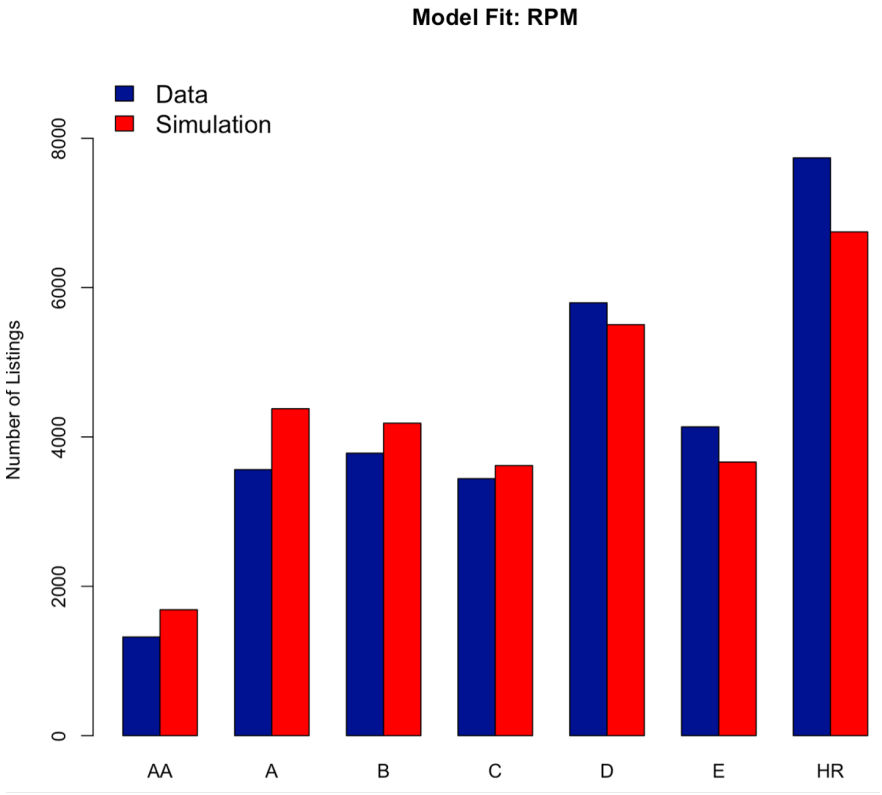
Notes: Red bars represent the expected rating distribution if Prosper put equal weights on all the historical observations.

Figure 10: Rating Distribution: MWM



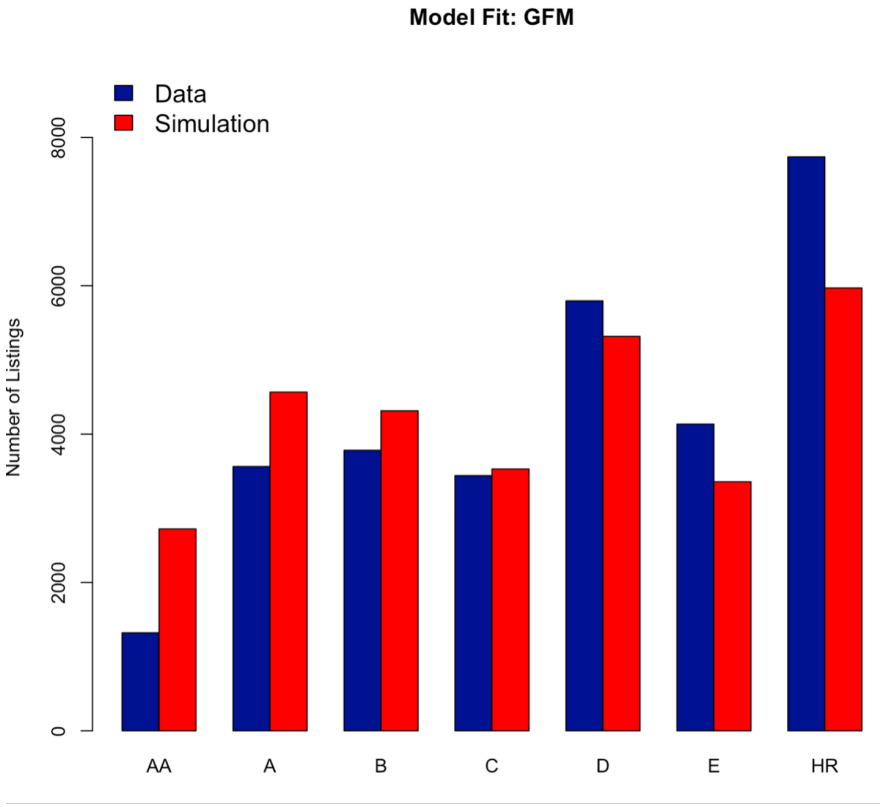
Notes: Red bars represent the expected rating distribution if Prosper used moving window method (MWM) to make rating assignment.

Figure 11: Rating Distribution: RPM



Notes: Red bars represent the expected rating distribution if Prosper used recession probability method (RPM) to make rating assignment.

Figure 12: Rating Distribution: GFM



Notes: Red bars represent the expected rating distribution if Prosper used gradual forgetting method (GFM) to make rating assignment.

Appendix

A Concept Drift Detection

We follow Gama, et al. (2004) to test for the existence of concept drift. As an example, let us consider how to test the existence of concept drift in lender's investing behavior. Consider a set of loan applications observations, in the form of pairs (x_i, y_i), where x_i represents borrower i's characteristics, and y_i takes value 1 if the loan application is funded and 0 otherwise. Suppose that we use a logistic regression model to predict each loan application's probability of getting funded. Let $\hat{y}_i$ denote the predicted funding outcome, with the value 1 if the loan application is funded and 0 otherwise. The event of false prediction, $\hat{y}_i \neq y_i$, is a random variable from Bernoulli trials. For a sequence of n_t observations, the number of false prediction events follows a Binomial distribution. Let p_t denote the percentage of false predictions (error rate) in period t, the corresponding standard deviation is $s_t = \sqrt{p_t(1-p_t)/n_t}$.

If lender's investment behavior is stationary, the error rate (p_t) of our model should remain stable across different periods. A significant change in the error rate suggests a change in lender's investment behavior. The drift detection method manages two threshold values, p_{min} and s_{min}, during the training of the learning algorithm. When $t = 1$, we set $p_{min} = p_1$ and $s_{min} = s_1$. As more loan applications arrive, these values are updated when $p_t + s_t$ is lower than $p_{min} + s_{min}$. In particular, if in period t we have $p_t + s_t < p_{min} + s_{min}$, then we update p_{min} and s_{min} as $p_{min} = p_t$ and $s_{min} = s_t$. For a sufficiently large number of observations (>30), the Binomial distribution in each period is closely approximated by a Normal distribution with the same mean and variance. Under the hypothesis that concept drift does not happen, the confidence interval for p_t can be approximated by $p_t \pm \iota \cdot \delta_t$. The parameter ι depends on the desired confidence interval. A commonly used confidence level for warning is 95% with the threshold $p_t + s_t \geq p_{min} + 2 \times s_{min}$, and for concept drift detection is 99% with the threshold $p_t + s_t \geq p_{min} + 3 \times s_{min}$. To be specific, for period t, our concept drift detection system will be in one of the following three states:

(1) $p_t + s_t < p_{min} + 2 \times s_{min}$. The system is stable. That is, we cannot reject the hypothesis that lender's investing behavior does not change.

(2) $p_t + s_t \geq p_{min} + 3 \times s_{min}$. The error rate has increased significantly. With a very high probability that concept drift is detected.

(3) $p_{min} + 2 \times s_{min} \leq p_t + s_t < p_{min} + 3 \times s_{min}$. This is the warning state, which is in between of the two previous states. Intuitively, there is some evidence that concept drift might happen, but the evidence is not particularly strong yet.

Suppose the warning level is reached in period t_w and concept drift is detected in period t_d. Then we believe the underlying data generation process for our observations has changed from t_d onwards. Gama, et al.

(2004) then propose using observations between t_w and t_d to train our model and make predictions for future observations, until another concept drift is detected.

B Update Naive Bayes Classifier in Each Period

In the actual implementation, we keep updating the prior ω_k and posterior P_k over time. We use $\boldsymbol{\Omega_t} = \{\omega_{1t}, \omega_{2t}, ..., \omega_{Kt}\}$ represent prior belief at time t and $\boldsymbol{P}_{kt} = (P_{k1t}, P_{k2t}, ..., P_{kmt})$ represent the joint distribution of all m characteristics in class k at time period t.

Let $\boldsymbol{\Omega_t} = \{\omega_{1t}, \omega_{2t}, ..., \omega_{Kt}\}$ represent prior belief at time t and $\boldsymbol{P}_{kt} = (P_{k1t}, P_{k2t}, ..., P_{kmt})$ represent the joint distribution of all m characteristics in class k at time period t. In particular, $P_{kjt} = (p_{kj1t}, ..., p_{kjn_jt})$ represent characteristic j's distribution in class k at the beginning of period t. The prior belief of class k in period t can be calculated as the empirical frequency of observing class k up to time t:

$$\omega_{kt} = \frac{1}{\sum_{z=1}^{t} I_z} \sum_{z=1}^{t} \sum_{i=1}^{I_z} \mathbf{1}[y_i = k], k = 1, 2, ..., K \tag{12}$$

where I_z represents the number of listings in period z. The probability of observing characteristic j's h level in class k in period t, p_{kjht}, can be approximated by the empirical frequency of observing X_{ijh} up to time t,

$$p_{kjht} = \frac{\sum_{z=1}^{t} \sum_{i=1}^{I_z} \mathbf{1}[y_i = k] \cdot X_{ijh}}{\sum_{z=1}^{t} \sum_{i=1}^{I_z} \mathbf{1}[y_i = k]}, \tag{13}$$

where $k = 1, 2, ..., K; j = 1, 2, ..., m; h = 1, 2, ..., n_j$.

C Estimation

C.1 Logistics Regression with A Forgetting Factor

This section demonstrates the estimation procedure of the funded and withdrawal probabilities using GFM. Given the assumption that the residual terms follow type I extreme value distribution in equation 1, we basically need to estimate two logistic regression models in each period on the borrower and lender sides. The estimation process could be time consuming because we need to update the weight on each observation and re-estimate the model in each period. To make the estimation more efficient, we employ an estimation scheme proposed by Balakrishnan and Madigan (2008). This method is based on a quadratic Taylor approximation to the log-likelihood. A forgetting factor can be easily incorporated into this estimation scheme and the model can be recursively estimated, which significantly reduces the computational burden.

With a slight abuse of notation, we present the details of adaptive logistic regression below. Suppose the data set we have is $\{X, Y\} = \{X_i, y_i\}_{i=1}^{n}$. y_i is the class label which takes value 0 or 1. Under the logistic

regression model, the log-likelihood function is:

$$\log L(\beta|X,Y) = \sum_{i=1}^{n} f_i(\beta^T X_i) \tag{14}$$

where

$$f_i(\beta^T X_i) = \begin{cases} \log \Phi(\beta^T X_i), & \text{if } y_i = 1 \\ \log(1 - \Phi(\beta^T X_i)) & \text{o.w.} \end{cases} \tag{15}$$

Assume that our current estimate of β is $\widetilde{\beta}$, Balakrishnan and Madigan (2008) and Anagnostopoulos, et al. (2009) point out that each f_i may be replaced by the first few terms of its Taylor expansion around the current location $z_i = \beta^T X_i$. Adding up all f_i's, we get

$$\log L(\beta|X,C) \approx \frac{1}{2}\beta^T \Psi(\widetilde{\beta})\beta - \beta^T \theta(\widetilde{\beta}) \tag{16}$$

where

$$\Psi(\widetilde{\beta}) = \sum_{i=1}^{n} a(\widetilde{\beta}^T X_i) X_i X_i^T, \quad \theta(\widetilde{\beta}) = \sum_{i=1}^{n} b(\widetilde{\beta}^T X_i, y_i) X_i \tag{17}$$

and

$$a(\widetilde{\beta}^T X_i) = -\Phi(\widetilde{\beta}^T X_i)(1 - \Phi(\widetilde{\beta}^T X_i))), \quad b(\widetilde{\beta}^T, y_i) = \Phi(\widetilde{\beta}^T X_i) - y_i + \widetilde{\beta}^T X_i a(\widetilde{\beta}^T X_i) \tag{18}$$

We can get a better estimation of β by maximizing the log-likelihood function with respect to β:

$$\widetilde{\beta}^* = \underset{\beta}{\operatorname{argmax}} \left(\frac{1}{2}\beta^T \Psi(\widetilde{\beta})\beta - \beta^T \theta(\widetilde{\beta}) \right) \tag{19}$$

Anagnostopoulos, et al. (2009) modify this estimation scheme to make it recursively update its parameter estimates upon the arrival of new data points. Assume our current estimate of β is $\widehat{\beta}_t$ and the new data point arrives is (X_{n+1}, y_{n+1}), our new estimate of β is updated as follows:

$$\widehat{\Psi}_{n+1} = \widehat{\Psi}_n + a(\widehat{\beta}_n^T X_{n+1}) X_{n+1} X_{n+1}^T \tag{20}$$

$$\widehat{\theta}_{n+1} = \widehat{\theta}_n + b(\widehat{\beta}_n^T X_{n+1}, y_{n+1}) X_{n+1} \tag{21}$$

$$\widehat{\beta}_{n+1} = \widehat{\Psi}_{n+1}^{-1} \widehat{\theta}_{n+1} \tag{22}$$

Now, it is straightforward to introduce a forgetting factor into the recursions. The idea is to put less weight on more distant observations. For a forgetting factor $0 < \lambda \leq 1$, equations (11) and (12) can be revised as:

$$\widehat{\Psi}_{n+1} = \lambda \widehat{\Psi}_n + a(\widehat{\beta}_n^T X_{n+1}) X_{n+1} X_{n+1}^T \tag{23}$$

$$\widehat{\theta}_{n+1} = \lambda \widehat{\theta}_n + b(\widehat{\beta}_n^T X_{n+1}, y_{n+1}) X_{n+1} \tag{24}$$

The weights this model puts on historical observations are discounted exponentially at rate λ. When λ equals 1, it is equivalent to a binary logistic regression model.

We estimate the adaptive Naive Bayes model following the procedure described in section 6.1.2. We update the labels of all the loan in each period because in each period some borrowers may fail to repay their monthly installments and make their non defaulted loan become defaulted.

Given the parameter estimates on the borrower, lender sides and risk assessment model, we can compute the expected funding probability, withdrawal probability, default probability and LGD for each listing. Then we maximize the likelihood function (7) by choose the optimal δ.

C.2 Naive Bayes Classifier with A Forgetting Factor

On the risk assessment side, we incorporate a forgetting factor to introduce temporal adaptivity in the naive Bayes model in the following way: let $\lambda \in [0,1]$ be the user-defined forgetting factor. Let $\widetilde{\boldsymbol{\Omega}}_{\boldsymbol{t}} = \{\widetilde{\omega}_{1t}, \widetilde{\omega}_{2t}, ..., \widetilde{\omega}_{Kt}\}$ and $\widetilde{\boldsymbol{P}}_{kt} = (\widetilde{P}_{k1t}, \widetilde{P}_{k2t}, ..., \widetilde{P}_{kmt})$ denote the corresponding prior probability and feature distribution at time t, where $\widetilde{P}_{kjt} = (\widetilde{p}_{kj1t}, ..., \widetilde{p}_{kjn_jt})$ represent feature j's distribution in class k at the beginning of period t. Let t_i denote the period that listing i comes in, $v_i = \lambda^{t-t_i}$, $V_t = \sum_{z=1}^{t}\sum_{i=1}^{I_z} v_i$, and I_z denotes all the listings in period z. Similar to equations (12) and (13), we have the following expressions:

$$\widetilde{\omega}_{kt} = \frac{1}{V_t}\sum_{z=1}^{t}\sum_{i=1}^{I_z} v_i \cdot \mathbf{1}[y_i = k], \tag{25}$$

$$\widetilde{p}_{kjht} = \frac{\sum_{z=1}^{t}\sum_{i=1}^{I_z} v_i \cdot X_{ijh} \cdot \mathbf{1}[y_i = k]}{\sum_{z=1}^{t}\sum_{i=1}^{I_z} v_i \cdot \mathbf{1}[y_i = k]}. \tag{26}$$

where $k = 1, 2, ..., K; j = 1, 2, ..., m; h = 1, 2, ..., n_j$

Similar to equation (6), we have

$$p(X_i|y_i = k) \propto \prod_{j=1}^{m}\prod_{h=1}^{n_j} \widetilde{p}_{kjht}^{X_{ijh}}, \tag{27}$$

According to Bayes' rule, we have:

$$\begin{aligned} P_i^k = p(y_i = k|X_i) &= \frac{p(y_i = k, X_i)}{p(X_i)} \\ &\propto p(X_i|y_i = k) \cdot p(y_i = k) \\ &= p(X_i|y_i = k) \cdot \widetilde{\omega}_{kt_i} \\ &\propto \prod_{j=1}^{m}\prod_{h=1}^{n_j} \widetilde{p}_{kjht}^{X_{ijh}} \cdot \widetilde{\omega}_{kt_i}, \end{aligned} \tag{28}$$

D Receiver Operating Characteristic

In a classification problem, we need to build a mapping between instances and certain categories. If the output of the classifier happens to be continuous, the classification boundary between classes must be determined by a threshold value. For example, in a binary classification problem, the categorical outcome can be either positive (P) or negative (N). If the predicted outcome is P and the actual value is also P, then it is called a true positive (TP); however, if the actual value is n then it is said to be a false positive (FP). Conversely, a true negative (TN) has occurred when both the prediction outcome and the actual value are N, and false negative (FN) is when the prediction outcome is n while the actual value is p.The true positive rate (TPR) is defined as the number of true positive divided by the number of positive and the false positive rate (FPR) is defined as the number of false positive divided by the number of negative.

To draw an ROC curve, only the TPR and FPR are needed. The TPR defines how many correct positive results occur among all positive samples available during the test. FPR, on the other hand, defines how many incorrect positive results occur among all negative samples available during the test.

An ROC space is defined by FPR and TPR as X and Y axes, respectively, which depicts relative trade-offs between true positive (benefits) and false positive (costs). Intuitively, the more the ROC curve to the upper left, the more prediction power the corresponding classifier has. The best possible prediction method would yield a point at coordinate (0,1) of the ROC space, representing no false negatives and no false positives. The (0,1) point is also called a perfect classification. A random guess would give a point along a diagonal line from the left bottom to the top right corners.

E Prediction Performance of Different Methods

This section presents the prediction performance comparison between different methods. All the results are summarized in Table 6.

E.1 Equal Weight Model

Figure A1 compares the prediction performances of EWM and E-RPM. E-RPM outperforms EWM in all four prediction tasks. The corresponding AUCs for funding, withdrawal and default predictions of EWM are 0.712, 0.534 and 0.597, respectively, while the corresponding AUCs of E-RPM are 0.854 and 0.617, respectively. Moreover, EWM gives a MSE of LGD prediction at 0.045, while the MSE of the E-RPM model is 0.035.

E.2 Moving Window Model

Figure A2 compares the prediction performances of MWM and E-RPM. E-RPM outperforms MWM in funding, default and LGD predictions. The corresponding AUCs for funding and withdrawal predictions of

MWM are 0.847 and 0.609, respectively, and MWM gives a MSE of LGD prediction at 0.092.

E.3 Recession Probability Method

Figure A3 shows the ROC curves for RPM and E-RPM. E-RPM significantly outperforms RPM in predicting a listing's funding, withdrawal and default outcomes. RPM's AUC for the three prediction tasks are 0.796, 0.587 and 0.603 respectively. The MSE of LGD prediction is 0.037 using the RPM model, which is larger than the MSE of the E-RPM model.

E.4 Gradual Forgetting Method

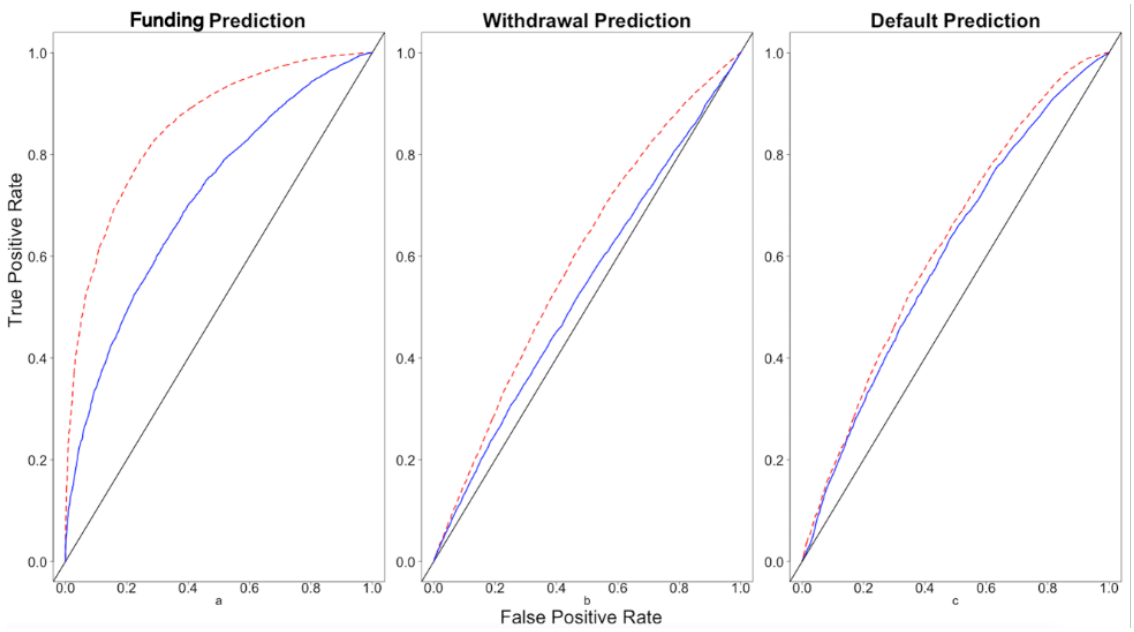
Figure A4 shows the ROC curves for GFM and E-RPM. E-RPM significantly outperforms GFM in predicting listing's funding and withdrawal outcomes. They have similar prediction powers as for the default prediction. GFM's AUCs for the three prediction tasks are 0.638, 0.535, and 0.615, respectively, The MSE of LGD prediction is 0.095 using the GFM model.

Table A1: Summary Stats of Recession Probability (%)

Mean	Median	SD	Max	Min
0.36	0.32	0.097	0.62	0.26

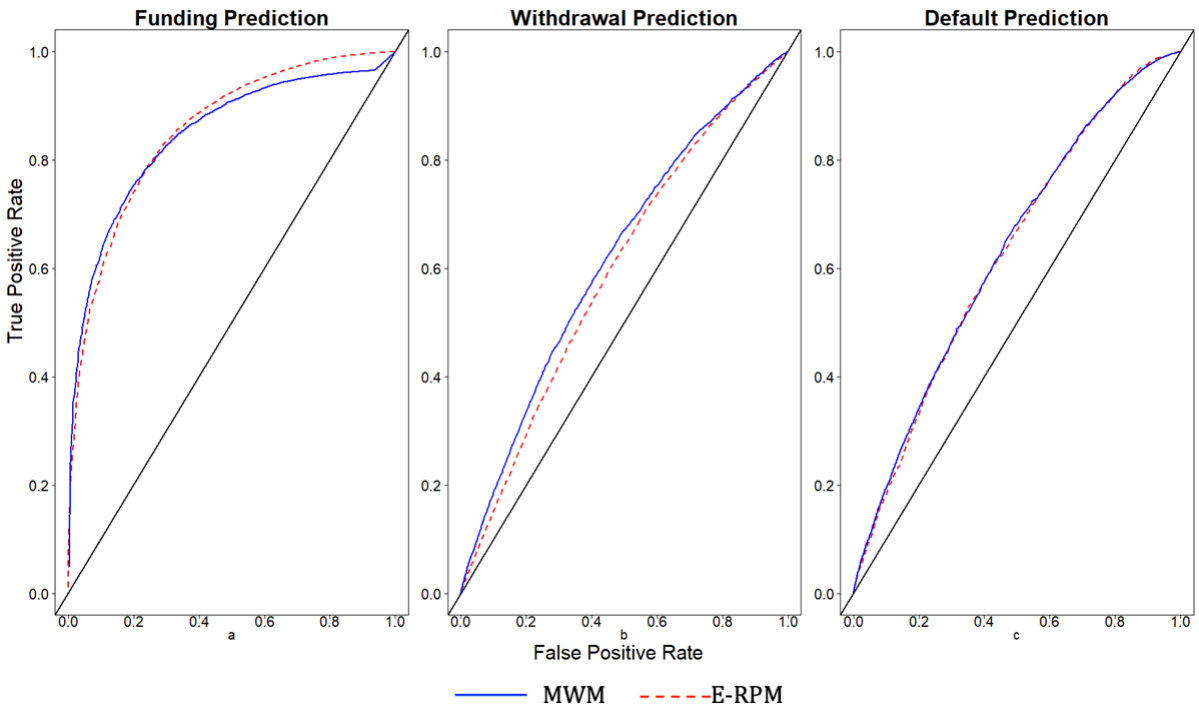
Notes: Summary statistics of the recession probability from 2011 to 2012.

Figure A1: E-RPM vs. EWM



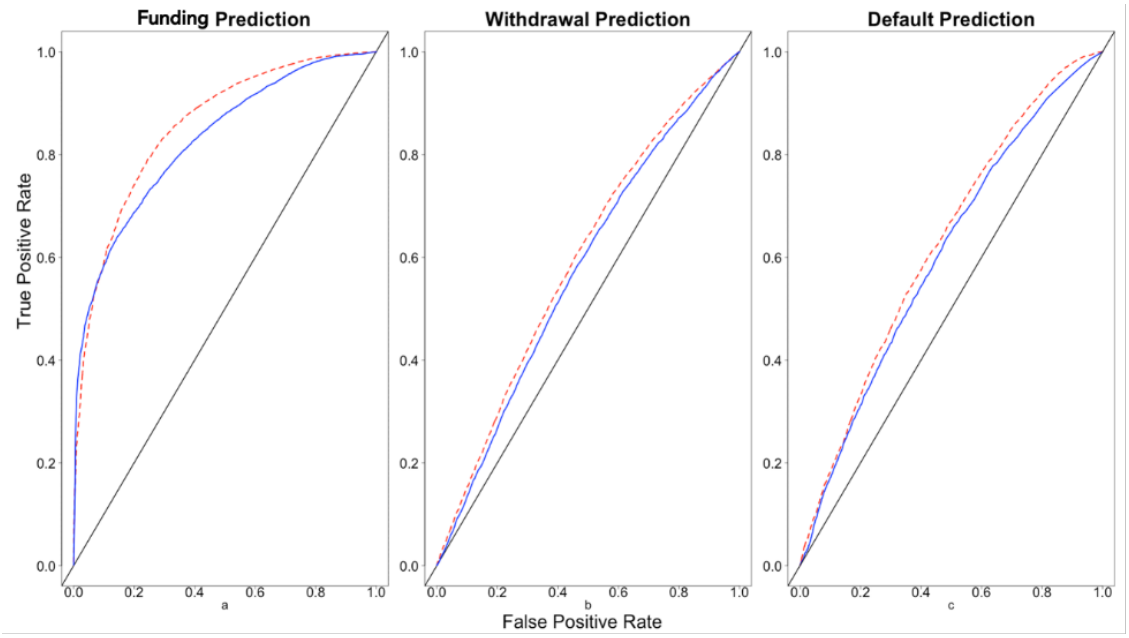
Notes: The larger the area under ROC is, the better prediction performance the model has. E-RPM outperforms EWM in predicting funding, withdrawal and default outcomes.

Figure A2: E-RPM vs. MWM



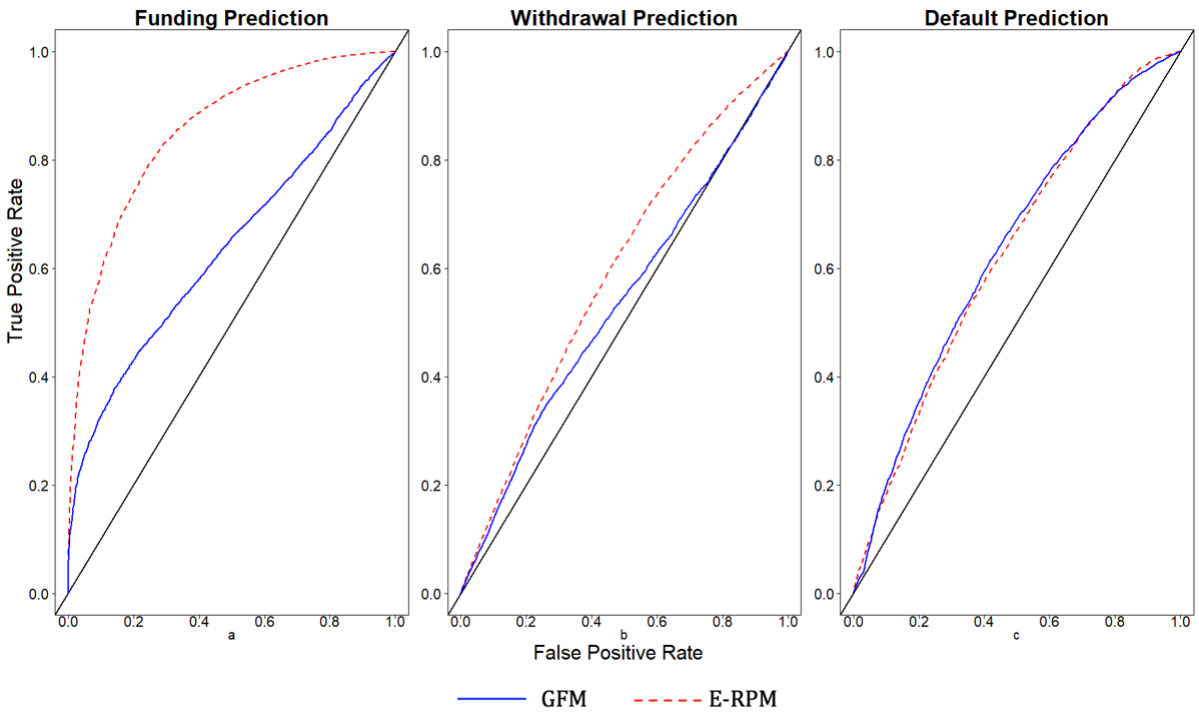
Notes: The larger the area under ROC is, the better prediction performance the model has. E-RPM outperforms MWM in predicting funding outcome. MWM outperforms E-RPM in withdrawal prediction. The two algorithms are very close when predicting default.

Figure A3: E-RPM vs. RPM



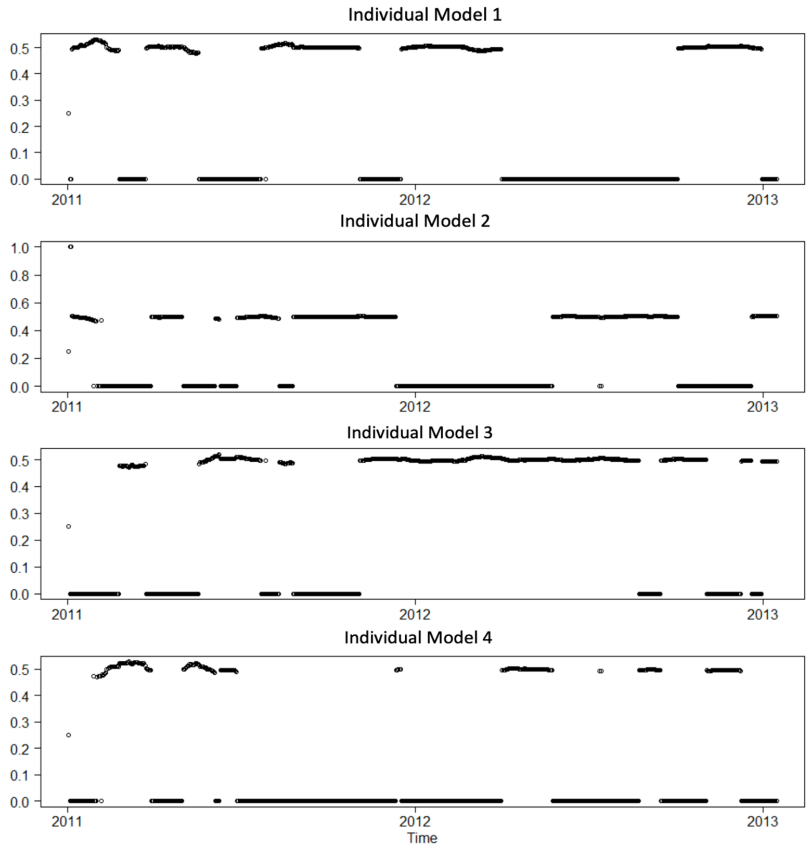
Notes: The larger the area under ROC is, the better prediction performance the model has. E-RPM outperforms RPM in predicting funding, withdrawal and default outcomes.

Figure A4: E-RPM vs. GFM



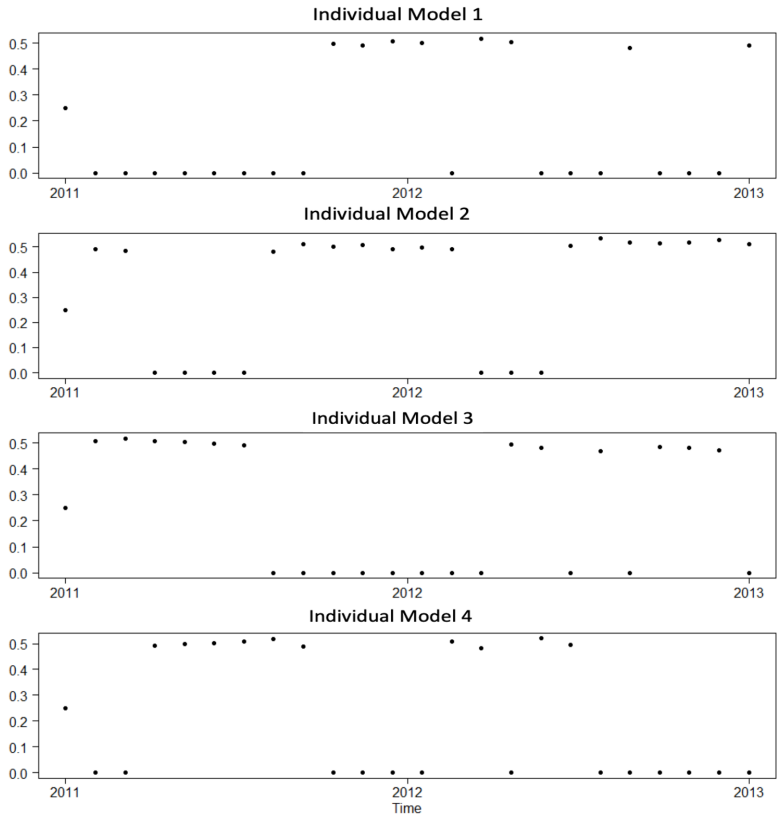
Notes: The larger the area under ROC is, the better prediction performance the model has. E-RPM outperforms GFM in predicting funding, and withdrawal outcomes. The two algorithms are very close when predicting default.

Figure A5: Weights on Individual Withdrawal Model in E-RPM



Notes: This figure shows the weights E-RPM puts on each individual model on withdrawal predictions. The weights are updated every day.

Figure A6: Weights on Individual Withdrawal Model in E-RPM



Notes: This figure shows the weights E-RPM puts on each individual model on default predictions. The weights are updated every month.