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

Xinlong Li and Andrew T. Ching*

July 8, 2020

Abstract

In a rapidly changing world, older data is not as informative as the most recent data. This is known as a *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 recover the way the firm uses the past data to make business decisions. We apply this approach to study how Prosper Marketplace, an online P2P lending platform, adapts in order to address the concept drift problem. More specifically, we develop a two-sided market model, where Prosper uses the past data and machine learning techniques to assess borrowers' and lenders' preferences, borrowers' risks, and then set interest rate for their loans to maximize his expected profits. By observing his interest rate choices over time and using this structural model, we infer that Prosper assigns different weights to past data points depending on how close the economic environments that generate the data are to the current environment. In the counterfactual, we demonstrate that Prosper may not be using the past data optimally, and it could improve its revenue by changing the way it uses data.

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

*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 at Carey Business School, Johns Hopkins University.

1 Introduction

What is the hottest item one day later may become just a fad. Nowadays, individual's behavior can be changing rapidly due to technological innovations which have revolutionized the world. Therefore, treating all historical data equally informative when building an analytical model to guide business decisions could lead to very misleading results. More formally, the problem can be framed as follows. Suppose the independent variables are denoted by X (e.g., consumer characteristics) and the dependent variable is denoted by Y (e.g., consumer choice). The relationship between X and Y can be formulated as $Y = F(X, \theta_t)$. Note that θ_t may change over time. If the change in θ_t is unaccounted for, the inferences we make out of the model will be biased. Researchers in statistics and machine learning call this the *concept drift* problem.

Concept drift is of increasing importance as more data arrive in a stream. Some common examples of this kind of data include online reviews, website clicks, mobile phone apps, credit card transactions, E-commerce purchases, and social networks. Hereafter, we refer this type of data to *streaming data*. Businesses such as credit card companies, e-commerce, etc. are harnessing streaming data to radically change the way they run their businesses.

However, streaming data is very likely to bear the pitfall of concept drift. Models estimated using older data could quickly become obsolete over time. Therefore, conclusions drawn from streaming data analysis will be questionable if the concept drift problem is not accounted for.

Does a firm attempt to address the concept drift problem in a rapidly changing environment? If so, how does the firm make use of data to adapt to the changes? Could a firm better predict consumer behavior by utilizing historical data in different ways? These are important fundamental questions. In particular, from the viewpoint of market intelligence, it will be very useful for a company to know how its competitors use the past data to make decisions.

In this paper, we provide answers to these questions in the context of Prosper Marketplace, one of the largest online Peer-to-peer (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 grades: AA, A, B, C, D, E and HR, where AA denotes the lowest risk and HR stands for the highest risk. Each rating corresponds to a certain interest rate. When setting the rating, Prosper faces a delicate balancing task. The interest rate should balance the demand (investors/lenders)

and supply (borrowers) of loans. If the interest rate is set too high, the supply of loans will drop because more borrowers would withdraw their loan applications. If it is set too low, it cannot attract enough investors to finance the available loans and that could discourage borrowers to return.

Our key insight is that Prosper's decision on classifying risk categories will not only reveal her objective function, but also the way Prosper discounts past data when estimating borrower and investor preferences and the default risk of a loan. Specifically, we develop a two-sided market model to describe Prosper's decision process. We assume that Prosper selectively uses past data to calibrate borrowers' and investors' preferences and the default risk of a loan. Prosper then take the calibrated borrowers' and investors' preference, and loan default risks as given, and assign each loan to a risk category to maximize her profits. Because these calibrated components are a function of the way Prosper weighs past data, Prosper's choice in classifying loans should reveal their data weighing mechanism. Hence, we argue that by observing Prosper's choices, one can infer how Prosper uses the past data. We call this the *generalized revealed preference approach*.

The challenging part is that it is very difficult to parameterize how one selects past data. There are infinitely many non-nested methods that a firm could use to weigh the past data. Hence, we restrict our attention to a set of plausible methods in handling the concept drift problem. We estimate a series of structural models, one for each data selection method. Then we compare their goodness-of-fit. The data selection method which allows the structural model to generate the best model fit will be treated as the most likely method adopted by Prosper among our consideration set.

We study a dataset consisting of 31,807 unsecured personal loans from Dec 2010 to Dec 2012 provided by Prosper Marketplace. Using this structural approach, we find an ensemble model method, which assumes that Prosper builds multiple individual models and makes decisions by taking the weighted average of those individual models, best describes its decision process.

In the counterfactual study, we consider Prosper switches to a different way of utilizing historical data. In particular, we assume Prosper first uses the hidden Markov model to recover the unobserved economic states, then pools together all the data from similar hidden economic states to make rating assignments. Our results show that Prosper can expect a 7.23% increase in its revenue in this "what-if" scenario.

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. In Section 4, we show some reduced form evidence that concept drift exists in our context and the firm updates its algorithms over time. The model framework is summarized in Section 5. We introduce different algorithms of using historical data in Section 6. Section 8 presents the estimate results and counterfactual experiment. Concluding remarks can be found in Section 9. Details about the estimation procedure are presented in the Appendix.

2 Literature

This paper makes a contribution to the recent literature that examines the P2P lending market. Wei and Lin (2017) investigate the effect of market mechanism change on Prosper. They find that when the interest rate is determined by the firm, loans are funded with higher probability, but the pre-set interest rates are higher than borrowers' starting interest rates. Moreover, loans funded under posted prices are more likely to default in comparison to when the "crowd" determines the price of the transaction through an auction process, given all else equal. Lin and Viswanathan (2016) find evidence that home bias, the tendency for transactions to be more likely to occur between parties in the same geographical area rather than outside, exists in the P2P lending market. Iyer et al. (2009) investigates whether lenders in the P2P lending markets are able to use borrower information to infer creditworthiness. They find that lenders in these markets mostly rely on standard banking variables to draw inferences on creditworthiness. At the same time, they also use non-standard or soft sources of information (e.g., pictures, text descriptions, or friend endorsements) in their screening process, especially in the lower credit categories. Lin et al. (2013) show similar findings that the online friendships of borrowers act as signals of credit quality. Friendships increase the probability of successful funding and lower interest rates on funded loans, and they are associated with lower ex post default rates. Zhang and Liu (2012) find evidence of rational herding among lenders. Well-funded borrower listings tend to attract more funding after they control for unobserved listing heterogeneity and payoff externalities. Moreover, instead of passively mimicking their peers (irrational herding), lenders engage in active observational learning (rational herding); they infer the creditworthiness of borrowers by observing peer lending decisions and use publicly observable borrower characteristics to moderate their inferences. Freedman and Jin (2011) find that learning by doing plays an important role in alleviating the information asymmetry

between market players in the P2P lending market. Early lenders did not fully understand the market risk but lender learning is effective in reducing the risk over time. As far as we know, our paper is the first that uses Prosper's choice to infer how she may select the past data in a concept drift environment.

In the machine learning and statistics literature, many studies have shown the existence of concept drift and how it may bias analysis results. Schlimmer and Granger (1986) first noticed the phenomenon of concept drift. 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. Adams et al. (2010) find that a model that adapts to concept drift can outperform models that do not in classifying credit applicants as a good or bad risk. 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 an over time changing rating average of a user to capture the consumer's drifting preference. A review about related works on concept drift can be found in Hoens et al. (2012). We propose a new method to deal with concept drift problem by using the hidden Markov model (HMM). HMM is widely used in speech recognition (Rabiner 1989, Leggetter and Woodland 1995) and computational biology (Borodovsky and McIninch 1993, Durbin et al. 1998). Within the marketing literature, HMM has been used to predict consumers' online purchase decisions by analyzing their web path (Montgomery et al. 2004), study the impact of the family lifecycle on families' budgetary allocations (Du and Kamakura 2006), estimate cross-promotion effects by imputing the level of competitors' promotions (Moon et al. 2007), and model the dynamics of customer relationships in the context of alumni gift-giving (Netzer et al. 2008).

In the economics literature, the concept drift problem is most related to the strand of research that examines structural change. Structural change refers to the fundamental changes in the ways a market or economy functions or operates. Such changes can occur due to a major change in government policy, a war, natural disasters or technological revolutions. Structural changes can deteriorate a regression model's structural stability, i.e., the time-invariance of regression coefficients. Chow (1960) proposes a testing procedure to split the sample into two subperiods, estimates the parameters for each subperiod, and then tests the equality of the two sets of parameters using a classic F statistic. Bai and Perron (1998) develop tests for multiple structural changes.

McConnell and Perez-Quiros (2000) test for the stability of the volatility of U.S. GDP growth rates and find overwhelming evidence of a substantial decrease in volatility around 1984. Other related works include Andrews (1993), Andrews and Ploberger (1994), Stock and Watson (1996) and Ben-David and Papell (1998). Compared with concept drift, structural change mainly refers to sudden changes in the market, whereas concept drift is a more general concept that captures not only sudden changes, but also gradual changes. There is evidence that our economic system changes slowly over time (e.g., Hyndman, 2014), and only occasionally with sudden discontinuous change. Hence, the general framework we propose here could be applied to areas other than P2P lending platform.

3 Institutional Background and Data

Propser.com is the first platform to provide P2P lending service in America. Since its inception, Prosper has been growing fast. The organization facilitated over $11 billion of unsecured loans by 2017. 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.[1] Figures 1 and 2 show the growing trend of Prosper from 2011 to 2016.

In the early years of Prosper's practice, an eBay style auction system that allowed lenders and borrowers to determine the interest rates was used. Applying this business model earned Prosper the name "eBay of Loans." On December 20, 2010, the platform switched to a post-price model with pre-sets interest rates for each loan application (called a listing).[2] In this paper, we focus on listings initiated during the two-year window after the regime change.

To register as a borrower on Prosper, some basic identity information like names, Social Security number, addresses, and telephone numbers should be provided. If that information 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 per loan on Prosper and choose to repay over the 36- or 60-month amortization periods. We only consider loans with a fixed loan length of 36 months since most borrowers choose to repay the loan on a 36-month basis and the

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

[2] A loan application is called a listing on Prosper. We will use these two terms, loan application and listing, interchangeably in this paper.

interest rate for 60-month loans is determined quite differently from that of 36-month loans.

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. Loan applications can only be approved if the borrower's credit score is above a certain threshold.[3] Prosper then uses its risk assessment algorithm to assess each borrower's estimated loss rate. The risk assessment algorithm is trained based on the historical performance of Prosper loans with similar characteristics and adjustment can be made to reflect anticipated deviations from historical performance that may exist due to the current macro-economic or competitive environment. Based on the estimated loss rate, Prosper assigns each listing a rating, and the rating is what determines the listing's interest rate. The rating system has 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 a unique interest rate. The interest rate of each rating level occasionally changes over time. Once the loan application is approved, it is open for investment for 14 days. Moreover, Prosper reports a certain estimated loss rate for all the listings with the same rating. So the estimated loss rate is not listing specific, but rating specific.

Prosper charges the borrower an origination fee on each completed loan. The origination fee is a percentage of the funded loan amount and varies across different Prosper ratings. Prosper charges lenders a 1% annual loan service fee based on the current outstanding loan principal lenders hold. The average interest rate, average annual return, origination fee rate and reported estimated loss rate for each rating level are shown in Table 1. It is worth noting that, on average, lower rated loans have higher annual returns. Specifically, rating E has the highest annual return.

Borrowers make repayments of equal amounts on a monthly basis. If a borrower misses four repayments in a row, the loan is marked as "Defaulted." Most of the loss is taken by lenders because lenders will lose their outstanding principal in those defaulted loans. Defaulted loans hurt Prosper as well. First, Prosper will lose the annual service fee on the unpaid principal of defaulted loans. Second, Prosper has to pay the operating cost to run a department that is in charge of loan collections. Third, default rate is one of the main factors that affects lenders' investment decisions, and it is widely discussed online.[4] A high default rate could generate negative image

[3] The threshold is 600 in our sample period

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

for Prosper and in turn affect Prosper's profit. So the credit screening process is very important to Prosper.

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 on the borrower side as described above. Prosper also provides all the historical loan performance data on its website. Sophisticated lenders can better understand the market by analyzing that data.

The dataset we compile spans from Feb 2006 to Dec 2012. But we are mainly interested in looking at how Prosper adaptively learns the market after the regime change. Thus this subsection only provides a description of all the loans initiated after Prosper switched to posted price business model, which makes the sample period range from Dec 2010 to Dec 2012. Since Prosper did not start to provide rating for each listing until July 2009, we use data between July 2009 and Dec 2010 as the initial period to compute our initial conditions in our model.

Our dataset consists of 31,807 listings. Among those 31,807 listings, 22,277 (70.04%) loan applications get funded and become loans, 5,825 (18.31%) are withdrawn by the borrowers, and 3,705 (11.65%) expired. For each loan application, we observe all the credit history variables, as well as the Prosper rating, interest rate, monthly loan payment amount, and whether this loan application is completed, withdrawn, or expired. For loan applications that originated as loans, we observe the borrowers' full repayment history. For defaulted loans, we observe the percentage of principal loss. Table 2 provides a breakdown of all the listings based on their Prosper ratings. Listings that are rated D and lower account for 59.44% of the total listings. The completed percentage of HR listings is significantly lower than listings in all the other rating levels. It seems to indicate that the interest offered by HR listings cannot offset their risks, so lenders are less interested in investing in those listings. Summary statistics about some key variables are given in Table 3. Figures 3, 4, and 5 show how the number of listings, funded percentages and withdrawn percentages change over time for different rating levels. Generally speaking, loans in this platform become risker and risky loans are slow more likely to be withdrawn over time. Lenders also become more interested in loan applications in better rating categories over time.

Among those 22,277 originated loans, 3,529 (15.84%) defaulted. Table 4 summarizes the number of loans that defaulted at different repayment stages. Almost half (45.43%) of those loans defaulted within 10 repayment cycles. Figure 6 shows how the default rates changes over time for loans with different Prosper ratings. On average, loans with better Prosper ratings are less likely to default,

with the exception that in 2011, loans with rating E are more likely to default than loans with rating HR.

4 Reduced Form Evidence

This section provides some preliminary evidence of the concept drift problem in the online P2P lending industry and how the firm updates its algorithms to adapt to the market changes.

4.1 Evidence of Concept Drift

Concept drift is likely an issue in the context of online P2P lending because lenders' investment preference and borrowers withdrawal and default behavior may change over time. DellaVigna (2009) provides evidence that an individual's risk preference depends on a reference point. If changes in the economic environment affect lenders' reference points, their risk preferences will change as well. The changed risk preferences will affect their investment preferences accordingly. The default behavior of borrowers might be drifting as well. Borrowers who work in the finance industry are usually very capable of repaying their loans when the economy is good because of their high income. However, in 2008 when the economic crisis happened, finance industry practitioners were among the most risky borrowers because they could easily lose their jobs and stop repaying their loans. We can see that even the same occupation information can contribute differently to a borrower's level of risk in different periods due to changes in the economic environment. This is a typical concept drift phenomenon.

To provide evidence to show that concept drift is present, we run three sequences of logistic regressions at different time points by using data within the six-month window prior to that certain time point. The interval between two time points is a day. Dependent variables for the three sequences of regressions are three dummy variables funding, withdrawal and default, which represent whether a listing is funded, withdrawn or defaulted, respectively. For instance, on Jan 1st, 2012, we run three regressions using funding, withdrawal and default as dependent variables. The data we use to fit these three regressions are from Jul 1st, 2011, to Jan 1st, 2012. On Jan 2nd, 2012, we re-estimate these three regressions by incorporating the new data points we got from Jan 2nd, 2012 and getting rid of data points from Jul 1st, 2011. Generally speaking, the regressions using funded or not as the dependent variable explore whether lenders' investment preference is changing over time, while the regressions using withdrawn or not and defaulted or not as the

dependent variables can explore borrowers withdrawal and default behavior over time. We include 29 independent variables in our regressions. Some key variables include amount requested, credit lines, current delinquencies, monthly debt, income range, ect. If concept drift does not exist in our context, we should observe similar coefficient estimates from these three regressions over time. Given the space limitation, we show six key variables which change most over time for funding and withdrawal regressions in Figures 8 and 9 and four key variables for the default regressions in Figure 10. Significant changing trends can be observed for other parameter estimates. These results provide evidence of gradual concept drift in customer behavior in this market.

4.2 Evidence of Firm Adapting

Evidence provided in Section 4.1 suggests that concept drift is present in the P2P lending market. Does Prosper update its model to adapt the change of the market over time? We next conduct reduced form analysis to investigate this. For each loan application, Prosper reports the corresponding estimated loss rate, which is determined based on the borrower's characteristics. If Prosper does not adapt to the fast changing market, the weight Prosper puts on each borrower characteristic should not change too much over time. To test this, as in the previous section, we run a sequence of linear regressions at different time points by using a six-month moving window method. The interval between two time points is a day. The dependent variable in these regressions is estimated loss rate. Figures 8 presents the daily level coefficient estimates over time. All the six coefficient estimates change significantly over time, suggesting that Prosper actively changes its algorithm to adapt to the market changes.

5 Model

We model Prosper as an adaptive decision maker. In each period t, there will be new loan applications arriving. Prosper uses their characteristics to predict loan funded, withdrawal and default probabilities, and loss given default accordingly at the beginning of the current period. At the end-of-period t, Prosper observes loans' status and then updates its model accordingly. In our model, each period is a day. Since Prosper makes money from a loan application only when it is funded, Prosper should try to strike a balance between each loan application's funding probability, withdrawal probability, and at the same time take into account the loan application's level of risk.

Our model has four parts. The first part describes the borrower side model. The second part describes the lender side model. The third part describes how Prosper updates its risk assessment model and the fourth part is Prosper's objective function. We consider six different ways that Prosper may utilize the past observations to deal with the concept drift problem, and we investigate which one can best describe Prosper's decision process. These mechanisms include some common approaches in the machine learning literature (e.g., gradual forgetting, moving window, etc.) and some intuitive methods proposed by us. The details of those algorithms can be found in Section 6.

5.1 Borrower Side

If a borrower with a very good credit history was charged a high interest rate, she may easily find better terms with another lender and withdraw her loan application. That means the borrower's fair market interest rate is lower than what Prosper assigns to her. So it is natural to assume a borrower's outside option is a function of her characteristics.

Let X_i denote the observed characteristics of borrower i. X_i consists of the borrower's financial history information, such as credit score, monthly income, number of delinquencies, prior Prosper loans, and total inquiries, as well as amount requested, stated monthly income, etc. Let Z_{il} collect Prosper-determined variables like Prosper rating, interest rate, estimated loss, etc. Let R_l represents the interest rate Prosper charges on rating l listings.[5] Let E_t represents macro-environmental variables, which include S&P 500 closing quotes, the TED spread, and the U.S. 30-year mortgage rate.[6] Borrower i's utility from getting a rating l loan is

$$\begin{aligned} U_{il}(X_i, Z_{il}; \Gamma_t) &= \gamma_{lt} + \gamma_{1t} \cdot C_l \cdot M_i + \gamma_{2t} \cdot MP_{il} + \gamma_{3t} \cdot M_i + \epsilon_{il}, \\ &= f_1(Z_{il}; \Gamma_t) + \epsilon_{il}, \quad l \in \{AA, A, ..., HR\} \end{aligned} \tag{1}$$

where Γ_t represents the borrower side coefficient vector in period t; C_l is the origination fee rate for a rating l loan; MP_{il} is the monthly payment borrower i has to repay. It is a function of the rating l interest rate and the amount the borrower requested. For a 36-month loan, monthly payment is calculated as follows:

[5] Prosper changes the interest rate for each rating level very occasionally in our sample period. For simplicity, we do not denote the interest rate as a function of t.

[6] S&P 500 closing quotes and the TED spread are daily measurements, while the U.S. 30-year mortgage rate is a weekly measurement.

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

where $R^*_{il} = R_l/12$; M_i is the amount requested by borrower i; ϵ_{il} is an idiosyncratic utility term that follows the extreme value I distribution.

As mentioned above, heterogeneity exists in the borrower's fair market interest rate. Some borrowers may have better outside options. Hence, the outside option is modeled as a function of the borrower's characteristics. Moreover, the macro-environment will also affect an individual's participation propensity in the P2P lending. So we model the outside option as a function of a borrower's characteristics and the macro environment.

$$U_{i0} = f_0(X_i, E_t; \Gamma_t) + \epsilon_{i0}, \tag{2}$$

Assuming type-I extreme value distribution for the idiosyncratic errors ϵ_{il} and ϵ_{i0}, we have borrower i's withdrawal probability for a rating l loan as

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

5.2 Lender Side

A listing's funding probability is defined as follows:

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

where the definitions of X_i, Z_{il} and E_t are the same as in the previous subsection; $\boldsymbol{\beta_t}$ represents the lender side coefficient vector.

Notice that we allow all the coefficients to be a function of t. This allows Prosper to update its model in each period to adapt to the non-stationary environment. This is one of the key insights of this paper. We consider the firm needs to use new information in each period in order to update its models to make decisions. This is especially important when the market is changing rapidly nowadays. Aguirregabiria and Jeon (2018) provide a comprehensive review about how firms learn about demand, costs, or the strategic behavior of other firms in the market. The adaptive learning method in the economics literature bears the similar idea (Doraszelski et al. 2018, Jeon 2016, Evans and Honkapohja 2001, Sargent 1993).

5.3 Firm's Belief on Market Risk

In addition to adaptively updating beliefs on the borrower and lender sides, Prosper also re-evaluates its risk assessment model in each period. This subsection presents how Prosper revises its risk assessment model over time.

Risk assessment models have become increasingly important with the dramatic growth in consumer credit. In addition to classifying loans by risk categories, Prosper also needs to estimate the loss given default (LGD), which is the share of a loan amount that is lost if a borrower defaults. Calculating LGD is a common practice for most financial institutions. Various classification algorithms have been applied to this area (Baesens et al. 2003). In practice, a naive Bayes classifier often performs very well compared with other machine learning algorithms. Viaene et al. (2002) compare different classification methods in the context of expert automobile insurance claim fraud detection and find that the naive Bayes method shows excellent overall prediction compared with more complicated techniques. Wang et al. (2003) show that naive Bayes outperforms decision trees in dealing with the concept drift problem. Other works include Albashrawi and Lowell (2016); Zareapoor and Shamsolmoali (2015); Ngai et al. (2011); Domingos and Pazzani (1997); Duda, Hart and Stork (2001); Friedman, Geiger and Goldszmidt (1997) and Webb (1999). In addition, naive Bayes model can be re-estimated very quickly in each period by incorporating new data points. This feature is attractive for dealing with concept-drifting datasets. We now proceed to describe our modeling framework in more detail.

5.3.1 Naive Bayes Classifier

Assume we have a labeled dataset $\{(X_i, c_i), i = 1, 2, ..., n\}$, where X_i is a multivariate feature vector of listing i, and $c_i \in \{1, 2, ..., K\}$ represents the ith observation's label.

Suppose each listing has m features (independent variables), which can be represented by $X_i = (X_{i1}, X_{i2}, ..., X_{im})$. Given X_i, the probability of assigning listing i to class k is given by

$$p(c_i = k|X_i) = \frac{p(c_i = k) * p(X_i|c_i = k)}{p(X_i)},$$

where $p(c_i = k)$ is the prior probability of a listing belonging to category k. For instance, when we use Naive Bayes classifier to predict default, c_i can take two values: default or not default. The equality follows the Bayes rule. Now the "naive" conditional independence assumption comes into play: given the category c_i, each feature X_{ij} is conditionally independent of every other feature.

That means

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

Therefore, we have

$$\begin{aligned} p(c_i = k|X_i) &\propto p(c_i = k) * p(X_i|c_i = k) \\ &\propto p(c_i = k) * \prod_{j=1}^{m} p(X_{ij}|c_i = k), \end{aligned}$$

we allow each of the conditional distributions $p(X_{ij}|c = k)$ to follow a multinomial distribution with the number of trials equal to 1. For example, assume the jth feature is income and there are three income levels: 1, 2, and 3. If borrower i has income level 2, then $X_{ij} = (X_{ij1}, X_{ij2}, X_{ij3}) = (0, 1, 0)$.

Let the vector $p_{kj} = (p_{kj1}, ..., p_{kjn_j})$ denote the probability distribution of feature j in category k. We have

$$p(X_{ij}|c_i = k) \propto \prod_{h=1}^{n_j} p_{kjh}^{X_{ijh}}, \tag{5}$$

where n_j represents the number of discrete values feature j can take.

Moreover, we have

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

Assume the total number of observations we have is n. Let $\mathbf{\Omega_t} = \{\omega_{1t}, \omega_{2t}, ..., \omega_{Kt}\}$ represent prior belief t and $P_{kjt} = (p_{kj1t}, ..., p_{kjn_jt})$ represent feature j's distribution of class k at the beginning of period t. ω_{kt} and p_{kjht} can be updated as follows:

$$\omega_{kt} = \frac{1}{\sum_{z=1}^{t} \sum_{i=1}^{I_z} 1} \sum_{z=1}^{t} \sum_{i=1}^{I_z} \mathbf{1}[c_i = k], k = 1, 2, ..., K \tag{7}$$

$$p_{kjht} = \frac{\sum_{z=1}^{t} \sum_{i=1}^{I_z} \mathbf{1}[c_i = k] \cdot X_{ijh}}{\sum_{z=1}^{t} \sum_{i=1}^{I_z} \mathbf{1}[c_i = k]}, k = 1, 2, ..., K \tag{8}$$

where I_z represents the number of listings in period z.

The naive Bayes classifier assumes every feature is independent conditional on category, which is not always the case in reality. Although this assumption seems quite strong, the naive Bayes classifier usually works surprisingly well in practice. Actually, there are many papers showing that

the "naive" independence assumption is not as strong as it seems to be. Hand and Yu (2001) argue that because the naive Bayes model is more parsimonious compared with other models that assume dependence between features, it can result in lower variance for the estimates of classification probabilities. Often, the reduction in variance resulting from the relatively few parameters involved in the independent model will more than compensate for any increase in bias due to the naive independence assumption.

We apply a naive Bayes classifier to estimate each listing's default probability and loss given default. When estimating a listing's default probability, the classes are default and no default. A listing's probability of belonging to the default class is the estimated default probability. When estimating a listing's loss given default, we discretize the numeric target value, the loss rate, into four groups using the 25%, 50% and 75% quantiles, and we apply the standard naive Bayes classifier to the discretized data to get each listing's probabilities of belonging to each loss rate group. At the same time, we compute the average loss rate, $\overline{LR}_{kt}$ in each loss group k and in each period t. Then, the predicted loss rate for listing i with rating l, LR_{il}, is defined as the weighted average of $\overline{LR_{kt}}$, $k = 1, 2, 3, 4$. The weight for $\overline{LR}_{kt}$ is the probability that listing i belongs to loss rate group k.

An important issue worth mentioning about the risk assessment model is that at each time step, the class label arrives after the feature vector. That is, the outcome of a loan is revealed after its features are observed. For instance, we cannot observe a loan's default outcome until the repayment day arrives. Moreover,

- We define a loan as defaulted if a borrower misses 4 repayments in a row.[7]
- A loan is considered to be a good performance loan only if the borrower has made at least 3 repayments in a row.
- A loan will be excluded from the good performance loan category once it defaults.

5.4 Firm's Objective Function

The firm's objective is to assign a rating l to each loan application to maximize the expected profit from that loan application, constrained by truthfully reporting the loan application's risk. The decision is based on the state vector $S_t = (\Gamma_t, B_t, \Omega_t, P_{1t}, .., P_{Kt})$. Γ_t and B_t are borrower and lender

[7] This is consistent with how Prosper defines a defaulted loan.

side parameters, respectively. $\Omega_t, P_{1t}, .., P_{Kt}$ denote the risk assessment model parameters. Notice that interest rates at each rating level are determined outside of our model. In our sample period, Prosper seldom changes the interest rate associated with each risk category. This study will focus on modeling Prosper's rating assignment decisions and take interest rates as exogenously given.

The objective function is given as

$$\pi_{il}(X_i, Z_{il}|S_t) = (C_l + L_{il}) \cdot M_i \cdot F_{il} \cdot (1 - W_{il}) + \delta \cdot |EL_{il} - D_{il} \cdot LR_{il}| + \epsilon_{il}, \tag{9}$$

C_l is the origination fee rate for a rating l listing as defined in the borrower side model; L_i denotes the expected annual loan service fee rate Prosper can charge the lenders;[8] M_i is the amount requested by listing i; D_{il} is the expected default probability by assigning listing i with rating l; F_{il} and W_{il} are the funding probability and withdrawal probability, which are calculated in sections 5.1 and 5.2, respectively; EL_{il} is the estimated loss rate at rating l provided by Prosper; δ captures the cost of misreporting the listing's risk; ϵ_{il} is a mean zero idiosyncratic shock to the objective function, which follows a type I extreme value distribution. The estimated loss rate is a discrete measure determined by rating. It occasionally changes over time to adjust to changes in the economic environment. The misreporting term, $|EL_{il} - D_{il} \cdot LR_{il}|$, is important since there is no third party credit rating agency in the P2P lending market other than the platform itself . Most lenders participating in the P2P lending practice are non-professionals and therefore rely heavily on the platform's reported rating to get an idea of a loan application's level of risk. If Prosper deviates from truthfully reporting a listing's risk and only focuses on increasing a listing's funding probability and decreasing its withdrawal probability, it may hurt Prosper's long-term reputation and credibility. So we add this misreporting term to quantify how costly it is for Prosper to deviate from truthfully reporting a

[8] 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 Proser 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 $PT_{il} = \textbf{round}((1\text{-LossRate}_{il})/\frac{1}{36}) = \textbf{round}((1\text{-LossRate}_{il}) \cdot 36)$. 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_{il} is calculated as

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

listing's risk.

To estimate $F_{il}, W_{il}, D_{il}, LR_{il}$, we assume Prosper first uses a particular method (which we will detail in the next section) to select the past data and apply it to the models described in the previous subsections. Then taking the estimates of $\hat{F}_{il}, \hat{W}_{il}, \hat{D}_{il}, \hat{LR}_{il}$ as given, Prosper maximizes this objective function by choosing an optimal rating for each loan listing.

$$l^* = arg\max_l \left[(C_l + L_{il}) \cdot M_i \cdot \hat{F}_{il} \cdot (1 - \hat{W}_{il}) + \delta \cdot |EL_{il} - \hat{D}_{il} \cdot \hat{LR}_{il}| + \epsilon_{il}\right], \tag{10}$$

Given the demand, supply, risk assessment estimates and the type I extreme value distribution assumption on ϵ_{il}, we can write the likelihood function in the closed form:

$$L(X, Z|\mathbf{S}) = \prod_i \prod_l \left(\frac{\exp(\pi_{il}(X_i, Z_{il}|S_i))}{\sum_{j=1}^{7} \exp(\pi_{ij}(X_i, Z_{ij}|S_i))}\right)^{\mathbb{1}\{\text{Rating}_i = l\}}, \tag{11}$$

where $\mathbf{S}$ denote the state variables in all periods; Rating_i is the rating Prosper assigns to listing i. We estimate δ in the objective function via maximum likelihood.[9]

It is worth highlighting the difference between our approach and the traditional structural modeling approach. Unlike traditional structural modeling approach, which typically assumes a firm knows everything about the demand side and supply side when making decisions, we assume that the firm needs to first update its demand side and supply side models in each period, before making decisions (e.g., pricing decision). Therefore, in our case, firm's choice is also a function of how she uses the past data to estimate the demand and supply side to better adapt to market changes. It is this insight that allows us to shed light on which mechanism Prosper may use to select or weigh the past data. In the Results section, we will discuss how we conduct model comparison to let the data reveal which data selection method best describes Proper (subject to the set of methods we consider).

6 Adaptive Learning Algorithms

One common approach to handle the concept drift problem is to "forget" the outdated observations. The idea is that as an observation ages, it becomes less relevant, and more recent observations tend to provide more accurate information about the current world. Among many, *gradual forgetting* and

[9] Note that we would have already estimated $F_{il}, W_{il}, D_{il}, LR_{il}$ before we form this likelihood. Hence the only structural parameter left to be estimated is δ.

moving window are the two most widely used forgetting mechanisms. Gradual forgetting is a full memory approach, which means that no observations are completely discarded from the memory. Observations in memory are associated with weights that reflect their age. In contrast, moving window is a partial memory method, which means a given observation is either inside or outside the training window.

In this section, we first present these two popular forgetting algorithms, gradual forgetting and moving window. We then propose three other methods, the recession probability method (RPM), ensemble recession probability method (E-RPM) and ensemble hidden Markov model method (E-HMM), to help address the concept drift problem.

6.1 Equal Weight Method

As in most studies, we use all available observations to estimate our model and put the same weight on each observation in this method. We call this method equal weight method (EWM).

6.2 Gradual Forgetting Method (GFM)

The gradual forgetting method (GFM) is a full memory approach. The model is trained by using all the observations, and each observation is assigned a weight determined by its age. Many studies show that gradual forgetting is able to improve the model's adaptability to drifting concepts. (Koychev 2000, Klinkenberg 2004, Koren 2010, Helmbold and Long 1994).

In our study, we follow Klinkenberg (2004) and weight observations according to their age using an exponential aging function. Specifically, if the forgetting parameter is $\lambda \in (0, 1)$, then all the observations in period t will be assigned with weight, λ^t. To estimate the logistic regression models for the borrower side and lender side, 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. The estimation details of this method can be found in Appendix A.

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 $\widetilde{\omega}$ and $\widetilde{p}$ denote the corresponding prior probability and feature distribution estimates. Analogous

to equations 7 and 8, we have the following expressions

$$\widetilde{\omega}_{kt} = \frac{1}{V_t} \sum_{z=1}^{t} \sum_{i=1}^{I_z} v_i \mathbf{1}[c_i = k], \quad k = 0, 1, ..., K \tag{12}$$

$$\widetilde{p}_{kjht} = \frac{\sum_{z=1}^{t} \sum_{i=1}^{I_z} v_i X_{ijh} \mathbf{1}[c_i = k]}{\sum_{z=1}^{t} \sum_{i=1}^{I_z} v_i \mathbf{1}[c_i = k]}, \quad k = 0, 1, ..., K \tag{13}$$

where $v_i = \lambda^{t-t_i}$ and V_t is a normalization parameter given by $V_t = \sum_{z=1}^{t} \sum_{i=1}^{I_z} v_i$; t_i represents the period that listing i comes in; I_z denotes all the listings in period z. The other notations are the same as in section 5.3.1.

Analogous to equation 6, we have

$$p(X_i|c_i = k) \propto \prod_{j=1}^{m} \prod_{h=1}^{n_j} \widetilde{p}_{kjht}^{X_{ijh}}, \tag{14}$$

According to Bayes' rule, we have:

$$\begin{aligned} P_i^k = p(c_i = k|X_i) &= \frac{p(c_i = k, X_i)}{p(X_i)} \\ &\propto p(X_i|c_i = k)p(c_i = k) \\ &= p(X_i|c_i = k)\widetilde{\omega}_{kt_i} \\ &\propto \prod_{j=1}^{m} \prod_{h=1}^{n_j} \widetilde{p}_{kjht}^{X_{ijh}} \widetilde{\omega}_{kt_i}, \end{aligned} \tag{15}$$

As in section 5.3.1, we still apply the adaptive forgetting naive Bayes classifier to each listing i to estimate two things: the rating-specific default probability D_{il} and the rating-specific estimated loss rate LR_{il}. It is straightforward to implement the adaptive forgetting naive Bayes classifier for the first task because it is a binary classification problem. The dependent variable in the second task is continuous. To make our classifier suitable for the second task, we discretize the loss rate into four categories as what we did in section 5.3.1. We then employ a grid search over the forgetting parameter and find that when the forgetting parameter equals to 1, GFM gives us the best fit compared with other forgetting parameter values. The result suggests that Prosper is unlikely to use a gradual forgetting method in its practice.

6.3 Moving Window Method (MWM)

Unlike gradual forgetting, the moving window method (MWM) is a partial memory approach. A set of observations defines a window and the model is trained only using observations within the window. Studies have shown that the moving window method can successfully deal with concept drift (Pechenizkiy et al. 2009, Forman 2006, Widmer and Kubat 1996). A large window makes this approach more conservative and a short window makes it more reactive. The estimation for this method is straightforward. We pre-determine the size of the window and train the model using data within the window.

We experiment with different window sizes (1-12 months) and find that an adaptive learning model with a window size equal to 6 months provides the best fit to the data compared with other window sizes.[10]

6.4 Recession Probability Method (RPM)

It is plausible that lenders' investment preferences and borrowers' withdrawal and default behavior depend on the economic environment. To define similar economic environments, we leverage the Smoothed U.S. Recession Probabilities released by the Federal Reserve Bank of St. Louis. 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 in Chauvet (1998) and has been used to study stock market volatility business cycle turning points in many studies like Sornette (2017), Kim and Nelson (1998), Hamilton and Lin (1996), etc. We call this model the recession probability method (RPM). We assign months with similar recession probabilities (differences in recession probabilities smaller than 0.2%) into one group.[11] For example, if the recession probability for December 2011 is 0.4%. Then all the data in the past with recession probabilities between 0.2% and 0.6% are used to build the model for December 2011.

Unlike GFM and MWM, RPM does not forget older information in a mechanical way. Instead,

[10] At first, our MWM results might seem at odd with our finding that there is "no forgetting" in GFM. But we should highlight that GFM impose a specific gradual forgetting process to discount older data. If an agent actually uses MWM to select data, it is possible that we find "no forgetting" is better than some forgetting in GFM because any $\lambda < 1$ will force the agent to discount data in the actual window being used.

[11] The RPM model gives the best log-likelihood when the probability interval is chosen to be 0.2%.

it utilizes older information in a more complicated way, the rationale behind this method 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.

6.5 Ensemble Recession Probability Method (E-RPM)

In all the previous exercises, we try to create a single classifier 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. This is the so called *ensemble* method. For example, random forest is one of the most popular and most powerful machine learning algorithms. It is a type of ensemble machine learning algorithm that operates by constructing a multitude of decision trees outputting the average of the predictions from each individual tree.

In our case, we can also take advantage of the ensemble method. The idea is to allow particular models to specialize in understanding consumers' behavior in certain economic environments. For instance, we can train an individual model which only specializes in understanding consumers' behavior in recession periods. We can rely more on this individual model when the current period is in recession. Similarly, we can train other individual models which specialize in other economic environments and rely on them more when understanding the current economy is their expertise. In other words, we do not expect to get a model which specializes in every economic environment. Instead, we want to train different individual models, which all have their own expertise. We put more weight on an individual model when it is good at understanding the current market.

To get individual models that are experts in different economic environments, we first use the recession probability index to divide the whole dataset 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 boom group. We then estimate a model using each data group respectively. For instance, an individual model trained using data from the recession group is an expert that specializes in understanding consumers' behavior in recession. Then we take a weighted average of each individual model to get our ensemble model. We call this method the

ensemble recession probability method (E-RPM). As an example, we will use how E-RPM computes withdrawal probability, default probability and loss given default are similar.

(i) Initial individual models: We use the data from February 01, 2007 to December 19, 2010 as our initial dataset. We first divide the data into four sub-samples, D_{10}, ..., D_{40}, using the first, second, and third quantiles of the data's recession probability. Then we train four individual models C_{10},...,C_{40}, using each sub-sample, respectively. Notice the specifications of the four individual models are the same. The only difference between them is the dataset on which it is built. We can think C_{10},...,C_{40} as four experts that 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_{1t},...,C_{4t}. Those four individual models are trained using datasets D_{1t-1}, ..., D_{4t-1}, respectively. D_{1t-1} indicates the dataset in which individual model j is trained at the end-of-period $t-1$. For each new loan application i with characteristics X_i, Z_{li} (rating specific characteristics) and current macro environment index E_t at time t, we compute the weighted funding probability by combining the predicted funding probabilities output by every individual model:

$$C_i = \sum_{j=1}^{4} h_{jt} C_{jt}(X_i, Z_{li}, E_t), \tag{16}$$

where h_{jt} is the weight we put on individual model j at time t; C_{jt} represents the jth individual model at time t. It is a function of t because it updates by combining more data points over time.

(iv) Update weights: Following Wang et al. (2003), we define the weight we put on each individual model as the inverse of each individual model's prediction mean squared error (MSE). The MSE for individual model j is given by

$$B_{jt} = \sum_{i=1}^{n_{t-1}} (y_i - C_{jt-1}(X_i, Z_{li}, E_{t-1}))^2, \tag{17}$$

where y_i takes value 1 if listing i is indeed funded and 0 otherwise; C_{jt-1} is individual model $j-1$ at time $t-1$, and n_{t-1} represents the set of listings that are funded or expired in time $t-1$. That is to say, we measure an individual model's prediction performance using the model's prediction accuracy in the last period. We want to make each individual model focus on predicting the right

answer for the cases where it is doing better than the other individuals. 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. That is

$$h_{jt} = \begin{cases} 0 & \text{if individual model } j \text{ has the largest or second largest MSE} \\ [B_{jt}]^{-1} & \text{otherwise} \end{cases}$$

We then normalize h_{jt} to make sure $\sum_j h_{jt} = 1$.

(v) Update individual models: Let D_t denote the whole dataset we have at the end-of-period t. We divide the data into four sub-samples, D_{1t}, ..., D_{4t}, using the first, second, and third quantiles of D_t's recession probability. We train individual model C_{jt} using data D_{jt}.

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

We show how to implement E-RPM using the example of estimating funding probabilities. When applying E-RPM to estimate default probabilities and loss given default, all the procedures are the same as described in this subsection. The only difference is that the period is defined as one month because we do not often observe loans defaulting. If we still used a day as a period, we would have no data points for most of our periods. A month is a reasonable definition of a period in this case.

6.6 Ensemble Hidden Markov Model (E-HMM)

In RPM and E-RPM, we take advantage of the Smoothed U.S. Recession Probabilities to define similar economic environments. The rationale behind those methods is that we expect consumers to behave similarly in similar economic environments. Specifically, in E-RPM, we train multiple individual models, which specialize in different economic environments. Each individual model is trained using data generated from a certain economic environment. The validity of E-RPM is based on two fundamental assumptions. First, the economic environments can be defined using the recession probability index. Second, borrowers' and lenders' behavior can be determined by economic environments. However, these two assumptions are questionable. First, the economic environment is complicated, it is unlikely to be defined by a single index. Second, there are other factors, e.g., Prosper's competitor's policy change, that might affect borrowers' and lenders' behavior and not be captured by the recession probability index.

In this subsection, we propose another method that can overcome the shortcomings of E-RPM. We do not use recession probability index to define economic environments. Instead, we can think lenders' and borrowers' behavior as being generated by some unobserved states. In different states, lenders and borrowers tend to have different investing, borrowing, and defaulting behavior. To recover the unobserved states and model the transactions between those states, we could take advantage of the hidden Markov model (HMM). An HMM is a Markov process with hidden states. In our context, the hidden states represent a finite set of lenders' and borrowers' behavior states. Take lenders' investing behavior for example. For simplicity, assume two lender-behavior states. At the low investing intention state, lenders do not invest a lot on Prosper, possibly because the stock market is more rewarding. Consequently, the listings' funded rate on Prosper is low. In contrast, at high investing intention state, lenders invest more frequently and the overall funding rate on Prosper is high. Lenders stochastically transition among these hidden states. The hidden states the lenders are at and the corresponding transition probabilities can be recovered from the platform's overall funded rates over time. Similar arguments apply when we analyze borrowers' withdrawal and default behavior.

HMM has two main properties. First, it assumes the observation at period t, x_t, is generated from an unobserved state S_t. Second, it assumes the transaction probabilities between unobserved states satisfy the Markov property. That is, the current state S_t depends only upon the value of S_{t-1}. A third assumption of HMM is that the hidden state variable is discrete. To be specific, an HMM should have the following components:

- A sequence of unobserved states $\mathbf{S}_T = \{s_1, s_2, ..., s_T\}$, each one drawn from a set of K values $\{q_1, q_2, ..., q_K\}$.
- A transition probability matrix: $\mathbf{A} = \{a_{11}, a_{12}, ..., a_{KK}\}$, where a_{ij} denotes the probability of moving from state i to state j.
- A sequence of T observations: $\mathbf{X}_T = \{x_1, x_2, ..., x_T\}$. Each observation takes value from $\{0, 1, 2, ..., L\}$.
- Emission probabilities: $\mathbf{E} = \{e_{11}, e_{12}, ..., e_{KL}\}$, where e_{ij} denotes the probability of observation value j being generated by state i, where $j = 1, 2, ..., L$.
- An initial probability distribution over states. $\boldsymbol{\pi} = \{\pi_1, \pi_2, ..., \pi_K\}$.

To illustrate HMM, let's consider lenders' investment decisions. We define each week as a period

and use s_t to present period t's unobserved state. We further assume there are two ($K = 2$) unobserved state values, q_l and q_h. q_l represents the low investing intention state and q_h represents the high investing intention state. The observation in period t, x_t, can take a value of 0 or 1. x_t equals to 1 if the average funding probability in period t is higher than the median funding probability from period 1 to period $t-1$. Otherwise, x_t takes value 0. The initial probability distribution satisfies $\pi_1 = \pi_2 = 1/2$.[12] Then, given the observation sequence $\boldsymbol{X}_T$, we want to learn the most probable sequence of states $s_1, s_2, ..., s_T$, the transition probability matrix $\boldsymbol{A}$, and the emission probabilities $\boldsymbol{E}$.

The joint distribution of a sequence of states and observations can be written as follows

$$P(\boldsymbol{S}_T, \boldsymbol{X}_T) = P(s_1)P(x_1|s_1)\prod_{t=2}^{T} P(s_t|s_{t-1})P(x_t|s_t),$$

The standard algorithm to estimate HMM is the Baum-Welch algorithm (Baum, 1972), which is a special case of the expectation-maximization (EM) algorithm (Dempster et al. 1977). This algorithm works by computing an initial estimate for the probabilities, then using those estimates to compute a better estimate, iteratively improving the probabilities it learns.

After the underlying states are recovered, we follow the ensemble model method described in the previous section to train multiple individual models using data from each underlying state. Notice here, each individual model is an expert in a certain unobserved state, while in E-RPM, each individual model is an expert in a certain economic environment. Unobserved states in HMM are recovered by analyzing consumer behavior data. We do not reply on any external information, e.g., recession probability index, anymore. Then we pool together those individual models by taking a weighted average. The weight on each individual model is determined by the individual model's data fit performance in the previous period. We call this method the ensemble hidden Markov model method (E-HMM).

7 Identification

In this section, we provide some intuitions about our model identification, which can be summarized in two parts.

[12] As long as we have enough observations, the initial probability distribution does not affect the final results.

First, we need to identify all the parameters in our model. The lender side and borrower side parameters can be identified from the variations in lenders' investment decisions and borrowers' withdrawal decisions. Similarly, parameters in the risk assessment model can be identified from variations in borrowers' default behavior. Once these parameters are identified, the only parameter left in our model is δ, which penalizes Prosper from misreporting the risk of each loan application. δ can be identified from ways in which Prosper's rating assignment decisions deviate from a model without the reputation term. For simplicity, assume there is a loan application with very high risk. Assume further that a good rating assigned by Prosper will make this loan application achieve very high funding probability and relatively low withdrawal probability. Then, it is in Prosper's interest to assign a good rating to this loan application, despite it is very risky, if Prosper only maximizes its profit. But if Prosper's rating assignment decision is at the same time constrained by truthfully reporting the loan application's risk, Prosper may not assign a very good rating to such a loan application. In other words, δ can be identified from the trade-off between maximizing profit and maintaining reputation.

Second, we need to identify how Prosper uses historical data to make decisions. We will use a hypothetical example to show the intuition. Suppose that the average default rate for borrowers from California is 50% in the last two years, while only 6% of borrowers from California default in the last 6 months. The default rate changes are shown in Figure 11. At the very beginning of 2013, Prosper needs to determine what rating to assign to a borrower from California. If Prosper uses the last two years data to inform its decision making, then Prosper will believe that a borrower from California is very risky and they need to assign a low rating to such a borrower, given the other borrower characteristics are the same. However, if Prosper only uses the last six months data to learn about the market, it may believe that borrowers from California are quite safe 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 hence lead to different rating assignment decisions. Therefore, by analyzing Prosper's rating assignment decisions, we can identify how Prosper uses historical data in its practice.

8 Results

8.1 Estimate of δ

For each method of selecting past data, we estimate δ. We will select the method best describe Prosper's behavior based on the model fit w.r.t. the observed risk category assignments. The estimates of δ, log-likelihood and root-mean-square-deviation (RMSD) of rating predictions are shown in Table 5. To calculate the RMSD of each model, we first map the letter ratings into numeric values. Ratings AA, A, ..., HR correspond to numbers 1, 2,..., 7, respectively. Comparing those six methods, E-RPM gives the best log-likelihood and RMSD. δ estimates are significant in all of the six models. In E-RPM, the estimate of δ equals to -14.372. That means a 1% deviation from the true estimated loss rate will result in a \$143.7 loss in Prosper's revenue.[13] For instance, the average reported loss rate is 3.03% for listings with rating A. If Prosper's algorithm predicts a listing's loss rate to be 1.70%, which is in the middle of average reported loss rates for AA listings and A listings, it will cost Prosper (3.03-1.70)* \$143.7= \$191.12 to report this listing as an A listing. At the same time, Prosper can only charge a \$300 origination fee for a \$10,000 listing that is rated as A. So the penalty term is large. The result indicates that Prosper tends to report the estimated loss rate quite accurately.

According to the results in Table 5, it is most likely that Prosper is using E-RPM in its practice. E-RPM has the largest log-likelihood and smallest RMSD among the six algorithms of using historical data [14]. Figures 12, 13, and 14 show the weights E-RPM puts on each of its sub-models in predicting funding, withdrawal, and default probabilities. We can find that as the economic environment evolves, E-RPM adjusts the weights it puts on each sub-model to adapt to the changing environments.

8.2 Alternative Prediction Method Performance

The estimation results in the previous section suggest that among the set of data selection methods being considered, Prosper is more likely using an E-RPM (ensemble recession probability model). However, this does not mean E-RPM is the best way of using historical data to understand lenders and borrowers' behavior. It is interesting to examine whether Prosper can better predict each loan

[13] Loan amount is in thousand dollars.

[14] We need to interpret the log-likelihood with caution here since these models are not nested. RMSD might be a better measure here.

application's funding probability, withdrawal probability, default probability and loss given default by switching to the other four ways (GFM,[15] MWM, RPM, and E-HMM) of using historical data. If any of those four methods can outperform E-RPM in those prediction tasks, Prosper can better understand lenders' and borrowers' behavior by using them in its practice.

8.2.1 Gradual Forgetting Method

We compare GFM and E-RPM by looking at their out of sample prediction performances on listing's funding, withdrawal, default, and loss given default outcomes. We use the receiver operating characteristic (ROC) curve as the measurement metric for the first three predictions and use MSE for loss given default prediction.

The ROC curve is a widely used metric in binary classification problems. The rationale behind this method is to show in a graphical way the trade-off between the true positive rate and false positive rate for every possible cut-off for a model. In an ROC curve figure, the x-axis represents the false positive rate and the y-axis represents the true positive rate. A point in ROC space is better than another if it is to the northwest of the other. The area under ROC (AUROC) represents the model's ability to separate positive examples from negative examples. The larger the area is, the better the model's prediction performance is. We explain more details about ROC curve in Appendix B.

To implement GFM, we first need to find the optimal discount factor. We do a grid search from 0.1 to 1 with step size 0.01 and find discount factor 1 gives the best prediction performance in terms of maximizing the AUROCs for the first three prediction tasks and minimizing the MSE for the loss given default model.

Figure 15 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. Table 6 presents the AUROCs and MSEs for different algorithms. GFM's AUROCs for the three prediction tasks are 0.712, 0.534, and 0.597, respectively, while the corresponding AUROCs for E-RPM are 0.854, 0.595, and 0.617, respectively. The MSE of loss given default prediction is 0.045 using the GFM model, while the MSE of the E-RPM model is 0.035.

[15] The equal weight method is a special case of GFM when the forgetting parameter equals to 1.

8.2.2 Moving Window Model

Figure 16 compares the prediction performances of MWM and E-RPM. E-RPM outperforms MWM in all the four prediction exercises. The corresponding AUROCs for funding, withdrawal and default predictions for MWM are 0.800, 0.614 and 0.595, respectively. Moreover, MWM gives a MSE of loss given default prediction at 0.040, which is larger than that of E-RPM.

8.2.3 Recession Probability Method

To implement RPM, we first need to find the optimal probability intervals to define similar economic environments for the borrower side, lender side, default prediction, and loss given default models. We do a grid search over different recession probability intervals from 0.1% to 10% and find that 0.06%, 0.26%, and 0.14% give the best prediction performances for the first three models in terms of maximizing the AUROC, and 0.06% gives the smallest MSE for the loss given default model.

Figure 17 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 AUROC for the three prediction tasks are 0.827, 0.587 and 0.603 respectively. The MSE of loss given default prediction is 0.037 using the RPM model, which is larger than the MSE of the E-RPM model.

8.2.4 Hidden Markov Model

E-HMM gives the best prediction performance among all the methods according to Figure 18. E-HMM's AUROCs for the funding, withdrawal and default predictions are 0.871, 0.635, and 0.623, respectively, which are the best among all the methods. Moreover, the MSE of loss given default prediction produced by E-HMM is 0.034, which is again the best among all the methods.

8.3 Counterfactual Experiment

Comparing the prediction performances of different models, we find that E-HMM can best predict borrower and lender behavior. Therefore, it is interesting to investigate if Prosper can increase its revenue if it indeed employs the E-HMM method in its practice. In the following simulation exercise, we assume Prosper uses E-HMM in its practice to facilitate decision making. That is, we assume Prosper uses E-HMM to calibrate its borrower side, lender side and risk assessment models. In each period, a listing's funding, withdrawal, default, and loss rate outcomes are simulated according to its estimated funding, withdrawal, default probabilities, and loss rate given default

using E-HMM. Prosper then takes into account those newly revealed loan application statuses and updates its model using E-HMM. We run the counterfactual using data from January 1, 2012 to December 31, 2012 and use data from December 19, 2010 to December 31, 2011 as the initial period for the model to learn about the market.

Table 7 summarizes the counterfactual results. The current revenue Prosper makes is $5,854,886, which is recovered from the data. The withdrawal rate, funding rate, default rate and loss give default are 15.98%, 72.50%, 13.01%, and 85.55%, respectively. While under the counterfactual scenario, the revenue Prosper makes is $6,278,026, which is 7.23% higher than Prosper's actual revenue. The corresponding withdrawal rate, funding rate, default rate and loss given default are 13.91%, 75.56%, 13.90%, and 75.17%, respectively. Notice the funding rate increase under the counterfactual scenario. The actual number of listings that get funded is 12,962 under the counterfactual scenario, which is 525 more than the number of listings funded in the data. The average funded amount is $6,592.01 in the counterfactual, which is almost the same as the average funded amount in the data, $6,593.82. Table 8 summarizes the detailed rating distributions for the funded listings in both the data and the counterfactual experiment. One thing worth noting is that there are more A and B level ratings under the counterfactual scenario. The higher the rating is, the lower the interest rate a borrower has to pay and the less likely a borrower will withdraw his loan application. This is why the withdrawal rate drops in the counterfactual scenario. A higher rating may have mixed effects on a listing's funding probability. On the one hand, a higher rating means a lower interest rate, which will discourage potential lenders. On the other hand, a higher rating can increase lenders' investment confidence and in turn increase the listing's funding probability. The counterfactual results show that assigning more listings higher ratings can indeed increase the overall funding rate.

9 Conclusion

In this paper, we study how a FinTech company adaptively learns the market in a changing world. We first provide evidence showing that in the P2P lending market, consumers' borrowing and lending behavior is changing over time. Then we provide evidence to support our hypothesis that Prosper uses the data selectively to adapt to the non-stationary environment. Based on these pieces of evidence, we develop a structural model to capture Prosper's decision process. We propose the generalized revealed preference approach and show that by analyzing the choices made by Prosper,

we recover not only the parameters in her objective function, but also the way Prosper selects data to make decisions. We take advantage of a naive Bayes classifier to build the risk assessment part of our structural model and compare different methods of using historical data. Different measurement metrics show that the method that assumes Prosper trains multiple sub-models and adjusts the weights it puts on each sub-model over time to adapt to the changing market can best describe Prosper's rating assignment decisions. In our counterfactual, we consider Prosper first divides the whole dataset into multiple subsets according to the unobserved states recovered from a hidden Markov model and then combines those individual models together by taking the weighted average to form an ensemble model. The counterfactual exercise shows that Prosper's revenue will increase under this "what-if" scenario.

To the best of our knowledge, this is the first study using structural modeling and machine learning techniques to investigate how a FinTech firm deals with the concept drift problem in complex and dynamic settings. Addressing the concept drift problem is of increasing importance because this problem exists in a wide variety of data such as online reviews, E-commerce purchase data, web clicks, mobile phone click data, etc. Although the research context is P2P lending, we believe the method developed in this study can be generalized to other settings in which concept drift exists.

References

1. Aguirregabiria, Victor, and Jihye Jeon. "Firms' Beliefs and Learning: Models, Identification, and Empirical Evidence." (2018).

2. Andrews, Donald W.K. 1993. "Tests for Parameter Instability and Structural Change with Unknown Change Point." Econometrica. July, 61:4, pp. 821–56.

3. Andrews, Donald W.K. and Werner Ploberger.1994. "Optimal Tests When a Nuisance Parameter is Present Only Under the Alternative." Econometrica. November, 62:6, pp. 1383–414.

4. Antonakis, A. C., and M. E. Sfakianakis. "Assessing naive Bayes as a method for screening credit applicants." Journal of applied Statistics 36.5 (2009): 537-545.

5. Bai, Jushan and Pierre Perron. 1998. "Estimating and Testing Linear Models with Multiple Structural Changes." Econometrica. January, 66:1, pp. 47–78.

6. Baum, Leonard. "An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process." Inequalities 3 (1972): 1-8.

7. Bell, Robert M., Yehuda Koren, and Chris Volinsky. "The bellkor 2008 solution to the netflix prize." Statistics Research Department at AT&T Research 1 (2008).

8. Ben-David, Dan and David H. Papell. 1998. "Slowdowns and Meltdowns: Postwar Growth Evidence from 74 Countries." Review of Economics and Statistics. November, 80:4, pp. 561–71.

9. Bera, Anil K., and Sung Y. Park. "Optimal portfolio diversification using the maximum entropy principle." Econometric Reviews 27.4-6 (2008): 484-512.

10. Borodovsky, Mark, and James McIninch. "GENMARK: parallel gene recognition for both DNA strands." Computers and chemistry 17.2 (1993): 123-133.

11. Brocato, Joe, and Steve Steed. "Optimal asset allocation over the business cycle." Financial Review 33.3 (1998): 129-148.

12. Chauvet, Marcelle. "An econometric characterization of business cycle dynamics with factor structure and regime switching." International economic review (1998): 969-996.

13. Chow, Gregory C. 1960. "Tests of Equality Between Sets of Coefficients in Two Linear Regressions." Econometrica. 28:3, pp. 591–605.

14. Clogg, Clifford C., Eva Petkova, and Adamantios Haritou. "Statistical methods for comparing regression coefficients between models." American Journal of Sociology 100.5 (1995): 1261-1293.

15. 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.

16. Crespo, Fernando, and Richard Weber. "A methodology for dynamic data mining based on fuzzy clustering." Fuzzy Sets and Systems 150.2 (2005): 267-284.

17. DellaVigna, Stefano. "Psychology and economics: Evidence from the field." Journal of Economic literature 47.2 (2009): 315-72.

18. 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) (1977): 1-38.

19. Domingos, Pedro, and Michael Pazzani. "On the optimality of the simple Bayesian classifier under zero-one loss." Machine learning 29.2 (1997): 103-130.

20. Doraszelski, Ulrich, Gregory Lewis, and Ariel Pakes. "Just starting out: Learning and equilibrium in a new market." American Economic Review 108.3 (2018): 565-615.

21. Durbin R, Eddy S, Krogh A, Mitchison G. Biological Sequence Analysis. Cambridge, UK: Cambridge University Press; 1998.

22. Dzyabura, Daria, and John R. Hauser. "Active machine learning for consideration heuristics." Marketing Science 30.5 (2011): 801-819.

23. Evans, George W., and Seppo Honkapohja. Learning and expectations in macroeconomics. Princeton University Press, 2012.

24. Fitzpatrick, Trevor, and Christophe Mues. "An empirical comparison of classification algorithms for mortgage default prediction: evidence from a distressed mortgage market." European Journal of Operational Research 249.2 (2016): 427-439.

25. Forman, George. "Tackling concept drift by temporal inductive transfer." Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2006.

26. Friedman, N., D. Geiger, and M. Goldszmidt, 1997, Bayesian Network Classifiers, Machine Learning, 29: 131-163.

27. Garber, Tal, Jacob Goldenberg, Barak Libai, and Eitan Muller. "From density to destiny: Using spatial dimension of sales data for early prediction of new product success." Marketing Science 23, no. 3 (2004): 419-428.

28. Hamilton, James D., and Gang Lin. "Stock market volatility and the business cycle." Journal of applied econometrics 11.5 (1996): 573-593.

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

30. Helmbold, David P., and Philip M. Long. "Tracking drifting concepts by minimizing disagreements." Machine learning 14.1 (1994): 27-45.

31. Hyndman, Robert (2014) "Structural Breaks," https://robjhyndman.com/hyndsight/structural-breaks/

32. Jeon, Jihye. Learning and investment under demand uncertainty in container shipping. NYU Stern working paper. Available at http://www. jihyejeon. com, 2016.

33. 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.

34. 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.

35. 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.

36. Koren, Yehuda. "Collaborative filtering with temporal dynamics." Communications of the ACM 53.4 (2010): 89-97.

37. Kolter, Jeremy Z., and Marcus A. Maloof. "Dynamic weighted majority: A new ensemble method for tracking concept drift." Data Mining, 2003. ICDM 2003. Third IEEE International Conference on. IEEE, 2003.

38. Koychev, Ivan. "Gradual forgetting for adaptation to concept drift." Proceedings of ECAI 2000 Workshop on Current Issues in Spatio-Temporal Reasoning, 2000.

39. Lazarescu, Mihai M., Svetha Venkatesh, and Hung H. Bui. "Using multiple windows to track concept drift." Intelligent data analysis 8.1 (2004): 29-59.

40. Leggetter, Christopher J., and Philip C. Woodland. "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models." Computer speech and language 9.2 (1995): 171-185.

41. 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.

42. Lin, Mingfeng, and Siva Viswanathan. "Home bias in online investments: An empirical study of an online crowdfunding market." Management Science 62.5 (2015): 1393-1414.

43. McConnell, Margaret M. and Gabriel Perez- Quiros. 2000. "Output Fluctuations in the United States: What Has Changed Since the Early 1980s?" American Economic Review. Decem- ber, 90:5, pp. 1464–76.

44. Moon, Sangkil, Wagner A. Kamakura, and Johannes Ledolter. "Estimating promotion response when competitive promotions are unobservable." Journal of Marketing Research 44.3 (2007): 503-515.

45. Netzer, Oded, James M. Lattin, and Vikram Srinivasan. "A hidden Markov model of customer relationship dynamics." Marketing science 27.2 (2008): 185-204.

46. Pechenizkiy, Mykola, et al. "Online mass flow prediction in CFB boilers with explicit detection of sudden concept drift." ACM SIGKDD Explorations Newsletter 11.2 (2010): 109-116.

47. 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).

48. Rabiner, Lawrence R. "A tutorial on hidden Markov models and selected applications in speech recognition." Proceedings of the IEEE 77.2 (1989): 257-286.

49. Sargent, Thomas J. "Bounded rationality in macroeconomics: The Arne Ryde memorial lectures." OUP Catalogue (1993).

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

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

52. Srinivasan, Venkat, and Yong H. Kim. "Credit granting: A comparative analysis of classification procedures." The Journal of Finance 42.3 (1987): 665-681.

53. Steenburgh, Thomas J., Andrew Ainslie, and Peder Hans Engebretson. "Massively categorical variables: Revealing the information in zip codes." Marketing Science 22.1 (2003): 40-57.

54. Stock, James H. and Mark W. Watson. 1996. "Evidence on Structural Instability in Macroeconomic Time Series Relations." Journal of Business and Economic Statistics. July, 14:3, pp. 11–30.

55. Stutzer, Michael. "A simple nonparametric approach to derivative security valuation." The Journal of Finance 51.5 (1996): 1633-1652.

56. 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.

57. 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.

58. Webb, A., 1999, Statistical Pattern Recognition (London: Arnold).

59. Wei, Zaiyan, and Mingfeng Lin. "Market mechanisms in online peer-to-peer lending." Management Science 63.12 (2016): 4236-4257.

60. Widmer, Gerhard, and Miroslav Kubat. "Learning in the presence of concept drift and hidden contexts." Machine learning 23.1 (1996): 69-101.

61. Yeh, I-Cheng, and Che-hui Lien. "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients." Expert Systems with Applications 36.2 (2009): 2473-2480.

62. 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%
Estimated Loss	1.42%	3.03%	5.56%	7.94%	10.83%	14.41%	17.08%

Note: 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 estimated loss in this Table is reported by Prosper. For loan applications with the same rating, Prosper reports the same estimated loss. The reported estimated loss is not necessary to coincide with the real loss rate. This is why the the average annual return and the estimated loss 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

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

Table 3: Variables: All Listings (2011-2013)

Variable	Mean	SD	Max	Min
Panel A: Loan Features				
Amount Requested ($1000)	6.829	0.095	25	2
Amount Funded ($1000)	4.975	4.584	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
Funding Threshold	0.78	0.13	1	0.70
Panel B: Borrower Credit Variables				
Bank Card Utilization (%)	51.50	33.12	223	0
Home Owner (0/1)	0.488	0.500	1	0
Estimated Loss (%)	10.53	5.29	20.3	0.49
Credit Score	707	44.19	780	610
Current Credit Lines	9.11	5.39	56	0
Credit Lines Last 7 Years	25.76	13.93	120	2
Current Delinquencies	0.476	1.32	27	0
Delinquencies Last 7 Years	0.476	1.32	27	0
Monthly Income ($1000)	5.667	13.837	1,750	0
Income Range	4.04	1.30	7	1
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	0.327	0.694	7	0
Prior Prosper Loans Active	0.145	0.352	2	0
Prior Loans Ontime Payments	23.99	18.66	110	0
Prior Loans Late Cycles	0.914	3.31	43	0
Monthly Debt ($1000)	0.873	1.346	100.3	0
Real Estate Balance ($1000)	107.6	162.8	3,830	0
Group (0/1)	0.033	0.18	1	0
Panel C: Market Level Variables				
Mortgage Rate (%)	4.380	0.507	5.330	3.580
TED Spread (%)	0.323	0.107	0.570	0.140
Adjusted Closing Price	1322	78.3	1466	1099

Table 4: Defaulted Loans

Months	# of Defaulted Loans	Average Loss Rate
1	113 (2.31%)	95.1%
2-5	1,136 (23.25%)	94.8 %
6-10	971 (19.87%)	85.5%
11-15	675 (13.81%)	73.8%
16-20	402 (8.23%)	61.3%
21-25	193 (3.95%)	47.4%
>25	40 (0.82%)	32.1%
Total	4,887 (100%)	58.59%

Note: The first column represents the number of repayment cycles. Loss rate is defined as the unpaid principal divided by the total amount borrowed.

Table 5: Log-Likelihood and RMSD

	$\hat{\delta}$	Log-Likelihood	RMSD
EWM	-1.273^{***} (0.116)	-57715.10	2.605
GFM	-1.273^{***} (0.116)	-57715.10	2.605
MWM	-7.460^{***} (0.118)	-55373.54	2.179
RPM	-7.196^{***} (0.014)	-55456.76	2.219
E-RPM	-14.372^{***} (0.145)	-51595.39	1.814
E-HMM	-2.487^{***} (0.113)	-57232.12	2.589

Note: For GFM, the forgetting factor takes value 1. In MWM, the window size is set to be 6 months. Standard errors of $\hat{\delta}$ are in brackets. *p<0.1; **p<0.05; ***p<0.01.

Table 6: Comparison of Prediction Results

	GFM	MWM	RPM	E-RPM	E-HMM
Funding AUROC	0.712	0.800	0.827	0.854	0.871
Withdrawal AUROC	0.534	0.614	0.587	0.595	0.635
Default AUROC	0.597	0.595	0.603	0.617	0.623
Loss Given Default MSE	0.045	0.040	0.037	0.035	0.034

Note: AUROC represents area under ROC. We know that the more the ROC to the upper left corner, the better prediction performance the corresponding method has. So the larger the AUROC is, the better prediction performance the corresponding method has.

Table 7: Counterfactual Results

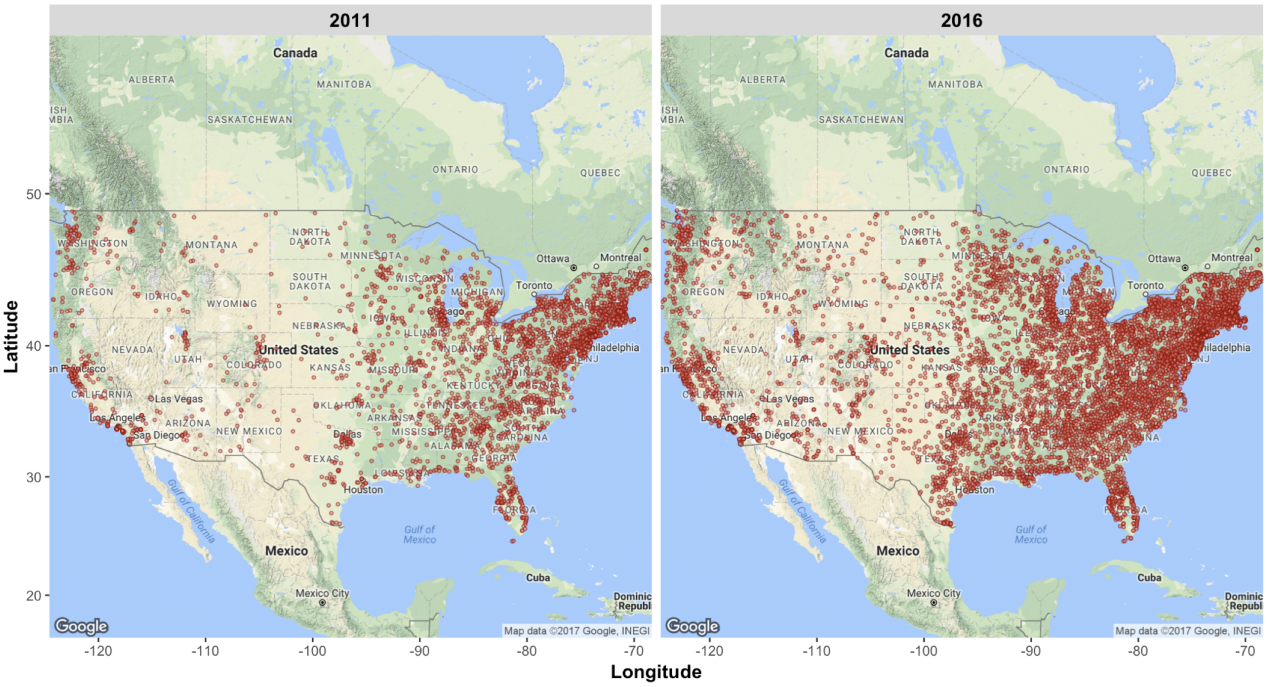
	Data	E-HMM
Funding Rate	72.50%	75.56%
Withdrawal Rate	15.98 %	13.91%
Default Rate	13.01%	13.90%
Loss Given Default	85.55%	75.17%
Average Funded Amount	$6,593.82	$6,592.01
Revenue	$5,854,886	$6,278,026

Table 8: Rating Distribution

Ratings	AA	A	B	C	D	E	HR	Total
Real Data	659	1,545	1,664	2,021	1,761	1,218	3,607	12,475
E-HMM	152	4,261	3,593	1,039	727	494	2,695	12,961

Note: This table presents the simulated rating distribution in the counterfactual scenario.

Figure 1: Cities With $10K+ Loan Originated



Notes: Each red dot represents a city in which the total amount of loans originated is worth more than $10,000.

Figure 2: Amount Originated Across States

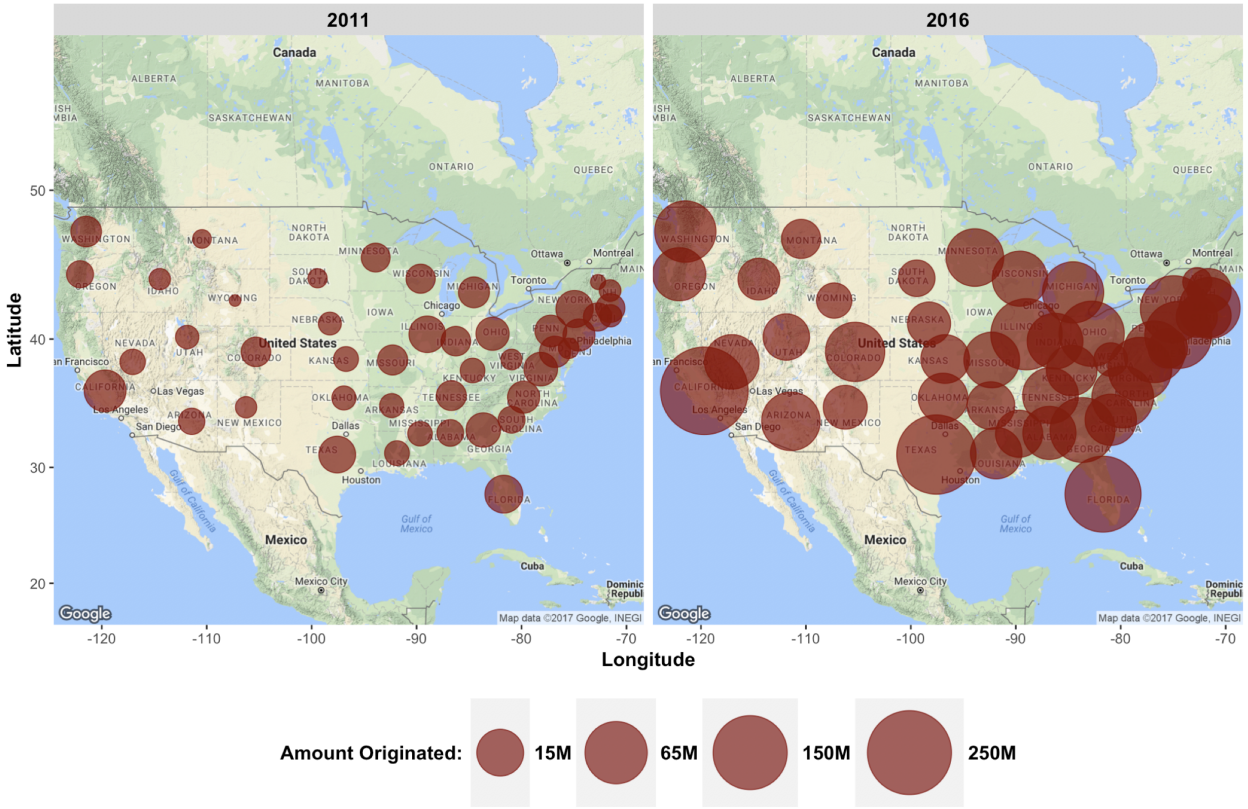
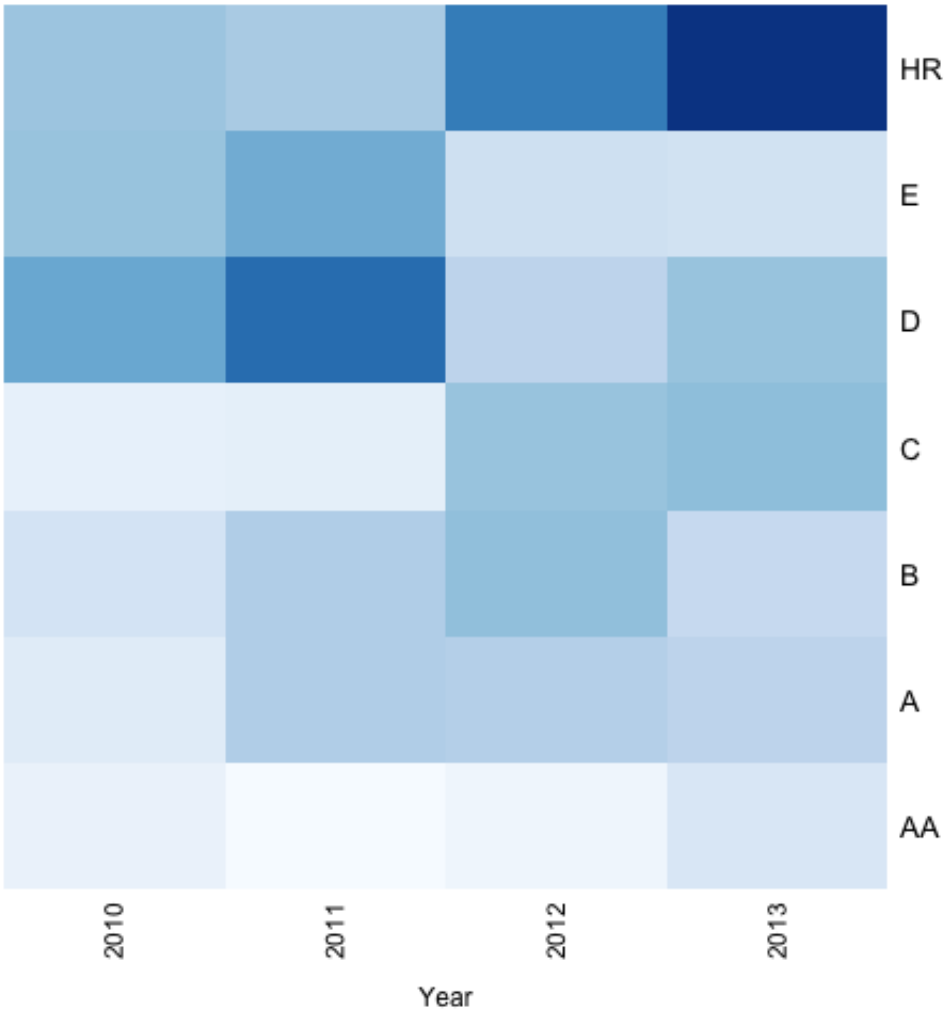
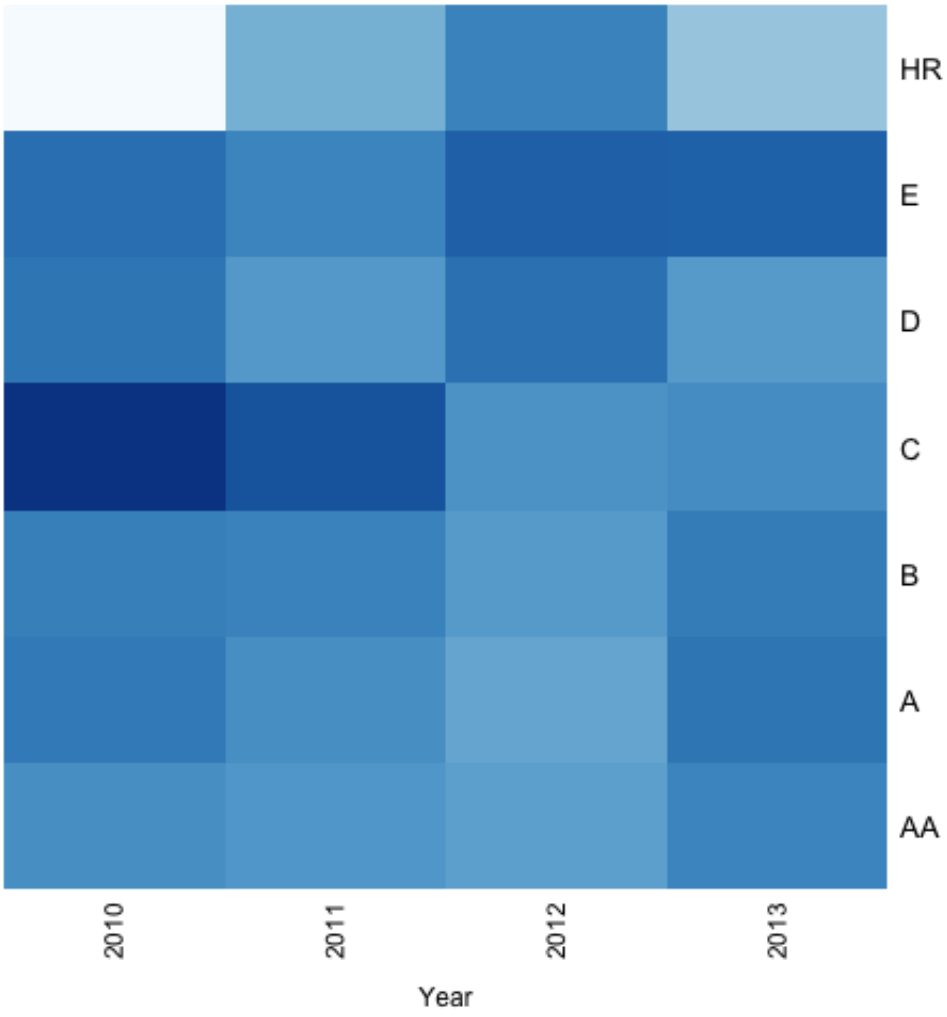


Figure 3: Number of Listings



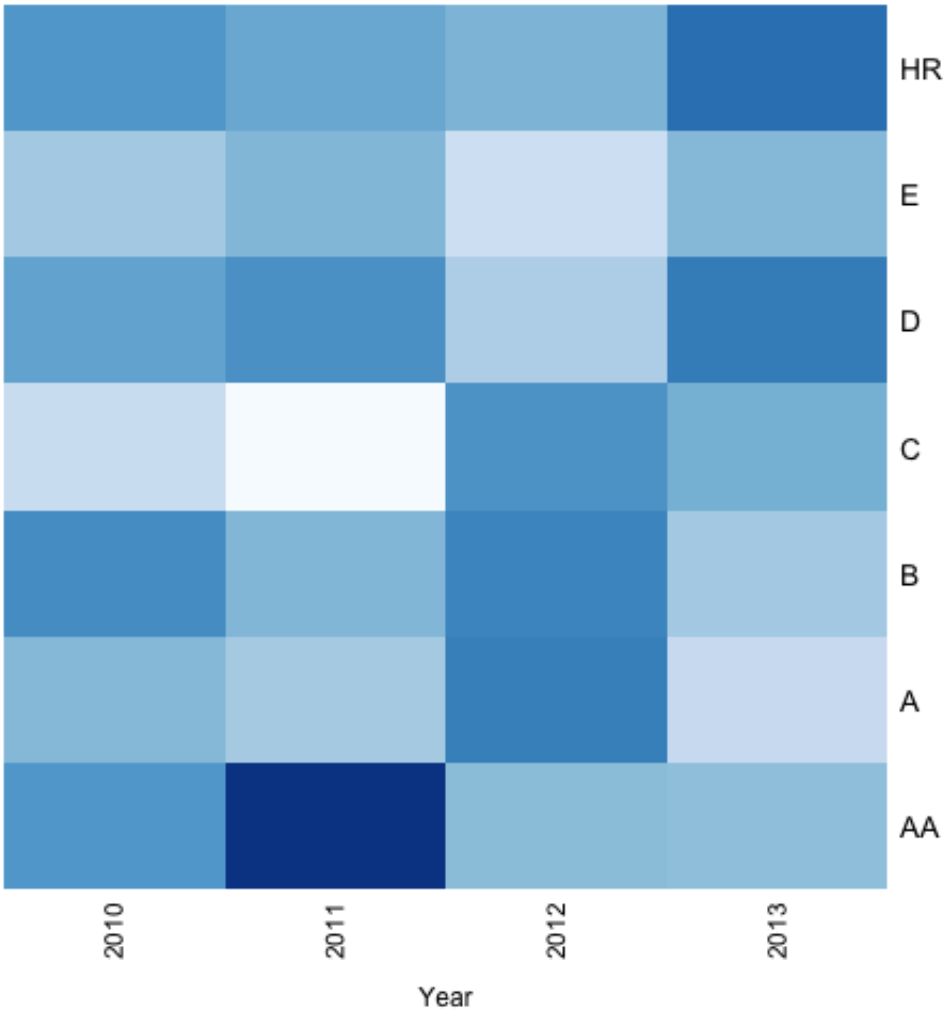
Notes: Figure 3 is a heatmap of the number of listings distributed across different ratings in different years. The darker a category's color is, the more listings fall into this category. We can notice that most listings on Prosper.com are assigned with ratings from D to HR. Especially, from 2012, HR became the largest rating category.

Figure 4: Percentage of Funded Listings



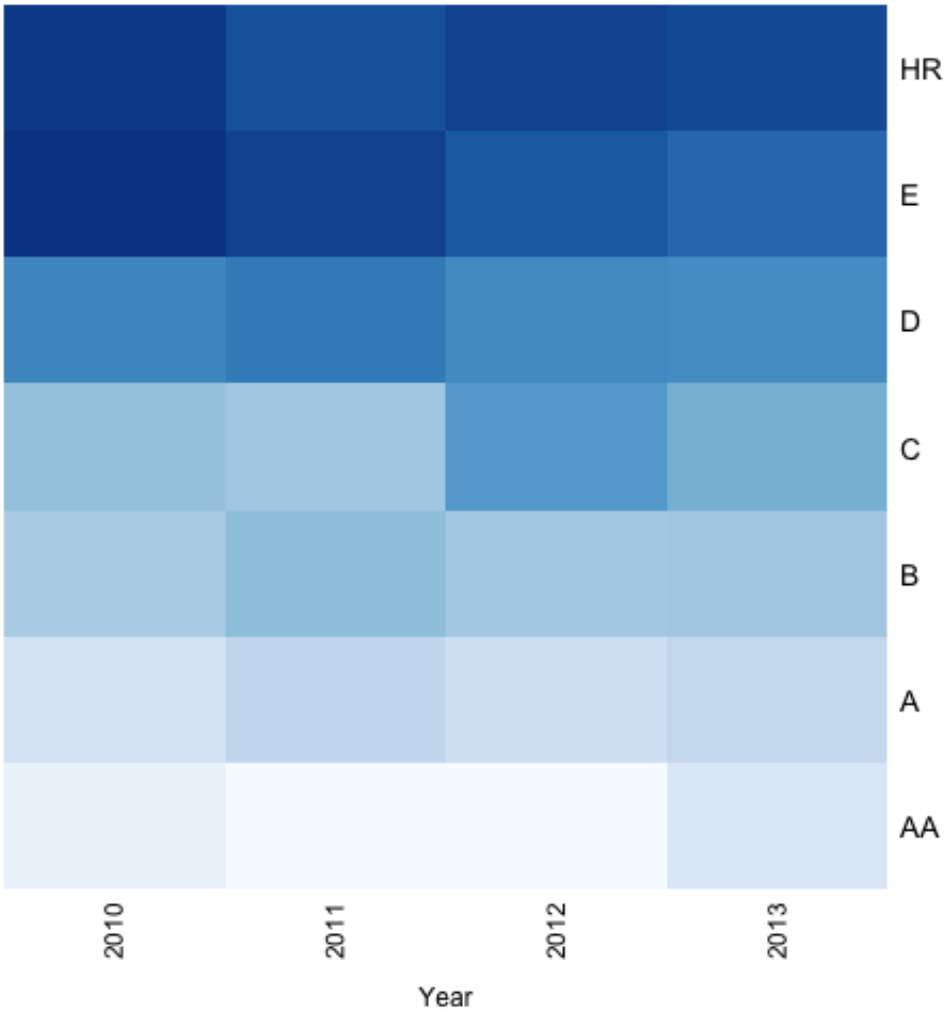
Notes: Figure 4 is a heatmap of the funded percentage distributed across different ratings in different years. The darker a category's color is, the larger percentage of listings are funded in this category. In 2010 and 2011, listings with rating C are most likely to be funded by lenders. While from 2012, rating E becomes the most funded category.

Figure 5: Percentage of Withdrawn Listings



Notes: Figure 5 is a heatmap of the withdrawn percentage distributed across different ratings in different years. The darker a category's color is, the larger percentage of listings are withdrawn in this category. In 2011, listings with rating AA are most likely to be withdrawn by borrowers. While in 2013, listings with ratings D to HR are most likely to be withdrawn.

Figure 6: Percentage of Defaulted Loans



Notes: Figure 6 is a heatmap of the default rate distributed across different ratings in different years. The darker a category's color is, the larger percentage of loans are defaulted in this category. Overall speaking, loans with worse ratings are more likely to default. This is true all the years. But it is worth noticing that in 2011, E rating loans are more likely to default than HR rating loans.

Figure 7: Firm Adapting

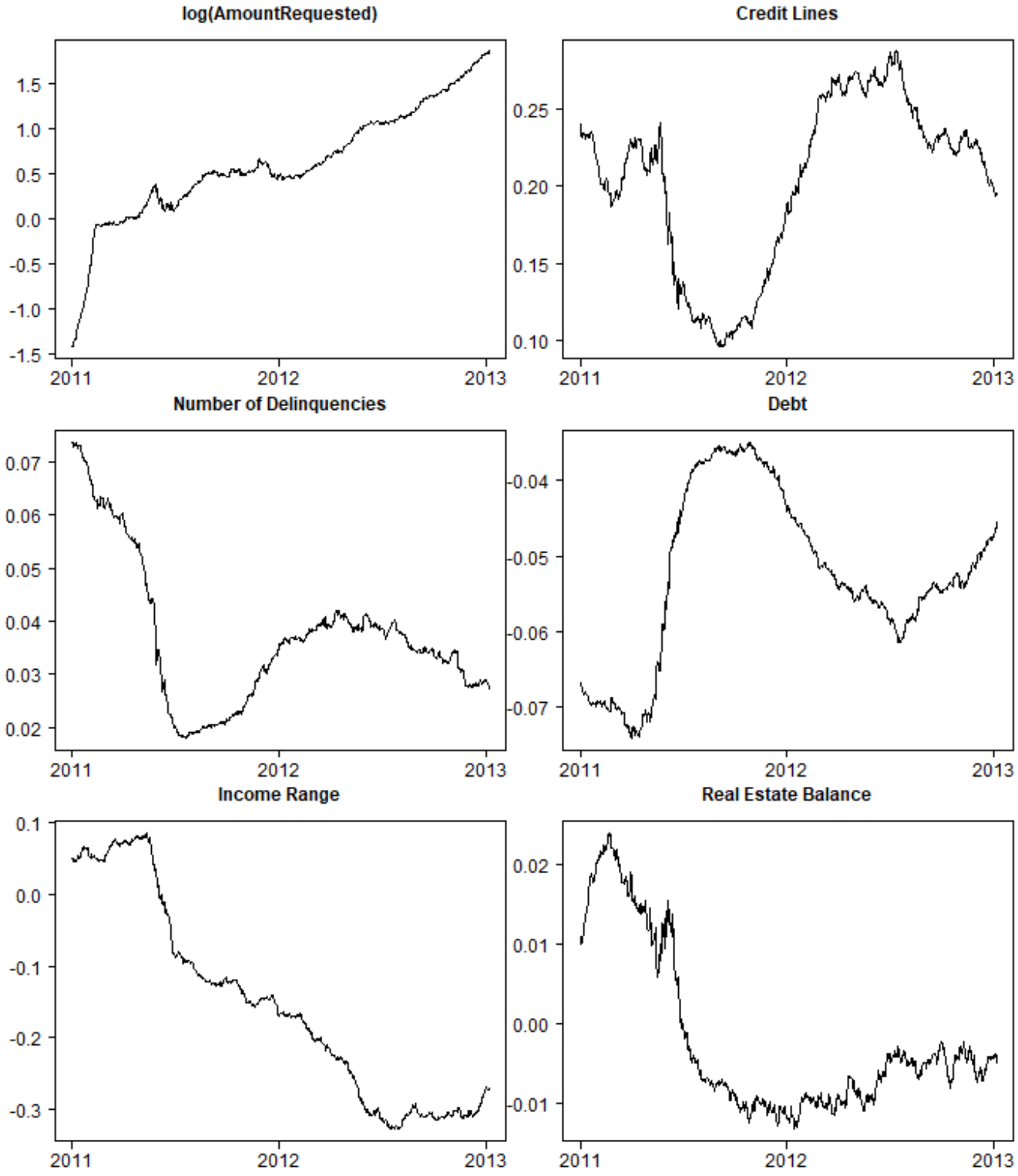


Figure 8: Change in Coefficients Over Time: Funded

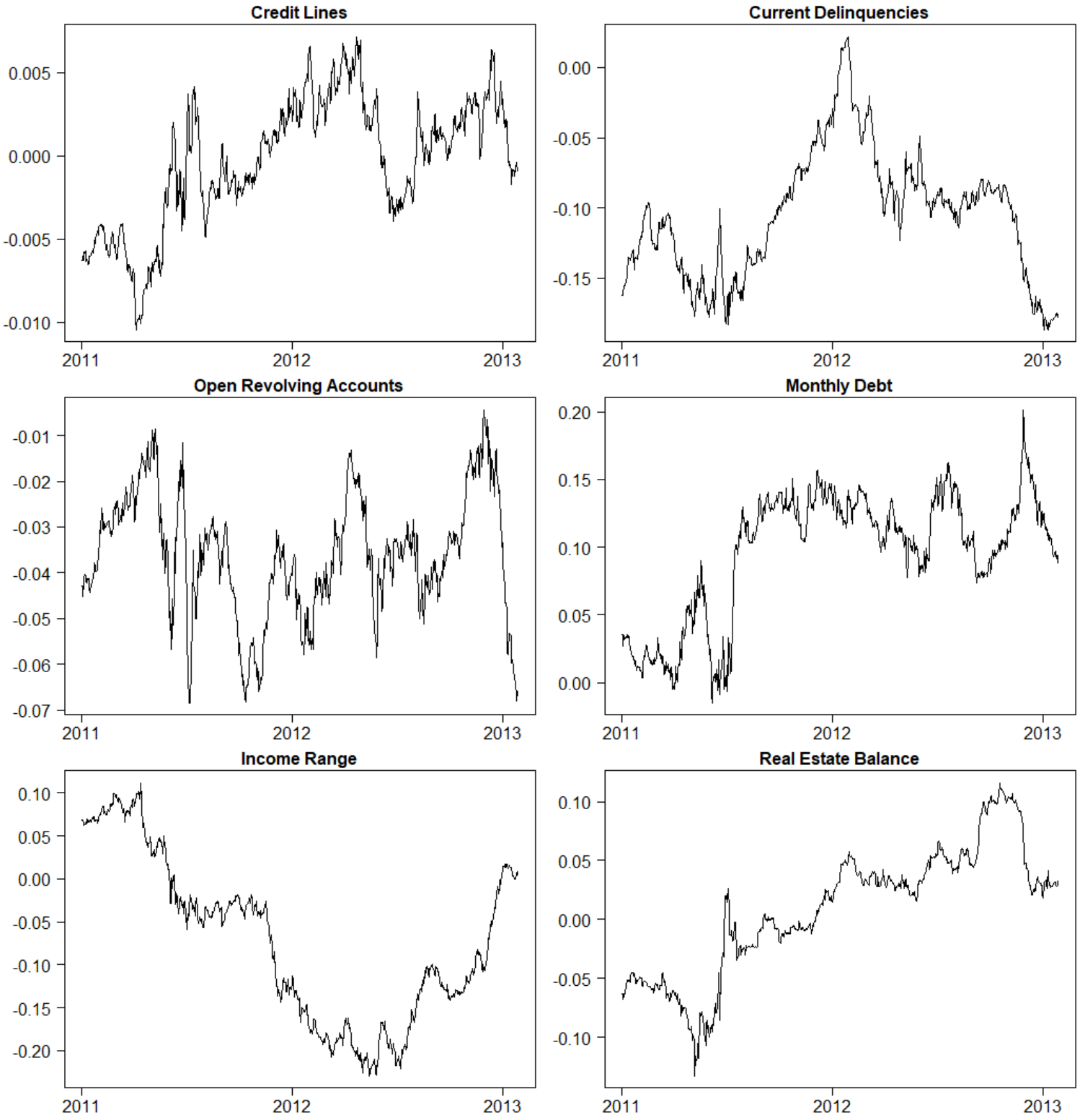


Figure 9: Coefficients Change Over Time: Withdrawal

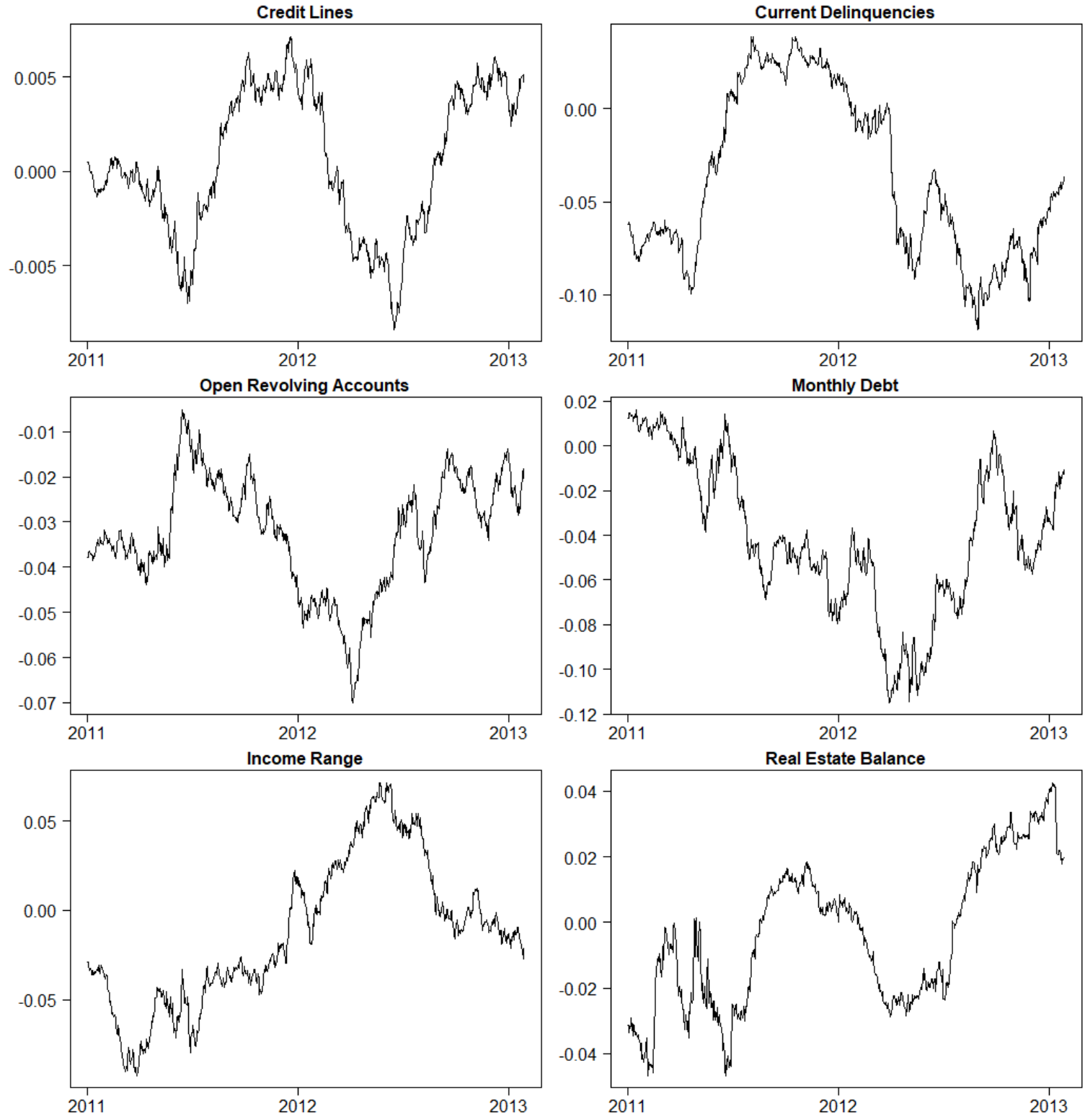
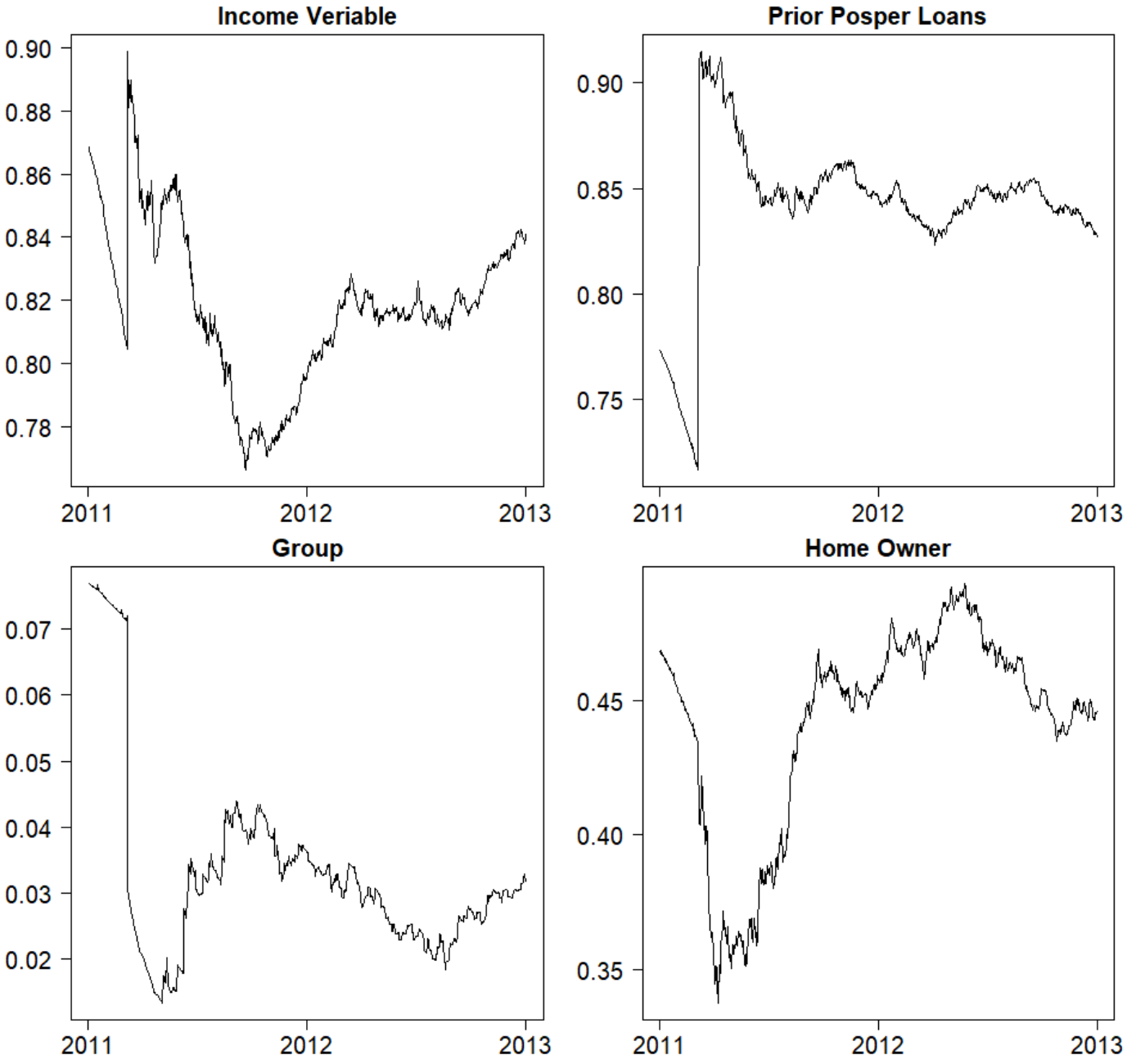
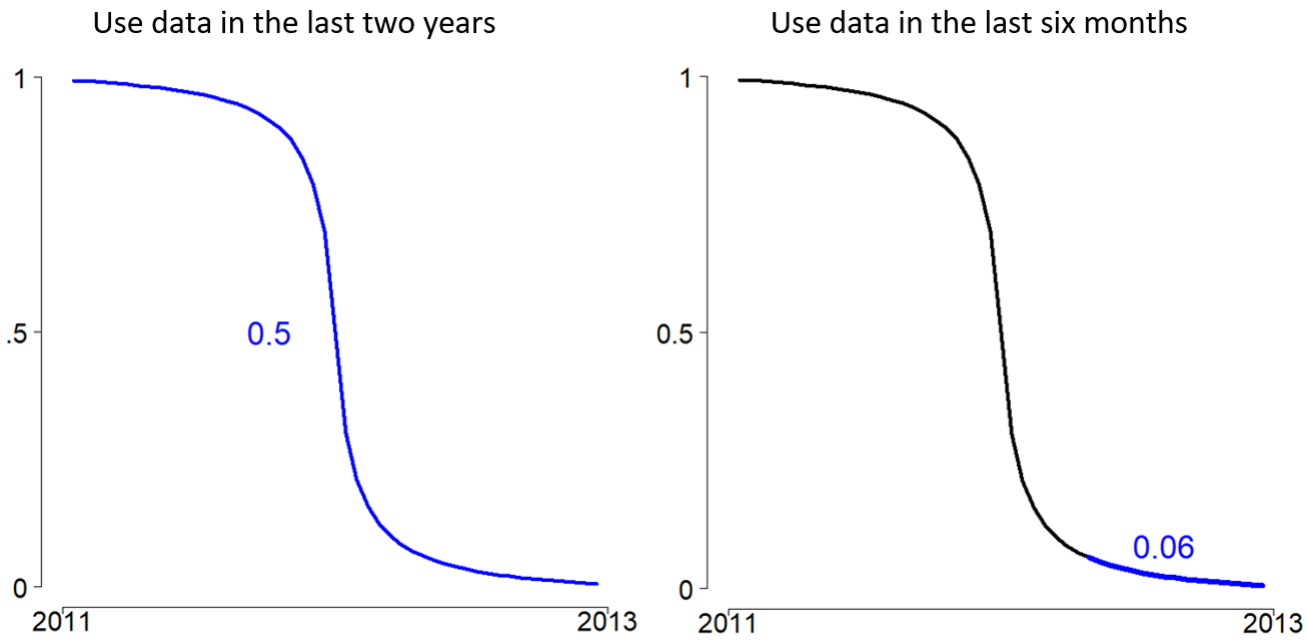


Figure 10: Coefficients Change Over Time: Default



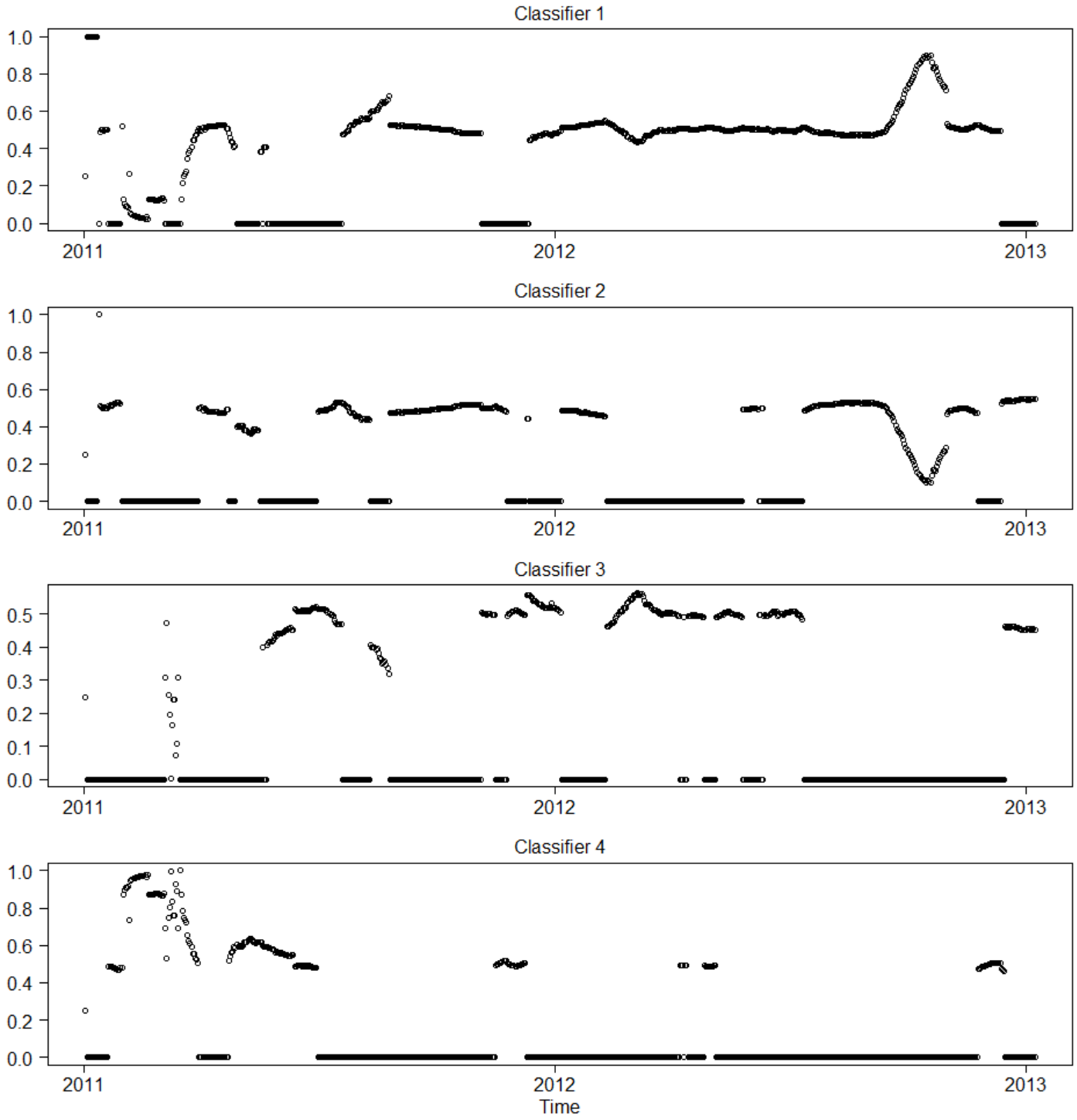
Notes: The reason that we see sudden changes in early 2011 is probably because we do not observe enough defaulted loans. Readers could refer to the coefficient changes after June 2011 for a better idea about the reduced form evidence. We will consider only using data after June 2011 in the next version.

Figure 11: Average Default Rate of Borrowers from California



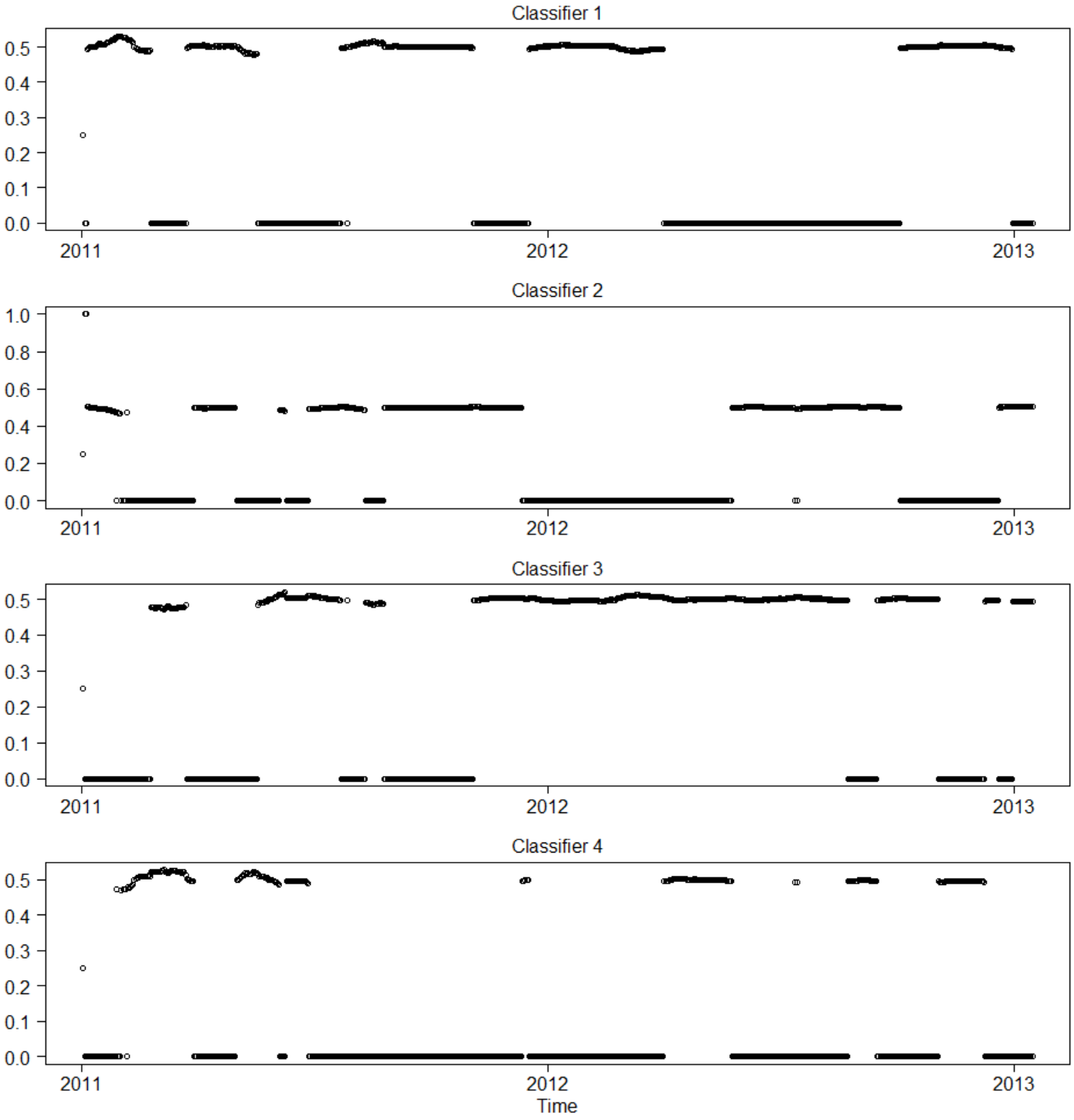
Notes: In this example, the average default rate for borrowers from California is 50% in the last two years, while only 6% of borrowers from California default in the last 6 months.

Figure 12: Weight of Each Individual Model in E-RPM on Funding Probability Prediction



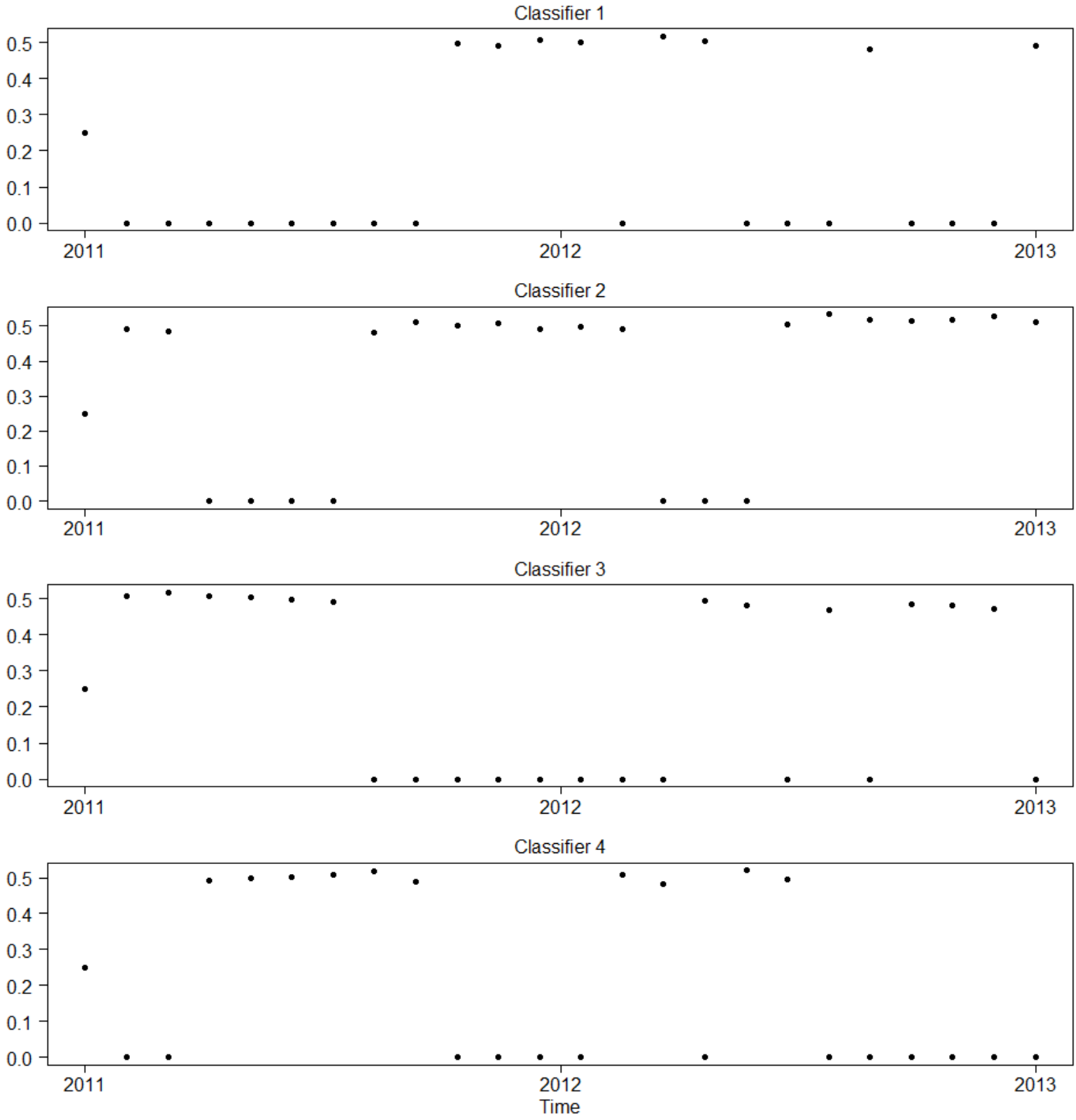
Notes: This figure shows the weights E-RPM puts on each individual model on the funding probability predictions. The weights are updated every day.

Figure 13: Weight of Each Individual Model in E-RPM on Withdrawal Probability Prediction



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

Figure 14: Weight of Each Individual Model in E-RPM on Default Probability Prediction



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

Figure 15: GFM vs. E-RPM

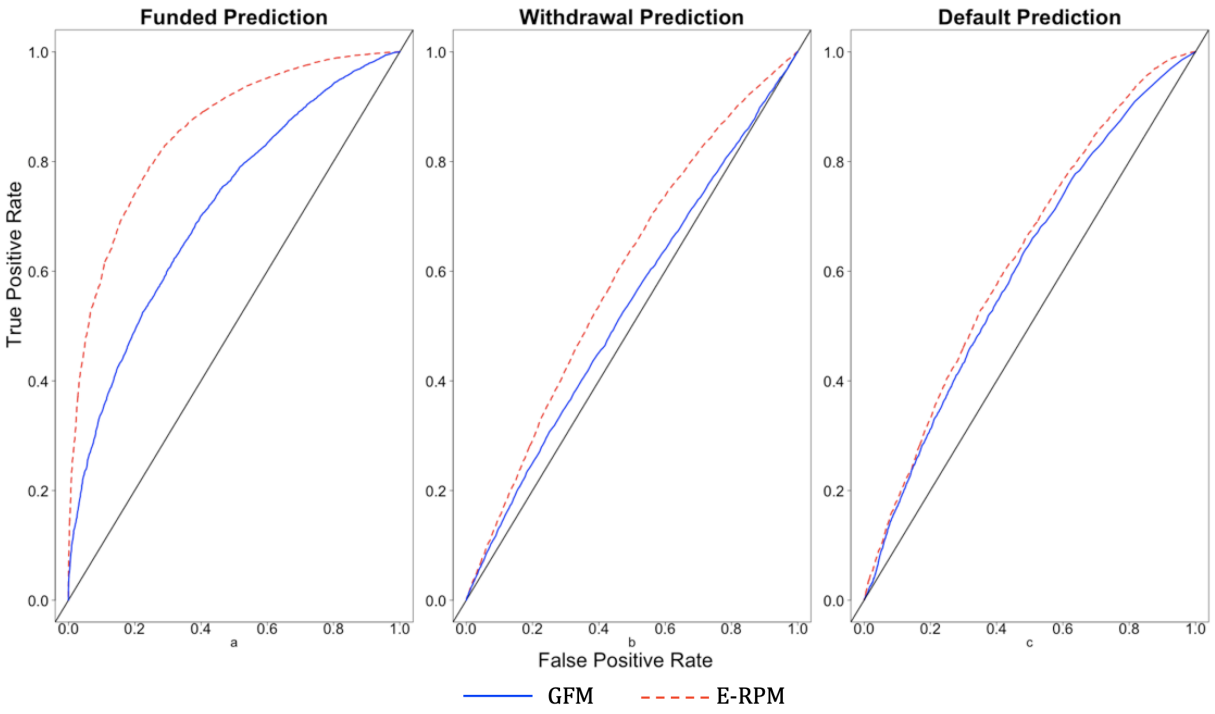


Figure 16: MWM vs. E-RPM

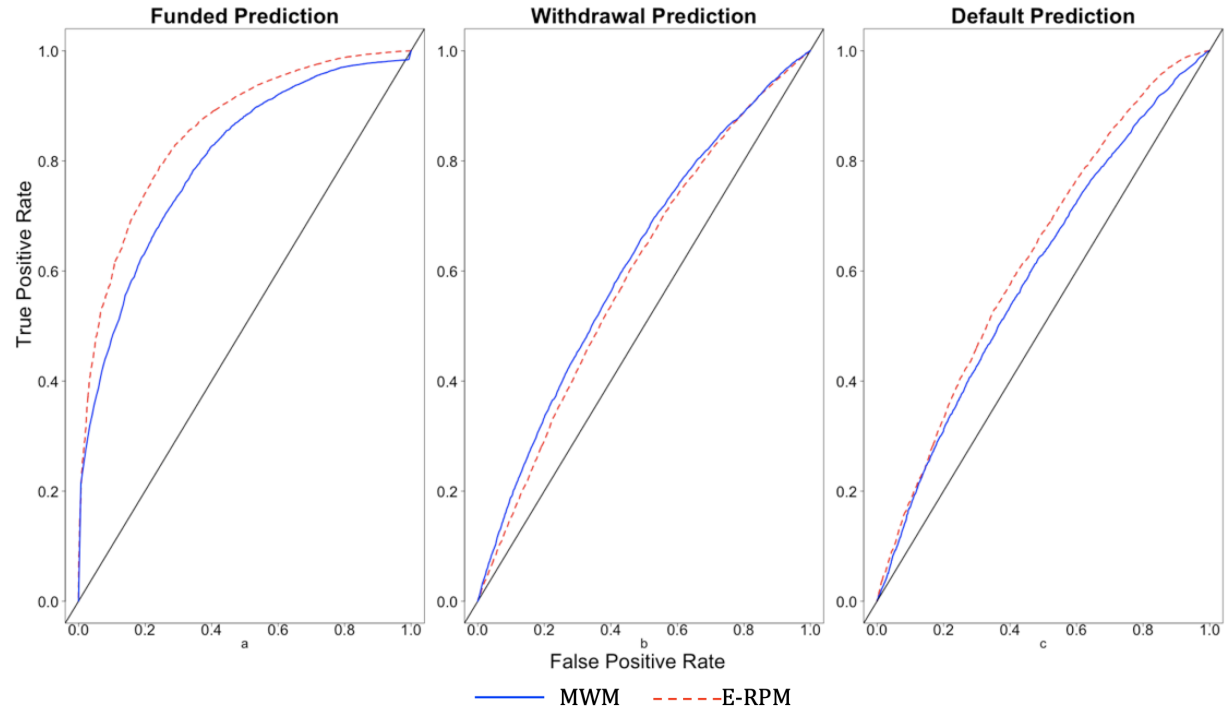


Figure 17: RPM vs. E-RPM

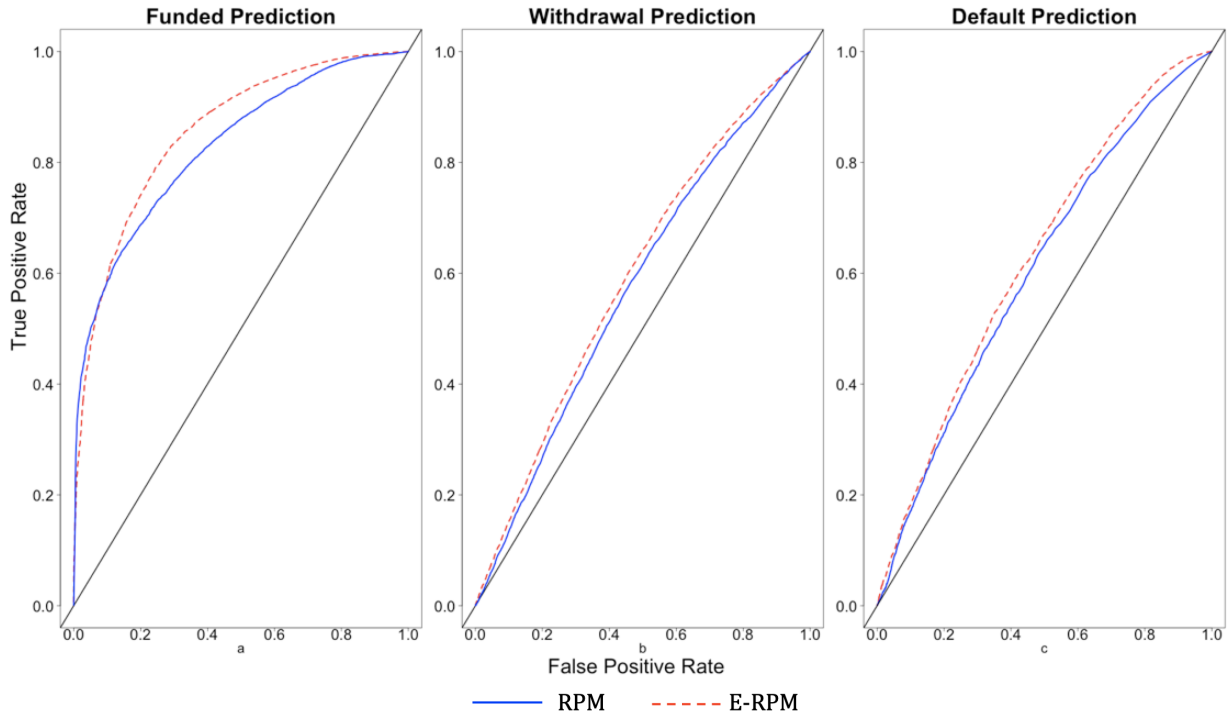
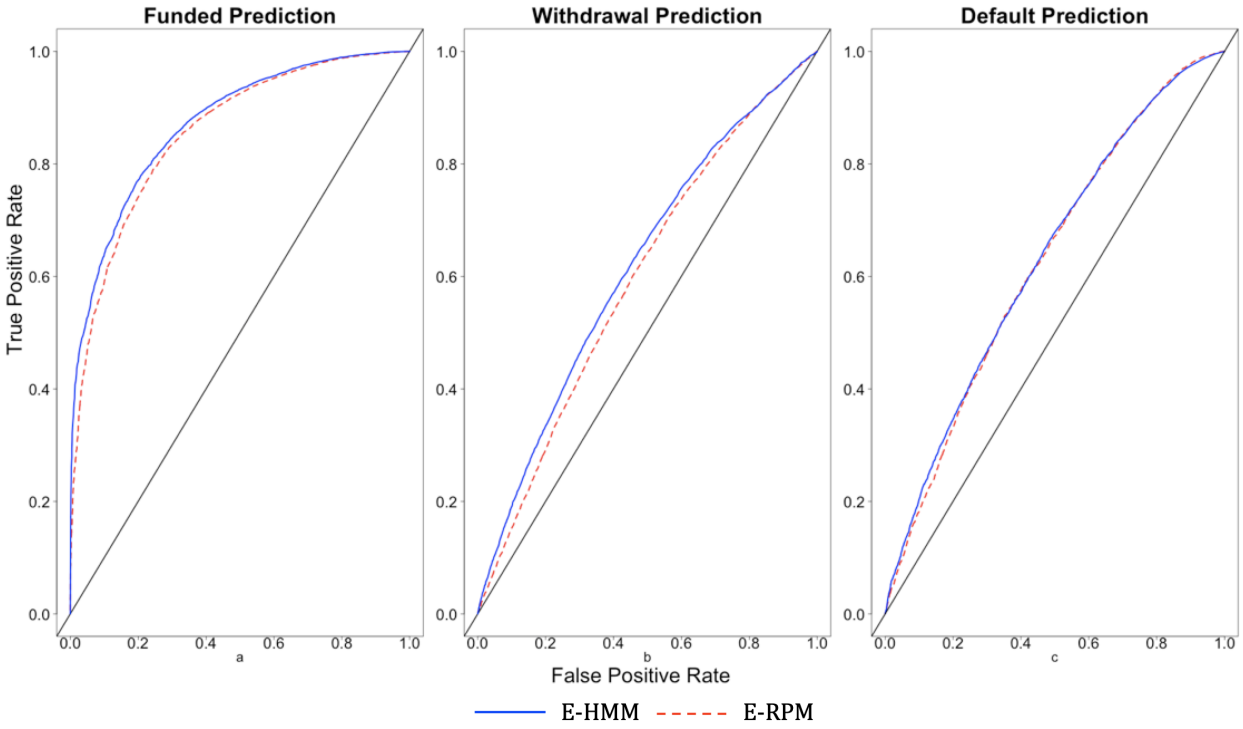


Figure 18: E-HMM vs. E-RPM



Appendix

A Estimation

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 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 dataset we have is $\{X, C\} = \{X_i, c_i\}_{i=1}^n$. c_i is the class label which takes value 0 or 1. Under the logistic regression model, the log-likelihood function is:

$$\log L(\beta|X, C) = \sum_{i=1}^{n} f_i(\beta^T X_i) \tag{18}$$

where

$$f_i(\beta^T X_i) = \begin{cases} \log \Phi(\beta^T X_i), & \text{if } c_i = 1 \\ \log(1 - \Phi(\beta^T X_i)) & \text{o.w.} \end{cases} \tag{19}$$

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

$$\log L(\beta|X, C) \approx \frac{1}{2}\beta^T \Psi(\widetilde{\beta})\beta - \beta^T \theta(\widetilde{\beta}) \tag{20}$$

where

$$\Psi(\widetilde{\beta}) = \sum_{i=1}^{n} a(\widetilde{\beta}^T X_i) X_i X_i^T, \quad \theta(\widetilde{\beta}) = \sum_{i=1}^{n} b(\widetilde{\beta}^T X_i, c_i) X_i \tag{21}$$

and

$$a(\widetilde{\beta}^T X_i) = -\Phi(\widetilde{\beta}^T X_i)(1 - \Phi(\widetilde{\beta}^T X_i))), \quad b(\widetilde{\beta}^T, c_i) = \Phi(\widetilde{\beta}^T X_i) - c_i + \widetilde{\beta}^T X_i a(\widetilde{\beta}^T X_i) \tag{22}$$

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

$$\widetilde{\beta}^* = \underset{\beta}{\operatorname{argmax}} \left(\frac{1}{2} \beta^T \Psi(\widetilde{\beta}) \beta - \beta^T \theta(\widetilde{\beta}) \right) \tag{23}$$

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}, c_{n+1}), our new estimate of β is updated as follows:

$$\widehat{\Psi}_{n+1} = \widehat{\Psi}_n + a(\widehat{\beta}_n^T X_{n+1}) X_{n+1} X_{n+1}^T \tag{24}$$

$$\widehat{\theta}_{n+1} = \widehat{\theta}_n + b(\widehat{\beta}_n^T X_{n+1}, c_{n+1}) X_{n+1} \tag{25}$$

$$\widehat{\beta}_{n+1} = \widehat{\Psi}_{n+1}^{-1} \widehat{\theta}_{n+1} \tag{26}$$

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:

$$\widehat{\Psi}_{n+1} = \lambda \widehat{\Psi}_n + a(\widehat{\beta}_n^T X_{n+1}) X_{n+1} X_{n+1}^T \tag{27}$$

$$\widehat{\theta}_{n+1} = \lambda \widehat{\theta}_n + b(\widehat{\beta}_n^T X_{n+1}, c_{n+1}) X_{n+1} \tag{28}$$

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. In our model, each period is defined as a day.

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

B 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.

Table A1: Summary Stats of Recession Probability (%)

Mean	Median	SD	Max	Min
0.358	0.320	0.097	0.620	0.260

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

Figure A1: Recession Probability Over Time (%)