Does Overconfidence Affect Financial Behaviors? Evidence from Retirement Readiness, Precautionary Savings, and Financial Market Participation

Minghao Yang[*]

June 2, 2020

Abstract

This paper investigates whether overconfidence in financial literacy affects financial behaviors of households with similar true financial literacy. To measure overconfidence of households, I utilize the National Financial Capability Study (NFCS) data to train six machine learning classifiers which learn from demographic characteristics, perceived financial literacy, and true financial literacy of households. Support Vector Machine (SVM) and random forest classifier outperform the others in terms of cross-validation mean squared errors (MSE). Using the probability of overconfidence predicted by SVM and random forest, I find that overconfidence in financial literacy has significantly positive effects on retirement readiness, precautionary savings, and financial market participation. Moreover, the effects are more evident in households with low level of true financial literacy.

keywords: household finance, overconfidence, machine learning, financial literacy.

JEL classification: G53, D90

[*] MA Program of Computational Social Science, University of Chicago

1 Introduction

Households do not always make wise financial decisions. No more than one third of US households plan for their retirement, no more than one half of US households set aside rainy funds, and no more than one third of US households participate in financial markets. These echoes the "investment mistakes" mentioned in Campbell (2006). Traditional economic literature (Lusardi and Mitchell, 2011b) points out that the lack of financial literacy contributes to the mistakes. Meanwhile, behavioral economics and psychology literature suggests that people tend to be overconfident (De Bondt and Thaler, 1995, and Brenner et al., 1996). The overconfidence bias can lead to persistent mistakes because households rely heavily on heuristics to make decisions but the feedbacks of their heuristics are lagged and incomplete. Although a large quantity of studies have empirically test the effect of financial literacy (see Lusardi and Mitchell, 2014 for a summary), limited literature has investigated whether overconfidence matters and whether the effects are positive or negative. In this paper I interrogate how overconfidence in financial literacy affect households' financial behaviors through the lens of retirement readiness, precautionary savings, and financial market participation. Moreover, I further investigate whether the effects are heterogeneous among households with different levels of true financial literacy.

The data of this paper comes from the 2012, 2015, and 2018 National Financial Capability Study (NFCS). The NFCS directly asks questions regarding retirement readiness, precautionary savings, and financial market participation. It also provides a rich set of demographic characteristics. In terms of financial literacy, it asks the households to self assess their math capability and financial knowledge, which could be used to measure perceived literacy. It also contains the "Big Five" questions defined in Lusardi and Mitchell (2017) regarding interest rate, inflation, risk diversification, mortgage payment, and bond price, which could be used to measure true literacy. The answers along with the demographic characteristics enable me to measure overconfidence using different machine learning classifiers. Besides, I also construct a measure for true literacy from the "Big Five" questions using factor analysis so that the effects of overconfidence could be examined true literacy being controlled.

The results show that overconfidence in financial literacy contributes to households' retirement readiness, precautionary savings, and financial market participation, with all else being equal, especially the true financial literacy. In addition, the effects are more evident in households with low level of true literacy. This might suggest that overconfident households do have previous exposures to finance, which helps them make better decisions and enhances their confidence in financial literacy, but the exposures are far from enough such that it is impossible for them to answer the “Big Five” questions correctly.

The potential contributions of this paper include using machine learning classifiers to construct overconfidence measures from survey data, and embedding behavioral economics in household finance. Feature-based machine learning classifiers are widely used in multiple disciplines to yield reliable predictions, including vision identification and content analysis. However, they are not commonly used in the field of economics. This paper applies the classifiers to build measures for overconfidence, which has never been done before. Behavioral economics provides fruitful insights into overconfidence and suboptimal choices, but they are rarely applied in household finance. I attempt to fill this gap by exploring the effects of overconfidence on retirement readiness, precautionary savings, and financial market participation. There are also two limitations. Firstly, overconfidence is not quite common among households, leaving me an unbalanced classification of overconfident households in the learning set. In addition, the learning set is relatively small compared with the whole dataset. These might lead to a relatively low accuracy of the out-of-sample predictions, and thus threaten the main conclusions of this paper. Secondly, the standard errors are not clustered at household level since the NFCS does not track their observations, so the results might not be robust.

The following sections are organized as follow: section 2 provides a thorough literature review; section 3 describes the data in detail; section 4 introduces the machine learning classifiers, the factor analysis technique, as well as the regression model that is used throughout the paper; section 5 summarizes the results from machine learning, factor analysis, and regressions; section 6 concludes.

2 Literature Review

2.1 Demographic Characteristics that Affect Financial Literacy and Overconfidence

To train machine learning classifiers, it is important to establish the feature space, i.e. to determine which features should be used to train the classifiers. Given that overconfidence in financial literacy is a new concept that consists of both “overconfidence” and “financial literacy”, it is beneficial to examine what demographic characteristics might affect each of them.

A large amount of literature has worked on disaggregating financial literacy. Lusardi and

Mitchell (2014) build a life cycle model with endogenous financial literacy investment and find that young and old people tend to have low financial literacy. This finding is verified in Agarwal et al. (2009), who employ data from Health and Retirement Study (HRS) and find that the youth and the old are more likely to make suboptimal decisions. Hsu (2016) constructs a life cycle model with marriage decision to explain why women tend to be less financially literate, which is supported by Hung et al. (2009), Lusardi et al. (2010) and Lusardi et al. (2014) among both the youth and the old using data from different national surveys. Education also has an impact on financial literacy. Lusardi and Mitchell (2007a) use HRS data and find that high education is correlated with high financial literacy. The questions in HRS regarding financial literacy are also implemented in other surveys all over the world gradually. Lusardi and Mitchell (2011b) make use of these surveys and find the same pattern. However, they also emphasize that education alone is not a good proxy for financial literacy. Several studies also suggest that race (Lusardi and Mitchell, 2007a, Lusardi and Mitchell, 2007b, and Lusardi and Mitchell, 2011a) and residential region (Bumcrot et al., 2013, and Fornero and Monticone, 2011) should matter, but the effects might not be as solid as other factors.

Less literature examines the demography of overconfidence. Bhandari and Deaves (2006) use a survey with nearly 2,000 defined contribution pension plan members and conclude that the male and the well-educated are more likely to be overconfident. Lin (2011) conducts a questionnaire survey in Taiwan and suggests that the male, the youth, and the old are more vulnerable to overconfidence bias.

In sum, the factors that affect true financial literacy and overconfidence are similar, with age, gender, and education influencing both of them. Therefore, I treat all the characteristics mention above as the features used to predict overconfidence in financial literacy.

2.2 How Financial Literacy and Overconfidence Affect Financial Behaviors

Limited literature has shed light on how overconfidence in financial literacy affects financial behaviors. Hence, it is necessary to review the effects of financial literacy and overconfidence separately. To begin with, what is the effect of financial literacy on financial behaviors?

The lack of financial literacy makes it difficult for households to figure out their optimal choices. Hence, we may observe suboptimal behaviors. The life cycle model in Lusardi and Mitchell (2014) also implies that individuals with low financial literacy choose to receive lower returns and thus accumulate less wealth during their life. There are also empirical evidences that support this argument. Calvet et al. (2007) and Calvet et al. (2009) make use of Swedish households data to examine the relationship between financial literacy and investment mistakes. Although they do not measure financial literacy directly, they find that proxies such as wealth, income, education, and immigration status are associated with mistakes. Lusardi and Mitchell (2007b), Lusardi and Mitchell (2011a), and Lusardi and Mitchell (2017) formally construct a measure of financial literacy according to different surveys and find that financial literacy is positive related to retirement readiness.

On the other hand, how will overconfidence affect financial behaviors? People tend to be overconfident in their ability (Brenner et al., 1996), which leads to misperception of optimum. Odean (1998) examines overconfidence bias in stock markets. Under an extensive form game with traders, insiders, and marketmakers, he finds that overconfident players tend to receive lower returns due to their excess trades. Empirically, Barber and Odean (2001) use gender as a proxy of overconfidence and find that male traders trade more and thus receive less. With financial markets becoming more accessible in this time and age, households are expected to be affected by their overconfidence as well.

The reason why few previous studies investigate the effect of overconfidence in household finance, even though the data corresponding to both perceived and true financial literacy are available in several national surveys, is that the questions measuring perceived and true financial literacy are not counterparts to each other, so a direct measure is hard to construct. The only study to my knowledge that tries to combine overconfidence with household finance is Anderson et al. (2017). They conduct a survey on LinkedIn to capture true financial literacy and perceived financial literacy so that they become comparable. Their findings suggest that households tend to overestimate their financial literacy and both true and perceived financial literacy affect their financial behaviors. Nevertheless, they do not directly examine the effect of overconfidence, and the external validity might be weak given that LinkedIn data could not represent the national population well.

I will solve the problem of unmatched questions by using feature based machine learning classifiers to construct the overconfidence measures. Hence, it then becomes feasible to empirically examine the effects of overconfidence.

3 Data

The data of this paper come from the National Financial Capability Studies (NFCS). I use the 2018 tracking dataset which contains all the observations from the 2009, 2012, 2015, and 2018 study.

I drop the observations in the 2009 study because the education data cannot be merged. After that, there are 80,164 observations left in the final dataset. The following sections describe the key variables of financial behaviors, demographic characteristics, perceived financial literacy, and true financial literacy.

3.1 Financial Behaviors

The NFCS covers a series of questions regarding financial behaviors of households. This paper makes uses of the questions related to retirement readiness, precautionary savings, and financial market participation. The questions are as follow:

- Have you ever tried (Before you retired, did you try) to figure out how much you need(ed) to save for retirement?
- Have you set aside emergency or rainy day funds that would cover your expenses for 3 months, in case of sickness, job loss, economic downturn, or other emergencies?
- Not including retirement accounts, do you have any investments in stocks, bonds, mutual funds, or other securities?

The questions are all “Yes or No” type, while ”Don’t know (DK)” and ”Prefer not to say (R)” are allowed. For each question, I construct an indicator which equals one if the answer is “Yes”, so that the three indicators could reflect whether a household is ready for retirement, has precautionary savings, or participates in the financial market. The weighted summary statistics are provided in Panel A of Table 1. It seems that households do not always make wise decisions. To be concrete, only 30.9% of them plan for retirement, only 44.9% of them set aside emergency funds, and only 31.4% of them invest in financial markets.

3.2 Demographic Characteristics

The NFCS also provides a rich set of demographic characteristics. For this paper I mainly use age, gender, race, income, education, marital status, and residential state data. For age and income, the NFCS only gives a range for each household, so I use the group mean as the imputed age or income. For education, I construct two dummies representing high school graduates and college graduates. For residential state, I create 51 dummies representing the 51 states included in the NFCS. Panel B of Table 1 provides the weighted summary statistics for demographic characteristics.

Table 1: Summary statistics: Financial behaviors and demographic characteristics

Panel A: Financial behaviors

Variables	10^{th} pct	Median	90^{th} pct	Mean	S.D.	#Obs.
Retirement Readiness	0	0	1	0.309	0.462	80164
Precautionary Savings	0	0	1	0.449	0.497	80164
Financial Market Participation	0	0	1	0.314	0.464	80164

Notes: The variables are extracted from the 2012, 2015, and 2018 NFCS. The sample weights in the NFCS are used to calculate the statistics.

Panel B: Demographic characteristics

Variables	10^{th} pct	Median	90^{th} pct	Mean	S.D.	#Obs.
Female	0	1	1	0.514	0.500	80164
Age	20	50	70	46.34	16.52	80164
Nonwhite	0	0	1	0.350	0.477	80164
Married	0	1	1	0.523	0.499	80164
Income	7500	42500	125000	62054.3	49231.7	80164
High School	1	1	1	0.954	0.210	80164
College	0	0	1	0.355	0.479	80164

Notes: The variables are extracted from the 2012, 2015, and 2018 NFCS. The sample weights in the NFCS are used to calculate the statistics.

3.3 Perceived Financial Literacy

The NFCS asks the following two questions to capture perceived financial literacy of each household:

- How strongly do you agree or disagree with the following statements? - I am pretty good at math.

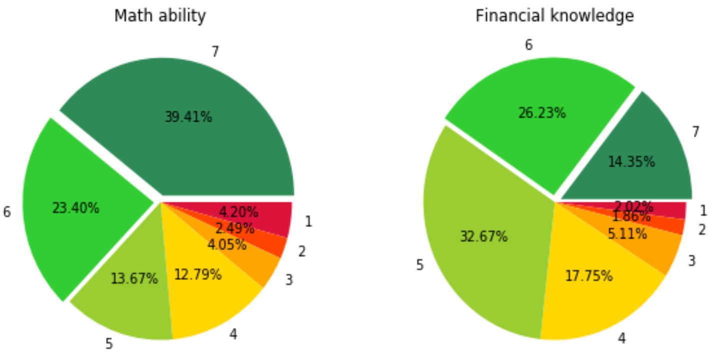
- How would you assess your overall financial knowledge?

These two questions indicate math capability and self-assessed financial knowledge of each households, which are two important aspects of financial literacy. The answers are scaled from 1 to 7, where 1 means strongly disagree or very low and 7 means strongly agree or very high. Figure 1 summarizes the weighted answers of the households. More than 60% households choose 6 or 7 for the math question and more than 40% households choose 6 or 7 for the financial knowledge question. Hence, households are generally confident in their financial literacy.

3.4 True Financial Literacy

Lusardi and Mitchell (2014) summarizes the "Big Three" questions regarding interest rate, inflation

Figure 1: Summary statistics: Perceived financial literacy



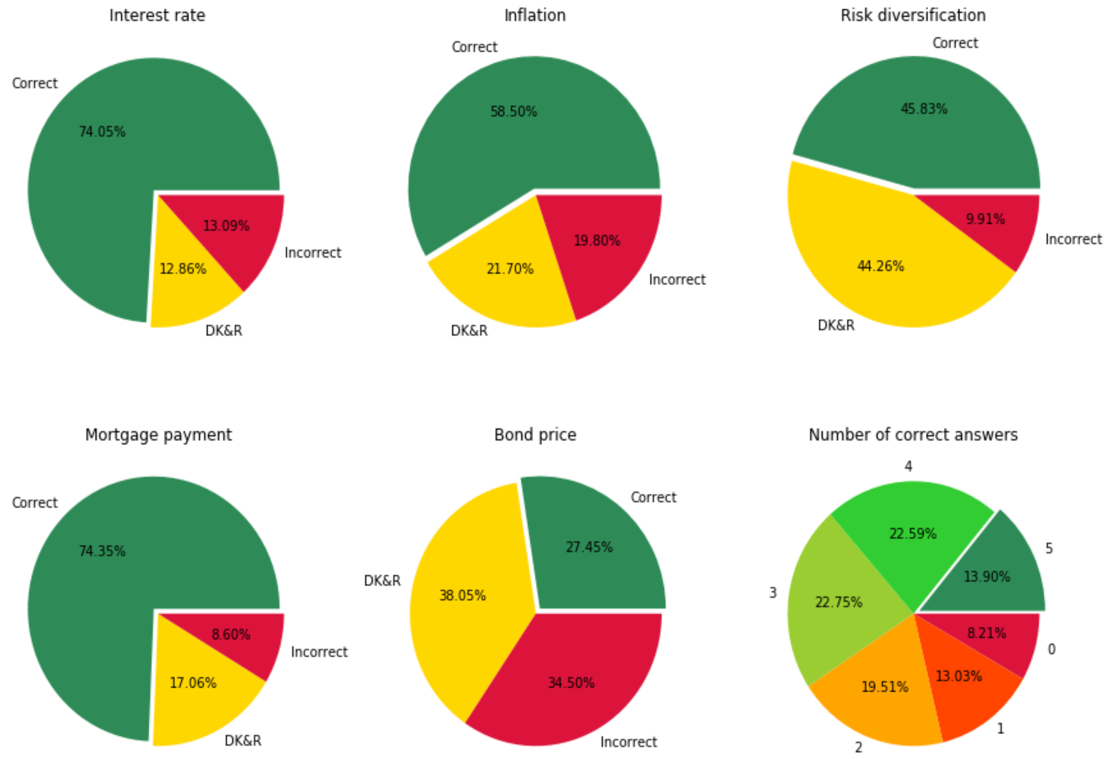
and risk diversification. Other than the "Big Three" questions, the NFCS further asks two questions regarding mortgage payment and bond price. These five questions can reflect the true financial literacy of households and are referred to as the "Big Five" questions (Lusardi and Mitchell, 2017). The households are allowed to reply "Don't know (DK)" or "Prefer not to say (R)" when they answer the questions. The questions are summarized in Table 2, with the correct answers in bold. Figure 2 shows the weighted proportion of correct, don't know or prefer not to say, and incorrect responses of each question, as well as the weighted proportion of households who give different numbers of correct answers. It is surprising that only 13.59% households could answer all the questions correctly, compared with the high proportion of households who are confident in their financial literacy.

Table 2: Questions regarding true financial literacy

Q1: Suppose you had $100 in a savings account and the interest rate was 2% per year. After 5 years, how much do you think you would have in the account if you left the money to grow?		
More than $102	Exactly $102	Less than $102
Q2: Imagine that the interest rate on your savings account was 1% per year and inflation was 2% per year. After 1 year, how much would you be able to buy with the money in this account?		
More than today	Exactly the same	**Less than today**
Q3: Buying a single company's stock usually provides a safer return than a stock mutual fund.		
True	**False**	
Q4: A 15-year mortgage typically requires higher monthly payments than a 30-year mortgage, but the total interest paid over the life of the loan will be less.		
True	False	
Q5: If interest rates rise, what will typically happen to bond prices?		
They will rise	**They will fall**	They will stay the same

Notes: The questions are from the 2012, 2015, and 2018 NFCS Questionnaires.

Figure 2: Summary statistics: True financial literacy



4 Methods

4.1 Constructing Overconfidence Measures

To train the machine learning classifiers, I first construct a learning set where the households can be unambiguously defined as overconfident or not. To be specific, the households who give incorrect answers to all the “Big Five” questions but choose six or seven in the two self-assessed questions are hard coded as overconfident. On the other hand, the households who give correct answers to all the “Big Five” questions and choose six or seven in the two self-assessed questions, as well as the households who give incorrect answers to all the “ Big Five” questions and choose one or two in the two self-assessed questions are hard coded as not overconfident. The above coding rule yields 858 overconfident households and 7,506 not overconfident households. Then I use demographic characteristics, answers to the two self-assessed questions, and answers to the “Big Five” questions as the inputs of the classifiers. After fitting all the classifiers with the optimal parameters from randomized search cross validation, the mean squared errors (MSE) are calculated for model selection. In the end, I get the out-of-sample predictions of the remaining 71,800 observations for the outstanding classifiers as the overconfidence measures used in the main analyses. Rather than use the predicted class, I use the predicted probability as the overconfidence measure so that it is continuous on range $[0, 1]$. I train six classifiers, namely logistic regression classifier, random forest classifier, Support Vector Machine (SVM) classifier, Naive Bayes (NB) classifier, K-nearest Neighbors (KNN) classifier, and Multi-layer Perception (MLP) classifier. The following sections describe how I train the random forest and SVM classifier in detail because they yield the smallest MSEs (see section 5.1). The description of other classifiers are shown in Appendix A.1.

4.1.1 Support Vector Machine Classifier

The binary Support Vector Machine (SVM) classifier divides the observations into two classes by constructing a surface in the feature space such that the margin (the distance between the surface and the closest data point) is maximized. Mathematically, the SVM classifier solves the following problem:

$$\begin{aligned} \max_{\beta_1,\beta_2,...\beta_P,M} \quad & M \\ \text{s.t} \quad & \sum_{j=1}^{P} \beta_j^2 = 1 \\ & y_i \left(\beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + ... + \beta_P X_{P,i}\right) \geq M \quad \forall i = 1, ...N \end{aligned} \tag{1}$$

where M is the maximized margin, y_i is the overconfidence indicator, and $X_{p,i}$ is the pth feature of household i. Empirically, I tune the regulation parameter C in randomized search cross validation. I utilize the RBF kernel to run the model because of its high computational speed.

4.1.2 Random Forest Classifier

The building block of any random forest classifier is the decision tree classifier. It divides the feature space into multiple subspaces and uses the mean value of each subspace as the prediction for that space. Mathematically, let R denotes the feature space. The decision tree classifier divides it into J subspaces. Then for household i whose features are in space R_j, the probability of overconfidence is given by

$$\Pr(y_i = 1 | \mathbf{X_i} \in R_j) = \frac{1}{\# R_j} \sum_{\mathbf{X_j} \in R_j} y_j \tag{2}$$

where y_i is an indicator of overconfidence for household i, and $\mathbf{X_i}$ is the feature matrix of household i. The random forest classifier is an ensemble method that combines multiple decision trees by bootstrap aggregation and feature randomness. Therefore, the prediction would become more robust than any single tree. For this paper each decision tree takes random features from feature space and make a prediction on whether a household is overconfident. After that, the random forest classifiers select the prediction that most trees give as the final output. The number of trees, the maximum depth of each tree, the minimum number of samples required to split an internal node, the minimum number of samples required to be at a leaf node, and the number of features to consider when looking for the best split are tuned in randomized search cross validation.

4.2 Constructing True Financial Literacy Measure

The measure is constructed following Lusardi and Mitchell (2017). In order to combine the answers of the "Big Five" questions into a single measure, I generate an indicator for each question which equals one if the household answers the question correctly. After that, I perform a factor analysis on these five indicators using the principal component factor method so that questions accounting for higher variance are assigned with higher weights. The factor score is then calculated and normalized to range $[0, 1]$ in order to match the scale of the overconfidence measures. The normalized factor score is used as the measure for true financial literacy. The factor loads and uniqueness as well as the summary statistics are provided in Appendix A.3.

4.3 Investigating the Effect of Overconfidence

To see whether overconfidence in financial literacy influence financial behaviors of households, especially those with similar true financial literacy, I run the following logit regression:

$$\begin{aligned} &\Pr(y_i = 1 | \mathrm{Overconfidence_i}, \mathrm{True\_Literacy_i}, \mathbf{X}_\mathrm{i}^\mathrm{D}, \mathbf{X}_\mathrm{i}^\mathrm{YS}, \varepsilon_i, \beta_0, \beta_1, \beta_2, \beta_3, \beta_4) \\ &= F(\beta_0 + \beta_1 \mathrm{Overconfidence_i} + \beta_2 \mathrm{True\_Literacy_i} + \mathbf{X}_\mathrm{i}^\mathrm{D} \beta_3 + \mathbf{X}_\mathrm{i}^\mathrm{YS} \beta_4 + \varepsilon_i) \end{aligned} \tag{3}$$

where $F(x) = e^x/(1 + e^x)$. y_i represents the indicators for retirement readiness, precautionary savings, and financial market participation; $\mathrm{Overconfidence_i}$ stands for the overconfidence measures generated by machine learning classifiers; $\mathrm{True\_Literacy_i}$ is the true financial literacy measure given by factor analysis; $\mathbf{X}_\mathrm{i}^\mathrm{D}$ is a matrix of demographic characteristics including age, age squared, log income, log income squared, gender, race, marital status, and education; $\mathbf{X}_\mathrm{i}^\mathrm{YS}$ is a matrix of state and year dummies.

5 Results

5.1 Overconfidence Measures from Different Classifiers

The MSEs of different classifiers are presented in Figure 3. It demonstrates that the SVM and random forest classifiers have much smaller MSE. I also plot the confusion matrices of all the classifiers. The confusion matrices of SVM and random forest classifiers are presented in Figure 4(a), and those of other classifiers are presented in Figure 4(b). Figure 4 suggests that SVM and random forest classifiers outperform the others because they can predict the true overconfident households better given the highly unbalanced classification in the learning set. The out-of-sample predictions of SVM and random forest for the remaining 71,800 observations are used in the main analyses given their high accuracy. Table 3 presents the weighted summary statistics for the two overconfidence measures over the full sample. The summary statistics for other measures could be found in Appendix A.2. The overconfidence measure from SVM has both a larger mean and a larger variance compared with that from random forest.

Figure 3: MSEs of different classifiers

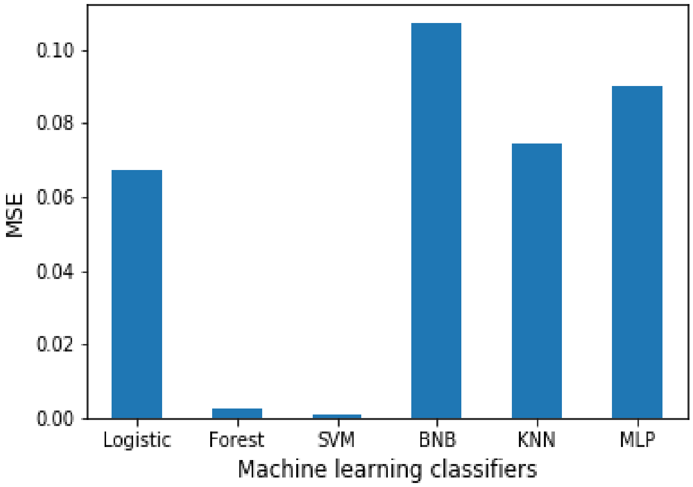
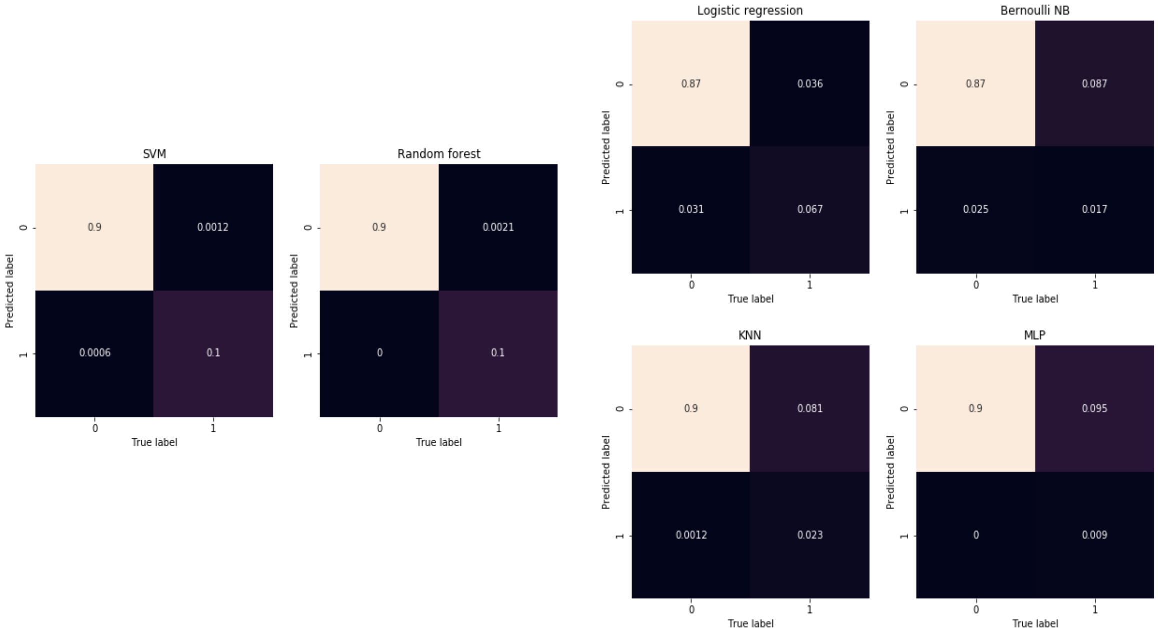


Table 3: Summary statistics: Overconfidence measures from SVM and forest

Classifiers	10^{th} pct	Median	90^{th} pct	Mean	S.D.	#Obs.
SVM	4.37e-05	0.133	1.000	0.392	0.426	80164
Forest	0.029	0.203	0.455	0.234	0.170	80164

Notes: The overconfidence measures are predicted by SVM and random forest classifiers as described in section 4.1. The sample weights in the NFCS are used to calculate the statistics.

Figure 4: Confusion matrices of different classifiers



(a) SVM and random forest classifiers

(b) Other classifiers

5.2 The Effects of Overconfidence on Financial Behaviors

In this section I only use the overconfidence measures given by SVM and random forest to run regression (3). The regression results with other overconfidence measures given by different classifiers are provided in Appendix A.4.

5.2.1 Baseline Results

Table 4 shows the results from regression (3) using the SVM based overconfidence measure, where the coefficients are the *average marginal effects* rather than the *log odds ratios*. Demographic characteristics and year dummies are controlled in all columns, while column (2), (4), and (6) further control state dummies. In column (1) and (2), the dependent variable is the indicator for retirement readiness; in column (3) and (4), the dependent variable is the indicator for precautionary savings; in column (5) and (6), the dependent variable is the indicator for financial market participation. The average marginal effects of overconfidence on retirement readiness, precautionary savings, and financial market participation are all significantly positive even if the true financial literacy is controlled, which suggests that overconfidence does affect financial behaviors of households. Concretely, a standard deviation increase (0.43 in Table 3) in overconfidence will promote the probability of retirement readiness by around 6.3%, promote the probability of precautionary savings by around 7.1%, and promote the probability of financial market participation by around 6.3%, with all else being equal. The increases in probabilities are quite decent given that only 30.9% households prepare for retirement, only 44.9% households have precautionary savings, and only 31.4% households participate in financial market (see Table 1). It is not surprising that the average marginal effects of true financial literacy are also significantly positive, indicating that a higher level of true financial literacy contributes to a higher probability of retirement readiness, precautionary savings, and financial market participation. This echoes the findings in Lusardi and Mitchell (2017).

Table 5 shows the results from regression (3) using the random forest based overconfidence measure. The setting of Table 5 is exactly the same as Table 4. The average marginal effects of overconfidence are still positive, but with larger scales. This is because the standard deviation of the random forest based overconfidence measure is much smaller. Again from standard deviation perspective, a standard deviation increase (0.17 in Table 3) in overconfidence will promote the probability of retirement readiness by around 7.7%, promote the probability of precautionary savings by around 8.0%, and promote the probability of financial market participation by around 7.9%, with all else being equal. Hence, the effects of overconfidence on financial behaviors become even stronger. On the other hand, true financial literacy still has a positive effect on retirement readiness, precautionary savings, and financial market participation.

5.2.2 Heterogeneous Effects of Overconfidence

It is of great interest to interrogate whether overconfidence in financial literacy has heterogeneous effects on households with different levels of true literacy, especially those with low and high literacy. To answer the above question, I restrict my sample to the households whose true financial literacy is either 0 or 1. The former subsample stands for households with low literacy and the latter stands for high literacy. Then I rerun the regression (3) with these two subsamples and compare the coefficients of overconfidence.

Table 6 presents the results using the SVM based overconfidence measure. The significantly positive coefficients of overconfidence only remain in the low literacy subsample, while they become insignificant in the high literacy subsample. Hence, overconfidence in financial literacy only benefits the households with low level of true financial literacy. Table 7 displays the results using the random forest based overconfidence measure. Again only the coefficients from low literacy subsample are significantly positive. Moreover, those from high literacy subsample are significantly negative, suggesting overconfidence might even lead to suboptimal decisions.

The above findings suggest that overconfidence only benefits households with low level of true financial literacy. A potential explanation would be that these households do learn something about finance from their friends or finance advisors, or simply from televisions, newspapers, or the Internet. The knowledge not only helps them make better decisions in terms of retirement readiness, precautionary savings, and financial market participation, but also enhances their confidence in financial literacy. However, they did not receive any formal educations in finance, so they could not answer the "Big Five" questions correctly. Hence, these households are overconfident but behave somewhat more close to the optimum. On the other hand, most households with high level of true financial literacy received formal educations in finance, so overconfidence becomes unimportant and even harmful.

6 Conclusions and Discussions

The results of my paper shows that overconfidence in financial literacy has significantly positive effects on households' financial behaviors. Households who are more overconfident are also more likely to prepare for their retirement, set aside precautionary savings, and participate in the financial market. The magnitudes are economically important. With all else being equal, a standard deviation increase in overconfidence will lead to a 6 - 8% increase in the probability of retirement readiness, precautionary savings, or financial market participation. Additionally, The effect is more evident in households with low level of true financial literacy, which might suggest that the overconfidence in financial literacy comes from their shallow exposures to materials related to finance.

The potential contributions of my work are constructing measures for overconfidence from survey data using machine learning classifiers, and connecting behavioral economics to household finance. Using data from National Finance Capability Studies (NFCS), I construct six measures for overconfidence with different machine learning classifiers and use the measures from SVM and random forest to investigate the effects of overconfidence given their better performance in predicting overconfident households. To my limited knowledge, no researchers have tried to measure overconfidence by machine learning, and few have ever explored whether overconfidence matters or not when households' make their decisions.

This paper also has several limitations. Firstly, the classification of overconfident and not overconfident households is highly unbalanced in my learning set. Moreover, the size of the learning set is not large enough. It is natural because only a small proportion of households can be clearly categorized as overconfident or not overconfident. However, the performance of the machine learning classifiers might be impaired, as logistic regression, Bernoulli NB, KNN, and MLP classifiers fail to predict overconfident households well. Future work includes choosing an appropriate class weight or trying other classifiers that are robust to unbalanced classes. Secondly, the standard errors in the regressions are not clustered at household level because the NFCS do not provide a tracking ID for each household included in multiple waves of surveys. Therefore, the robust standard errors might be much larger than the ones presented in the tables.

Table 4: Logit regression on overconfidence (SVM based) and true financial literacy

Dependent Variables	(1) Readiness	(2) Readiness	(3) Precaution	(4) Precaution	(5) Participation	(6) Participation
Overconfidence	0.146***	0.147***	0.164***	0.164***	0.145***	0.148***
	(0.00637)	(0.00637)	(0.00657)	(0.00656)	(0.00648)	(0.00647)
True Literacy	0.341***	0.339***	0.317***	0.319***	0.374***	0.375***
	(0.00993)	(0.00995)	(0.0101)	(0.0101)	(0.0102)	(0.0102)
Observations	80,164	80,164	80,164	80,164	80,164	80,164
Demo. chars.	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
State dummies	No	Yes	No	Yes	No	Yes
Pseudo R-squared	0.136	0.137	0.148	0.150	0.187	0.190

Notes: The results are from regression (3). Overconfidence measure is predicted by the SVM classifier. True financial literacy is calculated via factor analysis. Demographic characteristics and year dummies are controlled in all columns. Column (2), (4), and (6) further control state dummies. In column (1) and (2), the dependent variable is the indicator for retirement readiness; in column (3) and (4), the dependent variable is the indicator for precautionary savings; in column (5) and (6), the dependent variable is the indicator for financial market participation. Observations are weighted by the NFCS sample weights. Standard errors are in the parentheses. The symbols *, **, and *** denote significance at the 10%, 5% and 1% levels respectively.

Table 5: Logit regression on overconfidence (random forest based) and true financial literacy

Dependent Variables	(1) Readiness	(2) Readiness	(3) Precaution	(4) Precaution	(5) Participation	(6) Participation
Overconfidence	0.452***	0.455***	0.473***	0.466***	0.466***	0.470***
	(0.0246)	(0.0246)	(0.0268)	(0.0268)	(0.0240)	(0.0240)
True Literacy	0.422***	0.420***	0.390***	0.388***	0.459***	0.459***
	(0.0146)	(0.0147)	(0.0155)	(0.0155)	(0.0147)	(0.0147)
Observations	80,164	80,164	80,164	80,164	80,164	80,164
Demo. chars.	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
State dummies	No	Yes	No	Yes	No	Yes
Pseudo R-squared	0.133	0.134	0.145	0.146	0.185	0.187

Notes: The results are from regression (3). Overconfidence is predicted by the random forest classifier. True financial literacy is calculated via factor analysis. Demographic characteristics and year dummies are controlled in all columns. Column (2), (4), and (6) further control state dummies. In column (1) and (2), the dependent variable is the indicator for retirement readiness; in column (3) and (4), the dependent variable is the indicator for precautionary savings; in column (5) and (6), the dependent variable is the indicator for financial market participation. Observations are weighted by the NFCS sample weights. Standard errors are in the parentheses. The symbols *, **, and *** denote significance at the 10%, 5% and 1% levels respectively.

Table 6: Heterogeneous effects of overconfidence (SVM based)

Dependent Variables	(1) Readiness	(2) Readiness	(3) Precaution	(4) Precaution	(5) Participation	(6) Participation
Overconfidence	0.0891***	4.572	0.142***	-8.442	0.0554***	-0.973
	(0.0173)	(8.692)	(0.0193)	(11.09)	(0.0154)	(10.82)
Observations	5,886	12,539	5,886	12,539	5,886	12,539
Sample	Low Lit.	High Lit.	Low Lit.	High Lit.	Low Lit.	High Lit.
Demo. chars.	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
State dummies	Yes	Yes	Yes	Yes	Yes	Yes
Pseudo R-squared	0.107	0.166	0.0914	0.129	0.164	0.106

Notes: The results are from regression (3) with low and high literacy subsamples. Overconfidence measure is predicted by the SVM classifier. Demographic characteristics, year dummies, and state dummies are controlled in all columns. Column (1), (3), and (5) use the low literacy subsample, while column (2), (4), and (6) use the high literacy subsample. In column (1) and (2), the dependent variable is the indicator for retirement readiness; in column (3) and (4), the dependent variable is the indicator for precautionary savings; in column (5) and (6), the dependent variable is the indicator for financial market participation. Observations are weighted by the NFCS sample weights. Standard errors are in the parentheses. The symbols *, **, and *** denote significance at the 10%, 5% and 1% levels respectively.

Table 7: Heterogeneous effects of overconfidence (random forest based)

Dependent Variables	(1) Readiness	(2) Readiness	(3) Precaution	(4) Precaution	(5) Participation	(6) Participation
Overconfidence	0.224***	-1.895***	0.313***	-1.986***	0.146***	-2.041***
	(0.0216)	(0.263)	(0.0262)	(0.231)	(0.0177)	(0.258)
Observations	5,886	12,539	5,886	12,539	5,886	12,539
Sample	Low Lit.	High Lit.	Low Lit.	High Lit.	Low Lit.	High Lit.
Demo. chars.	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
State dummies	Yes	Yes	Yes	Yes	Yes	Yes
Pseudo R-squared	0.127	0.171	0.106	0.137	0.182	0.112

Notes: The results are from regression (3) with low and high literacy subsamples. Overconfidence measure is predicted by the random forest classifier. Demographic characteristics, year dummies, and state dummies are controlled in all columns. Column (1), (3), and (5) use the low literacy subsample, while column (2), (4), and (6) use the high literacy subsample. In column (1) and (2), the dependent variable is the indicator for retirement readiness; in column (3) and (4), the dependent variable is the indicator for precautionary savings; in column (5) and (6), the dependent variable is the indicator for financial market participation. Observations are weighted by the NFCS sample weights. Standard errors are in the parentheses. The symbols *, **, and *** denote significance at the 10%, 5% and 1% levels respectively.

References

Agarwal, Sumit, John C Driscoll, Xavier Gabaix, and David Laibson, "The age of reason: Financial decisions over the life cycle and implications for regulation," *Brookings Papers on Economic Activity*, 2009, *2009* (2), 51–117.

Anderson, Anders, Forest Baker, and David T Robinson, "Precautionary savings, retirement planning and misperceptions of financial literacy," *Journal of Financial Economics*, 2017, *126* (2), 383–398.

Barber, Brad M and Terrance Odean, "Boys will be boys: Gender, overconfidence, and common stock investment," *The quarterly journal of economics*, 2001, *116* (1), 261–292.

Bhandari, Gokul and Richard Deaves, "The demographics of overconfidence," *The Journal of Behavioral Finance*, 2006, *7* (1), 5–11.

Bondt, Werner FM De and Richard H Thaler, "Financial decision-making in markets and firms: A behavioral perspective," *Handbooks in operations research and management science*, 1995, *9*, 385–410.

Brenner, Lyle A, Derek J Koehler, Varda Liberman, and Amos Tversky, "Overconfidence in probability and frequency judgments: A critical examination," *Organizational Behavior and Human Decision Processes*, 1996, *65* (3), 212–219.

Bumcrot, Christopher, Judy Lin, and Annamaria Lusardi, "The Geography of Financial Literacy," *Numeracy*, 2013, *6* (2).

Calvet, Laurent E, John Y Campbell, and Paolo Sodini, "Down or out: Assessing the welfare costs of household investment mistakes," *Journal of Political Economy*, 2007, *115* (5), 707–747.

_ , _ , and _ , "Measuring the financial sophistication of households," *American Economic Review*, 2009, *99* (2), 393–98.

Campbell, John Y, "Household finance," *The Journal of Finance*, 2006, *61* (4), 1553–1604.

Fornero, Elsa and Chiara Monticone, "Financial literacy and pension plan participation in Italy," *Journal of Pension Economics & Finance*, 2011, *10* (4), 547–564.

Hsu, Joanne W, "Aging and strategic learning: The impact of spousal incentives on financial literacy," *Journal of Human Resources*, 2016, *51* (4), 1036–1067.

Hung, Angela, Andrew M Parker, and Joanne Yoong, "Defining and measuring financial literacy," *RAND Working Paper Series 708*, 2009.

Lin, Huei-Wen, "Elucidating the influence of demographics and psychological traits on investment biases," *International Scholarly and Scientific Research & Innovation*, 2011, *5* (5), 424–429.

Lusardi, Annamaria and Olivia S Mitchell, "Baby boomer retirement security: The roles of planning, financial literacy, and housing wealth," *Journal of monetary Economics*, 2007, *54* (1), 205–224.

_ and _ , "Financial literacy and retirement preparedness: Evidence and implications for financial education," *Business economics*, 2007, *42* (1), 35–44.

_ **and** _ , "Financial literacy and retirement planning in the United States," *Journal of Pension Economics & Finance*, 2011, *10* (4), 509–525.

_ **and** _ , "Financial literacy around the world: an overview," *Journal of pension economics & finance*, 2011, *10* (4), 497–508.

_ **and** _ , "The economic importance of financial literacy: Theory and evidence," *Journal of economic literature*, 2014, *52* (1), 5–44.

_ **and** _ , "How ordinary consumers make complex economic decisions: Financial literacy and retirement readiness," *Quarterly Journal of Finance*, 2017, *7* (03), 1750008.

_ **,** _ **, and Vilsa Curto**, "Financial literacy among the young," *Journal of consumer affairs*, 2010, *44* (2), 358–380.

_ **,** _ **, and** _ , "Financial literacy and financial sophistication in the older population," *Journal of pension economics & finance*, 2014, *13* (4), 347–366.

Odean, Terrance, "Volume, volatility, price, and profit when all traders are above average," *The Journal of Finance*, 1998, *53* (6), 1887–1934.

A Appendix

A.1 Constructing Overconfidence Measures (Con't)

This section describes the other machine learning classifiers that are used to construct the overconfidence measures. However, because of their large MSEs and failure to predict overconfident households, the out-of-sample predictions are not used in the main analyses.

A.1.1 Logistic Regression Classifier

The logistic classifier fits the linear regression in a sigmoid function such that the probability will not exceed the range $[0, 1]$. Formally,

$$\Pr(y_i = 1|\mathbf{X_i}, \varepsilon_i, \beta_0, \beta_1) = F(\beta_0 + \mathbf{X_i}\beta_1 + \varepsilon_i) \tag{4}$$

where $F(x) = e^x/(1+e^x)$, y_i is an indicator of overconfidence, and $\mathbf{X_i}$ represents the feature matrix. In this paper, the inverse of regulation strength C is tuned in randomized search cross validation.

A.1.2 Naive Bayes Classifier

The Naive Bayes (NB) classifier is an application of Bayes rule. Given overconfidence indicator $y \in \{0, 1\}$, features X_1 through X_P, and the naive conditional independence assumption, Bayes rule gives:

$$\Pr(y|X_1, \cdots, X_P) = \frac{\Pr(y)\prod_{p=1}^{P}\Pr(X_p|y)}{\Pr(X_1, \cdots, X_P)} \tag{5}$$

Since $\Pr(X_1, \cdots, X_P)$ is constant, we can use Maximum A Posteriori (MAP) estimation to estimate the probability of household i to be overconfident, which is

$$y_i = \arg\max_{y\in\{0,1\}} \Pr(y) \prod_{p=1}^{P} \Pr(X_{p,i}|y) \tag{6}$$

Since most features in this paper are binary or categorical, I use the Bernoulli kernel to estimate the model. The additive smoothing parameter α is tuned in randomized search cross validation.

A.1.3 K-nearest Neighbors Classifier

The K-nearest Neighbors (KNN) classifier is a non-parametric model, which simply uses the data in the neighborhood of each data point to predict type. Concretely, the probability of household i to be overconfident is given by

$$\Pr(y_i = 1|\mathbf{X_i}) = \frac{1}{K}\sum_{k\in\mathcal{N}_0} y_k \tag{7}$$

where $\mathcal{N}_0$ is the set of K nearest neighbors. In this paper I use Euclidean distance to find the nearest neighbors. The number of neighbors K is tuned in randomized search cross validation. To avoid overfitting, I set K to be larger than 50.

A.1.4 Multi-layer Perceptron Classifier

The Multi-layer Perceptron (MLP) classifier is a neural network model with multiple hidden layers and nodes. It ensembles nonlinear functions of linear functions of features. To get the nodes at layer j, the MLP classifier estimates the following function:

$$Z_{m,j} = f_j(\alpha_j + \sum_{k \in \mathcal{N}_{j-1}} \beta_{k,j} Z_{k,j-1}) \tag{8}$$

where $Z_{m,j}$ denotes the mth node at layer j, $Z_{k,j-1}$ denotes the kth node at layer $j-1$, $\mathcal{N}_{j-1}$ denotes the set of nodes at layer $j-1$, and $f_j(\cdot)$ denotes the nonlinear activation function. In this paper I use rectified linear unit (reLU) function to estimate the model. The hidden layer sizes and the L2 penalty parameter α are tuned in randomized search cross validation. Given that the learning set is quite large, I set the initial learning rate at 0.02 so that the classifier does not always give the same prediction for every observation.

A.2 Overconfidence Measures from Other Classifiers

Table A1 shows the weighted summary statistics of overconfidence measures from other machine learning classifiers that are not presented in Table 3.

Table A1: Summary statistics: Overconfidence measures from other classifiers

Classifiers	10^{th} pct	Median	90^{th} pct	Mean	S.D.	#Obs.
Logistic	0.002	0.113	0.789	0.264	0.301	80164
Bernoulli NB	0.010	0.187	0.591	0.247	0.231	80164
KNN	0	0.155	0.464	0.203	0.200	80164
MLP	0.011	0.169	0.353	0.189	0.154	80164

Notes: The overconfidence measures are predicted by logistic regression, Bernoulli NB, KNN, and MLP classifiers as described in Appendix A.1. The sample weights in the NFCS are used to calculate the statistics.

A.3 True Financial Literacy Measure from Factor Analysis

Panel A of Table A2 shows the factor loads and uniqueness of the correct indicator for each "Big Five" question. Panel B presents the weighted summary statistics of the constructed measure.

A.4 The Effects of Overconfidence on Financial Behaviors (Con't)

A.4.1 Baseline Results

Table A3 - A6 show the results from regression (3) using overconfidence measures given by other classifiers. Table A3 uses the logistic regression based overconfidence measure; Table A4 uses the Bernoulli NB based overconfidence measure; Table A5 uses the KNN based overconfidence measure; Table A6 uses the MLP based overconfidence measure. The setting of the tables are the same as Table 4. No matter which measure I use, the average marginal effects of overconfidence are always positive and significant, except for those in Table A4 column (5) and (6). This is because the NB classifiers always have a pretty good in-sample fit, but the out-of-sample predictions of them cannot

be seriously treated given its easy setup. In addition, given our unbalanced learning set, it fails to predict a large quantity of overconfident households. Overall, the relationship found in the main analyses is robust.

A.4.2 Heterogeneous Effects of Overconfidence

Table A7 - A10 present the results from regression (3) with low and high literacy subsample using overconfidence measures given by other classifiers. Table A7 uses the logistic regression based overconfidence measure; Table A8 uses the Bernoulli NB based overconfidence measure; Table A9 uses the KNN based overconfidence measure; Table A10 uses the MLP based overconfidence measure. The setting of the tables are the same as Table 6. The average marginal effects are consistent with the ones from SVM and random forest classifiers.

Table A2: Measure for true financial literacy: Factor loads and summary statistics

Panel A: Factor loads and uniqueness for the "Big Five" questions						
Question	Interest Rate	Inflation	Risk Diversification	Mortgage Payment	Bond Price	
Loads	0.6435	0.7315	0.4972	0.6508	0.6824	
Uniqueness	0.5859	0.4649	0.7528	0.5765	0.5344	
Panel B: Summary statistics of the true financial literacy measure						
	10^{th} pct	Median	90^{th} pct	Mean	S.D.	#Obs.
True Literacy	.214	.630	1	0.580	0.299	80164

Notes: I perform a factor analysis on the correct indicators of the "Big Five" questions. Panel A displays the factor loads and uniqueness. The measure for true financial literacy is constructed as the normalized factor score. Panel B presents the summary statistics using the sample weights from the NFCS.

Table A3: Logit regression on overconfidence (logistic regression based) and true financial literacy

Dependent	(1)	(2)	(3)	(4)	(5)	(6)
Variables	Readiness	Readiness	Precaution	Precaution	Participation	Participation
Overconfidence	0.311***	0.309***	0.436***	0.435***	0.320***	0.323***
	(0.0139)	(0.0139)	(0.0158)	(0.0158)	(0.0140)	(0.0141)
True Literacy	0.248***	0.245***	0.223***	0.224***	0.273***	0.272***
	(0.00741)	(0.00745)	(0.00759)	(0.00761)	(0.00721)	(0.00725)
Observations	80,164	80,164	80,164	80,164	80,164	80,164
Demo. chars.	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
State dummies	No	Yes	No	Yes	No	Yes
Pseudo R-squared	0.136	0.137	0.152	0.154	0.188	0.190

Notes: The results are from regression (3). Overconfidence measure is predicted by the logistic regression classifier. True financial literacy is calculated via factor analysis. Demographic characteristics and year dummies are controlled in all columns. Column (2), (4), and (6) further control state dummies. In column (1) and (2), the dependent variable is the indicator for retirement readiness; in column (3) and (4), the dependent variable is the indicator for precautionary savings; in column (5) and (6), the dependent variable is the indicator for financial market participation. Observations are weighted by the NFCS sample weights. Standard errors are in the parentheses. The symbols *, **, and *** denote significance at the 10%, 5% and 1% levels respectively.

Table A4: Logit regression on overconfidence (Bernoulli NB based) and true financial literacy

Dependent	(1)	(2)	(3)	(4)	(5)	(6)
Variables	Readiness	Readiness	Precaution	Precaution	Participation	Participation
Overconfidence	0.0417***	0.0613***	0.0506***	0.0305*	-0.0507***	-0.0452**
	(0.0150)	(0.0164)	(0.0155)	(0.0172)	(0.0159)	(0.0177)
True Literacy	0.189***	0.189***	0.152***	0.150***	0.204***	0.203***
	(0.00708)	(0.00711)	(0.00745)	(0.00746)	(0.00692)	(0.00694)
Observations	80,164	80,164	80,164	80,164	80,164	80,164
Demo. chars.	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
State dummies	No	Yes	No	Yes	No	Yes
Pseudo R-squared	0.128	0.129	0.140	0.142	0.179	0.182

Notes: The results are from regression (3). Overconfidence measure is predicted by the Bernoulli NB classifier. True financial literacy is calculated via factor analysis. Demographic characteristics and year dummies are controlled in all columns. Column (2), (4), and (6) further control state dummies. In column (1) and (2), the dependent variable is the indicator for retirement readiness; in column (3) and (4), the dependent variable is the indicator for precautionary savings; in column (5) and (6), the dependent variable is the indicator for financial market participation. Observations are weighted by the NFCS sample weights. Standard errors are in the parentheses. The symbols *, **, and *** denote significance at the 10%, 5% and 1% levels respectively.

Table A5: Logit regression on overconfidence (KNN based) and true financial literacy

Dependent Variables	(1) Readiness	(2) Readiness	(3) Precaution	(4) Precaution	(5) Participation	(6) Participation
Overconfidence	0.179***	0.178***	0.189***	0.188***	0.219***	0.219***
	(0.0160)	(0.0160)	(0.0170)	(0.0170)	(0.0163)	(0.0163)
True Literacy	0.234***	0.231***	0.196***	0.197***	0.271***	0.269***
	(0.00813)	(0.00815)	(0.00842)	(0.00843)	(0.00802)	(0.00805)
Observations	80,164	80,164	80,164	80,164	80,164	80,164
Demo. chars.	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
State dummies	No	Yes	No	Yes	No	Yes
Pseudo R-squared	0.130	0.131	0.142	0.143	0.182	0.184

Notes: The results are from regression (3). Overconfidence measure is predicted by the KNN classifier. True financial literacy is calculated via factor analysis. Demographic characteristics and year dummies are controlled in all columns. Column (2), (4), and (6) further control state dummies. In column (1) and (2), the dependent variable is the indicator for retirement readiness; in column (3) and (4), the dependent variable is the indicator for precautionary savings; in column (5) and (6), the dependent variable is the indicator for financial market participation. Observations are weighted by the NFCS sample weights. Standard errors are in the parentheses. The symbols *, **, and *** denote significance at the 10%, 5% and 1% levels respectively.

Table A6: Logit regression on overconfidence (MLP based) and true financial literacy

Dependent Variables	(1) Readiness	(2) Readiness	(3) Precaution	(4) Precaution	(5) Participation	(6) Participation
Overconfidence	0.247***	0.245***	0.194***	0.192***	0.120***	0.120***
	(0.0198)	(0.0198)	(0.0214)	(0.0214)	(0.0198)	(0.0197)
True Literacy	0.231***	0.228***	0.183***	0.184***	0.234***	0.232***
	(0.00776)	(0.00779)	(0.00821)	(0.00822)	(0.00764)	(0.00767)
Observations	80,164	80,164	80,164	80,164	80,164	80,164
Demo. chars.	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
State dummies	No	Yes	No	Yes	No	Yes
Pseudo R-squared	0.130	0.131	0.141	0.143	0.180	0.182

Notes: The results are from regression (3). Overconfidence measure is predicted by the MLP classifier. True financial literacy is calculated via factor analysis. Demographic characteristics and year dummies are controlled in all columns. Column (2), (4), and (6) further control state dummies. In column (1) and (2), the dependent variable is the indicator for retirement readiness; in column (3) and (4), the dependent variable is the indicator for precautionary savings; in column (5) and (6), the dependent variable is the indicator for financial market participation. Observations are weighted by the NFCS sample weights. Standard errors are in the parentheses. The symbols *, **, and *** denote significance at the 10%, 5% and 1% levels respectively.

Table A7: Heterogeneous effects of overconfidence (logistic regression based)

Dependent Variables	(1) Readiness	(2) Readiness	(3) Precaution	(4) Precaution	(5) Participation	(6) Participation
Overconfidence	0.150***	-0.116*	0.213***	-0.0864	0.0929***	-0.0796
	(0.0165)	(0.0656)	(0.0198)	(0.0580)	(0.0118)	(0.0722)
Observations	5,886	12,539	5,886	12,539	5,886	12,539
Sample	Low Lit.	High Lit.	Low Lit.	High Lit.	Low Lit.	High Lit.
Demo. chars.	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
State dummies	Yes	Yes	Yes	Yes	Yes	Yes
Pseudo R-squared	0.121	0.166	0.101	0.130	0.178	0.106

Notes: The results are from regression (3) with low and high literacy subsamples. Overconfidence measure is predicted by the logistic regression classifier. Demographic characteristics, year dummies, and state dummies are controlled in all columns. Column (1), (3), and (5) use the low literacy subsample, while column (2), (4), and (6) use the high literacy subsample. In column (1) and (2), the dependent variable is the indicator for retirement readiness; in column (3) and (4), the dependent variable is the indicator for precautionary savings; in column (5) and (6), the dependent variable is the indicator for financial market participation. Observations are weighted by the NFCS sample weights. Standard errors are in the parentheses. The symbols *, **, and *** denote significance at the 10%, 5% and 1% levels respectively.

Table A8: Heterogeneous effects of overconfidence (Bernoulli NB based)

Dependent Variables	(1) Readiness	(2) Readiness	(3) Precaution	(4) Precaution	(5) Participation	(6) Participation
Overconfidence	0.135***	-0.296***	0.188***	-0.330***	0.0887***	-0.288***
	(0.0152)	(0.0570)	(0.0198)	(0.0489)	(0.0117)	(0.0537)
Observations	5,886	12,539	5,886	12,539	5,886	12,539
Sample	Low Lit.	High Lit.	Low Lit.	High Lit.	Low Lit.	High Lit.
Demo. chars.	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
State dummies	Yes	Yes	Yes	Yes	Yes	Yes
Pseudo R-squared	0.119	0.169	0.0963	0.135	0.177	0.109

Notes: The results are from regression (3) with low and high literacy subsamples. Overconfidence measure is predicted by the Bernoulli NB classifier. Demographic characteristics, year dummies, and state dummies are controlled in all columns. Column (1), (3), and (5) use the low literacy subsample, while column (2), (4), and (6) use the high literacy subsample. In column (1) and (2), the dependent variable is the indicator for retirement readiness; in column (3) and (4), the dependent variable is the indicator for precautionary savings; in column (5) and (6), the dependent variable is the indicator for financial market participation. Observations are weighted by the NFCS sample weights. Standard errors are in the parentheses. The symbols *, **, and *** denote significance at the 10%, 5% and 1% levels respectively.

Table A9: Heterogeneous effects of overconfidence (KNN based)

Dependent Variables	(1) Readiness	(2) Readiness	(3) Precaution	(4) Precaution	(5) Participation	(6) Participation
Overconfidence	0.142***	-0.309***	0.204***	-0.264***	0.0905***	-0.320***
	(0.0163)	(0.0912)	(0.0201)	(0.0762)	(0.0123)	(0.0975)
Observations	5,886	12,539	5,886	12,539	5,886	12,539
Sample	Low Lit.	High Lit.	Low Lit.	High Lit.	Low Lit.	High Lit.
Demo. chars.	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
State dummies	Yes	Yes	Yes	Yes	Yes	Yes
Pseudo R-squared	0.118	0.167	0.0986	0.131	0.176	0.107

Notes: The results are from regression (3) with low and high literacy subsamples. Overconfidence measure is predicted by the KNN classifier. Demographic characteristics, year dummies, and state dummies are controlled in all columns. Column (1), (3), and (5) use the low literacy subsample, while column (2), (4), and (6) use the high literacy subsample. In column (1) and (2), the dependent variable is the indicator for retirement readiness; in column (3) and (4), the dependent variable is the indicator for precautionary savings; in column (5) and (6), the dependent variable is the indicator for financial market participation. Observations are weighted by the NFCS sample weights. Standard errors are in the parentheses. The symbols *, **, and *** denote significance at the 10%, 5% and 1% levels respectively.

Table A10: Heterogeneous effects of overconfidence (MLP based)

Dependent Variables	(1) Readiness	(2) Readiness	(3) Precaution	(4) Precaution	(5) Participation	(6) Participation
Overconfidence	0.145***	-0.559***	0.201***	-0.590***	0.0943***	-0.566***
	(0.0145)	(0.0879)	(0.0191)	(0.0735)	(0.0111)	(0.0825)
Observations	5,886	12,539	5,886	12,539	5,886	12,539
Sample	Low Lit.	High Lit.	Low Lit.	High Lit.	Low Lit.	High Lit.
Demo. chars.	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
State dummies	Yes	Yes	Yes	Yes	Yes	Yes
Pseudo R-squared	0.124	0.169	0.100	0.136	0.181	0.110

Notes: The results are from regression (3) with low and high literacy subsamples. Overconfidence measure is predicted by the MLP classifier. Demographic characteristics, year dummies, and state dummies are controlled in all columns. Column (1), (3), and (5) use the low literacy subsample, while column (2), (4), and (6) use the high literacy subsample. In column (1) and (2), the dependent variable is the indicator for retirement readiness; in column (3) and (4), the dependent variable is the indicator for precautionary savings; in column (5) and (6), the dependent variable is the indicator for financial market participation. Observations are weighted by the NFCS sample weights. Standard errors are in the parentheses. The symbols *, **, and *** denote significance at the 10%, 5% and 1% levels respectively.