Predicting Retirement and Social Security Claiming Decisions using Machine Learning

Alexander Kwon [*] Lilia Maliar [†]

January 3, 2024

Abstract

We show that machine learning can help to improve predictions of individual retirement and Social Security (SS) claiming decisions. When predicting the number of people receiving SS, we achieve an error of less than 1%, while the benchmark model employed by the Social Security Administration results in over a 4% error. When forecasting SS claiming decision, we attain a low error of 0.2%, whereas the benchmark SSA model produces an error rate exceeding 2%. The set of important variables selected by our model differs from that of the SSA model. We use Shapley values to evaluate the non-linear contributions of variables to predictive outcomes.

[*]Department of Economics, Graduate Center, CUNY
[†]Department of Economics, Graduate Center, CUNY

1 Introduction

According to the most recent projections by the Treasury Department, the Old-Age & Survivors Insurance (OASI) Trust Fund will be depleted by around 2033.[1] Therefore, it is important to obtain accurate forecasts of individual retirement and Social Security claiming decisions to correctly predict the date of depletion. The accuracy of prediction is relevant for policy reforms that solve this problem.

In this paper, we use a machine learning technique called a gradient boosted tree to obtain accurate predictions of retirement and Social Security (OASI) claiming probabilities across individuals. We use gradient boosted tree because it can handle datasets with a large number of variables. Moreover, given that not all variables might be important for prediction, we need a machine learning model which is robust to the inclusion of unimportant variables. The gradient boosted tree can easily work with such big data in which not all variables might be important. In comparison to other widely used machine learning models, including random forest, neural networks, and lasso estimation, we find that the gradient boosted tree yields the highest performance on our test data (comparison is provided in Appendix B).

Another advantage of using machine learning, rather than using classical econometric tools or economic theory, is that researchers can avoid making assumptions about which variables are important for prediction. The machine learning model itself can discern between important and unimportant variables for prediction, rather than relying on the researcher's decision.

We compare our model's performance with a model used by the Social Security Administration (SSA) which is called "Model of Income in the Near Term" (MINT); see Smith et al. (2021). SSA uses the MINT as a tool to project future retirement incomes, as well as the distributional effects of Social Security reform proposals. To do so, MINT uses probit estimation for predicting retirement probabilities and logistic regression for Social Security take-up.

Using the Health and Retirement data from 1992 to 2018, we identify which variables are important for the prediction of retirement and Social Security claiming decisions, respectively. Although there are a large number of previous studies that discuss the determinants of retirement and Social Security claiming decisions, there is no consensus regarding the variables that are important for prediction. Our paper selects the variables that are important for prediction from a dataset that contains over 580 variables. We also

[1] Trustees Report, Treasury Department in June of 2023, `https://home.treasury.gov/news/press-releases/jy1381`

show that the gradient boosted tree outperforms the commonly used logit and probit estimation methods in predicting retirement and Social Security claiming decisions.

The results of our analysis show that, for the retirement case, variables related to respondents' job histories are selected as important in our model, but are not used in MINT. Conversely, MINT incorporates demographic characteristics, education, and spouse-related variables, which our model finds not significant for prediction. In the case of Social Security, our analysis selects job history, health insurance, and bequest-related variables as important, while they are not used in MINT. However, MINT incorporates earnings, self-employment status, and education-related variables, which our model finds not significant. These disparities provide insights into areas where improvements in prediction can be made for each case.

To compare the performance of the two models we use the area under a receiver operating characteristic curve (AUC). For the retirement case, our model has an AUC of 0.8173, while the probit estimation used by MINT has an AUC of 0.7436. A higher AUC of our analysis suggests that machine learning can help us improve the prediction of retirement. However, the improvement for the Social Security case is smaller, with our model at 0.9589 and MINT at 0.9397.

We further do a two-step prediction exercise similar to MINT. We first predict retirement and then predict Social Security beneficiary status using the predicted retirement. Our model yields less than 1% error in the percentage of Social Security beneficiaries, while MINT yields over 4% error. Taking into account of the number of beneficiaries and the average monthly benefit, we show that the 3% difference in prediction amounts to 39.6 billion dollars a year, which is approximately 3% of the total benefit payments in 2023.

Next, we predict the Social Security beneficiary rate with the above two-step procedure for year 2020 with data before year 2020. We consider the COVID-19, which hit the US at 2020, as an exogenous shock and check the external validity of our model. Our model yields an error rate of 0.2%, while the MINT model yields an error rate exceeding 2%. The results show that our models performance is robust to the COVID-19 shock.

We further do robustness checks for three potential drawbacks and provide suggestive evidence that our results are not affected by those potential drawbacks. First, one possibility is that the information of the non-selected variables is already reflected in the selected variables. When we fit the gradient boosted tree, we use residuals after fitting the selected variables with the non-selected variables using a polynomial of degree 2 to see if the left-over variation in the selected variables help predictions. Second, common time effects might affect our results. We demean the data for each interview period to

eliminate the common time effect of a given interview period. Third, our results might be affected by the fact that the definition of retirement is different with the MINT model. We redefine retirement as in MINT and see the performance of our model. We observe that our method outperforms the MINT model after considering different specifications.

Finally using Shapley values from cooperative game theory, we show how the selected variables contribute to the prediction of retirement. Some of the highlight of our results are as follows: The contributions of the continuous variables are highly non-linear. For the retirement case, having pension and annuity income is an important variable in the prediction of retirement rather than the detailed type of pension. The fact that the respondent has pension or annuity income helps the model predict higher retirement probability for this respondent. Our results show that there is evidence of job transition before retirement. The results also show evidence of liquidity constraint. For the Social Security case, age and birth cohort are the most important variables contributing to the prediction. Also, there is evidence that health insurance is an important factor of receiving Social Security.

The paper is organized as follows. Section 2 contains a brief literature review related to this paper. Section 3 presents the methodology of our analysis. Section 4 introduces the data set. Section 5 shows the results. Section 6 shows the robustness of our results. Section 7 compares our results with the previous literature. Section 8 concludes.

2 Literature review

To compare with our method, we consider Version 8 of the MINT microsimulation model. MINT encompasses various economic and demographic projections, such as employment, individual earnings, periods of unemployment, and contributions to pension plans. In particular, our project is focused on predicting retirement and Social Security claiming within Version 8 of MINT. [2]

The paper contributes to the literature that identifies the incentives of retirement and Social Security claiming decisions by giving suggestive evidence on what variables are important for prediction and how they contribute to predictions. The incentives are well documented in Chapter 8 of *Handbook of Economics of Population Aging* by Blundell et al. (2016). According to Blundell et al. (2016), the incentives for retirement and claiming Social Security are closely related to each other since the decisions are not independent of each other.

A summary of the literature review is provided in Table 8 in section 7. For example,

[2]The MINT retirement model that we compare with are built on the work of Gustman and Steinmeier (2001).

Rust and Phelan (1997) illustrate that those who report having a health problem, which limits work, are nearly twice as likely to receive Social Security at 62 than at 65, which is the early retirement age and normal retirement age for the sample. As they show with a structural model, Medicare provides a large incentive to retire for those who are not covered by employer-provided health insurance after retirement; see also French and Jones (2011). Some papers focus on the joint household decision. These papers are motivated by the fact that couples' retirement roughly coincides (Hamermesh, 2000).

Hurd et al. (2004) and Coile et al. (2002a) provide a theoretical discussion and empirical evidence on the determinants of Social Security claiming. Coile et al. (2002a) find support for predictions from theory: men with longer life expectancy have longer delays in claiming; wealth has an inverse U-shape effect on delay; men with younger spouses delay more. Hurd et al. (2004) show that subjective probability of survival is an important determinant of when to claim Social Security. They find that the lower the subjective survival probability, the more likely the individual will claim early. However, the effect is not large. They also find that a high level of pension savings increases the likelihood of early claiming. Haurin et al. (2022) show that financial stress can increase the probability of delayed claiming at age 62. Other papers that focused on the determinants of early claiming include von Wachter (2009), Li et al. (2008), Glickman and Hermes (2015), Butrica and Karamcheva (2018), Shoven (2018), and Huang et al. (2022).

Moreover, we are related to the literature that uses machine learning for prediction in economics. Some noticeable examples that use machine learning for answering economic questions are as follows: Fouliard et al. (2020) use machine learning to predict the financial crisis. Albanesi and Vamossy (2019) use deep learning to predict consumer default. This paper is similar to the aforementioned papers in that it will use machine learning to improve prediction.

3 Methodology of our analysis

There are two steps in our analysis. First, we fit a gradient boosted tree with all available variables to select variables important for prediction. Second, we fit the gradient boosted tree with only the selected variables and check how each variable contributes to the prediction.

We briefly discuss what a gradient-boosted tree is by following Hastie et al. (2009).[3] A tree of a given data point x, $T(x;\Theta)$, can be expressed as

[3]Section 10.9

$$T(x;\Theta) = \sum_{j=1}^{J} \gamma_j I(x \in R_j),$$

in which $\Theta = \{R_j, \gamma_j\}_1^J$; R_js are disjoint partitions of the space of predictors (x) and J is the number of disjoint partitions (number of terminal nodes); $I()$ is an indicator function; and γ_j is the predicted value when $I(x \in R_j) = 1$, meaning $f(x) = \gamma_j$. The parameters are found by minimizing a loss function,

$$\hat{\Theta} = \arg\min_{\Theta} \sum_{j=1}^{J} \sum_{x_j \in R_j} L(y_j, \gamma_j).$$

A boosted tree model is a sum of trees,

$$f_M(x) = \sum_{m=1}^{M} T(x;\Theta_m),$$

in which M is the number of trees and m represents the tree index. The next tree is searched by a step-wise procedure by solving

$$\hat{\Theta}_m = \arg\min_{\Theta_m} \sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + T(x_i;\Theta_m))$$

for the region set and constants $\Theta_m = \{R_{jm}, \gamma_{jm}\}_1^{J_m}$ of the next tree, given $f_{m-1}(x)$, considering all data points $i = 1, \ldots, N$. By using the Taylor approximation on $L()$, the gradient-boosted tree updates the previous $f_{m-1}(x)$ using gradient of the loss function $L()$ with respect to $f_{m-1}(x)$.

We consider 3 ways of selecting variables and a variable is selected as important if it enters at least twice of three different lists of important variables. In this way, we are being conservative. The three important variable lists are based on three different criteria: a mean decrease in impurity, a permutation-based method, and Shapley values. The three lists contain the top 25 important variables for each criterion.

The mean decrease in impurity takes the weighted average of the decrease of the splitting criteria when a certain variable is included. *The permutation-based method* randomly shuffles the value of a variable of interest and checks the loss of the model. If the loss increases more as we shuffle a certain variable, that variable is considered more important. *The Shapley value method* compares the mean absolute value of Shapley values of a variable.

The Shapley value for a single observation is the weighted average of all possible differences of a prediction model that includes j with the one that excludes j, $f_{S\cup\{J\}}(x_{S\cup\{J\}}) - f_S(x_S)$, where S is a subset of the feature space F, x_S represents values of input features in set S, and $f_S(x_S)$ represents the model trained with feature S. In particular, we have

$$\phi_j = \sum_{S\subseteq F\setminus\{j\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f_{S\cup\{J\}}(x_{S\cup\{J\}}) - f_S(x_S)].$$

Intuitively, if we think of the prediction as a game and features as players, Shapley values can be thought of as the weighted average marginal contribution of a single player to every possible subset of players (Kumar et al., 2020).

Shapley values are calculated with SHapley Additive exPlanations (SHAP). In particular, since we are using a gradient boosted tree as our model, we use the TreeSHAP. Sundararajan and Najmi (2020) show that the desirable properties of Shapley values, which are presented in Lundberg and Lee (2017), might not hold when the computation method uses conditional expectations, which is the case of TreeSHAP. Among the failure of desirable properties, the failure to give zero Shapley values to variables that do not contribute might be problematic for our analysis (failure of dummy). However, this problem is circumvented by only looking at the top 25 important variables.

The three different lists do not have the same interpretation and all the lists have their advantages and disadvantages. Below, we explain the difference in interpretation and the advantages and disadvantages of each method. We then explain why we selected variables if a variable enters at least twice of three lists.

First, the impurity-based variable importance list is based on the training data and shows what variables the model depends on when making predictions (Molnar, 2022). However, the important variables based on the training data might not be important in the test data. Moreover, the impurity based method is known to favor high cardinality variables over low cardinality variables such as binary features. The permutation-based and Shapley value methods do not have these problems. Thus, finding the intersection of the impurity-based list and the other lists can alleviate the problem that the impurity-based list possesses.

Second, permutation-based variable importance should be interpreted as the increase in loss when a variable is shuffled randomly. There are a couple of problems with permutation-based variable importance. First, the permutation process might yield unrealistic data that we might not observe in reality. For instance, let one variable be the child's age and the second variable be the father's age. If we shuffle the father's

age, it is possible to obtain a shuffled data point such as the father's age being 20 and the child's age being 20 (only thinking of the biological father). Training the model with this unrealistic data point might create bias in computing the true variable importance. The TreeSHAP uses the conditional expectation for calculating the Shapley values. This implies that the TreeSHAP uses the given data instead of using unrealistic data when computing variable importance. Thus, taking the intersection with the permutation-based method and the Shapley value method can alleviate the problem of the permutation-based method potentially relying on unrealistic data points.

A further possible problem with the permutation-based method is that if a variable has a high correlation with another variable, the importance can be split into the two variables' importance measure, leading to a decreased importance measure for a highly important variable. The selection process in this paper truncates the important variable lists to the top 25. This might omit important variables that are ranked lower than the top 25 due to a high correlation with other variables. We cannot solve this problem but expect that the highly important variables are mostly included since the performance measure only slightly decreases after using the selected variables compared to the case of using all the available variables.

Third, the Shapley values do not show the difference of the predicted values before and after removing a certain variable but represent the contribution of a variable to the difference between the actual and mean/baseline prediction (Molnar, 2022). Thus, if you want to know what variables are important to the correct prediction, Shapley values might not be the values you should rely on. Instead, the permutation-based method might be more suitable for this case. Therefore, taking the intersection with the permutation-based method and Shapley values method is similar to combining the two interpretations: We want to select variables that have high contributions to the prediction and are important for the correct prediction.

When using the TreeSHAP, there are further problems as pointed out in Sundararajan and Najmi (2020). The TreeSHAP uses the conditional expectation in the context of a decision tree or an ensemble tree which makes some of the desirable properties of the Shapley values not hold: dummy, linearity, symmetry, and demand monotonicity.[4] We do not check how much the properties are violated in our cases. However, we expect that our analysis is not affected by the violation of the desirable properties since our analysis needs a measure of importance that captures the contribution of a variable to

[4]This is problematic since it is known that Shapley values are the unique payment attribution method that satisfies a subset of desirable properties including dummy, linearity, and symmetry. This implies that, in some cases, the TreeSHAP is producing something else than the theoretical Shapley values.

the prediction of the model and do not need the properties to always hold.[5]

The second step starts by cleaning the data again with the selected variables. We do so since the variables in our dataset contains reasons for missing data which is not a numerical value. For example, if the respondent refuses to answer a survey question which leads to the related variable to have a missing value, the variable is usually assigned with a character ".r". Previously, a large negative value was assigned to each missing value. This is to consider the case in which there is a systematic pattern for the missing values that is different from the non-missing values. In other words, we are giving a new category for each reason for "missing"; see Hastie et al. (2009) (page 311). The gradient boosted tree will split the data space between missing and non-missing if there is a systematic pattern for the missing values that help predictions. Before the second step, we check each variable and assign a reasonable value to the missing value. This process is important because we want to interpret the feature contribution. With the re-cleaned data, we fit a gradient-boosted tree again and use Shapley values to see how the selected variables contribute to the prediction of the dependent variable.

4 Data

"Our data are sourced from the Health and Retirement Study (HRS), specifically the *RAND HRS Longitudinal File 2018*. This dataset is a cleaned version of the HRS Core and Exit interviews processed by the RAND organization. The HRS data contain samples of the old-aged population in the U.S. We only include individuals whose age is equal to or over 55. We include waves from 4 to 14. Waves 1,2 and 3 are not included because they have different interview terms depending on the individual birth cohort. Using waves 4-14 allows us to keep the data-cleaning process consistent throughout the waves. We clean the sample separately for the retirement and Social Security cases.

For the retirement case, the dependent variable is a dummy variable that is 1 if the respondent considers oneself as completely retired. The way we pooled the data is as follows. First, we get the non-retired from wave 14. The implicit assumption is that individuals who consider themselves completely retired do not rejoin the labor force. However, this may not be true. Some individuals report that they retired at a certain wave but say that they are not or partly retired in later waves. We do not address this issue in our analysis.

Next, we get retired individuals from each wave. To be specific, individuals who

[5]Further research is needed to incorporate other methods, such as integrated gradient method, to compute Shapley values from a gradient boosted tree and obtain Shapley values that hold desirable properties.

report that they consider themselves completely retired and report their retirement year between 2016 and 2018 are considered as individuals who retired in wave 14. We do this process for each wave. We construct the data in this way because variables usually ask about the status of the individual in the previous 2 years. Thus, wave 14 variables should be related to individuals who either retired in wave 14 or who did not. However, we use the previous wave data for earnings, hours worked, and whether an individual is self-employed. This is because retirement usually mechanically decreases the earnings and the hours worked, and retired individuals are mostly not self-employed after retirement.

We drop individuals who receive Social Security Disability Insurance (SSDI) or Supplemental Security Income (SSI) benefits because the rules for those programs are different with OASI. We exclude variables that directly contain information about retirement. We further exclude variables that contain errors, and we exclude variables that are not in all waves from 4 to 14. There are 584 variables left in the sample. Finally, we only keep individuals who are married. More details about the data cleaning process is presented in Appendix D. 11,230 respondents are in the sample and 5,926 consider themselves completely retired while 5,304 do not. When we do the second step, we re-clean the data by assigning specific values for missing values or dropping observations that contain missing values that are not possible to assign an intuitive number. There are 7,966 respondents and 4,284 consider themselves completely retired in the married couple data.

For the Social Security case, we use the same set of variables but exclude 2 more variables that directly contain the information about Social Security. There are 11,230 observations and 8,665 receive Social Security. The re-cleaned sample for the second step has 10,870 observations and 8,599 receive Social Security.

5 Numerical Results

In this section, we present our numerical results. The hyper-parameters used for the gradient boosted trees are reported in Appendix C. In Section 5.1, we show the selected variables as a result of our analysis. Section 5.2 shows that our gradient boosted tree outperforms the MINT models for both the retirement case and Social Security case. In Section 5.3, we show that our gradient boosted tree method produces less error than the MINT models when we do two-step prediction.

5.1 Selected variables

Table 1 shows the selected variables for the retirement and Social Security cases and compares the variables with the MINT variables. During the selection process, variables highly correlated with another variable are selectively dropped, assuming that the two variables have roughly similar information. Detailed selection process is in Appendix F.

For a summary of the retirement case, our results show that variables related to the job history of the respondents are selected as important in our analysis but are not used in MINT. In contrast, demographic characteristics, education, and spouse-related variables are not selected as important for prediction in our model, while MINT incorporates those variables.

Intuitively, variables related to the job history of an individual are informative about the retirement of the individual. For example, *years at the longest reported job* can be important since the Social Security benefits are provided if, according to the law, "you have earned an average of one work credit for each calendar year between age 21 and the year in which you reach age 62".[6] Individuals who want to claim benefits should satisfy a certain work credit to be qualified. This work credit depends on the years of work. Thus, years at the longest reported job can be informative about the work credit an individual earned.

Interestingly, demographic characteristics, education, and spouse-related variables are not selected as important variables in our analysis. However, we should be careful interpreting the results. The results do not suggest that demographic characteristics, education, and spouse-related variables do not have a causal effect on the retirement decision. Instead, the results show that when it comes to prediction, the aforementioned variables are less important than the selected variables. Because the gradient-boosted tree is not looking for a causal relationship, we acknowledge the fact that the selected variables might already contain information about demographic, educational, and spouse-related information. Further discussion is provided in Section 6.

For a summary of the Social Security case, our analysis selects variables related to job history, health insurance, and bequest-related variables as important but are not used in MINT. However, earnings, self-employment status, and education-related variables are not selected as important in our analysis while MINT uses those variables.

Medicare and employer provided health insurance are considered as an important variable in numerous previous studies (Rust and Phelan, 1997; French and Jones, 2011). Blundell et al. (2016) point out that the provision of health insurance by the employer is

[6]Further information about the rules that determine eligibility for Social Security benefits are summarized in Appendix A

an important factor that explains US labor supply decisions. Those who do not qualify for Medicare due to not being old enough might not be able to retire because they will lose their health insurance once stop working. Considering this fact, it is understandable that Medicare is an important indicator variable in the prediction of Social Security claiming, intuitively through the link with retirement. Moreover, the same logic can be applied why the variable that indicate whether the individual is covered by employer provided health insurance (*Covered by employer/Hlth insurance*) and the indicator variable for whether the employer provided health insurance covers retirees (*Employer-provided health plan covers retirees*) are selected as important variables.

Industry code of the longest job is included as an important variable for predicting Social Security claiming. The lifetime labor supply behavior can be different depending on the industry code which is based on the 1980 census code. For instance, those who work in "Professional and related services", which include includes physicians, legal services, education services, social services, engineering, architectural, accounting services, and more, might have a different retirement age with those in "Agriculture, forestry, fishing" sector.

A bequest-related variable is selected as important for the Social Security case which is the probability of leaving bequest over $100,000 (*Probability to leave bequest over 100k*). Although Social Security cannot be bequeathed, the paid benefit can be subject to bequest. According to Mukherjee (2022), Mukherjee (2018), and Lee and Tan (2023), there is a non-trivial bequest motive for Social Security benefits. Our finding is consistent with the results of the aforementioned studies in the context that bequest motive might be an important factor for Social Security benefits.

Further comparison with previous literature is given in Section 7.

If we compare the selected variables between retirement and Social Security, health insurance, spouse, demographic, and bequest related variables are important for the Social Security case but not for the retirement case. Meanwhile, household business assets are only important for the retirement case. Since retirement and Social Security claiming are not independent to each other, it is possible that the important variables in each case are important for both cases. Our analysis at least shows that the priority of the important variables is different for the prediction of retirement and Social Security claiming.

In the next section, we use the selected variables in Table 1 to fit the gradient-boosted tree. For the retirement and Social Security case, we add the replacement rate and the logarithm of earnings to be comparable with MINT. Replacement rate, which is the percentage of how much your benefits will make up your earnings, summarizes the rules related to Social Security. Thus, it can be an important variable to know the effect of So-

cial Security rules on decisions. If replacement rate and log earnings are not important in prediction, then the gradient-boosted tree will not pick those variables to split. If they are important, then a better prediction will be possible. We find that and including the two extra variables do not significantly change the results.

Table 1: Selected variables: a comparison between MINT and our analysis

Category	MINT	Retirement	Social Security
Age	Age dummies(52 to 63) Born 1938-1942 Born 1943-1954 Born 1955 or later	Age in months Respondent birth cohort	Age in months Respondent birth cohort
Health	Fair/poor health	Health problem limits work	Doctor visit/ dummy
Income/Wealth	Per capita family wealth/AWI Homeowner	Household total income Net wealth/exclude secondary residence Value of other debt	Household total income
Earnings	Logarithm of earnings		Earnings/prev. wave
Retirement	Retired		Retired
Job history		Years at longest reported job Hours worked per week in -1st job in previous wave Hours worked per week in -2nd job in previous wave Number of jobs over -employment history	Years at longest reported job Industry code of longest job
Self-employment	Self-employed	Household business assets	
Social Security benefits		Receive Social Security benefits Social Security income	
Private pension	DB plan DC plan CB plan	Pension/Annuity income Pension income/dummy Number of pensions currently receiving	Pension/Annuity income
Health insurance			Medicare Covered by employer/Hlth insurance Employer-provided health plan covers retirees
Household joint decision	Age difference to spouse Log spouse earnings Present value of spouse lifetime earnings divided by the cohort average Spouse black Spouse Hispanic Spouse two year age dummies (45-66) Spouse age 67 or higher Spouse DB plan Spouse DC plan Spouse CB plan Spouse self-employed		Receive Social Security/spouse Imputed wage rate/spouse
Demographics	Black Hispanic Foreign born		Birth place
Education	Less than high school Some college College More than college		
Bequest			Probability to leave bequest over 100K
Related to Social Security rules	Replacement rate and squared New age 64 to 68		

5.2 Performance comparison

In this section, we present the performance of the gradient boosted tree model, where the variables are those selected in the previous section. Table 2 compares the performance of each model for the retirement case. First, we replicate the probit model used by MINT, producing an AUC of 0.7436. If we select variables and run a gradient-boosted tree, the

AUC increases to 0.8173. Since the MINT variables do not include Social Security-related variables, we also exclude them—specifically, *Receive Social Security* and *Social Security income*—and run the gradient-boosted tree again. The resulting AUC is still 0.8086, which is higher than the MINT probit model.

Table 2: Performance comparison/ Retirement case

	584 variables	MINT variables		Selected variables		No SS variables
Method	Gradient boosted	Gradient boosted	Probit	Gradient boosted	Probit	Gradient boosted
AUC curve	0.8481	0.7569	0.7436	0.8173	0.7878	0.8086

Table 3: Performance comparison/ Social Security case

	582 variables	MINT variables		Selected variables		No earnings & replacement rate
Method	Gradient boosted	Gradient boosted	Logistic	Gradient boosted	Logistic	Gradient boosted
AUC curve	0.9677	0.9341	0.9397	0.9598	0.9580	0.9603

Table 3 compares the performance of each model for the Social Security case. The MINT logistic model already gives a high AUC, which is 0.9339, meaning that the MINT logistic model can accurately discern individuals receiving Social Security from those who do not. The AUC for the gradient-boosted tree with the selected variables improves slightly to 0.9598 compared to the MINT logistic model. The last column reports the AUC when we only include the selected variables and exclude earnings and replacement rate which were arbitrarily included to match the MINT project. The AUC is still high with a value of 0.9603.

5.3 Two-step prediction

In this section, we perform a two-step prediction process to check how much the prediction can be improved when we do the prediction sequentially. We follow the MINT model and predict retirement first and then predict Social Security status using the selected variables and predicted retirement status with the gradient boosted tree. The results are summarized in Table 4. As we can see in the second column of Table 4, our model creates less than 1% error in the test data. However, as presented in the third column of Table 4, the MINT model creates more than 4% error in the test data.

To know the significance of a difference of 3% error, we perform a simple calculation. Based on a report published by the SSA on September 2023 [7], the number of people receiving only Social Security was 63 million in August 2023 [8]. A 3% error in prediction

[7]Named "Monthly Statistical Snapshot"

[8]Need to be careful since those who receive Social Security because of being disabled is included

means misclassifying 1.89 million people. The average monthly benefit was around 1,758 dollars for OASI beneficiaries in August 2023. If we multiply 1.89 million with 1,758, it amounts to 3.3 billion dollars. If we convert it to one year, the 3% difference in prediction amounts to 39.6 billion dollars, which is approximately 3% of the total benefit payments in 2023.

Table 4: Two-step prediction

	GB tree with selected variables	MINT model
True average beneficiary rate	78.04%	77.19%
Predicted average beneficiary rate	79.08%	81.93%

The percentage of beneficiaries differ between data because the cleaning process was redone for each case.

We also predict the beneficiary rate between 2018 and 2020 using the two-step prediction by training the gradient boosted tree with data before 2018. Given that 2020 is the year when COVID-19 hit the U.S., there is a possibility that our model, trained with data before 2018, might predict poorly on 2020 data, indicating low external validity. However, the error is only 0.2% for our model. In contrast, the error is around 2% for the MINT model. Even though we know that there was an exogenous shock at 2020, our model still yields highly accurate predictions.

Table 5: Two-step prediction for 2020 data

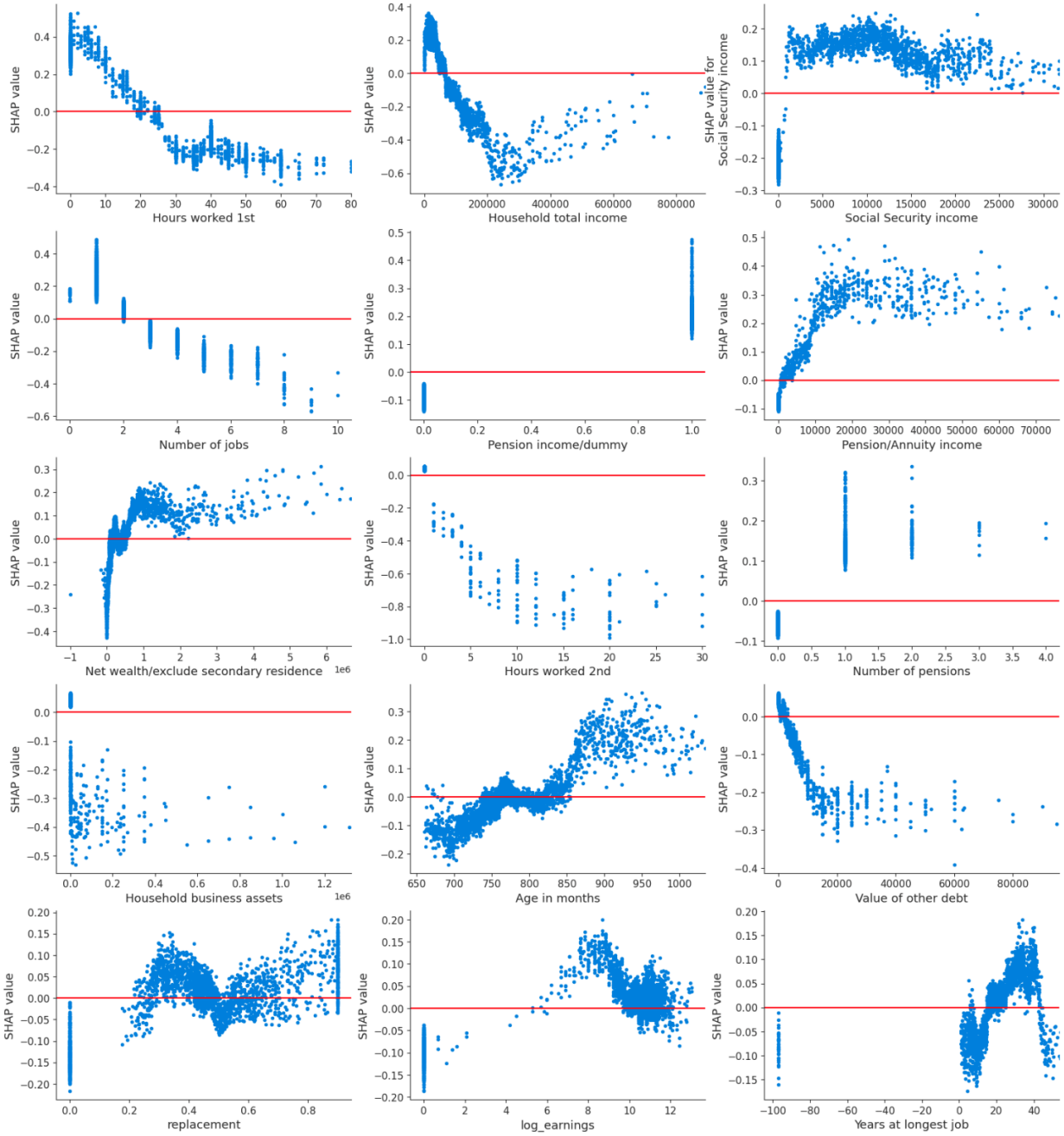
Percentage of beneficiaries	27.41%
Predicted with MINT	25.25%
Predicted with GB	27.68%

In Section 6, we check the robustness of our results against three possible problems. First, the important information in the non-selected variables from the MINT model might be already included in the selected variables. Second, the time effects that are not considered in our model might cause a problem. Third, the definition of retirement in the MINT model is whether the individual worked less than 20 hours which is different from our definition. This might be the cause of the better performance of our model.

5.4 Contributions to prediction

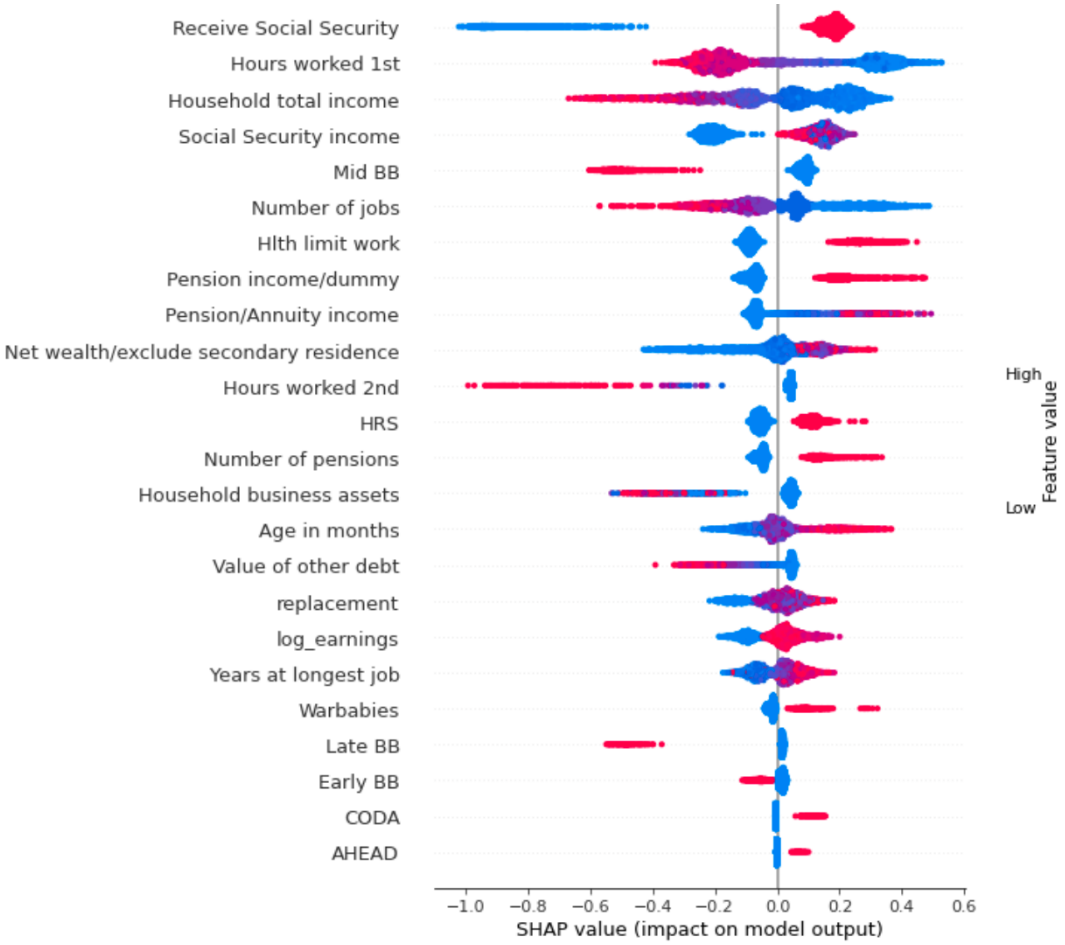
In this section, we study how the variables affect the prediction.

Figure 1: Dependence plot of the retirement case



Note: The horizontal line is the line that represents zero Shapley values. In the years at longest job, the value -100 represents the missing value for this variable. The reason for missing is expected to be no jobs during the respondent's lifetime.

Figure 2: Summary plot of the Shapley values for the retirement case



5.4.1 Retirement case

By looking at Figure 1 and Figure 2, we can see how the variables are affecting the prediction. If the data point gives a positive Shapley value for a certain variable, it means that the value of the data point contributes to an increase in the predicted probability of retirement by including that variable in the model. Figure 1 plots all the Shapley values of different data points while Figure 2 shows a summary of Shapley values. The following is a brief explanation how to understand Figure 2. The horizontal axis represents the Shapley value of each given variable listed in the left part of the panel. The red dots represent a higher value of the feature values. Thus, if the red dots are distributed on the positive side of Shapley values while the blue dots are distributed on the negative side, the variable increases the probability of retirement according to the model. For example, Early Baby-Boomers (Early BB), Mid Baby-Boomers (Mid bb), Late Baby-Boomers (Late BB) are variables that indicate the respondent's birth cohort. The Early BB represent those

who are born between 1948 to 1953, the Mid BB represents those born between 1954 to 1959, and the Late BB represents those born between 1960 to 1965. Detailed definitions of other birth cohorts are given in Appendix E in Table E.3. Figure 2 shows that the Shapley values for the Early Baby-Boomers (Early BB), Mid Baby-Boomers (Mid bb), and the Late Baby-Boomers (Late BB) have red dots distributed in the negative domain of the Shapley values. This means that if the respondent is born after 1947, this fact contributes to a lower probability of retirement because, for those whose birth cohort is after 1947, the Shapley values are mostly negative when the dummy variables are 1.

Figure 1 shows the Shapley values of variable with more than two values including continuous variables. A general pattern we can find from Figure 1 is a highly nonlinear pattern of Shapley values. For example, if we look at the last panel in Figure 1, there is a W like pattern in the Shapley values of positive years at longest job. Although Shapley values do not revel causation, we can see that there is a highly non-linear relationship between the years at the longest job and the predicted probability of retirement.

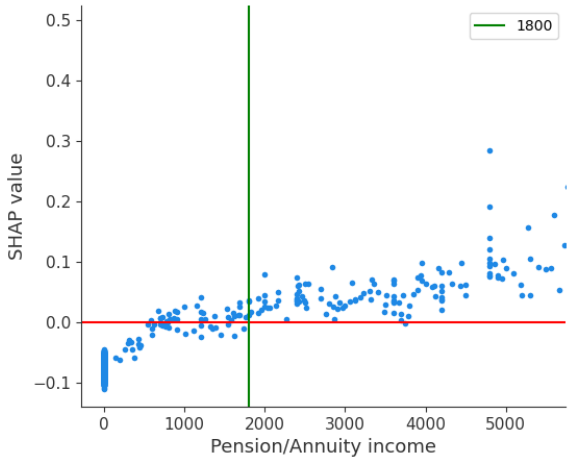
Bellow, we summarize some key findings. First, the third panel in the first row of Figure 1 shows that higher net wealth (excluding secondary residence) contributes to a higher probability of retirement. Conversely, those with low net wealth have a low predicted probability of retirement. This suggestively confirms the liquidity constraint argument of retirement in which individuals with insufficient savings cannot retire when they want (Blundell et al., 2016; Gustman and Steinmeier, 2005).

Second, our results show that the detailed pension type is not important for prediction but the information whether the respondent has a private pension is important. HRS data is designed to collect detailed information on individuals' pension plans. For example, the survey asks the respondent, if any, what type of pension plan does the respondent hold. If the respondent holds more than one, there are information for each pension plan up to 4 plans. However, the selected variables in our analysis are only the pension income dummy variable (*Pension income/dummy*) and the Amount of pension or annuity income (*Pension/Annuity income*). As we can see from the third panel in the second row of Figure 1, as the pension or annuity income becomes positive, the Shapley values are mostly distributed in the positive domain.

If we look closely to the Shapley values of pension or annuity income variable, we can check that the Shapley values are mostly positive after the annual value equal to \$1,800. If we cut the top 20 % of pension or annuity income variable, the result is as in figure 3. We can observe that the bottom 80 % of the data have pension or annuity income below \$6,000 annually and the Shapley values are positive after \$1,800.

Private pensions can differ greatly by plans and individuals but are commonly catego-

Figure 3: Shapley values of pension or annuity income after cutting the top 20%



rized as two groups: Defined benefit (DB) and defined contribution (DC) plans. DB plans are similar to the US Social Security where benefits are typically a function of earnings on the job, years at the job, and age (Blundell et al., 2016). DC plans are more like subsidized savings plans. Blundell et al. (2016) note that DC plans are gaining popularity over DB plans due to increased life expectancy. It is documented that DC plans provide less incentive to exit the labor force than the DB plans. However, when it comes to prediction, our results show that the information about the detailed plan is less important then the information about whether one has a private pension.

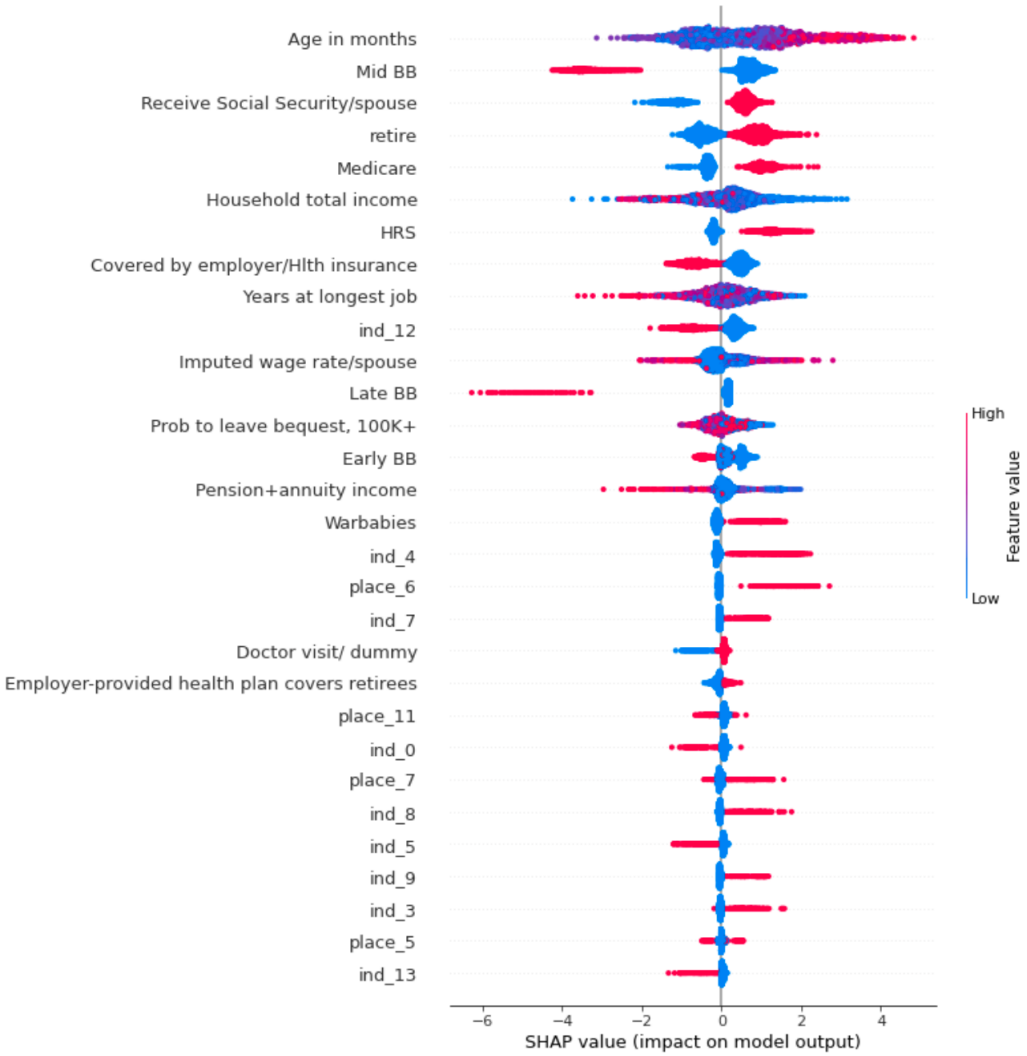
Third, our results give suggestive evidence about job transition. For the main job of the respondent, those who work less than 20 hours of work per week in the previous wave (*Hours worked 1st*) have mostly positive Shapley values. This means that higher values hours of work in the main job contribute to a lower probability of retirement in the next wave. For the second job (*Hours worked 2nd*), those who work more hours per week in the previous wave have lower predicted probability of retirement. Combining the two pattern implies that those who work less hours per week have lower predicted probability of retirement. Although the two pattern do not establish causation, they suggest that there are people in job transition, i.e., individuals sequentially work full-time, part-time, and then become fully retired. Maestas (2010) shows that there is a large portion of retirees following a nontraditional retirement path that involves partial retirement.

Some of the selected variables are related to household business assets and debt. Having business assets decreases the probability of retirement. We can observe in Figure 2 that red dots are distributed on the negative side of the Shapley values for *Household business assets*. Figure 1 also shows that individuals with a positive amount of business assets

mostly have a negative Shapley value. These findings provide suggestive evidence that self-employed individuals have a different pattern of retirement (Blundell et al., 2016). Our results show that self-employed seem to retire later on average than those who are not self-employed. Meanwhile, a positive amount of value of "other debts" gives negative Shapley values (*Value of other debt*). This result is intuitive since those who have more "other debts" would need to work more to pay back their debts and delay retirement. Other debts include credit card balances, medical debts, life insurance policy loans, loans from relatives, and so forth.

5.4.2 Social Security case

Figure 4: Summary plot of the Shapley values for Social Security case



Birth cohort and age are the most important variables that predict Social Security sta-

tus. If a respondent is born after 1948, it contributes to a lower probability of receiving Social Security.

Among the industry codes, "Professional and Related Services" is the most important dummy variable that contributes to the prediction of Social Security. That is, if the respondent worked at "Professional and Related Services" at the longest reported job, then this fact contributes to a lower probability of receiving Social Security. "Professional and Related Services" includes physicians, dentists, legal services, Elementary and secondary schools, education services, social services, engineering, architectural, accounting services, and more.

Being covered by health insurance from her/his current or previous employer (*Covered by employer/Health insurance*) contributes to a lower probability of receiving Social Security. This is consistent with the view that those who lose their health insurance when retired would prefer to retire later than would otherwise be the case. This is further confirmed by another finding: if the employer's health insurance covers retirees (*Employer-provided health plan covers retirees*), it is more likely that this respondent is receiving Social Security. This implies that those who have an employer-provided health plan covering retirees are not constrained to their job due to health insurance. Thus, they can retire and claim Social Security when they want, relative to those who do not have such an employer-provided health plan.

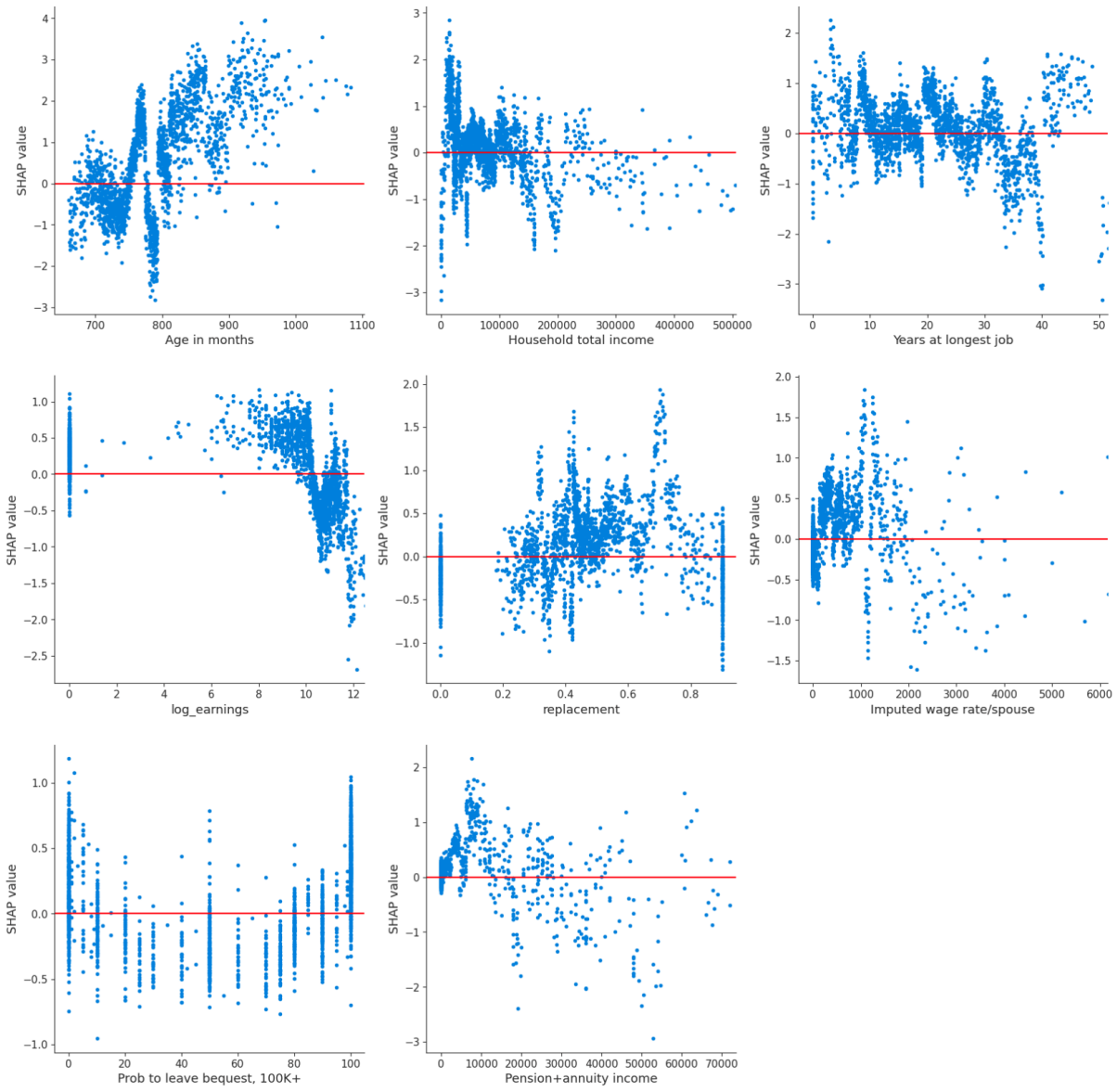
Figure 5 shows a dependence plot for the Social Security case for variables with more than two values. As we can see from Figure 5, it is hard to find a systematic pattern of Shapley values of the continuous variables. However, there are some noticeable parts in Figure 5. The upper left panel shows that the Shapley values increase sharply before age 65 (780 months) and drop sharply around age 66 (800 months). After age 70, most of the data points give positive Shapley values, indicating no incentive to delay Social Security after age 70. There is a sudden jump to positive Shapley values in the years of work at the longest job around 20, 30 years and after 40 years. Further research is needed to know the relationship between the Social Security claiming decision and the aforementioned peculiarities.

6 Robustness checks

In this section, we discuss three potential drawbacks of our results. We also provide suggestive evidence that our results are not affected by those potential drawbacks.

First, one possibility would be that the important information in the non-selected variables of the MINT model is already reflected into the selected variables. For example,

Figure 5: Dependence plot of the Social Security case



Note: The horizontal line is the line that represents zero Shapley values.

education or race of the respondent might be already influencing the household total income or years worked at the longest reported job. Second, the time effects, such as year specific events, are not considered in our model and they might be potentially important. Third, the definition of retirement in the MINT model and our dataset is different. In the MINT model, the individual who worked less than 20 hours per week is defined as retired. However, the retirement is defined as those who are considered as themselves as retired. The difference in definition might be the cause of difference in performance.

We check whether our results change after attempting to address three potential drawbacks. The first drawback is assessed by fitting each selected variable with the non-selected variables from the MINT model, using only the residuals for prediction. When fitting each selected variable, we employ polynomials up to the power of 2 for each non-selected MINT variable and include their interactions. Our goal is to determine whether the remaining variation in the selected variables contributes to prediction. The second drawback is checked by demeaning the data for each interview period to eliminate a common time effect. The third problem is checked by changing the definition of retirement to those who work under 20 hours per week as in MINT.

Our method outperforms the MINT model after considering different specifications. The results are summarized in Table 6 and 7. The column named "Residuals" in Table 6 contains the result when we use the residuals of the selected variables after fitting with the non-selected MINT variables. The difference between the true and predicted beneficiary rates in the test data is 2.89 for our method, which is smaller than that of the MINT model (4.74). This provides suggestive evidence that there is information in our selected variables rather than the non-selected variables that help prediction of retirement status and Social Security status.

The column of named "Demeaned" in Table 6 shows the result when we demean the variables to eliminate common time effects. The result of our gradient-boosted tree method is better when we demean the data, yielding a difference of 0.4 between the true and predicted beneficiary rates. The last column of Table 6 shows the result when we demean the variables and use the residuals. Our model still outperforms the MINT model.

Table 7 shows the result when we redefine retirement following MINT. We define an individual as retired if the individual works less than 20 hours per week. Our gradient-boosted tree method yields 0.81 difference between the true and predicted beneficiary rate while the MINT model yields 4.30. The last column of Table 7 shows the result when we use only the residuals and redefine retirement as MINT. Our model still outperforms the MINT model even after redefining retirement.

Table 6: Robustness check with two-step prediction

	MINT	Original GB	Residuals	Demeaned	Demeaned and residuals
% of beneficiaries	77.19%	78.04%	78.07%	79.46%	79.46%
Predicted % of beneficiaries	81.93%	79.08%	80.96%	79.86%	81.36%
Difference	-4.74	-1.04	-2.89	-0.40	-1.90

The percentage of beneficiaries differ between data because the cleaning process was redone for each case.

Table 7: Two-step prediction after redefining retirement

	GB	MINT	Residuals
% of beneficiaries	78.58%	76.53%	78.58%
Predicted % of beneficiaries	79.39%	80.83%	80.46%
Difference	-0.81	-4.30	-1.88

The percentage of beneficiaries differ between data because the cleaning process was redone for each case.

7 Comparison with the previous literature

In this section, we compare our results with the previous literature. Table 8 summarizes the selective studies, while Table 9 compares the selected variables in our analysis to those in the previous literature. Since retirement and Social Security are not independent decisions, we do not distinguish whether the variables are affecting only retirement or Social Security claiming decisions.

First, age and birth cohort are widely accepted features for both empirical and theoretical research related to retirement and Social Security claiming. This is because age and birth cohort convey the information about the eligibility for Social Security. For example, the rule for eligibility changes as the birth cohort changes. For those who are born between 1943 to 1954, the normal retirement age is set to age 66. Then, the normal retirement age gradually increases by 2 months until it reaches the age of 67 for those who are born in 1960 or later. The differential policies between different birth cohorts are expected to yield different retirement and Social Security claiming patterns.

The two selected variables related to health are first, whether health problem limits work and second, whether the respondent visited a doctor during the past 2 years. There is a lot of previous literature that points to health as an important factor that affects either retirement or Social Security claiming decisions, although the degree of significance differs (Shoven, 2018; French, 2005; Blundell et al., 2016). For example, Li et al. (2008) show that health limitation at work increases the probability of early claiming. This is consistent with our results in the retirement case where observations that experience health limitation at work help the model to predict higher retirement probability.

Among the variables related to income, the importance of household total income is widely emphasized (von Wachter, 2009; Glickman and Hermes, 2015; Rust and Phelan, 1997; Coile et al., 2002b; Shoven, 2018). Also, the importance of wealth is widely recognized in the form of liquidity constraint faced by those who want to retire (Blundell et al., 2016; Shoven, 2018; Haurin et al., 2022).

Our analysis suggests that the value of other debts decreases the probability of re-

tirement. This is consistent with Butrica and Karamcheva (2018) that show mortgage or credit card debt decreases the probability of claiming.

Our results related to the job history are similar to Glickman and Hermes (2015), who show that job history is an important factor that determines the timing of claiming. They show that those who worked for at least 35 years are more likely to claim early. In Figure 1, which is the case for retirement, more years at the longest reported job seems to contribute to an increase in the probability of retirement before 40 years of work. Moreover, in Figure 5, which is the case of Social Security, there is a sudden jump in Shapley values to all positive values between 40 years to 50 years, indicating that years of work between 40 to 50 make the model increase the probability of claiming.

However, Figure 1 shows that the Shapley values are negative after 40 years of work at the longest job for the retirement case while Figure 5 shows mostly positive Shapley values for the Social Security case between 40 and 50 years of work at the longest job. This implies that those between 40 and 50 years of work tend to not retire and claim Social Security. This is a pattern that needs further investigation. It is worth noting that since we are only looking at the contributions of each data point of a certain variable, we can only show a tendency of this behavior.

Furthermore, the selected variable that indicates whether the respondent's longest job was in "Professional and Related Services" is mentioned as a significant variable in Glickman and Hermes (2015). They show that those who work in less physically demanding jobs, such as the category "Professional and Related Services", tend to not claim Social Security early. Our results also show that those who worked or work in "Professional and Related Services" contribute to a lower probability of claiming Social Security.

The hours worked per week in 1st job in the previous wave give consistent results with French (2005). The life cycle model in French (2005) shows that there is a sharp decline in hours worked after age 59, which is consistent with our results. In Figure 1, the first panel shows that as the hours worked decrease, it is more likely for the model to predict a higher probability of retirement.

Two variables, *hours worked per week in 2nd job in previous wave* and *the number of jobs over employment history* need further research to know the theoretical and empirical effects on retirement and Social Security claiming decisions. As mentioned in Section 5.4.1, hours worked per week in the 2nd job might indicate job transition before retirement. The number of jobs might indicate that the respondent did not have a stable job over the respondent's employment history. Our results show that a larger number of jobs over the life history leads the model to predict a lower probability of retirement.

As mentioned in Section 5.4.1, our results suggest that having pension or annuity

income is important information for predicting retirement and Social Security status. If a respondent has a pension or annuity income, this fact contributes to an increase in the predicted probability of retirement. This is consistent with the results in Hurd et al. (2004) who show that a high level of pension savings increases the likelihood of early claiming. However, it is worth noting that, by looking at the Shapley values, our results for the Social Security case do not show a systematic impact of pension or annuity income on the predicted probability.

The importance of Medicare and employer-provided health insurance in explaining the retirement and Social Security claiming decisions is widely documented in previous studies (Rust and Phelan, 1997; French and Jones, 2011; Blundell et al., 2016). Our results are consistent with previous literature, indicating that individuals with employer-provided health insurance, yet ineligible for Medicare, tend to delay both retiring and receiving Social Security benefits. Furthermore, our findings show that individuals covered by Medicare contribute to an increased probability of Social Security claiming, while those with employer-provided health insurance contribute to a decreased probability of Social Security claiming. Furthermore, we find that those who have an employer-provided health plan that covers retirees help the model increase the probability of claiming. These results give suggestive evidence that there are respondents who cannot retire because their health insurance is tied to their work status and they are not eligible for Medicare.

The variable related to the bequest motive is widely recognized as an important factor in life cycle models (Modigliani, 1988; Kotlikoff and Summers, 1981). Moreover, Coile et al. (2002a) show that bequest motives make individuals delay claiming Social Security for further wealth. Although our results do not show a systematic patter in Shapley values, such a variable is still considered as important for the prediction of Social Security claiming in our analysis.

8 Conclusion

Our results show that machine learning, particularly gradient boosted trees, helps improve predictions of individual retirement and Social Security claiming decisions. The dataset is from the Health and Retirement Study, which comprises over 580 variables. In terms of predicting the percentage of people receiving Social Security, our model achieves an error rate of less than 1%, while the benchmark model employed by the Social Security Administration results in a 5% error rate. Moreover, when forecasting Social Security claiming behavior between 2018 and 2020 based on data collected before 2018, our proposed model achieves a low error rate of 0.2%, whereas the benchmark model produces

Table 8: Literature review

Retirement	
Paper	Summary
Blundell, French, and Tetlow (2016)	Summary of research on retirement incentives and labor supply. Couple of factors that affect retirement: Health, substitution effect, wealth effects, incentives from public pensions, incentives from private pensions, incentives from health insurance, behavioral aspects.
Rust, Phelan (1997), French, Jones (2011)	Social Security and Medicare provides a large incentive to retire for those who are not covered by employer-provided health insurance after retirement.
Hamermesh (2002)	Discovers that couples' retirement roughly coincides.
Social Security	
Paper	Summary
Coile et al. (2002)	More likely to delay if longer life expectancy; and claiming delays follow a U-shape pattern in wealth (more wealth delays claiming at first but causes reduction in delays for further wealth due to bequest motive); having a pension result in shorter delay
Hurd et al. (2004)	High level of pension savings increases the likelihood of early claiming. Lower subjective probability of survival increases early claiming and retirement but not large effect.
Von Wachter (2009)	Low-income households might claim early due to high replacement rate.
Li et al. (2008)	Health limitation at work increase the probability of early claiming.
Glickman and Hermes (2015)	Blue-collar jobs claim earlier. Those out of the workforce or had longer work histories more likely to claim early. Lower life expectancy leads to higher probability of claiming early. Greater income and wealth lead to more delayed claiming
Butrica and Karamcheva (2018)	Mortage or credit card debt decreases the probability of claiming.
Shoven et al. (2018)	Ask the reason of early claiming: work, health, liquidity, and expectation related reasons. Also, claiming was affected by behavioral aspects: normative retirement age, financial literacy
Huang et al. (2022)	Change in house value did not have an effect except the 2002-2006 housing boom period: more housing wealth increased delayed claiming.
Haurin et al. (2022)	Financial stress increases the probability of delayed claiming at age 62.

an error rate exceeding 2%. Our results in the retirement case reveal that variables related to respondents' job histories are important, yet they are not considered in the benchmark model. Conversely, demographic characteristics, education, and spouse-related variables are not identified as significant predictors in our model, whereas the benchmark model incorporates these variables. In the Social Security case, our analysis highlights the importance of variables related to job history, health insurance, and bequest-related factors, which are not accounted for in the benchmark model. However, earnings, self-employment status, and education-related variables are not found to be significant in our analysis, whereas the benchmark model includes those variables. Additionally, our study employs Shapley values to evaluate the non-linear contributions of variables to predictive outcomes, underscoring the complex nature of the relationships between predictors and

Table 9: Comparison between the previous literature and selected variables

Category	Selected variables	Retirement or SS	Previous literature
Age	Age in months	Retirement, SS	All
	Respondent birth cohort	Retirement, SS	All
Health	Health problem limits work	Retirement	Li et al. (2008)
	Doctor visit/ dummy	SS	Shoven (2018)
Income/Wealth	Household total income	Retirement, SS	von Wachter (2009); Glickman and Hermes (2015)
	Net wealth/exclude secondary residence	Retirement	Shoven (2018)
	Value of other debt	Retirement	Butrica and Karamcheva (2018)
Earnings	Earnings/prev. wave	SS	Blundell et al. (2016)
Retirement	Retired	SS	Glickman and Hermes (2015)
Job history	Years at longest reported job	Retirement, SS	Glickman and Hermes (2015)
	Hours worked per week in -1st job in previous wave	Retirement	French (2005)
	Hours worked per week in -2nd job in previous wave	Retirement	
	Number of jobs over -employment history	Retirement	
	Industry code of longest job	SS	Glickman and Hermes (2015)
Self-employment	Household business assets	Retirement	Blundell et al. (2016)
Social Security benefits	Receive Social Security benefits	Retirement	
	Social Security income	Retirement	
Private pension	Pension/Annuity income	Retirement, SS	Hurd et al. (2004)
	Pension income/dummy	Retirement	Hurd et al. (2004)
	Number of pensions currently receiving	Retirement	Hurd et al. (2004)
Health insurance	Medicare	SS	Rust and Phelan (1997); French and Jones (2011)
	Covered by employer/Hlth insurance	SS	Rust and Phelan (1997); French and Jones (2011)
	Employer-provided health plan covers retirees	SS	Rust and Phelan (1997); French and Jones (2011)
Household joint decision	Receive Social Security/spouse	SS	Hamermesh (2000)
	Imputed wage rate/spouse	SS	
Demographics	Birth place	SS	
Bequest	Probability to leave bequest over 100K	SS	Coile et al. (2002a); Mukherjee (2022, 2018); Lee and Tan (2023)

predictions. Although this paper primarily focuses on prediction, it provides suggestive evidence for future research in retirement modeling, both empirically and theoretically.

References

Albanesi, Stefania and Domonkos F. Vamossy (2019) "Predicting Consumer Default: A Deep Learning Approach," 10.48550/ARXIV.1908.11498.

Azinovic, Marlon, Luca Gaegauf, and Simon Scheidegger (2022) "DEEP EQUILIBRIUM NETS," *International Economic Review*, 10.1111/iere.12575.

Blundell, R., E. French, and G. Tetlow (2016) "Chapter 8 - Retirement Incentives and Labor Supply," 1 of Handbook of the Economics of Population Aging, 457–566: North-Holland, https://doi.org/10.1016/bs.hespa.2016.10.001.

Butrica, Barbara A. and Nadia S. Karamcheva (2018) "In Debt and Approaching Retirement: Claim Social Security or Work Longer?" *AEA Papers and Proceedings*, 108, 401–06, 10.1257/pandp.20181116.

Coile, Courtney, Peter Diamond, Jonathan Gruber, and Alain Jousten (2002a) "Delays in claiming social security benefits," *Journal of Public Economics*, 84 (3), 357–385, https://doi.org/10.1016/S0047-2727(01)00129-3.

——— (2002b) "Delays in claiming social security benefits," *Journal of Public Economics*, 84 (3), 357–385, `https://ideas.repec.org/a/eee/pubeco/v84y2002i3p357-385.html`.

Fouliard, Jeremy, Michael Howell, and Hélène Rey (2020) "Answering the Queen: Machine Learning and Financial Crises," Working Paper 28302, National Bureau of Economic Research, 10.3386/w28302.

French, Eric (2005) "The Effects of Health, Wealth, and Wages on Labour Supply and Retirement Behaviour," *Review of Economic Studies*, 72 (2), 395–427, `https://EconPapers.repec.org/RePEc:oup:restud:v:72:y:2005:i:2:p:395-427`.

French, Eric and John Bailey Jones (2011) "THE EFFECTS OF HEALTH INSURANCE AND SELF-INSURANCE ON RETIREMENT BEHAVIOR," *Econometrica*, 79 (3), 693–732.

Glickman, Mark M. and Sharon Hermes (2015) "Why Retirees Claim Social Security at 62 and How It Affects Their Retirement Income: Evidence from the Health and Retirement Study," *The Journal of Retirement*, 2 (3), 25–39, 10.3905/jor.2015.2.3.025.

Gustman, Alan L and Thomas L Steinmeier (2001) "Retirement and Wealth," Working Paper 8229, National Bureau of Economic Research, 10.3386/w8229.

Gustman, Alan L. and Thomas L. Steinmeier (2005) "The social security early entitlement age in a structural model of retirement and wealth," *Journal of Public Economics*, 89 (2), 441–463, https://doi.org/10.1016/j.jpubeco.2004.03.007.

Hamermesh, Daniel S (2000) "Togetherness: Spouses' Synchronous Leisure, and the Impact of Children," Working Paper 7455, National Bureau of Economic Research, 10.3386/w7455.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009) *The elements of statistical learning: data mining, inference, and prediction*, New York, NY, USA.

Haurin, Donald, Stephanie Moulton, and Caezilia Loibl (2022) "The relationship of financial stress with the timing of the initial claim of U.S. Social Security retirement income," *The Journal of the Economics of Ageing*, 21, 100362, https://doi.org/10.1016/j.jeoa.2021.100362.

Huang, Naqun, Jing Li, and Amanda Ross (2022) "Housing wealth shocks, home equity withdrawal, and the claiming of Social Security retirement benefits," *Economic Inquiry*, 60 (2), 620–644, 10.1111/ecin.13058.

Hurd, Michael D., James P. Smith, and Julie M. Zissimopoulos (2004) "The Effects of Subjective Survival on Retirement and Social Security Claiming," *Journal of Applied Econometrics*, 19 (6), 761–775, http://www.jstor.org/stable/25146321.

Kotlikoff, Laurence J. and Lawrence H. Summers (1981) "The Role of Intergenerational Transfers in Aggregate Capital Accumulation," *Journal of Political Economy*, 89 (4), 706–732, http://www.jstor.org/stable/1833031.

Kumar, I. Elizabeth, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler (2020) "Problems with Shapley-value-based explanations as feature importance measures," in III, Hal Daumé and Aarti Singh eds. *Proceedings of the 37th International Conference on Machine Learning*, 119 of Proceedings of Machine Learning Research, 5491–5500: PMLR, 13–18 Jul, https://proceedings.mlr.press/v119/kumar20e.html.

Lee, Siha and Kegon T.K. Tan (2023) "Bequest motives and the Social Security Notch," *Review of Economic Dynamics*, 51, 888–914, https://doi.org/10.1016/j.red.2023.09.001.

Li, Xiaoyan, Michael Hurd, and David Loughran (2008) "The Characteristics of Social Security Beneficiaries Who Claim Benefits at the Early Entitlement Age."

Lundberg, Scott M and Su-In Lee (2017) "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, 30.

Maestas, Nicole (2010) "Back to Work: Expectations and Realizations of Work After Retirement," *The Journal of human resources*, 45, 718–748, 10.1353/jhr.2010.0011.

Modigliani, Franco (1988) "The Role of Intergenerational Transfers and Life Cycle Saving in the Accumulation of Wealth," *Journal of Economic Perspectives*, 2 (2), 15–40, 10.1257/jep.2.2.15.

Molnar, Christoph (2022) *Interpretable Machine Learning*, 2nd edition, https://christophm.github.io/interpretable-ml-book.

Mukherjee, Anita (2018) "Time and Money: Social Security Benefits and Intergenerational Transfers," *AEA Papers and Proceedings*, 108, 396–400, 10.1257/pandp.20181115.

——— (2022) "Intergenerational Altruism and Retirement Transfers," *Journal of Human Resources*, 57 (5), 1466–1497, 10.3368/jhr.58.1.0419-10140R3.

Rust, John and Christopher Phelan (1997) "How Social Security and Medicare Affect Retirement Behavior In a World of Incomplete Markets," *Econometrica*, 65 (4), 781–831, http://www.jstor.org/stable/2171940.

Shoven, Slavov S.N. Wise D.A., J.B. (2018) "Understanding Social Security claiming decisions using survey evidence.," *J. Financial Planning*, 31 (11), 35–47.

Smith, Karen E., Aaron Williams, and Stipica Mudrazija (2021) "Modeling Income in the Near Term-version 8," final report, Urban Institute.

Sundararajan, Mukund and Amir Najmi (2020) "The Many Shapley Values for Model Explanation," in III, Hal Daumé and Aarti Singh eds. *Proceedings of the 37th International Conference on Machine Learning*, 119 of Proceedings of Machine Learning Research, 9269–9278: PMLR, 13–18 Jul, https://proceedings.mlr.press/v119/sundararajan20b.html.

von Wachter, T. (2009) "The effect of labor market trends on the incentives to incidence for claiming social security benefits early.," Working Paper RRC NB09-06, National Bureau of Economic Research, 10.3386/w8229.

Appendix A Social Security benefits

In this section, we summarize the rules that determine the eligibility and the amount of Social Security retirement benefits.

You may be eligible for monthly Social Security benefits if you are: a retired insured worker age 62 or over, a disabled insured worker who has not reached full retirement age, a spouse of a retired or disabled worker entitled to benefits, a divorced spouse of a retired or disabled worker entitled to benefits, the dependent of a retired, disabled, or deceased insured worker. To receive benefits based on your earnings, you should be 62 or older and worked and paid Social Security taxes for 10 years or more. The detailed requirements for other categories are in the SSA handbook.[9] One example of a detailed category is for spousal benefits given to divorced spouses. To receive benefits based on your ex-spouse's earnings, you need to be married to the worker for at least 10 years.

The Social Security benefits are based on "average indexed monthly earnings" (AIME), which is the monthly average over 35 years of a worker's indexed earnings. The earnings are indexed to reflect the general rise in the standard of living over the years. The indexing is based on the national wage indexing series. Prior to the year that an individual reaches 60, the earnings are indexed to the year that the individual reaches 60. On and after the year that the individual reaches 60, the earnings are taken at face value. For example, if the individual is 60 in 2022, the earnings in 2021 are multiplied by the ratio of 63,795.13, which is the average wage index for 2022, and divided by 60,575.07, which is the average wage index for 2021.

AIME averages a maximum of 35 years of earnings. Among the years of earnings, SSA chooses those years with the highest indexed earnings and divides the total amount by the number of months in those years. This is the so-called AIME of the individual. The benefits that are paid to an individual are based on the primary insurance amount (PIA), which is a function of AIME. The portion that an individual can receive from the AIME is based on two "bend points". The "bend points" are based on the year the individual first becomes eligible. For example, if an individual first becomes eligible in 2024, the PIA is the sum of 90 percent of the first $1,174 of his/her AIME and 32 percent of his/her AIME over $1,174 and through $7,078 and 15 percent of his/her AIME over $7,078. There is a maximum for the monthly benefit and a cost-of-living adjustment over the years.

The actual amount of benefit may differ from PIA since the individual can retire early or later than the normal retirement age. The normal retirement age is increasing from 65 to 67 gradually and differs by birth cohort. For example, those born prior to or in 1937

[9] https://www.ssa.gov/OP_Home/handbook/handbook.html

are subject to the normal retirement age of 65 while those born after or in 1960 are subject to the normal retirement age of 67. If the individual chooses to retire early and receive benefits, the benefit is reduced by each month before the normal retirement age. For example, an individual can choose to retire at the age of 62 in 2024 and receive monthly benefits that are reduced by as much as 30%. On the other hand, the individual can choose to delay receiving Social Security over the normal retirement age and receive a credit that increases the monthly benefit by 8 percent per year for those born after 1942. There is no delayed credit given after age 69.

Appendix B Comparison of different machine learning techniques

In this section, we present the performance of other machine learning models, which include random forest, lasso, and neural networks, and compare it with the performance of the gradient boosted tree.

A random forest is a bagging method applied to decision trees. The idea is to make many bootstrap sample and train a decision tree on each bootstrap data to decrease the variance of the overall model. If there are B bootstrap sample, we grow a random forest tree T_b for each $b = 1$ to B with randomly selected m variables from all the variables. For classification problems, the overall prediction is made with majority voting over B trees.

Estimators from lasso estimation method are obtained by solving the following problem in which the number of variables is p and the sample size is N. Lasso estimation method shrinks the coefficients and can set the coefficient of some variables to 0.

$$\hat{\beta}^{lasso} = argmin_{\beta} \left[\frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right]$$

To represent the neural network with notations, I borrow the notations from Azinovic et al. (2022). Let K be the number of layers of the network, m_i be the number of neurons inside layer i, and τ_i be the activation function of layer i. Let $\rho = \{W_1, W_2, \ldots, W_K, b_1, b_2, \ldots, b_K\}$ be the trainable parameters that the neural network will update. Given hyper-parameters $\{K, \{m_i\}_{i=1}^{K}, \{\tau_i\}_{i=1}^{K}\}$, and trainable parameters

$\rho = \{W_1, W_2, \ldots, W_K, b_1, b_2, \ldots, b_K\}$, a neural network $\mathcal{N}_\rho$ is a map,

$$x \to \mathcal{N}_\rho = \tau_K \left(W_k \ldots \tau_2 (W_2 \tau_1 (W_1 x + b_1) + b_2) \cdots + b_K \right) \tag{1}$$

The object of the neural network is to minimize the loss/cost function, $l(p)$.

The package RandomForestClassifier from scikit-learn is used to fit a random forest. The maximum depth is set to 6 and the criterion is set to entropy. The other hyper-parameters are set to the default value. Next, the neural network structure reported here has two hidden layers with number of neurons equal to 128 for each hidden layer. The max iteration is 500 and the other hyper-parameters are set as the default value. The package is from scikit-learn named MLPClassifier. Lastly, the lasso regression has an alpha equal to 1 and maximum iteration of 3,000. When doing the prediction after selecting variables from lasso, we did not estimate the coefficients on the training data again.

Table B.1: Performance comparison between different machine learning methods: retirement case

Method	Gradient boosted tree	Random forest	Lasso	Neural network
AUC	0.8392	0.8107	0.7839	0.6258

Appendix C Selected hyper parameters

Here we report the hyper parameters used in our main analysis. Table C.2 shows the selected hyper parameters of the gradient boosted tree for each retirement and Social Security case. The hyper parameters are selected after 5-fold cross validation using the scoring method as AUC.

Table C.2: Gradient-boosted decision tree hyper-parameters

hyper-parameters	Retirement case	Social Security case
Maximum depth of a tree	5	3
Minimum sum of instance weight (hessian) needed in a child	1	1
Number of trees	500	500
Subsample ratio of the training instances	0.5	1
subsample ratio of columns when constructing each tree	1	1
evaluation metric	logloss	logloss
gamma	1	0
lambda	3	1
learning rate	0.01	0.3

Appendix D Data cleaning process

In this section, we summarize the data cleaning process. The codes used for data cleaning contain further details.

1. Obtain respondents that retired for a specific interview period, which is called a wave. For instance, the interview period of wave 14 is from 2016 to 2018. We use waves from 4 to 14. The sample covers the period from 1996 to 2018.

2. Obtain not-retired respondents from the wave 14 (most recent interview period).

3. Among the variables, flu shot, cholesterol, mammogram, Pap smear, Prostate, and back variables are only interviewed on odd number waves. For even number waves, the respondents can answer that the question was asked in the previous wave. If the respondents answered that they were asked in the previous wave, we use the previous wave answers for the variables.

   The occupation code for the longest job and the industry code for the longest job is not consistent. For example, the respondent's occupation code for the longest job can be in the 1980 census code, 2000 census code, or 2010 census code. Meanwhile, the AHEAD cohort is classified with another 9 categories. We try our best to match the occupation code to the 1980 census code with the provided data.

4. We keep the longest job occupation code and longest job industry code recorded in the 1980 census code.

5. If the respondent answers ".b", which represents that the occupation or industry code is in other census code, we replace the ".b" with the other census code.

6. By using the detailed classification of occupation code and industry code, we match similar names between the 1980 census code and other census codes. For instance, "Management occupations" in the 2000 census code is matched to "Managerial specialty operation" in the 1980 census code.

7. If there is no match for the name in other census codes, we categorize it in two ways. Let there be a category in another census code that does not have an exact match in names with the 1980 census code. For example, "financial specialists" have no exact match in the names of the 1980 census. We find respondents who reported their occupation code as "financial specialists" and reported their occupation code in the 1980 census code in wave 14. We check if there is a category in the 1980 census code

that most of those respondents are classified. For example, "financial specialists" are mostly "Managerial specialty operation" in the 1980 census code. We categorize "financial specialists" as "Managerial specialty operation". If there is no one category that most of those respondents are categorized, we split those respondents randomly into the 1980 census code as we observe from those who reported in both the 1980 census code and the other census code. For example, for those who reported "craftsmen/foremen/kindred workers", we split those respondents into "mechanics/repair", "construction trade/extractors", and "precision production" with the probability of 52/165, 50/165, and 63/165.

8. When we merge the waves from 4 to 14, some variables are not available among all 11 waves. We drop those variables.

9. We drop variables that contain information about the interview waves.

10. We drop variables that have a erroneous skip pattern specific to 2018 HRS survey.

11. We drop respondents who receive SSDI or SSI.

12. We only keep respondents who are married.

13. We keep respondents whose age are greater or equal to 55.

14. We change the missing reasons to a large negative value. We check and change the value if the negative value overlaps with the variables that can have negative values such as the net value of different assets.

Appendix E The definition of categorical variables

Here, we present the definition of Birth cohort, birth place, and industry code. All the categories are from the HRS.

Table E.3 shows the Birth cohort defined by HRS. For example, the hrs cohort represents the respondents who were born between 1931 to 1941. Table E.4 represents where the respondent is born. More detailed birth place is a masked information that can be obtained after request to HRS. Table E.5 shows the industry code for the longest worked job of the respondents. The industry codes are following the 1980 census code. Some of the respondents do not have their industry code in 1980 census code. We manually match the other year census code into 1980 census code.

Table E.3: Birth cohort

cohort name	year
ahead	before 1924
coda	1924 to 1930
hrs	1931 to 1941
warbabies	1942 to 1947
early_babyboomers	1948 to 1953
mid_babyboomers	1954 to 1959
late_babyboomers	1960 to 1965
not_in_any	after 1965
unmarried	

Table E.4: Birth place

cohort name	year
place_1	New England
place_2	Mid Atlantic
place_3	EN Central
place_4	WN Central
place_5	S Atlantic
place_6	ES Central
place_7	WS Central
place_8	Mountain
place_9	Pacific
place_10	US/NA Division
place_11	Not US/ inc US terr

Appendix F Variable selection process

In this section, we describe the variable selection process in detail. The three important variable lists are shown in Table F.6 and F.7.

A variable is selected as important if the variable enters the important variable list at least twice out of the three lists. The lists of selected variables are given in Table F.8.

Variables with a correlation higher than 0.6 are selectively dropped. This is because we assume that the variables with high correlations have similar information about retirement or Social Security status.

For the retirement case, the following variables are selected under the situation when two or more variables are highly correlated to each other.

- ”Net wealth/ exclude secondary residence” and ”Household non-housing wealth” ”Net non-housing financial wealth” are highly correlated. We only keep ”Net

Table E.5: Industry Code / 1980 census code

cohort name	year
ind_1	Agriculture, Forestry, Fishing
ind_2	Mining and Construction
ind_3	Manufacturing: Non-durable
ind_4	Manufacturing: Durable
ind_5	Transportation
ind_6	Wholesale
ind_7	Retail
ind_8	Finance, Insurance, and Real Estate
ind_9	Business and Repair Services
ind_10	Personal Service
ind_11	Entertainment and Recreation
ind_12	Professional and Related Services
ind_13	Public Administration

wealth/ exclude secondary residence".

- "Weeks worked main job" is highly correlated with "Hours worked 1st job". We only keep "Hours worked 1st job".
- "Weeks worked 2nd job" is highly correlated with Hours worked in the 2nd job. We only keep "Hours worked 2nd job".
- The "Hours worked 1st job" is highly correlated with "Whether self-employed" and the latter is dropped. The information about self-employment is thought to be contained in "Household business assets".
- The "Age when started to receive SS" is highly correlated with "Receive Social Security". We only keep "Receive Social Security".
- "Birth year" is highly correlated with "Birth cohort". We keep the "Birth cohort".

Next, we include one more variable which is "Age in months" to create the list for the retirement case in 1.

For the Social Security case, the following variables are selected under the situation when two or more variables are highly correlated to each other.

- "Covered by Gov plan" is highly correlated with "Medicare". We only keep "Medicare".

- ”Number of doctor visits” is highly correlated with ”Doctor visits (dummy)”. We keep ”Doctor visits (dummy)”.
- ”Birth year” is dropped and we keep ”Birth cohort”.

Table F.6: Variable importance lists for retirement case: Descending order

Impurity base	Permutation base	Shapley values
Receive Social Security	Household total income	Receive Social Security
Birth cohort	Hours worked 1st job	Birth year
Medicare	Whether self-employed	Household total income
Birth year	Number of jobs	Hours worked 1st job
Social Security income	Household business assets	Social Security income
Pension income/dummy	Hlth plan covers retirees #1	Birth cohort
Spouse Birth cohort	Employer-provided health plan covers retirees	Pension/Annuity income
Number of pensions	Health limit work	Number of jobs
Age when started to receive SS	Net non-housing financial wealth	Health limit work
Covered by Gov plan	Receive Social Security	Age when started to receive SS
Spouse had lung disease	Total wealth excluding IRAs	Pension income/dummy
Weeks worked 2nd job	Current pension type #1	Net non-housing financial wealth
Spouse labor force status	Net wealth/exclude secondary residence	Household capital income
Hours worked 2nd job	Spouse birth date	Whether self-employed
Health limit work	Change in total non-housing wealth	Number of pensions
Spouse current balance of DC plan	Weeks worked main job	Household business assets
Whether self-employed	Earnings/2 previous waves	Pension income
Hours worked 1st job	Household non-housing wealth	Household non-housing wealth
Pension/Annuity income	Hlth plan covers retirees/summary	Weeks worked 2nd job
Receive Social Security/Spouse	Value of other debt	Hours worked 2nd job
Household business assets	Month last worked/ spouse	Value of other debt
Pension income	Height	Weeks worked main job
Some difficulty get in/out bed	Spouse Birth place	Age in months
Employer contribution for #3 DC	Mother age	Years at longest reported job
Spouse birth year	Years at longest reported job	Net wealth/exclude secondary residence

Table F.7: Variable importance lists for Social Security case: Descending order

Impurity base	Permutation base	Shapley values
Birth year	Birth year	Birth year
Birth cohort	Retired	Receive Social Security/Spouse
Covered by employer/Hith insurance	Birth cohort	Retired
Retired	Receive Social Security/Spouse	Birth cohort
Covered by Gov plan	Medicare	Age in months
Receive Social Security/Spouse	Medicare/Medicaid HMO monthly premium	Covered by Gov plan
Covered by spouse's Hlth insurance	Household total income	ind code_1980
Gets help dressing	ind code_1980	Medicare
Spouse checked prostate	Income/earnings	Years worked
Status of job history	Imputed wage rate/spouse	Household total income
Spouse age in months	Earnings/2 previous waves	Transportation assets
Hlth plan covers retirees #1	Weeks worked main job	Years at longest reported job
Some difficulty walking several blocks	Chg live 80-100: R/LfTab ratio	Years of Education
Net value of 2nd home	Father age	Pension/Annuity income
Some difficulty dressing	Covered by employer/Hlth insurance	Birth place
Spouse some difficulty lifting/ 10 lbs	Spouse number of marriages	Covered by employer/Hlth insurance
Hlth limit work	High BP	Total wealth excluding IRA
spouse jobs with missing dates	Years at longest reported job	Mother's years education
Doctor visits (dummy)	Chg live 75+: R/LfTab ratio	Prob leave bequest 100K+
Household total income	Spouse earnings/ 2 previous waves	Out of pocket medical expense
Employer-provided health plan covers retirees	Hours worked 1st	Number of times widowed
Spouse Some difficulty walking blocks	Birth place	Change in total non-housing wealth
Spouse covered in retirement	Employer-provided health plan covers retirees	Doctor visits (dummy)
Spouse pension income	Number times divorced	Imputed wage rate/spouse
Pension/Annuity income	Prob leave bequest 100K+	Income/earnings

Table F.8: Variables that enter twice out of three important variables lists: The cross is marking the variables that are dropped due to high correlation with other variables

Retirement case	Social Security case
Birth cohort	Age in months
Health limit work	Birth cohort
Household total income	Doctor visits (dummy)
Net wealth/ exclude secondary residence	Household total income
Value of other debt	Income/earnings
Years at longest reported job	Retired
Hours worked 1st job	Years at longest reported job
Hours worked 2nd job	industry code_1980
Number of jobs	Pension/Annuity income
Household business assets	Medicare
Receive Social Security	Covered by employer/ Hlth insurance
Social Security income	Employer-provided health plan covers retirees
Pension/Annuity income	Receive Social Security/Spouse
Number of pensions	Imputed wage rate/spouse
Pension income/dummy	Birth place
~~Birth year~~	Prob leave bequest 100k+
~~Pension income~~	~~Number of doctor visits~~
~~Household non-housing wealth~~	~~Birth year~~
~~Weeks worked 2nd job~~	~~Covered by Gov plan~~
~~Age when started to receive SS~~	
~~Weeks worked main job~~	
~~Net non-housing financial wealth~~	
~~Whether self-employed~~	