Article

Not peer-reviewed version

Targeted Retirement and Investment Plans for Under-Saved Families: A Machine Learning Approach Using Survey of Consumer Finance Data

Srimanvas Veludandi *

Posted Date: 28 November 2025

doi: 10.20944/preprints202511.2282.v1

Keywords: retirement savings; machine learning; XGBoost; random forest; support vector machine; financial planning; predictive modeling; survey of consumer finance; under-saved families; targeted interventions; cluster analysis; scenario modeling; feature importance; housing debt; financial literacy

 doi:10.20944/preprints202511.2282.v1

Article

Targeted Retirement and Investment Plans for Under-Saved Families: A Machine Learning Approach Using Survey of Consumer Finance Data

Srimanvas Veludandi

Harrisburg University of Science and Technology; srimanvas@gmail.com

Abstract

This research addresses the critical problem of retirement under-saving among American families by developing a comprehensive predictive machine learning system to identify at-risk households and recommend targeted intervention strategies. The retirement savings crisis affects millions of Americans, with nearly 50% of adults aged 60 and above having income below basic needs thresholds. Traditional approaches to this problem have been largely descriptive, identifying the scope of the crisis without providing actionable tools for intervention. This study bridges that gap by combining advanced machine learning techniques with economic analysis to create a practical decision-support system. Using longitudinal data from the Survey of Consumer Finances (SCF) spanning 1989 to 2022, I developed and validated multiple classification models including Support Vector Machines (SVM), Random Forest, and Extreme Gradient Boosting (XGBoost). The dataset encompasses 72 observations across six age groups and 12 survey waves, with over 40 financial variables including income, assets, debts, and derived financial ratios. The XGBoost model emerged as the best performer, achieving 100% test accuracy on optimal configuration and 93.33% mean accuracy across rigorous multi-seed validation with a ROC-AUC of 0.982, demonstrating excellent discriminative ability and model stability. The study makes three major contributions to the literature. First, feature importance analysis revealed that Home-Secured Debt (mortgage debt) is the strongest predictor of retirement preparedness with an importance score of 0.35, followed by Non-financial Assets (0.18) and Debt-to-Income Ratio (0.10). This finding challenges conventional wisdom that income level is the primary determinant of retirement readiness and suggests that housing decisions critically affect long-term financial security. Second, K-Means cluster analysis identified three distinct under-saver segments requiring different intervention strategies: AT RISK families needing automatic enrollment, MODERATE families requiring debt consolidation, and CRITICAL families needing urgent catch-up strategies. Third, scenario modeling quantified the value of early intervention, demonstrating that families under 35 can accumulate approximately $708,000 more in retirement wealth through aggressive saving strategies compared to maintaining current trajectories. The practical implications are significant for financial advisors, employers, and policymakers. This research provides a validated, deployable system for identifying at-risk families, segmenting them into actionable groups, and calculating the return on investment for various intervention strategies. The findings suggest that the retirement savings crisis, while severe, is addressable through targeted, data-driven interventions that account for heterogeneity in family financial situations.

Keywords: retirement savings; machine learning; XGBoost; random forest; support vector machine; financial planning; predictive modeling; survey of consumer finance; under-saved families; targeted interventions; cluster analysis; scenario modeling; feature importance; housing debt; financial literacy

 doi:10.20944/preprints202511.2282.v1

1. Introduction

Background and Motivation

Retirement security represents one of the most pressing financial challenges facing American families in the twenty-first century. The traditional three-legged stool of retirement income—Social Security, employer pensions, and personal savings—has become increasingly unstable as defined-benefit pension plans disappear, Social Security faces long-term funding challenges, and personal savings rates remain inadequate for millions of households. According to the National Council on Aging (2024), the average income of nearly 50% of adults aged 60 and above falls below the standard needed to afford their basic necessities, including housing, food, healthcare, and transportation.

The magnitude of this crisis is staggering. AARP (2023) reports that nearly half of older Americans rely almost entirely on Social Security benefits, which average only approximately $1,900 per month—far below typical living expenses in most American communities. The Center for Retirement Research at Boston College estimates that more than half of working-age households are at risk of being unable to maintain their pre-retirement standard of living in retirement. These statistics represent millions of individuals facing financial insecurity, reduced healthcare access, housing instability, and diminished quality of life during what should be their golden years.

The retirement crisis is not merely a personal financial problem but a significant macroeconomic and social policy challenge. Inadequately prepared retirees place increased demands on social safety net programs, healthcare systems, and family support structures. The economic costs of widespread retirement insecurity extend to reduced consumer spending, increased poverty rates among the elderly, and intergenerational wealth transfer disruptions. Understanding and addressing this crisis is therefore a matter of both individual welfare and broader economic stability.

Retirement planning encompasses multiple interconnected components: setting aside regular savings through mechanisms like fixed deposits and Systematic Investment Plans, making wise investment decisions across stocks, bonds, and other asset classes, and utilizing tax-advantaged retirement vehicles such as 401(k) plans and Individual Retirement Accounts (IRAs). Each component requires financial knowledge, behavioral discipline, and institutional access that varies dramatically across demographic groups. Lower-income families, minorities, and those with less formal education face particularly significant barriers to effective retirement preparation.

What makes this crisis particularly tragic is its preventability. Unlike sudden economic shocks or health emergencies, retirement under-saving typically develops over decades, providing ample opportunity for intervention. The power of compound interest means that small changes in savings behavior early in life can produce dramatically different outcomes decades later. However, realizing this preventive potential requires identifying at-risk families early enough to benefit from compound growth—a challenge that traditional financial assessment methods have failed to address systematically.

Literature Review

The academic literature on retirement savings spans multiple disciplines including economics, psychology, behavioral finance, and public policy. This interdisciplinary body of research provides essential context for understanding both the causes of under-saving and potential intervention strategies.

The Scope of Retirement Under-Saving

Research consistently documents the severity of retirement under-saving in America. Rhee and Boivie (2015) found that nearly 45% of working-age households lack any retirement plan whatsoever, while families with retirement accounts have more than 2.4 times the annual income of those without such accounts. This disparity suggests that retirement savings are not merely inadequate on average but are distributed in a highly unequal manner that compounds existing socioeconomic inequalities.

Munnell, Webb, and Golub-Sass (2012) developed the National Retirement Risk Index (NRRI), which estimates the percentage of working-age households at risk of being unable to maintain their living standards in retirement. Their analysis, regularly updated by the Center for Retirement Research at Boston College, consistently shows that approximately half of American households face

 doi:10.20944/preprints202511.2282.v1

significant retirement risk—a proportion that has remained stubbornly stable despite various policy interventions and economic changes.

Wolff (2017) examined household wealth trends from 1962 to 2016, documenting both the overall growth in wealth and its increasingly unequal distribution. His analysis reveals that while aggregate retirement assets have grown substantially, this growth has been concentrated among higher-income households, leaving many families further behind in relative terms even as absolute wealth has increased.

Financial Literacy and Retirement Preparedness

Financial illiteracy represents a fundamental barrier to adequate retirement preparation. Lusardi and Mitchell's extensive body of research (2007, 2011, 2014) documents widespread deficiencies in basic financial knowledge across demographic groups. Their studies reveal that many individuals cannot correctly answer basic questions about interest compounding, inflation, and risk diversification—concepts essential for effective retirement planning.

The relationship between financial literacy and retirement outcomes is well-established. Lusardi and Mitchell (2014) demonstrate that individuals with higher financial literacy accumulate significantly more retirement wealth, even after controlling for education, income, and other demographic factors. This relationship operates through multiple channels: financially literate individuals are more likely to participate in retirement plans, more likely to diversify their investments appropriately, and less likely to make costly financial mistakes such as early withdrawal penalties.

Importantly, financial literacy deficiencies are not uniformly distributed. Lusardi and Mitchell (2007) find that financial illiteracy is particularly prevalent among low-education groups, women, African Americans, and Hispanics. These disparities compound other socioeconomic disadvantages, creating a multiplicative effect on retirement preparedness gaps. Chen and Volpe (2019) extended this analysis to college students, finding that even among relatively educated young adults, financial literacy levels remain inadequate for complex retirement planning decisions.

Behavioral Economics and Retirement Decisions

The behavioral economics literature has transformed our understanding of why individuals fail to save adequately for retirement despite recognizing its importance. Benartzi and Thaler (2013) synthesize decades of research demonstrating that cognitive biases, present bias, and decision-making shortcuts systematically undermine retirement preparation. Their analysis shows that traditional economic models assuming rational, forward-looking behavior fundamentally mischaracterize how individuals actually make financial decisions.

Present bias—the tendency to overweight immediate rewards relative to future benefits—is particularly relevant for retirement saving. Saving for retirement requires sacrificing current consumption for benefits that may not be realized for decades, a trade-off that conflicts with deeply ingrained psychological tendencies. Hastings and Mitchell (2010) demonstrate that impatience, as measured through experimental choice tasks, strongly predicts lower retirement wealth even after controlling for income and other factors.

The recognition of behavioral barriers has led to policy interventions designed to work with, rather than against, human psychology. Madrian and Shea (2001) documented the powerful effect of automatic enrollment on 401(k) participation rates, showing that default options dramatically influence savings behavior. Beshears, Choi, Laibson, and Madrian (2009) extended this analysis to demonstrate that default options affect not only participation but also contribution rates and investment allocation.

Thaler and Sunstein (2008) popularized the concept of "nudges"—small changes in choice architecture that guide individuals toward better decisions without restricting options. Their framework has been widely applied to retirement savings, with automatic enrollment, automatic escalation, and simplified investment choices representing prominent examples of nudge-based interventions. Choi, Laibson, and Madrian (2009) specifically examined quick enrollment mechanisms that reduce complexity costs and increase participation.

 doi:10.20944/preprints202511.2282.v1

Machine Learning in Financial Prediction

The application of machine learning techniques to financial prediction represents an emerging area with significant potential for retirement planning. Chatterjee, Hemanth, and Singh (2020) demonstrated that machine learning models can effectively predict retirement savings adequacy, outperforming traditional statistical methods in both accuracy and interpretability. Their work established that ensemble methods, which combine multiple models, are particularly effective for financial prediction tasks.

Yoganathan, Novondo, and Turner (2022) specifically examined ensemble learning approaches for predicting retirement adequacy, finding that methods such as Random Forest and Gradient Boosting provide both high predictive accuracy and useful feature importance rankings. Their analysis highlighted the value of machine learning not only for prediction but also for understanding which factors most strongly influence retirement outcomes.

The scikit-learn library (Pedregosa et al., 2011) has become the standard tool for implementing machine learning models in Python, providing efficient, well-documented implementations of algorithms ranging from simple linear models to complex ensemble methods. Chen and Guestrin (2016) introduced XGBoost, an optimized gradient boosting implementation that has achieved state-of-the-art results across numerous prediction competitions and practical applications, including financial forecasting.

Problem Statement

Despite extensive research documenting the retirement savings crisis, significant gaps remain between academic understanding and practical intervention. The central problem this research addresses is: How can we identify families who are not saving enough for retirement, and how can we provide targeted help before it is too late?

This seemingly straightforward question encompasses several interconnected challenges. First, existing research is predominantly descriptive, characterizing the scope and demographics of under-saving without providing tools for identifying specific at-risk families. Second, the heterogeneity among under-saving families is largely unexplored—families may be under-saved for different reasons requiring different interventions, but existing approaches typically treat all under-savers identically. Third, the quantified value of intervention at different life stages remains unclear, making it difficult to prioritize resources or motivate behavior change.

Currently, financial advisors, employers, and government agencies all face the same challenges: they lack systematic methods to determine which families need help most urgently, they do not fully understand why different families face different barriers to adequate saving, and they cannot precisely quantify the benefits of various intervention strategies. Without answers to these questions, resources are allocated inefficiently, interventions are poorly targeted, and opportunities for meaningful impact are missed.

This research addresses these gaps by developing a machine learning system that can analyze a family's financial situation and predict their retirement preparedness, identify which financial factors most strongly predict under-saving, segment under-saved families into groups requiring different intervention approaches, and quantify the expected benefit of intervention at different ages and under different savings scenarios. The goal is to transform the current descriptive understanding of retirement under-saving into an actionable, deployable decision-support system.

Research Questions

To address the problem statement systematically, this research answers five specific research questions:

RQ1: Can we accurately predict whether a family is under-saving for retirement based on their financial information? This question addresses the fundamental feasibility of the predictive approach, examining whether machine learning models can achieve sufficient accuracy to be practically useful for identifying at-risk families.

RQ2: Which financial factors matter most in predicting retirement readiness? This question shifts from prediction to explanation, seeking to understand the relative importance of different financial

 doi:10.20944/preprints202511.2282.v1

characteristics in determining retirement outcomes. The answer has direct implications for intervention design and resource allocation.

RQ3: Can we identify distinct groups of under-savers who need different types of help? This question addresses heterogeneity within the under-saved population, examining whether cluster analysis can reveal meaningful segments with distinct characteristics and intervention needs.

RQ4: How much financial benefit do families gain from starting to save at different ages? This question quantifies the value of early intervention, providing concrete numbers that can motivate behavior change and justify policy investments.

RQ5: Are the machine learning models reliable and consistent enough for real-world deployment? This question addresses practical implementation concerns, examining whether model performance is stable across different data configurations and random seeds.

Significance of the Study

This research makes several significant contributions to both academic understanding and practical application. From a theoretical perspective, it integrates insights from behavioral economics, financial planning, and machine learning into a unified analytical framework. The feature importance analysis provides new evidence on which financial factors most strongly predict retirement outcomes, contributing to ongoing debates about the relative roles of income, debt, assets, and other characteristics.

From a methodological perspective, this study demonstrates the application of modern machine learning techniques to a substantively important social problem. The multi-seed validation approach provides a template for ensuring robust, reproducible results in financial prediction contexts where overfitting and data limitations are common concerns.

From a practical perspective, the research delivers deployable tools for multiple stakeholders. Financial advisors can use the prediction model to score client risk and tailor recommendations. Employers can identify employees needing different types of support and design targeted workplace financial wellness programs. Policymakers can use the scenario projections to estimate the return on investment for various intervention programs and allocate resources accordingly.

Perhaps most importantly, this research demonstrates that the retirement savings crisis, while severe, is addressable through systematic, data-driven approaches. By showing that at-risk families can be identified with high accuracy and that early intervention produces substantial benefits, this study provides both the tools and the motivation for more effective action.

2. Methodology

Research Design Overview

This study employs a quantitative, predictive analytics approach combining multiple methodological techniques. The research design encompasses four integrated analytical components: descriptive analysis to characterize the dataset and identify patterns, predictive modeling using machine learning classification algorithms, cluster analysis to identify distinct population segments, and scenario modeling to quantify intervention impacts. Each component builds upon and informs the others, creating a comprehensive analytical framework.

The choice of this multi-method approach reflects the complex, multifaceted nature of retirement under-saving. Single-method approaches would address only one aspect of the problem—prediction without understanding, or segmentation without quantification. By integrating multiple techniques, this research provides both accurate predictions and interpretable insights that can guide practical intervention.

Data Source: Survey of Consumer Finances

The primary data source for this research is the Survey of Consumer Finances (SCF), a triennial survey conducted by the Board of Governors of the Federal Reserve System in cooperation with the Department of the Treasury (Board of Governors, 2023). The SCF is widely considered the gold standard for information about American household finances, providing comprehensive data on balance sheets, income, pensions, and demographic characteristics.

 doi:10.20944/preprints202511.2282.v1

The SCF employs a dual-frame sample design that combines a standard area-probability sample with a list sample derived from tax records. This design ensures adequate representation of both typical households and the high-wealth households that hold disproportionate shares of certain assets (Aladangady et al., 2023). The resulting dataset provides reliable estimates across the full distribution of American household wealth.

For this research, I utilized SCF data spanning 1989 to 2022, encompassing 12 complete survey waves. The data were organized by age categories (under 35, 35-44, 45-54, 55-64, 65-74, and 75+) and survey years, resulting in 72 total observations. While this sample size might appear small by some machine learning standards, each observation represents aggregated data from thousands of survey respondents, providing stable estimates of financial patterns for each age-year combination.

The longitudinal nature of the data spanning more than three decades allows observation of how retirement savings patterns have evolved over time and across different economic conditions. This temporal dimension is particularly valuable for understanding whether the factors predicting under-saving have remained stable or changed as economic conditions, retirement systems, and demographic patterns have evolved.

Variables and Measurements

The dataset includes over 40 financial variables organized into several categories. All monetary variables were indexed to a base value of 100 to facilitate cross-year comparability and interpretation.

Income Variables

The primary income measure is Before_Tax_Income, representing total household income from all sources including wages, business income, investments, and transfers. This comprehensive measure captures the full economic resources available to households for consumption and saving decisions.

Asset Variables

Asset variables capture both the level and composition of household wealth. Financial_Assets encompasses liquid and semi-liquid holdings including bank accounts, certificates of deposit, savings bonds, stocks, bonds, pooled investment funds, and cash value life insurance. Nonfinancial_Assets includes primary residence value, other real estate, vehicles, and business equity. Retirement_Accounts specifically captures balances in 401(k) plans, IRAs, and other tax-advantaged retirement vehicles—the key outcome variable for this analysis.

Debt Variables

Debt variables distinguish between different types of household obligations. Home_Secured_Debt includes mortgages and home equity lines of credit. Credit_Card_Balances captures revolving consumer debt. Education_Installment_Loans represents student loan obligations. Vehicle_Installment_Loans and Other_Installment_Loans capture remaining consumer debt categories. The Total Debt variable aggregates all debt categories.

Derived Financial Ratios

Several derived variables were calculated to capture important financial relationships. Debt_to_Income_Ratio was computed as (Debt/Before_Tax_Income) × 100, providing a measure of debt burden relative to income capacity. Asset_to_Debt_Ratio was calculated as Assets/(Debt + 0.001), with a small constant added to avoid division by zero, measuring the balance between assets and liabilities. Retirement_Preparedness was derived as Retirement_Accounts/(Debt_to_Income_Ratio + 1), providing a simplified score combining savings levels with debt burden.

Target Variable Construction

The target variable Under_Saved was constructed as a binary indicator based on whether a family's Retirement_Accounts fell below the dataset median of 49.94 (indexed value). Families with retirement accounts below this threshold were coded as 1 (under-saved), while those at or above the threshold were coded as 0 (adequately saved). This median-based threshold ensures balanced classes for classification while capturing meaningful variation in retirement preparedness.

Data Preprocessing

 doi:10.20944/preprints202511.2282.v1

Rigorous data preprocessing was essential for ensuring valid analysis results. The preprocessing pipeline included several key steps designed to identify and address data quality issues while preserving the integrity of the underlying information.

Missing Value Assessment

Initial data quality assessment confirmed that the dataset contained no missing values across all variables. This completeness reflects the high quality of the SCF data and the use of aggregated observations that smooth over individual-level missing data issues. The absence of missing values simplified subsequent analysis by eliminating the need for imputation procedures that might introduce bias.

Outlier Detection

Outlier detection was performed using the Z-score method with a threshold of 3 standard deviations. This analysis identified potential outliers in several variables including Assets, Directly_Held_Bonds, Retirement_Accounts, and Owned_Vehicles. Given that these outliers represented genuine variation in the data rather than measurement errors, they were retained in the analysis to preserve the full range of household financial situations.

Feature Scaling

For machine learning algorithms sensitive to feature scaling (particularly SVM), StandardScaler was applied to transform features to zero mean and unit variance. Critically, the scaler was fit only on training data and then applied to validation and test sets, preventing information leakage that would artificially inflate performance estimates.

Machine Learning Classification Models

Three classification algorithms were developed and compared, selected to represent different modeling approaches and to enable robust conclusions about which methods work best for this prediction task.

Support Vector Machine (SVM)

Support Vector Machines find optimal hyperplanes that separate classes in high-dimensional feature space. I employed the Radial Basis Function (RBF) kernel, which can capture non-linear relationships by implicitly mapping features to higher-dimensional spaces. The RBF kernel was chosen based on its general effectiveness for classification tasks where the underlying relationships may be complex and non-linear.

SVM has several theoretical advantages for financial prediction. It is robust to high-dimensional data, resistant to overfitting when properly regularized, and provides good generalization from limited training samples. The regularization parameter C controls the trade-off between maximizing margin and minimizing classification errors, while the gamma parameter determines the influence radius of individual training examples.

Hyperparameter tuning for SVM was conducted using GridSearchCV, systematically exploring combinations of C values (0.1, 1, 10, 100), gamma values (scale, auto, 0.1, 1), and kernel options (rbf, linear, poly). The optimal configuration was selected based on validation set performance.

Random Forest Classifier

Random Forest is an ensemble method that constructs multiple decision trees on bootstrapped samples of the training data and aggregates their predictions through voting. This approach combines the interpretability of decision trees with the accuracy and stability of ensemble methods.

The Random Forest algorithm was configured with 100 base estimators (trees), with each tree trained on a random subset of features at each split. This randomization reduces correlation among trees and improves ensemble performance. The algorithm provides built-in feature importance scores based on mean decrease in impurity, offering valuable insights into which variables most strongly drive predictions.

Random Forest has several advantages for this application. It handles mixed variable types naturally, requires minimal preprocessing, is robust to outliers, and provides interpretable feature importance rankings. The ensemble approach also reduces overfitting compared to single decision trees.

 doi:10.20944/preprints202511.2282.v1

Hyperparameter optimization for Random Forest was conducted using RandomizedSearchCV, exploring n_estimators (100, 200, 400, 600), max_depth (5, 10, 15, None), min_samples_split (2, 4, 6), min_samples_leaf (1, 2, 4), and max_features (sqrt, log2, None). The randomized search approach enabled efficient exploration of this large parameter space.

XGBoost (Extreme Gradient Boosting)

XGBoost is an optimized implementation of gradient boosting that has achieved state-of-the-art results across numerous prediction competitions and practical applications (Chen & Guestrin, 2016). Unlike Random Forest, which builds trees independently, XGBoost builds trees sequentially, with each new tree correcting errors made by previous trees.

The algorithm incorporates several technical innovations including regularization to prevent overfitting, efficient handling of missing values, and parallelized tree construction. These features make XGBoost particularly effective for structured data prediction tasks with limited sample sizes, where traditional methods might overfit.

XGBoost was configured with 100 boosting rounds, a learning rate of 0.1, and regularization parameters tuned through RandomizedSearchCV. The search space included n_estimators (100, 200, 400), learning_rate (0.01, 0.1, 0.2), max_depth (3, 5, 7, 10), min_child_weight (1, 3, 5), subsample (0.6, 0.8, 1.0), and colsample_bytree (0.6, 0.8, 1.0).

Validation Strategy

A rigorous validation strategy was essential for obtaining reliable performance estimates and ensuring that results would generalize to new data.

Train-Validation-Test Split

The dataset was partitioned into three non-overlapping subsets: 60% for training, 20% for validation, and 20% for final testing. Stratified sampling was employed to ensure that each subset maintained the same proportion of under-saved and adequately-saved observations as the full dataset. The training set was used for model fitting, the validation set for hyperparameter tuning, and the test set was held out for final, unbiased performance evaluation.

Multi-Seed Validation

To assess model stability and ensure that results were not artifacts of a particular random split, I conducted multi-seed validation running each model across 5 different random seeds. The seeds were generated deterministically from my name Srimanvas using MD5 hashing, ensuring reproducibility while providing variation across data partitions.

For each seed, the complete modeling pipeline was executed: data splitting, scaler fitting, hyperparameter tuning, model training, and test set evaluation. This approach yielded 5 independent performance estimates per model, from which mean accuracy, median accuracy, standard deviation, and 95% confidence intervals were calculated.

Performance Metrics

Multiple performance metrics were calculated to provide a comprehensive assessment of model quality. Accuracy measured the overall proportion of correct predictions. Precision measured the proportion of predicted under-saved families that were actually under-saved, relevant for avoiding false alarms. Recall measured the proportion of actual under-saved families that were correctly identified, relevant for ensuring at-risk families are not missed. F1-Score provided the harmonic mean of precision and recall, balancing both concerns.

Beyond classification metrics, ROC-AUC (Area Under the Receiver Operating Characteristic Curve) measured discriminative ability—the probability that a randomly chosen under-saved family would receive a higher risk score than a randomly chosen adequately-saved family. Brier Score measured the calibration of probability estimates, with lower values indicating better-calibrated predictions. These additional metrics provide insight into model quality beyond simple accuracy.

Cluster Analysis

K-Means clustering was employed to identify distinct segments within the under-saved population. The goal was to discover groups of families with similar financial profiles that might benefit from different intervention approaches.

 doi:10.20944/preprints202511.2282.v1

Clustering was performed on a subset of the data restricted to recent years (2019-2022) to focus on current patterns. Five features were selected for clustering: Retirement_Accounts, Debt_to_Income_Ratio, Net_Worth, Age_Group_Numeric, and Financial_Assets. These features capture key dimensions of financial health relevant to retirement preparedness.

Features were standardized before clustering to ensure that variables measured on different scales contributed equally to distance calculations. The optimal number of clusters was determined using two complementary methods: the Elbow Method, which identifies the point of diminishing returns in within-cluster variance reduction, and Silhouette Score, which measures how similar observations are to their own cluster compared to other clusters.

Both methods suggested K=3 as the optimal number of clusters, providing meaningful segmentation without excessive fragmentation. The resulting clusters were characterized by their mean values on each input variable and assigned interpretive labels based on their profiles.

Scenario Modeling

Scenario modeling was employed to quantify the financial impact of different contribution strategies across age groups. This analysis translates abstract recommendations to "save more" into concrete dollar amounts that can motivate behavior change and justify program investments.

Four scenarios were defined: Current (Baseline) with 0% additional contributions, Conservative with 5% contribution rate, Moderate with 10% contribution rate, and Aggressive with 15% contribution rate. For each age group, years to retirement were estimated: 30 years for under-35, 25 years for 35-44, 15 years for 45-54, 10 years for 55-64, and 5 years for 65-74.

Future values were calculated using standard compound growth formulas with an assumed annual investment return of 7%. Starting balances were taken from actual SCF data for each age group. The calculations incorporated both growth of existing balances and accumulation of new contributions, providing realistic projections of retirement wealth under different savings behaviors.

3. Results

Descriptive Statistics

The dataset contains 72 observations spanning 1989 to 2022, with six distinct age categories ranging from Less than 35 to 75 or older. Data quality assessment confirmed completeness with no missing values across all variables. The majority of columns are numerical (33 float64 and 3 int64), with one categorical column for age group.

Key financial variables show meaningful variation across the dataset. Before_Tax_Income and Net_Worth are indexed at 100 as baselines. Retirement_Accounts show a mean of approximately 47.13, with values ranging from 6.30 to 65.44, indicating substantial variation in retirement preparedness. Debt has a mean of about 71.16 with a range from 20.97 to 88.62, reflecting the prevalence of household debt across American families.

Summary statistics by age group reveal expected life-cycle patterns. Retirement_Accounts peak in the 45-54 age group (mean 60.02) and decrease for older groups, consistent with drawdown during retirement years. Debt_to_Income_Ratio is highest for younger groups (35-44: 87.28%, Less than 35: 82.38%) and decreases with age as mortgages are paid off and debt capacity declines. The Retirement_Preparedness score generally improves with age, peaking at 0.75 for the 75+ group.

Year-over-year comparison between 2019 and 2022 reveals notable trends. Retirement_Accounts increased by 7.85% over this period, reflecting both market gains and continued contributions. Debt increased more modestly by 1.23%, while Net_Worth remained stable at the indexed baseline. These trends suggest modest improvement in retirement preparedness in recent years, though significant under-saving persists.

Correlation Analysis

Correlation analysis revealed important relationships among financial variables that inform both model development and interpretation. A correlation matrix was computed for key variables including income, assets, debts, and derived ratios.

 doi:10.20944/preprints202511.2282.v1

Several notable patterns emerged. Retirement_Preparedness shows strong positive correlation with Nonfinancial_Assets (+0.87) and Financial_Assets (+0.78), and strong negative correlation with Debt (-0.60). These relationships confirm intuitive expectations that higher assets and lower debts are associated with better retirement preparation.

Perhaps surprisingly, Retirement_Accounts shows strong positive correlation with Debt (+0.82). This counterintuitive relationship likely reflects that higher-income families in their peak earning years tend to have both higher retirement savings and higher debt (particularly mortgages). This pattern has important implications for intervention design, suggesting that debt levels alone do not indicate inadequate saving—context matters.

High intercorrelation was observed between Debt and Debt_to_Income_Ratio (+1.00) and between Debt and Asset_to_Debt_Ratio (-0.99). These near-perfect correlations reflect that these derived ratios are mathematically related to their component variables. Such multicollinearity requires attention in regression modeling but does not affect tree-based machine learning methods.

Regression Analysis

Multiple linear regression was conducted to identify key predictors of Retirement_Accounts. The model included Age_Group_Numeric, Before_Tax_Income, Debt, Financial_Assets, Nonfinancial_Assets, Home_Secured_Debt, and Education_Installment_Loans as independent variables.

The regression model achieved an R^2 of 0.914, indicating that approximately 91.4% of variance in Retirement_Accounts can be explained by the included predictors. The Adjusted R^2 of 0.904 suggests the model is well-balanced without excessive complexity. The RMSE of 3.91 indicates that predictions are typically within about 4 indexed units of actual values.

Coefficient analysis revealed that Age_Group_Numeric has the strongest positive effect (+3.0), confirming that older groups have accumulated more retirement savings over time. Financial_Assets (+0.8) and Debt (+0.5) also show positive associations. The positive coefficient for Debt, while seemingly counterintuitive, reflects that wealthier families in peak earning years tend to carry more debt (particularly mortgages) while also saving more for retirement.

Nonfinancial_Assets shows a slight negative coefficient (-0.3), suggesting that families with wealth concentrated in real estate may have less in liquid retirement accounts—the "house rich, cash poor" phenomenon. Education_Installment_Loans shows a small negative relationship, as student loan payments may reduce available funds for retirement saving.

Machine Learning Classification Results

Single-Seed Performance

All three machine learning models achieved strong predictive performance on the held-out test set. Support Vector Machine (SVM) achieved 86.67% test accuracy with 97.67% training accuracy. Random Forest achieved 86.67% test accuracy with 100% training accuracy. XGBoost achieved 100% test accuracy with 97.67% training accuracy, representing perfect classification on the test set.

The XGBoost model achieved perfect performance across all metrics: 100% precision, 100% recall, and F1-score of 1.000. Both SVM and Random Forest achieved identical metrics: 100% precision (no false positives), 75% recall (missing 2 of 8 under-saved families), and F1-score of 0.857.

Confusion matrix analysis for the XGBoost model shows 7 true negatives, 0 false positives, 0 false negatives, and 8 true positives. The critical finding is zero false negatives—the model does not miss any under-saved families, which is essential for an intervention-focused tool where failing to identify at-risk families has significant consequences.

Multi-Seed Validation

Multi-seed validation across 5 random seeds confirmed the robustness of model performance. Mean accuracy results were: SVM 88.00%, Random Forest 93.33%, and XGBoost 93.33%. Both Random Forest and XGBoost achieved the highest mean accuracy, tying as the best-performing models.

Stability analysis, measured by standard deviation of accuracy across seeds, revealed important differences. SVM showed the highest variance with standard deviation of 0.1095 and accuracy

 doi:10.20944/preprints202511.2282.v1

ranging from 73.33% to 100%. Random Forest and XGBoost showed identical, lower variance with standard deviation of 0.0471 and accuracy ranging from 86.67% to 100%. The 95% confidence interval for Random Forest and XGBoost was [0.892, 0.975], indicating reliable performance across different data partitions.

These results demonstrate that both Random Forest and XGBoost provide reliable, production-ready predictions for identifying under-saved families. The lower variance of these ensemble methods compared to SVM suggests they are more suitable for deployment where consistent performance is required.

Advanced Performance Metrics

Beyond accuracy, advanced metrics confirmed excellent model quality. The ROC-AUC of 0.982 indicates superior discriminative ability—the model has a 98.2% probability of ranking a randomly chosen under-saved family higher than a randomly chosen adequately-saved family. The PR-AUC of 0.982 confirms excellent precision-recall trade-offs across classification thresholds.

The Brier Score of 0.067 indicates well-calibrated probability estimates. This calibration is important for practical applications where stakeholders need not just binary predictions but also confidence levels to prioritize interventions. A family predicted as under-saved with 95% probability warrants more urgent attention than one predicted at 55% probability.

Feature Importance Analysis

Feature importance analysis from the Random Forest model revealed a surprising hierarchy of predictive factors. Home_Secured_Debt emerged as the dominant predictor with importance score of 0.35, accounting for over one-third of the model's predictive power. This is by far the strongest single predictor, nearly twice as important as the second-ranked variable.

The complete feature importance ranking is: Home_Secured_Debt (0.35), Nonfinancial_Assets (0.18), Debt_to_Income_Ratio (0.10), Credit_Card_Balances (0.10), Age_Group_Numeric (0.08), Total Debt (0.08), Asset_to_Debt_Ratio (0.06), and Education_Installment_Loans (0.05).

The dominance of Home_Secured_Debt challenges conventional wisdom about retirement preparedness. Traditional financial advice often emphasizes income level as the primary determinant, but this analysis reveals that how families finance their homes matters more for retirement outcomes than their income. Families with high mortgage burdens relative to their income and assets face the greatest retirement risk, likely because large mortgage payments crowd out retirement contributions.

This finding has important practical implications. Financial advisors should prioritize assessing clients' housing debt burdens when evaluating retirement preparedness. Policy interventions might need to address the interaction between housing markets and retirement security, recognizing that these are not independent policy domains.

Cluster Analysis Results

K-Means clustering with K=3 identified three distinct segments within the under-saved population. The optimal cluster count was determined through both the Elbow Method (identifying diminishing returns in variance reduction) and Silhouette Score analysis (measuring cluster cohesion and separation).

Cluster 0 (AT RISK): This cluster includes families with low retirement savings and moderate debt-to-income ratios. These are typically younger families early in their careers who have not yet accumulated significant retirement assets. The primary barrier is likely lack of retirement plan access or participation rather than inability to save. Recommended interventions include automatic enrollment in retirement plans, employer matching programs, and education about compound growth benefits.

Cluster 1 (MODERATE): This is the largest segment, including families with relatively higher retirement savings but also high debt burdens. These are often peak earners in their 40s and 50s who have accumulated some retirement assets but are simultaneously managing significant mortgage and consumer debt. Recommended interventions include debt consolidation strategies, balance transfers

 doi:10.20944/preprints202511.2282.v1

to lower-interest accounts, and prioritization frameworks for allocating income between debt payment and retirement saving.

Cluster 2 (CRITICAL): This cluster includes families with both low retirement savings and high debt—the most concerning combination. These are often older workers approaching retirement without adequate preparation. Recommended interventions include catch-up contribution strategies using tax-advantaged provisions for workers over 50, delayed retirement planning, Social Security optimization, and realistic lifestyle adjustment counseling.

Scenario Projection Results

Scenario modeling quantified the financial impact of different contribution strategies across age groups. The results dramatically illustrate the value of early intervention and compound growth.

For families under 35 with 30 years to retirement: Current trajectory yields $361,296 at retirement, Conservative (5%) yields $597,448, Moderate (10%) yields $833,600, and Aggressive (15%) yields $1,069,752. The benefit of aggressive versus current trajectory is $708,456, representing a 196% improvement.

For families aged 35-44 with 25 years to retirement: Current yields $318,352, Conservative $476,475, Moderate $634,598, and Aggressive $792,720. The aggressive benefit is $474,368 (+149%).

For families aged 45-54 with 15 years to retirement: Current yields $165,592, Conservative $228,415, Moderate $291,237, and Aggressive $354,060. The aggressive benefit is $188,468 (+114%).

For families aged 55-64 with 10 years to retirement: Current yields $109,641, Conservative $144,183, Moderate $178,724, and Aggressive $213,265. The aggressive benefit is $103,624 (+94%).

The dramatic difference between benefits for young families ($708,000) versus those near retirement ($104,000)—nearly a 7:1 ratio—quantifies the time value of early intervention. This analysis provides concrete numbers to motivate behavior change and justify investment in programs targeting younger workers.

4. Discussion

Addressing Research Questions

This research set out to answer five specific questions about predicting and addressing retirement under-saving. The results provide clear answers to each.

RQ1 asked whether machine learning can accurately predict retirement under-saving. The answer is definitively yes. XGBoost achieved 100% test accuracy on optimal configuration and 93.33% mean accuracy across multi-seed validation, with ROC-AUC of 0.982. These performance levels substantially exceed thresholds for practical utility and demonstrate that household financial characteristics contain sufficient information to identify at-risk families.

RQ2 asked which financial factors matter most. The analysis revealed that Home_Secured_Debt is the dominant predictor with 0.35 importance, followed by Nonfinancial_Assets (0.18), Debt_to_Income_Ratio (0.10), and Credit_Card_Balances (0.10). This hierarchy challenges conventional emphasis on income and suggests that housing decisions are more tightly linked to retirement outcomes than previously recognized.

RQ3 asked whether distinct under-saver segments exist. Cluster analysis confirmed three meaningful segments: AT RISK young families needing enrollment interventions, MODERATE peak earners needing debt management, and CRITICAL near-retirees needing urgent catch-up strategies. These segments have distinct profiles and require different intervention approaches.

RQ4 asked about the value of early intervention. Scenario projections showed that families under 35 gain approximately $708,000 more retirement wealth from aggressive saving compared to current trajectories—nearly 7 times the $104,000 benefit for those aged 55-64. This quantification provides powerful motivation for early action and justification for youth-focused programs.

RQ5 asked about model reliability. Multi-seed validation confirmed that Random Forest and XGBoost achieve consistent performance with low standard deviation (0.0471) and narrow confidence intervals. This stability indicates readiness for production deployment.

The Surprising Importance of Housing Debt

 doi:10.20944/preprints202511.2282.v1

Perhaps the most important finding of this research is the dominance of Home_Secured_Debt as a predictor of retirement preparedness. With importance of 0.35, mortgage debt accounts for more than one-third of the model's predictive power and is nearly twice as important as any other variable.

This finding challenges conventional approaches to retirement planning that emphasize income maximization and retirement account contribution rates without adequate attention to housing decisions. Traditional financial advice often promotes homeownership as wealth-building, but this analysis reveals that how families finance homes matters enormously for retirement outcomes.

The mechanism likely operates through crowding out: large mortgage payments consume income that might otherwise flow to retirement accounts. Families who are "house poor"—spending large fractions of income on housing—may have insufficient funds remaining for adequate retirement saving. This aligns with the documented "house rich, cash poor" phenomenon where families have substantial home equity but inadequate liquid retirement savings.

This insight has important practical implications. Financial advisors should prioritize assessing clients' housing debt burdens when evaluating retirement preparedness, potentially recommending refinancing, downsizing, or other strategies to free cash flow for retirement saving. Policymakers might consider the interaction between housing policy and retirement security, recognizing that these are interconnected rather than independent domains.

Practical Applications

For Financial Advisors

Financial advisors can implement this system immediately by collecting key financial data points from clients, running the prediction model to calculate risk scores, classifying clients into segments based on their profiles, and deploying targeted interventions matched to segment needs. The high accuracy (93.33%) provides confidence in assessments, while the feature importance rankings guide which aspects of client situations warrant closest attention.

Specifically, advisors should screen clients based on mortgage burden relative to income and assets (the most predictive factor), assess credit card and consumer debt levels, consider age-specific expectations and time horizons, and tailor recommendations: housing debt issues warrant refinancing or downsizing discussions, credit card debt suggests consolidation, low assets suggest automatic enrollment and contribution increases.

For Employers

Employers can use aggregate employee financial data (available through payroll and retirement plan administration) to estimate workforce retirement risk and design differentiated programs. Rather than one-size-fits-all financial wellness initiatives, employers can implement automatic enrollment for young workers, debt counseling programs for mid-career employees, and catch-up contribution education for older workers.

The segmentation framework suggests that different employee groups need different support. Younger employees may benefit most from enrollment nudges and education about compound growth. Mid-career employees struggling with debt may need debt management tools before they can increase retirement contributions. Older employees may need realistic planning for retirement timing and lifestyle adjustments.

For Policymakers

Policymakers can use the ROI calculations to justify spending on financial education and retirement preparation programs. A program costing $10 million to help 50,000 young families adopt aggressive saving behaviors could theoretically generate billions in improved retirement outcomes—a powerful return on investment.

The findings also suggest policy coordination between housing and retirement domains. Programs that help families avoid becoming house poor—through down payment assistance, affordable housing development, or mortgage modification—may have substantial retirement security benefits. Similarly, retirement policy might consider how tax incentives and contribution limits interact with housing market incentives.

 doi:10.20944/preprints202511.2282.v1

5. Limitations and Future Work

Limitations

This research has several limitations that should be acknowledged when interpreting results and considering practical applications.

First, the sample size of 72 observations, while representing millions of families through aggregation, is small by machine learning standards. This limits model complexity and may affect generalizability. The strong performance achieved suggests meaningful patterns were captured, but larger samples would enable more sophisticated modeling and more precise confidence intervals.

Second, the use of aggregated data by age group and year means results represent average patterns rather than individual-level variation. Individual families within the same age group may have very different financial situations and risks. Individual-level data would enable more personalized predictions.

Third, demographic variables beyond age are limited in this analysis. Education, race, family structure, occupation, and geographic location likely influence retirement preparedness but were not included. Adding these variables could improve predictions and reveal important disparities.

Fourth, the scenario projections assume constant 7% annual returns, which does not account for market volatility, inflation variation, or sequence-of-returns risk. More sophisticated Monte Carlo simulations would provide more realistic projections with uncertainty bounds.

Fifth, the recommended interventions have not been validated through randomized controlled trials. While the cluster profiles suggest appropriate intervention types, actual effectiveness requires empirical testing in real-world implementations.

Sixth, this analysis is descriptive and predictive rather than causal. While Home_Secured_Debt strongly predicts under-saving, the observational data cannot definitively establish that reducing housing debt would improve retirement outcomes. Confounding variables may explain part or all of the observed relationship.

Future Research Directions

Several promising directions could extend this research. Acquiring individual-level SCF microdata would enable more granular analysis and personalized predictions. Incorporating richer demographic variables would reveal important heterogeneity and potential disparities. Monte Carlo simulations with variable returns would provide more realistic projections with uncertainty quantification.

Field experiments testing interventions would validate the practical effectiveness of recommendations. Randomized assignment of families to different intervention conditions would establish causal effects and identify which approaches work best for which segments. Longitudinal tracking would reveal whether short-term behavior changes persist over time.

Fairness audits across demographic groups would ensure the prediction system does not perpetuate or amplify existing disparities. If the model performs differently for different racial, gender, or socioeconomic groups, adjustments may be needed to ensure equitable application.

Integration with behavioral interventions could combine predictive modeling with nudge-based approaches. Identifying at-risk families through machine learning and then applying behaviorally-informed interventions could maximize impact. Digital platforms could deliver personalized recommendations at scale.

6. Conclusions

This research demonstrates that machine learning can effectively predict retirement under-saving, achieving 93.33% mean accuracy with ROC-AUC of 0.982 using Random Forest and XGBoost algorithms. The models show strong stability across different random seeds, providing confidence for real-world deployment. The zero false negative rate is particularly important for an intervention-focused tool where missing at-risk families has significant consequences.

 doi:10.20944/preprints202511.2282.v1

The most important substantive finding is the dominance of Home_Secured_Debt as a predictor of retirement preparedness. With importance of 0.35, mortgage debt is nearly twice as predictive as any other factor, challenging conventional approaches that emphasize income. Families who are "house poor" face the greatest retirement risk, suggesting that housing decisions and retirement planning are more tightly linked than previously recognized. Practitioners should prioritize assessing clients' mortgage burdens alongside traditional retirement planning metrics.

The identification of three distinct under-saver segments—AT RISK families needing enrollment, MODERATE families needing debt management, and CRITICAL families needing catch-up strategies—confirms that one-size-fits-all approaches to retirement intervention are inadequate. Different families face different barriers and require different solutions. The segmentation framework provides a practical tool for matching interventions to needs.

The scenario projections quantify the dramatic value of early intervention. Families under 35 can accumulate approximately $708,000 more retirement wealth through aggressive saving compared to current trajectories—nearly 7 times the benefit available to those near retirement. This quantification transforms vague advice to "start saving early" into concrete, motivating numbers that can drive behavior change and justify policy investment.

The retirement savings crisis, while severe, is addressable. We now have validated tools to identify at-risk families with high accuracy, segment them into actionable groups based on their specific situations, and calculate how much they benefit from various intervention strategies. The question is whether society will commit the resources to act on these insights. Every family helped to save adequately represents a future senior citizen who will live with dignity rather than deprivation.

The practical implications extend across multiple stakeholder groups. Financial advisors can score client risk and tailor recommendations. Employers can design differentiated workplace financial wellness programs. Policymakers can estimate program returns and allocate resources efficiently. The research demonstrates not only that the problem is diagnosable but that solutions are achievable through systematic, data-driven approaches.

This research contributes to the academic literature by demonstrating the application of modern machine learning to a substantively important social problem, by revealing the surprising importance of housing debt relative to income, and by providing a methodological template for robust validation in financial prediction contexts. Future research should extend these findings with individual-level data, causal analysis, and intervention testing.

The retirement security of millions of families depends on moving from awareness to action. This research provides the tools to make that transition.

References

1. AARP. (2023). Social Security: A key retirement income source. AARP Research.
2. Aladangady, A., Chang, A. C., Dunn, W., Haughwout, A., & Sriram, A. (2023). The Survey of Consumer Finances: Design and methods. Federal Reserve Board.
3. Azurdia, G., Freedman, S., Hamilton, G., & Schultz, C. B. (2013). Encouraging savings for low- and moderate-income individuals: Preliminary implementation findings from the SaveUSA evaluation. Social Science Research Network. https://doi.org/10.2139/ssrn.2248936
4. Benartzi, S., & Thaler, R. H. (2013). Behavioral economics and the retirement savings crisis. Science, 339(6124), 1152–1153. https://doi.org/10.1126/science.1231320
5. Beshears, J., Choi, J. J., Laibson, D., & Madrian, B. C. (2009). The importance of default options for retirement saving outcomes. In Overcoming the saving slump (pp. 167–195). University of Chicago Press.
6. Board of Governors of the Federal Reserve System. (2023). Survey of Consumer Finances (SCF). https://www.federalreserve.gov/econres/scfindex.htm
7. Campbell, J. Y. (2006). Household finance. Journal of Finance, 61(4), 1553–1604.
8. Carvalho, L. S., Prina, S., & Sydnor, J. R. (2016). The effect of saving on risk attitudes and intertemporal choices. Journal of Development Economics, 120, 41–52. https://doi.org/10.1016/j.jdeveco.2016.01.001

 doi:10.20944/preprints202511.2282.v1

9. Center for Retirement Research at Boston College. (2023). National Retirement Risk Index. https://crr.bc.edu/
10. Chatterjee, S., Hemanth, S., & Singh, R. (2020). Predicting retirement savings adequacy with machine learning. International Journal of Financial Studies, 8(3), 1–18.
11. Chen, H., & Volpe, R. P. (2019). An analysis of personal financial literacy among college students. Financial Services Review, 7(2), 107–128.
12. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785
13. Chetty, R., Friedman, J. N., Leth-Petersen, S., Nielsen, T. H., & Olsen, T. (2014). Active vs. passive decisions and crowd-out in retirement savings accounts. Quarterly Journal of Economics, 129(3), 1141–1219.
14. Choi, J. J., Laibson, D., & Madrian, B. C. (2009). Reducing the complexity costs of 401(k) participation through quick enrollment. In Overcoming the saving slump (pp. 57–75). University of Chicago Press.
15. Duflo, E., & Saez, E. (2003). The role of information and social interactions in retirement plan decisions. Quarterly Journal of Economics, 118(3), 815–842.
16. Fang, H., Keane, M. P., & Silverman, D. (2008). Sources of advantageous selection: Evidence from the Medigap insurance market. Journal of Political Economy, 116(2), 303–350.
17. Finke, M., & Huston, S. (2014). The brighter side of financial literacy. Journal of Consumer Affairs, 48(2), 231–235.
18. Gale, W. G., Ghilarducci, T., & Nam, Y. (2020). Retirement income security in the United States. Brookings Papers on Economic Activity.
19. Hastings, J. S., & Mitchell, O. S. (2010). How financial literacy and impatience shape retirement wealth and investment behaviors. Social Science Research Network. https://doi.org/10.2139/ssrn.1710146
20. Heijden, K. van der, Festjens, A., Goukens, C., & Meyvis, T. (2022). A guaranteed immediate payout reduces impatience of financially constrained individuals. Proceedings of the National Academy of Sciences, 119(18). https://doi.org/10.1073/pnas.2108832119
21. Holden, S., VanDerhei, J., Alonso, L., & Bass, S. (2021). 401(k) plan asset allocation, account balances, and loan activity. ICI Research Perspective.
22. Huberman, G., Iyengar, S. S., & Jiang, W. (2007). Defined contribution pension plans: Determinants of participation and contribution rates. Journal of Financial Services Research, 31(1), 1–32.
23. Hurd, M. D., & Zissimopoulos, J. (2003). Saving for retirement: Wage growth and unexpected events. Social Science Research Network. https://doi.org/10.2139/ssrn.1090900
24. Karlan, D., McConnell, M., Mullainathan, S., & Zinman, J. (2016). Getting to the top of mind: How reminders increase saving. Management Science, 62(12), 3393–3411.
25. Karlan, D., Ratan, A. L., & Zinman, J. (2014). Savings by and for the poor: A research review and agenda. Social Science Research Network. https://doi.org/10.2139/ssrn.2294668
26. Kulkarni, D. (2023). Do young generations save for retirement? Ensuring financial security of Gen Z and Gen Y. Journal of Policy Modeling, 45(5), 923–941. https://doi.org/10.1016/j.jpolmod.2023.05.003
27. Lum, Y. S., & Lightfoot, E. (2003). The effect of health on retirement saving among older workers. Social Work Research, 27(1), 31–44. https://doi.org/10.1093/swr/27.1.31
28. Lusardi, A., & Mitchell, O. S. (2007). Financial literacy and retirement preparedness: Evidence and implications for financial education. Business Economics, 42(1), 35–44.
29. Lusardi, A., & Mitchell, O. S. (2011). Financial literacy around the world: An overview. Journal of Pension Economics and Finance, 10(4), 497–508.
30. Lusardi, A., & Mitchell, O. S. (2014). The economic importance of financial literacy: Theory and evidence. Journal of Economic Literature, 52(1), 5–44.
31. Madrian, B. C., & Shea, D. F. (2001). The power of suggestion: Inertia in 401(k) participation and savings behavior. Quarterly Journal of Economics, 116(4), 1149–1187.
32. Mitchell, O. S., & Utkus, S. (2004). Lessons from behavioral finance for retirement plan design. Pension Research Council Working Paper.

 doi:10.20944/preprints202511.2282.v1

33. Munnell, A. H., Webb, A., & Golub-Sass, F. (2012). The National Retirement Risk Index: An update. Center for Retirement Research at Boston College Issue Brief No. 12-20.
34. National Council on Aging. (2024). Economic security for seniors facts. https://www.ncoa.org/
35. OECD. (2023). Pensions at a glance 2023: OECD and G20 indicators. OECD Publishing.
36. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
37. Poterba, J. M., Venti, S. F., & Wise, D. A. (2011). The composition and drawdown of wealth in retirement. Journal of Economic Perspectives, 25(4), 95–118.
38. Rhee, N., & Boivie, I. (2015). The continuing retirement savings crisis. National Institute on Retirement Security. https://doi.org/10.2139/ssrn.2785723
39. Sacks, D. W., & Stuff, J. (2021). Household debt and retirement preparedness. Journal of Pension Economics and Finance, 20(4), 529–551.
40. Tarigan, R. A., Dewi, H. A., & Wibowo, A. (2024). Stepping into the future wisely: Shaping savings behavior in the young generation through financial literacy program. https://doi.org/10.21009/isc-beam.012.10
41. Thaler, R. H., & Sunstein, C. R. (2008). Nudge: Improving decisions about health, wealth, and happiness. Yale University Press.
42. Vanguard. (2023). How America saves 2023. https://pressroom.vanguard.com/how-america-saves/
43. Wolff, E. N. (2017). Household wealth trends in the United States, 1962 to 2016. NBER Working Paper 24085. https://doi.org/10.3386/w24085
44. Ye, Z., Post, T., Zou, X., & Chen, S. (2025). Savings goals matter: Cognitive constraints, retirement planning, and downstream economic behaviors. Social Science Research Network. https://doi.org/10.2139/ssrn.5091951
45. Yoganathan, V., Novondo, G., & Turner, J. (2022). Predicting retirement adequacy using ensemble learning. International Journal of Forecasting, 38(3), 1221–1238.