Article

Exploring Life Insurance Purchases: A SHAP-Based Feature Importance Analysis Using SCF 2022 Data

Zin Mar Oo[1,2]

Abstract

This study employs machine learning to analyze life insurance ownership decisions using data from the 2022 Survey of Consumer Finances (SCF). We compare the predictive performance of Logistic Regression, a conventional model, with three advanced machine learning models—Gradient Boosting Machine (GBM), Extreme Gradient Boosting (XGBoost), and Artificial Neural Networks (ANN)—to determine which best captures the complexities of life insurance ownership. To enhance interpretability, we apply SHapley Additive exPlanations (SHAP) to assess feature importance. Our findings indicate that ANN performs best for predicting term life insurance, while Logistic Regression excels for cash value life insurance, highlighting the importance of model selection based on dataset characteristics. SHAP analysis reveals that economic factors such as income and financial assets, are key drivers of term life insurance, whereas financial assets, age, and bequest preferences strongly influence cash value insurance. Psychographic factors play a greater role in cash value decisions. Additionally, SHAP helps uncover not only the key factors driving insurance purchases but also why some individuals choose not to buy life insurance, offering valuable insights for insurers and policymakers.

Keywords: life insurance, machine learning, SHAP analysis, Survey of Consumer Finances, feature importance.

1. Introduction

Life insurance is an important financial product that provides security and peace of mind to policyholders and their beneficiaries. Despite declining mortality rates since the 1970s,

[1]Doctoral Program Student, Graduate School of Economics, Ritsumeikan University, Japan. Email: gr0602sk@ed.ritsumei.ac.jp, zinmaroo@iuj.ac.jp.
[2]Budget Department, Ministry of Planning and Finance, Myanmar

recent studies highlight the persistent risk of death between ages 25 and 65 in the U. S., emphasizing the importance of life insurance in mitigating financial loss. The National Academies report (2021) notes that while overall life expectancy in the U. S. has improved, mortality rates among adults aged 25–64 remain troublingly high (National Academies of Sciences and Medicine, 2021). The premature death of a family head can severely impact surviving members by causing a loss of primary income and leaving unfulfilled financial obligations like supporting dependents and repaying a mortgage. Life insurance is crucial for financial security, serving as a hedge against income loss from the premature death of an earner (Garman and Forgue, 2018). This underscores the critical role life insurance plays in providing financial security and protecting against the economic impacts of an untimely death.

Life insurance can be broadly categorized into term life insurance and cash value life insurance. Term life insurance provides coverage for a specific period and allows renewal without new proof of insurability, though premiums rise with age. Benefits are paid only if the insured dies during the policy term. Cash value life insurance includes both a death benefit and savings component, with level premiums throughout the policy's duration. Early excess premiums are invested to cover future risks, and the policyholder can borrow against or surrender the cash value without tax consequences. Understanding these structural differences is crucial for interpreting the distinct drivers behind the ownership of each type.

Explaining what drives households to purchase life insurance is complex. Previous studies, such as the one by Li (2008), have explored various demographic, economic, and psychographic determinants of life insurance demand using conventional models. However, these studies often rely on assumptions, such as the inclusion of age squared terms to account for potential non-linear effects or suggesting the incorporation of interaction terms based on assumed relationships. While these approaches are insightful, they require assumptions about the nature of the relationships between variables. In contrast, machine learning models, which do not require such assumptions, can account for non-linear effects and interactions more naturally.

This study builds on the foundations laid by previous research by employing a combination of conventional and advanced machine learning techniques, including Logistic Regression, XGBoost, GBM, and ANN, to analyze the key factors influencing life insurance ownership decisions. These models accommodate different levels of complexity in life insurance decision-making, from structured relationships to intricate interactions. Rather than focusing solely on model performance, this study prioritizes feature importance analysis, using SHAP-based global and local interpretations to uncover the most influential predictors of life insurance purchases.

According to SCF (2022), 36% of households own only term life insurance, 12.6% own

only cash value life insurance, and 8.6% own both types. It is evident that over 42% of households have no life insurance. Coe et al. (2016) found that behavioral factors such as inertia significantly influence life insurance decision-making, often causing delays in determining appropriate coverage. This study aims to enhance the understanding of household life insurance purchase decisions, evaluate whether machine learning models can better capture complex relationships in life insurance demand, and determine whether simpler or more complex machine learning models offer superior predictive performance.

This study utilizes data from the 2022 Survey of Consumer Finances (SCF), a comprehensive dataset that provides detailed financial information from U. S. households, including life insurance ownership. The analysis focuses on two binary outcome variables: whether a household owns term life insurance and whether a household owns cash value life insurance. To understand the key factors influencing these ownership decisions, we employ SHAP-based feature importance analysis, examining a range of demographic, economic, and psychographic features. These include demographic factors such as age, gender, education, marital status, number of children, employment status, health condition, and race; economic factors such as income, financial assets, debt, and homeownership; and psychographic factors, including risk aversion and attitudes toward leaving a bequest.

While conventional models like Logistic Regression have long been utilized for predicting life insurance ownership due to their interpretability and simplicity, they often require predefined assumptions about the relationships between variables, such as linearity or the inclusion of interaction terms. In contrast, machine learning models, such as ANN, offer greater flexibility by automatically capturing complex, non-linear relationships without the need for such assumptions.

To analyze the complex relationships between these features and life insurance ownership, this study uses SHAP-based interpretation methods. First, global SHAP analysis identifies the most influential predictors for term and cash value life insurance ownership. Next, local SHAP interpretation is used to examine individual cases, illustrating how specific features contribute to household-level predictions. This approach enhances both transparency and interpretability, making machine learning insights more accessible for financial decision-making.

Our findings reveal that the effectiveness of machine learning models in predicting life insurance ownership varies depending on the type of insurance and the characteristics of the dataset. For term life insurance, where the dataset is relatively balanced, ANN demonstrated the strongest predictive performance, achieving the highest AUC-ROC score. In contrast, for cash value life insurance, where the dataset is more imbalanced, Logistic Regression outperformed more complex models in terms of AUC-PR, suggesting that simpler models may be more effective in handling imbalanced datasets. These findings emphasize the importance of selecting models based on data characteristics rather than solely relying

on complexity.

SHAP-based feature importance analysis further highlights the distinct factors driving term and cash value life insurance ownership. Economic factors, including income, financial assets, and debt, play a dominant role in predicting term life insurance purchases, aligning with its function as an income protection tool. In contrast, cash value insurance ownership is more influenced by financial assets, age, and attitudes toward inheritance, reflecting its role in long-term financial planning and wealth transfer. Additionally, psychographic factors, such as risk aversion and bequest preferences, exhibit greater importance in cash value insurance decisions than in term insurance. These insights underscore the nuanced differences in consumer decision-making processes for life insurance and highlight the value of machine learning in uncovering complex, non-linear relationships in household financial behavior.

This paper makes three main contributions. First, it applies interpretable machine learning techniques to the analysis of life insurance ownership, offering a complementary perspective to traditional econometric approaches. Second, it evaluates the performance of both simple and complex models across different types of life insurance, showing how model effectiveness varies depending on data characteristics such as class imbalance. Third, by leveraging SHAP-based feature importance analysis, the study uncovers distinct demographic, economic, and psychographic factors that drive ownership decisions for term and cash value life insurance, highlighting the nuanced and differentiated nature of consumer preferences in this domain.

Following this introduction, the Data and Methodology section provides a comprehensive overview of the dataset and machine learning models used in this study. The Results section begins with an evaluation of prediction performance, followed by global SHAP feature importance and local SHAP interpretations for individual households. Finally, the paper concludes with a summary of the study's key insights, suggestions for future research, and reflections on the implications of feature importance analysis for understanding life insurance purchase decisions.

2. Data and Methodology

This section outlines the data and methodologies used in the study. We describe the dataset, including demographic, economic, and psychographic factors, and discuss the application of machine learning algorithms such as Logistic Regression, XGBoost, GBM, and ANN. Additionally, we explain the use of SHAP-based interpretation methods, which provide both global insights into feature importance and local explanations of individual predictions. The section also details the processes for data preparation, hyperparameter tuning, and model

training, improving understanding of life insurance purchase decisions.

2.1 Data

This study utilizes data from the 2022 Survey of Consumer Finances (SCF), a triennial survey conducted by the Board of Governors of the Federal Reserve System. The SCF is a vital resource for analyzing household finances in the U.S., offering comprehensive data on various financial attributes, including detailed records on life insurance, assets and liabilities, and demographic profiles. The 2022 wave of the SCF includes interviews with 4,595 families. To address missing data, the SCF dataset employs multiple imputation techniques, resulting in five implicates for each survey wave. Multiple imputation replaces missing entries with several plausible values, generating five distinct implicates for each missing data point (Kennickell, 2017). For our analysis, these five implicates are combined into a single observation by averaging numerical features and using the mode for categorical features.

The outcome variables are binary, indicating whether a household owns term life insurance (1 for yes, 0 for no) and whether they own cash value life insurance (1 for yes, 0 for no). For reference, term life insurance typically offers temporary coverage over a fixed period, whereas cash value life insurance combines lifelong coverage with a savings or investment component. The analysis includes three categories of features: demographic, economic, and psychographic.

Demographic features include the age of the household head, gender of the household head (1 for male, 0 for female), education level of the household head (15 levels), marital status of the household head (1 for married, 0 for otherwise), number of children in the household, employment status of the household head (categorized as employed, self-employed, unemployed, and other—retired, student, disabled, etc.), health condition of the household head (measured on a 4-point scale: excellent, good, fair, poor), and race of the household head. **Economic features** include household income, household financial assets, and household debt—each measured in dollar terms and log-transformed to account for skewness in distribution—as well as homeownership, represented as a binary variable (1 for homeowners, 0 otherwise). **Psychographic features** involve the household head's level of risk aversion (measured across four levels: high, moderate, low, and no risk) and the household head's attitude towards bequest (5 levels, where 1 is very important and 5 is not important).

The selection of features is guided by empirical findings in the literature, such as those outlined in previous studies by Lewis (1989), Li (2008), and Bhatia et al. (2021). Summary statistics are presented in Table 1 and detailed descriptions of the data are in Appendix Table A1.

Table 1: Summary Statistics

	Observations	Mean	Std. Dev.	Min	Max
Outcome					
Has_LI_Cash	4,595	0.212	0.409	0	1
Has_LI_Term	4,595	0.446	0.497	0	1
Demographic Features					
Age	4,595	54.469	16.190	18	95
Gender	4,595	0.761	0.426	0	1
Education	4,595	10.330	2.807	0	14
Marital_Status	4,595	0.632	0.482	0	1
Child	4,595	0.739	1.108	0	10
Health_Condition	4,595	2.039	0.807	1	4
Race	4,595	0.598	0.490	0	1
Work_Status	4,595				
Employed		0.496	0.500	0	1
Self_employed		0.217	0.413	0	1
Non_working_other		0.248	0.432	0	1
Unemployed		0.039	0.194	0	1
Economic Features					
Log_Income	4,595	11.685	1.983	0	19.920
Log_Total_Debt	4,595	8.190	5.268	0	18.645
Log_Total_Fin_Asset	4,595	11.171	3.839	0	21.410
Home_Ownership	4,595	0.677	0.468	0	1
Psychographic Features					
Risk_Aversion	4,595	2.982	0.871	1	4
Attitude_Inherit	4,595	2.625	1.497	1	5

2.2 Machine Learning Algorithms

Logistic Regression is a straightforward statistical method for binary classification, mapping predictor variables to a binary outcome using a logistic function. Its simplicity, interpretability, and efficiency make it a common baseline for classification tasks. In this study, Logistic Regression serves as a benchmark against which the performance of more complex models is evaluated.

Gradient Boosting Machine (GBM) similarly constructs an ensemble of decision trees, optimizing a loss function at each stage. While akin to XGBoost, GBM focuses on flexibility and accuracy, making it suitable for modeling highly non-linear relationships (Friedman, 2001), and GBM's versatility across different loss functions further enhances its applicability in various predictive analytics tasks.

Extreme Gradient Boosting (XGBoost) is an advanced gradient boosting technique designed for speed and performance. By building an ensemble of decision trees, XGBoost sequentially improves model accuracy. Its ability to handle sparse and imbalanced data, combined with regularization techniques like L1 and L2, makes it a robust choice for complex predictive tasks (Chen and Guestrin, 2016).

Artificial Neural Networks (ANN) are powerful algorithms recognized for their pattern recognition and classification capabilities. Comprising multiple layers of interconnected neurons, ANNs optimize weights and biases during training to minimize classification errors. The use of SeLU and sigmoid activation functions, along with the Adam optimizer, ensures high classification accuracy (Klambauer et al., 2017; Kingma and Ba, 2014). This study fine-tunes ANN architecture to align with the specific characteristics of the dataset.

Overall, these models were selected to account for different assumptions and complexities in life insurance ownership predictions. Logistic Regression provides an interpretable structure, tree-based models like XGBoost and GBM capture feature interactions without predefined assumptions, and ANN identifies highly non-linear patterns. This selection enables a robust comparison of predictive performance across varying levels of complexity.

2.3 SHapley Additive exPlanations (SHAP)

To ensure model interpretability, this study employs Shapley Additive exPlanations (SHAP), a widely used interpretable machine learning method that quantifies each feature's contribution to model predictions. SHAP applies game theory principles to distribute the predicted outcome among input features, allowing for a consistent and fair attribution of feature importance (Lundberg & Lee, 2017). By computing SHAP values, we can enhance transparency through both global and local interpretation of model predictions.

Specifically, global interpretation identifies the most influential predictors for life insurance ownership across the entire dataset. Using SHAP summary plots, we rank feature importance and assess whether each factor increases or decreases the likelihood of owning term or cash value life insurance. This global interpretation helps align findings with economic theories and prior research on life insurance determinants.

In contrast, local interpretation applies SHAP at the individual case level. By utilizing SHAP force plots and decision plots, we analyze specific households to understand how different features contribute to their life insurance ownership predictions. This local analysis provides personalized insights, identifying key drivers for specific cases—such as correctly classified, misclassified, or borderline households.

SHAP has been shown to enhance model interpretability in finance, improving accuracy and transparency in applications such as credit scoring (Ariza-Garzón et al., 2020; Hjelkrem & de Lange, 2023), and uncovering complex relationships in studies of corporate financial insolvency (Yıldırım et al., 2021) and financial distress (Zhang et al., 2022). The use of SHAP allows this study to bridge the gap between black-box machine learning models and explainable financial decision-making, making our findings both predictively powerful and interpretable.

2.4 Data Preparation and Model Training

Before initiating model training, it is essential to preprocess the dataset to ensure that the features are in an optimal format for the machine learning algorithms. For numerical features such as income, assets, debt, and age, we applied z-score normalization. This technique standardizes the values by subtracting the mean and dividing by the standard deviation, which helps in scaling the data and improving the performance of the model.

Categorical variables were encoded using different strategies depending on their nature. Binary categorical variables, including gender, race, and homeownership, were transformed using binary encoding. For ordinal variables, such as health condition, risk aversion, and attitudes toward bequest, ordinal encoding was applied to preserve the inherent order of the categories. Variables with more than two categories, such as work status, were encoded using one-hot encoding to ensure that the model could interpret these categorical distinctions effectively without implying any ordinal relationship.

To evaluate the performance of our model and prevent overfitting, we employed a k-fold cross-validation method with k set to 5. This approach divides the dataset into five subsets, where the model is trained on four subsets and tested on the remaining one, rotating this process until each subset has been used as a test set. This ensures that the model's performance is robust and not overly dependent on any specific subset of the data.

For hyperparameter tuning, we utilized Grid Search in conjunction with stratified k-fold cross-validation. Grid Search systematically tests a predefined range of hyperparameters to identify the optimal combination that yields the best performance. Stratified k-fold cross-validation ensures that the folds used during Grid Search maintain the same proportion of class labels, which is particularly important in the context of imbalanced datasets. Table 2 reports the hyperparameter grid and model selection.

Table 2: Hyper-parameters

	Grid search	Choice model
Artificial Neural Network (ANN)		
No. of hidden layers	2	2
No. of nodes in 1st hidden layers	8, 12, 16, 32, 64	12
No. of nodes in 2nd hidden layers	6, 8, 12, 16, 32	6
Dropout rate	0.3, 0.2, 0.1	0.1
XGBoost		
No. of trees (n_estimators)	50, 100, 200	200
Learning rate	0.01, 0.1, 0.2	0.01
Max depth of trees (max_depth)	3, 5, 7	5
Min child weight	1, 3, 5	1
Gradient Boosting Machine (GBM)		
No. of boosting stages (n_estimators)	50, 100, 200	200
Learning rate	0.01, 0.1, 0.2	0.01
Max depth of trees (max_depth)	3, 5, 7	5

Once the optimal hyperparameters were identified, the model was trained on the training set using these parameters. The model's performance was then evaluated on the respective testing sets using two established evaluation tools: the Receiver Operating Characteristic (ROC) curve and the Precision-Recall (PR) curve. The ROC curve provides a graphical representation of the trade-off between sensitivity and the false positive rate, where the Area Under the ROC Curve (AUC-ROC) serves as a key performance indicator. An AUC-ROC score of 1.0 indicates perfect classification, while an AUC of 0.5 suggests no better accuracy than random guessing. According to Hosmer Jr et al. (2013), AUC-ROC values between 0.7 and 0.8 are acceptable, between 0.8 and 0.9 are excellent, and above 0.9 signify outstanding performance.

Given the imbalanced nature of some of our data, the Precision-Recall (PR) curve was also utilized to evaluate the model's ability to correctly identify positive instances. The AUC-PR is particularly informative in imbalanced datasets, where a higher value reflects better precision and recall. For the 'Has LI Cash' variable, with a positive class proportion of 21%, an AUC-PR greater than 0.21 indicates that our model performs better than random chance. Similarly, for 'Has LI Term', with a positive class proportion of 45%, an AUC-PR greater than 0.45 suggests superior model performance.

3. Results

This section presents the results of our analysis, evaluating the performance of different machine learning models in predicting life insurance ownership and analyzing the key factors driving these predictions using SHAP-based interpretation. The findings are structured into three parts: first, we assess the predictability of life insurance purchases using performance metrics; second, we examine global feature importance using SHAP; and third, we provide local interpretation for individual predictions using SHAP force and decision plots.

3.1 Predictability of Life Insurance Purchase

In this study, we evaluate the performance of four learning algorithms across two types of life insurance: term life insurance and cash value life insurance. The key evaluation metrics include the Area Under the ROC Curve (AUC-ROC) and the Area Under the Precision-Recall Curve (AUC-PR), as these metrics provide critical insights into model performance, particularly in different data balance scenarios. For term life insurance, which has a relatively balanced class distribution (45 : 55 positive to negative), AUC-ROC is the primary metric for model selection. Conversely, for cash value life insurance, which exhibits a more imbalanced class distribution (21% positive cases), AUC-PR is used as the key metric. In addition to these key metrics, other performance metrics such as Precision, Recall, F1

Score, and Matthews Correlation Coefficient (MCC) are presented in Table 3 to provide a comprehensive assessment of model performance.

Table 3: Model Performance on Testing Set

	Precision	Recall	AUC-ROC	AUC-PR	F1	MCC
Term Insurance						
ANN	0.637	0.720	0.756	0.689	0.676	0.387
XGBoost	0.655	0.635	0.752	0.690	0.644	0.366
GBM	0.628	0.698	0.748	0.686	0.661	0.362
Logistic	0.610	0.680	0.719	0.657	0.643	0.328
Cash Value Insurance						
Logistic	0.349	0.697	0.733	0.384	0.465	0.287
ANN	0.333	0.737	0.727	0.371	0.458	0.277
XGBoost	0.468	0.079	0.719	0.374	0.134	0.119
GBM	0.326	0.718	0.717	0.370	0.447	0.259

For term life insurance, with a nearly balanced dataset, the performance was primarily evaluated using AUC-ROC. The ANN model demonstrated the highest AUC-ROC at 0.756, making it the most effective model for this task. The ANN also achieved strong performance in other metrics, including F1 Score (0.676) and MCC (0.387), further confirming its robustness. XGBoost and GBM followed closely, with AUC-ROC scores of 0.752 and 0.748, respectively, but their slightly lower F1 scores and MCC indicated a lesser ability to balance precision and recall compared to ANN. Logistic Regression, while still performing reasonably well with an AUC-ROC of 0.719, lagged behind the other models, especially in more complex performance metrics, as indicated by its MCC of 0.328.

In the case of cash value life insurance, where the dataset is more imbalanced, AUC-PR was the key metric for evaluating model performance. Surprisingly, Logistic Regression outperformed the other models, achieving an AUC-PR of 0.384, suggesting that it was better suited to handling the imbalanced nature of the data. It also achieved a balanced F1 Score of 0.465 and an MCC of 0.287, reinforcing its effectiveness in this scenario. While ANN demonstrated a higher recall (0.737) and was more successful in identifying positive instances, its AUC-PR of 0.371 and F1 Score of 0.458 did not surpass the performance of Logistic Regression. XGBoost and GBM, although typically strong performers, struggled with the imbalanced data, as reflected in their lower AUC-PR scores (0.374 and 0.370, respectively).

The results of this study reveal that while machine learning models like ANN, XGBoost, and GBM often outperform simpler models in various contexts, the best model choice depends significantly on dataset characteristics and analysis requirements. For term life insurance, which features a more balanced dataset, ANN proved to be the most effective model, as indicated by its superior AUC-ROC score. However, for cash value life insurance, where the data was more imbalanced, Logistic Regression, a simpler benchmark model, outper-

formed the more complex models in terms of AUC-PR. This finding underscores the importance of selecting models that align with the dataset's characteristics, demonstrating that in some scenarios, simpler models can be more appropriate.

3.2 Global Interpretation Using SHAP

Understanding the factors that drive life insurance ownership is crucial for both financial institutions and policymakers. While predictive accuracy is important, interpreting the underlying model and identifying the most influential features provide valuable insights into consumer behavior and decision-making processes. In this subsection, we delve into the feature importance analysis using SHAP, which enables us to assess the ranking of features and how they influence the model's predictions.

A SHAP summary plot provides a visual representation of the impact of each feature on model predictions. In these plots, each point represents an observation in the dataset, showing how a particular feature affects the model's output. The x-axis indicates the SHAP value, where positive values increase the likelihood of life insurance ownership, while negative values decrease it. The color gradient represents the feature's value, with darker shades indicating higher values and lighter shades indicating lower values. While SHAP summary plots provide useful insights into feature importance and general patterns in model predictions, they do not reveal the exact functional form of these relationships. For a precise characterization, additional methods such as SHAP dependence plots or partial dependence plots would be required. However, since our objective is not to establish causality, we do not employ these additional methods here. We emphasize that SHAP values do not imply causal effects but quantify each feature's contribution to the model's output based on correlations. Nevertheless, they offer a valuable foundation for enhancing our qualitative understanding of how machine learning models generate predictions.

Figure 1: SHAP summary plot for term life insurance ownership using the ANN model. Features are ranked by their overall impact on predictions; higher SHAP values indicate greater predictive importance.

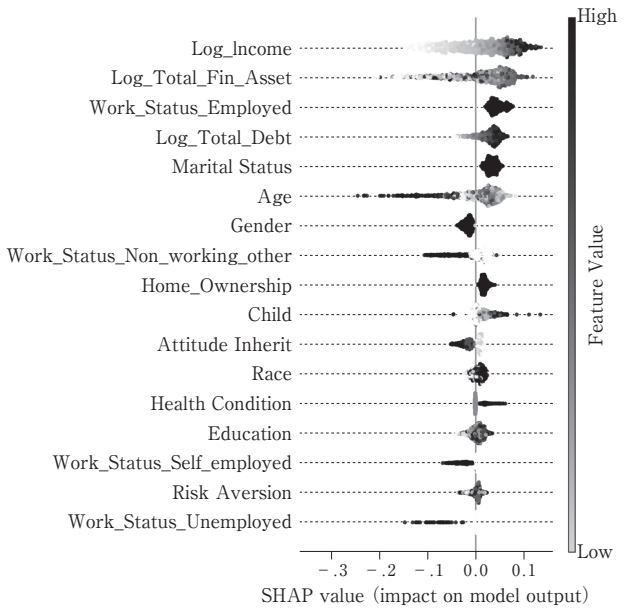


Figure 1 presents the SHAP summary plot for term life insurance ownership, analyzed using the ANN model. The top five features—income, financial assets, employment status, debt, and marital status—emerge as the most significant predictors. These features exhibit high SHAP values, indicating their substantial influence on the model's predictions. Higher income is generally associated with an increased likelihood of term life insurance ownership, reflecting its affordability among individuals with greater financial resources. The distribution of SHAP values for financial assets suggests a potential non-linear pattern in its contribution, as the magnitude and direction of its influence vary across different observations.

Figure 2: SHAP summary plot for cash value life insurance ownership using the Logistic Regression model. Features are ranked based on their overall impact; higher SHAP values indicate greater predictive significance.

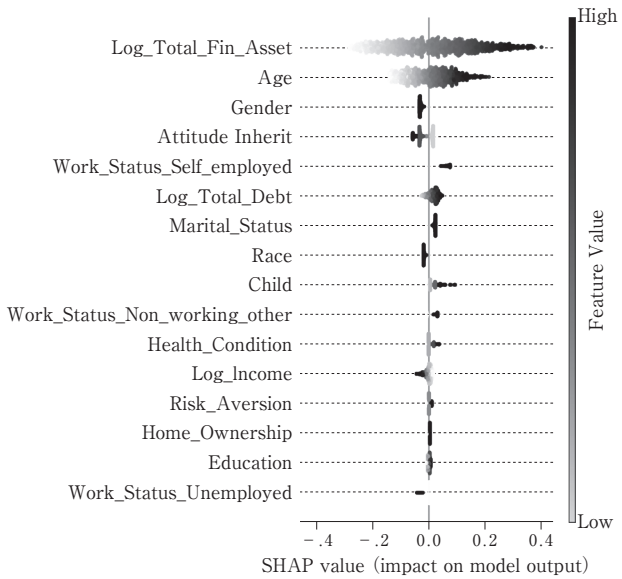


Figure 2 presents the SHAP summary plot for cash value life insurance ownership, generated by the Logistic Regression model. Unlike term life insurance, where income is the dominant predictor, financial assets and age emerge as the strongest predictors for cash value life insurance ownership. These two features significantly outweigh others in their contribution to the model's output, reflecting the different financial motivations underlying cash value life insurance. Households with greater financial assets are more likely to own cash value policies, as these policies are often viewed as investment vehicles with savings components. Age also plays a crucial role, reinforcing the idea that cash value life insurance aligns with long-term financial planning and wealth transfer objectives.

Differences in feature importance between term and cash value life insurance highlight the distinct financial considerations and life stages associated with each product. Term life insurance tends to appeal to individuals in active employment who prioritize income protection, making income, financial assets, and debt management key determinants. Conversely, cash value life insurance is more attractive to those focused on long-term financial planning, where financial assets and age play a more significant role. Attitude toward inheritance also emerges as an important factor for cash value life insurance, suggesting that individuals who prioritize wealth preservation and estate planning are more likely to opt for these policies.

Figure 3: Relative importance of three feature groups in predicting life insurance ownership. The left panel shows the ANN model's predictions for term life insurance ownership, while the right panel presents the Logistic Regression model's predictions for cash value life insurance.

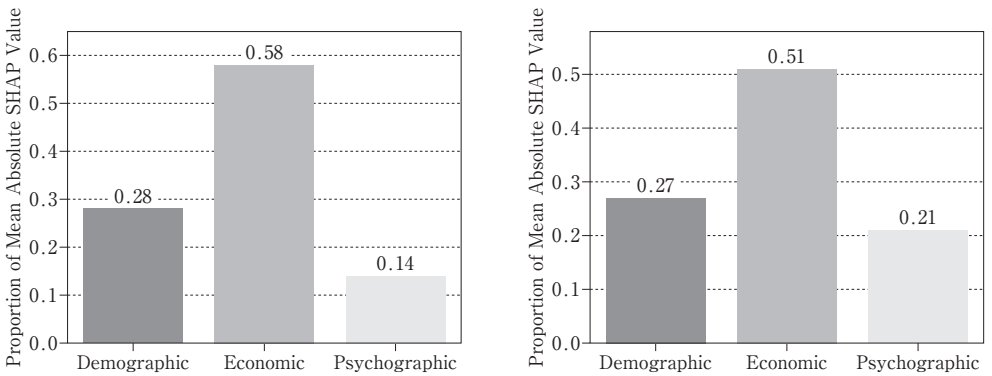


Figure 3 compares the relative importance of economic, demographic, and psychographic feature groups for predicting life insurance ownership. The proportions are calculated based on the aggregated mean absolute SHAP values for features within each group.

Across both types of insurance, economic factors emerge as the most influential. In term life insurance predictions (ANN model), economic variables contribute 58% to the model's decision-making, whereas in cash value life insurance (Logistic Regression model), they contribute 51%. The slightly higher importance of economic features in term life insurance aligns with the notion that term policies are often purchased for income protection, where income, financial assets, and debt management play a crucial role. Demographic features contribute similarly across both models, accounting for 28% in term life insurance and 27% in cash value life insurance. Notably, psychographic factors have a greater influence on cash value life insurance (21%) than on term life insurance (14%), highlighting the role of attitudes, beliefs, and risk preferences in cash value life insurance decisions. This may be due to the long-term savings and investment component of cash value life insurance, which tends to attract individuals with a stronger focus on financial planning, inheritance considerations, and risk aversion.

These findings reinforce that while economic stability is a key driver of both types of life insurance ownership, psychological factors play a disproportionately larger role in cash value life insurance decisions.

3.3 Local Interpretation Using SHAP

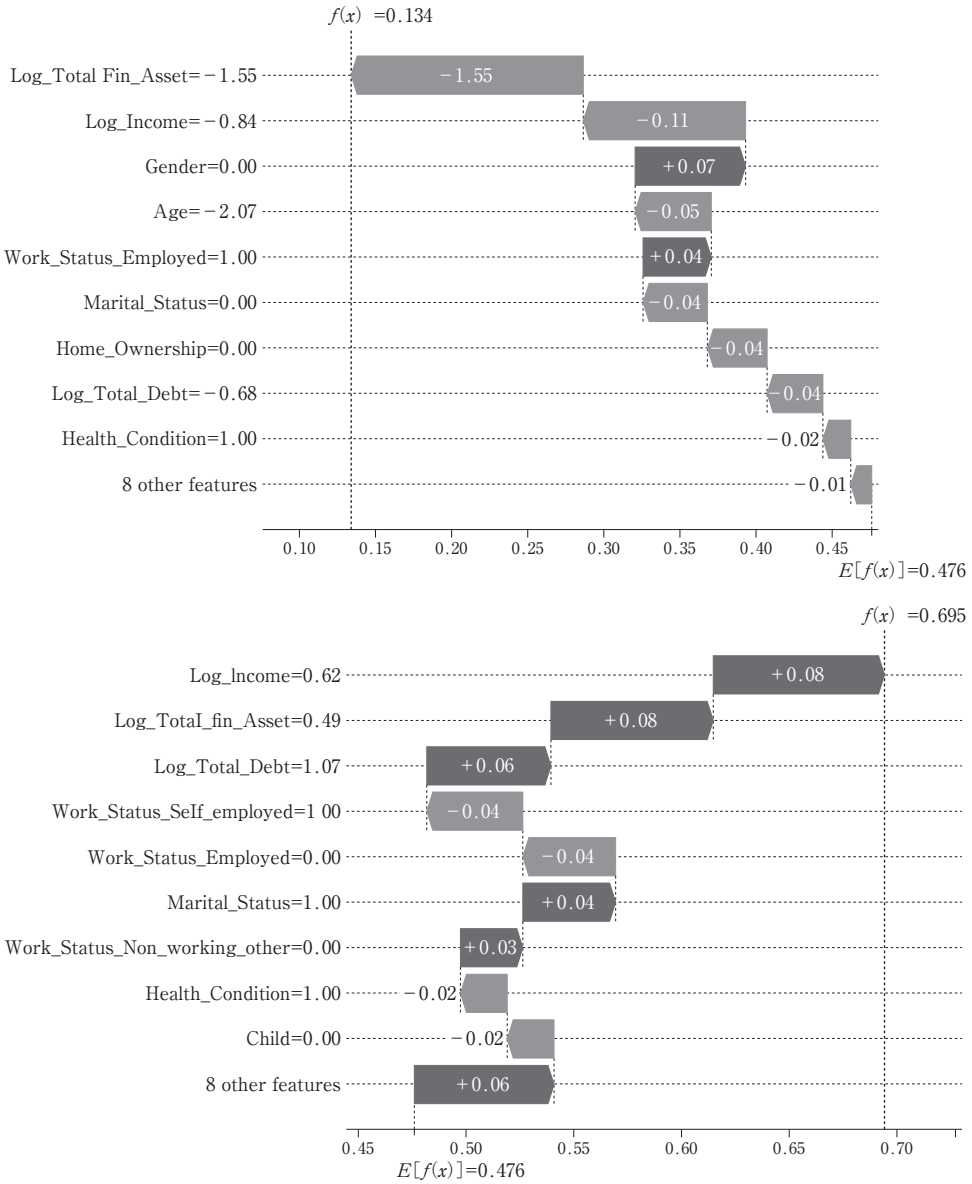
SHAP enables a detailed examination of individual predictions by showing how each feature contributes to the model's decision. To illustrate this, we use waterfall plots, which visually represent the cumulative impact of different features on a specific prediction. In these plots, positive contributions (increasing the likelihood of life insurance purchase) are

shown in black, while negative contributions (reducing the likelihood) are shown in grey. The features are ranked by their contribution, with the most influential ones listed on the left. Each plot starts with a baseline probability—the model's average prediction before considering individual characteristics—and then shows how each feature shifts the final probability for a particular household.

To gain a deeper understanding of the model's decision-making process, we analyze four representative cases, capturing both high and low probability predictions as well as correct and incorrect classifications. By doing so, we highlight which factors strongly influence life insurance ownership at an individual level and how their contributions vary across different households. This case-by-case analysis provides insights into the role of economic, demographic, and psychological factors in shaping model predictions.

Figures 4 and 5 present SHAP waterfall plots for two representative cases of term and cash value life insurance ownership, respectively: one correctly classified as a True Positive (TP) and another as a True Negative (TN). These cases provide deeper insights into how individual features contribute to the model's predictions across different types of life insurance. Since numerical features are normalized using the z-score method, we can easily interpret their levels: a value of 0 represents the mean, while negative values indicate below-average levels and positive values indicate above-average levels.

Figure 4: SHAP waterfall plots illustrating local interpretations of term life insurance predictions by the ANN model. The upper panel (True Negative) shows an individual correctly predicted not to purchase term insurance, while the lower panel (True Positive) shows an individual correctly predicted to purchase term insurance.

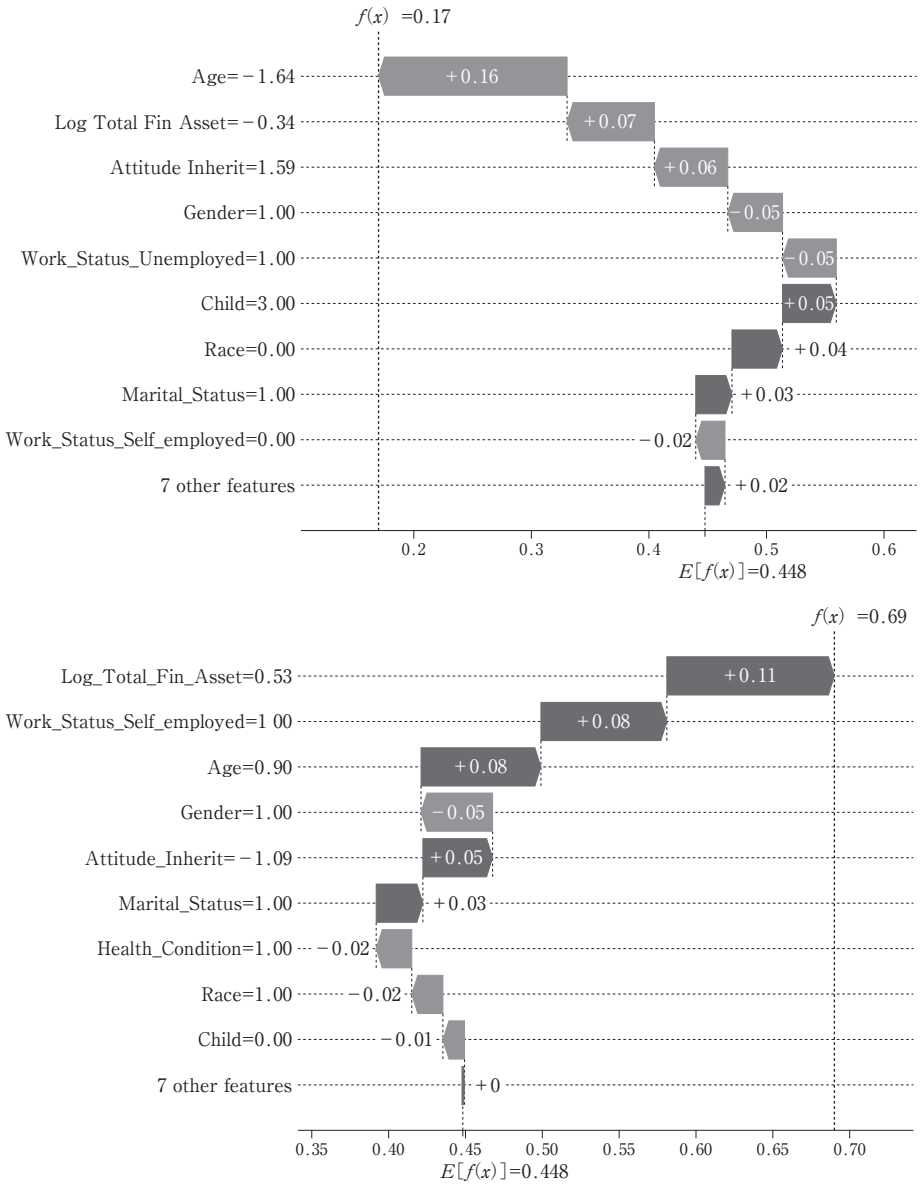


The upper panel of Figure 4 (TN case) illustrates an individual correctly predicted not to purchase term life insurance. The most influential factors driving this prediction are low financial assets (−1.55) and low income (−0.84), both significantly decreasing the likelihood of purchasing term life insurance. Additionally, age (−2.07) further reduces the probability, suggesting that this individual may be below the typical purchasing age for term life insurance. On the other hand, being female (0) and being employed (1) slightly

increase the probability of ownership, but their impact is insufficient to offset the strong negative contributions of financial assets, income, and age. As a result, the overall predicted probability remains below 0.5 (0.134), leading to a final binary classification of non-ownership (0) and a correct True Negative prediction.

The lower panel of Figure 4 (TP case) shows a household correctly predicted to purchase term life insurance. Higher income and financial assets increase the likelihood, reflecting the affordability aspect of term life insurance. Greater financial obligations further contribute to ownership, suggesting a protective motive. Conversely, self-employment slightly decreases the probability, possibly due to alternative risk management strategies.

Figure 5: SHAP waterfall plots illustrating local interpretations of cash value life insurance predictions by the Logistic Regression model. The upper panel (True Negative) shows an individual correctly predicted not to purchase cash value life insurance, while the lower panel (True Positive) shows an individual correctly predicted to purchase cash value insurance.



The upper panel of Figure 5 illustrates an individual correctly predicted not to own cash value life insurance. The key factors reducing the likelihood of ownership include young age, low financial assets, and unemployment, all of which contribute to a lower probability of purchasing cash value life insurance. Additionally, the individual's perception that inheritance is not important, along with being male, further decreases the probability. These

combined influences lead to a correct True Negative classification.

The lower panel of Figure 5 presents an individual correctly predicted to own cash value life insurance. The most influential factors driving this prediction are high financial assets, self-employment status, and older age, emphasizing the role of investment and long-term financial planning in cash value life insurance decisions. Further positive contributions from marital status and a strong belief in the importance of inheritance reinforce the likelihood of ownership, leading to a correct True Positive classification.

These examples illustrate the diverse financial and demographic factors shaping life insurance decisions. While economic stability is a key driver for both term and cash value life insurance ownership, factors such as long-term financial planning and wealth transfer considerations play a more significant role in cash value life insurance. Our analysis provides a case-by-case examination of how different features influence individual life insurance ownership decisions. To gain deeper insights into specific interest groups—such as younger individuals, lower-income households, or highly educated populations—we can leverage SHAP values. By analyzing aggregate measures such as the sum, mean absolute, or mean SHAP values, we can identify which features are most influential in driving life insurance purchase decisions within these groups.

3.4 Discussion

The SHAP-based global and local interpretations reveal that term life insurance ownership is strongly associated with income, financial assets, employment status, debt levels, and family structure. These findings support several key economic theories of household behavior. First, the liquidity constraint hypothesis (Zeldes, 1989) is reflected in the fact that households with low income and limited financial assets are less likely to purchase term life insurance—indicating that present financial limitations restrict their ability to engage in forward-looking financial planning, even when the need for protection exists. Second, the precautionary savings theory (Deaton, 1991) is supported by the observation that higher-income and asset-rich households are more likely to purchase term policies, suggesting that they treat life insurance as a tool to mitigate income risk and uncertainty. Third, the results align with the life-cycle hypothesis (Modigliani & Brumberg, 1954; Ando & Modigliani, 1963), which emphasizes that working-age individuals—especially those with dependents—are more likely to seek income protection through insurance. SHAP values highlight the importance of marital status, number of children, and employment in predicting term life insurance ownership, consistent with the view that individuals tailor financial decisions to their needs at different life stages.

For cash value life insurance, the results also support the life-cycle hypothesis, but reflect a different phase of the life course. SHAP findings indicate that financial assets, age, and attitudes toward bequests are the most influential predictors of cash value insurance

ownership. These features suggest that households purchasing cash value insurance are generally older, wealthier, and more focused on long-term financial planning and intergenerational wealth transfer. Within the life-cycle framework, this reflects a natural progression: as individuals move beyond the income-protection phase of their lives, their financial objectives shift toward saving, preserving wealth, and planning for bequests. Cash value insurance supports these goals by combining protection with a savings component and favorable tax treatment. Thus, while term and cash value insurance serve different purposes, both types of ownership patterns are consistent with life-cycle models that link financial behavior to evolving needs over time.

In addition to aligning with theoretical frameworks, our SHAP-based findings are broadly consistent with earlier empirical studies on life insurance ownership, though many of these studies do not differentiate between term and cash value insurance. Income and financial assets consistently emerge as key predictors in our models, and prior research has similarly found that higher-income households are more likely to purchase life insurance due to greater affordability and a higher opportunity cost of premature death (Duker, 1969; Truett & Truett, 1990; Gandolfi & Miners, 1996). Li (2008) distinguishes between policy types and finds a significant positive effect of income on term life insurance and a weaker effect on cash value insurance. Regarding wealth, our SHAP values indicate a positive influence of financial assets, which aligns with Hau (2000) and Li (2008), who found that asset-rich households are more inclined to purchase life insurance for protection and planning purposes. At the same time, our findings also reflect the nuanced view in the literature that wealthier households may self-insure beyond a certain point (Fortune, 1973; Lewis, 1989). Age also plays a key role: our results show that term insurance is more common among younger individuals, while cash value insurance increases with age—patterns that echo Li (2008) and Baek & DeVaney (2005), who note that age influences insurance decisions differently depending on the product type. While most of the cited studies do not explicitly differentiate between types of life insurance, our analysis contributes a more detailed perspective by linking specific predictors to term and cash value policies separately.

4. Conclusion

This study enhances the understanding of life insurance ownership by applying both conventional and advanced machine learning models to data from the 2022 Survey of Consumer Finances (SCF). Our results highlight that model performance depends on both the type of insurance and dataset characteristics: ANN demonstrated the highest predictive accuracy for term life insurance, while Logistic Regression performed better for cash value life insurance, particularly due to the dataset's imbalance.

Beyond predictive performance, this study emphasizes interpretability by leveraging SHapley Additive exPlanations (SHAP) to analyze feature importance at both global and local levels. SHAP-based global interpretation reveals that economic factors—such as income, financial assets, and debt—are key drivers of term life insurance ownership, while financial assets, age, and bequest preferences are more influential in cash value life insurance decisions. Additionally, psychographic factors play a greater role in cash value insurance, aligning with its function as a long-term financial planning tool. Local SHAP interpretation further illustrates how individual characteristics shape household-level predictions, providing granular insights into why some individuals choose to purchase—or forgo—life insurance.

While this study demonstrates the advantages of machine learning in capturing complex, non-linear relationships, it also highlights the importance of selecting models suited to dataset characteristics and interpretability needs. Although tree-based models and neural networks improve predictive power, their "black-box" nature makes SHAP essential for uncovering the reasoning behind predictions. This approach bridges the gap between accuracy and explainability, making machine learning insights more accessible for financial decision-making.

Nevertheless, certain limitations remain. The analysis relies on U.S. data, which may limit its applicability to other economic contexts. Future research could explore intensive margin decisions—how much life insurance is purchased—rather than focusing solely on the extensive margin of ownership. Additionally, further refinement of interpretability techniques could provide deeper insights into feature effects beyond SHAP, particularly for informing policy recommendations.

Acknowledgment

I extend my sincere thanks to Professor Makoto Kakinaka, Graduate School of Economics, Ritsumeikan University, and Professor Ching-Yang Lin, School of Liberal Arts and Sciences, Musashi University, for their invaluable advice and feedback on this manuscript.

References

Anderson, D. R. and Nevin, J. R. (1975). Determinants of young marrieds' life insurance purchasing behavior: An empirical investigation. Journal of Risk and Insurance, pages 375–387.

Ando, A., & Modigliani, F. (1963). The "life cycle" hypothesis of saving: Aggregate implications and tests. The American Economic Review, 53(1), 55–84.

Ariza-Garzón, M. J., Arroyo, J., Caparrini, A., and Segovia-Vargas, M.-J. (2020). Explainability of a machine learning granting scoring model in peer-to-peer lending. IEEE Access, 8: 64873–64890.

Baek, E. and DeVaney, S. A. (2005). Human capital, bequest motives, risk, and the purchase of life

insurance. Journal of Personal Finance, 4(2): 62–84.

Berekson, L. L. (1972). Birth order, anxiety, affiliation and the purchase of life insurance. Journal of Risk and Insurance, pages 93–108.

Bhatia, R., Bhat, A. K., and Tikoria, J. (2021). Life insurance purchase behaviour: A systematic review and directions for future research. International Journal of Consumer Studies, 45(6): 1149–1175.

Chen, R., Wong, K. A., and Lee, H. C. (2001). Age, period, and cohort effects on life insurance purchases in the U. S. Journal of Risk and Insurance, pages 303–327.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 785–794.

Coe, N. B., Belbase, A., and Wu, A. Y. (2016). Overcoming barriers to life insurance coverage: A behavioral approach. Risk Management and Insurance Review, 19(2): 307–336.

Deaton, A. (1991). Saving and Liquidity Constraints. Econometrica, 59(5), 1221–1248.

Duker, J. M. (1969). Expenditures for life insurance among working-wife families. Journal of Risk and Insurance, pages 525–533.

Fortune, P. (1973). A theory of optimal life insurance: Development and test. The Journal of Finance, 28(3): 587–600.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of Statistics, 29(5): 1189–1232.

Gandolfi, A. S. and Miners, L. (1996). Gender-based differences in life insurance ownership. Journal of Risk and Insurance, pages 683–693.

Garman, E. T. and Forgue, R. E. (2018). Personal Finance. Houghton Mifflin Company, Boston, 13th edition.

Hau, A. (2000). Liquidity, estate liquidation, charitable motives, and life insurance demand by retired singles. Journal of Risk and Insurance, pages 123–141.

Hjelkrem, L. O., & Lange, P. E. D. (2023). Explaining deep learning models for credit scoring with SHAP: A case study using Open Banking Data. Journal of Risk and Financial Management, 16(4), 221.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). Applied Logistic Regression, volume 398. John Wiley & Sons.

Kennickell, A. B. (2017). Multiple imputation in the survey of consumer finances. Statistical Journal of the IAOS, 33(1): 143–151.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv: 1412.6980.

Klambauer, G., Unterthiner, T., Mayr, A., and Hochreiter, S. (2017). Self-normalizing neural networks. Advances in Neural Information Processing Systems, 30.

Lewis, F. D. (1989). Dependents and the demand for life insurance. The American Economic Review, 79(3): 452–467.

Li, M. (2008). Factors influencing households' demand for life insurance. University of Missouri-Columbia.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

Modigliani, F., & Brumberg, R. (1954). Utility analysis and the consumption function: An interpretation of cross-section data. Franco Modigliani, 1(1), 388–436.

National Academies of Sciences, E. and Medicine (2021). High and Rising Mortality Rates Among Working-Age Adults. The National Academies Press, Washington, DC.

Sianipar, A. S. and Hutagalung, A. Q. (2021). The determinants of life insurance ownership. JurnalKeuangan dan Perbankan, 25(1): 93–103.

Truett, D. B. and Truett, L. J. (1990). The demand for life insurance in Mexico and the United States: A comparative study. Journal of Risk and Insurance, pages 321–328.

Yıldırım, M., Okay, F. Y., and Özdemir, S. (2021). Big data analytics for default prediction using graph theory. Expert Systems with Applications, 176: 114840.

Zeldes, S. P. (1989). Consumption and liquidity constraints: an empirical investigation. Journal of Political Economy, 97(2), 305–346.

Appendix

Table A1: Data Source and Measurements of Variables

Abbreviation	Variable	Description
Outcome Variables		
Has_LI_Cash	Cash Value Life Insurance Ownership	Binary variable indicating whether the household owns cash value life insurance (1 = yes, 0 = no)
Has_LI_Term	Term Life Insurance Ownership	Binary variable indicating whether the household owns term life insurance (1 = yes, 0 = no)
Demographic Features		
Age	Age of the household head	Continuous variable representing the age of the household head
Gender	Gender of the household head	Binary variable where 1 = male and 0 = female
Education	Education level of the household head	Categorical variable with 15 levels representing education attainment
Marital_Status	Marital status of the household head	Binary variable where 1 = married, 0 = otherwise
Child	Number of children in the household	Count variable indicating the number of children in the household
Health_Condition	Health condition of the household head	Ordinal variable on 4 levels: 1 = excellent, 2 = good, 3 = fair, 4 = poor
Race	Race of the household head	Binary variable where 1 = White, 0 = non-white
Work_Status	Employment status of the household head	Categorical variable: 1 = employed, 2 = self-employed, 3 = non-working other (retired, student, disabled), 4 = unemployed
Economic Features		
Log_Income	Log of household income	Continuous variable for log-transformed annual household income
Log_Total_Debt	Log of total household debt	Continuous variable representing the log-transformed total household debt
Log_Total_Fin_Asset	Log of total financial assets	Continuous variable for log-transformed total household financial assets
Home_Ownership	Home ownership status	Binary variable where 1 = homeowner, 0 = otherwise
Psychographic Features		
Risk_Aversion	Risk aversion level	Ordinal variable with 4 levels: 1 = high, 2 = moderate, 3 = low, 4 = no risk
Attitude_Inherit	Attitude towards inheritance	Ordinal variable with 5 levels: 1 = very important, 5 = not important