Predicting saving behavior with demographic characteristics, psychological characteristics, and situational factors using machine learning models

Students name: Vitali Tsentsiper

Supervisor: Guy Hochman

Institution: Reichman University (IDC) M.A Behavioral Economics

Introduction:

The current thesis aims to examine the determinants of consumer saving behavior using the 2018 National Financial Capability Study (NFCS). We plan to combine household characteristics, situational factors, and demographic data to understand their compound, moderating, and mediating effects on consumers' saving behavior. As the financial markets develop and products become more complex (Arora et al., 2011), the importance of understanding economic behavior increases.

Financial behavior can be defined broadly as money management (Xiao, 2008), or more specifically, as monitoring individual economic conditions, purchasing, savings, and investments activities (Ozer & Mutlu, 2019). Financial behavior is often categorized into subcategories of cash-flow management, credit management, savings, and investing (Hilgert & Hogarth, 2003). To measure saving behavior, the literature suggests several common methods, such as asking about the difference between current expenditures and income, measurement of the amount saved or invested in the past 12 months (Jayathirtha & Fox, 1996), measuring the difference in wealth (Kennickell 1995), and asking respondents about saving habits, usually when using cross-sectional data (Hogarth & Anguelov, 2003).

A considerable number of academic studies show the importance of psychological characteristics, demographics, and situational factors in saving behavior. In the following section, we present the main categories in each one of the factors and their relationships to saving behavior. Under psychological factors, we present personality traits, self-control, risk aversion, planning horizon, self-perception, locus of control, and income uncertainty. Under demographics, we present the income, race, age, gender, and marriage status. Finally, under situational factors, we present changes in unemployment status, the number of new dependent children, and health status.

Psychological Characteristics

Consumer behavior, which includes purchasing and savings behavior, is deeply affected by personality traits, that is, individuals' thoughts, feelings, time preferences, and intentions (Mowen, 2000). We will investigate psychological characteristics like personality traits that can be mainly divided into The Big Five personality traits (Digman, 1990): Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Researchers have found a link between these traits and saving behavior. Conscientiousness and self-control were found to be positively connected with attitudes toward savings (Brandstatter, 2005) and increased savings (Kausel et al., 2016), and negatively correlated with compulsive buying (Mowen & Spears, 1999). Agreeability was positively related to compulsive buying (Mowen & Spears, 1999) having less savings and more debt (Nyhus & Webley, 2001).

Self-control can be defined as the ability to control impulsive behavior when facing temptations to achieve a certain goal (Willems et al., 2019) or the struggle between hedonic and willpower forces, and the strategies we developed to overcome time-inconsistency choices which leads to impulsive buying and less savings (Hoch & Loewenstein, 1991; Cronqvist & Siegel, 2015). Risk attitude is an essential part of an individual personality and can be represented by the Big Five traits (Ahmad, 2020). Risk aversion is positively related to saving behavior as it motivates precautionary saving motives (Bommier et al., 2012). On the other hand, it is related to lower saving rates which can decrease savings over time (Chen & Huang, 2007). Prior research found that longer planning horizons increase the probability of saving (Rha et al., 2006), and the longer the horizon, the greater are the probabilities for savings (Fisher & Montalto, 2010). Self-perception can be thought of as how people perceive their control over their lives. Individuals with an external locus of control tend to believe that luck and chances are the main factors of achieving goals. In contrast, individuals with an internal locus of control tend to believe that their actions are responsible for goals achievements. Research shows that external locus of control is negatively related to responsible financial behavior as savings (Perry & Morris, 2005). Future income uncertainty is widely shown to be positively related to saving behavior, as uncertainty about future income increases, we see an increase in savings (Juster & Taylor, 1975). This can be explained as precautionary savings (Guariglia & Kim, 2004).

Based on the behavioral life-cycle theory (Shefrin & Thaler 1988), which claims that our self-control, feelings, and emotions are an essential part of decision-making, we assume that psychological characteristics will assist us in predicting saving behavior.

Demographics

For a household to save, the household members need the willingness and the ability to do so (Gerhard et at., 2018). While willingness might be affected by psychological factors, the ability mainly depends on sociodemographic factors such as household size, number of family members, age, race, etc., and financial characteristics such as income, home-ownership, self-employed, etc. We refer to all these factors as demographics. Demographics are an essential part of economic research that can mediate or moderate the results. For example, income has a strong positive relationship with saving, and it is one of the prime factors affecting saving behavior (Hefferan 1982, Chang 1994). Race also relates to savings, and white families tend to save more than any other ethnic household (Lee & Hanna, 2015, Rha et al., 2006). Age is positively related to savings when all other variables are held constant (Chang, 1994). However, young households (e.g., with heads of the family under 30 years) are more likely to save than households with older heads of the family (Yuh & Hanna, 2010). Interestingly, exposing participants to future selves reduced their discount rates and increased total savings

(Hershfield et al., 2011). Finally, while gender has no direct effect on savings, it has a moderating effect on race, age and unemployment rate (Whitaker et al., 2013). Married couples tend to have better financial conduct, especially if the husband has better wages, consistent participation in the labor market, and employer benefits (Wilmoth & Koso 2002).

Based on the life cycle theory (Modigliani & Brumberg, 1954) that claims that households tend to smooth consumption over a lifetime, we believe that demographics will strongly affect the total current household wealth and will help us predict saving behavior.

Situational factors

The situation has a strong effect on human behavior (Ross & Nisbett, 2011), sometimes even beyond one's dispositional variables (Darley & Batson, 1973; Milgram 1975; Darley & Latane, 1968). In general, the literature divides situational factors into three main categories: Cues, Characteristics, and Classes. Cues are objective descriptions of the environment and can be represented as interactions, objects, location, activities, and time. For example, the situational cues for a party contain a description of what people are doing, the location, the date, who is attending, the weather at that time, and how these objective cues affect the behavior. Characteristics reflect the subjective perception of events and cues such as conflict, pleasant, negative, intellectual, social, etc., and can describe a situation, while Classes condense the entire situation into a class based on cues and subjective perception as party situation, trading situation, working situation, conflict situation, etc. (Rauthmann et al., 2015). In the current thesis, we refer to situational factors as Cues as it is less susceptible to differences in interpretation.

Sudden changes in a situation, such as unemployment or health, can be referred to as shock, which can affect people's consumption (Mullainathan & Shafir, 2009). Prior research finds that a change in the unemployment rate in a state is expected to have a positive effect on savings due to an increase in expected uncertainty regarding future income (Juster & Taylor 1975), and at a household level, a change in unemployment status has a negative relation to saving behavior (Fisher 2006). An increase in the number of dependent children in a household is negatively related to savings (Lugauer et al., 2019), and emergency health events might create financial stress that leads to an increase in household debt and lower savings (Babiarz et al., 2013).

Based on the economic Permanent Income Hypothesis (Friedman, 1957) and the Prospect theory (Kahneman and Tversky, 1979) that claim that the change in consumption and saving behavior is also affected by situations, contexts, and our expectations, we believe situational factors will assist us in predicting saving behavior.

Theoretical background

The life cycle hypothesis (Modigliani & Brumberg, 1954) is an economic theory that assumes that individual consumers try to maximize their utility by smoothing consumption over the expected lifetime. At every point in life, people have their present value of wealth, which is their current and expected future discounted sources. This affects purchase and saving behavior through increased savings when income is high and credit and borrowing otherwise.

The permanent Income Hypothesis (Friedman, 1957) is an economic theory that assumes that households try to smooth their consumption over the entire lifetime and to be prepared for income decline. The theory claims that change in consumption is less volatile than the change in income. The household will change its consumption when there is a sudden longterm shock to their permanent income, contrary to unexpected shocks to their temporary income.

The behavioral life-cycle theory (Shefrin & Thaler, 1988) is a behavioral finance theory that assumes that although households want to smooth their consumption, they are affected by self-control, emotions, and cognitive biases, which challenge them to save and tempts them to consume and differentiate between assets which can be labeled as under high temptation to spend on consumption and lower temptation to spend on consumption.

Prospect theory (Kahneman & Tversky, 1979) is a behavioral economics theory of decisions under risk developed following criticism and as an alternative to expected utility theory (Modigliani & Brumberg, 1954). The researchers argue that people presented with alternatives involving risk, will assess different probabilities of gains and losses relative to a reference point. The reference point serves as a status quo, while the value function is different between gain and loss domains. Researchers claim that due to an asymmetrey between gain and losses, people behave differently when there is a potential loss and a potential gain. In a gain domain, the value function is concave due to risk aversion, and in the loss domain, the value function is convex due to risk-seeking behavior. In addition, the function is steeper for losses. When a household encounters a new situation, its decisions and behaviors are affected by its reference point. Situations can be presented and perceived as an opportunity relative to the reference point, or as unfavorable when presented as a loss relative to the reference point and lead to different behaviors.

Research question: What are the main factors contributing to the prediction of household saving behavior?

Hypotheses:

Psychological factors:

H1: Self-control is positively related to saving behavior

H2: Risk aversion is positively related to saving behavior

H3: External locus of control is negatively related to saving behavior

H4: Future income uncertainty is positively related to saving behavior

Demographics factors:

H5: Income has a positive relation to saving behavior

H6: White families tend to save more than any other ethnic household

H8: Married status is positively related to saving behavior

Situational factors:

H9: Change in unemployment status has negative relation to savings

H10: Increase in the number of dependent children in a household negatively related to saving behavior

H11: Emergency health event is negatively related to saving behavior

Method

Our main purpose in this thesis is to build a model that will predict saving behavior of households using demographics, psychological characteristics, and situational factors based on theoretical and empirical findings. In some cases, research in psychology measures the accuracy of a model based on the fitness between the model and the sample data, or whether the size or direction of a coefficient matches the theoretical explanation, but lacks the accuracy when it comes to predicting behavior. Using machine learning techniques that are testing out of sample data, we will be able to increase the understanding of the existing theories and findings (Yarkoni & Westfall, 2017). In the following sections, we present our data, the collection method, and machine learning algorithms that will be used for prediction.

Data set

We will use the 2018 National Financial Capability Study Well-Being Survey (NFCS) data, collected by the Consumer Financial Protection Bureau. In the survey data, saving behavior (target variable) will be measured by the amount of money in the savings account reported by the respondents in one of the seven categories: $0, $1-99, $100-999, $1,000-4,999, $5,000-19,000, $20,000-74,999, $75,000 or more. We will use machine learning models and techniques to predict this target. The survey was conducted in English and Spanish between October 21, 2016, and December 5, 2016, in the 50 states of the U.S and Washington D.C. Only adults aged 18 and above were sampled, and only one panelist per household. The final survey includes 6,394 participants who answered a total of 217 questions. The survey was done by GfK Group using KnowledgePanel, the largest U.S. probability-based non-volunteer internet panel, which allows for stratification of the sample with oversampling to subgroups with low representation. The randomization of the sample was done using address-based sampling (ABS). To ensure that the collected data represents U.S. population segments, weights are reported for each group by age, sex, race, poverty, and education. Although weights are very important for explaining relationships, they might be insignificant for prediction purposes. The purpose of this survey is to measure the current state of financial well-being of American adults among subpopulations. The questions in this survey represent information about individuals, households, and families, income and employment, saving and safety nets, financial experience, financial behavior and attitude, financial knowledge, social context, and personal traits. We will recognize psychological features using a combination of answers based on literature and inventories, while demographics and situational factors can be identified by feature name. Some features include answers such as “Prefer not to say” and “I don’t know”. these answers will be removed from the data or imputed.

Machine learning algorithms

Using the questions as features, we will use supervised machine learning algorithms to predict the class of a household saving amount on an out-of-sample dataset. We will examine the possibility of using two-stage prediction pipelines. First, predicting saving vs. non-saving behavior, and then predicting specific classes of saving behavior (Osman, 2019). Using feature engineering, we will extract the features based on literature that are most important for the prediction of saving behavior. Following that, we will use algorithms such as Support Vector Machine (SVM), Decision Tree, Random Forest, and gradient boosting models such as LightGBM and XGBoost to predict the class of saving behavior. These models are used in prior research for behavior prediction (Levantesi & Zacchia, 2021; Osman, 2019). The results will be compared to Logistic Regression performance as a sanity check, because behavior tends to be nonlinear and complex (Jenkins et al., 2017). To evaluate the results, we will use evaluation metrics such as Precision, Recall, F1 Score, and ROC AUC. Finally, we will examine which features are the most important for behavior prediction. In the following section, we will elaborate on machine learning models and evaluation metrics.

Machine learning models

a. Logistic Regression

Logistic Regression is a statistical model which assesses the probability of a certain class by using sigmoid function on linear regression to turn it into logistic regression.

b. Support Vector Machine

Support Vector Machine (SVM) is a supervised learning algorithm that tries to maximize the margin distance between data points (support vectors) and hyperplanes, which are decision boundaries that help us classify data into certain classes.

c. Decision Tree

Decision Tree is a nonparametric supervised learning algorithm with nodes representing the features, branches representing the decision rules and leafs representing the output. The model learns how to classify the target into a certain group by decision rules learned from the training data and can be thought of as a sequence of "if"s. The model finds the best attributes of features to divide classes into homogeneous groups with low impurity or noise inside classes.

d. Random Forest

Random Forest is an ensemble method, non-parametric supervised learning algorithm which builds many decision trees, the "forest", in parallel. The Decision is built using subsamples from the training set with replacement and the output is aggregated using majority vote. In such a way the model reduces the variance in the base estimator (Decision tree), and results are more generalized and less overfitted.

e. XGBoost

XGBoost is a gradient boosting model. Gradient boosting technique is an ensemble learning which uses weak learners as tree-based algorithms in a sequential order to convert them to strong learners by fixing their predecessor's error. In each iteration the algorithm tries to minimize the loss function of a weak learner using gradient descent optimization.

f. LightGBM

LightGBM is a gradient boosting model which uses gradient-based one-side sampling (GOSS). GOSS technique, selects the splits that affect the loss function by choosing the instances with the highest gradients and combining them with a random sample of instances with lower gradients which will contribute to larger information gain.

Evaluation metrics

a. Precision is the fraction of positive predictions that were correctly classified from all predicted positive cases for each class. This can be expressed as TP/(TP+FP). Where TP is true positive (number of cases predicted positive and were actually positive) and FP is false positive (number of cases predicted positive but were actually negative). Precision is used when our goal is to minimize errors even in the cost of some cases going undetected.

b. Recall is the fraction of positive predictions that were correctly classified from all actual positive cases for each class. The recall can be expressed as TP/(TP+FN). FN stands for false negative (number of cases predicted negative and were actually positive).

c. F1 score is a harmonic mean of recall and precision. The formula for F1 score is 2*(precision*recall)/(precision + recall). This metric will be high when both recall and precision are high.

d. ROC AUC measures a two-dimensional area under the curve of TPR and FPR that represents the performance of all possible classification thresholds. Where TPR stands for true positive rate (as recall) and FPR stands for false positive rate, probability for false alarm. TPR is expressed as TP/(TP+FN) and FPR is expressed as FP/(FP+TN). TN stands for true negative (number of cases predicted negative and were actually negative). As AUC is reflecting the overall ranking performance of a model, it is appropriate for algorithm comparison.