Capstone Project #2: Predicting Future Financial Well-Being using CFPB Survey Data

By: Justin Hughes-Coleman

Project Proposal

The Consumer Financial Protection Bureau(CFPB) published a survey in September 2017 that measured the financial well-being of Americans. The survey was split into two categories, present and future financial well-being.The survey produces a score from zero to 100 with a higher score meaning better financial security.Currently the CFPB can't use the study to measure time series differences but they are looking for benchmarks that can be established.

The CFPB defines financial well-being as "a state of being wherein a person can fully meet current and ongoing financial obligations, can feel secure in their financial future, and is able to make choices that allow them to enjoy life. " (page 7 of cfpb_financial-well-being-scale-technical-report). There are four elements of financial well-being; present and future security and freedom of choice. The logic follows that someone that is in good financial standing in the present (they can exercise financial freedom and are currently financially secure) this should lead to good financial well-being in the future.

The goal of this project is to use the present financial well-being data to predict the future financial well-being data. The project will have a user answer the questions provided by the CFPB that accompanied the survey and will use those answers to predict a person's financial future. This project can be used to further map a specific income based on the scores that are produced by the survey. The survey data will be analyzed to find the most significant factors in understanding financial stability. Then these features will be measured to find their predictability for future stability.

We have data regarding prospective finances that can be used in this analysis. After comparing present and future variables, a model will be built that uses the data from the current to predict expected financial stability. If a person's financial well-being can be successfully predicted, then further actions in helping that person or additional analyses can be made.

The utility in this project can be exercised when an individual can accurately map their financial stability and can make decisions that impact their financial future. For instance, an individual could decide they need to save more money in order to increase they financial prospects. Though someone might have their assumptions about their future prospects, since we have data from a wide range of incomes, we can help those individuals see the long-term trends of their current financial well-being and can make changes to alter the course for the future. It will help those who do better with a plan and like to see everything laid out in front of them.

The aforementioned benchmarks could be tied to specific survey scores in order to help future analysis. Let's say for scores 30-50, it is better for them to cut down debt that simply get more income, that could be a helpful gauge of someone's standing. These guidelines could be measured over time to indicate to measure they effectiveness in predicted financial prosperity.

Data Wrangling

The dataset is provided by the CFPB on their website in a multitude of formats. Alongside the dataset are three documents that explicitly explain the purpose and response percentages for every variable. As mentioned above, the survey has questions regarding different aspects of the participants' lives in order to get a better picture of their financial well-being of the participants. For the sake of this project, the focus will be on the explicit socio-economic standings of the participants and their responses to the financial knowledge questions that were apart of the survey. The dataset contains 6394 samples from participants of the survey.

The socio-economic survey items centered on ethnicity, gender, marital status, age group and generation (boomers, gen x, millennials, etc.), highest level of education completed, current employment status, household size, number of children, household income, and whether the participant is living above the federal poverty line. Since this information was vital for accurate analysis the CFPB went through great lengths to make sure to get an accurate sample across income brackets and attempt to get more minority participants. The ratio roughly aligns with the national demographics (page 9 of cfpb_nfwbs-puf-user-guide).

The other aspect of the data that is relevant to this analysis in the financial knowledge questions. Over the course of the survey three different financial knowledge tests are asked to the participants and a score is given based on the number of questions answered correctly in each questionnaire. The questionnaires each have a different focus for instance, the Knoll and Houts survey asks questions regarding interest rates, stocks and bonds. The FINSOC questions are regarding financial knowledge either taught or discussed by the participant and their household. (page 4 of cfpb_nfwbs-puf-codebook).

There are many more variables in the dataset, many of which are relevant to the question being asked in this report. However, they are either redundant (a discrete variable for a set of boolean variables) or erroneous for this study (price brackets for mortgage payments when there is a variable regarding present financial liquidity). The variables selected for this analysis should provide the most succinct answer to the question of predicting future financial security. Through further analysis, both statistical and machine learning, the weight each of these variables plays will become illuminated.

The following variables were categorical. As such, they were one-hot encoded for further analysis. "FS1_1", "FS1_2", "FS1_3", "FS1_4", "FS1_5", "FS1_6", "FS1_7", "FS2_1", "FS2_2", "FS2_3", "SUBKNOWL1", "KHscore", "KIDS_1", "KIDS_2", "KIDS_3", "KIDS_4", "EMPLOY1_1", "EMPLOY1_2", "EMPLOY1_3", "EMPLOY1_4", "EMPLOY1_5", "EMPLOY1_6", "EMPLOY1_7", "EMPLOY1_8", "agecat", "generation", "PPEDUC", "PPETHM", "PPGENDER", "PPHHSIZE", "PPINCIMP", "fpl", "PPMARIT". The remaining variables were converted to integers. Now the entire dataset is in integers and the categorical variables won't be tied to their numerical values.

Missing Values

Outlined in the CFPB User Guide for the dataset, there are some variable inputs that are nonsubstantive. The values -4, -1 and 99 are all in the selected variables and their corresponding rows will be deleted. They refer to non-responses in the dataset. The proportion of these responses is never more than 10% with the vast majority being less than 0.5%. After the null values were dropped there was a total of 5736 samples left to analyze. This was a little less than 10%. The reason for such a big portion being dropped is the question KIDS_NoChildren almost 10% of respondents refused to answer the question. This leaves a large portion of data incomplete. In order to complete the analysis these missing values will be dropped. Over 90% of the data remains and no other data needs to be removed. In the future, I would like to explore a method where I can distribute the null values proportionately among the different discrete values of the variable. For now, the values will be dropped.

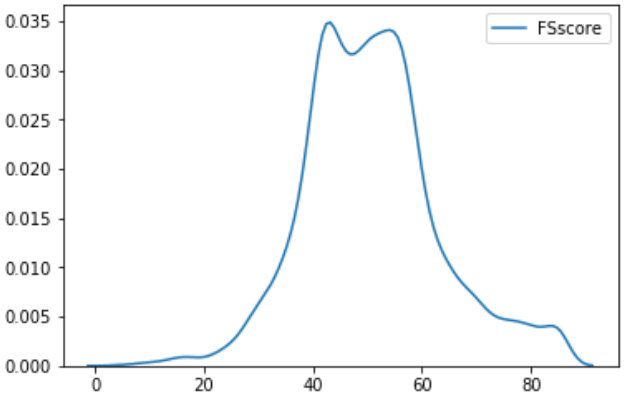
Exploratory Data Analysis

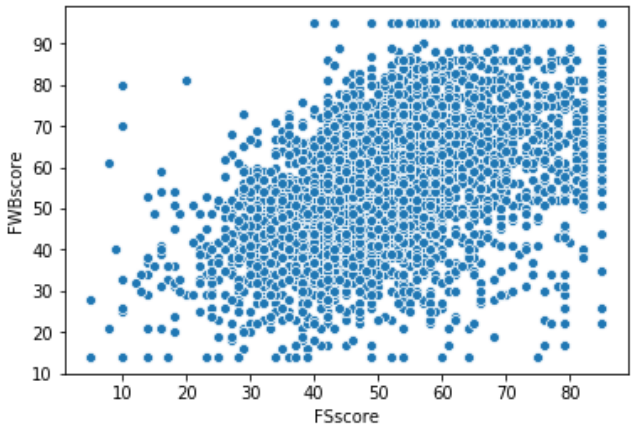
The target variable will be the FWBscore. This score is a composite between a questionnaire asking the participants about their financial well-being and the number of questions they answered correctly in the same survey. This is the main variable that is used to assess a participant’s financial well-being; a higher score means they are more financially well-off than someone with a lower score. The purpose of this project is to assess whether financial literacy and socio-economic background plays a part in predicting the financial well-being of an individual. Furthermore, if these variables do factor into well-being, to what extent. The focus of this project will be to find the most significant factors in predicting financial well-being with respect to financial literacy and socio-economic standing.

The null hypothesis is financial well-being is not sufficiently determined by financial literacy and socio-economic background. This will be tested by finding a correlation between the FWBscore target variable and any of the socio-economic or financial literacy features. The alternative hypothesis is that there is a strong correlation between socio-economic standing and financial literacy in regards to predicting financial well-being. This will be tested by splitting the data into their respective types of data; continuous and discrete, and evaluating various measures.

The discrete variables are all the variables except the target variable, FWBscore and a financial literacy feature FSscore. FSscore is based on a series of statements asked to the participants regarding financial skills. These statements were rated either yes or no by the participants and range between knowing how to follow-through on financial agreements to understanding the terms of a credit card. The more the participants' answered in the affirmative, the higher their FSscore. A KDE plot of the FSscore illustrates that there are the curve is skewed to the right with two humps near the middle of the distribution. This means the typical FSscore is slightly

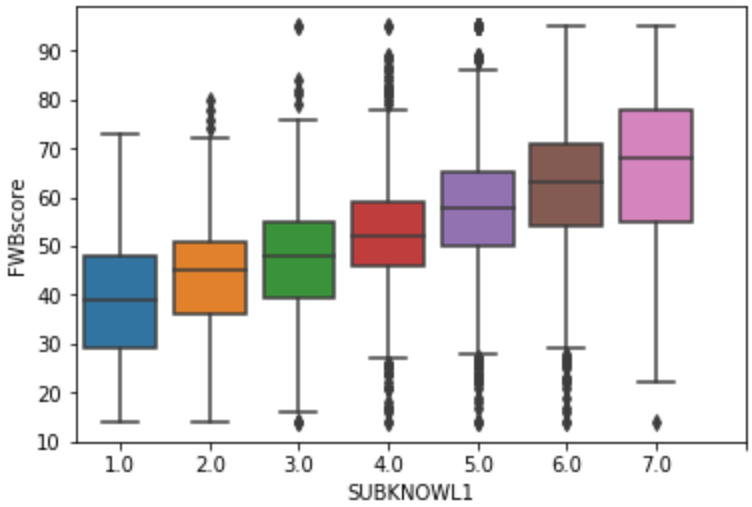
higher than the halfway score for the majority of participants. Next to this plot is a scatter plot to show the correlation between FSscore and FWBscore. As one might expect there seems to be some correlation between the variables. One is a culmination of all the financial literacy data gathered by the researchers and the other is the score after participants answered a series of financial statements. A pearson correlation test revealed there is a correlation between the two variables of 0.49 and a p-value of 0.0. Pearson was chosen because FSscore is used to calculate FWBscore so a shift if the prior should lead to a proportional shift in the latter. This failed to reject the null hypothesis which means there is some correlation between FSscore and FWBscore.





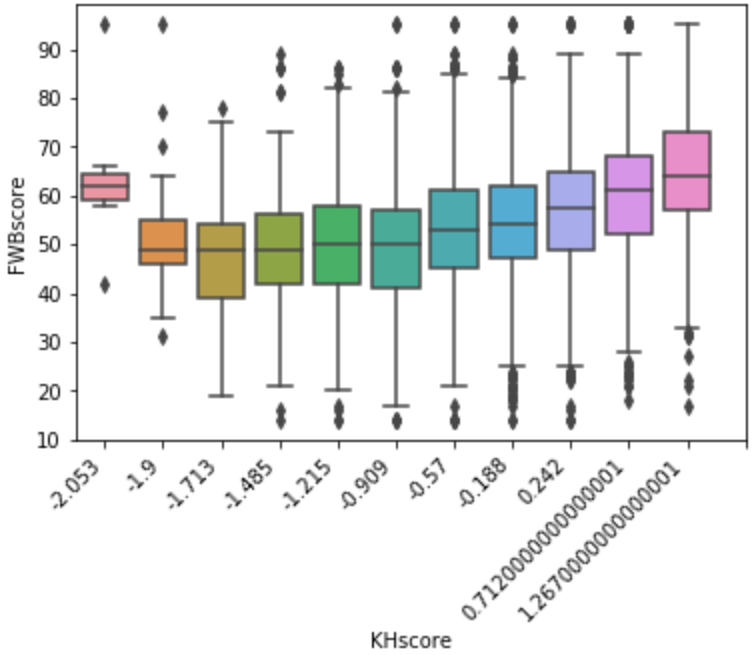
As previously mentioned, the majority of the features are discrete variables. They are the socio-economic characteristics of the participants. As such, they are categorical in nature. The other financial literacy tests that were administered were also scored on an ordinal basis so everyones' scores were grouped into a set amount of scores.

SUBKNOWL1 is a statement asked to participants asking them to assess their overall financial knowledge. They were given a scale from 1-7 with 1 being very poor. The following boxplot shows the distribution of responses compared to the FWBscores associated with each response. As one would expect, the higher the SUBKNOWL1 score the higher the FWBscore.

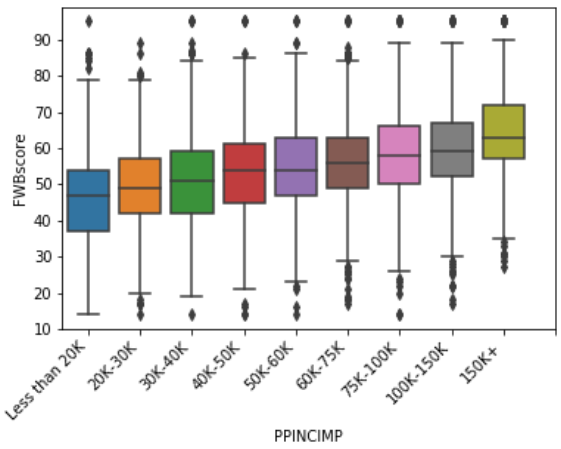


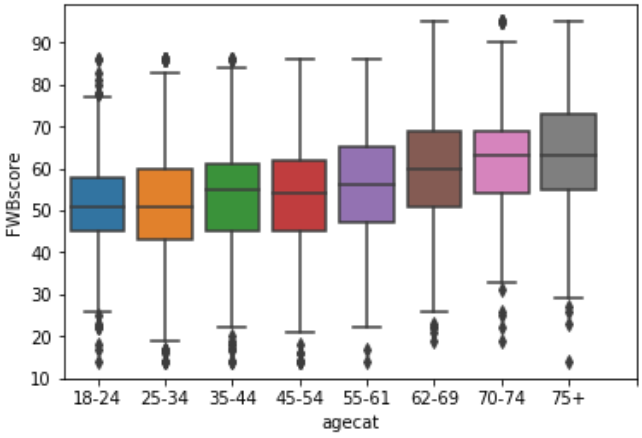
There are some peculiarities with this variable. There are many outliers in the 4-6 range that actually score pretty low on the FWBscore. This means a significant amount of participants are overestimating their financial knowledge.

The KHscore is the culmination of correct answers to the Knott and Houts financial literacy test. Though the test is administered as a continuous variable, for the sake of the survey, the results have been compiled into a discrete variable. The values of the scores are essentially from low financial knowledge to high financial knowledge. The following chart illustrates the distribution compared against the FWBscore. Similar to the FSscore, the higher the KHscore the higher the FWBscore. Also similar to the FSscore, a significant amount of people who have answered the questions correctly about half to two-thirds of the time don't have very good FWBscores. This may be do to participants knowing sound financial advice but not following it themselves.



The socio-economic categories covered areas such as age, ethnicity, gender, income, household size ,education level and number of children. Number of children, household size, gender, and ethnicity didn't show any significant changes between the categories. All were more or less centered around the median FWBscore of 50. Income showed similar results to the financial literacy questions. Below shows a general positive correlation between income and FWBscore; the higher the income the higher the score. Also in the $60K-$150K range there seem to be many outliers that have low FWBscores.



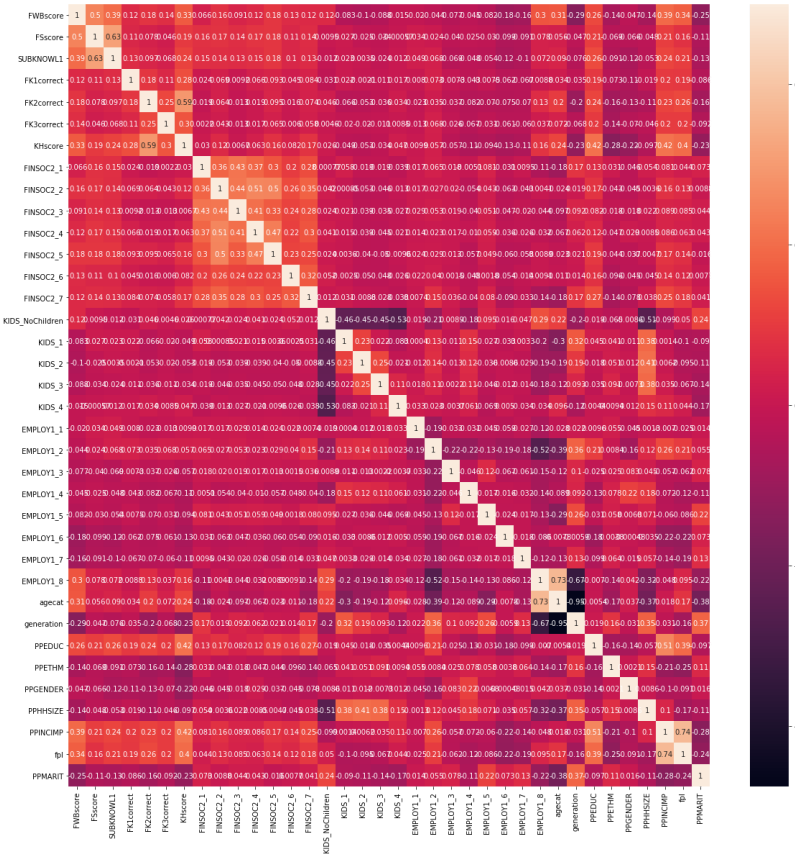


Similarly, the age categorical variable shows results in line with the income variable. The older someone gets the higher FWBscore with some noticeable outliers.

Feature Selection

Due to the 55 variables that have been selected for exploratory analysis, it is necessary to reduce the number of variables in order to get the most accurate results from training the model in the following step. First, the question responses from the financial questionnaire will be removed. They are directly used to calculate the cumulative score for each group (i.e. FS1_1 - FS2_3 are all used to calculate FSscore). These represent collinearity in the dataset and will greatly hamper the performance of the model. These questions are essential to helping users of this project map their financial knowledge but this can be applied after they participate. If they get a less than desirable FWBscore, then they should take the associated financial literacy tests, see what areas they are lacking and improve from there. The question results themselves don't need to be included in the dataset because their values are already accounted for in the score variables.

After the questions are removed a heat map will help determine which variables should be kept. Below is the heat map with the remaining variables. There are a few highly correlated variables; generation and agecat as well as fpl and PPINCIMP. This makes sense, the prior are age groups and the latter are income groups. For this project, generation and PPINCIMP will be kept since this project deals directly with generational changes in income; habits users can do to make an impact in 10-15 years.


FWBscore
FSscore
SUBKNOWL1
FK1correct
FK2correct
FK3correct
KHscore
FINSOC2_1
FINSOC2_2
FINSOC2_3
FINSOC2_4
FINSOC2_5
FINSOC2_6
FINSOC2_7
KIDS_NoChildren
KIDS_1
KIDS_2
KIDS_3
KIDS_4
EMPLOY1_1
EMPLOY1_2
EMPLOY1_3
EMPLOY1_4
EMPLOY1_5
EMPLOY1_6
EMPLOY1_7
EMPLOY1_8
agecat
generation
PPEDUC
PPETHM
PPGENDER
PPHHSIZE
PPINCIMP
fpl
PPMARIT
0.8
0.4
0.0
−0.4
−0.8