PREDICTING FINANCIAL WELL-BEING: A MACHINE LEARNING APPROACH

Final Project for Course PPOL 565: Data Science II – Applied Statistical Learning, Spring 2020

Andy Green
Georgetown University
May 4, 2020

Table of Contents

Executive Summary

This report explores the various factors that help explain Americans' financial well-being. Specifically, it investigates the demographic, behavioral, and skill/knowledge-based factors that contribute to an individual's sense of financial well-being. The goals of the report include:

- Predicting which individuals are most at risk for being unable to achieve financial well-being
- Identifying which factors are most effective in predicting an individual's financial well-being

In order to investigate these questions, I utilize data from the Consumer Financial Protection Bureau's National Financial Well-Being Survey. The dependent variable used to measure an individual's financial well-being is the CFPB Financial Well-Being Scale Score. The scale is designed to measure the following four components (Consumer Financial Protection Bureau, 2017b):

1. The ability to be in control of daily and monthly finances
2. The ability to withstand a financial shock
3. The ability to stay on track in pursuing financial goals
4. The freedom from financial burdens that prevent one from enjoying life

In order to predict an individual's financial well-being, I focus on 10 independent variables that were identified by a forward stepwise selector algorithm to be the most effective variables for prediction purposes. These variables represent concepts including:

- Possessing strong financial skills, including being able to process financial information, control spending, and make financial decisions
- Maintaining sound financial habits, such as paying off monthly credit card balances and regularly adding money to savings
- Following through on financial goals
- General stress levels
- Being retired
- Household income

After identifying the most effective independent variables, I utilize both ordinary least squares linear regression (OLS) and a k-nearest neighbors regression algorithm (KNN) to predict financial well-being. While OLS initially appears to outperform KNN in terms of predictive accuracy, the two methods ultimately end up producing nearly identical results after standardizing the variable scales and optimizing relevant hyper-parameters in the KNN model.

OLS produces an average cross-validated R^2 of 0.598, and predicts values that are 6.85 points, or 13.9%, away from the true value of the dependent variable, on average. KNN produces an average cross-validated R^2 of 0.603, and predicts values that are 6.78 points, or 14.1% away from the true value of the dependent variable, on average.

Background

In a recent report, the Federal Reserve estimated that roughly 40% of American adults would struggle to cover an unexpected expense of $400, being forced to go into debt or sell something to cover it (Federal Reserve, 2019). Additionally, they estimate that even without the presence of an unforeseen expense, 17% of adults are unable to pay off their bills in full in a given month (Federal Reserve, 2019). Such statistics speak to the relative fragility of Americans' financial lives, despite the fact that typical economic indicators pointed to a generally strong economy prior to the outbreak of the COVID-19 epidemic. The unemployment rate was at near-historic lows (Bartash, 2020), with GDP growth hovering around 2-3% in recent years (Cox, 2020).

This report explores the various factors that help explain Americans' financial well-being. Specifically, it investigates the demographic, behavioral, and skill/knowledge-based factors that contribute to an individual's sense of financial well-being. The goals of the report include:

- Predicting which individuals are most at risk for being unable to achieve financial well-being
- Identifying which factors are most effective in predicting an individual's financial well-being

One useful measure of financial well-being is the CFPB Financial Well-Being Scale. This scale was created by the Consumer Financial Protection Bureau in an effort to develop an objective and consumer-centric measure of financial well-being that could be used widely within the field of consumer finance. The scale is designed to measure the following four components (Consumer Financial Protection Bureau, 2017b):

1. The ability to be in control of daily and monthly finances
2. The ability to withstand a financial shock
3. The ability to stay on track in pursuing financial goals
4. The freedom from financial burdens that prevent one from enjoying life

Through developing a better understanding of Americans' financial well-being, including determining who is most at risk for being unable to achieve financial well-being, and identifying the factors that are most useful in predicting financial well-being, we will be in a better position to help all Americans secure a path to a more promising future. Policymakers can use this information to identify which populations could benefit most from further education aimed at improving financial literacy, as well as what skills and behaviors are most important for ensuring financial well-being.

Data

In 2016, the Consumer Financial Protection Bureau fielded a survey, the National Financial Well-Being Survey, which aimed to measure the financial well-being of American citizens. The survey also included a number of questions about behavioral, skills-based, and demographic factors that may be related to an individual's financial well-being. Results from the survey were combined with information about respondents collected prior to the survey, as well as some information from the Census Bureau's American Community Survey to form a cohesive dataset (Consumer Financial Protection Bureau, 2017b). The Consumer Financial Protection Bureau made this dataset available to the public in September 2017, along with a user guide and codebook.

The dataset is comprised of 217 variables, and it has 6,394 observations (Consumer Financial Protection Bureau, 2017c). For the purposes of this analysis, I will be focusing on just 11 variables – the dependent variable and 10 independent variables. (An in-depth discussion of how these variables were chosen using forward stepwise selection can be found in the Methodology section of this report.) After removing all observations with a missing value on any of the 11 chosen variables, we are left with 6,320 observations.

Descriptive statistics and visual summaries are provided for each of the 11 variables below. For the two continuous variables, a table of summary statistics (mean, standard deviation, minimum, and maximum) and a histogram of the variable's distribution are provided. For the nine binary or ordered variables, a table showing the relative frequency of each value, as well as the average "FWBscore" (the dependent variable) associated with each value is provided. The latter metric helps to provide some insight as to how that variable serves to predict an individual's financial well-being.

All variable names and descriptions provided in quotations below are from the dataset's codebook (Consumer Financial Protection Bureau, 2017a).

Dependent Variable

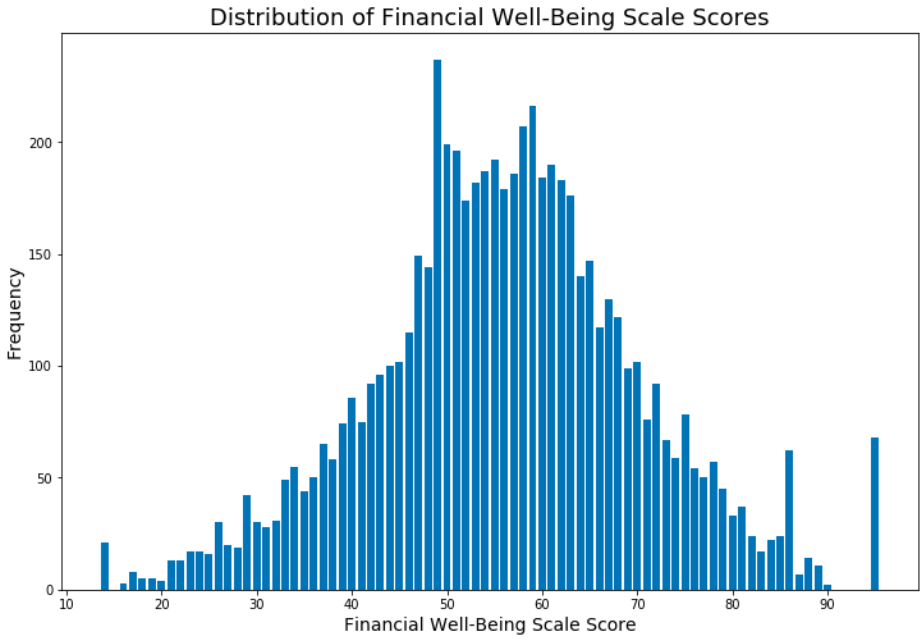
The dependent variable of the analysis is "FWBscore", or the Financial Well-Being Scale Score. This score was constructed using Item Response Theory techniques, based on respondents' level of agreement with the following prompts (Consumer Financial Protection Bureau, 2017a):

- "I could handle a major unexpected expense" (FWB1_1)
- "I am securing my financial future" (FWB1_2)
- "Because of my money situation…I will never have the things I want in life" (FWB1_3)
- "I can enjoy life because of the way I'm managing my money" (FWB1_4)
- "I am just getting by financially" (FWB1_5)
- "I am concerned that the money I have or will save won't last" (FWB1_6)
- "Giving a gift…would put a strain on my finances for the month" (FWB2_1)
- "I have money left over at the end of the month" (FWB2_2)
- "I am behind with my finances" (FWB2_3)

- “My finances control my life” (FWB2_4)

Summary statistics and a histogram of the variable’s distribution are provided here:

Metric	
Variable Name	FWBscore
Short Description	Financial Well-Being Scale Score
Mean	56.12
Standard Deviation	14.07
Minimum	14
Maximum	95



As demonstrated in the histogram above, this variable follows a relatively clean normal distribution.

Independent Variables

- ACT1_2 – “I follow through on my financial goals I set for myself”

ACT1_2	Mapped_Label	Frequency	Percentage	Average_FWBscore
1.0	Not at all	98	1.6%	37.13
2.0	Very little	461	7.3%	42.25
3.0	Somewhat	2200	34.8%	50.17
4.0	Very well	2544	40.3%	60.48
5.0	Completely	1017	16.1%	66.21

- DISTRESS – "Lot of stress in respondent's life"

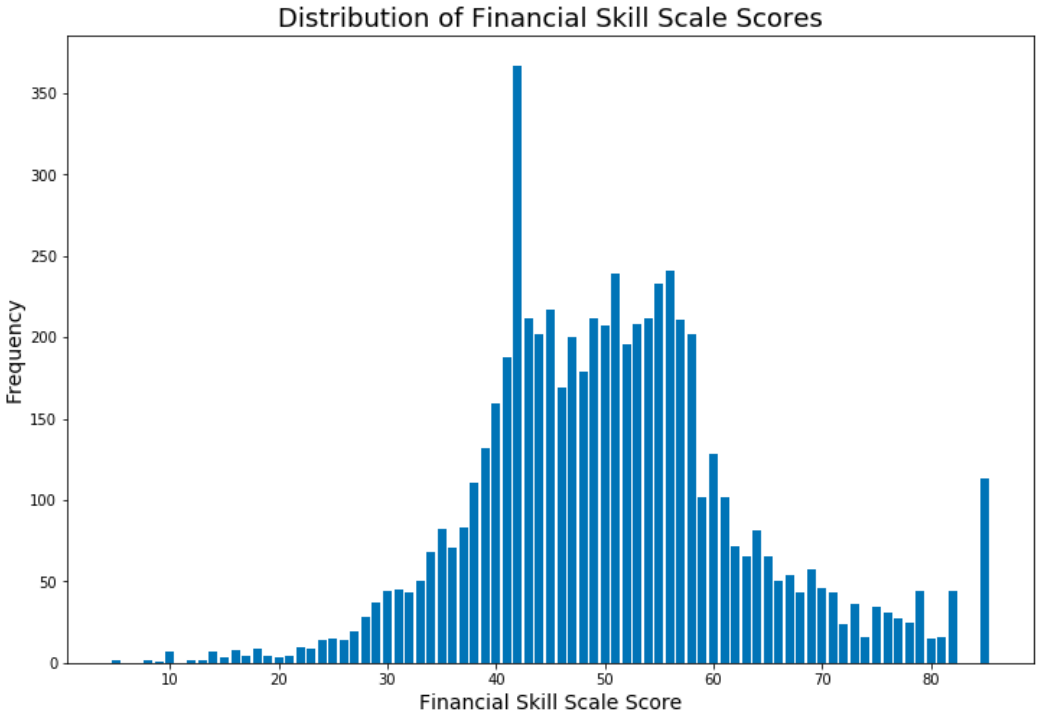
DISTRESS	Mapped_Label	Frequency	Percentage	Average_FWBscore
1.0	Strongly disagree	391	6.2%	66.57
2.0	Disagree	1465	23.2%	63.01
3.0	Neither agree nor disagree	1989	31.5%	56.99
4.0	Agree	1779	28.1%	52.09
5.0	Strongly agree	696	11.0%	43.57

- EMPLOY1_8 – Retired

EMPLOY1_8	Mapped_Label	Frequency	Percentage	Average_FWBscore
0	No	4438	70.2%	53.34
1	Yes	1882	29.8%	62.7

- FSscore – "Financial Skill Scale Score" – a score constructed using Item Response Theory techniques, based on a series of questions where respondents were asked to rate their ability to make financial decisions, control their spending, seek advice, process financial information, and more ("FS1_1" – "FS1_7", "FS2_1" – "FS2_3").

Metric	
Variable Name	FSscore
Short Description	Financial Skill Scale Score
Mean	50.78
Standard Deviation	12.52
Minimum	5
Maximum	85



As demonstrated in the histogram above, this variable follows a relatively normal distribution, with an anomaly around a score of 42, and slightly higher frequencies among the right tail of the distribution than the left tail.

- MANAGE1_3 – "Paid off credit card balance in full each month"

MANAGE1_3	Mapped_Label	Frequency	Percentage	Average_FWBscore
1.0	Not applicable or never	1196	18.9%	45.33
2.0	Seldom	635	10.0%	49.14
3.0	Sometimes	841	13.3%	51.78
4.0	Often	825	13.1%	56.1
5.0	Always	2823	44.7%	63.57

- MATHARDSHIP_1 – "Worried whether food would run out before got money to buy more"

MATHARDSHIP_1	Mapped_Label	Frequency	Percentage	Average_FWBscore
1.0	Never	5199	82.3%	59.42
2.0	Sometimes	834	13.2%	42.88
3.0	Often	287	4.5%	34.88

- MATHARDSHIP_4 – "Any household member couldn't afford to see doctor or go to hospital"

MATHARDSHIP_4	Mapped_Label	Frequency	Percentage	Average_FWBscore
1.0	Never	5301	83.9%	58.76
2.0	Sometimes	791	12.5%	43.94
3.0	Often	228	3.6%	37.15

- PPINCIMP – Household Income

PPINCIMP	Mapped_Label	Frequency	Percentage	Average_FWBscore
1	Less than $20,000	701	11.1%	46.04
2	20,000 to 29,999	501	7.9%	49.65
3	30,000 to 39,999	607	9.6%	51.45
4	40,000 to 49,999	458	7.2%	53.78
5	50,000 to 59,999	501	7.9%	55.74
6	60,000 to 74,999	650	10.3%	56.76
7	75,000 to 99,999	942	14.9%	58.67
8	100,000 to 149,999	1105	17.5%	60.27
9	$150,000 or more	855	13.5%	64.34

- PROPPLAN_1 – “I consult my budget to see how much money I have left”

PROPPLAN_1	Mapped_Label	Frequency	Percentage	Average_FWBscore
1.0	Strongly disagree	200	3.2%	56.62
2.0	Disagree	722	11.4%	56.93
3.0	Neither agree nor disagree	1305	20.6%	55.82
4.0	Agree	2754	43.6%	55.64
5.0	Strongly agree	1339	21.2%	56.9

- SAVEHABIT – “Putting money into savings is a habit for me”

SAVEHABIT	Mapped_Label	Frequency	Percentage	Average_FWBscore
1.0	Strongly disagree	295	4.7%	40.01
2.0	Disagree	607	9.6%	45.68
3.0	Disagree slightly	653	10.3%	48.88
4.0	Agree slightly	1310	20.7%	53.46
5.0	Agree	1717	27.2%	59.24
6.0	Strongly agree	1738	27.5%	64.15

Limitations of Dataset

It's worth briefly discussing a few limitations of the dataset before moving on to the Methodology section. One limitation is in regard to the household income variable (PPINCIMP) that I will be using in the analysis. This variable was coded such that the income ranges associated with each value are not consistent – most are $10,000 ranges (e.g. 4 = $40,000 - $49,999), but one value represents a $50,000 range (8 = $100,000 - $149,999), and the top range is open-ended. Unfortunately, the dataset only comes with the income values coded this way, which will make it more difficult to interpret the regression results directly.

Another limitation of the dataset is that it does not include any information on debt levels. With U.S. household debt reaching an all-time high of $14 trillion in 2019, including about $1.5 trillion in student loan debt, about $1.3 trillion in auto debt, and just shy of $1 trillion in credit card debt, it's very likely that debt levels are weighing significantly on Americans' financial well-being (Tanzi, 2020).

Methodology

The first supervised learning technique that I utilize is ordinary least squares linear regression (OLS). Often referred to as the "workhorse of social science" (Brodnax, 2020c), OLS is the logical choice to start with in tackling a project like this. OLS works by finding the line that minimizes the sum of the squared residuals, where a residual is defined as the distance between an actual observation and its corresponding fitted value (Bailey, 2020). In other words, OLS works by finding the line that best fits the data.

OLS' parametric nature and ease of interpretability make it an excellent candidate for inferential questions (Brodnax, 2020a, 2020c). One interesting application of OLS for inferential purposes can be found in a study from Clark and Arel-Bundock (2013). In this study, the authors utilize OLS to evaluate whether the incumbent president's political party has an effect on how the Federal Reserve manages interest rates in the run-up to an election. The authors find that the Federal Reserve cuts interest rates in the run-up to an election when there is a Republican incumbent and raises interest rates when there is a Democratic incumbent. OLS is useful for inferential purposes in this case, as it can help determine if the incumbent president's party truly does have an effect on the Federal Reserve's management of interest rates, and if so, what the magnitude of the effect is.

In addition to the fact that OLS is highly effective for inferential questions, it can also be useful for prediction purposes (Brodnax, 2020c). An interesting example of the effectiveness of OLS for prediction can be found in a study from Kibekbaev and Duman (2016). In this study, the authors compare the effectiveness of 16 different linear and non-linear models on predicting credit card applicants' incomes. They find that a simple OLS model performs on par with a variety of more complex models in terms of predictive accuracy. This study serves to reinforce the value of OLS for prediction purposes and underscores the idea that OLS can be effectively utilized for both prediction and inferential questions. As such, I will use OLS both as a means of

predicting financial well-being and as a means of understanding the relative importance of different factors in that prediction process.

In order to complement the use of OLS, I also decided to implement a k-nearest neighbors regression algorithm (KNN). As a non-parametric, flexible approach, KNN has the potential to provide better predictive accuracy than a more restrictive model like linear regression (Brodnax 2020a). The k-nearest neighbors algorithm works by identifying the k observations that are most similar to each observation, and then returning the mean of the dependent variable for those k observations as the predicted value for the observation in question (Harrison, 2018). The algorithm does not fit a model to the entire dataset and thus does not have defined parameters, but rather just analyzes each observation individually, identifying the other observations that have the most similar values on the various features in the model (Brodnax, 2020b).

An interesting application of a k-nearest neighbors algorithm can be found in a study from Alizadeh et al. (2018). In this study, the authors utilize KNN to predict monthly water flows based on rainfall and other factors. The authors indicate that having accurate predictions of future water flows can be widely useful for managing droughts, controlling floods, and various other applications. The authors implement a variety of different models and find that KNN is among the most effective at predicting monthly water flows. This study serves to reinforce the idea that KNN can be a powerful technique for predictive purposes.

In comparing the relative merits of OLS and KNN, it can be helpful to think about the trade-offs present with each technique. KNN's flexibility relative to OLS gives it the potential to fit the training data better, resulting in lower bias (Brodnax, 2020a). However, the downside of KNN's flexibility is that it may lead us to overfit the data, whereby the model is fit too closely to the training data and does not perform as well on the test data as a result (Brodnax, 2020a). This means that KNN can suffer from higher variance than OLS. An additional downside of KNN's flexible, non-parametric nature is that it is difficult to interpret relative to OLS, and thus, it is better suited for prediction tasks than inferential purposes (Brodnax, 2020a). As a result, I will use KNN as a supplement to OLS in my task of predicting financial well-being, but rely on OLS to provide a deeper understanding of the relative importance of different factors in that prediction process.

Analysis

Feature Exploration

With 217 variables in the dataset, spanning diverse topics from demographic factors to financial skills to individual values and beliefs, I wanted to start out by exploring the dataset in order to get a better understanding of everything it contains. In order to do so, I decided to classify each variable into broad categories (e.g. financial goals, hardships suffered, financial products used, etc.) based on the descriptions provided in the codebook. I settled upon 10 different categories, with 110 different potential independent variables among them. Many of the variables that were

not included were components of variables that were included (e.g. the "FSscore" variable discussed above has 10 component variables that were not directly included). Other variables were excluded for redundancy or lack of relevance to the research question at hand. Finally, I excluded a few variables that had a high volume of missing observations in order to avoid significantly reducing the size of the total dataset.

The 110 variables that were included, as well as the groups that I classified them into, are shown below. The meanings of any individual variables not discussed in the body of this report can be found by reviewing the codebook linked in the References section of the report.

```
demo_vars = ['PAREDUC', 'EMPLOY1_1', 'EMPLOY1_2', 'EMPLOY1_3', 'EMPLOY1_4', 'EMPLOY1_5', 'EMPLOY1_6',
               'EMPLOY1_7', 'EMPLOY1_8', 'EMPLOY1_9', 'agecat', 'PPEDUC', 'PPETHM_1', 'PPETHM_2', 'PPETHM_3',
               'PPETHM_4', 'PPGENDER_1', 'PPGENDER_2', 'PPHHSIZE', 'PPMARIT_1', 'PPMARIT_2',
               'PPMARIT_3', 'PPMARIT_4', 'PPMARIT_5', 'PPMSACAT', 'PPREG4_1', 'PPREG4_2', 'PPREG4_3', 'PPREG4_4']
skills_vars = ['FSscore', 'SUBKNOWL1', 'LMscore', 'KHscore', 'FINSOC']
goals_vars = ['ACT1_1', 'ACT1_2', 'FINGOALS', 'PROPPLAN_1', 'PROPPLAN_2', 'PROPPLAN_3', 'PROPPLAN_4', 'SCFHORIZON',
              'SELFCONTROL_1', 'SELFCONTROL_2', 'SELFCONTROL_3', ]
habits_vars = ['MANAGE1_1', 'MANAGE1_2', 'MANAGE1_3', 'MANAGE1_4', 'SAVEHABIT', 'FRUGALITY', 'ASK1_1', 'ASK1_2']
products_vars = ['PRODHAVE_1', 'PRODHAVE_2', 'PRODHAVE_3', 'PRODHAVE_4', 'PRODHAVE_5', 'PRODHAVE_6', 'PRODHAVE_7',
                 'PRODHAVE_8', 'PRODHAVE_9', 'PRODUSE_1', 'PRODUSE_2',  'PRODUSE_3',  'PRODUSE_4',  'PRODUSE_5',
                 'PRODUSE_6',  ]
beliefs_vars = ['CHANGEABLE', 'MATERIALISM_1','MATERIALISM_2', 'MATERIALISM_3', 'PEM']
shocks_vars = ['SHOCKS_1', 'SHOCKS_2', 'SHOCKS_3', 'SHOCKS_4', 'SHOCKS_5', 'SHOCKS_6', 'SHOCKS_7', 'SHOCKS_8',
               'SHOCKS_9', 'SHOCKS_10', 'SHOCKS_11', 'SHOCKS_12', ]
hardships_vars = ['MATHARDSHIP_1', 'MATHARDSHIP_2', 'MATHARDSHIP_3', 'MATHARDSHIP_4', 'MATHARDSHIP_5',
                  'MATHARDSHIP_6', 'REJECTED_1', 'REJECTED_2']
benefits_income_vars = ['EARNERS','VOLATILITY','BENEFITS_1', 'BENEFITS_2', 'BENEFITS_3', 'BENEFITS_4',
                        'BENEFITS_5', 'COVERCOSTS_1.0', 'COVERCOSTS_2.0', 'COVERCOSTS_3.0', 'COVERCOSTS_4.0',
                        'BORROW_1', 'BORROW_2','PPINCIMP']
health_vars = ['HEALTH', 'MEMLOSS', 'DISTRESS']
```

After classifying the variables into these groups, I decided to do some preliminary diagnostics to see how the different groups of variables stack up in terms of their ability to predict financial well-being. In order to do so, I wrote a loop that cycles through each of the variable groups, and uses only those variables as the independent variables in a linear regression model. Within the loop, it fits the model, does 5-fold cross-validation, and then reports the average cross-validated R^2 for the model with that group of variables. The results for the 10 groups are shown below:

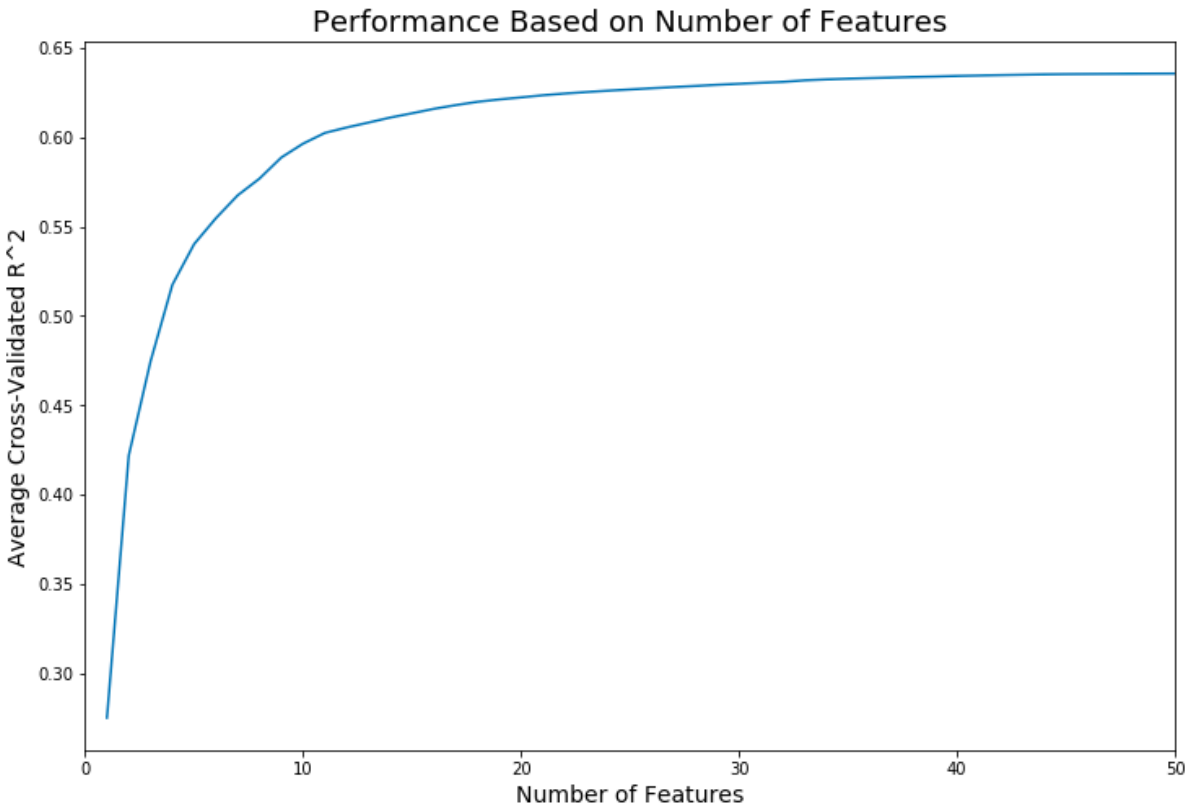
Variable Group	Cross-Validated R^2
Financial Habits	0.39
Financial Goals/Planning	0.38
Hardships Suffered	0.32
Financial Skills/Knowledge	0.31
Financial Products	0.26
Health	0.24
Demographics	0.22
Benefits and Income	0.22
Shocks	0.08
Financial Beliefs	0.07

To be sure, this is a relatively rough diagnostic and certainly should not be construed to imply causal inference. The groupings were subjective at my discretion and comparing between the groupings is complicated by differing number of variables, amongst other factors. Additionally, the relative explanatory power of certain variables or groups of variables may completely change when combined with other groups of variables. Finally, it obscures the role of individual variables, where one or two variables in a group may be doing all of the work while the other variables are providing little value, or a particular variable in a low-performing group may actually be quite powerful.

With all of that being said, it does provide some insight into the relative effectiveness of different conceptual factors in predicting financial well-being. The "Financial Habits" group – which includes things like making a habit of paying bills on time, making a habit of regularly shifting money into savings, etc. – scored the highest of all the groups, followed by the "Financial Goals/Planning" group – which includes things like making and consulting budgets. One thing this exercise does make clear is that financial well-being goes well beyond just an individual or household's income level, though that is certainly an important component.

Feature Selection

After completing the initial diagnostics discussed above, I began the process of selecting the specific variables to use in the model. In order to do so, I utilized the "Sequential Feature Selector" functionality from the "mlxtend" package in Python to conduct forward stepwise selection on the 110 potential independent variables. Behind the scenes, this functionality would continuously add one variable at a time to a linear regression model, always selecting the specific variable that would most improve the fit of the model when included (James et al., 2017). At each step, the package reports which specific variable was added, as well as the average R^2 from 5-fold cross-validation with that variable included. The average cross-validated R^2 performance for each number of variables is shown below.



The graph above demonstrates that adding additional variables to the model significantly improves model performance initially, before leveling off around 10 variables (with R^2 around 0.6). From there, the R^2 only increases very slightly as more variables are added. I completed this process using values of both 50 and 100 as the hyper-parameter for maximum number of variables to be considered, with the graph for the specification allowing up to 50 variables shown above. The graph for the process allowing up to 100 variables is largely similar, with the trajectory remaining essentially level before beginning to drop off very slightly as you reach about 65 variables.

Based on these results, I decided to move forward using only 10 independent variables in the model. I believe this provides the best balance between predictive performance and interpretability. If we solely cared about predictive performance, we could pick the specification where the cross-validated R^2 was maximized (with 65 variables and R^2 of 0.636). However, by going this route, we forfeit a degree of interpretability that is available in a model with fewer variables. Other specifications that would slightly improve performance over the 10 variable model and remain relatively interpretable include 15 variables (R^2 of 0.613) and 20 variables (R^2 of 0.622).

One matter worth discussing here is that there were two variables that I initially included in this forward stepwise selection process, but ultimately decided to remove from the model. Upon initially running the stepwise selection package, these two variables ("ENDSMEET" and "GOALCONF") were identified as the two variables that most improved the model fit out of all 110 possible variables (i.e. they were selected as the variables in the first and second stages of

the stepwise selector, respectively). As such, including them did improve the maximum fit of the overall model in the stepwise selection process (with maximum R_2's in the 0.65+ range, as compared with 0.636 without them). However, upon closer inspection of the meaning of those variables and the component variables of the dependent variable, I felt that they were too similar to the dependent variable to provide legitimate insight – or in other words, I felt that it was "cheating" to include them in the model. For example, the "ENDSMEET" variable ("Difficulty of covering monthly expenses and bills") is fairly similar to some of the component variables of the dependent variable (e.g. "I have money left over at the end of the month" or "I am just getting by financially") (Consumer Financial Protection Bureau, 2017a). As a result, I felt that it was better to exclude these variables from the model altogether.

The table below contains the 20 "best" variables as selected by the forward stepwise selection process, not including the two variables discussed above. They are displayed in the order in which they were chosen by the stepwise selector – so the first 10 listed are the ones I have chosen for the model.

Variable_Order	Variable_Name
1	ACT1_2
2	MATHARDSHIP_1
3	DISTRESS
4	MANAGE1_3
5	PPINCIMP
6	EMPLOY1_8
7	SAVEHABIT
8	FSscore
9	PROPPLAN_1
10	MATHARDSHIP_4
11	SCFHORIZON
12	FRUGALITY
13	MANAGE1_2
14	PRODHAVE_6
15	PROPPLAN_2
16	PEM
17	REJECTED_1
18	MATERIALISM_1
19	ASK1_1
20	agecat

To be sure, this should certainly not be construed to imply causal inference – investigating and rooting out endogeneity is beyond the scope of this analysis. Additionally, it should be noted that since this analysis was conducted with a forward stepwise selector, as opposed to a best subset selector, we can't say for certain that these variables are definitively the best possible variables

for the model (James et al., 2017). However, this process does give a good indication that the variables shown above can be considered important and effective factors in predicting an individual's financial well-being.

Modeling – Linear Regression

As discussed above, I decided to move forward with the 10 "best" independent variables as selected by the forward stepwise selection process. In order to evaluate the effectiveness and insightfulness of a linear regression model with these 10 independent variables, it is helpful to consider three things:

- The average cross-validated R^2 from the model
- How close the predicted values are to the actual values of the dependent variable on average
- The coefficients on the independent variables, interpreted appropriately relative to their scales

After fitting a linear regression model with the 10 independent variables, I used 5-fold cross-validation to obtain the average cross-validated R^2 from the model ($R^2 = 0.598$), which is consistent with the results of the stepwise selection process above.

Next, I used 5-fold cross-validated prediction to obtain predictions of the dependent variable for every observation. After arranging the true values of the dependent variable and their corresponding predicted values into a dataframe, I calculated the residual for each observation. The average of the absolute value of the residuals was 6.85 points, where the variable has a theoretical scale of 0-100. In other words, this means that the linear regression model's predictions were off by 6.85 points, or 13.9%, on average.

Finally, I ran the linear regression model on the full dataset in order to understand the impact of each of the independent variables. The regression results are reported below – the first table contains the results without accounting for survey weights, while the second table contains the results with survey weights accounted for. The survey weights are included as a way of making the sample better reflect demographic factors of the population at large (Consumer Financial Protection Bureau, 2017b).

OLS Regression Results

Dep. Variable:	FWBscore	R-squared:	0.601
Model:	OLS	Adj. R-squared:	0.601
Method:	Least Squares	F-statistic:	951.4
Date:	Mon, 20 Apr 2020	Prob (F-statistic):	0.00
Time:	09:38:45	Log-Likelihood:	-22773.
No. Observations:	6320	AIC:	4.557e+04
Df Residuals:	6309	BIC:	4.564e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
EMPLOY1_8	4.1806	0.265	15.775	0.000	3.661	4.700
PPINCIMP	0.7847	0.047	16.535	0.000	0.692	0.878
FSscore	0.1832	0.013	14.379	0.000	0.158	0.208
ACT1_2	2.3388	0.187	12.514	0.000	1.972	2.705
PROPPLAN_1	-1.5277	0.116	-13.214	0.000	-1.754	-1.301
MANAGE1_3	1.2089	0.088	13.677	0.000	1.036	1.382
SAVEHABIT	1.3303	0.099	13.504	0.000	1.137	1.523
MATHARDSHIP_1	-3.8221	0.284	-13.437	0.000	-4.380	-3.265
MATHARDSHIP_4	-3.3320	0.283	-11.794	0.000	-3.886	-2.778
DISTRESS	-2.4828	0.114	-21.729	0.000	-2.707	-2.259
intercept	44.7570	0.910	49.176	0.000	42.973	46.541

Omnibus:	123.707	Durbin-Watson:	2.036
Prob(Omnibus):	0.000	Jarque-Bera (JB):	262.759
Skew:	0.015	Prob(JB):	8.76e-58
Kurtosis:	3.998	Cond. No.	439.

WLS Regression Results

Dep. Variable:	FWBscore	R-squared:	0.579
Model:	WLS	Adj. R-squared:	0.579
Method:	Least Squares	F-statistic:	868.8
Date:	Mon, 20 Apr 2020	Prob (F-statistic):	0.00
Time:	09:38:45	Log-Likelihood:	-23227.
No. Observations:	6320	AIC:	4.648e+04
Df Residuals:	6309	BIC:	4.655e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
EMPLOY1_8	3.6940	0.295	12.538	0.000	3.116	4.272
PPINCIMP	0.7585	0.046	16.399	0.000	0.668	0.849
FSscore	0.1482	0.013	11.641	0.000	0.123	0.173
ACT1_2	2.1975	0.183	12.030	0.000	1.839	2.556
PROPPLAN_1	-1.3664	0.117	-11.697	0.000	-1.595	-1.137
MANAGE1_3	1.0939	0.086	12.650	0.000	0.924	1.263
SAVEHABIT	1.4356	0.098	14.673	0.000	1.244	1.627
MATHARDSHIP_1	-3.9947	0.263	-15.190	0.000	-4.510	-3.479
MATHARDSHIP_4	-3.1685	0.261	-12.122	0.000	-3.681	-2.656
DISTRESS	-2.5356	0.113	-22.433	0.000	-2.757	-2.314
intercept	46.5816	0.870	53.557	0.000	44.877	48.287

Omnibus:	348.306	Durbin-Watson:	2.023
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1264.536
Skew:	-0.155	Prob(JB):	2.57e-275
Kurtosis:	5.169	Cond. No.	412.

Interpretations are provided below for each variable, with the values from the survey-weighted regression provided in parentheses. All of the coefficients discussed below are statistically significant at all conventional significance levels.

- ACT1_2 – "I follow through on my financial goals I set for myself"
  - Variable Values: 1-5, ranging from "Not at all" to "Completely"
  - The results indicate that a 1 point increase on the ACT1_2 variable is associated with a 2.34 (2.20) point increase on the dependent variable on average, with all else equal.
- DISTRESS – "Lot of stress in respondent's life"
  - Variable Values: 1-5, ranging from "Strongly disagree" to "Strongly agree"
  - The results indicate that a 1 point increase on the DISTRESS variable is associated with a 2.48 (2.54) point decrease on the dependent variable on average, with all else equal.
- EMPLOY1_8 – Retired
  - Variable Values: 0-1, binary No/Yes
  - The results indicate that being retired is associated with a 4.18 (3.69) point increase on the dependent variable on average, with all else equal.
- FSscore – "Financial Skill Scale Score" – a score constructed using Item Response Theory techniques, based on a series of questions where respondents were asked to rate

their ability to make financial decisions, control their spending, seek advice, process financial information, and more ("FS1_1" – "FS1_7", "FS2_1" – "FS2_3").
    - Variable Values: 5-85, continuous scale
    - The results indicate that a 1 point increase on the FSscore variable is associated with a 0.18 (0.15) point increase on the dependent variable on average, with all else equal.
- MANAGE1_3 – "Paid off credit card balance in full each month"
    - Variable Values: 1-5, ranging from "Not applicable / never" to "Always"
    - The results indicate that a 1 point increase on the MANAGE1_3 variable is associated with a 1.21 (1.09) point increase on the dependent variable on average, with all else equal.
- MATHARDSHIP_1 – "Worried whether food would run out before got money to buy more"
    - Variable Values: 1-3, ranging from "Never" to "Often"
    - The results indicate that a 1 point increase on the MATHARDSHIP_1 variable is associated with a 3.82 (3.99) point decrease on the dependent variable on average, with all else equal.
- MATHARDSHIP_4 – "Any household member couldn't afford to see doctor or go to hospital"
    - Variable Values: 1-3, ranging from "Never" to "Often"
    - The results indicate that a 1 point increase on the MATHARDSHIP_4 variable is associated with a 3.33 (3.17) point decrease on the dependent variable on average, with all else equal.
- PPINCIMP – Household Income
    - Variable Values: 1-9, where each value is associated with an income range (e.g. 1 = Less than $20,000, while 9 = More than $150,000). It's worth noting here that the ranges are not consistent – most are $10,000 ranges (e.g. 4 = $40,000 - $49,999), but one value represents a $50,000 range (8 = $100,000 - $149,999), and the top range is open-ended. Unfortunately the dataset only comes with the income values coded this way, which makes it more difficult to interpret the regression results directly.
    - The results indicate that a 1 point increase on the PPINCIMP variable is associated with a 0.78 (0.76) point increase on the dependent variable on average, with all else equal.
- PROPPLAN_1 – "I consult my budget to see how much money I have left"
    - Variable Values: 1-5, ranging from "Strongly disagree" to "Strongly agree"
    - The results indicate that a 1 point increase on the PROPPLAN_1 variable is associated with a 1.53 (1.37) point decrease on the dependent variable on average, with all else equal.
- SAVEHABIT – "Putting money into savings is a habit for me"
    - Variable Values: 1-6, ranging from "Strongly disagree" to "Strongly agree"
    - The results indicate that a 1 point increase on the SAVEHABIT variable is associated with a 1.33 (1.44) point increase on the dependent variable on average, with all else equal.

On account of the fact that the independent variables have different scales, it can also be useful to run the regression model with standardized variables. This will allow us to directly compare the magnitude of their effects. Standardizing the variables means that each variable is transformed such that the variable values indicate how many standard deviations that observation is above or below the variable's mean (Bailey, 2020). The below table includes the coefficients from the regression models that were run with standardized variables – the first column is from a model without the survey weights included, and the second column is from a model with the survey weights included. The table has been sorted by the absolute value of the first column, in order to display the coefficients in order of the magnitude of their coefficients, regardless of whether the variable has a positive or negative effect on the dependent variable.

	Standardized_Coefficient	Standardized_Coefficient_Weighted
DISTRESS	-2.70	-2.75
FSscore	2.29	1.86
PPINCIMP	2.09	2.02
ACT1_2	2.09	1.96
MATHARDSHIP_1	-1.96	-2.05
SAVEHABIT	1.94	2.10
EMPLOY1_8	1.91	1.69
MANAGE1_3	1.90	1.72
MATHARDSHIP_4	-1.60	-1.52
PROPPLAN_1	-1.57	-1.41

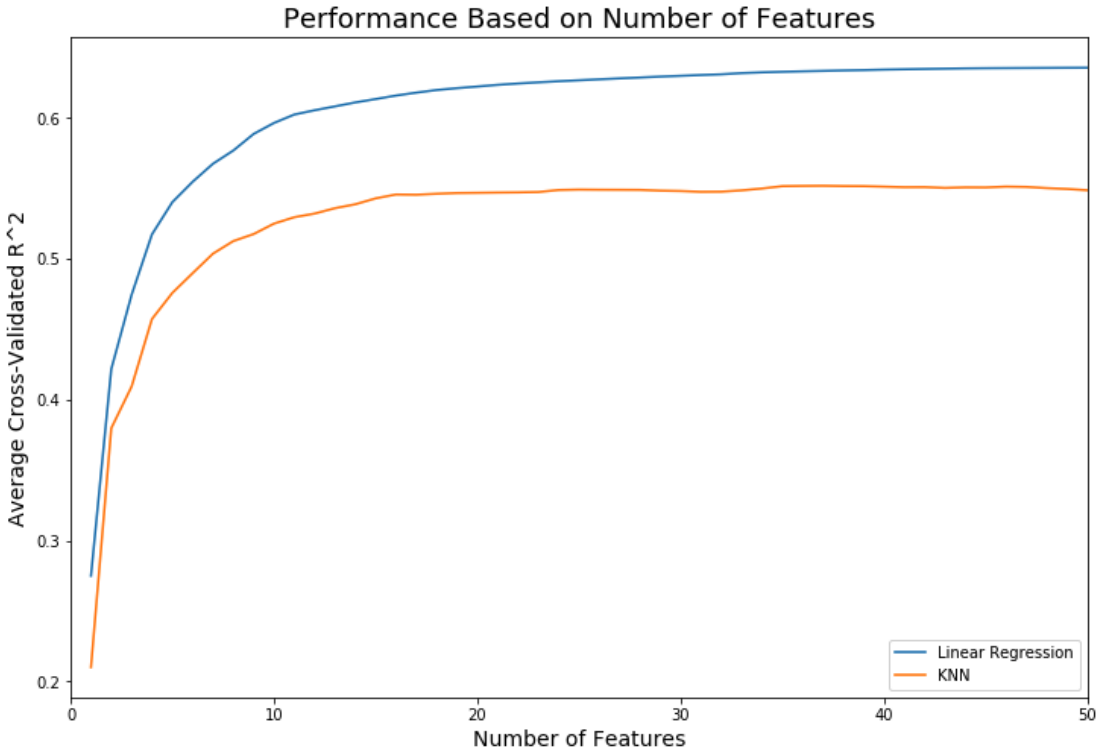
These coefficients indicate the increase/decrease of the dependent variable that is associated with a one standard deviation increase on each of the independent variables, on average, with all else equal. The results indicate that there aren't massive differences among the 10 variables, with effects ranging from about 1.6 – 2.7 points in absolute value. Additionally, the effects from the model with survey weights included seem to differ a fair amount from their non-survey-weighted counterparts within the context of the effect ranges we're seeing here. Finally, it should be reiterated here that these results should not be construed to imply causal effects, but rather just an indication of how big of an effect each variable is having on the predicted value.

With all of that being said, it's still worth discussing a few of the variables that seem to have relatively large effects. The results indicate that experiencing a high degree of general stress in one's life (DISTRESS) and having concerns about food running out before the end of the month (MATHARDSHIP_1) are particularly impactful negative factors on financial well-being predicted values. On the other hand, having a high level of financial skills (FSscore), having higher income (PPINCIMP), following through on financial goals (ACT1_2), and making a habit

of putting money into savings (SAVEHABIT) are among the most impactful positive factors on financial well-being predicted values.

Modeling – K-Nearest Neighbors

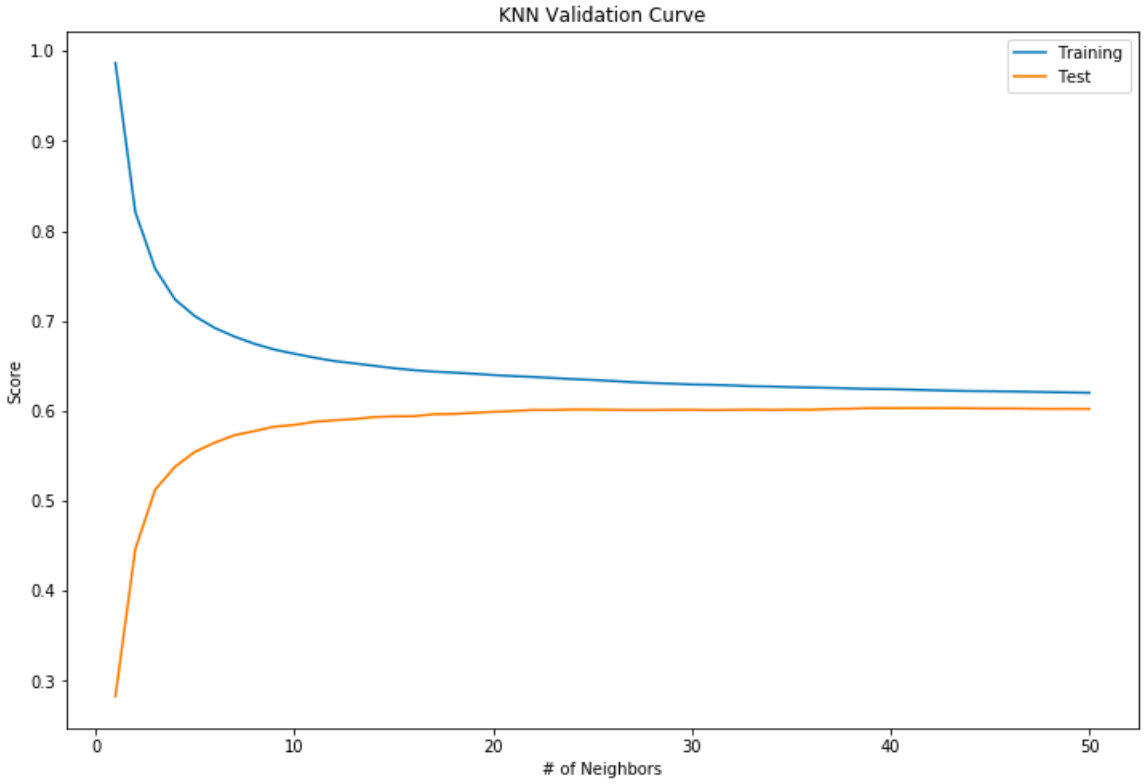
As discussed in the Methodology section, I also decided to implement a k-nearest neighbors regression algorithm in an attempt to improve upon the predictive accuracy of the linear regression models discussed above. Upon initially implementing the KNN model, the results indicated that it was performing significantly worse than the linear regression model. The graph below shows the performance of the forward stepwise selector using KNN as compared to OLS; it indicates that KNN underperformed relative to OLS for every number of features. At the 10 variable mark, KNN has a cross-validated R_2 of 0.525, as compared with 0.597 for OLS.



One likely reason that KNN did not perform well initially is that the distance calculations that KNN relies upon can be distorted by variables with different scales, as is the case with this dataset (Brodnax, 2020b). In order to address this issue, I decided to utilize the standardized versions of the variables as discussed in the linear regression section above. After re-running the model with the standardized variables included, KNN performance increased significantly (more details provided later).

Another important component of the k-nearest neighbors algorithm is the hyper-parameter k that is selected, which reflects the number of neighbors the algorithm uses. In order to ensure that I am working with the optimal value of k, I decided to make use of a validation curve. The

validation curve below reflects the average training and test scores of a k-nearest neighbors algorithm with 5-fold cross-validation for values of k between 1 and 50. The goal will be to select the number of neighbors k that leads to the highest possible test score.



The graph above demonstrates that as the number of neighbors k increases, model performance significantly improves initially, before leveling off around 10 neighbors (with R^2 around 0.58). Upon closer inspection of the performance values for each number of neighbors, we find that the R^2 score is technically maximized with 41 neighbors (with R^2 around 0.6). Since there aren't any noticeable costs to interpretability or speed by using 41 neighbors as opposed to 10, I decided to proceed with a value of 41.

Finally, using the standardized versions of the 10 independent variables discussed throughout this report, and a value of 41 neighbors for k, I implemented a KNN model to predict financial well-being. The average cross-validated R^2 from 5-fold cross-validation was 0.603, which is almost identical to the equivalent cross-validated R^2 from OLS (0.598). I also tried running the model without the one binary independent variable included, but doing so resulted in a lower R^2 than when it was included.

After following the exact same procedure for generating prediction values that I discussed above in the linear regression section, I found that KNN produced predictions with an average residual of 6.78 points (absolute value), or 14.1%. These values are very similar to those produced by OLS and discussed above (6.85 points and 13.9%).

Thus, while KNN did not materially improve upon the predictive accuracy of OLS, by fine-tuning the hyper-parameter k and standardizing the independent variables, KNN was able to produce predictive performance almost identical to that of OLS in the context of this analysis.

Conclusion / Discussion

This report identified 10 factors that, when taken together, are highly effective at predicting an individual's financial well-being. Unsurprisingly, one of these factors is household income, but factors like financial skills, budgeting/saving habits, and general stress levels are shown to be important predictors as well.

Policymakers could use this report to better identify which individuals are unable to achieve financial well-being, going beyond simply relying on measures like household income. Additionally, the report gives a very rough indication of how financial well-being may be affected by changes in various financial skills and habits, though further analysis with a focus on causal inference would need to be conducted to produce definitive conclusions and estimates of such effects.

An important ethical factor for policymakers to consider in conjunction with this report is that not all individuals and communities have equal access to the resources and opportunities needed to improve financial well-being. A significant and persistent racial wage gap results in black and Hispanic men earning 73% and 69%, respectively, of the median hourly earnings of white men (Patten, 2016). The situation is even more dire for women, with white, black, and Hispanic women earning 82%, 65%, and 58%, respectively, of the median hourly earnings of white men (Patten, 2016). These racial and gender wage gaps make it more difficult for women and communities of color to build wealth and achieve financial well-being.

Additionally, there is a significant racial gap in terms of access to banking and financial institutions, with 47% of black households and 43% of Latino households considered unbanked or underbanked (Brown et al., 2019). This means that these communities have less access to savings accounts, lines of credit, and financial planning tools that can help individuals achieve financial well-being.

Of course, there are many more obstacles to achieving financial well-being that have disproportionate impacts along racial, gender, and class lines than the few that have been discussed here. It is imperative that policymakers take a holistic view of the challenges facing different individuals and communities when designing policies to improve financial well-being. With an enhanced understanding of the different factors that are important in predicting financial well-being, combined with a deep understanding of the unique needs of different communities, policymakers will be in a better position to take effective action aimed at improving Americans' financial well-being.

References

Alizadeh, Z., Yazdi, J., Kim, J., & Al-Shamiri, A. (2018). Assessment of Machine Learning Techniques for Monthly Flow Prediction. *Water*, *10*(11), 1676. doi: 10.3390/w10111676

Bailey, M. A. (2020). *Real econometrics: the right tools to answer important questions*. New York: Oxford University Press.

Bartash, J. (2020, February 7). U.S. Adds 225,000 Jobs in January as Hiring Speeds Up Again - Labor Market 'Astounding'. Retrieved from https://www.marketwatch.com/story/us-adds-225000-jobs-in-january-as-hiring-speeds-up-unemployment-rises-to-36-2020-02-07

Board of Governors of the Federal Reserve System. (2019, May). Report on the Economic Well-Being of U.S. Households in 2018 - May 2019. Retrieved from https://www.federalreserve.gov/publications/2019-economic-well-being-of-us-households-in-2018-dealing-with-unexpected-expenses.htm

Brodnax, N. M. (2020a, February 3). PPOL 565: WK 4 – Regression.

Brodnax, N. M. (2020b, February 10). PPOL 565: WK 5 – Classification.

Brodnax, N. M. (2020c, March 23). PPOL 565: WK 10 – Model Selection.

Brown, C., Torres, M., & Loya, R. (2019). *The Future of Banking: Overcoming Barriers to Financial Inclusion for Communities of Color*. UnidosUS. Retrieved from http://publications.unidosus.org/bitstream/handle/123456789/1955/future_of_banking_52419_v3.pdf?sequence=1&isAllowed=y

Clark, W. R., & Arel-Bundock, V. (2013). Independent but Not Indifferent: Partisan Bias in Monetary Policy at the Fed. *Economics & Politics*, *25*(1), 1–26. doi: 10.1111/ecpo.12006

Consumer Financial Protection Bureau. (2017a, September). National Financial Well-Being Survey: Public Use File Codebook. Retrieved from https://files.consumerfinance.gov/f/documents/cfpb_nfwbs-puf-codebook.pdf

Consumer Financial Protection Bureau. (2017b, September). National Financial Well-Being Survey: Public Use File User's Guide. Retrieved from https://files.consumerfinance.gov/f/documents/cfpb_nfwbs-puf-user-guide.pdf

Consumer Financial Protection Bureau. (2017c, September). Financial Well-Being Survey Data. Retrieved from https://www.consumerfinance.gov/data-research/financial-well-being-survey-data/

Cox, J. (2020, January 30). Fourth-Quarter GDP Rose Only 2.1% and Full-Year 2019 Posts Slowest Growth in Three Years at 2.3%. Retrieved from https://www.cnbc.com/2020/01/30/us-gdp-q4-2019-first-reading.html

Harrison, O. (2018, September 10). Machine Learning Basics with the K-Nearest Neighbors Algorithm. Retrieved from https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning: with applications in R*.

Kibekbaev, A., & Duman, E. (2016). Benchmarking regression algorithms for income prediction modeling. *Information Systems*, *61*, 40–52. doi: 10.1016/j.is.2016.05.001

Patten, E. (2016, July 1). Racial, gender wage gaps persist in U.S. despite some progress. Retrieved from https://www.pewresearch.org/fact-tank/2016/07/01/racial-gender-wage-gaps-persist-in-u-s-despite-some-progress/

Tanzi, A. (2020, February 11). U.S. Household Debt Exceeds $14 Trillion for the First Time. Retrieved from https://www.bloomberg.com/news/articles/2020-02-11/u-s-household-debt-exceeds-14-trillion-for-the-first-time