***Project Report***

**Financial Well-being Evaluation**

**Based on Multiple**

**Self-Reported Financial Metrics**

Long Nguyen, Kadiatou Sogodogo, Srujana Bele

 University of South Florida

ISM6136: Data Mining

**Introduction**

In this project, we will explore the topic of financial well-being in the United States using the 2016 National Financial Well-Being Survey prepared by the Consumer Financial Protection Bureau's (CFPB).

6,394 surveys were completed, and upon answering the questions provided, participants were given a score ranging from 0 to 100. The results are used to measure and examine the financial well-being of participants. We will utilize data mining techniques to unravel the metrics that could help determine whether a person is likely to achieve financial well-being in his lifetime.

1. **Problem Statement:**

In the complex word of personal finance, understanding the factors that contribute to financial well-being is crucial. Despite the abundance of resources, many individuals struggle to achieve financial stability. This project aims to identify the key metrics that influence financial well-being in the United States. By applying classification and regression methods to the data from the 2016 National Financial Well-Being Survey, we aim to unravel patterns and correlations that could help determine an individual’s likelihood of achieving financial well-being. These models could potentially serve as valuable tools for individuals and financial advisors in financial planning and decision-making processes.

1. **Data Source**

As previously mentioned, the primary data source for this project is the 2016 National Financial Well-Being Survey conducted by the CFPB. This survey provides a comprehensive snapshot of the financial well-being of U.S. consumers, capturing a wide range of financial behaviors, attitudes, and experiences.

The dataset includes responses from diverse demographic groups, making it a rich resource for understanding the financial well-being of different segments of the population. It includes various aspects such as income, employment, financial experiences, and attitudes towards money, among others.

1. **Preprocessing**

**In the preprocessing phase of our project, we have carefully refined our dataset which set a solid foundation for the following stages of our project. There were certain financial metrics in the survey that consists of multiple survey questions, each one scored on a scale of 1 to 5, with 1 representing the most negative or lowest outcome, and 5 indicating the most positive or highest outcome. We consolidated these categorical variables by calculating the mean value and that provided us with a comprehensive measure for each of those metrics. We then proceeded to select the variables that were most relevant to our study. Some of these variables were renamed to more intuitive and descriptive labels which improves the understanding of our data. We also identified and removed invalid data points. Specifically, responses such as 99, 98, -1, -2, -3, -4, or -5 which indicated that the survey participants either refused to answer or did not have an exact answer to the questions.**

**Our target variable is the “FWBSCORE”, a composite score representing an individual's overall financial well-being. It is calculated based on various financial aspects such as financial security, financial behavior, and financial literacy. A higher FWBscore indicates better financial well-being, encompassing aspects like the ability to meet financial obligations, manage finances effectively, and plan for the future. we introduced a ranking system for financial well-being using the “FWBscore”. This score ranged from 0 to 100, with different ranges representing varying levels of financial well-being: 0 to 25 was categorized as poor, 26 to 50 as fair, 51 to 75 as good, and 76 to 100 as excellent. This ranking system allowed us to effectively categorize and analyze the financial well-being of the survey participants.**

**Below are the variables of interest selected for our project separated into quantitative and qualitative variables:**

**Quantitative Variables**

* **Financial Knowledge “FINKNOWL”**
* **Household income “INCOME”**
* **Benefits**
* **Materialism**
* **Self-control**
* **Material Hardship “MATHARDSHIP”**

**Qualitative Variables**

* **Financial Goals “FINGOALS”**
* **Education**
* **Employment Status “EMPLOY”**
* **Mortgage**
* **Health Status “HEALTH”**
* **Savings**

**This** **represents the state of our primary dataset following the preprocessing stage**

A table of numbers with black text

Description automatically generated

1. **Methodology**
2. **Classification**

To start the classification modeling process, we must first identify target and predictor variables from the dataset. Within our dataset (fin\_df), the column ‘FWBgrade’ serves as the target variable which is a categorical variable showing the financial well-being grade of individuals. As the central focus of our analysis, our goal is to predict or categorize financial well-being based on several factors and the predictor variables are the features or attributes used to predict the target variable (FWBgrade). These are obtained by excluding the FWBgrade and FWBscore columns from the dataset. After this step the dataset is divided into training and testing subsets where we have allocated 70% of the data for training and 30% of the data for testing.

To start the process of modal building, we have employed Cross-validation technique which uses K-Fold method. In machine learning, cross validation is a technique used to decide a model's performance on unseen data. The available data is divided into several folds or subsets; one is used as a validation set, while the other folds are used to train the model. This technique is used to prevent the modal by over fitting. Here we have used 5-fold cross-validation as an effective way to evaluate different categorization methods. With this approach, the dataset is divided into five folds of equal size. Four of the folds are used for training, and the fifth fold is used for validation. The validation set is carefully rotated through each fold. Finally, the results from each validation step are averaged to produce a more robust estimate of the model’s performance.

**1.1. Random Forest Classifier**

This is the most popular supervised machine learning modal used for classification, regression, and other tasks using decision trees. This modal is well suited for handling large and complex datasets which also deals with high-dimensional feature spaces. The Random Forest Classifier will generate decision trees from the randomly selected subset of the training dataset and then it collects the votes from different decision trees and then results in prediction of final decision tress. Upon conducting cross validation for the random forest, the cross-validation score is 0.73.

**1.2. K-Nearest Neighbors (KNN) Classifier**

K-Nearest Neighbors (KNN) Classifier is one of the straightforward yet effective Machine learning modal which classifies data points nearer to its proximity. The K-NN algorithm will work by finding the nearest K neighbors to a given data point based on a distance metric, such as Euclidean distance. We have taken “n\_neighbors=71” and metrics as Euclidean.



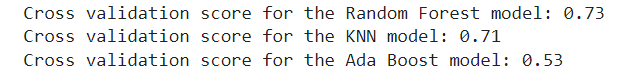
We studyn the model's performance through cross-validation. The calculated cross-validation score for the KNN model is {knn\_score:,.2f} which is 0.71 which is lesser than Random Forest modal.

**1.3. Ada Boost Classifier**

The AdaBoost Classifier is a powerful ensemble learning method for classification. Using decision stumps as weak learners, it creates a robust model by sequentially combining them. A specific random state ensures reproducibility. Its performance is assessed through 5-fold cross-validation, with the mean accuracy denoted as adb\_score. AdaBoost is effective in handling complex datasets by iteratively refining predictions based on misclassified instances, leading to improved accuracy.

**1.4. Cross-Validation scores**

The cross-validation scores for the Random Forest, K-Nearest Neighbors (KNN), and AdaBoost models are 0.73, 0.71, and 0.53, respectively. These scores represent the mean accuracy of each model across multiple folds of cross-validation. The Random Forest model achieved the highest cross-validation score of 0.73, indicating its superior performance in accurately predicting the target variable compared to the other models. The KNN model follows closely behind with a score of 0.71, showcasing its effectiveness in capturing patterns within the data. However, the AdaBoost model yielded the lowest cross-validation score of 0.53, suggesting that it may struggle to generalize well to unseen data or adequately capture the underlying patterns in the dataset. The cross-validation score for the Random Forest model is 0.73 which is the best out of three models.

Fig. 1 Cross validation scores of classification models

**1.5. Hyperparameter tuning**

After training, the model predicts the test data's output, and various metrics such as accuracy, recall, precision, and F1 score are computed and printed for analysis. These metrics gauge the classifier's effectiveness in correctly categorizing instances, providing essential insights into its performance.

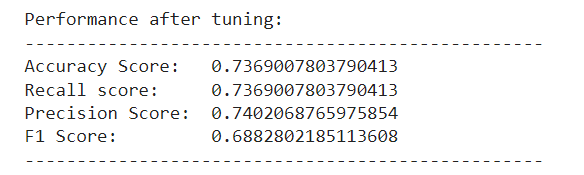
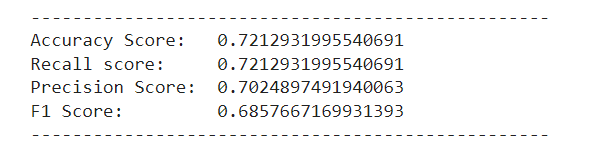


Fig. 2 Performance metrics before and after tuning

Hyperparameter tuning using grid search systematically explores combinations of hyperparameter values to find the best settings for a machine learning model. By evaluating each combination's performance through cross-validation, it helps maximize the model's effectiveness without overfitting or underfitting. In this project, we selected n\_estimators, max\_depth, max\_leaf\_nodes, min\_samples\_split, and min\_samples\_leaf as the main parameters that would affect its performance. Using GridSearchCV from scikit learn package, the grid search returns the optimized parameter values as below.

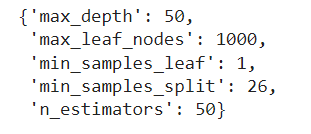


Fig. 3 Best parameters retrieved from GridSearchCV

After tuning, as seen in fig. 2, there is a slight enhancement in performance, with scores showing a modest increase of 1%. Considering working with data about financial well-being, which is more of a social study about people’s perceptions of financial well-being, 73.7% accuracy is an impressive number, indicating that our random forest model is useful in predicting FWBgrade.

1. **Regression**

**2.1. Multiple linear regression model**

Before building the regression model, we started by splitting the data set into training and testing portions, with the testing size equal to 30% of the data set. Following the same steps as in building the classification model, the target variable is “FWBscore”, and the predictors are the remaining features except for “FWBgrade.” Because the regression model accepts continuous dependent variable, we preferred FWBscore over FWBgrade.

To begin, we made use of the LinearRegression function from scikit learn to develop our regression mode. We fitted our training predictor and target variables into the model and generated the beta coefficients of all features.

Afterwards, we used the remaining testing data to evaluate the performance of the model. Mean squared error and R-squared score were calculated using the functions from scikit learn. Below are the results of our regression model:

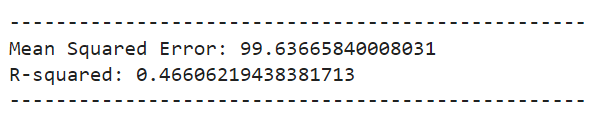
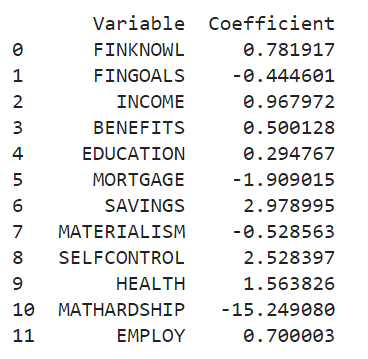


Fig. 4 The coefficients, MSE, and R-squared from the multiple regression model

As a result, many variables show to have a positive relationship with financial well-being, including financial knowledge, income, benefits, education, savings, self-control, health, and employment. Financial goals, mortgage, materialism, and material hardship have a negative correlation with FWBscore. Specifically, a level increase in one’s material hardship would significantly decrease their financial well-being, while their education level has minimal impact.

However, our regression model returns an R-squared of 0.466, showing that the model only explains 46.6% of the variation in FWB score when validating it with the testing data. Our regression model shows a bad fit and our independent variables are not particularly useful in predicting FWB score. For that reason, we continued to try other methods to improve the fit of the regression model.

**2.2. Ordinary Least Squares (OLS) Regression Model**

OLS regression model, implemented in the statsmodel package, is another method to establish a linear relationship between the independent variable and the predictor variables. The main difference between the two methods is that OLS regression is to estimate the coefficients that best fit the observed data. OLS method focuses on minimizing the sum of the squared differences between the observed values of and the values predicted by the model:

Where: are the observed target values

are the predicted values of y based on the model

A screenshot of a computer

Description automatically generated

Fig. 5 OLS regression results

Due to the low p-value of 0.00, we can conclude that there is sufficient evidence to say that some variables are useful in explaining the variation in one’s FWBscore. We can use this regression result to explain some of the relationships between the predictors and FWbscore. Based on the t-test results of the variables, we can see that financial goals, benefits, and materialism are not significant enough in explaining FWBscore.

**2.3. Multicollinearity and Variance Inflation Factor (VIF)**

One important aspect of this data set is that we recognized many of these financial evaluation metrics might be correlated with each other. For example, it can be expected if EDUCATION and FINKNOWL are correlated, as a person with a higher level of education would usually have a better knowledge in financial subjects. The multicollinearity among variables in the data set might cause inaccurate representation of the coefficient values due to inflated standard errors of the magnitude of regression weights. Therefore, to detect multicollinearity, we drew a correlation matrix and used VIF to evaluate our coefficients and removed those that cause the problem.

A chart with red and blue squares

Description automatically generated

Fig. 6 Correlation matrix with FWBscore and all predictor variables

A table of numbers and symbols

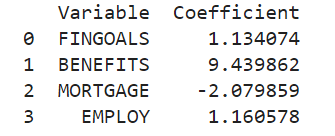
Description automatically generated

Fig. 7 Variance Inflation Factor (VIF) of each variable

According to fig. 6, most variables do not have a strong correlation with each other, pairwise. That includes FWBscore, as only SAVINGS and MATHARSHIP have a decent linear relationship with it. This might explain why the regression model has a bad MSE and poor fit when we ran the model the first time.

However, when we look at the VIF values from fig. 4, most variables have extreme multicollinearity issues, and their coefficients are inflated in the model. Small pairwise correlations but high VIF values might indicate that these variables are correlated with all other variables in the data set instead of just with each individual variable.

To reduce the effect of multicollinearity, we decided to remove variables with VIF above 10, which show extreme, not tolerable correlation of the model predictors. This left us with FINGOALS, BENEFITS, MORTGAGE, and EMPLOY as the dependent variables. Then, we ran another linear regression with these variables, and the results are as below:



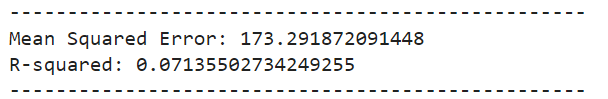
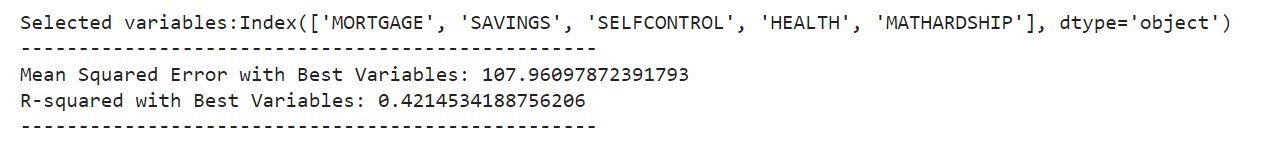


Fig. 8 The coefficients, MSE, and R-squared from the multiple regression model with reduced predictor variables

After rerunning the regression, we could see that these variables now have different coefficients with higher absolute values than before, especially BENEFITS. However, other than a different perspective, this reduced model might not be as practical due to the extremely high MSE and 0.07 R-squared.

**2.4. Recursive Feature Elimination (RFE)**

Our last attempt to improve the performance of our regression model is to use the RFE function from the scikit learn package. RFE is a feature selection method that identifies the most important features in a dataset and removes unwanted ones from a machine learning model. In this last step, we selected the number of features to be 5, thus returning MORTGAGE, SAVINGS, SELFCONTROL, HEALTH, and MATHARDSHIP as the chosen variables. We ran the regression model again, and the results can be found in the figure below. The reduced model has a higher MSE and lower R-squared, which indicates a decrease in fit. However, the model’s complexity improved as the number of predictor variables reduced from 12 to 5, which could make it worth the slight decrease in accuracy.

Fig. 9 Results of reduced regression model after using RFE

1. **Conclusion**

To conclude, our comprehensive analysis of the 2016 National Financial Well-Being Survey presented insightful findings and revealed some interesting patterns pertinent to financial well-being. The dataset revealed that most participants have a Financial Well-Being (FWB) score approximating 59, signifying a degree of stability in their finances. We were able to identify factors that contribute to the financial well-being of U.S. consumers, as well as elements that adversely impact their financial status. Notably, aspects of life such as earning a substantial income, maintaining high savings, and not enduring material hardship constitute significant contributors to financial well-being.

Regarding our data modeling, the application of regression models presents certain challenges, primarily due to the complex and non-linear nature of human behavior. This resulted in low R-squared values, indicating a weak fit and suggesting that the model was unable to adequately explain the variability of the response data. The financial well-being of individuals is influenced by a multitude of factors that may not be fully represented in the dataset or may interact in ways that the regression models may fail to capture. It could have been helpful to perform some variable transformation to address non-linearity. On the other hand, the classification methods used provided great performance with our dataset since they have the capacity to handle non-linear relationships and interactions between variables. Our models segregated the data into distinct classes based on specific criteria which indicated that these approaches may be more suitable for patterns in human behavior.