Asking Consumers about their Finances

Kimberly Kreiss       Mike Zabek

Federal Reserve Board
1 Motivation and Overview

2 Data and Sample

3 Methods

4 References

5 Appendix
The views expressed in this presentation are those of the authors and does not reflect the position of the Federal Reserve Board of Governors or its staff.
Each year the Federal Reserve conducts the Survey of Household Economics and Decisionmaking (SHED)
Each year the Federal Reserve conducts the Survey of Household Economics and Decisionmaking (SHED)

- Nationally representative survey
SHED

- Each year the Federal Reserve conducts the Survey of Household Economics and Decisionmaking (SHED)
  - Nationally representative survey
  - Focuses on the financial lives and experiences of U.S. individuals and households
Each year the Federal Reserve conducts the Survey of Household Economics and Decisionmaking (SHED)

- Nationally representative survey
- Focuses on the financial lives and experiences of U.S. individuals and households

Ask people questions on a range of topics:

- Economic Wellbeing
- Financial Fragility
- Student loans and education
- Income and employment
- Credit and banking experiences
- Housing, neighborhoods, and living situations
- Retirement
Most of our questions are multiple choice questions, with the exception of one open-ended response.
In this year’s survey, we ask respondents to tell us how they are doing financially, and then we ask them to briefly explain why they selected that answer:
In this year’s survey, we ask respondents to tell us how they are doing financially, and then we ask them to briefly explain why they selected that answer:

**General Well-Being Section**
Base: All respondents

[S]
B2. Overall, which one of the following best describes how well you are managing financially these days:

4. Living comfortably
3. Doing okay
2. Just getting by
1. Finding it difficult to get by

Base: B2 ne Refused

[Textbox, 500 characters]

[O]

B2a. In a sentence or two, please describe why you are [IF B2=1 SHOW: living comfortably / IF B2=2 SHOW: doing okay / IF B2=3 SHOW: just getting by / IF B2=4 SHOW: finding it difficult to get by ]?

[Textbox, 500 characters]
The Problem

- We have over 10,000 write-in responses to our question
- It is costly to analyze these data...
The Problem

- We have over 10,000 write-in responses to our question
- It is costly to analyze these data...
- What tools and techniques can we use to systematically extract information from these write-in responses?
The Problem

- We have over 10,000 write-in responses to our question
- It is costly to analyze these data...
- What tools and techniques can we use to systematically extract information from these write-in responses?
- Why is it important?
Why is this important?

- We have a lot of information on consumers’ finances, but less information on how consumers *feel* about their finances.
We have a lot of information on consumers’ finances, but less information on how consumers feel about their finances.

Consumer attitudes and beliefs drive decisionmaking, decisionmaking affects outcomes.
Why is this important?

- We may be able to identify emerging trends or issues that are not captured in multiple choice data
Why is this important?

- We may be able to identify emerging trends or issues that are not captured in multiple choice data
- We may be able to see cyclical patterns in what people are talking about
Why is this important?

- We may be able to identify emerging trends or issues that are not captured in multiple choice data.
- We may be able to see cyclical patterns in what people are talking about.
- We can use responses here to see if the rest of our survey captures these topics or if we should include new modules.

Why is this important?

- We may be able to identify emerging trends or issues that are not captured in multiple choice data.
- We may be able to see cyclical patterns in what people are talking about.
- We can use responses here to see if the rest of our survey captures these topics or if we should include new modules.
Why is this important?

- We may be able to identify emerging trends or issues that are not captured in multiple choice data.
- We may be able to see cyclical patterns in what people are talking about.
- We can use responses here to see if the rest of our survey captures these topics or if we should include new modules.
- We can collect information on what it means to be doing well financially or doing poorly financially in the words of consumers.
Research Question

What circumstances, life events, and attitudes are important for people when asked to describe why they view their financial wellbeing in a positive or negative light?
Research Question

What circumstances, life events, and attitudes are important for people when asked to describe why they view their financial wellbeing in a positive or negative light?
Data and Sample
Data

We have 11,316 respondents in this year’s survey, just about all of them tell us how they are doing financially.
Financial wellbeing distribution

- Finding it difficult to get by
- Just getting by
- Doing okay
- Living comfortably
Who answers?

- We have a very high response rate for our follow up open-ended question—over 90% of people write in an answer
- Sample size is 10,440 write-in responses
Who answers?

- We have a very high response rate for our follow up open-ended question—over 90% of people write in an answer
- Sample size is 10,440 write-in responses
- Median word count is 9, mean word count is 12
In a sentence or two, please describe why you are “struggling to get by”?

- "No work in rural community!"
- "My hospital bills"
- "Bad job prospects, student loans."
- "I am disabled and make $831.00 a month so I have to pick meds & food vs rent & bills"
- "Single mother of two. My job is not giving me enough hours. I've applied to different job openings and haven't heard back."
In a sentence or two, please describe why you are “struggling to get by”?

- “No work in rural community!”
In a sentence or two, please describe why you are “struggling to get by”?

- “No work in rural community!”
- “My hospital bills”
In a sentence or two, please describe why you are “struggling to get by”?

- “No work in rural community!”
- “My hospital bills”
- “Bad job prospects, student loans.”
In a sentence or two, please describe why you are “struggling to get by”?

- “No work in rural community!”
- “My hospital bills”
- “Bad job prospects, student loans.”
- “I am disabled and make $831.00 a month so I have to pick meds & food vs rent & bills”
In a sentence or two, please describe why you are “struggling to get by”?

- “No work in rural community!”
- “My hospital bills”
- “Bad job prospects, student loans.”
- “I am disabled and make $831.00 a month so I have to pick meds & food vs rent & bills”
- “Single mother of two. My job is not giving me enough hours. I’ve applied to different job openings and haven’t heard back.”
In a sentence or two, please describe why you are “living comfortably”?

- "Good job, no debt but the mortgage, and the savings are growing."
- "Two experienced full time wage earners - two good incomes, low credit card debt."
- "We have retirement savings and do not worry about our finances. House is paid for and pay off credit cards in full. Travel when we want to without touching any of our savings."
- "I retired with $1.8 million in investments and a fully paid-for home."
In a sentence or two, please describe why you are "living comfortably"?

- "Good job, no debt but the mortgage, and the savings are growing."
In a sentence or two, please describe why you are “living comfortably”?

- “Good job, no debt but the mortgage, and the savings are growing.”
- “Two experienced full time wage earners”
In a sentence or two, please describe why you are “living comfortably”?

- “Good job, no debt but the mortgage, and the savings are growing.”
- “Two experienced full time wage earners”
- “two good incomes, low credit card debt”
In a sentence or two, please describe why you are “living comfortably”?

- “Good job, no debt but the mortgage, and the savings are growing.”
- “Two experienced full time wage earners”
- “two good incomes, low credit card debt”
- “We have retirement savings and do not worry about our finances. House is paid for and pay off credit cards in full. Travel when we want to without touching any of our savings.”
In a sentence or two, please describe why you are “living comfortably”?

- “Good job, no debt but the mortgage, and the savings are growing.”
- “Two experienced full time wage earners”
- “two good incomes, low credit card debt”
- “We have retirement savings and do not worry about our finances. House is paid for and pay off credit cards in full. Travel when we want to without touching any of our savings.”
- “I retired with $1.8 million in investments and a fully paid-for home.”
Wide range of topics come up across all categories, but how they are talked about varies substantially:
Wide range of topics come up across all categories, but how they are talked about varies substantially:

- Income
- Debt
- Health
- Work
Methods
Applications

- We have a lot of responses, analyzing by hand is time-consuming
Applications

- We have a lot of responses, analyzing by hand is time-consuming
- What tools can we use to efficiently analyze these responses?
Tools

- Text mining
- Simple machine learning and regression models
- Natural language processing
Tools

- Text mining
- **Simple machine learning and regression models**
- Natural language processing
Extensions

- We have a unique set up with our data: responses that are labelled with an outcome variable
We have a unique set up with our data: responses that are labelled with an outcome variable.

<table>
<thead>
<tr>
<th></th>
<th>B2a</th>
<th>not_okay</th>
</tr>
</thead>
<tbody>
<tr>
<td>After I pay all of my bills, I still have mone...</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Bills get paid</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>All money going to medical and pills bills</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>I have enough to pay my bills plus put some in...</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>After being laid off twice and then going on</td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>
Extensions

- We have a unique set up with our data: responses that are labelled with an outcome variable

  After I pay all of my bills, I still have mone...
  Bills get paid
  All money going to medical and pills bills
  I have enough to pay my bills plus put some in...
  After being laid off twice and then going on ...

  B2a  not_okay

  0
  0
  1
  0
  0

- This allows us to use some simple supervised machine learning algorithms and regression techniques

- Text data in responses are features, outcome is a binary “doing okay” or “not doing okay”
We aren’t really that interested in prediction, why use machine learning?
We aren’t really that interested in prediction, why use machine learning?

Certain machine learning algorithms can give information about feature importance
Machine Learning & Regression techniques

- Random forest
- Logistic regression with l1 penalty (Lasso)
Random forest

- Ensemble method: aggregated predictions of many different algorithms to increase accuracy
- Random forest is an ensemble method of decision trees
Decision Trees

- Decision trees are used for both classification and regression.
- Flowchart with similar structure to a tree, where each node denotes a decision.
Decision Trees

- Decision trees are used for both classification and regression.
- Flowchart with similar structure to a tree, where each node denotes a decision.

```
Is it round?  
  Yes     No
  
Is it red?  
  Yes     No
  Apple

Banana

Orange
```
Decision Trees: how a tree decides where to split

- Can choose from among several different decision criteria for deciding when to split
Decision Trees: how a tree decides where to split

- Can choose from among several different decision criteria for deciding when to split
- Decision trees split on the sub-node that makes the sample the most homogenous
Decision Trees: how a tree decides where to split

- Can choose from among several different decision criteria for deciding when to split
- Decision trees split on the sub-node that makes the sample the most homogenous
- We use gini impurity
Decision Trees: how a tree decide where to split

- You have 4 apples and 4 oranges & want to build a model that will predict whether the fruit is an apple or orange
Decision Trees: how a tree decide where to split

- You have 4 apples and 4 oranges & want to build a model that will predict whether the fruit is an apple or orange

You have 4 apples and 4 oranges & want to build a model that will predict whether the fruit is an apple or orange.
Decision Trees

- Can be unstable
Decision Trees

- Can be instable
- Can lead to overfitting
The idea behind a random forest is to average multiple decision trees to build a more robust model that is less susceptible to overfitting.
The idea behind a random forest is to average multiple decision trees to build a more robust model that is less susceptible to overfitting.

Test Sample Input

```
Tree 1

Prediction 1

Tree 2

Prediction 2

( . . . )

Tree 600

Prediction 600

Average All Predictions

Random Forest Prediction
```
Random Forest

1. “Draw a random bootstrap sample of size n
2. Grow a decision tree from the bootstrap sample. At each node:
   1. Randomly select $d$ features without replacement
   2. Split the node using the feature that provides the best split according to the objective function, i.e., maximizing information gain
3. Repeat the steps $1-2k$ times.” (Mirjalili, Raschka, 2017)
Random Forest: feature importance

- For each tree, gives a weighted calculation for how much each split decreases the impurity
- Averaged across all the trees
- Features with this highest value are the “most important”
Logistic Regression

- Used for prediction when the outcome variable is binary
- Used to estimate the probability that a particular outcome will occur
l1 penalty (Lasso)

- Imposes a penalty that will lead to zero coefficients for some variables
- Generally features not shrunk towards zero can be interpreted as the more important features
Interpretation of coefficients

- Can be generally hard to interpret
- “Expected change in the log odds of not doing okay”
- Think about in terms of signs and magnitude in relation to each other:
  - Positive coefficient with higher magnitude means that the presence of that phrase is associated with a higher probability of not doing okay
Preprocessing of Data

```python
from nltk.stem import PorterStemmer

#convert to lower case
B2a['B2a'] = B2a['B2a'].apply(lambda x: " ".join(x.lower() for x in x.split()))

# Removing punctuation
# finds any character that is not a word or white space and replaces with ''
B2a['B2a'] = B2a['B2a'].str.replace('[^\w\s]', '')

# Stem words
st = PorterStemmer()
B2a['B2a'] = B2a['B2a'].apply(lambda x: " ".join([st.stem(word) for word in x.split()]))
```
Preprocessing of data

“Living on Social Security and a small retirement which has not kept up with the cost of living.”

“live on social secur and a small retir which ha not kept up with the cost of live”
Using the model

```python
# split the data into test and training
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(B2a['B2a'], B2a['not_okay'], 
    test_size=.3, random_state=200)

# Initialize a random forest classifier
from sklearn.ensemble import RandomForestClassifier
forest = RandomForestClassifier(n_estimators = 900, max_depth = 100)
```
# Now need to transform the input data into something the model can handle.

```
vectorizer = CountVectorizer(analyzer = "word",
    ngram_range = (2,2),
    tokenizer = None,
    preprocessor = None,
    stop_words = None,
    max_features = 500)
```
Using the model

```python
# transform the training features
vector = vectorizer.fit_transform(x_train)

# make into array for forest.fit()
train_data_features = vector.toarray()

# transform the testing features
vector1 = vectorizer.transform(x_test)

# change test features to numeric input
test_data_features = vector1.toarray()
```
Using the model

```python
# transform the training features
vector = vectorizer.fit_transform(x_train)

# make into array for forest.fit()
train_data_features = vector.toarray()

# transform the testing features
vector1 = vectorizer.transform(x_test)

# change test features to numeric input
test_data_features = vector1.toarray()

# train a random forest
rf = forest.fit(train_data_features, y_train)  # train the model
y_pred_rf = rf.predict(test_data_features)  # use the model on the testing data set
y_pred_score_rf = rf.predict_proba(test_data_features)
```
Accuracy Metrics

```python
print("Classification Report: ")
print(classification_report(y_test,y_pred))
print("\n")
print("Accuracy : ", accuracy_score(y_test, y_pred) * 100)
print("\n")
print("ROC_AUC : ", roc_auc_score(y_test,y_pred_score[:,1]) * 100)
```
Accuracy Metrics

```python
print("Classification Report: ")
print(classification_report(y_test, y_pred))
print("\n")
print("Accuracy : ", accuracy_score(y_test, y_pred) * 100)
print("\n")
print("ROC_AUC : ", roc_auc_score(y_test, y_pred_score[:,1]) * 100)
```

<table>
<thead>
<tr>
<th></th>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.81</td>
<td>0.94</td>
<td>0.87</td>
<td>2257</td>
</tr>
<tr>
<td>1</td>
<td>0.73</td>
<td>0.42</td>
<td>0.53</td>
<td>875</td>
</tr>
<tr>
<td>micro avg</td>
<td>0.79</td>
<td>0.79</td>
<td>0.79</td>
<td>3132</td>
</tr>
<tr>
<td>macro avg</td>
<td>0.77</td>
<td>0.68</td>
<td>0.70</td>
<td>3132</td>
</tr>
<tr>
<td>weighted avg</td>
<td>0.79</td>
<td>0.79</td>
<td>0.77</td>
<td>3132</td>
</tr>
</tbody>
</table>

Accuracy : 79.46998722860792

ROC_AUC : 84.63834419899993
Visualization

```python
## plot feature importances
# get feature importances
importances = forest.feature_importances_

# convert the importances into one-dimensional 1darray
# with corresponding column names as axis labels
f_importances = pd.Series(importances, vectorizer.get_feature_names() )

# sort the array in descending order of the importances
f_importances.sort_values(ascending=False, inplace=True)

# make the bar Plot from f_importances
f_importances.plot( x='Features', y='Importance', kind='bar', figsize=(12, 9), rot=90, fontsize=)

# show the plot
plt.tight_layout()
plt.show()
```
Visualization

Figure 1: Feature Importance
# Initialize a logistic classifier
from sklearn.linear_model import LogisticRegression
logistic = LogisticRegression(penalty='11')

# train a logistic regression
lasso = logistic.fit(train_data_features, y_train)
y_pred_lasso = lasso.predict(test_data_features)
y_pred_score_lasso = lasso.predict_proba(test_data_features)

# Calculate accuracy metrics
accuracy_metrics(y_test, y_pred_lasso, y_pred_score_lasso)
Lasso Application

```python
# Initialize a logistic classifier
from sklearn.linear_model import LogisticRegression
logistic = LogisticRegression(penalty='l1')

# train a logistic regression
lasso = logistic.fit(train_data_features, y_train)
y_pred_lasso = lasso.predict(test_data_features)
y_pred_score_lasso = lasso.predict_proba(test_data_features)

# Calculate accuracy metrics
accuracy_metrics(y_test, y_pred_lasso, y_pred_score_lasso)
```

Classification Report:

<table>
<thead>
<tr>
<th></th>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.82</td>
<td>0.93</td>
<td>0.87</td>
<td>2257</td>
</tr>
<tr>
<td>1</td>
<td>0.72</td>
<td>0.46</td>
<td>0.56</td>
<td>875</td>
</tr>
<tr>
<td>micro avg</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>3132</td>
</tr>
<tr>
<td>macro avg</td>
<td>0.77</td>
<td>0.70</td>
<td>0.72</td>
<td>3132</td>
</tr>
<tr>
<td>weighted avg</td>
<td>0.79</td>
<td>0.80</td>
<td>0.78</td>
<td>3132</td>
</tr>
</tbody>
</table>

Accuracy : 79.94891443167306

ROC_AUC : 86.09980378504969
Visualization
### Comparison of Results

<table>
<thead>
<tr>
<th>Word/phrase</th>
<th>Relative Feature Importance</th>
<th>Word/phrase</th>
<th>LASSO Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>not enough</td>
<td>.031</td>
<td>no debt</td>
<td>-3.54</td>
</tr>
<tr>
<td>cost of</td>
<td>.024</td>
<td>not live</td>
<td>-3.47</td>
</tr>
<tr>
<td>abl to</td>
<td>.022</td>
<td>can pay</td>
<td>-3.40</td>
</tr>
<tr>
<td>go up</td>
<td>.021</td>
<td>do ok</td>
<td>-2.78</td>
</tr>
<tr>
<td>we have</td>
<td>.021</td>
<td>have good</td>
<td>-2.73</td>
</tr>
<tr>
<td>fix incom</td>
<td>.020</td>
<td>have an</td>
<td>-2.68</td>
</tr>
<tr>
<td>dont make</td>
<td>.017</td>
<td>do okay</td>
<td>-2.59</td>
</tr>
<tr>
<td>of live</td>
<td>.015</td>
<td>roof over</td>
<td>-2.59</td>
</tr>
<tr>
<td>disabl and</td>
<td>.014</td>
<td>plenti of</td>
<td>-2.45</td>
</tr>
<tr>
<td>low pay</td>
<td>.013</td>
<td>good health</td>
<td>-2.43</td>
</tr>
<tr>
<td>on social</td>
<td>.012</td>
<td>not enough</td>
<td>2.06</td>
</tr>
<tr>
<td>have enough</td>
<td>.011</td>
<td>not make</td>
<td>2.11</td>
</tr>
<tr>
<td>low incom</td>
<td>.011</td>
<td>fix incom</td>
<td>2.23</td>
</tr>
<tr>
<td>we are</td>
<td>.011</td>
<td>go up</td>
<td>2.32</td>
</tr>
<tr>
<td>too mani</td>
<td>.011</td>
<td>do not</td>
<td>2.41</td>
</tr>
<tr>
<td>live on</td>
<td>.011</td>
<td>to find</td>
<td>2.67</td>
</tr>
<tr>
<td>can pay</td>
<td>.009</td>
<td>live is</td>
<td>2.75</td>
</tr>
<tr>
<td>to work</td>
<td>.009</td>
<td>disabl and</td>
<td>3.14</td>
</tr>
<tr>
<td>and have</td>
<td>.008</td>
<td>dont make</td>
<td>3.21</td>
</tr>
<tr>
<td>to find</td>
<td>.008</td>
<td>low incom</td>
<td>3.44</td>
</tr>
<tr>
<td>on fix</td>
<td>.008</td>
<td>low pay</td>
<td>3.66</td>
</tr>
</tbody>
</table>
Results: disability

- Mentioning disability is extremely predictive of poor financial wellbeing (although there is a low incidence overall)
  - Being on disability and not having enough to live on
  - Work and disability
Results: health

- There seems to be a relationship between health and financial wellbeing
  - People talk about not being able to pay medical bills
  - On the other hand, many mention “good health” as an important reason for their financial wellbeing (can refer to both good health or good health insurance)
Results: debt

- People value not having any debt
- People seem to have aversions to certain types of debt more than others: people mention high credit card debt for as an important factor for poor financial wellbeing or not using credit cards as a sign of good financial wellbeing
Issues and next step

- Does not currently incorporate weights—not nationally representative
- How frequent are each of these across responses? Across categories?
  - Likely picking up words that are very common in one category and not the other (there is still value in that)
- Other models or techniques?
- Deeper dives on these topics
Conclusions

- One of many ways to analyze text data
- Most techniques won’t completely automate analysis for us, but can give us some information about how to think about our responses.
- Our results suggest some factors that are important for determining financial wellbeing:
  - disability/health
  - debt
  - income
  - work
Questions / comments

Email:  kimberly.kreiss@frb.gov

Code:  https://github.com/kimberlykreiss/text_analytics_and_nlp

Slides:  https://kimberlykreiss.github.io/projects_and_code/GASP_slides.pdf
References
References

- Random forest diagram: https://towardsdatascience.com/random-forest-and-its-implementation-71824ced454f
Appendix
Text mining

- Provides some simple tools that are useful for exploring and understanding data
- Can only paint a simple picture
Text mining

- Word frequency
- Bigram network analysis
Word Frequencies in Write-in Responses

- bills
- money
- pay
- enough
- income
- able
- debt
- good
- living
- paid
- can
- job
- time
- live
- savings
- retirement
- expenses
- need
- save
- work
- little
- month
- food
- make
- paying
- left
- get
- home
Word Frequencies by Group

- Asking Consumers about their Finances
  - Kimberly Kreiss, Mike Zabek
- Motivation and Overview
- Data and Sample
- Methods
- References
- Appendix
Text mining

- Helpful for exploring data, but can be somewhat limited in terms of what we learn
- Can get creative with visuals
- Can show different visuals by group
- Can expand to include dictionary-based sentiment analysis
Text mining

- Used many techniques from Julia Silge’s book *Tidy Text Mining with R* found on https://tidytext.com
- Code: https://github.com/kimberlykreiss/text_analytics_and_nlp
Asking Consumers about their Finances

Kimberly Kreiss, Mike Zabek

Motivation and Overview

Data and Sample

Methods

References

Appendix

Natural Language Processing

- Topic modeling (Latent Dirichlet Allocation)
Unsupervised machine learning technique to identify topics in documents and words in each topic
Topic models

• Don’t work all that well here:
  • Responses tend to be pretty short, makes it hard for the model to work
  • Code/visuals from this soon to be on github page