**Shock Prophets**

Final Report

By Thomas Seay, Qinya Pang, Lucas Melo, & Ming Li

**Abstract**

In the academic literature, any significant change in an individual's income, whether positive or negative, is known as an "income shock." But not all shocks are created equal: While a positive income shock may be cause for celebration, a negative income shock could lead to painful reductions in living standards or even the loss of a car or home.

To better prepare individuals to deal with negative shocks, this project aims to predict a person's risk of suffering such a shock (which we define as a reduction in real personal earnings of at least 20%) in the next two years. We use data from the two major cohorts of the National Longitudinal Survey of Youth to uncover factors correlated with income shock risk, and we explore the effectiveness of machine learning algorithms in predicting that risk.

Ultimately, we find that a machine-learning model can accurately characterize a person's likelihood of suffering a negative income shock, suggesting that such models could help individuals understand and potentially improve their financial health.

**Hypothesis**

We hypothesize that, using data related to an individual's financial situation (such as their age, income, and education level) and broader economic conditions (such as unemployment rate and inflation rate), we can accurately classify a person's risk of suffering a negative income shock over the next two years.

**Applications or motivation for a good result**

We believe that effective predictions of income shock risk could allow individuals to improve their financial stability and prepare for possible income loss.

For instance, an individual who learns that he or she is at a high risk for a negative income shock might:

* Increase their "rainy day" or emergency fund to cover a potential spell of unemployment;
* Update their resume or begin searching for new jobs;
* Seek to reduce their risk of income loss, perhaps by pursuing further education or training or by moving to a region with more promising job opportunities;
* Rebalance their investment portfolio away from risky assets such as stocks and toward safer assets, such as bonds or cash; and/or
* Reconsider making long-term financial commitments, such as signing a lengthy automobile lease or taking out a new mortgage.

On the other hand, those who find their income shock risk is relatively low might feel greater confidence in taking on long-term financial commitments or investing in risky assets such as stocks.

(We emphasize, however, that because of the limitations of our model and the fact that even relatively low-risk individuals may still suffer income loss, our calculator should be viewed only as a general educational tool. We are not offering financial advice to any individual, and any financial decisions should be made in consultation with appropriate experts.)

One note on terminology: Because most individuals would welcome a positive income shock but fear a negative shock, we focused our analysis exclusively on negative income shocks. As such, the term "income shock" in this paper refers only to negative income shocks, except where otherwise specified.

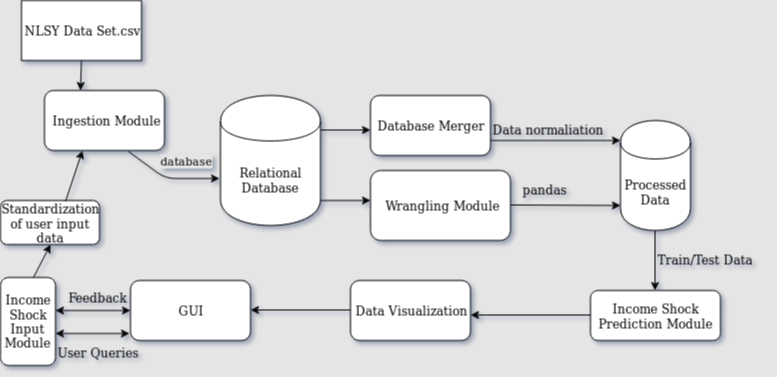
**Steps taken to perform the analysis**

In preparation for our analysis, we identified variables that could potentially predict income shock risk through both literature review and a careful examination of our data.

We began by identifying a formal definition of an "income shock," which we borrowed from Krueger and Perri (2011). In their study, Krueger and Perri used the Italian Survey of Household Income and Wealth from 1987 to 2008 to identify income shocks, which they defined as a one-year reduction in income of at least 20% or €2,000. We opted for a slightly simpler version of this definition, characterizing an income shock as a loss in inflation-adjusted income of at least 20%. Further, due to limitations in our data, we decided to evaluate income shocks over a two-year, rather than a one-year, time period.

Next, we reviewed further literature to identify factors related to income and income shock risk. This helped especially in identifying the link between income change and macroeconomic variables, such as unemployment rate and inflation rate.

The flowchart below, which was created using Draw.io, illustrates the steps we then took to ingest and wrangle data, explore it using statistical tools, and model income shock risk.

****

**Data sources, ingestion, and wrangling**

Our primary data source was the [National Longitudinal Survey of Youth](https://www.nlsinfo.org/content/cohorts/nlsy79) (NLSY), a survey conducted by the U.S. Bureau of Labor Statistics that has followed two cohorts of young Americans for up to 40 years—repeatedly asking about characteristics such as their income, employment, demographics, and more.

The first cohort began in 1979 and included about 13,000 people who were then between ages 14 and 22. (A subset of participants later were dropped from the survey, leaving approximately 10,000 respondents in later years.) These respondents have been surveyed 27 times, including in annual surveys from 1979 to 1994 and then biannual surveys from 1996 to 2016. The second cohort began in 1997 and included approximately 9,000 respondents who were then between ages 12 and 17. These individuals have been surveyed 17 times to date.

To complement these data, we also incorporated data on broader economic conditions, such as unemployment rate, inflation rate, and GDP growth rate.

In the first phase of data ingestion, we imported the initial data into an SQLite database. We then converted the data from wide format to long format, which better reflected the longitudinal nature of the data (that is, reflecting that the same questions were asked repeatedly of the same individuals over many years).

Next, we normalized responses to adjust dollar figures for inflation and to account for changes in how survey questions were asked or coded over time. For example, the NLSY in various years recorded occupation and industry data using codes defined in the 1970, 1980, 1990, and 2000 Censuses. All of these responses had to be translated into standardized codes to allow comparison across time and cohorts—a challenging and time-consuming process. A further complication was that 1979 and 1997 cohorts used different coding for many questions, so we had to do significant work to translate the two cohorts' data into the same language.

We then identified instances of income shocks by comparing individuals' real (that is, inflation-adjusted) income in one year to their income two years later. The processed data were then exported into a format suitable for data exploration.

**Exploratory data analysis**

We conducted exploratory data analysis to (1) understand the variables' type and measurement scale, (2) examine missing values to determine the best method for handling missingness, (3) perform univariate analysis to derive early insights, (4) summarize the main features of the dataset using tables and other visualizations, and (5) perform bivariate association analysis to identify variables related to income shock.

Missingness in our data resulted from survey respondents' failing to participate, refusing to respond, not knowing the answers to a particular question, or because the interviewer skipped a question for either valid or invalid reasons.

We found that missing percentages across the variables ranged from approximately 1% to 60% (with the high end of that range primarily due to increasing numbers of respondents dropping out of the survey over time). We then identified appropriate methods for imputing missing values.

Given the longitudinal nature of our data, we determined that for many variables the best approach was simply to assume that any missing values had not changed since the respondent's previous interview. For instance, if someone indicated in a particular year that their highest grade completed was 16 (representing a college undergraduate degree), we "carried forward" that value to future years—until or unless that respondent reported a different value.

Some variables, however, could not reasonably be imputed using this approach, such as hours worked in the last calendar year. For these variables, we typically imputed the median value of all reported instances.

We also conducted univariate analyses. Specifically, for continuous variables, we examined the mean, median, standard deviation, skewness, and kurtosis. We also identified and dealt with outliers not only numerically but graphically. For example, we found a few reported numbers of hours worked that dramatically exceeded other reports. To deal with these extreme outliers, we capped the number of hours worked at the mean plus three standard deviations of all reported instances.

Then, we binned certain discrete variables into categories. For instance, "highest grade completed" is a discrete variable representing years of schooling, ranging from 0 (no education) to 20. Although the variable can take on any value in that range, certain values have distinctive semantic meaning (for instance, 12 ordinarily means that an individual has graduated from high school), and we anticipated that these "cut-off" values might represent inflection points that would not be well captured by a linear function. As such, we grouped the year-by-year levels into the new categories of "less than elementary school" (0-4), "elementary school" (5-7), "middle school" (8-11), "high school" (12), "some college" (13-15), "four-year college degree" (16), and "graduate school" (17-20).

We also collapsed categorical variables where appropriate, especially to deal with categories that contained very few samples. For example, rather than separately modeling all of the individual industry codes in our data—many of which appeared only rarely—we instead grouped those responses into the bins utilized in the 1990 Census data (for example, grouping together all codes representing "AGRICULTURE, FORESTRY, AND FISHERIES"). We similarly grouped occupation data into the bins utilized in the 2010 Census data (such as "MANAGEMENT, BUSINESS, SCIENCE, AND ARTS").

We then examined bivariate relationship between pairs of variables, using both graphics and statistical tests such as Chi-square—not only to identify potential predictors but to foresee possible methodological issues in our modeling. In some cases, we found that two variables were highly correlated. For example, individuals who reported that their health limited the *kind* of work they could perform often also said their health limited the *amount* of work they could perform—so we combined these two variables into a single variable representing either variety of health limitation.

We also hypothesized that income might have a tendency to revert to the mean—in other words, that positive income shocks in one period might correlate with negative income shocks in the next period. As such, we constructed an "income change" variable representing the percentage change in a respondent's income over the prior two years.

Our exploratory data analysis revealed that a few of our proposed variables did not exhibit any statistically significant relationship with income shock—most notably, whether the respondent lived in an urban or rural area. After preliminary modeling confirmed that these variables did not have any meaningful explanatory power, we removed them from consideration.

Ultimately, we were left with a list of predictors that, after dummy encoding, yielded 71 total features.

**Computation, analyses, and modeling**

After the two cohorts' data sets were ready, we merged them, which resulted in a total number of 180,104 instances. We then used 12-fold cross-validation to estimate the performance of various machine learning models.

Given that our target variable was whether or not someone experiences an *income shock*, ours was a classification problem. Therefore, all of the algorithms we explored were classification algorithms—and specifically those that support probabilistic estimates. We considered Stochastic Gradient Descent (SGD) Classifier, Logistic Regression, Multilayer Perceptron (MLP) Classifier, Gradient Boosting Classifier, Random Forest Classifier, Bagging Classifier, and Gaussian Naïve Bayes (GaussianNB). We compared all machine learning algorithms against a baseline "dummy" predictor, which simply assigns every instance an income shock risk equal to the population mean.

All other candidate classifiers are powerful in their own way. For example, SGD Classifier is a fundamental technique for general classification problems. Logistic Regression and Gaussian NB algorithms are well-established and often produce relatively high classification accuracy, and they are easy to understand. Random Forest Classifier can handle large data sets with high dimensionality, and it can cope with missing data while maintaining accuracy. Gradient Boosting Classifier often improves upon the random forest approach and may yield high accuracy, fast training, and speedy prediction time. MLP Classifier has remarkable ability to derive meaning from complicated or imprecise data. And Bagging Classifier may help solve for problems of overfitting.

**Evaluation of classification algorithms and hyperparameter tuning**

To evaluate our classifiers' effectiveness, we first had to identify an appropriate evaluation metric. Because we believe many income shocks are inherently unpredictable (since even someone who appears to be at a low risk of an income shock could still lose their job unexpectedly), we determined that, rather than simply predicting that someone *will* or *will not* suffer an income shock, we should instead attempt to predict their *probability* of an income shock. As such, although we made some limited use of metrics such as F1 score that evaluate a model's classification accuracy, we preferred metrics such as log loss score and Brier score that evaluate the accuracy of probabilistic predictions.

Log loss is calculated for each predicted probability by taking into account the difference between the predicted probability and the actual class (e.g., 0 or 1 for binary classification). The reported log loss score is simply the average log loss across all observations. For any given model, a lower log loss score means better prediction, with a log loss score of 0 meaning that a model performs perfectly. One problem with log loss, though, is that it may not work well with test data that—like ours—reflects a large imbalance between the two classes.

An alternative is Brier score, which is calculated from the mean squared error between predicted probabilities and the expected values of actual classes. The reported Brier score is the average Brier score across all instances and takes a value between 0 and 1, with a lower value always preferred and a value of 0 suggesting a perfect forecaster. Although both log loss score and Brier score penalize estimated probabilities that deviate further from the expected value, Brier score experiences less impact from imbalanced data. Therefore, we used Brier score as our primary criterion to evaluate model performance.

Table 1 below shows the estimates of F1, log loss, and Brier score for each classifier, as well as the time used to train each fold.

Table 1: Classifier Performance Estimates

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Classifier** | **F1** |  | **Log Loss** | **Brier Score** | **Time to Train Each Fold**  **(Seconds)** |
| Baseline Dummy | 0.662 |  | 0.546 | 0.180 | -- |
| SGD | 0.691 |  | 0.539 | 0.175 | 2.10 |
| Logistic Regression | 0.691 |  | 0.518 | 0.169 | 1.48 |
| Random Forest | 0.698 |  | 0.513 | 0.168 | 41.99 |
| Bagging | 0.700 |  | 1.027 | 0.189 | 26.10 |
| Gaussian NB | 0.591 |  | 3.184 | 0.365 | 0.90 |
| MLP | 0.712 |  | 0.538 | 0.176 | 73.12 |
| Gradient Boosting  (with Default Configuration) | 0.687 |  | 0.510 | 0.167 | 33.24 |
| Gradient Boosting  (Hyperparameter Tuning) | 0.709 |  | 0.504 | 0.165 | 25.28 |

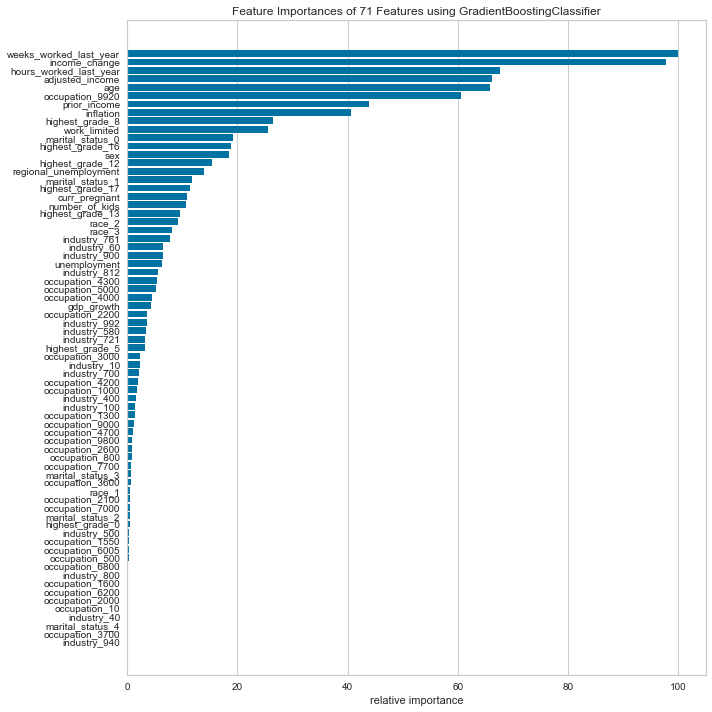
We noted that Gaussian NB Classifier had the highest (and thus worst) Brier score of 0.365, far inferior to the "dummy" predictor—likely due to highly correlated predictors or highly skewed continuous variables.

We also observed that MLP Classifier achieved the highest F1 score of any tested model, even though its Brier score was relatively poor. In other words, its absolute predictions ("shock" vs. "no shock") were impressive, but its probabilistic estimates were poorly calibrated. This suggests that a fruitful avenue for future research might be improving the calibration of the MLP Classifier's estimates.

We found that Gradient Boosting had the lowest Brier score among the untuned models at 0.167. We then conducted hyperparameter tuning using a five-fold cross-validated grid search, yielding an improved Brier score of 0.165 with an acceptable training time of 25.28 seconds. Therefore, we chose a tuned Gradient Boosting Classifier as our final model.

Although the improvement in Brier score from the dummy model (at 0.180) to our final model (at 0.165) may seem minor, it has significant real-world implications. Specifically, while the population-wide income shock risk in our data set is about 24%, our final model characterizes one-quarter of instances as having a risk of 13% or less—and another one-quarter as having a 31% risk or greater. This suggests that our model could provide well-differentiated (and, importantly, well-calibrated) risk estimates to different individuals at different points in their lives, enabling them to make smarter financial decisions.

The below chart illustrates the importance of all features in the final model. Notably, the variables most strongly associated with income shock ("weeks worked last year," "income change," and "hours worked last year") all relate to *instability* in a person's employment or financial situation. For example, in instances where "weeks worked last year" was less than 52 (indicating that the respondent had experienced a spell of unemployment), the risk of an income shock was dramatically elevated. Similarly, respondents who experienced a significant *increase* in income in one period were far likelier to experience a significant *decrease* in the next period.



**Discussion of the final product**

Our ultimate aspiration is to develop a web-based app that allows individuals to estimate their own risk of suffering an income shock.

Such an app could integrate visualizations to make the prediction easier to understand (for example, to make clear how an individual's personal risk compares to the population-wide risk). We have created a simple, illustrative version of such a visualization in the "Income Shock Predictor.ipynb" in this project's Github repository.

A more sophisticated app also could suggest ways for users to reduce their income shock risk or better prepare to deal with an income shock. It might, for instance, provide expert guidance on how to build an emergency fund, search for a new job, or budget on a variable income.

**Python packages used**

We made use of the following Python packages in our work:

ipython==7.4.0  
ipython-genutils==0.2.0  
ipywidgets==7.4.2  
jsonschema==3.0.1  
jupyter==1.0.0  
jupyter-client==5.2.4  
jupyter-console==6.0.0  
jupyter-core==4.4.0  
jupyterlab==0.35.4  
jupyterlab-server==0.2.0  
Markdown==3.0.1  
MarkupSafe==1.1.1  
matplotlib==3.0.3  
notebook==5.7.8  
numpy==1.16.2  
numpydoc==0.8.0  
pandas==0.24.2  
pandocfilters==1.4.2  
partd==0.3.10  
path.py==11.5.0  
pep8==1.7.1  
pexpect==4.6.0  
pickleshare==0.7.5  
Pillow==5.4.1  
pipreqs==0.4.9  
scikit-image==0.14.2  
scikit-learn==0.20.3  
scipy==1.2.1  
seaborn==0.9.0  
sklearn==0.0  
yellowbrick==0.9.1  
packaging==19.0

**Conclusion, avenues for future research, and lessons learned**

Working with large survey data within a short period of time is challenging. With more time, more issues could be addressed, which likely would improve the performance of our model.

First, our data set includes a significant amount of missing data, and dealing with missingness usually takes a great deal of effort. We used comparatively straightforward and simple methods to impute missing values; with more time, we expect that more sophisticated approaches, such as multiple imputation, could have achieved better results. Missing data also leads to a significant risk of nonresponse bias. If nonrespondents are disproportionately likely (or unlikely) to experience an income shock, that could skew our results—a concern that we have not had time to fully address.

A second concern is that our data set contains a significant class imbalance, with "shocks" reported in only about 24% of instances. If we had more instances of "shock," or perhaps had identified other data sets with further examples of income shocks, we would have been able to better train the classifiers.

Third, although we conducted some hyperparameter tuning, with additional time and computing power we likely could improve our model's accuracy. For example, we could have fine-tuned many hyperparameters of the MLP neural network, such as by adding further layers to the network topology—perhaps helping the neural network classifier offer better-calibrated predictions, and potentially outperforming our current model.

Fourth, we could use other measures to evaluate our model's performance. Although Brier score and log loss score offer useful measures of the accuracy of probabilistic predictions, other measures such as ROC AUC score or ROC curve (a plot of true positive rate versus false positive rate) might prove even more valuable—especially given our data's class imbalance.

And finally, we could explore integrating other variables from the NLSY data set or creating our own derivative variables. It's noteworthy that many of the variables most correlated with income shocks relate to instabilityin a person's life. Future research could build upon this insight by constructing derivative variables to indicate, for instance, whether a respondent has recently had a child or experienced a divorce. It's plausible that such life changes could correlate highly with income shock risk.

But even though these and other steps could certainly improve our model, we don't believe income shocks will ever be perfectly predictable. There are simply too many random factors affecting any one person's income to ever allow for flawless predictions. Even so, this pilot serves as a solid starting point from which we can explore further machine learning algorithms and refine our existing model.

**Location on Github of the code related to the project**

<https://github.com/georgetown-analytics/Shock-Prophets>