Shock Prophets

Will You Experience A 20% Income Shock Within 2 Years?

Thomas Seay, Ming Li, Lucas Melo, Qinya Pang

OUR PROBLEM

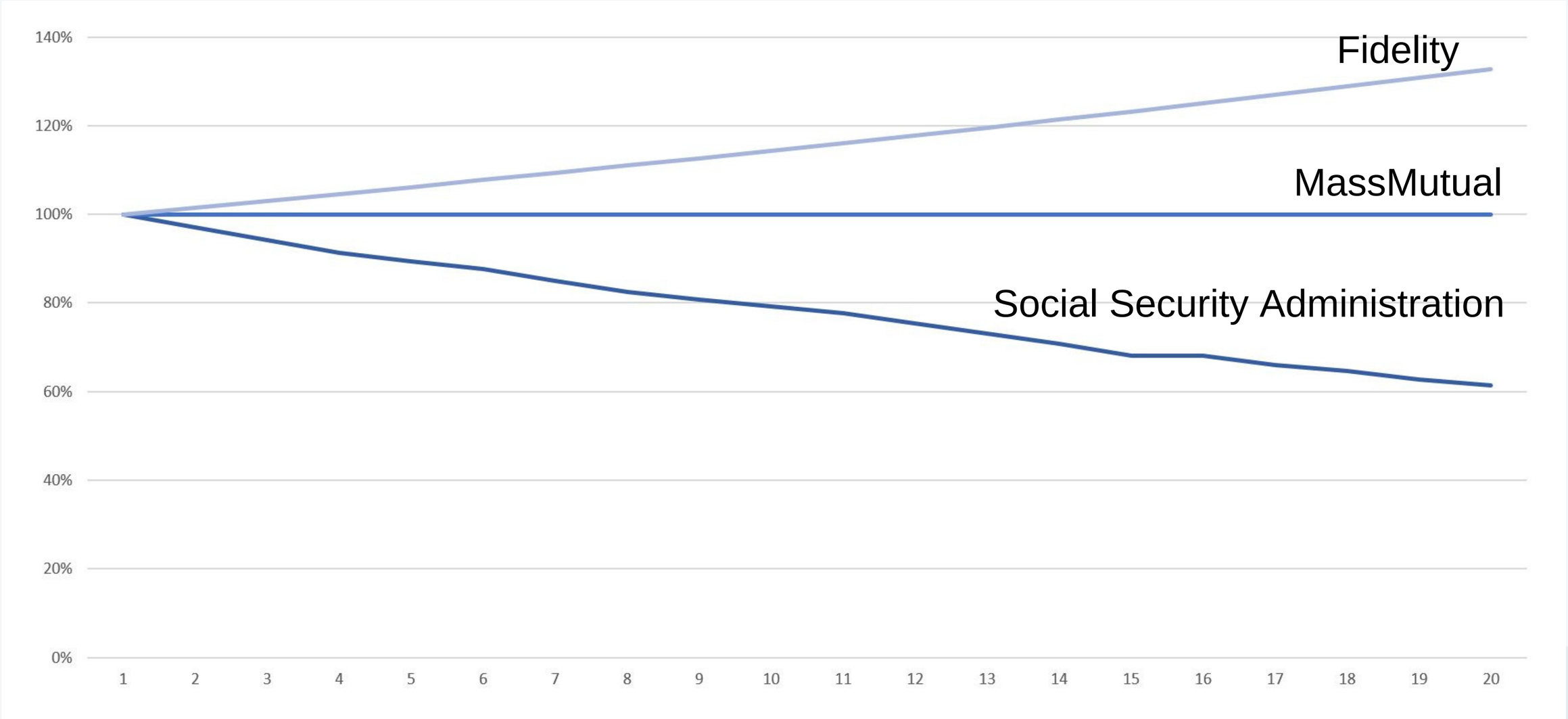
As modern adults, we constantly must make financial decisions that depend not only on how much we earn today, but on how much we'll earn in the future.

STUDENT LOAN

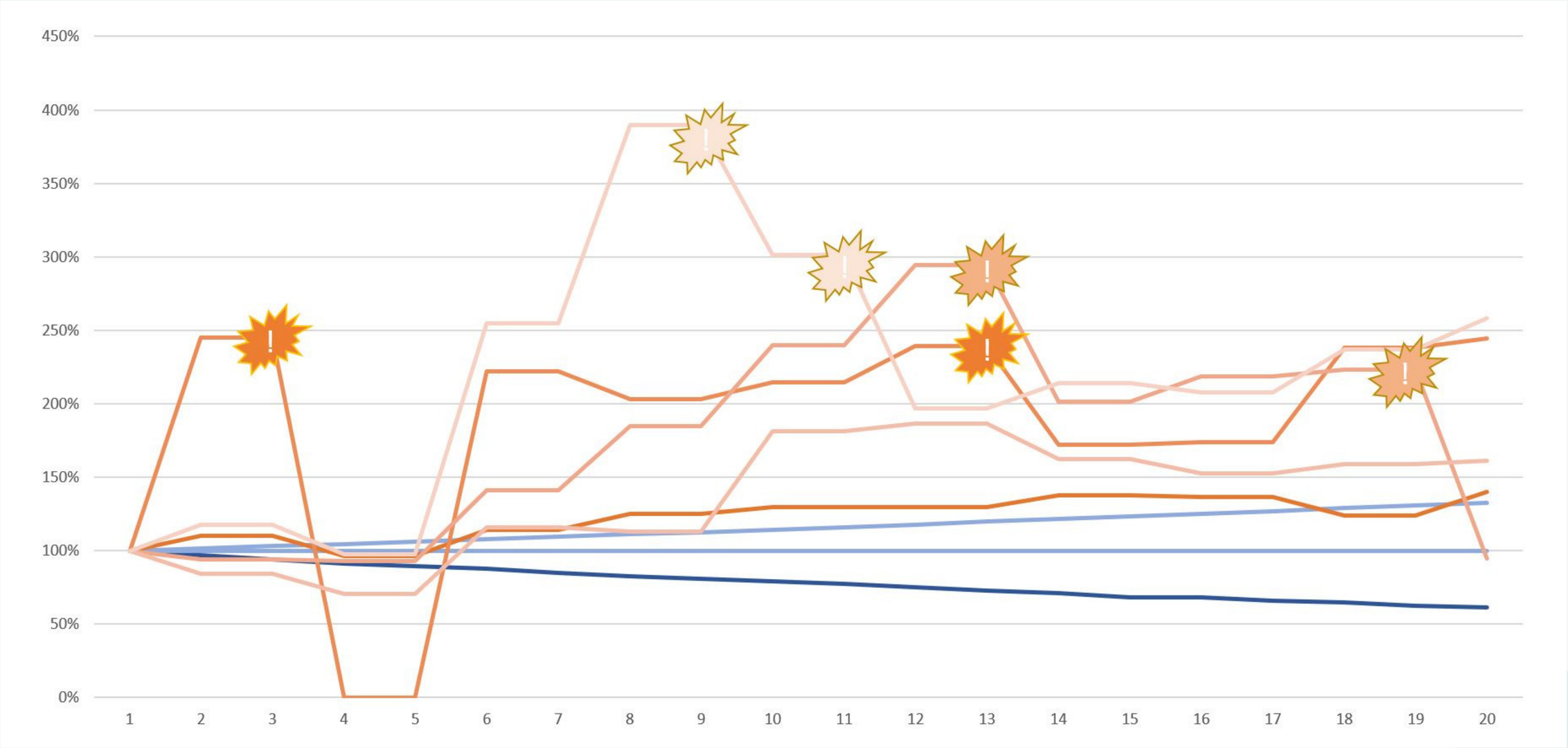
CAR LEASE/MORTGAGE

HOME MORTGAGE

Existing Projectors of Income



The Reality of Income Trends



Can Income Shocks Be Predicted?

OUR THEORY IS, YES, AT LEAST TO SOME DEGREE.

About Our Data Sources

NATIONAL LONGITUDINAL SURVEY OF YOUTH (NLSY)

- COHORT 79

The NLSY79 Cohort is a longitudinal project that follows the lives of a sample of American youth born between 1957-64. The cohort originally included 12,686 respondents ages 14-22 when first interviewed in 1979; after two subsamples were dropped, 9,964 respondents remain in the eligible samples.

- COHORT 97

The NLSY97 Cohort is a longitudinal project that follows the lives of a sample of American youth born between 1980-84; 8,984 respondents were ages 12-17 when first interviewed in 1997.

MACRO ECONOMIC FACTORS

- GDP GROWTH (BOTH REGIONAL AND NATIONAL) Bureau of Economic Analysis
- INFLATION RATE Bureau of Economic Analysis
- UNEMPLOYEMENT RATE (BOTH REGIONAL AND NATIONAL) Bureau of Labor Statistics

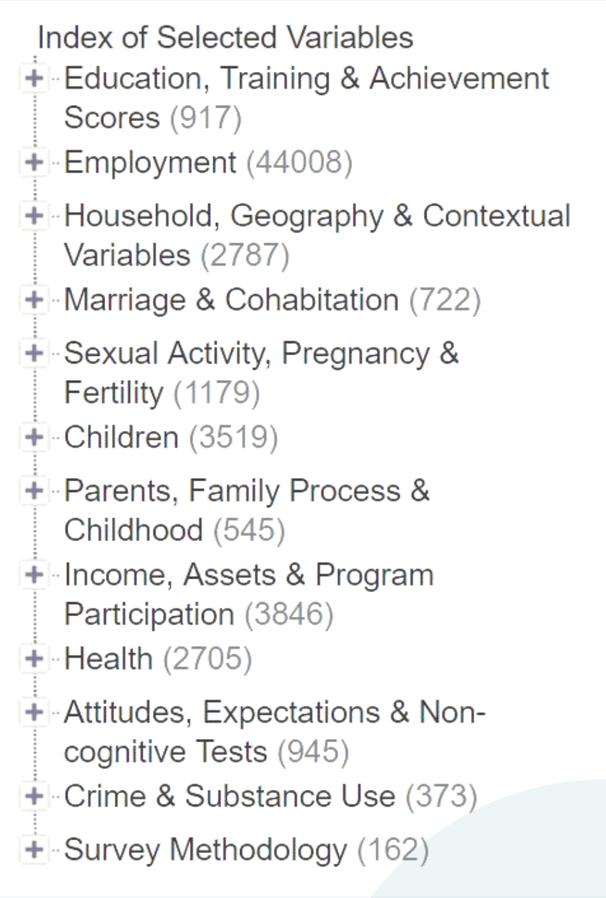
Data Ingestion and Wrangling

Issues with the NLSY Data

- Extremely high dimensionality
- Large amount of null and error codes
- Inconsistent variables and coding

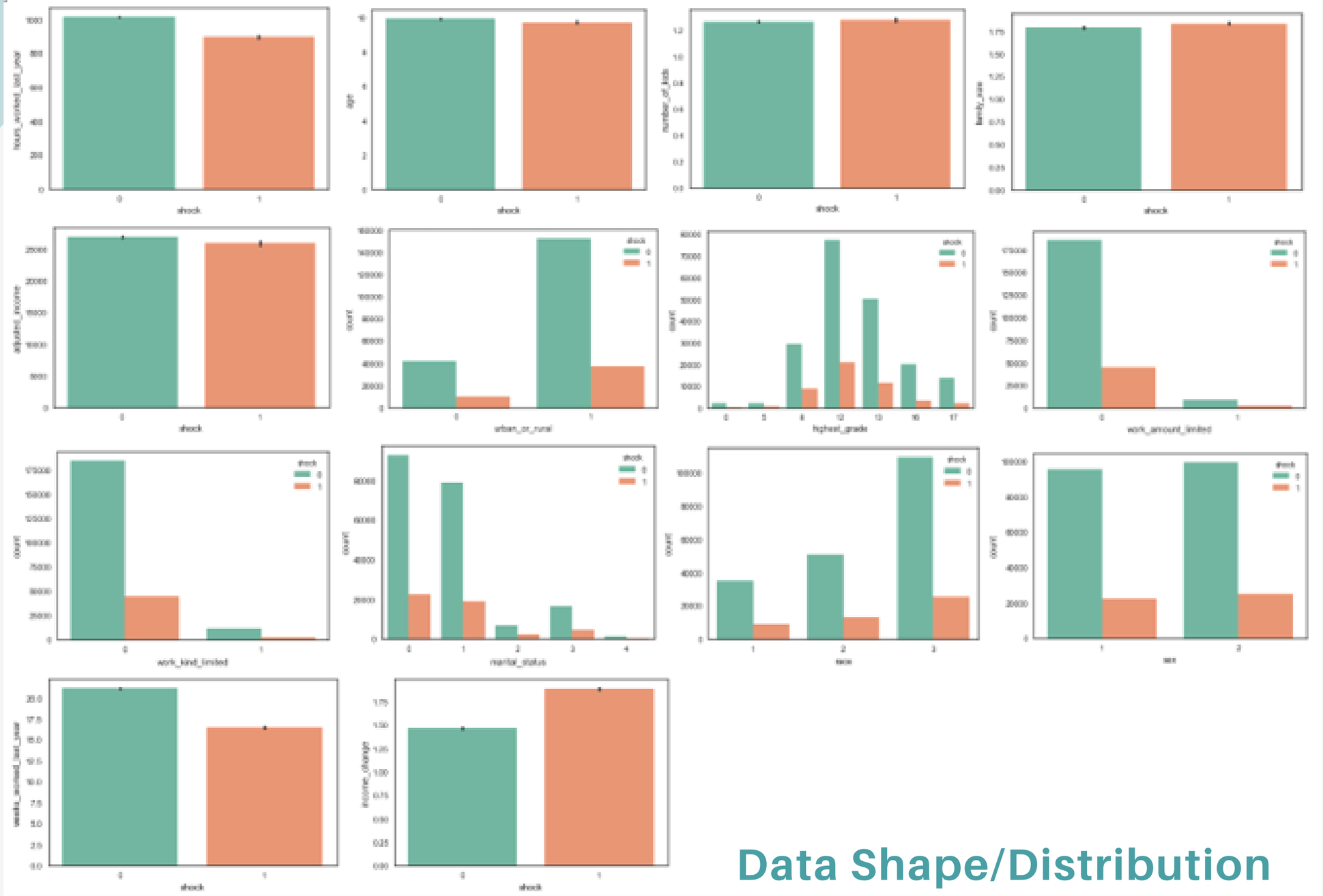
Data Wrangling Procedures

- Identified variables of interest through literature review
- Excluded inconsistent variables across cohorts and years
- Excluded variables that potential users wouldn't know
- Imputed missing data to be used in later analyses



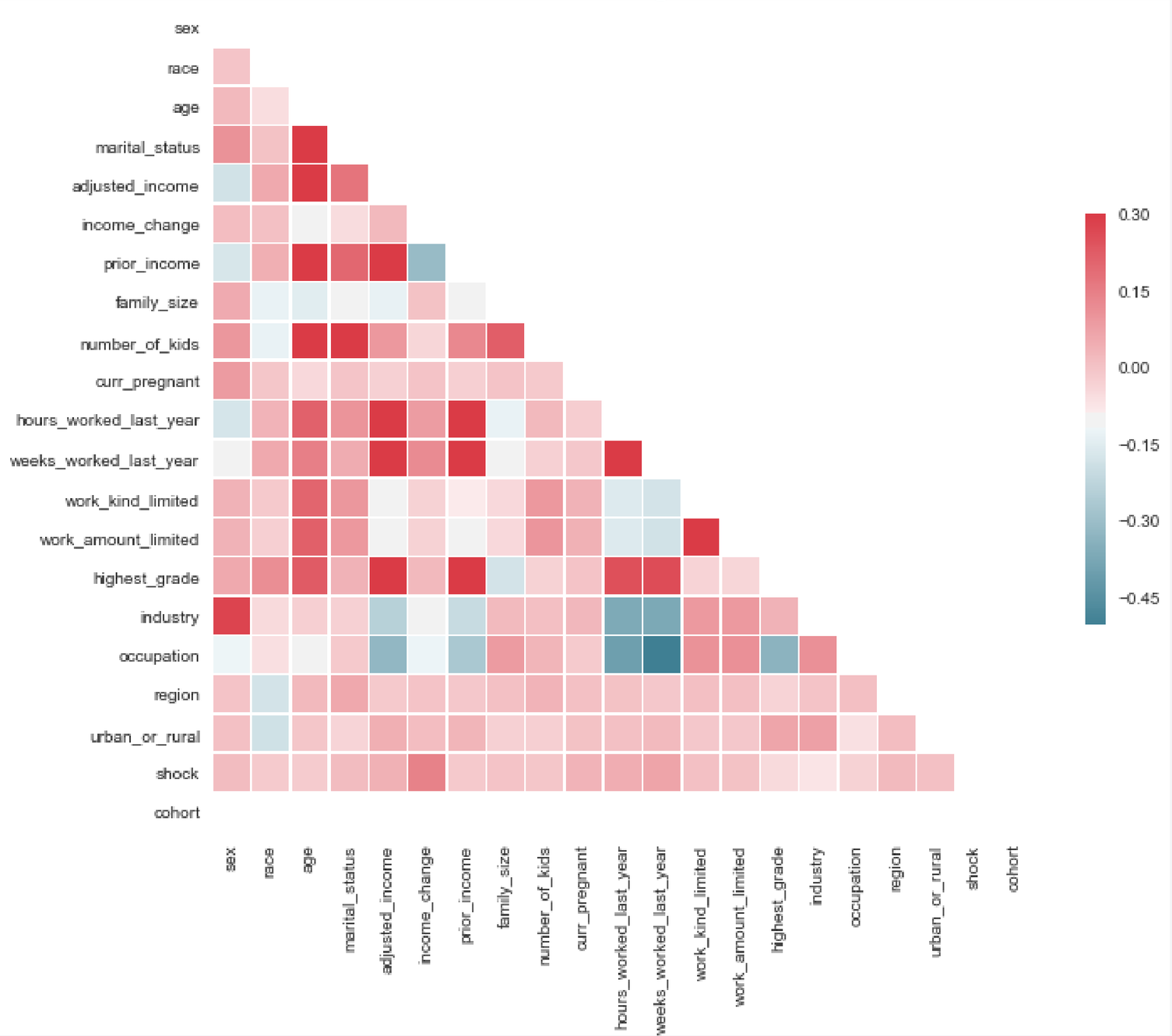
Computation and Analysis

- Exploratory Data Analysis (EDA)
- Feature Analysis, Selection, and Engineering
- List of Features Selected



Data Shape/Distribution

Correlations Among Variables



Feature Analysis, Selection and Engineering

List of Features Selected

sex
race
hours
age
marital_status
adjusted_income
current pregnancy
hours worked last year
weeks worked last year
number of kids

prior income
unemployment
gdp growth
inflation
regional unemployment
income change
work_limited
highest grade
industry
occupation

Instances: 180,104

Features(including dummy variables): 71

Modeling and Application

POTENTIAL MODELS

Due to a large sample size, and the fact that we want to predict probability rather than classify whether a person will experience income shock, the models that we explored are as below:

- SGDC Classier
- Logistic Regression
- MLPClassifer
- Gradient Boosting Classier
- Random Forest Classifier
- Bagging Classifer
- GaussianNB

Modeling and Application

F1 SCORE

F1 Score was used to filter bad-performing models

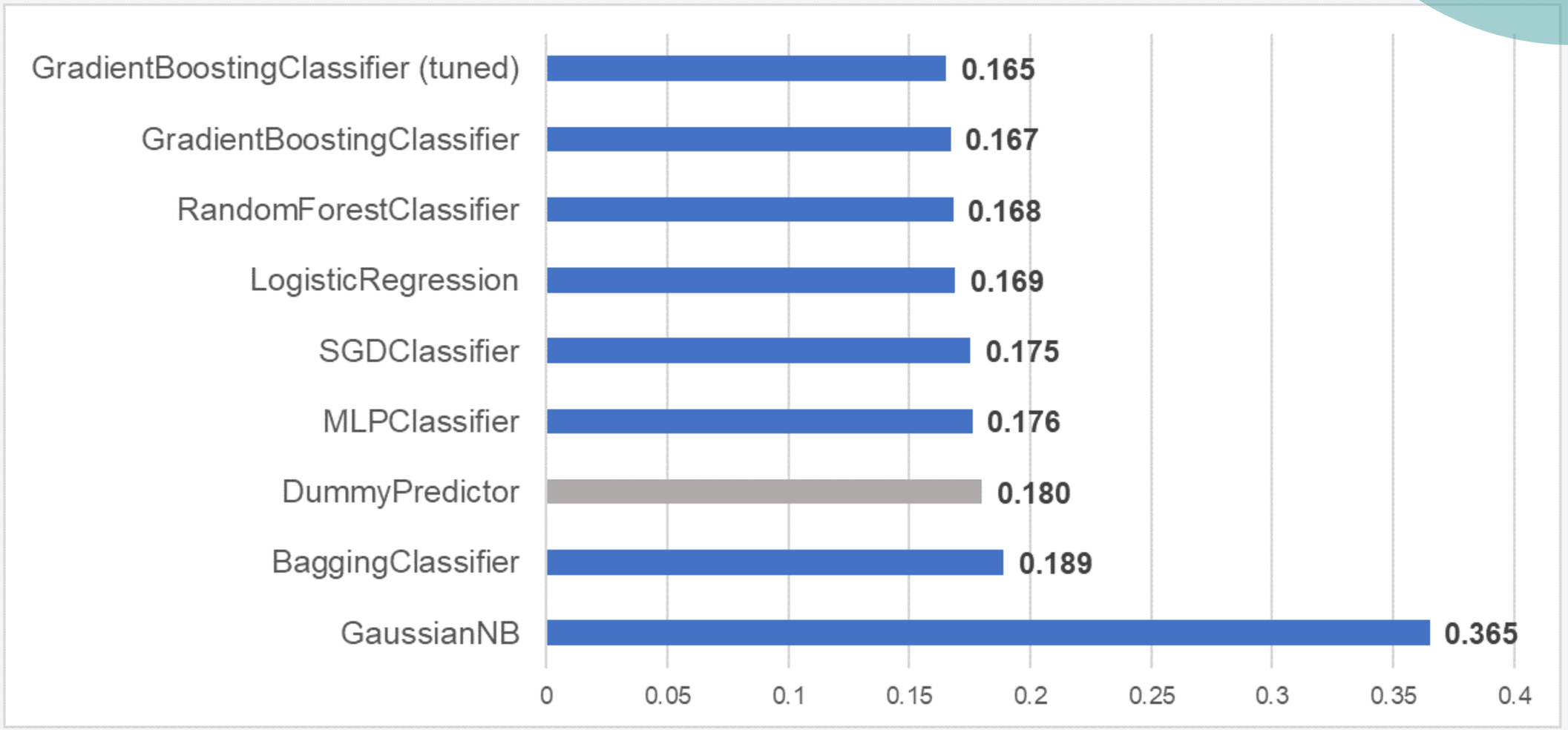
BRIER SCORE

Brier score is proper score function that measures the accuracy of probabilistic predictions. The lower the Brier score is for a set of predictions, the better the predictions are calibrated.

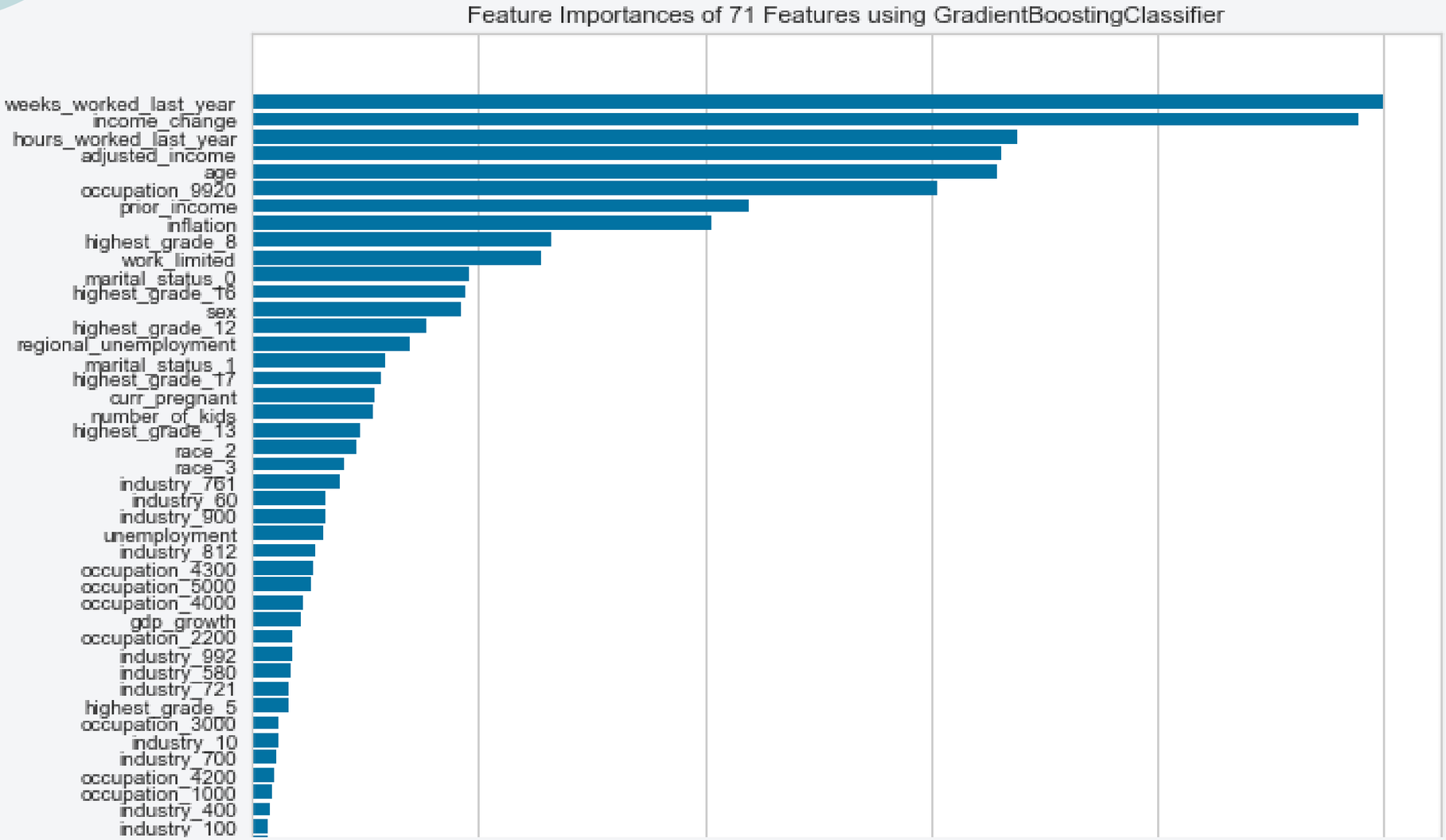
CROSS ENTROPY (LOG LOSS)

Cross entropy can be used to define a loss function in machine learning and optimization. The true probability is the true label, and the given distribution is the predicted value of the current model.

Model Selection

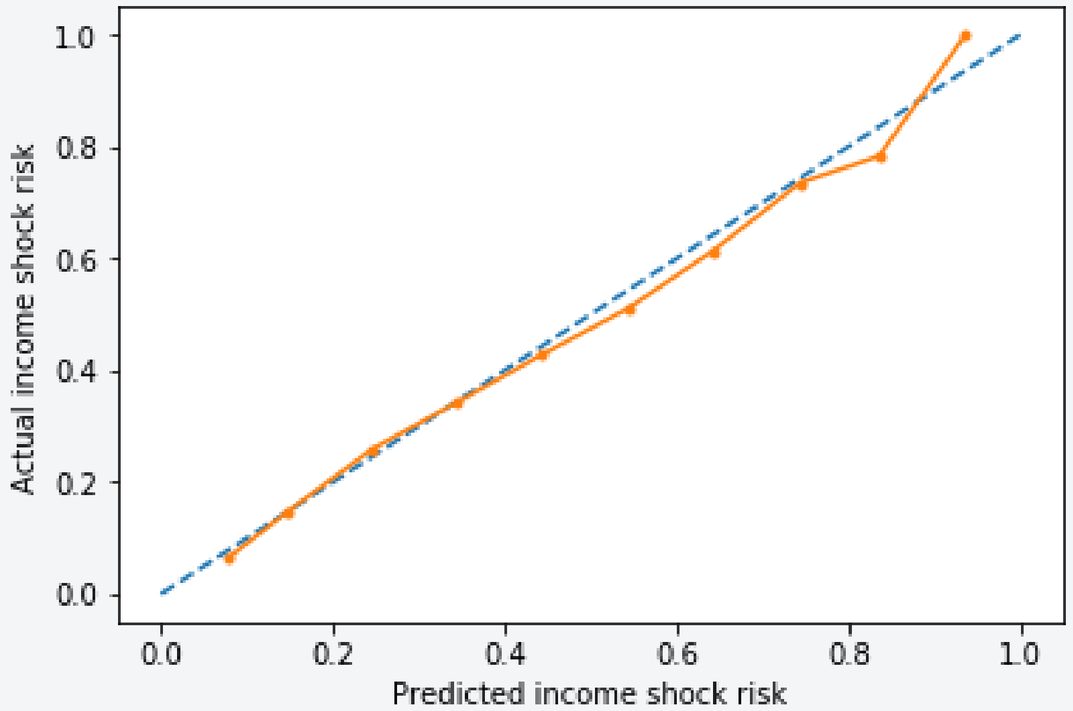


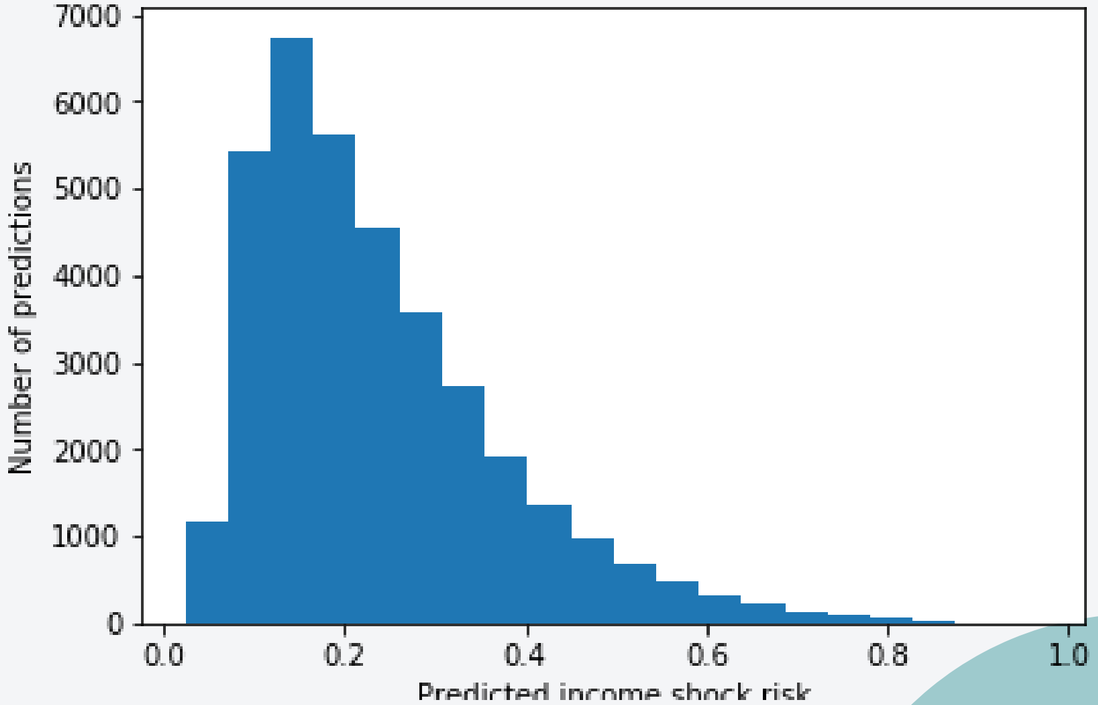
Feature Importance



Calibration Chart

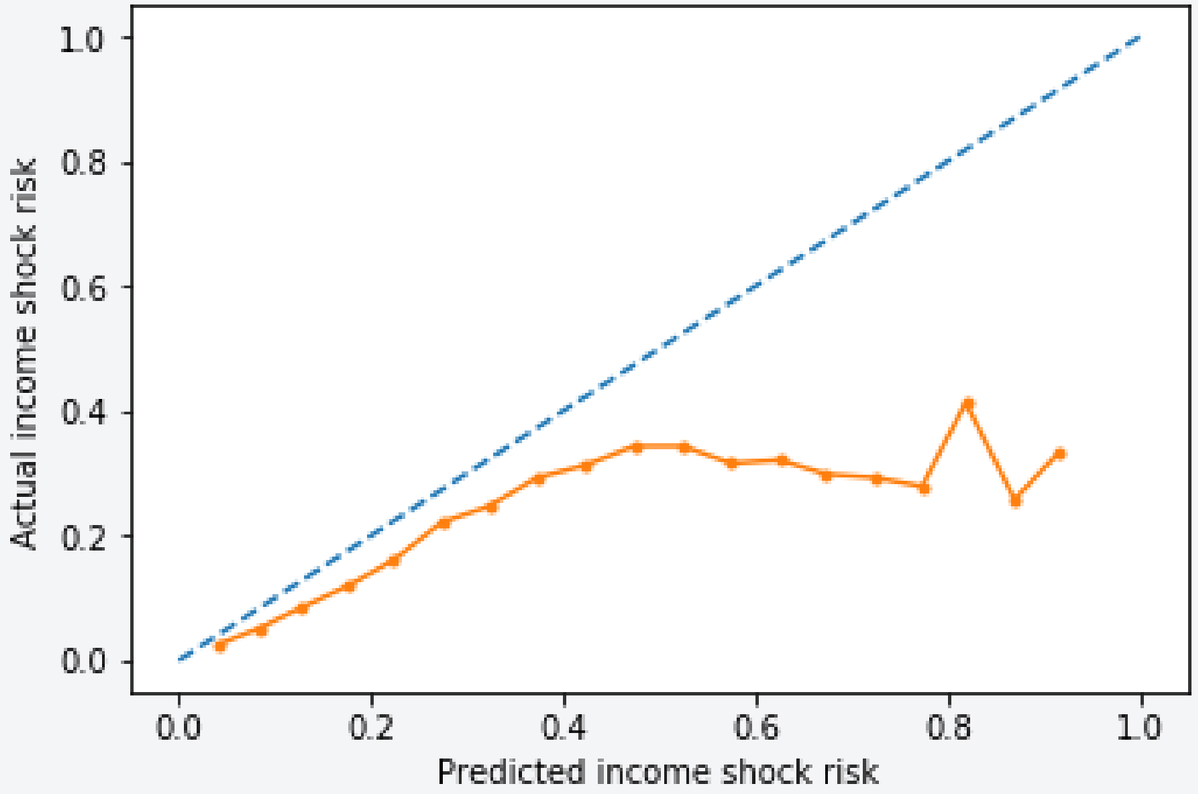
Trained on 80% full sample, tested on 20% full sample

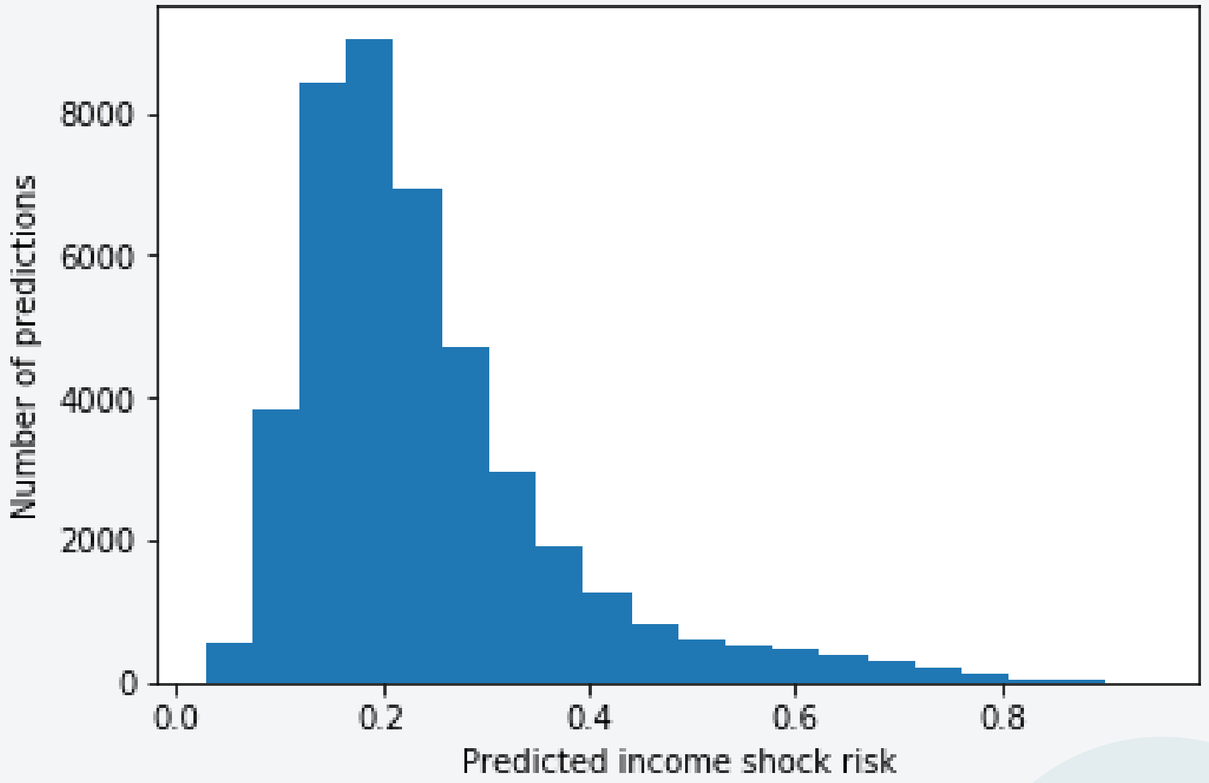




Calibration Chart

Trained on cohort 79, tested on cohort 97





Tools and Packages

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

SHOCK PROPHETS

Let's try it out!

Potential Next Steps

- STREAMLINE GDP, INFLATION, UNEMPLOYMENT DATA
- ADD IN MORE DERIVED VARIABLES TO REFLECT CHANGES WITHIN LAST TWO YEARS (CHANGE OF INDUSTRY, HAD KIDS, ETC)
- ADD IN FINANCIAL TOOLS TO HELP PEOPLE INTERPRET THE SCORE ESTABLISH AN SAVING/INVESTMENT PLAN

ANNUAL INCOME TWENTY POUNDS, ANNUAL EXPENDITURE NINETEEN SIX, RESULT HAPPINESS. ANNUAL INCOME TWENTY POUNDS, ANNUAL EXPENDITURE TWENTY POUND OUGHT AND SIX, RESULT MISERY. --CHARLES DICKENS

SHOCK PROPHETS