Predicting Credit Card Defaults

Introduction – The Project

- Economic downturns represent significant operating risk to the lending industry. They are also notoriously hard to predict
- Rather than attempting to predict economic downturns, this project seeks to predict the outcome that lenders care about: Credit Card Default Rates
- Publically available macroeconomic data is used to predict the future path of the aggregate credit card default rate, as recorded by the Federal Reserve

Sourcing the Data

- Nearly a hundred economic datasets were selected from the Federal Reserve Economic Data (FRED), a publically maintained repository for economic datasets.
- Variables from several categories were selected:
  - Aggregates designed to predict macroeconomic outcomes
  - Measures of employment
  - Metrics for aggregate and sector-specific economic production
  - Interest Rates
  - Stock Indices
  - Measures of Inflation
  - Measures of Consumer and Producer Sentiment
  - And other popular measures on FRED

Wrangling the Data

- The data required significant manipulation to process it into a usable form:
  - Daily and weekly data were resampled into monthly features, including average monthly day-over-day and week-over-week changes of each variable, the median value for each variable each month, and many more calculations
  - Quarterly and annual data were imputed to monthly data using a variety of methods
  - Month-over-month and year-over-year changes in each variable were added as separate features

Exploratory Analysis

- Several manipulations were made to nearly 100 features, making exploration of each feature impractical
- Exploratory analysis was performed on some of the more promising variables. The analysis was designed to validate that:
  - Some variables track closely with the outcome variable, the Credit Card Default Rate. If no variables can be found that correlate strongly with the default rate, then it will be difficult for a model to predict the default rate.
  - At least some of the variables *lead* changes to the Credit Card Default Rate. The goal of the model is to *predict* defaults, which will be impossible if variables lag or directly overlap the default rate.

Exploratory Analysis – Industrial Production

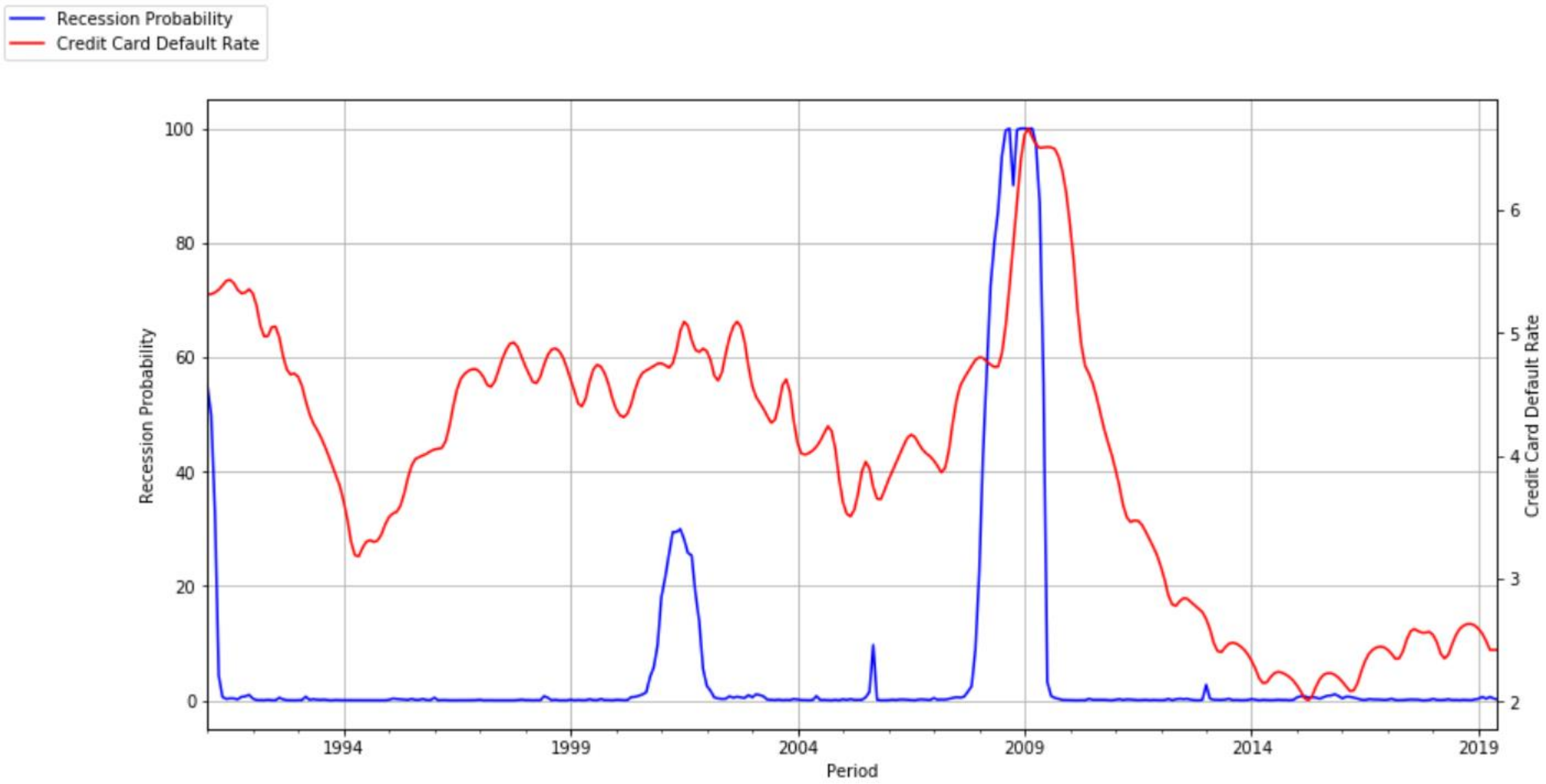
- The Industrial Production Index measures all industrial production within the United States, including manufacturing, some natural resource production, and some utilities.
- Industrial production is generally known to track closely with economic output, as manufacturing is a large component of total economic activity. Industrial Production steadily increases over time as the economy grows, so the *year-over-year change* in industrial production is considered.
- The next slide shows the year-over-year change in industrial production graphed as a time series with the aggregate credit card default rate.
  - As expected, Industrial Production has a very strong looking inverse relationship with credit card defaults.
  - The variable looks as though it might begin to change before credit card defaults - for example, Industrial Production rises sharply in 1991, *after which* credit card defaults appear to fall.

Exploratory Analysis – Industrial Production

Exploratory Analysis – Recession Probability

- The start date of a recession is often not identified until several months after its start. For this reason, there is value to predicting whether or not the economy is currently in recession.
- The Smoothed U.S. Recession Probabilities metric is a measure of the likelihood that a recession would occur during each month. As an economic recession typically severely increases credit card default rates, it is expected that a rise in the recession probability will be associated with an increase in defaults.
- The next slide shows the relationship between recession probabilities and the Credit Card Default Rate. The probability of a recession is typically zero, but when it increases it is often associated with an increase in the rate of credit card defaults, and it seems to be a very strongly leading indicator.

Exploratory Analysis – Recession Probability



Exploratory Analysis – Consumer Sentiment

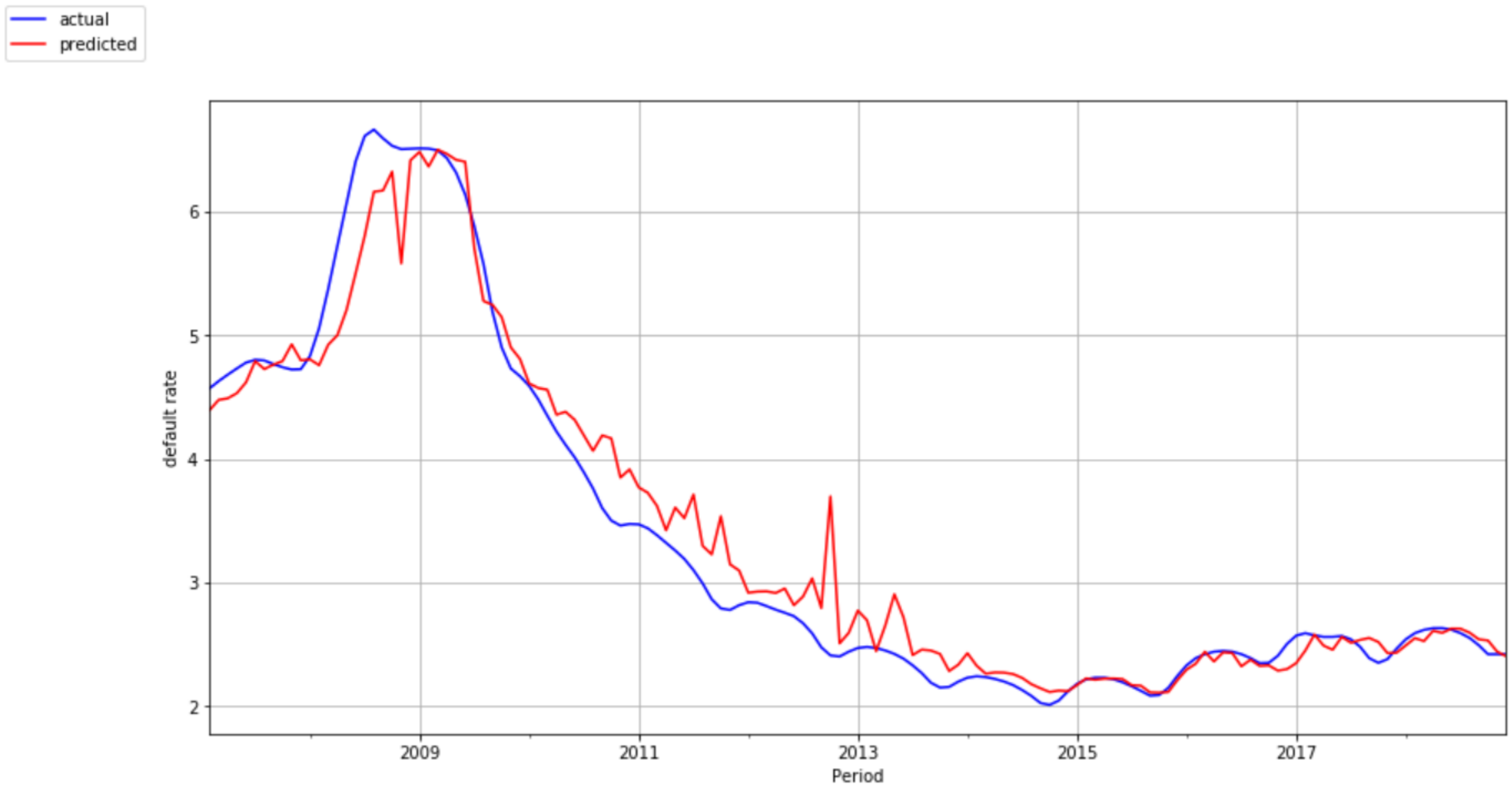
- The University of Michigan conducts monthly surveys of consumers to evaluate how they perceive the health of the economy, the business environment, and their own personal finances.
- Results are aggregated into a single index that is designed to have a value of 100 during 1966Q1 - values greater than 100 represent sentiment that is better than this benchmark, and values less than 100 represent sentiment that is worse.
- The next slide shows consumer sentiment plotted with the credit card default rate. It is very common, as expected, for an increase in consumer sentiment to be associated with a decrease in the credit card default rate shortly after, which makes consumer sentiment a very good leading indicator.

Exploratory Analysis – Consumer Sentiment

Machine Learning – Introduction

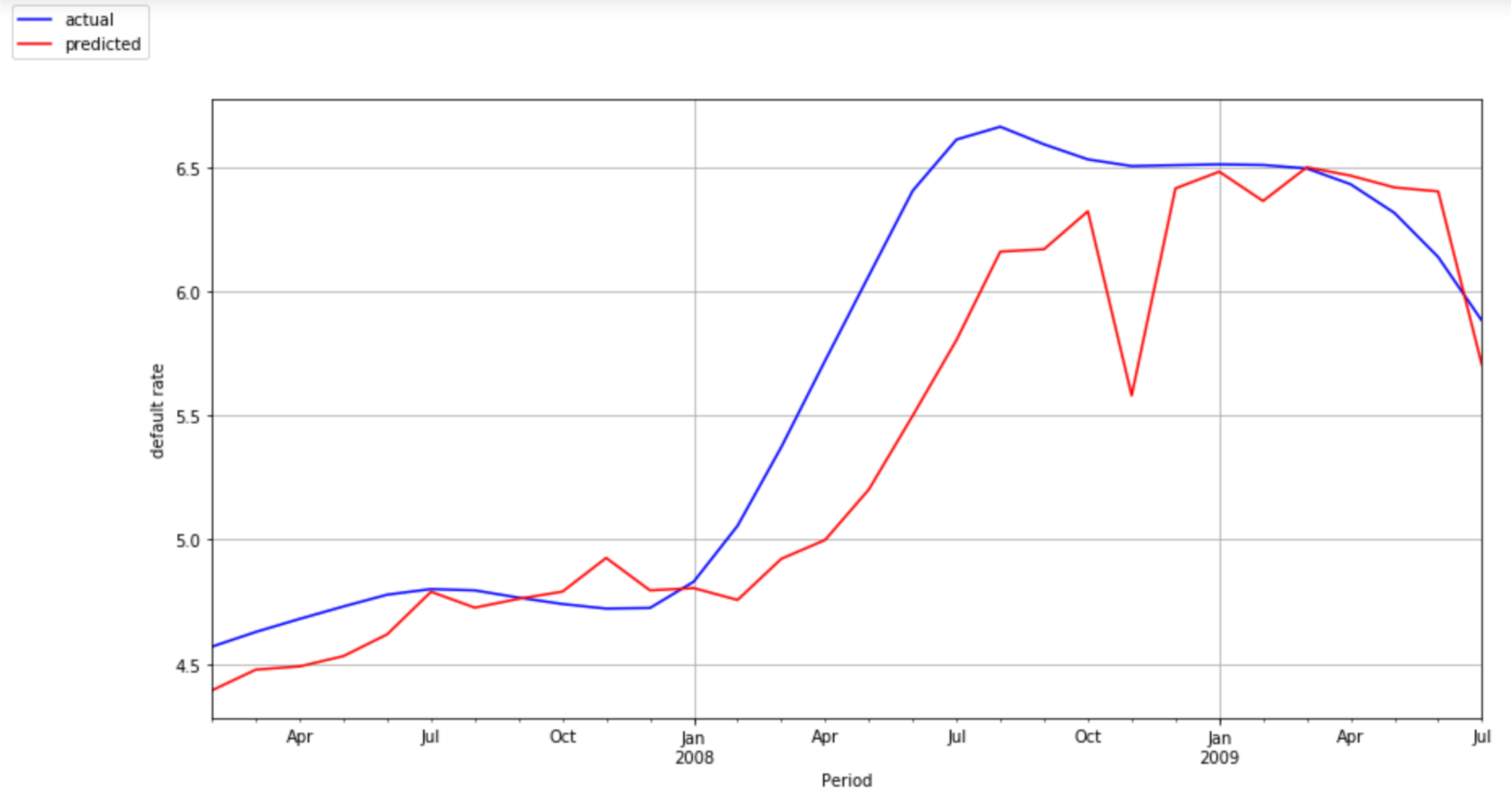
- We have seen that there are economic variables that closely predict the credit card default rate, some of which might change *before* changes in the credit card default rate occurs. This should provide sufficient information to train a Machine Learning model to predict defaults.
- Two models were tested – Random Forest was used to iteratively evaluate and improve feature selection and engineering. Gradient-boosted random forest using XGBoost was compared to unboosted Random Forest and improved predictions significantly, so it is the model that was used.
- Several iterations of models were used – the best result was found with XGBoost predicting the default rate with a Root Mean Squared Error (RMSE) of 0.284. This is a high degree of accuracy for predicting a series with a standard deviation that is higher than 1.

Machine Learning – Results



- The graph above shows the predicted value of the default rate in red and the actual value in blue. There is a lag in predictions when the default value is changing rapidly, but the lines are in relatively close alignment.

Machine Learning – Results



- Charted above is the same view but including only February 2007 through July 2009. The lag is pronounced, but there is still significant ability to predict downturns – the model predicted a ~100bps increase in the default rate from Dec 2007 to May 2008. The actual increase was around ~175, so while the model failed to gauge the magnitude, it got the direction correct a full year before the 2008 recession was officially declared.

Machine Learning – Feature Importance

- **Feature Importance** can be extracted from the model to determine the contribution to predictions from each independent variable.
- A table of the top 15 features is shown to the right. The values are roughly as expected – broad macroeconomic indicators, household debt burden, and business-related measures dominate the most important features in determining the default rate.

Description	Avg. Feat. Imp.
Quarterly Nominal GDP with linear interpolation.	0.078156
The year-over-year percent change in the number of heavy trucks purchased nationwide.	0.073253
The change from 9 months prior in quarterly real GDP.	0.062874
The Consumer Price Index for urban consumers, a measure of inflation.	0.054636
Household debt service payments as a percent of disposable personal income.	0.045195
Year-over-year percent change in the average sale price of homes.	0.041364
Year-over-year change in the median monthly difference between the 10-year and 2-year treasury rates.	0.035186
The year-over-year absolute change in the number of heavy trucks purchased nationwide.	0.025162
The year-over-year percent change in quarterly profits.	0.024532
The change from 12 months prior in quarterly real GDP.	0.023567
The minimum value of the daily Volatility Index during the month.	0.022132
The mean value of the daily Volatility Index during the month.	0.021893
The percent change from 9 months prior in the level of unemployment.	0.019394
The change from 9 months prior in the Effective Federal Funds rate.	0.019266
The year-over-year change in the absolute S&P/Case-Shiller Home Price Index.	0.018480

Conclusion

- The model has been demonstrated to predict the aggregate level of credit card defaults 6 months hence with relatively high accuracy.
- While predicted *large changes* in the default rate appear to be small relative to actual changes, the model does appear to accurately predict that the changes will occur before they would otherwise be known to do so.
- Further development of the model could increase its accuracy. This could include:
  - Adding new macroeconomic time series
  - Performing different types of feature engineering on the macroeconomic data
  - Evaluating different machine learning models

Thank You