## Credit Card Aggregate Default Prediction

### I. Introduction

Economic downturns are a significant operating risk for many industries. The lending industry is particularly impacted, as economic downturns not only reduce demand for their services, but they also impose credit losses when borrowers impacted by the downturn are no longer able to service their debts.

Lending companies nearly always respond to recessions by tightening credit standards - refusing to lend to riskier borrowers who they might lend to in better times, or lending less to each borrower. There is tremendous value, then, in identifying downturns as soon as possible, including forecasting downturns. This project uses publically available variables, such as retail sales aggregates and consumer sentiment metrics, to attempt to predict downturns. There are several firms, including Moody's Analytics, that attempt to predict the probability of a recession occurring - rather than attempting to classify future periods by whether or not they will be recessions, this project predicts the outcome variable of interest most relevant to credit card lenders - the aggregate credit card default rate, which is tracked quarterly by the Federal Reserve. The potential clients for this project include lenders of all types, who could save money by tightening credit in anticipation of a pending increase in defaults.

### II. Data Sourcing

A host of economic factors are likely to be predictors of loan delinquencies and defaults. Classic leading indicators of economic downturns, such as the yield curve, manufacturing indices, and jobless claims, would be useful predictors of economic downturns, which in turn cause credit deterioration. Often, credit deterioration is at its worst *in the late stages of a recession*, so even lagging predictors of recessions could be used to predict delinquencies.

For the initial stage of this project, 94 different economic indicators are sourced from the Federal Reserve Economic Data (FRED), a repository of publicly available economic datasets. These indicators were selected based on their expected ability to predict credit deterioration, and included variables in the following categories:

1. **Variables that Predict Recessions Directly**: Economic contractions, which can be protracted and often impact all industries, consistently involve deterioration in credit performance. Several variables are useful for recession prediction but do not necessarily

impact credit performance directly. One such class of variables is indicators of *yield curve inversion*, which is historically a leading indicator that a recession will occur. Another such variable is the *Smoothed U.S. Recession Probabilities* metric, which was created to attempt to identify recessions as soon as they occur.

2. **Unemployment Measures:** Unemployment is a common reason that borrowers are unable to service their debts. Multiple measures of unemployment are sourced from FRED, including the standard headline Unemployment Rate, calculated as the proportion of the active labor force that is employed, and the Labor Force Participation Rate, which is the proportion of the entire population that is part of the active labor force.
3. **Production Measures:** Measures of production, including an Industrial Production index and economy-wide GDP, were included.
4. **Macroeconomic Variables Associated with Downturns:** There are many macroeconomic outcomes associated with poor economic performance. A list of these outcomes, and some of the variables selected to represent them, follows:
    a. **Interest Rates**: Reducing interest rates is one of the primary levers utilized by the Federal Reserve to respond to economic contractions. Several metrics for interest rates, including the 1-, 2-, 10-, and 30-year treasury rates, were selected.
    b. **Stock Indices:** Stocks fluctuate rapidly with changing economic conditions, making them a potentially leading indicator of the state of the economy. Multiple stock indices and aggregates are tracked.
    c. **Inflation:** The inflation rate tends to fall when economic contractions occur. A measure of inflation and a measure of expected future inflation are both included.
    d. **Consumer and Producer Sentiment:** Qualitative consumer and producer sentiment surveys are produced monthly. Their results are often leading predictors of economic health, as consumers and producers might identify softness in their respective industries before they appear in other economic measures.
    e. **Financial System Stability:** Aggregates that relate to the stability of the financial system, such as the St. Louis Federal Reserve Financial Stress Index.
5. **Other Metrics:** A host of other metrics that may be associated with economic well-being are included. Many of these were selected based on their being popular indicators pulled from FRED, which might indicate that others find them to be useful economic measures.

**III. Data Wrangling**

The 94 data series are all sourced from FRED using the open-source *fredapi* package. The outcome variable of interest is tracked monthly, while the dependent variables are tracked as frequently as daily or as sparsely as annually. For this reason, significant manipulation is required to wrangle each dataset into a set of monthly data. Additionally, a host of feature engineering steps are performed to ensure that the maximum amount of information is extracted from each variable. The manipulations and feature engineering performed on each variable depend on its frequency as follows:

1. **Daily:** Several summary statistics were used to convert daily data to monthly granularity. The maximum, minimum, mean, and median of each variable is calculated for each month, as different measures of the month's "value" for the variable. The standard deviation of each variable is also taken for the month, as it is expected that volatility of some measures may be a good predictor of economic outcomes. For each of the measures mentioned above, the difference between the variable and a prior value for the variable is taken - for example, the maximum, minimum, mean, median, and standard deviation of the 1-day *difference* in the variable is calculated. Similar calculations are performed for different date differences between 1 and 365, such that week-over-week, month-over-month, and year-over-year changes in each variable are included in the final dataset.
2. **Weekly:** Weekly data is calculated nearly identically to daily, except that all data summarization is based on each of the weeks that fall within the month. For example, if 4 weeks' worth of data is available during a month, then the minimum, maximum, and other summary statistics for each of those 4 weeks represents that week's values. Differences in each summary metric are also calculated, but are based on weekly instead of daily differences - for example, maximum week-over-week, 4-week, and 52-week change in each variable is inluded.
3. **Monthly:** Monthly data requires the least manipulation. Differences are calculated for monthly data similar to daily and weekly data above - the 1-month and 12-month difference in each variable, for example, is included as a feature.
4. **Quarterly:** Quarterly data requires imputation. A host of imputation methods, including linear, polynomial, quadratic, and cubic, are used to create different features. Additionally, once the imputations are performed, the same difference-from-prior-months calculations described for monthly data are applied to the imputed data.

5. **Annually:** The process for annual data is identical to that for quarterly data.

Additionally, where required and not mentioned above, forward-filling of missing data is performed where required. For example, stock indices do not have values on weekends and holidays, when stock markets are closed, so the prior day's value is input.

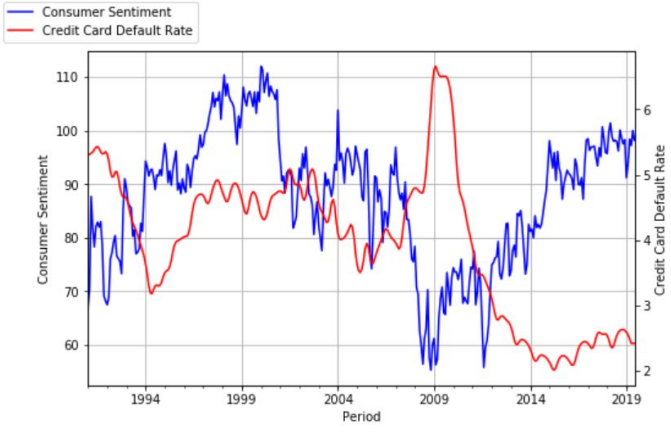
## IV. Exploratory Analysis

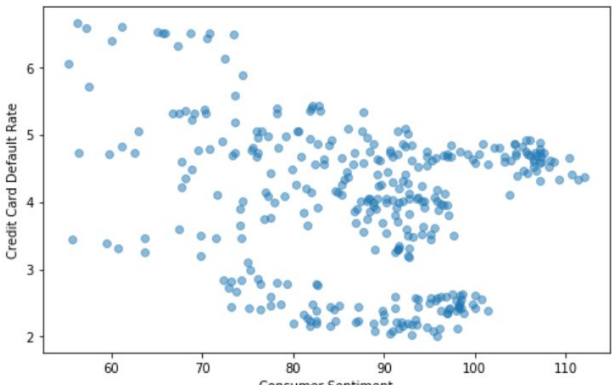
Next, the data is explored to evaluate whether a selection of the chosen metrics seems to track with the credit card default rate as expected. While more variables were selected and explored, 3 of the 94 variables that seemed highly likely to have relationships to the outcome variable were chosen to evaluate in greater detail below:

### University of Michigan Consumer Sentiment Index:

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. It is expected that consumer sentiment and the credit card default rate are negatively correlated.

The two charts below show different visualizations of the relationship between Consumer Sentiment and the Credit Card Default Rate. The line chart shows how the two are related over time - 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. The inverse is also true. The scatter plot appears to show a negative relationship between the value of the two variables plotted independent of the time that they occurred. A linear regression was performed to verify this observed relationship - the 95% confidence interval for the relationship is -0.030 to -0.009, suggesting that an increase in Consumer Sentiment by 1 is associated with a statistically significant ~0.02% decrease in the rate of credit card defaults.

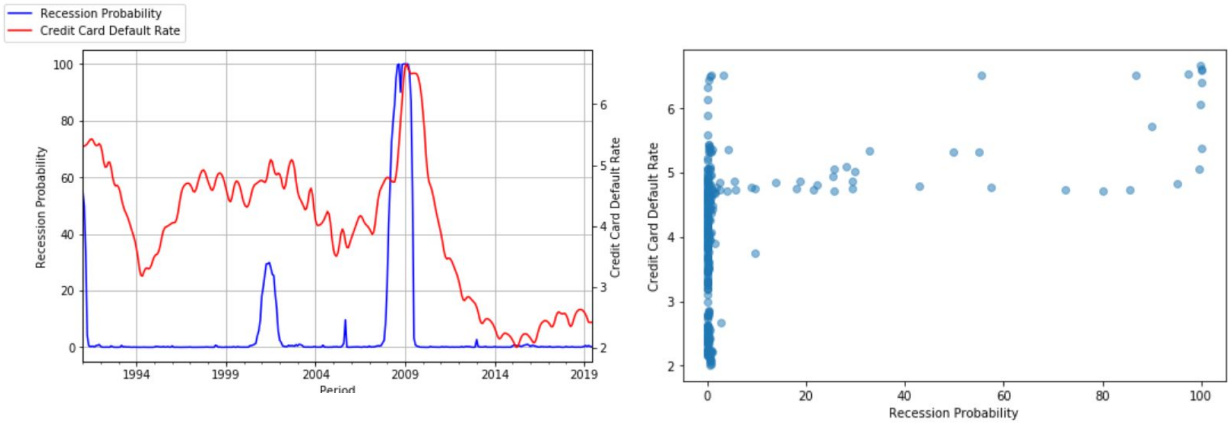


**Smoothed U.S. Recession Probabilities:**

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 two charts below show 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 leading indicator. The scatter plot appears to show a negative relationship between the value of the two variables, but it is difficult to know how strong the relationship is as most values are clustered around a zero recession probability. A linear regression was performed to evaluate the relationship - the 95% confidence interval is 0.019 to  0.027, suggesting that an increase in the recession probability of 1% is associated with a statistically significant ~0.023% increase in the rate of credit card defaults.




**V. Machine Learning**

Given the observed relationship between a variety of macroeconomic variables and the Credit Card Default Rate, it appears that defaults are likely able to be predicted with some level of accuracy. The next step is to use a variety of Machine Learning models to attempt to predict the default rate, trying different models, hyperparameters, and features to arrive at the most predictive model. The standard deviation of the dependent variable was taken as a benchmark,

and found to be **1.118** - this is the benchmark to which all models will be compared, as it is the baseline variation in the data.

The goal of the model is to predict the credit card default rate six months in the future. This would allow for predictions with a distant enough lag that a lender could modify underwriting thresholds or marketing strategies to ensure that those with an elevated default likelihood are not able to open accounts ahead of a broad deterioration in credit quality.

As the outcome variable is a continuous number rather than a category, regression models are tested. The **root mean squared error (RMSE)** was selected as the regression metric - large prediction errors could cause significant business errors, so selecting a metric such as RMSE that disproportionately penalizes large errors is advantageous.

Although all of the data used is time series data, each sample is treated as independent for the purpose of training and testing machine learning models - each month's set of variables is treated as a separate independent sample that predicts that month's credit card default rate. However, because each record of the time-series datasets would only be known at the month of its release, a typical random train-test split was not performed. Instead, the following sequence was performed:

1. Lag the dependent variable by six months. This means that each month's worth of independent variable observations aligns with the dependent variable *in six months*. Put another way, each test will attempt to predict the credit card default rate in six months, using only the data that would have been known six months prior to the release of the credit card default rate.
2. Using the single most recent known observation as the test set and the remainder of the observations as the training set, train the Machine Learning model on the training set and record the predicted and observed values.
3. Eliminate the most recent record from both the dependent and independent variables and repeat step (2). This is performed for each month of the prior 12 years to ensure that performance over a recession is included.
4. The RMSE is calculated based on the 144 observations tested in steps (1) through (3).

The primary advantage conveyed by this method is that the results are very close to those that would have been calculated in the past using the same methodology with all of the information known at the time. Each result is effectively the result that the model *would have produced with information available at the time.*

**Random Forest**

The first machine learning model that was tested was a Random Forest Regressor. It was expected that a variety of features will be added, so choosing a relatively fast model like Random Forest was desirable. Additionally, the extensive feature engineering that was performed resulted in an enormous number of features, with Random Forest models perform well on.

Grid Search Cross Validation was used to tune hyperparameters. Because Random Forest is a relatively fast model to tune, hyperparameter tuning was performed on each of the 144 iterations.

As mentioned, Random Forest was partially chosen due to its relatively fast processing time. This is because a large volume of tests would be performed to fine-tune the model. Features were progressively added, tweaks were made to the number of periods for which period-over-period differences were calculated for each variable, and other modifications were performed. A portion of this process of iteratively improving model performance is described below.

The first test was performed before the calculations performed on monthly, quarterly, and annual data were defined, so only daily and weekly features were used. This iteration resulted in a RMSE of **0.491**, significantly smaller than the standard deviation of the dependent variable.

Next, monthly, quarterly, and annual data were added. Monthly data was imputed from the quarterly and annual datasets through simple forward-filling - all of the months between the months when quarterly or annual values are recorded are set as the prior recorded value. Additionally, the change-from-prior-period-calculations described in Section III were performed at this stage. This resulted in a modest RMSE improvement to **0.475**.

Next, a variety of different imputation methods were added to each of the quarterly and annual data pulls; forward-filling was replaced with linear, quadratic, cubic, order-one spline, akima, order 5 polynomial, and order 7 polynomial interpolation. This improved the RMSE to **0.461.**

Finally, the outcome variable was imputed. Several tests were performed to determine which imputation method best fit the data - the quadratic imputation method seemed to provide the imputation that seemed to best align with the true expected path of default rates through the known quarterly data points.

**Gradient Boosted Decision Trees (using XGBoost)**

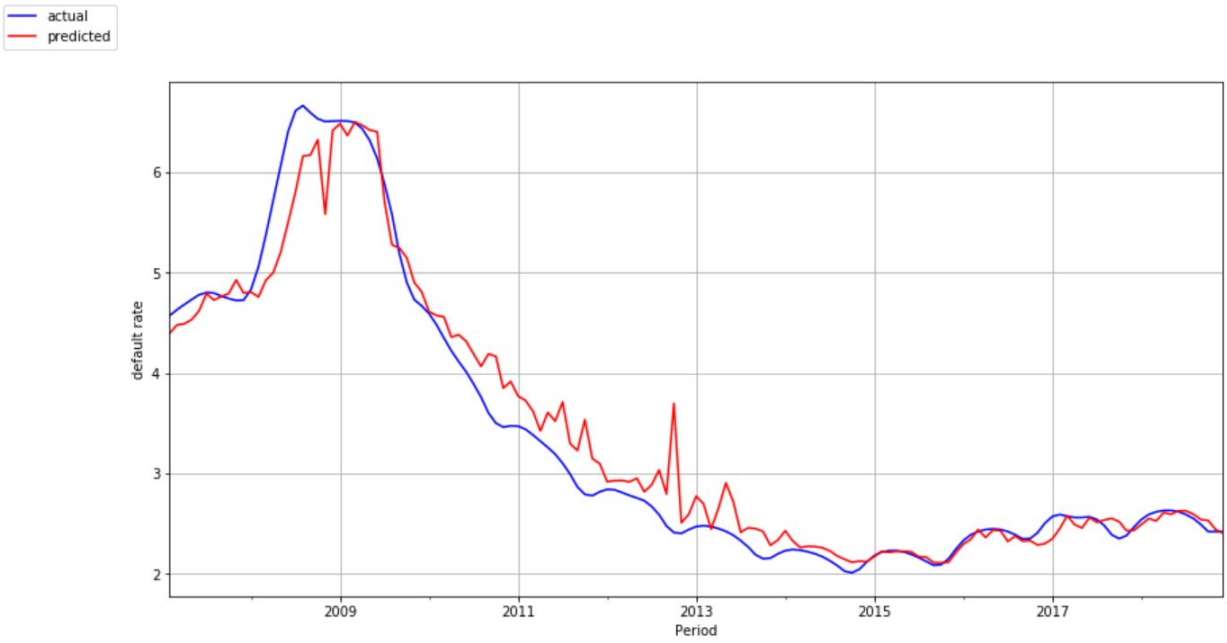
The next step is to use a different machine learning algorithm. XGBoost was observed to consistently perform well in Kaggle competitions, so it was selected as the algorithm to compare to Random Forest.

As with the Random Forest model, Grid Search Cross Validation was used to tune hyperparameters. However, due to XGBoost's greater computing demands, hyperparameter tuning was performed on only the first of the 144 tests - the parameters found to be optimal for the first test were used for each subsequent test. Using the model in this way, a significant improvement over the results generated by Random Forest was observed - the RMSE of the 144 tests performed using XGBoost was **0.284**. This is a small enough error that the results are very likely to be valuable for lenders - predicting the loan default impact of changing macroeconomic conditions with a six-month lag and within 0.284 percentage points is likely to be valuable.

**Results - Predictions**

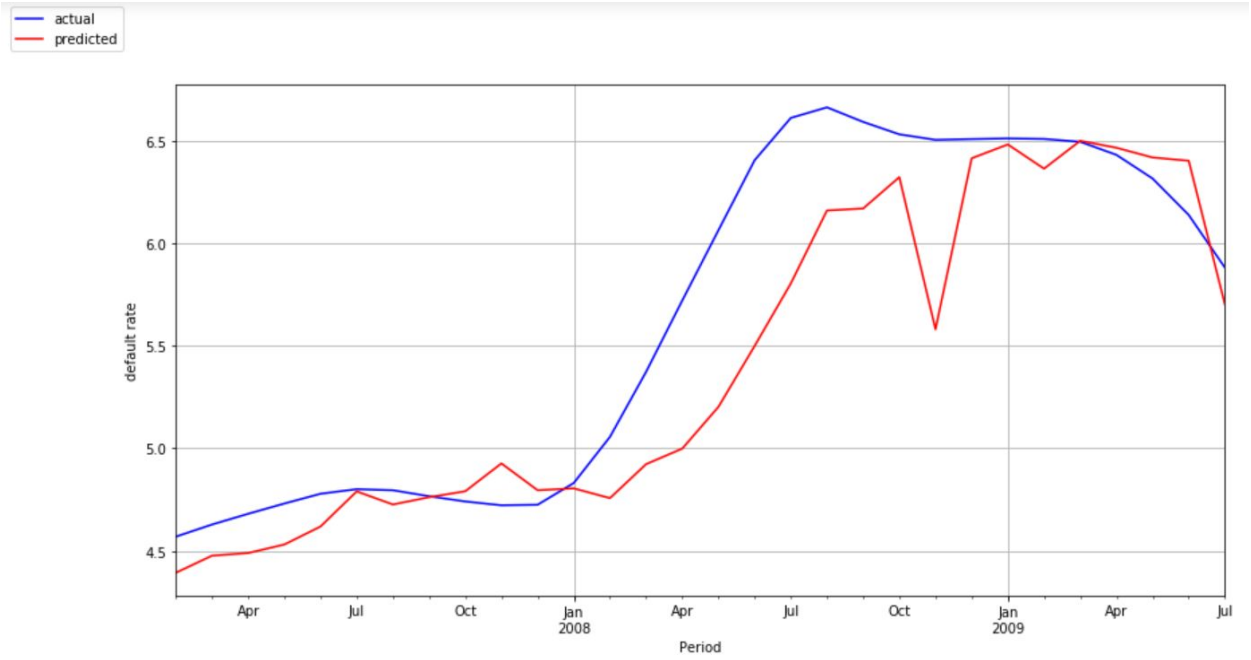
An RMSE of 0.284 is not enough to conclusively determine whether the predictions are of value - it is possible, for example, that the model accurately predicts defaults when they are relatively stable and does not accurately predict large *changes* in the default rate, which would make it of little use to lenders.

The graph below shows the actual vs. predicted values for each month - the blue line shows the actual rate, and the red line the rate predicted six months prior:



From this view, the predicted value appears to significantly lag the actual value when the default rate is changing sharply; when the rate rises in 2009, the predicted rate does not predict that the rise will be as rapid as it is. Similarly when the rate falls throughout the economic recovery, from roughly 2009 through 2015.

Because of the importance of accurate predictions during recessions, the period from February 2008 through July 2009 was evaluated more closely:




While it is true that the actual rise in the default rate is higher than predicted, it is notable that a rise was predicted at all. The recession officially began in December 2007, but was not officially declared for a full year, making December 2007 among the earliest times that one could know that the economy would dramatically falter. At that time, the model would have predicted an increase in the default rate of approximately 100bps in six months. The default rate actually rose closer to 175bps - while the model failed to accurately predict the magnitude of the increase, it did accurately predict that an increase would happen *before* it was known that the country was entering a recession. This is a significant finding, and would have allowed a lender using this model to more rapidly respond to the impending crisis in a way that could materially impact profitability.

**Results - Feature Importance**

Feature importance for each of the 144 runs was aggregated and averaged to evaluate which inputs seem to be of greatest impact. Table 1 in the appendix shows these average feature importances. The values are roughly as expected - values that are known to be accurate measures of in-period economic quality appeared, such as Nominal GDP the change in Real GDP, a measure of the yield curve, a measure of unemployment, and a measure of the inflation rate. Values that predict reductions in business performance, including the year-over-year change in large truck purchases and the year-over-year change in the level of corporate profits, also featured heavily. The level of household debt was found to be extremely significant, with the level of aggregate debt payments appearing as the 5th most important predictor of defaults.

**VI. Conclusions**

Accurately predicting the aggregate level of credit card defaults in six months is a potentially very valuable exercise to lenders seeking direction in their broad underwriting strategies and reactions to macroeconomic risks. This project demonstrates that, using Machine Learning, one can relatively accurately predict the direction in which the credit card default rate will go in six months, potentially at times when others would not predict significant macroeconomic deterioration. Improvements are certainly possible through using other machine learning models, adding or modifying feature engineering methods, and adding to or subtracting from the features initially included. But while more fine-tuning and accuracy can likely be added to the model, it seems predictive enough to be very useful for lending businesses in its current state.

## Appendix

### Table 1 - Top 15 Most Important Features

| Name of Field | Description | Avg. Feat. Imp. |
|---|---|---|
| GDP_interp_linear_GDP | Quarterly Nominal GDP with linear interpolation. | 0.078156 |
| HTRUCKSSAAR_12mth_per_ch | The year-over-year percent change in the number of heavy trucks purchased nationwide. | 0.073253 |
| GDPC1_9mth_val_ch | The change from 9 months prior in quarterly real GDP. | 0.062874 |
| CPIAUCSL_value | The Consumer Price Index for urban consumers, a measure of inflation. | 0.054636 |
| TDSP_value | Household debt service payments as a percent of disposable personal income. | 0.045195 |
| MSPUS_12mth_per_ch | Year-over-year percent change in the average sale price of homes. | 0.041364 |
| T10Y2Y_365day_val_ch_median | Year-over-year change in the median monthly difference between the 10-year and 2-year treasury rates. | 0.035186 |
| HTRUCKSSAAR_12mth_val_ch | The year-over-year absolute change in the number of heavy trucks purchased nationwide. | 0.025162 |
| CP_12mth_per_ch | The year-over-year percent change in quarterly profits. | 0.024532 |
| GDPC1_12mth_val_ch | The change from 12 months prior in quarterly real GDP. | 0.023567 |
| VIXCLS_value_min | The minimum value of the daily Volatility Index during the month. | 0.022132 |
| VIXCLS_value_mean | The mean value of the daily Volatility Index during the month. | 0.021893 |
| UNEMPLOY_9mth_per_ch | The percent change from 9 months prior in the level of unemployment. | 0.019394 |
| FEDFUNDS_9mth_val_ch | The change from 9 months prior in the Effective Federal Funds rate. | 0.019266 |
| CSUSHPINSA_12mth_val_ch | The year-over-year change in the absolute S&P/Case-Shiller Home Price Index. | 0.018480 |