THE GRADUATE CENTER, CITY UNIVERSITY OF NEW YORK

Predicting Economic Recessions using Machine Learning

Working with Data Fundamentals, Spring 2020

Sheridan Kamal

5/18/2020

Using the supervised machine learning methods logistic regression and decision tree classification, we will determine the best method to predict economic recessions. To do this, we will be using train_test_split to create training and testing sets for the hold-out validation method and TimeSeriesSplit for the walk-forward cross validation method. Each machine learning method will be trained on both scaled and unscaled data and will be performed using both validation methods so that there will be four models of each method. The best model using the hold out validation method is the logistic regression model using scaled data with an accuracy of 93.08% and the best model using the walk forward cross validation method is also the logistic regression model using scaled data with an accuracy of 93.56%.

Introduction

Due to the pandemic caused by the novel Coronavirus (COVID-19), you may have noticed that the stock markets have been turbulent as a result of fears that the pandemic will have on different industries and trade. These fears have caused financial markets to tumble at an alarming rate and at times triggering the "circuit breakers" to halt trading. The triggering of these "circuit breakers" seems to occur more frequently with the duration of the pandemic. It is safe to say if the pandemic is prolonged then the negative impacts on the financial markets and the economy as a whole could push us into a recession, which we have not been in for 11 years. It is precisely these thoughts which drove me to decide on my project topic.

In finance, a central idea is “the best indicator of future performance is past performance”, which is why many forecasting models are heavily dependent on historical data. For my project, I will be using the supervised machine learning methods logistic regression and decision tree classification to predict recessions. By using machine learning methods we may be able to get a sense of where the market is heading even if we cannot predict a recession, which will be useful for the Federal Reserve and policy makers to know because if the market is heading towards a recession they may be able to provide a boost to the economy and in turn the financial markets. This exploration is also useful for traders because if it looks like we may be heading towards a recession then they may want to shift their money from equities into other financial instruments.

Related Work

Azhar Iqbal and Kyle Bowman (2018)[1] decided to use various machine learning models and statistical data mining to determine if this would improve recession prediction accuracy. In their paper, they decided to use gradient boosting, random forest, data mining (logit/probit), and benchmark-probit models for their analysis. For each model, they generated a ROC curve for the in-sample and out-sample data as well as the AUC for the in-sample and out-sample data. They determined that the machine learning models (gradient boosting and random forest models) provided more accurate results than the statistical data mining models.

Although I will be creating machine learning models as in the research mentioned above, I will be using logistic regression and decision tree classification models rather than gradient boosting and random forest models as a way to expand on their research. I also am using validation methods and accuracy scores for my model evaluation rather than relying completely on the ROC curve and AUC scores as my cross validation method does not allow for the generation of an ROC curve or an AUC score.

[1] Iqbal, Azhar, and Kyle Bowman. “Can Machine Learning Improve Recession Prediction Accuracy?” *Journal of Applied Economics and Business*, vol. 6, no. 4, Dec. 2018, pp. 16–34.

Data

The dataset that I will be using is monthly data from a combination of economic and financial data. Since the features I need for my dataset are not conveniently included in a downloadable dataset, I downloaded each feature separately and combined them together into one data frame in Python using an inner join. The best place to pull each economic feature from is the FRED (Federal Reserve Economic Data from the Federal Reserve Bank of St. Louis) because I can be sure the data is reliable and can be sure of the quality of the data. To pull the economic data, I will be using Quandl, which also has the added bonus of automatically calculating selected transformations or different data frequencies if I chose to do so. The financial feature is downloadable from Yahoo! Finance so I can download the dataset and create any transformed variables in Excel and import the dataset as a CSV in Python. I will also create Recession labels from a list of start and end dates found from the NBER (National Bureau of Economic Research). After creating the data frame of the features (including the transformed features) and the recession labels, I created a correlation heat map to determine which features will make up my final dataset.

Figure 1. Correlation heat map for final dataset feature selection.

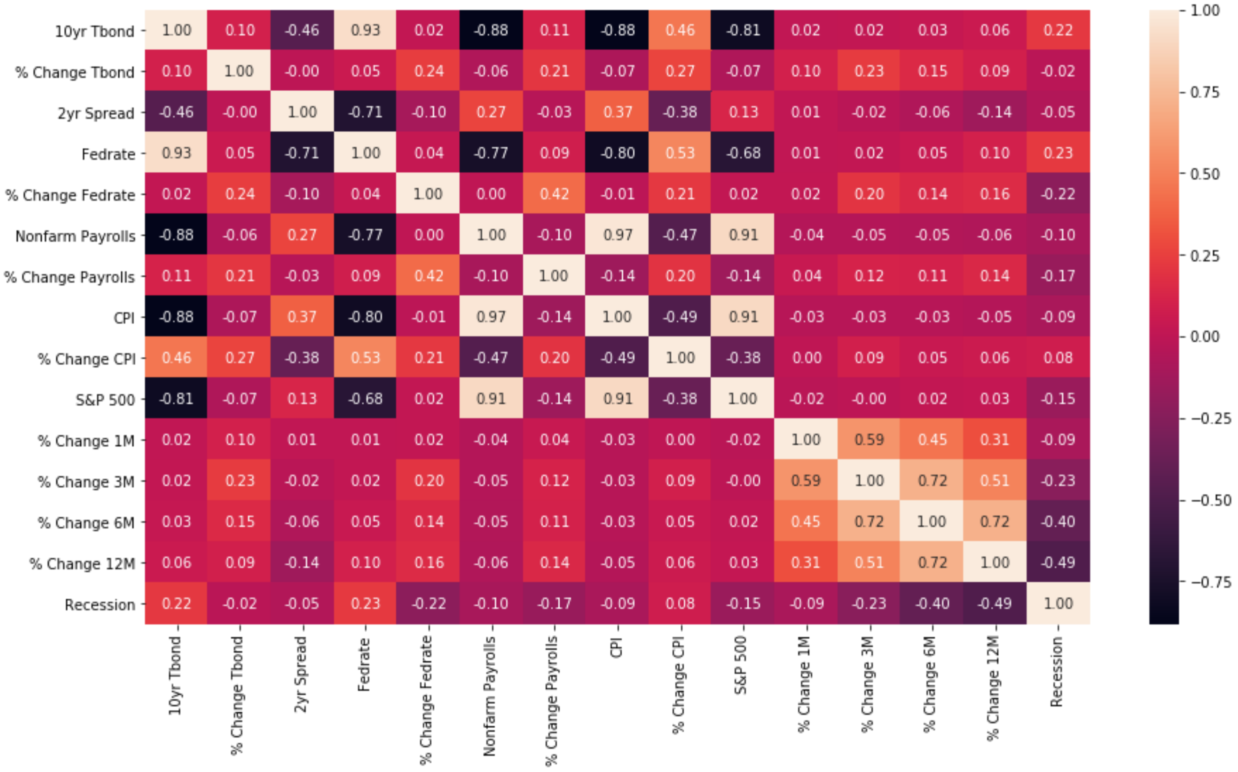
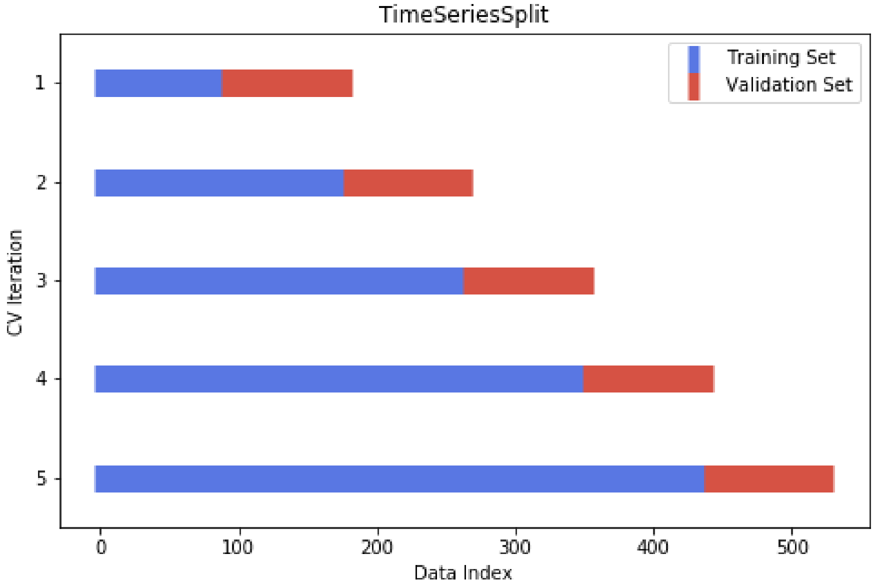


Table 1. Summary statistics table for the chosen 6 features from the original 14 features, 06-01-1976 to 04-01-2020.

Variable	Count	Mean	STD	Min	Max
10yr Bond	527	6.177287	3.25887	0.66	15.32
2yr Spread	527	0.931366	0.912993	-2.13	2.83
Fedrate	527	4.890266	4.039102	0.05	19.1
% Change Payrolls	527	0.000975	0.006252	-0.135484	0.012404
CPI	527	163.281934	57.315716	56.7	259.05
% Change 12M (S&P 500 Index)	527	0.078462	0.152531	-0.593415	0.42489

While train_test_split divides the data into a training and testing set once with fixed training and testing sizes, TimeSeriesSplit divides the data into training and testing sets five times with the size of the training set increasing with each split.

Figure 2. The length of the training set (blue) increases through each iteration while the validation set (red) has a set size through each iteration.

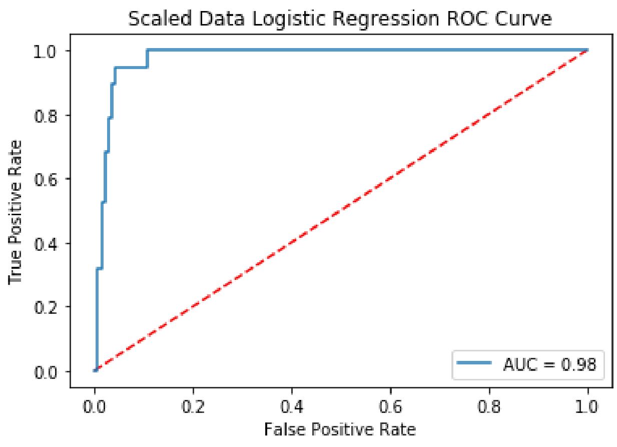


Results

Table 2. The results of the machine learning models using the hold-out validation method.

Logistic Regression – Hold-out Validation		
	Accuracy	ROC AUC
Unscaled	88.05%	0.99
Scaled	93.08%	0.98
Decision Tree Classification – Hold-out Validation		
	Accuracy	ROC AUC
Unscaled	89.94%	0.62
Scaled	91.82%	0.70

Table 2 summarizes the results of the models using the hold-out validation method. As we can see, the best prediction model using the hold-out validation method is the logistic regression model using scaled data because it has a higher accuracy than the logistic regression model with the unscaled data even though it has a higher ROC AUC (by only 0.01) followed by the decision tree classification model using scaled data.



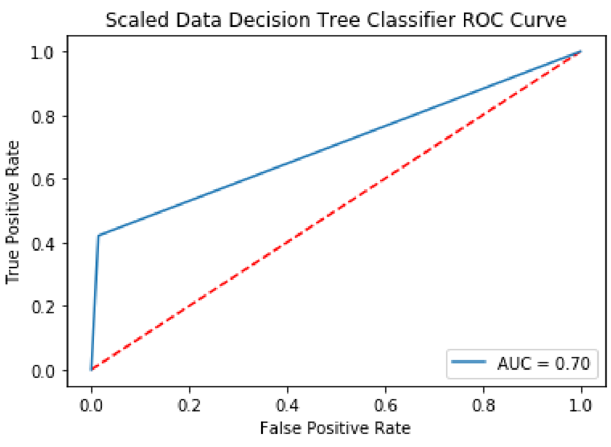


Table 3. The results of the machine learning models using the walk-forward cross validation method.

Logistic Regression – Walk-forward Cross Validation	
	Accuracy
Unscaled	90.11%
Scaled	93.56%
Decision Tree Classification – Walk-forward Cross Validation	
	Accuracy
Unscaled	79.54%
Scaled	92.64%

Table 3 summarizes the results of the models using the walk-forward cross validation method. As you may have noticed, this method is only evaluated based on the accuracy of the model because the evaluation metric is actually the mean of the scores for all five of the iterations and for iteration 2 and 5 the ROC AUC score is unable to be computed so we cannot take the mean of all the ROC AUC scores. Regardless, the accuracy of the models using this method supports the results of the hold-out validation method that shows the best prediction model is the logistic regression model using scaled data followed by the decision tree classification model using scaled data.

Conclusion

The purpose of this paper was to use machine learning methods to predict economic recessions. To do this we used logistic regression models and decision tree classification models and ran both with scaled and unscaled data as well as with a hold-out validation method and a walk-forward cross validation method.

It was determined that the best model to predict economic recessions is the logistic regression model using scaled data and using the walk-forward cross validation method. The second best model to predict economic recessions is the logistic regression model using scaled data and using the hold-out validation method. Through the results of our models, it is safe to say that the best type of model to use to predict economic recessions is a logistic regression model as they had higher accuracies and ROC AUC scores (using the hold-out method) than the decision tree classification models. It is also safe to say that the best validation method to use to predict economic recessions is the walk-forward cross validation method as it improved the accuracy of three out of four (it did not improve the unscaled decision tree classification model) of the machine learning models.