Predicting Recessions: A Machine Learning Approach

By: Shu-Chen Tsao, Kazuki Ueda, and Mark Vandre

Abstract

Business cycles can be either wonderfully creative or terribly destructive depending on the direction they turn. Many people devote their careers to moderating or even predicting these cycles for the purposes of economic stability or arbitrage. This paper seeks to identify an optimal algorithm for predicting recessions. Using data from the Federal Reserve Bank of St. Louis Economic Database, three primary algorithms are tested: logistic regression, k-nearest neighbors, and a support vector machine. From the results of this paper, a linear support vector machine is better for predicting recessions with an accuracy rate of 93.2% overall. These results suggest a potentially useful avenue for further developments in using machine learning to predict economic phenomena.

I. Introduction

Founded in 1920, the National Bureau of Economic Research (NBER) has been classifying recessions and examining business cycles for nearly 100 years.[1] While the NBER engages with a wide variety of macroeconomic research, it is their work on recessions that is probably best known. Other areas of study included national income accounting (done by Simon Kuznets) and demand for money (done by Milton Friedman).[2]

Contrary to popular belief, the NBER's definition of a recession is not a specific number of consecutive quarters of decline in real Gross Domestic Product (GDP). Instead, the NBER classifies a recession as a "significant decline in economic activity spread across the economy, lasting more than a few months," which they say is influenced by factors such as retail sales, production, employment, real GDP and real income.[3] While this distinction may seem minimal to some, this way of stating the categorization process opens the door to some subjectivity rather than being a hard and fast rule that is applied automatically.

It is this element of subjectivity that opens the door of opportunity for a machine learning approach to predict recessions. If the NBER process were a well defined and automatic rule, predicting recessions would be as simple as tracking the financials of leading industries to see declines in real GDP. It is because of the opening to interpretation that leveraging machine learning can be extremely useful in not only understanding better what the NBER will classify as a recession, but also forecasting whether a recession will happen or not.

This project attempts to identify various metrics that are released in a timely manner and can be used to predict the occurrence of a recession in advance of it happening. The methods of prediction used will be machine learning algorithms such as k-nearest neighbors and logistic regression.

The remainder of the paper will be formatted in the following way. Section I has covered an introduction to the topic being examined and has presented the task this paper attempts to accomplish, section II will cover the data used in this paper to include necessary definitions and descriptions, section III will outline the methodology used and the associated challenges and

---

[1] https://nber.org/info.html

[2] https://nber.org/info.html

[3] https://nber.org/cycles/cyclesmain.html

solutions, section IV will present the results to include evaluation metrics, major results, and analysis, and section V will wrap up with a conclusion and discussion of future works.

II. Data Description

The data for this project comes from the Federal Reserve Bank of St. Louis Economic Database (FRED). FRED is a desirable data source due to the ease with which various series from several different government agencies can be exported all at once in a single data file. The data covers every month from January of 1960 to December of 2018 and contains 708 observations. The response variable is, as expected, an indicator variable for NBER recessions. The dependent variables include consumer price index (CPI), civilian unemployment rate, industrial production index (IPI), weekly hours worked in the US manufacturing sector, new housing starts, real disposable personal income, 10-year treasury bond yields, and the business confidence index. Further descriptions of each variable and a table of five-number summaries follow below.

The recession indicator variable is a straightforward variable that equals 1 if the period is a recession and a 0 if the period is not in a recession. This is, of course, the variable we are attempting to predict in future periods. Just as the NBER determines the value of this variable we hope to also determine what the value of the variable will be by predicting the NBER's decision.

The consumer price index is a measure of the general price level calculated by the Bureau of Labor Statistics (BLS). The BLS chooses a base year and sets the CPI in that year equal to 100 with inflation or deflation being percent changes in relation to the base year. To determine the CPI, the BLS creates a representative basket of goods by examining the Consumer Expenditure Survey, and by reviewing the journal entries from thousands of consumers who keep detailed records of purchases they make.[4] This consumption information is then used to weight each category of good or service according to how much it represents overall consumption.

The civilian unemployment rate is also straightforward. It is the percent of individuals in the labor force who are looking for a job, but do not have one yet.[5] This variable is often cited and under close observation during recessions or economic expansions. Generally speaking, when increases in this variable are usually associated with slowing economic activity and possibly recessions.

The IPI is one of several indicators of production activity in the United States. It measures manufacturing and industrial output with a base reference year set to be 2012.[6] Given that it is a measure of productivity for the United States, it stands to reason it would serve as a good predictor of economic slowdown and possibly recessions. Similarly, weekly hours worked in the manufacturing sector is related to IPI in that is shows, from a different view, the amount of productive activity in the United States at any given time.

New housing starts are often considered an important indicator for economic growth. The logic behind this variable as a possible predictor of recessions is that it shows the confidence or ability of consumers in purchasing new homes, which is less likely to happen when consumers expect to earn less. This variable serves as a proxy for consumer financial health. Real disposable personal income also serves as a proxy for consumer financial health. This variable would be a consumer's income less taxes. Together these variables measure a consumer's ability

[4] https://www.bls.gov/cpi/overview.htm

[5] https://www.bls.gov/cps/

[6] https://www.federalreserve.gov/releases/g17/IpNotes.htm

to consume indicating the ability of consumers generally to drive economic growth. If both these variables begin to stagnate or decline, it would seem reasonable to assume a recession may not be far off.

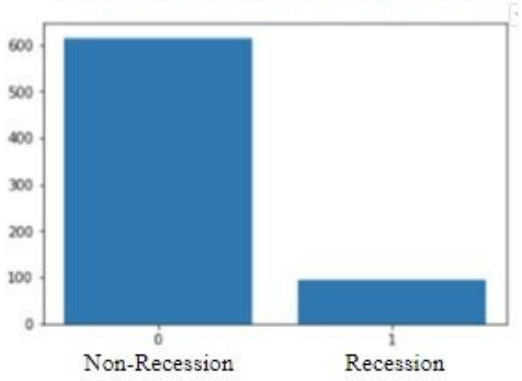
The 10-year treasury bond yield helps to control for changes in the yield curve, which have often proven to be useful indicators for business cycle fluctuations.[7] During times of economic uncertainty, the yields on long-term treasury bonds tend to decline relative to short-term treasury bonds.[8] This is often interpreted as indicating a greater amount of risk in short term investments rather than the standard case of longer term investments being riskier. The drop in short-term confidence can be considered a possible predictor of recessions. Similarly, the business confidence index measures how good or bad producers feel about the economy, which is yet another probable indicator of changes in the business cycle.

Summary statistics are shown in Table 1 and a visual representation of the recession indicator are shown in Figure 1.

Table 1: Summary Statistics for Explanatory Variables

	CPI (index)	Unemployment (%)	Production (index)	Hours Worked
Min	29.37	3.4	22.76	37.3
Q1	50.9	4.9	45.41	40.3
Med	124.3	5.7	63.32	40.7
Q3	187.18	7.1	94.23	41.3
Max	252.79	10.8	110.08	42.4
Mean	126.13	6	67.05	40.76
	Housing Starts	**Personal Income**	**10-year Bond Yield**	**Business Confidence (index)**
Min	478	2269	1.5	95.45
Q1	1177	4242	4.04	99.29
Med	1455	6680	5.81	100.11
Q3	1669	10533	7.75	100.81
Max	2494	14636	15.32	103.75
Mean	1430	7284	6.12	99.97

Figure 1: Recession Indicator Variable



[7] https://www.investopedia.com/terms/i/invertedyieldcurve.asp

[8] https://www.treasury.gov/resource-center/data-chart-center/interest-rates/Pages/yieldmethod.aspx

In order to use these variables to properly predict recessions, we transform CPI, production, and personal income into growth rates. These variables specifically have a general trend of growth over time with momentary dips and spikes. Using the growth rates allows us to remove the positive overall trend and focus on the period-to-period changes in the variables that better indicate business cycles over time. In calculating the growth rates, however, we do lose the first observation so we end up with a data set of 707 observations across all variables.

In addition to transforming the variables, we also split the data into a testing and training set. We randomly specify two-thirds of our data to be training data and one-third of our data to be testing data. Doing this allows us to evaluate our model once it is trained with the training data to simulate predicting out-of-sample observations.

III. Methodology

In order to determine the best method of prediction, three different prediction algorithms were used. The first algorithm tested is a logistic regression from the scikit-learn package, the second algorithm is a k-nearest neighbors algorithm from the scikit-learn package, and the final one tested is a support vector machine from the scikit-learn package. Each of these models are trained using the training data set, and the results are compared in terms of their respective abilities to predict the testing data set.

In order to ensure an optimal prediction model, the first task was to determine how many of the features are relevant to recessions. In order to do this, an ordinary least squares regression was fit on the data with the recession indicator variable as the outcome variable and each of the features as the independent variables. The regression algorithm to do this came from the statsmodel.api package in Python. The results for this process are shown in Table 2 below.

Table 2: OLS Regression for Feature Selection

Variable	Coefficient	Standard Error	t-value	P > |t|	Conf Int Lower	Conf Int Upper
Constant	13.8556	0.968	14.315	0.000	11.955	15.756
CPI	5.0278	3.443	1.460	0.145	-1.732	11.788
Unemployment	-0.0396	0.007	-5.287	0.000	-0.054	-0.025
Production	-9.5170	1.471	-6.469	0.000	-12.405	-6.629
Work Hours	-0.1884	0.018	-10.522	0.000	-0.224	-0.153
Housing Starts	-0.0002	3.25E-05	-7.094	0.000	0.000	0.000
Personal Income	-1.6398	1.319	-1.243	0.214	-4.230	0.951
Long Term Yield	0.0032	0.004	0.740	0.460	-0.005	0.120
Business Confidence	-0.0549	0.010	-5.443	0.000	-0.075	-0.035
F-Statistic: 85.22		Sample: 707				
Prob F-Statistic: 4.56E-98		Adj. R-Squared: 0.494				

As Table 2 shows above, most of the features are strongly associated with recessions. The only exceptions to this are CPI, personal income, and long term yield. Due to the statistical insignificance of these three variables, they will be excluded from the data set. This allows the fitting of a logistic regression using only the most relevant variables. For the second model, we fit a k-nearest neighbors model that minimizes euclidean distance. The k-value chosen was $k = 10$.

For the third model, a support vector machine was trained to predict the recessions. The first step in training this model was to test the model errors over various c-values. The second step was to test the model over each of three different kernel types: linear, polynomial, and rbf. Figures 2 and 3 graphically show the results of the selection process. Figure 2 reveals that a c-value of two or four would be preferable. The c-value chosen in the end was four. For the kernel testing, Figure 3 shows that a linear kernel is clearly the preferred kernel to use.

Figure 2: SVM C-Value Selection

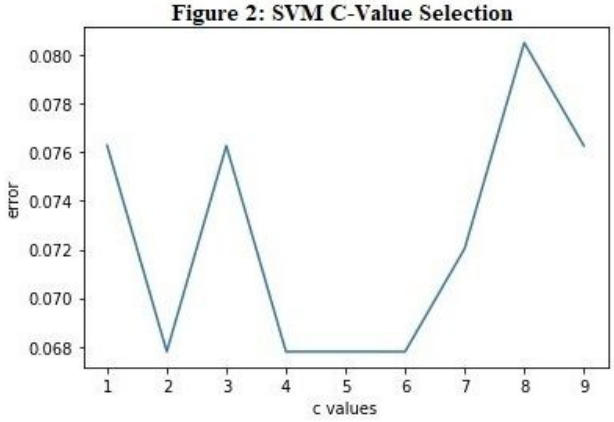
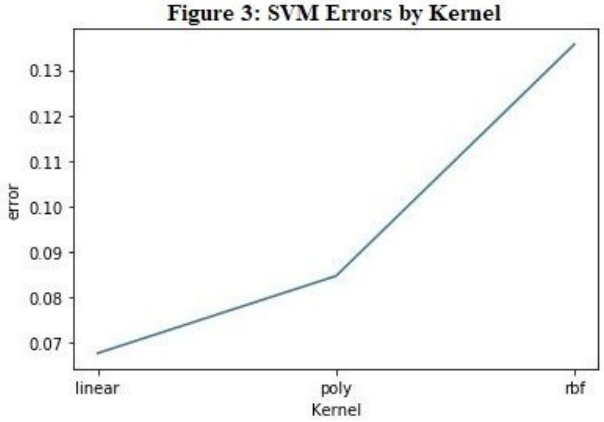


Figure 3: SVM Errors by Kernel



IV. Results

The main results of the models are shown in Tables 3 and 4 below. In terms of overall accuracy, the best model was the linear SVM with a c-value of four. The logistic regression came in as a close second to the SVM model. The KNN model, however, was noticeably worse relative to the other two. In order to determine the most desirable model, it is important to examine more closely the types of errors made by each. Additionally, Figures 4 through 6 below provide visual representation of the prediction results.

The k-nearest neighbors classifier performed the worst in nearly all respects. The model had an overall accuracy of about 87.7%. The recession precision and recall were 100% and 3.33% respectively. The non-recession precision and recall were 87.7% and 100% respectively. The primary issue with the results for this model is how infrequently it correctly predicted a recession. It only predicted one quarter of a recession correctly. Most of the accuracy for this model came from its ability to correctly predict a non-recession. While this is important in some sense, the goal of this paper was to find a model that predicts recessions with reasonable accuracy, which the k-nearest neighbors algorithm clearly did not.

Table 3: Confusion Matrices for Various Model Fits

Logistic Confusion Matrix				KNN Confusion Matrix				SVM Confusion Matrix			
		Recession Prediction				Recession Prediction				Recession Prediction	
		0	1			0	1			0	1
Recesssion Truth	0	195	11	Recesssion Truth	0	206	0	Recesssion Truth	0	197	9
	1	7	23		1	29	1		1	7	23

The next best algorithm was the logistic regression model. This model had an overall accuracy of about 92.4%. The recession precision and recall were 67.6% and 76.7% respectively. The non-recession precision and recall were 96.5% and 94.7% respectively. This model performed reasonably well at predicting recessions. Even though the k-nearest neighbors model committed significantly fewer false positives compared to the logistic regression, the logistic regression predicted significantly more true positives. Additionally, the k-nearest neighbors model performed significantly more false negatives. This essentially shows the k-nearest neighbors model to be correct in nearly the same fashion as a broken clock, which is correct twice a day.

Table 4: Accuracy, Precision, and Recall

Model	Accuracy	Recall 0	Recall 1	Precision 0	Precision 1
Logistic Regression	0.924	0.947	0.767	0.965	0.676
KNN; k = 10	0.877	1	0.033	0.877	1
Linear SVM; C = 3	0.932	0.956	0.767	0.966	0.719

The final, and best performing model, was the linear SVM model. This model had an overall accuracy of 93.2%. The recession precision and recall were 71.9% and 76.7% respectively. The non-recession precision and recall were 96.6% and 95.6% respectively. This model and the logistic regression model performed similarly well. Both models were equal in terms of correctly predicting recessions. The aspect of the SVM result that differentiated it from the logistic model as that if committed fewer false positives. Otherwise, the models performed identically.

Conclusion

As discussed previously, the SVM model performed best within the context examined. It had an overall accuracy of 93.2%, outperforming the logistic regression by 0.8 percentage points and the k-nearest neighbors model by 5.5 percentage points. While the accuracy could be improved upon, this model does a reasonably decent job of predicting recessions. For this reason, such a model could prove useful to policy makers as a way of making predictions about business cycles that entail a degree of confidence based on the historical accuracy of the model.

There is still more work that could be done, however. A useful task for future work would be to identify new and creative variables associated with recessions to make better predictions. Examples could include using Google search term trends, more industry specific macroeconomic variables, or social media data to predict recessions. It would also be useful to see if some combination of novel and traditional variables would improve prediction accuracy.

This paper has demonstrated that recessions can be predicted with a reasonable amount of accuracy. This knowledge could help policy makers to improve economic policies or anticipate needed changes in policy. Several avenues for further research have also been identified. This subject remains rich for further consideration.

Figure 4: Logistic Regression Predictions

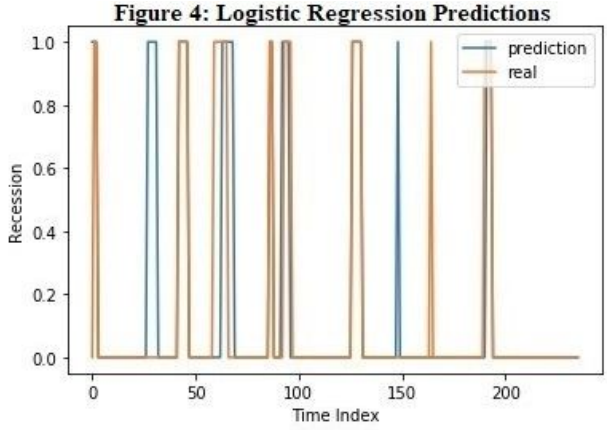


Figure 5: KNN Predictions (k = 10)

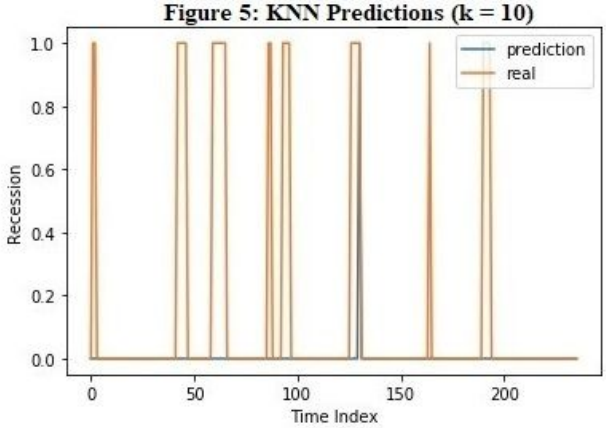
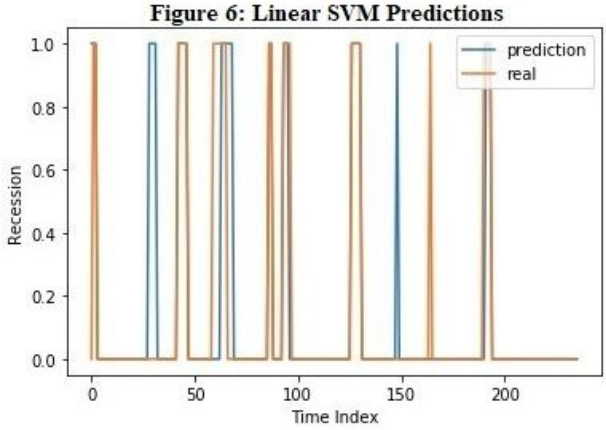


Figure 6: Linear SVM Predictions



References

"Board of Governors of the Federal Reserve System." The Fed - Industrial Production and Capacity Utilization - G.17. Accessed March 21, 2019. https://www.federalreserve.gov/releases/g17/IpNotes.htm.

Chen, James. "Inverted Yield Curve." Investopedia. March 12, 2019. Accessed March 21, 2019. https://www.investopedia.com/terms/i/invertedyieldcurve.asp.

"Consumer Price Index for All Urban Consumers: All Items." FRED. March 12, 2019. Accessed March 21, 2019. https://fred.stlouisfed.org/series/CPIAUCSL.

"CPS News Releases." U.S. Bureau of Labor Statistics. Accessed March 21, 2019. https://www.bls.gov/cps/.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay. *Scikit-learn: Machine Learning in Python,* Journal of Machine Learning Research, 12, 2825-2830 (2011)

John D. Hunter. *Matplotlib: A 2D Graphics Environment, Computing in Science & Engineering*, 9, 90-95 (2007), DOI:10.1109/MCSE.2007.55. https://aip.scitation.org/doi/abs/10.1109/MCSE.2007.55

Stéfan van der Walt, S. Chris Colbert and Gaël Varoquaux. *The NumPy Array: A Structure for Efficient Numerical Computation*, Computing in Science & Engineering, 13, 22-30 (2011), DOI:10.1109/MCSE.2011.37

"U.S. Department of the Treasury." Treasury Yield Curve Methodology. Accessed March 21, 2019. https://www.treasury.gov/resource-center/data-chart-center/interest-rates/Pages/yieldmethod.aspx.

Wes McKinney. *Data Structures for Statistical Computing in Python*, Proceedings of the 9th Python in Science Conference, 51-56 (2010). http://conference.scipy.org/proceedings/scipy2010/mckinney.html

Statement of Contribution

- Shu-Chen Tsao: Performed and coded data acquisition and preparation. Provided critical edits to the final report.
- Kazuki Ueda: Coded the prediction algorithms and results computations. Provided critical edits to the final report.
- Mark Vandre: Prepared the final report. Provided edits to the data preparation code and the prediction algorithms code.