Predicting US Recessions using Machine Learning Algorithms

ECON 425 Final Project

Anshika Sharma(305488635), Noriyasu Kaneda(305635354)

Abstract

The report attempts to study if the optimal algorithm used by Tsao, Ueda and Vandre for predicting recession would reflect the shock to the economy caused by the pandemic, Covid-19. This paper attempts to rework the aforementioned algorithm to improve recession prediction. Data for the period January 1990 to December 2020 from Federal Reserve Bank of St. Louis Economic Database, Yahoo Finance and OECD was used for the study. New variables were identified in order to improve the accuracy of recession prediction. Out of the three machine learning algorithms, it was found that the Logistic regression model was the best model.

Introduction

Predicting business cycles is challenging as well as one of the important activities performed by various economic institutions. Information about the future state of the economy is essential for policy makers, central and commercial banks, private sector and households. Since 1929, the NBER has charted the U.S. business cycles. NBER defines a recession as a “significant decline in activity spread across the economy, lasting more than a few months, visible in industrial production, employment, real income, and wholesale-retail trade.” The NBER uses employment, industrial production, real personal income less transfer payments, and the volume of sales of the manufacturing and trade sectors to define recessions but there are no hard defined metrics.

Early work of forecasting recession includes the studies of Stock & Watson (1989) and Diebold & Rudebusch (1989). These studies did not apply Probit models but alternative methods. Most studies use either a static Probit model to estimate the recession probabilities. Probit models have been a standard tool to predict the business cycle states ever since the study of Estrella & Hardouvelis (1991). A study by Yazdani, 2020 predicts recession in the U.S using the following machine learning models-PROBIT, SVM, Random 7Forest and NNET. RF showed the highest AUC of 95%, precision of 61% and accuracy of 61%. SVM had an accuracy of 90% and the AUC was also 90%, the precision was 59%. The Probit model had an accuracy and AUC of 87% and 89%, respectively. The precision was 53%.

In a 2019 report titled, “Predicting Recessions: A Machine Learning Approach” by Tsao et al, an attempt was made to identify an optimal algorithm for predicting recessions. Data from 1960 to 2018 was acquired from the Federal Reserve Bank of St Louis Economic Database (FRED), and three primary algorithms were tested:

logistic regression, K-nearest neighbours, and support vector machine. Out of the three algorithms, the report concluded that a linear SVM was best suited for predicting recessions.

In the former report, a high accuracy rate (93.2%) of the recession prediction algorithm on the testing data suggested that the model performed well on new data. However, when the existing dataset in the study was updated to include 2020 data, the algorithm was unable to generalise the model to the new data. More specifically, it was seen that the algorithm could not capture the effect of the COVID-19 shock on the economy. Given such limitations, attempts have been made in the following report to rework the existing algorithm to better model the variables which may be used in predicting recessions. A major part of updating the algorithm includes identifying and creating new variables which may further improve the recession prediction. Thus, the report acknowledges and further builds on the work of Tsao, Ueda and Vandre.

The report is structured in the following manner:

- Section I provides a brief description of the data used to build, train and test the model. Sources from which the data was acquired, and necessary definitions of the variables are provided here.
- Section II outlines the methodology used to build and train the model. Challenges encountered in the process, and their solutions are highlighted here.
- Section III summarises the results of the models. This includes confusion matrix, and accuracy, precisions and recall scores for all models.

Finally, the report is concluded with suggestions on how it can be further improved.

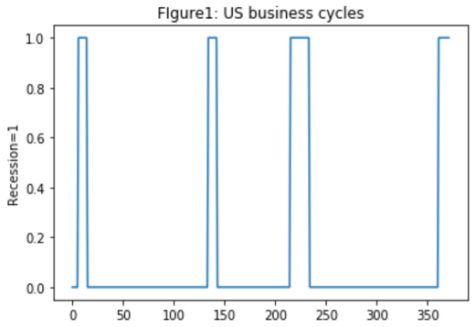
Section I: Data Description

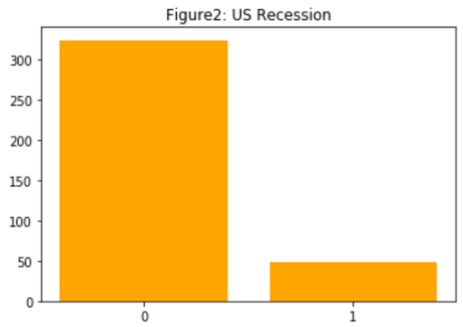
Since our objective is to predict U.S. recessions, our model uses economic and financial indices that are considered to reflect economic cycles from January 1990 to December 2020 (all are converted to monthly data). We obtain the data from the OECD website, the Federal Reserve Bank of St. Louis Economic Database (FRED) and Yahoo Finance.

Target Variable (Explained variable)

The recession was taken as the indicator variable - taking a value of 1 during the recession,"Duration, peak to trough", otherwise taking 0- based on the U.S. business cycles judged by NBER. A value of 1 is thus a recessionary period, while a value of 0 is an expansionary period. For this time series, the recession begins the first day of the period following a peak and ends on the last day of the period of the trough. The

graph below plots the recession variable used for the study, 0 represents “No Recession” and 1 represents “Recession”.





Explanatory variables

Sentiment Index

U.S. Business Confidence Index[1] is an indicator of current and future trends in production, orders, and inventories of finished goods, and can be used to predict turning points in economic activity; the U.S. Consumer Confidence Index[2] provides an indication of future trends in household consumption and savings. For both indexes, a value below 100 indicates a pessimistic attitude toward the future development of the economy, while a value above 100 indicates increasing confidence among businesses and consumers.

Economic Indicators[3][4]

Unemployment rate is one of the most important indicators for monetary policy, and it receives a lot of attention when expanding economic decisions. Also, we use Initial Claims which is the number of new unemployment claims filed by individuals seeking to receive unemployment benefits, and can be an important employment indicator.

In general, production activity is strongly correlated with the business cycle and can be useful in predicting economic downturns. So, we use Industrial Production (IPI), which is an indicator of U.S. production activity. Weekly worked hours in the manufacturing sector is an important indicator not only for employment but also production activity. New housing starts is the number of all housing units in a multifamily building being started, and are an indication of consumers' purchasing attitudes toward new home purchases and are often considered a key indicator for economic growth. Consumer Price Index (CPI) is a Lapsers index of the average monthly change in the price for goods and services paid by urban consumers and a

[1] Business Confidence Index: https://data.oecd.org/leadind/business-confidence-index-bci.htm#indicator-chart
[2] Consumer Confidence Index: https://data.oecd.org/leadind/consumer-confidence-index-cci.htm#indicator-chart
[3] Economic indicators: https://fred.stlouisfed.org. All indicators are seasonally adjusted.
[4] The real disposable personal income variable used in previous work was dropped because of no significance for the recession variables. During the COVID-19 shock, the movement in personal income to economic fluctuations seems to be quite different from the past due to historically large scaled support programs.

Lapsers index, which is one of important indicators for monetary policy, relating to economic cycles.

Financial Indicators[5]

Financial indicators could reflect market participants' views on the economic cycles, and in this case, we use the S&P 500 and the VIX. Also, Spread between 10-Year and 2-Year US Treasury is calculated as the spread between 10-Year and 2-Year US Treasury yield. The long/short spread in the U.S. bond market is considered to be an indicator of the future economic outlook, and a negative long/short spread (an inversion between long-term and short-term bond' yield) is generally considered to be a possible signal of a recession, and is one of the indicators that market participants look at.

To properly predict recessions, we take the log-difference for IPI, CPI and S&P 500. These variables show an upward trend and temporary spikes. Using the growth rate instead of the level data allows us to remove the positive trend and focus on the period-to-period changes in the variables that better indicate the business cycle over time. Also, we take the difference for long/short spread.

Summary of fundamental statistics of explanatory variables are shown in Table 1.

Table1:Fundamental Statistics

	ConsumerCf	BusinessCf	Unemployment	IC	WorkHour	Production	HousingStart	CPI	SP500	VIX	US10Y
count	372.000000	372.000000	372.000000	3.720000e+02	372.000000	372.000000	372.000000	372.000000	372.000000	372.000000	372.000000
mean	100.078265	99.819405	5.914785	3.862515e+05	41.175000	0.001516	1304.255376	0.001956	0.005247	19.480753	0.002144
std	1.576410	1.136221	1.738426	3.167538e+05	0.633259	0.012683	398.972841	0.002605	0.042792	7.775444	0.140790
min	96.262330	95.598437	3.500000	2.114000e+05	38.500000	-0.172286	478.000000	-0.017900	-0.217377	10.130000	-0.718434
25%	98.940184	99.200789	4.675000	3.051875e+05	40.700000	-0.002500	1055.750000	0.000700	-0.018660	13.930000	-0.071355
50%	100.443552	99.902404	5.500000	3.422000e+05	41.300000	0.001700	1288.500000	0.002000	0.009729	17.545000	-0.006674
75%	101.119332	100.573616	6.825000	3.982500e+05	41.700000	0.006401	1568.500000	0.003200	0.030666	23.215000	0.071268
max	103.060286	102.223694	14.800000	5.040250e+06	42.300000	0.075472	2273.000000	0.013700	0.119464	62.640000	0.591818

Section II: Methodology

Splitting the data into training and testing set

The first task in any model building process is to split the data into a training and testing set. Two-thirds of the data was randomly assigned to the training set, and the remaining one-third was added to the testing set manually. This allows for evaluating the model after it has been trained on the training data in order to simulate the prediction of out-of-sample observations and understand how well the model generalises to new data.

[5] S&P500 and VIX: https://finance.yahoo.com, spread between 10- and 2-Year US Treasury: https://fred.stlouisfed.org

The minmax scaler was then applied to the training features. This estimator scaled and translated each feature individually such that it was in the given range on the training set, between zero and one.

Feature selection:Linear Regression

Linear regression was used for identifying optimal features which would best predict recessions. Thus, 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 table below summarises the results of the OLS regression model. It is seen that the variables CPI, SP500 and long/short spread are not statistically significant, and are not strongly associated with recession variables. These variables are thus excluded from the dataset. Only models with strong correlation to recession were selected.

OLS Regression Results

Dep. Variable:	Recession	R-squared:	0.574
Model:	OLS	Adj. R-squared:	0.561
Method:	Least Squares	F-statistic:	44.02
Date:	Sun, 14 Mar 2021	Prob (F-statistic):	5.65e-60
Time:	22:01:58	Log-Likelihood:	37.255
No. Observations:	372	AIC:	-50.51
Df Residuals:	360	BIC:	-3.484
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	20.3914	1.606	12.700	0.000	17.234	23.549
ConsumerConfidence	-0.1303	0.013	-9.657	0.000	-0.157	-0.104
BusinessConfidence	-0.0296	0.016	-1.839	0.067	-0.061	0.002
Unemployment	-0.1412	0.017	-8.354	0.000	-0.174	-0.108
IC	5.299e-07	6.04e-08	8.776	0.000	4.11e-07	6.49e-07
WorkHour	-0.0888	0.034	-2.619	0.009	-0.155	-0.022
Production	4.4114	1.235	3.573	0.000	1.983	6.840
HousingStart	-0.0001	5.83e-05	-2.223	0.027	-0.000	-1.49e-05
CPI	3.2002	4.997	0.640	0.522	-6.627	13.027
SP500	0.2312	0.296	0.780	0.436	-0.352	0.814
VIX	0.0089	0.002	4.319	0.000	0.005	0.013
US10Y	0.1229	0.088	1.389	0.166	-0.051	0.297

Omnibus:	58.079	Durbin-Watson:	0.623
Prob(Omnibus):	0.000	Jarque-Bera (JB):	110.416
Skew:	0.865	Prob(JB):	1.06e-24
Kurtosis:	5.032	Cond. No.	2.17e+08

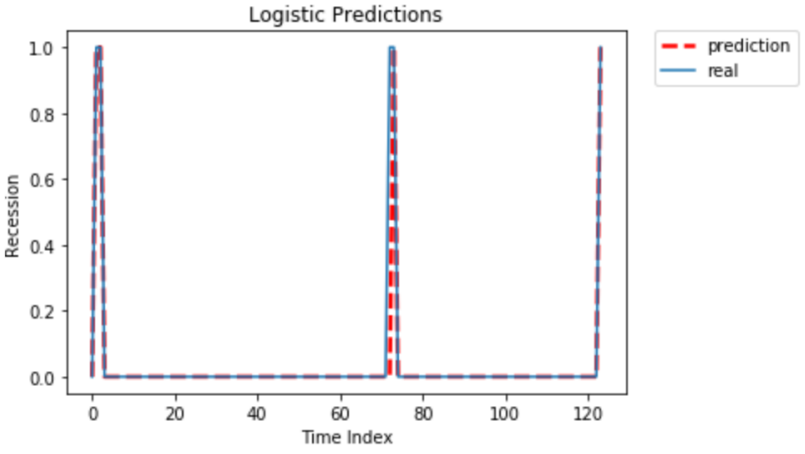
From the correlation matrix, it is seen that for the most part, all of the features are uncorrelated.

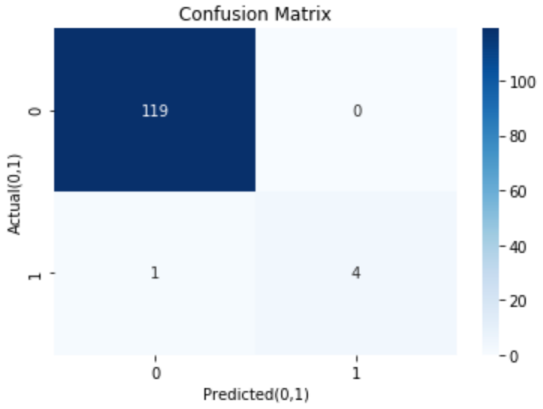
For the purposes of this study, three primary machine learning classification algorithms were tested on the data: **Logistic regression, k-Nearest Neighbors (kNN) and Support Vector Machines (SVM)**. In order to do this, third party modules like the scikit-learn package were used. Each of these models was trained using the training data set, and the results are compared in terms of their respective abilities to predict the testing data set. All three models, along with their performance on the testing set have been discussed below.

Section III: Findings and Analysis

Model 1: Logistic regression

Logistic regression models are extensively used for classification and predictive analysis. They are easy to implement, interpret and very efficient to train. This was true for the recession model as well. The graph below shows that the model fit the data quite well, and the accuracy and precision scores were very high. Logistic regression is especially appropriate when the number of observations is more than the number of features, which holds in case of the regression data. The accuracy is 0.99, and the recession precession and recall rates are 80% and 100%, respectively.



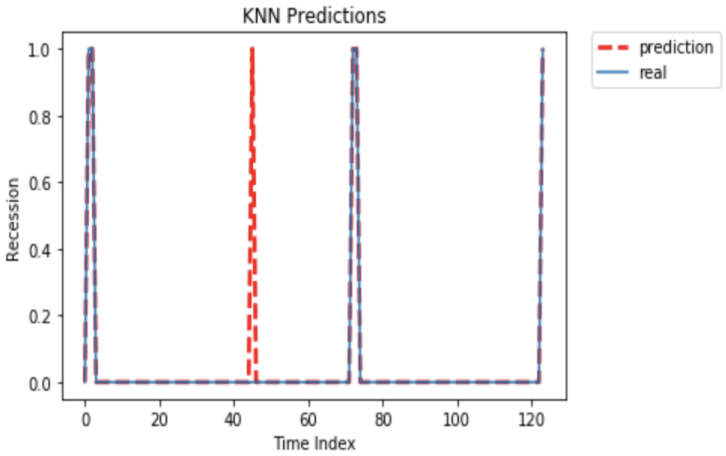


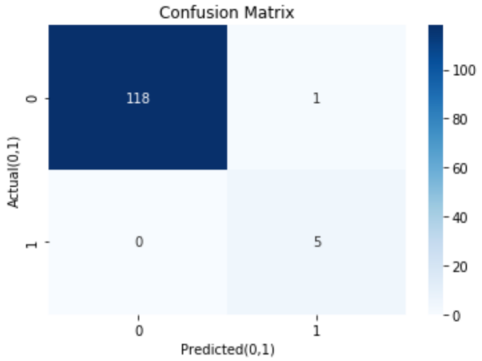

```
Logistic Regression:

Accuracy: 0.9919354838709677
Recall: [1.  0.8]
Precision: [0.99166667 1.        ]
```

Model 2: k Nearest Neighbour (kNN)

The k-NN algorithm was also used. The selection of k, or the number of neighbours, is crucial in training the model. Lower k results in a decision boundary that closely follows the training data. A larger k value leads to a smoother decision boundary which is important for generalization of the model. For the purposes of recession classification, 10 neighbors are taken into account, and K=5 is chosen. From the graph below, the accuracy is 0.99, and the recession precession and recall rates are 100% and 83%, respectively.



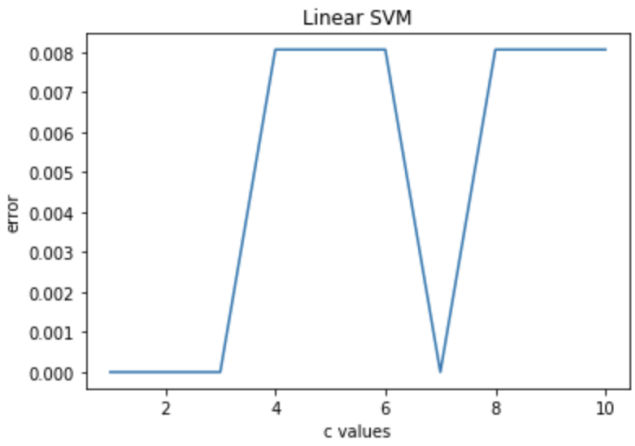


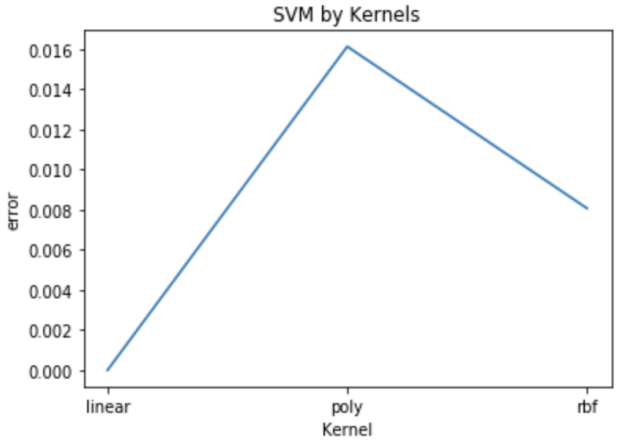

```
KNN predictions:

Accuracy: 0.9919354838709677
Recall: [0.99159664 1.        ]
Precision: [1.         0.83333333]
```

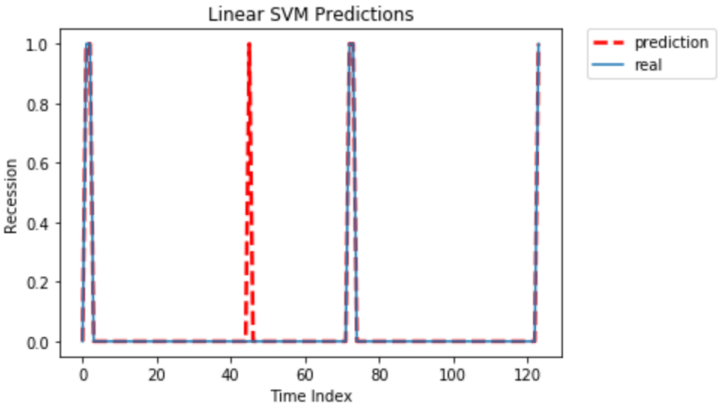
Model 3: Support vector machines (SVM)

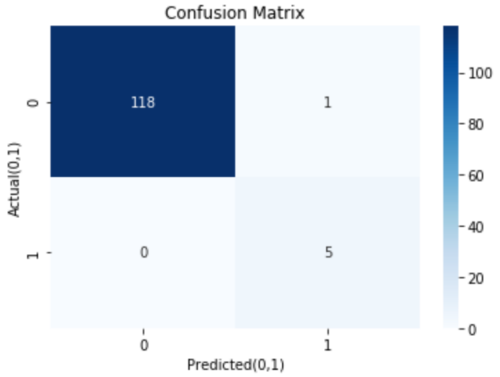
The next model that was used was the Support vector machines (SVM). The previous literature on predicting recession also suggests that after Random forests, this model has proven to be very effective in forecasting recession with a very high accuracy rate. In the first step, the kernel type was selected. From the diagram below, it is evident that the best kernel type is “linear”. The next step is to choose a c value. This was done by developing a for-loop and setting the range of c-value from 1 to 10. C-value of 1, 2, 3 and 7 were chosen since it was the one with minimum errors. In this case, the best model to be selected from SVM is the one with linear kernel and the c value = 7 (other candidates can be chosen as the c value).





The prediction from this model can be shown in the graph of the next page. As seen from the graph and confusion matrix plotted, the SVM model fits the testing data quite well, and it is able to predict the recessions over the sample periods appropriately. The accuracy is 0.99, and the recession precession and recall rates are 100% and 83%, respectively.






```
SVM:

Accuracy: 0.9919354838709677
Recall: [0.99159664 1.        ]
Precision: [1.         0.83333333]
```

Cross validation

As is evident from the analysis section, three models showed high accuracy scores (0.99) equally, and overall the recession precession and recall rates are very high**.** We conduct the K-fold cross-validation(K=5) for the three models using the cross_val_score function from sklearn to assess how the results of our analysis can generalize to an independent data set. As can be seen from the figures below, Logistic regression presents a little bit higher accuracy score than SVM (with linear kernel and c value 7) or kNN.

```
<Cross-validation score> Logistic:0.892, SVM (with linear kernel, c=7):0.882, kNN:0.828
```

Conclusion & Future works

As is evident from the analysis section and the cross validation scores, the three models successfully predict the US recession and the accuracy score is 0,99. **The Logistic regression (model1) performed best within the context examined.**The sample period includes the recession associated with COVID-19. The model maintains a stable forecast accuracy and is quite robust in light of the cross-validation. Therefore, we conclude that machine learning models show a high degree of accuracy in terms of predicting the U.S. recession. The main improvement from the previous reports is as follows: (i) scaling of features,(ii) adding new variables useful for prediction. Although this model is considered to be somewhat useful for economists to forecast US business cycles, it may be useful to use a longer sample period in order to confirm the possibility of overlearning the data.

A future challenge is to identify new variables that are useful for predicting recessions to make better predictions. For example, it is conceivable that financial variables such as long-term interest rates as well as social media data such as Google Trends, could be used to better predict recessions.

Statement of contribution

All the group members (Anshika Sharma and Noriyasu Kaneda) contributed equally towards the project in the terms of collecting data, coding, analysis and drawing conclusions of the results obtained and compiling the report.

References

Diebold, Francis. and Rudebusch, Glenn. (1989) Scoring the Leading Indicators. The Journal of Business, 1989, vol. 62, issue 3, 369-91.

Estrella, Arturo. and Hardouvelis, A, Gikas. (1991) The Term Structure as a Predictor of Real Economic Activity. The journal of finance, Volume46, Issue2 June 1991, 555-576.

Liu, Weiling. and Moench, Emanuel. (2014) What Predicts U.S. Recessions? Federal Reserve Bank of New York Staff Reports, no. 691, https://www.newyorkfed.org/medialibrary/media/research/staff_reports/sr691.pdf

Malik, Lukas. (2020). Using Machine Learning to Predict Recessions Replicating "Machine Learning Prediction of Recessions" (Yazdani, 2020). towards data science, article on September 9, https://towardsdatascience.com/replicating-machine-learning-prediction-of-recessions-yazdani-2020-9b9500131c71

Nyman, Rickard. and Ormerod, Paul. (2020) Understanding the Great Recession Using Machine Learning Algorithms. a keynote presentation at the Bank of England/Federal Reserve conference, https://arxiv.org/ftp/arxiv/papers/2001/2001.02115.pdf

Rudebusch, D, Glenn. (2001) Has a Recession Already Started? FRBSF Economic Letter on October 19, Federal Reserve Bank of San Francisco, https://www.frbsf.org/economic-research/publications/economic-letter/2001/October/has-a-recession-already-started/

Stock, James. and Watson, Mark. (1989) New Indexes of Coincident and Leading Economic. A chapter in NBER Macroeconomics Annual 1989, Volume 4, 1989, 351-409.

Tsao, Shu-Chen., Ueda, Kazuki., and Vandre, Mark. (2019) Predicting Recessions: A Machine Learning Approach

Yazdani, Alireza. (2020) Machine Learning Prediction of Recessions: *An Imbalanced Classification Approach. T*he Journal of Financial Data Science Fall 2020, 2 (4) 21-32, https://doi.org/10.3905/jfds.2020.1.040