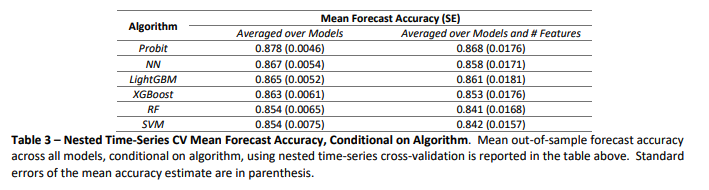
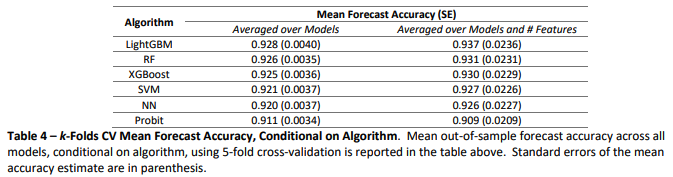
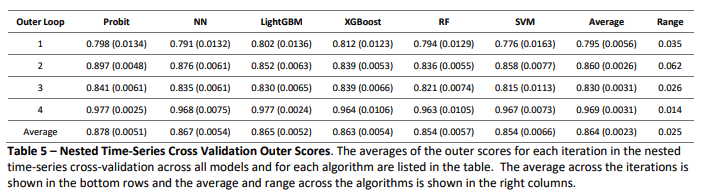
**Machine Learning, the Treasury Yield Curve and Recession Forecasting – Notes:**

* *Consistent with the existing literature we find that, in the time series setting, forecast accuracy estimates derived from k-folds are biased optimistically, and cross-validation strategies which eliminate data “peeking” produce lower, and perhaps more realistic, estimates of forecast accuracy.*
  + **Data Peeking:** when data from the testing dataset are intentionally/unintentionally used in the training dataset
  + In k-folds the data is split into training and test sets k times and each is used as a training and testing set once
    - This can lead to peeking whereby the all the data is used as both training and testing sets
* *That is, while a k-folds cross-validation indicates that the forecast accuracy of tree methods dominates that of neural networks, which in turn dominates that of probit regression, the more conservative cross-validation strategy we propose indicates the exact opposite, and that probit regression should be preferred over machine learning methods, at least in the context of the present problem.*
  + So, the tree methods dominate the neural networks, and the neural networks dominate the probit regression models but if their strategy is used the probit models perform better than all of them
* *This latter result stands in contrast to a growing body of literature demonstrating that machine learning methods outperform many alternative classification algorithms and we discuss some possible reasons for our result.*
  + Their findings contrast from many other findings which indicate that machine learning methods are better
* *We also discuss techniques for conducting statistical inference on machine learning classifiers using Cochrane’s Q and McNemar’s tests; and use the SHapley Additive exPlanations (SHAP) framework to decompose US recession forecasts and analyze feature importance across business cycles*
  + **Cochrane’s Q Test:** [Cochran's Q Test - Statistics How To](https://www.statisticshowto.com/cochrans-q-test/)
  + **McNemar Test:** [McNemar Test Definition, Examples, Calculation - Statistics How To](https://www.statisticshowto.com/mcnemar-test/)
  + **SHAPLEY:** [Welcome to the SHAP documentation — SHAP latest documentation](https://shap.readthedocs.io/en/latest/)
* *It is well documented that an inverted Treasury yield curve is a strong signal of recession in the United States. In May 2018, the term spread between the 3-month Treasury bill discount and 10-year on-the-run yield-to-maturity – one of several common measures of the yield curve’s slope - narrowed to less than one percentage point for the first time since the Great Recession. The spread continued to decline and in May 2019 the yield curve inverted; that is to say, the 10-year yield-to-maturity fell below level of the 3-month bill discount. At time of writing (October 2019), the yield curve’s slope is 11 basis points.*
* Past analysis usually uses a 1-4 quarter horizon
* Probit models worked really well during the Volker era
* *Virtually every study of this topic following those first published in the 1990’s has used some version of the probit method to investigate the problem, each extending the original studies in a new direction*
* *In addressing the shortcomings of the standard probit method, all previous efforts essentially contend with one simple fact, which is that the probit framework is rigid. That rigidity can be credited to the probit link function. In order to classify data, the probit method attempts to draw a hyperplane through the feature space by minimizing a cost function. In doing so, the optimization must trade off goodness of fit in some areas of the feature space (by fitting the hyperplane separator well in those regions) against poorness of fit in other areas.*
* *In addressing the shortcomings of the standard probit method, all previous efforts essentially contend with one simple fact, which is that the probit framework is rigid. That rigidity can be credited to the probit link function. In order to classify data, the probit method attempts to draw a hyperplane through the feature space by minimizing a cost function. In doing so, the optimization must trade off goodness of fit in some areas of the feature space (by fitting the hyperplane separator well in those regions) against poorness of fit in other areas.*
* Machine learning models are popular choices for extending traditional probit models since they are more flexible and generally more accurate when used correctly
* *that said, machine learning methods have the tendency to over-fit data, unless controlled sufficiently via a set of hyperparameters, which is not a problem that arises when probit methods are applied. Thus, the flexibility of machine learning methods also presents the researcher with new difficulties, which are managed via a bias-variance tradeoff.*
* *Specifically we study artificial neural networks, support-vector machines and the 2 Random Forest, XGBoost and LightGBM algorithms (the latter three being presently some of the most popular tree ensemble methods in use). In service of this goal, we propose a novel strategy for conducting cross-validation on classifiers trained with macro/financial panel data of monthly frequency and compare the results to those obtained from standard k-folds cross-validation. We also discuss techniques for conducting statistical inference on machine learning classifiers using Cochrane’s Q and McNemar’s tests; and use the SHapley Additive exPlanations (SHAP) framework of Lundberg and Lee (2017) to decompose US recession forecasts and analyze feature importance across business cycles.*
* *In a slight preview of our results we find that, consistent with established results, forecast accuracy estimates derived from k-folds cross-validation in time series settings are biased optimistically.*
* *For longer horizons, however, they conclude that “the slope of the yield curve emerges as the clear individual choice and typically performs better by itself out-of-sample than in conjunction with other variables.”*
* Over time, researchers continuously add more and more credit related inputs to the models and this generally improves their predictive ability
* *In contrast to Holopainen and Sarlin’s work, in this paper we propose a very conservative nested time-series cross-validation procedure and explore strategies for contending with the time-series properties of macro panel data containing multiple structure breaks.*
* *For a measure of terms spreads, or yield curve slope (Slope hereafter), we use the monthly average of the 10-year Treasury spot yield of Gurkaynak, Sack, and Wright (2006) less the monthly average of the 3-month Treasury bill discount in the Federal Reserve’s H.15 series.*
* Use the log returns of the sp500 as the target variable
* *In addition to bagging, the Random Forest method employs another process called feature bagging to further decorrelate the trees. Feature bagging is the selection of a random subset without replacement of features to be considered in the fitting process for each decision tree. That is, only bagged features are considered at each split in the tree.*
* **Gini Impurity:**
  + Gini impurity is a specific type of loss function used in decision trees.
  + It measures how often a randomly chosen data point would be misclassified. Lower Gini impurity means better classification.
  + Each time the tree makes a split (asks a question), it aims to reduce Gini impurity, making the classification task clearer.
* **Binary Cross-Entropy Loss:**
  + Binary Classification:
    - Used when there are two classes (0 and 1).
  + Definition:
    - For a single example, the binary cross-entropy loss is calculated as follows: L(y,y^)=−(y⋅log(y^)+(1−y)⋅log(1−y^))L(y,y^)=−(y⋅log(y^)+(1−y)⋅log(1−y^))
    - Here, yy is the true label (0 or 1), and y^y^ is the predicted probability of belonging to class 1.
  + Interpretation:
    - The loss penalizes the model more if the predicted probability (y^y^) deviates from the true label (yy).
* *The data set is serially correlated, which violates the i.i.d. assumption required to use many of the cross-validation methods most popular in the machine learning literature, such as k-folds, leave-p-out, etc. Furthermore, because the data is time-series and not cross-sectional in nature, attempting k-folds cross validation on the data set would result in “data peeking” and overly optimistic estimations of forecast performance.*
* Page 17 outlines many of the issue with conducting the analysis
* *Our strategy for contending with these features of the data is to implement a 4-fold nested time-series (NTS) cross-validation to estimate out-of-sample forecast performance for each combination of the 106 models and 6 algorithms under study. Originally designed by Varma and Simon (2006) for use on small datasets to address the unique difficulties they pose, according to Raschka (2018) nested crossvalidation “shows a low bias in practice where reserving data for independent test sets is not feasible.” We augment the standard nested cross-validation strategy to incorporate several features that make it more amenable to conducting time-series analysis on the small macro/financial panel dataset under study in this paper, and to address the serial correlation, indicator imbalance and structural breaks that 18 are present (hence the name nested time-series cross-validation). In particular, we overlay standard nested cross validation with an expanding window, so as to respect the time ordering of the data and prevent future leakage. We add one wrinkle to this feature in that, rather than forward chaining the window over a single data point or a fixed-size block of data points, we forward chain the outer loop of the NTS cross validation over business cycles*



* *The worst performing models generally include the effective federal funds rate.20 Even when combined with other features that together perform well, the federal funds rate appears to diminish the power of a model, particularly if estimated by probit regression.*
* *Accuracy increases generally in the number of features used, up to 7 features, then declines. It seems that 4 features is better than 1, but it is not clear that 7 is necessarily better than 4, after taking into account classifier variance. This topic will be addressed in more detail shortly.*
* *The accuracy estimates from NTS cross-validation, conditional on algorithm and number of features have much higher variance than k-folds cross-validation. This calls into question whether the k-folds cross-validated classifiers, which have been allowed to “peek” at future data before forecasting a sample that is held out, are truly able to forecast recession or have merely memorized the joint time-series of recession and the features.*





* Leading recession indicators seems to work well and add to the overall analysis
* Seems that for larger datasets machine learning models are better but that small datasets are still best reserved for probit models
* *Complicating our situation further is the fact that the major cross-validation methodologies (such as k-folds, leave-p-out, etc.) commonly used to estimate models using machine learning algorithms and manage the bias-variance tradeoff are also built for application on i.i.d. data, rather than serially correlated time-series data. Though expanding and sliding windows cross-validation are available for the latter type of data, and are likely work well on time-series of daily or higher frequency and many years or decades in length, in cases such as ours these methods presents difficulties.*
* Page 55 explains more about business cycles and how they affect the results
* *As term spreads flatten, as the credit and financial conditions tighten or macro conditions deteriorate… recession probabilities rise across all of our models and methods. We also showed that the machine learning methods are able to identify characteristics of the empirical distribution of our features over recession that the probit method cannot.*

**Modeling and Predicting U.S. Recessions Using Machine Learning Techniques**

* *The results strongly support the application of machine learning over more standard econometric techniques in the area of recession prediction. More specifically, the analysis indicates that penalized Logit regression models, k-nn method, and Bayesian generalized linear models largely outperform ‘original’ Logit/Probit models in the prediction of U.S. recessions, as they achieve higher predictive accuracy across long, medium and short-term forecast horizons.*
* It is very hard to predict recessions since the causes are often different and the economy changes over time
* *The main findings can be summarized as follows. The predictive power of the yield curve over a 12-month ahead horizon is reconfirmed.*
* *Moreover, the analysis reveals that the yield curve remains one of the most robust and important predictors of upcoming recessions, as it appears to be a significant predictor for all different forecast horizons employed in the analysis. Having said that, the inclusion of additional variables alongside the term spread can enhance the underlying predictive accuracy. More importantly, however, the implementation of various machine learning techniques increases accuracy even further. Penalized likelihood binary Logit models (LASSO and Elastic Net) and the k-nn method seem to be the most consistent across short, medium, and long-term forecast horizons, followed by the regularized discriminant analysis and Bayesian techniques, as well as Random Forest, especially for shorter-term forecasts. Overall, there is strong evidence supporting the implementation of machine learning models for economic activity prediction. To our knowledge, this is the first comprehensive, comparative study of machine learning techniques in the area of economic recession forecasting.*
* *They find that one of the most successful models, is that containing the yield curve slope or term spread (10-year Treasury yield and the 3-month Treasury bill spread) as the sole explanatory power for forecasting a recession four quarters ahead. The estimated probability of recession from the specific model is 10% when the spread averages 0.76% over the quarter, and rises to 90% when the spread averages -2.40%. Apart from the yield curve slope, Estrella and Mishkin (1998) include interest rates, equity market indices, monetary aggregates, macroeconomic and leading indicator variables as inputs, but conclude that the yield curve slope is the single-most powerful predictor of U.S. recessions in the medium-term.*
* *Apart from the yield curve slope, Estrella and Mishkin (1998) include interest rates, equity market indices, monetary aggregates, macroeconomic and leading indicator variables as inputs, but conclude that the yield curve slope is the single-most powerful predictor of U.S. recessions in the medium-term.*
* *Dueker (1997), for example, includes a lagged recession parameter, i.e. a lagged dependent variable, in the specification of the Probit model and finds that the inclusion of the lagged dependent variable appears to complement the explanatory power of the slope of the yield curve*
* *Moreover, he extends the dynamic Probit model by allowing for some time variation in the structure of the model. More specifically, the coefficients are allowed to change values based on a latent binary state variable that follows a Markov process; in this case, the coefficients take either of the two values, depending on the value of the state variable, altering the shock necessary to induce a recession. Dueker (1997) finds that the Markov process assists the forecast, as it is able to predict the length of the recession, once it is ongoing. He concludes that it is important to incorporate dynamic serial correlation in the Probit model, while time variation in the coefficients is not particularly significant for short-term horizons.*
* Used a confusion matrix

Predicted Class

| Positive | Negative |

Actual Class |----------|----------|

Positive | TP | FN |

Negative | FP | TN |

* *More specifically, the set of predictive variables contains 56 macroeconomic and financial market related indicators, the majority of which are widely-followed by both policy makers and practitioners, and have been used in the existing literature for predictability of recession. Apart from the ’usual suspects’, i.e. predictive variables such as the yield curve, the default spread, financial market and economic activity related indicators, a number of less studied indicators, such as the change in the ratio of residential investment to GDP, the change in the ratio of short-term household liabilities to disposable personal income, heavy truck sales and financial conditions indices are included in the analysis*
* *In general, the forecasting variables are representative of categories related to output and income, the labor market, the housing market, orders and inventories, money and credit, interest rates, prices, and the financial markets.*
* ***Natural cubic spline interpolation*** *is a method for constructing a smooth curve (spline) that passes through a set of given data points. The cubic spline is a piecewise-defined curve that consists of cubic polynomials in each interval between consecutive data points. The term "natural" refers to the boundary conditions imposed on the spline, which ensure a smooth and natural behavior at the endpoints of the curve.*
* Since the data is so unbalanced, there is a good reason to increase the threshold above 50% for deciding whether the economy is in recession or not
* They think that using a threshold of 50% is too high since the data is so unbalanced that they settled on 33% for the threshold
* *It is evident from the preceding section that the implementation of numerous machine learning techniques in U.S. recession forecasting substantially improves the underlying predictive ability and, thus accuracy, relative to more standard and widely-used econometric models (Logit, Probit models). In addition, it is a fact that there is a certain degree of divergence in the forecasting ability of the underlying machine learning techniques. Having said that, it would be rather cumbersome in practice for both economists and financial analysts to apply numerous machine learning techniques relative to more established econometric techniques, as they are computationally intensive and time consuming.*
* *A practical way to address these issues would be to use a machine learning technique that can perform effective variable selection and use the selected predictive variables as inputs in a more standard econometric framework. As a practical example for the design of an early warning recession framework, the LASSO model is used for out-of-sample variable selection over the long-term forecast horizon (12-months)*
* *The main findings of this study can be summarized as follows; first, consistent with the existing literature, the ability of the yield curve to act as an early warning system for predicting U.S. recessions is reconfirmed. More specifically, the yield curve remains a consistent and reliable recession predictor at the 12-month forecast horizon. It is thus a rational and pragmatic choice to treat forecasts generated by the specific predictor as benchmark or reference models. Having said that the addition of alternative macroeconomic and financial market related predictors enhances recession predictability compared to forecasts based solely on the term spread*
* *Third, as anticipated, there is a certain degree of divergence in the predictive accuracy of the underlying machine learning models. Model accuracy is not consistent across forecast horizons, some techniques, however, exhibit more consistent relative performance. Penalized likelihood binary models (LASSO, Elastic Net, Ridge), the k-nn method, and the Bayes GLM technique, seem to generate the most reliable forecasts across different forecast horizons, while the Random Forest model is strong in medium and short-term horizons.*

**FORECASTING ECONOMIC RECESSIONS USING MACHINE LEARNING: AN EMPIRICAL STUDY IN SIX COUNTRIES**

* *This paper proposes a methodology for forecasting economic recessions using Machine Learning algorithms. Among the methods examined are Artificial Neural Networks (ANN), Support Vector Machines (SVM) and Random Forests. The datasets analysed refer to six countries (Australia, Germany, Japan, Mexico, UK, USA) and cover a time span of more than 40 years. All methods are compared against each other in terms of six evaluation metrics on their out-of-sample performance. In contrast to most similar empirical studies, the methodology developed focuses on the timepoints of the last four quarters before a recession begins rather than on those of a recession per se. It has been found that the SVM method tends to outperform the others, as it classified correctly at least 75% of the pre-recessionary periods for half of the countries, with mean overall classification accuracy around 90% in these cases. Moreover, for all the countries under study, the traditional Logit and Probit models are always inferior to at least one Machine Learning-based model. Additionally, it turns out that macroeconomic variables representing a kind of debt – such as, household debt – are most frequently considered as important across the six datasets, in terms of the Mean Decrease Gini measure.*
* *business cycles are not caused by exogenous shocks, as the popular opinion holds, but they are created endogenously due to the stochastic behavior of economic systems.*
* *One of the biggest challenges in Economics is the accurate prediction of some measures of interest, such as the Gross Domestic Product (GDP), for various reasons, e.g., policy making, financial speculation, etc. However, long-term predictions in chaotic systems are impossible, due to the inherent property of systems’ sensitive dependence on initial conditions.*
* *Since it is not clear whether economic systems are truly chaotic systems or not, the objective of the present paper is to propose a Machine Learning-based methodology, the goal of which is to provide reliable short-term predictions of economic recessions. The methodology proposed focuses on the signs that precede significant downturns of economic activity. In other words, our goal is to capture the dynamics of some important macroeconomic factors before a recession occurs, in order to use these signs as indicators for upcoming recessions.*
* *According to the International Monetary Fund, there is no official definition of the term economic recession. However, a practical definition that seems to be widely accepted is the following: “Recession is a period of two consecutive quarters of decline in a country’s real GDP” (Claessens & Kose, 2009, p. 52). This definition is also accepted in the framework of this paper. A special case of recessions are the so-called depressions. A depression is a severe and long-lasting recession (Hall & Lieberman, 2013, p. 125). Although there is no general consensus regarding the magnitude and the duration that labels a recession as depression, most analysts make this distinction if the decline in real GDP exceeds 10% (Claessens & Kose, 2009, p. 53). Generally speaking, depressions are very rare, and, thus, we do not study them separately in the models presented below.*
* There is no real consensus about what actually causes recessions and there are various theories
* *It is well known that, during recessionary periods, a characteristic situation in the economy is low profitability of the firms. Adam Smith (1723–1790), a social philosopher considered to be the father of Classical Economics, mentions three reasons that cause low profitability: (a) competition in the labour market, which leads to higher wages, and, therefore, decreased profits; (b) competition in the capital market, which leads to higher prices of capital goods, and (c) competition in the consumer goods market, which forces capitalists to sell at cheaper prices, which also diminishes profits (Smith, 1776 [1977], pp. 129, 469). These reasons are linked to macroeconomic variables like unemployment or inflation, which may be found useful for the models of this paper.*
* They decided to include demographic variables as well for predicting recessions
* *A mathematical explanation for this concept can be found in Tsoulfidis (2010, pp. 119-120). According to his analysis, the absolute over-accumulation of capital happens when the elasticity of profit rate with respect to capital (denoted as ) is –1. This suggests that er,c is likely to be a good predictor of economic crises and – therefore – recessions.*
* Use money supply as a potential input for predicting recession
* *Irving Fisher (1867–1947) concluded that the two factors with prevailing impact on the evolution of business cycles are over-indebtedness and deflation (Fisher, 1933, p. 341). According to Fisher, debt and price levels are primary variables when studying business cycles, in the sense that other similarly important variables are affected by them.*
* *Christiansen (2013) used a Probit model in order to examine the forecasting ability of yield curve spreads in simultaneous recessions of six countries (Australia, Canada, Germany, Japan, United Kingdom and United States). She considers a recession as ‘simultaneous’ if it occurs in at least half of the countries studied. She found that, at short horizons, only the German yield spread was significant in explaining future simultaneous recessions, but, at long horizons, both U.S. and German spreads were such.*
* Used data from the OECD and only collected data from OECD countries so that they could do that
* Created their own new algorithm called an average tree algorithm
* *The following metrics were used for evaluating each method: 1) Classification Accuracy, 2) Sensitivity, 3) Precision, 4) Specificity, 5) False Alarm and 6) F1 -Score. For each method, the results from the K-fold cross-validation are summarized by calculating a weighted average on each metric. The weights are proportional to test set sizes. It is expected that the K-th test set is usually of smaller size than the previous K-1 test sets, because the division of observation number by K is likely to give a non-zero remainder. The last test set may show, for example, 100% Classification Accuracy, since it consists of only one (correctly classified) observation. But it is obvious that such results do not have the same significance as those from the other K-1 test sets. Thus, the impact of the last test set should be shrunk proportionally to the number of cases used for evaluating a classifier on this set. This detail is taken care of by using the weighted average mentioned above.*
* *the ten methods evaluated are the following28: Average Trees (both variants of the algorithm; i.e., K-fold sampling and random sampling with replacement), Decision Trees, Random Forests, Logit, Probit, k-NN, Boosted Regression Trees (Logistic Regression model), Support Vector Machines, and Artificial Neural Networks (single layer, feed-forward).*
* Used principal component analysis to reduce the number of inputs from over 100 to around 10
* Exports seems to be the biggest driver of recessions in Mexico
* All the methods seemed to be most accurate for US-data
* K-fold validation didn’t perform better than other validation methods for the average trees algorithm
* Somewhat conflicting results since some of the algorithms performed better on different countries than others
* *Especially in our problem, which holds that class “0” appears much more often than class “1” due to the rarity of economic recessions, a classifier could possibly achieve more than 80% Classification Accuracy just by correctly predicting only class “0”.*
* In the author’s opinion, a model can be considered good in the framework of this paper, if its *Classification Accuracy is at least 85%, its Sensitivity and Precision at least 70%, and its False Alarm at most 10%. Of course, this is a reasonable, albeit subjective, choice.*
* Random forests performed best in-sample bur inline with pretty much everyone else out of sample
* *Another question was whether a set of general rules that lead to recessions exists, at least for the countries studied. From the findings of this paper the answer is no. Important variables differ among countries, so pre-recessionary conditions are necessarily different according to our findings. We can argue that there are some country-specific macroeconomic conditions often preceding recessions – this is the output the Average Trees algorithm provides – but we cannot support the opinion that national economies operate in a way that consistently follows such simple rules. At a second glance, we could say that these rules – if they exist – need to be represented by a more complex concept than a tree-like one.*
* *Fisher’s debt-deflation theory. Can we say that this is the theory which best corresponds to reality? The answer is maybe. While we could agree with Fisher that factors related to debt have an important impact on the evolution of business cycles44, we completely miss the ‘deflation’ part in our analysis. In the context of this paper, deflation means GrowCPI < 0. Apart from the USA, there is no other country for which an inflation-related variable is characterized as important. Therefore, we cannot say that the debt-deflation theory is truly verified through this paper. Alternatively, we could say that it will be a rather beneficial decision to include debt-related variables in similar future works*
* *One innovation of this paper is focusing on the short period before a recession begins and not on the recession per se. The advantage of this choice is that the predictions resulting from it refer to potentially pre-recessionary periods. This means it is very likely that if a policymaker takes them into account, they have the time to design a proper policy and, ultimately, intervene in the economy.*
* *Probably the most important innovation of this paper is the Average Trees algorithm. It generally achieved better out-of-sample performance than classic Decision Trees, while its good interpretability remained unchanged*.