Ensemble of Machine Learning Algorithms for Economic Recession Detection (October 2018)

First R. do Ó

***Abstract*— This work intends to solve one of the most desired questions in economics, predicting when an economic recession will happen. The study uses the Unites States of America economy to develop its propositions, since it remains the most powerful economy in the world. It intends to solve this problem by using several macroeconomic indicators and applying Machine Learning algorithms to them. Three algorithms were used, the linear Logistic Regression and two nonlinear algorithms, the Random Forest Classifier and the XGBoost. Also used was an equally weighted average of these three algorithms to improve on their results. This work proved that with these models, with some transformations to the macroeconomic and recession signals, it is possible to predict recessions up to eighteen months in advance and with high accuracy. The best predictions were obtained using the Models' Average, where above 50% of probability the results show residual amounts of false positives. These results were tested using the NBER recession dates and metrics such as the F1-Score, and the ROC curve with its AUC.**

***Index Terms*— Recession, Unites States of America, Macroeconomy, Indicators, Machine Learning, Classification, Binary, Logistic Regression, Random Forrest, XGBoost, NBER, F1-Score, ROC curve, AUC.**

I. Introduction

A common perception of working modern economies shows that they work around a trend rate of growth, with expansions and recessions phases. The expansion periods bring economic growth, and usually increasing standards of living, such as purchasing power, increased salaries and the ability to afford good healthcare and education. As for the recession periods the opposite tends to occur, as the economies halt their growth, so do the salaries tend to stabilize, and with it, a reduction of purchasing power that sometimes can lead to difficulties in accessing basic healthcare and education. These phases of growth affect not only individuals but also businesses, with reduced demands, reduction of profits and profitable economic opportunities. Evidently that these shifts in economic growth also affect governments, as macroeconomic agents, with an increased pressure from people and businesses for help in reversing these downward trends.

In the light of these statements, it becomes obvious the importance that people, businesses, and governments give to discovering when these shifts are bound to happen. This way, the primary concern of this paper is to find a way to predict these events with a considerable advance so that these agents can take the right precautions. Since the analyses of all global economies is impractical to perform due to structural differences between them, the United States of America (US) economy must be the one to focus since it is still the strongest economy in the advanced economies as for the latest International Monetary Fund (IMF) report [1], meaning that a recession in the US is more likely to affect all other global economies.

The use and classification of economic variables to infer economic downturns has been used for a long time in economic research, at least since Burns & Mitchell [2]. Throughout time there have been used several distinct variables, or in these case, economic indicators, that are recognized as useful indicators of an economic turning point, but the most acknowledge remains the yield curve [3]–[5]. Nevertheless most economists follow a broad variety of indicators which its usability cannot be determined to assess the future state of the economy, since until now none of them as proved completely reliable in the past.

So, this paper intends to develop Machine Learning (ML) models, for binary classification, that analyze diverse economic indicators and signal the possibility of an inversion point of the economic growth, in this case, the beginning of a recession phase through a variety of different time horizons.

II. Background & State of the Art

This section provides the fundamental concepts of economic recessions and ML algorithms required to better understand this work.

A. Economic Recessions

The economy can usually fluctuate between two distinct states, expansion or recession, simplified as economic growth or economic shrinkage, respectively. This cyclical behavior of the economy has been studied for several years and by various economists and institutions, but like all cyclical behavior studies, their interests focus mainly on the inversion points, not the stable phases. It is after these shifting periods that critical measures and different economic methods must be applied, resulting in structural changes in the economy, usually after a recession. Therefore, the prediction of an inversion point to a recession phase is one of the most coveted goals in economics.

These changes to the economy have a propensity to occur after a recession, due to their detrimental effects on society. Recessions weigh very heavily on families, mainly due to unemployment, loss of purchasing power, the decline in welfare

and education conditions, which ultimately tend to lead to earlier deaths, suicides, depressions, lower school achievement, child poverty and decreased natality rates [6].

A broad definition used by many economists to describe a recession is that it occurs when there are two consecutive quarters of Gross Domestic Product (GDP) decline, but the National Bureau of Economic Research (NBER)[1] uses a more extensive range of parameters to define each cycle [7]. This is important because the NBER has been widely accepted as the leading entity on the definition of the business cycles dates and economic research, providing a standard in this area of expertise.

To detect these economic turning points, economists use a set of macroeconomic related with consumption, production and employment. But this set of indicators is vast and presents several challenges, such as contradicting signals, survey's sampling errors, and the attributed weight of each indicator [8]. However, there are some indicators that surface in numerous literature as very strong predictors, like the yield curve, initial claims of unemployment insurance, new housing permits, monthly employment gains, or industrial production [3]–[5], [9]–[11].

As for the methods used to detect these turning points, current economist and researchers mainly use probabilistic models as they are currently the best predictive tools available. Probit models, a specific type of binary classification models, are the most commonly used probabilistic models to estimate the probability that an observation fits into one of the two categories available, in this case, recession or non-recession.

In conclusion, the detection of inflection points in business cycles is made using several macroeconomic indicators, mainly in the areas of consumption, production and employment, and the variation of these signals. The methods used to achieve these predictions are based on probabilistic models with an emphasis on binary classification models

B. Time Series Analysis

Time Series analysis, modeling, and forecasting has been applied to several problems in the last decades and has been the object of several studies and research. Its objective is to create and develop a model that can accurately describe a series throughout time and into the future using predictions or forecasts. This forecasting is usually described as the act of predicting the future by looking into the past and understanding it [12]. As so, time series analysis has become very important in the fields of engineering, economics, and finances. This last two are also considered to be the most challenging to work with, due to their noisy, non-stationary and deterministically chaotic characteristics, as mentioned in [12].

Given the importance of a time series analyses to comprehend the behavior of financial and economic signals[2], it is required to understand their components and how do their models work in forecasting their future behaviors.

Time series are non-deterministic by nature, meaning that they cannot predict the future with certainty. It is presumed that they follow a probability model, which describes the distribution of a random time-dependent variable, and the mathematical expression that describes it is referred to as a stochastic process [13]. They usually also have four main components [14]:

- Trend: the monotonic change in the average level of the time series.
- Trade Cycle: A long wave in the time series. The object of study of this paper.
- Seasonality: Fluctuation in the time series that recur during specific periods of less than a year and usually caused by factors as weather, vacations or holidays.
- Residues: Represents all the influences on the time series that are not explained by the previous components.

C. Binary Classification

Converting the problem of economic recession detection to a mathematical and probabilistic one requires an examination of the possible and favorable outcomes. By reviewing the economic overview, it is easy to conclude that there are only two possible outcomes available, there is either an economic recession or there isn't. This leads to the postulation that this problem is not only a classification one, i.e. the outcome belongs to a specific category, but it is also binary since there are only two possible categories to which the results can be distributed.

The leading area of study on the inferring process of transformation of an input signal to an output one is considered to be ML[15]. Mainly machine learning techniques use the available data to infer on some other information, which in the study of this paper presents itself as a good solution for discovering recession dates based on economic signals. To achieve this goal only two multivariate[3] types of models are needed, the linear and nonlinear.

1) Linear Approach

A model is defined as linear when each of its terms is either a constant or the product of a parameter with a predictor variable [16]. These models tend to follow an equation of the type,

$$Y(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \qquad (1)$$

where β_0 is a constant, β is a parameter and X is a predictor variable. Following this concept, it may appear that the linear models are uncapable to fit curves but that is not the case, since the linearity is set in the parameters and not on the predictor variables, so the model could be,

$$Y(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2^2 \qquad (2)$$

[1] The NBER is a private, non-profit, non-partisan organization dedicated to conducting economic research and to disseminating research findings among academics, public policy makers, and business professionals. [31]

[2] For the purposes of this paper an economic or financial signal can be understood as a macroeconomic indicator.

[3] Multivariate analysis is the analysis of three or more variables.

and still be linear [16]. This sets that the linear models can be applied to a binary classification problem even with some modifications. One of these models that is interesting to study is the Logistic Regression.

a) Logistic Regression

This model is suited to make a binary classification analysis since it uses several continuous values to determine an outcome that can only be one or zero[17].

Considering a model with a certain number of predictors, for instances, macroeconomic indicators, x_1, x_2, …, x_p, that take continuous values can be displayed in a logistic function like,

$$l(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad (3)$$

very much like the equation (1), where β are the models parameters, confirming that this is a linear model. The corresponding odds can be given by,

$$o = b^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p} \quad (4)$$

where b is the base of the logarithm and exponent. Now to predict the probability of finding a recession, meaning, $o = o : 1$, it can be given by,

$$p = \frac{b^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{b^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}+1} = \frac{1}{1+b^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p)}} \quad (5)$$

where β_0 is the y-intercept[4], and can be interpreted as log-odds, if all the predictors are zero. Also, the β_1, β_2 to β_p can be interpreted as the effects of x_1, x_2 to x_p since they increase the odds by b^{x_i}. This model can be used for several purposes and with numerous predictors, which for this paper is advantageous, since there are various macroeconomic indicators that can be used the discover the binary US recession signal. In conclusion, the Logistic Regression model can be described as [18],

$$logit\ p = \frac{e^{l(x)}}{e^{l(x)}+1} \quad (6)$$

substituting b, from equation (5), with the exponential e and using $l(x)$ from equation (3).

2) Nonlinear Approach

The nonlinear approach follows a very different methodology than the linear one. These models use nonlinear regressions to fit signals, and are usually preferred for fitting curved ones.

The nonlinear models can be divided into two categories, one is nonlinear in the variables and linear in the parameters, and the other is nonlinear in the parameters [19]. One example of a nonlinear model would be the Cobb-Douglas production function,

$$Y = \alpha L^{\beta} K^{\gamma} \quad (7)$$

Using the logarithms yields in (7),

$$\ln(Y) = \delta + \beta \ln(L) + \gamma \ln(K) \quad (8)$$

where $\delta = \ln(\alpha)$. This function is nonlinear in the Y, L and K variables and is linear on the parameters δ, β and γ. This shows that the nonlinear models can take different forms and are therefore able to fit very different kinds of signals, proving to be very adaptable, and probably a good solution for a binary classification system. There many of methods for implementing nonlinear models, but two of the most used and interesting are the *Bootstrapping Aggregating* and the *Gradient Boosting*.

a) Bootstrapping Aggregating (Random Forest)

Bootstrapping aggregating, also called bagging, is a meta-algorithm created to improve the stability and accuracy of ML algorithms, being considered a special case of the model averaging approach[5], mainly using decision classification trees.

The bagging was proposed by Leo Breiman in 1994, where he stated that [20],

"Bagging predictors is a method for generating multiple versions of a predictor and using these to get an aggregated predictor."

Basically, Breiman proposed that using several bootstrap replicates of the datasets, running them in classification trees and averaging their results would improve the accuracy of the model instead of using only one dataset. Afterward, in 1995, Tin Ho developed the concept of random decision forests. They were devised because decision trees could not grow in complexity, and therefore had a loss in accuracy for unseen data, and suboptimal accuracy in training data. Breiman then used the concept devised by Ho and joined his bagging proposition and developed what now is one of the most used models in ML, the *Random Forests*. Breiman defined random forests as [21],

"A random forest is a classifier consisting of a collection of tree structured classifiers $\{h(x, \Theta_k), k = 1, \ldots\}$ where the $\{\Theta_k\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x."

Essentially, Breiman constructs several trees that have random vectors or datasets. Each tree gets several results that it groups by class, and then votes on the most occurring class for a specific input. This strategy works as an effective prediction tool because it doesn't overfit due to the Law of Large Numbers.

b) Gradient Boosting (XGBoost)

As with bootstrapped aggregating, gradient boosting is an ML technique for regression and classification problem. It mainly uses decision classification trees.

Boosting is an ML algorithm designed to reduce bias and variance. Schapire and Freund describe Boosting as [22],

[4] The point that intercepts the function in the Y axis.

[5] Model Averaging or Ensemble Learning, uses multiple learning algorithms to obtain better predictive performance than only using one single algorithm.

"…converting a 'weak' PAC[6] learning algorithm that performs just slightly better than random guessing into one with arbitrarily high accuracy."

They also go on to describe it as,

"*Boosting refers to this general problem of producing a very accurate prediction rule by combining rough and moderately inaccurate rules-of-thumb*".

Basically, *Boosting* is the process of using predictors that are known to be imperfect, combining them to generate an improved predictor capable of producing better predictions on the target. This procedure uses a set of labeled training examples, (x_1, y_1), …,(x_N, y_N), where y_i is the label associated with x_i. Each round $t = 1, \dots, T$ the booster devises a distribution D_t over the set of examples and request for a weak hypothesis, or a rule-of-thumb, h_t, with low error, ε_t, with respect to D_t. This way, distribution D_t specifies the relative importance of each example for each round. After the T rounds, the procedure combines the weak hypothesis into a single prediction rule [22]. This describes the basic functioning of the *Boosting* algorithm

In 1999 Jerome Friedman officially proposed the first gradient boosting algorithms and *Tree Boosting*. He considered that the gradient boosting of regression and classification trees produced competitive and highly robust procedures for regressions and classifications [23]. The idea behind gradient boosting was to follow the concept of the *Boosting*, combining the 'weak' learners into a strong one, and finding an approximation to the function that minimized the expected value of some specific loss function. Therefore, to achieve this he used a Gradient Descent[7] that gave the name to the algorithm.

Friedman also developed a special modification to use with decision classification trees, so that they could better fit each base learner[8]. Its main method was to partition the input space by the number of leaves in the tree, J_m, into disjointed regions $R_1, \dots, R_{J_m}$, and then predict a constant value in each region. With this alteration to the input space he could then change its algorithm so that he could add a multiplier for each of the tree's regions instead of the entire tree. This was called Tree Boosting.

But the work on gradient boosting trees has not stopped since Friedman. One of the latest models that uses this concept is the Extreme Gradient Boost (XGBoost), proposed by Tianqi Chen and Carlos Guestrin [24]. This model follows the previous notions of Tree Boosting and Gradient Boosting but uses different weight distribution procedures for sparse datasets. This model was specifically designed for computational use, with improvements on the former models in terms of computational time and memory allocation, but still using the same principles

D. State of the Art

The prediction of an economic recession is a long time desired goal by many economists, but also by many other areas of study, including engineering and computer science. With the development of the computational power of Personal Computers (PC) some old mathematical models became increasingly easy to apply and use on great amounts of information. This development led to the booming of areas of study like ML, Data Mining (DM), Neural Networks (NN), that began providing answers to long sustained questions in our societies, like the works of Rui Neves and Nuno Horta [25]–[28]. The combinations of these two ideas guided many scientists and economists to the computer realm in order to predict economic recessions. The state of the art for this thematic can be divided into three areas of relevance, the data, the models, and the results & metrics.

1) Data

Each author uses several different indicators, but as stated in the economic background, some areas appear of bigger relevance than others. For Instances Arturo Estrella and Frederic Mishkin, some of the first to explore this theme, use variables more connected with interest rates and money availability, rather than production and working conditions [3]. Other authors like Travis Berge, use more macroeconomic indicators like the industrial production, the average weekly hours of the manufacturing industry or the initial claims of unemployment [11]. These approaches follow a smaller set of indicators, usually between ten and twenty different ones, but not all authors use this low information method.

Chikako Baba and Turgut Kışınbay of the IMF use up to 166 macroeconomic indicators divided into five categories: income and output, employment, construction, interest rates and finally nominal prices and wages [10]. Also, the Wells Fargo Securities Economics Group uses a high-level number of indicators approach. They start by retrieving 500,000 indicators from the Federal Reserve Bank of St. Louis (FRED), then by only using the start date of 1972 reduce the dataset to 5,889 variables and eventually using other procedures they reduce the dataset to a final count of 192 different indicators [29].

The most commonly used methods are the probit and logit models, they are considered good solutions for binary classification problems and therefore many of the authors used them, if not to develop the final forecast then as part of their system [9], [10].

2) Models

There are several mathematical models proposed to solve this type of binary classification problems. Most of the literature follows the same basic principles on this matter, adding some alterations in handling the data and the forecasting methods, but almost always using multivariate models.

The most commonly used methods are the probit and logit models, as they are considered good solutions for binary classification problems and therefore many of the authors used them, if not to develop the final forecast then as part of their system [9], [10]. But even with the good results shown by these models, the community kept on trying different approaches to

[6] Probably Approximately Correct is a framework for mathematical analysis that receives samples and selects a generalization function from a certain class of possible functions [32].

[7] Gradient Descent is a first-order iterative optimization algorithm that finds a minimum of a function.

[8] Weak learner.

the problem and lately, companies like Wells Fargo also tried Random Forest approaches and Gradient Boosting to deal with their data [29]. These models are the current the state of the art not only in the recession detection area but in almost every problem that deals with binary classification.

Apart from the models themselves, the authors have used several other strategies to improve their results by meddling with the data or even with the models. One of the strategies for forecasting recessions with some time advance was to make a recession prediction, using some specific model, and then using its results and forecasting them with a new univariate model [4]. But not only on the future prediction methods are these changes applied, Travis Berge for instances also applies several model averages on his forecasting models to try and improve their power. He uses a normal weighted average, a different weighted average called Bayesian Model Average, and also an alternative solution of choosing the better model through a boosting algorithm [11].

3) Results & Metrics

The only benchmark used for the US recession dates is the one provided by the NBER. These dates provide a term of comparison in all literature and are widely accepted as the signal to beat. As for the metrics used, they tend to vary from author to author, with different explanations being used.

The mainly used metrics tend to be the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC). Authors like Weiling Liu and Emanuel Moench, use these metrics to evaluate their work [9], retrieving results for their several tests with AUC values around 0.8. Also the Wells Fargo Group uses this metric with AUC results around the 0.9 values, even though these values don't seem to fit the graphics they present [29].

Older works like the Estrella and Mishkin, and the Baba and Kışınbay used other types of metrics to evaluate their works. Estella uses the Pseudo R^2 and Baba the Quadratic Probability Score (QPS) and the Log Probability Score (LPS) [3], [10]. These metrics are not used very often in more recent works in the area.

III. System Architecture

The design of the solution for this work is based on a layered scheme architecture, where each layer of the system provides a logical division between the storage, the transformation and the forecasting of the data. This segmentation has a very straightforward approach, but it is abstract enough so that future changes in the system can be applied with no major complications.

There are four main layers in this system architecture, the Database Layer, Data Transformation Layer, Classification Layer, and the Validation Layer, and each have several components, as shown in Figure 1. It can clearly be noted that the CSV Files do not enter any of these layers, mainly because they were retrieved directly from the internet with no resource to any tool designed for this paper.

The process begins by retrieving the CSV Files from the FRED website [30], and by sending them to the Database Layer. This layer stores the information sent by the CSV Files where it can be analyzed, for instances, by its granularity and correlation[9], then directed to the next step, the Data Transformation Layer. Here it receives these signals and reshapes them accordingly to the users' desires, changing their memory, granularity, or even adding lag, so that this information is processed by the Classification Layer in the desired manner. Then the Classification Layer takes these signals and runs them through the chosen models to produce the final results that are sent to the Verification Layer to be observed, evaluated, and compared to the benchmark performance.

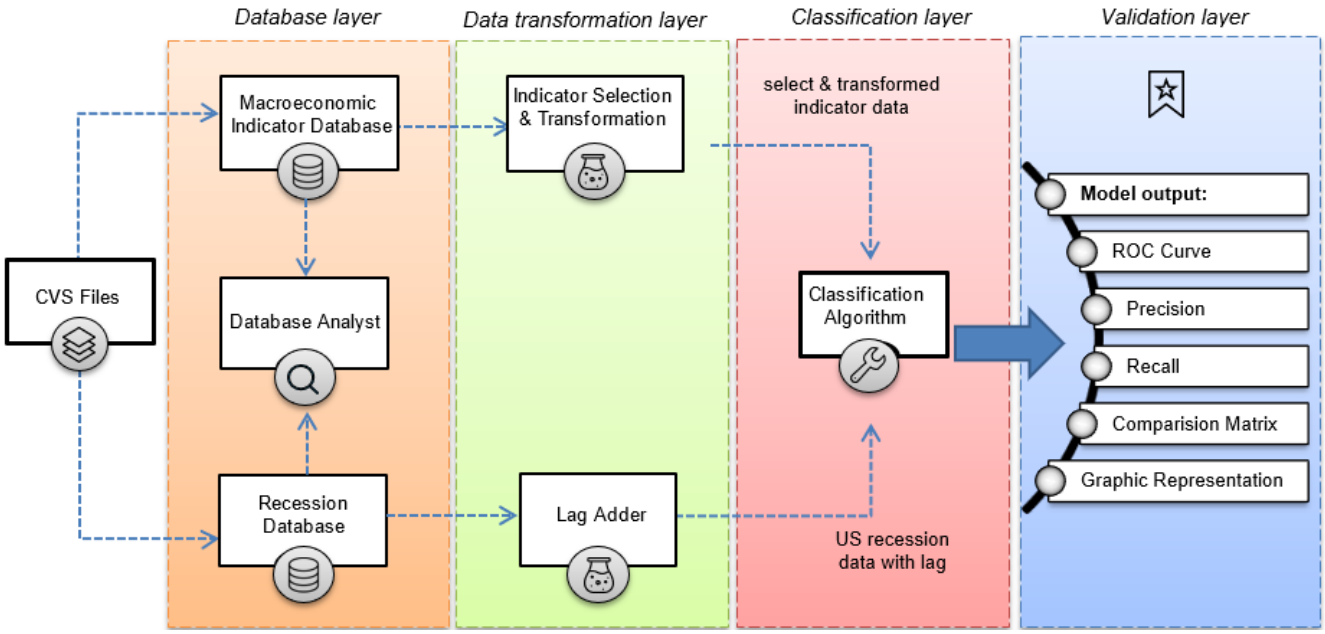

Figure 1. System Architecture.

[9] Correlation is the statistical relationship between two signals.

A. Database Layer

This Layer, as seen in Figure 1, has two main components, each in a single branch that extends until the Classification Layer, the Macroeconomic Indicator Database and the Recession Database, and an analytic component, the Database Analyst, that can receive data from both databases and provide simple analyses.

The Macroeconomic Indicator Database stores and handles the macroeconomic indicators retrieved from the FRED and provides access to the Data Analyst and the Indicator Selection & Transformation components. The data stored in this database has very different granularities, starting and ending dates, units, and seasonalities. That leads to a difficult analysis and the requirement to transformed them, so they can be used in the Classification Layer, as it will be demonstrated in the System Implementation chapter.

The Recession Database has the information about US recession dates with a monthly granularity, and its instances can be used in the Data Analyst component like the Macroeconomic Indicator Database and in the Lag Adder. This database may seem unnecessary, but it is separated from the Macroeconomic Indicator Database for abstraction purposes, guarantying that in the future other economic recession signals can be added and, analyzed between each other and against indicators from different countries.

The last component of the Database Layer is the Data Analyst which is responsible for the analysis on each individual indicator or recession, the combination of data from within each database and the combination of data between both databases. The Data Analyst does evaluations on the granularity, number count, verification of missing data, histogram studies, and a primary forecasting for each signal and analyzes the correlation between signals from the same and different databases. This allows for a more comprehensive study of the available signals in order to help the user choose which signals to use. This reveals importance because the signals can be highly correlated, that would lead to overfitting, or even to reduce the size of the dataset if computational power is not available. In conclusion, this layer has to be able to compartmentalize the data and grant the user the ability to analyze it.

B. Data Transformation Layer

The Data Transformation Layer like the Database Layer has two main branches, mainly differentiated by the content of the data. One deals with the macroeconomic indicators and the other uses the economic recessions, this is required since the data has different specialties, different methods to change it and two different goals to achieve.

The Indicator Selection & Transformation component's main goal is to supply the user with the ability to choose which indicator to use, according to its analyses made in the Data Analyst, and transformations to use in the Classification Algorithm. This component also guarantees that all the indicators have the same granularity, end and start dates, and can be processed by the Classification Algorithm. For the architecture design phase, this component was simplified, not showing the entirety of the processes within it, but it will be further explored in the Data Transformation Layer Implementation subchapter.

Like with the Indicator Selection & Transformation the Lag Adder was also simplified for this stage and will be developed in further chapters. But its main goal is to add a lag to the economic recession signal in order to "trick" the Classification Algorithm into "thinking" that a recession begins before it actually does. It is expected that with this technique the chosen models would be able to detect a recession with some advance and therefore accomplish the purpose of this paper

C. Classification Layer

This layer only has one component, the Classification Algorithm, that receives two datasets, one from the Indicator Selection & Transformation and the other from Lag Adder, with the chosen macroeconomic indicators and the lagged economic recession signals, respectively. The purpose of this layer is to produce forecasts of the US economic recessions using the macroeconomic indicators, with the most precision and recall possible.

The Classification algorithm component uses several models that receive the selected macroeconomic signals and their transformations to produce a binary response that assimilates to the US recession signals with their several lags. The component also returns other validation responses that are fundamental for the user to assess the confidence that can be handed to the model and its responses. This segment of the system may require more computational power and therefore, can and must be worked by the user in order to make it viable for running in the testing machine.

D. Validation Layer

This layer has a very different behavior from the previous ones, since it does not have any component, and only shows the results from the Classification Layer to the user so that they can be analyzed and concluded upon. This way the Verification Layer can be considered an abstract layer or a ghost layer since it doesn't really exist but makes no logical sense to join it with Classification Layer.

The Verification Layer presents to the user, the plots returned from the used models for a graphic evaluation and the ROC curve, the precision, recall, f1-score, and confusion matrix for an analytic examination of the results. This layer can work as a powerful iterative tool to understand possible problems with the chosen models, their definitions or even with the indicators themselves, enabling the user to adjust the system for better results. The explanation and evaluation of this results will be presented in the chapter Results & Validation.

IV. Results & Metrics

A. Metrics

To evaluate the quality of the different approaches some global metrics must be applied. They measure the performance of the several models and their variations, following the state of the art metrics for a time series binary classification analyses. For the following examination, the resulting signal provided by the models' probability prediction is averaged and is classified as a recession or not, i.e. the result value is one (recession), for recession probabilities above 50%, and zero (no recession) otherwise.

1) Confusion Matrix

The Confusion Matrix is a table composed of two rows and two columns that cross the results of the prediction model and the targeted values.

Table 1. Confusion matrix table for the US recession prediction

		Actual Recession	
		True	False
Predicted Recession	True	True Positive	False Positive
	False	False Negative	True Negative

The Table 1 represents the confusion matrix of this paper and presents the relation between the predicted results and targeted results. From this crossing, four concepts arise, the True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN).

- True Positive:

Is an outcome where the model correctly predicts the positive class.

- False Positive:

Is an outcome where the model incorrectly predicts the positive class.

- False Negative:

Is an outcome where the model incorrectly predicts the negative class.

- True Negative:

Is an outcome where the model correctly predicts the negative class.

The positive class, represents the case of a recession really happening, and the negative class the opposite. The matrix of Table 1, returns the amount of each of these concepts present in the results of each model.

2) Precision

The Precision metric represents the ratio between the correctly predicted *positive* observations and the total of positive observations. This metric answers the question "of all recessions labeled as true, how many actually happened?".

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

The Equation 1 presents the calculus formula to asses this ratio where a return of 1 represents a good quality model and 0 a bad quality model.

3) Recall

The Recall, or sometimes called Sensitivity, metric represents the ratio between the correctly predicted positive observations and all the observations of that class. This metric answers the question "of all the recessions that truly happened, how many did we labeled?" where a return of 1 represents a good quality model and 0 a bad quality model.

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

[10] Accuracy is the ratio between the correctly predicted observations and the total number of observations. $\frac{TP+TN}{TP+FP+FN+TN}$

The Equation 2 presents the calculus formula to asses this ratio where a return of 1 represents a good quality model and 0 a bad quality model.

4) F1-Score

The F1-Score metric represents the weighted average of the Precision and Recall metrics. This metric takes both the False Positives and the False Negatives into account. It may not be of easy understanding, but this metric acts as a kind of accuracy[10], but better suited for unbalanced datasets. In view of these assertions, this metric presents itself as the best for verifying the quality of this paper models' predictions.

$$F1 - Score = \frac{2TP}{2TP + FP + FN} \quad (3)$$

The Equation 3 presents the calculus formula to asses this ratio, where a return of 1 represents a good quality model and 0 a bad quality model

5) ROC Curve

The Receiver Operating Characteristic (ROC) curve is a graphical plot that illustrates the power of a binary classification system with different rates of TP and FP. The ROC curve is created by plotting the *Recall* against the *False Alarm Rate*[11] and as the curve approaches the left and the top side of the graphic the more accurate the test is. By contrary, if the curve approaches the 45º degree the less accurate it is.

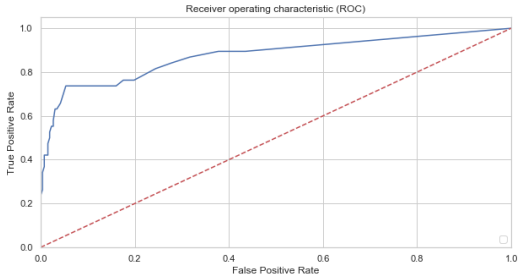

Figure 2. ROC curve of a tested model not used in the paper.

Figure 2 presents an example of a ROC curve that appears to have a good accuracy since the curve is somehow near to the top left of the plot, but just by analyzing the graphic no conclusion can be reached if compared with another model. To solve this issue another metric can be used, the Area Under the Curve (AUC).

a) Area Under the Curve (AUC)

The AUC measures the area beneath the ROC curve by making an integral of the signal. If the AUC returns near the value 1, then it represents a perfect test and if near 0.5 a worthless one. The following ranges are a crude way of assessing the strength of the model:

- 1 - 0.9 = Excellent
- 0.9 – 0.8 = Good
- 0.8 – 0.7 = Fair
- 0.7 – 0.6 = Poor
- 0.6 – 0.5 = Fail

[11] The False Alarm Rate is rate of false positives among all the cases that should be negative. $\frac{FP}{FP+TN}$

So, applying this method to Figure 2 we have:

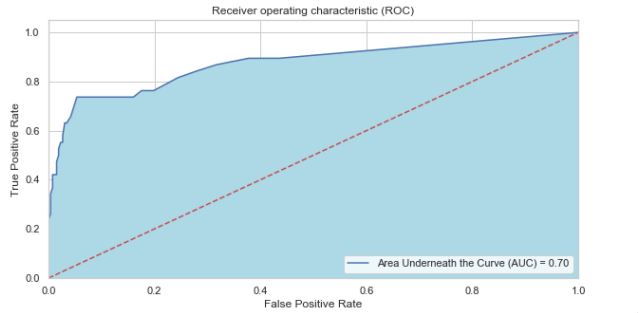

Figure 3. ROC curve with AUC of a tested model not used in the paper.

Now it becomes clear that even though the curve is near the top left of the plot it can only be considered a fair model. This metric system allows for a more analytic comparison between models, and since it tests for all probability thresholds, i.e. the threshold defined by the user to differentiate above which probability value is it considered a recession or not, it can improve on the previous metrics.

B. Case Studies

This paper's results will be benchmarked with the dates of the US recessions provided by the NBER, the respective lags, and validated with the previously stated metrics. The most positive results, in a best-case scenario, would be to beat the benchmark and the most recent state of the art results, providing a clear and reliable US recession forecaster.

This work is divided into four case studies, to compare their different solutions on achieving the goal of detecting and predicting a US recession. In all case studies, there are used four different US recession lags and three different input signals, two of them transformed.

- Case Study A:

A linear approach is taken, using the *Logistic Regression* model, normalized data, and monthly granularity.

- Case Study B:

A nonlinear approach is taken, using the *Random Forest Classifier* model, absolute value data, and monthly granularity

- Case Study C:

A nonlinear approach is taken, using the *XGBoost* model, absolute value data, and monthly granularity.

- Case Study D:

A model averaging approach is taken, where the models' prediction results used in the previous case studies are averaged by a simple, same weight average.

To evaluate the results of this systems and to be able to compare the different case studies, the chosen metric for this paprer was the AUC, since the F1-Score results varies to much, due to the probability threshold of 50%.

1) Case Study A

This case study follows a linear approach to the modeling and uses normalized data to achieve it. For this study several US recession lags were used, the six months lag, twelve months lag, eighteen months lag and the signal with no lag. This allows forecasting in four different time spans and at maximum trying to detect a recession one year and a half before of time. Also, in this case study, three types of input signals were used, the three month memory transformation, the six month transformation and the original signal with no transformation.

Table 2. AUC results from the ROC curves of the Logistic Regression *Model*

X	Logistic Regression - Area Under the Curve (AUC)			
	No Month Lag	Six Month Lag	Twelve Month Lag	Eighteen Month Lag
No Memory	0,89	0,81	0,83	0,82
Three Month Memory	0,9	0,82	0,82	0,82
Six Month Memory	0,89	0,83	0,82	0,86

Using Table 2 it becomes clear that this model behaves exactly like expected, increasing its predictive power with the increase of memory and decreasing the predictive power with the increasing of lag.

For visual comparison between case studies, the models' results can be plotted. Following an average lag of twelve with a full six-month memory approach, the models that will be compared in all case studies.

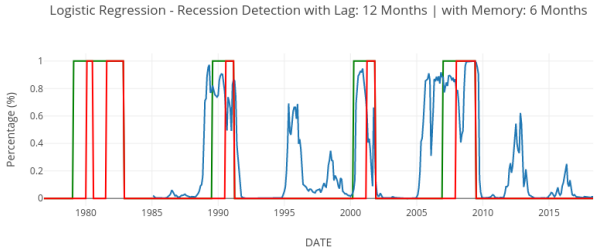

Figure 4. The probability of a US recession happening with twelve months of lag and six months of memory calculated by a Logistic Regression *algorithm.*

Figure 4 displays the probability of a recession happening with a twelve month advance, where the green lines represent the recession signal lagged by twelve months, the red, the actual recessions, and the blue, the probability of a recession occurring calculated by the model.

2) Case Study B

This case study follows a nonlinear approach and uses the *Random Forest Classifier* model to achieve it. The data, contrary to the Case Study A, uses absolute values, but like the previous case study, it uses all the US recession lags and all the memory transformations provided by the system.

Table 3. AUC results from the ROC curves of the Random Forest Classifier *Model.*

X	Random Forest Classifier - Area Under the Curve (AUC)			
	No Month Lag	Six Month Lag	Twelve Month Lag	Eighteen Month Lag
No Memory	0,83	0,88	0,88	0,9
Three Month Memory	0,84	0,88	0,88	0,9
Six Month Memory	0,87	0,87	0,89	0,89

Table 3 shows that the Random Forest Classifier algorithm didn't behave as expected, since the predictive power didn't increase with the memory and didn't decrease with the lag. Even though, the results are very similar, and little conclusion can be derived by them if not that the increase in memory does not help this model, and the increase in lag also doesn't hurt it.

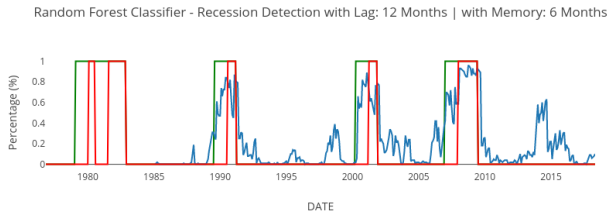

Figure 5. The probability of a US recession happening with twelve months of lag and six months of memory calculated by a Random Forest Classifier *algorithm.*

Figure 5 appears to have more volatile results than in Case Study A, but on the other hand, it fits the recessions better and doesn't raise such higher FP probabilities. Both case studies provide very good and reliable results but clearly show that they work in a very different manner providing distinct results.

3) Case Study C

This case study, like the Case Study B, presents a nonlinear approach, but instead of using the *Random Forest Classifier* model it uses the *XGBoost* model. Also, like the Case Study B, it also uses absolute values and all the US recession lags, and all the memory transformations provided by the system.

Table 4. AUC results from the ROC curves of the XGBoost *Model.*

X	XGBoost - Area Under the Curve (AUC)			
	No Month Lag	Six Month Lag	Twelve Month Lag	Eighteen Month Lag
No Memory	0,77	0,87	0,83	0,8
Three Month Memory	0,82	0,83	0,8	0,8
Six Month Memory	0,83	0,81	0,81	0,78

Unlike the results from Table 3, the results from Table 4 show that most of the iterations decrease their predictive power with the increasing of the lag. But on the other hand, they tend to decrease their predictive power with the increasing of the memory, raising the point that the *XGBoost* model does not deal well with the increasing of the memory of the signal.

The results of this case study are not as good as the ones from the Case Study B but follow a more expected path. As for the Case Study A, it loses in most of the iterations, but at the six month lag presents better results.

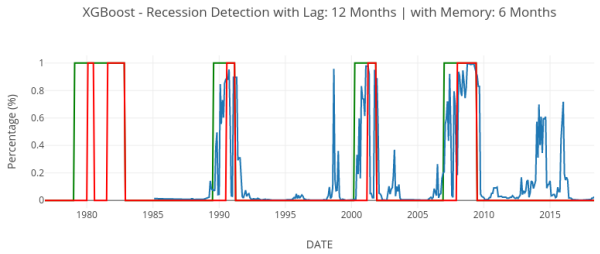

Figure 6. The probability of a US recession happening with twelve months of lag and six months of memory calculated by an XGBoost *algorithm.*

Figure 6 presents much more volatile results as the ones from the previous case studies, as expected, since the results from Table 3 aren't as strong as their previous counterparts. Even though, the results are also very well fitted to the US recessions and follow the same dates as Case Study A and B for its FP.

4) Case Study D

The final case study of this paper is by far the most different of all since it isn't based on a single model, but the average of the three previous ones. So, it uses the joint power of the linear and nonlinear approach to try and compensate for the flaws of each model.

Table 5. AUC results from the ROC curves of the Model´s Average

X	Models Average - Area Under the Curve (AUC)			
	No Month Lag	Six Month Lag	Twelve Month Lag	Eighteen Month Lag
No Memory	0,81	0,88	0,86	0,87
Three Month Memory	0,86	0,86	0,87	0,9
Six Month Memory	0,87	0,86	0,89	0,9

Table 5 demonstrates that even though the predictive power increases with the lag they have very similar values and very high results. That is expected when averaging the three models since it covers the weak points of each model providing a more stable solution. This solution presents itself has the strongest amongst all case studies, since it has strong combination of stability and high values of AUC than its predecessors.

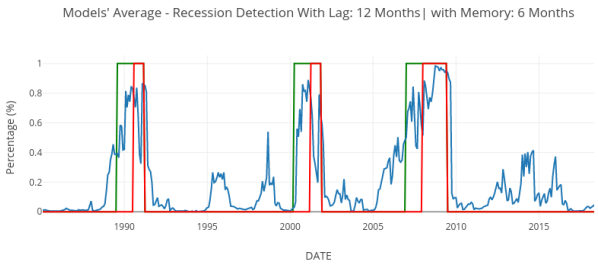

Figure 7. The probability of a US recession happening with twelve months of lag and six months of memory calculated by an average of all the models' algorithms.

Figure 7 presents very stable and well fitted results, demonstrating that the average of these models is very

important in the construction of a reliable solution for this paper's problem.

With all the case studies reviewed and analyzed it is easy to conclude that all of them have strengths and weaknesses. But, even with some failures, almost all of them present very high levels of AUC's, and their visual representations are clear enough to retrieve some conclusions.

V. Conclusion

By reviewing the previous chapter, it can be concluded that the main objectives of this paper were accomplished bearing good results. There were used three distinct models, with distinct capabilities and formulations, and all bared good results, each with its strengths and weaknesses. For instances the Logistic Regression model was best at predicting recessions with no lag. The Radom Forest Classifier presented the best results for the six, twelve and eighteen months of anticipation and the XGBoost model was the most all rounded model maintaining the quality of its predictions no matter the changes. This can be confirmed by the usage of the Case Study D, where all the previous models' results were averaged and produced very high-quality predictions for all the transformations and lags. In conclusion, the best model to use is the Models' Average, where for a probability above 50% the user might clearly be alerted for a possibility of a recession.

On an ending note, this work devised several solutions, all thoroughly evaluated using state of the art metrics, and provided good results. It demonstrated that it is possible to detect US recessions with a certain confidence, with the usage of several macroeconomic indicators assigned to specific areas of the economy. It also proves that this problem can be solved with public available information and using a personal computer to produce the necessary computations for each model.

References

[1] IMF, "World Economic Outlook, April 2018: Cyclical Upswing, Structural Change," 2018.

[2] A. Burns and W. C. Mitchell, "Measuring business cycles," *Q. J. Econ.*, vol. 64, no. 2, pp. 311–318, 1946.

[3] A. Estrella and F. S. Mishkin, "Predicting U.S. Recessions: Financial Variables as Leading Indicators," *Rev. Econ. Stat.*, vol. 80, no. 1, pp. 45–61, 1998.

[4] H. Kauppi and P. Saikkonen, "Predicting U.S. Recessions with Dynamic Binary Response Models," *Rev. Econ. Stat.*, vol. 90, no. 4, pp. 777–791, 2008.

[5] G. D. Rudebusch and J. C. Williams, "Forecasting recessions: The puzzle of the enduring power of the yield curve," *J. Bus. Econ. Stat.*, vol. 27, no. 4, pp. 492–503, 2009.

[6] V. Adrian, S. Coontz, and P. Education, "The Long Range Impact of the Recession on Families," 2010.

[7] NBER, "The NBER Business Cycle Dating Committee," *NBER Website*, 2010. [Online]. Available: https://www.nber.org/cycles/recessions.html.

[8] C. Council of Economic Advisors, "Assessing the State of the Economy in Real Time Using Headline Economic Indicators," 2017.

[9] W. Liu and E. Moench, "What predicts US recessions?," *Int. J. Forecast.*, vol. 32, no. 4, pp. 1138–1150, 2016.

[10] C. Baba *et al.*, "Predicting Recessions: A New Approach For Identifying Leading Indicators and Forecast Combinations IMF Working Paper Monetary and Capital Markets Department Predicting Recessions: A New Approach for Identifying Leading Indicators and Forecast Combinations," 2011.

[11] T. J. Berge, "Predicting Recessions with Leading Indicators: Model Averaging and Selection over the Business Cycle," *J. Forecast.*, vol. 34, no. 6, pp. 455–471, 2015.

[12] L. Cao and F. E. H. Tay, "Application of support vector machines in financial time series forecasting," *Omega*, vol. 29, no. 4, pp. 309–317, 2001.

[13] R. K. R. A. Agrawal, *An Introductory Study on Time Series Modeling and Forecasting*. 2013.

[14] P. Desikan and J. Srivastava, "Time Series Analysis and Forecasting Methods for Temporal Mining of Interlinked Documents," 2015.

[15] E. Alpaydın, *Introduction to machine learning*. 2014.

[16] J. A. Nelder and R. W. M. Wedderburn, "Generalized Linear Models," *J. R. Stat. Soc. Ser. A*, vol. 135, no. 3, pp. 370–384, 1972.

[17] D. R. Cox, "The Regression Analysis of Binary Sequences," *J. R. Stat. Soc. Ser. B*, vol. 20, no. 2, pp. 215–242, 1958.

[18] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*. 2000.

[19] R. C. Hill, W. E. Griffiths, and G. G. Judge, *Undergraduate econometrics*. Wiley, 2001.

[20] L. Breiman, "Bagging predictors: Technical report no. 421," 1994.

[21] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[22] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, 1997.

[23] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001.

[24] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *KDD '16 Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 785–794, 2016.

[25] A. Gorgulho, R. Neves, and N. Horta, "Applying a GA kernel on optimizing technical analysis rules for stock picking and portfolio composition," *Expert Syst. Appl.*, no. 38, pp. 14072–14085, 2011.

[26] A. Canelas, R. Neves, and N. Horta, "A SAX-GA approach to evolve investment strategies on financial markets based on pattern discovery techniques," *Expert Syst. Appl.*, no. 40, pp. 1579–1590, 2013.

[27] A. Silva, R. Neves, and N. Horta, "A hybrid approach to portfolio composition based on fundamental and technical indicators," *Expert Syst. Appl.*, no. 42, pp. 2036–2048, 2015.

[28] B. Jubert de Almeida, R. Ferreira Neves, and N. Horta, "Combining Support Vector Machine with Genetic Algorithms to optimize investments in Forex markets with high leverage," *Appl. Soft Comput. J.*, no. 64, pp. 596–613, 2018.

[29] J. E. Silvia, "Can Machine Learning Improve Recession Prediction? Does machine learning and statistical data mining improve recession prediction accuracy? | Economics Group Special Commentary," 2018.

[30] "Federal Reserve Economic Data | FRED | St. Louis Fed." [Online]. Available: https://fred.stlouisfed.org/.

[31] The National Bureau of Economic Research, "About the NBER," 2017. [Online]. Available: http://papers.nber.org/info.html.

[32] L. G. Valiant, "A theory of the learnable," *Commun. ACM*, vol. 27, no. 11, pp. 1134–1142, 1984.

First R.do Ó was born in Pragal, Almada, Portugal in 1992. He received his M.S. degree in Electrotechnical Engineering and Computer Science, from Instituto Superior Técnico from the University of Lisbon, in Lisbon 2018.

From June 2015 to November 2016 he was the President of the Students' Association of Instituto Superior Técnico, and from April of 2017 to October of 2018 he was Counselor at the General Council of the University of Lisbon.