Candidate Number: 36044

Supervisor: Dr Raúl Aldaz

Date: 20/08/2020

Project Title:

The Use of Shrinkage Methods to Predict US Economic Recessions.

Word Count: 9662

Abstract:

This study's aim is to test if the use of shrinkage methods can improve the out-of-sample performance of the prediction of recessions in the United States. The benefits of this family of models include the reduction of the issues of overfitting even when the model includes a large number of explanatory variables. To do so the study provides an in-depth analysis of the models previously used to predict recessions and the potential for improvement using shrinkage methods. Subsequently, it tests the performance of a number of model fitted across 5 different forecasting horizons (0, 3, 6, 9, 12 months in advance) using shrinkage methods against the predictions of the Survey of Professional Forecasters and the model specification defined by Estrella and Mishkin (1998). The various models differ in the dependent variable that is being predicted since recession detection can be interpreted both as a binary classification problem or as regression problem. The results show that shrinkage methods are successful in predicting recession for the classification problem up to 12 months in advance, while for the recessions problem there are no accurate results past the 3 months in advance forecasts.

Section I: Introduction (626 words, 319 already used)

Since the beginning of the study of econometrics, one of the most fundamental questions has been forecasting the future economic performance of a nation. The advantages of being able to do so are numerous and depend on the point of view of the stakeholder. For example, the most obvious advantage for the government would be being able to adopt the most appropriate monetary and fiscal policy to dampen the fluctuations of the business cycle (Valckx, de Ceuster and Annaert, 2002, p.1). While instead, an investor could change their portfolio to a more or less risky position depending on the forecasted economic growth (Sornette, 2004, p.1-2). Many researchers and organizations such as the Conference Board or the Federal Reserve Bank of Philadelphia developed and refined over the years statistical models to predict the probability of recession and their magnitude. However, despite the advances in forecasting economic growth, one problem that has persisted is producing a reliable recession forecast far enough in advance to allow policymakers to act (Gogas *et al.*, 2015, p.635-636). As Zarnowitz and Braun (1993) point out even while the 1973 and 1980 recessions were taking place in the US, they were not still unanimously recognized as such. The issue was that initially these studies were conducted to understand the mechanisms behind recession more than to provide reliable predictions (Estrella & Mishkin, 1998, p.46). However, as more data was gathered and the methodology shifted from optimizing the in-sample error to the out-of-sample error[1], the prediction power of these methods improved steadily, but always to find an indicator or a small number of variables that could produce consistent predictions as shown by Filardo (1999) comprehensive comparison. Recently, machine learning methods were also implemented in prediction models, with very encouraging results but still with a very limited number of variables (Gogas et al., 2015; Nyman & Ormerod, 2016). Moreover, the machine learning methods implemented, such as Support Vector Machines or Random Forest were not linear which are indeed more flexible but cause the final results to be less interpretable (James et al., 2013).

Therefore, this study will test if introducing a wide array of economic and financial indicators in a linear machine learning method capable of conducting variables selection produces more accurate recession forecasts than what has been attempted by the cited literature. It will do so by two using model regularizations methods, the LASSO, and a Logistic regression with a penalty term, to predict GDP growth and the probability of recession respectively. This class of machine learning methods introduces a penalty term for the size of each coefficient to the standard OLS regression or logistic regression, thus conducting variable selection by bringing

[1] In-sample and out-of-sample is the terminology used in the literature to describe what in machine learning would be called the training sample and test sample.

the coefficient of certain variables to zero and maintaining the number of the explanatory variables in the model low (James *et al.*, 2013, p.219–220). As far as I am aware these techniques have not been applied to this field before and using this method should construct a model more detailed than the ones reported in the literature but without incurring in the risk of overfitting even when a large number of explanatory variables are included (James *et al.*, 2013, p.214–215). The model's results will then be compared against two benchmark models: the Survey of Professional Forecaster and a model constructed following the specification detailed by Estrella and Mishkin (1998). To do so this study will first examine the existing literature relating to the topic of predicting recessions in section II. In section III it will discuss the data collection strategy, subsequently section IV will discuss the data analysis methodology. Section V will analyse the results and compare them with the benchmark models. Lastly, Section VI concludes by discussing the results and limitations of the approach as well as potential directions for further research.

Section II: Literature review (2102 words, 1876 already used)

The earliest studies that attempted to predict recession were usually mostly focused on linking specific financial indicators to recession. For instance, Hamilton (1985) and Mills (1988) focused on oil shocks and stock prices, respectively, as possible leading indicators for recessions. The former only managed to prove the existence of a correlation between high oil prices and recession, while the latter showed that a model to predict recessionary quarters solely based on stock prices produces a great deal of false alarms and missed forecasts even with manipulation of the threshold probability to trigger recession predictions. Both studies, as it was common for the majority of the studies before the 1990s, focused mostly on in-sample analysis, in an attempt to better understand the cause of each recession (Estrella & Mishkin, 1998, p.46). However, models uniquely aimed at improving in-sample analysis usually suffer from severe overfitting as they attempt to create model that fits the available data extremely well by adding explanatory variables (Filardo, 1999, p.39). This makes the predictions on out-of-sample data worse as the model becomes very susceptible to random noise in the sample (ibid.).

Recently, however, more and more studies have adopted test error minimization as the principal criterion of predictive accuracy, as shown for instance in Estrella and Hardouvelis' (1991) seminal paper, which also first introduced the 10 year bond – 3 months treasury bill spread (yield spread) as an indicator to predict Gross National Product (GNP). This estimator has proved extremely successful in predicting all of the recessions in the US since data collection begun, hence it is implemented in some form in the models of most recent research as well as many organizations devoted to the forecasting of the business cycle (Filardo, 1999, pp.36-39). The motivation behind the power of this estimator is that a single measurement includes information about current monetary policy in the form of short term interest rates and the expectation of the market regarding the health of the economy in the future (Wright, 2006, p.1)[2].

The methodology used in the literature focused on forecasting recessions can essentially be divided in two: studies using regression models such as OLS to predict the future level of GNP or GDP, or using classification models, such as probit or logistic regression, to focus purely on the probability of the economy being in a recession at a certain forecasting horizon. The

[2] The 10-year bond interest rate is mostly dictated by the market and fluctuates depending on the expectation of future economic growth and interest rates, while the 3-month treasury bill is mostly determined by the federal funds rate (Estrella & Hardouvelis, 1991, p.566). Hence, a positive spread indicates that the market expects the economy to grow in the future, which would cause expansionary monetary policy, an increase in inflation and in base interest rate from the central bank, hence the yields of the 10 year bond would be much higher than the 3 month treasury bill. A negative spread instead indicates the exact opposite, the market expects a recession, contractionary monetary policy, possibly disinflation and a reduction in base interest rate (ibid.).

literature for both methodologies uses very similar model specifications since the theory behind the explanatory power of said variables is the same. Often the exact same indicators are used for both methodologies in the same paper, and despite some clear difference in the significance of each indicators and the accuracy of the forecasts, the result usually tend to show similar overall results (Estrella & Mishkin, 1998; Gogas et al., 2015; Valckx et al., 2002).

An example of this is Estrella and Hardouvelis' (1991) study, which showed that the yield spread was a significant estimator for GNP growth forecasts up to 16 quarters in advance using OLS, with the greatest explanatory power around 5-7 quarters, and up to 4 quarters in advance for the probability of recession using a probit model. Moreover, to test the true explanatory power of this estimators Estrella and Hardouvelis tested models with the yield spread and other indicators strongly connected to recessions[3]. For GNP forecasts, despite the significance of the additional indicators at shorter forecasting horizons, the coefficient and significance of the yield spread estimator did not change substantially (Estrella & Hardouvelis, 1991, pp.570-572). For the Probit model the yield spread remained significant while none of the additional indicators in the model proved significant (Estrella & Hardouvelis, 1991, p.572). This evidence, together with the limited gains in out of sample performance of models with additional variables compared with those with solely the yield spread, supported the articles' thesis this estimator alone had great potential for forecasting GNP or recession well into the future (Estrella & Hardouvelis, 1991, pp.574-575).

Since the yield spread alone proved to be a powerful indicator, other studies started including more explanatory variables into their recession forecasting models to better understand the explanatory power of the indicator and improve the forecasting power of their models. For instance, Wright (2006) included federal funds rate and return forecasting factor[4] to confirm that the yield spread was not simply a proxy for monetary policy. The results showed that the inclusion of monetary indicators slightly reduced the size of the coefficient of the yield spread, especially 2 quarter ahead forecasts, but did not affect the significance of the term (Wright, 2006, pp.13-14). Moreover, the study showed that the model with the best fit overall and the best prediction performance was one including the yield spread and the nominal federal funds rate (Wright, 2006, p.9).

[3] Full list in the appendix.
[4] This is a measurement of the premium that longer term bonds have that was first developed by Cochrane & Piazzesi (2005).

Estrella and Mishkin (1998), also attempted to generate a better recession model by comparing large number of variables[5] either individually or with the yield spread in probit models forecasting from 1 to 8 quarters ahead. The aim of this methodology was to produce the model with the best forecasting performance while maintaining the number of variables as small as possible, thus avoiding overfitting (Estrella & Mishkin, 1998, p.46). The model that produced the best results for every single horizon is composed by the yield spread and the NYSE composite price index (also known as the EM model) which shows that there is additional information regarding stock prices not included in the yield spread. However, it is important to notice that the overall explanatory power of this model is still low, with a maximum pseudo[6]-R^2 of 0.367 at 3 quarter horizon, and that at longer time horizon the difference between the yield spread alone and the EM model is very small. Hence the EM model could still be improved upon to avoid underfitting and producing better long-term forecasts.

This is exactly what Valckx, de Ceuster and Annaert (2002) did by including volatility terms in the form of mean absolute deviation of both the yield spread and NYSE composite index. The authors' aim was to provide empirical evidence that volatility can be a precursor of recession as it increased uncertainty, thus decreasing consumption and investments and possibly leading to an economic contraction (Valckx et al., 2002, pp.1–2). Moreover, the authors also included an 8 month[7] lagged term to the EM models to monitor if the other variables contribute additional information other than the previous state of the economy (Valckx et al., 2002, pp. 4–5). The model that had the best out of sample recession forecasting performance across all horizons taken into account (3,6,9,12 months in advance, with 9 months being the best most accurate) included all of the indicators mentioned above, but it still provided an unsatisfactory forecasting accuracy in terms of large numbers of false positives (Valckx et al., 2002, p.11). Moreover, the same model did not perform equally well when attempting to forecast GDP growth, essentially not improving the out of sample error of the simpler EM model (Valckx et al., 2002, p.12).

Another interesting variable to be included in recession forecasting models has been produced by Christiansen (2012), who included a dummy variable (DR-9) to indicate if three or more selected economies[8] where in a recession at the same time 9 months before the prediction is made. This proved to be a better indicator than the US yield spread alone up to 5 months

[5] Full list in appendix.
[6] Pseudo R^2 from Estrella (1998)
[7] This time period was chosen because usually 8 months are required on average to officially confirm the estimates of GDP and officially determine if a month was in an expansionary or contractionary period (Valckx et al., 2002, pp. 4–5).
[8] Australia, Canada, Germany, Japan, United Kingdom, United States

ahead forecasts, even though with a very small pseudo-R^2 with a maximum of 0.047. For longer forecasts the model including both the US yield spread and DR-9 became dominant (Christiansen, 2012, p.1040). Interestingly, however, the inclusion of the German yield spread into the variable model greatly increased the pseudo-R^2 across all horizons tested (3, 6, 9, 12 months in advance) to reach a maximum of 0.36 for the 6 and 9 months forecasts, and made DR-9 not significant except for the 9 and 12 months ahead forecasts (Christiansen, 2012, p.1042). This is an example of how an underfitting model can actually cause the variables to behave differently as shown by DR-9's change of explanatory power from the short run to the long run with the inclusion of the German yield spread.

However, as previously mentioned the inclusion of additional variables can cause overfitting and produce worse out-of-sample forecasts. For instance Filardo (1999), included AAA/BBB corporate bond spread and the Consumer Board's Leading Economic Index (LEI) into the EM model, but produced worse forecast accuracy compared to the EM model, missing 2 or 3 recession with a long-term horizon and also producing numerous false signals for shorter horizons. Therefore, this clearly shows that even though there is a benefit in increasing the number of variables, overfitting is very much an issue using a simple probit model.

A solution to overfitting comes from machine learning methodologies that use cross-validation to reduce the out-of-sample error as attempted by Gogas *et al.* (2015), who employed Support Vector Machines (SVM) with a radial basis function to produce a model that had a perfect record in forecasting recessions using a 1 quarter ahead horizon. The results of this study, however, are severely limited and hard to compare because of the extremely small test sample (2007:Q3 to 2011:Q4), which caused the results to be extremely skewed in favour of recession, predicting 14 recessionary quarter but only having a sample size of only 18 (Gogas et al., 2015, p.641). This meant that out of 10 total expansionary quarters only 4 were correctly qualified, which is equivalent to a specificity of 40%, and there were 6 False negatives, which can be quite troubling for any government attempting to use this model as guidance to determine the monetary policies to implement. Moreover, the results based on monthly predictions were even worse, always predicting growth for the whole test period (Gogas et al., 2015, p.643). The very low accuracy was probably caused by the limited type of explanatory variables used since the only indicators were bond interest rates of different length, which together with the lack of any forecasting past 3 months ahead, makes the results of the implementation of SVM in predicting recessions quite inconclusive.

A more comprehensive implementation of a machine learning methodology was conducted by Nyman and Ormerod (2016), who successfully implemented a random forest algorithm to

produce a model able to predict GDP growth 6 quarters in advance without a single false alarm or reduction in accuracy. However, also this approach has drawbacks: usually the start of the recessions is missed by a quarter or two, the size of both contractionary and expansionary periods is usually underestimated, and because the number of explanatory variables used is limited[9], it doesn't fully test the possibility that the models produced are underfitting (Nyman & Ormerod, 2016, pp. 6–9). Moreover, considering that random forest models are much harder to interpret since they do not produce sparse models and that most of the literature on the topic supports the existence of a linear relationship between GDP and common recession estimators, it would be more appropriate to also adopt a linear machine learning model (James, Witten, Hastie, & Tibshirani, 2013, p. 320). However, is also important to note that the results of Nyman and Ormerod's random forest model are more accurate than the Survey's of Professional Forecasters predictions 3 quarters ahead, thus strongly suggesting that considerable improvement using machine learning methods are possible.

[9] Full list in appendix

Section III: Data Collection Strategy (words 1434)

Two datasets were created for this study due to the different nature of the classification and regression models' dependent variable. For classification, a dataset with monthly observations was created since the NBER recession dates are officially set giving the first and last month of each recession (NBER, 2020). Instead, for regression, the dataset had quarterly observations since the GDP growth rate estimates are published every quarter (U.S. Bureau of Economic Analysis, 2020a). Both datasets contain the same set of independent variables, which is available in full in the appendix, with the exceptions of the lagged dependent variable terms.

Most of the variables taken into consideration in this study are available through the Federal Reserve Economic Data (FRED) portal maintained by the Federal Reserve Bank of St Louis and are provided by a range of departments and bureaus of the US government as reported in table 1 in the appendix. The exceptions to this are:

- The international business cycles dates, which are instead calculated by the Economic Cycle Research Institute as this is what is wildly used in the literature focused on non-US recessions (Christiansen, 2012; Nyberg, 2010; Schrimpf & Wang, 2010)
- The PMI index, Manufacturing Deliveries, and Manufacturing New Orders which instead are calculated by the Institute of Supply Management (Institute for Supply Management, 2020a).
- The S&P 500 index which was instead collected through yahoo financial data and is instead calculated by Standard and Poor's (Yahoo Finance, 2020).
- The Consumer sentiment index, which is instead calculated and provided by the University of Michigan (University of Michigan, 2020).
- The Manufacturers' Unfilled Orders for Durable Goods which was instead provided by the US Census Bureau (U.S. Census Bureau, 2020).

It's important to mention that the data from FRED consist of revised figures, and these get updated over time. Hence it is not possible to have a perfect snapshot of the original figures published at a certain time. The Archival version of FRED, ALFRED, which allows users to retrieve economic data as it was available on a specific date, unfortunately does not have an option to retrieve only the original estimate and the vintage range is much more limited. Hence, this model will assume that the revision of the FRED time series are generally unbiased and of minor entity, thus not causing major changes to the model results.

The first observation taken into account is January 1962 or Q1 1962 for quarterly data. This date was chosen because the daily data regarding bond yields, necessary to construct the mean absolute deviations measure of volatility described by Valckx, de Ceuster and Annaert, (2002), is only available up to the beginning of 1962. Some data from before January 1962 is also present in the form of 3 months or 6 months lagged values of the dependent variable. Hence, for the earliest observations, the lagged dependent variables are actually from July 1961 or Q3 1961. All the independent variables' time series were collected until December 2018 or Q4 2018, while the dependent variable data was collected until December 2019 or Q4 2019. This was done to have the largest amount of available data while also avoiding having the effect of a black swan event, such as the Covid-19 pandemic, affect the accuracy of the model. Moreover, the discrepancy between the dependent and independent dataset length is caused by the need to have one additional year of data regarding independent variables to make predictions with a horizon of a year. Hence, if we want to test the accuracy of 12 months ahead prediction made in December 2018 or Q4 2018, it is necessary to have the dependent

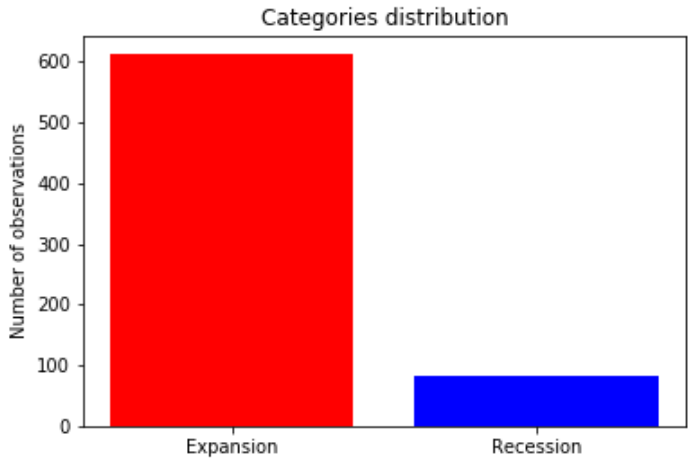

Figure 1

variables until a year later. This produces a dataset with 7 distinct recessionary periods totalling 83 recessionary months out of 684 observation, which is quite unbalanced as shown in figure 1. This can cause issues in measuring the accuracy of the predictions as a model might achieve extremely high accuracy simply by always predicting the dominant class. Unfortunately, class rebalancing techniques commonly used to address these issues, such as oversampling or stratified cross validation are not applicable with time series, as it would involve using future data to predict previous events, which defeats the purpose of the prediction. Therefore, instead of rebalancing the data, this study will use classification scoring methods such as Log Loss and Quadratic Probability Score, that take into account the prediction probability rather than the accuracy of the classification (Moneta, 2005, p. 280)

All the time series that use dollars as units were collected in nominal terms to assure that the inflation correction was conducted using the same base year and CPI deflator[10]. For this study 2012 was chosen as base year and the Consumer Price Index for All Urban Consumers was chosen as deflator (U.S. Bureau of Labor Statistics, 2020b). This choice was also dictated by the date range of the time series as other deflators were either discontinued or started to be calculated later than the earliest measures taken into consideration. Some of the data using dollars or millions of dollars as units were also changed to billions of dollars to simplify interpretation, table 1 in the appendix provides the original unit of measure for all the time series taken into consideration. Lastly, some time series with strong seasonal effects, in particular those related to labour, were also seasonally corrected to avoid the noise caused by the various fluctuations as reported in table 1.

The various time series also differed due to their frequency of publication. Some, such as the yield spread or the S&P 500 composite, were published daily and were, therefore, averaged to produce monthly results. Most of the series was instead published monthly, hence these series were averaged to produce the quarterly data set but were left untouched for the monthly dataset. The Consumer Sentiment Index instead had a slightly more complicated calculation for the monthly data as it was released on a quarterly basis until January 1978 when it started being released on a monthly basis. Therefore, the monthly time series has been constructed by repeating the quarterly measurements for each month of the quarter until January 1978.

Since only the data which is available at the time of prediction should be taken into consideration for the model, the publication lag of each time series also plays a crucial factor in the construction of the dataset. Some of the time series, for instance bond yields and stock composite index, can be calculated in real time due to the fact that they are available on a daily basis. Other series, however, have one or two months of publication lag due to the fact that some time is required once the month ends to calculate and publish the data. For instance, let's say that at the end of July 2015 an investor wants to predict the probability of recession for August 2015 using the PPI of copper base scrap. This estimate for July has not been calculated yet, hence the best the investor can do is to use the June estimate. This would be considered as a 1-month publication lag. Hence, the data has been shifted accordingly for the series that require this correction[11]. A separate discussion is needed for the publication lag of the International business cycle dates because they vary substantially (NBER, 2020). This study will assume that the date has a nine months publication lag since these series are not

[10] Some time series, such as the various money bases, are present both in the real and nominal version as both were used in the cited literature.
[11] Table 1 in appendix has the exact lags for each series.

used individually in the model specification but are combined to create the simultaneous recession variables as described by Christiansen (2012), which uses this lag length in her study.

Despite the aim to include most if not all the economic and financial indicators used in the cited literature, certain series were not possible to include due to issues with the length of the time series or their availability. A list of these series is included in the appendix together with the reason for their exclusion. Lastly, despite some of the cited literature including leading indicators in the model specification, I decided not to include them because the components of these indicators often overlap in part or completely with the individual time series which are going to be included in the model specification (Filardo, 1999, p. 37). Hence, to include them it could cause these indicators to reduce the explanatory power of the individual components, without however having the same granular understanding of using the individual indicators. Moreover, the leading indicators can also provide an interesting comparison of the prediction accuracy of the model. Therefore, I will use the Survey of Professional Forecasters' Leading Economic Indicator as one of the benchmark models to compare the performance of the model specification created by the shrinkage methods.

Section IV: Data Analysis Methodology (1674, 19 already used)

As mentioned in Section 1, this study implements shrinkage methods, which is a family of techniques that fit models while shrinking the coefficients of certain explanatory variables towards zero to reduce the model's variance (Hastie et al., 2009, p. 61). The shrinkage is conducted by a penalty term (such as an L1 penalty[12], and L2 penalty[13], or a combination of both) added to function to be minimized, for instance the OLS residual sum of squares (ibid, p. 68). The strength of the penalty is determined by a tuning parameter λ, which is selected during the cross validation to minimize the validation error (James *et al.*, 2013, p.219). One of the most commonly used shrinkage methods is the Lasso, which is defined by minimizing the following equation for all $\beta_j \in R^p$:

$$f(\beta_i, \dots, \beta_p) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\,\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

which is just the OLS residual sum of squares with the addition of a L1 penalty term (highlighted in green). Shrinkage methods can be applied to any linear model by using

[12]This is the absolute value of the sum of the coefficients.
[13] This is the squared value of the sum of the coefficients.

Generalized Linear Models (GLM). Hence, the equation to minimize a logistic regression with an L2 penalty term would be:

$$f(\beta_0, \beta) = \frac{1}{n}\sum_{i=1}^{n} l(y_i, \beta_0 + \beta^T x_i) + \lambda(\beta_j)^2$$

With $l(y_i, \beta_0 + \beta^T x_i)$ being the negative binomial log-likelihood and λ the tuning parameter computed using cross-validation (Hastie & Qian, 2014).

Before, applying any of these methods however, normalization was conducted on all the explanatory variables to avoid having large variables, such as the various money bases, overly penalised by shrinkage methods not because of the correlation to the dependent variable but simply due to their unit of measurement. This is true also for logistic regression, which can be performed without normalization even with input variables of different scales, however when including an L1 or L2 penalty this is not the case as the solutions are not scale invariant (Hastie et al., 2009, pp. 63–64). As a consequence, all the input variables for both the monthly and quarterly dataset were also standardized before running the logistic regression models. This also has the added benefit of improving the computation efficiency of the algorithm as it should allow for faster convergence (ibid.).

The λ parameter could not be determined using standard cross validation techniques, such as standard k-fold cross validation, due to the fact that these techniques mix the order of the

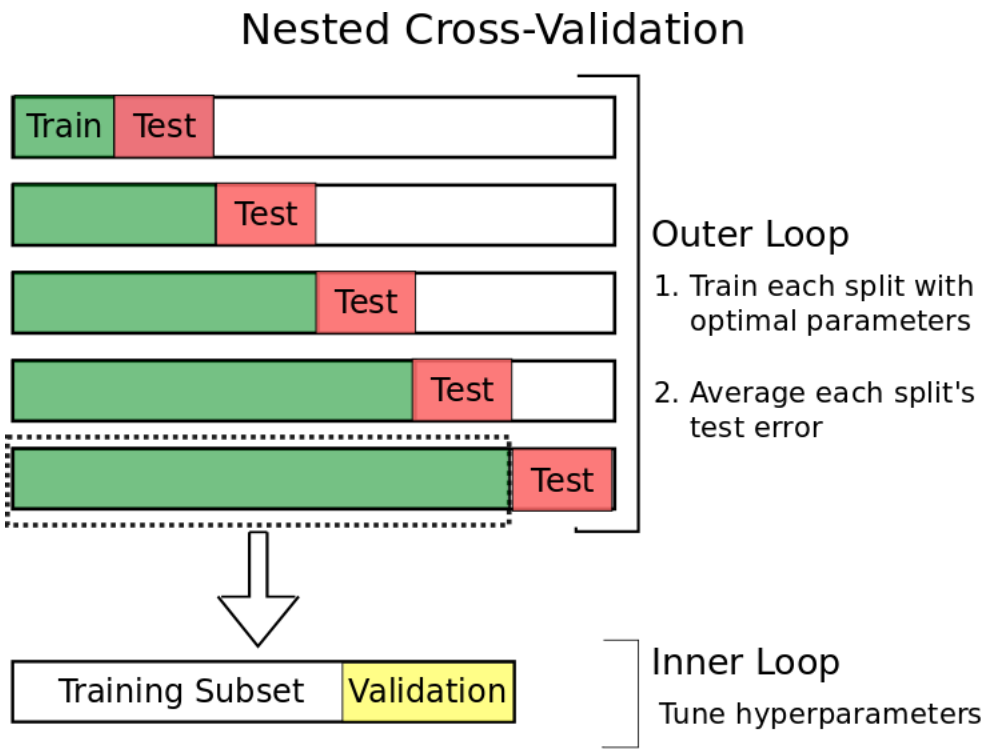

Figure 2

data, which would cause future data to be implemented to predict past events. Hence, nested cross validation was used instead, this process is clearly represented in figure 2[14]. Nested cross validation consists in progressively shifting the validation and test set along the time series, while incorporating the previous validation and test sets into the training set. Hence, a model is constructed for each inner loop and the error is then calculated against its individual test set. When the whole outer loop has been completed the test error can then be averaged to produce an estimate which is almost unbiased (Varma & Simon, 2006). Despite ideally wanting to use as much data as possible for the testing process, this study will use the data up to January 1990 or Q1 1990 uniquely as training data, with testing starting from February 1990s or Q2 1990 and shifting one month or quarter at the time for each subsequent inner loop. This produces a training set containing four recessionary periods for the training set and three recessionary periods for the test set.

To tune λ for each inner loop the training set was divided into the training subset and the validation subset. This division was conducted either by using only the most recent observation in the training set as the validation set (single nested cross validation), or by conducting a 10-fold secondary nested cross validation (double nested cross validation). In other words, the training set was divided into 10 equally sized time periods, with the validation set initially being the second period and shifting progressively after each model is fit as shown in figure 2. These two validation methods were chosen because the first one would tune a λ that would fit best only the most recent data, which should produce the most accurate predictions of the following observation. Instead, the secondary nested cross validation would tune a model that should better represent the overall data, at the expense of producing the most accurate predictions.

As the cited literature conducted predictions over various horizons, this study will also predict quarter over quarter GDP growth rates and the probability of recession for the current observation time period and for observations 3, 6, 9, and 12 months ahead. These forecasts horizons were chosen because they would coincide for both the monthly and quarterly dataset and these were also some of the most frequently used intervals in the literature (Christiansen, 2012; Nyman & Ormerod, 2016; Valckx et al., 2002). Moreover, the Survey of Professional Forecasters also uses the same forecasting horizons but it only produces predictions up to a year ahead (Federal Reserve Bank of Philadelphia, 2020). Hence producing predictions further than 12 months ahead, as done by Estrella and Hardouvelis (1991), would mean to deprive the prediction of one of the benchmark models tests.

[14] This figure was taken from (Cochrane, 2018).

Since multiple forecasting horizons were predicted at the same point in time, multiple output regression or classification were needed. Not all the scikit-learn functions[15] used in this study's analysis natively supported multi target results, hence I preferred to implement a wrapper function to consistently produce results using the same method. Each subsequent forecast could either be considered as independent results or could also be taken into consideration together with all the other available independent variables to predict the subsequent forecast. Therefore, both methods were attempted to assess which wrapper function would produce the best out-of-sample performance.

The out-of-sample performance was calculated with different measures depending on the dependent variable used in the model:

For the classification model, most of the cited literature uses the pseudo-R_2 developed by Estrella (1998), which should closely match the interpretation of the usual OLS R_2. However, the coefficient of determination is not a very accurate out-of-sample performance indicator, especially in an unbalanced prediction task, as a model that always predicts expansionary periods would actually have a very high R_2. log loss and quadratic probability score (QPS) will be used instead, since these measures take into account not only the actual classification but also the level of certainty of the predictions. The first is a very commonly used metric in machine learning for the test error performance of classification task, while the second was developed by Diebold & Rudebusch (1989) with the explicit aim of measuring the accuracy of probability predictions. For both metric the more accurate the model the closer the score will be to zero, however for QPS the worst possible score is 2, while log loss does not have an upper bound. This study will also take into account overall accuracy, sensitivity[16], and the number of false positive and false negatives for each different forecasting horizon as in a real-life application these numbers are crucial in determining whether a government or a private investor would trust the model's results more than a machine learning metrics like QPS or log loss.

Instead, for the regression model, this study will use the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE), two very commonly used test error metrics that were also used in the cited literature (Christiansen, 2012, p.1039; Estrella & Hardouvelis, 1991, p.573). Both measures essentially take an average of the error between the predicted value and the

[15] In particular sklearn.linear_model.LogisticRegressionCV and sklearn.linear_model.LassoCV did not support multi target results.

[16] Also known as recall or true positive rate (TP/TP+FN)

real test values. Hence for both metrics the closer the value is to zero the more accurate is the model's out-of-sample result. These two metrics differ from the penalty inflicted to larger errors, with the RMSE providing a much harsher penalty since each error is squared. Moreover, for the regression models the study will also include the R_2 to have a better idea of how much variation is actually explained by the model. Lastly, for both the classification and regression model there is also going to be a visual assessment using graph to better understand the actual output of the models.

Lastly, for the classification model it could be possible to tweak the threshold probability that determine when a specific observation is classified as a recession to improve the out-of-sample performance as done in the literature (Filardo, 1999; Valckx et al., 2002). However, this study will not alter the default threshold of 50% probability because any tweak can is essentially an alteration using the test data to fit the data better, thus undermining the whole concept of out-of-sample testing.

As previously mentioned, two benchmark models were also constructed to compare the performance of the shrinkage methods. The first benchmark is simply a model with the same specification and methodology detailed in Estrella and Mishkin (1998). Hence, this model will includes only the 10y-3m yield spread and the Real S&P500 composite index[17] as explanatory variables and will be fitted using logistic regression for the classification model and OLS for the regression model. The second benchmark is the predicted probability of recession or the predicted GDP quarter over quarter growth rate published by the Survey of Professional Forecasters (SPF). Since the SPF directly publishes their forecast figures there is no need for additional model fitting, however the probability of recession predicted by the survey is published on a quarterly rather than monthly basis. This meant that to calculate the performance metrics there was a need for a quarterly true recession value. Therefore, the US Recession monthly dates were altered by taking only the end value (in this case the last month of each quarter) to create a quarterly series. The end value was used instead of an average as it was done for other indicators due to the need of maintaining the binary indicator for the performance metrics to be comparable.

[17] Estrella and Mishkin (1998) actually use the NYSE composite index rather than the S&P500, however due to the issues in finding the NYSE historical data before 1965 it was not possible to do the same.

Section V: Results

Part 1: Classification Model

The first model to be tested was fitted using a logistic regression classifier with an L1 penalty with a single nested cross validation (SNCV) using a wrapper function with independent predictions of the various horizons (Model 1). The L1 penalty was the first one to be tested because, differently from L2, it allows for the shrinkage of the independent variables to zero, hence It should be particularly effective since there are 49 independent variables in the model.

Table 1:

Model 1	**QPS**	**Log Loss**	**Accuracy**	**Sensitivity**	**False Pos**	**False Neg**
hrz_0	0.099144	0.221543	0.968300	0.764706	3	8
hrz_3	0.117198	0.264280	0.936599	0.411765	2	20
hrz_6	0.140130	0.275449	0.927954	0.294118	1	24
hrz_9	0.144356	0.285472	0.925072	0.225806	2	24
hrz_12	0.134005	0.267977	0.933718	0.250000	2	21

As you can see from table 1 the results are already quite encouraging with both QPS and log loss quite close to 0 and with only minor increments as the horizon is expanded. Interestingly the 12 months ahead forecast has a slightly better out-of-sample performance than even the 6 and 9 months ahead forecast. When considering the number of False positive and False negatives is quite encouraging to see that of the 335 tested observations only, there were never more than 26 misclassified months, which means an accuracy always over 92.5%, with also an extremely low number of False positive. Unfortunately, the number of False Negative is never below 20 for any forecast other that the current observation, which is also reflected by the dramatic worsening of the sensitivity score from the 3 months horizon onwards. Despite the good overall accuracy and out-of-sample performance of model 1, only a minority of the recessionary months are incorrectly predicted at any forecasting horizon other than the current observation.

An improvement to model 1 could be the use of the double nested cross validation (DNCV) instead of SNCV with all the other specification of model 1, as this type of CV validation should provide a result with a better overall representation of the data. As you can see from table 2, Model 2 has an improvement across all horizons in terms of QPS and log loss, thus showing that compared to Model 1 the classification is conducted with a higher degree of certainty. However, as you can see from the greater number of False negative and the worse sensitivity for all horizons (except the 3 months ahead forecast) the majority of the recessions are still

being misclassified with especially the 9 months ahead forecast almost always predicting expansionary months. Table 2:

Model 2	**QPS**	**Log Loss**	**Accuracy**	**Sensitivity**	**False Pos**	**False Neg**
hrz_0	0.056963	0.109837	0.959654	0.647059	2	12
hrz_3	0.102512	0.203227	0.933718	0.500000	6	17
hrz_6	0.121361	0.226909	0.925072	0.235294	0	26
hrz_9	0.117690	0.183779	0.916427	0.096774	1	28
hrz_12	0.108121	0.178051	0.927954	0.250000	4	21

This can also be seen from figure 3, where the 9 months horizon forecast barely clears the 50% threshold throughout the test period taken into consideration.

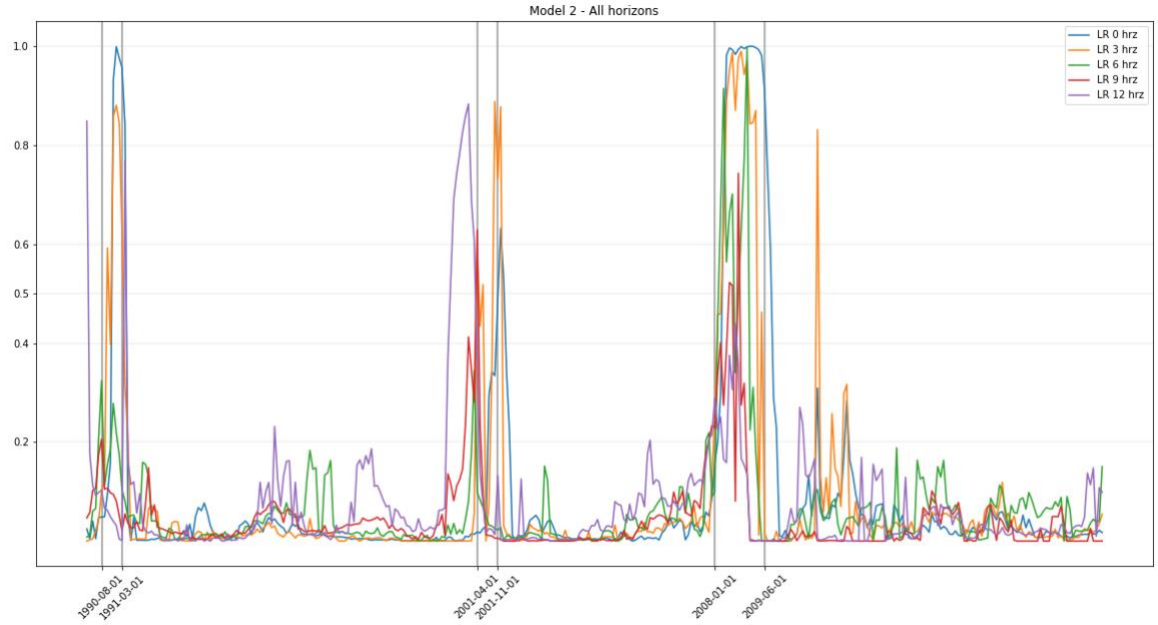

Figure 3

From Figure 3 it is also possible to notice how despite the high number of False Negatives and low sensitivity all the various horizon forecasts of model 2 have a spike in the probability of recession during or close to the actual recession dates. The only exception seems to be the 3 months ahead forecast which also has a spike in recession probability around 2011, probably due to some indicators more strongly affected by the global market as many other nations around the world suffered a recession during that period. From figure 3 is also possible to notice the fact that, with the exception of the 12 month forecast, the increases in the recession probability do not seem to be leading the recessions. Undoubtedly, further improvements are needed to produce a result that could be used to reliably predict recessions in advance.

One possibility to further improve model 2 is the implementation of the chained classifier wrapper function to take into account all the previous prediction in the chain while making the prediction for the following horizon, which should allow for the later predictions to build upon the information provided in the previous forecasts in the chain.

Table 3:

Model 3	**QPS**	**LogLoss**	**Accuracy**	**Sensitivity**	**False Pos**	**False Neg**
hrz_0	0.056949	0.109842	0.959654	0.647059	2	12
hrz_3	0.131900	0.266133	0.927954	0.441176	6	19
hrz_6	0.145516	0.253569	0.913545	0.205882	3	27
hrz_9	0.147503	0.252476	0.904899	0.096774	5	28
hrz_12	0.140047	0.251673	0.922190	0.071429	1	26

It is quite clear by the results reported in table 3 that for every single metric taken into consideration the out-of-sample metrics of model 3 are worse of both model 2 and even model 1. Clearly the chained wrapper function did not produce the desired effect despite the potentially greater amount of information available using the chained classifier wrapper function.

Another possibility to improve the results is to use an L2 penalty rather than an L1 penalty for the Logistic regression classifier, which instead of producing sparse models progressively shrinks the coefficients towards zero, without actually reaching the value (Hastie et al., 2009, p.63). Model 4 was constructed with the same specification as model 2 but with an L2 penalty instead of an L1. As you can see in table 4, model 4 has comparable QPS score and log loss score to model 2 for the most part with the two longest horizons exhibiting even lower out-of-sample error metrics than the previous model.

Table 4:

Model 4	**QPS**	**LogLoss**	**Accuracy**	**Sensitivity**	**False Pos**	**False Neg**
hrz_0	0.058890	0.114602	0.965418	0.705882	2	10
hrz_3	0.096467	0.177474	0.933718	0.441176	4	19
hrz_6	0.118933	0.212597	0.919308	0.235294	2	26
hrz_9	0.101401	0.163641	0.933718	0.322581	2	21
hrz_12	0.100146	0.156906	0.927954	0.250000	4	21

At the same time, model 4 also has a higher sensitivity than and less False Negatives than model 2 and for certain horizons (3 and 6 months ahead) even lower than model 1. Checking

the results graphically from figure 4 it seems to show that with model 4's specification all horizons clear the 50% threshold at least once during each recessionary period. Moreover the 12 and 9 months ahead predictions are rapidly increasing before the recession, hence showing the recession leading characteristics necessary to predict recessions in advance.

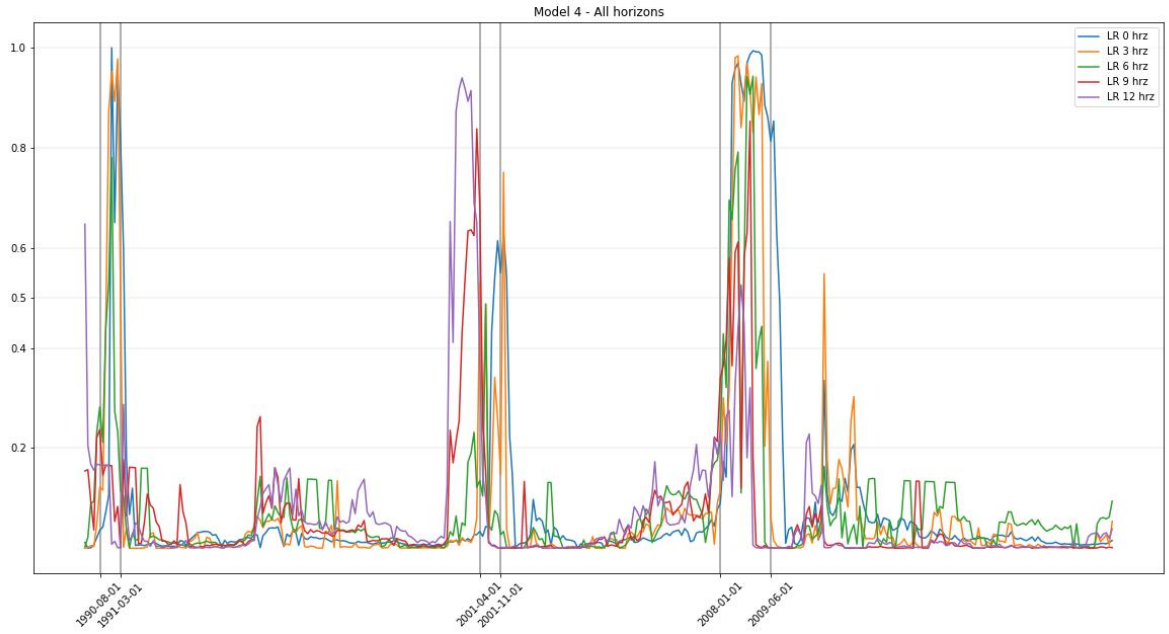

Figure 4

Of all the models tested, model 4 seems to be the best performing both in terms of out-of-sample metrics and for accuracy and timeliness of recession predictions. This study will now turn towards the benchmark models to see how model 4 compares.

Part 2: Classification Benchmark Comparison

The first model benchmark model is the EM model run using a simple Logistic Regression defined by Estrella and Mishkin (1998). However, a wrapper function is still necessary for the multiple horizon fitting. Therefore, this study conducted both the independent horizon prediction and the chained horizon predictions. The results are reported in table 5.

Table 5:

EM independent	**QPS**	**LogLoss**	**Accuracy**	**Sensitivity**	**False Pos**	**False Neg**
hrz_0	0.173002	0.311130	0.902017	0.0	0	34
hrz_3	0.174592	0.314473	0.902017	0.0	0	34
hrz_6	0.170527	0.300855	0.902017	0.0	0	34
hrz_9	0.148528	0.261011	0.910663	0.0	0	31
hrz_12	0.129941	0.234101	0.919308	0.0	0	28

EM chained	QPS	LogLoss	Accuracy	Sensitivity	False Pos	False Neg
hrz_0	0.172985	0.311089	0.902017	0.0	0	34
hrz_3	0.185847	0.372080	0.902017	0.0	0	34
hrz_6	0.178769	0.328429	0.902017	0.0	0	34
hrz_9	0.158290	0.282216	0.910663	0.0	0	31
hrz_12	0.141645	0.253210	0.919308	0.0	0	28

As clearly shown in table 5, the out-of-sample metrics are worse than the model 4 one for both instances of the EM model. However, the biggest issue is the fact that all the recessionary periods were classified as expansionary, thus sensitivity to be equal to 0. Only by graphically checking the probabilities, as shown in figure 5, is possible to see that the issue is caused by the recession probability spikes that happen concurrently with the recession never pass the 50% threshold necessary to be classify the observation as a recessionary period. With a different threshold, maybe even as low as 15%, some of the longer horizons forecast could produce observations classified as recession. However, in doing so it would also probably cause, or risk causing in the future, a higher level of false positives, which is usually inevitable when a model has so little fluctuations.

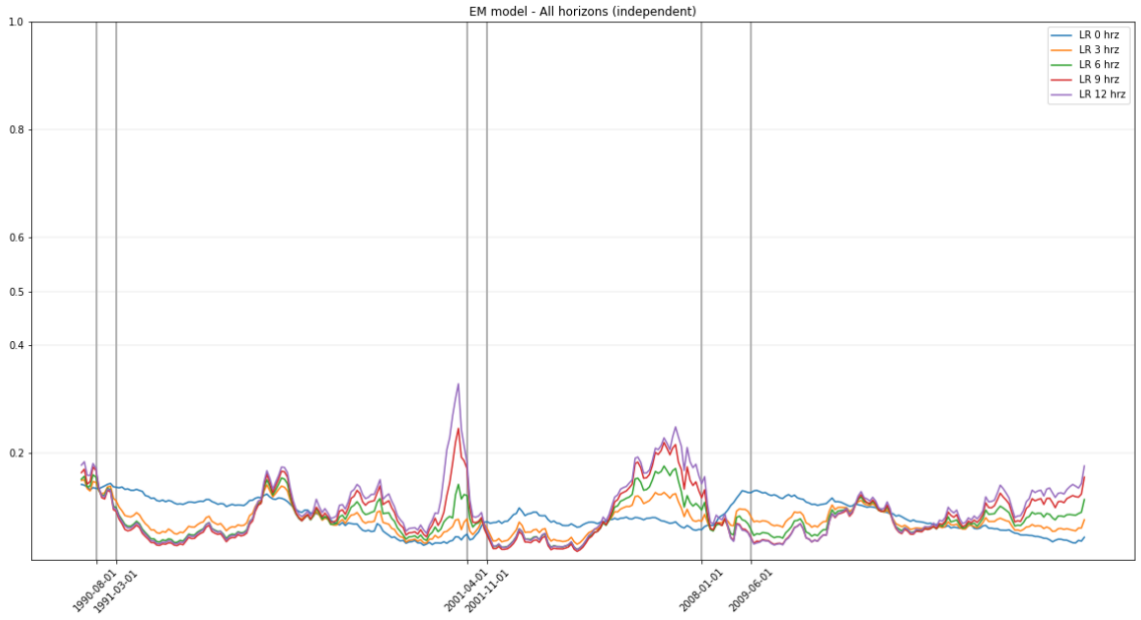

Figure 5

However, at least for the two most recent predictions, there is a clear increment in the probability of recession for the 3, 6, 9, and 12 months ahead predictions well before the recession starts followed by the increment of the current observation prediction during the recession. This is quite impressive considering the simplicity of the model and the speed it can be calculated. It definitely provides a good starting point for more complex model. However, in this case due to the better out-of-sample performance metrics, greater sensitivity,

and larger fluctuation, model 4 can be considered to perform better than the EM model in classifying recessionary and expansionary months.

The second benchmark model is the SPF probability of recession estimate. This estimate only provide a probability, without any classification for each observation. Hence, this study employed the same classification threshold level as the previously constructed models (50%). The classification is necessary to calculate the accuracy and sensitivity measurements. The results are reported in table 6

Table 6:

SPF	**QPS**	**LogLoss**	**Accuracy**	**Sensitivity**	**False Pos**	**False Neg**
hrz_0	0.135475	0.235617	0.896552	0.272727	4	8
hrz_3	0.114391	0.225481	0.922414	0.272727	1	8
hrz_6	0.123611	0.244609	0.913793	0.090909	0	10
hrz_9	0.148526	0.279307	0.905172	0.000000	0	11
hrz_12	0.171475	0.314259	0.905172	0.000000	0	11

Since the SPF results are based on quarterly observations rather than monthly, the false positive and false negative number are not comparable as there are only one third of the observations for the SPF test set. Nevertheless, by comparing the QPS and log loss to the results from table s4 it seems that model 4 has a marginally better out-of-sample performance. The superior performance of model 4 is ever clearer by checking the sensitivity and the probability graph in figure 6.

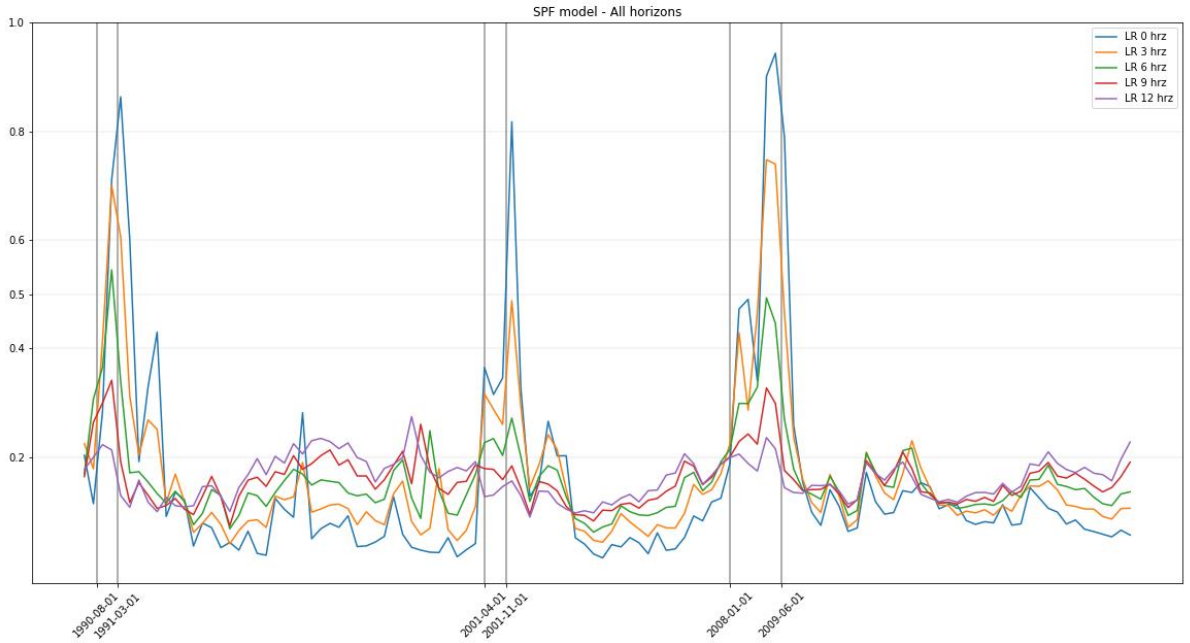

Figure 6

Essentially, only the two predictions with the shortest horizons are able to clear the threshold that enables the model to classify quarters as recessionary. In addition, the spikes in probability, instead of leading the recessions, seems to be lagging, which makes it impossible to use the SPF results to detect possible recession with any kind of warning. Therefore, the impossibility of the SPF predictions to anticipate recessions together with the worst performance of the out-of-sample performance indicators shows that the model 4 is a better performing model.

Part 3: Regression Model

The first model to be tested was an OLS regression with an L1 penalty term (also known as the LASSO), using SNCV and a wrapper function with independent predictions of the various horizons (Model 5). Once again, this study started from the L1 penalty because of the possibility of generating sparse models, which is a desirable property considering the large number of explanatory variables. The results are reported in table 7.

Table 7:

Model 5	**MAE**	**MSE[18]**	**RMSE**	**R_2**
hrz_0	0.435990	0.372240	0.610115	-0.082010
hrz_3	0.426890	0.301336	0.548941	0.123321
hrz_6	0.451163	0.355218	0.596002	-0.040828
hrz_9	0.430914	0.340213	0.583277	-0.060214
hrz_12	0.452189	0.388188	0.623048	-0.250091

The results are not very encouraging to begging with, due to the presence of significant negative R_2, which suggests that the model fits the data so poorly that using a horizontal line with a value equivalent to the mean value of GDP growth rate would produce better results. The only forecasting horizons that seems to provide some explanatory power is the 3 months ahead prediction, which coincidently also has the lowest out-of-sample prediction metrics. Nevertheless, an improvement is necessary to produce a more reliable predictor of GDP growth rate.

As before with the classification method, the next step could be to implement DNCV instead of SNCV, while maintaining all the other parameters the same as model 5.

[18] The Mean Squared Error (MSE) is not really used as an out-of-sample performance metric in this study as it conveys almost the same information as the Root Mean Squared Error (RMSE). Nonetheless, I included this measure in the regression results tables as it is sometimes preferred to RMSE.

As shown by table 8, model 6 produces better overall results for the first two forecasting horizons (0 and 3 months ahead), but it still suffers from the issue of having the remaining three forecasting horizons with a negative R_2.

Table 8:

Model 6	**MAE**	**MSE**	**RMSE**	**R_2**
hrz_0	0.386625	0.251968	0.501964	0.267591
hrz_3	0.406987	0.301116	0.548741	0.123961
hrz_6	0.437207	0.377307	0.614253	-0.105550
hrz_9	0.417459	0.344932	0.587309	-0.074921
hrz_12	0.412623	0.349103	0.590849	-0.124223

This is probably due to the lack of fluctuations that can be seen in figure 7. In fact, the predicted GDP growth rate quarter over quarter seems to reduce the width of the fluctuations the longer the interval of prediction, with the 12 months ahead forecasts even having periods of stationarity around the 0.75% growth rate even during the 2008 financial crisis. It is, therefore, no surprise the R_2 is negative considering these results.

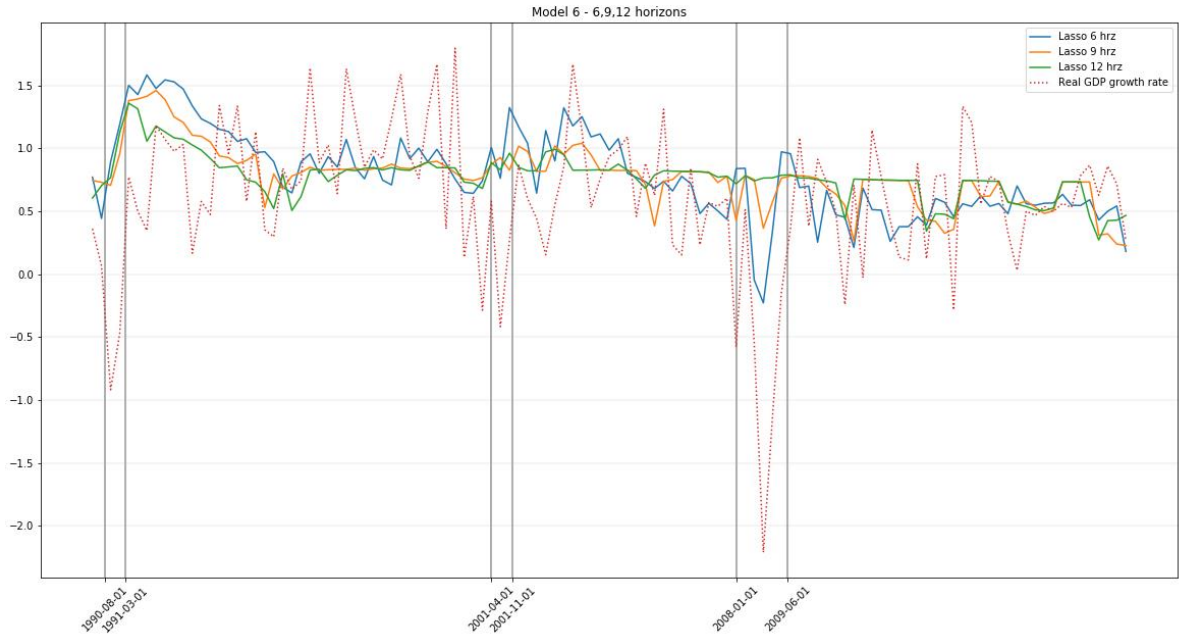

Figure 7

Despite this negative result, both the MAE and RMSE exhibit improvements across all the forecasting horizons taken into account. In fact, by graphically checking only the predictions of only the 0 and 3 months ahead predictions (figure 8) it is clear that for these horizons there is some out-of-sample prediction power, even though the fluctuations remain less extreme than the real GDP growth rate.

Nonetheless, further improvements are needed to have a satisfactory GDP growth rate prediction model across as many horizons as possible.

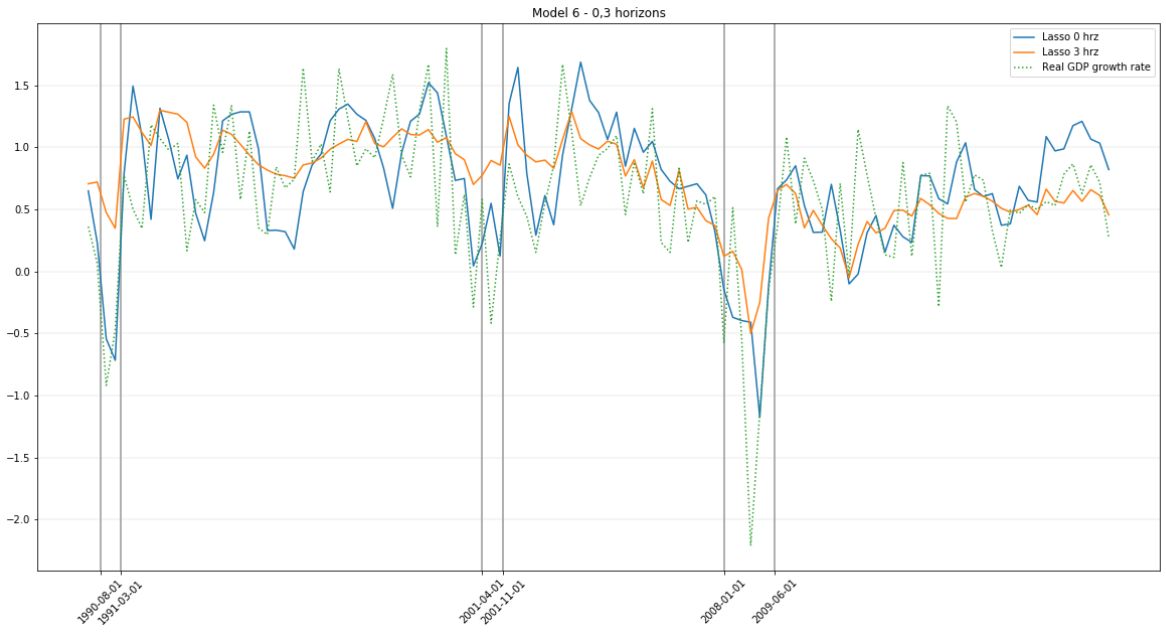

Figure 8

For this purpose, model 7 implements the chained regressor wrapper function to test the possibility that taking into account the previous growth rates could produce better overall results, especially for the longer forecast horizons. As you can see from table 9, there is a marginal improvement for MAE, RMSE, and R_2 for all horizons taken into consideration with the exception of the three months ahead prediction.

Table 9:

Model 7	**MAE**	**MSE**	**RMSE**	**R_2**
hrz_0	0.386625	0.251968	0.501964	0.267591
hrz_3	0.411825	0.314128	0.560471	0.086107
hrz_6	0.427750	0.359723	0.599769	-0.054026
hrz_9	0.391515	0.327609	0.572371	-0.020936
hrz_12	0.389015	0.318861	0.564678	-0.026835

Nonetheless, the three 6, 9, and 12 months ahead forecasts remain a worse fit than a horizontal line through the mean of the test values, essentially providing no useful information for the prediction. This can also be seen from figure 9, where the 6, 9, and 12 months ahead forecasts only have minor fluctuations even during recessionary periods. However, the 0 and 3 months ahead predictions definitely follow the peaks and through of the real GDP growth rate more closely.

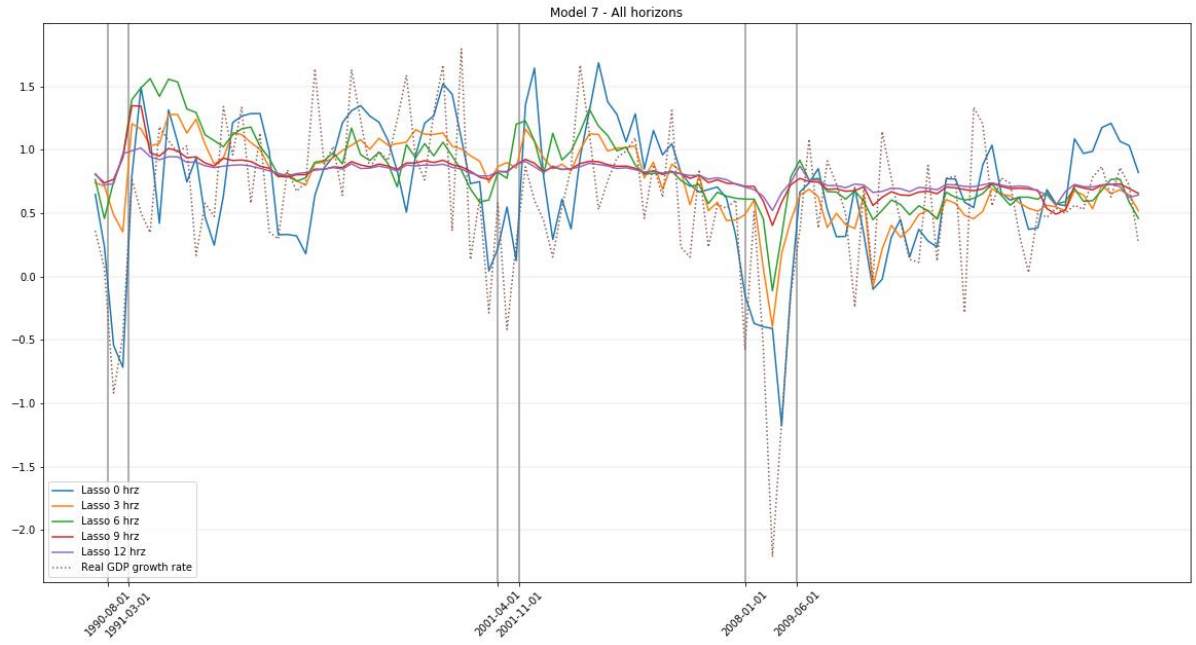

Figure 9

Therefore, model 8 changed the penalty term used in the regression from a L1 to and L2 penalty, thus conducting a Ridge Regression using the same specification and parameters as model 7. As in the case of classification, it is possible that a non-sparse solution provide a better prediction performance.

Table 10:

Model 8	**MAE**	**MSE**	**RMSE**	**R_2**
hrz_0	0.443285	0.321588	0.567087	0.065222
hrz_3	0.458613	0.331343	0.575624	0.036022
hrz_6	0.475667	0.398617	0.631361	-0.167990
hrz_9	0.467115	0.385249	0.620684	-0.200562
hrz_12	0.454610	0.386371	0.621588	-0.244240

Unfortunately, as shown by table 10, this is not the case with all the performance metrics being worse than the ones produced by model 7. Hence, of all the models tested, model 7 seems to be the best performing both in terms of out-of-sample metrics and regarding the goodness of fit. This study will now turn towards the benchmark models to see how model 7 compares.

Part 4: Regression Benchmark Comparison

The EM model specification using a simple OLS is the first benchmark model tested. As with the classification a wrapper function needs to be implemented for the multiple output regressions. Hence, both the independent and chained horizons wrapper functions are reported in table 11.

Table 11:

EM independent	**MAE**	**MSE**	**RMSE**	**R_2**
hrz_0	0.510022	0.491769	0.701263	-0.429454
hrz_3	0.523580	0.521184	0.721931	-0.516282
hrz_6	0.510275	0.508273	0.712933	-0.489295
hrz_9	0.460901	0.404167	0.635741	-0.259516
hrz_12	0.438066	0.354638	0.595515	-0.142048
EM chained	**MAE**	**MSE**	**RMSE**	**R_2**
hrz_0	0.510022	0.491769	0.701263	-0.429454
hrz_3	0.523580	0.521184	0.721931	-0.516282
hrz_6	0.510275	0.508273	0.712933	-0.489295
hrz_9	0.460901	0.404167	0.635741	-0.259516
hrz_12	0.438066	0.354638	0.595515	-0.142048

The most striking result is that the two different wrapper functions produced exactly the same out-of-sample performance metrics, thus the exact same predictions were made. This could only be possible if the previous horizons' predictions did not contain any additional information. However, considering the high values of the MAE and RMSE together with the negative R_2 for all horizons, it might be the case that the possibility of using the data relative to previous predictions did not affect the EM models GDP growth rate forecasts. Table 11 also clearly shows that model 7 has a better out-of-sample performance than the EM model.

The second benchmark model is the SPF GDP growth rate prediction. The SPF actually publishes the predicted GDP for each quarter rather than the quarter over quarter GDP growth rate. Nevertheless, it is fairly straight forward to calculate the growth rate as shown in the code[19] The out-of-sample performance of this model is reported in table 12

Table 12:

SPF	**MAE**	**MSE**	**RMSE**	**R_2**
lag_0	0.558409	0.485463	0.696752	-0.411124
lag_3	0.626697	0.624711	0.790387	-0.817474
lag_6	0.666248	0.699514	0.836369	-1.049649
lag_9	0.678905	0.708930	0.841980	-1.209259
lag_12	0.681429	0.709566	0.842357	-1.285029

[19] The code is available from https://github.com/Davide-Bestagno/Master_Dissertation_Public

Table 12 clearly shows that the SPF out-of-sample performance is worse than the model 7 performance according to every single metric and for every single horizon. In fact, the SPF out-of-sample performance is even worse than the EM model performance. As you can see in greater detail from figure 10, this is due to the constant overestimation of the GDP growth rate as well as the lack of any anticipation of the recessions. Similarly, to the SPF's recession forecasts, even the SPF's GDP growth rate forecasts seem to be lagging behind the recessions. Overall, it is clear that model 7 outperforms the benchmark models according to every single metric taken into consideration. Nonetheless, the results of the implementation of shrinkage method in forecasting quarter over quarter GDP growth rates is not as satisfactory as for the classification problem.

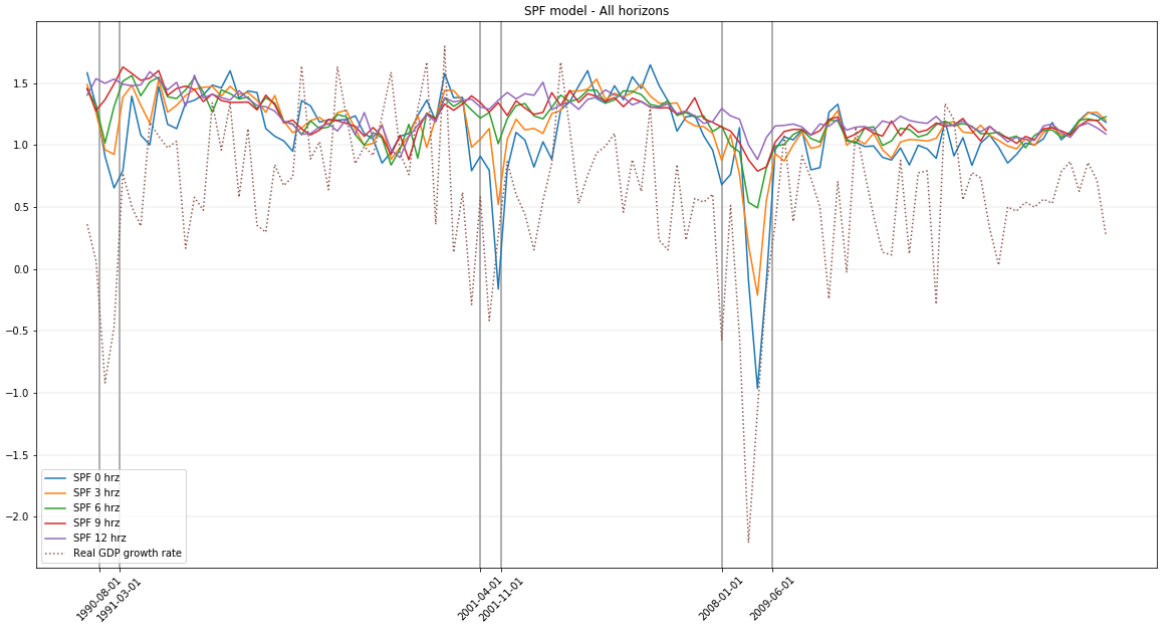

Figure 10

Section VI: Conclusion

The shrinkage methods clearly outperformed both benchmark models taken into consideration for both the recession forecasting and GDP growth rate prediction problem. However, the absolute performance in the two different problems was extremely dissimilar.

For the classification problem, model 4's specification detected the all three periods of recessions at every forecasting horizon taken into consideration with a minimal number of false alarm. Moreover, both the 9 and 12 months ahead predictions were partially leading the recessions, which would have given some time to the government or individual investors to act in preparation of at least the 1990 and 2001 recessions. The relatively low QPS and log loss scores for all horizons also shows how model 4 was accurately classifying the test observation with a high degree of confidence. Lastly, the model seemed to maintain the same level of accuracy for all the out-of-sample metrics as the prediction horizon lengthened, with the two longest horizons actually outperforming the 6 months ahead forecast. Therefore, I believe that the implementation of shrinkage methods for the problem of predicting recessions has been moderately successful as it provides a model with enough warning time for the government or individuals to act to dampen the recession and with very few false positives.

On the other hand, for the regression problem, even the most accurate model (model 7) was not capable of producing a satisfactory forecast after the 3 months horizon, as showed by the persistent negative R_2. The mean absolute error shows that the model 7's GDP growth forecasts were on average around 0.4% off the true value, which certainly is a positive sign for the 0 and 3 months ahead forecasts. However, as seen in figure 7, this is also due to the lack of fluctuations following the real GDP growth rate for the longer-term prediction. Therefore, this should not be entirely seen as a positive result. Overall the implementation of shrinkage models for the forecast of quarter over quarter GDP growth rates is only partially successful for the two shortest predictions horizons and overall it is not as successful as the for the classification problem, since it does not provide information with enough warning for individuals or the government to act accordingly.

It is complex to talk exactly about the model specification created by the models due to the fact that with nested cross validation, there is a new model being fitted and tuned for each new inner loop. The final model coefficients produced using the specification of model 4 and model 7 are reported in table 3 in the appendix, but not much insight is possible due to the fact that scikit-learn functions are meant for machine learning and do not report p values or standard errors. Moreover, a possible avenue for further research would be to determine how the

magnitude and the significance of the coefficients has changed over time, since it has already been noted that the yield spread has been losing predictive power since the 1980s (Wright, 2006, p. 8).

Since the model was constructed using mostly data from the United States, another possible direction for further research would be the application of these shrinkage methods to other economies such as Japan, Germany or the United Kingdom following the examples of Nyman & Ormerod (2016) and Valckx, de Ceuster and Annaert (2002). This could potentially allow for the discovery of specific differences or similarities between recessions in different nations

Lastly, is always important to remember that despite the advances in econometrics even the most state-of-the-art models have the same warning: past performance does not guarantee future results. As Filardo (1999) shows, a mix of different methods all in agreement could be better than blindly trusting one single model or indicator just because it has been correct in the past.

Appendix

The code for this study is available from:
https://github.com/Davide-Bestagno/Master_Dissertation_Public

Indicators employed by Estrella and Hardouvelis's (1991):
10 year US bond – 3 months US treasury bill spread
Real federal funds rate
Growth index of leading economic indicators
one quarter lagged GNP
annualized rate of inflation (quarter on quarter)

Indicators employed by Estrella and Mishkin (1998)

Interest Rates and Spreads
- 10-year – 3-month Treasury spread
- AAA 6 months Commercial paper – 6-months treasury bill spread
- 3-month Treasury bill market yield
- 10-year Treasury bond

Stock Prices
- Dow Jones industrials
- NYSE composite
- S&P 500 composite

Monetary Aggregates
- Monetary base
- M1
- M2
- M3
- Monetary base deflated by CPI
- M1 deflated by CPI
- M2 deflated by CPI
- M3 deflated by CPI

Individual Macro Indicators
- Lagged Growth in real GDP (previous quarter)
- Consumer price index
- PMI composite index
- Vendor Performance, Percent Reporting Slower Deliveries for United States
- New Orders and Contracts for Plant and Equipment, Value for United States
- New Private Housing Units Authorized by Building Permits
- University of Michigan: Consumer Sentiment
- Trade-weighted exchange value of U.S. dollar versus G-10 countries
- Change in Manufacturers' Unfilled Orders, Durable Goods Industries for United States

Indexes of Leading Indicators
- The Conference Board leading economic index
- Stock and Watson (1989) leading index
- Stock and Watson (1993) leading index

Indicators employed by Wright (2006):
10 year US bond – 3 months US treasury bill spread
Nominal Federal Funds rate
Real Federal Funds rate
Excess Bond Return Forecasting Factor
Indicators employed by Filardo (1999):
10 year US bond – 3 months US treasury bill spread
AAA-BBB corporate bond spread
S&P 500 stock returns
The Consumer Board Leading Economic Indicator

Indicators employed by Nyman and Ormerod (2016):
3 months treasury bill rate
10 year US bond yield
S&P500 quarterly percentage changes
lagged GDP term

Dependent Variable to include in Shrinkage models:

Regression dataset: Real Gross Domestic Product quarterly change
Classification dataset: Binary dummy variable following NBER dates

Explanatory Variables to include in Shrinkage models:

Interest Rates and Spreads

1. 10year treasury bond yield
2. 3month treasury bill yield
3. 1year treasury bond yield
4. 6months treasury bill yield
5. Moody's AAA Corporate bond yield
6. Moody's BAA Corporate bond yield
7. Moody's AAA Corporate bond – 10year treasury bond yield spread
8. Moody's BAA Corporate bond – 10year treasury bond yield spread
9. 10year–3month Treasury spread
10. 10year–1year Treasury spread
11. Effective Federal Funds Rate

Stock Prices

12. Real S&P 500 composite index
13. Real S&P 500 quarterly percentage change

Commodity prices

14. Producer Price index by Commodity for Metals and Metal Products: Copper base scrap
15. Producer Price Index by Commodity for Metals and Metal Products: Iron and Steel
16. WTI Crude deflated by CPI

Monetary Aggregates

17. Monetary base

18. M1
19. M2
20. M3
21. Monetary base deflated by CPI
22. M1 deflated by CPI
23. M2 deflated by CPI
24. M3 deflated by CPI

Individual Macro Indicators

25. Consumer price index
26. Unemployment Rate
27. PMI composite index
28. Purchasing Managers' Composite Index
29. Manufacturing Supplier Deliveries Index
30. Manufacturing New Orders Index
31. New Private Housing Units Authorized by Building Permits
32. University of Michigan: Consumer Sentiment
33. Change in Manufacturers' Unfilled Orders, Durable Goods Industries for United States
34. Employment Level: Part-Time for Economic Reasons, Slack Work or Business Conditions, All Industries
35. Simultaneous recession dummy variable (DR-9) as defined by (Christiansen, 2012)
36. No. of advanced economies in recession (CR-9) as defined by (Christiansen, 2012)
37. Australia recession dummy variable
38. Canada recession dummy variable
39. Germany recession dummy variable
40. Japan recession dummy variable
41. UK recession dummy variable
42. Total Non-farm payroll employment
43. Index of industrial production
44. Real personal income excluding current transfer payments

Volatility Indicators

45. Mean absolute deviation of 10year–3month Treasury spread
46. Mean absolute deviation of 10year–1year Treasury spread
47. Mean absolute deviation of Real S&P500 composite

Lagged indicators

48. One quarter (3 months) lagged dependent variable
49. Two quarter (6 months) lagged dependent variable

Table 1: Time series used and description

Series code	Series name	Lag	Unit	Seasonally Adjusted?	Source
USREC	US Recession dates	Variable[20]	Binary	N	(FRED Federal Reserve Bank of St. Loius, 2020)

[20] Business Cycle dates are not published regularly. Only once a consensus is reached the dates are published which can happen also a year or more after a peak or through is reached.

GDPC1	Real Gross Domestic Product	1 quarter lag[21]	Billions of Chained 2012 $	Y	(U.S. Bureau of Economic Analysis, 2020a)
GS10	10-years Treasury Constant Maturity Rate	No lag[22]	Percent	N	(Board of Governors of the Federal Reserve System (US), 2020b)
TB3MS	3-months Treasury Bill	No lag	Percent	N	(Board of Governors of the Federal Reserve System (US), 2020c)
GS1	1-year Treasury Constant Maturity rate	No lag	Percent	N	(Board of Governors of the Federal Reserve System (US), 2020a)
AAA	AAA Corporate Bond Yield	No lag	Percent	N	(Moody's, 2020a)
DTB6	6-months Treasury Bills	No lag	Percent	N	(Board of Governors of the Federal Reserve System (US), 2020d)
BAA	BAA Corporate Yield	No lag	Percent	N	(Moody's, 2020b)
FEDFUNDS	Effective Federal Funds Rate	No lag	Percent	N	(Board of Governors of the Federal Reserve System (US), 2020e)
S&P500	S&P 500 monthly average	No lag	Index	N	(Yahoo Finance, 2020)
WPU102301	PPI of Copper Base Scrap	1-month lag	Index 100 = 1982	N	(U.S. Bureau of Labor Statistics, 2020d)
WPU101	PPI of Iron and Steel	1-Month lag	Index 100 = 1982		(U.S. Bureau of Labor Statistics, 2020e)
WTISPLC	WTI – Crude oil price	No lag	$ per barrel	N	(Federal Reserve Bank of St. Louis, 2020)
BOGMBASE	Monetary Base (M0)	1-Month lag	Millions $	N	(Board of Governors of the Federal Reserve System (US), 2020i)
M1NS	M1 Money Stock	1-Month lag	Billions $	N	(Board of Governors of the Federal Reserve System (US), 2020g)
M2NS	M2 Money Stock	1-Month lag	Billions $	N	(Board of Governors of the Federal

[21] During Q3, Q1 is available.

[22] During July, June is available.

					Reserve System (US), 2020h)
CPIAUCNS	Consumer price index	1-Month lag	Index June 2012 = 100	N	(U.S. Bureau of Labor Statistics, 2020b)
UNRATE	Unemployment Rate	1-Month lag	Percent	Y	(U.S. Bureau of Labor Statistics, 2020f)
MAN_PMI	Purchasing Managers' Composite Index	1-Month lag	Percent (no of answers that reported an improvement)	N	(Institute for Supply Management, 2020d)
MAN_DELIV	Manufacturing Supplier Deliveries Index	1-Month lag	Percent (no of answers that reported an improvement)	N	(Institute for Supply Management, 2020c)
MAN_NEWORDERS	Manufacturing New Orders Index	1-Month lag	Percent (no of answers that reported an improvement)	N	(Institute for Supply Management, 2020b)
PERMIT	New Private Housing Units	1-Month lag	Thousands of Units	Y	(U.S. Census Bureau & Development, 2020)
CS_m	Survey of Consumer Sentiment	1-Month lag	% of unfavourable consumer replies from the % of favourable one	N	(University of Michigan, 2020)
Unfilled Orders	Manufacturer's Unfilled Orders, Durable Good Industries	2-Month lag (In July we have May)	Millions $	N	(U.S. Census Bureau, 2020)
LNS12032195	Employment Level: Part-Time	1-Month lag	Thousands of Individuals	Y	(U.S. Bureau of Labor Statistics, 2020c)
PAYEMS	Total Nonfarm Payroll	1-Month lag	Thousands of Individuals	Y	(U.S. Bureau of Labor Statistics, 2020a)
INDPRO	Industrial Production Index	1-Month lag	Index 100 = 2012	Y	(Board of Governors of the Federal Reserve System (US), 2020f)
W875RX1	Real personal income excluding	2-month lag	Billions of 2012 chained $	Y	(U.S. Bureau of Economic Analysis, 2020b)

	current transfer receipts				
MABMM301USM189S	M3 Money Stock	2-month lag	Nominal $	Y	(Organization for Economic Co-operation and Development, 2020)
Various names	International Business cycle dates	Variable	Binary	N	(Economic Cycle Research Institute, 2020)

Table 2: Time series used in literature but not in this study

Time series	Reason for exclusion
Real Manufacturing and Trade Industries Sales	Time series only starts in January 1967
Public Debt as % of GDP	Time series only starts in January 1966
NYSE composite monthly average	Issues in finding historic data before December 1965
Dow Jones industrial average	Issues in finding historic data before 1985 in part probably due to changes in composition of the index
NASDAQ composite index	Established in February 1971 – no data available beforehand
BRENT Crude	Only started extraction in 1976 – no data available beforehand
Trade weighted dollar	Time series only started to be calculated in post-Bretton Woods era and it has changed significantly in the early 2000s with the introduction of the Euro
Stock–Watson (1989) leading index	Not available and the components of the index (Filardo, 1999, p.39) are included in the aforementioned variables
Stock–Watson (1993) leading index	Not available and the components of the index (Filardo, 1999, p.39) are included in the aforementioned variables
The Conference Board leading economic index	Not accessible with LSE resources
Survey of Professional Forecasters Leading Economic Indicator	Time series only starting from 1968

Table 3: Model coefficients

Time series	Coefficient Model 4	Time series	Coefficient Model 7
GS10	0.293614	**GS10**	0.000000
TB3MS	0.351066	**TB3MS**	0.000000
GS1	0.328259	**GS1**	0.000000

AAA10YM	1.077953	**AAA10YM**	-0.250057
DTB6	0.227050	**DTB6**	0.000000
BAA10YM	0.770681	**BAA10YM**	-0.000000
FEDFUNDS	0.697304	**FEDFUNDS**	0.000000
WPU102301	0.339769	**WPU102301**	-0.000000
WPU101	-0.054266	**WPU101**	-0.000000
BOGMBASE	-0.468109	**BOGMBASE**	-0.000000
M1NS	-0.302458	**M1NS**	-0.000000
M2NS	-0.027652	**M2NS**	-0.044340
UNRATE	-1.131717	**UNRATE**	0.311733
PERMIT	-1.597640	**PERMIT**	0.233541
LNS12032195	-0.020292	**LNS12032195**	-0.000000
PAYEMS	0.355038	**PAYEMS**	-0.178045
INDPRO	0.442868	**INDPRO**	-0.000000
MAN_PMI	-1.084541	**MAN_PMI**	0.000000
MAN_DELIV	0.640046	**MAN_DELIV**	-0.047706
MAN_NEWORDERS	-1.606986	**MAN_NEWORDERS**	1.916326
MABMM301USM189S	-0.033476	**MABMM301USM189S**	-0.062564
Australia	-0.921065	**CPIAUCNS**	-0.023995
Canada	0.433542	**CS_m**	0.906082
Germany	-0.097121	**10Y-3M**	0.000000
Japan	0.198822	**10Y-1Y**	0.000000
UK	-0.360568	**AAA**	0.000000
CPIAUCNS	0.136162	**BAA**	0.000000
CS_m	-1.740490	**CR**	0.000000
10Y-3M	-0.170677	**DR**	0.152492
10Y-1Y	-0.140225	**deflator**	-0.000000
AAA	0.555460	**R_WTISPLC**	0.000000
BAA	0.618913	**R_BOGMBASE**	-0.000000
CR	-0.342431	**R_M1NS**	-0.000000
DR	-0.573258	**R_M2NS**	-0.000000
deflator	0.148085	**R_Unfilled Orders**	-0.000000
R_WTISPLC	0.439436	**R_W875RX1**	0.000000
R_BOGMBASE	-0.488651	**R_MABMM301USM189S**	-0.000000
R_M1NS	-0.498310	**R_S&P500**	-0.000000
R_M2NS	-0.062042	**R_S&P500_change**	0.222327
R_Unfilled Orders	-0.343692	**USREC_lag6**	0.000000
R_W875RX1	0.889738	**mad_10Y-3M**	-0.237343
R_MABMM301USM189S	-0.053617	**mad_10Y-1Y**	-0.000000
R_S&P500	-0.074134	**mad_S&P500**	-0.000000
R_S&P500_change	-0.629203	**GDPC1_lag1**	-0.069525
USREC_lag3	2.543596	**GDPC1_lag2**	0.000000
USREC_lag6	-0.041910		
mad_10Y-3M	0.815515		
mad_10Y-1Y	0.479253		
mad_S&P500	0.612325		

References

Board of Governors of the Federal Reserve System (US). (2020a). *1-Year Treasury Constant Maturity Rate (GS1)*. https://fred.stlouisfed.org/series/GS1

Board of Governors of the Federal Reserve System (US). (2020b). *10-Year Treasury Constant Maturity Rate (GS10)*. https://fred.stlouisfed.org/series/GS10

Board of Governors of the Federal Reserve System (US). (2020c). *3-Month Treasury Bill: Secondary Market Rate (TB3MS)*. https://fred.stlouisfed.org/series/TB3MS

Board of Governors of the Federal Reserve System (US). (2020d). *6-Month Treasury Bill: Secondary Market Rate (DTB6)*. https://fred.stlouisfed.org/series/DTB6#0

Board of Governors of the Federal Reserve System (US). (2020e). *Effective Federal Funds Rate (FEDFUNDS)* . https://fred.stlouisfed.org/series/FEDFUNDS

Board of Governors of the Federal Reserve System (US). (2020f). *Industrial Production Index (INDPRO)*. https://fred.stlouisfed.org/series/INDPRO

Board of Governors of the Federal Reserve System (US). (2020g). *M1 Money Stock (M1NS)*. https://fred.stlouisfed.org/series/M1NS

Board of Governors of the Federal Reserve System (US). (2020h). *M2 Money Stock (M2NS)*. https://fred.stlouisfed.org/series/M2NS

Board of Governors of the Federal Reserve System (US). (2020i). *Monetary Base; Total (BOGMBASE)* . https://fred.stlouisfed.org/series/BOGMBASE

Christiansen, C. (2012). Predicting severe simultaneous recessions using yield spreads as leading indicators. *Journal of International Money and Finance*, *32*(1), 1032–1043. https://doi.org/10.1016/j.jimonfin.2012.08.005

Cochrane, C. (2018). *Time Series Nested Cross-Validation*. Towards Data Science. https://towardsdatascience.com/time-series-nested-cross-validation-76adba623eb9

Cochrane, J. H., & Piazzesi, M. (2005). Bond risk premia. *American Economic Review*, *95*(1), 138–160. https://doi.org/10.1257/0002828053828581

Diebold, F. X., & Rudebusch, G. D. (1989). Scoring the Leading Indicators. *The Journal of Business*, *62*(3). https://doi.org/10.1086/296467

Economic Cycle Research Institute. (2020). *International Business Cycle Dates*. https://www.businesscycle.com/ecri-business-cycles/international-business-cycle-dates-chronologies

Estrella, A. (1998). A new measure of fit for equations with dichotomous dependent variables. *Journal of Business and Economic Statistics*, *16*(2), 198–205. https://doi.org/10.1080/07350015.1998.10524753

Estrella, A., & Hardouvelis, G. A. (1991). The Term Structure as a Predictor of Real Economic Activity. *The Journal of Finance*, *46*(2), 555–576. https://doi.org/10.1111/j.1540-6261.1991.tb02674.x

Estrella, A., & Mishkin, F. S. (1998). Predicting U.S. Recessions: Financial variables as leading indicators. *Review of Economics and Statistics*, *80*(1), 45–56. https://doi.org/10.1162/003465398557320

Federal Reserve Bank of Philadelphia. (2020). *Historical Data Files for the Survey of Professional Forecasters*. https://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/data-files

Federal Reserve Bank of St. Louis. (2020). *Spot Crude Oil Price: West Texas Intermediate (WTI) (WTISPLC)*. https://fred.stlouisfed.org/series/WTISPLC

Filardo, A. J. (1999). How Reliable Are Recession Prediction Models? *Economic Review, Federal Reserve Bank of Kansas City*, *84*(QII), 35–55.

FRED Federal Reserve Bank of St. Loius. (2020). *NBER based Recession Indicators for the United States from the Period following the Peak through the Trough (USREC)*. https://fred.stlouisfed.org/series/USREC

Gogas, P., Papadimitriou, T., Matthaiou, M., & Chrysanthidou, E. (2015). Yield Curve and Recession Forecasting in a Machine Learning Framework. *Computational Economics*, *45*(4), 635–645. https://doi.org/10.1007/s10614-014-9432-0

Hamilton, J. D. (1985). Historical Causes of Postwar Oil Shocks and Recessions. *The Energy Journal*, *6*(1), 97–116.

Hastie, T., & Qian, J. (2014). *Glmnet Vignette*. https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html#intro

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Institute for Supply Management. (2020a). *ISM Report On Business®*. https://www.ismworld.org/supply-management-news-and-reports/reports/ism-report-on-business/

Institute for Supply Management. (2020b). *Manufacturing New Orders Index*. https://www.quandl.com/data/ISM/MAN_NEWORDERS-Manufacturing-New-Orders-Index

Institute for Supply Management. (2020c). *Manufacturing Supplier Deliveries Index*. https://www.quandl.com/data/ISM/MAN_DELIV-Manufacturing-Supplier-Deliveries-Index

Institute for Supply Management. (2020d). *PMI Composite Index*. https://www.quandl.com/data/ISM/MAN_PMI-PMI-Composite-Index

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to Statistical Learning: with Applications in R. In *Springer* (7th ed.). https://doi.org/10.1007/978-1-4614-7138-7

Mills, L. (1988). Can Stock Prices Reliably Pedict Recessions? *Business Review, Federal*

Reserve Bank of Philadelphia, *3*((September/October)), 3–14.

Moneta, F. (2005). Does the yield spread predict recessions in the Euro area? *International Finance*, *8*(2), 263–301. https://doi.org/10.1111/j.1468-2362.2005.00159.x

Moody's. (2020a). *Moody's Seasoned Aaa Corporate Bond Yield (AAA)*. https://fred.stlouisfed.org/series/AAA

Moody's. (2020b). *Moody's Seasoned Baa Corporate Bond Yield (BAA)*. https://fred.stlouisfed.org/series/BAA

NBER. (2020). *US Business Cycle Expansions and Contractions*. https://www.nber.org/cycles.html

Nyberg, H. (2010). Dynamic probit models and financial variables in recession forecasting. *Journal of Forecasting*, *29*(2115–2230). https://doi.org/10.1002/for.1161

Nyman, R., & Ormerod, P. (2016). *Predicting Economic Recessions Using Machine Learning Algorithms*. http://arxiv.org/abs/1701.01428

Organization for Economic Co-operation and Development. (2020). *M3 for the United States (MABMM301USM189S)* . https://fred.stlouisfed.org/series/MABMM301USM189S#0

Schrimpf, A., & Wang, Q. (2010). A reappraisal of the leading indicator properties of the yield curve under structural instability. *International Journal of Forecasting*, *26*(4), 836–857. https://doi.org/10.1016/j.ijforecast.2009.08.005

Sornette, D. (2004). Why Stock Markets Crash. *New Thesis*, *1*(1), 5–18. https://doi.org/10.1515/9781400885091

Tashman, L. J. (2000) 'Out-of-sample tests of forecasting accuracy: An analysis and review', *International Journal of Forecasting*, 16(4), pp. 437–450. doi: 10.1016/S0169-2070(00)00065-0.

U.S. Bureau of Economic Analysis. (2020a). *Real Gross Domestic Product (GDPC1)* . https://fred.stlouisfed.org/series/GDPC1

U.S. Bureau of Economic Analysis. (2020b). *Real personal income excluding current transfer receipts (W875RX1)* . https://fred.stlouisfed.org/series/W875RX1

U.S. Bureau of Labor Statistics. (2020a). *All Employees, Total Nonfarm (PAYEMS)* . https://fred.stlouisfed.org/series/PAYEMS

U.S. Bureau of Labor Statistics. (2020b). *Consumer Price Index for All Urban Consumers: All Items in U.S. City Average (CPIAUCNS)*. https://fred.stlouisfed.org/series/CPIAUCNS

U.S. Bureau of Labor Statistics. (2020c). *Employment Level - Part-Time for Economic Reasons, Slack Work or Business Conditions, All Industries (LNS12032195)*. https://fred.stlouisfed.org/series/LNS12032195

U.S. Bureau of Labor Statistics. (2020d). *Producer Price Index by Commodity for Metals and Metal Products: Copper Base Scrap (WPU102301)* . https://fred.stlouisfed.org/series/WPU102301

U.S. Bureau of Labor Statistics. (2020e). *Producer Price Index by Commodity for Metals and Metal Products: Iron and Steel (WPU101)* . https://fred.stlouisfed.org/series/WPU101

U.S. Bureau of Labor Statistics. (2020f). *Unemployment Rate (UNRATE)*. https://fred.stlouisfed.org/series/UNRATE

U.S. Census Bureau. (2020). *US Census Bureau Manufacturers' Shipments, Inventories, and Orders - Historical Data*. https://www.census.gov/manufacturing/m3/historical_data/index.html

U.S. Census Bureau, & Development, U. S. D. of H. and U. (2020). *New Private Housing Units Authorized by Building Permits (PERMIT)*. https://fred.stlouisfed.org/series/PERMIT

University of Michigan. (2020). *Surveys of Consumers*. http://www.sca.isr.umich.edu/

Valckx, N., de Ceuster, M. J. K., & Annaert, J. (2002). Is Financial Market Volatility Informative to Predict Recessions? In *DNB Staff Reports* (Issue 93). https://doi.org/10.2139/ssrn.429482

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*. https://doi.org/10.1186/1471-2105-7-91

Wright, J. H. (2006). The Yield Curve and Predicting Recessions. In *Finance and Economics Discussion Series*. https://doi.org/10.17016/feds.2006.07

Yahoo Finance. (2020). *S&P 500*. https://finance.yahoo.com/quote/%5EGSPC/history?p=%5EGSPC

Zarnowitz, V., & Braun, P. a. (1993). Twenty-two years of the NBER-ASA quarterly economic outlook surveys: Aspects and comparisons of forecasting performance. In J. H. Stock & M. W. Watson (Eds.), *Business Cycles, Indicators, and Forecasting* (1st ed., pp. 11–94). University of Chicago Press.