# Predicting Recessions Using Boosting and Bayesian Model Averaging

Marthe Elisabeth Aastveit


Thesis submitted for the degree of
Master in Economic Theory and Econometrics
30 credits

Department of Economics
Faculty of Social Sciences

UNIVERSITY OF OSLO

May 2019

# Predicting Recessions Using Boosting and Bayesian Model Averaging

Marthe Elisabeth Aastveit

Predicting Recessions Using Boosting and Bayesian Model Averaging

http://www.duo.uio.no/

Printed: Reprosentralen, University of Oslo

# Preface

I would like to thank my supervisor Leif Anders Thorsrud for his encouragement, interest and investment in this project. His help and support have been invaluable. I would also like to thank my oldest brother, Knut Are, for reading through my thesis and providing excellent comments. Finally I would like to thank the rest of my family and friends for moral support and encouragement. A special thanks to my boyfriend Joakim for support, discussions about machine learning and for allowing me to run loops on his computer whenever I wanted.

The analysis and implementation are done in R. Code examples are provided in Appendix B. All remaining errors are my own.

# Summary

Economic recessions are costly, and are among other things associated with high unemployment rates, low wage growth, low investment spending and a higher number of bankruptcies. Whether the economy is in a recession or an expansion is important for economic policy decisions. To make accurate predictions of the state of the economy is then of key importance for policymakers.

In this thesis I compare the performance of two data-driven methods for predicting US recessions. The methods I use are called boosting and Bayesian model averaging (BMA). Boosting is a machine learning technique that can be used for both classification and regression problems. Because it is flexible, boosting have been used for a wide set of applications in many fields and is considered one of the most powerful learning ideas in the last twenty years (Hastie et al., 2008, p. 337). Bayesian model averaging is a Bayesian method that accounts for model uncertainty. BMA is a framework for both model selection and model combination. As boosting, is this method also rather flexible and can be applied to address many different questions. In recent years, has BMA also gained popularity in economics, especially for macroeconomic forecasting.

My main objective in this thesis is to predict recessions in the US. Since recessions can be viewed as rare events, is it important to use data that covers a long time span. The dataset I use consist of data for the US from January 1959 to November 2018. During this time period, has there been eight recessions, where the Great Recession is the most recent one. There is a large amount of papers that propose different methods and predictors that can be useful for predicting recessions. In order to use as much information about the economy as possible, I use a dataset consisting of 128 different economic and financial variables. An advantage of using boosting and BMA compared to standard methods used in economics, is that they can handle a large amount of data, i.e. high dimensional data.

I evaluate in-sample and out-of-sample performance of different boosting and BMA specifications for predicting recessions six months ahead. To assess the forecast accuracy, I calculate the receiver operating characteristic (ROC) curve. The forecast performance is evaluated by the integrated area under the ROC-curve (AUROC). The in-sample results for boosting and BMA show that both methods predict recessions well, with

AUROC-values of above 0.9 The model with the highest AUROC-value is a boosting model which has an AUROC-value of 0.972. In general, I find that boosting has somewhat higher AUROC-values than BMA. This AUROC-value is also high compared to previous published papers which use more traditional econometric models.

The out-of-sample results are more mixed. As expected, are the AUROC-values a bit lower for the out-of-sample analysis compared to the in-sample analysis. In contrast to the in-sample results, provides BMA more accurate out-of-sample forecasts than boosting. The AUROC-values are in most of the cases between 0.85 and 0.90. The model with the highest AUROC-value is a BMA-model with an AUROC-value of 0.892. The out-of-sample AUROC-values that I obtain for BMA and boosting are in line with AUROC-values found in earlier published papers that use more traditional econometric models.

Finally, although both BMA and boosting allow for including a large set of predictors, I find that only a few predictors are important for predicting recessions. Most of these variables are well-known for being informative about future recessions. Particularly, I find that different interest rate spreads are the most important predictors for US recessions. In contrast to earlier studies, which typically have only included one specific interest rate spread as a predictor, I show that combinations of various interest rate spreads have high predictive power together for predicting a future recession. This is an interesting result, which indicates that various spreads are not mutually exclusive.

# Contents

# List of Figures

# List of Tables

# 1 Introduction

The National Bureau of Economic Research (NBER) maintains the chronology of the business cycle in the United States. The business cycle consists of recessions and expansions. A recession is the period between a "peak" and a "through", while an expansion is the period between a "through" and a "peak" (NBER, 2010). Recessions are costly and often associated with high unemployment and stagnant wage growth, in addition to decreasing economic opportunities and lower investment spending (Berge, 2015). The most recent recession in the US was the Great Recession. The strains in the market in August 2007 were the beginning of the longest recession in recent history. The catalyst was the collapse of Lehman in September 2008, which led to panic in the financial markets and a big decline in the economic activity within weeks. The credit access dropped, the growth rate became low, the real wages were stagnant and there were higher volatility in consumption, investment, output and inflation (Ng & Wright, 2013).

All of the recessions since 1985 have had origins in the financial market (Ng & Wright, 2013). The nature of recessions with origin in financial markets, are different from the ones where the financial markets play a passive role. Ng & Wright (2013) highlight five differences in the recessions after 1985:

1. Long, but weak expansions
2. Weakened procyclicality of labor productivity
3. Jobless recovery: The labor markets have been slowly improving during the last three recessions
4. Pronounced leverage cycle: The ratios of assets to liabilities of household and firms have a downward trend
5. Tight availability of credit, which leads to headwinds to the recovery

These points also highlight the key challenges for forecasting recessions; the important predictors have changed over time. The recessions from 1960 to 1985 had different origins than the recessions from 1985 and onwards. It is therefore a challenge for classical econometric models, such as VARs, which are limited to only include a small set of predictors to capture all of the different warning signs for recessions. One reason is that many classical econometric models have difficulties with incorporating

information from a large amount of data. As a result, important information may be excluded from the model. One possible solution to these problems is to rely on methods and algorithms that can incorporate information from many predictors at the same time. Moreover, by sequentially updating the forecasts from these methods over time, the algorithm can learn about which predictors have been important for capturing previous recessions.

The recent development in computer performance, machine learning, artificial intelligence and the use of Big Data, have suggested new ways of handling large amount of data. One of the main differences between these new methods and traditional mathematical or statistical methods is that they are much more data-driven. The use of these data-driven approaches in economics is still mostly unexplored, but they are starting to gain more popularity also within economics. Athey (2018) states that

> I believe that machine learning (ML) will have a dramatic impact on the field of economics within a short time frame. Indeed, the impact of ML on economics is already well underway, and so it is perhaps not too difficult to predict some of the effects. (p. 1)

This is also the motivation for writing this thesis, namely to explore some of these new data-driven methods and apply them to the question of how to predict US recessions. Predicting recessions by using these methods is a way of contributing to how data-driven methods in economics can be used. The collection of the predictors can then be combined in ways that has not yet been covered and give additional information and insights about the state of the economy.

The two data-driven methods I use are called boosting and Bayesian model averaging (BMA). To illustrate what boosting is, I present an example from Freund & Schapire (1997). The example starts with a horseracing gambler. The problem for this gambler is that he looses a lot, even though many of his friends win considerably more. He then decides to allow a group of his gambling friends to make bets on his behalf. He have a fixed sum in each race and divides them between his friends, first equally, then according to who wins the most. He does not know which of his friends who wins the most before he allocates his money. In order to get the most money in the end, he tries to allocate each race's wager in a way that the total number of wins will be approximately close to what he would have won, had he bet all with friend who is the luckiest. The boosting algorithm solves how he should allocate his money in order to earn approximately the same amount of money as if he had bet only on his luckiest friend.

A more formal way to express boosting is that it is a method that combine weak learners to a strong learner (Mayr et al., 2014). A weak learner is a classifier that can predict an event only a bit better than

random guessing. This is the friends of the gambler in the example. The weak learners are combined on modified versions of the data many times. Each time, the weak learners get updated. In the example, this is the combination of how much money the friends get to bet on a horse in each wager. How the weak learners change during the iterations, differs for different types of boosting. In the end, these weak learners are combined to a new learner that can predict the outcome almost perfectly (Hastie et al., 2008, p. 337-338). This is called a strong learner (Mayr et al., 2014). This is illustrated in the example above by how the man can allocate each race's wager in a way that the total number of wins will be approximately close to what he would have won by only betting on the luckiest of his friends. Summing up, boosting is an algorithm which learns from the iterative process of the weak learners and uses this information to combine it to an accurate classification (Mayr et al., 2014).

The boosting method I use in the analysis is called gradient boosting. This is a type of statistical boosting. The method is then developed from a statistical perspective, which have some advantages compared to pure machine learning methods. Mayr et al. (2014) points at these advantages; (i) their ability to combine variable selection that is automated and model choice in the fitting process, (ii) how flexible they are of the type of predictor effect that is possibly included in the final model and (iii) how stable they are in cases with high dimensional data where it might be more possible variables than observations.

To illustrate the concept of Bayesian model averaging (BMA), I present an example from Hoeting et al. (1999). The example starts with a researcher that gather data for cancer in the esophagus. The number of patients is big, but she has gathered information about demographic and medical risk factors and patient's survival status, for each of these patients. The researcher wants to specify the size of the predictors' effect in order to predict the survival time. She first uses a classical regression model to analyze the data and then conduct a data-driven search for this regression model. The final model which fits the data well is called M. Suppose that there exists an alternative model called M*, which almost fits the data equally well, but leads to different important predictors and different predictions. Which model should she choose? And should she ignore the results from the other model?

Bayesian model averaging is a method that takes model uncertainty into account and provides a way around the problem stated above (Hoeting et al., 1999). It does so by averaging over all of the possible models and weights them. The estimates of the model given the data is then a weighted average of the parameter estimates from the different models (Amini & Parmeter, 2011). Taking this insecurity into account when finding the most accurate model is an advantage in situations with a large number of predictors. This is because the existence of many combinations of the predictors makes it hard to find which model predicts the response

most accurate. The results from the other models should not be ignored either, because they may contain important information. Incorporating information from all of the different possible models is then the problem that BMA solves.

The choice of boosting and BMA as methods, is inspired by three papers that have done similar research – Ng (2014), Berge (2015) and Döpke et al. (2017). Ng (2014) and Berge (2015) have investigated boosting for predicting recessions in the US. Berge (2015) has also compared boosting with BMA, using a considerably smaller dataset consisting of leading indicators. Döpke et al. (2017) performed similar analysis using German data. My thesis build on these studies, but also differ in some important aspects as both my dataset and model specifications differ. In section 2, I will provide more details on how my analysis differ from the mentioned studies.

I evaluate both in-sample and out-of-sample performance for different boosting and BMA specifications for predicting US recessions 6 months ahead. My data sample is from January 1959 to November 2018 and the out-of-sample predictions are evaluated from November 1977 to November 2018. To measure how well the models are at predicting US recessions, I use the area under the receiver operating characteristic (AUROC) curve. ROC is a probability curve, where in this analysis, the x-axis is the probability to falsely predict a recession (false positive rate), while the y-axis shows the probability of predicting a recession when there is a recession (true positive rate). To summarize the implied forecast performance by each curve, I integrate the area under the ROC. The higher the AUROC-values are, the better the model is at predicting recessions (Fawcett, 2006).

My main finding is that both boosting and Bayesian model averaging predicts recessions fairly well. I find that the most important predictors are the interest rate spreads and building permits in different areas of the US. The important predictors in the in-sample results and out-of-sample results are mostly the same. The interest rate spreads are often considered important together, which means that they are not mutually exclusive and have high predictive power together and not separately. This separates my analysis from previous studies, because they tend to consider one spread variable at a time (Ng, 2014).

The in-sample results show clear spikes in the probability around the recessions for both boosting and BMA. The results from boosting indicates that the recessions before 1990 are predicted almost perfectly. The recessions after 1990 are also predicted well, but they have lower spikes around the recession dates. The in-sample results for BMA are also in most cases accurate. However, there are also some spikes between the recessions. While these spikes seem to be lower than the ones during recession periods, they still provide weak signals of ”false” recessions. In general, the AUROC-values from the in-sample analysis are high. The

model with the highest AUROC-value is a boosting model that obtains an AUROC-value of 0.972. This is a high number also compared to other studies that have used more traditional econometric models for predicting US recessions.

When it comes to the out-of-sample results, the results from the different models are more mixed. In most of the cases, there are spikes when there is a recession, but also here are there periods with weak signals of "false" recessions. For most model specifications, the AUROC-values exceed 0.85 and are in some cases close to 0.90. The AUROC-values from boosting and BMA in the out-of-sample analysis lie in the same area, around 0.8 to 0.9, compared to results in other studies that use more traditional econometric models. I compare my results to a simple benchmark probit model with the Treasury term spread as predictor. In both the in-sample analysis and the out-of-sample analysis, the various specifications for boosting and BMA have higher AUROC-values than this simple benchmark model.

The rest of the paper is structured as follows; section 2 describes the existing literature and how my thesis contribute to the literature on predicting recessions. Section 3 describes the methods I use in the analysis, with a focus on the theoretical framework. The data and the experimental design is described in section 4. This section focuses on the empirical framework since the theory has previously been described in section 3. It will especially focus on how the methods are implemented and used in the packages that I use in R. Section 5 presents results for the in-sample and out-of-sample analysis for both boosting and BMA. The methods will also be compared in this section. In section 6, there is a discussion about the methods and the results, with a focus on advantages, disadvantages and future usage of the methods. The thesis ends with a conclusion in section 7.

# 2 Literature and Contribution

There is a large amount of literature for predicting recessions. One of the most common ways of predicting recessions is to use the yield curve. Estrella & Mishkin (1996) investigates whether the spread between the interest rates on the ten-year Treasury and three-month Treasury bill can predict recessions. Their results show that the yield curve contains important informations for predicting recessions, especially one to two years ahead. Probit models using the yield curve to forecast recessions are examined in Wright (2006). He finds that models that use the level of the federal funds rate combined with the term spread give better in-sample and out-of-sample predictions than models that use the term spread alone.

Forecasting recessions using a probit model is for example done in Fossati (2015). He uses a large amount of data and estimates three factors, namely a bond and exchange rates factor, a stock market factor and a real activity factor. He has three main results. The first is that models that use only financial indicators performs worse after 2005. The second result is that models that use factors give better fit than the models where the indicators are used directly. Third, he finds evidence that the individual indicators affect the factors more than data revisions.

Liu & Moench (2016) predict recessions in the US both in-sample and out-of-sample at various horizons, from three months ahead to two years ahead. They do this using different well-known leading indicators, but they use the Treasury term spread as a benchmark. They consider both univariate and multivariate probit models and evaluate the performance of the predictions using AUROC. Their findings are that adding lagged observations of the term spread improves the predictions in the short run. Adding the annual return on S&P 500 index with the term spread improves the predictions even more for a time horizon shorter than one year. New orders of capital goods for the manufacturers and balances in Broker-Dealer margin accounts increases the prediction precision when forecasting more than one year ahead.

Chauvet (1998) empirically characterize business cycles with a dynamic factor model with regime switching. She captures how the macroeconomic variables comove by an unobservable dynamic factor. The asymmetries for the business cycles are captured by allowing the factor to switch regimes. Her results shows that the method makes it possible to analyze business cycles in real time. An example is if a recession is close by, it can be found by inferred probabilities or by the implied coincident indicator. This can be done at the same time as the macroeconomic variables are signaling a recession.

Chauvet & Piger (2008) compare two multivariate well-known business cycle dating approaches, both a nonparametric algorithm and a parametric Markov-switching dynamic factor model. Their results show that both of the approaches can identify turning points in real time quite accurately. The dynamic factor Markov-switching model identifies the turning points from NBER more accurate and the business cycle throughs with more lead than the nonparametric algorithm.

Chen et al. (2011) forecasts the probability of a US recession with a probit and dynamic factor modeling approach. They do this by using a large set of explanatory variables to model and forecast the probability of a recession. Their results show that the recessions since 1980 is captured by their model. The model also catches the Great Recession one year before the formal declaration from NBER. Their model outperforms many recession forecast models, both in-sample and out-of-sample. This paper is an example of predicting recessions using a data-rich environment.

The papers presented above are examples of more traditional econo-

metric methods. Typically these studies use either a single predictor or compress the information from a set of variables into a few common factors. Moreover, these models are typically either logit/probit models or models that allows for regime switching, such as Markov Switching models. One reason to investigate a new type of methods with roots in machine learning, is to incorporate the large amount of information we have in our data. These methods can then find new combinations of variables, which have not been investigated before. Many of the papers I presented above have given accurate predictions of recessions, but there is still room for improvement. Since recessions are severe events, can small improvements in forecast accuracy actually be quite important. In my thesis, I therefore aim to analyze what some newer machine learning techniques can add to the existing literature on predicting recessions. So far, there is only a limited number of studies that have explored these techniques for predicting recessions.

Berge (2015) uses 19 predictors for the US and compare how four type of methods, equally weighted forecasts, forecasts from BMA and forecasts from two different boosting specifications, predict US recessions. His analysis shows that equal weighted forecasts perform relatively badly. Both boosting and BMA are more successful in terms of predicting recessions. He finds that for shorter forecasting horizons, the most informative predictors are real economic activity variables, while variables for the housing market and the financial market are the most informative predictors at longer horizons. Moreover he finds that yield curve in general is a good predictor, but it did not provide a strong signal for the two last recessions (the ones starting in 2001 and 2007).

The boosting method in Ng (2014) is similar to what I use in this thesis. She uses boosting to screen up to 1500 potentially relevant predictors that consists of 132 real and financial time series and their lags. Even though she uses a large combination of variables, her results indicate that there are less than 10 important predictors. She also finds that there are different variables that are important before and after the mid 1980s. Her rolling window estimation indicates that how important the term and default spreads are depends on the recession. The analysis also reveal that the boosting model provided signals of an upcoming recession in the middle of 2006.

Ng (2014) models the log-odds ratio as a non-parametric function of the predictors, where the weak learner is a two-node decision tree. On the other hand, Berge (2015) takes an approach that is analogous to a logistic model where the log-odds is assumed to be linear in each predictor. He also goes further to include nonlinearity, where he uses smoothing splines[1] as weak learners.

[1]Smoothing splines will not be covered in this thesis, but Berge (2015) refers to Eilers & Marx (1996) for details.

Döpke et al. (2017) use boosted regression trees (BRT) to look at the usefulness of selected leading indicators for predicting recessions in Germany. Their results show that measures of the short-term interest rate and the term spread are important leading indicators. The relative importance of the short-term interest rate has, however, decreased over time, while for the term spread it has increased. The BRT approach also shows better out-of-sample results than the ones for standard probit models. They also argue that the BRT approach is a technique that can be useful for analysis of economic policy. The reason is that the relative importance of the short-term interest rate as a leading indicator has decreased and this may have implications for monetary policy.

In addition to Ng (2014), Berge (2015) and Döpke et al. (2017) recent research conducted by Raffinot & Benoit (2018) investigate other alterinative machine learning techniques for predicting recessions. Raffinot & Benoit (2018) use random forest and boosting to detect economic turning points in the US and the Eurozone.

Berge (2015) and Ng (2014) are the studies that are the most closely related to my thesis. I would therefore like to highlight the differences between their studies and my thesis.

First, to implement a boosting analysis I need to define a weak learner. Berge (2015) uses smoothing splines as weak learners, while Ng (2014) uses decision trees. I also use decision trees, but my trees are different from the ones in Ng (2014)[2]. Ng (2014) use a tree depth of 1, which is called a decision stump. In many applications, using a decision stump is considered insufficient. In stead a depth between 3 and 7 are often preferred for boosting applications (Hastie et al., 2008, p. 363). This is the reason why I allow for a larger tree depth than 1. My choice of tree depth is discussed in section 4.2.

Second, one of the goals with this thesis is to study the usefulness of boosting and BMA for prediction in a data-rich environment. While Berge (2015) uses a somewhat limited dataset consisting of 19 variables, I instead use a large dataset that consists of 128 variables[3]. Ng (2014) also study boosting in a data-rich environment using 132 variables and their lags. Moreover, it is not straightforward to use BMA on a dataset consisting of 128 variables, because it is not possible to evaluate all of the different models[4]. To solve this issue, I rely on using Markov Chain Monte Carlo (MCMC) methods.

Third, I study four different BMA model specifications and six different boosting model specifications in the in-sample analysis. In the out-of-sample analysis, I study two different boosting model specifications and two different BMA model specifications. All of these different model

[2]Decision trees are introduced in section 3.1.

[3]For boosting are also some lags included.

[4]The total number of possible models is $2^{128}$. Section 3.2 explains this further.

specifications are done on the same forecast horizon of six months. Berge (2015) and Ng (2014) have not studied different model specifications for any of the methods for the same forecast horizon.

Fourth, an essential part of this thesis is to compare the in-sample and out-of-sample performance between boosting and BMA and also to a simple benchmark model. Both Ng (2014) and Berge (2015) evaluate the predictive power from their models, but they have not compared their results from these models with results from alternative models[5]. I compare the AUROC-values that I obtain for the various models with Liu & Moench (2016). I also compare both my in-sample and out-of-sample results with the results from a simple probit model with the Treasury term spread as predictor.

Finally, I also use a longer data sample than Berge (2015) and Ng (2014). The sample that I use starts in January 1959, while both Berge (2015) and Ng (2014) start their sample later. This means that the analysis Berge (2015) covers six recessions, while the analysis from Ng (2014) covers seven recessions. Moreover, this may affect both the in-sample and out-of-sample predictions in addition to the most important predictors. I have also extended the sample with about five years forward. That the time period is extended forward is an advantage especially for boosting, because there are more data to train and test on.

# 3 Methods

Both boosting and Bayesian model averaging are data-driven methods with roots in statistics, which means the methods have a theoretical foundation. This section gives insight to the theoretical characteristics of the methods. These characteristics make it possible to interpret the most important predictors, which results in predictions in the end. The methods will be presented separately by first introducing them at a general level, before moving over to the details.

## 3.1 Boosting

Boosting is a machine learning technique, where you build simple base learners, called weak learners, and combine them to a strong learner in an iterative and stagewise process (Döpke et al., 2017). In a binary setting, a weak learner is defined to have a classification rate that is a bit better

[5]Berge (2015) compare his predictions from the 2001 and 2007 recessions with Chauvet & Piger (2008) for their nowcast and a univariate logit model with the slope of the yield curve at a forecast horizon of one year.

than random guessing. A strong learner on the other hand, should be able to predict the response accurately because the classification rate is high. Since it is easier to find weak learners, the idea of boosting is to combine these weak learners into a strong learner. This strong learner can be used for prediction. Weak learners vary in different types of boosting, but can typically be decision trees (both classification and regression trees), linear models and smoothing splines. The only restriction is that they are weak, in a way that they should not have too complex solutions in one iteration (Mayr et al., 2014).

A decision tree is the weak learner used later on in the analysis. A simple example of a decision tree is illustrated in figure 3.1. The tree shows the decisions being made according to preferences, in this case education. The decision that needs to be made is whether to ask a person to a job interview or not. Since education is the most important decision for whether the person should get an interview or not, it is at the top of the tree. To find out if the person is suitable for the job, some questions needs to be answered. A decision tree makes decisions about the outcome of a variable based on the data of the predictors by asking different questions which leads to a decision in the end. The questions are if one event occurs, then another event will follow based on the data. In figure 3.1 one example of a question will be ”Does the person have job experience?”. If yes, the person gets an interview, if no, the next question is asked. In the end, there will be a specific decision with an answer to the original question.

First of all, there are some terms for decision trees that should be defined. The first is the definition of different nodes. There exist three types of nodes. The first node is called a root node/decision node. This represents the first choice that will split in two or more internal nodes/chance nodes. The internal nodes then represent the possible choices at that point in the structure. The final node in the tree is called the leaf nodes/end nodes and represent the final result which consists of the combination of the decisions made previously (Song & Ying, 2015). An example of a root node in figure 3.1 is the question about education, while an internal node whether the person has experience or not. The end node is whether the person gets an interview or not.

Moreover, to make decision trees interpretable in boosting, they need to be based on data. The idea behind a decision tree with data is to build classification or regression models in a tree-structured form. To break down the data to smaller subset, the space is split into two regions and the response is modeled in each region. The splitting can happen again with the response. This continues until a stopping rule is applied to prevent the model from becoming too complex. An example of a stopping rule is the depth of the tree. The region that represents the decision in the end is, as stated above, the end node (Song & Ying, 2015). In a boosting setting, the individual decision trees are the weak learners. Combining these decision trees in an additive way results in a strong learner. This ensemble is used

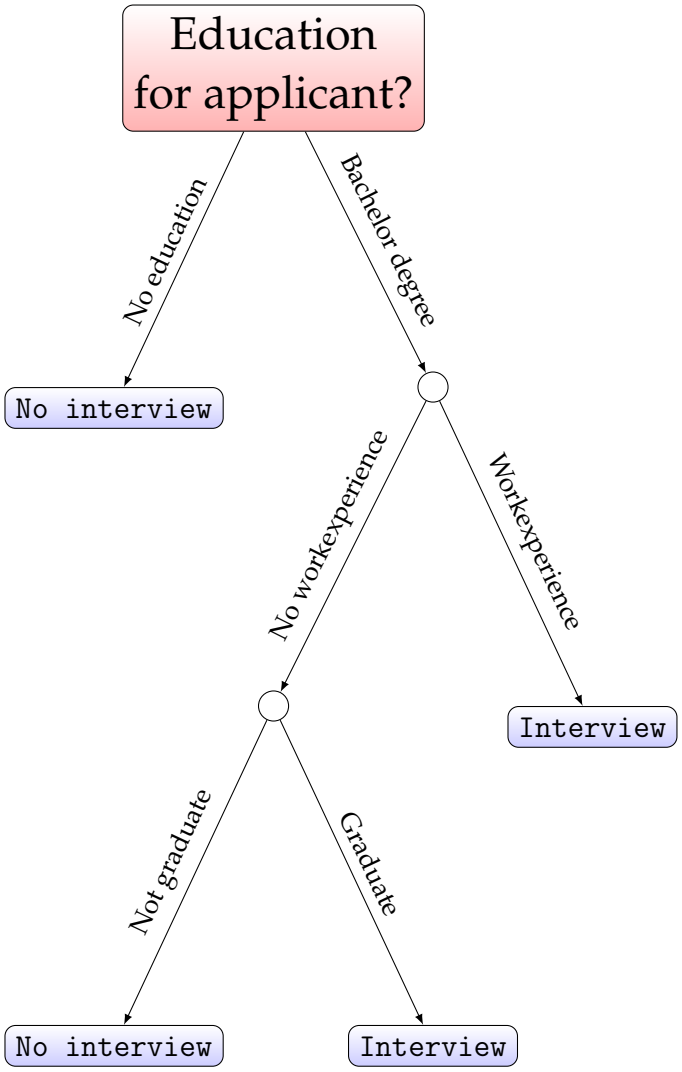



Figure 3.1: Decision tree

to forecast recessions (Döpke et al., 2017).

There are at least two points of view when it comes to the theory of boosting. The first one is to look at boosting from a machine learning perspective. The focus of this perspective is boosting as a machine learning algorithm. The other perspective is called statistical boosting (Mayr et al., 2014). This perspective focuses on presenting boosting as an algorithm with roots in statistics. It is mainly this statistical perspective that is presented in this section.

The theory of combining weak learners into a strong learner with good prediction accuracy was developed by Schapire (1990) and Freund (1995). Further Freund & Schapire (1997) developed the historically most popular boosting algorithm, called "AdaBoost.M1". The concept for the algorithm is, as stated above, to learn from the iterative process with a weak learner and combine it to classification. This procedure would not work if the

observations were trained over and over again on the same dataset. The solution to this is to do modifications on the data. The data is modified by re-weighting the observations in the training data during the iteration process (Mayr et al., 2014). This means for each iteration, the weights on the observations are modified and the algorithm is applied again on the weighted observations. The observations that were misclassified by the learner at the previous step gets higher weight and the weights will decrease for those that were correctly classified (Hastie et al., 2008, p.338). This forces the algorithm to focus on the objects that are hard to classify. In the final step, the results for the weak learners are combined to a more accurate prediction. This is done by increasing the iteration-specific coefficient of the solutions that performs better. This coefficient depends on the misclassification rate. The weak learners in "AdaBoost.M1" are often simple classification trees (Mayr et al., 2014).

The next step in the history of boosting was expanding the method with a higher focus on statistical and mathematical interpretations. In this category, the algorithms are used to estimate quantities in general statistical models (Mayr et al., 2014). J. Friedman et al. (2000) looked at boosting from a statistical perspective. One reason to look at boosting from a statistical perspective is that general machine learning algorithms often have a "black box" interpretation. In these algorithms, only the result matter, the underlying data structure is not relevant and how the predictors contribute to the final solution is not known (Mayr et al., 2014). J. Friedman et al. (2000) then provided statistical tools to be able to understand and interpret the boosting algorithm. It will be this group of boosting algorithms that will be the focus from now on.

In J. H. Friedman (2001) gradient boosting was developed. The main idea in this boosting category is to fit the weak learner, not to the re-weighted observations as in "AdaBoost.M1", but to the negative gradient vector of the loss function evaluated at the previous iteration. The gradient is a mathematical vector which gives information of how fast and in which direction a function changes (Holden, 2018). Both "AdaBoost.M1" and gradient boosting increase the performance of the weak learner by focusing on the observations, which are hard to predict. While "AdaBoost.M1" do this by up-weighting the observations that were classified wrongly, gradient boosting find the difficult observations by using the negative gradient evaluated in the previous iteration (Mayr et al., 2014). This means that the empirical risk is minimized. For each iteration, the models get strengthened because the fitted regression function is updated. In the end, the optimal fitted model is found and can be used for predictions.

Summing up, the data is modeled by using weak learners and minimizing the errors. These errors also find the datapoints, which are hard to fit. The models get updated in a way that focuses particularly on the datapoints that were hard to fit. In the end, these predictors will be

combined to the final model (Mayr et al., 2014).

Since gradient boosting is the method that I use in this thesis, I will describe this method and the algorithm in more detail. However, I first need to define some parameters. These parameters must be tuned according to getting the best results but at the same time avoid overfitting. The first is the depth of each tree, J. This means how many nodes the tree consists of, where a node is the end of a branch of a tree, as previously presented (Greenwell et al., 2019). The branch are the line segments in figure 3.1. In figure 3.1, the tree depth is 3 and the end node is whether the person is invited to an interview or not. The next tuning parameter is the number of boosting iterations, M. This means how many iterations the weak learner, the tree, goes through. It is important to have the right number of iterations to avoid overfitting (Hastie et al., 2008, p. 364). $\nu$ is defined as shrinkage, also called the learning rate, which controls the learning rate of the boosting procedure. The learning rate measures how much each tree contribute when it is added in the approximation of the strong learner. It is a way of slowing down the learning by scaling the contribution of each tree when it is added to the approximation. Smaller values of $\nu$ indicates more shrinkage and higher training risk for the same number of iterations, which means that that the adaption of the model to the data is slowed down (Hastie et al., 2008, p. 364-365). On the other hand, higher values of $\nu$ then means lower accuracy since there are higher steps so the optimization becomes less precise compared to if $\nu$ had been small. Another tuning parameter is the subsampling rate. At each iteration, a fraction of the training observations is sampled and the next tree grows using the subsample. It reduces the computing time and often produces a more accurate model (Hastie et al., 2008, p. 365).

Ridgeway (2019) has given a schematic overlook of the gradient boosting algorithm that is implemented in the package that is used in the analysis. $\hat{f}(x)$ is a regression function, while $L(y_i, f(x))$ is a loss function. The algorithm is presented in Algorithm 1.

**Algorithm 1** Gradient Tree Boosting in the gbm-package adapted from Ridgeway (2019)

Select:

- A loss function (distribution)
- The number of iterations, M (n.trees)
- The depth of each tree, J (interaction.depth)
- The shrinkage (or learning rate) parameter $\nu$
- The subsampling rate, $p$ (bag.fraction)

Initialize $\hat{f}(x)$ to be a constant, $\hat{f}(x) = argmin_\gamma \sum_{i=1}^{N} L(y_i, \gamma)$. This is the optimal constant model and contains only one single end node tree. For $m = 1, \ldots, M$:

1. Compute the negative gradient as the working response:

$$r_i = -\Big[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\Big]_{f(x_i)=\hat{f}(x_i)}$$

2. Select $p \times N$ random cases from the dataset. N is the number of observations in the training sample.

3. Fit a regression tree with J end nodes. This tree is fitted only using randomly selected observations

4. Compute the optimal end node predictions, $\gamma_1, \ldots, \gamma_J$ as:

$$\gamma_j = argmin_\gamma \sum_{x_i \in S_j} L(y_i, \hat{f}(x_i) + \gamma)$$

   where $S_j$ is the set of $x$s that define end node j. This step uses only the randomly selected observations.

5. Update $\hat{f}(x)$ as

$$\hat{f}(x) \leftarrow \hat{f}(x) + \nu\gamma_{j(x)}$$

   where $j(x)$ indicates the index of the end node into which an observation with features $x$ would fall

In step 1. the negative gradient is calculated. The negative gradient is evaluated for $f(x_i) = \hat{f}(x_i)$, which means that it is evaluated for the previous regression function. Step 2. selects a random number of observations from the dataset. How many observations are selected, depends on the randomization parameter $p$. The regression trees are fitted to all the end nodes in step 3. This step also depends on the selected observations. Then in step 4. the optimal end node predictions are calculated. It depends on which predictors are defined in that specific end node. In the last step, the regression function is updated. This depends

on the previous regression function, the learning rate and the index of the terminal node for the predictors. This means that $\gamma_j$ is included for the $x$s that are included in the terminal nodes. These steps are repeated for all of the iterations.

The two loss functions that are implemented in this analysis is the Bernoulli deviance and the AdaBoost exponential loss function. However, it is important to note that this AdaBoost exponential loss function is not the same as "AdaBoost.M1". It can be shown[6] that "AdaBoost.M1" is equivalent to a boosting approach, called forward stagewise additive modeling, with a loss function

$$L(y, f(x)) = exp(-yf(x)) \tag{3.1}$$

(Hastie et al., 2008, p. 343). This exponential loss functions is called AdaBoost from now on.

The output of the algorithm determines which of the variables that are most important and will be used for prediction. In many applications, only a few variables matter for the prediction and the rest is irrelevant. The output for gradient boosting is the relative importance (Hastie et al., 2008, p. 367-368). The relative influence for one variable $x_j$ is

$$\hat{I}_j^2 = \sum_{\text{splits on} x_j} I_m^2 \tag{3.2}$$

$I_m^2$ shows the empirical improvement of splitting $x_j$ at that specific point and at stage $m$ (Ridgeway, 2019). This means that for one variable, the squared relative importance is the sum of the squared improvements for all of the internal nodes that were chosen as the splitting variable (Hastie et al., 2008, p. 368). The way of getting the relevant importance for that variable for all of the iterations, is to average the relative influence of that variable for all of the trees that has been generated by the boosting algorithm. The equation becomes

$$\hat{\hat{I}}_j^2 = \frac{1}{M} \sum_{m=1}^{M} \hat{I}_j^2 \tag{3.3}$$

This means that the relative importance is then the average of $\hat{I}_j^2$.

There is no straightforward interpretation of relative importance in boosting. It is based on how many times a variable is selected over the M steps and weighted by the squared improvement. However, the sum of the relative importance for all of the different variables is 100, more specific, the sum of $\hat{\hat{I}}_j^2$ for all of the different predictors is 100. This means that a higher value of relative importance indicates a more important variable. If the variable is almost never selected, it has a relative influence of zero (Ng, 2014)

[6]See Hastie et al. (2008, p. 343-344) for details

## 3.2 Bayesian Model Averaging

In order to explain the theoretical part of Bayesian model averaging, some terms in Bayesian statistics need to be introduced. In general, Bayesian statistics has a different approach to statistical problems compared to traditional frequentist statistics. It is based on Bayesian interpretations, where you look at problems with a focus on the probability of an event. The probabilities can change as you gather new information. The starting point is that there might be an idea of the distribution of the parameter. This is called the prior distribution, $p(\beta)$. Then you have a distribution of the data given the parameter. This is called the sampling/data distribution, $p(y \mid \beta)$ and is the likelihood function. Using Bayes rule:

$$p(\beta \mid y) = \frac{p(\beta)p(y \mid \beta)}{p(y)} \tag{3.4}$$

where $p(y)$ is the sum (discrete distribution) or integral (continuous distribution) over all the possible values of $\beta$. $p(\beta \mid y)$ is called the posterior distribution and is the desired outcome (Gelman et al., 2013, p. 6-7).

Moreover, there are also some other arguments that needs to be clarified in this analysis. The first is that the model used for the binary data is the logistic regression. The formula for this equation is

$$log\Big(\frac{pr(y_t = 1)}{pr(y_t = 0)}\Big) = \beta_0 + \beta_1 X_{1,t-h-s} + \beta_2 X_{2,t-h-s} + \cdots + \beta_K X_{K,t-h-s} \tag{3.5}$$

where $y_t = 1$ if there is a recession in period $t$ and $y_t = 0$ if there is not a recession in period $t$. $X_{1,t-h-s}$ is variable 1 in period $t - h - s$, where $h$ is the forecast horizon, and $s$ is either 0, 1, 2 or 3 in this analysis. s is then the lag in addition to the forecast horizon. The $\beta$'s are the coefficients for the variables. This equation is called the log odds-ratio. Equation 3.5 is the main regression equation that the next part of the analysis is based on. This is one of the equations that needs to be specified in order for the analysis to be on binary response data (Agresti, 2015, p. 168).

BMA is method which accounts for model uncertainty. This is important because when there is a large set of explanatory variables, which might have an influence on the outcome, it is hard to know which variables are important and which are irrelevant. Traditionally, this is solved by doing a sequence of tests to find the best model where the irrelevant predictors are omitted (Koop, 2003, p. 267). As the number of tests increases, the probability of a mistake being made increases. One example of a mistake is if the researcher rejects the model considered ”better” for the one that is ”not so good”. The second problem is that even though the best procedure is being chosen, there is still a problem with ignoring the results and evidence from the models that are not the

best one. In this way, the model uncertainty is being ignored. This means that the researcher does not know what the parameters of the model are and which model is correct (Koop, 2003, p. 267).

While general model averaging only takes the average of the models being considered, BMA takes the average of the posterior distribution for each of the models being considered and weight them by the posterior model probability (Hoeting et al., 1999). Taking the average of the models is a way of finding the variables that are most relevant in the data generating process. Each of the models then get a weight and the estimates are then a weighted average of the parameter estimates from the models (Amini & Parmeter, 2011). All of the variables are included in the analysis, but the impacts of certain variables are almost 0.

The next step is to have a closer look at the properties of BMA. Assume we have a set of M models that is estimated to produce a forecast $y_t$ which will result in $\{\hat{y}_{1t}, \hat{y}_{2t}, \ldots, \hat{y}_{Mt}\}$. Then assume that there are K predictors. The total number of models are then $M = 2^K$. The reason for this number is that the models are defined by inclusion or exclusion of each of the explanatory variables (Koop, 2003, p. 268). The equation for BMA is defined as

$$y_t = Pr(\boldsymbol{\Delta} \mid D) = \sum_{i=1}^{M} Pr(\boldsymbol{\Delta} \mid M_i, D) Pr(M_i \mid D) \tag{3.6}$$

(Hoeting et al., 1999). $\boldsymbol{\Delta}$ is the quantity of interest, which might be an effect size or a future observable. In this case, the quantiy is whether there is a recession in period $t$ or not. The empirical question is then "What is the posterior probability that we are in a recession in period t?". Equation (3.6) shows the average of the posterior distributions under the models considered, but weighted by the posterior probability of the model considered.

The posterior probability for model $M_i$ is

$$Pr(M_i \mid D) = \frac{Pr(D \mid M_i) Pr(M_i)}{\sum_{j=1}^{M} Pr(D \mid M_j) Pr(M_j)} \tag{3.7}$$

where

$$Pr(D \mid M_i) = \int Pr(D \mid \beta_i, M_i) Pr(\beta_i \mid M_i) d\beta_i \tag{3.8}$$

is the integrated likelihood of model $M_i$ and $\beta_i$ is the vector of the parameters in model $M_i$ (Hoeting et al., 1999). $Pr(\beta_i \mid M_i)$ is then the prior density of $\beta_i$. The integral must also be solved and this calculation can be demanding, because it is not necessarily possible to solve the integral directly. The solution is to approximate the integral using a computational method (Hoeting et al., 1999). This is done directly in the package I use for the BMA-analysis in this thesis.

Hoeting et al. (1999) points at another problem with solving the equations in BMA. The number of terms in equation (3.6) might be enormous. This makes it hard to find the final solution. In this case, the number of potential models is $M = 2^K = 2^{128}$. This number is so high that it is impossible to try all of the different combinations of the variables. One possible solution to solve the sum is to use an algorithm to carry out BMA without evaluating every possible model. An algorithm that does this is the Markov Chain Monte Carlo (MCMC). This group of algorithms takes draws from the parameter space in order to mimic draws from the posterior. This is done by taking many draws from regions of the parameter space where the posterior probability is high, while the draws are few from the regions where the posterior probability is low (Koop, 2003, p. 272). This means that MCMC focus more on regions where the posterior probability is high and less on the regions where the posterior probability is low. This is the standard definition of MCMC with a parameter focus. Since the focus in BMA is models and not parameters will the MCMC algorithms behave a bit differently in this setting. The MCMC algorithm in a BMA setting then draws from the model space, not the parameter space, and focus on the models with high posterior model probability (Koop, 2003, p. 272-273). This is called Markov Chain Monte Carlo Model Composition ($MC^3$). $MC^3$ is based on a MCMC algorithm called Metropolis-Hastings algorithm which stimulates a chain of models. It draws different candidate models from a distribution over the model space and with a certain probability accepts them. If the candidate models is not accepted, the chain does not go forward, but instead remains at the current model (Koop, 2003, p. 273).

The specific $MC^3$ model considered is the Random Walk Chain Metropolis-Hastings. In the region of the model space, it draws in the neighborhood of the current draw. An alternative model then exists, namely the neighborhood model. This candidate model is then proposed and drawn randomly with equal probability from the set of models. It includes the current model, the models with one explanatory variable deleted and all the models with one explanatory variable added. The result is an acceptance probability, which indicates which model is being accepted (Koop, 2003, p. 273). The method used in the analysis is this $MC^3$ combined with a random swap where it swaps a variable included with a variable that is currently excluded (Clyde, 2018). Updating one at a time, might be a poor mixing with variables that are correlated, so one consider an additional update proposal. The additional update selects a variable included in the current model randomly and swaps it with a variable that is randomly excluded from the model (Clyde et al., 2011). This means that we have a new state using the $MC^3$ algorithm with a probability and uses the swap proposal with one minus the probability of using the $MC^3$ algorithm (Clyde et al., 2011).

In order to find and interpret the most important variables in the BMA-

analysis, a value called posterior/marginal inclusion probability is used. It is a weighted average of the posterior probabilities for all the models that include predictor j.

$$PIP(\beta_j) = Pr(\beta_j \neq 0) = \sum_{M_i : \beta_j \in M_i} p(M_i \mid D) \tag{3.9}$$

(Berge, 2015). This shows the probability of that specific predictor to be included in the final model and that this is based on the posterior probability for the models.

# 4 Data and Experimental Design

One advantage of using databased methods is that they can handle high dimensional data. The dataset I use in the analysis is from Federal Reserve Economic Data (FRED), more specifically from McCracken (2019), and contains a large selection of common macroeconomic and financial indicators. It is an easily accessible dataset because it is open-source. The analysis on these data has been implemented in R by using one package for boosting and one for Bayesian model averaging. The package chosen for boosting is called ”gbm” and the package chosen for Bayesian model averaging is called ”BAS”. This section gives an overview of the dataset and dig deeper into the empirical framework of the methods.

## 4.1 Data

Since the FRED dataset is updated every month, the dataset I use is from December 2018. The dataset consists of raw data but I follow McCracken & Ng (2016) and transform each variable to induce stationarity. I provide details about all the variables in the dataset and the transformations I use in Appendix A.

The dataset covers 128 US variables in the period January 1959 until November 2018. The variables cover a broad range of the US economy and are divided into categories, where the categories are (i) output and income, (ii) labor market, (iii) housing, (iv) consumption, orders and inventories, (v) money and credit, (vi) interest and exchange rates, (vii) prices and (viii) stock market (McCracken & Ng, 2016). McCracken & Ng (2016) point at some advantages of this dataset. First of all, it is updated every month. It is then possible to update the analysis easily and follow the development of different variables in the economy. Secondly, it is publicly accessible, which means that it is easy to replicate and confirm empirical work. Third, it will relieve researchers from managing changes in the dataset

and revisions. One challenge with collecting a dataset that spans a long time period is that definitions of variables have often changed over time. It can, for example, be hard to find exactly the same series back in time. Using a prepared dataset is then a big advantage. Another advantage of using this dataset in this analysis is the time frame of the dataset. Since the dataset starts in 1959, there is enough data to predict a rare event, which recessions are.

Even though there are many advantages of using the FRED dataset in this analysis, there is also at least one problem. Since the dataset is updated every month, observations for some variables may be revised over time. This is typically the case for National Account variables, which in some cases can undergo substantial revisions. For these series, it means that the value of for instance January 2018 is different if the data was collected in February 2018 compared to being collected in December 2018. This is, however, only a concern for variables that undergo revisions. Many variables such as financial markets data and price data are either not revised or only undergo minor revisions. This may affect the analysis since the actual value is available at a different point in time than presented in this analysis. One solution is to make a new dataset by going through all of the datasets back in time and type the values for that specific month. Fortunately does not most of the variables change back in time, but it is still a weakness with this analysis.

The National Bureau of Economic Research (NBER) has a formal declaration of recessions and these recession dates are found in NBER (2012). Recessions are binary variables, which means if there is a recession in period t it is denoted by 1 and if there is not in a recession it is denoted 0. In the period from January 1959 to November 2018, there are eight recessions. The Great Recession is the longest recession in this time frame and the shortest recession was the first recession in the 1980s.

In the analysis, recession or not is considered as the binary response variable. It is only the predictors from the FRED dataset that is included as predictors and not the lags of whether there has been a recession or not. The reason is that it often takes time before NBER announces a recession. Since this information will be available too late, it is not realistic to include it in the model.

## 4.2 Experimental Design

This section explains the empirical framework and how the different specifications in the algorithm will affect the outcome. An example of an important specification is which loss-functions are used for boosting. Different loss functions will give different important variables and different predictions. The rest of this section is structured by first looking at the empirical framework that the methods have in common, in both the

in-sample and out-of-sample analysis. Then the specific implementation for both boosting and BMA will be discussed. In the end, I introduce what the AUROC-values are.

The first challenge is to make the forecasting analysis as realistic as possible is the data availability. If the goal is to forecast a recession 6 months ahead, then there is only data available now that can be used. In more formal terms, the data must be available in time period $t - h$ when the goal is to forecast $h$ months ahead. Further, it may be the case that it is some data that were available in period $t - h - 1$ that are more important. This suggests that in addition to the forecasting period, the data should also be lagged. I therefore introduce $s$, which stands for how many lags are included in addition to the forecast horizon and $s = \{0, 1, 2, 3\}$. The data with lags will be analyzed together in the boosting case and separately in the BMA case. The forecast horizon in both the in-sample and out-of-sample analysis is 6 months. In the in-sample cases, the predictors are lagged according to $h + s$. So the dataset consist of a $Y_t$ which indicates recession or not in time period t and predictors 6 months back in time. In the out-of-sample cases, the predictors are lagged according to $s$ in the dataset and forecasted 6 months forward.

The estimation method to get out-of-sample results is called rolling estimation. The window starts with observations from $t_1 - h$ = January 1960 and $t_2 - h$ = May 1977. The rolling forecasts are constructed in this way:

1. Initialize $t_1$ and $t_2$
2. For $m = 1, \ldots, M$, follow algorithm 1 using the predictors in $X_{t-h}$. t $\in [t_1 - h, t_2 - h - 1]$
3. For each of the predictors, j = 1 ... N, record relative importance for boosting or posterior inclusion probability for BMA in the interval
4. Construct the predicted probability $\hat{p}_{t_2} = \hat{P}(y_{t_2+h} = 1 \mid X_{t_2})$. Increase $t_1$ and $t_2$ by one.

(Ng, 2014). There are 493 rolling regressions. Like I stated above, $t_1 - h$ = January 1960 and $t_2 - h$ = May 1977. The first forecast is then made for $t_2$ = November 1977. The next round of forecasts is based on training and estimation from $(t_1 - h, t_2 - h)$ = (February 1960, June 1977) and forecasts for December 1977 and so it continues. The windows rolls forward and the predictions for next period are in the end gathered and presented in figure 5.3.

The most important variables in the out-of-sample case are found by taking the average of the variables in all of the rolling estimations. For boosting it is the average of the relative importance, while for BMA it is the average of the posterior inclusion probability. The top ten predictors with highest average are reported in section 5.2.

### 4.2.1 Boosting

The first part of Algorithm 1 is to select a loss function/distribution, the number of iterations, the depth of each tree, the shrinkage parameter and the subsampling rate. The first choice is the loss function/distribution. In cases with binary response, it is most common to use the the deviance (Hastie et al., 2008, p. 360). The Bernoulli deviance is recommended (Ridgeway, 2019). In addition to the Bernoulli deviance, I also look at the AdaBoost exponential loss function[7]. Even though "Bernoulli" is in general recommended, can "AdaBoost" still be more appropriate in some settings. That is the reason why both of these loss functions are used. The method is still gradient boosting as presented in section 3.1.

The tuning parameters are important for the final results in boosting. The goal is to find the combination of tuning parameters, which give the most accurate predictions but at the same time avoid overfitting. The problem with overfitting is that it may represent misleading predictions and weights for the coefficients, resulting in misleading conclusions in the analysis. Possibly the most important tuning parameter is the number of iterations. There are different ways of finding the optimal choice of this parameter. One method is to use cross-validation[8]. This is a method to test the model out-of-sample to find how the data will perform on an independent dataset. The disadvantage of this method is that it only finds the optimal number of iterations given the other tuning parameters. That is the reason why the method used in this analysis is based on another method that is found in University of Cinncinati (2018). The setting of this method is to use the train fraction. A train fraction of 70 % means that 70 % of the first rows of the observations are used to fit the model and the rest are used to compute out-of-sample estimates for the loss function (Greenwell et al., 2019). This is presented in figure 4.1. When the train fraction is 70%, the training part in figure 4.1 is 70%, while the validation part is 30%. The optimal combination of the tuning parameters is found by making a grid search of all the possible combinations and minimizing the validation error. A grid search is an iteration process for tuning the parameters. In this case, it searches through all of the different combinations of the parameters and find the optimal combination in this case decided by the minimized validation error. Taking the square root of the minimized validation error gives the lowest root mean square error (RMSE). If the root mean square error is low, the predicted values are almost equal the actual values. This is because the RMSE measures the error between the predicted and observed values (Chai & Draxler, 2014). In this case, it is the error between the predicted and observed recessions. The combination of the chosen tuning parameters are decided by the square root of the minimized validation error, which is the same as the

[7]Details about both of these functions are discussed in Ridgeway (2019).
[8]This method is used in Ng (2014).

RMSE.

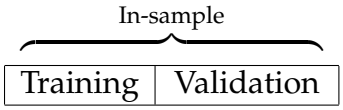



Figure 4.1: Obtaining in-sample results

This training fraction is used both in the in-sample analysis for finding the tuning parameters but also in the actual fitting of the boosted model. This is because the in-sample fitting is done on the same dataset as the predictions. By using a train fraction of less than 100 % prevents overfitting because it also tests how the model will perform on new data.

The tuning parameters are especially important for avoiding overfitting. The included tuning parameters are the number of trees to fit, the maximum depth of each tree, the minimum number of observations in the end nodes of the trees (n.minosbsinnode) and the learning rate (Greenwell et al., 2019). The rest of the parameters in the package are set as the default in Greenwell et al. (2019). I have also implemented three different training fractions, 30%, 50% and 60%. This is to look at whether there are big differences in the results for the training fractions. This parameter is also important for overfitting, because training everything on the same data and predicting on the same data without looking at validation, may lead to overfitting (University of Cinncinati, 2018).

Table 4.1: In-sample tuning parameters using the Bernoulli deviance

| **Tuning parameters** | **Alternatives** | **Result train = 0.3** | **Result train = 0.5** | **Result train = 0.6** |
|---|---|---|---|---|
| Shrinkage ($\nu$) | 0.001, 0.005, 0.01 | 0.001 | 0.01 | 0.01 |
| Interaction depth | 3, 4, 5 | 5 | 3 | 3 |
| n.minobsinnode | 5, 10 | 5 | 10 | 10 |
| Number of trees | up to 3000 | 1759 | 117 | 320 |

Table 4.1 and 4.2 shows all of the different possibilities of the tuning parameters and the chosen tuning parameters for the in-sample analysis. The tuning parameters are chosen by the method described above. In the case of Bernoulli deviance, the chosen tuning parameters when training 30 % of the data, are then a shrinkage of 0.001, an interaction depth of 5, minimum 5 observations in the end nodes of the trees and 1759 tress.

Both of the tables show that the chosen number of trees are a lot smaller than 3000. The only case were it is above 400 is for the Bernoulli deviance when the training fraction is 30%. The reason may be that empirically smaller values of $\nu$ require larger values of trees (Hastie et al., 2008, p. 365).

Table 4.2: In-sample tuning parameters using the AdaBoost exponential loss function

| **Tuning parameters** | **Alternatives** | **Result train = 0.3** | **Result train = 0.5** | **Result train = 0.6** |
|---|---|---|---|---|
| Shrinkage ($\nu$) | 0.001, 0.005, 0.01 | 0.01 | 0.01 | 0.005 |
| Interaction depth | 3, 4, 5 | 3 | 5 | 3 |
| n.minobsinnode | 5, 10 | 10 | 10 | 10 |
| Number of trees | up to 3000 | 133 | 79 | 265 |

In all of the different training fractions, table 4.2 has always a smaller number of trees compared to 4.1. These are then the parameters that will be used in section 5.1. A code snippet, which shows an example of the implementation, is found in Appendix B.

The construction of the out-of-sample experimental design is a bit differently. In addition to training and validation, there is also a testing part. This is data that has not been used in the in-sample analysis. The

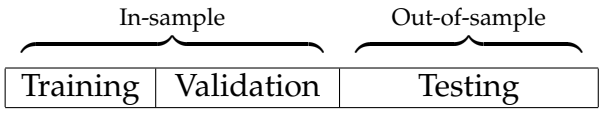



Figure 4.2: Obtaining out-of-sample results

procedure is to tune the data to the variables in $X_{t_1-h}$, $X_{t_2-h-1}$ for $Y_{t_1}, Y_{t_2-1}$. This is done by using 60% of the data for training and then 40% for validation. This is in order to avoid overfitting when finding the tuning parameters. Then the boosting models are fitted. Since the predictions are made for $Y_{t_2}$ using data from $X_{t_2-h}$, and the training fraction is 100% in the fitting process. Then the predictions out-of-sample are made out-of-sample by using the in-sample fitting process. There is a code example of this in Appendix B.

### 4.2.2 Bayesian Model Averaging

There are also some choices that need to be made when doing the BMA-analysis in the BAS-package. Logistic regression is the most important model for binary data. It is then usual to assume a binomial distribution with a logit link (Agresti, 2015, p. 165). This is the specification that needs to be made to get a logistic regression model in the BAS-package.

One of the choices that need to be made in Bayesian analysis are the priors. The prior on the coefficients for the model is called the betaprior and in equation (3.8) it is $Pr(\beta_i \mid M_i)$. I have chosen the Bayesian

information criterion (BIC) as the betaprior in this thesis. This is because it is considered a consistent estimate of the marginal likelihood and is frequently used in applied work (Berge, 2015).

The modelprior is the family of the prior distribution on the models. In this case, I have used a uniform prior. This means that all of the models are equally likely to be drawn. The reason why this is chosen is that there is no information which indicates that one model is better than the other.

The method that I use to manage the summation is MCMC, which is covered in section 3.2. The BAS-package use a combination of the random walk Metropolis-Hastings algorithm with a random swap. This means that the variable is swapped with a variable included with a variable that is currently excluded (Clyde, 2018). The number of iterations is 100000 in the in-sample part and 1000 in the rolling estimation part. The rest of the specifications in the algorithm are set as default parameters in the BAS package and is found in Clyde (2018).

### 4.2.3 AUROC

To compare the classification abilities of my models I use the Receiver Operating Characteristic curve, which plots the full mapping of the false positive rate, across different values of the threshold parameter. To assess the recession classification abilities of the various models, I follow Berge & Jordá (2011) and Liu & Moench (2016) and calculate the Area Under Receiver Operating Characteristic (AUROC), which takes every point on the ROC-curve into account. The ROC-curve is a curve plotting the true positive rate against the false positive rate. This is illustrated in figure 4.3. True positive rate in this setting means that the recession is predicted correctly. False positive rate means cases where there is predicted a recession, but there is actually not a recession. The AUROC statistic has a lower bound of 0 and an upper bound of 1, where higher values indicate better classification. If the AUROC-value is 0.5, it predicts as good as random guessing. This is because random guessing produces a diagonal line in the ROC-curve, where the area under the curve is 0.5 (Fawcett, 2006). This is the diagonal line indicated in figure 4.3. A model with an AUROC-value of below 0.5 performs worse than random guessing and should not be used. Figure 4.3 shows the ROC-curve for a case where the AUROC-value is 0.916.

The choice of using AUROC as a measure of forecasting performance is based on having a measure which is independent of the incidence of the forecasting event. The AUROC-value is independent of the incidences which makes it a good measure of classification abilities when forecasting recessions (Berge, 2015).

The interpretation of the AUROC-value of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive case

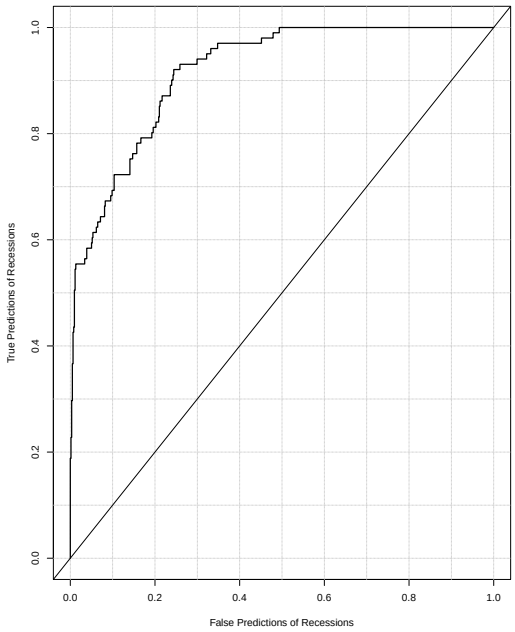



Figure 4.3: ROC-curve

higher than a randomly negative case (Fawcett, 2006). If I mix the order of recessions and non-recessions and the classifier should rank them, then the probability that a recession is ranked higher than a non-recession is determined by the AUROC-value. When the AUROC-value is 50%, the probability of ranking a recession higher than a non-recession is 50%, which is random guessing.

# 5 Results

This section presents the results for both the in-sample and out-of-sample analysis for boosting and BMA. I will first present the results for boosting and BMA, respectively. I will then compare the results from the two methods. When presenting the results, I will also show and discuss what are the most important predictors for the two methods.

## 5.1 In-sample Results

### 5.1.1 Boosting

As I stated in section 3.1, is there no straightforward interpretation of relative importance in boosting. The most important empirical property

is that the sum of the relative importance for all of the different variables is 100. This means that a higher value of relative importance indicates a more important variable for the final predictions (Ng, 2014).

Table 5.1 shows the predictors with the highest relative importance for the various boosting specifications. Table (a), (b) and (c) shows the Bernoulli deviance as a loss function, while table (d), (e) and (f) shows the results when using the AdaBoost exponential loss. The predictors are sorted according to the top ten variables with the highest relative influence. $s + h$ indicates at which lag the variable is chosen from. The forecast horizon is six months ahead, which means that $h$ is always 6. Behind the top ten most influential variables, it can potentially be a combination of one variable at four different lags. The reason why the top ten variables with higher relative importance is chosen is to have a closer look at the reasonably more important variables. The rest of the variables are also included in the predictions, but the relative importance of many of the variables are very small and will therefore have a small effect on the final predictions.

Table 5.1: Most important in-sample predictors, boosting

| **Method** | **Variable (Code)** | **h+s** | **RI** |
|---|---|---|---|
| (a) Train = 30%, Bernoulli | 3mo-FF spread (TB3SMFFM) | 6 | 9.31 |
| | Baa-FF spread (BAAFFM) | 6 | 6.92 |
| | Aaa-FF spread (AAAFFM) | 6 | 6.10 |
| | Real M2 stock (M2REAL) | 9 | 5.20 |
| | Aaa-FF spread (AAAFFM) | 7 | 4.04 |
| | 3mo-FF spread (TB3SMFFM) | 7 | 3.72 |
| | Real M2 stock (M2REAL) | 8 | 3.59 |
| | Baa-FF spread (BAAFFM) | 8 | 3.59 |
| | CP-FF spread (COMPAPFFx) | 9 | 3.31 |
| | CP-FF spread (COMPAPFFx) | 8 | 3.09 |
| (b) Train = 50%, Bernoulli | 6mo-FF spread (TB6SMFFM) | 6 | 11.48 |
| | Aaa-FF spread (AAAFFM) | 7 | 10.62 |
| | 1yr-FF spread (T1YFFM) | 6 | 9.37 |
| | 1yr-FF spread (T1YFFM) | 7 | 7.20 |
| | Aaa-FF spread (AAAFFM) | 6 | 5.37 |
| | 6mo-FF spread (TB6SMFFM) | 7 | 3.70 |
| | 3mo-FF spread (TB3SMFFM) | 6 | 2.90 |
| | Aaa-FF spread (AAAFFM) | 9 | 2.49 |
| | Baa-FF spread (BAAFFM) | 6 | 2.44 |
| | Hous. Permit MW (PERMITMW) | 9 | 2.29 |
| (c) Train = 60%, Bernoulli | Aaa-FF spread (AAAFFM) | 6 | 8.95 |
| | Aaa-FF spread (AAAFFM) | 7 | 8.81 |
| | Aaa-FF spread (AAAFFM) | 7 | 5.37 |
| | Aaa-FF spread (AAAFFM) | 8 | 5.04 |

| | | | |
|---|---|---|---|
| | VXO-index (VXOCLSx) | 9 | 4.25 |
| | VXO-index (VXOCLSx) | 8 | 3.27 |
| | VXO-index (VXOCLSx) | 6 | 3.23 |
| | 3mo-FF spread (TB3SMFFM) | 6 | 2.80 |
| | VXO-index (VXOCLSx) | 7 | 2.77 |
| | 10yr-FF spread (T10YFFM) | 6 | 2.19 |
| (d) Train = 30%, AdaBoost | 3mo-FF spread (TB3SMFFM) | 6 | 11.97 |
| | Baa-FF spread (BAAFFM) | 6 | 9.43 |
| | Real M2 stock (M2REAL) | 9 | 7.38 |
| | Real M2 stock (M2REAL) | 7 | 6.34 |
| | Hous. Permit W (PERMITW) | 9 | 5.47 |
| | 3mo-FF spread (TB3SMFFM) | 7 | 4.82 |
| | Aaa-FF spread (AAAFFM) | 6 | 4.39 |
| | Aaa-FF spread (AAAFFM) | 7 | 4.28 |
| | Real M2 stock (M2REAL) | 8 | 4.01 |
| | CP-FF spread (COMPAPFFx) | 8 | 2.91 |
| (e) Train = 50%, AdaBoost | 1yr-FF spread (T1YFFM) | 6 | 13.01 |
| | 1yr-FF spread (T1YFFM) | 7 | 9.35 |
| | 6mo-FF spread (TB6SMFFM) | 6 | 8.00 |
| | Aaa-FF spread (AAAFFM) | 7 | 7.34 |
| | Aaa-FF spread (AAAFFM) | 6 | 6.72 |
| | Aaa-FF spread (AAAFFM) | 9 | 3.84 |
| | 6mo-FF spread (TB6SMFFM) | 7 | 3.77 |
| | Aaa-FF spread (AAAFFM) | 8 | 3.23 |
| | 3mo-FF spread (TB3SMFFM) | 6 | 2.64 |
| | Hous. Permits (PERMIT) | 9 | 2.48 |
| (f) Train = 60% AdaBoost | Aaa-FF spread (AAAFFM) | 7 | 10.70 |
| | Aaa-FF spread (AAAFFM) | 9 | 8.19 |
| | Aaa-FF spread (AAAFFM) | 6 | 7.94 |
| | VXO-index (VXOCLSx) | 6 | 5.67 |
| | VXO-index (VXOCLSx) | 8 | 5.01 |
| | Aaa-FF spread (AAAFFM) | 8 | 4.42 |
| | VXO-index (VXOCLSx) | 9 | 3.25 |
| | 3mo-FF spread (TB3SMFFM) | 6 | 3.24 |
| | VXO-index (VXOCLSx) | 7 | 3.22 |
| | 10yr-FF spread (T10YFFM) | 6 | 2.85 |

Note: RI stands for relative importance

Table 5.1 (a) shows the predictors with the highest relative importance when the training period consists of 30% of the data and using the Bernoulli deviance as the loss function. The most important variable in (a) is the three month Treasury spread when $s = 0$, which has a relative importance of 9.31. The rest of the important variables consists mainly of Aaa, Baa and 3-month commercial paper interest rate spreads in addition

to the real M2 money stock at different lags. In table 5.1 (b) 50% of the data is used as training sample. 6-month Treasury bill spread is the most important variable in table 5.1 (b). This has a relative importance of 11.48. In addition to spreads, is new private housing permits in the Midwest also included among top ten predictors in (b). In table 5.1 (c), where the training sample is 60%, the four variables with highest relative influence is the Aaa federal funds rate spread at different lags. The first two has a relative influence of around 8-9, while when $s = 2$ and $s = 3$ it is a bit higher than 5. Another important predictor is the VXO index. This is the stock market volatility and may be a measure of the uncertainty for the financial market. This predictor is also included at different lags.

Moreover, the results for the AdaBoost exponential loss function are presented in table 5.1 (d), (e) and (f). In table 5.1 (d), where the training fraction is 30%, has the 3-month Treasury bill minus the effective federal funds rate when $s = 0$ a relative influence of almost 12. In table 5.1 (e), where the training fraction is 50%, the interest rate spread for 6-month Treasury bill and the interest rate spread for 1-year Treasury constant maturity at two different lags have all a high relative influence of above 8. In table 5.1 (f), where the training fraction is 60%, are all of the different lags for the Aaa spreads are included among the top ten predictors. The Aaa spread with the highest relative influence is 10.70 and the Aaa federal funds rate spread at different lags is the top three predictors with highest relative importance. These predictors with highest relative importance are similar to table 5.1 (c). The VXO index is also included in this figure at many different lags. The conclusion of (f) is then that there are few included predictors but each predictor is included at many different lags.

The Aaa spread is among the top 10 variables for all the six model specifications. This indicates that the Aaa spread carries important information for predicting recessions. The Aaa spread is also often included in the same model at many different lags. Another variable that is also always included among top ten predictors is the 3-month Treasury minus the effective federal funds rate. These are the most important variables for the predictions.

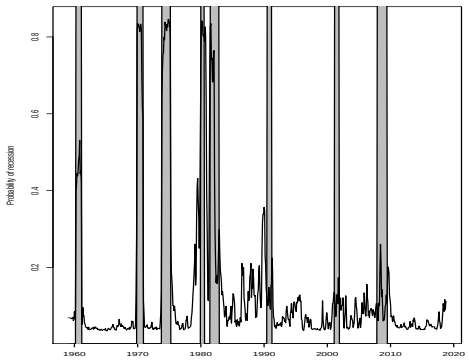
Treasury term spreads and credit/default spreads are in general considered as useful predictors for forecasting recessions. I also find that these variables are important for predicting recessions, as they have a high relevance for all my models. A large amount of previous studies tend to consider one spread variable at a time (Ng, 2014). The results in table 5.1 shows that many of the spreads have high relative importance together. They are in addition often important at different lags. This means that the spreads are not mutually exclusive and can have high predictive power together.

Two predictors that are to a large extent included as the most influential predictors are the Aaa and Baa yield minus the effective federal funds rate. From the data in McCracken (2019) it can be seen that Aaa yield has

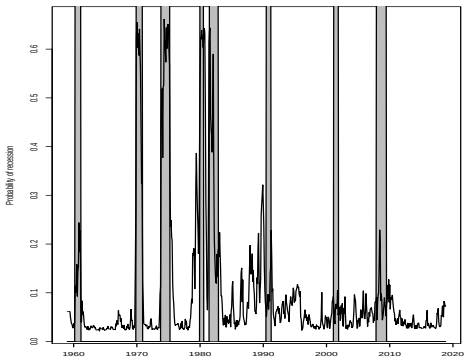
previously been stable during recessions, especially in the recessions after 1990. The reason is that these bonds have the highest rating from credit rating agencies and is considered to have low risk. These bonds also have low yield. Just before a recession, the effective federal funds rate often reaches a top, since the economy is in an expansion. This means that the spread is at it most negative. When the bottom is reached, the spread starts to grow again and that is when the recession often hits. It can be sign of recession when the spread between these two has gone from decreasing to suddenly increasing. The Baa bonds on the other hand are a bit riskier but the curvature on the spread for Baa is similar to the one for Aaa. A bottom is reached right before the recession hits. During the Great Recession, the data from McCracken (2019) shows the spread was below 2% in November 2007, while it was above 8% in October 2008.

Another variable that to a large extent is included in the tables is the permit to build new houses in different parts of the US. The reason can be that when the economy is in a good state, there are more investments and optimism so people buy more houses. It is then profitable to build new houses. When the housing permits decrease, it can be a sign of declines in jobs in the construction sector in addition to pessimism of the private economy in the future. The permit to build new houses therefore often slows down right before the recession hits. The housing market, more specifically residential investment, has also previously been considered an important indicator of measuring the state of the economy (Aastveit et al., 2018).

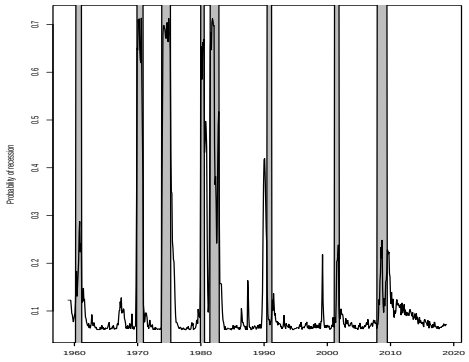
In general, table 5.1 shows that the influential predictors in the various boosting specifications are predictors that earlier have been considered as useful for predicting recessions. These variables consists primarily of different spreads, real M2 money stock, the VXO-index and housing permits. There are not big differences between the predictors chosen in the different models. The relative influence in Ng (2014) at forecast horizon six months also consists of many of the same variables, especially the different spreads. This also confirms that these predictors are important for forecasting recessions.



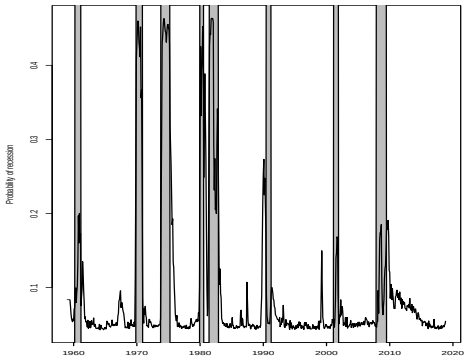

(a) Train=30%, Bernoulli distribution



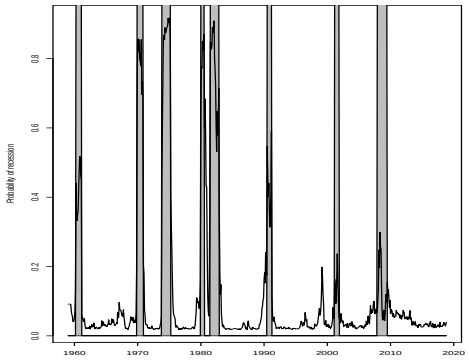

(b) Train=30%, AdaBoost exponential loss



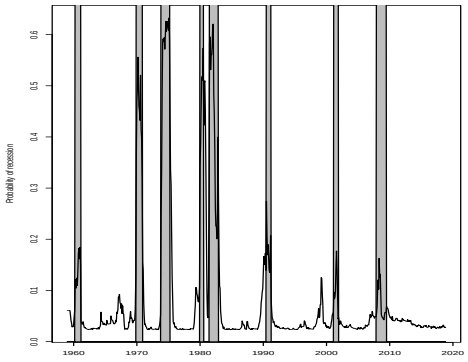

(c) Train=50%, Bernoulli distribution




(d) Train=50%, AdaBoost exponential loss




(e) Train=60%, Bernoulli distribution




(f) Train=60%, AdaBoost exponential loss

Figure 5.1: In-sample predictions, boosting

Figure 5.1 shows the predictions made in-sample for the various boosting specifications. The grey shaded areas indicate recession periods defined by NBER. The main conclusion from these plots is that all of the models predict the recessions until the early 1990s very well. When a higher fraction of the data is used for training, the predictions after 1990 also get better. This is natural because the data is trained on a higher fraction of the data so the algorithm has learned more. The predictions are more accurate for the boosting models with a training sample that consists of 60% of the sample. This is also confirmed by the AUROC-values discussed in section 5.1.3, where the models are compared to each other.

For figure 5.1 (c), (d), (e) and (f), it is only one place where a "false" spike in the probability is evident, in the late 1990s. The spike, however, is always below 20% and occurs just prior to the start of the recession in the early 2000s.

The boosting and the data that I use are similar to Ng (2014). She also finds that boosting is better at predicting the earlier recessions than the last three recessions. One reason for this may be which variables that are important for predicting recessions have changed over time. The specification of the algorithm in my thesis is trained on the first 30%, 50% or 60% of the data, which means the algorithm is unable to catch whether the important predictors has changed over time. Ng (2014) solves this by splitting the dataset in two parts and looks at the variables chosen in the two subsamples. In her analysis, the interest rate spreads are the most important predictors in the first subsample, while the real activity variables becomes more important in the second subsample. There are few variables that have a high relative importance in both of the samples. Her conclusion is that the important predictors have changed over time.

### 5.1.2 Bayesian Model Averaging

The variables I present for Bayesian model averaging are sorted according to the posterior inclusion probability (PIP). PIP is the weighted average of the posterior probability for each of the models that includes a specific predictor (Berge, 2015). If the PIP is high, it is considered important in explaining the predictions. On the other hand, the probability may be small if the predictors are highly correlated (Clyde, 2018). Another difference between the results for BMA and boosting is that the variables are only included one lag at a time. This means that there are four different sets of results. One set of results for $s = 0$, one for $s = 1$, one for $s = 2$ and one for $s = 3$.

Table 5.2: Most important in-sample predictors, BMA

| Method | Variable (Code) | PIP |
|---|---|---|
| (a) BMA, s=0 | Baa-FF spread (BAAFFM) | 0.77 |
| | Cap. util. manufac. (CUMFNS) | 0.57 |
| | Hous. Permit MW (PERMITMW) | 0.56 |
| | S&P price index (S&P 500) | 0.56 |
| | VXO-index (VXOCLSx) | 0.55 |
| | IP: Manufac. (IPMANSICS) | 0.55 |
| | Emp: Service (SRVPRD) | 0.44 |
| | Emp: Wholes. trade (USWTRADE) | 0.39 |
| | S&P industry (SP indust) | 0.32 |
| | FF (FEDFUNDS) | 0.28 |
| (b) BMA, s=1 | Emp: Wholes. trade (USWTRADE) | 0.95 |
| | Hous. Permit MW (PERMITMW) | 0.91 |
| | Baa-FF spread (BAAFFM) | 0.88 |
| | S&P price index (S&P 500) | 0.81 |
| | S&P div yield (S&P div yield ) | 0.57 |
| | S&P PE ratio (S&P PE ratio) | 0.55 |
| | Hous. Permit S (PERMITS) | 0.30 |
| | Hous. Permit (PERMIT) | 0.24 |
| | IP:Manufac. (IPMANSICS) | 0.22 |
| | S&P industry (S&P indust) | 0.21 |
| (c) BMA, s = 2 | S&P price index (S&P 500) | 0.84 |
| | Hous. Permit MW (PERMITMW) | 0.65 |
| | 10yr-FF spread (T10YFFM) | 0.52 |
| | S&P PE ratio (S&P PE ratio) | 0.46 |
| | Emp: Fin. act. (USFIRE) | 0.45 |
| | Baa-FF spread (BAAFFM) | 0.39 |
| | Emp: Wholes. trade (USWTRADE) | 0.38 |
| | Nonrev. con. cred. (CONSPI) | 0.36 |
| | Hous. Permit S (PERMITS) | 0.33 |
| | S&P div yield (S&P div yield) | 0.28 |
| (d) BMA, s=3 | Emp: Fin act. (USFIRE) | 0.63 |
| | Nonrev. con. cred. (CONSPI) | 0.58 |
| | Hous. Permit S (PERMITS) | 0.55 |
| | 10yr-FF spread (T10YFFM) | 0.52 |
| | Switz./US exch. rate (EXSZUSx) | 0.37 |
| | Aaa-FF spread (AAAFFM) | 0.36 |
| | Hous. starts (HOUSTS) | 0.35 |
| | S&P price index (S&P 500) | 0.28 |
| | Hous. Permit MW (PERMITMW) | 0.26 |
| | Emp: Wholes. trade (USWTRADE) | 0.25 |

Note: PIP stands for posterior inclusion probability

Table 5.2 presents the most important predictors chosen by BMA at different lags. In (a), where $s = 0$, the predictor with the highest posterior inclusion probability is the spread between Moody's Seasoned Baa Corporate Bond and the federal funds rate. The posterior inclusion probability for this predictor is 77%. The second most important predictor is the capacity utilization for manufacturing. The capacity utilization is defined as the ratio of actual production to maximum sustainable production and is an important measure of slack in the economy (Pierce & Wisniewski, 2018). If the capacity utilization is below 100%, it means the factories are producing less than they can. This can be an indication that the demand in the economy is low. Other than that, it is a combination of financial and macroeconomic variables that are considered most important in this model.

The most important predictor in table 5.2 for (b), where $s = 1$, is the employed in the wholesale trade. This is an important predictor for NBER when they announce a recession. NBER (2008) state

> A recession is a significant decline in economic activity spread across the economy, lasting more than a few months, normally visible in real GDP, real income, employment, industrial production, and wholesale-retail sales.

This predictor has negative growth during recessions and the turning point for the growth is often close to the recession start.

A difference between the models when $s = 1$ and $s = 0$ is that the posterior inclusion probabilities are often higher when $s = 1$ compared to if $s = 0$. The wholesale trade has a posterior inclusion probability of 95% when $s = 1$, which means that this variable is almost always included in the different models. The four most important predictors have a posterior inclusion probability of above 80% when $s = 0$. It is much higher compared to model (a) where the predictor with the highest posterior inclusion probability is only 77%.

S&P 500 index is the predictor with the highest posterior inclusion probability in (c), where $s = 2$. This index is included among top ten posterior inclusion probability in all of the other BMA model specifications as well. S&P price-earnings ratio is the fourth most important predictor. This is the ratio of the market price of the company's stocks to its earnings per share. It is then a measure of the optimism of the market when it comes to the growth prospects for the firms. The growth prospects are often low during recessions and that is why this may be an important predictor for recessions.

In (d), where $s = 3$, the most important predictor is the number of employed in financial activities. Since the employment goes down and the financial markets have tough times during recessions, a decline in the employment in this sector is expected. Nonrevolving consumer credit to personal income is the predictor with the second highest PIP in

(d). Nonrevolving loans are often flat during recessions. This is because car loans and home mortgages takes time to pay off and can therefore not adjust personal financial problems by reducing these loans (Federal Reserve Bank of St.Louis, 2010). The personal income is high during expansions and lower during recessions. The ratio of these becomes high.

Moreover, figure 5.2 shows plots of the in-sample probabilities of recession along with the NBER recession dates marked by the grey shaded areas. The plot consist of four subplots, one for each value of $s$, where (a) is when $s = 0$, (b) is when $s = 1$, (c) is when $s = 2$ and (d) is when $s = 3$. The predictions are more accurate before the early 1990s, than after. However, the BMA models seem to more frequently predict "false" recessions than the boosting models.

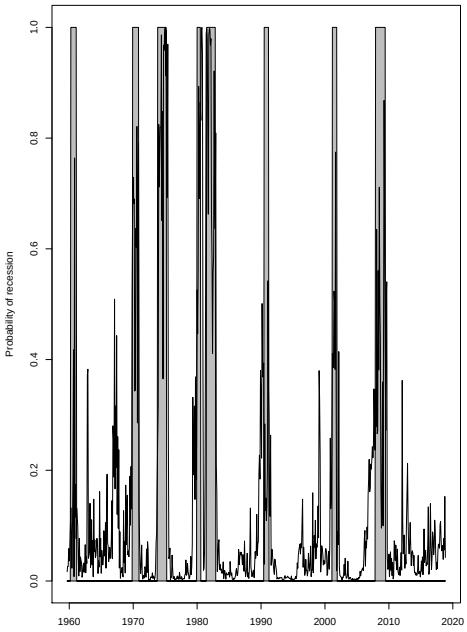
In figure 5.2 (a), when $s = 0$, the model seems to predict the recessions quite well. However, the model seems to provide somewhat noisy signals during the 1960s, indicating wrongly a high probability of a recession occurring for several periods. In the period before the recession in the beginning of the 1990s, the probability of a recession increases before the actual recession. The same happens with the Great Recession. This shows that the model has predictive power.

The predictions in figure 5.2 (b) are also in most cases accurate. The first recession in the 1960s is predicted too late. Other than that, when the probability of a recession is above 50%, it is a clear warning sign of a recession. Before the Great Recession, the probability is nicely building its way up. This happens to most of the recessions in this figure, which means that the model has predictive power.

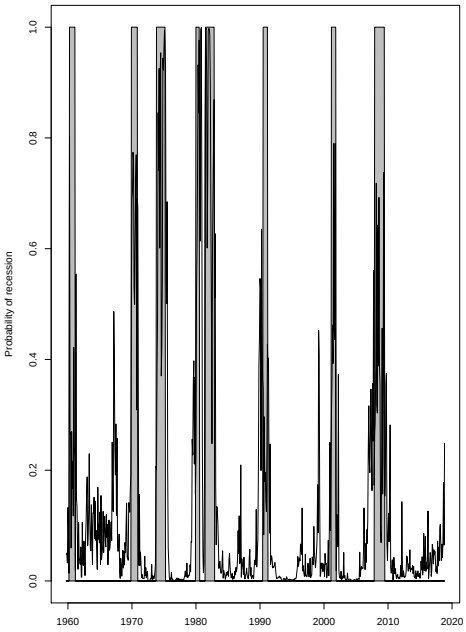
Figure 5.2 (c) shows that this model predicts recessions reasonably well. The spike before the early 2000 recession is quite high, around the same height as during the recession in the 1990s. This is then a wrongly predicted recession. As in figure (a) and (b), this model also predicts the Great Recession very well. The probability increase more and more before the recession hits. The disadvantage with the model is that it also wrongly specifies recessions and the probabilities for the wrongly specified recessions are too high.

The last subplot (d) in figure 5.2 has many spikes in the probabilities that are not recessions. The spikes are around 30%. This model also predicts the Great Recession accurately, but there are too many warning signs between recessions, which makes the model not good enough.

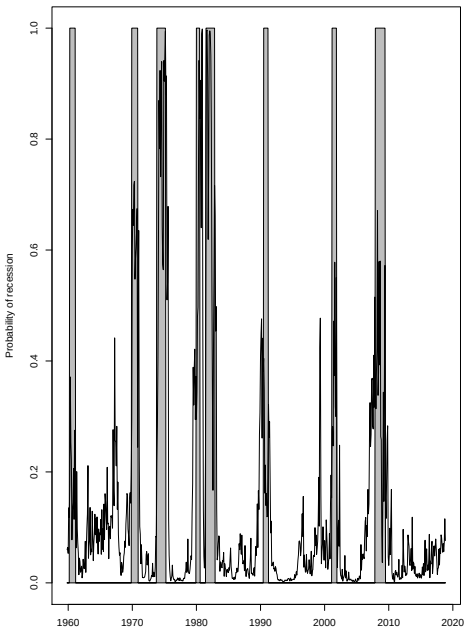
Summing up figure 5.2, the best predictive power seems to lie in (b). The spikes that is not in a recession, are either small or works as a warning sign of a recession that is on its way. For (c) and (d), the warning signs are too severe and can probably not be used for prediction.



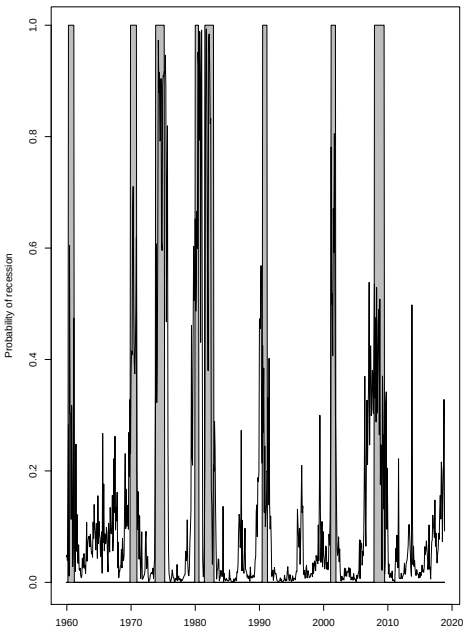

(a) $s = 0$




(b) $s = 1$




(c) $s = 2$




(d) $s = 3$

Figure 5.2: In-sample predictions, BMA

### 5.1.3 Comparison of the In-sample Results

Table 5.3: AUROC-values for in-sample analysis

| Model | AUROC-value |
|---|---|
| Boosting, train = 30 % , Bernoulli | 0.916 |
| Boosting, train = 50 % , Bernoulli | 0.932 |
| Boosting, train = 60 % , Bernoulli | 0.972 |
| Boosting, train = 30 % , AdaBoost | 0.895 |
| Boosting, train = 50 % , AdaBoost | 0.911 |
| Boosting, train = 60 % , AdaBoost | 0.965 |
| BMA, $s = 0$ | 0.933 |
| BMA, $s = 1$ | 0.953 |
| BMA, $s = 2$ | 0.944 |
| BMA, $s = 3$ | 0.924 |

Table 5.3 shows the AUROC-values for the in-sample predictions for all boosting and BMA specifications. In general, both boosting and BMA have high AUROC-values, where almost all of the specifications obtain AUROC-values above 0.9. The model with the highest AUROC-value is the Bernoulli deviance with a training fraction of 60%. The AUROC-value in this case is 0.972. The exponential AdaBoost loss function with a training fraction of 60% has the second highest AUROC-value of all the different models. BMA when $s = 1$ follows close behind. These models with the top three values have a very high AUROC-value, since all of them have an AUROC-value of above 0.95.

I create a univariate probit model as a benchmark in order to compare these values[9]. The probit model is constructed as the spread between the 3-month Treasury and 10-year Treasury[10]. The in-sample analysis is on the same sample as the boosting and BMA analysis. I then get an AUROC-value of 0.786, which is smaller than all of the values I have gotten from BMA and boosting. The boosting and BMA results in this analysis are then higher than the benchmark model when forecasting six months ahead.

In order to interpret how high these AUROC-values are compared to traditional methods, some AUROC-values from Liu & Moench (2016) are presented in table 5.4. The method used in Liu & Moench (2016) is probit regressions. The AUROC-values presented in table 5.4 are in-sample results for spread and spread-lagged-spread models in addition to the three best performing models with an additional variable for the spread-lagged-spread model. The best-performing model in table 5.4 has an AUROC-value of 0.965. This model has S&P, 1y% change as an additional hand-picked variable for predictions.

[9]The probit model is explained in Estrella & Hardouvelis (1991).
[10]This data is calculated using the dataset from McCracken (2019).

Table 5.4: AUROC-values for in-sample analysis in paper

| Paper | Variables | AUROC-value |
|---|---|---|
| Liu & Moench (2016, p.1144) | Spread(t) | 0.769 |
| | Spread(t) + Spread(t-6) | 0.878 |
| | Building permits | 0.922 |
| | S&P, 1y% change | 0.965 |
| | 5-year - FF spread | 0.933 |

Note: This is the mean of the AUROC-values in the paper. The spread is the Treasury term spread. The time horizon is January 1959 to December 2011

Because of the time frame, I can not give clear conclusions of whether boosting and BMA performs better than these traditional models. However, the results from this thesis, shows that the models in table 5.3 in some cases have higher AUROC-values than the models in table 5.4. My results then lie in the same region as the the models in Liu & Moench (2016). Both boosting and BMA have then given promising results in-sample. It is still important to note that this are only AUROC-values from one paper with different specifications of the probit models and not traditional models in general.

## 5.2 Out-of-sample Results

This section presents the results from the out-of-sample exercise. As for the in-sample exercise, I will present the results for both boosting and BMA, in addition to a discussion of what are the most important predictors. I present two different specifications of the boosting analysis and two different specifications of the BMA analysis. The predictors I present are the average of the predictors selected when I perform the rolling estimation analysis. As in the in-sample case, I will focus on the top ten predictors, those with the highest relative importance for boosting, and those with the highest posterior inclusion probability. The in-sample results are relatively good, but it does not necessarily reflect the out-of-sample predictive ability of the various models. If predictions is the key objective, which is the case in this thesis, it is the out-of-sample results that really matters.

Table 5.5: Most important out-of-sample predictors

| Model | Variable (Code) | $h + s$ | RI/PIP |
|---|---|---|---|
| (a) Boosting, Bernoulli | 10yr-FF spread (T1YFFM) | 6 | 4.94 |
| | Aaa-FF spread (AAAFFM) | 9 | 4.36 |
| | 1yr-FF spread (T1YFFM) | 9 | 4.25 |
| | 3mo-FF spread (TB3SMFFM) | 6 | 4.19 |

| | | | |
|---|---|---|---|
| | 6mo-FF spread (TB6SMFFM) | 6 | 4.02 |
| | Aaa-FF spread (AAAFFM) | 6 | 3.91 |
| | Aaa-FF spread (AAAFFM) | 8 | 3.37 |
| | Aaa-FF spread (AAAFFM) | 7 | 3.16 |
| | 10yr-FF spread (T10YFFM) | 9 | 2.54 |
| | 1yr-FF spread (T1YFFM) | 7 | 2.53 |
| (b) Boosting, AdaBoost | Aaa-FF spread (AAAFFM) | 9 | 5.84 |
| | 1yr-FF spread (T1YFFM) | 6 | 5.75 |
| | 1yr-FF spread (T1YFFM) | 9 | 5.30 |
| | 3mo-FF spread (TB3SMFFM) | 6 | 4.80 |
| | 6mo-FF spread (TB6SMFFM) | 6 | 4.67 |
| | Aaa-FF spread (AAAFFM) | 8 | 3.66 |
| | Aaa-FF spread (AAAFFM) | 6 | 3.41 |
| | Aaa-FF spread (AAAFFM) | 7 | 3.37 |
| | 10yr-FF spread (T10YFFM) | 9 | 2.94 |
| | 3mo-FF spread (TB3SMFFM) | 7 | 2.65 |
| (c) BMA, $s = 0$ | Baa-FF spread (BAAFFM) | 6 | 0.39 |
| | CP-FF spread (COMPAPFFx) | 6 | 0.32 |
| | 10yr-FF spread (T10YFFM) | 6 | 0.30 |
| | Aaa-FF spread (AAAFFM) | 6 | 0.29 |
| | 5yr-FF spread (T5YFFM) | 6 | 0.28 |
| | 3mo-FF spread (TB3SMFFM) | 6 | 0.28 |
| | 6mo-FF spread (TB6SMFFM) | 6 | 0.27 |
| | 1yr-FF spread (T1YFFM) | 6 | 0.23 |
| | S&P PE ratio (S&P PE ratio) | 6 | 0.22 |
| | Switz./US exch. rate (EXSZUSx) | 6 | 0.21 |
| (d) BMA, $s = 1$ | Baa-FF spread (BAAFFM) | 7 | 0.36 |
| | 10yr-FF spread (T10YFFM) | 7 | 0.32 |
| | 5yr-FF spread (T5YFFM) | 7 | 0.30 |
| | 3mo-FF spread (TB3SMFFM) | 7 | 0.30 |
| | Aaa-FF spread (AAAFFM) | 7 | 0.28 |
| | VXO-index (VXOCLSx) | 7 | 0.27 |
| | 6mo-FF spread (TB6SMFFM) | 7 | 0.27 |
| | 1yr-FF spread (T1YFFM) | 7 | 0.25 |
| | Switz./US exch. rate (EXSZUSx) | 7 | 0.25 |
| | Hous. Permit MW (PERMITMW) | 7 | 0.21 |

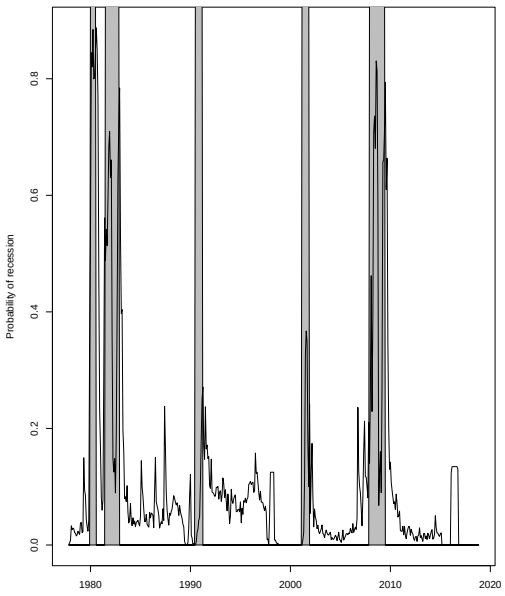
Note: RI stands for relative importance and PIP stands for posterior inclusion probability. RI is reported in (a) and (b), while PIP is reported in (c) and (d)

Table 5.5 presents the ten most relevant predictors for two boosting specifications and two BMA specifications, respectively. In table 5.5 (a) and (b), I report the relative importance for boosting, where the loss function in (a) is the Bernoulli deviance, while in (b) it is the AdaBoost exponential loss. The predictor in table 5.5 (a) with the highest relative

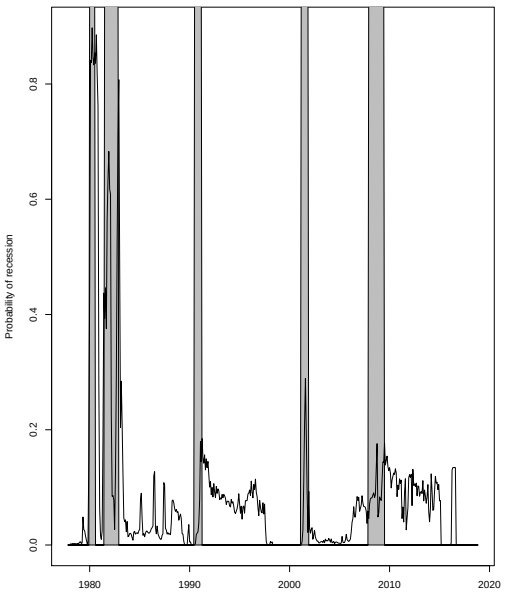
importance is the spread between 1-year Treasury constant maturity and the federal funds rate. This predictor is also one of the most important predictors in the in-sample cases as well. In table 5.5 (a), the spreads are prominent. Especially the Aaa federal funds rate spread, which is included at all lags. In fact, there are only spreads among the top ten variables with the highest relative importance. The levels of the relative importance for the predictors are a bit lower than for the in-sample cases. Table 5.5 (b) also has the Aaa spread as the predictor with the highest relative importance. The important predictors are almost as in (a), but the ordering has changed. The relative importance values are also smaller here than in the in-sample cases.

In table 5.5 (c) and (d) I report the posterior inclusion probabilities BMA, where (c) is when $s = 0$ and (d) is when $s = 1$. The posterior inclusion probability are for both model specifications considerably smaller than those obtained in the in-sample analysis. This is probably because what are the important predictors have changed during this time period. This will affect the mean of the posterior inclusion probabilities. The federal funds rate spreads dominates the list of the top ten predictors in both (c) and (d) with highest posterior inclusion probability. The predictor with the highest posterior inclusion probability in both (c) and (d) is the spread between Moody's Seasoned Baa Corporate Bond and the federal funds rate, the same as in table 5.2. In the BMA model when $s = 1$ is also housing permits in the Midwest included among the top ten predictors. One surprising variable in (c) and (d) is the exchange rate with Switzerland.

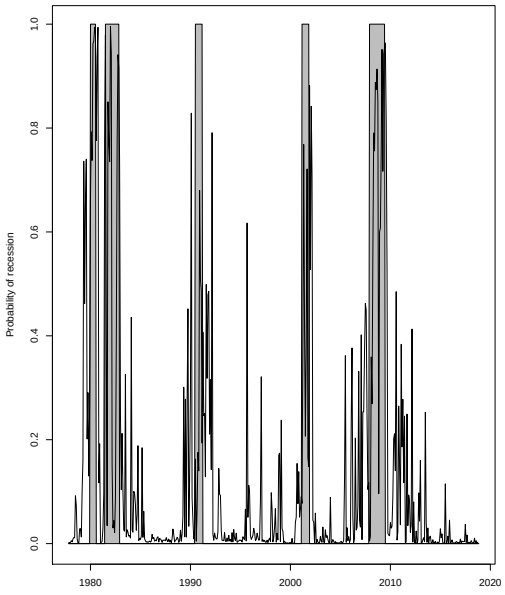
To sum up, various interest spreads are the most important predictions in the out-of-sample analysis.



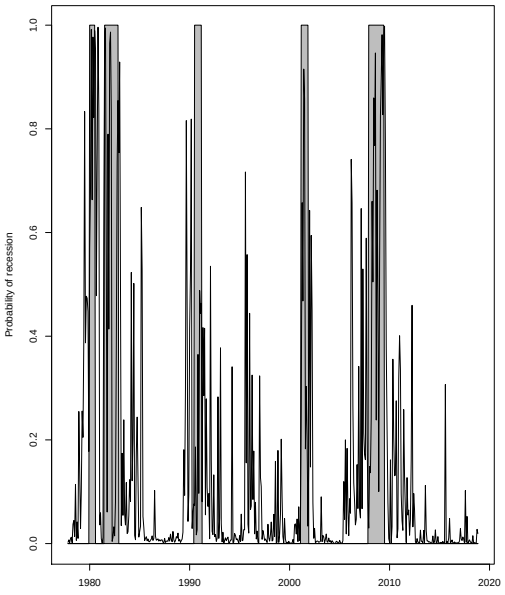

(a) Boosting, Bernoulli




(b) Boosting, AdaBoost




(c) BMA, $s = 0$




(d) BMA, $s = 1$

Figure 5.3: Out-of-sample predictions

Figure 5.3 (a) and (b) show the plots for the out-of-sample predictions for boosting, while (c) and (d) show plots of the out-of-sample predictions for the various BMA specifications. The NBER recession dates are marked by the grey shaded areas. Figure (a) shows the plot for the predictions using the Bernoulli deviance in the boosting approach. The results show that the first two recessions are predicted reasonably well. Between the last recession in the 1980s and the recession in the early 1990s, there are small spikes in the probability of a recession, but they barely exceed 20%. It is a spike right before the recession in the early 1990s, but the probability of a recession decreases again after this spike. In the middle of the recession, the probability of a recession increases again. If the spike before the recession had been a bit higher, it could have been a warning signal of a recession. Since it is not, it means that the recession in the early 1990s is predicted a bit late. The model predicts the recession in the early 2000s, but this is also predicted a bit late. On the other hand, the Great Recession is very well predicted with this model. The probability of a recession starts to increase some months prior to the recession. By the time the US economy enters into the recession in December 2007 has the model already provided a strong signal of a recession. In general, and perhaps as expected, the results from the in-sample analysis are somewhat stronger than those for the out-of-sample analysis. However, the model still provides in most cases clear spikes in the recession probabilities around recession dates.

The out-of-sample predictions for the AdaBoost exponential loss functions is plotted in 5.3 (b). The accuracy of the predictions is more mixed for this model specification. While the model captures the two recessions in the 1980s and the early 2000 recession quite well, it struggles to provide a timely signal for the early 1990 recession. In contrast to all other models that I consider, it does not capture the Great Recession. There are some small spikes in the periods between recessions, but they are in general small. Compared to boosting with the Bernoulli deviance, the boosting with the AdaBoost seems worse at predicting recessions.

Figure 5.3 (c) shows results for BMA when $s = 0$. This model seems to predict recessions well. However, compared to the two boosting models, the BMA model also seem to provide more noisy signals, with some false spikes. It wrongly predicts a recession in the beginning of the sample and there are some spikes between the latest recession in the 1990s and the recession in the early 2000s. There is also a peak of the predictions with a probability of recession of 60% between the recession in the 1990s and the early 2000s. On the other hand, the Great Recession is predicted early and the predictions are nicely building its way up. Figure 5.3 (d) shows results for BMA, $s = 1$. In general this model provided noisy signals with several wrongly predicted recessions. There are spikes both before and after some of the recessions have occurred. The Great Recession is also predicted accurately in this model specification.

### 5.2.1 Comparison of the Out-of-sample Results

Table 5.6: AUROC-values for out-of-sample analysis

| Model | AUROC-value |
|---|---|
| Boosting; Bernoulli | 0.855 |
| Boosting; AdaBoost | 0.780 |
| BMA; $s = 0$ | 0.892 |
| BMA; $s = 1$ | 0.869 |

The AUROC-values for the out-of-sample predictions are shown in table 5.6[11]. BMA with $s = 0$ has the highest AUROC-value. Among the boosting models, boosting with the Bernoulli deviance has the highest AUROC-value. Both of the BMA specifications have a higher AUROC-value than for boosting. The AUROC-values for boosting with the Bernoulli deviance and the BMA analysis are all above 0.85. The differences between the AUROC-values for these models are in general small. The conclusion is that except for boosting with the AdaBoost loss function, the AUROC-values are almost the same.

To get a sense of whether these AUROC-values are high or not, I also conduct out-of-sample predictions using a probit model[12] for the spread between the 3-month Treasury and 10-year Treasury as predictor. The out-of-sample analysis is then done using rolling estimations in the same way as for all the other models. For this benchmark probit model I achieve an AUROC-value of 0.700, which is considerably smaller than the AUROC-values from the boosting and BMA models.

Liu & Moench (2016) report AUROC-values for probit models with the spread and the difference between the spread and lagged spread using a probit model. They first estimate these models based on a sample from January 1959 to August 1985. Based on the estimated parameters, they then construct recursive out-of-sample forecasts for the period September 1985 to December 2011. They obtain an AUROC-value of 0.842 for the spread and 0.910 for the spread-lagged-spread difference.

Aastveit et al. (2018, Appendix D) show values for predicting US recessions using panel probit models with residential investment, term spread, stock prices, consumer confidence survey and oil price, respectively, as regressors. Their out-of-sample forecast period is from 1990Q1 to 2014Q4. Their out-of-sample analysis is based on an expanding window estimation. They estimate a probit model using data from 1960Q1 to 1989Q4. The parameters that are estimated are then used to predict recessions over the next 6 quarters. Next, they reestimate the models, using data until 1990Q1 and predict recessions again. In their out-of-sample analysis, the

[11]The AUROC-values for BMA when $s = 2$ is 0.852 and 0.861 when $s = 3$.

[12]Same construction of the probit model as in the in-sample analysis.

probit model with residential investment has the highest AUROC-value of 0.901. For specifications that use other predictors, the AUROC-values are mostly between 0.8 and 0.9 in their analysis.

Although my main results are not directly comparable with Aastveit et al. (2018) and Liu & Moench (2016) as the time periods differ, the AUROC-values that I obtain in my analysis are in the same region as what these papers report.

# 6 Discussion

Machine learning techniques have over the past couple of decades become popular in statistics and computer science. Recently, these techniques have also gained popularity in other fields, for example in medicine, ecology, meteorology and economics. There are reasons to believe that these techniques will continue to grow in popularity as the data availability increases. In this thesis, I have applied one particular machine learning technique, boosting, in order to predict recessions in the US. I have also compared boosting to a more traditional model, Bayesian model averaging. In this thesis has BMA also been implemented in a data-rich environment.

Chen et al. (2011) points at two problems in the existing literature for predicting recessions; not enough explanatory variables and too restrictive specifications. Boosting is a method that can select from a large number of explanatory variables. The variables are picked and weighted before they are combined to a final model. The variables that are never picked, do not contribute in the final model used for predictions. In addition, boosting is also able to select different variables at different lags without causing overfitting. In the gbm-package in R that I use for the boosting analysis, there is a function called the validation error and the training error. The validation error shows the value of the loss function for each of the boosting iterations evaluated at the validation data (Greenwell et al., 2019). The training error shows the same, but on the training data. The number of iterations is decided by the minimized validation error. This is the solution for boosting in a data-rich environment to avoid overfitting in this thesis. Boosting then picks the most important variables and combine them to model used for predictions. This means that boosting addresses two of the limitations with the existing literature on predicting recessions.

Machine learning techniques have often been criticized for being ”black box” methods where the final results are the only thing that matters. As a result it is often not very transparent how these results have been obtained and how to interpret the mechanisms lying behind the results. This is not the case for boosting, as the predictors that contribute

the most to the predictions can be analyzed in a very transparent way. As a result, it is easy to study what are the variables that contributes the most to improving the forecast accuracy.

BMA, which is a frequently used method in economics, also has some advantages in this analysis. The results are obtained a lot faster compared to boosting. The main reason is that the parameters do not need to be tuned as they do for boosting. Another advantage is that this method requires less data. While boosting requires additional data observations for training and validation, BMA can be implemented straightaway.

However, the main reason for using BMA is to take into account model uncertainty. Selecting one particular model can for example lead to riskier decisions because the model uncertainty is being ignored. This can be especially relevant in economic applications where there is a large number of potential explanatory variables. For predicting recessions, a large amount of predictors have been proposed by earlier literature. It is then an advantage that BMA takes this model uncertainty into account.

Moreover, I would like to highlight that the most important predictors for both methods in the data-rich environment are previously known predictors considered useful for predicting recessions. This is reassuring and a reason to further investigate these methods in a data-rich environment.

There are also some limitations with the methods that I use in this thesis that are important to be aware of. A limitation of boosting is that it requires a large amount of data. Boosting requires training and validation to find the preferred model and there are clear limitations when boosting can be used. Since recessions are rare events, there is a limitation to what countries one can apply the method of boosting for predicting recessions. One of the reasons for using US data, is that data exists for a long time period. For most other countries, the data sample covers a substantially smaller time period, making it infeasible to use boosting.

Ng (2014) also discussed how well boosting predicts recessions. She states that the fitted probabilities are not sufficiently persistent because the model dynamics is only driven by the predictors. The predictors are frequently selected at isolated lags, but the lags are not consecutive. She argues further that by improving the model dynamics will lead to better predictions and it will probably not change which variables are considered most important. She also argues that incorporating dynamics further is a topic for future research.

It has been claimed that BMA is robust to overfitting problems. Buntine (1992) argued that BMA removes the overfitting problem by "canceling out" the effects of the overfitted models. Domingos (2000) on the other hand, claims that BMA is very sensitive to overfitting. The conclusion of whether BMA is robust to overfitting or not is therefore unclear.

When it comes to future usage of boosting, Döpke et al. (2017) stated that

> From the point of view of applied business-cycle forecasting, machine-learning techniques are not a substitute for experience in business-cycle forecasting in general, and in interpreting changes in estimated recession probabilities in particular. (p. 755)

On the other hand, they also state that their boosting approach can be a useful technique for analyzing economic policy. Their main view of future usage of boosting in practical business cycle environment is that it is limited to a complement of the probit model approach.

However, in the future, detailed data of human behavior, details of people and different societies and so on may be available. As the data availability grows, it becomes more and more important to incorporate the big amount of information. As an example of this in the context of recessions is that before a recession, it is likely that people in general change their behavior, for instance by changing their spending behavior. Can this in someway predict a recession? Can it be the case when the human prospect change, the probability of a recession increases? Or is the human behavior a result of media or other external factors? As the amount of information grows, it is important to be able to handle this new information. Boosting, and possibly BMA, are methods that can handle this big amount of information and may therefore be important for future prediction methods. Machine learning technique are therefore likely to also be highly relevant in economics, a hot topic for future research.

# 7 Conclusion

In this thesis, I apply the methods of boosting and BMA for predicting US recessions. In doing so, I consider 128 different predictors. I perform both in-sample and out-of-sample predictions, where the forecasting horizon is six months ahead. One advantage of using boosting and BMA is that they can incorporate all of the information from this big dataset. Many traditional econometric methods have a problem handling high dimensional data. This may result in a model which does not include enough explanatory variables.

I consider six different boosting specifications and four different BMA specifications in the in-sample analysis and two different boosting specifications and two different BMA specifications in the out-of-sample analysis. For the in-sample analysis, I find that the AUROC-values for most of the specifications exceeds 0.9. For the out-of-sample analysis, the results are somewhat more mixed. The specifications for BMA and one boosting specification provides fairly accurate forecasts with an AUROC-value of above 0.85. However, the other boosting specification provides

forecasts with a lower AUROC-value.

For both of the methods I also investigate what are the most important predictors. The predictors found to be most important are well known economic and financial indicators, which have commonly been used for predicting recessions in previous studies. I find that the most important predictors are the Treasury term federal funds rate spreads at different time horizons and housing permits for various regions in the US. I also find that the most important predictors are mostly the same for both in-sample and out-of-sample analysis. In my analysis, I also find that several different interest rate spreads seem to be important predictors both in-sample and out-of-sample. An interesting finding is that the spreads are often included together, which means that they are not mutually exclusive, but have high predictive power together. This differ from most of the previous studies, which typically only include one spread as a predictor.

The goal of this thesis has been to apply machine learning techniques for predicting recessions. I find that these methods provide promising results for forecasting US recessions. These methods should therefore be further explored for predicting recessions, as well as other economic variables. I leave this as a topic for future research.

# Bibliography


Aastveit, K. A., Anundsen, A. K., & Herstad, E. I. (2018). Residential investment and recession predictability. *International Journal of Forecasting*. doi: https://doi.org/10.1016/j.ijforecast.2018.09.008

Agresti, A. (2015). *Foundations of linear and generalized linear models*. Hobroken, NJ: John Wiley & Sons.

Amini, S. M., & Parmeter, C. F. (2011). Bayesian model averaging in R. *Journal of Economic and Social Measurement*, *36*(4), 253–287. Retrieved from `https://content.iospress.com/articles/journal-of-economic-and-social-measurement/jem00350`

Athey, S. (2018). The impact of machine learning on economics. In *The economics of artificial intelligence: An agenda* (pp. 507–547). University of Chicago Press. Retrieved from `http://www.nber.org/chapters/c14009`

Berge, T. J. (2015). Predicting recessions with leading indicators: Model averaging and selection over the business cycle. *Journal of Forecasting*, *34*(6), 455–471. doi: doi.org/10.1002/for.2345

Berge, T. J., & Jordá, O. (2011). Evaluating the classification of economic activity into recessions and expansions. *American Economic Journal: Macroeconomics*, *3*(2), 246–277. Retrieved from `http://www.aeaweb.org/articles?id=10.1257/mac.3.2.246`

Buntine, W. L. (1992). *A theory of learning classification rules* (Doctoral dissertation, The University of Technology Sydney). Retrieved from `https://pdfs.semanticscholar.org/f44f/bc2fb3df4425654ae429c6cd1e175c3a522d.pdff`

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250. Retrieved from `https://www.geosci-model-dev.net/7/1247/2014/gmd-7-1247-2014.pdf`

Chauvet, M. (1998). An econometric characterization of business cycle dynamics with factor structure and regime switching. *International Economic Review*, *39*(4), 969–996. Retrieved from http://www.jstor.org/stable/2527348

Chauvet, M., & Piger, J. (2008). A comparison of the real-time performance of business cycle dating methods. *Journal of Business & Economic Statistics*, *26*(1), 42–49. doi: 10.1198/073500107000000296

Chen, Z., Iqbal, A., & Lai, H. (2011). Forecasting the probability of US recessions: a probit and dynamic factor modelling approach. *Canadian Journal of Economics/Revue canadienne d'conomique*, *44*(2), 651–672. doi: 10.1111/j.1540-5982.2011.01648.x

Clyde, M. (2018). BAS: Bayesian Variable Selection and Model Averaging using Bayesian Adaptive Sampling [Computer software manual]. Retrieved from https://cran.r-project.org/web/packages/BAS/index.html (R package version 1.5.3)

Clyde, M., Ghosh, J., & Littman, M. (2011). Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics*, *20*(1), 80–101. doi: 10.1198/jcgs.2010.09049

Domingos, P. (2000). Bayesian averaging of classifiers and the overfitting problem. In *ICML* (Vol. 2000, pp. 223–230). Retrieved from https://homes.cs.washington.edu/~pedrod/papers/mlc00b.pdf

Döpke, J., Fritsche, U., & Pierdzioch, C. (2017). Predicting recessions with boosted regression trees. *International Journal of Forecasting*, *33*(4), 745–759. doi: doi.org/10.1016/j.ijforecast.2017.02.003

Eilers, P. H. C., & Marx, B. D. (1996). Flexible smoothing with *B*-splines and penalties. *Statistical Science*, *11*(2), 89–102. Retrieved from http://www.jstor.org/stable/2246049

Estrella, A., & Hardouvelis, G. A. (1991). The term structure as a predictor of real economic activity. *The Journal of Finance*, *46*(2), 555–576. Retrieved from http://www.jstor.org/stable/2328836

Estrella, A., & Mishkin, F. S. (1996). The yield curve as a predictor of US recessions. *Current Issues in Economics and Finance*, *2*(7), 1–6. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1001228

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. Retrieved from http://www.sciencedirect.com/science/article/pii/S016786550500303X

Federal Reserve Bank of St.Louis. (2010). Economic snapshot: Consumer credit. *Inside the Vault*, *14*(2), 9. Retrieved from https://www.stlouisfed.org/~/media/files/pdfs/publications/pub_assets/pdf/itv/2010/itv_fall_10.pdf

Fossati, S. (2015). Forecasting US recessions with macro factors. *Applied Economics*, *47*(53), 5726–5738. doi: 10.1080/00036846.2015.1058904

Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Information and Computation*, *121*(2), 256–285. Retrieved from http://www.sciencedirect.com/science/article/pii/S0890540185711364

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139. Retrieved from http://www.sciencedirect.com/science/article/pii/S002200009791504X

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Annals of Statistics*, *28*(2), 337–407. doi: 10.1214/aos/1016218223

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. Retrieved from http://www.jstor.org/stable/2699986

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (Third ed.). Abingdon on Thames: Taylor & Francis.

Greenwell, B., Boehmke, B., Cunningham, J., & GBM developers. (2019). Package gbm: Generalized boosted regression models [Computer software manual]. Retrieved from https://cran.r-project.org/web/packages/gbm/index.html (R package version 2.1.5)

Hastie, T., Tibishirani, R., & Friedman, J. (2008). *The elements of statistical learning* (Second ed.). New York, NY: Springer Science+Business Media.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*(4), 382–401. Retrieved from http://www.jstor.org/stable/2676803

Holden, H. (2018). gradient - matematikk. In *Store norske leksikon.* Retrieved 14.04.2019, from https://snl.no/gradient_-_matematikk

Koop, G. (2003). *Bayesian econometrics* (First ed.). Chichester: John Wiley & Sons Ltd.

Liu, W., & Moench, E. (2016). What predicts US recessions? *International Journal of Forecasting*, *32*(4), 1138–1150. Retrieved from http://www.sciencedirect.com/science/article/pii/S0169207016300279

Mayr, A., Binder, H., Gefeller, O., & Schmid, M. (2014). The evolution of boosting algorithms – from machine learning to statistical modelling. *arXiv preprint arXiv:1403.1452*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.759.6052&rep=rep1&type=pdf

McCracken, M. W. (n.d.). *FRED-MD Updated Appendix.* Retrieved 08.04.2019, from https://s3.amazonaws.com/files.fred.stlouisfed.org/fred-md/Appendix_Tables_Update.pdf

McCracken, M. W. (2019). *FRED-MD (2018-12.csv)* [Data file]. Available from: https://research.stlouisfed.org/econ/mccracken/fred-databases/.

McCracken, M. W., & Ng, S. (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, *34*(4), 574–589. doi: 10.1080/07350015.2015.1086655

NBER. (2008). The NBER's recession dating procedure. Retrieved from https://www.nber.org/cycles/jan08bcdc_memo.html

NBER. (2010). The NBER's Business Cycle Dating Committee. Retrieved from https://www.nber.org/cycles/recessions.html

NBER. (2012). US Business Cycle Expansions and Contractions. Retrieved from https://www.nber.org/cycles/US_Business_Cycle_Expansions_and_Contractions_20120423.pdf

Ng, S. (2014). Boosting recessions. *Canadian Journal of Economics*, *47*(1), 1–34. doi: doi.org/10.1111/caje.12070

Ng, S., & Wright, J. H. (2013). *Facts and challenges from the great recession for forecasting and macroeconomic modeling* (Working Paper No. 19469). National Bureau of Economic Research. doi: 10.3386/w19469

Pierce, J., & Wisniewski, E. (2018). Some characteristics of the decline in manufacturing capacity utilization. FEDS Notes. Washington: Board of Governors of the Federal Reserve System. Retrieved from https://doi.org/10.17016/2380-7172.2162

Raffinot, T., & Benoit, S. (2018). Investing through economic cycles with ensemble machine learning algorithms. Retrieved from https://ssrn.com/abstract=2785583

Ridgeway, G. (2019). Generalized boosted models: A guide to the gbm package. Retrieved from http://ftp5.gwdg.de/pub/misc/cran/web/packages/gbm/vignettes/gbm.pdf

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, *5*(2), 197–227. doi: 10.1007/BF00116037

Song, Y.-Y., & Ying, L. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, *27*(2), 130–135. doi: 10.11919/j.issn.1002-0829.215044

University of Cinncinati. (2018). Gradient boosting machines. Retrieved 02.04.2019, from http://uc-r.github.io/gbm_regression

Wright, J. H. (2006). *The yield curve and predicting recessions* (Finance and Economics Discussion Series No. 2006-07). Board of Governors of the Federal Reserve System (US). Retrieved from https://ideas.repec.org/p/fip/fedgfe/2006-07.html

# Appendices

# A Transformations and Definitions of the Variables

The column with transformation denotes the following data transformation for a series x:
(1) no transformation
(2) $\Delta x_t = x_t - x_{t-1}$
(3) $\Delta^2 x_t = x_t - 2 * x_{t-1} + x_{t-2}$
(4) $log(x_t)$
(5) $\Delta log(x_t) = log(x_t) - log(x_{t-1})$
(6) $\Delta^2 log(x_t) = log(x_t) - 2 * log(x_{t-1}) + log(x_{t-2})$
(7) $\Delta(x_t/x_{t-1} - 1.0) = (x_t/x_{t-1} - 1.0) - (x_{t-1}/x_{t-2} - 1.0)$

| Variable | Explanation | Transformation |
|---|---|---|
| RPI | Real personal income | 5 |
| W875RX1 | Real personal income excluding transfer receipts | 5 |
| DPCERA3M086SBEA | Real personal consumption expenditures | 5 |
| CMRMTSPLx | Real Manufacturing and Trade Industries Sales | 5 |

| | | |
|---|---|---|
| RETAILx | Retail and Food Services Sales | 5 |
| INDPRO | IP Index | 5 |
| IPFPNSS | IP: Final Products and Non-industrial Supplies | 5 |
| IPFINAL | IP: Final Products (Market Group) | 5 |
| IPCONGD | IP: Consumer Goods | 5 |
| IPDCONGD | IP: Durable Consumer Goods | 5 |
| IPNCONGD | IP: Nondurable Consumer Goods | 5 |
| IPBUSEQ | IP: Business Equipment | 5 |
| IPMAT | IP: Materials | 5 |
| IPDMAT | IP: Durable Materials | 5 |
| IPNMAT | IP: Nondurable Materials | 5 |
| IPMANSICS | IP: Manufacturing (SIC) | 5 |
| IPB51222S | IP: Residential Utilities | 5 |
| IPFUELS | IP: Fuels | 5 |

| | | |
|---|---|---|
| CUMFNS | Capacity Utilization: Manufacturing | 2 |
| HWI | Help-Wanted Index for United States | 2 |
| HWIURATIO | Ratio of Help Wanted/Number of Unemployed | 2 |
| CLF16OV | Civilian Labor Force | 5 |
| CE16OV | Civilian Employment | 5 |
| UNRATE | Civilian Unemployment Rate | 2 |
| UEMPMEAN | Average Duration of Unemployment (Weeks) | 2 |
| UEMPLT5 | Civilians Unemployed - Less Than 5 Weeks | 5 |
| UEMP5TO14 | Civilians Unemployed for 5-14 Weeks | 5 |
| UEMP15OV | Civilians Unemployed - 15 Weeks & Over | 5 |
| UEMP15T26 | Civilians Unemployed for 15-26 Weeks | 5 |

| | | |
|---|---|---|
| UEMP27OV | Civilians Unemployed for 27 Weeks and Over | 5 |
| CLAIMSx | Initial Claims | 5 |
| PAYEMS | All Employees: Total nonfarm | 5 |
| USGOOD | All Employees: Goods-Producing Industries | 5 |
| CES1021000001 | All Employees: Mining and Logging: Mining | 5 |
| USCONS | All Employees: Construction | 5 |
| MANEMP | All Employees: Manufacturing | 5 |
| DMANEMP | All Employees: Durable goods | 5 |
| NDMANEMP | All Employees: Nondurable goods | 5 |
| SRVPRD | All Employees: Service-Providing Industries | 5 |
| USTPU | All Employees: Trade, Transportation & Utilities | 5 |

| | | |
|---|---|---|
| USWTRADE | All Employees: Wholesale Trade | 5 |
| USTRADE | All Employees: Retail Trade | 5 |
| USFIRE | All Employees: Financial Activities | 5 |
| USGOVT | All Employees: Government | 5 |
| CES0600000007 | Average Weekly Hours: Goods-Producing | 1 |
| AWOTMAN | Average Weekly Overtime Hours: Manufacturing | 2 |
| AWHMAN | Average Weekly Hours: Manufacturing | 1 |
| Average HOUST | Housing Starts: Total New Privately Owned | 4 |
| HOUSTNE | Housing Starts, Northeast | 4 |
| HOUSTMW | Housing Starts, Midwest | 4 |
| HOUSTS | Housing Starts, South | 4 |
| HOUSTW | Housing Starts, West | 4 |

| | | |
|---|---|---|
| PERMIT | New Private Housing Permits (SAAR) | 4 |
| PERMITNE | New Private Housing Permits, Northeast (SAAR) | 4 |
| PERMITMW | New Private Housing Permits, Midwest (SAAR) | 4 |
| PERMITS | New Private Housing Permits, South (SAAR) | 4 |
| PERMITW | New Private Housing Permits, West (SAAR) | 4 |
| ACOGNO | New Orders for Consumer Goods | 5 |
| AMDMNOx | New Orders for Durable Goods | 5 |
| ANDENOx | New Orders for Nondefense Capital Goods | 5 |
| AMDMUOx | Unfilled Orders for Durable Goods | 5 |
| BUSINVx | Total Business Inventories | 5 |
| ISRATIOx | Total Business: Inventories to Sales Ratio | 2 |

| | | |
|---|---|---|
| M1SL | M1 Money Stock | 6 |
| M2SL | M2 Money Stock | 6 |
| M2REAL | Real M2 Money Stock | 5 |
| AMBSL | St. Louis Adjusted Monetary Base | 6 |
| TOTRESNS | Total Reserves of Depository Institutions | 6 |
| NONBORRES | Reserves Of Depository Institutions | 7 |
| BUSLOANS | Commercial and Industrial Loans | 6 |
| REALLN | Real Estate Loans at All Commercial Banks | 6 |
| NONREVSL | Total Nonrevolving Credit | 6 |
| CONSPI | Nonrevolving consumer credit to Personal Income | 2 |
| S&P 500 | S&P's Common Stock Price Index: Composite | 5 |

| | | |
|---|---|---|
| S&P: indust | S&P's Common Stock Price Index: Industrials | 5 |
| S&P div yield | S&P's Composite Common Stock: Dividend Yield | 2 |
| S&P PE ratio | S&P's Composite Common Stock: Price-Earnings Ratio | 5 |
| FEDFUNDS | Effective federal funds rate | 2 |
| CP3Mx | 3-Month AA Financial Commercial Paper Rate | 2 |
| TB3MS | 3-Month Treasury Bill | 2 |
| TB6MS | 6-Month Treasury Bill | 2 |
| GS1 | 1-Year Treasury Rate | 2 |
| GS5 | 5-Year Treasury Rate | 2 |
| GS10 | 10-Year Treasury Rate | 2 |
| AAA | Moodys Seasoned Aaa Corporate Bond Yield | 2 |

| | | |
|---|---|---|
| BAA | Moodys Seasoned Baa Corporate Bond Yield | 2 |
| COMPAPFFx | 3-Month Commercial Paper Minus FEDFUNDS | 1 |
| TB3SMFFM | 3-Month Treasury C Minus FEDFUNDS | 1 |
| TB6SMFFM | 6-Month Treasury C Minus FEDFUNDS | 1 |
| T1YFFM | 1-Year Treasury C Minus FEDFUNDS | 1 |
| T5YFFM | 5-Year Treasury C Minus FEDFUNDS | 1 |
| T10YFFM | 10-Year Treasury C Minus FEDFUNDS | 1 |
| AAAFFM | Moodys Aaa Corporate Bond Minus FEDFUNDS | 1 |
| BAAFFM | Moodys Baa Corporate Bond Minus FEDFUNDS | 1 |
| TWEXMMTH | Trade Weighted U.S. Dollar Index: Major Currencies | 5 |

| | | |
|---|---|---|
| EXSZUSx | Switzerland / U.S. Foreign Exchange Rate | 5 |
| EXJPUSx | Japan / U.S. Foreign Exchange Rate | 5 |
| EXUSUKx | U.S. / U.K. Foreign Exchange Rate | 5 |
| EXCAUSx | Canada / U.S. Foreign Exchange Rate | 5 |
| WPSFD49207 | Producer Price Index by Commodity for Finished Goods (Index 1982=100) | 6 |
| WPSFD49502 | Producer Price Index by Commodity for Finished Consumer Goods (Index 1982=100) | 6 |
| WPSID61 | Producer Price Index by Commodity Intermediate Materials: Supplies & Components (Index 1982=100) | 6 |

| | | |
|---|---|---|
| WPSID62 | Producer Price Index: Crude Materials for Further Processing (Index 1982=100) | 6 |
| OILPRICEx | Crude Oil, spliced WTI and Cushing | 6 |
| PPICMM | PPI: Metals and metal products | 6 |
| CPIAUCSL | CPI: All Items | 6 |
| CPIAPPSL | CPI: Apparel | 6 |
| CPITRNSL | CPI: Transportation | 6 |
| CPIMEDSL | CPI: Medical Care | 6 |
| CUSR0000SAC | CPI: Commodities | 6 |
| CUSR0000SAD | CPI: Durables | 6 |
| CUSR0000SAS | CPI: Services | 6 |
| CPIULFSL | CPI: All Items Less Food | 6 |
| CUSR0000SA0L2 | CPI: All items less shelter | 6 |
| CUSR0000SA0L5 | CPI: All items less medical care | 6 |
| PCEPI | Personal Cons. Expend.: Chain Index | 6 |

| | | |
|---|---|---|
| DDURRG3M086SBEA | Personal Cons. Exp: Durable goods | 6 |
| DNDGRG3M086SBEA | Personal Cons. Exp: Non-durable goods | 6 |
| DSERRG3M086SBEA | Personal Cons. Exp: Services | 6 |
| CES0600000008 | Average Hourly Earnings: Goods-Producing | 6 |
| CES2000000008 | Average Hourly Earnings: Construction | 6 |
| CES3000000008 | Average Hourly Earnings: Manufacturing | 6 |
| UMCSENTx | Consumer Sentiment Index | 2 |
| MZMSL | MZM Money Stock | 6 |
| DTCOLNVHFNM | Consumer Motor Vehicle Loans Outstanding | 6 |
| DTCTHFNM | Total Consumer Loans and Leases Outstanding | 6 |
| INVEST | Securities in Bank Credit at All Commercial Banks | 6 |

| | | |
|---|---|---|
| VXOCLSx | CBOE S&P 100 Volatility Index: VXO | 1 |

(McCracken, n.d.)

# B Code Examples

I provide some examples of the code used in the analysis. It is important to note that there are some randomness in the models. Running the model twice will then not necessarily result in exactly the same results, but they will similar overall results. This is boosting with the gbm-package, where the code is inspired by University of Cinncinati (2018):

```
# IN-SAMPLE ANALYSIS
# new_data_frame_d4 consists of data with for h+s
# Estimation for the boosting analysis
gbm_final_train_30_bern <- gbm(Y ~., data =
new_data_frame_d4[1:719,2:513],
distribution = "bernoulli",
n.trees = 1759, interaction.depth = 5,
shrinkage = 0.001, n.minobsinnode = 5,
bag.fraction = 0.5, train.fraction = 0.3,
n.cores = NULL, verbose = FALSE)
gbmTrainPredictions_train_30_bern =
predict(object = gbm_final_train_30_bern,
newdata = new_data_frame_d4[1:719, 2:513],
n.trees=1759, type="response")

# OUT-OF-SAMPLE ANALYSIS
# Tuning the parameters out-of-sample
# Defining the different choices of the
# tuning parameters
grid_search = expand.grid(
  shrinkage = c(.001, .005, .01),
  interaction.depth = c(3,4,5),
  n.minobsinnode = c(5, 10),
  optimal_trees = 0,
  min_RMSE = 0)

# Running the grid search
# new_data_frame consists of data
# lagged according to s
grid_search_new=NULL
for(j in 0:492){
  for(i in 1:nrow(grid_search)) {
```

```
    # reproducibility
    set.seed(1222)
    Y_new = Y[19:226+j]
    # train model
    gbm.tune <- gbm(
     formula = Y_new ~ .,
     distribution = "bernoulli",
     # Data frame consists of alle the
     # variables with lags
     data = new_data_frame[13:220+j,2:513],
     n.trees = 3000,
     interaction.depth =
     grid_search$interaction.depth[i],
     shrinkage = grid_search$shrinkage[i],
     n.minobsinnode =
       grid_search$n.minobsinnode[i],
     # Adds some randomness to the model:
     bag.fraction = 0.5,
     # Using a training fraction of 60 %:
     train.fraction = 0.6
     )

   # Optimal number of trees - minimizes
   # the validation error
   grid_search$optimal_trees[i] =
     which.min(gbm.tune$valid.error)
   grid_search$min_RMSE[i] =
     sqrt(min(gbm.tune$valid.error))
   }
   grid_combinations = grid_search %>%
     dplyr::arrange(min_RMSE) %>%
     # only using the optimal one
     head(1)
   frame_grid = data.frame(new_data_frame[227+j,1],
                           grid_combinations)
   grid_search_new = rbind(grid_new,
                                frame_grid)
}

# Making the actual out-of-sample analysis
# n_trees_bern, n_minobsinnode_bern,
# interaction_depth_bern, shrink_bern
# are vectors that comes from the
# grid search
full_data_proj = NULL
```

```
data_frame_check =
  variable.names(new_data_frame[2:513])
for_auc = NULL
for (i in 0:492){
  Y_new = Y[19:226+i]
  number_of_trees = n_trees_bern[1+i]
  number_minobsinnode = n_minobsinnode_bern[1+i]
  depth = interaction_depth_bern[1+i]
  number_shrink = shrink_bern[1+i]
  # This is the training data, train.fraction=100%
  gbm_final_train <- gbm(Y_new ~.,
  data = new_data_frame[13:220+i,2:513],
  distribution = "bernoulli",
  n.trees = number_of_trees,
  interaction.depth = depth,
  shrinkage = number_shrink,
  n.minobsinnode = number_minobsinnode,
  bag.fraction = 0.5,
  train.fraction = 1,
  n.cores = NULL, verbose = FALSE)
  rel_inf = summary(gbm_final_train,
  order = FALSE, plotit = FALSE)
  data_frame_check = data.frame(data_frame_check,
                                rel_inf[,2])
  # Predicting based on the analysis already
  # done but forward in time(out-of-sample)
  gbmTrainPredictions_train = predict(object =
  gbm_final_train,
  newdata = new_data_frame[221+i, 2:513],
  n.trees=number_of_trees, type="response")
  full_data_proj = rbind(full_data_proj,
        data_frame(new_data_frame[227+i,1],
        gbmTrainPredictions_train))
  # To bind the final predictions together
  for_auc = rbind(for_auc, gbmTrainPredictions_train)
}
```

BMA with the BAS-package:

```
# IN-SAMPLE ANALYSIS
# In-sample estimation and predictions for s=0
Y_new1 = Y[9:719]
# This dataframe consists of data
# lagged according to s+h,
# only using s=0
new_data_frame_lag6 =
```

```
  new_data_frame_d4[9:719, 2:129]

bas_recession_glm_lag6 = bas.glm(Y_new1~.,
family = binomial(link = "logit"),
data = new_data_frame_lag6,
betaprior = bic.prior(), modelprior = uniform(),
initprobs = "uniform",
method = "MCMC", MCMC.iterations = 100000)
prediction_bas_lag6 = predict(bas_recession_glm_lag6,
  newdata = new_data_frame_lag6, type="response")

# OUT-OF-SAMPLE ANALYSIS
# Out-of-sample predictions for s=0
full_data_proj_6 = NULL
for_auc = NULL
bas_reg = variable.names(bas_recession_glm_lag6)
for (i in 0:492){
  Y_new = Y[19:226+i]
  bas_rolling_glm_lag6 = bas.glm(Y_new~.,
  family = binomial(link = "logit"),
  data = new_data_frame[13:220+i,2:129],
  betaprior = bic.prior(),
  modelprior = uniform(), initprobs = "Uniform",
  method = "MCMC",
  MCMC.iterations = 1000)
  basPredictions_rolling_6 =
  predict(object = bas_rolling_glm_lag6,
  newdata = new_data_frame[221+i, 2:129],
  type="response")
  bas_reg = rbind(bas_reg,
        bas_rolling_glm_lag6$probne0)
  full_data_proj_6 = rbind(full_data_proj_6,
  data_frame(new_data_frame[227+i,1],
  basPredictions_rolling_6$fit))
  for_auc = rbind(for_auc,
        basPredictions_rolling_6$fit)
}
```