EC4308: Machine Learning and Economic Forecasting

Final Project Report
Predicting US Recessions

Devika Rastogi – A0176373X
Manan Mittal - A0176381Y
Maulik Jain - A0177164Y
Soumil Banerjee - A0176385R

INTRODUCTION

The 2008 financial crisis in the United States was the most severe economic disaster since the Great Depression of 1930 . It took huge taxpayer bailouts to shore up the financial industry . This crisis led to the Great Recession where the housing prices dropped more than the price fall during the Great depression and major consequences of the crisis were a major drop in international trade , rising unemployment and slumping commodity prices. In general, a recession is defined as a contraction in the business cycle when there is a substantial decline in economic growth for more than 2 consecutive quarters. A recession can be triggered by various events like the financial crisis of the United States in 2008 , an external trade shock, a supply shock like the oil crisis in the 1980's and a natural disaster like the Covid 19 pandemic in 2020.

A great deal of economic research attempts to answer the very important question , is it possible to forecast a recession accurately given the various leading and lagging indicators like the interest rates and stock prices. Majority of the earlier studies analyse the probability of recession in the coming quarters by using probit models in which the current and past values of financial data are considered as regressors. Recent studies like the Chauvet and Potter (2005) have even tried to apply more specifications to the probit model by adding autocorrelated errors and multiple break points across business cycles.

Our paper draws inspiration from Kauppi and Saikkonen (2008) paper to predict the probability of recessions in the United States with particular focus on the 2008 Financial Crisis. We will attempt to predict the 2008 crisis using the data from previous years employing different techniques like the Dynamic Probit model, Autoregressive Dynamic Probit model and the Random Forest Classifier. The mentioned paper makes extensions to the previously established ways of modelling and forecasting binary time series data.

We have conducted 1, 3, 6 and 12 step forecasts using these three methods mentioned above. Through our analysis we will compare the Autoregressive Dynamic Probit model , the Dynamic probit model and the random forest classifier and see which methods are better predictors of recession under different forecast horizons and addition of which variables make the models more efficient to predict the probability of recessions .

From our research , the Autoregressive Dynamic Probit model from the Kauppi-Saikkonen paper performs the best for the short term (1 and 3-step ahead forecast). However, when we lengthen our forecast horizon (6 and 12-step ahead forecast) the Random Forest forecaster proves to be more accurate in its prediction of the probability of recession

LITERATURE REVIEW

A great deal of recent empirical research has used financial and economic variables such as the interest rate spread, GDP values, and unemployment rates to provide useful analysis about the possibility of a recession in the coming months/quarters. While writing this paper we are attempting to use machine Learning methods and check whether they can predict recessions better than simple logit/probit modeling. The main paper we have used to support our analysis is the Kauppi and Saikkonen (2008) paper but to understand binary modeling better we have also tried to understand methods used in Serena Ng's paper on boosting(2014) and Davig Hall's (2019) Recession forecasting using Bayesian classification.

In the Kauppi and Saikkonen (2008) paper the author compares different probit models and different forecasting methods in terms of in-sample and out of sample performance by using the interest rate spreads as the only external predictor. According to the paper, the dynamic probit models perform better than the static model in terms of using both in sample and out of sample predictions. The dynamic models with lagged values of binary response variables perform better than probability models where dynamics enter only through lagged probability models. The paper also shows the importance of lag orders to forecast accurately. The paper shows that experimenting with various lags and using statistical modeling to choose the appropriate lags is better than choosing lags to match forecast horizons directly.

The Serena Ng Boosting (2014) paper explores the usefulness and effectiveness of boosting as a recession prediction tool. In this paper to facilitate boosting, the recession probability estimates are based on using logit models. The difference in approach by the author is that she uses one predictor in one model instead of clubbing many predictors in a single logit model. This is done to identify the key issues the author has defined – the most important one being, understanding which variables at which lags can provide the most amount of information about a recession. The author also emphasizes the fact that the relative importance of certain variables differs among different time horizons.

Davig Hall's (2019) Recession forecasting using Bayesian classification uses a Naïve Bayes framework as a recession prediction tool. The paper uses Bayes theorem in a manner to incorporate varied data sets with multiple lag structures and it attempts to capture the persistence of various business cycle phases. The paper uses a set of data to predict the recession at some point in the future using the binary framework What the paper does differently is that it uses NBER turning points as data to forecast past recessions, rather than as something to be inferred. The paper uses the approach mentioned in Diebold and Rudebusch (1989) by using a composite of indices using a Bayesian framework. The authors use methods to find probabilistic forecasts of future business cycle turning points and study how much lead time an index can provide to signal an upcoming peak or trough.

DATA USED

In predicting recessions, we source the US recession indicator (Binary Variable; 1 = recession, 0 = no recession), as provided by the NBER (US Recession). As for the variables used in forecasting, we take the dataset from FRED-MD. We stationarize all these variables through the methods mentioned in the FRED-MD Appendix 1. Once we have this, we create a new variable called the interest rate spread, which is defined as the 10-year Treasury Bond Rate minus the 3-month Treasury Bill Rate.

At this step, we perform some pre-processing to make the data more accessible. For the main models, we eliminate some time periods. We exclude the first period (January 1959) as it has a few NA values due to the stationary transformations. We also exclude all time periods of 2020 as this period is in the near past and might have some misreported values. As the final processing step, we remove all predictors that have NA values. These are mainly predictors with data only in the last few years and missing values before the 1990s. We lose about 10 out of 130 predictors at this step, which is a fair compromise.

METHODOLOGY

To forecast recession, we are using three key methods- Dynamic Probit, Autoregressive Dynamic Probit and the Random Forest Classifier. The first two approaches are drawn from the Kauppi-Saikkonen (2008) paper and are thus used as our benchmark for comparison. The Random Forest is a new method we are trying to implement and see how it stacks up against tried and tested methods.

1. Dynamic Probit and Autoregressive Dynamic Probit (Benchmarks)

We draw this technique from the Kauppi-Saikkonen (2008) paper. The mentioned paper makes extensions to the previously established ways of modelling and forecasting binary time series data. Under the dynamic probit method, we use the lagged value of interest rate spread as a predictor. The appropriate lagged value shifts according to the forecast interval. The estimation equation is shown below:

$$p_{t-1}(y = 1) = \phi(\omega + \beta x_{t-k})$$

Secondly, under the autoregressive dynamic probit method, the dynamic probit model is enhanced by adding a lagged value of the recession indicator as an additional regressor in the model. This method takes into assumption conditional probability, as we have shown in the equation below:

$$p_{t-1}(y = 1) = \phi(\omega + \delta y_{t-L} + \beta x_{t-k})$$

We use these two approaches, Dynamic Probit and Autoregressive Dynamic Probit, as our benchmark for the random forest, described as under.

2. Random Forest Classifier

Random forest is regarded as one of the most significant techniques of ensemble learning. This arises from the fact that it is able to deal with large data sets with higher dimensionality. In addition to that, its strength also lies in being able to predict an out of bag MSE (OOB MSE) since it uses only one third of the bootstrap sampling data as training dataset. It maintains accuracy even with a large proportion of missing values in the dataset. Lastly, it performs well for classification. Keeping this in mind, we apply random forest to our analysis and compare the results with the output of methods employed in the paper by Kauppi and Saikkonen. The intuition behind using Random Forest may seem confusing at first since the purpose of random forest is to de-correlate the data, however, this seems unlikely when working with time-series data since that is defined by serial dependence.

Nonetheless, we observe that with the transformation for stationarity we can effectively use the random forest for time series analysis. In addition to that, it is often observed that random forest time series prediction results in enhanced predictive ability.

We use all the variables in the FRED-MD dataset to perform random forest and find out which ones are good predictors of Recession. We also add the first 4 principal components to the predictor set. Similar to the methods used in Kauppi and Saikkonen, we use only one lag value to forecast ahead.

Our comparisons are done by 2 methods:

1) We make predictions of the probability of a recession $P(Y_t = 1)$ and then compare the methods based on RMSE. This shows us how close the prediction is to the actual value of Recession.
2) We make the actual binary predictions with a threshold value of 0.5 and compare the methods based on misclassification percentage. This gives us an accurate estimation of the correctness of the decision.

The time period we are particularly interested in is the 2008-09 Financial Crisis. The NBER officially announced a recession in the economy of the United States of America to begin from the last quarter of 2007. Thus, our training set is until the period of 2008, and the test set is the period of recession for 2007 up to 2019, ignoring the recent values of 2020. As best-practice for time-series forecasting we will be using the rolling window estimation and predicting periods ahead iteratively. This will ensure that we choose a time period which gives us a good enough understanding of the model performance.

As a matter of interest and testing our model further, we will also be looking at 2 different settings:

1) Predicting recessions without using the lagged value of Recession. This will try to model a situation where we are unsure if we are in a recession or not. We analyze this for 1-period and 3-period forecasts.
2) Looking into the Covid-19 pandemic period of 2020. Our training set would be till December 2019 and we will test the results for the available data of 2020. We wish to see if this was predictable or not.

ANALYSIS OF RESULTS

After running the models, we analyze our results for the three methods used, the dynamic probit, the autoregressive dynamic probit and random forest.

First, we will analyze the performance of these three decisions from a series of plots graphing the 1, 3, 6 and 12-step forecast and compare it to the actual outcome (Yt = 1 i.e. recession) with interest in the 2008 financial crisis.

First, we will look at the 1-step forecast -

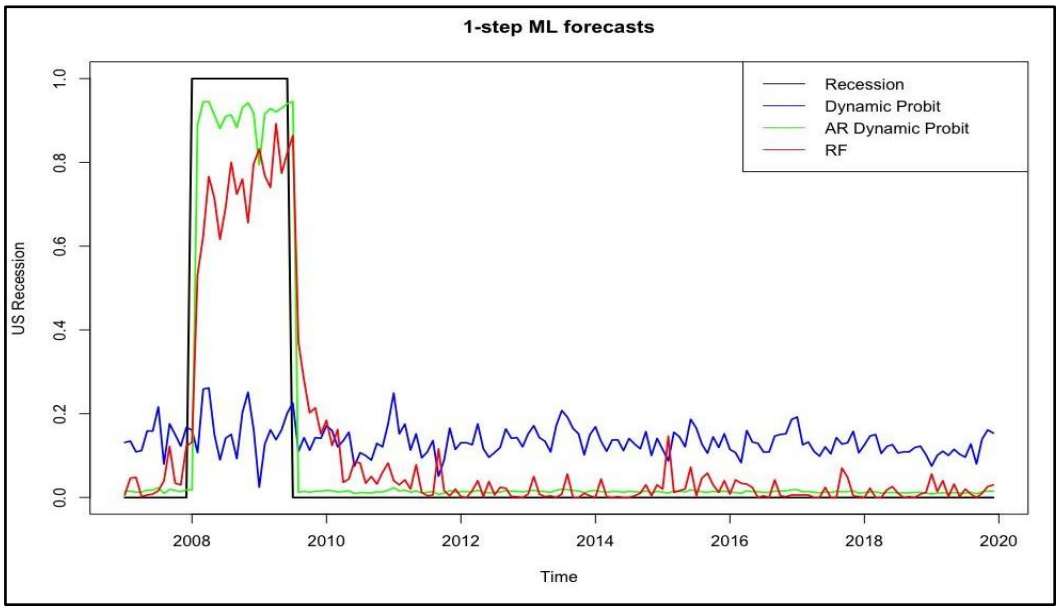


Figure 1

We see from Figure 1 that compared to the actual recession (shown by the black line), the autoregressive dynamic probit and random forest forecast quite accurately the probability of a recession. Specifically, we see that the AR dynamic probit performs slightly better than the random forest itself. The dynamic probit does not perform very well which is consistent with the results postulated by Kauppi-Saikkonen. They highlight the superiority of the AR dynamic probit to the dynamic probit thus showing the importance of auto-regressing on lagged values of Y.

This is showing a stronger result for the prediction of the 2008 financial crisis because it is dependent on only one lagged value, thus, we are almost in the recession stage. The usual leading economic predictors may have already given some indication of the financial crisis, this close to the actual recession.

Moving onto the 3-step forecast -

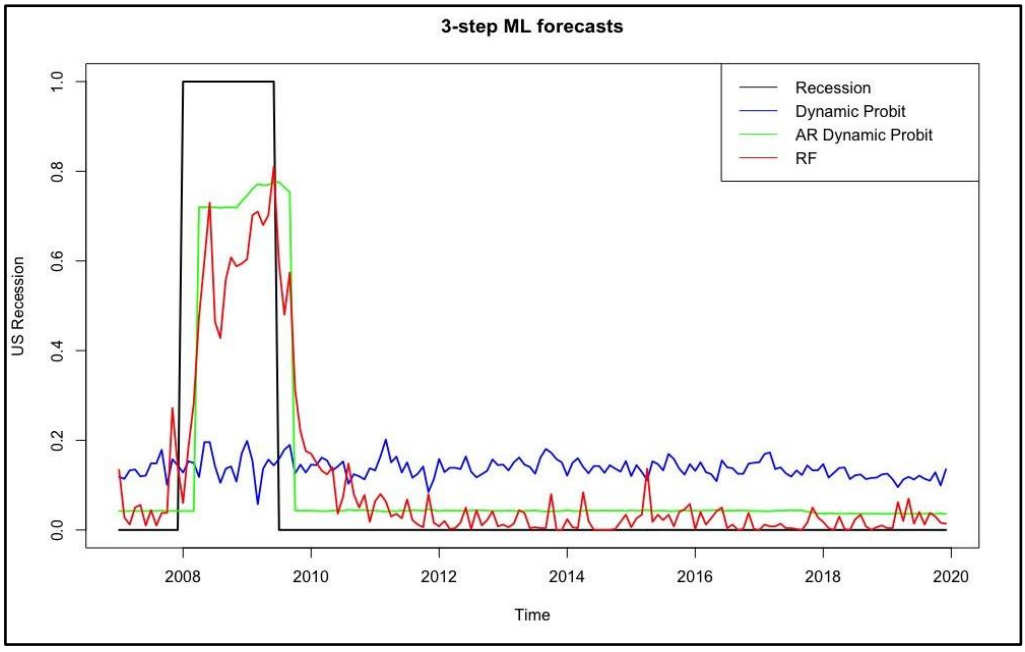


Figure 2

We see that while overall the results are consistent as in the case of 1-step ahead forecast, the accuracy with which the AR Dynamic Probit and random forest predict the probability of recession has decreased. However, it mostly lies above the probability threshold of 0.5. Here as well, like the previous case, since the duration is shorter to predict for the 2008 financial crisis, it may be the case that there were early warning signs which can be caught. This will allow for an easier prediction possibility.

Next, looking at our 6-step forecast -

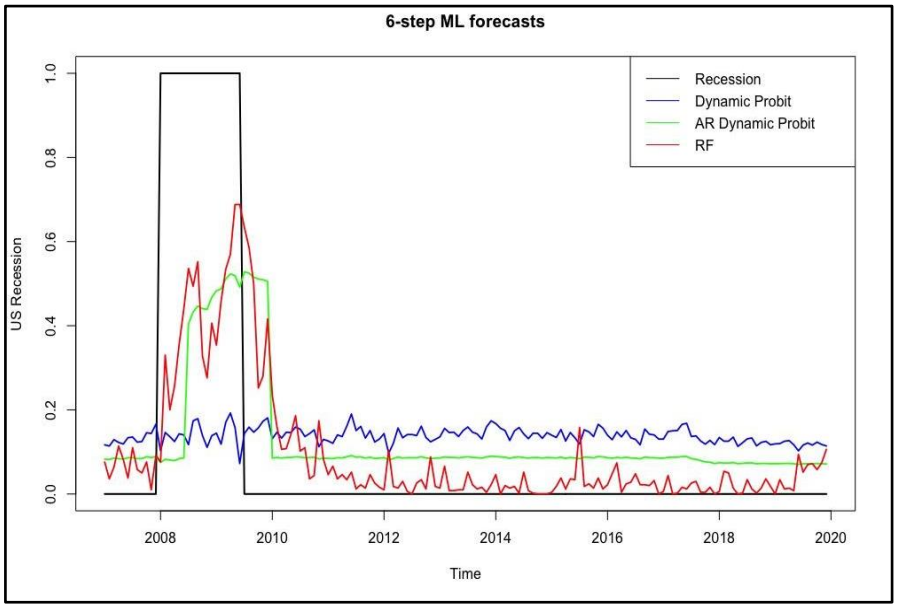


Figure 3

Keeping in line with the consistency of results and decrease in accuracy with an increase in time horizon, we see that the AR Dynamic Probit and Random Forest still predict quite well. However, something interesting to note here is that the random forest now appears to give a slightly better forecast than the AR dynamic probit.

Particularly for the period of 2008 Financial Crisis, we see a lagged effect under the AR dynamic Probit, as the recession indicator would be picked up with a lag. For the Random Forest, a little probability is picked up. These may be due to the important variables under the Random Forest that we are discussing in the next section. The analysis can be said for the 12-step forecast below as well.

Finally, looking at our 12-step forecast below-

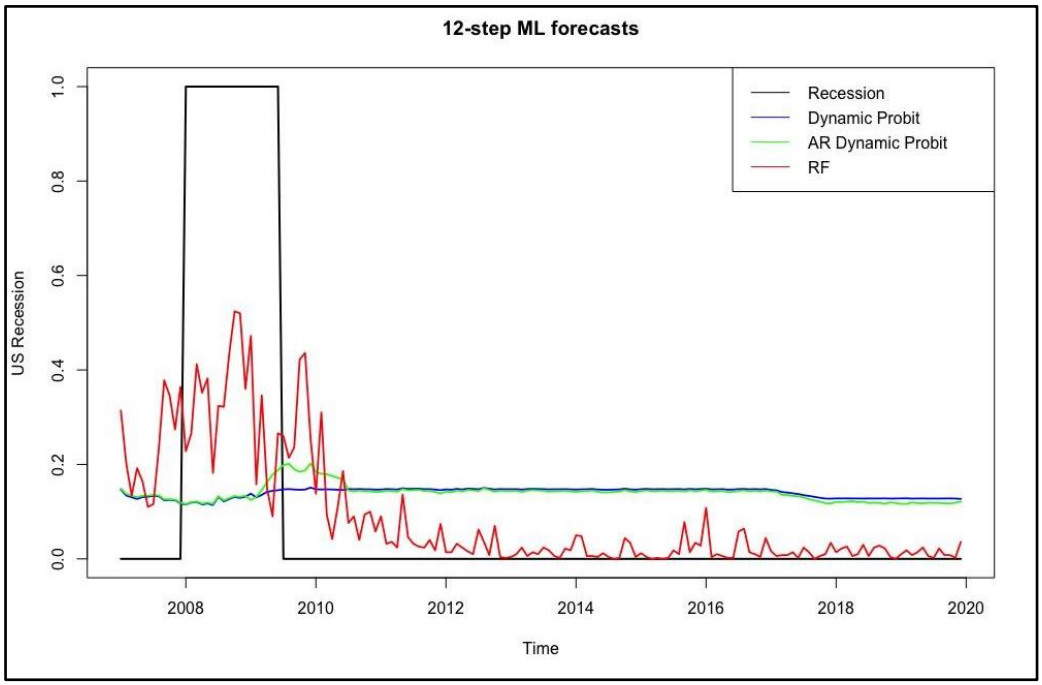


Figure 4

The results here are quite surprising. We see that for the 12-step prediction, the AR dynamic probit and dynamic probit give a similar outcome in terms of predicting the probability of recession which is not very accurate. However, what is surprising is that the random forest is still able to identify the probability of a recession and does slightly better (though still not very accurate) in terms of the prediction than the AR dynamic probit and the dynamic probit.

Thus, some of the key observations that arise from this analysis is as follows:

1. The dynamic probit is not very accurate in predicting the probability of recession irrespective of the timeframe (1, 3, 6 or 12-step ahead forecast) and hence it should not be used.

2. The AR Dynamic Probit and Random Forest are quite consistent in predicting the probability of recession. However, as the time-horizon of prediction increases, the Random Forest starts outperforming the AR Dynamic Probit. This shows the strength of random forest in terms of being able to handle large datasets and make accurate predictions for longer horizon time frames.
3. The accuracy of the three methods decreases with an increase in the time-horizon of prediction. Thus, the methods perform much better for the 1-step ahead forecast as compared to the 12- step ahead forecast. This is expected and is attributed to the fact that there is lesser uncertainty when it comes to shorter period predictions. It is natural to assume that uncertainty increases as the time frame of our prediction increases. This also shows that the forecast variance increases.
4. The AR Dynamic Profit is much more accurate than the Dynamic probit in terms of a comparison of our benchmarks. This can also be seen from the fact that when we check for variable importance in Random forest, lagged Y is seen as one of the most important predictors (further elaborated in a later section).

As we have established that Autoregressive Dynamic Probit is a better benchmark, we compare random forest with AR dynamic probit in terms of prediction loss.

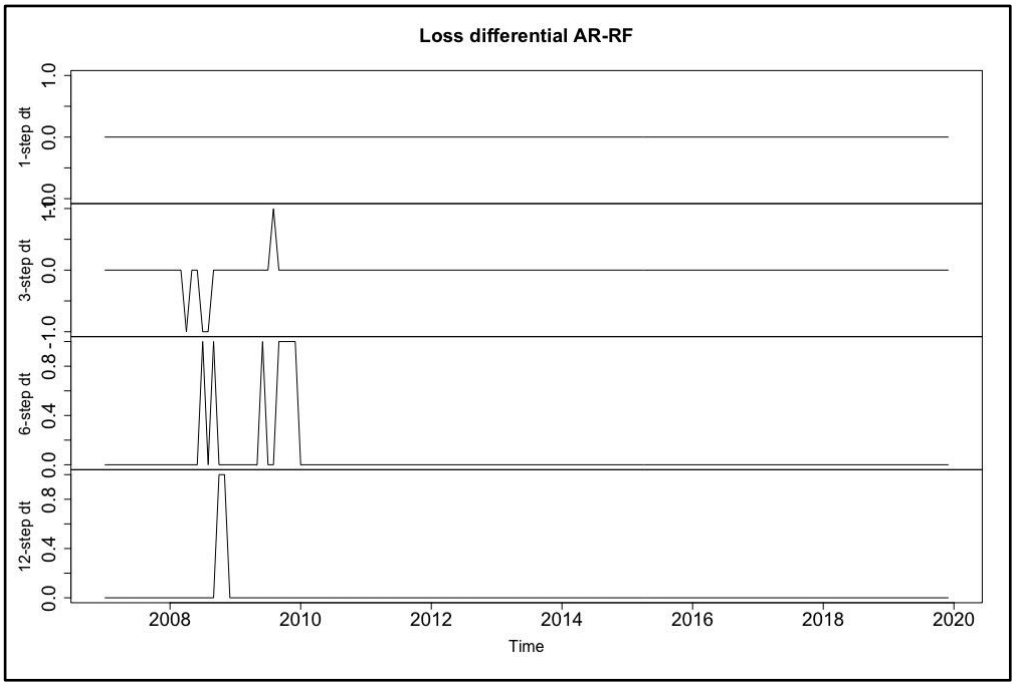


Figure 5

We analyze the loss differentials of AR-RF for the 1, 3, 6 and 12 step forecast differentials. As expected, the differentials occurred during the 2008 Financial Crisis. We see that for the 1 step forecast, the differential value is nil thus showing the consistency of both methods. For the 3-step ahead forecast, it is interesting to

note that initially it shows that the AR Dynamic Probit performs better, however, later favors the Random Forest. AR Dynamic Probit is initially doing very well and this fact could also be attributed to its heavy dependence on the lagged values. For the 6-step forecast we really see how Random Forest steps up over the AR Dynamic Probit and manages to predict much better. In the 12-step forecast, random forest is able to predict one period of recession, but it is not good enough in a real setting to get the right predictions. Thus, both methods perform poorly in a 12-step forecast. The plots showing losses over the prediction probabilities is shown in the Appendix.

Moving onto the Root Mean-Squared Error (RMSE Values) and the Misclassification Percentage; we present our findings in the form of the following tables:

TABLE 1

RMSE	1-step	3-step	6-step	12-step
Dynamic Probit	0.314	0.319	0.320	0.325
Autoregressive Dynamic Probit	0.1143	0.1924	0.2650	0.3230
Random Forest	0.1473	0.1945	0.2305	0.2560

As we can see, looking at the results from the perspective of RMSE over probabilities yields the same analysis. As mentioned previously, Autoregressive Dynamic Probit and Random Forest perform better than Dynamic Probit, with the performance of Random Forest improving with time horizon.

TABLE 2

Misclassification Percentage	1-step	3-step	6-step	12-step
Autoregressive Dynamic Probit	1.28%	3.85%	13.5%	11.5%
Random Forest	1.28%	5.13%	8.97%	10.3%

Here we report misclassification percentage over the binary predictions with threshold value as 0.5 for both methods. As we can see, looking at the results from the perspective of Misclassification percentage yields the same analysis. Autoregressive dynamic probit performs better for the short-term, with the performance of

the Random Forest improving with time horizon. For the time series graphs of the binary predictions, we can refer to the appendix.

Variable Importance in Random Forest

We have analyzed the results of the random forest and seen which methods work better in which forecast intervals. A key feature of the random forest model is the variable importance. For each of the forecast intervals, we extracted the top 5 variables in terms of Mean Decrease in Accuracy if the variable is removed. The results with Gini impurity were similar so we exclude that part.

The top 5 variables for the forecast intervals were as such [the coding of the variables are explained in the appendix]:

TABLE 3

	Ranking	1	2	3	4	5
1 - period	Variable	USREC	USGOOD	PAYEMS	DMANEMP	MANEMP
	MDA	21.17	8.75	8.37	8.01	7.93
3-periods	Variable	USREC	T1YFFM	TB6SMFFM	TB3SMFFM	T10YFFM
	MDA	16.56	9.91	9.64	9.62	7.50
6-periods	Variable	T1YFFM	T5YFFM	T10YFFM	AAAFFM	TB6SMFFM
	MDA	13.92	10.5	10.50	10.42	10.22
12-periods	Variable	T10YFFM	T5YFFM	T1YFFM	Comp 2	AAAFFM
	MDA	13.81	13.03	11.51	10.83	10.45

Through this we can see that for the 1 and 3 period forecasts, the lagged value of Recession is a very important predictor. For the 1 period forecast we see that no. good producing industries and level of employment (other 3 variables) are very important predictors. One of the key consequences of recession is an increase in unemployment and a decrease in the GDP. Therefore, this decrease in employment is a result of recession and is caused because companies usually lay off workers to cut down on their costs.

In the 3-period forecasts, we see that along with lagged recession, the interest rates start playing an important role, especially the short-term interest rates (less than 1 year).

As our forecast horizon increases, the recession indicator is no longer a good indicator as it is so far ahead. In the 6-period horizon, it is the longer interest rates (> 1 year) which play an important role in prediction of the recession. We also see the entry of the AA bond yield minus the Fed Funds Rate.

One possible reason for this can be that interest rates play an important role in acting as a leading indicator of a recession. This can be widely observed through the yield curve. When we observe an inverted yield curve i.e. the short-term rates higher than the long-term rates, it is believed to be a signal for recession in the future.

Finally, as we try to forecast a year in advance, the 10-year treasury bond rate minus the FedFunds rate is the most important predictor. As the model in general does not predict very well for this horizon, we are still quite unsure whether this is a good enough signal for recession.

Contrary to proposition Kauppi and Saikkonen's paper we see that the interest rate spread (10 Year Treasury rate - 3 month Treasury Bill rate) is not a very important predictor and over the last few years, using random forest we are able to find more important predictors.

EXTENSIONS

1] What if we are unaware that we are in a recession?

It is often the case that economists and policy makers are unaware of a recession till about 6 months after it has already hit. Thus, we try to use random forest to see if we can predict recession in a 1-period and 3-period horizon, without using the lagged value of Recession as a predictor.

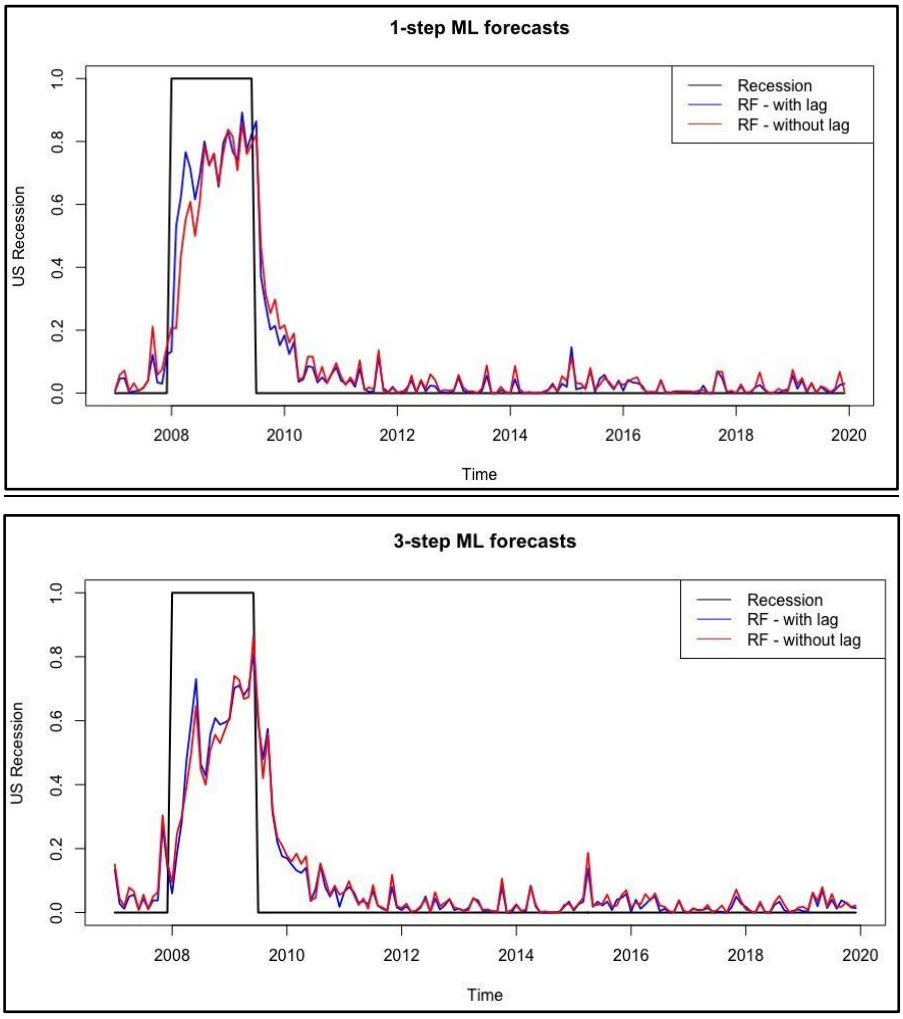


Figure 6

Surprisingly, this model works almost as well as the model with the lag value of recession and has very few errors. This seems to be unexpected as we postulated that the lagged values of recession would be the best signal in the 1 and 3 period forecasts. From the graph we can see that the forecast variance is slightly larger (peaks higher than before, troughs lower than before) and this is mostly due to the absence of the balancing act done by the lagged value of Recession. The RMSE is 0.1703 (compared to 0.1474 in the original model) for the 1-period forecast and 0.2003 (compared to 0.1945 in the original model) for the 3-period forecast.

The most important predictors are the same as the previous random forest model and it just takes the next 5 most important predictors. Thus, employment rates are important for the 1 period forecast while the interest rates are the most important predictors for the 3-period forecast.

2] Analyzing Predictions for COVID-19

We extend our original model to predict the recession in 2020. As we know, a recession was declared in March 2020. One would assume that this recession is a one off and there was no way to predict this as this recession was caused by a once in 100 years pandemic. We see the results of this model as follows:

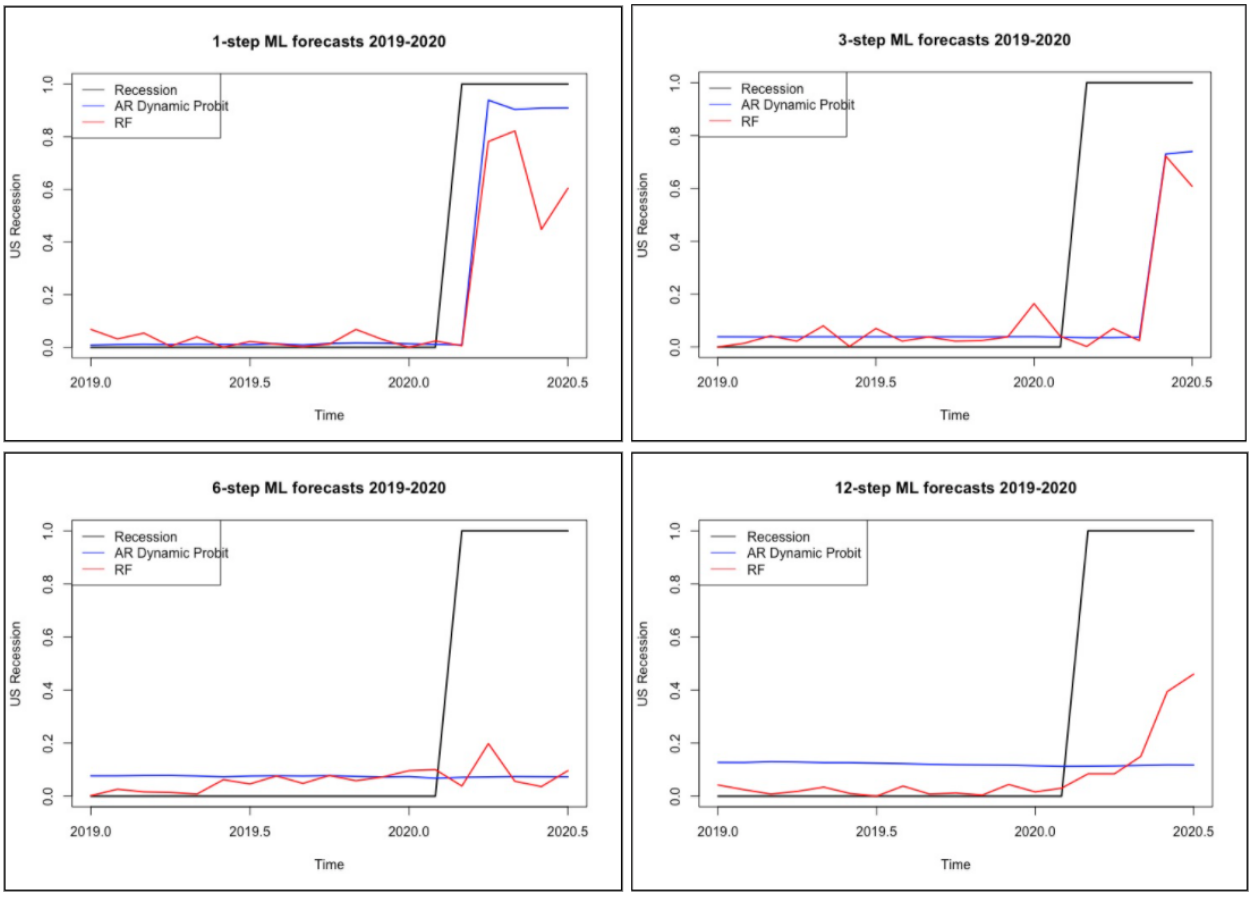


Figure 7

As expected, there is not enough evidence to predict a recession in March 2020. Almost all horizons are unable to predict it until March 2020 is used as a predictor for the future months. This is just due to the importance of the lag recession and no other variable moving significantly.
One interesting observation is the small spikes shown in all intervals which indicate back to a period of July-September 2019, which predicts a spike in probability of recession. This can be due to the several events taking place during that time. For one, the trade war was raging with both US and China raising tariffs on the other country's imports. In addition to that, the uncertainty was accentuated because of Brexit and finally

the increased geopolitical tensions also disrupted the energy prices. All of these may have signaled a potential recession.

Predicting recession due to pandemics is thus out of the ability of this model. It is possible that in the future, with pandemics being a part of the training time period, we would be able to predict a recession in such a setting.

CONCLUSION

In this paper, we have attempted to forecast the probability of a recession in the US Economy with specific focus on the 2008 Financial Crisis. For this we have used three primary methods - Dynamic Probit, Autoregressive Dynamic Probit (from Kauppi-Saikkonen (2008)) and Random Forest. We have used the dataset from FRED-MD and NBER. The key variables we have looked at are the US Recession Indicator and its lag as well as the interest rate spread (and consequent lags). For the Random Forest, we have used the lag of every variable from the FRED-MD dataset.

We conducted 1, 3, 6 and 12 step forecasts using the three methods. From our analysis, we concluded that the Autoregressive Dynamic Probit model from the Kauppi-Saikkonen paper is difficult to beat for the short term (1 and 3-step ahead forecast). However, when we expand our forecast horizon (6 and 12-step ahead forecast) the new method we are trying - Random Forest proves to be more accurate in its prediction of the probability of recession. This can be attributed to the inclusion of some other key variables in the analysis such as the different interest rate spreads (example T1YFFM, T5YFFM and T10YFFM).

Another notable observation is that the Dynamic Probit method is largely ineffective in predicting the US recession probability for any given forecast horizon. This points us to an important result, that the lag of the US Recession (Y_{t-1}) is a crucial indicator for prediction.

We also run these methods to observe the model performance for predicting the probability of the recession caused by the COVID-19 pandemic as an extension of our analysis. We observe from this that due to the absence of the recession lag indicator, the accuracy of the model is limited.

One of the key limitations we have identified is the over reliance on the US Recession lag. Further research and analysis can be conducted to find out dummies for the US Recession lag. This can be done through using variables outside of the FRED-MD database for example sentiment analysis. Another limitation identified is the delay in the announcement of the recession indicator by the NBER which prevents us from trusting the one-lagged value. We have attempted to tackle this problem of information lag by providing an extension to the paper which does not rely on the lagged values.

Through the course of this study, we have learnt the benefits of ensemble methods of prediction (random forest). Reading previously established literature and trying to work on their extensions can be helpful for understanding, solving and appreciating the problem.

BIBLIOGRAPHY

Davig, T., & Hall, A. S. (2019). Recession Forecasting using Bayesian Classification. International Journal of Forecasting, 35(3), 848-867. doi:10.1016/j.ijforecast.2018.08.005

Kauppi, H., & Saikkonen, P. (2008). Predicting U.S. recessions with dynamic binary response models. The Review of Economics and Statistics, 90(4), 777-791. doi:10.1162/rest.90.4.777

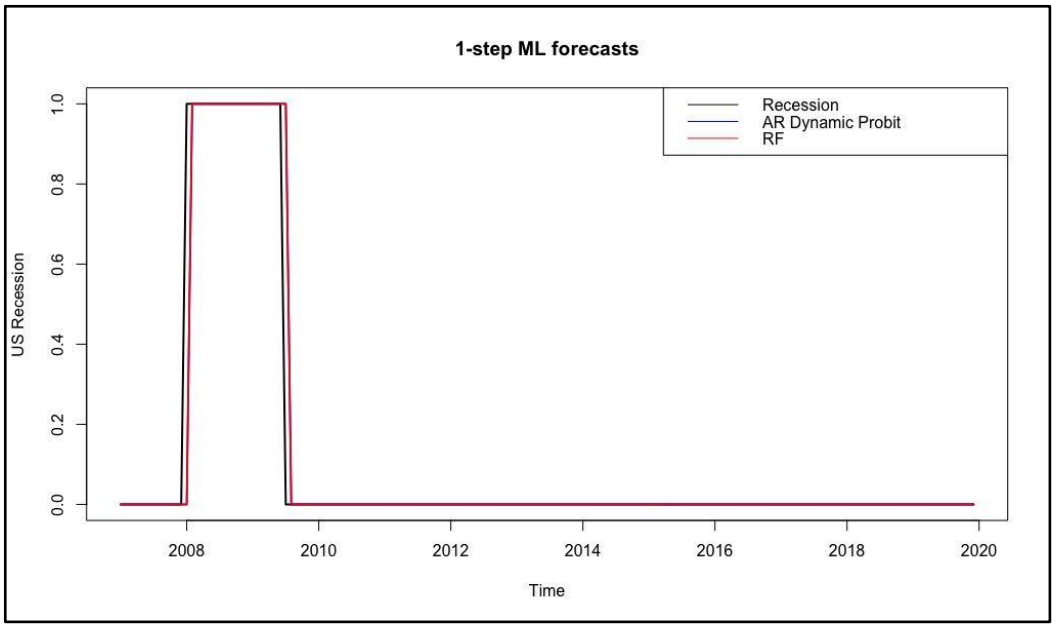
Ng, S. (2014). Viewpoint: Boosting recessions. The Canadian Journal of Economics, 47(1), 1-34. doi:10.1111/caje.12070

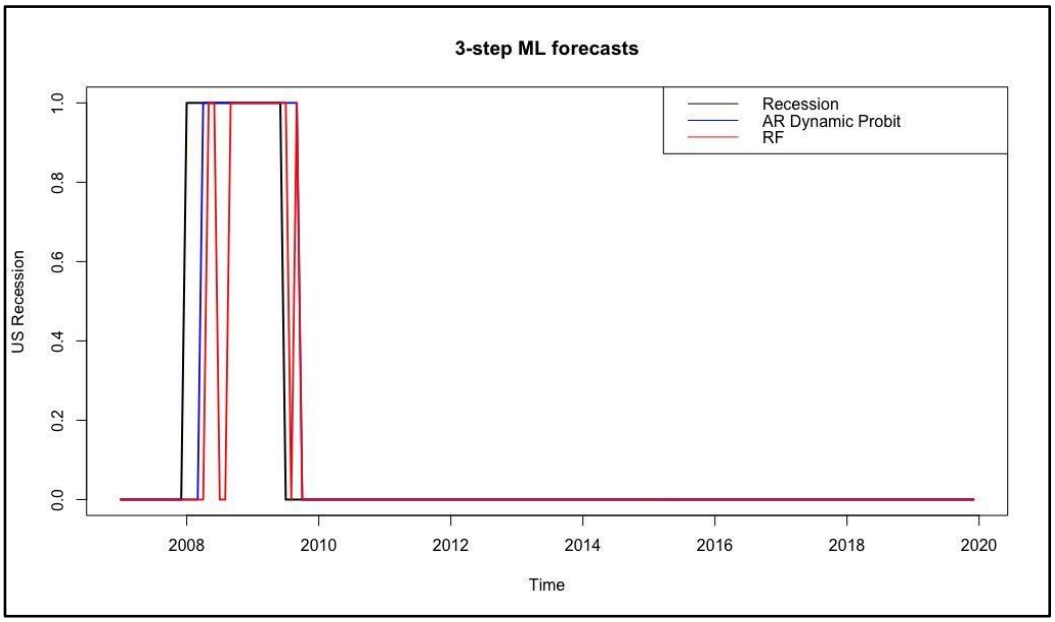
DATA SOURCES

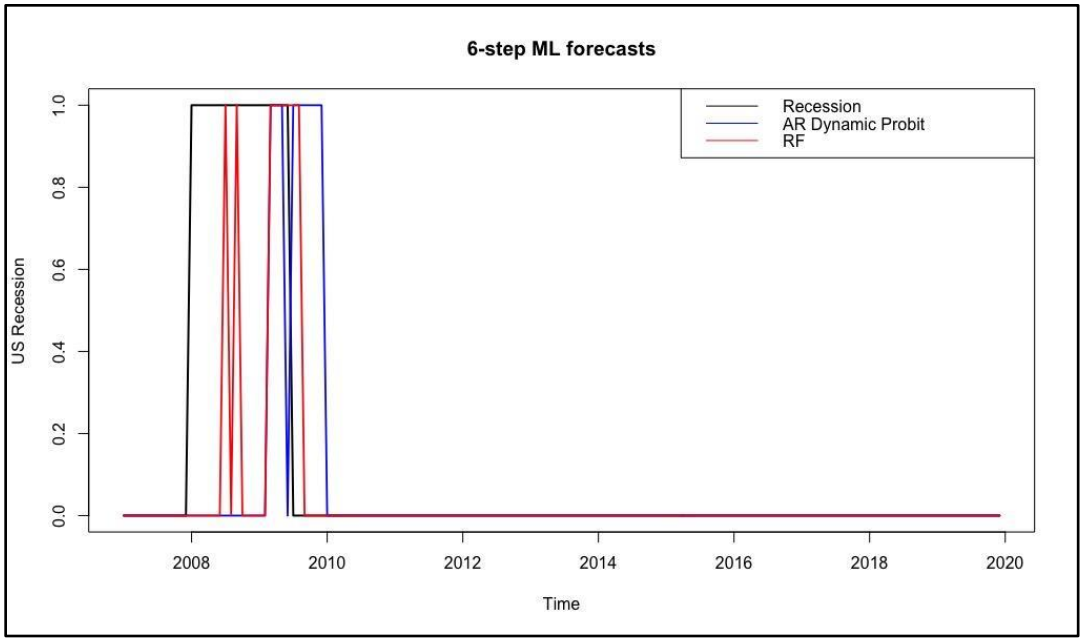
1. NBER - https://research.stlouisfed.org/wp/more/2015-012
2. FRED-MD - https://fred.stlouisfed.org/series/USREC

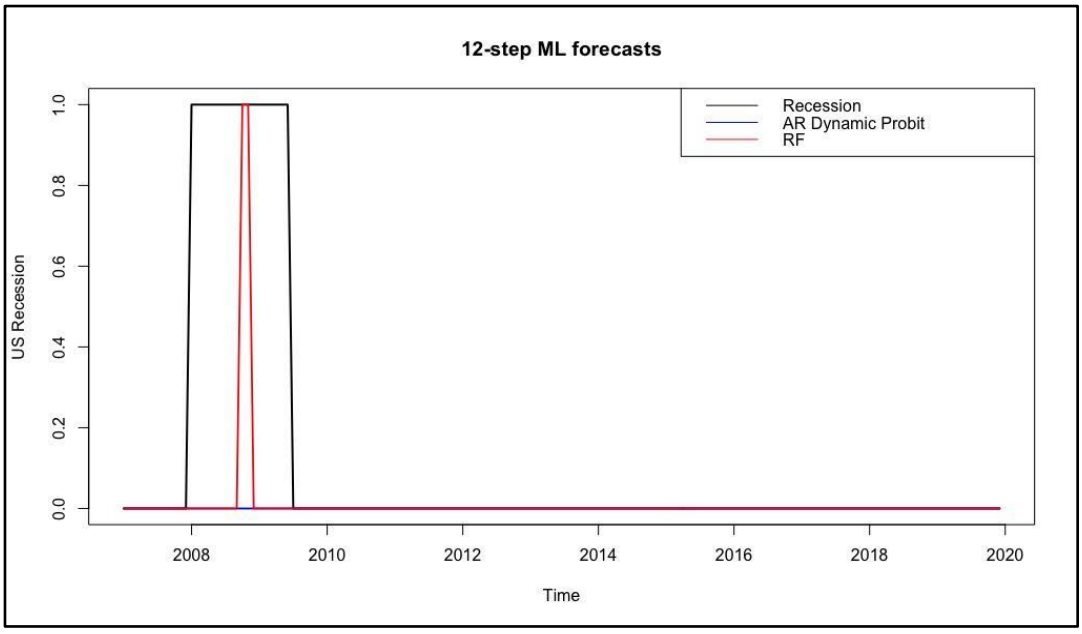
APPENDIX

1] Here, we show the random forecast predictions as binary, when we model the probabilities as higher than 0.5 to be converted to 1. The comparisons are shown as follows. The predictions for 1 step are identically good and it gets worse as we progress, but Random forest tends to outperform AR Dynamic Probit:

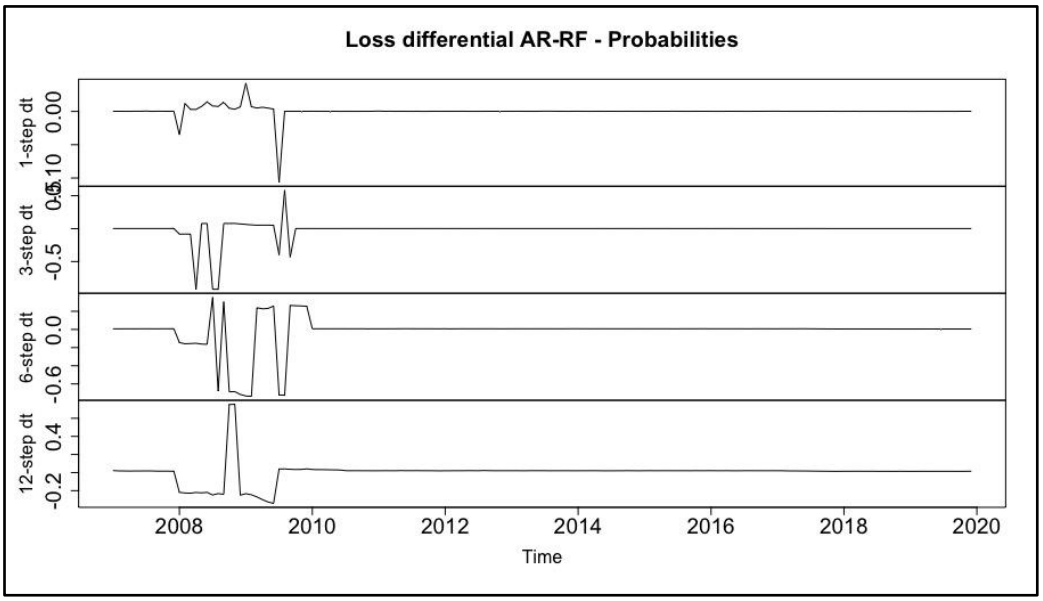





6-step ML forecasts
Recession
AR Dynamic Probit
RF
US Recession
0.0
0.2
0.4
0.6
0.8
1.0
2008
2010
2012
2014
2016
2018
2020
Time


12-step ML forecasts
Recession
AR Dynamic Probit
RF
US Recession
0.0
0.2
0.4
0.6
0.8
1.0
2008
2010
2012
2014
2016
2018
2020
Time

2] The loss of AR over RF in probabilities is shown below:



3] Coding of Variables

Variable Code	Description
USREC	US Recession Indicator
USGOOD	All Employees: Goods-Producing Industries
PAYEMS	All Employees: Total nonfarm
DMANEMP	All Employees: Durable goods
MANEMP	All Employees: Manufacturing
TB3SMFFM	3-Month Treasury C Minus FEDFUNDS
TB6SMFFM	6-Month Treasury C Minus FEDFUNDS
T1YFFM	1-Year Treasury C Minus FEDFUNDS
T5YFFM	5-Year Treasury C Minus FEDFUNDS
T10YFFM	10-Year Treasury C Minus FEDFUNDS
AAAFFM	Moody's AAA Corporate Bond Minus FEDFUNDS