Title
Unveiling Parallels: Analyzing Economic Factors from the 2008 Financial Crisis to the Present Day

Author
Hu, Amanda

Publication Date
2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Unveiling Parallels:
Analyzing Economic Factors from the 2008 Financial Crisis to the Present Day

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Applied Statistics and Data Science

by

Amanda Hu

2023

ABSTRACT OF THE THESIS

Unveiling Parallels:
Analyzing Economic Factors from the 2008 Financial Crisis
to the Present Day

by

Amanda Hu
Master of Applied Statistics and Data Science
University of California, Los Angeles, 2023
Professor Frederic R. Paik Schoenberg, Co-Chair
Professor Maryam Mahtash Esfandiari, Co-Chair

Given the current economic climate, many people believe we are currently in a recession, or at least that such a downturn is on the horizon. By using logistic regression and classification models including decision trees and random forest, we can confirm the economic factors that instigate a recession. In addition, I used machine learning methods such as support vector machine to reinforce the classification in a multivariate space. The structure of this thesis is mainly in a time series analysis format in order to compare the similarities of the decade before the Financial Crisis of 2008 and the past decade, which would ultimately suggest a recession by the end of 2023.

The thesis of Amanda Hu is approved.

Hongquan Xu

Yingnian Wu

Maryam Mahtash Esfandiari, Committee Co-Chair

Frederic R. Paik Schoenberg, Committee Co-Chair

University of California, Los Angeles

2023

TABLE OF CONTENTS

1 Introduction **1**

2 Literature Review **3**

3 Financial Definition **5**

3.1 Treasury Bonds 5

3.2 Federal Funds Rate 5

3.3 Housing Prices 6

3.4 Bank Credit 6

3.5 Unemployment Rate 6

4 Mathematical Theory and Definition **7**

5 Data Analysis **9**

6 Methodology **13**

7 Results **14**

7.1 Linear Regression 14

7.2 Logistic Regression 15

7.3 Probit Regression 17

7.4 Decision Tree and Random Forest 18

7.5 Support Vector Machine 22

7.6 Time Series Analysis 23

8 Forecasting **29**

9 Conclusion and Future Works **32**

References **35**

LIST OF FIGURES

1.1 Inverted Yield Curve [inv23] 2

5.1 Histogram of Predictor Variables 11

5.2 Box Plot For Federal Funds Rate and Recession 12

7.1 Summary Table For Linear Model 15

7.2 Autocorrelation From OLS 15

7.3 Summary Table For Logistic Model 16

7.4 Plot of Odds 17

7.5 Probit Model For Testing Set 18

7.6 Decision Tree For Testing Set 19

7.7 Pruned Decision Trees For Testing Set 19

7.8 Classifier Random Forest For Testing Set 20

7.9 Mean Decrease in Gini Coefficient For Testing Set 21

7.10 Histogram of Number of Tree Nodes For Testing Set 21

7.11 Support Vector Machine Plot 23

7.18 Confusion Matrices After LASSO 27

8.1 Bond Yield Prediction For Decade 1 30

8.2 Bond Yield Prediction For Decade 2 31

9.1 Summary Table of Forecasted Values 32

9.2 T10Y2Y Curve and Recession Indicator For Decade 1 33

9.3 T10Y2Y Curve and Recession Indicator For Decade 2 33

LIST OF TABLES

5.1 Definition of Predictor Variables 10

CHAPTER 1

Introduction

The economy is constantly fluctuating, following the nature of business cycles in expansions and recessions. After the Financial Crisis of 2008, both businesses and consumers learned to recognize the early stages of a potential recession. One of the most commonly analyzed curves is the US treasury bond yield curve, specifically when inverted. The yield curve can take on many forms such as normal, flat, steep, and humped. However, the inverted yield curve raises flags as an early indicator of a recession because it implies that short term investments have a higher yield than long term investments. The inverted yield curve does not need to be strictly downward sloping, but generally suggests that corporations are less likely to borrow money or hire new employees, which demonstrates slower economic activity. By analyzing the federal funds rate and bond yield curve, we can see the efficacy of monetary policy and how the economy is moving in accordance to the expectation hypothesis.

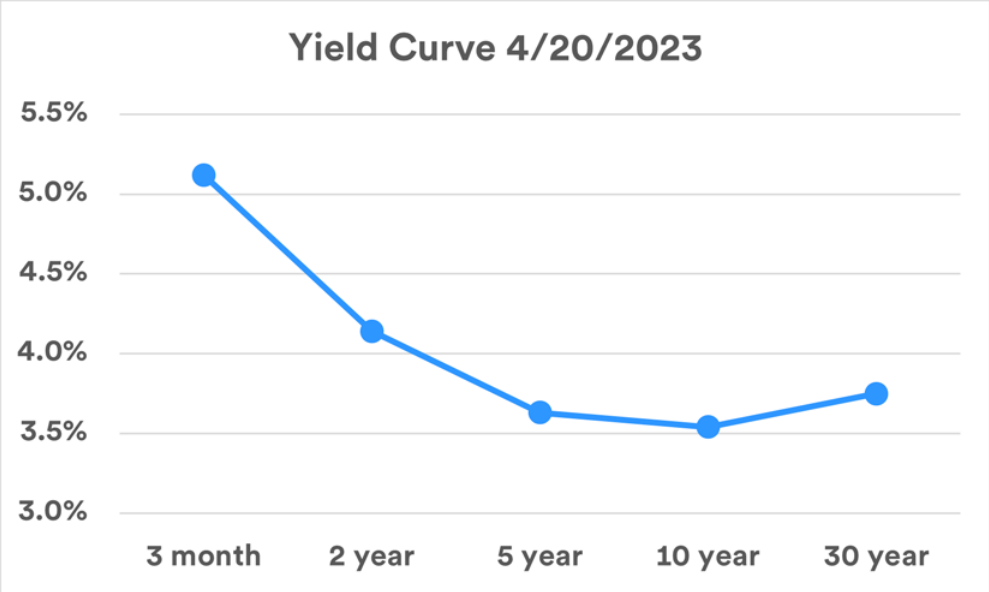


Figure 1.1: Inverted Yield Curve [inv23]

The expectation hypothesis states that long term investments will be less volatile compared to short term investments. In addition, monetary policy is unable to redirect the patterns of the open market. [Woo64] The hypothesis aims to predict the short term interest rates based on long term interest rates and revolves around the idea that an investor can earn the same amount by investing in two one-year bonds as they could with one two-year bond.

CHAPTER 2

Literature Review

In *The Yield Curve and Predicting Recessions* by Jonathan H. Wright [Wri06], the main approach in this paper was the probit model to analyze the probability of a recession. Wright used the definition of recession according to the National Bureau of Economic Research (NBER) — that measures the economic activity by depth, diffusion, and duration. Other variables included the three-month over ten-year spread, nominal federal funds rate, and term premium proxy. This first model proved to have the best in-sample fit and predictive power for out-of-sample data. Then, Wright tests different variations of the model such as changing the NBER variable to indicate a recession in the next t+1 to t+h quarters and including both real and nominal federal funds rates. Moreover, the expectation hypothesis and term premium will have different implications of long term growth. Therefore, Wright uses term spread, level of funds rate, and a forecasting factor as predictors of the model, which shows a statistically significant coefficient for federal funds rate and a negative coefficient for the forecasting factor after six quarters. The conclusion from this paper is the model that includes a binary variable for whether there was a recession, average three-month over ten-year constant maturity Treasury term spread, and nominal federal funds rate performed the best. [Wri06]

Similarly, another paper that looks into the relationship between yield curve and recession prediction is *What Does The Yield Curve Tell Us About GDP Growth?* by Andrew Ang, Monika Piazzesi, and Min Wei [And05]. They emphasize the importance of the yield curve slope as a predictor of the expected GDP growth. "The higher the slope or term

spread, the larger GDP growth is expected to be in the future." [And05] In addition, the term spread is also a strong indicator of the economic cycle. In the past, recessions after the 1960s had a downward sloping yield curve within six quarters of the event. The paper uses ordinary least squares (OLS) to model the relationship between various bond yields and GDP. Specifically, the authors attempt to simplify the variables down while preserving the effect of the yield curve. Furthermore, the model aims to predict out-of-sample data to forecast the GDP growth.

CHAPTER 3

Financial Definition

3.1 Treasury Bonds

Treasury bonds are debt securities issued by the government. Typically, treasury bonds tend to be long term investments that are bought with maturities of more than twenty years. Investors earn interest every period until the maturity date is reached, which is also when they are paid the principal amount. Treasury bonds are relatively low risk because they are backed by the U.S. government and can be fulfilled by increased taxes to its citizens. Therefore, when short term bonds have a higher yield than that of long term bonds, it is a warning sign of an incoming recession.

3.2 Federal Funds Rate

The federal funds rate is defined as the rate that banks borrow from each other in order to manage the money supply and, in turn, adjusts the interest rate. The overnight loans are used between banks to borrow reserves among other commercial banks after the market closes. The federal funds rate highlights inflation trends because they generally move in the same direction. According to the Federal Reserve, the target inflation rate is 2% in order to stabilize prices of goods and services.

3.3 Housing Prices

To continue, housing prices can be helpful to suggest changes in mortgages and the loaning process between consumers and banks. The Housing Price Index (HPI) is released every quarter to report the overall economic activity and market growth. During a recession, housing prices generally fall, which will also affect consumer beliefs of the economy and thus decrease demand for houses and even jobs in the real estate industry.

3.4 Bank Credit

Furthermore, bank credit is an effective measure of the amount available for investors or businesses to borrow from banks based on their ability to repay the loans. The bank uses the funds from clients' checking accounts, savings accounts, or other financial instruments to give out to other investors.

3.5 Unemployment Rate

Lastly, the unemployment rate is frequently used to indicate the percentage of the population that does not currently have a job. Specifically, the segment of the labor force that is actively looking to find employment. Updated every first Friday of the month, the unemployment rate is a lagging indicator which can increase or decrease depending on the current conditions of the economy.

CHAPTER 4

Mathematical Theory and Definition

Logistic regression, also known as logit regression, is used when the outcome variable is binary. The model outputs 1 if the probability of being classified in that category is greater than or equal to 0.5 and 0 if not. Similarly, the probit model will predict the inverse standard normal of the outcome probability. Typically, we assume that the conditional probability is linear. Therefore the logit and probit models can use non-linear methods to improve the interpretation of the probability by setting the outcome as a binary variable. [Gar21]

To continue, decision trees are used not only for classification but also for prediction. Each tree is composed of multiple nodes that are chosen based on the outcome that tests for specific attributes and then creates a label as it continues along the chart. The learning process of a decision tree is generated by creating splits at certain values where the data can be effectively grouped into similar subsets.

Additionally, random forest is a collection of decision trees that takes votes based on the classification decision or average. Random forest applies to both regression and classification cases. Within random forest, there are two different methods to combine models: bagging and boosting. Bagging is when the data is split into a training set with replacement and then the outcome with the majority votes is chosen. Boosting is when many weak models are used continuously to improve the prediction power. Support vector machine (SVM) can also be used for advanced classification models to locate the hyperplane that most accurately categorizes new data according to the existing clusters.

My research question can be best answered by looking at a time series comparing the

two different decades leading up to a recession. For non-stationary models, autoregressive integrated moving average (ARIMA) model is used to better capture the data. This method allows for the dependent variable to be influenced by the past values of the independent variables and, at times, the past values of the dependent variable. [Rob16] The noise in the regression is no longer assumed to be w_t and is instead a linear function of the past values. The autoregressive model (AR) is defined as

$$x_t = \psi_1 x_{t-1} + \psi_2 x_{t-2} + ... + \psi_p x_{t-p} + w_t \tag{4.1}$$

where x_t is stationary, p is the number of steps in the past to obtain the current forecasted value, w_t is the Gaussian white noise, and ψ is the weight as a constant not equal to zero.

$$x_t - \nu = \psi_1(x_{t-1} - \nu) + \psi_2(x_{t-2} - \nu) + ... + \psi_p(x_{t-p} - \nu) + w_t \tag{4.2}$$

The AR model can be interpreted as the first-order model where there are k backward iterations to eventually obtain the linear representation by

$$x_t = \sum_{j=0}^{\infty} \psi^j w_{t-j} \tag{4.3}$$

Therefore the stationary solution of the model above is

$$\sum_{j=0}^{\infty} \psi^j w_{t-j} = \psi(\sum_{k=0}^{\infty} \psi^k w_{t-1-k} + w_t) \tag{4.4}$$

Then the AR process is stationary

$$E(x_t) = \sum_{j=0}^{\infty} \psi^j E(w_{t-j}) = 0 \tag{4.5}$$

CHAPTER 5

Data Analysis

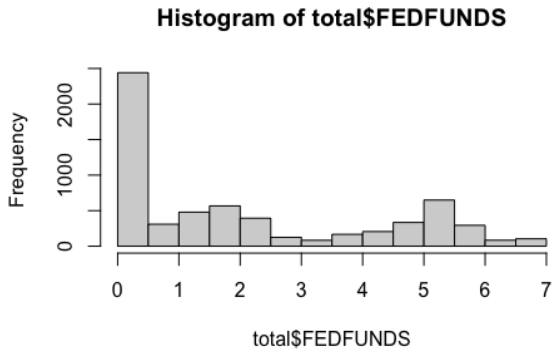
The data used in this analysis are sourced from the Federal Reserve Economic Data (FRED). With the aim of predicting a recession by the end of 2023, I chose to analyze the decade leading up to a recession and will then be used to compare the past ten years. Specifically, this time frame goes back to the year 1997 and extends to 2007. **Research Question:** The main research question is whether economic factors in the ten years leading up to the Financial Crisis of 2008 is repeating and could suggest an approaching recession by the end of 2023. Therefore, we can compare the years 1997 to 2007 and then the year 2012 to 2022 to see if there are parallels between the behaviors of the predictor variables and if it will signify a high probability of a recession.

Data processing makes sure that the data is clean and prepared for the analysis. This means that the data should be good quality with no mismatched data types, mixed data values, data outliers, or missing data. In order to make the data sets in the same frequency, I made the data in daily terms. For example, the unemployment rate is published every month; however, the bond yield is constantly fluctuating. Therefore, I replicated the unemployment rate to populate the data set for each day. The purpose of doing so is to combine tables of the same dimensions and ultimately build a comprehensive data set. In addition, the range of the dates start from the beginning of the year to the end of the year. Assuming that a decade is sufficient to speculate an incoming recession, I assigned two decades which would lead to a recession in the succeeding year. To continue, I created a new variable called *tested* which stems from the variable *USRECD* with values of either TRUE or FALSE indicating

Table 5.1: Definition of Predictor Variables

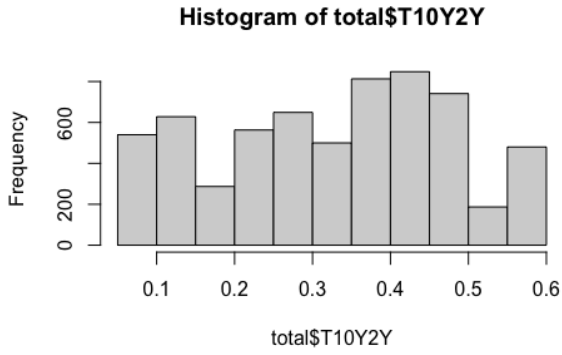
Name	Definition
Federal Funds Rate	The interest rate set by the federal reserve for banks to meet the overnight reserve requirements
T10Y2Y	The 10-year treasury constant maturity minus the 2-year treasury bond constant maturity
T10Y3M	The 10-year treasury constant maturity minus the 3-month treasury bond constant maturity
Unemployment Rate	The percentage of the labor force without employment
Home Price Index	Index that shows the average movement of property prices
Bank Credit	The demand for loans and ability of an investor to take loans from the bank
Recession	The time when economic activity is slow and is typically confirmed with two consecutive quarters with negative Growth Domestic Product (GDP)

if there was a recession on that specific date. The purpose for creating this variable is solely for clarity and making the decision tree nodes easier to interpret with additional information including the number of observations.

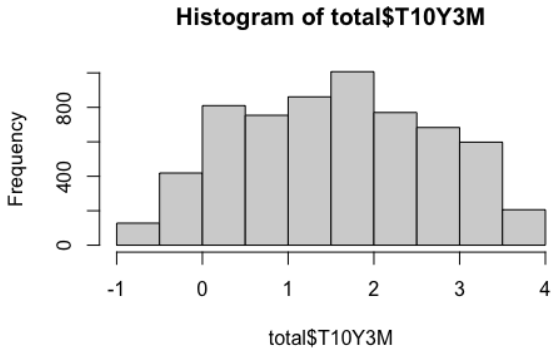


(a) Histogram of Fed Funds Rate

(b) Histogram of Home Prices



(c) Histogram of T10Y2Y



(d) Histogram of T10Y3M

Figure 5.1: Histogram of Predictor Variables

The box plot below shows that when there is no recession, the federal funds rate has a wider range with a median around 1.5%. On the other hand, during a recession, the federal funds rate has a smaller range and a median just above 2%. During a recession, corporations are less likely to borrow money and as a result, the revenue and stock values will be impacted. Thus, consumers will have a negative view on the market and slow down economic activity.

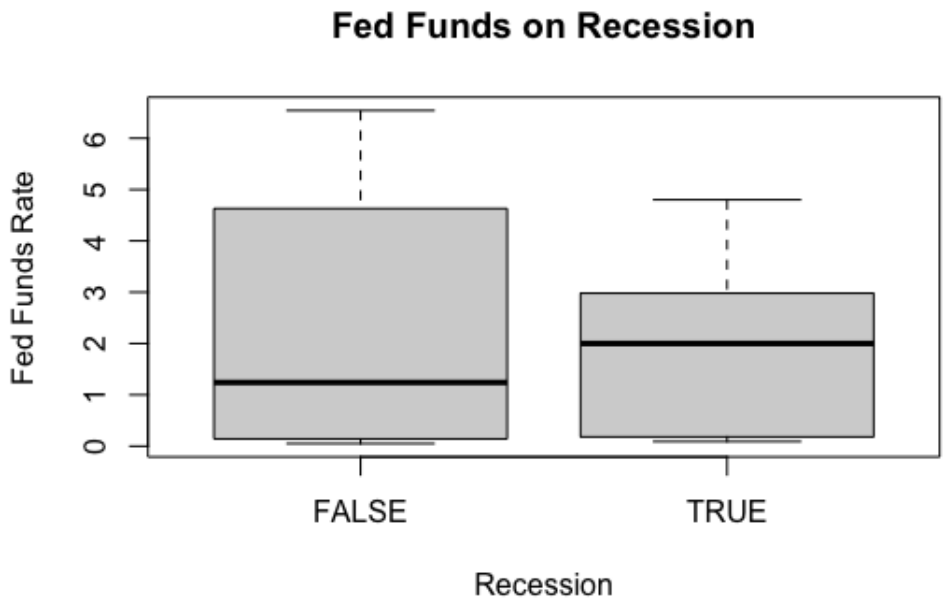


Figure 5.2: Box Plot For Federal Funds Rate and Recession

CHAPTER 6

Methodology

The main research question of this thesis is to compare the decade leading up to a recession and see the commonalities in economic factors such as federal funds rate, bond yield curves, unemployment rate, bank credit, and housing prices. To solve this classification problem, I used logistic regression, decision trees, random forest, support vector machine, and time series analysis.

In general, it is challenging to attempt to model financial data because of the unpredictability of the economy. Therefore, the analysis will take on more of a forecasting direction rather than an exact prediction of the next recession. The models will classify each day as a recession or no recession as well as the confusion matrix to evaluate the accuracy of the model. The emphasis is not on model selection, but rather the comparison between the two decades leading up to a recession in the sense of the magnitude of the coefficient estimates and the distribution between the outcome classifications. Then, the time series analysis will illustrate the projection of the next few years after December 2022.

For regularization of variables, I used the LASSO technique which incorporates a penalty term to induce shrinkage of the coefficients to drive the value closer to zero. This is due to the possibility of multicollinearity between the predictor variables. The economy moves in cycles where there are patterns of growth and decline from a network of contributing factors.

CHAPTER 7

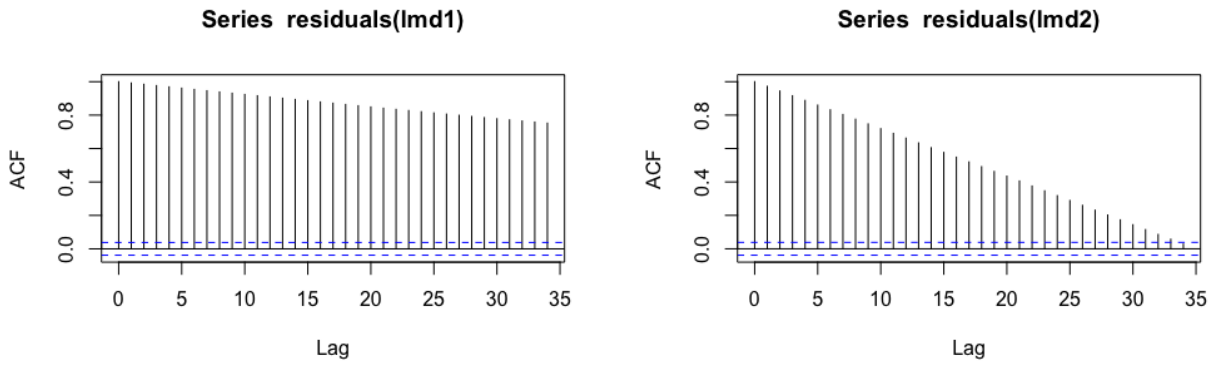
Results

7.1 Linear Regression

Based on the linear regression summary, we can see that all the predictor variables are statistically signification except for the federal funds rate. The T10Y3M bond yield is relatively less statistically significant, but enough to include in the final model. The T10Y2Y estimate is negative whereas the T10Y3M is positive. This supports the conjecture there is an inverted yield curve and long term investments have lower returns as compared to short term investments. The federal funds rate is also negatively related to the recession because as the rate increases, economic activity will slow down. Likewise, the bank credit is also predicted to be have a negative coefficient because even with a higher bank credit, banks do not have enough available funds which tightens lending standards.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.978e+00  1.482e-01  20.100   <2e-16 ***
T10Y2Y      -1.420e+00  7.700e-02 -18.439   <2e-16 ***
T10Y3M       1.486e-02  5.926e-03   2.508   0.0122 *
FEDFUNDS    -6.694e-03  4.923e-03  -1.360   0.1739
HOMEPRICE    2.070e-03  1.537e-04  13.472   <2e-16 ***
BANKCRED    -6.638e-04  2.998e-05 -22.140   <2e-16 ***
UNRATE       3.919e-02  2.839e-03  13.805   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2739 on 6229 degrees of freedom
Multiple R-squared:  0.1172,    Adjusted R-squared:  0.1163
F-statistic: 137.8 on 6 and 6229 DF,  p-value: < 2.2e-16
```

Figure 7.1: Summary Table For Linear Model



(a) Series Residual For Decade 1

(b) Series Residual For Decade 2

Figure 7.2: Autocorrelation From OLS

7.2 Logistic Regression

The **glm** function in R is similar to the **lm** function; however, with the specification of **family = binomial**, R understands that it will be performing a logistic regression. Using this model to make predictions on a specified day, the function **predict()** shows the pre-

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.924e+01  2.551e+00  15.380   <2e-16 ***
T10Y2Y      -2.184e+01  1.369e+00 -15.949   <2e-16 ***
T10Y3M       2.845e-01  9.373e-02   3.035   0.0024 **
FEDFUNDS     1.126e-01  7.795e-02   1.444   0.1487
HOMEPRICE    2.975e-02  2.783e-03  10.688   <2e-16 ***
BANKCRED    -9.671e-03  5.414e-04 -17.862   <2e-16 ***
UNRATE       5.205e-01  3.489e-02  14.919   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3877.6  on 6235  degrees of freedom
Residual deviance: 3107.6  on 6229  degrees of freedom
AIC: 3121.6

Number of Fisher Scoring iterations: 6
```

Figure 7.3: Summary Table For Logistic Model

dictions from the training data to categorize the data as TRUE or FALSE depending on if the predicted probability is greater than or less than 0.5. Then, the next step is to make a final prediction in terms of the chosen outcome TRUE or FALSE. The result is a correlation matrix where the diagonal shows the correct predictions and the off-diagonals are incorrect predictions. [Gar21] Most times the results will be optimistic because the logistic model uses the same data to train and test. Hence, the bootstrap method splits the data set by a ratio of 8:2 for training and testing respectively.

The logistic regression summary also confirms that all of the predictor variables are statistically significant except for the federal funds rate. Otherwise, the signs of the coefficient estimates are the same as those in the linear model. The plot of odds shows the relationship between each predictor variable and the odds of a recession occurring. It also illustrates the magnitude and trend across variables. In this case, *BANKCRED* and *HOMEPRICE* have the highest odds in predicting a recession.

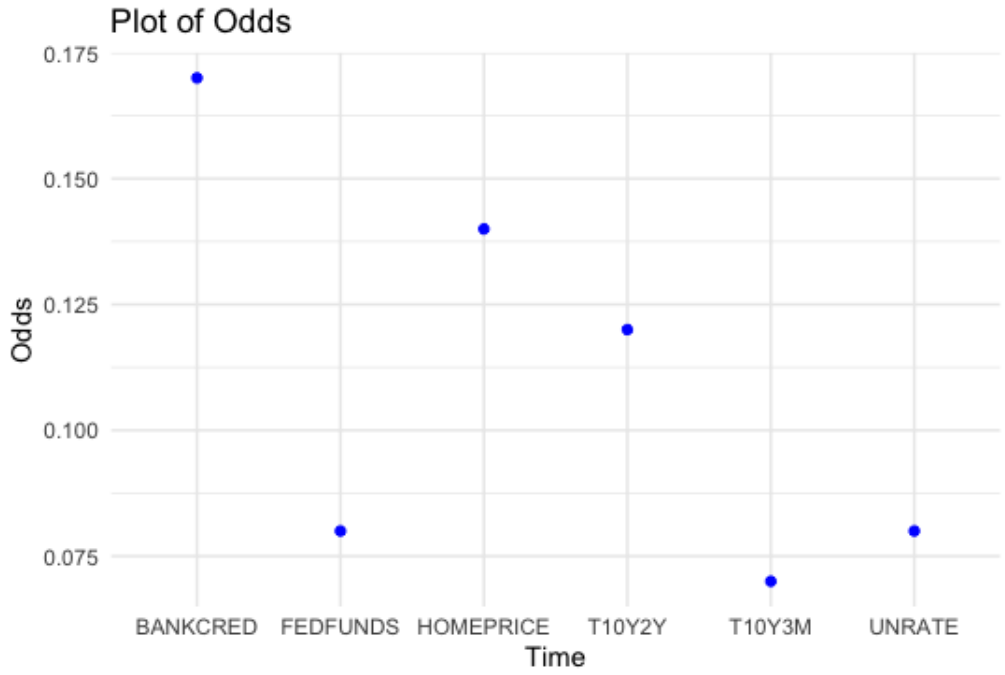


Figure 7.4: Plot of Odds

7.3 Probit Regression

In probit regression, the main assumption is that

$$E(Y|X) = P(Y = 1|X) = \phi(\beta_0 + \beta_1 X) \tag{7.1}$$

$$\phi(z) = P(Z \leq z), Z \ N(0, 1) \tag{7.2}$$

Then, the coefficient β_1 is the change in variable z when there is a one unit change in the variable X. Therefore, the coefficient for X does not have a direct interpretation and will need to be computed through the **pnorm()** function.

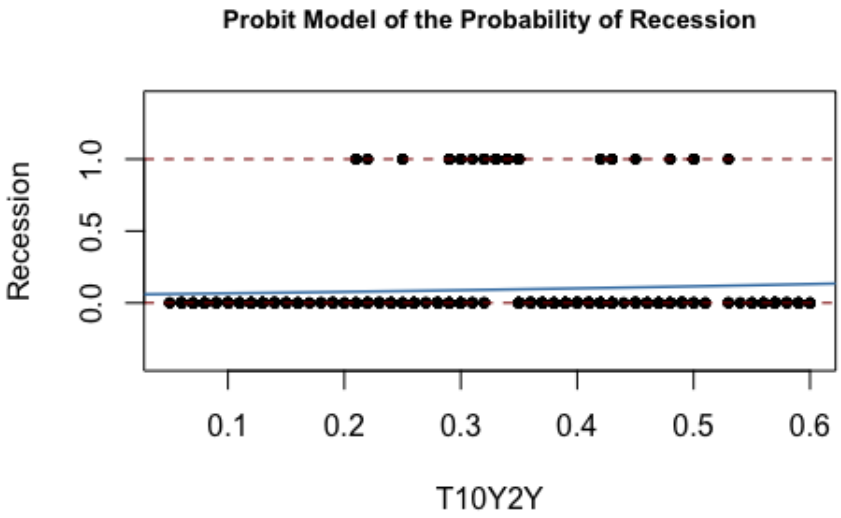


Figure 7.5: Probit Model For Testing Set

This plot shows the probability of a recession from the testing set. From this model, there is a low probability of a recession.

7.4 Decision Tree and Random Forest

By splitting the data to an 8:2 ratio of training and testing sets respectively, we can create decisions trees as seen in the following plot.

The original plot is too cluttered and can be simplified down to fewer trees. Therefore, I created a new factor *tested* that defines the variable *USRECD* as TRUE or FALSE to indicate whether there was a recession or not. In addition, the new plot prunes the decision trees based on the best complexity parameter value.

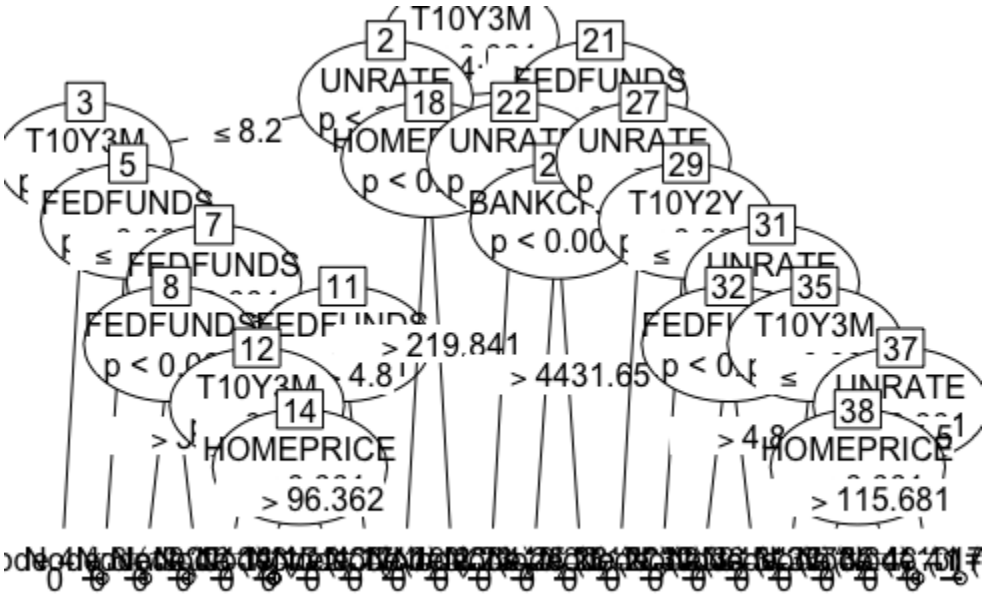


Figure 7.6: Decision Tree For Testing Set

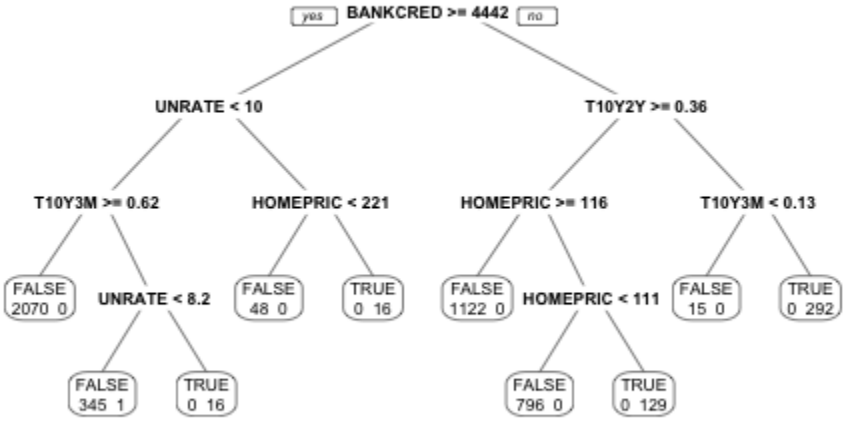


Figure 7.7: Pruned Decision Trees For Testing Set

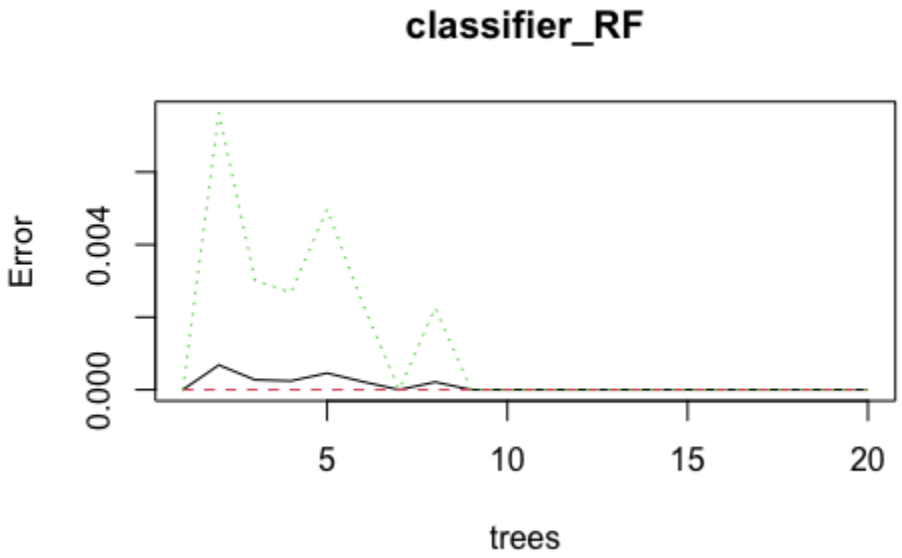


Figure 7.8: Classifier Random Forest For Testing Set

```
predict_model
  FALSE TRUE
0  1253    3
1     9  121
```

The model correctly predicts no recession 1253 times and predicts a recession 121 times. There are 9 false positive cases and 3 false negative cases. Ultimately, the model has an accuracy of 99%. This gives us confidence that the chosen predictor variables and parameter estimates can effectively explain the data.

The mean decrease in Gini coefficient describes the importance of the variable in the homogeneity of the decision trees. This value should be higher to show it is an important predictor variable to the model. From the plot above, we can see the Gini coefficient starts to level off and the histograms are evenly distributed, which means that 20 decision trees should be adequate for this data set.

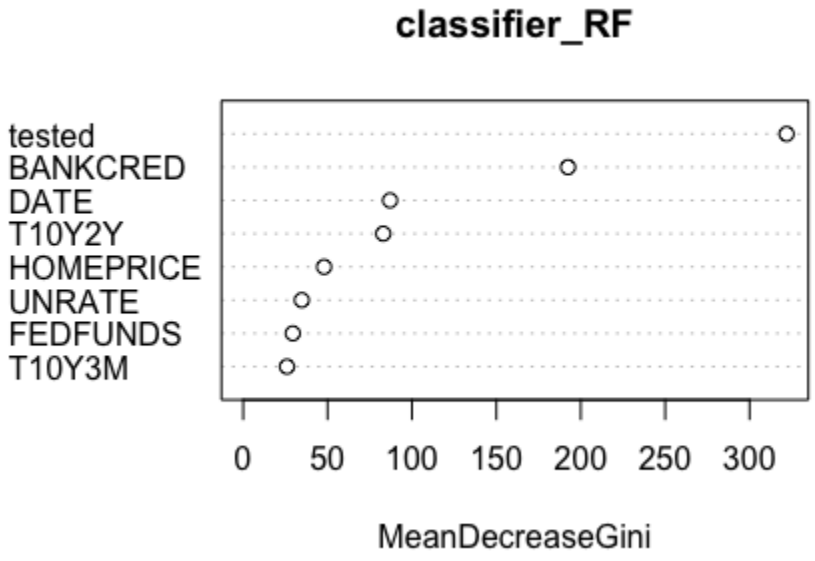


Figure 7.9: Mean Decrease in Gini Coefficient For Testing Set

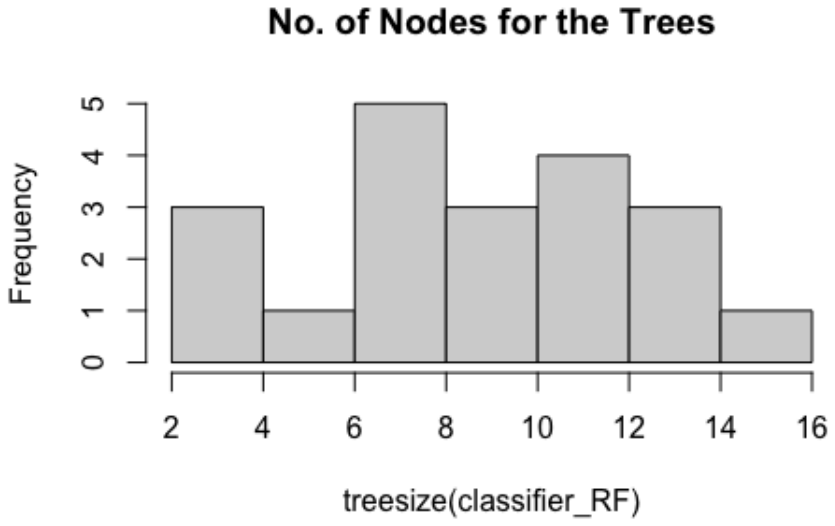


Figure 7.10: Histogram of Number of Tree Nodes For Testing Set

7.5 Support Vector Machine

Support vector machine (SVM) is suitable for binary classifications. The idea is to find an optimal hyperplane that would separate the data into two categories. A hyperplane is of dimension p-1 in a p-dimensional space with equation

$$\beta_0 + \beta_1 X_1 + \beta_2 X2 + ... + \beta_p x_p = 0 \tag{7.3}$$

Depending on whether the value X is greater than or less than the hyperplane equation, the SVM model classifies which side the point lies in relation to the hyperplane. In matrix form, the hyperplane has equation

$$x_1 = \begin{bmatrix} x_{11} \\ \vdots \\ x_{1p} \end{bmatrix}, \ldots, x_{np} = \begin{bmatrix} x_{n1} \\ \vdots \\ x_{np} \end{bmatrix} \tag{7.4}$$

Therefore, X can either be on the left or right side of the hyperplane. The farther away X lies from the hyperplane, the more certain we can be about the accuracy of the classification, whereas the closer it is to the hyperplane, the more uncertain the classification. The maximal margin classifier is the hyperplane that will maximize the margin between the training set. The largest perpendicular distance from the training observation to the hyperplane is the maximal margin classifier. A notable property of this method is it only depends on the support vectors. Thus, the classification would not change with any movement of the other observations.

The advantage of using SVM is its applicability in high-dimensional spaces. However, the kernel is sensitive to the chosen parameters when the number of features is larger than the training dataset. This would result in poor classifications because the model is attempting to find an optimal hyperplane with new samples added.

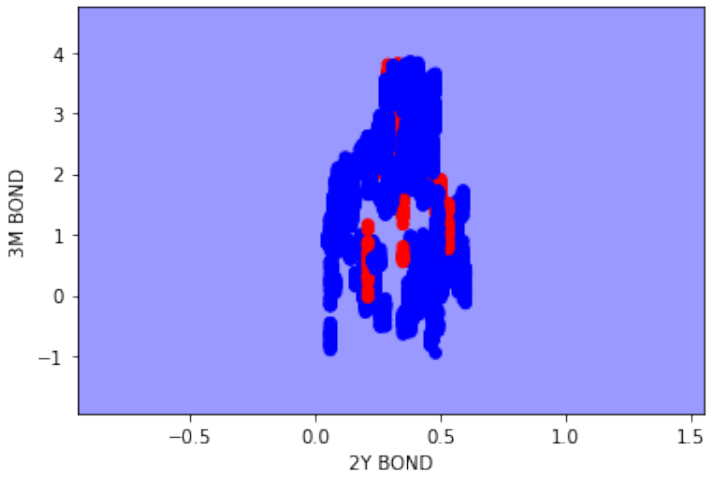


Figure 7.11: Support Vector Machine Plot

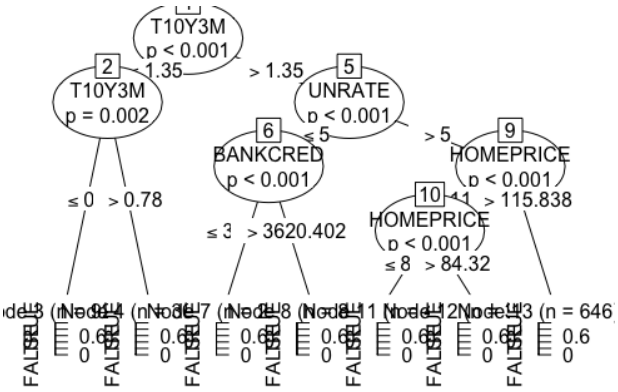
The plot above shows the classification from using the support vector machine method. When the boundaries overlap as so, it means that the data set is likely more complex than a linear model. However, SVM can still be performed on non-linear data with the help of non-linear kernels. Possible reasons why the boundaries overlap with the given dataset is because of the inherent complexity of modeling financial data from noise to time covariates. Specifically, one of the assumptions of SVM is having a balanced class distribution. The current dataset has more data classified as no recession because the nature of my research question is to analyze the trends leading up to a recession. Therefore, the SVM model may need further training with more instances of a recession or more model features.

7.6 Time Series Analysis

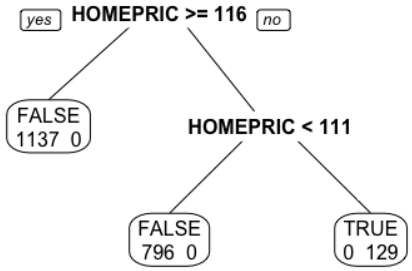
The first decade from 1997 to 2007 would ideally illustrate the trends leading up to the 2008 Financial Crisis, which we know to be true. Based on the decision tree of decade 1, the model predicts 34 cases of a recession on a daily basis with an accuracy of 99%. Likewise, the second decade spans from 2012 to 2022 and suggest a recession in the year 2023. The second decade correctly predicts a recession in 9 cases which means there are early signs of

a recession.

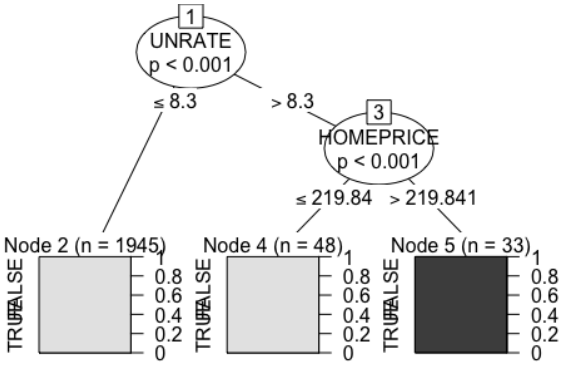
With the first decade, the probit model suggests that all predictor variables are statistically significant. Whereas in the second decade, the probit model highlights only the federal funds rate, home prices, and bank credit as statistically significant.



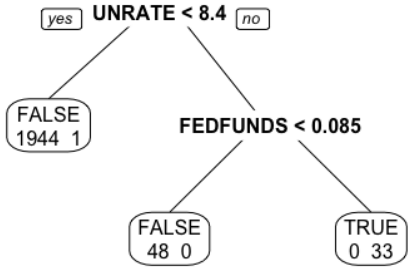
(a) Decision Tree For Decade 1



(b) Pruned Decision Tree For Decade 1



(a) Decision Tree For Decade 2



(b) Pruned Decision Tree For Decade 2

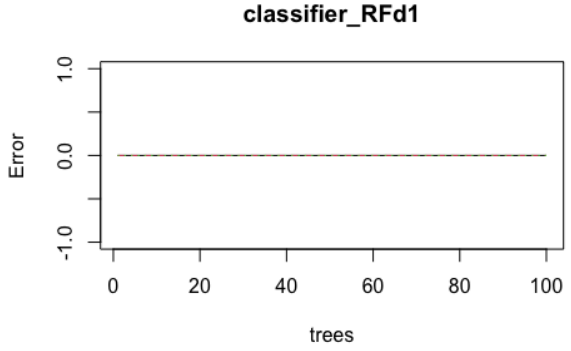
```
Call:
 randomForest(x = train_datad1[-9], y = train_datad1$tested, ntree = 20)
               Type of random forest: classification
                     Number of trees: 20
No. of variables tried at each split: 2

        OOB estimate of  error rate: 0%
Confusion matrix:
      FALSE TRUE class.error
FALSE  1933    0           0
TRUE      0  129           0
```

(a) Confusion Matrix For Decade 1

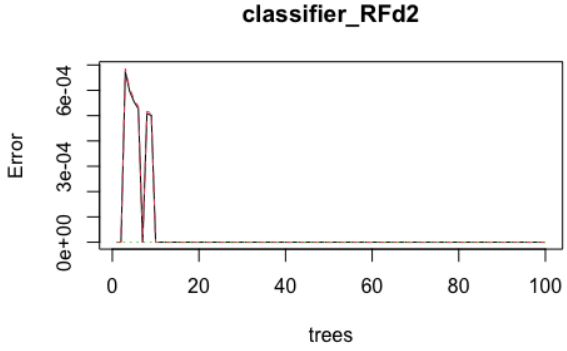
```
Call:
 randomForest(x = train_datad2[-9], y = train_datad2$tested, ntree = 20)
               Type of random forest: classification
                     Number of trees: 20
No. of variables tried at each split: 2

        OOB estimate of  error rate: 0%
Confusion matrix:
      FALSE TRUE class.error
FALSE  1992    0           0
TRUE      0   34           0
```

(b) Confusion Matrix For Decade 2

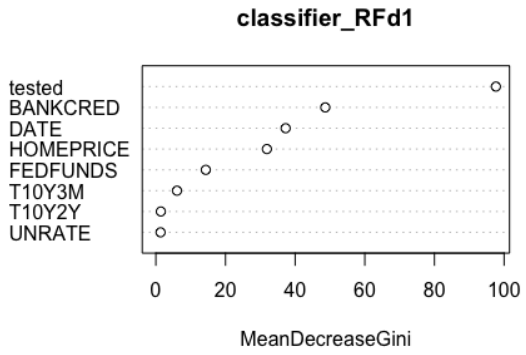
During the training of the random forest model, the model suggests that in the first decade, there are 129 cases of the model accurately predicting a recession. The plot shows that the error is zero, which may imply that there is overfitting or multicollinearity in the model because of the perfect separation between classes. Furthermore, the model accurately predicts the case of a recession 34 times in the second decade.



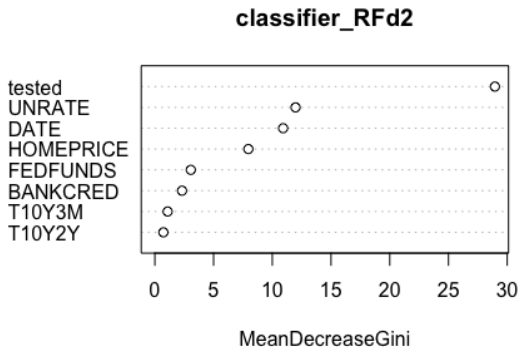
(a) Classifier Error For Decade 1



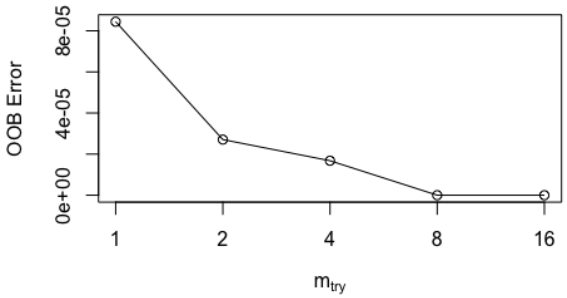
(b) Classifier Error For Decade 2



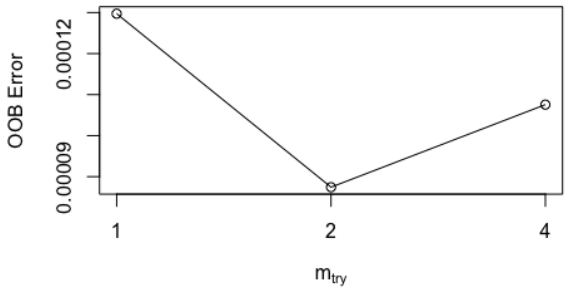
(a) Mean Decrease Gini Coefficient For Decade 1



(b) Mean Decrease Gini Coefficient For Decade 2



(a) Out-of-bag Error For Decade 1



(b) Out-of-bag Error For Decade 2

From these plots, we can see that first decade and second decade have similarities in terms of how the predictor variables decrease the mean Gini coefficient by improving the homogeneity of the decision tree nodes. Therefore, the random forest method concludes that bank credit and 10-year treasury constant maturity minus 2-year treasury constant maturity yield are the most significant variables for the classification model as proven by the flattened curve for the mean decrease Gini coefficient.

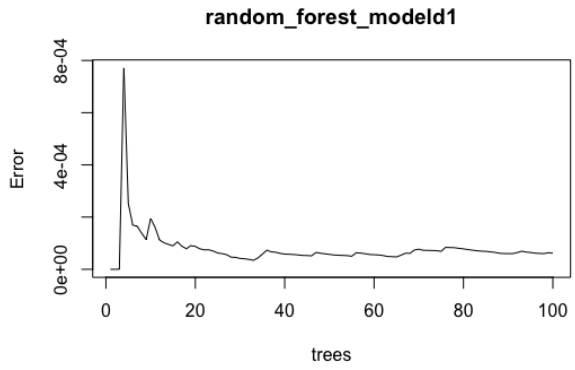
On the other hand, there is likely overfitting in the model because of the zero error. To solve this issue, I used the least absolute shrinkage and selection operator (LASSO) method

to stabilize the coefficients. As a result, the coefficients show a shrinkage towards zero with a penalty term that decreases the coefficient magnitude. Using the adjusted coefficients, the random forest in the first decade now accurately predicts 28 cases of a recession and 3670 cases of no recession. The model does not predict a recession 418 times when there was, in fact, a recession.

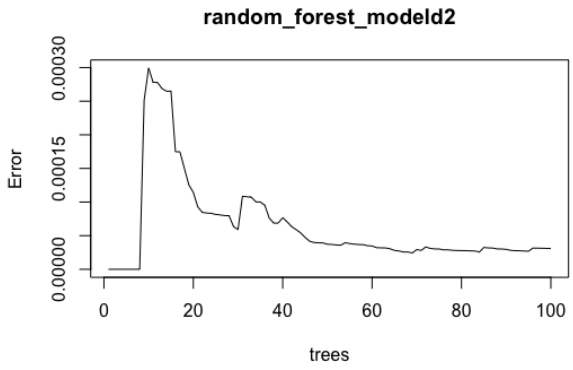
Similarly, after performing LASSO to the second decade, the model accurately predicts 33 cases of a recession and 3702 cases of no recession. The error for this model is 418 cases of no recession when it should have been classified as a recession. The plots now show that home price and bank credit are the most important variables for the predictive power of the model.

```
> confusion_lassod1
   binary_predictionsd1
       0    1
  0 3670    0
  1  418   28
> confusion_lassod2
   binary_predictionsd2
       0    1
  0 3702    0
  1  418   33
```

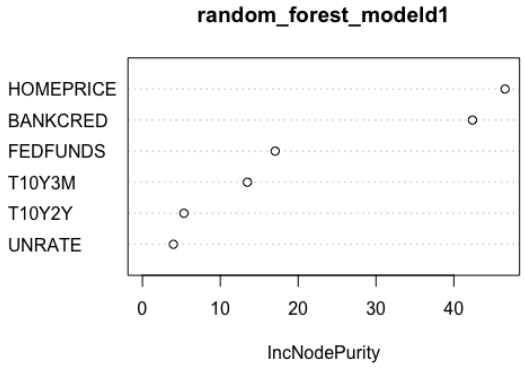
Figure 7.18: Confusion Matrices After LASSO



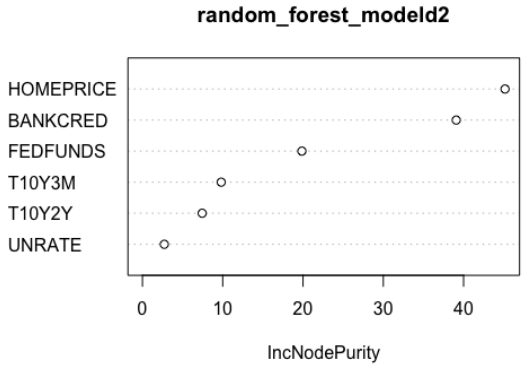
(a) Classifier Error For Decade 1



(b) Classifier Error For Decade 2



(c) Mean Decrease Gini Coefficient For Decade 1



(d) Mean Decrease Gini Coefficient For Decade 2

CHAPTER 8

Forecasting

Generally, forecasting aims to predict the future trends and values of a time series. This comes with the assumption that x_t is stationary and the parameters are known. [Rob16] The following equations show the minimum mean square error predictor of x_{n+m}. When the predictor minimizes the mean square error, the predictor is also known as the best linear predictor (BLP). In fact, under the Gaussian process, the minimum mean square error predictor and best linear predictor are equivalent.

$$x_{n+m}^n = E(x_{n+m}|x_{1:n}) \tag{8.1}$$

$$E[x_{n+m} - g(x_{1:n})]^2 \tag{8.2}$$

$$x_{n+m}^n = \alpha_0 + \sum_{k=1}^{n} \alpha_k x_k \tag{8.3}$$

The best linear predictor notation is as follows.

$$x_{n+m}^n = \nu + \sum_{k=1}^{n} \alpha_k (x_k - \nu) \tag{8.4}$$

Using the method from Section 7.5, we can forecast the bond yield for both decades in order to understand the relationship that it may have on the model. The plot of the bond yield prediction in the first decade shows a downward sloping trend with a particularly low point in the year 2045. Moreover, the plot for second decade also shows a low point shortly after the year 2040. Although there are many factors that could change the behavior of any economic index, according to the current state of the 10-year treasury constant maturity minus 2-year treasury constant maturity yield curve and the limitations of the ARIMA model, the model

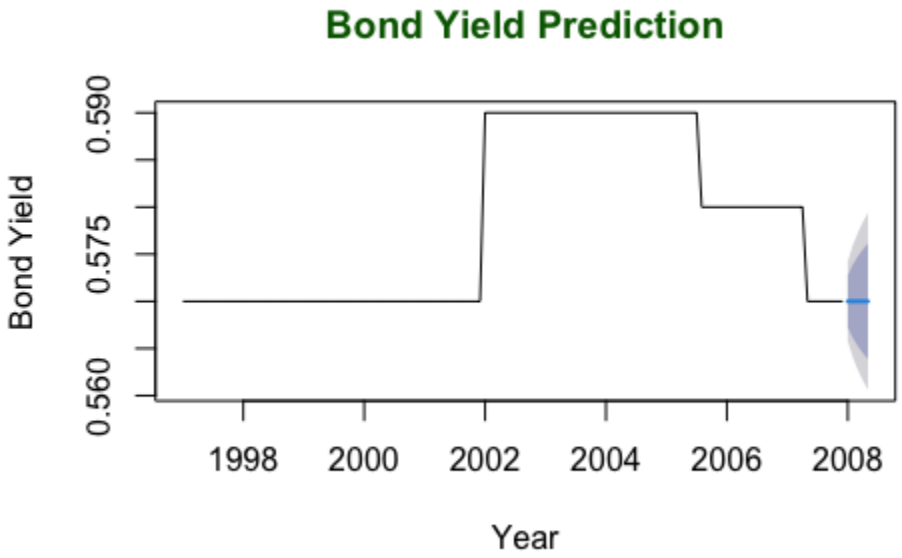


Figure 8.1: Bond Yield Prediction For Decade 1

suggests that there will likely be a consequential drop between the years 2040 and 2045.

The importance of these plots lies in the estimated region on the right. This provides some evidence that there will be a recession because, as mentioned before, the bond yield curve is commonly used as an early sign of a recession. Looking at the predicted values of the bond yield, we can see that in 2008, the bond yield curve started to decrease significantly creating an inverted curve over time. Right before 2020, the bond yield decreased which could also suggest a brief recession. However, soon after the year 2020, the bond yield increased again. Although there are other predictor variables that are statistically significant from the previous models, based on the 10-year treasury constant maturity minus 2-year treasury constant maturity, the forecasting results do not suggest a recession as of now.

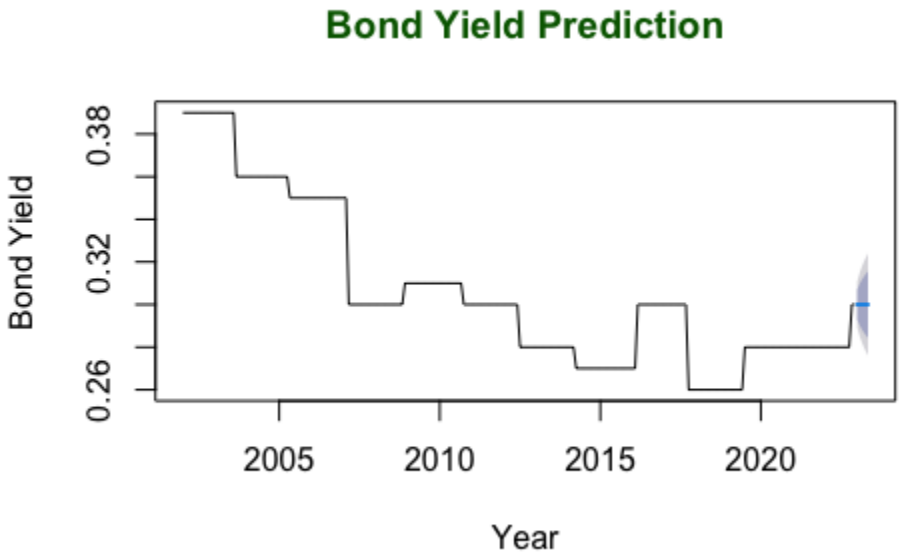


Figure 8.2: Bond Yield Prediction For Decade 2

CHAPTER 9

Conclusion and Future Works

As a result of the time series analysis, the plots above illustrate a comparison of the 10-year treasury constant maturity minus 2-year treasury constant maturity yield and the outcome variable *USRECD* which is equal to 1 if there was a recession and 0 if not. From the first decade, the projection is that there was a brief recession around 2018-2019 before the pandemic. This is backed by the inverted yield curve above showing a negative relationship between bond yield and time. The second decade shows that there will be a recession around the year 2047. The following summary table shows the next five predictions of the 10-year treasury constant maturity minus 2-year treasury constant maturity yield.

However, it should be noted that there is a tendency for the model to recognize the sharp drops in bond yield to return an indication of a recession. But as seen throughout history, it takes time for various economic factors to accumulate into a full recession.

In the future, further research can be done to improve the recession and bond yield prediction model by adding more economic factors as predictor variables. In addition, more

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
2047.806	0.38	0.3750023	0.3849977	0.3723567	0.3876433
2047.825	0.38	0.3729322	0.3870678	0.3691908	0.3908092
2047.845	0.38	0.3713438	0.3886562	0.3667615	0.3932385
2047.864	0.38	0.3700047	0.3899953	0.3647135	0.3952865
2047.883	0.38	0.3688249	0.3911751	0.3629091	0.3970909

Figure 9.1: Summary Table of Forecasted Values

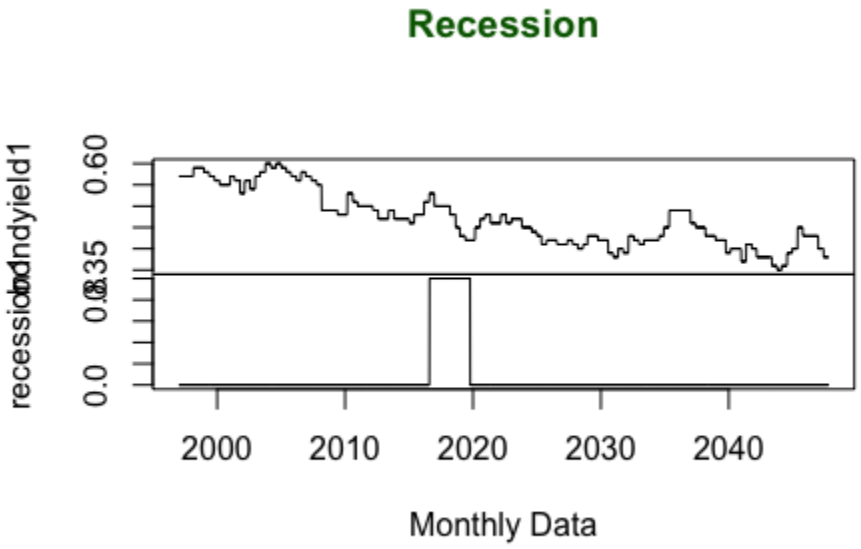


Figure 9.2: T10Y2Y Curve and Recession Indicator For Decade 1

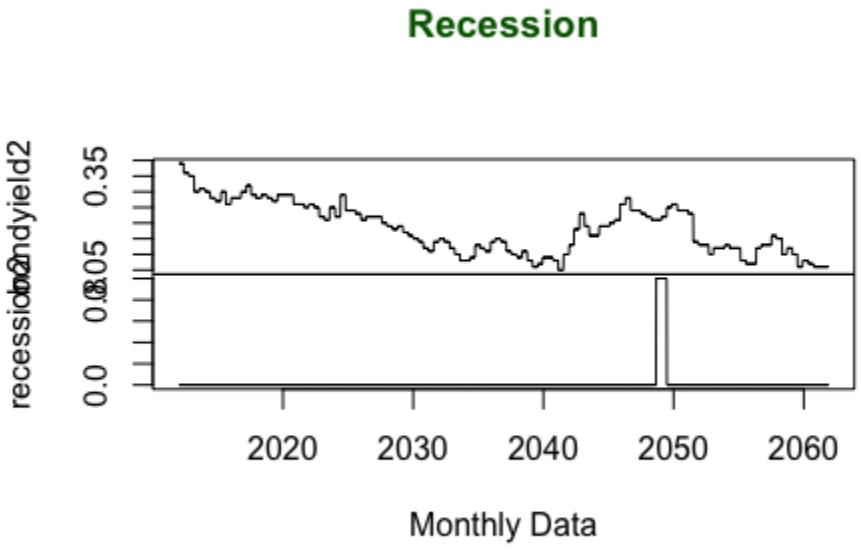


Figure 9.3: T10Y2Y Curve and Recession Indicator For Decade 2

analysis could be done on previous recessions besides the Financial Crisis of 2008. To continue, different machine learning methods can be used to effectively build and learn a model that recognizes the patterns of a recession. The methods used in this paper do not necessarily aim to predict the next recession but instead, to identify the important factors that affect the economy especially during a decline. The models are meant to show a comparison of the patterns witnessed in 2008 and present day.

REFERENCES

[And05] Min Wei Andrew Ang, Monika Piazzesi. "What Does The Yield Curve Tell Us About GDP Growth.", 3 2005.

[Gar21] Trevor Hastie Gareth James, Daniela Witten. *An Introduction to Statistical Learning*. Springer, 8 2021.

[inv23] *Treasury Yields Invert As Investors Weigh Risk of Recession*. US Wealth Management, 4 2023.

[Rob16] David S. Stoffer Robert H. Shumway. *Time Series Analysis and Its Applications*. Springer, fourth edition, 9 2016.

[Woo64] John H. Wood. *The Expectations Hypothesis, The Yield Curve, and Monetary Policy*. Oxford University Press, 8 1964.

[Wri06] Jonathan H. Wright. "The Yield Curve and Predicting Recessions.", 2 2006.