Hussam Taj

Darrin Weber

STAT 493

26 April 2020

Analyzing the Next Economic Recession Using Machine Learning

**Abstract**

This paper explains the effects of an economic recession on a country’s economy and its financial, psychological, social, and mental repercussions on its society. It will be imperative to counter these issues by getting a prediction or an indication for an upcoming recession. This way people can get ahead prepare better for tough financial times and feel more comfortable about their financial situation. Although it is impossible to accurately predict the next economic recession because there are so many factors that cannot be foreseen through computational techniques such as a pandemic like COVID-19. There are five variables that have historically provided some precise signs and indications for an upcoming recession which are the yield curve, unemployment rate, gdp output, consumer confidence and the stock market. I used three of these variables – the yield curve, unemployment rate, and GDP output – and machine learning techniques including Generalized Additive Models and Random Forest to analyze and indicate the next economic recession.

**An Economic Recession**

Today we are living in one of the most vulnerable economic times in the history. Due to the outbreak of the coronavirus (COVID-19), the whole world is facing a very tough financial situation. Many countries have been in lockdown for over a month with very slow economic growth. Unfortunately, this situation is expected to get worse as the International Monetary Fund predicts the worst downturn since the Great Depression in the 1930s (Rappeport and Smialek, 2020). More than 70% of economists believe that a US recession will strike by the end of 2021 because the yield curve inverted in August of 2019 (Heeb, 2019). The yield curve is a line that plots interest rates (yields) of bonds that have equal credit quality but differing maturity dates. The slope of this line givens an idea of future interest rates and economic activity (Chen, 2020). Every time short-maturity rates exceed long-maturity rates, meaning inversion of the curve, a recession occurs in the near future. In addition, after the pandemic, the new figures explain that 38% of economists expect a recession in 2020 while 34% still expect it in 2021 (Heeb, 2019). These expectations are influenced by many different economic and social factors as well as intuition – which is based on experience of the economist.

A recession is declared when the economic growth is negative for two or more consecutive quarters. This is measured by the gross domestic product, which is the total monetary or market value of all finished goods and services produced within a country’s borders in a specific time period (Chappelow, 2020). During a recession, unemployment rates tend to rise very quickly as companies face increased costs, falling revenue, and high financial pressure to lay off workers to cut costs. This leads to longer average unemployment periods for people who are searching for jobs. Due to this, people stop spending money on unnecessary things and make big personal budget cuts. One of the major issues rises with mortgage payments, with lesser income and high unemployment, people find it harder to pay their monthly mortgage payments. This leads to many homeowners starting to default on payments, houses to lose value, and the housing market to start to crashing. In addition, due to less cash flow, the Stock Market takes a major blow as stocks start to devalue. All of these factors create a very intense living environment for people.

According to an article by BMC Public Health, a research provided evidence that economic recession and mediators such as unemployment, income decline, and unmanageable debts are significantly associated with poor wellbeing, increased rates of common mental disorders, substance-related disorders and suicidal behavior (Frasquilho, D., Matos, M., Salonna, F., 2015). A recommended solution was taking cost-effective measures that can possibly reduce the occurrence of negative mental health outcomes on the population. These issues are becoming increasingly popular in the new generations and they get more severe during the financial crises. It is important to understand the effects an economic recession and its long-term financial and personal implications. If we can indicate an upcoming recession, we can prepare for it beforehand by building up an emergency fund, start to budget spending early, get ahead of any debt, maintain regular investment, and refine and diversify our skill set. This will help people in surviving the financial crisis without any serious consequences.

I decided to create a model that would use the yield curve, unemployment rate, GDP output, consumer confidence and the stock market as my variables to get an indication of an upcoming recession. I decided to use the yield curve as my response as it has historically proved to be the most accurate indicator. I had to make some assumptions to be able to create this model. One of major assumptions I had to make was that a future recession would be similar to the historic recessions. This assumption makes my model very generalized and vague because each recession in the past occurred due to unique circumstances and the economy recovered in a unique way. For example, after the 9/11 terrorist attacks in New York City, the U.S economy went into a brief recession, but that had no relation with unemployment rate or the GDP output. Similarly, in 2009, the government bailed many banks out and saved them from defaulting and those steps were taken due to specific circumstances. These factors make getting an accurate prediction or an indication of an upcoming recession almost impossible. However, there are some probability models that have been implemented by the New York Fed and Rabobank, but they only have a 12-month forecast and only rely on one variable. Guggenheim Partners and Wells Fargo Economics also have recession indicators that use economic and market data, but they have a limited time frame for the forecast (Zhang, 2019). After understanding these factors and considering the assumptions, I decided to change my approach towards the solution.

I decided to first understand the relationship between my variables and then get a prediction of a near future. For example, if the economy already had a negative growth in last two quarters, by looking at the unemployment rate and GDP output rate, we can get an idea of the yield curve and if it is going to invert soon. This way we can start preparing for tough financial times and get ahead on budgeting, debts, and investments.

**Data**

I attained the yield curve data from the Federal Reserve Economic Data (FRED) website. This is an online database that consists of hundreds of thousands of economic data time series extracted from national, international, public and private sources. It was created and maintained by the Federal Reserve Bank of St. Louis. Series are calculated as the spread between 10-year Treasury Constant Maturity and 3-month Treasury Constant Maturity. The time series graph from 1982 to present is represented on Figure 1. As we can see, the yield curve inverted in June 1989, November 2000, July 2006 till July 2007, and August 2019. This is a good way to keep track of the yield curve since the data is constantly updated by the Federal Reserve.

A close up of a map

Description automatically generated

Figure 1 - The time series graph from 1982 to present

I got the unemployment rate data from the U.S Bureau of Labor Statistics. The data is collected by the Bureau daily, but it is averaged to get the monthly figures. Figure 2 shows the monthly unemployment rates from 1982 to 2020. As we can see, the unemployment rates rose in the year following every yield curve inversion. By looking at the graphs, we can see a correlation between both variables.

A picture containing table, man

Description automatically generated

Figure 2 - Monthly unemployment rates from 1982 to 2020

The gross domestic product data was also obtained from the FRED website. The real potential GDP is an estimate of the output the economy would produce using high rate of its capital and labor resources. The GDP is reported quarterly. Figure 3 represents the graph between billions of chained 2012 dollars each year from 1982-2020. Although the earnings keep increasing in a quite linear fashion, it has a strong relationship with economic progress and recession indication. The grey chunks on the graph represent negative growth which happened closer to recession times. The grey chunk between 2001-2002 represents the 9/11 terrorist attacks.

A screenshot of a cell phone

Description automatically generated

Figure 3 - Billions of chained 2012 dollars each year from 1982-Present

The consumer confidence data was attained from the Conference Board Website. It is a non-profit organization that consists of think-tank members that provide trusted insights on “what’s ahead”. The data is based on business cycle and labor trends as well as structural underpinnings of sustainable growth. I acquired the data from 1982 till 2010 because that was the latest available data. The data was in a unique format (Figure 4) and cleaning it would take a very long time and unfortunately due to time constraints, I had to remove it from analysis. For the same reasons, I also eliminated the Stock Market data from the analysis.

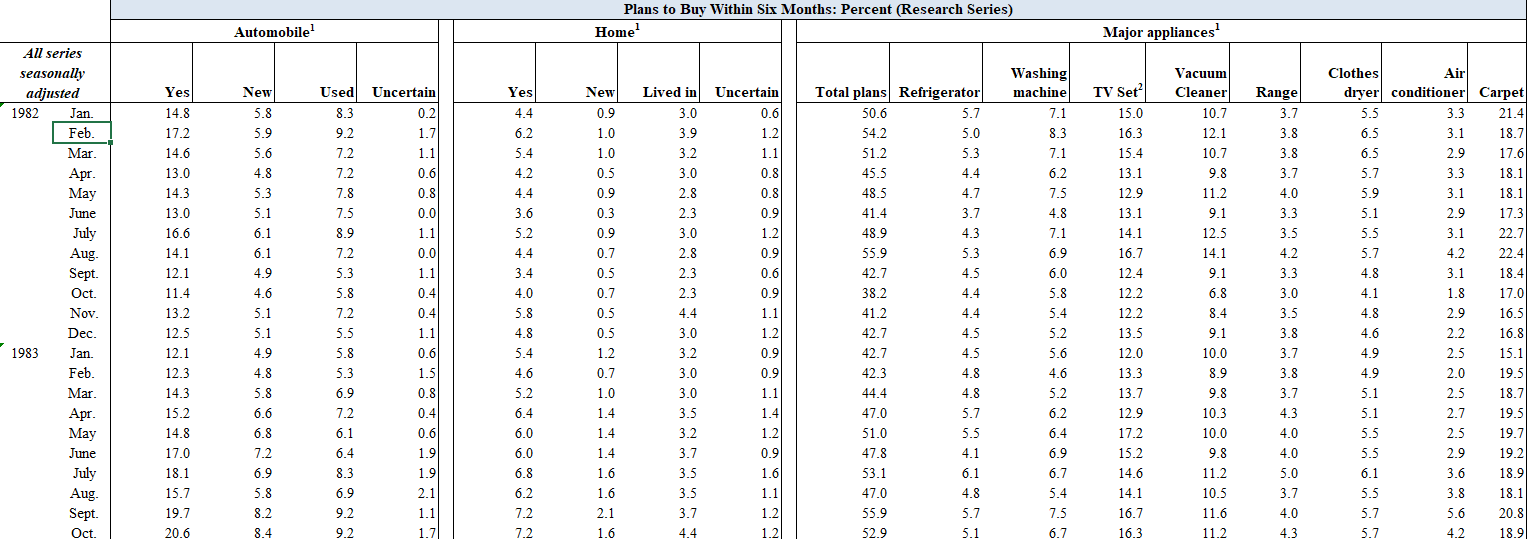
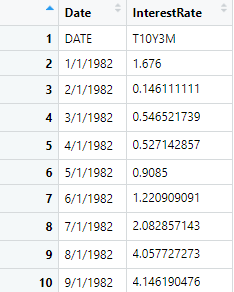
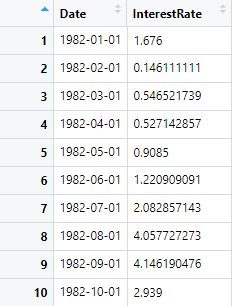


Figure 4 – Original format of consumer confidence dataset

**Explanatory Data Analysis**

I started my analysis with the yield curve data. After reading in the data into R, I realized that the date was in month-day-year format and had to be changed to year-month-day format for R to read in as date. Also, the date was in mm/dd/yyyy format and it was separated by forward slash and it had to be separated by hyphen to be in a constant format with the other variables. So, using format function I was able change the data into the right format as displayed below.

 A close up of a logo

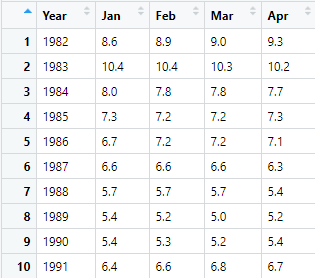
Description automatically generated 

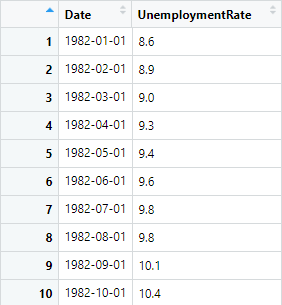
After cleaning the yield curve dataset, I started to clean the GDP output dataset. After reading the data in R, I had to do the same cleaning procedures as the yield curve because both datasets were extracted from the same source and had the same format. I also realized that the GDP dataset was recorded in quarters, while the rest of my variables were in a monthly format. I only had observations from January, April, July and October. To get a monthly data, I used cubic splines interpolation. I used the spline function in R to convert my data; this function inputs n quarterly data points which means we have n-1 spaces between them. Across each space, a unique cubic polynomial is drawn to connect two points; this is called piecewise polynomial function. For our connection between the line to be smooth, we force our first and second derivatives to be continuous and equal at each side. After all the requirements are met, we get a linear system that can be solved for n-1 coefficients to get weekly, monthly or yearly values. The transformation of the dataset can be seen below.

 A close up of a logo

Description automatically generated 

Lastly, I had to clean the unemployment rates dataset to have all my data ready to combine and analyze. After reading the data into R, I found that the monthly data was in column format with each column representing a month named “Jan”, “Feb” and so on, and each row representing a year. I had to transform the columns into rows. I used the “pivot\_longer” function in R to perform that. Then I had to convert character “Jan” to yyyy-mm-dd. After several tries to convert character (Jan 1982) to a date (1982-01-01), I had to change my strategy to move on. I decided to assign the date column from the unemployment dataset to the date column in yield curve dataset. The final dataset transformation is shown below. Now all my datasets were ready to merge, and I was could start the analysis.

 A close up of a logo

Description automatically generated 

After merging all my datasets, I decided to investigate the relationship between the variables to further understand my data. This would help me determine which modeling technique to choose for this problem. I used the ggpairs function to get a matrix plot, which is shown in figure 5. The relationship between interest rate (which is interest rate of bonds from the yield curve) and the unemployment rate had a correlation coefficient of 0.618 and the relationship appears to be quite linear. However, the relationship between the interest rate and GDP appeared non-linear and the correlation coefficient was -0.109. Likewise, the relationship between the GDP and the unemployment rate also appeared non-linear and the correlation coefficient between the variables was -0.285. This was expected because that’s how each variable behaved throughout history. Due to presence of both linear and non-linear variables, I decided to implement generalized additive models.

A close up of a map

Description automatically generated

Figure 5 – Matrix Plot of all the variables in the dataset

**Generalized Additive Model**

Before fitting a generalized additive model, I performed cross-validation to split my data into training and testing datasets. I used 75-25 split for my testing and testing datasets, respectively. After splitting my data, I fitted a generalized additive regression model on my training dataset using default k values (k = 10). I got p-values of 0.01403 and 2e-16 for GDP and unemployment rate, respectively, which meant that none of the variables needed to be eliminated from the model. The median for deviance residuals was 0.03797. This seemed promising until I got predictions and got a test mean squared error of 0.74 which indicated that this model leads to test predictions that are within around 0.86 of the true median. It was quite prominent that I was overfitting my model because it is a high error rate. This led me to try other modeling techniques such as random forest to get predictions that would get lower test mean squared error.

**Random Forest**

After getting a high test mean squared error with the generalized additive models, I decided to use random forest regression tree method in hopes of getting an improved result. Using the randomForest function from the randomForest library, I fit my model. I used default number of predictor (m = ) and trees (500). The plot shows that the error is lowest around 50 tress which is approximately 0.10 (Figure 6). After getting the predictions, the test mean squared error attained was 0.12. This indicated that this model leads to test predictions that are within around 0.347 of the true median. This was significantly smaller than the error rate obtained using generalized additive model. However, I also wanted to build a model using classification method to see if there was a better performance. Unemployment rate was the most important variable because highly correlated with the response.

A screenshot of a cell phone

Description automatically generated

Figure 6 – Error rate vs number of trees graph

Finally, to use the classification approach, I had to create an indicator variable for the response. I created IRIV variable as a factor representing 1 for negative interest rate values and 0 for positive interest rate values. After fitting the model, I got misclassification rate of 0.026 and overall accuracy of 0.974. This seemed very promising, but it was highly overfitting the data and there was extremely low flexibility. This was because the majority of my indicator variable values were 0s, and the model was being trained on them. Although these models produced some results that provide some information about this problem, these are far from explaining real world data. To be able to make a prediction of a specific timeline of when a recession might occur, we need to use forecasting. Therefore, these results so far are not applicable to the original problem. This is also because there were many general assumptions made and there are only two predictor variables. If there are more variables added to the model which were eliminated due to time constraint, a more flexible model can be created that can explain variation closer to the original problem. Even if there were not any concrete results found, I believe it is an ambitious start to get to the right solutions.

**Conclusion**

This project was a major learning curve for me because it was my first time taking on a project all by myself. Initially, I learned that theoretically there are so many solutions that seem reassuring, but when we try to implement them, there are so many complications that make it a lot harder to stay on the same path. I’m very passionate about economics and after doing some brief research, I got a general idea for this problem which was to use variables that provided precise signs and indications to get a prediction. After getting the datasets and cleaning them, I combined them for analysis. After understanding the relationship between the variables, I used machine learning techniques such as Generalized Additive Models and Random Forrest. Although I obtained some results, they were different than what I was expecting because of some assumptions made and absence of more variables in the model.

This project has helped me improve my data cleaning and analytics skills. Specially, working with different datasets has enhanced data formatting knowledge and helped me recognize the complications that arise in merging different datasets. It also improved my critical thinking skills because I have come to realize how complex a real-world problem is when it seems quite simple. I have understood that it is very crucial to create a road map on how to solve a complex problem and consult a superior each way through. Finally, I was also able to implement machine learning algorithms which increased my knowledge in this field. Although I just started to learn about machine learning, this project has advanced my passion to learn more and do more complex problems in the future.

**Bibliography**

Chappelow, Jim. “Gross Domestic Product – GDP.” *Investopedia*, Investopedia, 29 Apr. 2020, www.investopedia.com/terms/g/gdp.asp.

Chappelow, Jim. “Why Does Unemployment Rise During a Recession?” *Investopedia*, Investopedia, 21 Apr. 2020, www.investopedia.com/ask/answers/032515/why-does-unemployment-tend-rise-during-recession.asp.

Chen, James. “Yield Curve.” *Investopedia*, Investopedia, 2 Mar. 2020, www.investopedia.com/terms/y/yieldcurve.asp.

“Data: The Conference Board.” *Data | The Conference Board*, www.conference-board.org/data/.

Mda, et al. “Mental Health Outcomes in Times of Economic Recession: a Systematic Literature Review.” *BMC Public Health*, BioMed Central, 1 Jan. 1970, doi.org/10.1186/s12889-016-2720-y.

“More than 70% of Economists Think a US Recession Will Strike by the End of 2021 | Markets Insider.” *Business Insider*, Business Insider, markets.businessinsider.com/news/stocks/next-recession-trump-survey-shows-most-economists-predict-downturn-2021-2019-8-1028456445.

Rappeport, Alan, and Jeanna Smialek. “I.M.F. Predicts Worst Downturn Since the Great Depression.” *The New York Times*, The New York Times, 14 Apr. 2020, www.nytimes.com/2020/04/14/us/politics/coronavirus-economy-recession-depression.html.

Frasquilho, D., Matos, M., Salonna, F. et al. Mental health outcomes in times of economic recession: a systematic literature review. BMC Public Health 16, 115 (2015).

**Appendix**

Final Project

Hussam Taj

4/11/2020

## -- Attaching packages --------

## v ggplot2 3.2.1 v purrr 0.3.3  
## v tibble 2.1.3 v dplyr 0.8.3  
## v tidyr 1.0.0 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.4.0

## -- Conflicts -----------------  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

## Warning: package 'zoo' was built under R version 3.6.3

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

##   
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':  
##   
## date

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

##   
## Attaching package: 'GGally'

## The following object is masked from 'package:dplyr':  
##   
## nasa

## Warning: package 'gam' was built under R version 3.6.3

## Loading required package: splines

## Loading required package: foreach

##   
## Attaching package: 'foreach'

## The following objects are masked from 'package:purrr':  
##   
## accumulate, when

## Loaded gam 1.16.1

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:dplyr':  
##   
## combine

## The following object is masked from 'package:ggplot2':  
##   
## margin

Reading in the data

YC <- read\_csv("data/Yeild Curve.csv", col\_names = c("Date", "InterestRate"))

## Parsed with column specification:  
## cols(  
## Date = col\_character(),  
## InterestRate = col\_character()  
## )

UnE <- read\_csv("data/Unemployment.csv")

## Parsed with column specification:  
## cols(  
## Year = col\_double(),  
## Jan = col\_double(),  
## Feb = col\_double(),  
## Mar = col\_double(),  
## Apr = col\_double(),  
## May = col\_double(),  
## Jun = col\_double(),  
## Jul = col\_double(),  
## Aug = col\_double(),  
## Sep = col\_double(),  
## Oct = col\_double(),  
## Nov = col\_double(),  
## Dec = col\_double()  
## )

GDP <- read\_csv("data/GDPPOT.csv", col\_names = c("Date", "GDP"))

## Parsed with column specification:  
## cols(  
## Date = col\_character(),  
## GDP = col\_character()  
## )

Formatting the Yield Curve Dataset

YC <- YC[-1,]  
YC$Date <- format(strptime(YC$Date, format = "%m/%d/%Y"), "%Y-%m-%d")  
YC$Date <- as.Date(YC$Date)  
YC <- as.data.frame(YC)

Formatting GDP Output Data

GDP <- GDP[-1,]  
GDP$Date <- format(strptime(GDP$Date, format = "%m/%d/%Y"), "%Y-%m-%d")  
GDP$Date <- as.Date(GDP$Date)  
monthly <- seq(GDP$Date[1], tail(GDP$Date,1), by="month")  
GDP <- data.frame(Date = monthly, GDP = spline(GDP, method="fmm", xout = monthly)$y)

Formatiing Unemployment Dataset

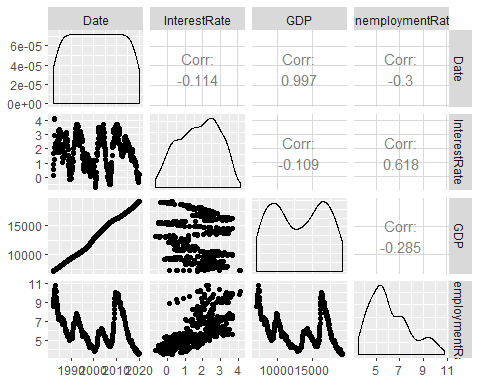
UnE <- UnE %>%   
 pivot\_longer(c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov", "Dec"), names\_to = "Months", values\_to = "UnemploymentRate")  
UnE <- as.data.frame(UnE)  
UnE <- UnE[-c(458:468), ]  
UnE$Months <- match(UnE$Months,month.abb)  
i1 <- grepl("^[0-9]$", UnE$Months)  
UnE$Months[i1] <- paste0("0", UnE$Months[i1])  
UnE <- unite(UnE, "Date", c(Year, Months), sep = "-", remove = TRUE, na.rm = FALSE)  
UnE$Date <- YC$Date

Megrging the datasets

Data <- merge(YC, GDP, by = "Date")  
Data <- merge(Data, UnE, by = "Date")

Matrix Plot

Data$InterestRate <- as.numeric(Data$InterestRate)  
ggpairs(Data)



Cross-validation

set.seed(39)  
Data <- Data %>% mutate(IRIV = ifelse(InterestRate >= 0, 0,1))  
Data$IRIV <- as.factor(Data$IRIV)  
Data <- mutate(Data, id = row\_number())  
train <- sample\_frac(Data, .75)  
test <- anti\_join(Data, train, by = "id")  
train <- train[c(-10)]  
test <- test[c(-10)]  
Data <- Data[ , !(names(Data) %in% c("id"))]  
train <- train[ , !(names(train) %in% c("id"))]  
test <- test[ , !(names(test) %in% c("id"))]

Fitting Generalized additive model

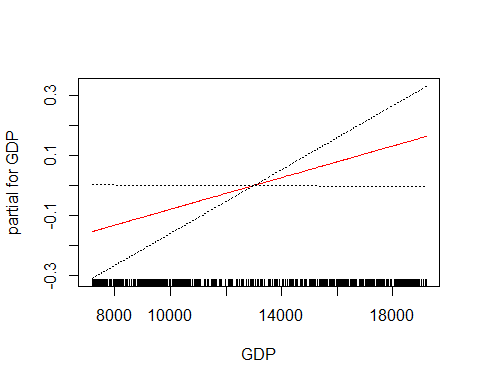
gam <- gam(InterestRate ~ GDP + UnemploymentRate, data = train)

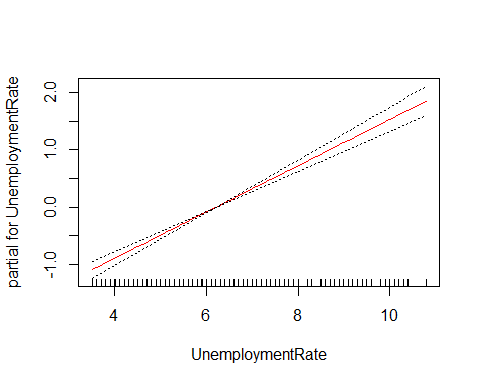
## Warning in model.matrix.default(mt, mf, contrasts): non-list contrasts argument  
## ignored

summary(gam)

##   
## Call: gam(formula = InterestRate ~ GDP + UnemploymentRate, data = train)  
## Deviance Residuals:  
## Min 1Q Median 3Q Max   
## -2.55386 -0.68649 0.03797 0.67818 2.11110   
##   
## (Dispersion Parameter for gaussian family taken to be 0.7721)  
##   
## Null Deviance: 428.0004 on 342 degrees of freedom  
## Residual Deviance: 262.5038 on 340 degrees of freedom  
## AIC: 889.6514   
##   
## Number of Local Scoring Iterations: 2   
##   
## Anova for Parametric Effects  
## Df Sum Sq Mean Sq F value Pr(>F)   
## GDP 1 4.708 4.708 6.0978 0.01403 \*   
## UnemploymentRate 1 160.789 160.789 208.2566 < 2e-16 \*\*\*  
## Residuals 340 262.504 0.772   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

plot.Gam(gam, se = TRUE, col = "red")





pred <- predict(gam, newdata = test)  
mean((pred - test$InterestRate)^2)

## [1] 0.7380704

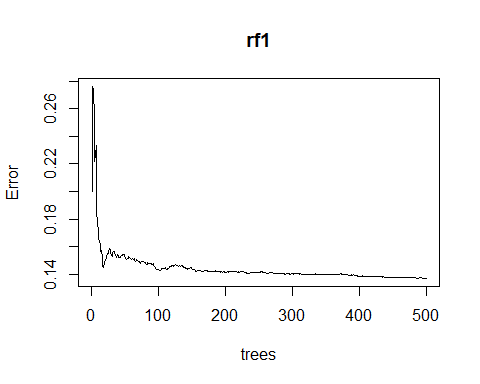
Indicates that this model leads to test predictions that are within around 0.8591102 of the true median.

Fitting Random Forest regression tree method

rf1 <- randomForest(InterestRate ~ GDP + UnemploymentRate,   
 data = train,  
 importance = TRUE)  
pred\_rf1 <- predict(rf1, newdata = test)  
summary(rf1)

## Length Class Mode   
## call 4 -none- call   
## type 1 -none- character  
## predicted 343 -none- numeric   
## mse 500 -none- numeric   
## rsq 500 -none- numeric   
## oob.times 343 -none- numeric   
## importance 4 -none- numeric   
## importanceSD 2 -none- numeric   
## localImportance 0 -none- NULL   
## proximity 0 -none- NULL   
## ntree 1 -none- numeric   
## mtry 1 -none- numeric   
## forest 11 -none- list   
## coefs 0 -none- NULL   
## y 343 -none- numeric   
## test 0 -none- NULL   
## inbag 0 -none- NULL   
## terms 3 terms call

plot(rf1)



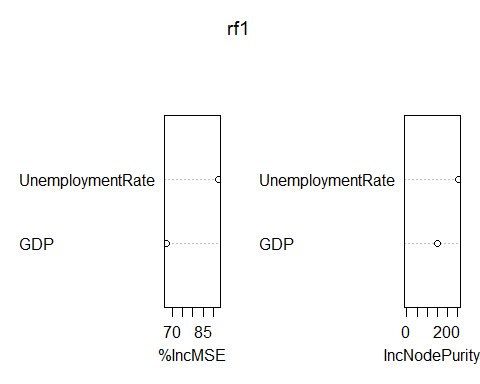
mean((pred\_rf1 - test$InterestRate)^2)

## [1] 0.1212357

importance(rf1)

## %IncMSE IncNodePurity  
## GDP 67.51874 150.4242  
## UnemploymentRate 92.47268 254.0372

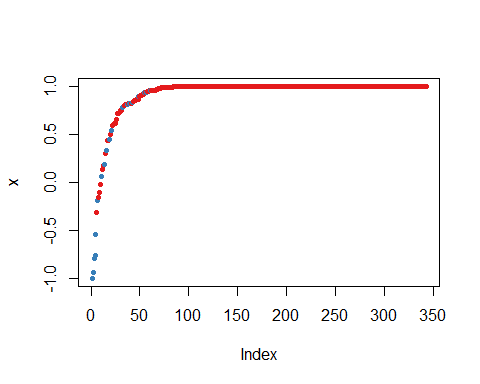
varImpPlot(rf1)

 Indicates that this model leads to test predictions that are within around 0.3474779 of the true median.

Fitting Random Forest Classification Tree

rf <- randomForest(IRIV ~ GDP + UnemploymentRate,   
 data = train,  
 importance = TRUE)  
pred\_rf <- predict(rf, newdata = test)  
tb <- table(pred\_rf, test$IRIV)  
mcr <- mean(pred\_rf != test$IRIV)  
plot(margin(rf, test$IRIV))

## Warning in RColorBrewer::brewer.pal(nlevs, "Set1"): minimal value for n is 3, returning requested palette with 3 different levels



accuracy <- (sum(diag(tb)))/sum(tb)

The model has a high overall accuracy that is 0.9736842.