UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

ASSIGNMENT 4

This assignment has two parts. In the first part, you will develop a number of time series models to understand the loans issued by LendingClub. Your main tasks include visualizing data and developing statistical models to help LendingClub management understand better the changes in the characteristics of loans issued in NY over time. You will develop models for the total dollar value of loans per capita, to guide LendingClub in its attempts to increase its market share in NY.

In the first part of the project, you will use two main data files: **lendingClub.csv** and **nyEcon.csv**. The first data file contains data for all the loans issued in the platform from June, 2007 to March, 2017. The data is aggregated to the state-month level. In peer-to-peer (P2P) lending platforms, consumers borrow from other consumers. The typical process is as follows: Consumers who are in need of borrowing money make a request by entering their personal information, including the SSN number, and the amount of money requested. If a request passes the initial checks, LendingClub's algorithm assigns a grade to the request, which translates into an interest rate (the higher the grade, the lower the interest rate). Other consumers who would like to invest into personal loans lend the money. For the most part, the lending is automated, so the P2P lending model is different from crowdfunding models. nyEcon.csv includes some economic indicators for NY for the same timeframe (from June, 2007 to March, 2017). You will be asked to join this dataset with the original dataset to use the variables in your models. You will also be asked to get the 2010 U.S. Census data for the population of each state (at the month level). Again, you will be asked to join this dataset.

Most business time series are not as good looking as some of the examples we used, or as some macroeconomic data. As you will see in the LendingClub data too, clear trends (incl. cycles) and seasonality may not exist. In the second part of this assignment, you will revisit a familiar dataset: retail sales. Remember that we looked at retail sales at the beginning of the course, when we did not have the tools for time series analysis. The large drop in retail sales after the 2008 crisis created a challenge in making predictions using a model trained in the past data. Unfortunately, a similar drop in retail sales is pending due to COVID-19, making this problem most timely. Now, using your new skills, you will revisit the retail sales data and apply the time series methods you have learned to make better predictions. In the second part of the project, you will use **retailSales.csv** which includes U.S. retail sales from January, 1992 to February, 2020.

Because this will be the last (required) assignment, I have added to it some elements I intended to include in Assignment 5. Because Assignment 5 is optional now, I would like you to gain some experience with data collection, formatting, and joining in this one. The tasks I have added are relatively simple, so, do not stress out about it but invest the time to work on this assignment.

Data Dictionaries

lendingClub.csv (All averages are the values averaged over the # of loans per state per month)

Variable	**Definition**
date	Monthly date
state	State abbreviation
Loans (avg and total)	The amount of loan issued in dollars
term (average)	The period in which the number of payments made are calculated (months)
intRate (average)	Interest rate on the loan (in percentages)
grade (average)	Loan grade assigned by the algorithm (A=1, B=2, C=3, D=4, E=5, F=6)
empLength (average)	Employment length of the borrower (in years)
annualInc (average)	The self-reported annual income provided by the borrower during registration
verifStatus (average)	Indicates if the income is verified by LendingClub (Verified=1, Not Verified=0)
homeOwner (average)	The home ownership status provided by the borrower during registration or obtained from the credit report (OWN=1, RENT OR OTHERWISE=0)
openAcc (average)	The number of open credit lines in the borrower's credit file
revolBal (average)	Total credit revolving balance
revolUtil (average)	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit
totalAcc (average)	The total number of credit lines currently open in the borrower's credit file
countOfLoans	The number of loans per month per state *(tally taken during aggregation)*

nyEcon.csv

Variable	**Definition**
date	Monthly date
NYCPI	Consumer price index in New York
NYUnemployment	Unemployment rate in New York -Seasonally adjusted
NYCondoPriceIdx	Condo price index in New York -Seasonally adjusted
NYSnapBenefits	Number of SNAP benefits recipients in New York

retailSales.csv

Variable	**Definition**
date	Monthly date
sales	U.S. retail sales in million dollars

usEcon.csv

Variable	**Definition**
date	Monthly date
income	Personal income (in billions of dollars) -Seasonally adjusted
unemployment	Unemployment rate -Seasonally adjusted

tenYearTreasury	10-Year treasury constant maturity minus 2-Year treasury constant maturity
CPI	Consumer price index
inflation	Inflation rate -Calculated from the consumer price index
vehicleSales	Total vehicle sale in the U.S. (in millions of units) -Seasonally adjusted
houseSales	New houses sold in the U.S. (in thousands) -Seasonally adjusted

Assignment Instructions - Part I (~60 points)

Predicting/forecasting the LendingClub loans

Before you start, load the following libraries in the order: *tidyverse, fpp3, plotly, skimr, lubridate*

1) (~10 points) Data processing

a) Load the LendingClub dataset into R and call it *tsLCOrg*.

b) Convert the dataset into a tsibble using date as index and state as key.

Hint: You might need lubridate's help for this.

c) Inspect and describe the data.

➔ *The data has 16 variables with 4943 observations. It is a time series data with covers the total loans among other things as an average value over each month. This data is divided state wise and each state is identified by a State code.*

d) Load the dataset with the NY Economy indicators.

Hint: You might need lubridate's help for this.

e) Visit the U.S. Census Bureau's data portal to download the population data for each state from the 2010 Census, and (i) add the population column to *tsLCOrg*. Then, (ii) calculate the loan amount per capita and add the new variable as *loansPerCapita*. (iii) Join it with the NY Economy data by date and state. Save the new tsibble as *tsLC*.

Hint: You might need to use the rowwise() function, and convert to tsibble again.

2) (~20 points) Exploratory analysis

a) Plot the loans per capita for the states within the top 10th percentile and bottom 10th percentile in terms of population. Compare the two plots and share your observations. What might be a (statistical) reason for the difference in variance?

➔ *We observe that the median Loans Per capita for the states with top 10th percentile in terms of population are close to each other and lower value as compared to the median loans per capita of the states in the bottom 10th percentile. This cannot be considered as an indicator of their borrowing preferences as the denominator population plays a role in the calculation of the loans per capita. We can say that among the less populated states North Dakota has an unusually high amount of loans as its median is quite high as compared to other readings.*

b) Create anomaly plots to compare the NY data with Massachusetts and Colorado. Use the STL decomposition and interquartile range to mark the anomalies. Compare the results. What are the differences across three states, and how do you explain them?

➔ *Based on the anomaly plots for the 3 states we see that there is an increase in loans per capita (LPC) between 2014 and 2016 in all the states. This increase is more than the expected value based on the trends and thus is being reported as anomalies. Since, the population considered for the analysis is just the 2010 census data the change in LPC is mainly due to the total loans. The increase in loans could be due to revised federal directives to allow financial institutions to lend more money and due to the rise in the non- bank lending companies. Beyond 2016, The LPC in NY behaves differently that CO and MA. While CO and MA reports reduction is total loans which is more than precedented, NY shows an increase anomaly. This could be due to the payday loan regulations which impacted the borrowing capacity of people in MA and CO while the real estate market kept growing in NY where more people wanted to own houses than rent them. Also people were willing to lend and invest in personal loans in NY*

c) Apply STL decomposition to the loan per capita in NY.

i) For the issued loans, identify/report the month in which the trend reverses.

➔ *In Oct 2015, the trend of the issued loans stops increasing. It flattens out and starts to decrease in April 2016.*

ii) What do you think is the reason for the change in trend in this month?

➔ *After the financial crisis in 2008 the loan markets were impacted. This changes slowly between 2011 and 2014 where people started borrowing again as the economy started growing. There was a lot of stimulus to make the economy grow faster and thus people were lending and borrowing more than usual. In 2015 this surge in loan activities started to level out and by April 2016 we can see that maybe it is lowering to go back to the normal levels of lending and borrowing.*

d) Create a seasonal plot and a seasonal subseries plot for NY. Share your observations. Do your observations change if you limit the data to the last three years?

➔ *The seasonal and seasonal subseries plots are essential tools in spotting the seasonality in the data and making inferences based on them. In case of the NY data we see that there is not much seasonality in the data over the years. Also, the subseries plot shows that the LPC values have increased every month and the mean values are almost the same irrespective of the months.*

When we consider only the past 3 years data, we can see some seasonality developing but not concrete. From the sub series plot we can see that the mean *values have been consistently higher in March, July and October and lower in May and September.*

e) Plot the autocorrelation function and partial autocorrelation function results for NY. What does the ACF plot tell you? What does the difference from the PCF plot tell?

➔ *The ACF plot is gradually decreasing with the highest spike with the data from closest lags. This indicates that the data has a positive trend which we can see from the data. The trend reverses at some point and thus the correlation also becomes negative beyond that point. There are no seasonal spikes in the ACF thus the data does not have clear seasonality. The PACF looks different from the ACF. Beyond the initial three spikes it has no significant spike, this is because the PACF deals with the residuals after removing the effects of the lags between the two points. This graph also indicates that there is no seasonality while there is some trend in the data.*

f) Create a lag plot for NY for the lags 1, 5, 10, 15, 20, 25. Discuss your observations.

➔ *The lag plot for NY shows us that the correlation is strong for the lower values of lags like lag 1 and lag 5 while the correlation gradually decreases as we increase the lag. This is in sync with what we observed from the ACF plot in the previous example. Also, we do not see the lag plots again showing strong correlation thus there is no seasonality but only a trend in the data.*

g) First, plot the loans per capita in NY over time. Then, create a fifth order moving average smoothing and plot the smoothed values on the actual loan data.

3) (~20 points) Modeling the loans issued in NY

a) Make a seasonal naive and drift forecast for NY data five years into the future, and display both models as visualizations. Discuss the results of these models. Do you think they capture the change in the amount of loans per capita? Why or why not?

➔ *Naïve and drift models work well when the time series follows a random walk or a simple increasing trend from the first observation to the last. But in the case of the LPC data it does not fall into this. Thus, we see that the range of the prediction is huge for 80 and 95 percentile values. Though these models tell us that the value would be in this range it is not very useful in our case.*

b) Build a time series regression using both the time trend and seasons, as well as other variables you can use to explain the loan issued per capita. Discuss the results of the regression, and what you can learn from the statistically significant coefficients.

➔ *The adjusted R squared value is 0.8949 which is high and a good indicator that the time series regression model can capture the variation in the data well. The statistically significant variables are the ones which have a low p-value. They are trend, average interest rate, the consumer price index in NY, the condo price index in NY and snap benefits to some extent. They tell us that seasons do not play much role in the time series and that the loans increase with decrease in interest rate. Also, the other economic indicators listed also have an impact on the loans per capita in NY.*

We cannot consider population, total loans, avg loans or the count of loans as variables in the model because these variables are highly correlated with each other and are directly involved in arriving at the value of LPC.

Hint: Note that you cannot use some of the variables to explain the loans per capita.

Hint: You might also need to remove the variables with any missing values.

c) Plot the fitted values from the model above and an alternative model excluding the time trend and seasons. Compare two plots and discuss your observations.

➔ *Firstly, we can see from the adjusted R squared and R squared values that the model with trend and seasons is slightly better than the model without them. Also, from the plot we can see that the line of fitted values is smoother when the trend and season are not included in the model. This can be explained using the variance bias tradeoff. When we include trend and seasons it increases the variance and thus reduces the bias while a model without trend and seasons increases the bias and reduces the variance.*

d) Create a predictive modeling plot using the model from (b) using two train/test splits. In the first split, use the data from 2014 and before for training, and in the second split, use the data from 2015 and before for training. Compare and discuss.

➔ *Based on the graph we can see that the model with less training data which is data before 2014 performs better than the data with more training data which is before 2015. We may feel that the model with more training data should perform better but in this case it is different. The data till 2014 sees a steady increasing trend, but in 2015 and beginning of 2016 the increase is steeper as the markets recover faster during this period. So a model training with this additional data predicts steeper increase than the model which sees only data till 2014. We must note that both models fail to predict the dip after 2016 as the markets stabilize.*

But because of the training characteristics the predictions made by the first model is better in terms of errors.

e) Check the residual diagnostics for the model from (b). Does it look fine? Discuss.

➔ *From the time series plot of residuals, we can see that the distribution of variation of the residuals is not uniform. We can see that the variation increases as we move ahead in time. This indicates heteroscedasticity in the residuals.*

The autocorrelation plot has spikes at lags 3, 7, 10, 15, 17. This means that the residuals themselves may be correlated and the accuracy of the prediction is impacted.

The histogram is right skewed slightly and though it is nearly normal distribution the predictions may be impacted

f) Build an ARIMA model using the same variables from (b) and using a grid search. What are the orders of the autoregressive model, differencing, and moving average model (p,d,q)? Which ones of the variables are significant? Are they the same as (b)?

➔ *The Grid Search can be done by not providing any pdq values to the ARIMA model. The model searches for the pdq values by itself and returns the most appropriate values. The values returned by the Grid search are (1,0,4).*

The variable significance can be calculated by calculating the p-value for each of the variables. Based on the p-value calculation we can see that the variables NY CPI, NY Unemployment and NY Condo price index are significant in explaining the data. This is different from the model 3(b) which has avg interest rate also as a significant variable and it does not consider unemployment as a significant variable.

g) Check the differencing suggested by the KPSS test. Does it align with the ARIMA model's differencing? *Answer the next question (h) only if your response is negative.*

➔ *On checking the differencing suggested by the Arima Grid search (which is 0 i.e. no differencing) we see that the p-value is low for KPSS test. This means that the Null hypothesis that the differencing is right must be rejected. Thus, we need to find the correct differencing level for this data.*

h) If KPSS suggests a different degree for differencing, repeat the grid search in ARIMA using the degree suggested by the KPSS test. What is the (p,d,q) of the new model?

   i) Compare the new model performance with the model from (f).

➔ *Based on unit root test we arrive at the value for ordinary differencing as 1. Thus we build a ARIMA model with d=1 while not providing p and q. This the GRID search returned the values (0,1,4) for the new model.*

We can say that this model performs better than the previous model based on the AIC and BIC values which are lower.

ii) What do you think is going on here? *(Research and)* discuss.

➔ *The p parameter controls the autoregressive part of the model. P=1 means that a lag of one period. The model will use one previous period to do the calculations. The d parameter is used to apply differencing to the data to convert it into stationary time series. When d=1 it means first differencing i.e. difference between the current time period and the previous time period. This action is equivalent to having p=1. Thus, when we provide d=1 to the ARIMA Grid search model based on the outcome of KPSS test, it no longer feels the need to have p=1. Thus, p is set to 0 when d is set to 1.*

Pro tip: You can run a constrained grid in ARIMA by presetting any of the parameters.

4) (~15 points) Predictive modeling of the loans issued in NY

a) Set the seed to 333 and split the data into training (earlier than March, 2016) and test sets (on and after March, 2016). Build and compare the performance of the following models. Based on RMSE, which model is the best forecasting model?

i) Time series regression with only trend and season

ii) Time series regression you built in 3(b)

iii) ARIMA grid search model without any other variables

iv) ARIMA grid search model you built in 3(f)

➔ *The model (iv) which is a ARIMA grid search with the variables we used in 3(f) turns out to be the best predictive model based on the RMSE values. Low RMSE value indicates better performance in predicting the test data. This is because while ARIMA accounts for the time component the variables explain the rest of the variation.*

b) Set the seed to 333 and split the data differently this time: training set (before April, 2016) and test set (on and after April, 2016). Build and compare the performance of the same models. Based on RMSE, which model is the best forecasting model now?

➔ *The model (i) which is the Time Series Regression model with only trend and season turns out to be the best model with the lowest RMSE value for the new Train and Test data.*

c) The only difference between the two sets of models (a) vs. (b) is that the second one uses one more month of data for training. How do you explain the resulting change?

➔ *The new one month of data i.e. April 2016 is where the trend of the time series reverses. This is the point where the loans reduce and thus when we give this additional piece of information to the model for training it performs much better with only trend and season. For ARIMA this new information is not captured as directly as it has converted it into a stationary time series before modelling. Thus,* *the RMSE for ARIMA is improved a bit but not as much as the model built using trend and season.*

Assignment Instructions - Part II (~40 points)

Predicting/forecasting the U.S. retail sales

1) **Preparation and exploration**
   a) Load the U.S. retail sales data into R and call it *tsRetail*.
   b) Convert the dataset into a tsibble using date as index.
   Hint: You might need lubridate's help for this.
   c) Plot the retail sales over time for (i) the full data, and for (ii) a subset starting from 2010. Share your observations.
   ➔ *(i) Based on the time series plot of the full data, we can see that there is a clear trend and seasonality in the data. Also, there is a sudden dip in the retail sales in the year 2008 which could be attributed to the recession and slowdown during that period.*
   (ii) For the plot of the subset starting from 2010, we can see the seasonality more clearly with the peaks located around the months of November and December which is the holiday season. Also, the black Friday and cyber Monday sales could cause this peak.

2) **Understanding the time series**
   a) Create a seasonal, and a seasonal subseries plot for the subset data starting from 2015.
   ➔ *We can observe that there is a seasonality component in this data based on these graphs*
   b) Create an STL decomposition plot (i) for the full data, and (ii) for a subset of the data between 2005 and 2015 (both bounds are inclusive). Compare and discuss.
   ➔ *(i) The STL decomposition plot for the full data suggests that most of the variation in the data is explained by the trend of the time series, some of it is explained by the seasons and the errors/noise in the data is least significant. This is because the full dataset has a clear indication of trend over a long period of time with a dip in the middle for 2008 which is also captured by trend.*
   (ii) The STL decomposition of the subset data suggests that the variation is explained by both trend and seasonality almost equally. This is because as the duration is less the seasonality in the data also becomes significant. Noise/error still plays only a negligible role in the data.
   c) Create an autocorrelation function plot and a partial autocorrelation function plot. What does the ACF plot tell you? What about the difference between the ACF and PCF plots?
   ➔ *ACF plot shows a gradually decreasing value of the correlation as the lags increase. This indicates an overall positive trend in the data series. The*

"scalloped" shape is seen because of the spikes at regular interval. These spikes indicate the seasonality in the data.

The PACF looks different as it is the correlation of the residuals after accounting for the correlation between the lags. The PACF has 3-4 spikes initially then it has no other significant spikes. Also, among the insignificant ones there is a pattern which indicates seasonality. It also shows the peaks and troughs in the data with the positive and negative correlations.

d) Plot the seasonally adjusted sales superimposed on the actual sales data. Use appropriate coloring to make both the seasonally adjusted and actual values visible.

e) Create a second order moving average smoothing and plot the smoothed values on the actual sales data. Use appropriate coloring to make both the smoothed values and actual sales data visible. What would you change in the moving average plot to achieve a plot similar to the one you created in 2(d)? Apply the change and share the outcome.

➔ *Moving average smoothing leads to a smooth curve which has less variance and more bias than the actual plot. This helps to adjust for the seasonal variations in the data and arrive at a smoother line. The moving average of order 2 means that it considers two adjacent values and takes and average to smooth the curve. But from the plot we can see that there is still a lot of variation in the curve, this is because the seasons are repeated every 12 months. If we take a moving average of order 12, we can achieve an almost smooth line. But to match the one with the previous seasonally adjusted graph we can try a moving average of order 7 and achieve the plot shown in the R notebook output.*

3) Modeling and analysis of the time series

a) Build a time series regression using the time trend and seasons. Report your output, and provide a short discussion of the results (e.g.coefficients). Check the residual diagnostics.

Btw, isn't that an impressive R-squared, achieved by using only the trend and seasons?

➔ *We achieve an R-squared of 0.9759 which means that the model performs well in terms of explaining the variation in the data. Based on the p-values we can see that both trend and season are significant variables.*

The residual diagnostics show some problems. Though the variation in residuals is equally distributed in most part it tends to increase a bit slightly towards the end. The ACF plot shows that there is a correlation in the residuals which is not desired and may impact model performance. Also, the residuals are not normally distributed.

b) Build an ARIMA model. Report your output, and provide a short discussion of the results. Check the residual diagnostics. How do you think the ARIMA model compares with the regression from 3(a)? What do the coefficients tell you in this case?

➔ *The parameters returned from the Grid search ARIMA model are pdq(1,0,1) and PDQ(2,1,2). Thus, the ARIMA calculates the optimal parameters for the seasonal and non-seasonal components.*
The residual diagnostics show improvement over the time series regression. Though there is a slight heteroscedasticity problem in the residuals, they are less correlated and are almost normally distributed. Thus, it has better performance.
The coefficients of the model output tell us the effect of the previous value of sales or the error on predicting the future value. It also tells us the effect of the previous value of sales or error from the same season on prediction the future values of the same season.

c) Run unit root tests to determine the amount of ordinary and seasonal differencing needed. Apply the suggested differencing, and run a KPSS test to check whether the KPSS test gives a pass on the stationarity of time series after the differencing applied. Finally, create two PACF plots for before vs. after differencing. Compare and discuss.

➔ *On running the unit root test we see that the data needs 1 level or ordinary differencing and 1 level of seasonal differencing to be converted into stationary time series. On differencing the data and running the KPSS test we see that the p-value is high enough to accept the null hypothesis. Thus, the test gives a pass on the stationarity of the time series after the differencing is applied.*
The PACF plots before and after the differencing shows much more stationarity in the data after differencing. It still shows some peaks and throughs but the value of these peaks and troughs have reduced from 0.9 to 0.4 which indicates a significant decrease in the correlation.

Hint: In some cases, if the seasonality is strong, applying the seasonal differencing first (and ordinary differences next) may help achieve a more stationary time series.
Hint: In inspecting PACF plots, don't forget to pay attention to the correlation values.

d) Set the seed to 333 and split the dataset into a training set (before 2011), and a test set (2011 and after). Test and compare the *ten-year* forecasting performance of a time series regression with trend and season, and an ARIMA model that uses a grid search. Which one is the better model for forecasting retail sales? Why?

➔ *The time series regression model with trend and season is the better model as it has the lower RMSE value.*

e) Set the seed to 333 and split the dataset into a training set (before 2016), and a test set (2016 and after). Test and compare the *five-year* forecasting performance of a time series regression with trend and season, and an ARIMA model that uses a grid search. Which one is the better model for forecasting retail sales now? Why?

➔ *With the additional training data, the ARIMA model with grid search is the better model based on the lower RMSE value.*

f) If your answers are different in 3(d) and 3(e), how do you explain the difference?
→ *ARIMA model converts the data into stationary time series and thus more the data available in training the better the model performance. While the additional data does not help the TSLM model much as the trend is already much clearer with the previous training set. Also, the new test set is smaller making the predictions to be less into the future thus helping ARIMA make less errors. Thus, ARIMA performs better than the TSLM in the second scenario*

4) **Checking for anomalies and reporting the results**

a) Run the anomaly detection algorithm GESD following the STL decomposition, as implemented in the anomalize library (using defaults). Report the plot and the list of observations marked as anomalies as a table. Is there an observation in the list that is different from others? If so, how do you explain *the outlier in the list of anomalies*?

- ➔ *There is an outlier in the list of anomalies which is the anomaly detected in Feb 2019. The other anomalies can be attributed to the recession and market crash in 2008. But the Feb 2019 is something else. Based on a report by CNBC "U.S. retail sales unexpectedly fell in February; the latest sign economic growth has shifted into low gear as stimulus from $1.5 trillion in tax cuts and increased government spending fades. The surprise drop in sales in February could partly reflect delays in processing tax refunds in the middle of the month. Tax refunds have also been smaller on average compared to prior years following the revamping of the tax code in January 2018. Cold and wet weather could also have hurt sales." This could also be attributed to the government shutdown that occurred around the period. (source:* https://www.cnbc.com/2019/04/01/retail-sales-february-2019.html)

b) For the models created in 3(d) and 3(e), create plots in which the actual values are shown against the predictions from the time series regression and ARIMA models. You will create two plots in total, and in each of the plots there will be *three lines (actual data, predictions from the regression, predictions from the ARIMA)*. Use appropriate coloring to make the actual, regression, and ARIMA model lines distinguishable. *In both plots*, limit the portion of data visible in plots to the last ten years starting from 2010.

Bonus questions (~2-3 points each):

(1) Quite a number of you seem to be interested in the analysis of financial time series. Here is an open ended question. See **usEcon.csv** and the data dictionary at the beginning of the assignment. Can you improve further the models you have built for the U.S. retail sales?
→ *Based on the new data available in the usEcon data set we can see that variables such as inflation, CPI, unemployment rate and income would have an impact on the purchasing power of the individuals. Thus, these variables can be used to explain the US retail sales variations.*

Thus, I created two models of TSLM and ARIMA by including these 4 variables. The model was trained on data till 2016 and rest was used as test data.
Based on the performance measure RMSE (decreased to 8782.29*) we can see that the ARIMA Grid Search model with the new variables performs much better now with additional data.*
(2) How can you incorporate the learning from the 2008 crisis in predicting future retail sales figures? You may also want to make predictions using your best model for March sales, and compare your results when the figures are released by the Census at 8:30am on April 15.
→
Due to the pandemic, the people will no longer be inclined towards buying non-essential goods. Also, the sale of essential goods would be low as many people have already stocked up their supplies ahead of time. Thus, the value of sales in March would see a steep decline.
Due to the slowdown in 2008 the US retail sales dropped by 2.8% in Oct 2008.
Considering this we can reduce/adjust the predicted value of sales by thrice this amount as the pandemic is much widespread and will have more effect than the 2008 event.
Since we do not have the values for other parameters in the data set and cannot verify the accuracy of external values, I did not implement a model to predict the value.
Also see “How the Virus Transformed the Way Americans Spend Their Money” at https://www.nytimes.com/interactive/2020/04/11/business/economy/coronavirus-us-economy-spending.html for early indicators of retail sales based on credit and debit card transactions.