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.

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 see a greater variance in loans per capita for the bottom 10th percentile states. This could be because since the population is constant throughout this time series and since in the bottom 10th percentile that number is lower, loans per capita fraction (total loans/population) would increase by a bigger margin for a higher total loans than it would for states in the top 10th percentile.

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?

The anomalies are closer to the actual data points in case of NY as compared to the other two states. The reason could be due to the populations of the latter two states being significantly lower than that of NY, causing the fractional change in loans per capita to increase more for them.

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.
   We see the trend change in December of every year.

   ii) What do you think is the reason for the change in trend in this month?
   This would be owing to the major holiday season where offices would be closed and people would be occupied with the festivities as well.

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?
We see an increase in variance in the latter years as compared to the former. There are clear spikes and troughs in the graphs for 2014 and beyond, but before that, we can see no significant variation. So yes, the observations do change if the data is limited to last three years.

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?
From the ACF plot, we can clearly see that the data has a trend since the autocorrelations are large and positive and also a scalloped shape due to seasonality. The correlation is highly significant, that is, loans per capita for the lower lag values is higher and this decreases as the lags increase. We can say that loans per capita values are dependent on the past values atleast until the lag value of 28.
From the PACF graph, it can be deduced that there is only statistically significant correlation in the smaller lags. This pattern indicates an autoregressive term of order 3.

f) Create a lag plot for NY for the lags 1, 5, 10, 15, 20, 25. Discuss your observations.
Linear patterns can be observed for lags 1 and 5, meaning a strong positive autocorrelation is present in data. In the graphs of the remaining lag values, we can observe seasonality and outliers.

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?
From the graphs, we can see that they do not completely capture the change in loans per capita. Naïve forecast does not capture trend and drift fails to capture seasonality, both of which are present in our data and are significant.

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.
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.
The variables chosen are trend() + season() + avgTerm + avgIntRate + avgAnnualInc + avgVerifStatus + NYCPI + NYUnemployment and achieved an R squared value of 0.9. The statistically significant coefficients here were the economic indicators in NY, specifically NYCPI and NYUnemployment, which would highlight the spending capabilities of consumers. The rest focused on the loan details and personal consumer characteristics, depending on which loans would be granted.

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.
Trend and seasonality play a very crucial role in our current data. As seen from the graphs, the model performs significantly worse when those two factors are not included. The above model is a better fit as well, as observed.

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.
The first model is able to predict future data much better than the second one. This is most likely down to the fact that the second model is encapsulates the highly increased trend observed in 2015 for its prediction, because of which it continues to predict a continuing trend in the test data. The model is unable to account for the sharp unexpected reversal in trend in 2016 and so predicts a continuing rise in trend. The first model though does not have this significant rise in trend of 2015 and predicts a deliberate rise in trend, similar to that of 2014, which is why it is able to better fit the data of the latter years, when the trend reverses in 2016.

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

The autocorrelation here is significant for a few lag values, meaning the forecasts are not that good. While the residuals are normally distributed at 0, the curve still continues a bit farther to the right to call it fully normal. The variance in the residuals increases sharply in the latter years and there are lot of outliers in those years, predictably so, as there is a very sharp increase in trend in data as well, making it harder for the model to capture that. Overall, the diagnostics are not fine and the model does not forecast well.

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 orders of (p,d,q) are (2,0,3). The standard errors of avgTerm, avgIntrate, NYCPI and NYUnemployment are the lowest and hence, more significant in explaining the model. The only difference here is avgTerm was not as significant in the earlier model.

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.*
No.

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?
(p,d,q) of the new model is (2,1,3)
    i) Compare the new model performance with the model from (f).
    The AIC is lower for the newer model at 28.08 as compared to 34.76 of the previous one and BIC has reduced to 63.99 from 68, which means the new model performs better.
    ii) What do you think is going on here? *(Research and)* discuss.

Pro tip: You can run a constrained grid in ARIMA by presetting any of the parameters.
When we applied differencing to the dataset, the trend and seasonality was eliminated leading to a stationary series, because of which the new ARIMA model was able to perform better and achieve lower AIC and BIC values.

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)

Based on RMSE, the best forecasting model is model 4, ARIMA grid search with all variables. They explain the data better than the other models and the variables are statistically significant.

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?
Now, the best forecasting model is model 1, Time series regression with only trend and season.

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?
This one difference is highly significant because this is the month that sees the trend reverse very sharply and the models are able to account for it. Earlier, this change in trend was not being captured by the training data and hence, the error rates were higher.

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.
The full data sees a slight reversal in trend in 2008-09, owing to the recession. In the subset though, the trend is ever rising and does not reverse. The seasonality, however, is constant for both plots.

2) Understanding the time series

a) Create a seasonal, and a seasonal subseries plot for the subset data starting from 2015.

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.

The seasonality of both plots remains about the same. There is a clear reversal of trend observed in the subset owing to the recession and the variance in data is not as significant as the full 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?
There is a significant correlation between data and its previous values (lags) owing to its large and positive autocorrelations. Furthermore, there is a clear pattern in seasonality observed as the autocorrelations spike after every 12th lag value.
In the PACF graph, there is statistically significant correlation almost throughout the plot. This pattern indicates an autoregressive term of high order value.

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.
Increasing the number of points used for the average calculation would help achieve a plot similar to the one in 2.d., meaning we have to increase the number of orders of moving average smoothing. Taking a 12th order moving average gives us a similar plot, a completely smoothened curve without seasonality.

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?*
All coefficients are significant in this model except the season value of 2 (February). The R-squared is very high and surprising, considering the variables used. This data however, is heavily seasonal and showcases trend throughout and that has been sufficiently explained by the model.
The variance in residuals for the year 2008 is quite high as compared to the rest of the years. The ACF plots are beyond the limits, meaning that the residuals are not noise and are significant and not captured by the model. Residual histogram is not completely 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?
Most of the variables used in the model are significant, owing to the low standard error. The residual plot has also improved on the previous one, as the variance in the residuals have reduced. Most of the autocorrelations are now within the limit, which was not the case previously. The residual histogram is now forming a bell-shaped curve, indicating a more normal distribution.

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.
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.
The PACF plot before differencing shows clear trend and seasonality owing to the autocorrelations spiking beyond the limits and at certain intervals. After differencing, most autocorrelations fall below the limit to be considered insignificant as trend and seasonality have been removed from data.

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?
Time series regression is a much better model for forecasting sales, as this has a much lower RMSE value than ARIMA. The significance of trend and seasonality in the data is captured by this model more than the latter does and therefore is able to forecast better, as sales would depend more on those two factors.

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?
The regression model is still better at forecasting but curiously, the RMSE value has increased for that but decreased for the ARIMA model.

f) If your answers are different in 3(d) and 3(e), how do you explain the difference?

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 are two hotspots for anomalies in the data, first being in 2008 and the other being in 2019. While the former is part of data that's on a downward trend, the one in 2019 is not following the upward trend pattern. This could be differentiated as an outlier amongst the other anomalies.

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?
The model used here is the time series regression with trend, seasonality, income, unemployment and CPI. This gives us a resulting R-squared value of 99.8%
This makes it very clear that adding the economic indicators of US has vastly improved the model as it explains most of the 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.
It is obvious that the CPI value would decrease, and unemployment and inflation rates would increase during a crisis situation such as the 2008 crash. So identifying such crisis moments and adjusting the model parameters to reflect the changes expected would help in forecasting retail sales in a much more accurate manner.

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.