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 is organized monthly, however each state do not have data present for each month like CO data is available for the months of June and July for the year 2007 but not for August. Further on inspecting the data shows very different number of loans and its amount issued for each state. Employment length of the borrowers is normally distributed but the total and average loans is skewed.

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?

The states in the bottom 10 percentile population have a larger variance than that in the top 10 percentile. This is possible due to the varied amount of population in the bottom 10 percentile states and very close population figures in the top 10 percentile states.

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?

Colorado captures more variance in seasonality than the other two states. In terms of trend, all 3 states have a similar plot, increasing after 2012. There's a random rise or drop in the total loans for a few days in Massachusetts around 2013. This can result to anomalies in Massachusetts, but nothing as such in New York and Colorado till 2014.

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.
   The trend reverses in February 2016.
   ii) What do you think is the reason for the change in trend in this month?
   The transition can be attributed to the extreme fluctuation in the stock market (drop) during that time resulting in the decrease of number of loans, thus loan per capita.

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?

On average, the mean of last 3 years is above than the mean of the whole dataset. It can also be observed that plots for these 3 years are similar with a rise in the 3rd month for 2016 and 2017, and drop in the 6th and the 9th month.

Loans per capita reach a peak during the 3rd month of 2016 and 2017, and 10th month of 2015. A pattern of decline can also be observed for these years during the 6th and the 9th month.

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?

With an overall declining ACF plot, a trend is confirmed in the data. The plot shows a negative correlation between the new and old values, thus the loan per capita decreases with each lag.

The PACF plot shows an increase in the earlier lags, with the following values spanned within the confidence interval. Thus confirms the independence of data.

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

As the lag increases the variance also increases in loans per capita. The relation is positive at the beginning but turns negative as the lag increases, that is the stochastic component increases.

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?
SNAIVE predictions have seasonality but no trend. Drift forecast predictions have an upward trend with no seasonality. They might not capture the change in amount of loans per capita because there can be an increase in population. This hinders with the ability of the model to see the amount component of the formula.

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.
Taking average grade, average interest rate, average annual income, average employment length, NYCondoPriceIdx and NYUnemployment the model can predict about 90% of the model correctly with trend, average interest rate and NYCondoPriceIdx being statistically significant. This shows that as the price of condos increases in the state the loans per capita increases on an average.
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.
The trend can be observed as statistically significant for data interpretation. This can be seen as the old model performs better as compared to the new one.

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.
For the given timeline, a larger training set is required for correct data prediction. This can be inferenced from the second plot predictions for loans per capita, which are better than the initial model since on an average the fitted values are nearer to the actual values.

e) Check the residual diagnostics for the model from (b). Does it look fine? Discuss.
The residual diagnostics show that the assumptions are not violated. There is no heteroskedasticity. There is only a little correlation in between which can be ignored.

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)?

Orders - 2, 0 and 3.

avgIntRate and NYCondoPriceIdx are statistically significant. The same variables were significant in 3(b) as well.

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.*
ARIMA model shows 0 order differencing while KPSS shows 1st order.

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?
Orders for this model are (3,1,3)

i) Compare the new model performance with the model from (f).
This model performs worse than the model in (f) as both the AIC and BIC values are higher for this model.

ii) What do you think is going on here? *(Research and)* discuss.
The model is performing worse than before as it took the data to be stationary already. Putting 1st order differencing has caused over differencing in the model.

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

3(f) model is the best as it has the lowest RMSE.

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?
Trend and season model performs the best according to the RMSE values.

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?
April data was important for proper training of trend and seasonality.

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.
Both the models have trend and seasonality. In addition, the full data also have a cyclic pattern created by the drop in sales in the year 2008. This can be explained by the recession of 2008.

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.
There is a cyclic pattern in the full data model which does not seem to be present in the partial data model. The former can explain season's variance better than the later model which can be seen by the larger left bar in the former season plot. Both analyze residual's variance effectively.
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?
A positive correlation can be seen between the lags. Spikes at regular intervals show seasonality in the data. The correlation decreases with subsequent lags. There is also a trend associated with the data.
PCF plot has spikes at the beginning showing the lags are highly correlated. This decreases later on. Except for a few spikes in between the plot is within the 95% confidence interval.
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.
Increase the size of the moving average to smoothen it further.

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.
Season and trend are significant in this model as they have very good p-values. The r-sqaured value is good suggesting the model is good. The lines out of the confidence

interval in the ACF plot shows that there is positive correlation which reduces gradually. Otherwise the model is good from the diagnostics view point.

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

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?
This model performs better than the regression model. The AIC value is good and the diagnostics show no assumptions are violated. There is very little correlation between the lags. Further, the coefficients tell us that there is seasonality in the model. There is no differencing required as the 1st order value is more than 0.05. The model is stationary.

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.
Unit tests shows a value of 1 for ordinary and seasonal differencing. After running the KPSS, it shows that the data is stationary as we get a value greater than 0.05. The model with differencing is better as is has very less correlation as compared to the other model. There are values outside the confidence interval but these are less.
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?
Trend and season model is better. The RMSE and MAE values for this model is very less. Smoothening of the fitted values in the ARIMA model can be a reason for the poor performance of the ARIMA model.

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.
Once again the trend and season model performs better. The RMSE and MAE values are very less as compared to that of ARIMA. This is again because the model might be smoothening the data more than it should to give errors in predictions.

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*?
An observation in 2019 is different from the others. While all other anomalies were seen during the recession time in 2009 this was seen recently. This observation had very low trend, season and residual values which was unexpected making it an outlier.

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?

(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.

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.