DA5020 – Assignment 9

This assignment provides you with an opportunity to create several forecasting models and evaluate the predictions. Now that you have completed the second practicum, reflect on the results for California. A possible next-step is to analyze the frequency of passengers at the various airports. Therefore, the purpose of this assignment is to analyze passenger activity at a specific airport and forecast the next time period.

The San Francisco International Airport reports monthly statistics on the passenger activity at its airport for each Airline. In this assignment, you will predict the passenger activity on a period-over-period basis. A period is composed of the total passenger activity for each month. A period-over-period analysis means that you are analyzing the data for the same period, over different years; for example, March 2009, March 2010, March 2011 etc. Click here to learn more about the data.

The data camps "Modeling with Data in the Tidyverse and "Modeling with tidymodels in R" may be helpful to complete before attempting this assignment and will count as bonus points if both are completed.

Bonus — (+5 bonus points)

1. Complete the data camps "Modeling with Data in the Tidyverse and "Modeling with Tidymodels in ". A link to Data Camp will be shared within the course, so that you can access the resource for free using your Northeastern email. Please submit the statement of accomplishment from both data camps to earn credit (must also have your name on it). (No partial credit)

Question 1 — (5 points)

Load the data into your R environment directly from the URL. Ensure that you inspect the data, so that you know how to identify the necessary columns.

Question 2 — (5 points)

Filter the dataset to extract all **domestic** passenger activity that occurred **each year, in the month of March**. After which **calculate the total passengers for each period**. Visualize the extracted data, using a line chart; indicate the year on the x-axis and the total passengers on the y-axis. Comment on the visualization. Note: the final/aggregated dataset should have one row for March of each year.

Use the extracted data to answer the questions below.

Question 3 — (5 points)

Forecast the total passenger activity for March 2019, using a simple moving average of the following time periods: 201603, 201703 and 201803. After which, calculate the error (i.e. the difference between the actual and the predicted values for March 2019). Evaluate the results; how does it compare to the actual value for the total passenger count in March 2019?

Question 4 — (5 points)

Forecast the total passenger activity for 2019, using a three year weighted moving average. Apply the following weights: 3, 5, and 7 for the respective time periods: 201603, 201703 and 201803. After which, calculate the error and evaluate the result from your prediction. How does it compare to the actual value for the total passenger count in March 2019?

Question 5 — (10 points)

Forecast the total passenger activity for 2019 using exponential smoothing (alpha is 0.7). Comment on the prediction for March 2019 with the actual value in the dataset. Note: use data from 2008 to 2018 to build your model.

Question 6 — (10 points)

Build a simple linear regression model using the year and total passenger activity for all data from 2008 to 2018. After which, forecast the total passenger activity for 2019 and 2020. Comment on the results. Note: Your predictions should be calculated using the coefficients. Do not use any libraries to make your predictions.

Question 7 — (10 points)

Calculate the mean squared error (MSE) for the models in (5 and 6) above based on the data from 2008 to 2018. Perform this step by step, using each model to make a forecast for each given time period, then calculate the squared error for each observation. After which average the squared errors. Which model has the smallest (MSE)? Note: do not use any libraries in your calculations.

Submission Details

- Your submission must contain two files: the .Rmd file and a knitted PDF or HTML (from the .rmd). Name your .Rmd file, DA5020.A9.FirstName.LastName.Rmd and your PDF/HTML DA5020.A9.FirstName.LastName.{pdf,html}, where *FirstName.LastName* is your first and last name.
- The .Rmd file must be fully commented and properly "chunked" R code and detailed explanations. Make sure that it is easy to recognize which question you answer and that your code runs from beginning to end (because that is how we will test it). Code that doesn't execute, stops, throws errors will receive no points. If the TAs have to "debug" your code or spend any effort getting it to run, substantial points will be deducted.
- Not submitting a knitted PDF or HTML will result in reduction of 30 points.
- Not submitting the .Rmd file (or both) will result in a score of 0.