STAT 153 Project

Assigned: October 25, 2019

Due: Friday, November 22, by 11:59pm

The main goal of this class is to be able to analyze and forecast a time series. Thus, your project is to analyze and forecast a time series of your choice! Broadly speaking, your task is to

1. Form a group of 3-5 people (and find ways to contribute equal work on this project!). There will be a brief activity for finding new group-mates before the end of October.
2. Choose one of the datasets. The datasets will be posted in .csv format on bCourses, along with brief descriptions.
3. Analyze and model the data, which will be written up in a professional report that you will turn in. This is described in detail below.
4. Forecast the next 10 observations that would have been observed in the time series, and turn these in as a text file. This is described in detail below.

1 Report

The main submission for each group is a professional quality report of your data analysis, modeling, and forecasting. This will be turned in on **Gradescope** as a .pdf file.The report should only contain relevant plots and R output. Include your R code as an Appendix to the report. The length of the report including all plots (excluding the R code) cannot exceed 6 pages (minimal page margins are one inch per side, minimal font point size is 12). Your report must be clearly structured into the following four parts.

1. Exploratory Data Analysis: describe the relevant features of your data.
2. Modeling time: pursue stationarity by remove trends and seasonalities, stabilizing the variance, and any other operation that makes sense for your dataset.
3. ARIMA Model selection: Provide convincing justification why a particular ARIMA (or SARIMA, etc.) is suitable for the data, and compare at least three models, choosing one.
4. Results: Estimate the parameters of your chosen model, and forecast appropriately.

This report should be written with a proper typesetting program, such as Word or LaTeX. We will provide a LaTeXtemplate on bCourses as a starting point, which also serves to clarify our expectations about the quality of your document.

2 Forecasts

Each group is required to turn in the predictions on **bCourses** as a .txt or .csv file. The text file should contain your predictions for the 10 time points, with each prediction on a new line. It should be named:

$$[DatasetNAME]\_[SID\#1]\_[SID\#2]\_[SID\#3]\_[SID\#4]\_[SID\#5].csv$$

For example, if we analyzed the "sales" dataset, and only had three group members with student ID's 123, 234, 345, then we'd title our file

$$sales\_123\_234\_345\_NA\_NA.csv$$

such that the student ID slots of members 4 and 5 in the filename are set to "NA". As an example, there is a sample submission for on bCourses: it is named $example\_123\_234\_345\_NA\_NA.csv$. We assume

that you submit your values in an increasing order: $X_{N+1}, X_{N+2}, ..., X_{N+10}$. There will also be a test-reader on bCourses that you can use to check whether your submission will be read correctly. Please be aware that your submission must be of the right form in order to be valid.

3 Grading

- 30 points - Technical details in report: Are the plots and numerical details interpreted properly? Were the modeling decisions reasonable/supported by the analysis?
- 30 points - Structure of report: figures/tables/plots are appropriately titled and captioned, the axes are labeled. Overall visual appealing and organization.
- 30 points - Quality of writing on report: Are the finding stated clearly and precisely? Is the order of ideas logical and easy to follow?
- 10 points - Forecasts: 8 points for submitting forecasts in the correct format, 2 points for how well compared to peers. Groups will be divided into "divisions" based on the dataset chosen and whether or not the methods used come from other courses. For example, a group uses machine learning methods to fit the trend, they would not be in the same division as a group that uses a parametric quadratic trend model, even if they use the same dataset. Within each division, the top 1/3 will get 2 points, the middle third 1 point, and the bottom third 0 points. The purpose of this small competition is to incentivize everyone to try and fit a great model instead of just a tolerable one.