STAT 153 Project

Assigned: February 4, 2021

Due: May 10, 2021, by 11:59pm
with checkpoints on
February 10, March 10, March 31, and April 21

The main goal of this class is to be able to analyze and forecast a time series. Thus, your final project (in place of a final exam) is to analyze and forecast a time series of your choice! This is an individual project. Broadly speaking, your task is to

1. Choose one of the datasets. The datasets will be posted in .csv format on bCourses, along with brief descriptions in README files that you will want to check out.
2. Complete a detailed analysis of the dataset.
   - Exploratory data analysis
   - Pursue stationarity two different ways (two signal models)
   - Model the plausibly-stationary noise from the two signal models, in two different ways each, resulting in four total models.
   - Appropriately select the "best" model (and you should discuss how you define "best").
   - Using this best model, forecast the next 10 time points. Turn in these 10 forecasted value as a csv file using the provided format. This is described in detail below.
3. Write up results in both a professional technical report (six page limit), written to a subject matter expert who has taken STAT 153 before, but it has been a while, so you'll need to talk through complicated points.

1 Report

The main submission is a professional quality report of your data analysis, modeling, and forecasting. There is a six (6) page limit, and it will be turned in on **Gradescope** as a .pdf file. The report should only contain relevant plots and R output. You do not need to include code, but you may include it as an appendix if you would like (it would not count against your page limit). This is written to a subject matter expert who has taken STAT 153 before, but it has been a while. Though you don't have to define things ("a correlogram is a plot of the ACF") you should explain your thinking so that the expert can see that you are competent in analyzing the problem ("the ACF correlogram shows large magnitudes of autocorrelation at lags 1, 2, and 5"). Note that tone is professional but "spikes at lags 1,2,5" is not.

This report should be written with a proper typesetting program, such as Word, LaTeX, RMarkdown, etc. We will provide LaTeXand RMarkdown templates on bCourses as starting points. You're free to use either or neither. Note the RMarkdown template is the example report.

Your report should be clearly structured into the following five parts.

1. Executive Summary: In the beginning of your report, in a short summary, summarize in a few (about 2-4) sentences which data set you analyzed, which model you used to make predictions, and briefly describe how your predictions address the scenario presented in the README file of your dataset. Anything *extraordinary* about your dataset and model can be mentioned here as well.
2. Exploratory Data Analysis: describe the relevant features of your data, including what features are preventing it from being stable/stationary? Are there any peculiar patterns (e.g. time series is only negative on Wednesdays)? Naturally, this section should contain a time series plot of the dataset your are describing.

3. Models Considered: FIRST, include details on your pursuit of stationarity: removing trends and seasonalities, stabilizing the variance, and any other operation that makes sense for your dataset. You should do this in two different ways (e.g. model A and model B).

   Then, SECOND, describe your ARMA model selections: provide convincing justification why a particular ARMA (or SARIMA, etc.) is suitable. You should do this in two different ways for each stationary series, resulting in four different models (e.g. models A1, A2, B1, B2). For example, the models could be

   - polynomial trend + ARMA(1,1)
   - polynomial trend + SARIMA(p=0,q=0,P=1,Q=2,S=12)
   - ARIMA(p=1,d=2,q=3) (second differences with ARMA(1,3))
   - SARIMA(p=0,d=2,q=0,P=1,S=12).

   You will *not* have enough space to show every single plot and detail of every model, so be selective in what needs to be shown! Two suggestions: first, anything common across many or all of the models can be addressed in the Exploration Data Analysis section. Second, be sure to describe your chosen model thoroughly, as that's the one people will care about! For each modeling step, you should minimally show two things through visualization: 1) what motivates your modeling decision (time series plot for signal model, ACF/PACF for ARMA model), and 2) that your model indeed fits well (residual plot for signal model; there's many options for showing ARMA model fit).

4. Model Comparison and Selection: compare these four models, and select the "best" one. Be sure to clearly describe why that model is "best". This may involve AIC, BIC, etc., but must include time series cross-validation.

5. Results: For your chosen model, do three things. First, write out the model mathematically. Second, estimate the parameters of your chosen model, probably in a table. If you have a lot of parameters to estimate, this table may be pushed to page 7 as an appendix, but only this table. Third, forecast appropriately and include a plot of your forecasted values appended to the end of your time series. This final plot is the star of the show, so do it well. One thought: how many datapoints should be on this figure? If your time series is 9,000 points long, the forecasted 10 points won't even be visible when appended to the end, so perhaps think about what subset makes sense.

2 Forecasts

Use your best model to forecast the next 10 points. For example, if you data set contains daily observations for January 1 through March 31, you should forecast the values of April 1 through April 10. Turn in the predictions on **bCourses** as a .csv file using the provided template by filling in the NA's with your predictions for the 10 time points, but nothing more. It should be named:

$$[DatasetName]\_[SID].csv$$

where the dataset name is "covid" or "stock". For example, if I analyzed sales.csv, and my student ID was 30333333, then it'd be called

sales_30333333.csv

There will 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

This project is 50% of your grade. Each checkpoint is worth 5%, and the final submission is worth 30%.

3.1 Checkpoints

The checkpoints' primary purpose is to spread out the workload, have you start early in the semester, and give you a chance for early feedback if you want it. As we have high expectations for your final report, and it will be graded accordingly, though these checkpoints will essentially be completion graded. Our hope is that your final project submission is fantastic and easily earn high grades! The secondary purpose is that this provides a direct application for what we are learning in class and helps to facilitate learning throughout the semester!

On each checkpoint, you will submit your report on Gradescope. These will be completion graded, for example, the first check point is to simply have a professional heading (title/name/date/etc.) and a plot of your chosen time series dataset, in PDF format of course. As long as you have these two things, you will get full credit. The details of the other checkpoints will be released closer to their due dates.

3.2 Final Submission

- 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: It looks like a professional report! Sections/subsections of the report are labeled. Everything fits on the page. The font size is appropriate. Figures/tables/plots are appropriately titled and captioned, the axes are labeled. Overall visual appeal and organization.
- 30 points - Quality of writing on report: It reads like a professional 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 in terms of root mean-squared prediction error (RMSPE). For the students who looked at a particular dataset, 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.
- A more-specific rubric will be provided later.