Final Project
MATH/STAT 4540/5540 Spr 2022 Time Series

Due date: Saturday, April 30, before 7 PM, on Canvas/Gradescope
Instructor: Prof. Becker
Theme: Real-world prediction
Last update: 4/11/22

Short version The goal of the project is to predict electric energy demand in the Xcel Energy grid in Colorado on Sunday May 1st from 5–6 PM[1]. You must turn in the project by Saturday, April 30 at 7 PM; late projects will not be accepted!

The person or team with the most accurate prediction will receive an automatic "A" in the entire course; the next two people/teams with the 2nd and 3rd most accurate predictions will receive an automatic 100% on the final project grade.

Long version: Details

1. Projects may be done individually or in groups of up to 2 students. If done in groups, both students are given the same grade.

2. Students enrolled in 5540 (rather than 4540) will be graded on a slightly stricter scale; if one student in the group is in 4540 and the other 5540, the project will be graded on the 5540 scale

3. Projects should be typeset using a computer and turned in as a PDF with about $2-4$ pages of text, not including any figures. Including figures and snippets of code, the project should be $3-10$ pages. The PDF should look professional and be ready to print (so $8\frac{1}{2} \times 11$ inches); do not include endless amounts of code and/or figures. Make sure that text in figures is readable. All plots should have correct labels, e.g., axes must be clearly labeled (e.g., x-axis should be a time/date when appropriate), and include units of measurement when appropriate.

4. The report should include an introduction section that gives background on the data and motivation for why we want to predict it.

5. Grading is based on more than just accuracy of the final prediction; see the grading rubric below. You must explain the methods you are using, writing out equations and/or code if appropriate, and should include a few figures. Your job is to convince the reader that you have used appropriate techniques.

6. To reiterate the above point, you receive credit for using appropriate techniques, and for explaining why you are using a given technique. There are no required techniques, as long as you justify the techniques that you do use.

7. You are allowed to use techniques not taught in the book, and you may use the internet to help you, but you may *not* use any fancy software package (if in doubt, ask the instructor) nor ask for explicit help from anyone outside this class (e.g., no posting questions on internet forums, other than basic questions about underlying math and/or coding).

    a) It is also permissible to use external datasets. For example, if you thought that energy demand and weather were correlated, you could use both historical weather data and weather forecasts (but this is by no means a requirement for the project). If you're interested in this approach, read about multivariate (vector) time series.

[1] i.e., for grading, I will look at the "6 PM" data from EIA, since the hourly data is an aggregate of the past hour.

8. Historical data on energy demand in Colorado is available from the US Energy Information Administration (EIA) at /www.eia.gov/opendata. Look for `EIA Data Sets > U.S. Electric System Operating Data > Demand > Balancing authorities > Public Service Company of Co` (PSCO the name of the Xcel Energy subsidiary that operates in Colorado). This data is refreshed constantly, so you may want to download a sample dataset early in order to plan your technique, and then download the most recent dataset on Saturday April 30th to make your final prediction.

9. If the project needs more details or clarifications, please check back to this PDF (it has a "last updated" date at the top so you can see if it has changed).

Background In mid-2015, the EIA began collecting hourly energy demand data across the contiguous United States. The data are separated by "balancing authority" (which means each smaller grid; there are about 56 of these though sometimes they change due to company mergers; grids are connected to each other). Why this data? Because you can probably find some motivations for why we want to predict it; it should be somewhat predictable (it's not as tied to financial markets compared to stock market data, not quite as volatile as weather, although it does depend on weather); it's freely available and constantly updated by the government EIA website; it has trends and seasonality of different types; while it may have outliers and occasional missing data, it's at least regular (recorded hourly 24-7).

Detailed grading rubric

APPM 4540/5540 Time Series, Spring 2022, Stephen Becker. Final project rubric

High-level Component	Component	Percent of grade	High standard	Medium standard	Low standard	*Grade*
Methods: math and programming (50%)	Preprocessing	10%	Data are preprocessed as needed, such as transformations, removing outliers; plotted for sanity checks	Follows high standard most of the time	Preprocessing not considered; data not downloaded correctly, or not loaded correctly	
	Correct usage of methods	10%	Appropriate methods are chosen, and used correctly. Coding and math are correct or mostly correct	Follows high standard most of the time	Inappropriate methods, or appropriate methods but used incorrectly. Major and/or repeated errors in coding and math	
	Justification of methods	10%	Clear justification is provided	Follows high standard most of the time	Not clear why the chosen methods are used; follows unmotivated formulat from textbook	
	Generates valid conclusion	10%	Discusses results and possible improvements and modeling errors; discusses uncertainty in data	Follows high standard most of the time	Lacks conclusions; may have hit an impasse and doesn't discuss possible reasons why or possible fixes	
	Accuracy of final prediction	10%	Highly accurate	About as accurate as guessing the value from the same time one week earlier	Less accurate than via simple methods, indicative of a major mistake or lack of diagnostics/debugging	
Presentation: written document (50%)	Coherent organization	10%	Includes clear intro and conclusion and other valid sections; logical flow, easy to follow	Follows high standard most of the time	Jumps around, no clear flow. Equations just stated without discussion	
	Writing, grammar	10%	Reads professionally, less than a handful of mistakes. Words are precise, jargon is appropriate	Mostly understandable but not senior college level	Not understandable in several or more places	
	Document is professional	10%	Typeset, figures laid out clearly. Appropriate length, correct size, font are reasonable	Follows high standard most of the time	Unprofessional; wrong length; hard to read fonts	
	Figures	10%	Appropriate number of figures included to justify methods; labels are clear; figures are legible, legends included as appropriate. Easy to distinguish line series (even if printed in black & white)	Follows high standard most of the time	Too many figures or too few; lacking many labels; needs legends; confusing to follow. Figures are not clearly related to the main text.	
	Math, code	10%	Math is easy to follow, equations incorporated into sentences as appropriate. Units are included as needed. Code snippets are included as appropriate to support the report	Follows high standard most of the time	No equations given when they would have been helpful; math hard to follow; missing units. Too little or too much code is provided, and/or hard to follow	
TOTAL		100%			*sub-total*	