Marko Suchy
3/8/24
Data Science Capstone
Reflective Essay

Crap in, Crap Out

Most of my education has revolved around modelling. Through fourteen physics classes, I've practiced modelling real-world phenomena via mathematical tools and complex theories. From applying equations of state to large thermodynamical systems in statistical physics, to exploring the probabilistic location of energy quanta in a single hydrogen atom, I've explored a wide range of physical models based on theoretical rules.

More recently, I have begun exploring the synergy between data and theoretical modelling. The synergy offers robust and useful capabilities for analysis and prediction. Consider regression, a fundamental tool in the world of policy setting. Or Generative-AI, which will continue to change the workflow of countless professionals, across almost every industry.[1] Each model pairs theoretical rules with data-reliant computational implementation. My education has positioned me to understand both the theory behind models, and the nuance behind their implementation.[2]

Of course, modeling cannot occur in a black box. Before one can access the synergy between data and modeling, there are several steps which occur:

1. Data must be collected in an ethical manner, with the method of collection thoroughly recorded (data must be well defined.)
2. Data must be prepared by cleaning and transformation into must some useful representation which can be understood by a human or a model.
3. Finally, the data must be interpreted and communicated, which may include visualizations, summary statistics, model implementation, or other methods.

In my progression throughout the data science minor at Washington and Lee University (and through some summer experiences with different institutions) I've leaned and practiced these three steps. In the following essay, I will review my 'data science journey' by presenting: a class list, to give context for my academic experiences; A note on data collection, which explains how I've come to formulate the first step as described above; and several project descriptions, which detail how I've put all 3 steps into practice in various contexts.

Class List

In chronological order, the classes I've taken which fulfil the minor requirements are as follows:

- MATH 222 - Linear Algebra
- CSCI 111 - Introduction to programming
- SOAN 265 - Exploring Social Networks

[1] I am currently studying both of these tools in my ongoing coursework

[2] I do not mean to leave out raw data analysis, which is an important part of data science. However, data analysis is almost never done without statistical, or other, tools acting as a theoretical framework. Although this theoretical framework may not match some definitions of a 'model,' for the purposes of a unified paper, I will consider them models here.

- PHYS 265 - Modeling and Simulation of Physical Systems
- BIOL 325 - Ecological Modelling and Conservation Strategies
- BIOL 201 - Stats for Biology and Medicine
- ECON 203 - Econometrics (ongoing)
- BUS 306D - User Generated Content: Analytics and insight (forthcoming)

I've also drawn on my experiences from some classes outside the minor to develop my intuitions and skills for data science. Those classes are:

- PHYS 207 - Electrical Circuits
- PHYS 295B - Advanced Lab
- CSCI 297D - Topics: Generative AI: Creating with Computation (ongoing)
- PHYS 345 - Statistical physics (ongoing)

A Note on Data Collection

One phrase which has come up many times throughout these classes, especially in SOAN 265, BIOL 325, BIOL 201, ECON 203, PHYS 207 and PHYS 295B: "crap in, crap out." The phrase refers to something of the utmost importance in data science: data quality. This is something I've learned first-hand, especially in my laboratory-based physics classes. Across disciplines, inconsistent data collection methods lead to difficulty in matching data to any sort of theory, be it mathematical, economic, sociological, or other.[3] Most classes requiring the collection of data have emphasized the importance of developing (and recording) quality collection methods. When using data from external sources, it is important to understand the nuance of data collection, so you actually be able be able to interpret the output of your model.

In their book *The Data Game* Maier and Imazeki claim "careful examination of the underlying data helps resolve the apparent statistical quandaries." They continue to explain the importance of considering the definition of data when drawing conclusions. They offer several examples of faulty conclusions generated across healthcare and education based on a lack of understanding of data definition. Without a solid definition, data cannot be quality, or useful in interpretation.

Projects

Throughout my data science journey, I've carried out a few projects which have required the use of all three steps I've described above. Those projects will be presented in chronological order here.

European Networks Project

In SOAN 265 "Exploring Social Networks" I conducted my first data driven project under Professor Eastwood, in W&L's sociology department. The project was a culmination of the course, which focused on classical social network analysis ideas: network homophily and heterophily, triadic closure, the strength of weak ties, and networks position. The project offered

[3] Poor data collection can become quite apparent in laboratory experiments, where your experiment can be visualized exactly versus theory, and error can be precisely calculated. It may not be so readily observed in other disciplines, where theory is less definitive, which is makes consistent collection even more important!

the opportunity to bring all these things together, and to attempt to re-create our graphs using various Exponential Random Graph Models (ERGMs.)

I used R for network analysis in this project, but I took the opportunity to practice my scraping in Python (as I much prefer Python.) I scraped web data using the "BeautifulSoup" and "requests" and "re" packages in Python. I scraped my data from the Correlates of War database[4], the Berlin Free Institute[5], and the World Integrated Trade Solution[6], which allowed me to create networks representing military alliances, trade relations, and cultural similarities. I also scraped several attribute lists for countries including various indicines of democracy, freedom, GDP, etc. from Wikipedia. In doing so, I learned a little about the structure of HTML tables, and how difficult it can be to locate tables on a website via HTML. I exported three edge lists (two of which were weighted) to csv files in my working directory, where they were ready to be used for network construction in R. This was my first foray into scraping and preparing data so that it could be used effectively by a model.

In R, I turned my scraped edge lists into networks, and visualized node attributes in color, and edge weights in opacity, along with distributions of centrality to understand actor prominence. I also fit null and alternate ERGMs to my networks and analyzed the outputs. I believe a major strength of this project was the effective use of network visualization techniques, especially the setting opacity of edges based on weight. This required a little extra code, but resulted in a very pretty graph. One area in which this project could be made better, is better understanding what exactly the definition of an alliance in the Correlates of War dataset. The lack of exploration of the types of relationships existing in the Correlates of War dataset made it difficult to interpret the alliance network (other than the NATO cluster, which was easily visible.) Furthermore, the 'strength' of these alliance was variable, but this was not captured in the network. More work could certainly be done here to better understand data definitions, and I should've considered using alternate sources.

The Future of Ursus Maritimus

In spring of 2023 I studied ecological modeling under Professor Humston in the Biology Department. The class focused on many types of models for population dynamics, ranging from simple exponential growth to stage-based methods using projection matrices (matrices of survival rates at varying stages of life.) The class used Excel, R, and Vensim (a simulation software.) The class also taught sensitivity analysis, which could be used to maximize population conservation efforts per resources allocated. Finally, we discussed dynamics of two-population systems, and the ways in which they could reach a stable (or unstable) state based on isoclines in phase plane plots. Ultimately the class explored how population dynamics modeling could be used to develop effective methods for conservation.

For the seminal project of the class, a partner and I studied the possible effects of a novel virus on the polar bear population in Alaska. We found data from Comadre[7] on the projection matrix of this population. We then implemented a deterministic, stage-based SIR (susceptible, infected, recovered) model in Vensim for a novel virus effecting fully grown bear, and used

[4] Note, this data didn't need to be scraped in the same way other data was. It was available for free download. (https://correlatesofwar.org/data-sets/formal-alliances/)
[5] https://userpage.fu-berlin.de/~jroose/index_en/main_indexvaluesaz.htm
[6] https://wits.worldbank.org/Default.aspx?lang=en
[7] https://compadre-db.org/Data/Comadre

Monte-Carlo methods to simulate results across different hypothetical virality and death rates. We exported data to Excel and fit a line to the logarithm of our population data,[8] then used the slope of the line to estimate the r value of our exponential decays (in all cases, the polar bears are dying.) We made a phase diagram across virality and death rates (using R) and found a well-defined relationship between rates.[9] Particularly, our model predicted that a novel virus with high virality was more threatening for polar bears than one with a high death rate.

One weakness of our project was that it did not account for the dependence on density in virus spreading. One may make the argument that, as a confounder, virality could capture density dependence. However, constant virality cannot account for changing population density as total population decreases. A strength in the project was its relevance to cute and fluffy things. When presented, the project had a strong emotional resonance with the audience, which was in large part due to the saddening reality that the polar bear population is declining, and will likely become extinct soon with or without a novel virus. This project taught me lots about the nuances in epidemiological models, and about presenting in way that is relevant and relatable to the audience. When you break your findings down in a simple way that is relevant to something the audience cares about, you will find great success.

The Leyburn Stair Exercise (LSE)

To fulfil the statistics requirement for the data science minor, I took Stats for Biology and Medicine with Dr. Toporikova this past fall. The class focused on classical statistical methods, with an emphasis on communicating statistical realities to normal people. Dr. Toporikova wanted her students to be able to carry out basic analyses, and clearly indicate the likelihood of outcomes in simple language. Dr. T maintained open book tests and assignments, encouraging students to use all available resources whenever possible. She also emphasized the creation of effective graphs and charts to understand statistical problems, and develop hypotheses. The class did not require use of any particular tool, however I often used Excel to carry out my analyses. For applying basic mathematical models to datasets, I used the "Distributions" app on my iPhone, along with several classic statistics formulas.

As a final test for our statistical prowess, Dr. T split the class into two groups, asked both groups to develop and test a hypothesis based on a biological measurement. Our group was assigned blood pressure (BP,) and given cheap BP measurement systems. We thoroughly researched prior experiments, along with biological pathways for change in BP, and subsequently developed our own. Our experiment was quite simple and designed such that our low fidelity tools for data collection could still capture changes in BP. We measured BP before and after students walked from the top floor of Leyburn Library, all the way to the bottom, and then back to the top. We kept our experimental group to be only within our class, so we did not have to go through the IRB review process. (We really would not have had time to go through this process.) It was important in our experiment to get all participants consent, and explain that as students we would be observing BPs. Although true blinding was not possible for such an experiment, we blinded names, and kept data private where we could. This was an excellent exercise for ethicality in data collection.

[8] Our population change looked a lot like exponential decay; thus the natural log of population change was roughly linear.

[9] When $virality \geq 0.125 + 1.25 * death\ rate$, a novel disease would result in a significantly faster decline in population.

Our analysis utilized a t-test and confidence interval to check that both men and women had a statistically significant increase in BP. We used a two-way ANOVA (begrudgingly carried out in R) to check if BP increase had effects from gender, or gender-LSE interaction, and found gender was not significant in BP increase. A major strength of the project was the simplicity in our experimental design, which allowed for effective use of classical statistical models. A weakness in the experiment was the low number of subjects (n = 9) due to constraints in class size.

Luddism on Diffusion Networks

This ongoing project began as a project for my modelling of physical systems class (PHYS 295) but has continued as fulfilment for my physics major's thesis requirement. The project is focused on understanding diffusion of innovation trough society, considering both adopters of an innovation and those who reject innovation (luddites.) The project focuses on two versions of similar models: a deterministic model based on coupled differential equations, much like an SIR style epidemiological model, and a stochastic network-based simulation which uses probabilistic rules to update individual agents in a system. So far, stochastic simulations based on Erdos-Renyi random graphs have matched rescaled deterministic simulations quite well on average.

The project is heavily reliant on Python. Through the project I have developed tens of scripts, with thousands of lines of code in total. I have utilized several packages including NumPy, SciKit, Matplotlib, and especially NetworkX and Pandas. The scale of the project has forced me to be well organized in the storage of both scripts and generated data. Because the stochastic model takes a long time to run on large networks, it is imperative to store metadata such as model's parameters, date of data creation, and more. For this, I have stored my data multi-dimensionally, in the form of .JSON files, with parameter dictionaries, dates, model rules, etc.. This has allowed for sooth analysis of the model's results across thousands of stochastic runs. Such analysis has taken form of histograms, box and whisker plots, phase plots (scatterplots of results across parameter space) and more.

One limitation of the model which I have developed is its lack of predictive power. As shown by Frank Bass (who my model is ultimately based off) fitting these types of models, which predict spread of one novel innovation through society, can be difficult in a noisy society. On the contrary, the model may offer a strong analytical framework for understanding radical anti-politics and polarization in a general way.

Conclusions

Throughout the data science minor, and experiences outside the minor requirements, I have honed my understanding from fundamental theoretical models of statistics to applied analysis of real-world data. I have practiced ethical collection of data with consistent methods, wrangling and preparing data such that it is useful for analysis, and effective interpretation and communication of analyses to a wide range of audiences. As I move into the working world I am excited to apply what I already know about data and continue learning more.

Work Cited

Maier, Mark, and Jennifer Imazeki. The data game: Controversies in social science statistics. Routledge, 2016.

Bass, Frank M. "A new product growth for model consumer durables." Management science 15.5 (1969): 215-227.