DATA SCIENCE PORTFOLIO

DS-401: Data Science Portfolio Class

Abstract

Jimmy Horvath is reflecting on his academic journey by connecting data science to other disciplines. In this project, he has compiled a portfolio showcasing his growth in the six core competencies of Data Science. Ultimately, this portfolio highlights his data skills and achievements at Washington & Lee University.

Jimmy Horvath

horvathj25@mail.wlu.edu

Introduction

Arriving at Washington and Lee University in September of 2021, I was unsure of what I wanted to study. As a child, I watched my parents work for brokerage firms and healthcare companies. Growing up idolizing my parents and how hard they worked, I often pictured myself in their shoes, in big business. Thus, after a year of liberal arts study at Washington and Lee, I declared an economics major. As I advanced in economics, I took statistics and econometrics classes which I really enjoyed. I figured if I could understand economic data, being able to represent it would make me more valuable to a business post-graduation. So, I added a data science minor to my economics degree in hopes of learning something new and honing skills.

In the minor, novice data scientists are required to take four types of classes and six classes overall. These types are called "Foundations," "Statistics," "Computing/Programming," and three "Electives." Atypically, my data science journey commenced during the second semester of my sophomore year. During Winter 2022, I began with the "Statistics" class called Economics-202. In Fall 2023, I took Economics-203: Econometrics and Business-306: Text Analysis for Business, both counting as "Electives." In Winter 2023, I again took two data science classes: Data Science-395: Statistics & Medicine and Cognitive & Behavioral Sciences-195: Data Trends Over Time. DS-395 functioned as my final elective, while I used CBSC-185 as my "Foundations" credit. Lastly, I finished the minor in Fall 2024 with Computer Science-111: Introduction to Computer Science. This class counted as my "Computing/Programming" credit.

ECON-202: Statistics for Economics

First, as part of the economics major, I was required to take Economics-202: Statistics for Economics. This served as my “Foundations” sector of the Data Science minor. While this topic is critical for success in the upper-level electives of the economics major, it is also important for any career that relies on empirical research in the social sciences.

Economics-202 focused on the fundamentals of probability, statistics, estimation, and hypothesis testing. It ended with an introduction to regression analysis. I learned how to think formally about data, uncertainty, and random processes, while learning hands-on methods to organize and analyze real data using modern statistical software. Specifically, through formal modeling, I gained an understanding of the nature of randomness and its implications for empirical investigation. Many examples came through hands-on labs using Stata.

Ultimately, Economics-202 culminated with a large project. As a result, I interpreted and critically evaluated statistical results derived from sampling distributions, hypothesis testing, and linear regression. Using Stata, we answered the question: “How important are factors like income, education, gender, and race/ethnicity on one’s self-reported health (SRH) condition?” These answers help health professionals prioritize creating healthier societies.

Our regression results determined that the main factors impacting SRH were education and gender. While all but one of the endogenous variables we looked at yielded statistically significant results, education was the only variable to yield clinically relevant results. Our research highlights the importance of providing equal access to educational opportunities to better health equity.

ECON-203: Econometrics

If Economics-202 is designed to develop a student’s ability to solve problems by supplementing common sense with simple quantitative models, Economics-203: Econometrics is designed to build upon the Econ-202 foundation. The class expands the list of concepts and techniques considered, thus broadening the types of problems that can be addressed using quantitative models. Additionally, the class considered methodological issues related to data quality. We constantly asked the question: “What is the use of fancy statistical/econometric models if the data are faulty?”

We had four goals for the class. First, Professor Blunch wanted us to understand and be able to critically discuss potential pitfalls of data analysis, especially pertaining to the analysis of education and health data. I found the health data portion the most interesting, and as such, took Economics-376: Global Public Health as a result. Secondly, we wanted to know the classical conditions and the Gauss-Markov Theorem, understanding the importance of these as crucial underpinnings of applied econometric analysis. Thirdly, my group understood and critically applied different estimation methods and corrections. These included OLS, correction for serial correlation and/or heteroskedasticity. Lastly, to culminate these thoughts, we formulated and conducted an empirical group project for a topic of our choice, based on economic theory. We used the Stata statistical package to do “real time” data analyses, helping develop our research.

I chose Maddie Weller as my project partner, as she was reliable and intelligent during the Economics-202 project. I also knew she enjoyed sports, as she’s a two-year captain on the volleyball team. I wanted to sprinkle sports onto my love for econometrics, and Maddie was on the same page. Thus, we set sail on a statistical voyage to explore the NFL Draft.

There are many variables influencing when a player is selected in the draft. Our paper analyzed these variables and answered two research questions. First, what are the most important factors in producing first-round draft picks in the NFL? Secondly, conditionally that a player is a first-round pick, what are the factors that affect where they're picked in the round?

Being college athletes, we understood better than non-athletic economists that the results have massive implications for colleges. With lucrative media contracts, sponsorship deals, and ticket sales, schools continue to find ways to maximize their revenue potential. For many large colleges, football is the largest revenue earner. Thus, programs focusing on the most influential variables can theoretically produce more first-round picks, earning more money for the school overall. Additionally, a program that produces more first-round picks can better recruit top-tier high school players, improving the team and enhancing the future of the college.

To answer the first question, we used the linear probability model to measure a player's likelihood of being picked in the first round. The coefficients suggested the expected change in percentage points that a player is selected in the first round, given the conditional explanatory variable.

To answer the second question, we used the traditional OLS estimator based on the given regressors in the theory section. It is essential to explain that a "higher" draft pick is desired; being drafted #1 overall versus #10 is a massive difference. Therefore, some of the coefficients produced by this model were negative. This means a player is drafted earlier and "higher," so they are closer to being the #1 pick, which is what we were looking to observe. The variables that yielded the most negative OLS estimators were the most relevant for answering this question. Ultimately, we titled this variable "pick height" in the first round.

Our regressions comprised ten variables: school winning percentage, school football expenses, race, height, weight, if a player played in the Southeastern Conference (SEC), if a player grew up in the Deep South area of the United States, and position.

Clinically, we were able to draw conclusions on statistically significant results, which only resulted with offensive positions. First, quarterbacks are 88% more likely to be first-round picks than second-round picks. On average, QBs move up 7 slots if they are first-round selections. Additionally, Black QBs are 3% more likely than white QBs to be a first-round pick than a second-round pick. Secondly, offensive linemen are 90% more likely to be first-round picks than second-round picks. Additionally, if a lineman is 23 pounds heavier than the average from 2013–2023, they're 62% more likely to be a first-round pick than a second-round pick. Ultimately, these findings confirmed our theory that the NFL is a passing league, and as such, general managers want quarterbacks who can throw, and heavy offensive linemen to protect them. Also, a significant result is that we're 90% confident that both school winning percentage and college football expenses impact a player's draft stock.

A few takeaways are critical to understand. First, college programs should emphasize their strength & conditioning and nutrition programs to improve where their players are selected. This is especially important for offensive linemen. Secondly, quarterbacks are the top draft priority to NFL teams. They are the "face of their franchise," increasing visibility of their alma mater to the entire nation. Thirdly, schools should spend more money on football, improving factors like strength and conditioning or facilities. Lastly and surprisingly, all defensive positions, the Deep South variable, and the SEC variable lacked significance, questioning the pertinence of these variables in draft tendency.

BUS-306: Text Analysis for Business

This class introduced me to text analysis, allowing for the conversion of raw text into data that may be explored. The goal of text analysis is to generate insights that guide business-related decisions. Additionally, I learned natural language processing, the basics of Python programming, and developed an understanding of how language and text are interpreted.

Throughout the course, I completed two projects that analyzed real-world textual data. First, in a three-person group, I completed an Earnings Report Project. During the project, we developed a tool that analyzes financial earnings reports to generate a prediction of the market reaction to each report. Those predictions were tested against actual market reactions. We started with $100,000 and ended with just over $103,000. We tokenized the MD&A (Management Discussion & Analysis) section of multiple 10-Ks into words and sentences for analysis.

The second group project was titled the Kickstarter Project. We developed a tool that analyzes the text from Kickstarter campaigns to generate a prediction of the likelihood that the campaign would be successful in receiving the required funding. Those predictions were then tested against whether the campaign was successful or not.

Throughout the course, we had four goals. First, we learned introductory Python programming. Secondly, we learned natural language processing and applications of text analysis and sentiment analysis, especially for business applications. Thirdly, we learned web scraping to collect textual data and generate usable information from web sources. Lastly, we discussed ethical concerns related to text analysis, especially with how rapid advances in AI technology are affecting text analytics.

DS-395: Statistics and Medicine

Prior to my junior year winter, I knew I had to take data science credits for the major. One of the classes offered was DS-395: Statistics and Medicine. My mom worked for Pfizer for two decades, and I took the class to get closer to her. Also, Professor Sybil Prince-Nelson was teaching it, and I had taken Calculus with her freshman year and enjoyed the course.

In Statistics and Medicine, I explored the current state of the pandemic. This included four main topics. First, I researched why COVID-19 continues to claim lives. Secondly, we learned how data science helps in tracking and mitigating its impact. Thirdly, I discovered what the enduring effects of the disease are, and the vaccines developed to fight it. Lastly, we questioned what the pandemic's trajectory over the past four years was and what the implications are for the future.

This course featured four goals. First, we aimed to build a strong foundation in R programming. Secondly, we wanted to master data analysis techniques for real-world applications. Thirdly, we were aware of ethical standards in data handling and analysis. Lastly, we wanted to develop effective data communication skills.

Project One was titled “Data Analysis of COVID-19 Vaccine Efficacy.” The objective was to analyze real-world data to assess the dynamic among different age groups being vaccinated. Our data analysis was consistent with prior research showing that the percentage of those who are fully vaccinated between the ages of 18–24 and 25–49 is significantly lower than those who are 65+. While we may not know the full reasoning behind this, it is an important observation when encouraging people to get vaccinated in the future for COVID-19 or other diseases.

Project Two was titled “Long-term Side Effects.” The goal was to investigate and report on the long-term side effects of COVID-19 vaccines using data science techniques. In a literature review titled “Analyzing Trends: The Impact of Long COVID on Various Bodily Systems,” we found that Long COVID drastically impacts many parts of the body over time. This includes thrombotic and cardiovascular systems, respiratory systems, neurological and olfactory functions, orthostatic challenges, and psychiatric symptomatology.

The final project was titled “Analyzing Global COVID-19 Vaccination Rates Using Chi-Square Tests.” We asked three questions. First, is there a significant difference in life expectancy before and after the pandemic? Second, how did the impact of the pandemic on life expectancy vary by group? Thirdly, are vaccination coverage and GDP good predictors of life expectancy post-pandemic? We used Levene’s test, a two-way ANOVA run in R, and linear regression to analyze GDP and life expectancy data from the World Bank. Ultimately, our model suggests that vaccination rates, economic standing (GDP percentile), and regional location all significantly impact post-pandemic life expectancy. Specifically, Europe has the highest expected life expectancy post-pandemic, followed by Asia and North America. Additionally, higher GDP and vaccination rates lead to higher life expectancy.

CBSC185: Data Trends Over Time

How can we map our feelings over time? How do various events impact our feelings, attitudes, and thoughts over the course of a year, or more? Can we effectively monitor our health behavior and choices and identify how they impact our mental and physical health? Can we assess employees' satisfaction in their jobs over time or student retention, attention, and learning over the course of a term?

CBSC185: Data Trends Over Time exposed me and my peers to these types of questions through data analysis. Throughout the course, we developed key skills in R. This included organizing and managing data, creating data visualizations, conducting statistical analysis, and undergoing best practices in data handling, management, and reproducibility. Goals for the course included developing a deeper understanding of data science and statistical concepts and developing a preliminary understanding of what longitudinal data is. Additionally, we learned how and when to conduct various analyses.

To achieve these goals, we had some outlines and objectives. First, we practiced and learned how to draw conclusions from various datasets. Secondly, it was important to further develop skills in presenting statistical findings both verbally and in written reports. Also, it was key to be an informed consumer of psychological findings in media and everyday life, especially the statistical and methodological information and conclusions drawn.

CSCI111: Introduction to Computer Science

Lastly and ironically, I finished the data science minor with an introductory course in the computer science department. However, CSCI111: Introduction to Computer Science is more than just an introduction to programming. Professor Tolley described it as “an exploration of the fundamental problem-solving techniques that are at the core of computer science.” Sure, we learned how to write code, but we also learned how to think like computer scientists.

In the course, we covered many broad topics. First, we designed and implemented algorithms for solving a wide variety of problems. Secondly, we got an introduction to the syntax, semantics, and pragmatics of Python and Linux. Thirdly, we had access to several programming applications, including numerical computation, text processing, graphics, and networking. Lastly, we developed systematic techniques for testing and debugging programs. This ensured that we not only wrote code but also understood how to make it robust.

For my first project, I created a movie review organizer. It was designed to manage and analyze user reviews for movies. It allows users to interact with the system via a menu, performing tasks like adding and displaying reviews, and calculating the average review length for each movie. Building off the movie review organizer, I created a library management system with three main classes: Book, Library, and Member. It allows managing books in a library, including adding, removing, and listing books, and enables library members to borrow and return books.

In the next project, I expanded my understanding of encryption by implementing the Vigenère cipher, a more advanced method compared to the Caesar cipher. I used a keyword to shift letters and applied modular arithmetic for string manipulation. I incorporated conditionals and loops to manage the encryption and decryption processes and used debugging techniques to ensure my code worked correctly. To make my implementation more efficient, I reduced redundant code by combining the encryption and decryption functionalities into one streamlined function.

The final project was titled “Superstore Sales Analysis.” I analyzed and extracted meaningful insights from a Superstore Sales dataset while practicing programming concepts like loops, lists, dictionaries, and classes. By the end of the project, I created a fully functional Python script and a professional report summarizing my findings.

Conclusion

Ultimately, through the completion of the capstone for the Data Science minor, I reflected on my academic journey. Additionally, I connected data science to other disciplines, eventually compiling this portfolio, which showcases my growth in the six core competencies of Data Science. My goals in this portfolio included reflecting on my development of Data Science core competencies, demonstrating my ability to collect, analyze, and communicate data effectively, connecting my Data Science coursework to broader academic and professional contexts, and compiling and presenting a portfolio of work that highlights my skills and achievements. I think I’ve done this well and am grateful for the teachers and classmates who helped me achieve my goals.