Applied Data Science Portfolio Summary Paper

Wesley Stevens

Table of Contents

Introduction 3

IST 659: Data Warehousing 3

Project Overview 3

Reflections & Learning Goals 3

IST 664: Natural Language Processing 4

Project Overview 4

Reflections & Learning Goals 4

IST 707: Data Analytics 4

Project Overview 4

Reflections & Learning Goals 5

Conclusion 5

References 6

Introduction

The Applied Data Science program at Syracuse University pushed me to utilize many data science techniques in unique ways on a variety of datasets. As a part of this program, I learned how to deal with data in all parts of its lifecycle and am now confident in my abilities to parse raw data, wrangle, pipeline, analyze, and create algorithms that will help drive business goals. In this paper, I will delve into three data science projects that I completed during my tenure here.

IST 659: Data Warehousing

Project Overview

This course was the highlight of my entire graduate experience, and I can say that the final project deliverable was the most rewarding to complete. In this course, we worked in teams, using the Kimball method, to build a data warehouse in order to combine two very different databases. The synthetic data came from two imaginary companies called Fudgemart and Fudgeflix (modeled after Walmart and Netflix). We identified five different business objectives that this warehouse could help achieve: Sales, Order Fulfillment, Inventory Levels, Customer Acquisition, and Reviews. For the sake of time, we limited our data warehouse creation and analysis to Order Fulfillment [1].

My team and I designed a database using a Bus Matrix and detail-level dimensional modeling, making sure to avoid extraneous information not related to order fulfillment. Then, in Visual Studio, we used SQL queries and ETL to automate moving the necessary data from the two exterior databases into a staging server and then into the data warehouse. Our programming was such that as new data entered the old databases, it would automatically be imported into the data warehouse without further effort or complication. Additionally, by using a data warehouse, we were able to ensure data privacy by only giving the warehouse the data it needs to evaluate the business processes.

Once the data warehouse was complete with dimensions and fact tables and filled out with data, we used Power BI to do some exploratory data analysis. We discovered some interesting trends in each of the companies, including that the lag time between the order and delivery of a product or service increased in the month of July. Another interesting insight we found was that order fulfillment time was directly proportional to the buying power of a client—meaning that these companies are catering to the wealthy. That in and of itself brought up some ethical questions, we would have brought up to the company.

Reflections & Learning Goals

Because of the lessons learned in this course, I understand how to collect and organize data in efficient ways and how to automate that process for a business. I also now have a deep understanding of how databases work, the importance of multi-threading (for speed), and the importance of data security. I now understand some of the many ethical practices to keep in mind when creating databases. Additionally, by using Power BI, I was able to identify trends and recommend data-driven business decisions for improving the order fulfillment process.

IST 664: Natural Language Processing

Project Overview

Natural Language Processing (NLP) has always been a topic of interest to me due to the amount of information contained in text that most data science algorithms can't capture well. So being able to learn and apply modern techniques was incredible. Though the course mainly used the Python package nltk, I chose to compare how Markov Chains compared to modern-day methods (GPT-2) on the task of Natural Language Generation (NLG) [2]. Markov Chains were the state-of-the-art method for Natural Language Generation for many years until computational limits were expanded in the early 2000's [3]. Once GPUs were invented and optimized, GPT-2 was created as the most intelligent neural network of our time [4].

I utilized corpuses from Project Guttenberg—an online repository of books whose copyrights have expired [5]. I pulled and cleaned four books written by Charles Dickens, then coded and trained each of these models with a single task in mind: Predict the next word. The results came out quite comical and interesting in some ways. The main difference between the two methods was that, although neither made grammatical sense, GPT-2 was able to grasp context much better than Markov Chains. Markov Chains could make 2- or 3-words sound like they could have been written by Charles Dickens, but GPT-2 could make a sentence or two that did the same and even fit coherently together.

I concluded that true general artificial intelligence as imagined by science fiction authors is still a ways off, but we have progressed by leaps and bounds in the past decade. The ethical implications of generating text in the voice of any author is also important to consider as it could violate intellectual property right laws, cause confusion on the internet, and more, especially as NLG becomes increasingly sophisticated.

Reflections & Learning Goals

Because of this course, I am confident in my abilities to collect, organize, and utilize any corpus. I have learned how to deal with data in the form of text, which is much more difficult to manage than numeric data, and I have learned how to identify the ethical implications of implementing data science solutions. This course has also allowed me to implement complex algorithms and use them in a unique way; I am confident in my ability to learn about and use new NLP strategies as research continues to progress.

IST 707: Data Analytics

Project Overview

For the final deliverable in this course, I designed and built a replacement for the Pokédex (originally from the fictional world of Pokémon) by using regression and classification techniques [6].

If you are not aware, a Pokédex is a large machine that Pokémon enthusiasts carry with them in the wilderness to catalogue Pokémon (i.e., fictional creatures) [7]. One must encounter, capture, and thoroughly scan a Pokémon with this machine to discover its species, legendary status, base statistics (e.g., attack and defense ability), and more. This process is difficult due to the low rates of quality Pokémon capturing, and can cost Pokémon enthusiasts a lot of time and money.

Using only the Pokémon's physical characteristics, I used a staged approach to first classify its legendary status (legendary Pokémon are rare and valuable). That was completed with an F-measure of 88% using an XGBoost classifier. Then, that data was used in 6 different regression models to predict the Pokémon's base statistics. The algorithms used were Linear Regression, XGBoost, and SVM Regression. I found that different algorithms worked better for different base Pokémon statistics. Each statistic had a MSE score of less than 0.02, meaning that the created models were an excellent fit. I then used that data to predict the Pokémon's type. This was done using an SVM classifier with an F-measure of 83%. Due to its simplicity, my method was shown to be 42 times faster than the traditional Pokédex at determining Pokémon attributes, albeit it did so with slightly less accuracy.

As a part of the deliverable, I described the data through visualizing various features. Of note was Pokémon Type, which revealed severe Type biases. I then developed a data splitting strategy so that I would end up with even amounts of data in each Type category, thereby enabling the predictive modeling to behave as expected.

Reflections & Learning Goals

This course taught me how to spot irregular data points, how to clean them, and how to use regression and classification methods to then create valuable systems. Through this project, I was able to solve many unique challenges posed by small datasets. I overcame the limitations of the data through innovative feature engineering, cross-validated splitting, and model selection. By simply stacking each of these predictive models, I arrived at a pipelined product that performed very quickly and extremely well.

Conclusion

These are but a few of the many interesting and unique projects I completed during my graduate experience. Through them, I have learned how all types of data must be dealt with at every point in its lifecycle, including collecting and organizing data, identifying patterns through visualization and statistical analysis, creating predictive models, developing data-driven recommendations for business decisions, and identifying and discussing ethical implications of data science products.

The ADS program has taught me how to store and wrangle data, how to generate business insights into any dataset, and how to create predictive products to further optimize a business's utility of its data. Because of the ADS program, I am confident that my abilities as a data scientist will allow me to provide value to any business from day one. I am excited for what opportunities the future will hold and the incredible applications I will make.

References

[1] W. Stevens, "GitHub," 9 1 2021. [Online]. Available: https://github.com/sirwes/MLResearchPortfolio/tree/master/IST%20722%20-%20Data%20Warehousing. [Accessed 5 4 2021].

[2] W. Stevens, "GitHub," 9 1 2021. [Online]. Available: https://github.com/sirwes/MLResearchPortfolio/tree/master/IST%20664%20-%20Natural%20Language%20Processing. [Accessed 5 4 2021].

[3] H. Maltby, W. Pakornrat and J. Jackson, "Brilliant," 5 4 2021. [Online]. Available: https://brilliant.org/wiki/markov-chains/#:~:text=A%20Markov%20chain%20is%20a,possible%20future%20states%20are%20fixed.. [Accessed 5 4 2021].

[4] "OpenAI," 14 2 2019. [Online]. Available: https://openai.com/blog/better-language-models/. [Accessed 5 4 2021].

[5] "Project Gutenburg," [Online]. Available: https://www.gutenberg.org/. [Accessed 5 4 2021].

[6] W. Stevens, "Github," 9 1 2021. [Online]. Available: https://github.com/sirwes/MLResearchPortfolio/tree/master/IST%20707%20-%20Data%20Analytics. [Accessed 5 4 2021].

[7] "Pokemon," [Online]. Available: https://www.pokemon.com/us/. [Accessed 5 4 2021].