Syracuse University

Portfolio Presentation

Master of Science
Applied Data Science

Michael Johnson

SUID: 516408416

https://github.com/AirJohnson3/Portfolio_Project

March 18, 2022

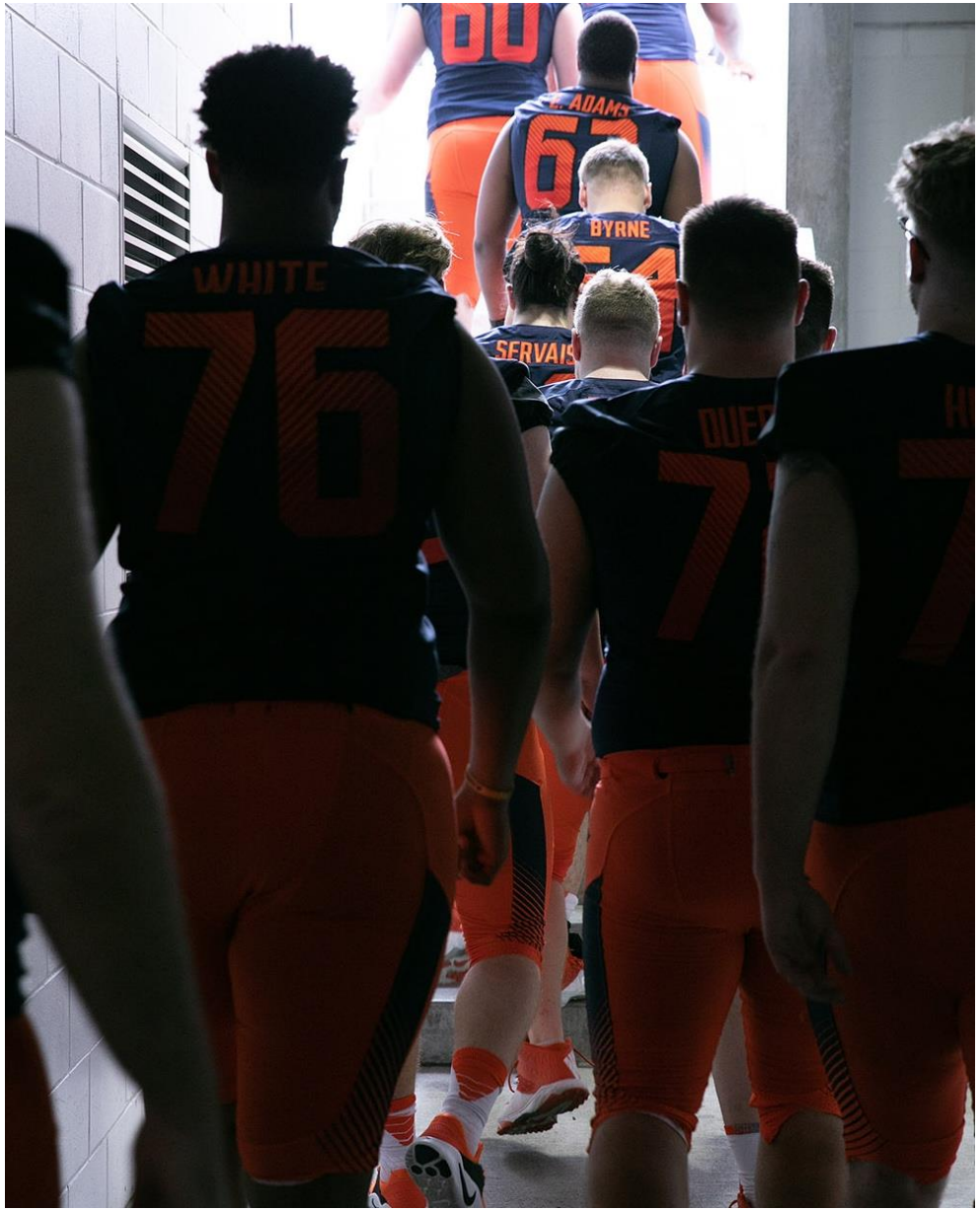
Introduction

- The Applied Data Science program at Syracuse University enabled students to identify actionable insights with data analytics through building visualizations and predictive models relying on a mixture of theory and application (Applied Data Science Master's Degree, 2022). The science behind understanding data drives some of the most important advancements through human history, from collecting the positions of stars to fighting the COVID-19 pandemic through information analysis.
- Understanding and leveraging the data science tools like Python, R programming, and Excel for analyzing data and building models was an integral part of succeeding in the program and will continue to be important for working in the field of data science.
- Coursework showcase:
    - Data Warehouse (IST 722) & Data Administration Concepts and Database Management (IST 659)
    - Natural Language Processing (IST 664)
    - Applied Machine Learning for Data Analysis (IST 707)

Syracuse University
School of Information Studies

Seven Learning Objectives

1. Describe a broad overview of the major practice areas in data science.
2. Collect and organize data.
3. Identify patterns in data via visualization, statistical analysis, and data mining.
4. Develop alternative strategies based on the data.
5. Develop a plan of action to implement the business decisions derived from the analyses.
6. Demonstrate communication skills regarding data and its analysis for relevant professionals in their organization.
7. Synthesize the ethical dimensions of data science practice.



Data Warehouse & Data Administration Concepts and Database Management

IST 722 & IST 659

Data Warehouse & Data Administration Concepts and Database Management Overview

- The IST 722 Data Warehouse course provided concepts, principles, and tools for designing, implementing, and using data warehouses with constructs such as Operational Data Store (ODS), data warehouse, and data marts. The course achieved that goal through defining roles and responsibilities in the design and implementation and stressing the importance of gathering requirements through well-defined guidelines that enable data analysis for an organization.
- IST 659 provided guidance into database management systems examining data structures, file organizations, concepts, and principles of database management systems (DBMS). Taking great care in database design, data modeling, database management, and database implementation enables organizations to employ data analysis techniques that lead to informed decision making.
- Both courses leveraged a portfolio of SQL Server tools that included SQL Server DBMS, SQL Server Integration Services (SSIS), and SQL Server Analysis Service (SSAS) alongside SQL database coding. In addition, the Data Warehouse course also required leveraging Microsoft Power BI for business intelligence to derive solutions to an enterprise process.

Syracuse University
School of Information Studies

Course Project

- The IST 659 project introduced students to developing conceptual and logical database models that drive database design and implementation. Those concepts were instrumental in Data Warehouse as the project centered on developing a data warehouse from two fictional companies merging databases. Successfully guiding the merge required ensuring minimal loss in business processes throughout the transition and that all aspects of the business remained operational.
- To achieve those goals, the project stepped us through understanding the company data using dimensional models that expressly defined the type of data and how that data could be used in the organization. SQL code was essential for creating the data warehouse and extract, transform, and load (ETL) techniques helped in populating the data warehouse from external data sources.
- Once the data was loaded into the data warehouse, we leveraged business intelligence solutions through Power BI to understand the performance details behind the order fulfillment business process. This analysis drove the final conclusions and we made suggestions to key business stakeholders to retain customers and save money through reduced shipment times.

Syracuse University
School of Information Studies

Reflection & Learning Outcome

- From a data engineering standpoint, the data warehouse and database management objectives listed above help in efficiently capturing, organizing, and promoting adequate data storage for a successful business warehouse. From a data science lens, knowledge around these objectives is critical for understanding data and performing successful data analysis.
- The projects in both classes also highlighted some important aspects to keep in mind when constructing a database or data warehouse:
    - Gathering information from all parties in the organization prior to starting a project helps in capturing every bit of the organization's information needs.
    - Conceptual and logical designs are critical to developing a well thought out database or data warehouse and serve as the linchpin for effectively communicating with both technical developers and business-focused individuals within an organization.
    - During the course projects, both professors instructed students to make assumptions about the data and how organizations use the data. However, effective communication with key stake holders is an absolute requirement during the development of databases and data warehouses in the real world.

Syracuse University
School of Information Studies

Natural Language Processing

IST 664

Natural Language Processing

- The aim of the Natural Language Processing (NLP) course is to develop an understanding of how to process written text and produce a linguistic analysis used in various applications. The course primarily covered the techniques of NLP in the levels of linguistic analysis, going through tokenization, word-level semantics, part-of-speech tagging, syntax, semantics, and the higher discourse level. The course also included NLP techniques, such as information retrieval, question answering, sentiment analysis, summarization, and dialogue systems, in applications.
- Goals for this course included demonstrating the levels of linguistic analysis and the computational techniques and process text through the Natural Language Toolkit (NLTK) as use for NLP in real-world applications.
- Python was instrumental in conducting linguistic analysis and was the primary tool throughout the course.

Syracuse University
School of Information Studies

Course Project

- The IST 664 project worked through the course objectives by having us walk through steps to process and analyze movie review text in a meaningful capacity to demonstrate knowledge in linguistic analytics and to attempt to classify movie reviews based on sentiment ratings.
- The first step involved processing, tokenizing, stemming, and lemmatizing the text through Python scripts that imported the text and automatically filtered based on various criteria that were compared against each other for the best model.
- The second step required producing the text features by writing feature functions in Python and using the NLTK Naïve Bayes classifier to train and test a classifier on the created feature sets. Cross-validation techniques to obtain precision, recall and F-measure scores helped in identifying the best model and feature set.
- In addition to the NLTK Naïve Bayes classifier, we also leveraged several different models from SKLearn: Decision Tree, Random Forest, Support Vector Classifier, Linear Support Vector, Logistic, Stochastic Gradient Descent, and Multinomial and Bernoulli Naïve Bayes models. The final models used a bigram bag-of-words and part of speech feature sets with Logistic Regression and Bernoulli Naïve Bayes.

Syracuse University
School of Information Studies

Reflection & Learning Outcome

- The NLP course provided a challenge for developing a unique and valuable skillset in a rising branch of data science designed to increase knowledge for the logistical properties in evaluating unstructured text data. NLTK as the base library for analysis requires users to define the text processing methods needed in generating insight from text mining. The final project in the course pulls elements from each lesson using a real-world example based on evaluating customer reviews for an organization. The project highlighted some important aspects to keep in mind when developing an NLP model:
    - Understand the text and methods needed for processing that text is integral to developing a solution. Having the end-state in mind is a necessary part before construction the solution and will save time in completing a project.
    - Identify the scale of the project prior to starting to set limitations in text analysis to improve on the ability to properly estimate the time needed to complete the project.
    - Experimenting with a variety of feature sets and models is necessary in getting to the best result. This sentiment holds true across all aspects of data analysis and evaluating unstructured data holds no differences in that regard.

Syracuse University
School of Information Studies

Applied Machine Learning for Data Analysis

IST 707

Applied Machine Learning for Data Analysis

- The Applied Machine Learning course was my first introduction to data mining techniques through getting familiar with real-world applications, challenges involved in these applications, and future directions of the field. This course covered popular data mining methods for extracting knowledge from data and include the principles and theories of data mining methods and applying data mining to problems. The focus of this course was to understand data and how to formulate data mining tasks to solve problems with machine learning.
- The course also includes understanding the key tasks of data mining: data preparation, concept description, association rule mining, classification, clustering, and data analysis.
- R programming was the primary tool used in this class with some open-source software packages to help with the modeling and visualizations.

Syracuse University
School of Information Studies

Course Project

- The project for this course required defining a problem on the dataset and describe it in terms of its real-world organizational or business application. We chose to work on a basic recommendation system to help with picking out movies based on the ratings and IMDb scores.
- Three primary techniques made up the basis for analysis, exploring the data set, and building a predictive model that centers on the IMDb score and a parental rating classifier.:
    - Clustering and Classification on the description text help derive the parental rating and serve as a basis for comparing different descriptions.
    - Association Rule Mining was used to suggest what genre of movies or TV shows a viewer should watch next based on their most recently viewed.
    - Additionally, we used a predictive solutions using Movie Length, Genre, Parental Rating, and TV Show or Movie as part of a Support Vector Machine (SVM), KNN, Random Forest, or Decision Tree model.

Syracuse University
School of Information Studies

Reflection & Learning Outcome

- The Applied Machine Learning course was the next step up in learning data science through the addition of various algorithms and data analysis techniques required for machine learning. Relying on R programming as the primary tool, the course highlighted methods to evaluate data and understand the type of analysis necessary for the specific data. The following were important aspects to keep in mind when evaluating and understanding data:
    - Ensure the data analysis has a logical flow from the start of evaluating and exploring a dataset to the end. Like many of the projects in this portfolio, having an idea in mind for the end solution elevates the exploration and analysis of data.
    - Breaking down the problem into smaller steps is an essential strategy for tackling larger problems. Knowing the breakdown for each problem helps in focusing efforts on stepping blocks to achieve the desired goal.
    - Along the same lines as the previous point, the intention for this project centered on a Netflix-like recommendation system but the data was not adequate to hit that lofty goal. Being able to pivot to a new goal after the exploration phase is a viable option and having a result not meeting the desired end state is perfectly fine if there are follow-up actions the organization can take to improve the analysis.

Syracuse University
School of Information Studies

Conclusion

- The reflections contained within this portfolio report exemplify the significant areas of knowledge gained throughout the Applied Data Science program at Syracuse University. Beyond the tools learned in this program, the lessons and strategies developed for achieving solutions derived from data analysis show in the careful development of projects integral for permanent success in the field.
- The projects outlined above represent major practice areas in the field of Data Science and highlight the skills obtained in using widely available tools in conjunction with statistical principles. Collection and organization of data using the data warehouse and database principles drive the subsequent analysis and visualization of data through displaying the necessary strategies for handling organizational data. Without these steps, businesses would struggle in achieving functional decision-making and would be unable to take advantage of key insights into business processes.
- The theory behind the methods within data science augment the application of data the science life cycle which include collection, exploration, visualization, analysis, and modeling for enterprise-level operations and is essential to successful decision-making.

Syracuse University
School of Information Studies

Syracuse University

Thank you!

Michael Johnson
SUID:
https://github.com/AirJohnson3/Portfolio_Project
March 18, 2022