A drawing of a cartoon character

Description automatically generated

Jeff Levesque

https://github.com/jeff1evesque/ist-exit-portfolio | Final Project

Data Science: Portfolio

2023

Table of Contents

[Introduction 2](#_Toc127804879)

[Capstone Projects 2](#_Toc127804880)

[Conclusion & Follow-up Interests 3](#_Toc127804881)

# Introduction

This paper provides a summary of requirements met for the MS in Applied Data Science. A collection of three projects was chosen to help portray the learning achievements made during the IST program at Syracuse University. Various techniques ranging from data mining, exploratory data analysis (EDA), natural language processing (NLP), as well as predictive and financial modeling were performed. This often entailed using languages such as python, R, hadoop, as well as basic systems and cloud engineering. Since each IST course was roughly 10 weeks long, performing a meaningful project was sometimes challenging. Nonetheless, each course and respective project provided valuable experience in making actionable insight from collected data.

# Capstone Projects

|  |  |  |
| --- | --- | --- |
| Course | Capstone Project | Skills |
| FIN-654  Financial Analytics | **Portfolio Analysis**  Chosen stock tickers were analyzed to determine optimal portfolio allocation, using financial and time series modeling | Python, R, time series analysis, financial analysis, Shiny Dashboard |
| IST-664  Natural Language Processing | **Chatbot**  EDA was initially performed to study the data distribution. A trained classifier was ensembled with an LSTM/NMT model to producer the overall chatbot experience | Python, MongoDB / Hadoop, Jupyter Notebook, Time series analysis, Classification analysis, Natural language processing |
| IST-736  Text Mining | **Stock Market Sentiment Analysis**  An attempt was made to determine whether sentiment from financial analysts can predict the stock market. Topic modeling was performed to determine most relevant stock tickers. Sentiment analysis was performed on the same financial analyst tweets and were coupled with corresponding stock ticker price using granger analysis to determine whether sentiment could predict stock price | Topic modeling, Sentiment analysis, Time series analysis, Classification analysis, Signal analysis, Data mining, AWS, Jupyter Notebook |

# Conclusion & Follow-up Interests

As discussed above, IST-664 involved an attempt to create a chatbot through numerous ensembled machine learning (ML) modeling on a personal laptop. While some parts were functional, the overall endeavor was greatly restricted by compute resources. On comparison, recent release of ChatGPT reportedly costs $100,000 a day to run[[1]](#footnote-1). Desired outcomes in “Data Science” are often at mercy of available data and compute resources. What I have learned generalized from the IST-664 chatbot vs. ChatGPT, my interests towards Cloud Native technologies supporting Big Data and streaming analysis have grown immensely. As retrospect, it’s interesting to see the evolution of “Data Science”. While practitioners perform varying degree of EDA to help contextualize problems for humans, we’re in the age where AI/ML frameworks are just beginning to dynamically solve problems orders of magnitude more sophisticated.

For these reasons, I have taken lessons learned from this program, and have expanded into the development of a platform (jefflevesque.com) to aggregate streaming data to facilitate generating and sharing ML models. For example, IST-736 have become a small part of the overall effort – just one stream, and one datalake of many that people can access. This segment actually ingests roughly 200 stock ticker price every minute during the business day, eventually consumed into a parquet partitioned datalake. To expand on FIN-654, candlestick analysis has been devised as an Apache Flink application on the same ingest stream. While the exact streaming codebase is private, an example demo codebase[[2]](#footnote-2) has been publicly released. Future plans may include integration of FIN-654 concepts including the Efficient Frontier as well as the Markowitz model. However, a greater desire of adding additional data streams or developing neural networks may take precedence.

Various fields within “Data Science” often try to visualize data to help contextualize a problem set. It will be interesting to see whether simple data science questions become less prevalent with time. In FIN-654, a staple component of the course was R with Shiny dashboard. However, it is not unimaginable that in the future, a ChatGPT equivalent can expose an API over the internet to directly answer the actual desired problem set. The IST program at Syracuse has afforded me foundational experience in applied Data Science, with a sharper sense of direction. T.S. Eliot once said ‘The journey not the arrival matters’. While numerous learning objectives have been met in this program, my journey as practitioner has only just begun.

1. https://www.ciocoverage.com/openais-chatgpt-reportedly-costs-100000-a-day-to-run/ [↑](#footnote-ref-1)
2. https://github.com/jeff1evesque/kinesis-analytics-demo [↑](#footnote-ref-2)