Learning Outcomes Writeup

Abhi Chakraborty
abchakra@syr.edu

SUID - 684856880

Applied Data Science Learning Outcomes and Course Deliverables

The Applied Data Science program at Syracuse University has equipped me with the skills and knowledge to analyze complex data, develop data-driven strategies, and apply machine learning and statistical techniques to solve real-world problems. This document links the learning objectives of the program with my coursework and projects from four selected courses: IST 687 (Introduction to Data Science), IST 686 (Quantitative Reasoning for Data Science), IST 664 (Natural Language Processing), and FIN 654 (Financial Analytics). I will discuss how my work in these courses contributed to my specialization, highlight my areas of strength and challenge, and outline my plan for lifelong learning in this evolving field.

IST 686 - Quantitative Reasoning for Data Science

IST 686 provided me with a strong foundation in quantitative reasoning, statistical modeling, and data-driven decision-making. This course emphasized the importance of statistical inference in real-world applications, covering both frequentist and Bayesian methodologies. Throughout the semester, I worked on assignments involving probability distributions, hypothesis testing, confidence intervals, regression analysis, and Bayesian inference, which played a crucial role in my understanding of statistical decision-making.

One of the key learning experiences in this course was working with real-world datasets, particularly those related to public health and policy decision-making. I became proficient in data wrangling and transformation using R, leveraging libraries like dplyr, ggplot2, and tidyverse for effective data manipulation and visualization. I also gained experience in designing and implementing hypothesis tests to validate statistical claims, ensuring data-driven insights were both meaningful and actionable.

For the final exam, I conducted an extensive statistical analysis on vaccination rates in California schools. This required implementing both frequentist and Bayesian inference methods to evaluate the impact of vaccination policies across different school districts. The project demonstrated my ability to apply the following concepts:

• **Data Cleaning & Preparation:** Processed large-scale vaccination datasets, handled missing values, and ensured data consistency using dplyr in R.

• **Exploratory Data Analysis (EDA):** Visualized data distributions, examined trends, and identified outliers to ensure the dataset was suitable for statistical testing.

• **Inferential Statistics:** Performed t-tests and ANOVA to compare vaccination rates across different school districts, identifying statistically significant differences.

• **Bayesian Analysis:** Applied Bayesian inference techniques to estimate probabilities and credibility intervals, ensuring robust conclusions despite limited data in some regions.

• **Decision Making Based on Data:** Developed policy recommendations based on statistical findings, helping to inform legislators on potential public health interventions.

A key challenge in the final project was ensuring the reliability of Bayesian inference, especially given the limited sample sizes in some districts. I overcame this by carefully selecting priors and using Markov Chain Monte Carlo (MCMC) methods to refine posterior distributions. Through this, I gained a deeper appreciation for the strengths and limitations of both frequentist and Bayesian approaches.

The final deliverable was a comprehensive report where I not only presented my statistical findings but also wrote the entire R code used for the analysis. This demonstrated my ability to integrate statistical methodologies with coding proficiency, producing a well-documented and reproducible data analysis workflow. The experience strengthened my ability to communicate complex statistical results effectively, a crucial skill for data scientists working in interdisciplinary fields.

Overall, IST 686 equipped me with the quantitative reasoning skills necessary to make data-driven decisions in a rigorous and statistically sound manner. The course has been instrumental in my ability to critically evaluate data, apply advanced statistical techniques, and contribute meaningfully to policy-related and real-world analytical challenges.

IST 664 - Natural Language Processing

IST 664 provided me with a deep understanding of natural language processing (NLP) techniques and their applications in text analysis. Throughout the course, I worked on various NLP tasks, including text preprocessing, feature engineering, and classification models, gaining hands-on experience with libraries such as NLTK and scikit-learn in Python. The course covered foundational concepts such as tokenization, stemming, lemmatization, part-of-speech (POS) tagging, and sentiment analysis, all of which were essential for my final project.

For my final NLP project, I conducted **sentiment analysis on Twitter data**, classifying tweets into positive, negative, and neutral sentiments. The project demonstrated my ability to apply multiple NLP techniques and align them with real-world text analysis challenges. My learning in the course was reflected through the following key aspects:

• **Data Processing:** Tokenized text, removed noise (URLs, Twitter handles, special characters), and converted data into structured formats to improve model performance.

• **Feature Engineering:** Experimented with n-grams (unigrams and bigrams) and POS tags to extract meaningful text features that contribute to sentiment classification.

• **Model Development & Evaluation:** Built and evaluated classification models using Naïve Bayes, logistic regression, and support vector machines (SVM). Precision, recall, and F1-score were used as performance metrics, with a focus on handling imbalanced datasets through weighted scoring.

• **Cross-Validation:** Used k-fold cross-validation to ensure robust evaluation of the sentiment classification model.

A key takeaway from the project was the importance of **feature diversity** in sentiment analysis. By combining different feature extraction techniques, I improved the classifier's accuracy and robustness. One major challenge was **handling imbalanced datasets**, where neutral sentiments were significantly overrepresented. To address this, I implemented **stratified sampling and class weighting**, ensuring the model effectively captured sentiment variations.

By the end of the course, I had gained proficiency in **text mining, classification, and evaluation of NLP models**, strengthening my ability to extract insights from textual data. The final report documented my findings along with the Python code used to preprocess data, build models, and evaluate performance, demonstrating my ability to integrate theoretical NLP concepts with practical applications.

FIN 654 - Financial Analytics

FIN 654 introduced me to **financial modeling and data-driven decision-making in investment analysis**. The course covered essential topics such as **portfolio optimization, risk management, quantitative trading strategies, and performance evaluation**. By working on real-world financial datasets, I learned how to apply statistical techniques and machine learning models to assess market trends and optimize investment portfolios.

A key component of the course was the **VSE (Virtual Stock Exchange) simulation**, where I managed a trading portfolio, executing investment strategies in real-time market conditions. This hands-on experience helped me develop critical skills in:

• **Portfolio Optimization:** Applied the **Efficient Frontier model** and **Modern Portfolio Theory (MPT)** to optimize risk-adjusted returns. Used historical market data to calculate expected returns, standard deviations, and correlations between assets.

• **Quantitative Trading Strategies:** Designed and tested trading strategies based on **sector trends, moving averages, and momentum indicators**. Implemented backtesting to evaluate strategy performance.

• **Risk Management:** Assessed portfolio volatility and implemented **stop-loss strategies** to mitigate downside risks. Analyzed beta values to understand asset sensitivity to market fluctuations.

• **Performance Analysis:** Measured investment performance using **Sharpe ratio, Sortino ratio, and transaction efficiency metrics**, refining strategies to maximize returns while controlling risk exposure.

One of the biggest challenges in the simulation was **managing short positions in volatile sectors**, requiring me to adapt strategies based on **real-time market fluctuations**. This experience taught me the importance of **dynamic strategy adjustments** and **risk mitigation in high-volatility environments**.

By the end of the course, I had gained expertise in **financial modeling, trading strategy design, and investment risk assessment**. The final project and report documented my **trading performance, risk-return tradeoffs, and data-driven investment decisions**, demonstrating my ability to apply analytical techniques to real-world financial markets.

IST 687 - Introduction to Data Science

IST 687 was instrumental in developing my **data science and analytical skills**, covering **data wrangling, exploratory data analysis (EDA), statistical modeling, and machine learning**. Throughout the course, I worked on multiple projects, including data visualization, regression analysis, and predictive modeling.

For the final project, my team and I worked on **'Data Analysis and Strategies for Sustainable Energy Usage'**, where we analyzed electricity consumption trends to help an energy provider, **eSC**, mitigate peak demand during summer months. The project integrated multiple datasets and statistical techniques to derive actionable insights. The key components of the project included:

• **Data Collection & Preparation:** Integrated multiple datasets, including **household energy usage, weather conditions, and static house data**. Preprocessing steps involved handling missing values, aligning timestamps, and filtering data to focus on **July, the peak energy demand month**.

• **Exploratory Data Analysis:** Examined key variables such as **temperature, humidity, wind speed, and solar radiation** to identify patterns in energy consumption. Used **scatter plots, histograms, and heatmaps** to visualize energy demand trends.

• **Model Building & Optimization:** Developed a **multiple regression model** to estimate energy demand based on **temperature, humidity, and radiation levels**. The model achieved an **R-squared value of 0.6749**, indicating that **67.49% of the variability in electricity usage** could be explained by environmental factors. Applied **polynomial terms and interaction effects** to improve prediction accuracy.

• **Predictions & Insights:** Simulated a **heatwave scenario** to project electricity demand under extreme weather conditions. To aid decision-makers, we built an **interactive Shiny application**, allowing users to adjust environmental factors and visualize their impact on energy demand.

• **Actionable Strategies:** Suggested targeted energy efficiency programs, **peak load management strategies, and policy recommendations** such as insulation incentives and green space integration to reduce electricity consumption.

This project significantly enhanced my ability to work with **large-scale time-series data, conduct regression modeling, and extract actionable insights**. A key challenge was ensuring the **robustness of our predictive model under different conditions**, requiring **iterative feature engineering and model refinements**.

By the end of the course, I had gained expertise in **data integration, predictive modeling, and energy demand forecasting**. The final project report documented our findings, methodologies, and the **R code used for data processing, modeling, and visualization**, showcasing my ability to apply data science techniques to solve real-world challenges.

Lifelong Learning and Future Goals

My studies in **Applied Data Science** have provided a solid foundation for **analyzing complex datasets, applying statistical techniques, and leveraging machine learning** to derive actionable insights. The technical skills and methodologies I learned throughout my coursework have been instrumental in my ability to **process, analyze, and model large datasets** in real-world applications.

Currently, in my role as a **Data Engineer / PySpark Developer Intern**, I actively apply **data processing, transformation, and machine learning techniques** learned during my academic journey. Working with **big data technologies such as Apache Spark and distributed computing frameworks**, I handle **large-scale ETL (Extract, Transform, Load) processes, optimize data workflows, and ensure efficient data pipelines for analytics and reporting**. My expertise in **R and Python, combined with SQL and Spark**, allows me to perform **data wrangling, feature engineering, and predictive modeling at scale**.

As I continue to develop my professional skills, the knowledge acquired through my Applied Data Science coursework will play a crucial role in shaping my career trajectory. If my current internship **transitions into a full-time opportunity**, the advanced **data processing, machine learning, and financial analytics techniques** I have mastered will further strengthen my ability to contribute effectively in a **data-driven organization**. The intersection of **big data, predictive modeling, and financial analytics** will allow me to **enhance decision-making strategies, improve data infrastructure, and optimize real-time data processing for business insights**.

To **stay at the forefront of data science and financial analytics**, I plan to continue learning through:

• **Advanced Financial Modeling:** Enrolling in **CFA Level 1** to deepen my understanding of **investment analysis, risk management, and quantitative trading strategies**. This will complement my data science expertise, allowing me to build **algorithmic trading models and optimize portfolio strategies** based on predictive insights.

• **Deep Learning for NLP:** Exploring **transformer-based models (BERT, GPT, LLaMA)** to advance my **natural language processing** capabilities. These techniques will help refine **sentiment analysis, information retrieval, and automated financial analysis**, which are crucial in algorithmic trading and market forecasting.

• **Big Data Processing & Optimization:** Enhancing my **proficiency in distributed computing frameworks like Apache Spark, Databricks, and cloud-based data platforms**. This will allow me to efficiently handle **high-volume financial transactions, large-scale time-series forecasting, and real-time data analytics** in enterprise environments.

Through **continuous self-study, professional certifications, and hands-on application in real-world scenarios**, I aim to **expand my expertise in data engineering, machine learning, and financial data analytics**. By integrating **data science principles with financial analytics and big data engineering**, I am positioning myself for a **long-term career in data-driven decision-making, quantitative finance, and AI-powered trading strategies**.