Syracuse University

School of Information Studies

Master of Science in Applied Data Science

**Portfolio Milestone – Spring 2024**

Blessy A. Thomas

SUID#: 520121045

Bthoma16@syr.edu

**Table of Contents**

1. Introduction .......................................................................................................... 3-4

2. IST 659: Data Administration Concepts & Management .......................................... 4-7

* 1. Project Overview

2.2 Real-World Application and Learning Outcomes

3. IST 707: Applied Machine Learning ..................................................................... 7-12

3.1 Project Overview

3.2 Real-World Application and Learning Outcomes

4. IST 652: Scripting for Data Analysis ................................................................... 12-17

4.1 Project Overview

4.2 Real-World Application and Learning Outcomes

5. IST 664: Natural Language Processing ............................................................... 17-21

* 1. Project Overview
  2. Real-World Application and Learning Outcomes

6. Conclusion .............................................................................................................. 21-22

1. **Introduction**

This reflective narrative encapsulates my academic journey as a graduate student in the Master of Science in Applied Data Science program at Syracuse University. As a student with a biology background, my transition into data science has been both demanding and fulfilling. The variety of projects and assignments I have engaged in throughout my studies have not only expanded my comprehension of the field but also endowed me with a unique skill set that I believe will be instrumental in addressing real-world, data-driven challenges.

My academic path has been characterized by a series of rigorous courses, each contributing to my burgeoning expertise in data science. In IST 659: Data Administration Concepts & Management, I delved into the complexities of SQL, establishing a robust foundation in database management. IST 707: Applied Machine Learning ushered me into the realm of R programming, where I learned to construct and apply machine learning models. In IST 652: Scripting for Data Analysis, I refined my Python skills, which are crucial for effective data analysis. Lastly, IST 664: Natural Language Processing further broadened my Python proficiency, specifically in the context of processing and analyzing human language data.

This portfolio is a testament to my academic journey, highlighting the various projects that have enabled me to apply and refine these skills. Upon completion of the program, I or any other student should be able to:

* Collect, store, and access data by identifying and leveraging applicable technologies
* Create actionable insight across a range of contexts (e.g. societal, business, political), using data and the full data science life cycle
* Apply visualization and predictive models to help generate actionable insight
* Use programming languages such as R and Python to support the generation of actionable insight
* Communicate insights gained via visualization and analytics to a broad range of audiences (including project sponsors and technical team leads
* Apply ethics in the development, use, and evaluation of data and predictive models (e.g., fairness, bias, transparency, privacy)

As I anticipate my graduation in Spring 2024, I am eager to share the insights I've acquired and the projects that have shaped my understanding of applied data science.

1. **IST 659: Data Administration Concepts & Management**
   1. **Project Overview**

In the Data Administration Concepts & Management course led by Mr. Chad Harper, students were tasked with creating a functional system complete with a database to address a practical issue. The project involved developing a robust relational database for a hypothetical grocery store, Whole-ER Foods, which was struggling to maintain its market position. Despite attracting customers with its eco-friendly and healthy product range, the store faced challenges in meeting the high demand. Whole-ER Foods, which restocked its inventory every week from a bulk supplier on Fridays, initially managed its inventory using Microsoft Excel. The goal of the project was to transition to a relational database system to enhance inventory management and meet consumer demand, reflecting a common real-world problem faced by grocery retailers. During the development of the database, the scope of the implementation was reduced to product inventory, shipments, transactions, and general customers. This required the collection of fictional data from sources such as store audits, warehouse receipts, invoices, POS (point of sale) data, and purchase orders. Examples of required data groups were products, vendors, orders, transactions, etc.

To manage the data relationships among various entities such as store, checked\_out, address, vendor, vendor\_product, product, purchase, purchase\_product, and product\_category, conceptual and logical models were constructed (Figure 1). Azure Data Studio, in conjunction with Microsoft SQL Server, was utilized to establish tables that enabled efficient data population and analysis. Furthermore, stored procedures were implemented to retrieve and display product details including names, prices, quantities, product\_id, and product\_category\_id within the database (Figure 2). The ability to search for products using their product\_id and product\_category\_id facilitated their identification in the database. These data fields are instrumental for Whole-ER Foods to gain insights into inventory levels and product availability, ensuring the store can meet customer demand effectively.

A diagram of a product

Description automatically generated

Figure 1: Logical Model

A screenshot of a computer

Description automatically generated

Figure 2: Relational Database

* 1. **Real-World Application and Learning Outcomes**

This project was instrumental in preparing me for real-world challenges by providing hands-on experience in designing and implementing a relational database system to address a practical business problem similar to those encountered by grocery retailers. Transitioning from a basic Microsoft Excel inventory management to a sophisticated SQL-based relational database, I learned to tackle inventory management and consumer demand issues, which are pivotal in the retail sector. My skills in data modeling were sharpened as I developed conceptual and logical models to manage data relationships among various entities, utilizing tools like Azure Data Studio and Microsoft SQL Server for database implementation. The creation of stored procedures for data retrieval and display enhanced my ability to perform efficient data analysis, a skill that is highly valued in many business contexts. This experience aligned with the course objectives, which included analyzing business problems, designing data-oriented solutions, constructing database objects and queries using SQL, and evaluating the effectiveness of database management systems (DBMS) in information systems. The project provided a thorough understanding of data management strategies and the practical application of database design principles, equipping me for future data administration challenges.

The project also met multiple program objectives by involving the gathering, storage, and retrieval of data using appropriate technologies. Data was sourced from simulated audits, receipts, invoices, point-of-sale records, and orders, and was stored in a relational database, marking a progression from the original Excel-based system. Access to this data was facilitated by stored procedures in Azure Data Studio and SQL Server, which enabled the display of product information. The project offered practical insights within a business context by utilizing the entire data science process. It examined the challenges faced by Whole-ER Foods, leading to the development and deployment of a data-driven solution that enhanced inventory management and met customer demand. The system's search functionality, using product\_id and product\_category\_id, allowed for efficient product location in the database, providing Whole-ER Foods with valuable information on stock levels and product availability. Additionally, the project's use of fictional data adhered to the educational goal of applying ethical principles in data and predictive model development, addressing key ethical issues such as fairness, bias, transparency, and privacy. The decision to use fictional data for Whole-ER Foods' database ensured the protection of individual privacy by mitigating risks associated with handling actual personal data.

1. **IST 707: Applied Machine Learning**
   1. **Project Overview**

In the Applied Machine Learning course overseen by Dr. Jeremy Bolton, students engaged in a practical exercise to tackle a real-world issue in data mining. Techniques such as association rule mining, clustering, support vector machines (SVM), etc. were applied. For this project, a dataset sourced from Kaggle was used, which comprised 33 variables across 395 entries. This dataset included details on student performance over three terms, demographic data, and various social and school-related attributes. The challenge was that schools were struggling to proactively identify students in need of extra support. To address this, two Portuguese schools, sharing similar class sizes, curricula, and weekly instructional hours, initiated an analysis with a focus on high school math students. The objective was to uncover factors leading to student failures and devise strategies for early identification and support of those at risk, specifically targeting students with grades below 55%. To facilitate this, a binary indicator was created within the study, marking a '1' for grades at or below 55% and a '0' for grades above this threshold.

The data was cleaned and pre-processed using R Studio. Despite the absence of missing data, certain variables such as school, sex, address, famsize (family size), pstatus (parent’s cohabitation status), medu (mother’s education), fedu (father’s education), schoolsup (extra educational support), famsup (family educational support), activities, nursery, higher, internet, romantic, famrel (quality of family relationships), freetime, gout (going out with friends), dalc (workday alcohol consumption), walc (weekend alcohol consumption), and health were transformed to facilitate comprehensive analysis. A secondary data set was also generated from the original, selecting only the top 10 variables based on feature importance. Subsequent exploratory data analysis yielded several insights. It was observed that both schools experienced a decline in grades and an increase in students scoring zero as the academic year advanced (Figure 3). Males appeared to have a marginally higher average than females across the schools. The proportion of students failing was roughly equal in both schools, indicating that the environment may not be a significant factor. The analysis also suggested that paying for additional classes did not necessarily improve performance.

A graph showing the slope downward

Description automatically generated

Figure 3: Decrease in Grades as the Academic Year Progressed (EDA)

Following initial findings, sophisticated analytical techniques such as decision trees, logistic regression, k-nearest neighbors (k-NN), association rule mining, support vector machines (SVM), and random forests were employed to tackle the problem. The initial decision tree, while visually cluttered and only marginally outperforming the no information rate (which assumed all students failed), saw a 12.46% boost in accuracy to 74.64% after pruning and assessing feature importance (Figure 4). Three logistic regression models were compared using only the top ten influential variables; the first model emerged as the most accurate with 71.74%. The k-NN approach, also limited to the top ten variables, selected k=11 based on the highest accuracy, achieving 68.84%. An SVM model, tuned with a radial kernel and a cost parameter of 0.95, reached an accuracy of 81.88% using the same subset of variables (Figure 5). The random forest algorithm, with parameters ntree=500 and mtry=4, and again using the top ten variables, achieved the highest accuracy at 99.28%, although this raised concerns about potential overfitting (Figure 6). Among all models, the random forest stood out for its accuracy, but with caution against overfitting, followed by the SVM as a strong predictor. Recommendations for the two Portuguese schools to better support struggling students include applying the random forest model after ensuring it is not overfit, and focusing on the top ten variables: failures, freetime, fedu, dalc, walc, schoolsup, absences, age, famrel, and famsize which significantly influence students' final grades.

**A diagram of a tree

Description automatically generated**

Figure 4: Decision Tree Model

**A screenshot of a computer

Description automatically generated**

Figure 5: SVM Model

**A screenshot of a computer

Description automatically generated**

Figure 6: Random Forest Model

* 1. **Real-World Application and Learning Outcomes**

This project prepared me for the real world by providing hands-on experience with the full data science life cycle, from data preprocessing to model evaluation. The use of real-world data to solve a practical problem and identify students in need of support mirrors the challenges faced in many professional settings. The project also emphasized critical thinking in model selection and the importance of ethical considerations, such as avoiding overfitting, which are essential skills in the responsible application of machine learning in real-world scenarios.

The project described above also aligns with several of the program’s learning goals. Firstly, the collection, storage, and access of data were demonstrated through the use of a Kaggle dataset, which was then cleaned and pre-processed using R Studio, reflecting the first learning goal. The creation of actionable insights was evident in the analysis of student performance data to identify at-risk students, aligning with the second learning goal. Visualization and predictive models were applied, as seen in the use of decision trees, logistic regression, and other sophisticated techniques, fulfilling the third learning goal. The use of R, a programming language, to support the generation of actionable insights was also highlighted, meeting the fourth learning goal. The ethical considerations of model development, such as concerns about overfitting with the random forest model, touch upon the sixth learning goal related to ethics.

1. **IST 652: Scripting for Data Analysis**
   1. **Project Overview**

In the Scripting for Data Analysis course led by Dr. Gregory Block, students were required to showcase their proficiency in Python scripting by collecting and analyzing data from various sources, both structured and unstructured. Students selected a specific research topic, gathered relevant data, and used it to create summaries, structures, and visualizations. The project involved employing multiple analytical methods to explore the chosen data set. For this project, a Kaggle HR analytics dataset with 38 variables was utilized, which included details like gender, department, education, role, and monthly income of employees within a company. The primary focus was on the 'Attrition' variable, which indicates whether an employee has left the company or remains employed. The analysis aimed to provide insights into human resource analytics and organizational work culture, with the ultimate goal of identifying strategies to reduce employee turnover, which can be costly for companies. The data helped to identify factors contributing to employee attrition and potential measures to mitigate it.

The data analysis project utilized Python for data cleaning, pre-processing, and analysis within Jupyter Notebook and Google Colaboratory environments. The investigation aimed to answer three key questions: the retention differences between younger and older employees, the impact of travel on employee retention, and the retention of top performers within the company. Python libraries such as pandas, matplotlib.pyplot, and numpy facilitated this process. Data exploration revealed issues like duplicate entries and 57 missing values in the 'YearsWithCurrManager' variable. These duplicates were removed, and missing values were imputed with the average of the respective age groups, assuming a correlation between age and years with the current manager. Further analysis corrected a typographical error in the 'BusinessTravel' variable. For the first question, a hypothesis was formed that younger employees were more likely to leave. A stacked bar chart (Figure 7) confirmed this, showing higher attrition rates among the 18-25 age group compared to older groups. The second question addressed the relationship between travel frequency and retention. Contrary to the hypothesis that more travel might lead to longer retention, the data (Figure 8) indicated that employees who traveled frequently had higher attrition rates. Lastly, to determine if top performers were being retained, the analysis focused on those in Research & Development and Sales. Visualizations (Figures 9 and 10) showed that approximately 84% of top performers remained with the company, supporting the hypothesis that top performers were indeed being retained effectively.

A screenshot of a video game

Description automatically generated

Figure 7: Attrition by Age Group

A screenshot of a graph

Description automatically generated

Figure 8: Attrition by Business Travel

A screenshot of a computer

Description automatically generated

Figure 9: Sales Attrition by Performance Level

A screen shot of a computer

Description automatically generated

Figure 10: Research & Development Attrition by Performance Level

A predictive model was developed, assigning 80% weight to certain features deemed most important (Figure 11). The model selected features with an importance score of at least 0.05, which included 'DistanceFromHome', 'PercentSalaryHike', 'JobLevel', 'NumCompaniesWorked', 'EducationField', 'YearsSinceLastPromotion', 'Attrition', and 'OverTime'. The model's accuracy was determined to be 0.826530612244898, marginally surpassing the baseline accuracy of 0.8197278911564626, which assumes no employee turnover. This slight improvement over the baseline suggests the model can provide insights into potential employee departures. Specifically, employees with 'OverTime' scores above 1.5, 'JobLevel' below 1.5, or 'DistanceFromHome' exceeding 17.5 are identified as having a higher likelihood of leaving. Subsequent analysis revealed the company's effectiveness in retaining top talent, with a retention rate above 80%. For frequent travelers, the analysis suggests a higher turnover rate, prompting a recommendation for the company to consider additional rest periods or regional travel assignments to prevent burnout. The data also indicates a trend of younger employees seeking opportunities elsewhere, suggesting that the human resources department could offer incentives like faster PTO accrual or educational benefits to retain this demographic. It is crucial to note the ethical considerations in utilizing this analysis. It should not be used in hiring decisions but rather as a tool to support current employees and ensure their satisfaction. Misuse of this data for hiring could lead to moral and legal repercussions for the organization.

A group of rectangular colored boxes

Description automatically generated with medium confidence

Figure 11: Predictive Model

* 1. **Real-World Application and Learning Outcomes**

In terms of preparation for the real world, this project provided me with a comprehensive experience in data science that is directly transferable to professional settings. The skills in data analysis, visualization, and predictive modeling are highly sought after in various industries. The emphasis on ethical considerations also prepared me to navigate the complexities of real-world data usage responsibly. The ability to derive meaningful insights from data and communicate them effectively is crucial in decision-making processes across business, societal, and political contexts, thus equipping me with the tools necessary for a successful career in data science.

Reflecting on the project, it is evident that the completion of the project aligns with several of the learning goals of the program. Firstly, the project successfully met the goal of collecting, storing, and accessing data by leveraging technologies such as Python, Jupyter Notebook, and Google Colaboratory. The use of a Kaggle HR analytics dataset demonstrates the ability to identify and utilize applicable technologies for data handling. The creation of actionable insights is also apparent, as the project's analysis provided strategies to reduce employee turnover, a valuable insight for business contexts. The investigation into factors influencing attrition rates and the development of a predictive model to forecast potential employee departures are clear examples of generating actionable insights across business contexts. Visualization and predictive modeling were applied effectively, fulfilling another learning goal. The use of Python libraries to create visualizations such as stacked bar charts and the development of a predictive model with an accuracy surpassing the baseline demonstrates the practical application of these techniques to generate insights. The project also showcased the use of Python to support the generation of actionable insights, meeting the learning goal related to programming language proficiency. The detailed data cleaning, pre-processing, and analysis conducted within the Python environment are a testament to this skill. Lastly, the project addressed ethical considerations in the use of data and predictive models. The caution against using the analysis for hiring decisions and the focus on supporting current employees' satisfaction align with the learning goal of applying ethics in data science practices.

1. **IST 664: Natural Language Processing**
   1. **Project Overview**

In the course on Natural Language Processing, led by Mr. Michael Larche, students demonstrated their skills in various aspects of linguistic analysis and natural language processing. Python was used to participate in a classification task, develop features, and conduct several experiments to determine the most suitable features for the selected data set. The data set used for this task was a movie review phrase data set from Kaggle, created for the Sentiment Analysis on Movie Reviews Kaggle competition. This competition used data from the sentiment analysis conducted by Socher et al, which was derived from the original Pang and Lee movie review corpus on the Rotten Tomatoes website. The data set comprised 156,060 phrases, each labeled with one of five sentiments: "negative", "somewhat negative", "neutral", "somewhat positive", and "positive." Despite the competition's conclusion, the training data "train.tsv" and some test data "test.tsv" remained accessible. The training data included phrases, their corresponding sentiment labels, and a SentenceId to identify phrases from the same sentence. The task's objective was to assign a sentiment label to each phrase, with the challenge being the selection of a suitable subset for processing and training. The sentiment labels were numerically encoded as follows: 0 for "negative", 1 for "somewhat negative", 2 for "neutral", 3 for "somewhat positive", and 4 for "positive."

The classification task employed various preprocessing techniques, including tokenization, filtering, and removal of non-alphanumeric words, alongside feature selection methods like bag of words, bigrams, unigrams, etc. Several classifiers were tested, including Naïve Bayes and Sci-Kit Learner algorithms like random forest and logistic regression, across different feature sets with and without preprocessing. The results indicated that preprocessing generally improved classification accuracy (Figure 12). For instance, the sentiment lexicon feature set scored slightly better than the unigram feature set with preprocessing. The Naïve Bayes classifier served as a baseline, with its unigram feature without preprocessing achieving an accuracy of 0.55. The study also explored cross-validation to enhance the reliability of the results, calculating precision, recall, and F-measure scores for a more comprehensive evaluation. A comparison of single-fold performances against all feature sets revealed that SVC and NuSVC classifiers displayed lower accuracies, leading to their exclusion from cross-validation testing. A detailed analysis of mean accuracies for various classifiers and feature sets was conducted, both with and without preprocessing (Figure 13). It was observed that bigram accuracy was almost equal to unigram accuracy for all classifiers, indicating no significant improvement. However, the average POS accuracy and subjectivity accuracy were slightly higher than the unigram accuracy, suggesting the effectiveness of these features. MultinomialNB, BernoulliNB, and logistic regression classifiers outperformed the Naïve Bayes classifier in terms of accuracy.

A table of numbers and lines

Description automatically generated with medium confidence

Figure 12: Comparison of Single-Fold Performances

A table of numbers and a few numbers

Description automatically generated with medium confidence

Figure 13: Summary of Mean Accuracies

* 1. **Real-World Application and Learning Outcomes**

This classification task served as a practical training ground for real-world preparation, offering valuable insights into managing and scrutinizing large data sets, formulating and evaluating predictive models, and effectively conveying the results. The detailed analysis highlighted the crucial role of preprocessing and feature selection in text classification tasks, illustrating the influence of various techniques and classifiers on the precision of sentiment analysis for movie reviews. These competencies are widely relevant in numerous sectors, such as data science, business analytics, and machine learning. Moreover, the task provided a hands-on comprehension of how sentiment analysis can be employed to understand public sentiment, a capability that can be utilized in diverse real-world scenarios, ranging from market research to monitoring social media.

Upon reviewing the classification task, it is evident that the task fulfills several of the learning goals outlined. Firstly, the project involved the collection, storage, and access of data, fulfilling the first learning goal. The data set used was a movie review phrase data set from Kaggle, which was accessed and stored for the task. Python was used to participate in a classification task, develop features, and conduct several experiments to determine the most suitable features for the selected data set, fulfilling the fourth learning goal of using programming languages such as Python to support the generation of actionable insight. Secondly, the project created actionable insight across a range of contexts, fulfilling the second learning goal. The task's objective was to assign a sentiment label to each phrase, which could be used to gain insights into the general sentiment of movie reviews. This could be applied in a business context, for instance, to understand how well a movie is being received by the audience. Thirdly, the project involved the application of visualization and predictive models to help generate actionable insight, fulfilling the third learning goal. The results of the classification task were visualized, and several classifiers were tested, including Naïve Bayes and Sci-Kit Learner algorithms like random forest and logistic regression.

1. **Conclusion**

My academic journey through the Master of Science in Applied Data Science program at Syracuse University has been a transformative experience that has not only enhanced my understanding and skills in data science but also equipped me with a comprehensive skill set to tackle the multifaceted challenges of the data-driven world. Through a series of challenging yet rewarding projects and meticulously designed courses, I have delved into the complexities of SQL in database management, applied machine learning models, refined my Python skills for effective data analysis, and explored the intricacies of natural language processing. As a result, I have successfully met the program’s learning outcomes. This program has allowed me to demonstrate proficiency in collecting, storing, and accessing data using cutting-edge technologies, create actionable insights across societal, business, and political contexts by engaging with the full data science life cycle, and apply visualization and predictive models to generate actionable insights. My ability to effectively utilize programming languages such as R and Python, communicate complex data into understandable and impactful narratives, and uphold ethical standards in the development, use, and evaluation of data and predictive models, ensuring fairness, bias mitigation, transparency, and privacy, has been honed. As I stand on the brink of graduation, I am confident that the skills and insights gained through this program will enable me to make meaningful contributions to the field of data science. This portfolio, a testament to my academic journey, showcases the knowledge and skills that I have acquired through this program as well as my readiness to tackle data-driven challenges and it underscores my commitment to continuous learning and professional development in the ever-evolving landscape of applied data science.