**Nolan Nicholas Arendt**

**Syracuse university | nnarendt@syr.edu | Student ID: 269836144**

Final Project Portfolio

m.S. in Applied data science



**Contents**

**Introduction**………………………………………………….…………………….……………..2

**Scripting for Data Analysis**……………………………………………………………………...2

Project Description………………………………………………………………………..2

Learning Objectives & Outcomes…………………………………………………………3

**Natural Language Processing**……..…………………………………………………………….4

Project Description…………………………………………..……………………………4

Learning Objectives & Outcomes…………………………………………………………5

**Introduction to Data Science**………..…..……………………………………………………….6

Project Description…………………..……………………………………………………6

Learning Objectives & Outcomes………………………….……………………………...7

**Deep Learning in Practice**……………………………………………………………………….8

Project Description……………………………………………...………………………...8

Learning Objectives & Outcomes………………………………………………………....9

**Applied Machine Learning**…………………………………………………………………….10

Project Description………………………………………………………………………10

Learning Objectives & Outcomes………………………………………………………..11

**Conclusion**.……………………………………………….…………………..…………………12

**Appendix**….……………………………………………….…………………………………….13

**Introduction**

In my final portfolio, I will be presenting a complete overview of my academic work and progress through the Applied Data Science program at Syracuse University. This portfolio will showcase my selected final projects from the following five courses: Scripting for Data Analysis, Natural Language Processing, Introduction to Data Science, Deep Learning in Practice, and Applied Machine Learning. These projects that I have selected will demonstrate my ability to collect, store, and access data by leveraging applicable technologies, demonstrating my ability to use various data sources with multiple programming languages and technologies to create actionable insights across various contexts, including societal, business, and political scenarios.

Through this collection of my work, I will also aim to showcase my experience and techniques in using various visualization techniques and predictive models to produce actionable insights that I can provide to businesses and individuals. My ability to use programming languages such as R and Python has been valuable in supporting these insights with code that can be reproduced and built upon for further analysis. In addition, I highlight my commitment to communicating these insights effectively to a broad range of audiences. As a result, this portfolio of my work reflects my dedication to the ethical application of data science principles, particularly in addressing issues related to fairness, bias, transparency, and privacy. **IST 652 - Scripting for Data Analysis**

**Project Description**

In my final project for Scripting for Data Analysis, "Analyzing Viewer Voting Trends in Classic and Modern Cinema," I developed a python script to analyze movie review data with a focus on viewer voting trends across various genres, countries of origin, and demographics. Utilizing datasets from Kaggle, I was able to collect semi-structured and structured data for movies produced between the years of 1894 and 2020. I filtered the data to focus on single-genre films for cleaner analysis and conducted exploratory data analysis (EDA) to identify trends within the data. This project aimed to provide movie producers with insights on historical successes and failures in the movie industry, helping them make informed decisions about future movies.

My analysis revealed several key findings. Drama and Comedy appeared to be one of the most popular genres, both in terms of the number of votes and average ratings. In comparison to other genres, Sci-Fi received lower average scores, indicating a potential area for improvement that directors could focus on. Sentiment analysis of plot descriptions showed that movies with positive sentiments, such as movies with happy endings, tend to receive higher votes, emphasizing the importance of a movie’s emotional engagement with audiences. Additionally, gender-based analysis highlighted that woman regularly rated movies higher than men, with Family movies showing the most significant difference of all genres when compared and analyzed. I was also able to discover a positive correlation between movie duration and average vote, regardless of what genre the movie was, meaning viewers enjoyed rich stories with longer run times. Finally, the country-specific analysis provided valuable insights into how cultural and geopolitical factors might influence movie ratings, offering a comprehensive view of global movie trends and preferences.

**Learning Objectives & Outcomes**

This project demonstrates my ability to collect, store, and access data by utilizing applicable technologies, aligning with the first program learning goal. By using Kaggle as a data source, I effectively gathered datasets and utilized Python's panda’s library to manage and thoroughly preprocess the data to perform further in-depth analysis. This process included data cleaning, feature engineering of columns, and exploratory data analysis, ensuring the dataset's reliability and suitability for further analysis. The project also highlights the creation of actionable insights, as we used the data science life cycle to analyze viewer voting trends and provide valuable recommendations for production executives and movie producers. These insights can help inform strategic business decisions, such as genre selection, plot sentiment, and marketing focus, highlighting the project's impact across societal and business contexts.

In addition, I was able to use visualization techniques and predictive models to generate actionable insights, meeting the third and fourth learning goals. Using Python libraries such as matplotlib and seaborn, I successfully created various visualizations, including bar graphs, scatter plots, and histograms to communicate my findings within the data and results. These visualizations made complex data trends more accessible to a broad audience that may not understand from just looking at the data, thus fulfilling the fifth learning goal. Additionally, my analysis considered ethical implications, such as avoiding bias in genre representation and country representation and ensuring transparency in the data processing methods I used. This approach aligns with the sixth learning goal, emphasizing the ethical development, use, and evaluation of predictive models and data analysis.

**IST 664 - Natural Language Processing**

**Project Description**

In the project "Detection of Spam in Emails," I explored the efficiency of various machine learning techniques in classifying spam from authentic emails using a dataset derived from the Enron corpus. The dataset, balanced with an equal number of spam and authentic emails, underwent several different preprocessing steps, including tokenization, random shuffling, and lowercasing words to ensure consistency across all emails in the dataset. My initial approach used the Naïve Bayes classifier to identify common patterns in the emails, followed by evaluation techniques to refine the model's performance such as grid searching. This phase of the project aimed to identify effective methods for accurately classifying emails to enhance email security in an age where security and data is incredibly valuable.

I extended our analysis to include several machine learning models, such as Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machines to compare their performance metrics in spam detection. Each model was tested and evaluated, with Logistic Regression and Gradient Boosting models achieving the highest accuracy rates of approximately 97.67%. The Gradient Boosting model showed perfect precision in classifying ham emails, making it a realistic option for enterprise-level applications where the accuracy of email classification is critical. This comprehensive examination of multiple models not only highlights my commitment to improving email security but also provides actionable insights for implementing effective spam detection systems in various real-world scenarios and businesses.

**Learning Objectives & Outcomes**

This project exhibits the first program learning goal by demonstrating my ability to collect, store, and access data using applicable technologies. By leveraging the Enron email corpus from Kaggle, I effectively gathered a dataset and used multiple Python packages/libraries for the data preprocessing and analysis. The thorough preprocessing steps, including data balancing, tokenization, and shuffling, ensured that the data was suitable for further and more in-depth machine learning tasks. Additionally, the project highlights my ability to create actionable insights across social and business contexts. The development and evaluation of multiple machine learning models provided valuable insights into the most effective techniques for spam detection, thus improving email security and theoretically informing business strategies for managing email communication systems.

Moreover, my project aligns with the fourth and fifth learning goals by applying visualization and predictive models to generate actionable insights and effectively communicating these insights to a broad audience. Using Python for this project and analysis, I developed several machine learning models, such as Naïve Bayes, Logistic Regression, and Gradient Boosting, and evaluated their performance using cross-validation and confusion matrices. These visualizations made complex data trends accessible and understandable, helping aid informed decision-making. Additionally, I was able to successfully address the ethical implications of my work by ensuring the models were evaluated for fairness and transparency, particularly in avoiding bias in classification. This ethical consideration demonstrates my commitment to the responsible development and use of predictive models, fulfilling the sixth learning goal.

**IST 687 – Introduction to Data Science**

**Project Description**

In the project "Ames IA Housing Market Analysis: 2006-2010," I was tasked with analyzing the Ames, Iowa housing dataset to identify key features that influence the return on investment (ROI) for different housing attributes. Using a dataset from Kaggle containing 2,930 observations and 80 variables, I focused on valuable features such as lot size, year built, neighborhood, exterior quality, basement quality, heating quality, kitchen quality, garage quality, deck square footage, and sale price. By running correlation and linear regression analyses, I aimed to determine the most significant variables influencing sale prices. My findings indicated that kitchen quality, basement height, and exterior quality were among the highest impact factors, while neighborhood and lot area had surprisingly weak correlations with sale price.

The analysis provided actionable insights for potential homebuyers and investors such as myself, highlighting specific improvements that could maximize a buyers return on their investment. For a specific example, upgrading kitchen quality from 'Typical' to 'Good' could increase the sale price by approximately $70,000 to $200,000. Similarly, enhancing exterior quality from 'Typical' to 'Good' could potentially yield an increase of around $90,000 in resale value. These insights help inform strategic investment decisions in the housing market, demonstrating the project's practical applications. The study also highlighted the importance of focusing on internal home updates over external home modifications, such as kitchen and basement improvements, which showed the most significant returns in resale value.

**Learning Objectives & Outcomes**

This project demonstrates my ability to collect, store, and access data by identifying and leveraging applicable technologies, fulfilling the first learning goal. In this project, I utilized the Ames Housing dataset from Kaggle, which provided broad data on housing features and sale prices. By employing R for data cleaning, feature engineering, and analysis, I effectively managed and transformed the data to suit our needs. This project also aligns with the second learning goal as we created actionable insights across a societal context, specifically in the real estate market. By identifying the most impactful features on sale prices, I was able to provide valuable information for homebuyers and investors to maximize their return on investment with the data and analysis to back up my conclusions.

Additionally, my project demonstrates the application of visualization and predictive models to generate actionable insights, addressing the third and fourth learning goals. Using R, I was able to develop various linear regression models and visualize the relationships between housing features and sale prices. These visualizations helped communicate the findings effectively to a broad audience, including potential investors and real estate professionals, meeting the fifth learning goal. Throughout the analysis, I remembered to ensure ethical considerations were met by clearly reporting our methods and findings, sticking to principles of fairness and bias reduction. This approach highlights my commitment to ethical data science practices, fulfilling the sixth learning goal.

**IST 691 – Deep Learning in Practice**

**Project Description**

In the project "Pistachio Type Classification," I explored the use of Convolutional Neural Networks (CNNs) to classify images of pistachios into two species: Kirmizi and Siirt. Utilizing a dataset with over 2,000 images, the primary goal was to develop an accurate model for pistachio species classification to aid in quality control within the food industry. I utilized data augmentation techniques, including zooming, shifting, and shearing, to increase the dataset's diversity and prevent overfitting. The dataset was split into training, validation, and test sets, with a slight imbalance that was within acceptable limits for effective model training.

I then implemented two CNN architectures: EfficientNetB0 and VGG16. EfficientNetB0, pre-trained on ImageNet, was modified with additional layers and dropout for regularization, achieving a test accuracy of 98.46%. However, it showed significant bias towards the Kirmizi species, with no actual predictions of the Siirt species. The VGG16 model, also pre-trained on ImageNet, was fine-tuned for deeper feature extraction, resulting in a test accuracy of 96.31%. While both models demonstrated high accuracy, VGG16 provided a better balance between the two species, highlighting the importance of continuous model evaluation and tuning to address biases and improve classification performance.

**Learning Objectives & Outcomes**

This project demonstrates my ability to collect, store, and access data by using the relevant applicable technologies, fulfilling the first program learning goal. I utilized a dataset of pistachio images from Kaggle and employed Python libraries such as TensorFlow and Keras for data preprocessing, augmentation, and model training. This approach ensured that I effectively managed and prepared the data for analysis. By implementing convolutional neural networks (CNNs) like EfficientNetB0 and VGG16, I applied advanced machine learning techniques to generate actionable insights, addressing the second learning goal. The project’s insights can improve quality control processes in the food industry, showcasing the practical applications of our analysis.

Additionally, my project highlights the application of visualization and predictive models to generate actionable insights, meeting the third and fourth learning goals. I utilized Python within Google Collab to develop and fine-tune the CNN models, providing a comprehensive evaluation of their performance through precision, recall, and F1-score metrics. These visualizations and performance metrics effectively communicated the findings to a broad audience, including industry professionals, fulfilling the fifth learning goal. Throughout the project, I maintained ethical considerations by ensuring transparency in our methods and addressing model biases, particularly in classifying the Siirt pistachio species. This commitment to ethical practices aligns with the sixth learning goal, emphasizing fairness and accuracy in my predictive models.

**IST 707 – Applied Machine Learning**

**Project Description**

In the project "Predicting COVID-19 Diagnosis," I developed predictive models to determine the likelihood of a positive COVID-19 diagnosis based on reported symptoms and behavioral variables. Using a dataset comprising 5,434 samples with 21 features, I was tasked with identifying patterns that could help in early detection and intervention of COVID-19. The dataset included symptoms like fever, dry cough, and breathing problems, as well as behaviors such as travel history and contact with COVID-19 patients. The preprocessing steps involved converting categorical variables into numerical values, addressing missing data, and balancing the dataset to handle potential bias. Next, I did exploratory data analysis to identify the most informative features, revealing strong correlations between certain symptoms and positive COVID-19 diagnoses.

I utilized several classification algorithms, including Logistic Regression, Random Forest, and Decision Tree, each adjusted to handle the class imbalance in the dataset with specified parameters. Logistic Regression achieved the highest accuracy at approximately 83.63%, followed by Random Forest and Decision Tree, both around 81%. Additionally, I explored clustering algorithms like K-Means and Hierarchical Agglomerative Clustering to identify potential data groupings, though these models showed moderate distinction with silhouette scores of 0.1169 and 0.1102, respectively. The findings highlight the challenges of working with imbalanced datasets and emphasize the importance of using a combination of supervised and unsupervised learning techniques to gain a comprehensive understanding of the data.

**Learning Objectives & Outcomes**

This project demonstrates my ability to collect, store, and access data by leveraging applicable technologies, fulfilling the first program learning goal. By using a thorough COVID-19 dataset and utilizing Python libraries for data preprocessing and analysis, I was able to effectively manage and prepare the data for the predictive models. This approach aligns with the second learning goal as I created actionable insights across a societal context. These predictive models can help assist healthcare providers in early detection and intervention, improving patient outcomes and resource allocation during the pandemic. These insights also inform public health initiatives, guiding targeted awareness campaigns to alleviate the spread of COVID-19.

Furthermore, my project highlights the application of visualization and predictive models to generate actionable insights, addressing the third and fourth learning goals. Using Python, I developed and evaluated various classification algorithms, such as Logistic Regression, Random Forest, and Decision Tree, and visualized their performance through confusion matrices and classification reports. These visualizations effectively communicated my findings to a broad audience, including healthcare professionals and policymakers, fulfilling the fifth learning goal. Throughout the project, I maintained ethical considerations by ensuring transparency in our methods and addressing the class imbalance in our dataset to reduce bias. This commitment to ethical practices aligns with the sixth learning goal, emphasizing fairness and accuracy in our predictive models.

**Conclusion**

In conclusion, this portfolio summarizes my academic journey and progression through the Applied Data Science program at Syracuse University. The projects showcased in my portfolio were carefully selected to accurately reflect a diverse range of applications in data science, from analyzing movie review trends and detecting spam emails to predicting housing market values, classifying pistachio species, and diagnosing COVID-19. Each project not only highlights my technical ability in utilizing several technologies for data collection, storage, and analysis but also proves my ability to generate actionable insights that can inform decision-making across various contexts, including societal, business, and healthcare sectors. My knowledge in using programming languages such as R and Python, paired with advanced machine learning and deep learning techniques, highlights my dedication and passion in pushing the boundaries of data science.

As a result, this portfolio highlights my dedication to ethical data practices, ensuring fairness, transparency, and bias reduction in all the projects I have completed. Effective communication of insights to both technical and non-technical audiences has been a foundation of my work, reflecting my ability to translate complex data findings into understandable and actionable recommendations that non data individuals can understand. As I transition from Syracuse University to the professional world, I am confident that the skills and experiences documented in this portfolio have equipped me with a strong foundation to handle real-world data challenges, drive innovation in data science, and contribute meaningfully work to the field of data science.

**Appendix**

1. **GitHub Repository**
   1. IST 782 **–** [**Applied Data Science Portfolio**](https://github.com/nolanarendt/IST782-Applied-Data-Science-Portfolio/tree/main)

**Individual Project Folders**

1. **Analyzing Viewer Voting Trends in Classic and Modern Cinema**
   1. IST 652 - [Scripting for Data Analysis](https://github.com/nolanarendt/IST782-Applied-Data-Science-Portfolio/tree/main/IST%20652%20-%20Scripting%20for%20Data%20Analysis)
2. **Detection of SPAM in Emails**
   1. IST 664 - [Natural Language Processing](https://github.com/nolanarendt/IST782-Applied-Data-Science-Portfolio/tree/main/IST%20664%20-%20Natural%20Language%20Processing)
3. **Ames, IA Housing Market Analysis**
   1. IST 687 - [Introduction to Data Science](https://github.com/nolanarendt/IST782-Applied-Data-Science-Portfolio/tree/main/IST%20687%20-%20Introduction%20to%20Data%20Science)
4. **Pistachio Type Classification**
   1. IST 691 - [Deep Learning in Practice](https://github.com/nolanarendt/IST782-Applied-Data-Science-Portfolio/tree/main/IST%20691%20-%20Deep%20Learning%20in%20Practice)
5. **Predicting COVID-19 Diagnosis**
   1. IST 707 - [Applied Machine Learning](https://github.com/nolanarendt/IST782-Applied-Data-Science-Portfolio/tree/main/IST%20707%20-%20Applied%20Machine%20Learning)