**Nolan Nicholas Arendt**

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

Final Project Portfolio

m.S. in Applied data science



**Contents**

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

**Key Skills and Strengths Developed**…………………………………………………………….2

**Scripting for Data Analysis**……………………………………………………………………...4

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

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

**Natural Language Processing**……..…………………………………………………………….6

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

*Learning Objectives & Outcomes*…………………………………………………………7

**Introduction to Data Science**………..…..……………………………………………………….8

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

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

**Deep Learning in Practice**……………………………………………………………………...10

*Project Description*……………………………………………...……………………….10

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

**Applied Machine Learning**…………………………………………………………………….12

*Project Description*………………………………………………………………………12

*Learning Objectives & Outcomes*………………………………………………………..13

**Conclusion**.……………………………………………….…………………..…………………14

**Appendix**….……………………………………………….…………………………………….16

**Introduction**

In this final portfolio, I will present a comprehensive 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 highlight my ability to collect, store, and access data by leveraging applicable technologies, demonstrating my ability to use data to create actionable insights across various real-world contexts, including societal, business, and political scenarios.

Through this collection of work, I will also illustrate my hands-on experience in applying visualization techniques and predictive models to generate actionable insights. My proficiency in programming languages such as R and Python, which I have already applied in real-world scenarios, has been valuable in supporting these insights. In addition, I emphasize my commitment to communicating these insights effectively to a broad range of audiences. As a result, this portfolio reflects my dedication to the ethical application of data science principles, particularly in addressing fairness, bias, transparency, and privacy issues.

**Key Skills and Strengths Developed**

Throughout the Applied Data Science program, I have gained a set of skills and strengths through various projects and coursework. In the Scripting for Data Analysis course, I learned how to effectively collect, clean, and manage data using Python and the Kaggle platform. This project honed my ability to leverage data for creating actionable insights, as demonstrated in my analysis of movie review data.

In the Natural Language Processing course, I improved upon programming skills by applying several machine learning models to classify spam emails. This project helped grow my understanding of the various and unique data preprocessing techniques and model evaluation, highlighting the importance of fairness and transparency in predictive modeling.

The Introduction to Data Science course allowed me to apply data science techniques to real-world datasets, such as the Ames Housing dataset. Through this project, I developed a new experience in using R for data analysis and visualization that I had not learned before, identifying key variables that influence housing prices and presenting these insights in a user-friendly manner.

In the Deep Learning in Practice course, I explored advanced machine learning techniques by developing Convolutional Neural Networks for image classification, a new process that I had never had experience with, both for deep learning and image data. This project not only increased my technical skill set but also highlighted the importance of continuous model evaluation and ethical considerations in reducing bias.

The Applied Machine Learning course provided practical experience in building predictive models to address real world problems, such as predicting COVID-19 diagnoses. Through this project, I learned to balance datasets, apply various classification algorithms, and communicate findings effectively to inform public health decisions.

Throughout the Applied Data Science program, I have developed a robust set of skills directly applicable to my role as a data scientist. My coursework has provided me with expertise in data collection, cleaning, and management using Python and R, along with a strong foundation in machine learning and natural language processing specifically. I have gained proficiency in implementing advanced techniques like Convolutional Neural Networks and have learned the importance of continuous model evaluation, fairness, and transparency in predictive modeling. Practical experience with real-world datasets has enhanced my ability to balance datasets, apply various algorithms, and communicate insights effectively. These comprehensive skills empower me to leverage data-driven methodologies to inform decision-making, optimize processes, and drive innovation in my current role as a data scientist.

**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 focusing 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 1894 and 2020. I filtered the data to focus on single-genre films for cleaner analysis and conducted exploratory data analysis to identify trends within the data. This project aimed to provide movie producers with insights into 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 are some of the most popular genres in terms of the number of votes and average ratings. Compared to other genres, Sci-Fi received lower average scores, indicating a potential area for improvement that directors could focus on. Sentiment analysis of each movie’s plot description 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, the 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**

In the "Analyzing Viewer Voting Trends in Classic and Modern Cinema" project, I demonstrated my ability to collect, store, and access data by leveraging Python and the Kaggle platform. I was responsible for gathering a comprehensive dataset of global movie titles produced between 1894 and 2020 from Kaggle. I utilized Python's pandas library to manage the data, performing tasks such as data cleaning, feature engineering, and preprocessing, including handling missing values and normalizing data formats. My thorough approach ensured that the dataset was suitable for following exploratory data analysis, allowing me to uncover meaningful trends and insights that informed the project's recommendations.

This project also showcased my ability to communicate insights gained via visualization and analytics to a broad range of audiences, including production executives and stakeholders in the movie industry. I used Python libraries such as matplotlib and seaborn to create various visualizations, including bar graphs, scatter plots, and histograms. These visualizations converted complex data trends into accessible and understandable formats, making it easier for non-technical stakeholders to understand the insights I found. For instance, my visualizations clearly highlighted the popularity of different genres, the impact of positive sentiment in plot descriptions on viewer votes, and gender-based rating trends. By effectively communicating these insights, I provided reassurance through valuable recommendations that could advise strategic decisions in movie production, marketing, and genre selection. My ability to communicate these findings through clear and impactful visualizations ensured that the analysis could be easily understood and acted upon by a diverse audience.

**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 project phase aimed to identify effective methods for accurately classifying emails to enhance email security in an age where security and data are precious, and the number of attempted data breaches continues to grow.

I extended my 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 carefully tested and evaluated, with Logistic Regression and Gradient Boosting models both achieving the highest accuracy rates of approximately 97%. The Gradient Boosting model showed perfect precision in classifying ham emails, making it a viable option for enterprise-level applications where the accuracy of email classification is critical. This comprehensive examination of multiple models not only underscores my dedication to improving email security but also provides actionable insights for implementing effective spam detection systems in various real-world scenarios and businesses, making the project's results highly relevant and valuable.

**Learning Objectives & Outcomes**

In the "Detection of Spam in Emails" project, I demonstrated my expertise using programming languages such as Python to generate actionable insights. I managed and analyzed the Enron email dataset using several of Python's libraries, including scikit-learn, pandas, and NLTK. I undertook several data preprocessing steps with the dataset, such as tokenization, random shuffling, and converting text to lowercase, ensuring consistency and suitability for machine learning tasks. I employed various machine learning models, including Naïve Bayes, Logistic Regression, Random Forest, and Gradient Boosting, to identify patterns and classify emails as spam or authentic. My specific use of these programming libraries allowed me to develop highly accurate models, particularly the Logistic Regression and Gradient Boosting models, which achieved accuracy rates of approximately 97%. These models provided critical insights for enhancing email security, making them instrumental for enterprise-level applications.

This project also underscored my commitment to ethical considerations in the development, use, and evaluation of predictive models. I ensured that the models were evaluated for fairness and transparency, particularly in avoiding bias during classification tasks. By using a balanced dataset with an equal number of spam and authentic emails, I mitigated the risk of bias that could have skewed the results. Additionally, my evaluation techniques, including cross-validation and the use of confusion matrices, provided a transparent evaluation of model performance, highlighting areas where false positives and false negatives occurred. This in-depth evaluation process not only confirmed that the models achieved high accuracy but also operated fairly and transparently, thus maintaining the integrity and reliability of our spam detection system. My ethical approach highlights the significance of responsible data science practices, aligning with the sixth learning goal and emphasizing the ethical integrity of the project.

**IST 687 – Introduction to Data Science**

**Project Description**

In the "Ames IA Housing Market Analysis: 2006-2010 project," I analyzed the Ames, Iowa housing dataset to identify key features that influence the return on investment 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. I ran correlation and linear regression analyses 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 features, while neighborhood and lot area had surprisingly weak correlations with the sale price.

The analysis provided actionable insights for potential homebuyers and investors such as myself, highlighting specific improvements that could maximize a buyer's return on their investment. For 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 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**

In the "Ames IA Housing Market Analysis: 2006-2010" project, I demonstrated not just my ability, but my proficiency in data science techniques. I thoroughly analyzed the Ames Housing dataset, which included 2,930 observations and 80 variables related to housing features and sale prices, reducing our scope to only the most influential features. Using the programming language R for data cleaning, feature engineering, and analysis, I identified features that significantly impact the return on investment for different housing attributes. By conducting correlation and linear regression analyses, I identified key variables such as kitchen quality, basement height, and exterior quality, which were among the highest impact factors on sale prices. These insights are not just findings, they are the result of a strong and consistent data science process. They provide valuable information for homebuyers and investors, guiding them in making strategic decisions to maximize their return on investment. The practical applications of my analysis in the real estate market illustrate the confidence you can have in the insights I generate through data science.

This project not only showcased my expertise, but my ability to transform complex data into understandable insights. Using R's ggplot2 and dplyr libraries, I created various visualizations, including scatter plots and bar graphs, to effectively communicate the relationships between housing features and sale prices. These visualizations are tools that make complex data trends simple and understandable, even for the individuals not as familiar with the data itself. They allow potential investors and real estate professionals to grasp the most impactful factors on housing values easily. Additionally, my linear regression models provided predictive insights into how specific improvements in housing attributes could influence sale prices. By clearly illustrating these relationships, I enabled stakeholders to make informed decisions based on my findings. This ability to transform data into actionable insights through visualization and predictive modeling is an example of the clarity and reliability my analysis showcased.

**IST 691 – Deep Learning in Practice**

**Project Description**

In the project "Pistachio Type Classification," I explored using 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 range and prevent overfitting. The dataset was split into training, validation, and test sets, with a slight imbalance within acceptable limits for practical 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 remarkable test accuracy of 98.46%. However, it showed a 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**

In the "Pistachio Type Classification" project, I demonstrated my ability to collect, store, and access image data by effectively leveraging relevant tools. I meticulously managed a comprehensive dataset containing over 2,000 images of pistachios, sourced from reliable databases in Kaggle. Using Python libraries such as TensorFlow and Keras, I handled the data preprocessing, including data augmentation techniques like zooming, shifting, and shearing, to enhance the dataset's diversity and prevent overfitting. I carefully split the data into training, validation, and test sets, ensuring robust and diverse data for model training and evaluation. This thorough data management process enabled accurate classification of pistachio species, demonstrating my proficiency in handling and leveraging large datasets for machine learning tasks.

This project also highlighted my commitment to ethical considerations in developing, using, and evaluating predictive models. Initially, one of the models I had built showed significant bias, predominantly predicting only one of the pistachio species. To address this issue, I implemented data augmentation and model fine-tuning techniques to balance the predictions and reduce bias. I improved the model's fairness and accuracy by adjusting the model architecture and using techniques like dropout regularization and class weighting. This process stressed the importance of ethical practices in machine learning, ensuring that my models provided reliable and unbiased classifications. This commitment to ethics aligns with the sixth learning goal, emphasizing fairness and transparency in my predictive modeling efforts.

**IST 707 – Applied Machine Learning**

**Project Description**

In the project "Predicting COVID-19 Diagnosis," I developed predictive models to determine the likelihood of a 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 aid with early detection and intervention of the COVID-19 virus. The dataset included symptoms like fever, dry cough, breathing problems, and 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 solid correlations between several specific 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 supervised and unsupervised learning techniques to understand the data in full.

**Learning Objectives & Outcomes**

In the "Predicting COVID-19 Diagnosis" project, I not only demonstrated my ability to apply visualization and predictive models to generate actionable insights but also showcased their practical application. Utilizing a dataset comprising 5,434 samples with 21 features, I conducted extensive exploratory data analysis to identify the most informative features correlated with positive and negative COVID-19 diagnoses. I developed several classification models, including Logistic Regression, Random Forest, and Decision Tree, to predict the likelihood of a COVID-19 diagnosis based on reported symptoms and behaviors. I evaluated these models using confusion matrices, which provided clear visual representations of their performance, while also highlighting areas for improvement and testing for biases. Logistic Regression achieved the highest accuracy of approximately 83.63%, demonstrating my capability to build effective predictive models. These visualizations and models, with their key insights into symptom patterns, have the potential to guide public health interventions and resource allocation, making the audience understand the relevance and impact of the project.

This project also demonstrated my commitment to ethical considerations in the development, use, and evaluation of predictive models. I ensured transparency in my methodology by clearly documenting each step of data preprocessing, model training, and evaluation within our report. Addressing the class imbalance in the dataset was an essential ethical concern, as imbalanced data could lead to biased predictions. I employed techniques such as under-sampling the majority class and setting class weights to balance the data, while also adjusting parameters for each model if available. This approach not only addressed the potential bias but also ensured that the models fairly represented both positive and negative COVID-19 cases. Additionally, I considered the implications of false positives and false negatives in the context of healthcare, aiming to minimize potential harm caused by misclassifications. By maintaining these ethical standards, I provided reassurance about the reliability and fairness of my predictive models, aligning with the sixth learning goal and ensuring the stakeholders feel confident in the project's outcomes.

**Conclusion**

In conclusion, this portfolio summarizes my academic journey and progression through the Applied Data Science program at Syracuse University. The projects showcased here have been carefully chosen to demonstrate a wide spectrum of my applications in data science. The projects chosen range from analyzing movie review trends and detecting spam emails to predicting housing market values, classifying pistachio species, and diagnosing COVID-19. Each project features my technical ability in utilizing various technologies for data collection, storage, and analysis. Additionally, they highlight my ability to generate actionable insights that can inform decision-making across diverse sectors, including societal, business, and healthcare. My proficiency in programming languages such as R and Python, together with advanced machine learning and deep learning techniques, feature my dedication and passion for pushing the boundaries of data science.

As a result, this portfolio not only highlights my dedication to ethical data practices but also reassures the audience about my integrity and responsibility. I have ensured 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. Often in real-world scenarios, bridging the gap between people who understand the data and those who do not can be the most difficult challenge of all, and my experience in these projects has given me newfound confidence and understanding of how to address that challenge. In addition, I have demonstrated my ability to work in teams, collaborating with professionals from diverse fields to tackle complex data problems. 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 to the field of data science.

**Appendix**

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)