Bryce Anthony

Applied Data Science Portfolio

This portfolio describes 6 assignments completed as a part of the curriculum for the Syracuse university master’s in applied data science program. The culmination of projects in this portfolio demonstrates that I have developed the advanced cognitive strategies required of a data scientist by mastering the 6 learning objectives outlined below:

* 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).

The assignments presented in this portfolio were completed for various courses as part of the Applied Data Science program at Syracuse University. Already having strong foundations in programming and Artificial intelligence concepts, from my bachelor’s degree in computer science, the focus for the projects in this portfolio is on leveraging modern technologies to solve practical problems. Each project discussed this portfolio begins with a brief description of its objectives, followed by an explanation of how the project aligns with and achieves specific learning outcomes.

**Deploying Artificial Intelligence Solutions** **(IST 615 -Cloud Management)**

The primary objective of this project was to implement and deploy custom-developed AI models into a cloud-based environment, emphasizing scalability, accessibility, and integration with modern data workflows. Three models were deployed; an exam score predictor implementing a regression algorithm to predict the exam score of a student using a mix of demographic information and past test scores, a deep learning model delivering AI powered predictions for regular season NBA matches, and an AI powered web application that brings the power of computer vision and large language models to calorie tracking and nutritional guidance. Through this project each of the artificial intelligence solutions that we had developed were made available to anyone via the cloud.

The development of the models was primarily carried out using Python and its ecosystem of libraries and frameworks for Artificial Intelligence and machine learning. The TensorFlow library was used in the development of the NBA match prediction model while the Sci-kit learn library was used in the development of the exam score predictor. To deliver the model outputs in a user friendly format the Fast API, Streamlit, and Flask packages were used. Fast API was used to make an API that allowed users to view the get spread and outcome predictions for games occurring on the user specified date and season. Flask was used to make a user friendly front end for the exam predictor model allowing users to input the demographic information along with their past test scores to get an exam score prediction from the model. Streamlit was used to make a user friendly front end for the For the Nutriscan application and the google Gemni pro API was used to enhance the capabilities of the model and allowing it to give nutritional guidance using natural language. Through the the use of programming languagues and graphical user interfaces, we were able to provide real-time prediction models with user interactivity allowing for actionable insight.

Amazon Web Services (AWS) played a central role in the collection, storage, and access of data for this project. To ensure scalability and real-world applicability, we deployed our solutions using AWS Elastic Compute Cloud (EC2) instances. These instances were customized for each application, with specific configurations tailored to the required instance size and storage needs. AWS Elastic Beanstalk was used to manage data security and access by utilizing IAM roles and security groups to maintain controlled access. Additionally, we used Amazon S3 to store source code and deployment artifacts, facilitating version control and deployment management. Each of these decisions was made after thorough research and evaluation of their potential impacts on the performance and scalability of the models.

The Exam Score Predictor provided actionable insights for educators by forecasting student outcomes based on various factors. The NBA Match Predictions API offered predictive insights into game outcomes for sports enthusiasts and analysts. NutriScan AI enabled personalized dietary guidance by analyzing food images for calorie tracking. Each of these applications was designed with user accessibility in mind, and their deployments were tailored around their diverse contexts, in education, sports, and health.

Ethical considerations were also integral to the project, with a focus on data security, fairness, and transparency. Security measures such as IAM roles and encrypted key-pair authentication were implemented for all EC2 instances. The NBA match predictor used publicly available data from a verified source to ensure reliability. Additionally, cost management for EC2 deployments was considered to ensure ethical resource usage.

This project successfully integrated cloud deployment and machine learning to provide real-world solutions. The use of AWS for infrastructure, the deployment of applications addressing varied domains (education, sports, and health), and the ethical considerations involved highlight the project’s contribution to the learning objectives. The project not only demonstrates technical expertise in AI deployment but also emphasizes the practical application of AI to solve real-world problems in a scalable, ethical manner.

**I**

**IST 659 (Data Administration Concepts and Database Management): Farm 2 Fork**

The *Farm 2 Fork* project was proposed as a solution to increase transparency in the food supply chain. The application was designed with a dual focus: allowing consumers to track the original source of the products they purchase, while enabling businesses throughout the supply chain to execute transactions more efficiently.

A relational database was designed and implemented using Azure's cloud infrastructure. The database included key tables for users (farmers, distributors, and customers), products, and transactions, with distinct user roles managed through structured relationships. To ensure data integrity and consistency across user roles and locations, identity columns and foreign key constraints were used. Real-time inventory tracking was automated using timestamping to capture updates across all user classes. A custom Python script was also used to generate realistic external data, populating the database with SQL commands for testing and demonstration purposes. This tested database was then used with a Microsoft PowerApp application that we developed.

The project provided tailored dashboards for each user type: farmers, distributors, and customers. Real-time inventory tracking and dashboards were provided, enabling users to monitor and act on trends within the supply chain. For farmers and distributors, key sales metrics were visually presented, facilitating operational planning and better decision-making. For customers the inventory tracking offered transparency by allowing them to track product origins throughout the supply chain. The integration of key metric visualizations addressed the diverse needs of the project stakeholders by offering actionable insights into inventory management (for farmers), transactional efficiency (for distributors), and supply chain transparency (for customers). Integration with Microsoft PowerApps allowed users to interact with the data through an intuitive interface, making inventory overviews and transactional insights easily accessible.

The *Farm 2 Fork* project effectively demonstrates the integration of database management concepts, cloud infrastructure, and data visualization techniques to enhance transparency in the food supply chain. By providing tailored solutions for farmers, distributors, and consumers, the project not only optimized inventory management and transactional efficiency but also empowered consumers to make informed decisions by tracing product origins. The use of Azure’s relational database infrastructure, coupled with real-time tracking and intuitive dashboards, ensured that the needs of all stakeholders were met.

**IST 687 (Introduction to Data Science): Energy Consumption Analysis**

For this project, eSC, a power company serving South Carolina and parts of North Carolina, sought to address the potential impact of rising temperatures due to global warming on its electrical grid in the coming summer. The goal of the project was to identify the key factors influencing energy consumption among eSC's customers, enabling the company to devise strategies that promote energy conservation. In addition, a predictive model was developed to forecast energy usage during unusually hot summers, allowing eSC to proactively plan and manage its resources more efficiently.

The project utilized data on energy consumption, housing characteristics, and weather to predict power usage across counties served by eSC. R was used throughout the project for data processing, predictive modeling, and developing the Shiny App; R was used to clean and merge datasets, reducing over 4 million records to a manageable sample. Using this data an XGBoost model was developed and optimized in R to forecast energy consumption using metrics like RMSE and MAE for accuracy. After the model was trained and evaluated, a 5-degree temperature increase was fed into the model to predict future electricity usage in addition to the factors most significantly influencing the expected energy demand.

Key insights were communicated through heatmaps highlighting hourly peak energy demand across regions and line plots displaying cyclical trends in energy usage, particularly the spikes in evening hours. These insights identified the most impactful things to focus on when trying to reduce energy costs leading to data driven and actionable recommendations for future energy conservation. These recommendations included: installing solar panels to offset energy usage, incentivizing insulation upgrades to reduce cooling costs, and promoting smart thermostats and energy-efficient lighting to optimize consumption.

A Shiny App was also created to provide an interactive dashboard showcasing our exploratory data analysis, model predictions, and recommendations for reducing energy usage. This combination of predictive modeling and interactive visualizations allowed stakeholders to easily access our insights and use them for better decision-making. The Shiny dashboard simplified complex analyses into three intuitive tabs:

* Data Overview: Provided daily and hourly energy consumption trends across counties.
* Model Predictions: Displayed future peak energy demands.
* Future Energy Use: Presented actionable recommendations for stakeholders.  
  Graphs and charts were used to visualize insights, making them accessible to both technical (e.g., energy providers) and non-technical (e.g., consumers) audiences. This ensured that the findings could be understood by a diverse range of stakeholders.

This project effectively demonstrates the practical application of data science techniques to address a critical societal challenge- managing energy consumption in the face of rising temperatures. By collecting and analyzing data on energy consumption, housing characteristics, and weather patterns, the project created actionable insights informing strategies for energy conservation. Predictive modeling using XGBoost, alongside visualizations such as heatmaps and line plots, enabled accurate forecasts of future energy demand and highlighted key factors influencing consumption. R was utilized throughout the project, from data processing to model development and the creation of an interactive Shiny App, showcasing proficiency in programming and the ability to communicate complex insights. The clear, accessible visualizations presented on the Shiny dashboard ensured that the findings could be understood by both technical and non-technical audiences. In meeting the learning objectives, this project successfully combined predictive modeling, visualization, and effective communication to provide meaningful insights that could contribute to resource management and energy conservation efforts.

**IST 707 (Applied Machine Learning): NBA Spread Prediction with Neural Networks**

The primary goal of this project was to provide AI-backed insights into the outcomes of NBA sporting events, enabling users—whether enthusiasts, analysts, pundits, or casual observers—to better understand individual games and their implications for the season. By giving users access to the predictive power of machine learning, the goal was to empower them to make more confident predictions and analyses.

Data for this project was collected using web scraping to gather detailed NBA game statistics, including player and team-level metrics from sources like Basketball Reference and the NBA Stats page. The data, spanning from 2015 to 2024, was organized into structured csv files using the pandas library in python. Features were engineered, such as rolling averages of game statistics, which allowed the model to incorporate trends and past performance. Extensive player and team statistics, including advanced metrics like PER (Player Efficiency Rating) and Win Shares, were utilized to create robust training datasets.

This project focused on predicting game outcomes (win/loss) and point spreads, providing valuable insights for stakeholders, such as sports bettors and analysts. The project aimed to address key challenges within the sports betting industry, where bookmakers rely on proprietary algorithms to set odds. In addition to predictions, the project also provided insights into player-level contributions (e.g., Defensive Win Shares, Box Plus-Minus) and team-level trends, helping fans, analysts, and casual enthusiasts better understand NBA dynamics.

Exploratory Data Analysis (EDA) was conducted to identify critical features by analyzing both team-level and player-level statistics. Visualizations, such as distribution trends, histograms, and confusion matrices, were employed to interpret model performance. A deep learning model was developed and fine-tuned using advanced techniques, including dropout layers and batch normalization to combat overfitting, as well as testing multiple activation functions (e.g., GELU, Tanh) and optimizers (e.g., Nadam). Feature selection was enhanced through L1 regularization and Principal Component Analysis (PCA). Python played a central role in the project, facilitating the construction and training of neural networks using TensorFlow and Keras. Various architectures and loss functions (e.g., Mean Squared Error, Huber) were explored, while custom preprocessing pipelines for feature engineering—including rolling averages, one-hot encoding, and genetic algorithms—were implemented. Hyperparameter tuning and model evaluation were automated using iterative loops and grid searches, showcasing proficiency in Python for developing scalable and efficient machine learning solutions. The resulting model achieved a win/loss prediction accuracy of approximately 57% and a spread prediction accuracy of 30% on unseen data.

Insights about the performance of the model for each of its tasks were communicated through charts and histograms. Charts created included validation loss curves, bar plots comparing hyperparameter configurations, and confusion matrices were displayed. Histograms were created Illustrated the differences between predicted and actual outcomes for both spread and win/loss predictions. Actionable takeaways were emphasized, such as the importance of player-level statistics and the optimal number of past seasons to take into consideration when trying to predict outcomes. The project was also transparent about the model’s limitations, such as struggles with spread prediction and signs of overfitting. While the model did not meet the metrics required for reliable profitability, it provided valuable insights into game outcomes and the complexity of sports betting when it comes to the NBA. The project also highlighted the importance of high-quality data and careful feature engineering to improve future performance.

This project applied machine learning to predict NBA game outcomes and point spreads, delivering insights for sports analysts and bettors. By leveraging web scraping to gather and preprocess player and team statistics, key learning objectives in data collection and analysis were achieved. Predictive modeling, supported by Python-driven preprocessing and hyperparameter tuning, generated actionable insights while visualizations, like validation loss curves and confusion matrices, effectively communicated results to diverse audiences. Ethical considerations, including transparency about model limitations and data quality, ensured responsible analytics practices. Despite room for improvement in spread prediction, the project successfully met learning goals in predictive modeling, data visualization, and communication, providing a strong foundation for further exploration in sports analytics.

**IST 736: Healthiness Evaluation of Recipes (Text Mining Final Project)**

The purpose of this project was to develop a tool that accurately evaluates the healthiness of recipes, empowering individuals and businesses to make informed dietary choices. By leveraging ingredient names and key nutritional metrics, the tool aims to promote healthier eating habits and support the health and wellness goals of its users. The project addressed a societal need by classifying recipes as healthy or unhealthy, enabling informed dietary choices. Machine learning models provided actionable insights for individual users, meal kit services, and the health and wellness industry. The insights from this project highlighted the importance of ingredient quantities and caloric content over ingredient types, offering a deeper understanding of recipe healthiness.

The data for this project was collected using the Spoonacular API, a reliable source for gathering recipe-related nutritional information. This approach ensured the collection of relevant, structured data for analysis. The recipes were stored in CSV files for efficient access and processing during text mining and machine learning tasks. Python was used extensively for: Data preprocessing (e.g., stemming, stop-word removal, vectorization), Machine learning workflows using libraries like scikit-learn, Visualization using matplotlib and other libraries. The use of Python showcases proficiency in programming to manage text mining tasks, train machine learning models, and visualize data.

Visualization tools (e.g., word clouds, PCA plots, SHAP dependence plots) were used to highlight important patterns, such as the overlap of ingredients in healthy and unhealthy recipes. Predictive models applied included: Logistic Regression, Support Vector Machines (SVM), Random Forest, Gradient Boosting, Decision Trees, and Naive Bayes. The Gradient Boosting and Linear SVM stood out for their high classification accuracy. The SHAP analysis identified influential features like fat, carbs, and protein in predicting calorie content, providing interpretability to the models. The combination of machine learning models and meaningful visualizations demonstrates the project's ability to generate and present actionable insights.

The project balanced technical depth with accessibility through intuitive visualizations (e.g., word clouds, PCA plots, confusion matrices) for non-technical audiences and detailed performance analysis (e.g., SHAP plots, accuracy, precision, recall, and F1-scores) for technical stakeholders. Potential applications, such as integrating health ratings into meal kit delivery services, highlighted the project’s cross-industry relevance. Biases in defining "healthy food" were addressed by curating a diverse, culturally inclusive dataset and ensuring a balanced representation of healthy and unhealthy recipes. Limitations, including potential overfitting in Gradient Boosting and Decision Trees, were acknowledged, enhancing transparency in model performance and results.

Python played a central role in the project, supporting data collection and processing, model development, and data visualization. The project’s clear communication of insights ensured that both technical and non-technical audiences could understand and act on the findings, while ethical considerations were also emphasized, focusing on fairness, transparency, and minimizing bias in the dataset. Overall, the project not only fulfilled the learning objectives but also demonstrated the power of data science to improve health outcomes and support wellness goals across various contexts.

**SCM 651 (Business Analytics): Book Emporium Sales Analysis**

For this assignment, I analyzed fictional sales data from the Book Emporium to help determine the optimal pricing strategy for the upcoming Harry Potter sequel. The dataset provided included weekly sales figures and customer purchase percentages for the hypothetical Harry Potter Book 7, with prices varying over time. My goal was to identify a demand curve for the book by examining how changes in pricing affected customer purchases and use the demand curve to recommend an optimal sale price for a sequel book that would maximize profits.

Using historical sales data for harry potter book sales, Multiple regression models—linear, exponential, power, and logarithmic—were employed to predict the relationship between price and purchase percentage. Scatter plots with trendlines and equations visualized the models, highlighting key relationships and the best fit using R² values. Optimization models illustrated profit-maximizing price points under various constraints. Predictive models provided accurate forecasts for key variables (e.g., predicted sales, revenue, and profit), while visualizations made these insights accessible to stakeholders.

Regression results were presented using visualizations, including scatter plots, optimization tables, and profit-maximization outputs. Each model's fit and its practical implications (e.g., revenue vs. price trade-offs) were articulated clearly to support stakeholder understanding. Optimization scenarios were summarized concisely to detailing how constraints like publisher discounts influence pricing strategy. Ethical considerations were embedded in the analysis by ensuring transparency in the assumptions (e.g., demand for the sequel will follow trends for Harry Potter Book 7). Biases were minimized by evaluating multiple regression models and selecting the best fit using statistical metrics like R². Constrained optimization scenarios balanced business goals with fairness, ensuring customers are not exploited with excessively high prices.

By clearly illustrating pricing trends and the impact of constraints, the project enabled informed decision-making and identified the price that would maximize profits for a hypothetical Harry Potter sequel. Utilizing key business metrics such as predicted sales, revenue, profit, and publisher discounts for higher sales volumes. This project enabled stakeholders to confidently set pricing strategies for the sequel, showcasing the practical application of data science to real-world business challenges.

The projects in this portfolio showcase the successful application of core data science learning objectives to real-world problems across diverse domains, including business, societal, and environmental contexts. Through assignments such as AI deployment in *Deploying Artificial Intelligence Solutions*, database management in *Farm 2 Fork*, and predictive analytics in *Energy Consumption Analysis*, I effectively applied data collection, processing, and modeling techniques using Python, R, and cloud technologies. Each project emphasized actionable insights generated through predictive models, visualizations, and ethical considerations such as fairness, transparency, and responsible analytics. By combining technical proficiency with clear communication of findings, each project has provided me with valuable experience and knowledge that aligns with the learning objectives of the program, showcasing my proficiency in data science methodologies and communication across diverse domains to leverage technology to solve practical problems.