**ADS Portfolio**

**By Mervin McDougall**

**Purpose of the Report**

This report reflects on whether I have met the learning goals of the MS Applied Data Science program and demonstrated competency in applying these goals throughout my academic career at Syracuse University. The program's learning goals are as follows:

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

To evaluate my progress, I will review three projects from three courses: IST692 Responsible AI, IST707 Applied Machine Learning, and IST718 Big Data Analytics. While some learning goals overlap across these projects, I will highlight the most significant contributions of each.

**Review of Projects**

**IST692: Mortgage Approval and Gender Bias Analysis**

**Objective:** Develop a predictive model for mortgage approvals while investigating gender disparities in the approval process.  
**Key Learning Goals:**

* Applied ethical considerations in data modeling.
* Created actionable insights across societal and financial contexts.
* Communicated findings effectively using visualizations.

**IST707: NBA Match Outcome Predictor**

**Objective:** Build a neural network model to predict NBA match winners and point spread.  
**Key Learning Goals:**

* Collected and processed data from multiple sources using applicable technologies.
* Used R and Python to develop predictive models.
* Communicated model performance insights effectively.

**IST718: Toxic Comment Discovery**

**Objective:** Develop a dataset to train models for detecting toxic comments in social media discussions.  
**Key Learning Goals:**

* Applied data collection and feature engineering techniques.
* Used visualization and predictive models to extract insights.
* Communicated findings effectively across audiences.

**Methodology**

To demonstrate how each learning goal was met, I will outline each goal and provide evidence from the three projects. This structured approach ensures clarity in mapping my academic experiences to program expectations.

**Learning Goals and Application**

**1. Collect, Store, and Access Data**

**Example: IST718 - Toxic Comment Discovery**  
**Technology Used:** Google Colab, Apache Spark (PySpark), NLTK, Spacy, Scikit-Learn

* Due to the large dataset size (3.23 GB), processing was offloaded to Google Colab.
* Apache Spark was employed for scalable data management.
* Memory management strategies were implemented to optimize resource usage.

**Key Takeaways:**

* Managing large datasets requires strategic resource allocation.
* Effective preprocessing techniques, such as feature engineering, enhance model performance.

**2. Create Actionable Insights**

**Example: IST692 - Mortgage Approval and Gender Bias Analysis**  
**Technology Used:** SHAP, LIME, Fairlearn, Evidently

* Identified gender-based disparities in mortgage approvals.
* Discovered that loan-to-value ratio and log-income were key predictors of approval.
* Suggested policy adjustments to improve fairness.

**Key Takeaways:**

* Ethical considerations must be embedded in data analysis.
* Data-driven insights can influence policy changes.

**3. Apply Visualization and Predictive Models**

**Example: IST707 - NBA Match Outcome Predictor**  
**Technology Used:** TensorFlow, Keras, Sci-kit Learn

* Conducted exploratory data analysis to identify feature correlations.
* Optimized model configurations using hypothesis testing on activation functions and learning rates.
* Used confusion matrices and histograms to assess model accuracy.

**Key Takeaways:**

* Visualizations help identify key model strengths and weaknesses.
* Systematic hypothesis testing improves model reliability.

**4. Use Programming Languages for Insights**

**Example: IST707 - Neural Network Training for NBA Predictions**  
**Technology Used:** Python, TensorFlow, Keras

* Implemented PCA to reduce feature dimensionality.
* Compared activation functions and regularization methods.
* Evaluated performance using validation loss and accuracy graphs.

**Key Takeaways:**

* Python provides a powerful ecosystem for predictive modeling.
* Feature selection significantly impacts model performance.

**5. Communicate Insights Effectively**

**Example: IST692 - Presentation on Gender Bias in Mortgage Approvals**  
**Tools Used:** PowerPoint, Data Visualization Libraries

* Developed clear, structured reports to present findings.
* Used bar charts, pie charts, and SHAP visualizations to illustrate key insights.
* Provided a video walkthrough explaining the significance of results.

**Key Takeaways:**

* Effective communication bridges the gap between technical findings and actionable insights.
* Audience-specific presentations improve comprehension.

**6. Apply Ethics in Data Science**

**Example: IST692 - Ethical Considerations in Loan Approvals**  
**Technology Used:** Fairlearn, SHAP, Data Anonymization

* Removed protected characteristics (gender, race) from model inputs.
* Identified dataset imbalances that could introduce bias.
* Recommended improvements for fairer lending practices.

**Key Takeaways:**

* Ethical analysis is crucial in model development.
* Fairness-aware algorithms can help mitigate bias.

**Conclusion**

The MS Applied Data Science program has significantly shaped my ability to work with complex datasets, extract meaningful insights, and develop ethical, data-driven solutions. Throughout these projects, I demonstrated proficiency in:

* Collecting and processing large-scale data efficiently.
* Creating predictive models to derive actionable insights.
* Applying visualization techniques to communicate findings clearly.
* Leveraging programming skills to enhance model performance.
* Implementing ethical best practices in data analysis.

Looking ahead, I aim to deepen my knowledge of statistical evaluation methods, explore alternative fairness-measuring libraries, and engage in a data science community to stay updated on emerging trends and technologies.

**References**

[1] Hanu, L. "Detoxify," 2022. Available: <https://huggingface.co/unitary/toxic-bert>.  
[2] Sports Reference. "Basketball Reference." Available: <https://www.basketball-reference.com/>.  
[3] NBA. "NBA Advanced Stats." Available: <https://www.nba.com/stats>.  
[4] Fairlearn. "Fairlearn Documentation." Available: <https://fairlearn.org/>.