**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 emphasizes the ability to collect, store, and access data by identifying and leveraging applicable technologies. It also focuses on creating actionable insights across various contexts, including societal, business, and political domains, using the full data science life cycle. Another key objective is to apply visualization and predictive models to generate insights, use programming languages such as R and Python for data analysis, and communicate findings effectively to both technical and non-technical audiences. Finally, the program emphasizes the importance of applying ethical principles in the development, use, and evaluation of data and predictive models, particularly in areas such as fairness, bias, transparency, and 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**

The first project, IST692: Mortgage Approval and Gender Bias Analysis, focused on developing a predictive model for mortgage approvals while investigating gender disparities in the approval process. This project involved applying ethical considerations in data modeling, generating actionable insights in the financial sector, and effectively communicating findings through visualizations.

The second project, IST707: NBA Match Outcome Predictor, involved building a neural network model to predict NBA match winners and point spread. This required collecting and processing data from multiple sources using appropriate technologies, employing R and Python to develop predictive models, and presenting insights on model performance.

The third project, IST718: Toxic Comment Discovery, aimed to develop a dataset to train models for detecting toxic comments in social media discussions. This involved applying data collection and feature engineering techniques, using visualization and predictive models to extract insights, and effectively communicating findings to a wide audience.

**Methodology**

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

**Learning Goals and Application**

One of the fundamental goals of the program is the ability to collect, store, and access data using applicable technologies. In the IST718 Toxic Comment Discovery project, data processing was particularly challenging due to the dataset’s large size of 3.23 GB. To manage this, I offloaded processing to Google Colab and used Apache Spark for scalable data management. Memory management strategies were implemented to optimize resource usage, ensuring that large-scale data operations were handled efficiently. This experience reinforced the importance of managing large datasets strategically and demonstrated how effective preprocessing techniques enhance model performance.

Creating actionable insights is another essential aspect of data science. In the IST692 Mortgage Approval and Gender Bias Analysis project, I applied various analytical techniques using SHAP, LIME, Fairlearn, and Evidently to identify gender-based disparities in mortgage approvals. The analysis revealed that loan-to-value ratio and log-income were key predictors of approval. This led to recommendations for policy adjustments to improve fairness, illustrating how ethical considerations must be embedded in data analysis and how data-driven insights can influence decision-making.

Visualization and predictive modeling play a significant role in generating insights. In the IST707 NBA Match Outcome Predictor project, I employed TensorFlow, Keras, and Sci-kit Learn to conduct exploratory data analysis, optimize model configurations, and test various activation functions and learning rates. Using confusion matrices and histograms, I assessed the model’s accuracy, which highlighted the importance of visualizing model performance to identify strengths and weaknesses. The process of systematic hypothesis testing demonstrated the value of continuously refining predictive models.

Programming languages such as R and Python were integral to these projects. In IST707, I used Python extensively for neural network training, implementing PCA to reduce feature dimensionality and experimenting with different model configurations. By comparing activation functions and regularization methods, I optimized the model’s performance. These experiments reinforced the importance of feature selection and demonstrated how Python provides a powerful ecosystem for predictive modeling.

Communicating insights effectively was a key component of all three projects. In IST692, I developed a presentation and an accompanying video to explain the findings of the gender bias analysis. Using PowerPoint and various data visualization tools, I structured the report clearly and used bar charts, pie charts, and SHAP visualizations to illustrate key insights. This experience emphasized how effective communication bridges the gap between technical findings and actionable insights, ensuring that complex data-driven conclusions are accessible to diverse audiences.

Ethical considerations in data science were particularly relevant in IST692. Although the dataset was anonymized, it included protected characteristics such as gender and race. Through exploratory data analysis, I identified imbalances in the data, including a significant underrepresentation of non-binary individuals. Models trained on this dataset exhibited a bias toward males, which was evident in the selection rate and false negative rate disparities. To mitigate this bias, I removed gender as a feature from the final dataset. Additionally, I recognized the importance of designing fairer lending models by incorporating fairness-aware algorithms. This project reinforced the need for ethical vigilance when handling sensitive data and demonstrated how fairness analysis can inform model adjustments.

**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, and implementing ethical best practices in data analysis.

Looking ahead, I plan to deepen my knowledge of statistical evaluation methods, explore alternative fairness-measuring libraries, and engage with the broader data science community to stay updated on emerging trends and technologies. The program has provided a strong foundation, and I am eager to continue refining my skills and applying them in real-world contexts.

**References**

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