# ADS Portfolio Reflection

By Mervin McDougall

Contents

[ADS Portfolio 1](#_Toc193651988)

[Purpose of the Report: 3](#_Toc193651989)

[Review of projects used in the report: 4](#_Toc193651990)

[Method: 5](#_Toc193651991)

[Learning Goals: 5](#_Toc193651992)

[Collect, store, and access data by identifying and leveraging applicable technologies. 5](#_Toc193651993)

[Create actionable insight across a range of contexts (e.g., societal, business, political) using data and the full data science life cycle. 5](#_Toc193651994)

[Apply visualization and predictive models to help generate actionable insight. 5](#_Toc193651995)

[Use programming languages such as R and Python to support the generation of actionable insight. 8](#_Toc193651996)

[Communicate insights gained via visualization and analytics to a broad range of audiences (including project sponsors and technical team leads). 10](#_Toc193651997)

[Apply ethics in the development, use, and evaluation of data and predictive models (e.g., fairness, bias, transparency, privacy) 12](#_Toc193651998)

[Conclusion 13](#_Toc193651999)

[References 14](#_Toc193652000)

## 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 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)

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 used in the report:

Before I delve into explaining the learning goals, it is important to provide some background about the projects that I will be reviewing in respect of these learning goals.

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

Objective: Develop a predictive model for mortgage approval while investigating gender disparities in the approval process.

**Learning goals**:

* Apply ethics in the development, use, and evaluation of data and predictive models.
* Create actionable insights across a range of contexts using data and the full data science life cycle.
* Communicate insights gained via visualizations and analytics to a broad range of audiences.

**IST707: NBA Match Outcome Predictor**

Objective: Build a neural network model to predict NBA match winners and point spread (the difference in the score)

**Learning goals**:

* Collect, Store, and access data by identifying and leveraging applicable technologies.
* Communicate insights gained via visualizations and analytics to a broad range of audiences.
* Use programming languages such as R and Python to support the generation of actionable insights.

**IST718: Toxic Comment Discovery**

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

**Learning goals:**

* Collect, Store, and access data by identifying and leveraging applicable technologies.
* Apply visualization and predictive models to help generate actionable insight.
* Communicate insights gained via visualization and analytics to a broad range of audiences.

## Method:

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.

When mentioning a project for the first time, I will provide a brief overview of the project and the technology used during the project. Thereafter, for any mention of that project, it will be assumed the reader is familiar with the project’s background.

## Learning Goals:

### Collect, store, and access data by identifying and leveraging applicable technologies.

#### IST718 – Toxic Comment Discovery

##### Project Background and Technology Used

Once a hypothesis has been set for a project, the next step always includes a search for data that can be used to answer that hypothesis. For IST718, the project aimed to develop a novel way of identifying toxic communication in social media posts, the result of which would be a dataset on which future models could be trained. Therefore, it would be important to exercise some feature engineering to enrich the dataset with new predictors, such as topics, by using LDA (Latent Dirichlet Allocation) from the discussion posts and a toxicity score that represents the probability of toxicity in the discussion posts.

After reviewing multiple data sources for social media datasets, the site Hugging Face was selected as it provided a large corpus of Reddit posts that could be used to develop the secondary dataset. The chosen dataset had 3.23 GB of data, making it one of the largest datasets I have had to work with and difficult to perform exploratory data analysis on a local machine. Consequently, the decision was made to offload the processing to a cloud service, Google Colab, for data processing. Google Colab provided a platform for collaborative development with other group members who worked on the project, freed up resources on the local machines of group members, had more processing power, and afforded the ability to run autonomously for up to 12 hours under their free tier without having to deal with interruptions that would be experienced on a laptop.

Given the volume of data that would be processed, it was decided to use Apache Spark and, by extension, PySpark for scalable data management. This would allow preparation and preprocessing of the data across multiple machines. Further, the Google Colab environment provided the option of choosing different processors. The T4 processors were chosen to accelerate the transformation of the dataset.

With the use of PySpark, Python became the language of choice for development, and all other libraries used in the project were selected based on this requirement. The data consisted primarily of text data. Therefore, libraries such as NLTK, Spacy, and Scikit-Learn were used to help filter stop words, generate word frequency distributions, and collect statistics about the corpus. The features of these libraries were applied by leveraging user-defined functions within PySpark, allowing the data transformations to occur without having to generate duplicates of the dataset.

##### Unique Approaches

Some unique approaches were developed to address unforeseen difficulties discovered while completing the project. The first was memory management. Even within Google Colab, there was a need for managing resources. The dataset was large, and the same can be said of the resulting dataset. Of the 1.8 million observations in the dataset, the data imported into the Google Colab environment was restricted to 50,000 observations.

Moreover, with various transformations applied to the dataset, it became necessary to delete intermediate data structures that still contained data; Otherwise, the Google Colab environment would frequently run out of memory. Many data structures in PySpark are immutable. Therefore, a change to the data required the creation of a new variable to hold the data. Unused variables were set to None to relinquish any resources tied to them and optimize resource usage. Further, many transformations were made within function calls to ensure that the intermediate variables were destroyed once the function was no longer in scope.

Long processing times became a recurring theme throughout this project. As stated before, two of the reasons Google Colab was chosen were that it freed up resources on local machines and could run autonomously for up to 12 hours. This was important because the record processing easily stretched across multiple 8-hour periods. To avoid restarting the project from the beginning when the processing time expired, data was saved to parquet files at specific stopping points, ensuring a record was readily available should the project need to be restarted at a future time or from a specific stopping point.

Although the dataset included separate features for the topic, body, and comments. The comments feature was complex. It consisted of a JSON string of a commentator key and comment value. The three features – topic, body, and comments – were combined to form a new feature, which was referred to as a thread. This was done for each observation, making it easier to perform LDA topic modeling and compare the topics across different threads.

To generate the toxic score of a thread, a transformer model known as ToxicBert [1] was utilized to rate the level of toxicity identified in the threads. This was particularly difficult to implement in PySpark as both PySpark and the ToxicBert transformer required significant amounts of memory. This often led to long-running times where the process would expire before the data was extracted. It was the only time that the PySpark data was extracted to a primitive, a Python list, for processing.

##### Insights gathered

The insights gathered were not groundbreaking; Rather, they were reinforcing. When topic modeling was performed on the dataset, it was discovered that there was alignment with the topics generated for the discussion threads and the thread title. Therefore, there was a lot of consistency in the discussion threads.

Further, discussion threads that had comparable topics frequently discussed similar themes. For example, threads that generated topics such as “government,” “tax,” and “party” revolved around themes regarding societal and financial issues. See Figure 1 for an excerpt of the thread topics, popularity score, and LDA topic words.

A close up of a letter

AI-generated content may be incorrect.

Figure

After generating toxicity scores with ToxicBert, it was discovered that posts that contained more explicit or aggressive language scored higher. However, the higher toxicity scores did not align with the higher post popularity scores that the Reddit commentators provided. Consequently, although there was higher engagement identified through the Reddit post popularity scores on some topics, this did not equate to a higher toxicity score.

##### Takeaway from the experience

This project tested my ability to think creatively about managing the data used for the project. I had to overcome issues with large volumes of data, manage the memory constraints of the chosen technology – in this case, Google Colab and PySpark- and combine the functionality of different libraries to surpass the limitations of using any one library.

It also taught me the importance of planning out a project and having alternative approaches ready to address some of the preconceived issues that could materialize.

#### IST707 – NBA Match Outcome Predictor

##### Project Background and Technology Used

In IST718, the data originated from a single source and underwent some feature engineering to extract insights. On the other hand, the goals of IST707 required the data be collected from multiple data sources and amalgamated into a singular dataset for training and testing different neural networks.

The goal of IST707 was to develop a neural network for predicting the winner of an NBA match and the resulting spread – the difference between the winner and loser of said match. To accomplish this, team, player, and previous game statistics had to be collected and combined to generate a dataset that could be utilized for training and testing a model.

Data was sourced from basketball-reference.com [2] and nba.com [3] and merged to produce the necessary dataset for training a neural network. Further, the data underwent transformations from categorical and other textual data to numerical data via pipelines for processing by the neural networks.

##### Unique approaches

This project made use of different neural networks, which had its challenges, the primary being there was a need for substantial amounts of data to train the model. This was facilitated by the historical statistical data gathered from the two sites.

Further, neural networks provide numerous configuration options to optimize the model performance. Consequently, the process of deciding on a model configuration consisted of a series of hypothesis tests on activation functions, hidden layer sizes and depths, learning rates, regularization methods, overfitting reduction methods, and other optimization approaches to discover what improved performance.

Google Colab was chosen as the development platform because it provided a selection of different free processors, allowed long processing times, and allowed collaborative work from other group members. Further, libraries such as TensorFlow, Keras, and Sci-kit Learn were installed and utilized to develop preprocessing pipelines and neural networks.

##### Insights gathered

An exploratory data analysis of the data was performed before beginning work on the neural network. Several correlations among the features of the combined datasets were discovered. Due to concerns over high dimensionality reducing the performance of neural networks, it was determined to include this as one of the hypothesis tests in the project. Through hypothesis testing, it was learned that utilizing a reduced dataset, which eliminated correlated fields, produced a model with lower validation loss. Figure 2 below illustrates a graph that compares the training and loss curves of the reduced dataset.

A graph of a train loss curves

AI-generated content may be incorrect.

Figure

Compare this to the same metrics when all the available data was used to train the model, as shown in Figure 3. Of the two, the reduced dataset provided the best performance curves.

A graph of loss curves

AI-generated content may be incorrect.

Figure

However, despite the reduction in the dataset, there was some overfitting observed when comparing the accuracy report of the training and testing results. At the time, it was not certain what caused overfitting but in retrospect, there may have been other features that became proxies for excluded data. Consequently, more feature reduction could have been used.

The resulting model performed well when identifying the winner of NBA matches, but it could not reliably predict the spread. Histograms depicting the predictions vs actual results are displayed below in Figures 4 and Figure 5.

A graph comparing predicted and actual winners and losers of different NBA matches
A graph of different colored lines

AI-generated content may be incorrect.

Figure Figure

##### Takeaways from the experience

The process of fine-tuning a neural network can be time-consuming and requires a lot of computing resources to prepare the data, train and test the models, and generate reporting. However, visualizations provide a quick way to assess the model’s strengths and weaknesses during development. Also, systematic hypothesis testing progressively improves model reliability and performance.

### Create actionable insight across a range of contexts (e.g., societal, business, political) using data and the full data science life cycle.

#### IST692 – Mortgage Approval and Gender Bias Analysis

##### Project Background and Technology Used

In IST692, I reviewed mortgage data collected from Rocket Mortgage and covered by the Home Mortgage Disclosure Act. The goal of this project was to identify whether there were any biases in the data and suggest remediations for this. This was accomplished through different techniques, such as checks for dataset imbalances, checks for the predictors of approval, fairness analysis, and checks for data drift.

Different Python-based libraries were used to achieve this. SHAP and LIME libraries were used for detecting the loan approval predictors; Microsoft Fairlearn [4] was used for comparing gender approval rates; Evidently was used for detecting data drift across the 2023 – 2024 calendar year.

##### Insights Gathered

In the process of analyzing the dataset, I discovered there were dataset imbalances among the genders. Most observations consisted of male applicants. However, the surprising discovery was that there was a small group of applicants for mortgage loans who checked both gender boxes on their application. I labeled these applicants as non-binary applicants since they identified as both male and female. The group only accounted for forty-four applicants, a small amount compared to those who identified as either male or female.

It came as no surprise that the resulting model would be biased toward males, given that the dataset consisted of a higher number of observations in favor of that gender. However, when performing checks for False Negative loan approval classifications (loans that should have been approved but were not), it was learned that more than half of the female applicants, 51%, fell into this category. On the other hand, the selection rate consisted of more than half of males, 53%.

Additionally, when examining the major predictors for mortgage loan approval, it was found that the loan-to-value ratio, preapproval, and log-income are significant predictors of whether an applicant is approved for a loan across the entire dataset.

Finally, adjusting the model to remove any protected characteristics and balancing the approvals and denials did not significantly change the predictive capabilities of the model.

##### Takeaways from the experience

The mortgage loan market is dominated by male applicants, and they have the highest approval ratings among genders. However, the underrepresented groups of females and non-binary individuals can be helped by improving the preapproval rate, the income, and the loan-to-value rates for these groups.

Further, the form used for applying for loans should be expanded to take into consideration those who identify as non-binary. It is possible that many who identify as non-binary may not have known how to report this and chose to identify with their gender assigned at birth. Only a select few were clever enough to check both boxes to communicate this. Having a designated field will go a long way to tracking the needs of this group and ensuring they are counted.

### Apply visualization and predictive models to help generate actionable insight.

#### IST718 – Toxic Comment Discovery

Despite the textual nature of the dataset used for this project, different approaches were used to quantify statistics about the text. Descriptive statistics were generated on the number of comments per title, the title word length, the body word length, answer length, and thread length. An example of this is shown in Figure 6 below. They helped inform about how taxing the dataset would be on the resources secured to process it.

A table with numbers and a number

AI-generated content may be incorrect.

Figure

Word clouds were used to communicate word frequency intuitively and clearly, where the frequently occurring words were displayed larger in the word cloud. The word cloud generated from the Reddit corpus used for this project is depicted below in Figure 7. It was created from the aggregation of all threads in the dataset.

A close up of words

AI-generated content may be incorrect.

Figure

The words know and time dominated the corpus content and suggested some of the views and concerns of the Reddit users who contributed to the corpus.

Frequency distributions were also created, which corroborated the information already depicted in the word clouds. However, they were mostly used to help identify words that could be added to a custom list of stop-words for preprocessing the dataset. For example, the frequency distribution bar chart in Figure 7 below identified the top fifty most common words in the corpus. The top word listed is “like.” This was subsequently added to the list of stop-words as the term on its own did not aid in better understanding the underlying themes of the corpus.

A graph of a graph showing a number of different levels

AI-generated content may be incorrect.

Figure

#### IST692 – Mortgage Approval and Gender Bias Analysis

For IST692, visualizations were applied extensively to gain insights into the nature of the mortgage loan dataset. It helped identify the imbalances in the approval and denial rates and gender imbalances in the applicants. For example, Figure 9 displays the proportion of approvals to denials of all the applications.

Figure

On the other hand, Figure 10 illustrates the gender imbalances captured in the data and how the non-binary group barely registers on the bar chart. Non-binary applicants made up only forty-four of all the applicants. Consequently, it was too small to register on the scale. This imbalance led me to integrate the non-binary group into the female group for continued assessment given the two are both underrepresented groups.

Figure

Eventually, it was decided to remove protected characteristics from the dataset used to train the model. This was done to test the hypothesis of whether application approval would increase if the gender of the applicant were not considered when applying for a loan. Through testing, it was determined that despite the removal of protected characteristics, there was not much change when predicting the number of applicants approved for applications. Consequently, other features may be acting as proxies for gender disparities.

### Use programming languages such as R and Python to support the generation of actionable insight.

#### IST707 – NBA Match Outcome Predictor

IST 707 made extensive use of visualizations generated specifically in Python to gain insights into the training of NBA predictor neural networks. There was an effort to systematically evaluate different configurations of the model to see what small tweaks boosted the performance of the model. An example of one such visualization was generated when experimenting with different activation functions for the model. This is depicted below in Figure 11, in which the different activation functions and their associated validation loss are graphed and compared.

A graph of different colored bars

AI-generated content may be incorrect.

Figure

Because of the graph above, tanh was selected as the activation function of the model.

Further, line charts generated with Python were used to compare the training and validation loss curves and accuracy curves when adjusting the regularization of the model. In Figure 12 below, two line charts are displayed, depicting the loss and accuracy when applying L1 regularization on the model.

A graph of training and validation

AI-generated content may be incorrect.

Figure

Moreover, when testing whether the application of PCA technique improved the outcome of match results, confusion matrices and histograms were generated using Python to get a sense of the effectiveness of this approach to fine-tuning the model. An example of the confusion matrix generated from the PCA analysis depicting the prediction of match winners is shown in Figure 13.

A chart of a number of colored squares

AI-generated content may be incorrect.

Figure

A corresponding histogram overlaying the observed over the predicted winners of matches is shown in Figure 14 below.

A graph of a number of values

AI-generated content may be incorrect.

Figure

In both graphs, the model does a reasonable job of predicting the winner of the match.

### Communicate insights gained via visualization and analytics to a broad range of audiences (including project sponsors and technical team leads).

In all three projects, PowerPoint slides were created to communicate the findings, and these were accompanied by a short presentation. In one case, a video presentation was provided explaining the content and context of the information presented on those slides.

#### IST692 – Mortgage Approval and Gender Bias Analysis

For IST692, a presentation and an accompanying video were created to communicate the insights learned in the process of conducting the project. The findings were arranged in a simple, developmental manner to ensure that they were coherent. Where necessary, bar charts, tables, and pie charts were provided to simplify the information.

For example, the IST692 project’s presentation began by providing a broad overview of the mortgage loan market and how the project was going to be evaluated. Having set the scope of the project, the findings of each of the evaluations were explained, and graphs were provided to simplify the results. For example, Figure 15 below displays the outcome of applying SHAP analysis to the mortgage loan dataset. The features of the dataset are displayed in descending order, illustrating the significance of that feature to being approved for a mortgage loan.

A graph with blue and white text

AI-generated content may be incorrect.

Figure

This finding was reiterated in the report for the project as some of the key ways to improve more equitable access to mortgage loans – the loan-to-value ratio, being preapproved for a loan, and the log income of the applicant.

#### IST707 – NBA Match Outcome Predictor

For IST707 a live presentation was performed among peers who were familiar with machine learning and neural network techniques utilized in the project. Consequently, the choice of visualizations was more technical than simpler bar charts and pie charts. A lot of the visualizations focused on comparing training loss vs validation loss, generating confusion matrices and histograms comparing observed values vs predictions.

Although not as visually attractive as bar charts or pie charts, the goal of these charts was to verify hypothesis tests, which were included as part of the project. These hypotheses asked various questions on how the performance of the model can be improved and therefore necessitated the need for more technical charts.

The most complicated of the charts involved a confusion matrix heat map that depicted the spread of different matches over an NBA season. This is depicted in Figure 16 below.

A diagram of a spreadsheet

AI-generated content may be incorrect.

Figure

Although the model worked well at determining the winner of the match, it did not perform as well when it came to predicting the spread.

#### IST718 – Toxic Comment Discovery

Finally, with IST718, a lot of effort went into making the insights clear and easily understood by all audiences. Consequently, more conventional and easily interpretable ways of conveying information were used. Tables, bar charts, and word clouds were used to convey information and insights.

### Apply ethics in the development, use, and evaluation of data and predictive models (e.g., fairness, bias, transparency, privacy)

All three projects involved some ethical considerations when developing the models. For IST718, the dataset was reviewed to ensure the information received for the Reddit posts was anonymized before use. For IST707, consideration was given to the possibility that individuals who make use of the completed model could potentially develop and suffer from a gambling addiction. This was noted in the project proposal. However, IST692 was the one project that encouraged the most ethical best practices.

#### IST692 – Mortgage Approval and Gender Bias Analysis

Although IST692 included anonymized data, it also included features that fell under the purview of protected characteristics. These were gender and race. Although the feature for race was removed early on, through the process of exploratory data analysis, I discovered various imbalances in the data, one of which was gender. I also learned there was one gender classification that was not accounted for: non-binary. Models were created to assess whether any biases existed based on gender, and it was demonstrated that the model had a high bias toward males. This is reiterated in Figure 17, where zero represents males and one represents females. The selection rate is higher among males than females. Further, the false negative rate is higher among females than males. This finding led to the exclusion of gender as a feature from the dataset.

A screenshot of a computer

AI-generated content may be incorrect.

Figure

Also, although it was not listed in the final project for IST692 because it was removed early on, race was removed from the final preprocessed dataset because of gender disparities.

A screenshot of a graph

AI-generated content may be incorrect.

Figure

## 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, 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 solid foundation, and I am eager to continue refining my skills and applying them in real-world contexts.

# References

|  |  |
| --- | --- |
| [1] | L. Hanu, "Detoxify," 2022. [Online]. Available: https://huggingface.co/unitary/toxic-bert. [Accessed 09 12 2024]. |
| [2] | Sports Reference, "Basketball Reference," Sports Reference, NA. [Online]. Available: https://www.basketball-reference.com/. [Accessed 03 2024]. |
| [3] | NBA, "NBA Advanced Sats," NBA, NA. [Online]. Available: https://www.nba.com/stats. [Accessed 03 2024]. |
| [4] | Fairlearn, "Fairlearn," Fairlearn, NA. [Online]. Available: https://fairlearn.org/. [Accessed 10 2024]. |