# ADS Portfolio

By Mervin McDougall

Contents

[ADS Portfolio 1](#_Toc191199546)

[Purpose of the Report: 3](#_Toc191199547)

[Review of projects used in the report: 4](#_Toc191199548)

[Method: 5](#_Toc191199549)

[Learning Goals: 5](#_Toc191199550)

[Collect, store, and access data by identifying and leveraging applicable technologies. 5](#_Toc191199551)

[Create actionable insight across a range of contexts (e.g., societal, business, political), using data and the full data science life cycle. 7](#_Toc191199552)

[Apply visualization and predictive models to help generate actionable insight. 7](#_Toc191199553)

[Use programming languages such as R and Python to support the generation of actionable insight. 7](#_Toc191199554)

[Communicate insights gained via visualization and analytics to a broad range of audiences (including project sponsors and technical team leads. 8](#_Toc191199555)

[Apply ethics in the development, use, and evaluation of data and predictive models (e.g., fairness, bias, transparency, privacy) 8](#_Toc191199556)

## Purpose of the Report:

The purpose of this report is to provide a reflection and review of whether the learning goals of the MS. Applied Data Science program were met and whether I have demonstrated competency in applying these goals throughout my academic career while attending Syracuse University. The learning goals of the program are listed 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 accomplish this, I will be reviewing three projects which have been taken from three classes. The classes in question are IST692 Responsible AI, IST707 Applied Machine Learning, and IST718 Big Data Analytics. It is expected there will be some overlap of learning goals, however, I will highlight the strongest combination of skills that were learned in the process or completing these projects.

## Review of projects used in the report:

As indicated in the purpose of the report, I will be reviewing three projects across three different classes. However, before I do, I am going to provide a brief overview of the learning goals that were achieved through the development of these projects.

* IST682: Mortgage Approval with a focus on Gender – The purpose of this project was twofold, to generate a model that could adequately predict whether a user will secure a mortgage, but it should also investigate and account for any gender disparities.

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 Predictor – The purpose of this project was to generate a model that could predict the winner of an NBA match and the 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.

Use of Python to support the generation of actionable insight

* IST718: Toxic Comment Discovery – The Toxic Comment Discovery is designed to find a novel approach to identifying toxic comments. Its primary purpose was to develop a dataset that could be used to train other models to identify toxic comments.

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 effectively communicate how I have achieved the requirements of this program, I will identify each of the learning goals and explain how they were achieved through the exercise of the three projects mentioned above. This will ensure consistency and will make it easier to follow how I have demonstrated command of the required learning goals and application of the data science skills acquired.

## Learning Goals:

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

#### IST718

##### Background and Technology Used

Before embarking on any data science project, it is important to determine a hypothesis. This guides the development of any project and more importantly what your data sources will be, and how you will collect, store, and access the data.

For IST718, the purpose of the project was to develop a novel way of identifying toxic communication in social media posts, and the result would be a dataset on which future models could be trained. The goal was to generate a dataset that included topics from the discussions as well as a toxicity score and a label that indicated whether the communication was toxic or non-toxic.

After reviewing multiple data sources, the site Hugging Face was selected as it provided a large corpus of Reddit posts that could be used to develop a dataset. The chosen dataset had 3.23 GB of data making it difficult for performing exploratory data analysis on my local machine. Consequently, Google Colab was selected for future collaborative development because it freed up resources from my local machine and the machines of other team members and afforded the ability to run autonomously for up to 8 hours.

Given the volume of data that would be explored, it was determined that it would be best handled by making use of Apache Spark and by extension PySpark for data management. This would allow preparation and preprocessing of the data across multiple machines. The Google Colab environment provided the option of choosing my processor and the T4 GPUs were chosen to expedite the process.

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 we were processing was text data. Therefore, we made use of libraries such as NLTK, Spacy, and Scikit-Learn to help filter our stop words, generate word frequency distributions, and collect statistics about the corpus. These were applied by leveraging user-defined functions within PySpark giving the ability to apply transformations to the data without having to generate duplicates of the dataset.

##### Unique Approaches

Some unique approaches were developed in the process of carrying out 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 was so for the resulting data structure. Of the 1.8 million observations in the dataset, the data imported into the Google Colab environment was restricted to 50,000.

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. These were set to None to relinquish any resources tied to variables used in the environment.

Long processing times became a recurring theme throughout this project. As stated before, two of the reasons Google Colab was chosen was that it freed up resources on local machines and could run autonomously for up to 8 hours. This proved to be an important choice as the record processing easily stretched across multiple 8-hour periods. To avoid restarting the project, data was saved to parquet files at specific stop 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 columns for the topic, body, and comments, the data was combined to form post threads. This made it easier when performing LDA topic modeling and comparing the topics across different threads.

To address the toxic scoring of a post, a transformer known as Toxic Bert was utilized to rate the level of toxicity identified in the threads.

##### 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 title of the thread. Therefore, there was a lot of consistency in the discussion threads.

Moreover, 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.

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 scores that the Reddit contributors provided. Consequently, although there was higher engagement, identified through their Reddit post scores, on discussion threads this did not equate to a higher toxicity score.

##### Takeaway from the experience

This project tested my ability to think outside the box while 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.

#### IST707

##### Background and Technology Used

In IST718, the data originated from one source and underwent some feature engineering to extract insights. The processing was done on one dataset. On the other hand, the goals of IST707 required the data be gathered from multiple data sources and amalgamated into a singular dataset for training and testing.

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, data about team statistics, and individual player statistics had to be collected and combined to generate a dataset that could be utilized for training and testing.

Data was sourced from basketball-reference.com and nba.com/stats and combined to produce the necessary dataset for training a neural network.

##### Unique approaches

This project made use of a neural network which had its challenges - the primary being there was a need for sufficiently substantial amounts of data to train the model. Further, different model configurations could be pursued to optimize the model’s performance.

Google Colab was chosen as the development platform because it provided a free selection of different processors and allowed long processing times and collaborative work from other group members.

Through the process of developing the neural network, various sub-experiments were performed to test different hypotheses on how to improve model performance.

##### Insights gathered.

An exploratory data analysis of the data was performed before beginning work on the neural network. Several correlations between the combined fields of the dataset were discovered. It was determined that these may reduce the accuracy of the results if the full dataset were included in the final model. This was tested as in the process of developing the model as one of the experiments and it was determined that utilizing a reduced dataset that eliminated correlated fields produced a model with lower validation loss.

A graph of a train loss curves

AI-generated content may be incorrect.

Figure

### 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)