# ADS Portfolio

By Mervin McDougall

Contents

[ADS Portfolio 1](#_Toc193025704)

[Purpose of the Report: 3](#_Toc193025705)

[Review of projects used in the report: 4](#_Toc193025706)

[Method: 5](#_Toc193025707)

[Learning Goals: 5](#_Toc193025708)

[Collect, store, and access data by identifying and leveraging applicable technologies. 5](#_Toc193025709)

[Create actionable insight across a range of contexts (e.g., societal, business, political), using data and the full data science life cycle. 10](#_Toc193025710)

[Apply visualization and predictive models to help generate actionable insight. 11](#_Toc193025711)

[Use programming languages such as R and Python to support the generation of actionable insight. 14](#_Toc193025712)

[Communicate insights gained via visualization and analytics to a broad range of audiences (including project sponsors and technical team leads. 16](#_Toc193025713)

[Apply ethics in the development, use, and evaluation of data and predictive models (e.g., fairness, bias, transparency, privacy) 18](#_Toc193025714)

[Conclusion 20](#_Toc193025715)

[References 20](#_Toc193025716)

## 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.

* IST692: 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 state each of the learning goals and explain how they were attained through the completion 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.

In the first instance of introducing a project, I will provide a project background and technology used section to acquaint the reader with the goals and expectations for the project.

## Learning Goals:

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

#### IST718

##### 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. The goal was to exercise some feature engineering to enrich the dataset with new features such as topics using LDA (Latent Dirichlet Allocation) from the discussion posts and a toxicity score which calculated the probability of toxicity in the discussion posts.

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 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 an external machine. Google Colab was selected for data processing. It provided a platform for collaborative development with other group members who worked on the project, 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 processed, it was decided to use Apache Spark and by extension PySpark for data management. This would allow preparation and preprocessing of the data across multiple machines. Further, the Google Colab environment provided the option of choosing my processor. The T4 processors were chosen to speed up transforming 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 data set. 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 requires the creation of a new variable to hold the data. These variables were set to None to relinquish any resources tied to variables used in the environment. 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 was that it freed up resources on local machines and could run autonomously for up to 8 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 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 features for the topic, body, and comments. The comments feature was a JSON string consisting of a commentator key and comment value. The three features 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 comparing the topics across different threads.

To generate the toxic score of a thread, a transformer 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 and was the only time that the PySpark data was extracted to 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.

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. 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 to manage 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

##### Project Background and Technology Used

In IST718, the data originated from a single 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 collected 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 [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 sufficiently 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 free selection of different processors, allowed long processing times, and allowed collaborative work from other group members. Further, libraries such as TensorFlow, Keras, and Sci-kit Learn were 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

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 performed the role of 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 3 and Figure 4.

##### 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 along the way.

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

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

#### IST692

##### Project Background and Technology Used

In IST692, I reviewed mortgage data collected from the 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. The goals for the project were approached through different techniques, such as checks for dataset imbalances, checks for the predictors of approval, farness 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. Males provided most observations. However, the surprising discovery was that there was a small group of applicants for mortgage loans who checked both gender boxes on their application. For these applicants, I labeled them 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 male or female.

It came as no surprise that the resulting model would be biased toward males given the dataset consisted of a higher number of observations. But, when performing checks for False Negative loan approval classifications (loans which 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%.

Also, when checking for the major predictors for mortgage loan approval, it was learned that the loan-to-value ratio, preapproval, and log-income are great predictors for whether an applicant is approved for a loan over the entire dataset.

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 males, and they have the highest approval ratings among genders. However, the smaller represented 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

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 5 below. They helped inform us of 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 6.

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.

A graph of a graph showing a number of different levels

AI-generated content may be incorrect.

Figure

#### IST692

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 rate and gender imbalances in the applicants. For example, Figure 8 displays the proportion of approvals to denials of all the applications.

Figure

On the other hand, Figure 9 illustrates the gender imbalances captured in the data and how the non-binary group barely registers on the bar chart. 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

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

#### IST707

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 10 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, python line charts were generated to compare the training and validation loss and accuracy when adjusting the regularization of the model. In Figure 11 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 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 12.

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 13 below.

A graph of a number of values

AI-generated content may be incorrect.

Figure

### 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

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. Figure 14 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 the selection of a mortgage loan.

A graph with blue and white text

AI-generated content may be incorrect.

Figure

This finding was reiterated in the conclusion as some of the key ways to improve more equitable access to mortgage loans.

#### IST707

For IST707 a live presentation was performed among peers who were familiar with machine learning and neural network development. 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 that depicted the spread of different matches over an NBA season. This is depicted in Figure 15 below.

A diagram of a spreadsheet

AI-generated content may be incorrect.

Figure

#### IST718

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.

### 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. IST718 ensured 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 model, once developed, could potentially develop and suffer from gambling addiction. This was noted in the project proposal. However, IST692 was the one project that encouraged the most ethical best practices.

#### IST692

Although IST692 included anonymized data, it also included data that fell under the purview of protected characteristics. These were gender and race. 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 test whether any biases existed on the basis of gender, and it was demonstrated the model had a high bias toward males. This is reiterated in Figure 16 where 0 represents males and 1 represents females. The selection rate is highest among males than females. Further, the false negative rate is highest among females than males. This finding lead 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 this part of my final project for IST692, race was removed from the final preprocessed dataset.

A screenshot of a graph

AI-generated content may be incorrect.

Figure

## Conclusion

The two-year-long program has been challenging in the way it has made me think about, and assess the data used to train and assess data-driven models, and what the interpretation of the output from those models means. Through it all, I aggregated and demonstrated skills that in aligned with the expected outcome of the program: collect, store, and access data by identifying and leveraging applicable technologies; create actionable insights across a range of contexts 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; and apply ethics in the development, use and evaluation of data and predictive models. Looking ahead, I plan to investigate more statistical methods of evaluating the quality of datasets. I would also like to investigate alternative libraries that help to evaluate the fairness of dataset content. Visualizing data is an important part of communicating findings. Therefore, I also plan to find alternative libraries that make it easier to visualize data. But more importantly, I want to get involved in a community that keeps current on the trends and techniques used in Data Science.

# 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]. |