Syracuse University, School of Information Studies

Master’s in Applied Data Science

Portfolio Milestone

Shashank D. Nagaraja

SUID: 862678552

Table of Contents

[Introduction 3](#_Toc75476220)

[Project: IST 719 Final Project (Netflix) 4](#_Toc75476221)

[Problem Solved 4](#_Toc75476222)

[Tools & Techniques 6](#_Toc75476223)

[Insight Gained 7](#_Toc75476224)

[Project: IST 707 Final Project (Fashion MNIST) 7](#_Toc75476225)

[Problem Solved 7](#_Toc75476226)

[Tools & Techniques 9](#_Toc75476227)

[Insight Gained 10](#_Toc75476228)

[Project 738 Final Project (Political Party Classifier) 11](#_Toc75476229)

[Problem Solved 11](#_Toc75476230)

[Tools & Techniques 12](#_Toc75476231)

[Insight Gained 15](#_Toc75476232)

[Learning Goals 16](#_Toc75476233)

[Conclusion 16](#_Toc75476234)

# Introduction

During my time as a graduate student in the Applied Data Science program at Syracuse University, I have been exposed to a wide variety of statistics, computer science and analytics techniques. The skills that I have developed during this program will allow me to better understand data sets, process data and convey useful business insight to stakeholders that include colleagues and customers. Specifically in the field of Bioinformatics, the advanced modeling techniques covered during the master’s program will serve as useful tools for identifying and modeling the patterns, dynamics, and interactions between genes and cells.

In this portfolio, I will present three milestone projects that show my proficiency with the four pillars of Data Science - domain expertise, statistics, programming and communication of insights. While these projects range across several different business and academic domains, they are deeply rooted in the central concepts of statistics. These projects will also demonstrate how to leverage this knowledge and communicate with members within the data science community, and to stakeholders that may not have a statistics background.

I believe that completing a master's in data analytics has armed with a powerful arsenal of tools that I can use in my future research in the biotechnology industry. As the biotech embraces predictive analytics in the place of traditional first-principal models, I anticipate these skills to become useful for developing research and research related insights.

## Project: IST 719 Final Project (Netflix)

### Problem Solved

Netflix has grown to become the one of the most popular services for content streaming that we know around the world. Home to more than 6000 movies and TV shows, Netflix has grown from being a DVD rental service to a subscription-based model with more than 203 million subscribers.

This dataset was obtained from Kaggle. Originally obtained using Netflix’s API, it consists of all the TV Shows and Movies available on Netflix as of 2019. Processing of the data was carried out using R including extraneous information relating to the API request. The sharp effect of the pandemic in 2021 (Fig. 1) which caused severe impacts to the theatrical productions is evident from the aggregated data.

Chart, histogram

Description automatically generated

Figure 1: Yearly aggregate shows sharp increase of titles added to the Netflix platform upto 2020. Netflix was not on track to equal their 2020 performance in 2021.

Netflix also shows a preference for adding movies that were 2-3 years old instead of the latest movies. In order to strike a balance between broadcast rights acquisition costs and delivering fresh, relatable content to its users, Netflix focuses on movies that are not very expensive to buy, but that are only a couple of years old. This awareness of their niche market also extends to the its focus on adult content (TV-MA). Unlike its counterpart Disney+, Netflix focuses on content for adult and teenage viewers with a small selection of children’s titles (Fig. 2). Since streaming services are not bound by the same laws as their TV Channel counterparts, Netflix is able to offer uncut, uncensored content to its vast majority of adult viewers. Some movies have an unknown MPAA rating; on closer inspection these movies are only available in countries outside the USA and have hence not been rated by the Motion Picture Association of America.

Chart, pie chart

Description automatically generated

Figure 2: Netflix shows a preference for content geared towards adults and older teenagers. Also note the sizable amount of content rated 'R'.

A wordcloud of the title descriptions (Fig. 3) shows that many movies & TV shows available are based on topics like “family”, “woman”, “young” and “friends”. In order to better understand the differences by genre, each genre has been color coded. For example, we see that Action and Adventure movies use words like “power”, “war”, “struggle” and “mysterious” while Family movies use words like “love” and “world” to draw their viewers in. Although this wordcloud only contained the words from English movies, it will be interesting to see how words in different languages and cultures are associated with movie genres.

Text

Description automatically generated

Figure 3: Wordcloud of titles available on Netflix, using description text that is shown to users along with the trailer.

### Tools & Techniques

Along with honing data munging skills, this project presented a unique opportunity to find ways to tell a data story through visualizations. The skills acquired included preparing, analyzing and visualizing data using R, the industry standard open-source language for statistics and data visualization. However, although this class encouraged the use of R, the focus was on effective data visualization rather than syntactic details of the language. This allowed a deeper understanding of how to present data to a viewer, allowing them to interact with the data and draw their own conclusions using simple yet powerful data visualizations. It was also a useful opportunity to demonstrate the elements of design we learned. For example, exploiting visual hierarchy to catch a viewer’s eye, or using special visualization techniques when the data calls for it (for e.g., Choropleth Maps or Alluvial Plots) are valuable techniques that will come of great use. Finally, a possible business strategy was explored that would allow Netflix to increase it’s share amongst its competitors by identifying areas for improvement. For example, it was seen that Netflix had a very sparse selection of PG-13 movies. By increasing the number of available PG-13 movies, Netflix could bring more value to existing customers and even entice churn from competitors like Disney+.

### Insight Gained

Data visualization cannot simply be reduced to a finite list of plot types or standard tools. Instead, it is the construction of effective data visualizations meant to interact with and encourage the viewer to explore the data themselves and come to their own conclusions. In order to do this, a data scientist may choose to use simple, time-tested visualizations or create their own custom visualization using the principles of design. Perhaps the most important part of this job is to balance between adding too much detail (cluttering the viewer’s visual space and therefore thinking) and providing enough detail to let the viewer make their own decision on the underlying trend. This class brought to light the fact that good data visualizations are calculated and customized based on sound principles of design, and not just a collage of graphics and plots.

## Project: IST 707 Final Project (Fashion MNIST)

### Problem Solved

The MNIST database is a modified version of the NIST (National Institute of Standards and Technology) database. MNIST is commonly used in the field of training image classification systems in Machine Learning. MNIST is a repository of 28X28 images of handwritten digits (numerals 0-9). It primarily serves as a benchmark dataset against which multiple models are trained and evaluated. MNIST is a particularly convenient dataset to work with, because (i) the pictures can easily be evaluated by hand (ii) models have been shown to achieve remarkable performance on this dataset with relative ease. In fact, convolutional neural networks have attained error rates below 1% (Romanuke et al., 2016); even a simple linear classifier has shown to achieve only a 12% error rate (LeCun et al., 2013). However, with increasing access to model-building frameworks (sklearn, tensorflow, etc.) and computational power, the MNIST dataset has been criticized for not being complex enough. Since the classification task is only to recognize handwritten digits, it does not accurately represent the complexity of modern Computer Vision tasks. Therefore, in order to benchmark algorithms against complicated computer vision tasks, Fashion MNIST is a proposed dataset that builds on MNIST’s drawbacks.

Fashion MNIST contains a training set of 60,000 images and a test set of 10,000 images obtained from the German fashion retailer Zalando. Fashion MNIST is built to be a drop-in replacement for MNIST, with the same image size and train/test split. The images themselves are pictures of articles of clothing, with class labels such as “T-shirt”, “Coat”, “Bag”, and “Sneaker”.

Shape

Description automatically generated

Figure 4: Sample images from Fashion MNIST and their respective class labels.

### Tools & Techniques

In this project, two main skills were developed: model training and model tuning. Two models (Keras Neural Network & Naïve Bayes Classifier) were deployed; however, both these models were tuned in different ways. The Keras model was hand-tuned, whereas an automated hyperparameter optimization was used to tune the Naïve Bayes Classifier. Even though the models were different and were tuned differently, they were evaluated using the same metrics of classification (Accuracy, F1-Score, Precision & Recall).

The Keras Neural Network showed the best performance, although several dropout layers had to be added to greatly reduce model overfitting.

Graphical user interface

Description automatically generated

Figure 5: Model overfitting gratly reduced after adding several dropout layers. An accuravy of ~90% was obtained within 60 epochs of model training, indicating that the model had the potential of improvement if trained with more data and for more number of epochs

Interestingly, both models (Keras Neural Network & Naïve Bayes Classifier) showed similar points of confusion in the classification of the following classes: “T-shirt/top”, “Shirt” and “Coat”. From a quick look at the sample pictures, it is evident that the pictures for these classes may be quite similar, resulting in the models to misclassify them (Fig. 4).

Graphical user interface

Description automatically generated

Figure 6: Both the CNN and MNB models have similar points of confusion. However the CNN performs slightly better that MNB w.r.t (Accuracy, Precision and Recall).

### Insight Gained

Using Neural Networks or machine learning methods often involves trying several different models and comparing their performance based on a common metric. However, it is important to keep in mind that each of these models can also be tuned in order to improve and optimize their performance. While tuning networks, both automated and manual methods may need to be pursued; both model architecture (convolutional layers, dropout layers, activation functions, etc.) and model hyperparameters (smoothing parameter, minimum categories, etc.) can be modulated in order to improve model performance. This improvement is only incremental and is expensive (in terms of compute power); the gain in model accuracy may not be worth the additional time & money required to tune the model. With increasing access to cloud computing resources, the cost of automated tuning of machine learning models is reducing. Frameworks such as RandomizedSearchCV and GridSearchCV simplify the hyperparameter tuning procedure. However, model architecture tuning (especially in the case of Neural Networks) is largely done manually.

This project helped paint a clearer picture of the trade-offs and limitation of model tuning in both manual and automated workflows. While it is tempting as a data scientists to deploy as many models as possible, care must be taken to balance model improvement with training cost. Thus, it is helpful to establish a default dataset like Fashion MNIST so that models can be compared using a complex computer vision task without breaking away from traditional MNIST.

## Project: IST 738 Final Project (Political Party Classifier)

### Problem Solved

Although the two major political parties in the United States are built on seemingly opposing principles of liberalism (Democrat) and conservatism (Republican), these lines have been increasingly blurred in recent times. Complex socio-political conundrums like abortion, the COVID-19 pandemic and President Trump’s election are hot topics of discussion, further amplified by social media like Twitter and Facebook. Politicians themselves have realized the benefits of pandering to their niche voter base – using their social media platforms to increase their reach amongst voters. Therefore, it is increasingly valuable to be able to classify tweet content in order to find potential leads and target audiences.

They primary stakeholders for such a classification would be companies involved in OSINT (open-source intelligence tool) like political consulting firms to gauge public sentiment. It could also be extended to classify target audiences by classifying users as Democrat or Republican. When integrated with other identifying information using cookies (location, browsing history, device, age, gender, etc.), political affiliation could be used effectively to target audiences with political ads, fake news, propaganda, and even ads for products that show a skew towards a particular party affiliation.

Several companies have already identified the value in such a prediction; in fact, Cambridge Analytica used similar models (albeit much more complex) in 2014 to harvest data and build voter profiles. Models deployed on large quantities of data can often yield actionable insight for political campaigns. The aim of this analysis was to evaluate the performance of several models on Twitter posts made by US Senators and politicians.

### Tools & Techniques

The most challenging (and expensive) problem in building an accurate model is obtaining and labeling a large dataset of Tweets to train them on. Since labelling such many tweets is a time-consuming and expensive process, an alternate approach was necessary to circumvent this challenge. Twitter handles of 581 politicians were scraped from three different websites. These websites also contained the political affiliation of the politicians. The TWINT API allowed the entire post history of each account to be harvested from just the twitter handle. The true label of each of the tweets was assumed to be the political party affiliation of the politician that posted them. The most challenging (and expensive) problem in building an accurate model is obtaining and labeling a large dataset of Tweets to train them on.

Approximately one million tweets were obtained using the method stated above to construct a dataset.

Figure 7: 900,000 tweets were collected in total, and were moderately balanced between the two parties.

All models were evaluated using 5-fold cross validation. True Positive Rate (TPR), False Positive Rate (FPR), F1-Score and Model Accuracy were collected from each fold and averaged across runs to evaluate long-run model performance.

Text

Description automatically generated

Figure 8: Word cloud for tweets sent by Democrat politicians in the United States.

Text

Description automatically generated

Figure 9: Word cloud for tweets sent by Republican politicians in the United States

Figure 10: Model performance for various types of classification algorithms.

### Insight Gained

Through this analysis, a large amount of publicly available text data was collected using an API and organized into two class labels. Various models were trained using the k-fold cross validation technique, and model performance was evaluated using various classification metrics. The models were saved in order to be deployed later. By conducting this analysis, a complete arc of data procurement, preparation, analysis, and insight identification was completed. This exercise also served the purpose of developing re-usable and clean code in Python. By using object-oriented programming (to customize of-the-shelf models), robust application structure (pipeline construction) and version control (using git), the pillars of data science software development learned during previous courses were put into practice. The ability of various popular algorithms to classify text data was evaluated with the intention of being able to save and deploy these models into a production environment. The ethics of profiling individuals was also considered – although the tweets obtained were publicly available (and made by government officials), all location and personal data collected along with the tweet was ignored and discarded in order to respect the privacy of individuals.

# Learning Goals

The overall goals of the master’s in applied data science include several objectives designed to prepare a data scientist to gain insight and solve problems in the real world. With each unique project (along with several others during the course), a different mix of objectives were met creating a portfolio.

# Conclusion

The deliverables elucidated in this portfolio have contributed to an understanding of the process of data analytics. By completing projects in the sub-fields of Data Visualization, Natural language Processing, Text Mining, Statistics and Neural Networks, I have had the chance to prove that I have the tools to further my journey on the path of data science practice. Several projects over the duration of this course have shown me the immense business value a data scientist can bring to an organization by collecting, analyzing and explaining data and its trends to other business units. By extracting insight from data, I am confident that the role of the data scientist will become even more important in the years to come.

\*\*\*