Abstract

Emilio Ramos Monzalvo, a student at Syracuse University, provides indisputable evidence to the Applied Data Science Graduation Panel on his knowledge and experience in the subject. The student shows clear understanding of the Data Science Pipeline by portraying relevant projects while in enrolled in classes from the Fall of 2019 to the Spring of 2022.

Applied Data Science Portfolio

Syracuse University

Emilio Ramos Monzalvo

ramosem97@gmail.com – eramosmo@syr.edu – (479) 925-8904

github.com/ramosem97/SyracusePortfolio

Table of Contents

[I. Introduction 3](#_Toc98066834)

[II. Personal Statement on Data Science 3](#_Toc98066835)

[III. Supporting Projects 5](#_Toc98066836)

[A. Historical Database of NFL Fantasy 5](#_Toc98066837)

[B. NFL Game Analysis and Prediction 6](#_Toc98066838)

[C. Public Sentiment Towards COVID-19 8](#_Toc98066839)

[D. Flower Image Classification 9](#_Toc98066840)

[E. Marvel Cinematic Universe Search Engine 10](#_Toc98066841)

# I. Introduction

I have been a Data Scientist since before I knew Data Science was an area of study. All it takes to call yourself one is a computer and data. Nevertheless, the bittersweet fact of data’s exponential growth in scale and complexity means I need to follow suit to keep up with a highly competitive area. The Applied Data Science Masters at Syracuse University has provided me with the necessary tools to keep chasing after data by exposing me to real-world problems and experiences. As supporting evidence of Syracuse’s influence and progress on my career as a Data Scientist, the project portfolio includes proof of the extensive experiences I had while being a student. In other words, the training through Syracuse’s course load shaped me into a Data Scientist ready for deployment to keep solving data problems.

# II. Personal Statement on Data Science

The idea of modeling population’s behavior was part of a science-fiction novel from 1942, *The Foundation* by Isaac Asimov, until recent history. The novel called the area of study ‘psychohistory’ which was used to correctly predict the inevitable end of the galaxy. Unlike the book, the real-life counterpart can be enclosed in the field known as Data Science. It was first theorized in 1962, but it did not come to life until the early 2000s. While there are clear differences in Data Science and ‘psychohistory’, *Isaac Asimov* gave civilization a hint of the powers that Data Science can have on the world.

Data Science has grown exponentially alongside, and thanks to, computers. Moore’s Law of the growth of computer capability has held true for the past 50 years, but it might not hold much longer now that microchip technology is reaching towards the quantum realm, an area that has yet to be a reality. Nevertheless, the capabilities of Data Science achieved since the 2000s have created endless applications. An application can range from improving wine by predicting the best time of year to harvest, to improving traffic lights using traffic data. The only application limited to Data Science is one where data is not available yet. Anyone with a hard drive full of data and a computer can become a Data Scientist.

The diversity of data and the number of growing applications has undoubtedly rocketed the research and development of new Machine Learning and Data Engineering technology. Unfortunately, some crucial areas of Data Science have fallen behind. Ethics in Data Science is one of them. The significant consequences a faulty algorithm can have on individuals, businesses or populations is not a standard practice in most applications.

Social Media creates a user’s feed using the output of a model that takes the historical interactions as the input. With high accuracy, the feed will likely include posts one is predicted to like; however, the same model can prevent the user from seeing other posts outside of their current interests. While users are more likely to come back, their likeliness to learn or experience something new from the same platform are lowered even if the information is out there.

It has become a reality that a Data Scientist’s algorithm controls what information is seen. Therefore, it is up to Data Scientist’s to not only know and apply the language of data, but to consider the ethical impacts that a model can have on the world.

# III. Supporting Projects

In the Applied Data Science Master’s Program at Syracuse University, one is presented with a course load of thought-provoking and application-based classes. Through these classes a student is not only challenged to learn about fundamental and complex topics, but they are also encouraged to apply them on real-life problems.

A project usually consists of a problem statement or a rubric, but the data and the objective questions are left for a student to gather and answer. With open-ended problem statements, the students can creatively find solutions while still applying techniques a Data Scientist needs to work in the real world. Furthermore, a couple of examples are illustrated which prove the Program’s effectiveness in teaching Data Science.

\* File Navigation in GitHub Repo: Projects / Quarter - Class Title - Class Name / {Project File(s)

## Historical Database of NFL Fantasy

**Problem Statement**

Each member of the Ramos family has been predicting the winner of every NFL game spanning from 2004 to 2021. Each season, there is a winner based on how many points they scored correctly. Now that more than 17 years have passed, an excel spreadsheet is lacking the ability to store and visualize the historical and current results.

**Approach**

To better store and manage the predictions, a database schema was constructed by considering the customer’s future need to analyze the data. This includes acknowledging a participant’s accuracy towards a specific team, home advantage, or the teams’ season statistics. The schema was then converted into a real database using the corresponding SQL language. Data was then added after each season’s predictions were processed in Python while pulling other information about the games from the web. Once the data was cleaned, the data was then inserted into the corresponding table in the database while considering the different table keys.

**Class Information:** IST659 – Data Admin & Management – 1st Quarter

**Technologies and Techniques:**

* Database Management: SQL Programming Language, Building DB Schema, DB Creation.

**Ethical Considerations:** The gathering, storing, and using of customer data can contain private information and cross data privacy laws.

**Project File(s):**

* Syllabus:
  + IST\_659\_Syllabus.pdf
* Project Rubric:
  + IST659+Project+Description.pdf
* Report:
  + Deliverable1\_Emilio\_Ramos\_Monzalvo.docx
  + Deliverable2\_Emilio\_Ramos\_Monzalvo.docx

## NFL Game Analysis and Prediction

**Problem Statement**

Each member of the Ramos family has been predicting the winner of every NFL game spanning from 2004 to 2021. Each season, there is a winner based on how many points they scored correctly. Now that more than 17 years have passed, an excel spreadsheet is lacking the ability to store and visualize the historical and current results. Historical data is often disregarded in current seasons, but to maintain bragging rights, the family is exploring using the data to expand their approaches to predicting NLF games correctly.

**Approach:**

To improve the predicting process, multiple hypotheses were put to the test including home-field-advantage, most offensive points per game, and least defensive points allowed. The dataset did not previously include the results of each NLF game, but instead it contained only the prediction on whether one team would win or lose. An API was used to pull data into python that could then be merged to the historical data in excel. Using Python’s Pandas and NumPy packages, the data was cleaned while also visualized to find important patterns in the data. The hypothesis was then tested using corresponding Hypotheses Testing methods like Chi-Squared. Once the best method was identified, it was put in place for the 2020 season and the SQL was computed to understand its effectiveness.

**Class Information:** MCB638 – Data Analysis and Decision Making – 1st Quarter

**Technologies and Techniques:**

* Programming: Data Frames, Python3, Pandas, NumPy, Jupyter Notebook.
* Data Visualization: Matplotlib and Seaborn.
* Statistics: Hypothesis Testing.

**Ethical Considerations:** Will the use of historical participant data provide other users with an unfair edge over the others.

**Project File(s):**

* Syllabus:
  + MBC\_638\_Syllabus.pdf
* Report:
  + StoryBoard.pptx

## Public Sentiment Towards COVID-19

**Problem Statement**

In early 2020, the COVID-19 had barely begun in the United States and news outlets were providing contradicting information regarding public opinion on the subject. The scale of the pandemic was still not known, but the working-class was now having to deal with new ways to work from home. Politicians were scrambling to figure out how their constituents felt about the subject to drive their actions based on public opinion.

**Approach:**

COVID-19 opinions were gathered, labeled, and posted on Kaggle.com for developers to test different NLP models that could be used to understand the public’s sentiment. The tweets were broken down into a training and testing set. The tweets were then cleaned by firstly tokenizing them to gather a list of each word, hashtag, or mention in the tweet. The words were then lemmatized to improve the model’s ability to compare words against each other from different tweets. The tweets were then evaluated by multiple algorithms including Random Forest and a Deep Neural Network. These models included other important features like favorites and retweets that could also infer the public opinion. The results in the test set were then compared using Confusion Matrices.

**Class Information:** IST664 - Natural Language Processing – 5th Quarter

**Technologies and Techniques:**

* Natural Language Processing: Sentiment Analysis, Text Processing, Tokenization, Lemmatization, Regular Expressions (Regex), Bag of Words Model.
* Modeling: Random Forest, TensorFlow, Neural Networks, Deep Learning, Model Comparison.
* Programming: Twitter API, JSON Format.

**Ethical Considerations:** The public’s understandings of COVID-19 might be misleading by false information on the Web. Similarly, a politician’s actions can affect the public’s beliefs in wearing masks or social distancing if the model provides misleading results.

**Project File(s):**

* Syllabus:
  + IST\_664\_Syllabus.pdf
* Project Rubric:
  + Final Project Rubric.pdf
* Report:
  + CovidSentiment.docx

## Flower Image Classification

**Problem Statement**

It is a well-establish fact that there are more plant and flower types than a human can remember. Flowers can be mislabeled and then undervalued or overvalued. I.e., a customer might pay more for a common plant while a seller might make a loss on a sale. Therefore, a flower should be easier to identify than by word of mouth.

**Approach**

An existing list of images was extracted from Kaggle.com containing classified images of flowers. The files contained more than 50 common flowering plants with at least 20 images each which exceeded eight gigabytes. The scale of the dataset led to the team to outsource the code to Google’s paid service of Google Colab. In this environment, the images were then processed, rescaled, amplified, and standardized to meet models’ input requirements. The model’s training and results were converted into a Python pipeline to prevent the environment to reach its RAM and CPU limits. Multiple models were chosen like Random Forest, Decision Tree, Naïve Bayes, and Inceptionv3, a pre-trained Neural Network. Each model was ran using the same image pipeline, train set and test set. The industry standard of the Inception pre-trained model topped the other’s by maintaining the highest accuracy across all of the other models.

**Class Information:** IST718 – Big Data Analytics – 4th Quarter

**Technologies and Techniques:**

* Data Prep: Image Data Wrangling
* Programming: Google Colab.
* Data Modeling: Computer Vision, Neural Networks, Deep Learning, Model Pipeline, Model Comparison, Pre-Trained Neural Networks, Inception Model.

**Ethical Considerations:** Public datasets can contain mistakes on manual image labeling. Mislabeling can mislead users of an app into buying or selling plants or flower without proper knowledge and price.

**Project File(s):**

* Syllabus:
  + IST\_718\_Syllabus.pdf
* Project Rubric:
  + final\_project\_report\_instructions.pdf
* Report:
  + FlowerRec.pdf
  + IST 718 Final Group Project Statement.dox

## Marvel Cinematic Universe Search Engine

**Problem Statement**

Phase one through three of the Marvel Cinematic Universe (MCU) consists of twenty-three movies which adds up to more than three thousand minutes of viewing. This, of course, makes it challenging for someone that wants to watch the historically relevant movies when a new one for phase four is about to come out at the cinema. If one had no previous knowledge of what movies might be relevant, there would be no choice but to view all three thousand minutes of these first three phases. Phase four of the MCU is going to consist of twelve movies and thirteen tv shows, so even further into the future, getting up to date with the MCU will get exponentially harder with so many movies coming out.

**Approach**

First challenge included scrapping the web for the movie scripts of the 23 movies in phase one through three. Each movie was found with distinct patterns like file formats or character line patterns, so a general approach to standardize the scripts was taken using Regex. Some. Scripts found as PDF’s were then converted into text documents using a PDF to Text python library. There were two models chosen: IMB’s Best Match 25 (BM25) and Google’s Universal Sentence Encoder (USE). The models were assessed based on real world application like its ability to show a user the best relating result while also making sure the search’s semantics were considered by the model. The best model was then used to create a Search Engine Web Application using python’s PyDash package.

**Class Information:** IST719 – Information Visualization – 8th Quarter

**Technologies and Techniques:**

* Text Mining: Google Universal Sentence Encoder (USE), Regular Expressions (Regex), BM25, Web Scrapping, Fuzzy Matching.
* Natural Language Processing: Tokenization, Lemmatization.
* Data Visualization: Plotly, PyDash, HTML, JSON.

**Ethical Considerations:** While the search engine is meant to inform Marvel fans, it might enrage some users by displaying spoilers without their consent.

**Project File(s):**

* Syllabus: IST\_736\_Syllabus.pdf
* Report:
  + Final Project Report.docx
  + IST-736-Final Project.pptx