**Portfolio Milestone**

Michael Carvalheiro

SUID: 968446790

Syracuse University

M.S. Applied Data Science

<https://github.com/jmike877/iSchool_Portfolio>

Prior to beginning the Applied Data Science program, I completed my undergraduate degree in Mathematics, during which I gained an extensive overview of theoretical and applied statistics and probability. However, I did not have an extensive background in computer science, I had only used Java to code briefly. Within the ADS program at Syracuse, I learned three different coding languages (Python, SQL, and R) to help me reach the multitude of goals the program looks to achieve to help a student become a strong data scientist. These goals were:

1. Describe a broad overview of the major practice areas of data science.
2. Collect and organize data.
3. Identify patterns in data via visualization, statistical analysis, and data mining.
4. Develop alternative strategies based on the data.
5. Develop a plan of action to implement the business decisions derived from the analyses.
6. Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization.
7. Synthesize the ethical dimensions of data science practice.

In this paper, I will discuss four projects that I completed during the ADS program which highlighted these goals.

In my final project for IST 719 Data Visualization, I analyzed a dataset of English Premier League soccer transfers from the past 15 years. This first entailed collecting and organizing the data, which was done by scraping the data using Python off the <https://www.transfermarkt.us/> website then cleaning the data (dropping unnecessary columns, dropping rows with NULL values, etc.) using R. The second step of the final project was exploratory data analysis, which was performed in R. The goal of this project was to produce a poster with various visualizations that told the story of the data and any patterns/trends that were uncovered. I presented this poster in front of my fellow students and staff members, where I was able to communicate effectively what I had uncovered via exploratory data analysis.

In my final project for IST 707 Applied Machine Learning, I utilized Python to create a Machine Learning model to predict whether it was optimal to bet on the home team to win the match in the English Premier League. The first step once again was to collect and organize the data, which was done by scraping the data using Python then cleaning the data (dropping unnecessary columns, dropping rows with NULL values, etc.) again using Python. While cleaning the data, I performed feature engineering to produce optimal feature columns that were used in my Machine Learning models to predict the outcome variable (home team win). I then performed a brief exploratory data analysis to uncover trends within the data. Then I implemented a Random Forest model, a Logistic Regression model, and a k-Nearest Neighbors model. After running each model, I decided to optimize the Logistic Regression model since it performed the best out of the three mentioned previously. I did this by running a Recursive Feature Elimination method to see if reducing the number of features in our models would result in a higher accuracy score. I ended up reducing the total number of features from 40 to 25, which increased our accuracy score by 0.003. This was all communicated effectively in my final project presentation, where I had to present my findings to my fellow students and my professor.

In IST 718 Big Data Analytics, my final project was to predict airline delays via Machine Learning. Again, the first steps were collecting and organizing data, feature engineering, and performing a brief exploratory data analysis. Utilizing Python’s Pyspark package, I implemented various machine learning algorithms, such as linear regression to predict how long a delay would last, and different classification algorithms (Random Forest, Logistic Regression, and Gradient Boosted Decision Trees) to predict whether there would be a delay. The Pyspark package was used in this class because we worked with very large datasets and Spark is able to do computations on these large datasets much quicker than Python. To optimize each model, I performed hyperparameter tuning and cross-validation. Again, this was all communicated effectively in my final project presentation, where I had to present my findings to my fellow students and my professor.

For IST 664 Natural Language Processing, my final project was to analyze movie review sentiment using various Natural Language Processing techniques such as tokenizing, pre-processing/filtering, bag-of-words, parts-of-speech tagging, subjectivity, and sentiment lexicon. Again, the first steps were collecting and organizing data which was fairly easy this time since there was an extensive movie review dataset on Kaggle. Then came implementing all the NLP techniques listed above in Python. Then I ran a Naïve Bayes classifier and a Logistic Regression classifier to predict sentiment of movie reviews. To optimize each model, I performed cross-validation. Again, this was all communicated effectively in my final project presentation, where I had to present my findings to my fellow students and my professor.