Syracuse University
MS Applied Data Science
Hwang, Kelly
SU ID: 361712622
Github: https://github.com/kehwang/SyracusePortfolio

Final Portfolio Milestone

Introduction

As our devices become more interconnected, having the ability to digest and parse data has never been more important. Big data is now entangled into every aspect of our lives, and business who learn to capitalize on this new avenue of information will come out on top. Data science is the multidisciplinary field that utilizes the scientific method, processes, algorithms, and frameworks to extract meaningful insights from structured and unstructured data. It's a unique area of study that blends the art of business intelligence and technical prowess. As a current professional in the technology industry, having the ability to collect, organize, and analyze large quantities of data is extremely vital; it will become a necessary skill for any serious analyst as technology continues to advance.

This driving force of data in everyday life is what compelled me to seek additional education and training. Since I already have experience working with data, I knew how much I didn't know. I absolutely needed to upgrade my skills to being able to make accurate predictions, and have hands on experience with machine learning. I also needed to improve upon the knowledge I already had, using this curriculum as a way to identify and fill in the gaps to my analytical foundation.

Program Learning Objectives

It's clear that collecting, organizing, and cleaning data is one of the most important aspects of data science. Every assignment, project, and deliverable for all my courses required many hours spent studying the data's structure and how to best process it to fit modeling requirements. Data munging can be quite challenging, and sometimes requires creativity to mold the data into what is needed. An effective data scientist should also be able to identify patterns in data, whether it be with visualization, exploratory analysis, or data mining techniques. Being hands on with Python and R has introduced a number of packages to choose from to aid in finding trends. GGplot, Seaborn, and Matplotlib are the main libraries I used for visualization purposes.

With programming languages comes an arsenal of machine learning algorithms for predictive analytics. Although overwhelming at first, with continued exposure it was clear that there are three main types of algorithms and use cases:

- Supervised Learning (Predictive): labeled data is given to the machine for training, and the machine will be able to classify new items with the same label. The algorithms learned for

classification are: linear regression, support vector machines, decision trees, naive bayes, and KNN (nearest neighbor).

- Unsupervised Learning (Descriptive): input data is given and the model is run on it. The input given is grouped together and insights on the inputs are the outputs. This is useful for clustering and anomaly detection, where relationships between data points need to be identified, not a predicted value. The algorithms learned for clustering are: k-means clustering and association rule mining.

- Reinforced Learning: the machine is exposed to an environment where it gets trained by trial and error, and then trained to make a specific decision. The machine learns from past experience and uses that to make accurate decisions in a feedback loop. This is where deep learning and neural networks come into play, and unfortunately the program wasn't able to cover this topic in detail.

But it's not enough to just be able to clean, manipulate, and model data. A data scientist needs to also understand the needs of the business and business context. In order to be successful at finding solutions, the right questions need to be asked so the correct problems can be identified. This is why having a defined framework for approaching problems is so important. One framework learned is DMAIC:

- Define: what is important, the goal, potential resources, business case, project scope, and project timelines.
- - Measure: what should be measured, how do we measure it, establish baselines, compare performance.
- Analyze: identify, verify, validate, and perform root cause analysis.
- Improve: identify, test, and implement a solution to the problem.
- Control: how to maintain the improvements, sustain the gains, and create a control plan.

Defining the problem, identifying the solutions, and effectively communicating the results to stakeholders: this is what it means to be a data scientist.

Portfolio Overview
Github: https://github.com/kehwang/SyracusePortfolio

The following projects were included in my portfolio to demonstrate the above learnings and outcomes from the program. Most notably these projects showcase a mix of data cleaning and manipulation of different types of data sets, exploratory analysis and visualization, supervised learning models, business framework implementation, and deriving actionable insights through the data.

- **Course: Data Analysis & Decision Making**
    - Skill Concentration: Microsoft Excel, Statistics in Business
    - Course Summary: Basic statistical techniques and their appropriateness to situations and assumptions underlying their use. Focuses on the concepts, principles, and methods to support a scientific approach to managerial problem solving and process improvement.
    - Project Deliverable: Process Improvement for Account Cancellations
    - Learning Applications: This project required understanding a business problem, gathering the required data, performing analysis, implementing solutions, and measuring any success of the outcomes. Using the framework taught in class, DMAIC executes techniques for defining, measuring, analyzing, implementing, and controlling process improvements. The data and business problem was gathered with permission from my employer, modified slightly for external academic use. Data analysis techniques were done with Excel and the final deliverable was a powerpoint presentation.

- **Data Mining**
    - Skill concentration: R
    - Course Summary: Data mining techniques for classification, clustering, association rule mining, descriptive and predictive analytics, communication skills.
    - Project Deliverable: Mushroom Classification
    - Learning Applications: This deliverable mainly demonstrated the need to study the data and figure out how to transform it into something a supervised learning model could use. The data attributes in their original form was letters, and using data manipulation techniques I transformed these letters into useable numbers. The machine learning algorithms used for classification were Naive Bayes and Random Forest. The data set was retrieved from Kaggle's mushroom classification competition. Although relatively short in terms of code-length compared to other projects, this was one of the first classes I completed during the program, and records a painful but necessary learning curve.
    - R libraries used: ggplot2, caret, randomforest, e1071, klar

- **Scripting for Data Analysis**
    - Skill Concentration: Python
    - Course Summary: Building data analysis pipelines. Acquiring, accessing and transforming data In the forms of structured, semi- structured and unstructured data.

- Project Deliverable: An Exploration of Yelp Restaurant Reviews

- Learning Applications: The majority of this project utilized data collection, parsing, cleansing, organizing, and extraction. The original format was JSON which had to be parsed into a more digestible format for Python. Using python as the language of choice, exploratory analysis was conducted on restaurant reviews. Data visualization, statistical analysis, and sentiment analysis (of unstructured text) were methods used in this project.

- Python tools used: JSON, pandas, re, itertools, numpy, textblob, scipy, statsmodels

- **Natural Language Processing**
  - Skill Concentration: Python

  - Course Summary: Foundations of natural language processing using Python. Focus on unstructured text analysis.

  - Project Deliverable: Kaggle Movie Reviews Text and Sentiment Classification

  - Learning Applications: Analysis of unstructured text from movie reviews. Training Naive Bayes with 3 different iterations of the data, to classify and predict positive or negative sentiment. Additional modeling done with sklearn Logistic Regression and Random Forest. Unstructured text analysis used techniques such as tokenization, frequency distribution, post processing (cleaning and removing stopwords), and generating feature sets for model processing. Models were evaluated using Precision, Recall and F1. All code was done in Python.

  - Python tools used: os, sys, random, NLTK, re, sklearn, sentiment lexicon (external source)

- **Big Data Analytics (Advanced Analytics):**
  - Skill Concentration: Python

  - Course Summary: Using case studies and analysis techniques to solve real world problems. Regression analysis, time series forecasting, image detection and classification.

  - Project Deliverables:
    - Case Study 1 - Salary Recommendation, NCAA Football Coach Analysis using OLS Regression
    - Case Study 2 - Predicting Best Investment Zip Codes With Time Series Analysis Using Prophet

  - Learning Applications: These case studies studies were an opportunity to take all knowledge gained from the core curriculum and demonstrate the ability to combine data

sets to produce meaningful analysis. The aim was to provide the decision maker with insights and understanding. It is a combination of all of the learning objectives of the program:

- **Obtaining** data and understanding data structures
- **Scrubbing** data using scripting methods
- **Exploring** data using qualitative analysis
- **Modeling** relationships between data matched to the information and needs of clients and users
- **Interpreting** the data, model, analysis, and findings
- **Communicating** in a meaningful way

- Python tools used: pandas, numpy, matplotlib, seaborn, fbprophet, scipy, statsmodels