M.S., Applied Data Science

**Portfolio Milestone Overview**

**Jagannathan Govindan**

[jgovinda@syr.edu](mailto:jgovinda@syr.edu)

Sep 2020

Syracuse University

MILESTONE PORTFOLIO

TABLE OF CONTENTS

**Overview**3

Broad View of major practice areas3

Collect and Organize Data3

Identify Patterns and data visualizations4

Develop alternative strategies4

Develop Plan of Action4

Develop communications skills using Data5

Synthesize ethical decisions5

**Key works and Projects**5

Game of Thrones Character Analysis5

Obamacare Healthcare Marketplace Analysis6

Multifactor Authentication and its future in Banking Industry6

Coach Salaries7

Zip code Investment7

Law and Weather8

**Overview:**

I got first introduced to Data Science and its emergence after attending the ‘Big Data day LA’ in 2016 and it provided me with the opportunity to visit University of Southern California (my first American campus visit) and also the insights from industry leaders like Netflix, Amazon, et., all. As an IT professional with over 8 years’ experience at that time, I understood the significant impact data is going to have for a lot of years to come. So in 2017 I decided to apply for Data Science courses and in 2018 I officially started my course in Syracuse University.

This course (MSADS) at Syracuse gave me introduction to tools and tips to handle data, model it for inference, and visualize to tell a story and help in decision making process. This portfolio will demonstrate the achievements of the overarching learning objectives laid in the program.

1. *Describe a broad overview of the major practice areas of data science*

Data Science as the name implies can be a use practice tool for almost any domains, from healthcare to politics as long as enough data and data scientists are available. Data Scientists collect, organize, clean, model and interpret data to bring insights as well as stories to help managers in the decision making process. It could be achieved by using add-ons in Excel to using models in programming languages like R or Python. The above steps will help not only the managerial decision making but ordinary users of data to pull relevant information which otherwise they would miss. The “Key works and Projects” section will demonstrate this broad overview of major practice area.

1. *Collect and organize data*

Collecting and organizing data is the fundamental and the most difficult part of a Data Scientist’s job. As the saying goes “Data is everywhere”, so where we collect became a significant part of our class conversations as well as how to interpret it (thru visualizations and other models tools) provide you with the opportunity to demonstrate your understanding of the data. We got the message that the interpretation of data is very subjective and small manipulation of data can provide a completely different recommendation to a data problem.

Only final note on this topic is data handling in cases of missing data or outliers. How these scenarios are handling can change from one organization to another so the Data Scientist should be well equipped with the organizational culture before dwelling deep into data to avoid inadvertent errors.

1. *Identify patterns in data via visualization, statistical analysis, and data mining*

This is the logical next step after organizing data and there are innumerable tools (open source or otherwise) to visualize, analysis and mine data. The thumb rule that this program demonstrated for me is to start with basic tools like Microsoft Excel which has an excellent source of inbuilt tools to visualize and to run statistical analysis for providing insights into the data. Running a simple statistical analysis in Excel will provide you the directions needed with respect to the available data and then regression techniques will help providing the significant variable(x) on the desired outcome(y). Based on the need, further data mining can be done using other techniques and plots like Boxplot, Whisker plot can help visualize the outliers.

1. *Develop alternative strategies based on the data*

Data Science provides the opportunity to be flexible in selecting a path of investigation of the available data. So when outcome from a model intended is not achieved or accuracy of the predictive analysis not satisfactory Data Scientist responsibility is to consider the alternative approaches to arrive at the best possible outcome. Some of the projects described in “Key works and Projects” will demonstrate the path the data took us rather than being rigid about a set methodology.

1. *Develop a plan of action to implement the business decisions derived from the analyses*

After collecting, organizing and modeling data (also considering alternative strategies) the next step is to provide the action plan for business decisions. This can come in the form for providing different marketing strategies or cost benefit analysis for business decisions. It will logically serve the time and effort involved in the prior steps in data handling.

1. *Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization*

After collect and organizing data, this is the next difficult step in Data Science. In any organization not all are equipped with statistics jargons or key words from Information Technology. So the responsibility of a Data Scientist is to articulate efficiently the analysis and path to the arrived decision by showing the professional communication skills to help understand the presentation from beginner level.

The project presentations in this programs are routinely insisted not to provide the destination (which is the arrived conclusion) but the journey the team taken to arrive at it.

So presentations were geared towards an audience with different skillsets in data science.

1. *Synthesize the ethical dimensions of data science practice*

Information Security is a field that is getting relevant day by day. I took two courses in this program to understand significant risks involved in handling data and privacy concerns. The past experiences of hacking and publishing of customer’s data is a good learning curve for a novice like me who is entering in the Data Science field.

**Key works and Projects:**

Game of Thrones Character Analysis| IST687

**Summary:** This group project analyzes the user ratings of the TV series Game of Thrones. Game of Thrones (GoT) is a fantasy TV-series that premiered on HBO in 2011; it is an adaption of the book series written by George R.R Martin. Due to the popularity and continued success of Game of Thrones, this predictive model will attempt to predict the demographic user ratings of GoT episode based on the characters and their associated screen time in the episode.

**Learning applications:** This was the first data analysis project in which we had to obtain data (scrubbing it from Internet Movie Database) and cleaning and then preprocessing for different data models using R, also had an interactive visualization and presentation using Shiny in R.

Overall, we realized that project was very ambitious especially in this short period time, however we believe that a prediction can be made based on what characters are in an episode and their screen time. While we had a great deal of characters and screen times, there were only 67 episodes and more episodes were required to enhance the models abilities to make a prediction.

**Learning Objectives:** 1, 2, 3, and 6.

Obamacare Healthcare Marketplace Analysis| IST565

**Summary:** This individual project looked into the open source data released by Centers for Medicaid and Medicare Services (CMS) and it contains the different insurance plans for medical and dental categories based on the corresponding network and geographic locations.

In addition to different visualizations, I also ran predictive algorithms to determine the statistically significant variables influencing the premium prices across different locations and created a model which predicts the future premium pricing based on geographic locations and demographic data.

**Learning applications:** Initial data cleaning and preparation took the bulk of the effort but visualizations revealed the story of how healthcare products are marketed in different geographic locations and the metal level marketing sales depended on demographics of the region. In this project learnt to visualize data to tell a story, generating a model requires multiple iterations of running algorithms to final get the best possible outcome.

This was also an ambitious project as I was learning the ropes with respect to tools in R and Python and the data was huge so lot of effort was needed in preparation and cleanup. However this project prepared me for the complex assignments in the future courses.

**Learning Objectives:** 1, 2, 3, 4, 5, 6, and 7.

Multifactor Authentication and its future in Banking Industry| IST704

**Summary:** This group project looked at the Information Security issues faced in Banking sector and the possibility of MFA (Multi Factor Authentication) to mitigate some of them. As part of this project, we analyzed the recent banking industry information security violations and their root cause for it, from it the alternate strategy (as a mitigation measure) we looked at multifactor authentication up to 4 levels of checks to reduce fraud.

**Learning applications:** As part of this, learnt the lessons about the application of information security and ethics for organizations as well as individuals in handling customer data. Information security is going to paramount given the data generated every day and we looked at one alternate strategy for banking sector to ensure better data security for their customers.

**Learning Objectives:** 1, 2, 4, 6, and 7.

Coaches’ Salaries | IST 718

**Summary:** Using coaches’ salary data provided by the professor, this case study recommends a base salary for the next head football coach at Syracuse University. The investigation uses supplemental data, like graduation rate of football players, stadium size of each school, in order to make an educated decision.

**Learning Applications:** This lab requires students to build a data frame, fit a regression model and find relevant predictors, and think creatively about how to organize and reorganize the data. The lab also introduces hypothetical scenarios to see if changing the setting would also alter the outcome, which prompts more creative data manipulation to answer the question.

**Learning Objectives:** 1, 2, 3, 4, 5, and 6

Zip Code Investment | IST 718

**Summary:** This case study investigates how we can predict three zip codes that provide the best investment opportunity for the Syracuse Real Estate Investment Trust (SREIT). Using historical data of housing prices, it will calculate the compound average growth rate of every zip code from 1997 - 2018 in order to identify the zip codes with the best CAGR. After the data has been down sampled, a time series regression is performed using FBProphet to predict future house price trends.

**Learning Applications:** This lab combines data modeling with some financial analytics. The data from Zillow is used to create a brand-new variable (CAGR) that is used to as a technique to down sample the large dataset. After performing the initial analysis and running the time series through FBProphet, the RMSE is also calculated and used as a variable to make the decision. This lab is unique in that it requires the student to use a new package and relies on Google Colab notebooks as a means of completing the modeling more efficiently, thus forcing students to branch out and look at the latest and greatest software.

**Learning Objectives:** 1, 2, 3, 4, 5, and 6

Law & Weather | IST 718

Summary: This is an investigation into the possibility of using weather data to predict crime incidents in two different cities. Crime and weather data are gathered for the city of San Diego, which is used as a “control city” since the temperatures in San Diego are relatively stable. Likewise, the data in New York is analyzed as the variant in order to truly see how weather may correlate with the number of crime incidents throughout the year.

**Learning Applications:** This was a very interesting study since it does A/B test some variables in a way. Weather data is gathered through the National Oceanic and Atmospheric Administration’s REST API, and crime data from government agencies. The data requires quite a bit of cleaning and manipulation in order to create a merged dataset and further wrangling in order to fit the rigid requirements of FBProphet, the method for time series modeling. Data visualizations such as bar graphs and heat maps are used initially to establish correlation between weather and crime.

**Learning Objectives:** 1, 2, 3, 4, 5, and 6