John Casola

17 March 2019

Portfolio Milestone Paper

I enrolled in the Applied Data Science program because I believed it would allow me to change myself for the better by providing me with skills to find patterns in data, understand why they occur, what they mean and explain my results to others. It would allow me to build on my problem solving and analytical abilities. Most of my experience with data was with Excel, but I wanted to stand out to employers and I believed this was the best way to accomplish that.

In my first quarter I took MBC 638 Data Analysis and Decision Making, SCM 651 Business Analytics, IST 659 Data Administration Concepts and Database Management, and IST 687 Introduction to Data Science. I was familiar with basic statistics and Excel, but SQL and R were completely new to me.

MBC 638 Data Analysis and Decision Making taught me about using statistics to convert data into information that allows people to make better informed decisions. The course taught me about the DMAIC framework which is an acronym for defining, measuring, analyzing, improving, and controlling. Defining involves describing the problem, process, and data with descriptive statistics. Measuring includes measuring the process, evaluating the measurement system, and measurement error. Analyzing uses inferential statistics to find root causes and identify critical predictors. Improving is where the solution is implemented. The control phase is where a plan is created to ensure the problem stays fixed. Some tools introduced in MBC 638 included the chi-square test, Sigma Quality Level, and confidence intervals. The chi-square test is a discrete data hypothesis test of whether two variables are independent or dependent. Sigma quality level is a quantitative measure of the capability of a process. A confidence interval is a range of uncertainty around a point estimate such as a mean. The course included inferential statistics such as regression, and time series forecasting. The objective of regression is to fit a line that minimizes the difference between the actual and predicted values of the dependent variable. It also introduced the bell-shaped normal distribution and the binomial distribution.

My MBC 638 project was about improving the process of weight lifting. This project was different from future projects because the data had to come from a process I could directly influence. I recorded the number of repetitions failed in a workout plan, and the predictors I used were hours slept, calories each day, water each day, and rest between sets in seconds. The workout plan involved 2 workout types alternated every other day, with the goal of completing every repetition to increase the weight lifted. I picked my topic late in the course so I had a small sample size of only 3-4 rows for each workout. Scatter plots, XmR charts, and correlation matrices were used to investigate which predictors were likeliest to influence the number of reps failed the most. The results indicated seconds of rest between sets and hours slept the night before had the most influence on the number of repetitions failed. I implemented this improvement by sleeping at least 8 hours the night before and resting at least 3 minutes between sets. This led to mixed results where the failure rate and sigma quality level improved for one workout type, but not the other. A chi-square test on workout type and failure rate suggested there is a relationship between failure rate and the type of the workout, but the results had to be regarded cautiously due to the very small sample size.

SCM 651 Business Analytics provided an introduction to several different types of software and how to use data with them to analyze trends, create predictive models for forecasting and optimizing business processes to enhance performance. In Excel I learned how to create and interpret pivot tables and pivot charts, use PowerPivot, slice and filter data, identify the direction and magnitude of correlations from a correlation matrix, identify statistically significant variables in regression models, interpret their coefficients, interpret the R-squared, the difference between Adjusted and Multiple R-squared, and perform a two-way sensitivity analysis. The first homework assignment in this class had a group of students use these tools to analyze real estate price and predictors such as square footage and the number of bedrooms and bathrooms. Google analytics was used to analyze data to identify patterns and opportunities using metrics such as cost, cost per click, number of sessions or users, average session duration, and bounce rate. A second group homework assignment used Google Analytics to examine Whitman Graduate Programs internet marketing campaigns and distribute a budget based on our findings. The course used R to perform logit and probit analysis on data to identify the likelihood of a dependent variable based on a number of independent variables, and the relationships between the variables, possible moderating effects, and perform sensitivity analysis on the model. R was also used to test whether regression models followed the assumptions of regression with linearity, collinearity, heteroscedasticity, serial correlation, and outlier tests. Decision trees and choice models were also introduced using R. Neural networks were introduced and used to create a prediction model in Excel and perform a sensitivity analysis on it. Microsoft Access and Tableau were demonstrated within the class and overviewed.

IST 659 Data Admin Concepts and Database Management provided me with useful SQL skills and knowledge of databases. This course taught me about relational databases, where my only experience was with tabular data. I learned how to model relational databases in Microsoft Access, normalize the relations between tables, and implement and manipulate them in SQL. I learned how to use a database management system to store and retrieve data in a database, limit access to the data, backup and restore the database, and manage its performance. In SQL I learned how to create tables, correctly assign primary keys as unique identifiers for rows of data, join tables, aggregate them, use queries to manipulate and view specific data from a database, and create functions.

My IST 659 project was to create a client company database for Reed Exhibitions, a trade show management company. The report I wrote outlined the database and its importance for a nontechnical audience. A logical model was used to show how the data and its relations are organized in the form of tables. Business rules were used to constrain and represent the structure of the organization. An Entity-Relationship Diagram was used to represent the logical model, an entity being a real-world item or concept represented by a table. Tables in the database included the relevant addresses, the client company, the exhibition event held by Reed, the contract between Reed and the client, the booth purchased by the client company for a particular exhibition at a particular venue, and the venue which is the location of an exhibition. The business rules constrained the relationships of the tables, and the logical model. The physical database was created with SQL code to create the tables and the data within them, forms that could be used to query the database, and code for the database administrator to manipulate the data.

IST 687 Applied Data Science was my introduction to R. I had no prior experience with R and very little with coding in general. The course began with how to manipulate vectors, data frames and their rows and columns, creating functions, and descriptive statistics. I learned about sampling, replicating samples, and the central limit theorem which states that over the long run, the mean of the sampling distribution will match the mean of the underlying population. The course taught me how to import data into R from spreadsheets, SQL, and JSON formats. I learned to visualize data with histograms and line charts, and to create and plot maps. Linear modeling using linear regression models was overviewed in this course as well. Data mining, the use of advanced algorithms to find associations and patterns in data, was touched upon in this course. Two types of data mining were used in this class, association rules mining and support vector machines. Support vector machines use supervised learning to split the data, the larger part is used to train the model and the smaller part is used to test the model and evaluate it. There was a module on text mining, using word clouds and sentiment analysis. Text mining analyzed free form text which is unstructured data, as opposed to structured data in the form of tables with rows and columns. Text mining seeks to expose patterns in the data by understanding frequency of words and sentences in the text.

My IST 687 project was a group project with the goal of finding predictors for sales at Dick’s Sporting Goods. The finished data set combined demographic data from 2015 from Kaggle, with store data from Dick’s provided by a group member employed by Dick’s at the time. Visualizations were used to inspect variation in sales by state, square footage, population, income, store size, population, and location. Successful stores were used to identify potential locations for lucrative stores with similar characteristics. The northern East Coast showed average to better than average sales. A linear model and a random forest model were created, both of which had an adjusted R-squared and % variation explained of about 52.3%. The models were used to predict the top seven sites for successful stores locations.

The next quarter I took IST 618 Information Policy, MAR 653 Marketing Analytics, IST 565 Data Mining, and IST 718 Advanced Information Analytics (Big Data Analytics).

IST 618 Information Policy combined economics, ethics, and current events. It involved writing, research, and cost-benefit analysis of decisions and policies. The course examined current events such as net neutrality and the importance of access to information. Intellectual property rights and regulations were touched upon, including copyrights and patents. The trade-off between privacy and security was described in the course. Freedom of speech and its exceptions under the First Amendment were defined. Lobbying, which is seeking to influence politics surrounding an issue, was discussed as well.

In MAR 653 Marketing Analytics I took data and turned it into actionable insight, and applied it to marketing. K-means clustering was used in Excel. The objective of k-means analysis is to assign data points to k number of clusters/segments. It starts by taking k random points and assigning data points to k groups depending on which random point they are closer to. Then it takes the mean of each group and reassigns the data points into clusters based on which cluster mean they are assigned to. It repeats this until no data points are reassigned to a different cluster/segment. Customer Lifetime Value was introduced as a metric for optimizing your marketing to maximize your investment by helping you understand long-term and short-term return on marketing investments. Linear regression was reviewed in this course as well as propensity models such as logistic regression. Propensity models predict likelihood of binomial dependent variables being true or not. Multinomial logit regression and conjoint analysis are similar, but predict multinomial outcomes as opposed to binomial. Collaborative filtering was discussed, which uses an ordinal logit where the dependent variable is an ordinal variable such as a 5-star rating system on Amazon.

The group project I did for MAR 653 was to analyze Airbnb listings in Boston, Massachusetts to define listing segments and understand variables that influence price. K-means clustering was used to define the segments based on listing characteristics and linear regression was used to identify market values and compare them to the listed prices. Segments were defined based on different review scores from renters. This had limitations due to the possible correlations between the review scores. Four segments were identified from the K-means clustering. The regression model was built using predictors such as cleaning fee, maximum nights, number of reviews, number of beds and bathrooms, years of experience as a host and more to predict price. The regression generated an adjusted R-squared of 0.608, so about 60% of the variation in price was explained by the independent variables. Our recommendation was to align prices to the prices predicted by our model to increase occupancy rates.

In IST 565 Data Mining, data mining was defined by Dr. Platetsky-Shapiro, a data mining pioneer, as the “non-trivial extraction of implicit, previously unknown and potentially useful information from data”. Data mining tasks derive patterns and predict values. Some of the major tasks included in data mining include classification, clustering, anomaly detection, and regression modeling.

Data preparation skills in R provided by this course included identifying types of data sets and attributes, converting those types, summary statistics, visualization, aggregation, and transformation. Association rule mining is used to find correlation between items bought together or what products people would buy if they bought a certain product. Classification is a type of supervised learning where data is randomly split into labeled training and test data, and the classification algorithm is used to train a model on the training data and the trained model is tested on the testing data. K-means clustering was performed using R and Weka to identify clusters and detect outliers. Decision tree algorithms splits the data into subsets based on how well an independent variable separates the training examples according to the target dependent variable. Bayes’ theorem lets us swap the order of dependence between events, and lets us calculate the probability of B given A, if we know the probabilities of A,B, and the probability of A given B. Naïve Bayes classifiers use Bayes theorem for prediction, but assumes independence between events to reduce computation cost. K-Nearest Neighbor (kNN) is a classification algorithm, that uses instance-based learning where the calculation for training doesn’t happen until the model is applied to test data. KNN compares the similarity to the test data compared to the stored training data and finds the k-nearest training examples and chooses the most common category among them. Support vector machine (SVM) is an algorithm that can solve linearly separable and inseparable problems. It chooses the hyperplane that maximizes the margin between the groups that the user wishes to separate. Support vectors are the training examples that are located on the margins of the separating hyperplanes. Models can be evaluated with hold-out tests or cross-validation. A hold-out test is where the data is split into training and testing subsets where the training trains the model and the trained model is tested on the testing data. Cross-validation, where N number of folds is determined by the training data, repeats a hold-out test N number of times, with each fold taking a turn being the testing set and the final accuracy being an average of the N folds’ accuracy. The class went further into text mining, by normalizing, tokenizing, vectorizing, and removing stop words.

The group project I completed for IST 565 sought to see what variables (beds, baths, crime, etc.) affect real estate price the most in Pittsburgh, Pennsylvania and how well real estate prices can be predicted based on those variables. The data was obtained from the Allegheny Information Portal, Zillow’s API and the Bureau of Labor Statistics. The price was skewed, so it was normalized and 70% of the data was used for training the data and 30% was used for testing. 5-fold cross validation was repeated 10 times using random samples. Decision tree, random forest, Naïve Bayes, SVM, and kNN models were created, with each having between 55%-65% accuracy.

IST 718 Advanced Information Analytics was probably the toughest class, given my lack of experience with Python, the number of courses I took that quarter, and the volume of work in the class itself. I used Anaconda Navigator to use Python within Jupyter notebooks. Learning Python while being new to R was challenging, especially considering I had no classes that provided a foundation in Python. The first lab in this class provided more on regression, transforming and manipulating data, and put emphasis on finding required data sets on your own. The second lab focused on multivariate time series analysis for housing values in four metro areas in Arkansas. The data was provided, but had to be manipulated and transposed into the correct format for time series analysis. The time series were tested for stationarity and had to be transformed to improve stationarity. ARIMA and linear regression models were used to predict prices in the future for the metro areas. The third lab made use of convolutional neural networks to classify images in the Fashion MNIST data set.

The goal of the group project for IST 718 was to find what variables determine the success of a Kickstarter project. The data was obtained from Kaggle. Success was determined by whether a project met or exceeded the amount of goal money desired or not. Logistic regression was run on whether a project was successful or not. The data was randomly split into training and testing data, with 75% designated as training data. The independent variables were the main category of the project, the number of backers, the amount of money pledged, the goal amount of money for the project and the launch difference which is the difference between the project launch and deadline dates. The logistic regression model had an accuracy of 98.9% on the testing data. A random forest model was created with the main category, amount of money pledged, the number of backers, the goal amount, and the launch difference. This model had an accuracy of 99.4% on the test data. A linear regression model was created with the amount of dollars pledged as the dependent variable and the independent variables chosen were the goal, the number of backers, the launch difference and the main category. This model only had an accuracy of 56.1%.

IST 652 Scripting for Data Analysis was the next class I took and it was the only class I took that quarter. It focused on coding in Python and using scripts to automate processes. Most of the course was taught using the Anaconda prompt and Notepad which was different from using Jupyter notebooks in IST 718. The course covered basic lists, loops, Booleans, and conditional statements. Lists and conditional statements were useful as they were not elaborated on in my previous classes. In the class I wrote reports and tables to files, read in semi-structured data from HTML, XML, JSON, and MongoDB. Some text mining was done by collecting tweets from the Twitter API, and performing sentiment analysis on unstructured text.

In the next quarter I took IST 623 Introduction to Information Security and IST 719 Information Visualization. IST 623 Introduction to Information Security provided me with information on security policies, more on the trade-offs between security and privacy, and various security models. Cryptography was explained as the process of hiding information. I learned about secret key cryptography, where both parties share an encryption and decryption key. I also learned about public key cryptography, where each party has a public and private key. The public key is used by anyone to encrypt information to send, and the private key is kept secret by the user to decrypt information that was encrypted by other parties using the user’s public key. I learned about the different types of internet security protocols. In the labs, I received hands-on experience identifying and removing malware, encrypting and decrypting information, and analyzing wireless traffic.

In IST 719 Information Visualization I used R, mainly with the ggplot2 package, and Adobe Illustrator to create and modify visualizations. I learned how to scrape data from tables on the internet and load them into R, create different types of visualizations, use the layout of a visualization to direct readers’ attention, and the effectiveness of mapping different variables to different visual attributes such as color, size and shape. I learned that correlation is different from causation, which means a connection between variables does not necessarily mean one causes the other. I used Illustrator to improve and annotate visualizations created in R. This course taught me how to code my own interactive visualizations using Shiny apps, which I thought was very valuable and entertaining.

My last course was IST 777 Statistical Methods in Information Science and Technology. This course provided some review on R code, shapes of basic distributions, basic visualizations, and descriptive statistics. I learned more about probability, and contingency tables. This class compared and contrasted the frequentist and Bayesian approaches to statistics and, most importantly, taught me how to synthesize both into information that I can communicate to a nontechnical audience. I learned how to create and interpret a confidence interval, analysis of variance (ANOVA), Bayes Factors, and Bayesian methods for logistic and linear regression. The course also provided review on time series, how to test and improve stationarity and how to perform change point analysis on the mean and variance of a time series. The course briefly touched on ARIMA models, and provided me with the auto-ARIMA function that tests a range of parameters and chooses the best combination for you. For the final exam, I created a report on vaccination data.

There are seven program learning objectives for the Applied Data Science Program. The objectives are to describe a broad overview of the major practice areas in data science, to collect and organize data, to identify patterns in data via visualization, statistical analysis, and data mining, to develop alternative strategies based on the data, to develop a plan of action to implement the business decisions derived from the analyses, to demonstrate communication skills regarding data and its analysis for relevant professionals, and to synthesize the ethical dimensions of data science.

I received education in several major areas of data science throughout the program in all the courses I have taken. SCM 651 provided me with a wide-ranging introduction to different tools available for data science such as Excel, R, and Tableau. In IST 565 and IST 777 I learned and used a variety of data mining and statistical techniques. For my IST 565 project I performed exploratory data analysis with summary statistics and visualizations, preprocessed the data, used machine learning algorithms to predict price based on possible predictors, and evaluated the models.

My MBC 638 project taught me how to record data myself and the dangers of small sample size. For other projects, data was primarily gathered from government sources, including the US Census and local government sites, or public data platforms such as Kaggle. The data for the Dick’s sporting goods project in IST 687 was downloaded from Kaggle and joined with data from Dick’s that was accessed by a group member who happened to be an employee. The data for my IST 565 project was downloaded from the Bureau of Labor Statistics, Zillow, the Western Pennsylvania Regional Data Center, and data.gov. The real estate price data from Zillow was very skewed because there were a number of very high prices, to correct for this the data was scaled and centered by subtracting the median value from all values and then re-centering the results around 0. Price was transformed into ordinal values of “low”, “mid” and “high”. The data was randomly partitioned into training and testing sets, with 70% of the data for training and 30% for testing.

Pattern identification through visualization, statistical analysis, and data mining was strengthened throughout the program. I created a poster featuring a collection of visualizations about arrest data in Pittsburgh. Arrests were higher in certain neighborhoods, and arrests spike between 3:00 pm to 8:00 pm. Racial proportion of arrests were stable over the 3 years of data available. The largest proportion of crimes were committed by young men between the ages of 18 to 29. The proportion of women arrested decreased as age increased. In my IST 687 project, we mapped Dick’s stores and the population and income of the counties they were located in, and compared them to counties without Dick’s stores as potential new store locations. The final exam for IST 777 used confidence intervals and Bayesian methods to produce high density intervals to test for credible differences in vaccination rates between public and private schools. The k-nearest neighbor model to predict real estate price for my IST 565 project mixed together low and mid-priced homes, which was because of the skewed pricing data. This suggests that mid-range neighborhoods have mid-range houses with similar characteristics to low-priced homes.

Most alternative strategies for the work I’ve done involved getting more data and/or better predictors for the dependent variable. In my MBC 638 project I had to propose different strategies in order to limit the number of repetitions failed in a workout based on variables I could control, such as sleep and rest between sets in a workout. In the future, a larger sample size would be needed to produce better results. The predictors for some of my projects, such as in IST 565 and IST 718, produced models that were overfitted or not generalizable enough. In these cases, additional predictors would have been useful.

I developed plans of action based on the data and results of my analyses. For my MBC 638 project, I hoped to limit repetitions failed by increasing the amount of sleep and rest I had before and during each workout. Based on the information presented by my IST 719 poster, policymakers and law enforcement agencies could allocate resources to neighborhoods with high arrests, particularly between 3:00 pm to 8:00 pm. The IST 687 project I was a part of applied the linear regression and random forest models we created to predict potential successful sites for new Dick’s stores.

My various projects were presented to classes and their reports were created for the intended audience. My IST 659 project was to create a database for a tradeshow management company, and the report was created for a non-technical audience to support its usefulness, importance, the information problem the company currently has, and how the database I proposed would solve the problem. My final exam for IST 777 used vaccination data and was formatted as a report for a scientifically knowledgeable staff member in a state legislator’s office.

IST 618 educated me on some of the ethical issues related to data science. In IST 618 I wrote two ethics related essays and took part in a debate about the regulation of fake news by the government. The first essay was on the negative impact digital only voting would have on impoverished people, particularly the homeless. The other essay was about solutions to address privacy concerns over the increased freedom of law enforcement to collect personal information following the September 11th terrorist attacks on the United States. One solution was to include personal information among the list of things protected by antidiscrimination legislation. In the debate, I argued that false information is easily disseminated through social media which leads to inaccurate beliefs, undermines the democratic process, causes the public to be more skeptical of legitimate news outlets, and makes it difficult for people to distinguish between legitimate and illegitimate news producers. Although false statements are protected as free speech, as ruled by the Supreme Court in the case United States v Alvarez 567 (2012), the government could apply regulation, fact checking, and demonetization of fake news to address the spread of false information.

Together, the classes and coursework I completed as part of the Applied Data Science Program have provided me with a range of useful tools and skills that I believe will enable me to act on my desire to analyze data and use it to create actionable information. I can now collect data, use descriptive statistics and visualization to explore the data, identify patterns, use inferential statistics and data mining to predict future values, assess the accuracy of those predictions, and communicate my findings in a way that my intended audience will understand.