Jeremiah Graff

9/3/20

Portfolio Submission

**Level of Knowledge**

My level of knowledge in the world of data science is much greater than when I started this program 18 months ago, and I still have a lot to learn. What I lack in technical skills, currently, is made up for in my desire to deliver an end product. That being said, I will break down the components of OSEMiN (Obtain, Scrub, Explore, Model, Interpret) to talk about the data science process, while highlighting which component of the 3-pronged approach to data science/statistics (Descriptive, Predictive and Prescriptive statistics) falls in with the OSEMiN phase to round out the knowledge that I have acquired throughout the program.

The first two steps of the data process are the most time consuming. One professor mentioned that ~70% of the time spent on any data science project lies within these two steps of the process (Obtaining & Scrubbing). Obtaining the data can be easy or quite difficult. The easy scenario entails receiving data that does not need to be Scrubbed or munged and one can jump straight into the Exploratory analysis, this is rarely the case. Thus, the harder and more normal side means there is a need to pull in multiple sources of data to complete the dataset being used to build models and provide insights.

After data has been Obtained, then comes the Scrubbing. Scrubbing, or munging, is the process of preparing data for Exploration and Modeling. It is here where you will vectorize text for text mining, change metric variables to categorical, find and replace values to simplify how the output reads, change the data type of variables and many other tasks for the sake of providing clarity in the Explore and Interpret phases of the data science process.

The Explore phase is where things start to get interesting. This is where you start gathering Descriptive statistics and plotting out the different variables in the search for understanding the data. Are there outliers? What does the distribution of the data look like? Is there a lot or little variation within the variables? Do we see trends or seasonality in the time series analysis? Do we find any notable insights when aggregating the data? Is there any correlation amongst the variables? There is a much more exhaustive list of questions that can be asked in this phase as you familiarize yourself with the data. This Exploratory work leads directly into the work done in the Modeling phase.

Modeling is where we shift from the Descriptive nature of data science into the Predictive and inferential side. It’s here that we start to build out Models to understand how the variables interact with one another. We can see: what level of variance is accounted for in a regression model by looking at the R^2 values, how many clusters can be found in a dataset, see the decision tree that shows the progressive line of thinking that goes into predicting a certain outcome, or the 95% confidence interval that helps indicate whether we should accept or reject the null hypothesis of a scenario, amongst many other things. Modeling is only as good as the data that feeds it, which is why the Obtain and Scrub phases are so critical to the data science process. The aim of Modeling is to provide outputs for Interpretation that help decision makers do what is best for the organization in response to the results provided from the Models.

The Interpretation is the last step of the data science process, not the final step, but the last, where Prescription comes into play. I make this distinction because of the iterative nature of data science. Once an Interpretation is provided, often is the case where the data scientist goes back and runs new data, adds new data to the existing data or recalibrates the Models with the addition of new data to see if the results changed. Just because we achieved 97.2761% accuracy with one model, doesn’t mean we will always see the same results. Change is constant and thus our models should be checked and rerun over time to ensure that we are operating with the right presuppositions.

Back to Interpretation, the last step of the data science process. Interpreting has 2 sub-steps: 1) technical interpretation; 2) non-technical interpretation. The results of the models need to be understood from a technical point-of-view, which will be seen in the data scientist’s reporting and ultimate recommendation based on those models. The technical interpretation is where the data scientist will break down the results by speaking about: cost applied to models, the presence of multi-collinearity that prevented them from using certain variables in regression analysis, the number of epochs or hidden nodes used in a neural network, the number of estimators used for the random forest algorithm, etc. These are all places where the data scientist should break down their technical findings so that other data scientists can validate or critique their work accordingly.

After providing technical results, the data scientist must then provide the non-technical results. This is where it’s important for the data scientist to provide a clear and understandable picture of their findings to the decision maker or subject matter expert. The SME or decision maker will then use the data scientist’s findings to make real-time decisions regarding their business. The non-technical interpretation generally requires: the removal of statistic-oriented speech, the removal of talking about specific algorithms or algorithm hyperparameters, and communicating some of the Obtaining and Scrubbing phases in a way that communicates what was done (where the data was acquired from, what was removed or cleaned up to clarify, etc.), not how it was done. These results must be clear and easily understood for people to make decisions that impact their business.

The last topic to briefly discuss in regard to becoming a data scientist comes down to ethics. It is essential that we work to preserve the privacy of the humans and entities behind the data with integrity. That goes further than not just sharing who the data is about, but also entails using data that we have the right to use legally, as well as making sure that we spend the time understanding how we are feeding the algorithms and models we are building. Are the underlying assumptions biased and therefore perpetrating unjust hiring practices? Are some people groups not given an equal chance because of a poor sample collection method? These are just a start to some of the questions that need to be addressed when practicing as a data scientist. This is a large topic that can be discussed in far more detail. IST 618 (Information Policy) was a helpful class on this topic to help me see how important it is to work to preserve the rights and dignity of the people around us and the entire world, especially as technology gets more advanced.

That is the process that I see as essential for the data scientist to follow in this profession. The process will look somewhat different with the nuance of every project, but OSEMiN is a great framework to follow as one pursues the story that data is waiting to tell.

**Which projects provided that knowledge?**

There are four projects that stand out over the course of this program for me that helped in acquiring the skills listed above: the IST 687 (Intro to Data Science) final project, the IST 736 (Text Mining) final project, the final project poster for IST 719 (Information Visualization), and the first lab from IST 718 (Big Data Analytics).

IST 687 was a great introduction into the various components that data science consists of. I got to learn and use basic syntax, literally, with a language that was new to me, let alone a whole new process. Beyond learning a new language, I got to learn to start visualizing, start to run some basic supervised learning algorithms (Naïve Bayes, SVM and Linear Regression) and I also got to learn to interpret the results of these models. These were mostly new procedures and this project provided a great starting point for me to get my feet wet in data science.

The end goal of this project was open-ended: to craft the most compelling and actionable insight we could find in a 129,000 row csv file regarding airplane flight satisfaction, as a consultant. I chose to find what insight would have the biggest impact to a company’s bottom line. So, I spent a lot of time aggregating up front to see which of the 14 airlines had: 1) the largest amount of flights; 2) the most delays/cancellation; 3) a satisfaction score that was out of the norm in tandem with the other 2 variables. What I found in this process was the 3rd largest airline had a satisfaction score that was lower than the others and had a completed & on-time flight % that was 2+ pts less than the next lowest airline. That was my target that was going to *pay* for my insights as a consultant. Beyond this, I was able to dig in and start to run my first algorithms and visualizations in R, which served as a good introduction to the predictive modeling techniques that are everyday practices for data scientists.

IST 736 was my second Python-heavy course. My first course was hard after having gone from 2 straight terms in R to Python and I struggled with the major differences in syntax and logic. This class was different and the learnings that converged on the final project were pretty impactful. I felt much more comfortable in Python by the end of this class and my contribution to the team project felt strong. The code I used was not efficient, it is clunky and long (~10K rows), but it works. Efficiency will come with time and I was more confident in my ability to use Python coming out of this course.

The goal of this project was to use various models on text to see if we could make any predictions. Our project aimed at seeing if we could use the presidential nomination acceptance speeches of every presidential candidate since 1900 to predict various attributes. Did they win? Republican or Democrat? Good or Bad performance of DJIA? Incumbent candidate? Incumbent Party? There were other factors, but this is a taste of what we were looking to build predictive models for. I was responsible for building SVM models for each of the 15 variables. I ended up building 15 SVM models for each of the 15 variables. The reason for so many models was that I built 9 of them for the base rbf kernel: 3 of which were normal term document frequency matrices (with differing costs), 3 were Boolean term document matrices (with differing costs) and 3 were tf-idf term document matrices (with differing costs). The last 6 models were 3 polynomial kernels (with differing costs) and 3 sigmoid kernels (with differing costs).

One of the trickiest tasks in this project was developing appropriate lists for each of the variables. Some of the variables weren’t measured until the 1940’s or later so I had to find ways to edit my data set, later in the project, to only include text from the years that were measured when the variables needed for prediction were documented. Again, I am still gaining my roots on the coding side, but I am always determined to find a solution and was able to do so on this project, which grew my interest in learning more Python. The final output of my various SVM models were: 1) an average accuracy of 69% for my best 2 models (57% overall); and 2) a highest-performing model with 92% accuracy all based on the text of the available presidential nomination acceptance speeches dating back to 1900.

Next, was the poster from IST 719. The purpose of the poster was to educate Major League Baseball (MLB) fans on the distribution of Wins Above Replacement (WAR) within the different leagues, divisions, teams and individual players of MLB for the 2019 baseball season. There were 1,400 rows of input per each of the players in the MLB. The poster was to be put together in a way that followed the learnings of the course: understanding how visualizations can help communicate a story, how colors/sizes/contrast/shape/etc. can be used to guide the readers’ eye through a visual and how science plays a factor in how we should consider visualizing an insight.

This project was enjoyable to me as both a baseball fan and as a data scientist/analyst. It focused more on the technical side of crafting an insight in a visual manner that will help communicate with fewer words and create something that will stick in the mind of the person receiving the insight. I learned a lot about how to layout the poster (rule of thirds) and how to use color and sizes to guide the audience through the poster. I was able to take some of the learnings from this course and apply it to my current job in insights, mainly by making a concerted effort to include more thoughtfully crafted visualizations in my presentations.

Lastly, the first lab from IST 718 was fun. I enjoyed every bit of the 30+ hours that I put into the project. I enjoyed the coding, the research, the roadblocks, and then seeing it all come together. This project made it clear for me that I want to be a data scientist *now*. I kept trying to write the report for my submission but couldn’t step away from making tweaks to the code to make it more efficient and make the insight delivery clearer.

The task for this project was to answer a few different questions: How much should the next Syracuse Football Coach make in the ACC/AAC/Big 10? How effective were my models? What variable made the largest impact on the coach’s predicted salary? There were other questions, but these were the model-focused ones. The data file had 129 inputs at the beginning with some basic information and we had the opportunity to collect as many relevant attributes as we wanted as well as the ability to drop any inputs if needed. The Exploratory Data Analysis for this project was fun to dig into. I built bar plots that aggregated the salary info by conference, created regression plots that communicated the r^2 value of the different variables against salary and built a correlation plot that noted the strength of relationship for each of the numeric variables.

Up next came the fun stuff, the predictive modeling part. I ran 2 different Ordinary Least Squares (OLS) linear regression models. These models both came out with adjusted r^2 values around 78.5%, which is mostly good, but the bigger learning that happened in this project was that I was able to learn how to access the parameters/coefficients of the variables in the regression equation and apply that directly to the answers that were needed to answer the questions for the assignment. I got the code to work, but I also understood what I was doing with the code.

**Conclusion**

This program has been impactful in helping me: 1) discover a new passion for data that Excel just can’t satisfy as well; 2) prepare to transition into a role that can use data to solve problems more effectively. I have a new skillset that can be used to drive change. There is still a lot for me to learn in the data science realm, but the training and skills acquired in this program, my determination to find answers that I do not have, and my professional experience in providing actionable business insights give me a great deal of confidence as I look to transition into data science. I look forward to the endless opportunities to learn in such an exciting field.

**Program Learning Goals**

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

The Level of Knowledge section above answers this.

1. Collect and organize data

Each of the projects discussed above required the collection and organization of data. The IST 687 project and the IST 718 project data files were both supplied by the professors of the course. IST 718 required the gathering of additional data.

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

Each of the 4 projects discussed, as well as other projects and assignments throughout the program required the 3 elements above. Some required more than others, but these were general principles throughout.

1. Develop alternative strategies based on the data

The IST 736 project required different processing of the data depending on when the variables that were being analyzed started getting collected. Some of them were collected from the beginning (1900), while others didn’t start until the 1940’s or so. This required providing a different text corpus for some of the variables to be matched up against when conducting the text mining labeling and analysis.

Additionally, it is best practice to run multiple models on the data to include different predictor variables given that the different models will provide different levels of accuracy. Multiple models means different iterations of the same type of model (i.e. linear regression with different predictor variables) as well as different types of models (linear regression + naïve bayes + svm, etc.).

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

The IST 687 used the Exploratory Data Analysis to hone in on a specific business problem. This led to a “recommendation” to the 3rd largest company to work towards completing more flights to increase their satisfaction scores.

Additionally, the IST 718 Lab ended with a suggested salary, thus informing the Athletic Department on a range of what they should look to pay the next head football coach. This salary suggestion was applied to various conferences and called upon the ability to develop alternative strategies (different salary suggestions depending on the different pay ranges within the alternative conferences).

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

The section in the Level of Knowledge that discussed the Interpret phase of OSEMiN addresses this. It’s important for the data scientist to know who they are talking to. Is it an individual who will understand the technical components or is it someone who needs the input in a non-technical way? The data scientist should adjust the vocabulary accordingly in those 2 scenarios.

1. Synthesize the ethical dimensions of data science practice

The second to last paragraph of the Level of Knowledge section discussed, briefly, the importance of understanding the ethical dimension of data science. There have been some interesting videos discussed throughout the program that have awakened the need to be aware of bias and potentially unethical scenarios that can occur when analyzing and modeling data. Truly, this is an important topic in the data science world that warrants much time and effort to be aware of when mining data.