Portfolio Milestone

Winston Vu

Syracuse University, iSchool

M.S Applied Data Science

**Introduction**

The iSchool provides a Masters of Science in Applied Data Science at Syracuse University. Within the Masters program, the classes provide students with information about the data science world. These classes taught students how to use relevant programs in Data Science, such as SQL and Python, mathematical insight on the properties of code, how to analyze a problem based on data, and how to collect data and its application to the analysis. Classes, such as Data Administration Concepts & Database Management (IST 659), Scripting for Data Analysis (IST 652), Text Mining (IST 736), and Big Data Analytics (IST 718) helped receive a perspective on how a data specialist would approach data and apply the information to business functions.

The Applied Data Science program enables users to learn and provide results of 7 learning objectives: developing an overview of major practices in data science, collect and organize data, identify patterns visually, statistically, and mining, developing alternative strategies, developing a plan of action, demonstrating communication, and learn ethical decisions of data science practice. These 7 objectives were exemplified throughout the course of the program. It was not an easy start because of the lack of programming experience, but throughout the program, the skills and knowledge increased as well as the confidence.

**IST 659: Database Administration & Database Management**

The first course ever taken at Syracuse University was Database Administration & Database Management. This class was taught by Chad Harper, and he taught about how to create conceptual models and logical models before developing in a SQL database. From there, the professor taught about how to develop code to create a dataset from the information collected and how to perform an analysis with the information imputed into the dataset.

Diagram

Description automatically generated

Text

Description automatically generated with medium confidence

Text

Description automatically generated

The project revolved around the top 10 draft picks in the NFL from the year 2010 to 2020. This project was to see whether or not there were any patterns involved with the top 10 draft picks and to apply it to potential business cases, such as college scouts or high school recruitment. The first step was to create a conceptual model and a logical model. The conceptual model created a baseline of the overview of thoughts and information that should be stored in a database, and the logical model expands the overview of the conceptual model. After setting up the models, the tables were created for each table column and the variables were added with specific identifiers, such as type of variable, constraints for keys, and whether to identify a variable as a key or item for analysis. After the tables were collected, the variables were added into each table, which were collected from the ESPN website. This information included name of player, height, weight, draft pick, draft year, conference, college, and other information.

‘What conference has the most players drafted in the top 10?’ Code and Table

Graphical user interface, application

Description automatically generated

Table

Description automatically generated

The analysis portion allowed the students to create their own questions in order to think like business people. This helps students integrate thoughts that would be applicable for business and how to solve them as a data specialist. The questions were either basic, such as the conference with the most draft picks, or more complex, such as which team chooses in the top 10 the most. The development of the questions were essential in thinking in the minds of business people and helps enable thought processes and steps in order to reach the answer.

Overall, the class provides insights into how SQL and its function for database administration and management is essential in the business world. In the realm of data science, SQL is one of the most important tools used in order to retrieve data and store data. Having the experience to use SQL and learn how collect, store, and analyze or find patterns in data helps with the future use and knowledge of this course.

**IST 652: Scripting for Data Analysis**

In Scripting for Data Analysis (IST 652), the class allowed students to get introduced to Python. This class was taught by Professor Debbie Landowski. Python is a program that is mainly used in many companies, and learning the program would provide great benefits. The class taught students the basics of using Python and how to write scripts. Scripts are essential in the data science world because they can be used as functions for later use. They can be definitions, loops, or simple functions that can be essential for the data science world.

For the project, a NCAA basketball dataset was used in order to see how the evolution of basketball has progressed throughout the decade, and to see what factors are best to look at when looking at a team’s winning percentage. The project consisted of preprocessing methods, such as changing column names and deleting and adding columns, performing linear regression models, and performing basic analysis.

A close-up of a document

Description automatically generated with low confidence Chart, treemap chart

Description automatically generated

The project also allowed students to create their own problems to solve while using Python. Figuring out questions can help with an exploratory data analysis, such as finding correlations of variables, and the creation of models. There were aspects used in this project, such as Variance Inflation Factor (VIF), linear regression, and other basic Python functions. Though the use was basic, it provided a baseline for the future use of Python, and how skills would be developed over time with work and future courses.

The class excelled in providing an introduction to Python. The program has been very popular among many companies, and it has a lot of capabilities to use. Learning how to create scripts can be simple or very complex because it depends on the model or function the user is trying to create. Also, running a project code through Python is a different mechanism compared to R. There are many different styles of Python that can be used, and it was interesting to have an attempt at each.

**IST 738: Text Mining**

In Text Mining students are able to learn how to use machine learning tactics and make predictions with text data. The class was taught by Norma Palomino-Grubb. Text data is a type of data that companies can use to help integrate new aspects of features, such as voice recognition or a classifier. The class allowed students to understand how different machine learning techniques can Affect text data as well as different preprocessing methods in order to prepare the data. These techniques were used and applied throughout the course and the final project.

Cleaned Dataframe of Lyric Data

Table

Description automatically generated

Unigram RandomForest: Unigram and Bigram RandomForest:

Table

Description automatically generated Table

Description automatically generated

For the final project, the team decided to create a lyric classifier that classifies a song’s lyrics to a genre. The data was collected manually from websites containing different genres for each song. For the preprocessing method, techniques, such as vectorization, tokenization, and train and test split, were used to prepare the data for the model. After completing the preprocessing section, the models, such as Multinomial Naïve Bayes and Random Forest, were used to create the models. The Random Forest model had the highest accuracy at 52.26% out of all the models after using a cross validation. The accuracy was not high, but according to the professor, it is close to real life experience. It is difficult to obtain a high accuracy in real life, and the project simulates real life in collecting the data, applying data science techniques to obtain the results, and drawing conclusions.

The class taught students a lot about the use of text data. Previously taking the Natural Language Processing class before Text Mining, it helped open up the basics of text data, while Text Mining helped with more of the complex applications. The course did teach a lot about how to approach text data in business world applications. There are components where text can help create new ideas for companies, and the processes of creating the machine learning model before its actual deployment.

**IST 718: Big Data Analytics**

In Big Data Analytics (IST 718), students are able to learn techniques and machine learning tactics in order to deal with big data. The class was taught by Professor John Fox. The class gave students the ability to approach the data in different ways. There were many ways to approach data, such as using forecast techniques to work with a time-series problem or using neural networks to solve a classification problem. Though the class had a variety of techniques to solve problems, each tactic settles to solve certain problems, and students were enabled to figure out which techniques are needed to solve a certain problem.

Chart, histogram

Description automatically generatedA picture containing timeline

Description automatically generated

Table

Description automatically generated

The project for this class focuses on temperature changes and insurance data. Due to the increasing temperatures, theories of global warming, and inflation, insurance prices could be altered due to these ideas. The project mainly focuses on the state of Texas due to the immense heat the state endures. The cities involved were Austin, Dallas, Houston, Amarillo, Brownsville, Corpus Christi, Laredo, and El Paso. The project uses the scikit-learn forecast packages in order to train the forecast, test the predictions, and apply the forecast for the next 5 years. The classifier that was used to train and test the forecast model was the Random Forest Classifier. This method was applicable to only the temperatures for each of the cities. The forecast package, Prophet, was used on the insurance data that was collected to see how the premiums increased and to create predictions of what the premiums might be during the next 5 years.

This project allows students to create a business problem, collect data, apply techniques learned from class, while allowing for alternative strategies, looking for patterns and creating visuals, and communicating with other group members. Instead of downloading a dataset from Kaggle, the team decided to obtain data from online sources by looking through government climate websites and documents pertaining to the data. The strategies learned through class were the Prophet model and ARIMA model. The data was multivariate since it contains precipitation and snow data, so the team had to find different types of forecasting models. The scikit-learn forecast was discovered and was able to work with multivariate data. Then, the team observed the visualizations to see if there were any trends or patterns involved with the increase in premiums. The project enabled students to work with other people in order to create a successful business plan in order to solve a problem that could be observational in the real world.

**Conclusion**

The courses provided knowledge about the learning objectives provided by the program. In each of the courses, the major practice areas of data science were applied. Each course provided a different insight into different programming styles and languages, such as managing databases in SQL, applying machine learning models in R and Python, or natural language processing. Each class also provided an insight into the collection of data. The data is often collected and researched through websites manually or there are datasets created through websites, such as Kaggle. Also, all the classes had alternative strategies when approaching data. There are many ways to solve a problem or there are additional strategies that should be added to improve the solution. For example, in IST 659, models were not learned in the class, and simple scripts were only learned. Preparing for future classes and possible job work, learning new algorithms while learning from the class helped grow the knowledge and drive necessary to become a successful data scientist. Lastly, these classes provided insight on how business decisions could be made with the projects. An objective that is required for each project is to determine how the data and solutions would affect a business, and it brings out more creative insights into how it could impact businesses.

Each class had a different way to implement visualization, analytics, and data mining. Visualizations can be created differently for each class. For IST 659, a logical model and conceptual model were created before creating the data tables, and for the analytical classes, graphs were created to easily visualize each of the variables, singular or comparison, as well as visualizing the machine learning models by accuracies, tables, or confusion matrices. Each class implemented analytics in some way or form. It was either basic, such as basic mathematics, or more complex, such as using machine learning models to create predictions. Data mining was used in each class in a way to solve problems. Each class had a different way of collecting, exploring the data, modeling, and creating conclusions. The different sets of insights helped provide different sets of creative mindsets in order to solve the problems.

With communication, there are different ways to communicate, with the team and the presentation. Meeting with the team is an essential part of communication because it helps the team design a strategy, implement the strategies, and drawing conclusions from the results. When presenting, the communication has to be clear and simple to understand. Having confusing terminology is not effective to pass through an entire audience, so having communication be clear and concise is best practice. With the ethics, each class provided an insight on how to secure the data and how to use it for good practice. To avoid any problems with data, the data would need to be kept private and only relevant sources of data should be used. There should not be any bias involved when selecting the data to create the analysis, and it would have major consequences if one would discover to be unethical.

Overall, the Applied Data Science program at Syracuse University helps students learn the methods in data science. The process of data science is a long and tedious one, but after learning various methods, the process becomes easier. The program has also provided different techniques to solve problems, whether it is basic analytics or using machine learning tactics in order to create predictions and forecasts towards the problem. After solving these problems, the solutions have to be brought out within reports and presentations in order to keep clients happy and the company happy. The school has provided the knowledge for my new job as a Junior Data Scientist, and the theory and creativity have provided new insights for my company.