APPLIED DATA SCIENCE – PORTFOLIO

Name: Sanman Yadav

SUID: 218223603

Email: [smyadav@syr.edu](mailto:smyadav@syr.edu)

INDEX

|  |  |  |
| --- | --- | --- |
| Sr. No. | Contains | Page No. |
| 1. | Introduction | 3 |
| 2. | Learning Outcomes of ADS Program | 4 |
| 3. | Coursework and Projects | 8 |
|  | 1. IST 687: Introduction to Data Science | 8 |
|  | 1. IST 718: Big Data Analytics | 9 |
|  | 1. IST 707: Data Analytics | 9 |
|  | 1. IST 659: Data Administration Concepts and Database Management | 10 |
|  | 1. IST 719: Information Visualization | 11 |
| 4. | Appendix | 12 |

1. Introduction

Data science, in its most basic terms, can be defined as obtaining insights and information, really anything of value, out of data. Data Science is an evolving field and it is evolving at a rapid speed. Data science is a cumulation of different strategies such as capturing data, processing data, maintaining data, analyzing, visualizing and communicating. As a data scientist in data science filed, the data scientist must ware different hats and be comfortable trying different things around. In future it is expected to have most jobs and the need for skilled data scientist is and will keep growing.

The ADS program at iSchool is one of the unique ones. It not only focuses on the coding and technology part of data science, but it also has the management aspect to it which makes it all rounded. Courses such as Business Analytics (SCM 651), Data Analysis and Decision Making (MBC 638), Marketing Analysis (MAR 653) really helped me look at the filed in a different light. I was able to from business questions and make right recommendations based on my findings. Forming and answering business questions helped me in my masters and helped me land an internship.

During my masters, because of ADS program at iSchool I got exposure to different technologies which are used by the companies to make business decisions. The different technologies that I got to use during my master’s are python, R, pyspark, MongoDB, MS SQL, MS Access, Excel, RShiny, Docker, Hadoop, Weka, SAS, Hadoop, Power BI and Tableau. The exposure to these current technologies has helped me expand my knowledge base and has made me an all-rounded Data Scientist. This course provided me with this opportunity to gain knowledge and skillsets in various practice areas of data science.

The best part of this program is the way it is designed. It touches base on all the different topics in data science and gives practical and the subjects in program give relevant knowledge. By giving me exposure to different technologies and techniques, the program has helped me think like a professional data scientist.

1. Learning Outcomes of ADS Program

The ADS program is very flexible and allowed me to choose my course from different colleges within the Syracuse University. Hence, I was able to pick courses from College of Engineering and Computer Science and Whitman School of Management. Due to this I was able to pick courses in the field of my interest and master the skills on offer through doing different projects. Here I am going to highlight how the courses and projects have helped me be in line with the learning outcomes of ADS program.

1. Broad overview of Major Practice Areas:

The first taste of Data Science I got was from the course Introduction to Data Science (IST 687). The course was designed perfectly in which I got to learn about different practices a data scientist must follow to get the desired results. The project presentation that I did with my group was top notch and helped me understand the life cycle of a data science project. The courses such as Scripting for Data Analysis (IST 652), Data Administration Concepts and Database Management (IST 659) and Data Warehousing (IST 722) helped me understand how to gather the data. Gathering and taking the relevant data is the most important step in any data science project. These courses also helped me learn the techniques for handling and processing of large amounts of data. The courses such as Big Data Analytics (IST 718) and Data Analytics (IST 707) were the courses that talked most about analyzing and building the models for prediction by using machine learning and analysis. These courses also taught me how to evaluate a model and built better ones. This were the courses in which I got to learn the most about analyzing the data and getting relevant results. The courses such as Business Analytics (SCM 651), Data Analysis and Decision Making (MBC 638), Marketing Analysis (MAR 653) helped me look at a problem from a business point of view. This made me look at the problem from the end user eyes. This helped me look at the results that will help me answer the business questions and not just give out the results. I leaned to give proofs for the results and recommendations that were given through the analysis which is the last step of a project and is an important one which people tend to forget. The course Data Visualization (IST 719) helped me improve the way I present my data and helped me make better visuals.

1. Collect and Organize data:

Collecting and organizing the data is the most important step in a data science project. If this step is not done properly it will affect the results even if the most tuned models are used for prediction or analysis. The courses in the ADS program did help me understand this step and taught me different ways in which one can collect and organize the data. The course Data Administration Concepts and Database Management (IST 659) was very help in this regard. In this I had to build a database from scratch, populate it and maintain it for other people to use. In this project I built a Gym Management database which was user oriented and helped them reach their health goals. The course Scripting for Data Analysis (IST 652) was one of the courses that I enjoyed the most. In this we had to scrap the data from the internet to look for patterns and make the recommendations based on those. In this we got an exposure to 3 different kinds of data: structured, semi-structured and unstructured. In this I got a chance to use python and many different libraries it had to offer for data collection. This course also gave me an exposure to No-SQL, which is used widely by the organizations around the world. In the course, Data Warehousing (IST 722) I got exposure to the ways an organization handles huge amount of data. In this I had to use data marts to collect and retrieve the data. I had a hands-on learning for consistency processes that one uses to keep the data consistent in the data warehouse.

1. Identify patterns via Visualization, Statistical Analysis and Data Mining:

In a data science life-cycle, data mining, statistical analysis and visualization from the crux of it. After successfully gathering the data the next important part is EDA (Exploratory Data Analysis). The course Big Data Analytics (IST 718) has all the above-mentioned parts in it. Form doing EDA on the data to visualizations to making business recommendations. According to me, this is the most rigorous and well-thought-out course in the program. In this I got to try my hands on pyspark, Docker, Hadoop and Databricks. For the project of this subject we took a very big dataset and had to run different analysis on it. This was my first experience handling huge amount of data. In this we course, we had to built models for prediction using machine learning. For the project, we had to present using a poster which the professor loved. This course also had statistical analysis element which helped us determine which way to procced in our analysis to get to the desired results. The statistical analysis mostly falls under EDA. The focus of this course on building different models was the most challenging part but this was also the part in which I got to learn the most. I felt the course Data Analytics (IST 707) was continuation of what I learned in Big Data Analytics (IST 718). This really help me to solidify all the concepts that I had learnt before. In the project we could use different technologies such Weka, R and Python to do our analysis which came in handy. Because of this I was able use different technologies in unison to report my analysis and I had to figure out how to use cross-platforms. For the project of this subject, we again decided to go for a large dataset as I had dwindled in big datasets before. The project for this subject also followed similar life-cycle to that of Big Data Analytics (IST 718) ranging from basic EDA to visualizations and recommendations. In the course Data Visualization (IST 719) we dived deeper into making cool and interactive visuals. The whole course revolved around making visuals that are presentation friendly and making them such that gets the whole crowd involved. In the project I had to do a poster which told a story of the dataset. I decided to show the history of Olympics and how it has changed form the olden times. The story that my poster told was a unique and was well received by the professor and the crowd. The labs and the quizzes of IST 719 were designed to test one’s knowledge and were very hands-on. This course also focused on critical thinking by asking ambiguous questions about the dataset. All the above mentioned courses helped me build a strong base on which I have continued to build.

1. Develop Alternative Strategies:

For developing alternate strategies, the courses that helped me do this were, Data Analysis and Decision Making (MBC 638) and Marketing Analysis (MAR 653). I think Data Analysis and Decision Making (MBC 638) was one of the subjects which made me grind and really made me work for my grade. The course revolved around essential concepts of data analysis, probability, and statistics necessary to understand and perform analysis of business and economic data. The course focused on topics such as descriptive measures, association and correlation, probability, normal distribution, sampling, hypothesis testing and time series and forecasting. The project we did for this topic was one of the most unique projects that I ever did. For the project we decide to see what external factors affect the stock prices of FANG. But for this there was no dataset out there was meeting our needs. So, we decided to make our own dataset to answer the business questions. To make sure we had the right variables for the analysis we used correlation between the variables. And the project was a huge success and the professor was impressed with the different factors that we had considered for the analysis. During the course of this project I had to meet the team on regular bases, since the project was unique, and it was necessary that everyone was on the same page. This was important because when it comes to developing strategies it is necessary to get everyone’s input as there is a chance in strategizing that the team might miss something important and if everyone is on the same page, at least one person can catch the miss. Marketing Analysis (MAR 653) is a course in which I got think about the marketing aspect of analysis. The course sometimes caught me off guard. This was a different experience for me as I got to interact with different people form different background. I got to use SAS form making the analysis. This focus primarily focused on PCA which helped me get a strong grasp on the concept. I now use PCA in basically in all of my projects.

1. Develop A Plan of Action to Implement the Business Decisions:

When a data scientist gives out recommendations based on his analysis, it is also important that he develops a plan of action to implement the business decisions. It is a part of data scientist’s job to give suggestions for how to address the business problem on the analysis. In the Business Analytics (SCM 651) the core focus was on developing the plan. In this course I got to use SAS, Tableau and Power BI. There tools are used by business analysts and data scientists to show the recommendations to the clients. The versatility of the tools makes them a perfect tool to develop plans and strategies for the business. As there are easy to use and understand when making plans, it makes the process of planning less tedious. In all the projects that I have done in my masters almost all them had a recommendation and planning aspect to it. This has helped me making my recommendations / suggestions in a concise way because the crowd that I will present to in the future will mostly be the higher management. It is the duty of the higher management to look at the feasibility of the plan / suggestions and implement them. The insights generation is the most critical part of any data science project and I can confidently say that this ADS program has definitely given me the exposure.

1. Communication Skills:

When it comes to communicating to higher management I would love to talk about my experience as a Data Science Intern with Gartner Inc. this summer. I learned about the importance of documentations, meetings, minutes of meetings and to translate expectations and to comprehensively understand negotiations between a client and employee. Understanding these expectations can determine where and to what degree the effort of an individual is concentrated and how it would benefit the objective of the project or the satisfaction levels of the client. I learned about the importance of clear and concise communication since the value of free time and time in general was made more evident to me after beginning 40-hour work weeks. Utilizing company resources to learn and get involved in the community beyond what is prescribed as a part of the internship experience is something, I learned that rewarded me intrinsically. Some other take-a- ways for me from the internship were, how to network with people within my organization, how to extend myself beyond my team and be able to understand holistically what kind of work is being pursued by the enterprise on a larger scale. This kind of information gave me the insight and incentive to go and look for a similar opportunity within another team once my current commitment ended, this gave me the chance to meet with and learn about what the Natural Language Processing team has been working on and how they move about their days, what kind of infrastructure they use and how they measure success. It was only because of the projects and the collaboration that I did in my masters that helped me communicate so effortlessly during my internship.

1. Ethical Dimensions of Data Science Practice:

During my internship I had access to private information of our clients. The ethical practices our organization observes has some elements in common with the practices followed by the organizations worldwide. In the organization, we were educated about security, data privacy, protocol and safe measures. Client information is very sensitive, and we have some policies and protocols we follow to ensure there is no identifiable data stored or used unnecessarily. We must clear outputs of code and make sure any presentations we do have no electronic records of any sort are displayed. We communicate on secure servers and send emails which are encrypted to protect client information. Gartner Inc. has a committee to oversee any breaches and malpractices and unethical practices being conducted by employees or managers. There is an anonymized way to report spam and phishing emails to our IT Department and there are ways to report instances where someone broke protocol and jeopardized the company’s brand value by misusing data. All these processes helped me understand the importance of privacy. If it was not for the program and my internship, I wouldn’t have been to understand the gravity of the topic: Privacy.

1. Coursework and Projects

In this section I am going to primarily focus on some of the subjects that I took and the projects that I did in them. I will also mention how have these things helped me achieve the learning goals of the ADS program.

1. IST 687: Introduction to Data Science

This is a core requirement for the ADS program. This is the first course that I had taken in my masters and helped me understand what data science is. It was an introduction to fundamentals about data and the standards, technologies and methods for organizing, managing, curating, preserving and using data. Discusses broader issues relating to data management, quality control and publication of data. The main technology used was R. In this we also got to use different types of data files for the analysis. Used ggplot for visualizations. In this I got to learn different methods such as SVM, Apriori, Association Rule Mining, Linear Regression, Logistic Regression and ways to do Text Mining. This course got me riled up about Data Science and I loved the degree of professionalism the professor portrayed. The principles and practices such as processing, linking, aggregating, summarizing, visualization and searching came in handy for me during my internship. This course also helped me interpret basic statistical measures.

**Hyatt Hotels Project->**

For this project on Hyatt Hotels, I cleaned and organized the data along with loading 1 million entries of data from a CSV format into R studio then proceeding to clean it and run descriptive statistics on the data and some visualizations to understand the data after which i filtered it and then worked on the part of the project involving applying supervised machine learning algorithms to test and train the algorithm to predict the likelihood to recommend of a customer, and other factors that go into predicting this metric, I also performed association rule mining to find correlation and impact of different variables upon a set of dependent variables we determined as a team. Proceeding to provide actionable insight and recommendations from a business perspective to improve the net promoter score. We also discovered other factors that go into predicting this metric, and impact of different variables upon a set of dependent variables we determined as a team. Proceeding to provide actionable insight and recommendations from a business perspective to improve the net promoter score. What makes me Proud of this project is the fact that i had no knowledge of R at the beginning of the project and accomplished all of the above mentioned tasks on R in the span of a few months, I grasped the key concepts and was very proud of executing a major part of this big data project and presented our findings to the client individually in an easy to understand and non-technical manner. This was the first ever project I did with a data scientist mindset ant the results were amazing.

GitHub:

1. IST 718: Big Data Analytics

In my masters one of the courses that really challenged me was Big Data Analytics (IST 718). This was the first time I got to play with big data. There were many aspects of this course that made it challenging. The way this course was designed, made sure that during al the labs and the project as well, the data science project pipeline was followed. This course had a broad introduction to analytical processing tools and techniques of data science. In this course I got a hands-on training on python and apache spark to build big data analytics pipeline. We got to study classic and state of the art machine learning techniques. In this course we also got to analyze big data, create statistical models and identify insights that can lead to actionable results. We also got an introduction to linear algebra, calculus and statistics. This course really set a path for me to use big data for analysis using Hadoop, MapReduce and Apache Spark.

**NYC Yellow Cab Taxis->**

Leverage the NYC Yellow cab data from Kaggle by extracting features to predict taxi trip duration and fare by building a predictive model. Tune model parameters to achieve maximum accuracy without overfitting data to assist Yellow Cab drivers by providing a pre-trip prediction of approximate duration and a price estimate. In this project since our target variable was a continuous variable, we used Linear Regression, Random Forest and XGBoost for prediction. The least RMSE that we got from was XGBoost which was about 1.87 $. In the later part of our prediction we used predicted time of the ride to predict the price of the ride. In this project different tools that we used were PySpark, XGBoost, Gradient Boosting, SparkML, PCA, Decision Tree, Regression.

GitHub:

1. IST 707: Data Analytics

This course was in tandem with the Big Data Analytics (IST 718) and this really helped me to solidify my knowledge of what I had learnt. This course gave a general overview of data mining techniques, familiarity with real-world applications, challenges involved in applications, and future directions of the field. The topics of the course will include the key tasks of data mining, including data preparation, concept description, association rules, classification, clustering, evaluation and analysis. Through the exploration of the concepts and techniques of data mining and practical exercises, helped me develop skills that can be applied to business, science or other organizational problems. This course also had a storytelling aspect to it which made me dive into different datasets to find useful patterns and articulate what the patterns have been found, how they are found and why are they valuable and trustworthy.

**Airbnb Data Analysis->**

The main aim of this project is to dig deep and analyze how we can benefit the consumers, host and even Airbnb to provide a better service. In this project we took data of Airbnb Listings in the

US and try to predict the price of stay in that listing. The data was of 6 cities across the US: NYC, LA, San Francisco, Washington DC, Boston and Chicago. During the EDA step we made some distribution plots to understand the data and look for the outliers. During the initial EDA is when we realized that the dataset had huge amount of NA’s. So, we took care of NA’s the way we were taught in the course and that really helped our prediction. The challenge here for us was to understand the broader purpose of the data. Then use our expertise to analyses the datasets, and to piece together the insights for consumption. Which type of visual analytics to use and selecting the best ways to crunch enormous volumes of data, select and present the data for meaningful interpretation. With data sources multiplying and complexity rising, the most common challenge was getting the relevant data. The challenge is mining the seemingly endless data sets, sifting and sorting it to get data that is valuable and useful. The Airbnb data that we had selected was diverse and had many possibilities.

GitHub:

1. IST 659: Data Administration Concepts and Database Management:

This was a course which gave me an introduction into database management systems. This course gave me knowledge about data structures, file organizations, concepts and principles of database management systems (DBMS); as well as, data analysis, database design, data modeling, database management, database implementation, hierarchical, network and relational data models; entity-relationship modeling; basics of Structured Query Language (SQL); data normalization; and database design. In this course I used Microsoft’s Access and SQL Server DBMS as implementation tools. This course also had introduction to advanced database concepts such as transaction management and concurrency control, distributed databases, multi-tier client/server architectures, web-based database applications, data warehousing, and NoSQL. One thing that I got hands-on during this course was to solve problems by constructing database queries using Structured Query Language (SQL). This course also helped me get insights into future data management tool and technique trends.

**Fitness Database ->**

As a project for this course I decided to explore the Health Industry. Hence, I decided to focus on building a database for a Gym. A Gym database has a lot of potential of helping a person to get healthy and fit. A normal gym with a traditional database can only help a person by giving him an area to workout. But my database did much more than that. It had the ability to capture the user’s everyday data. With this little modification to the old gym database, it was able possible for the individual to track his progress and to take the steps accordingly to improve. This database stored the everyday information of the workouts and nutrition one person did. In everyday workouts it will store the body part that is being exercised that day. Plus, on this one can store the kinds of exercise for that body part one is doing with information of sets and repetitions of that exercise. This database also stored the information about the food one eats. In everyday nutrition it stored which is the meal (Breakfast, Dinner etc.), the name of the dish and the calories in the food. This data will be stored on the everyday bases. This database also stores the information about the daily things like Blood Pressure, Heart Rate, Steps walked, Total Steps walked, total hours slept. This information will be stored on everyday bases. This is how I created and maintained a fitness database.

GitHub:

1. IST 719: Information Visualization:

This course was primarily taught on R. The main focus of the subject was to teach the students main types of visuals and how to make them eye-catching. I also taught us how to use Adobe Illustrator to make posters. The course taught me how to make stories about the data using the visuals and make the crowed understand what the data is talking about without actually looking at the data. Professor made this subject really fun and interactive. The labs, assignment and quizzes were challenging and the collaboration of all the people during the lectures made the lectures fun.

**Olympics: A Peep in the History ->**

The data that I used for the project was historical data of Olympic Games from 1896 to 2016. The data contained both the Winter and Summer Olympics. Olympics have a lot of history. It has survived to World Wars. Sports always brings the people together and this is what the Olympics does. This is what motivated me to investigate Olympics. The sports enthusiasts found it interesting to see the history of Olympics and all the changes that it has gone through. Plus, the section about Art Competitions will increase the curiosity. The findings form this project were interesting. It explored the data in which the Arts was a part of the Olympics but was later taken out because of some rule changes. In the end I used Abode Illustrator to create the final project poster which had all the graphs and insights.

GitHub:

1. Appendix

This section provides link to all the project from the courses mentioned in the above sections.

|  |  |
| --- | --- |
| Courses | Project Links |
| IST 687 |  |
| IST 718 |  |
| IST 707 |  |
| IST 659 |  |
| IST 719 |  |

GitHub link for all the projects:

----------------------\*----------------------\*----------------------\*----------------------\*----------------------