APPLIED DATA SCIENCE – PORTFOLIO

**Name**: Vishwanath Hegde

**SUID**: 764516311

**Email**: [vhegde@syr.edu](mailto:vhegde@syr.edu)

Index

[1. Introduction 2](#_Toc26304308)

[2. Learning Outcomes of ADS Program 2](#_Toc26304309)

[3. Coursework and Projects 2](#_Toc26304310)

[a. IST 687: Introduction to Data Science 2](#_Toc26304311)

[b. IST 707: Data Analytics 2](#_Toc26304312)

[c. IST 719: Information Visualization 2](#_Toc26304313)

[d. IST 659: Data Administration Concepts and Database Management 2](#_Toc26304314)

[4. Appendix 2](#_Toc26304315)

# Introduction

Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data. It is an interdisciplinary field consisting of subjects like statistics, computer science, visualization and big data. The Applied Data Science course at Syracuse University provided me with one such opportunity to gain knowledge and skillsets in various practice areas of data science such as data mining, text mining, big data, business analytics and information visualization to name a few. Throughout my masters I got exposure to different technologies and acquired a great amount of knowledge and project experience during course of my graduate program

# Learning Outcomes of ADS Program

The ADS program was flexible and allowed me pick my field of interest and master these skills through various courses and projects. In this section, I aim to highlight how these courses and projects helped me master the learning outcomes of this program:

* I first got a broad overview of the practice areas in data science through the course IST 687 – Introduction to Data Science. Through this course I learnt the fundamentals of data science and did my first data science project. Subsequently, through different courses such as SCM 651 - Business Analytics, IST 736 - Text Mining, MBC 638 - Data Analysis and Decision Making, IST 718 - Big Data Analytics, IST 719 – Information Visualization etc., I learnt about the different facets of data science and areas that come together to form this field. Through SCM 702, MBC 638 and SCM 651, I learnt how to perform statistical analysis on a given data and use formulas and technique to answer solutions to lingering questions on the supply chain or business side. Through IST 707 and IST 736, I learnt the working of different types of machine learning models and the basic math behind it.
* Collecting and organizing data is an essential part of every data analysis and I got the opportunity to learn this skill from almost all the courses in the ADS program through the projects I did. However, one course which gave me the maximum exposure and experience of collecting data was IST 652 – Scripting for data analysis. This is because, in this class we had to download or scrape data off different websites using various libraries from python and then manipulate the data to get a desired form. This helped me figure out various tricks and tips of scraping massive amount of data present on the web in structured or unstructured format, which can then be used for data analysis. I also got to learn a lot of data storing, cleaning and manipulation techniques that will aid in faster data analysis and data visualization
* Another important aspect of data analysis is identifying patterns through data via visualization, statistical analysis, and data mining. There were several courses that helped me learn and hone these skills. Through MBC 638 - Data Analysis and Decision Making, I learnt a great deal about statistical analysis and the basic statistical concepts required for a data science role. I also had a course IST 707 - Data Mining which helped me understand data mining concepts in more clarity and also the underlying working of these machine learning algorithms. I also took a Data Visualization course – IST 719, which further improved my knowledge of data visualizations and how to tell a compelling story through data. This subject required one to create a huge poster and present it to people during the poster presentation day. I still have this poster with me and hang it proudly in my room.
* Based on the data available, it is important to develop different strategies to analyze, understand and interpret data. It is very important to get a quick summary of the data in order to come up with strategies of how to analyze them and what outcomes to expect from them. I got to learn this through courses such as IST 687 – Introduction to data science and IST 707 Data Analytics. Through various classes and project, I learnt that it is very easy to do a very bad job at machine learning and data mining. So before we choose strategies for data analysis such as whether to use a supervised or unsupervised machine learning model we must first understand data and the business questions we are seeking to answer. Once we have understood the data we can then proceed to manipulate it and apply ML algorithms to get useful insights from it. I learnt these skills from subjects such as IST 687 and IST 707.
* After performing data analysis and obtaining insights, it is important to develop a plan of action to implement the business decisions derived from the analyses. We need to provide recommendations/ suggestions as to how to address the business problem based on our analysis. I especially got this experience while working on projects from IST 724 – Database Security, IST 707 and IST 718, as all these three courses involved providing suggestions/ recommendations at the end of the project. Through these experiences, I learnt that in order to provide useful and achievable plan of action, we need to thoroughly understand the outcome of our analysis and how the insights will impact the organization if they are implemented. What helps us form these suggestions are the insights we obtain from the exploratory data analysis, data modeling and factors we considered for forming a good prediction model.
* Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization. Most of my courses involved a project presentation at the end of the semester to professors, clients and students. This helped me hone my communication skills and learn the art of storytelling to cater to audiences from different backgrounds thereby preparing me for the corporate world experiences.
* Synthesizing the ethical dimensions of data science practice such as privacy. This is a very important point to keep in mind while approaching any data science project. It is important to not get carried away by different tools available in the market to extract data, some of which could be unethical to use. We were taught about who will be affected by this project, is it harming any group of people, who is in control of the data and to what extent etc. This is something that we have been made aware as a part of every course taught in the ADS program and every project I undertook.

# Coursework and Projects

In this section, I would like to briefly highlight some of my coursework and projects that helped me shape my data science portfolio and contributed to achieving the learning goals of the ADS program.

# IST 687: Introduction to Data Science

This course was an introductory course to the field of data science, and I had taken this in my first semester. It covered all the fundamental concepts of data science, served as an introduction to R programming and helped me understand some of the important supervised machine learning algorithms such as Linear Regression, Logistic Regression, Apriori, SVM and KSVM. Through this course I had the opportunity to understand how to collect, clean, organize, manage, transform, predict and visualize data. The course classes helped me learn new concepts and through assignments I was able to revisit the concepts and gain clarity. Finally, the course project helped me put my knowledge to good use. From this course, I started my first journey with a data science project, where I worked with a Hyatt Hotel dataset and produced insights from it. It helped me get an idea of what should be expected in a data science project and further increased my interest in this field.

**Hyatt Hotels Project:** For this project we worked in groups, on data from Hyatt Hotels in California for the month of August, 2014. Using R as a tool, our goal was to analyze this data and identify factors that affect Hyatt Hotel NPS (Net Promoters Score) and suggest possible actions the company can take to improve it. This was a good platform to learn how to use data analysis to help a business improve. After obtaining the dataset, we cleaned the data, performed descriptive analysis, found important factors which will help predict the NPS and predicted the NPS using different machine learning models – Association rule mining, SVM and Random forest. Finally, we visualized the results and presented it to the class.

# IST 707: Data Analytics

This course further improved my knowledge in R language and data mining. It also taught me how to covert the R code into a presentable format. The presentable format could be an html file or an app hosted on cloud. While introduction to data science focused on getting us well acquainted with data mining, this subject challenged us by taking us in depth into the whole process of data mining from start to end. Some assignments were challenging and required us to work with large Kaggle competition dataset. We were also encouraged to submit our Kaggle results and compete our models accuracy with the world. The project requirements needed one to create an app on cloud as well as create a poster to present it on poster presentation day.

**Costa Rica Poverty Level Prediction Project:** The main goal behind this dataset was to predict the household poverty levels in Costa Rica and determine the driving factors for that decision. Here too we used R as the programming tool for data mining and created a shiny app that was hosted in cloud so that anyone can access our app and get insights from it. We performed all the necessary steps required in datamining such as data extraction, data exploration, data cleaning, data selection and data modelling. To predict poverty we used algorithms such as Naïve Bayes, Logistic Regression and Random forests. The Random forest model gave us an accuracy of 94%. Finally we created an app that gives users the ability to visualize data, select data and run the above mentioned models. Here is a link to it <https://vishu133.shinyapps.io/Costarica/>

# IST 719: Information Visualization

This course was also in R programming but the main focus was on visualization and story telling. I learnt how to collect data from various source and manipulate it to produce the require visualization.

**Fighting the Obesity Epidemic Project:** The main goal of this project was to convey meaningful insights through visual communication. Data as raw numbers or text requires the audience to stop, focus, concentrate, and possibly calculate values to derive meaning. To express that value quicker and in a more memorable fashion the data can be arranged in symbols and colors. Humans can evaluate colors and symbols –sizes, positions, patterns—in a preattentive manner, meaning communication is instant and more pleasing to the audience.Various graphs and plots were created to give the story about the rise of obesity in USA for the years. Publicly available information was collected from CDC and tools like R, adobe illustrator was used to create the final project which was a poster filled with graphs and insights.

**Project Document:** [**https://github.com/vhegde13/Portfolio/blob/master/Projects/IST719/719poster.pdf**](https://github.com/vhegde13/Portfolio/blob/master/Projects/IST719/719poster.pdf)

# IST 659: Data Administration Concepts and Database Management

This course on collecting data and managing it on a system other than flat files taught me about how to create and handle databases. The data can be collected and organized in a database which makes it easier to access, read, organize and extract data for business and analytic purposes.

**Aggregator of Local Physical Stores Database System:** The proposed database project will provide users with information about types of products available in local stores along with their prices and description. It will also provide local stores with information regarding consumer behavior and address so that they can serve their customers better through offers and discounts. An aggregator platform that lists products available would make it easier for customers to save time, make informed decisions and plan their trip to the store accordingly. The platform would also help local stores get customer data and develop plans according to insights provided in the data. The database needs to be constantly updated using inputs given by the local stores so that customers get accurate data every time they access the platform. This database will power an online platform. The users will be customers and database administrator. For simplicity the administrator will provide information to local stores.

**Tools used:** Microsoft SQL server, MS access and MS visio

# Appendix

This section provides link to all the project from the courses mentioned in the above sections. Please click on the course name to access the project for that course (GitHub links have been embedded in the name of the course).

|  |  |
| --- | --- |
| Courses | GitHub Link |
| IST 687 | [Hyatt Hotels](https://github.com/vhegde13/Portfolio/tree/master/Projects/IST687) |
| IST 707 | [Costa Rica Poverty Level Prediction](https://github.com/vhegde13/Portfolio/tree/master/Projects/IST707) |
| IST 719 | [Fighting the Obesity Epidemic](https://github.com/vhegde13/Portfolio/tree/master/Projects/IST719) |
| IST 659 | [Aggregator of Local Physical Stores Database System](https://github.com/vhegde13/Portfolio/tree/master/Projects) |

Github link for all projects <https://github.com/vhegde13/Portfolio>