Syracuse University, School of Information Studies

M.S., Applied Data Science

Portfolio Milestone

Name: Haiwang Ge

SUID:606325845

Table of Contents

[1. Introduction 3](#_Toc120476587)

[2. IST 659 - Database Administration Concepts & Database Management 3](#_Toc120476588)

[a. Project Description 3](#_Toc120476589)

[b. Reflection & Learning Goals 4](#_Toc120476590)

[3. IST 707 - Applied Machine Learning 4](#_Toc120476591)

[a. Project Description 4](#_Toc120476592)

[b. Reflection & Learning Goals 5](#_Toc120476593)

[4. IST 719 - Information Visualization 5](#_Toc120476594)

[a. Project Description 6](#_Toc120476595)

[b. Reflection & Learning Goals 6](#_Toc120476596)

[5. IST 722 – Data Warehouse 6](#_Toc120476597)

[a. Project Description 7](#_Toc120476598)

[b. Reflection & Learning Goals 7](#_Toc120476599)

1. Introduction

The Applied Data Science program at Syracuse University’s School of Information Studies emphasizes the impact of data science and analytics on business practice and decision making and provides students the opportunity to collect, manage, analyze, and develop insights using data from a multitude of domains using various tools and techniques. In courses such as Database Administration (IST 659), Applied Machine Learning (IST 707), Information Visualization (IST 719), and Data Warehouse (IST 722) reports and presentations were developed to deliver insights using Microsoft Power BI, Microsoft Azure, SQL Server Management Studio, Python, R-studio, Excel and Adobe Illustrator.

1. IST 659 - Database Administration Concepts & Database Management

This class is going through definition, development, and management of databases for information systems. Learning data analysis techniques, building data modeling, and schema design. Introduction for SQL language, overview of file organization for databases. Understanding data administration concepts and applications.

1. Project Description

Through studying Database Administration under the direction of professor Jillian K. Lando, our group developed a database that used for recording students, teachers, and advisors’ statistics for each semester and applied in a User Interface that faces multiple user group. The user who are students can register courses, check waitlists, submit assignments, check grades. The user who are teachers can assign homework, grade homework, give final grades. The adviser user can approve students for registering courses.

We created the entities and attributes for the database, there are total 6 entities which are semesters, courses, users, sessions, assignment and assignment grades. Semesters represent each 5 semesters for a year, courses represent the course name and the course credits, users mean different types of users, sessions represent the same course have different sessions. Then, we made the ERD and logistic model diagram base on the relationships, and make the database system by the SQL query. Also connect the SQL and Power Apps to show the user interface.

1. Reflection & Learning Goals

During this project, the database will be heavy constrained since all the data will get into the user interface app, so we will make the data standard to unify all the information. Also, we used lots composite primary key in the database because there are many weak entities, so the user will use weak entities to search the information. The exercise of developing a data management solution revealed the significance of how the data is stored and accessed, which is imperative to analysts and data scientists.

1. IST 707 - Applied Machine Learning

This course general overview of industry standard machine learning techniques and algorithms. Student will focus on machine learning model building and optimization, real-world applications, and future directions in the field. Hands-on experience with modern data science packages.

1. Project Description

Through studying machine learning technique under the direction of professor Stephen Wallace, various data mining techniques were introduced which perform with varying precision and efficiency for applications in regression, classification, and clustering. For the final project, multiple classifications were implemented to compare recall and accuracy in predicting the heart disease from personal key indicators. The data we use is form Kaggle and python is the major programming language we use in this project.

This application required the cleaning and preprocessing of data, which sampling the data since people who don’t have the heart disease is much over than the people who have, the prediction will be overfit if we use the original data. Therefore, we split the data by the heart disease status, and make those two groups are balance which means same size. After the data preparation, dependent variables include four parts which are bad habits, possible related diseases, physical symptoms and personal information, predictor variable is about whether the observation has heart disease.

After split the data to training and testing data, logistic regression, SVM, decision tree, random forest, KNN and XGBooster were using for the prediction classifications. Since the model we retrieved from the machine learning package in python, to optimizing classifiers we need tuning parameters for each classification and use the ROC score to show the diagnostic ability of binary classifiers.

1. Reflection & Learning Goals

The best classification for this project is XGBooster which the accuacry is 76.3%, people who are over 60 years old and have high BMI, bad smoking habits and less sleep time have higher risk of getting heart disease. This project contributed to the successful application of the learning goals through the development of classifications based on the data, and the communication of observations which translate to actionable insights.

1. IST 719 - Information Visualization

This course will introduction to skills and techniques related to information visualization, through the R programming language, Adobe illustrator. These skills include data cleaning techniques, control of the R graphics environment, develop custom plots, visually explore data, use design concepts to visually communicate the story in the data, and discuss issues related to the ethics of data visualization. Conceptual themes will be presented alongside technical aspects of data visualization. Information visualization is the art of representing data in a way that it is easy to understand and manipulate. It can also help us make sense of information and thus make it useful in our lives like help people notice things that might go unnoticed if not visualized.

1. Project Description

Through studying information visualization technique under the direction of professor Jeff Hemsley, various plots for data visualization created by R-Studio were introduced, and grasp plot processing by Adobe illustrator. For the final project, multiple plots were created for show the story base on the data. ‘QS World University Rankings dataset’ is the data I used for this project from Kaggle.

This application required the cleaning and preprocessing of data, which get rid off the empty data. I only keep the name, country, city, university type, university size, university research and international student as variables to figure out the factor to affect the University QS score. I use the dplyr, ggplot2, tidyverse, RColorBrewer, ggalluvial to create word map, bar plot, circle bar plot and flow plot for the poster. The world map represents how many QS University in each continent, and the bar plot represent the value in different divisions, and the flow plot can present multiple dimensions by type, research and size.

1. Reflection & Learning Goals

According the poster, it shows the research is very important key to be a QS University, and also school size is very important factor to decide the research output. This course shows data visualization and retrieve information from data is essential for any data professional.

1. IST 722 – Data Warehouse

This course provides concepts, principles, and tools for designing, implementing, and using Data Warehouses. More specifically, we introduce database constructs such as Operational Data Store (ODS), Data Warehouse, and Data Mart, and related components. We study the differences between Ralf Kimball’s and Bill Inmon's approaches, roles and responsibilities in the design and implementation of a Data Warehouse, project management guidelines and techniques, requirements gathering, dimensional modeling, Extract Transform and Load (ETL) architecture, analytical reporting concepts, data governance and recent trends in the data warehouse domain.

1. Project Description

Through studying data warehouse concept, SQL, Microsoft Azure, and Microsoft Power BI technique under the direction of professor Humayun Khan. We will use fudgemart and fudgeflix database to build data warehouse based on the order fulfillment to figure out the shipping lag days.

The fact order fulfillment is evaluated by three detentions: customer, product, date. Then, loading the data into data warehouse from database by Microsoft Azure and doing the two types business analysis which are moalp from cube and rolap from the data warehouse.

1. Reflection & Learning Goals

According to the business analysis, Fudgemart need improve the shipping efficiency to reduce the lag day in some states like CA which have huge orders. For Fudgeflix company, some states in the US don’t have the service, so the company could evaluate the business to raise the benefits. This course represents how to use data warehouses as a business intelligence solution to make better organizational decisions.