Name: Yafei Lou

NetID: ylou05

SUID: 754244568

Portfolio First Draft

**Learning Objectives**

During the process of completing the applied data science program, now I am capable of developing a strategic plan to interpret the data using appropriate algorithms and BI tools. Most courses helped me to achieve the program learning goals, while other courses taught a different area of expertise such as finance. Therefore, I have an overview of the major practice areas of data science, identifying patterns in data via visualization, statistical analysis, and data mining, and developing alternative strategies based on the data. The various classes and projects in the program covered most fields of the data sciences and provided a good starting point to pursue further study in data science practices.

In addition, most projects provide an opportunity to practice in identifying the patterns, developing specific business strategies, and delivering the project results to the class. However, I did not get a chance to touch on the ethical dimensions of data science practice, despite the fact that the data privacy has been a controversial topic for the past years. For the next section, I will dive into some of the important projects and assignments in order to demonstrate what I learned from the courses and how to apply those skills in the real-world.

**Projects**

Airline Travel Company (IST 659): a metasearch engine website that monitors flight booking changes and cancellations during the covid outbreak

As covid-19 consumed the planet, there has been so many restrictions and international regulation on the international flights. In addition to these difficulties, some countries’ borders have been closed and some travel agencies charge a hefty fee to customers who wants to fly back to their home country. Therefore, this project aimed to sell international flight tickets and provides information on the latest travel policies.

The data is collected from some flight tracker websites using the web crawler and cleaned using Python. In this project, I used MSSQL, Visio, Access to design, build, and demonstrate the database. First, I used Visio to draw the ERD of the relational database, so I can visualize the conceptual design of the database. After designed the database in Visio, I used MSSQL to build the database in SQL. Lastly, I used Access to build a user interface, so users can perform different business activities, such as purchase the ticket and look up a particular flight’s past history. For the project deliverable, I presented the user interface in front of the class, used business story-telling for our project idea and explained our data warehouse structures.

**Yelp recommendation system (IST 718): recommending the restaurants by predicting the users’ rating based on users’ past visits**

The data provided by the Yelp website is a subset of the businesses, reviews and user data as JSON files. In addition, the yelp data contains many attributes such as hours, parking, availability and ambience. However, one of the challenges is that the data has exceeded 1 million rows, we have to optimize the model in order to simulate the model training process during presentation. In this project, the model was constructed on pyspark and Tensorflow and utilized pandas and sklearn packages to perform the preprocessing, descriptive analysis, and feature selection. More specifically, a LSTM model used to study users’ behavior, an artificial recurrent neural network architecture, maps a sequence of past observations as input to an output observation. As a result, given a new customer’ data input, we can predict what star this user would give to a restaurant. Almost 80% of our prediction about the potential customer are correctly predicted, so we can use the model to recommend the restaurants on users’ home page.

**Southwest Airline Customer Churn Rate (IST687)**

Since this project required us to find out why Southwest airline has such low customer churn rate, which is more like an open-end question, I came up some business questions that might help answering the question from exploratory analysis. Therefore, some of the business questions I came up with are what factors related to higher and lower satisfaction rate of customer, what routes produce particularly satisfied or dissatisfied customers and what effect the airline status has on customer satisfaction.

The data is provided, so there is no need to collect the data from the web. In the data cleaning step, I identified 26 observations with missing data throughout all categories and removed those from the data. I then converted suitable variables to factors for easier descriptive analysis and re-ordered those factors appropriately. I also generated new categorical variables, such as net promoter score, for a better measurement of likelihood to recommend the company to others. After doing the descriptive analysis and association rules analysis, it turns out that loyalty is not a particularly useful variable and is strongly negatively correlated with a higher number of flights likely simply because fewer flights mean a higher chance of having taken 2-3 with a specific airline. That said, this does show a number of loyal, regular customers (and even some frequent ones) with no status, which suggests room for action.

Based on these findings, I would make two business suggestions to the client. First, the Fly Fast Airways is causing significant issues with customers satisfaction. Therefore, the company should contact the Fly Fast Airways to improve its service. Second, the silver status is the single biggest positive factor affecting customer satisfaction. Lastly, the frequent flyers rank the airline poorly, so it will be worthy to discover why those customers have the complains, since they are the major source of revenue.

In conclusion, although there are other important courses not listed in details, such as data analysis and decision making - MBC 638, data mining – IST 707, the MBC 638 went through the fundamental statistics and IST 707 introduced practical data analytics techniques. Additionally, throughout the completion of applied data science program, I am familiarized with machine learning algorithms and be able to apply the data analysis knowledge learned from the courses on the real-world problems.