Name: Qingqing Hu

NetID: qhu100

SUID: 262500934

**Personal Portfolio**

**Learning Objectives:**

In the past year and a half, I am very honored to study in the Applied Data Science program in ISchool. Through this period of study in the Applied Data Science program, I have gained a deep cognition and understanding in this field. After learning the various courses, I have been able to independently formulate strategic plans and interpret data at the present stage using appropriate algorithms and various tools. During this period, the courses in Applied Data Science program have helped me understand relevant knowledge in this field and achieve the main learning objectives of this program. At the same time, I also gained expertise in other fields, such as finance and business analysis, through other courses. Therefore, in this paper, I will summarize my understanding of this professional field, the main practice areas of data science, determine patterns in data through visualization, statistical analysis, and data mining, and develop alternative strategies based on data. The Applied Data Science course covers most areas of data science and provides a good starting point for further research into data science practices.

At the same time, most courses have provided opportunities to practice identifying patterns, develop specific business strategies, and deliver project results to the class. So, in the next section, I'll delve into some important projects and assignments to show what I learned from the course and how to apply those skills in the real world.

**Projects:**

1. **Data collection: IST - 652: Scripting of data analysis**

**Stocks Prices and Finance Tweets Analysis**

Data collection is the foundation of every data science project, and it is a basic skill that we will need to complete every project in the future. In this section, our main goal is to understand how to use tools to collect and organize data. In this session, the first thing we do is gather the information or data set that meets our needs through the Internet or other channels. The next step is to organize the data and perform some basic processing on the data, including data cleansing and data filtering.

In IST-652: Data Analysis of script: In this project, we tried to study the effect of tweets sentiments on the stock prices of companies. We are analyzing five well-known companies (Amazon, Microsoft, Apple, Advanced Micro Devices (AMD), Tesla) for this study.

First, we found out and filtered tweets of above top 5 companies using regex patterns. Then we performed some descriptive analysis on these filtered tweets as well as lots of visualizations to convey the information to audience directly. In order to achieve the goal of the project, we performed various analyses on tweets as well as financial stock prices. We performed word frequency analysis to analyze the content present in the tweets. We used VADAR's Sentiment Intensity Analyzer on these tweets to calculate sentiment scores. We used these sentiment scores to analyze relationships between market sentiment and stock prices. In addition to this, we also tried to forecast the stock prices using a multivariate time series model, called Vector Auto Regression (VAR) which uses past values of a group of time-dependent variables to forecast future values. In this case we are using tweet sentiments and past stock prices of the companies to forecast future prices.

Finally, we performed multiple analyses and experiments on both tweets and stocks price data. From sentiment analysis on tweets, it was seen that there are low percentage counts of negative tweets (15%). In addition to that, stock of all companies has a positive trend, this hints that positive market sentiment has pushed up the stock prices.

Although from the correlation values, it was not very much evident that stock price and compound sentiment score are strongly correlated. This may be due to non-linearity in compound score as the correlation computation assumption signifies the linear vector. Forecasting using the VAR model with sentiment and stock prices resulted in acceptable error in all the companies except Tesla.

1. **Data analysis:** **IST-707 Data Analytics**

**Classification Analysis of Red Wine Quality**

Data collection is very basic work of data analysis project, while identifying patterns in data through visualization, statistical analysis and data mining is the bridge for us to understand and display data structure and data characteristics, so as to select the appropriate data analysis model. In the second part, we mainly study how to identify patterns in data through visualization, statistical analysis, and data mining. Data visualization and data analysis have been the focus of our research in the past two years. We need to know how to use Excel, R language, Python, Tableau, WEKA and other data analysis tools to classify, cluster and analyze data according to the characteristics of data, as well as data visualization.

In this project, we intended to figure out what factors influence the red wine quality mostly and build the model to impartially evaluate the wine quality and possibly find the pattern among the data points using the scientific method. Overall, Software R is used to predict the red wine quality via classification algorithms and logistic regression. Classification algorithms of Decision Tree, Naïve Bayes, and Association Data Mining are chosen to classify.

We used a data set downloaded from the UCI machine learning library. This data set is about the red variety of Portuguese "Vinho Verde" wine. The dataset consists of 1599 rows and 12 Input variables (based on physicochemical tests) and one Output variable (based on sensory data). Input variables include various of chemical elements, and quality (score between 0 and 10) is the Output variable.

First, we divided the wine into two classes, Good (7~8) with 217 rows and Bad (3~6) with 1382 rows, we set the quality column (score range 3~8) as the target variable on R. In the classification task, we analyzed and compared the mean distribution of the different attributes of the two grades to find the factors that have the most influence on the quality of wine. From the density plots, we knew the 6 factors reflecting bigger differences between the two classes of red wine.

Then, to find out which ingredients are the key to the quality of red wine, we selected data association mining to analyze the two levels separately. As a result, we found that the following components are within a certain range , volatile.acidity[0.12,0.43), total.sulfur.dioxide=[6,26), density=[0.99,0.996), sulphates=[0.68,2], alcohol=[10.8,14.9], the quality of red wine is better; we can also see that when the content of the following ingredients is in the corresponding range, such as volatile.acidity=[0.6,1.58], sulphates=[0.57,0.68), residual.sugar=[2.4,15.5], the quality of red wine may be poor.

In addition, we performed logistic regression model that has a great performance predicting the good and bad wine quality. To improve the model accuracy, we performed the stepwise regression model to include all the potential independent variables in the model and eliminate those that are not statistically significant. The model accuracy indeed has been improved to 88%.

Besides, the information gain contribution of each variable indicates that alcohol, sulphates, and volatile acidity are the top three most important factors of the wine quality assessment. We also discovered the level of alcohol is the root node by decision tree analysis.

Next, we used a naive Bayes model to analyze and predict the data. First, we build a model based on 30% (479 rows) of the train set and then make predictions on the remaining 70%(1174rows) of the test set. We used the confusion matrix to test the accuracy of the model. The accuracy of the model is 97.85%.

In the end, based on the results of four models (logistic regression, Decision tree, Naive Bayes, and association rule data mining) from the dataset, we found that some ingredients like volatile acidity, sulfates, and acidity are significant factors to identify whether the red wine is good or not. Furthermore, we can also conclude that when the Sulphates and Alcohol indexes tend to be large and the Volatile acidity index is small, the grade of red wine is higher. The analysis result of the model is almost consistent with our preliminary analysis. As for association rule mining, we can conclude that we can state which conditions have to be met to be considered as good wine, as seen in the above example. The algorithm depends on the support and confidence we set as well as the question we ask. A final reflection as a learning outcome, is that the model has better performance when predicting red wine with bad quality. Therefore, more factors need to be included to better distinguish the wine quality.

1. **Strategy and decision: IST-687** **Introduction to Data Science**

**Southwest Airline Customer Churn Rate**

In this section, we need to develop alternative strategies with our group members based on the data available. At this point, we need to have a full understanding of the data itself and the users of the data and build the required data model and give appropriate solutions to solve the problem according to the characteristics of the data itself and people's needs.

In the first semester, we learned these knowledges through the course of IST-687 Introduction to Data Science. After screening, cleaning, and analyzing the historical data of airlines, we develop corresponding strategies according to the needs of airlines, and then select several suitable models for analysis. Since this project required us to find out why Southwest Airlines had such a low customer churn rate, which was more of an open question, I posed 6 business questions that would help us answer this question from an exploratory analysis.

For data processing when preparing data, we substituted the ‘NA’ values in departure delays, arrival delays, and flight time with 0. If there was no value in the delay, we can assume there was no delay. If there was no value in flight time, we can assume the flight was cancelled. Lastly, we replaced missing values for the likelihood to recommend with the average value of likelihood to recommend. We also generated new categorical variables, such as the net promoter score, to better measure the likelihood of recommending the company to others.

First, we have a basic understanding of the data available by some visualizations, such as age, year of first flight, loyalty, type of travel. These direct information expressions provided us a guide for the business questions in next analysis. I will introduce briefly the six questions and model results for achieving our goal:

Q1: What are some common characteristics of detractors and promoters in relation to the available data?

We used association rule mining in the project, by placing recommender types on the right to identify the set of customer attributes that make customers become promoters and critics. From the result obtained after running the Associative Rule Mining model, based on the confidence level, we observed that personal flight customer whose age is over 70 with a Blue airline status and has both long flight and long delay is very likely to become a detractor; on the other hand, male customer whose age is between 30-50, with gold airline status and business travel sitting in the economic class and without long flight and long delay is very likely to become a promoter.

Q2: how can we expect passengers to rate their experience based on their age?

To model age, we divided our age data into 4 groups. Here we create 4 data frames for age groups: 30 and under, 31-50, 51-70, and over 70. Based on linear regression analysis with dummy variable and adjusted R square value, we can confidently say that we can expect older passengers to rate lower than younger passengers, with a cutoff age of around 60. So, we recommend that the airline develop a strategy to better accommodate older passengers. This may include offering a discount (off the flight or when checking a bag), training employees to better accommodate older passengers, allowing senior citizens to board the plane before other groups to allow them to settle, etc.

Q3: Can a promoter or detractor be predicted from just the passengers age, class, and type of travel?

After processed data with binary variable of promoters, we used support vector machine to train set and text set. We also used confusion matrix to test accuracy of 67%. Based on our SVM and linear modeling, we can say with confidence that these factors, like passengers age, class, of ticket and type of travel, are heavy contributors to the passenger’s likelihood to recommend.

Q4: How are blue status passengers rating exactly, as compared to those with silver, gold, or platinum status?

We use linear models to test whether Type of Travel and Airline Status are important to customer satisfaction rates. Dummy variables are created to classify and numerate different categories. We find that, by looking at p-values and adjusted R-square, our assumption about Type of Travel and Airline Status is backed by this linear model. From the linear model, we can tell those customers who have Blue Airline Status are more likely to have lower likelihood to recommend, compared to customers with Silver, Gold or Platinum status. Also, Business Type of Travel passengers have the highest likelihood to recommend flights to others.

Q5: Does the airport have a relation to the likelihood to recommend?

First, the average origin NPS and average Destination NPS mapped per airport. This is then plotted on a map of the United States. We showed the best and worst origin airports and mapped the air travel out of this region and peeled back the major travel to look at the regional travel of the bad airports. At last, we noticed that these are all run by the same partner airline: FlyFast Airways. It appears that the issue of lower ratings in the Texas region is not due to regional issues, but a partner issue where Flyfast airlines is underperforming compared to other partner airlines. Flyfast airlines is dragging down the total NPS score of Southeast Airlines. The relationship with this partner should be renegotiated or terminated.

Q6: What are some aspects that we can focus on to gain customer satisfaction?

We mainly analyzed this question by text mining with a word cloud, From the result, we suggest that the airlines can focus more on seat and delay because most customers are looking forward to these aspects and it can affect the promoting factors.

1. **Implementation:** **SCM-702 Principles of Management Science**

**Diet Planning**

In this step, we need to formulate an action plan to implement business decisions, adjust existing data and models by understanding customer needs and preferences, and then make further plans based on the adjusted models to ensure the smooth operation of the business.

In the SCM-702 Principles of Management Science course, the purpose of this course is to introduce students to the scientific aspects of management: the application of scientific methods in management and personal decision-making. Decision analysis includes defining problems, developing models, obtaining data, developing solutions, testing solutions, analyzing results, and implementing results.

In this project, the question we studied was what the minimum daily expenditure for an adult in Syracuse is to meet daily nutritional needs. First, we believe that total nutrition and all constraints can be expressed as a linear function of food, so we choose a linear programming model to solve this problem.

In order to obtain the data we need, we conducted a detailed investigation on the prices and nutritional components of 14 common foods. Combined with the nutritional requirements in the website's health department report, we have set intake limits for calories, fat, protein and dietary fiber. At the same time, according to the requirements of healthy living, we have constrained the intake of food (divided into four categories: bread, meat and eggs, vegetables, and fruits).

Next, we employed advanced MS-Excel spreadsheet modeling techniques to create useful decision-support tools, that is solver. In order to find the best solution, we constantly adjust the parameters. For very difficult issues, the professor also provided us with guidance in time. In the end, our project successfully solved the problem and found the optimal solution (Bread: wheat bread(80g), white bread (463.67g) Meat & Egg: ribs(80g), egg(80g) Vegetable: sprouts(80g), mushrooms (80g) Fruit: banana(80g), grapes(80g). The minimum daily cost in Syracuse is $4.61.

Even though our model figured out the most effective solution for us, we still point out some issues with the model. For example: Nutritional requirement not comprehensive: lack of vitamin & dietary fiber; food choices in real world scenario are more than 14 types. Food prices will change with seasons and market conditions, which will ultimately affect the model results.

**Conclusion**

To conclude, after studying at ISchool for one year, I have a deeper understanding of the field of applied data science. The various courses attended focus on the application of various tools, such as R language, SQL, python, database construction and management. More importantly, each course will combine the theoretical knowledge learned into real cases and project operations. From the initial problem definition to data acquisition, modeling, parameter adjustment and model prediction, even to analysis and interpretation and data visualization, the realization of each link is the consolidation of learned knowledge and the exploration of extracurricular knowledge. This learning experience is a very precious and valuable experience for me. I believe it will bring more inspiration to my future study and work.