Syracuse University, School of Information Studies

M.S. Applied Data Science

**Portfolio Milestone**

Xuanran Ji

SUID: 397655968

<https://github.com/jixuanran/Portfolio_Milestone_Applied_Data_Science>

**Introduction**

The applied data science program provides the student with the opportunity to learn a board area of skill that is required for work related to data science subjects. Collect, manage, clean, visualization, analysis, decision making, and so on. During the past year, I have learned many subjects such as natural language processing, text mining, big data analysis, data analysis, accounting analysis, business analysis, and so on. Each course has provided a unique insight into what data science is and how to make use of and draw useful insight from different data sets.

**Learning outcome for applied data science program**

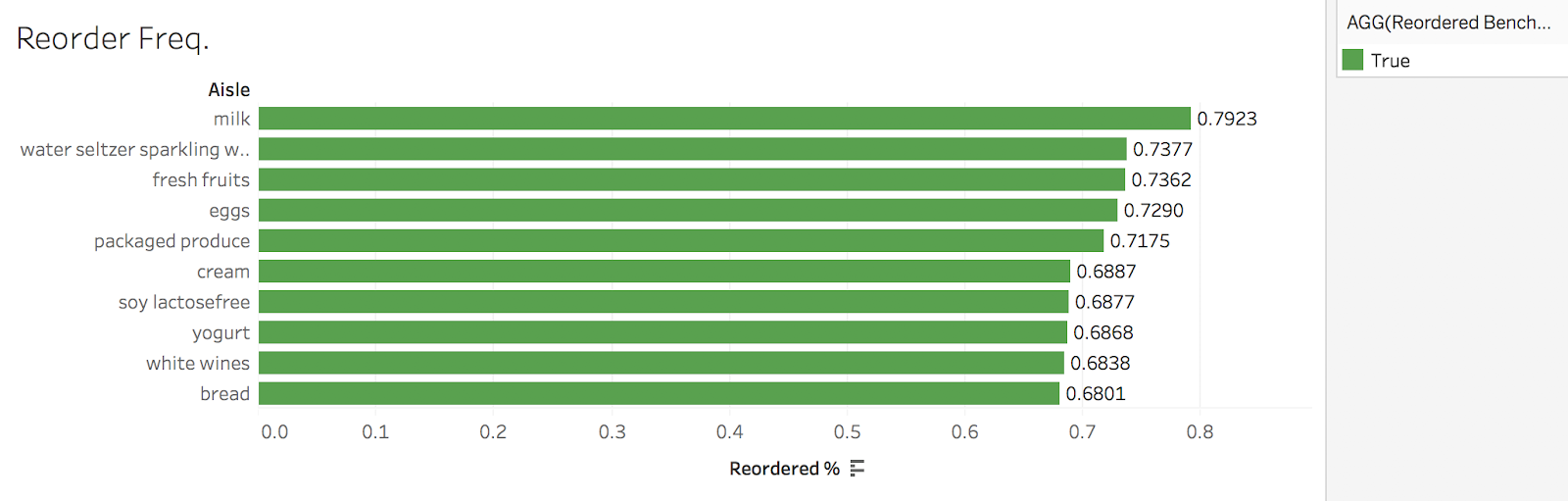
* Describe a broad overview of the major practice areas of data science.
* Collect and organize data.
* Identify patterns in data via visualization, statistical analysis, and data mining.
* Develop alternative strategies based on the data.
* Develop a plan of action to implement the business decisions derived from the analyses.
* Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization.
* Synthesize the ethical dimensions of data science practice (e.g., privacy).

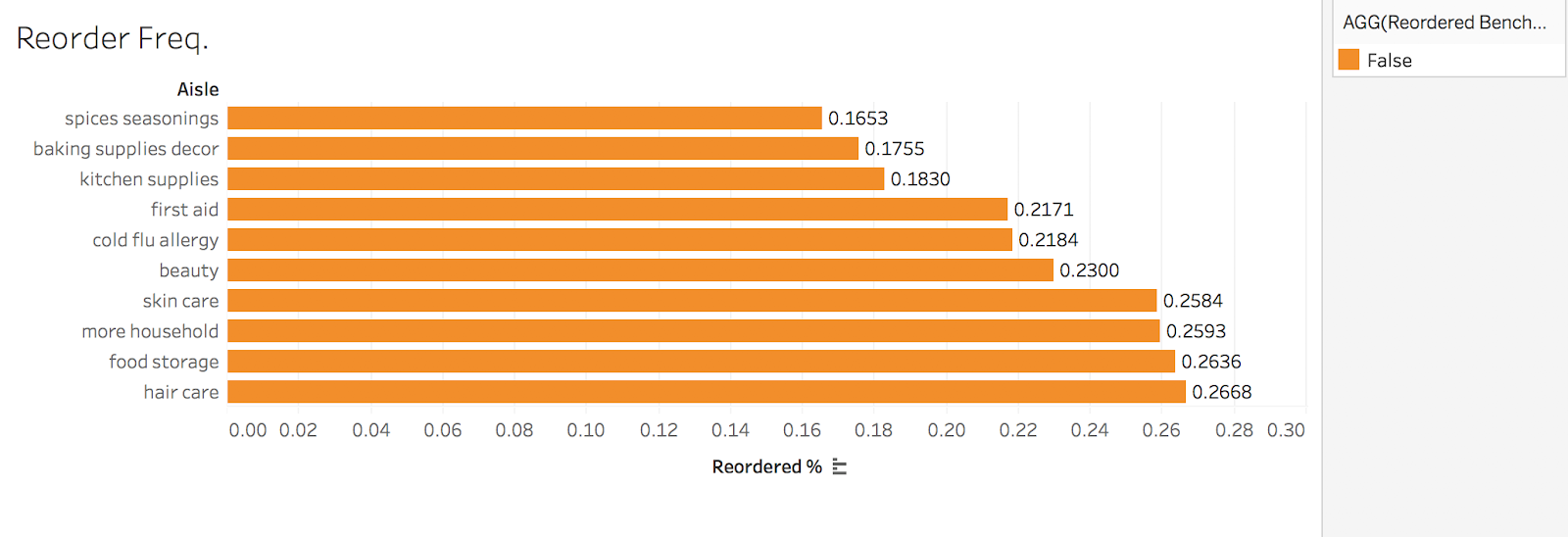
**ACC-652 Accounting Analysis (Instacart Market Basket Analysis)**

This final project was conducted in a 4 people’s group setting. We started the project by first define the questions that we are aiming to answer. We came up with the below questions to help Instacart uncover trends to improve their product and service.

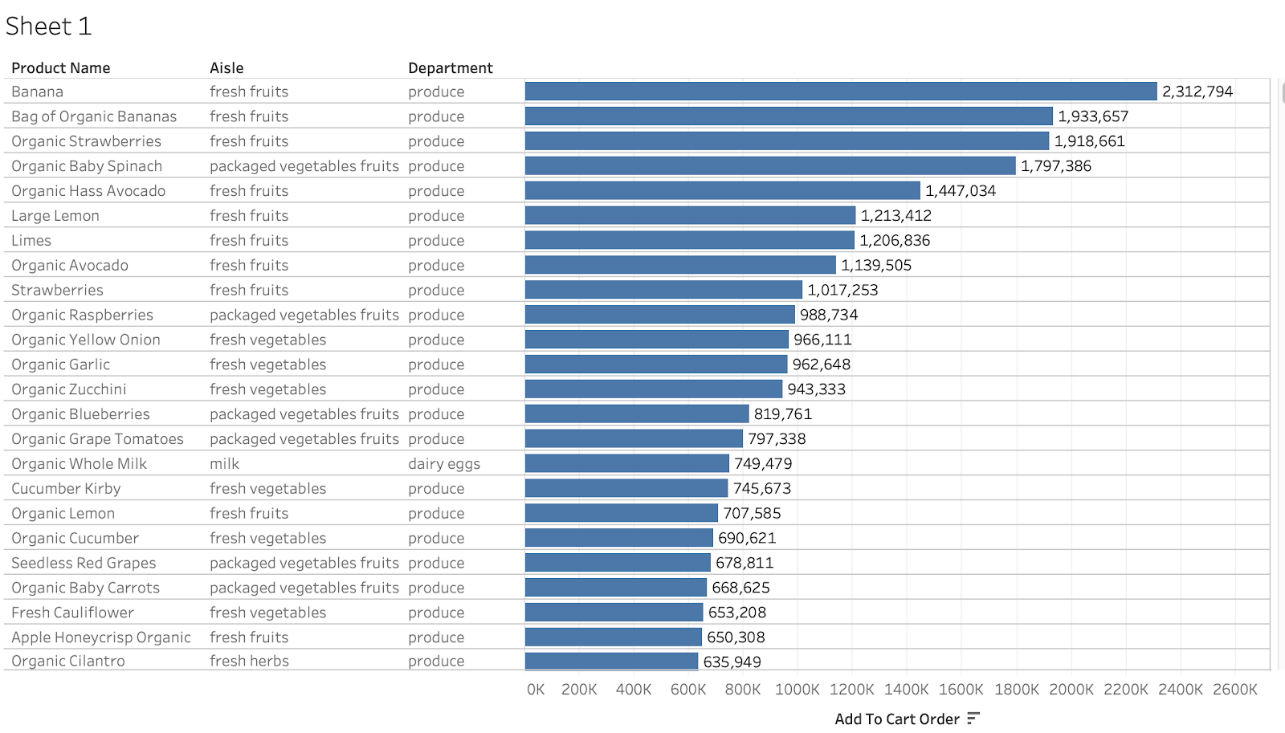
1. what types of products get reordered the highest and lowest percent and what category does it belong to? Find out reasons why certain products are not as popular as the product that got reordered.
2. Numbers of products that were ordered together. Day and time that customer would most likely to order product using Instacart service
3. How customer orders online and what products should be grouped for better sales?
4. Provide recommendations to Instacart based on add to cart order analysis. Find out the relation between add to cart order and reorder orders.

We started the data analysis by first collect data from multiple sources and merge those data set using R studio. Totally 6 files were used to merge and the data set was merged based on a common key from each file. Several files have more than 1 million rows. After the data is ready, we are ready to answer the questions that we asked at the beginning of the final project. We are using tableau for those questions. By creating a calculated field, filter, common baseline, and so on, we can answer those questions using different plot. Below are some of the examples of our final project.



This chart shows category of products get reordered the most by percentage.

This chart shows category of products get reordered the least by percentage.



This chart shows fresh vegetables from the produce department seem to be very popular and organic food is very popular VS non-organic food.

Our aim of conclusion to this final project is to help the vendor with our business recommendation. We suggest Instacart implement a recommendation system for their customer with a product that got reorder most. Because our analysis shows customers are more likely to order with the same products. We also recommend that certain product section should be placed next to each other with the market basket analysis like household closer to each other. We believe our recommendation will help Instacart to improve the sales of their products.

In this final project, I have gained a lot of experience with Excel and Tableau from the application aspect. More importantly, I have gained more understanding of developing a plan of action to implement the business decision derived from the analyses. Start from the beginning of the project. We first start to think about what questions would be helpful for the vendor to improve product sales. This is the key step to understand because data analysis projects start with a question, through data analysis, at the end to communicate what we have found from the data set. I have also gained more experience using tableau for data visualization and identify the key pattern so our recommendation for product sales is easy to understand and on point.

**Learning outcome**

This final project has provided a successful application to collect and organize data. The final data set for this project was obtained through multiple data sources with different information stored in each file. In order to prepare the data set ready for further analysis, the data set has to be organized. One of the important aspects of this final project is to think about the question in a business circumstance. We have answered our question by identifying patterns using data visualization but implement the business decisions derived from the analyses is more important. We have provided our suggestion, for example, implementing a recommendation system for the item that more likely to be added to the customer shopping cart.

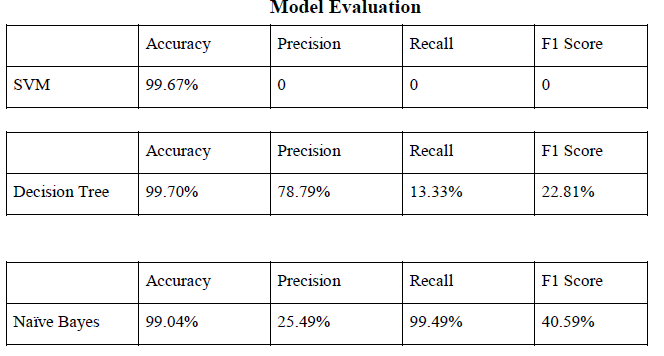
**IST-707 Data Analytics (Boston Crime data analysis)**

For me, IST 707 is the expanded course of IST 687. In IST 687, we have learned the basics of R language and studied the basic idea of several models such as linear regression, multiple linear regression, support vector machine, and so on. In IST 707, we jump into the deep end of statistical modeling. The goal of the course is to understand how different model works and how to measure the accuracy of each model. In the end we want to evaluate how model performs. Understand why a model is better than another and appropriately applies the model is the key to this class. In the final project of IST 707, I worked with another student study in business analytics from Whitemen business school. Our team has used K means cluster analysis and association rule analysis, decision tree, support vector machine and so on for the final project. The goal of the project is to help the police department to better allocate their limited resources. The main software used in this course is R studio.

We started the project by first think about what issue we want to resolve. After confirming the direction of the final project, we first clean the data. By dropping rows with NA and remove duplicates, we end up with data that has 319073 observations and 17 variables. We then further convert variables to its correct types to apply the model that we need in later analysis. For example, convert several numeric variables to factor variables. Start with some exploratory analysis, we found some interesting aspects of crime that happened in Boston. Such as the downtown area has the highest crime rate among other areas in Boston. We then implemented the Association rule and K means cluster analysis to see where and when are crime more likely to happen. We find that crimes like fraud, residential burglary, and violations are more likely to happen from Monday to Thursday. However, crimes like aggravated assault, vandalism, and property loss happen more frequently on Friday and weekends. Lastly, two models are build to predict crime involving shooting. We used the Support vector machine and decision tree model and the Naïve Bayes model to build the predictive model. All model has achieved great accuracy in terms of predict shooting crimes. In the end, we choose Naïve Bayes as the best model in terms of predict shooting crimes.



K means cluster analysis with optimal clusters.



In terms of predicting crime that involves shooting, we are using recall to access the accuracy of the model. In this case, the naïve Bayes model is the best in terms of predicting crime that involves shooting.

This project has provided a comprehensive application of several complex models. It has significantly improved our understanding of the different models. Determine the right model and compare the performance of each model consider to be the key skill for data science field. Start from the cleaning process and to the end model perform the evaluation. This course has provided a great guide for us to improve our understanding of different models. In our final presentation, we also gained more experience for our communication skill since our target audience are police department that does not know data science. As a data science, we must be able to present our analysis in a more understandable format. During the project I have also gained experience to apply different strategies based on different data set, and be able to process data from different resource.

This project was aim to find the most accurate prediction model by using different data mining techniques. By testing different models and compare different measurements of model accuracy we can make the final recommendation to predict future crimes. Therefore help police department to better manage their resources.

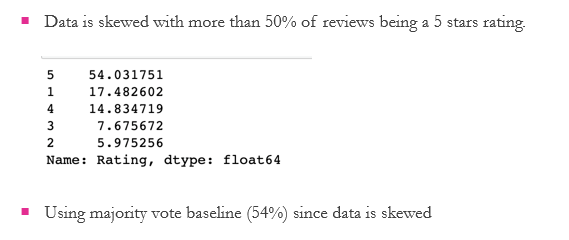
**Learning Outcome**

This final project has contributed to several learning outcomes. In the introduction, we described the areas where data science can be used the help the police department better allocate their resources. We collect and organize the data for further analysis and cleaning the data so our conclusion is accurate. We were deciding which model should be used for prediction analysis, we developed alternative strategies to look at the data set again and asked for a suggestion from the professor. Our target audience is a police officer with no knowledge of data science So our visualization and interpretation must be clear and easy to understand. We also need to demonstrate our communication skill with detailed explanation.

**IST-736 Text mining (Amazon unlocked phone review text mining)**

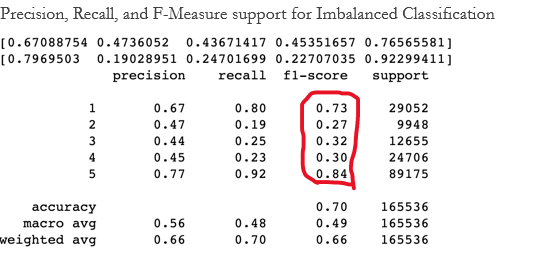
The goal of the text mining project is to classify the review with right rating because customer rating are not 100% accurate. Sometime people accidentally left bad ratings with good reviews and a good rating for a bad review. In this project, I applied several classifications to evaluate the customer review data set. Because reviews are not 100% accurate and customer reviews are the key for a company to learn and improve their products. By analysis the text data, I can help companies to improve their product sales. In this final project, the main software used is python.

Text data has its own specialty with big size and needs to be processed in order to adopt with machine learning technique. The data set for this project has more than 400000 rows and 6 columns. Before start the analysis, the data set has to be preprocessed using vectorization techniques. Vectorization is a common technique to extract feature from text to allocate words from text to a bucket called bag of words. Those techniques help us convert text sentences into numeric vectors. So, it can be used for further analysis.



Because data is skewed with most of the review rating being 5 star. When compare in the end, we will need to adjust our conclusion base on this.

**Multinomial Naive Bayes Model**



**Linear svc model**



Our conclusion, in the end, is that linear SVC performs better than Multinomial Naive Bayes Model and is more flexible to penalties and loss functions when there are large numbers of samples.

The text mining project has provided an opportunity to work with text data since text data are a little bit different and require additional process skills. Knowing how to deal with Text data, So useful insight can be drawn from text data such as product reviews, Facebook posts, and so on. I have experience with organizing and analyzing large collections of text allows for predictive analytics for text data. In this project, I have also considered the ethical dimension of data science since customer reviews are related to user privacy. It can be used to analyze user behavior.

Text data itself is important to study with its rich features. Understanding the source of text data and how text data are related to data science is also important. With more and more content being created in text format, the value of data has significantly increased. Throughout this final project, the experience of dealing with text data and understanding of text data has been greatly improved.

**Learning Outcome**

This project contributes to the application to apply data mining techniques to find out the pattern of text data. In error analysis, different strategies were thought to improve model performance. One important aspect of text data is ethical dimensions that are also considered when collecting user data from amazon.

**Conclusion**

This portfolio has demonstrated the required implementation of applied data science learning objectives and related areas in the data science field. It is also a reflection of what I have learned so far. The applied data science program has provided me with the opportunity to acquire the following key skill in data science

1. Define question
2. Data preparation
3. Data exploration
4. Model selection
5. Model evaluation
6. Result demonstration

The skill that I have learned can also be applied in a wide range of real-world applications. More importantly, knowing how to approach and why to approach and make use of data has greatly improved my learning and understanding of data science.