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 is data science and how to make use of and draw useful insight from different data sets.

**Learning outcome for applied data science program**

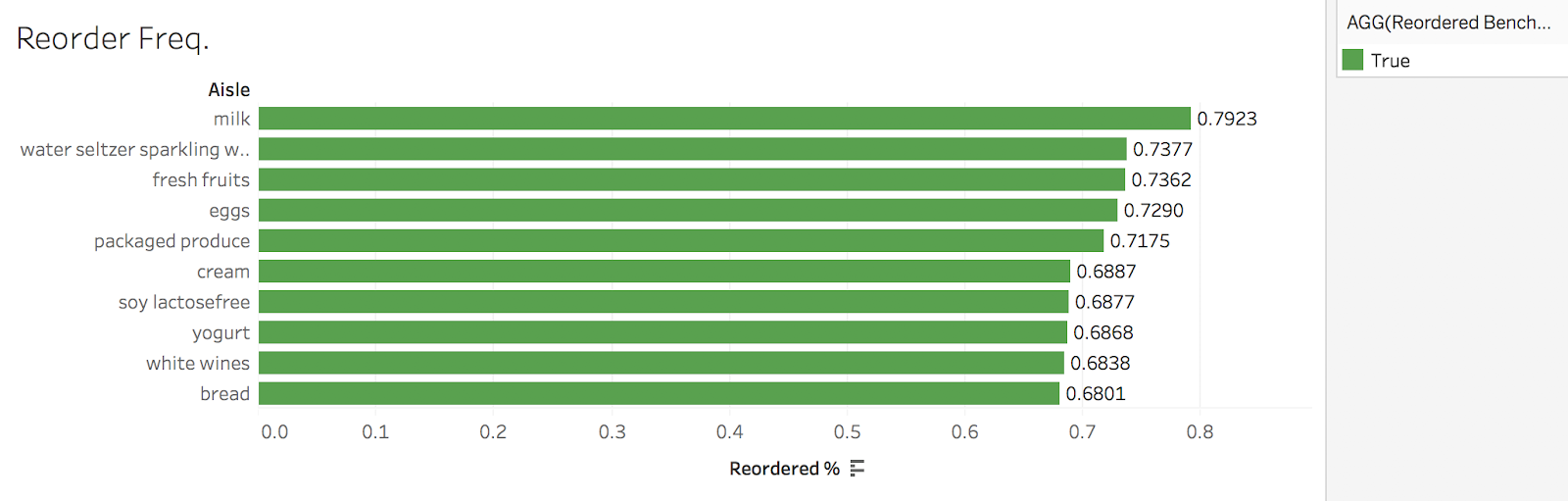
* Describe a broad overview of the major practice areas of data science.
* Collect and organize data.
* Identify patterns in data via visualization, statistical analysis, and data mining.
* Develop alternative strategies based on the data.
* Develop a plan of action to implement the business decisions 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**

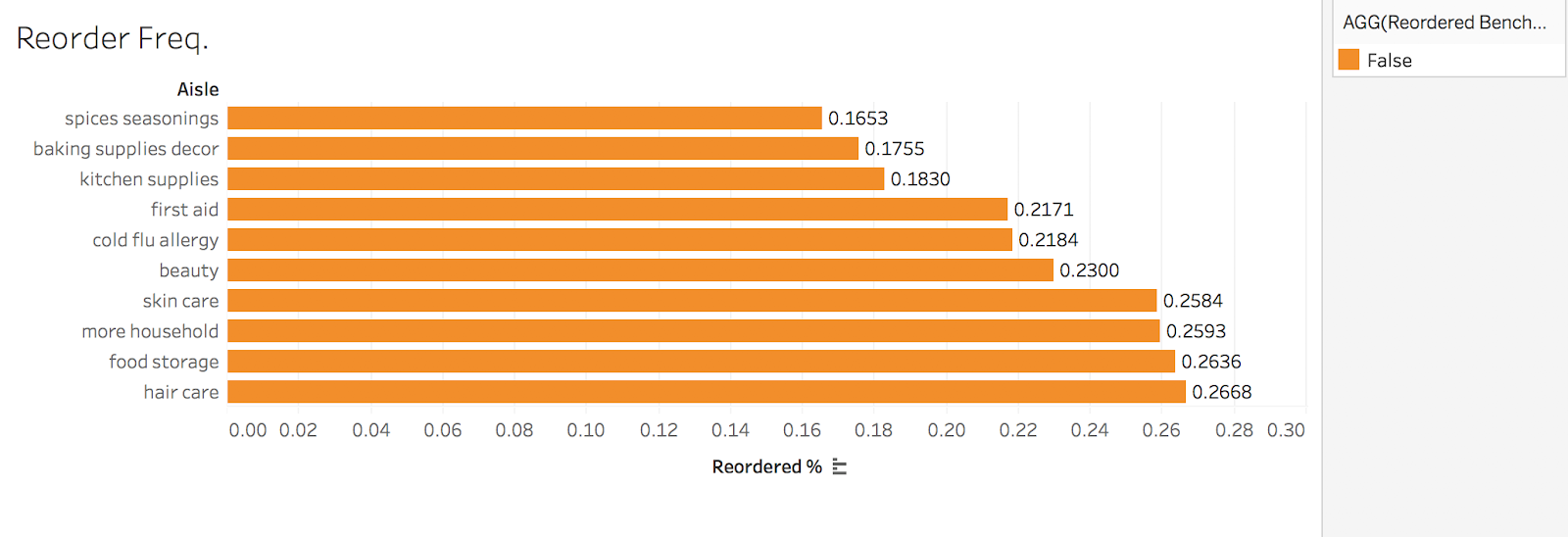
During the course, ACC 652, objects that we have learned are Benford's Law, outlier detection, correlation, and time-series detection, risk assessment and risk scoring, and so on. The main software used in this course is Excel and Tableau. In the final project, we are conducting a market basket analysis on frequently bought items by using Instacart service. The goal of this analysis is to answer 4 main data questions to help Instacart uncover trends to improve their product and service.

Below are the 4 questions

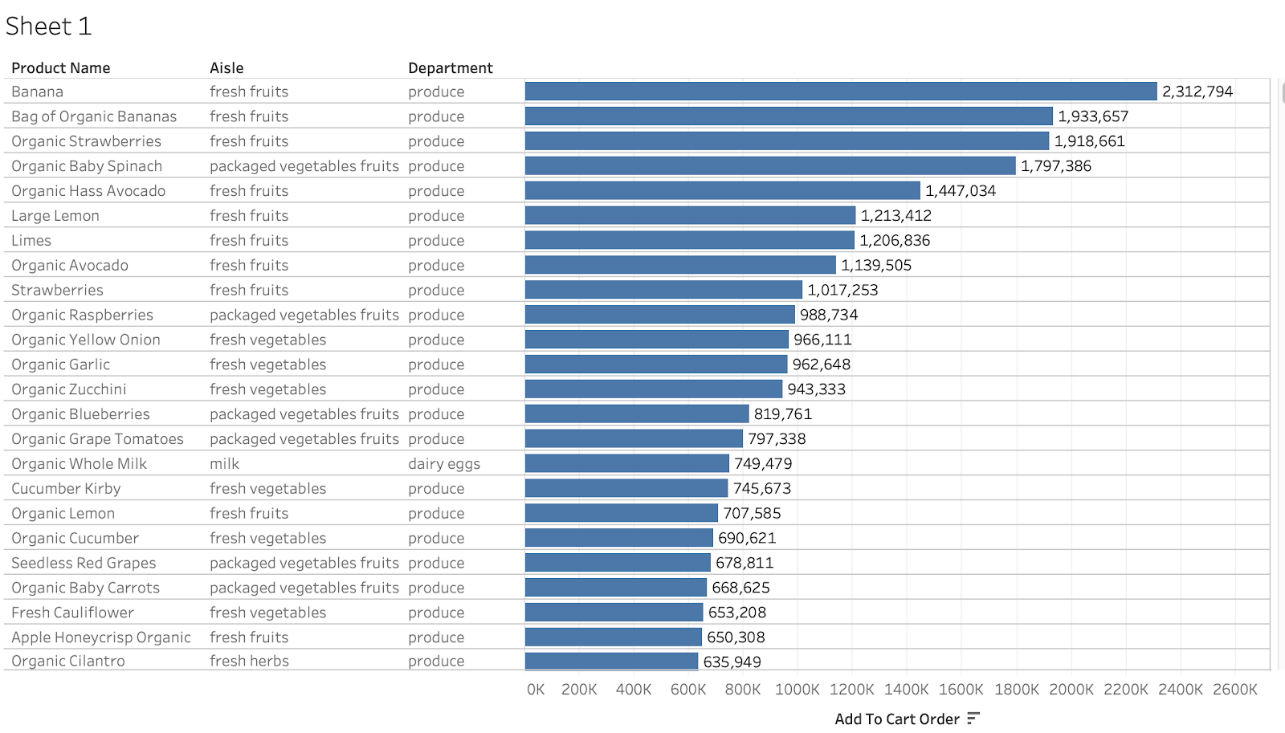
1. what types of products get reordered the highest and lowest percent and what category does it belong to ? Find out reasons why certain products are not as popular as the product that got reordered.
2. Numbers of products that were ordered together. Day and time that customer would most likely to order 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.



Products that got reordered most.



Products category with less reordered orders.



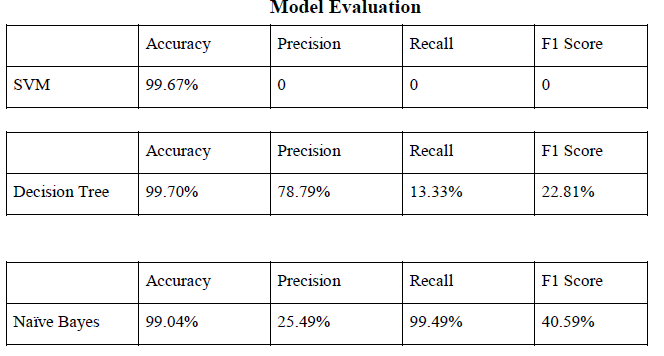
The final project of answering the four questions that we asked, in the beginning, can help Instacart to better serve the customer and potentially help the growth of the business. In the final project, we develop a plan of action that can be used to help business decisions which are mentioned in the learning outcome. This final project also helps with the understanding of what business needs by putting ourselves in the position of Instacart management.

**IST-707 Data Analytics**

During IST 707, we have been introduced to several data mining skills. Such as regression, classification, and cluster. In the final project, our team has used K means cluster analysis and association rule analysis. 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.

Our team start by clean the data. By dropping rows with NA and remove duplicates, we end up with clean data. 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 have to choose Naïve Bayes as the best model in terms of predict shooting crimes.





In the end, we provide our suggestion to the police department that Certain areas and certain times have higher risks of some crimes happening.

**Learning Goals**

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 a required skill for data science. 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.

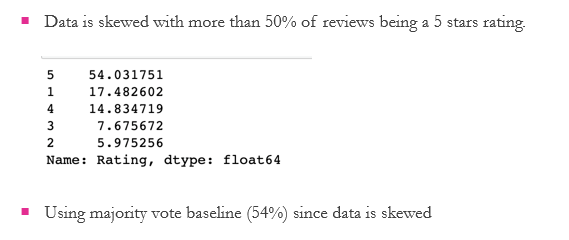
This project has contributed to the successful application to apply different strategies based on different data set, and the various to process data set. By apply data mining combine with data visualization skills, we are able to find out the interesting patterns in the data for better analysis.

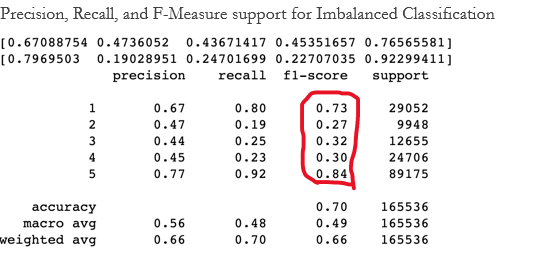
**IST-736 Text mining**

The final project of data mining was to apply several classifications to evaluate the customer review data set. Because reviews are not 100% accurate and the company needs to learn from customer reviews since customer reviews provide valuable feedback from customers. It can help companies improve their product sales. In this final project, the main software used is python.

The project also provides a useful insight into several python libraries and improves my Proficiency in the python language.

In the project, I have mainly used the Multinomial Naive Bayes Model and linear svc model after applying the concepts of Bag-of-words and, TF, TF-IDF to preprocess the text data. Both Bag-of-words and TF-IDF are techniques that help us convert text sentences into numeric vectors. It is essential to understand the concepts here.







This project has provided an opportunity to work with text data since text data are different when compare with other types of data. Text data is usually large and require additional process skills.

Data scientists need to know how to deal with Text data. Because many useful insight can be drawn from text data such as product reviews, Facebook posts, and so on. The ability to organize and analyze large collections of text allows for predictive analytics with text mining to be more efficient.

**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 an opportunity to collect, manage, analyze, and make a conclusion from any all types of data. 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. Thank you.