Syracuse University, School of Information Studies
M.S. Applied Data Science

Portfolio Milestone

Prasad Kulkarni
SUID: 928949766
https://github.com/prasadkulkarni01/SyracuseUniversityDataScience

Table of Contents

1. Introduction

The Applied Data Science program at Syracuse University's School of Information Studies 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 IST 659: Database Administration, IST 707: Data Analytics, IST 718: Big Data Analytics, and MAR 653: Marketing Analytics, reports and presentations were developed to deliver insights using Microsoft Access, SQL Server Management Studio, Python, R, Excel and Tableau. The skills developed at the School of Information Studies furnish data scientists focused in the field of marketing analytics with the ability to generate value within their organizations and produce actionable recommendations.

The Applied Data Science Program has seven learning objectives which were exemplified by the applications in this portfolio:

1. Describe a broad overview of the major practice areas in data science.
2. Collect and organize data.
3. Identify patterns in data via visualization, statistical analysis, and data mining.
4. Develop alternative strategies based on the data.
5. Develop a plan of action to implement the business decisions derived from the analyses.
6. Demonstrate communication skills regarding data and its analysis for relevant professionals in their organization.
7. Synthesize the ethical dimensions of data science practice.

2. IST 659: Database Administration : Healthcare Insurance Database

a. Project Description

Through studying Database Administration under the direction of Prof. John P Stinnett, a Healthcare Database was developed to store paid claims details which can provide ability to systematically access historical information for reporting purposes and risk analysis purposes.

In development and population of the database, the scope of the implementation was to implement database to accommodate: Member data, Provider Data, Episode Data, Claims Data & Payment Data. Conceptual and logical models were developed to organize the relationships between Member, Provider, Episode, Claims, Payment or Disposition (Fig. 1). Tables were created in SQL Server Management Studio while data population was accomplished using Microsoft Access, which also facilitated the exploration of data. Reports and stored procedures were created to display the sum of claim amounts for a family member, claims submitted by provider in state of Texas, high dollar (greater than $1,000) claims. These questions provide valuable insights healthcare spend by families in various states and also provider impacted by claims submission.

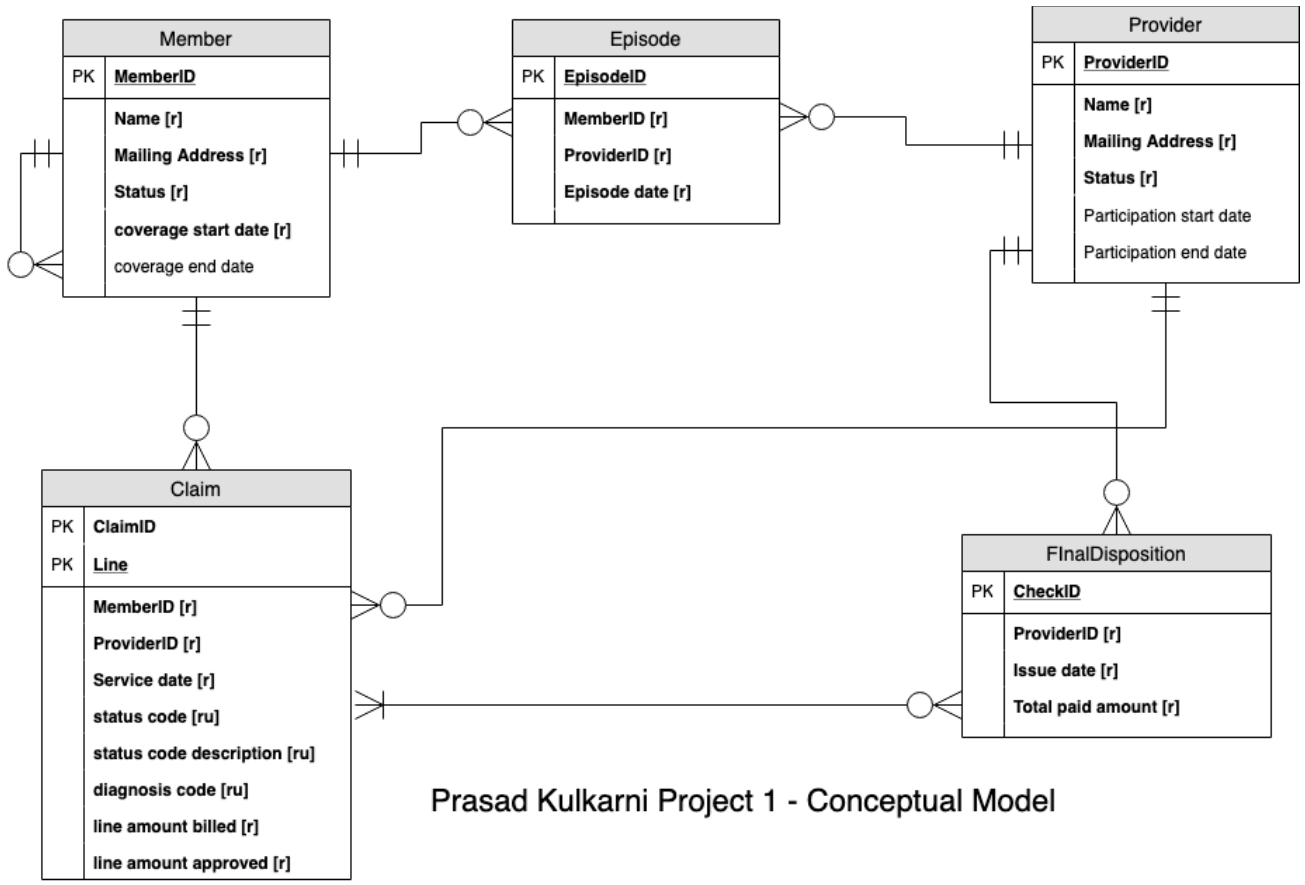

Fig. 1: Conceptual Model.

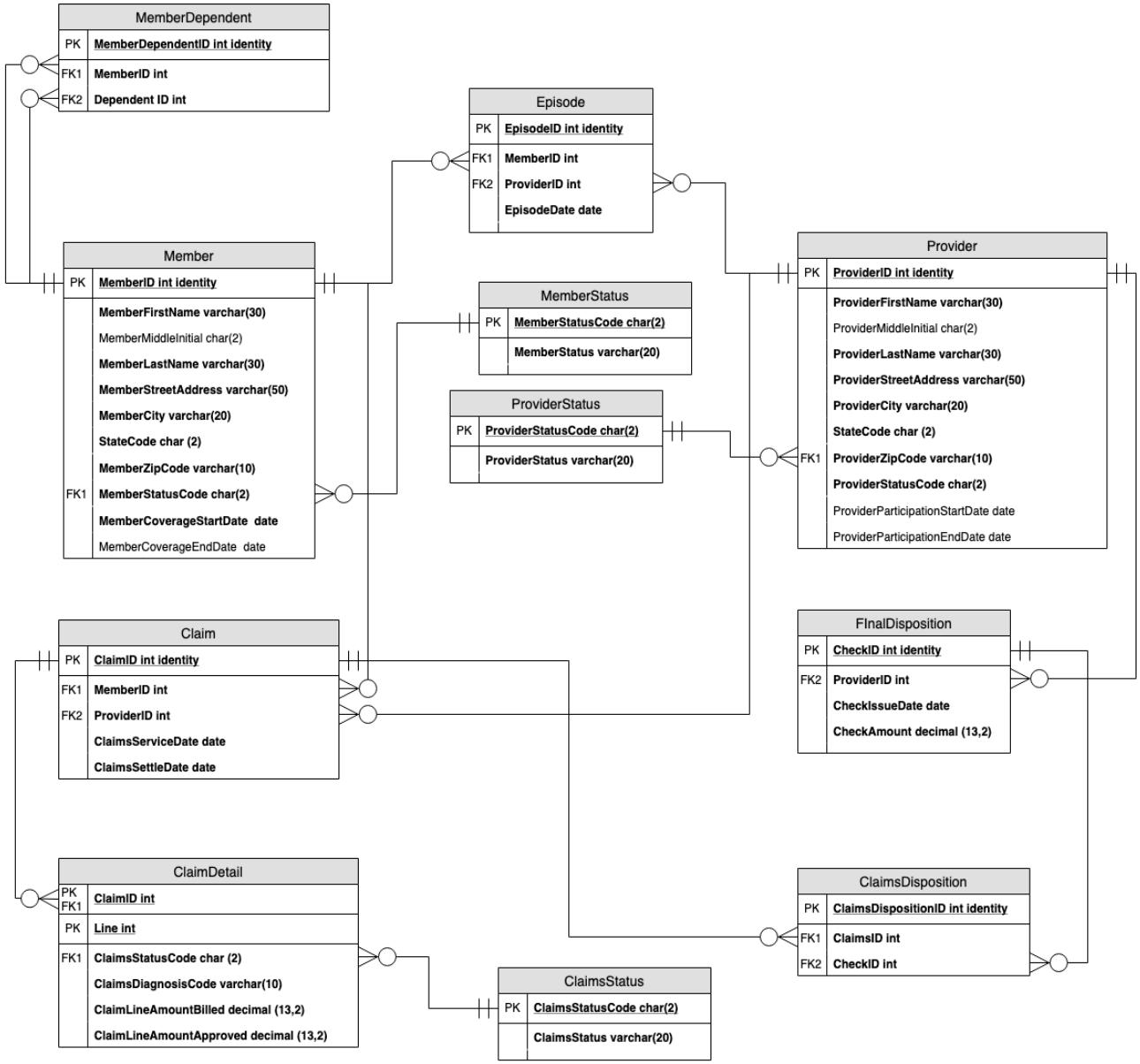

Fig. 2: Normalized Logical Model.

b. Reflection & Learning Goals

After working extensively on this project, I realized that designing, creating and implementing a database requires much more strategic thinking. Consistent focus on simplifying user/application interaction, at the same time achieving necessary data integrity, data security, and concurrent handling through proper designing techniques (such as Normalization, Security measures etc.) are key factors in implementing operational database.

In first part, during initial stage I had thought about capturing patient and provider interactions through “Episodes” table but later realized that it will make current normalization structure between Member, Provider and Claims very complicated. Instead, this can be also be achieved by creating views for specific requirements post implementation.

This project contributed to the successful application of the learning goals through the exercise of collecting and managing data, as well as the identification of patterns using statistical analysis; these observations are leveraged to reveal insights from within the music database which are delivered using reporting tools and are easily understood by relevant professionals.

3. IST 707: Data Analytics : NCAA March Maddness

a. Project Description

Through studying Data Analytics under the direction of Prof Lin, various data mining techniques were introduced which perform with varying precision and efficiency for applications in regression, classification, and clustering. In the final presentation, Random Forest, Support Vector Machine and Decision Tree Classification techniques were implemented to compare which model is capable of correctly predicting the score of the winning team for the 2017 women’s basketball season, an objective that is based on the Kaggle March Madness competition held in February 2017. R Studio is leveraged to conduct analysis within R using data from *Kaggle*, a public data mining resource.

The dataset was obtained from the 2018 Women’s NCAA March Madness Kaggle competition, and is conformed of 9 comma-separated value (CSV) files with a variety of data columns. However, for the purpose of this project, only relevant columns will be used to predict tournament results. Examples of variables that will be used include: season, day season started, day number, winning and losing teams, winning and losing team scores, seeds (for tournament games), number of overtimes played, cities were the game was played, and whether the winning

team won at home, away, or on neutral ground. Initially, we thought the best model would be the SVM, followed closely by the linear regression. It turns out that it was the other way round, with the SVM having a RMSE of 7.96, the linear regression a RMSE of 7.89, and the random forest a RMSE of 7.58.

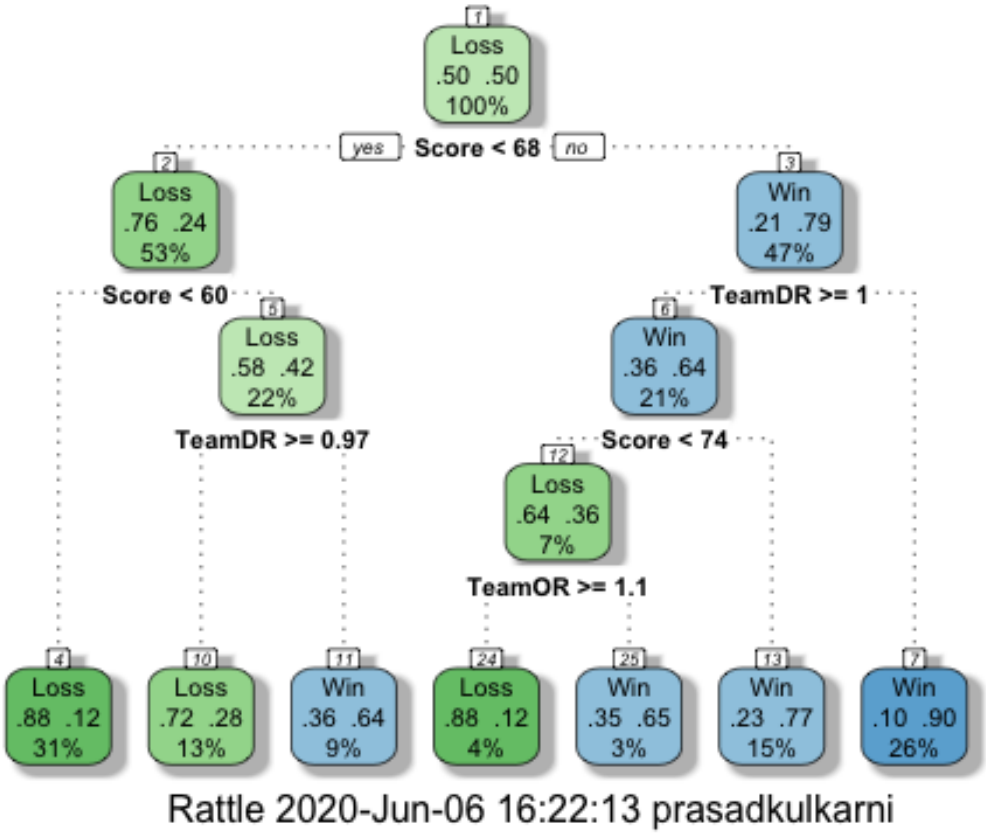

Fig. 3. Decision Tree Model.

b. Reflection & Learning Goals

Though these are just three of the models that we tested, we talked about building more models that could potentially improve the prediction. Models discussed were the popular Kaggle algorithms Generalized Additive Models and eXtreme Gradient Boosting, but due to time constraints we couldn't use them. However, we believe that it would be interesting to revisit these algorithms in the future. Another conclusion that we've drawn is that, even though we have a lot

of data, it would be interesting to see other additions to the dataset. Shooting percentage, fouls per game, number of triples and rebounds, are among the statistics that could improve the model, as they could give us a clearer idea on how a team both attacks and defends. Including more data into the training of the model could increase the accuracy of our results. Finally, of our three models, we were surprised that the random forest model was the most successful of the three.

This project contributed to the successful application of the learning goals through the development of alternative strategies based on the data, and the communication of observations which translate to actionable insights. Data mining was also used in conjunction with visualization to identify patterns in the data for use in classification tasks.

4. IST 718: Big Data Analysis : Google Football a Kaggle competition

a. Project Description

Through studying Big Data Analysis under the direction of Prof. Jon Fox, big data analysis techniques were introduced to win a soccer simulation through an automated agent coded in Python. This is a very important process as it gives us the ability to measure how well could an algorithm perform decisions without a human presence. Specifically, this exercise provided insights into strategies of the world's most-watched sport of soccer. The data consists of observations and actions. Observations are provided between each step and consist of the position and actions of each player on both teams. Actions are the options to simulate player movements. Based on the observations your code needs to provide a valid Action for each step.

We used modelling technique to create an agent which took arguments obs, player_x and player_y and decides a proper action based on set of rules. The rules mainly decide whether to select an action from an offense or a defense or goalkeeper category, based on whether player is in control of the ball or not. In each category, based on other multiple criteria like how far the player is from opponent player and/or how far player is from opponent's goal, we will decide the best

course of offensive action to select for the player. In defense category based on similar criteria we picked the suitable defensive action like to run towards the player with ball.

We tested 3 different types of agents –

1) **Rule based agent:** It used direct coding approach in taking players position and Ball's X and Y coordinate to determine based course of action.
2) **Neural Network:** Neural network model was designed using tensorflow and keras. This model used to train and create an agent which will predict player's next course of action
3) **Reinforcement Learning**: Using Impala distributed Architecture and SEED RL architecture model was generated which can successfully predict player's next move

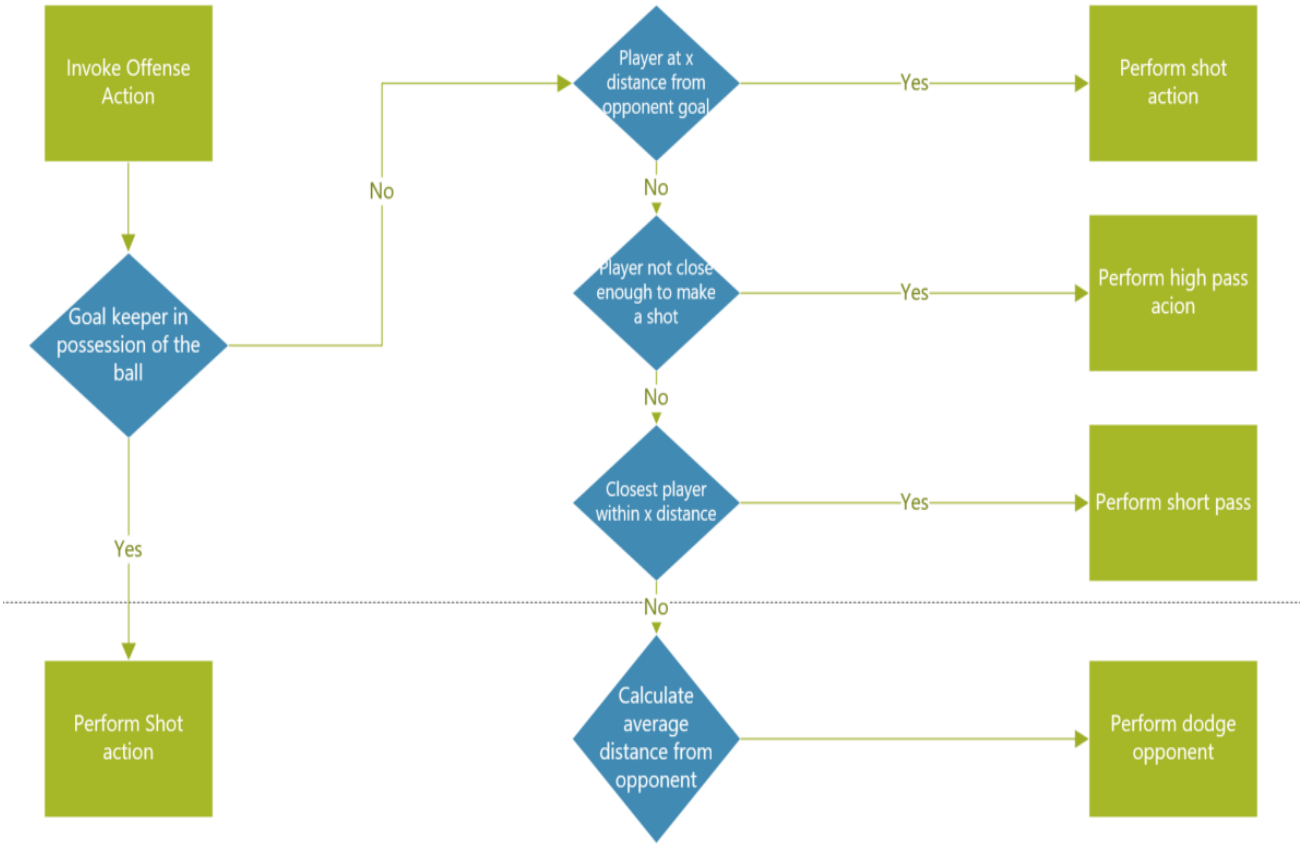

Fig. 4. Rule Based Agent.

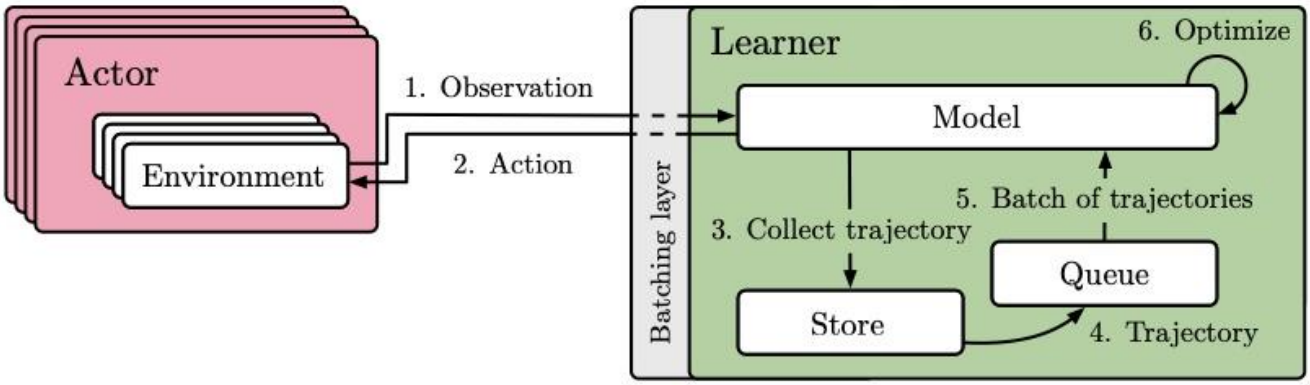

Fig. 5. SEED RL Architecture.

b. Reflection & Learning Goals

After extensive work with designing and creating 3 great models we pitched each of these models against Google's inbuild AI and then against each other to review pros and cons. Although very simple, rule based agent (created with direct coding approach) was more effective in predicting player's next move and performed very well in short duration. Reinforcement learning model needed more time to train the model in order to be able to perform well when against google's inbuilt AI or against rule based model

- Rule based.
    - Easy to reason about and inexpensive compute.
    - Limited agent expertise.
- Neural Network
    - Great potential for agent skill.
    - Lack of framework for learning.
- Reinforcement Learning
    - Closer to how humans learn.
    - More work to be done.
    - Learning from scratch every time..

5. MAR 653: Marketing Analytics

a. Project Description

Through studying Marketing Analytics under the direction of Prof. Rajkumar Venkatesan, data mining concepts specific to marketing were introduced which inspired the final presentation where yearly transaction information for households from a grocery store was used to identify a target group of customers using K-Means Clustering, and subsequently derive an optimal promotional offer using Apriori Rule Association and Sensitivity Analysis for use in a direct mail marketing campaign. The tools required include *Python* for clustering and rule association, while *Excel* was used for exploratory data analysis.

This exercise involved exploring and cleaning the data prior to clustering, by transforming categorical columns into one-hot encoded vectors. Segmentation was completed using the items purchased and pricing data, while profiling was accomplished using demographic information such as income, age, and household size; considerations are made to not cluster customers on demographic data to avoid biases being introduced to the models. The elbow method is revealed an optimal number of clusters of three; the first cluster contained the largest number of potential participants and used coupons the least, the second cluster was contained the fewest potential participants and were the most likely group to use coupons, while the third and selected cluster used coupons marginally more often than the average customer and contained nearly twenty percent of the entire customer base.

Three optimal carts are selected using Apriori Rule Association on the items purchased by customers within the selected group, with coupon promotions applied to two items (Fig. 6). These items are selected such that when items from *cart one* are purchased, items from *cart two* have a discount applied. This method was also successfully implemented by Grazyna Suchacka in predicting purchase behavior in an e-commerce setting; this study similarly selected a target subset

of their customers to improve the expected result (Suchacka & Chodak, 2016). The optimal discount was then calculated using the average discounts applied previously and varying the percent discount to not exceed a $10,000 liability margin (Fig. 7). This resulted in an expected 137% increase in gross revenue for the affected products.

Cluster	Cart1 ->	Cart2	conf	supp	lift	conv
2	drinks, frozen_pizza	meat	0.943	0.085	1.265	4.454
2	baking, food	food_add-ons	0.928	0.217	1.417	4.803
2	dessert, packaged_foods	meat	0.928	0.083	1.244	3.512

Fig. 6. Optimal Apriori Rules.

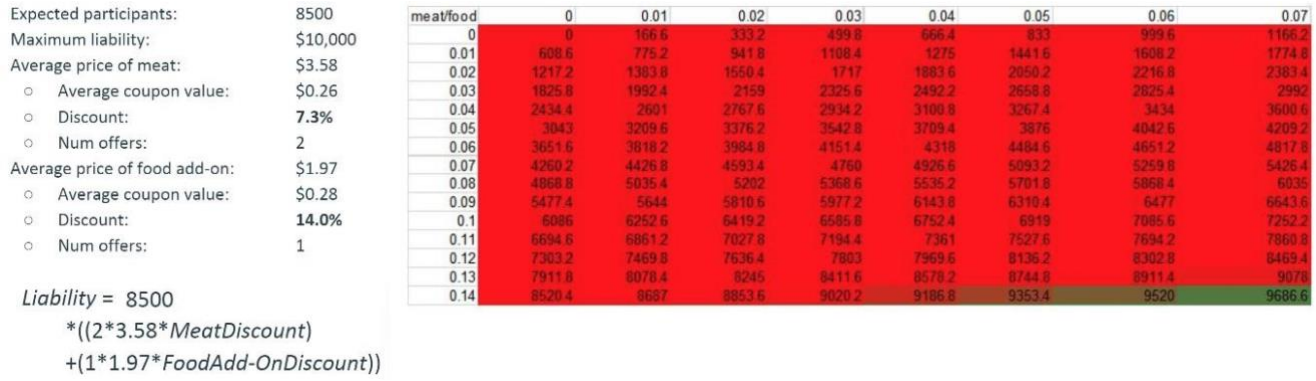

Fig. 7. Calculation of Optimal Coupon Value.

b. Reflection & Learning Goals

This project provided the opportunity to organize and analyze transaction information using data mining techniques, as well as visualization to identify patterns for customer targeting.

It was also necessary to develop a plan of action to quantify the insights developed in this analysis, which translates to measurable and actionable recommendations. Ethical considerations were also necessary to ensure that customer segmentation and profiling was free of bias, using demographic information to profile the previous behavior of a customer, rather than using said information to explain their behavior. This project allowed the data to guide the analysis, requiring alternative strategies to be developed as observations were made within the data.

6. Conclusion

This portfolio has demonstrated the successful implementation of these learning objectives and the major practice areas in data science. Data was collected and managed using web scraping and application programming interfaces in conjunction with database solutions to be analyzed using statistical methods and data mining techniques for tasks such as regression, classification, or clustering ("IST 659,"; "IST 707,"; "IST 718,"; "MAR 653,"). Various data visualizations were paired with clustering techniques to identify patterns which directed the respective analyses; actionable recommendations were developed to reflect tangible business decisions ("IST 718,"; "MAR 653,"). For example, the direct-mail coupon campaign was driven by the outcomes of the clustering and rule association mining, resulting in not only a recommended combination of items, but also the suggested discounts and expected increase in sales from the campaign ("MAR 653,").

Communications skills were developed and displayed in the organization and delivery of insights, expressing them in terms which could be simply understood and acted upon ("IST 659,"; "IST 707,"; "IST 718,"; "MAR 653,"). The ethical dimensions of data science practice were also reinforced in these applications by selecting only relevant data and considering user privacy when analyzing personally identifiable information ("IST 718,"; "MAR 653,").

Similarly, demographic information was excluded from segmentation analyses to avoid introducing bias to the implemented models ("MAR 653,"). These projects are representative of the successful execution of the learning objectives and have developed the necessary skills for practice in the field of data science.

Syracuse University’s School of Information Studies provides students the opportunity to synthesize the collection, management, and analysis of data, as well as the delivery of actionable insights using various data science techniques. Skills learned in the program have developed a multifaceted approach to solving structured and unstructured data problems, it has also cultivated

strategies that improve organizational efficiency. The program has fostered a practice of transparency, reproducibility, and ethical data management which promotes integrity and credibility within an organization's analytics team. Using the methods learned at the School of Information Studies, data scientists are equipped with the ability to tackle a wide range of problems and the resources to explain observations to a variety of stakeholders and business professionals.

References

IST 659: https://github.com/prasadkulkarni01/SyracuseUniversityDataScience/tree/main/IST%20659

IST 707: https://github.com/prasadkulkarni01/SyracuseUniversityDataScience/tree/main/IST%20707

IST 718: https://github.com/prasadkulkarni01/SyracuseUniversityDataScience/tree/main/IST%20718

MAR 653: https://github.com/prasadkulkarni01/SyracuseUniversityDataScience/tree/main/MAR%20653

Bhoi, A. K. (2017). Classification and Clustering of Parkinson's and Healthy Control Gait Dynamics Using LDA and K-means. *Int. J. Bio Automation*, 21(1), 19-30. Retrieved from http://www.biomed.bas.bg/bioautomation/2017/vol_21.1/files/21.1_02.pdf

Huang, J., Lu, J., & Ling, C. (n.d.). Comparing naive Bayes, decision trees, and SVM with AUC and accuracy. *Third IEEE International Conference on Data Mining*. doi:10.1109/icdm.2003.1250975

Pavlinek, M., & Podgorelec, V. (2017). Text classification method based on self-training and LDA topic models. *Expert Systems with Applications*, 80, 83-93. doi:10.1016/j.eswa.2017.03.020