Barbara Schweitzer

SU I.D. - 206499611

Portfolio Milestone

Summer 2021

Applied Data Science Program

Syracuse University iSchool

Table of Contents

[Introduction 3](#_Toc81928376)

[IST 707 - Data Analytics 4](#_Toc81928377)

[*Project Information* 4](#_Toc81928378)

[Figure 1. Tree Diagram from Decision Tree Model (Schweitzer, IST 707, 2021) 5](#_Toc81928379)

[*Program Learning Goals Reflection* 5](#_Toc81928380)

[MAR 653 – Marketing Analytics 6](#_Toc81928381)

[*Project Information* 6](#_Toc81928382)

[Figure 2. Logistic Regression on Demographic Information (Schweitzer, MAR 653, 2021). 7](#_Toc81928383)

[Figure 3. Logistic regression on Survey Questions (Schweitzer, MAR 653, 2021). 8](#_Toc81928384)

[*Program Learning Goals Reflection* 8](#_Toc81928385)

[IST 664 – Natural Language Processing 9](#_Toc81928386)

[*Project Information* 9](#_Toc81928387)

[Figure 4. Word Cloud from Reviews (Schweitzer, IST 664, 2021). 10](#_Toc81928388)

[Figure 5. Filtered, Unfiltered, and Partially Filtered Text Samples (Schweitzer, IST 664, 2021). 11](#_Toc81928389)

[Figure 6. Scores for 3 Naïve Bayes Models (Schweitzer, IST 664, 2021). 11](#_Toc81928390)

[*Program Learning Goals Reflection* 12](#_Toc81928391)

[IST 719 – Information Visualization (in progress) 12](#_Toc81928392)

[*Project Information* 12](#_Toc81928393)

[Figure 7. Final Poster (Schweitzer, IST 719, 2021). 14](#_Toc81928394)

[*Program Learning Goals Reflection* 15](#_Toc81928395)

[Conclusion 15](#_Toc81928396)

[References 17](#_Toc81928397)

# **Introduction**

Upon successful completion of the projects and courses in the Applied Data Science Program, students will learn how to successfully collect, analyze, present, and form plans of action based on data, using many different platforms and techniques. In courses including Data Analytics, Marketing Analytics, Natural Language Processing, and Information Visualization, students can utilize different programs and data sources to develop the skills necessary to become successful data scientists.

The Applied Data Science Program provided by the iSchool at Syracuse University has seven program learning goals that are demonstrated clearly through successful completion of the projects in this portfolio. The seven learning goals are as follows:

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 managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization.
7. Synthesize the ethical dimensions of data science practice (e.g., privacy).

# **IST 707 - Data Analytics**

## *Project Information*

For the final project in the Data Analytics course, a medical data set was chosen to be analyzed. After obtaining breast cancer data, the goal of this project was to create models to try to diagnose breast cancer based on different patient attributes. The original data set had 34 attributes with over 2,500 entries, with any identifying patient information removed before access. After the removal of missing data, information gain was utilized to find the importance level of each attribute, to aid in narrowing down the number of variables used for analysis.

After creating training and testing data sets, six models were created. These models were created using decision tree, Naïve Bayes, and k-Nearest Neighbor algorithms, with varying parameters. In the end, one decision tree model, two Naïve Bayes models, one with Laplace smoothing and the other without, and three k-Nearest Neighbor models, utilizing k values of 5, 15, and 30, were created. After reviewing each model, it was clear that the k-Nearest Neighbor models performed the worst, with low accuracy, precision, and recall scores among the three models. When choosing between the remaining three models, it was necessary to consider the medical nature of the data when deciding the most appropriate model. The decision tree model had the highest accuracy; however, it had a relatively high number of false negatives, which can be very dangerous when diagnosing cancers, and the model, when viewing the tree, was solely based upon variables that were a direct result of the outcome of the patient, as shown below (Figure 1).

Diagram, timeline

Description automatically generated

### Figure 1. Tree Diagram from Decision Tree Model (Schweitzer, IST 707, 2021)

The Naïve Bayes models both had relatively high accuracy and had more sound variable choices, however their precision and recall scores were quite low, and again the medical nature of the data needed to be considered.

In concluding the project, it was necessary to reflect on the less-than-ideal results and make conclusions on what could be done moving forward. These next steps included possibly creating more models utilizing Neural Networks, Support Vector Machines, or Ensemble Learning techniques (Schweitzer, IST 707, 2021).

## *Program Learning Goals Reflection*

Working with the medical data from this project highlighted the importance of proper data collection, as well as the necessity of being flexible when results do not turn out the way they were intended. Before collecting medical data, it is very important to follow the proper laws in removing any sensitive information that may compromise patients. Also, being able to listen to the data and change the analysis based on new discoveries is a key skill in data science.

This project not only helped to introduce one of the major practice areas of specialized data science, but also the importance of data collection and organization, development of new strategies based on data, and ethical restrictions in data science.

# **MAR 653 – Marketing Analytics**

## *Project Information*

For the final project in the Marketing Analytics course, voter survey data was used. In a four-person group, each person was responsible for exploratory data analysis and different analytical tests. The data set contained over 8,300 survey responses, most coming from younger voters, due to the electronic format of the survey. Before beginning the analysis, it was important to create a business question that could potentially be answered from the data exploration. The group found it most helpful to ask how the data could be used to encourage voter participation on college campuses and among younger people through certain marketing efforts. After formulating this business focus for the project, logistic regression was used to show how different categories of variables effected the survey participant’s propensity to vote.

The first category of variable used was demographic information provided by participants. This information included age, education, gender, race, and income level. After running a logistic regression model on this data, gender was removed, due to the insignificance of the variable on voter participation. The final regression is shown below (Figure 2).

Table

Description automatically generated

### Figure 2. Logistic Regression on Demographic Information (Schweitzer, MAR 653, 2021).

It can be concluded that education had the greatest effect on voter participation when comparing the demographic variables. From this regression, the regression equation was used to predict if a person was likely to vote based on their demographic information, and a hit rate was calculated. The hit rate calculation was 77%, therefore, using the given demographic variables, this regression equation could predict voter participation 77% of the time.

The next category of variable used were survey questions answered by participants. The questions were chosen based on an information gain function run to find the importance of each question. The questions all had categorical answers, which were used in the following linear regression (Figure 3).

Table

Description automatically generated

### Figure 3. Logistic regression on Survey Questions (Schweitzer, MAR 653, 2021).

From this model, the regression equation was again used to predict the voter participation, which resulted in a hit rate of 86%. Thus, the questions were more effective in predicting voter participation than the demographic information.

To conclude the project, all group members collaborated on marketing recommendations to help answer the business questions. The recommendations included making more information about political candidates, as well as voting information, more available on college campuses and inviting political candidates to campuses to speak with students (Schweitzer, MAR 653, 2021).

## *Program Learning Goals Reflection*

This project was able to reinforce each of the program learning goals by presenting a business and marketing approach to data science. This project allowed students to collect and research real world data of their choice, privacy restricted, while allowing them to make discoveries within the data and choose how to provide meaningful results. When the data appeared skewed due to the age gap in survey participants, students were encouraged to use this information to change the way it was analyzed and choose a relevant business decision based on the new analysis. When presenting the project, students were urged to present in the same way as someone would to a professional audience of businessmen and businesswomen.

# **IST 664 – Natural Language Processing**

## *Project Information*

For the final project in the Natural Language Processing course, Amazon musical instrument review data was used. The goal of this project was to extract useful information from text reviews to predict the ‘star’ rating a product was given by a consumer. To begin preprocessing the data, the unnecessary variables were removed, and the review text and ‘star’ rating were used for analysis. After the ‘star’ rating was simplified to a ‘good’ or ‘bad’ rating, the review text was analyzed.

To filter the review text, non-alphabetic characters and stop words were removed, and all letters were converted to lower cases. After lemmatizing and tokenizing the text, a word frequency distribution and word cloud were created (Figure 4).

A picture containing text, newspaper

Description automatically generated

### Figure 4. Word Cloud from Reviews (Schweitzer, IST 664, 2021).

To begin the analysis, the data was split by the rating level of ‘good’ or ‘bad’, put into a list, and tokenized. The words were combined into a list and paired with the rating of the review they came from, and finally they were shuffled. After splitting into training and testing sets, an overall number of 16 models were run. The first models were Naïve Bayes models run on completely filtered data, unfiltered data, and partially filtered data. Representations of the data sets are shown below.

Text

Description automatically generatedText, letter

Description automatically generatedText

Description automatically generated

### Figure 5. Filtered, Unfiltered, and Partially Filtered Text Samples (Schweitzer, IST 664, 2021).

For the Naïve Bayes models, none of the scores were very high for the positive reviews, due to the overall lack of existing data available. The scores of the most successful of the three models for each data set are represented below.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Accuracy** | **5-Fold CV Accuracy** | **10-Fold CV Accuracy** | **Precision** | | **Recall** | | **F-Measure** | |
|  | | | | **Pos** | **Neg** | **Pos** | **Neg** | **Pos** | **Neg** |
| **Model 1** | 85.50 | 85.65 | 85.67 | 0.016 | 0.994 | 0.293 | 0.859 | 0.030 | 0.922 |
| **Model 2** | 85.35 | 85.42 | 85.43 | 0.028 | 0.989 | 0.294 | 0.860 | 0.051 | 0.920 |
| **Model 3** | 85.33 | 85.46 | 85.46 | 0.028 | 0.995 | 0.475 | 0.861 | 0.054 | 0.923 |

### Figure 6. Scores for 3 Naïve Bayes Models (Schweitzer, IST 664, 2021).

Next, six Multinomial Naïve Bayes models were created utilizing filtered and unfiltered text data, along with base models, 10-fold cross validation, and 20-fold cross validation. The most successful model was the base Multinomial Naïve Bayes model created with the filtered data. This model had an accuracy of 88.67%, however the precision, recall, and F-measure scores were all zero for the positive reviews. Thus, this model was not very useful.

The final model was a Linear Regression model run on the filtered data. This model had an 89.64% accuracy, which was by far the best model, but the simplicity of this algorithm may not have been sufficient to pick up the subtleties in the text data. Thus, it was concluded that the data was not substantial enough to perform an effective analysis, due to the lack of positive review data (Schweitzer, IST 664, 2021).

## *Program Learning Goals Reflection*

This project allowed for the continued learning of the many varied areas of data science. Learning text processing and natural language processing is an invaluable resource when breaking into the data science field, especially during a social media dominant time such as now. This project encouraged students to collect data and identify important patterns by evaluating frequency distributions and word clouds made from the data. Due to the lack of complete data, other types of analysis needed to be considered to effectively analyze the data.

# **IST 719 – Information Visualization (in progress)**

## *Project Information*

For the final poster project in Information Visualization, the data set chosen was a Lichess.org chess game dataset. This data set included information about the players, their rating scores, number of turns in the game, moves, etc. Visualizations were created to show patterns and important information regarding the data set using both R and Adobe Illustrator. A distribution of the player ratings by the first move was created and combined with a barplot that shows the frequency of which the move was chosen. A distribution of the players ratings based on game piece color was also created. Next, a histogram of the number of turns per game was made, showing that the median number of turns played in the games was 55. A barplot of the type of game endings is shown, with a resignation being the most common. Finally, a scatterplot was added to show the correlation in rating matches in each game. Adobe Illustrator was used to alter the graphics to make them more detailed and more visually appealing. This project will be presented during a poster session and is intended to give students a chance to create professional level graphics to show a greater understanding of data visualization.

Timeline

Description automatically generated

### Figure 7. Final Poster (Schweitzer, IST 719, 2021).

## *Program Learning Goals Reflection*

This project gives the opportunity for further achievement of the program learning goals. From searching multiple different sources for data, to creating numerous different visualizations for the data, this project gives a chance to experience real world data presentation. Questions based on the data were formulated to give direction to the data analysis, allowing for more successful communication of the important trends in the data. By creating professional level graphics and poster layouts and presenting the final product, students will be able to demonstrate communication skills with fellow students and professors.

# **Conclusion**

Through the completion of these projects, in conjunction with the other assignments completed in the courses, the program learning goals have been successfully implemented, along with an understanding of the major practice areas of data science. Data collection was thoroughly practiced in each case, with different methods explored in each course. Using statistical testing, such as linear regressions, alongside visual representations of data and data mining techniques, a deeper understanding of data and the differences between each type of data was developed. By using different data sources for each project, an ability to rapidly change analysis was required to successfully follow the data and its story.

In multiple courses, such as Marketing Analytics and Information Visualization, along with others taken in the program, it was necessary to create business questions and to find ways to answer those questions to give business recommendations based on findings from the data. This newfound skill allowed students to develop methods of communicating the important information from data and its analysis.

By allowing students to explore data and its multiple sources and practice areas, skills were gained in understanding and recognizing the ethical implementations of data and its collection. Upon successful completion of the projects and courses in the Applied Data Science Program, students have learned how to successfully collect, analyze, present, and form plans of action based on data. Using these methods taught by Syracuse University, students can learn and develop their own skills to successfully implement in the world of data science.

# **References**

Schweitzer, B.J. (2021) IST 707: Data Analytics. Retrieved from

<https://github.com/bschwei/Syracuse_Portfolio_Milestone/tree/main/IST%20707%20-%20Data%20Analytics>

Schweitzer, B.J. (2021) MAR 653: Marketing Analytics. Retrieved from

<https://github.com/bschwei/Syracuse_Portfolio_Milestone/tree/main/MAR%20653%20-%20Marketing%20Analytics>

Schweitzer, B.J. (2021) IST 664: Natural Language Processing. Retrieved from

<https://github.com/bschwei/Syracuse_Portfolio_Milestone/tree/main/IST%20664%20-%20Natural%20Language%20Processing>

Schweitzer, B.J. (2021) IST 719: Information Visualization. Retrieved from

<https://github.com/bschwei/Syracuse_Portfolio_Milestone/tree/main/IST%20719%20-%20Information%20Visualization>