Syracuse University – School of Information
M.S. Applied Data Science

Portfolio Milestone

James Y. Eakins
SUID 586974381
https://github.com/jyeakins/MSADS_Portfolio

Table of Contents

1. Introduction ---------- 3
2. IST 659 ---------- 4
    a. Project Description ---------- 4
    b. Reflection & Learning Goal ---------- 6
3. IST 687 ---------- 7
    a. Project Description ---------- 7
    b. Reflection & Learning Goal ---------- 9
4. IST 707 ---------- 10
    a. Project Description ---------- 10
    b. Reflection & Learning Goal ---------- 13
5. IST 736 ---------- 14
    a. Project Description ---------- 14
    b. Reflection & Learning Goal ---------- 17
6. Conclusion ---------- 18
7. Reference ---------- 21

1. Introduction

Over the courses of Data Science Program, the objectives of the program are to learn how to use the techniques and adapt new innovative tools to solve problems by using the data. iSchool provides courses such as Data Admin Concepts and Database management (Eakins, “IST659”, 2019), Introduction to Data Science (Eakins, “IST687”, 2019), Data Analytics (Eakins, “IST707”, 2020) and Text Mining (Eakins, “IST738”, 2020) to develop proper skills to understand the problem, learn how to handle data for analysis and provide insightful conclusion by learning various tools (SQL, R, Python, Tableau, etc.) and techniques.

The Applied Data Science program has four major learning objectives which this portfolio met each criterion;

1. Data collection: using tools to collect and organize data
2. Data analysis: Identify patterns in the data via visualization, statistical analysis, and data mining
3. Strategy and decision: develop alternative strategies based on data
4. Implementation: develop a plan of action to implement the business decisions.

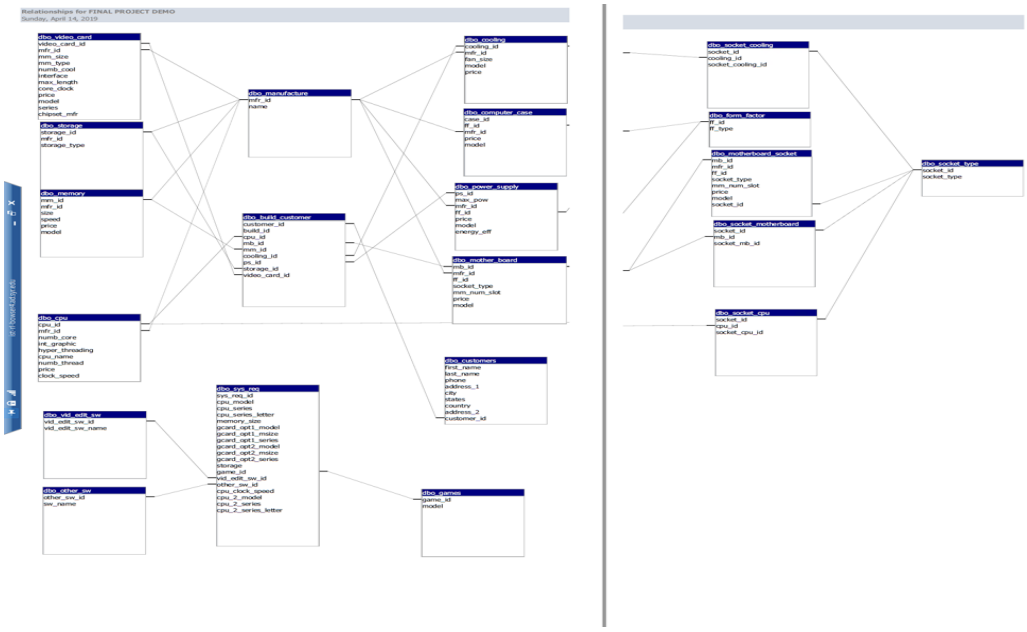
2. IST 659 – Data Admin Concepts & Database Management

    a. Project Description

Through the course of Database management, the database was built to give the user a list of parts, compatibility among the parts, and the total cost of the build. It also was to give the list of components to run any software that people want to use for their productivity or entertainment. The scope of the project is to give right tools to people who wants to build their own computer by only utilizing the major components that are popular among the computer enthusiasts community without wasting their resources, including, time, money and buying unnecessary parts that may not work with each other (Eakins, 'IST659', 2019); the application required to collect over hundreds of parts information with details which will be used to indicate whether the parts will work with one or another. Since not all components were not collected within the given time, there is an interface where user can enter the components as they want with details that are required.

The entity-relationship diagram was built from Microsoft Visio to show relationships between tables (Fig.1). There are total of 8 major tables that directly handles the items such as the computer components, software, brand of each items and etc. There are 13 more tables that handles linkages and relations of the

tables. The 13 tables take care of the compatibility issues between items and software requirements to give correct components to build for the users. The tables were made from SQL Server Management Studio and the Microsoft Access was used to store the data and create the interfaces for users to interact.



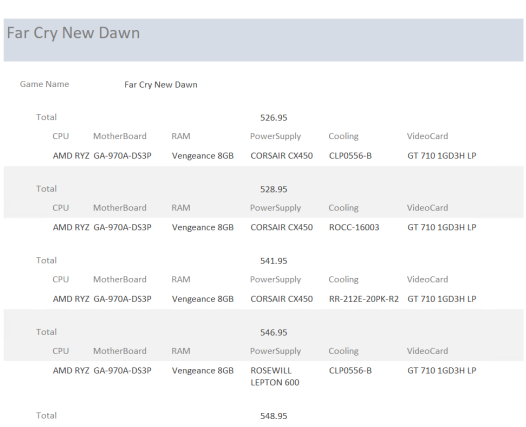
(Fig.1)

BUILD_FOR_GAME_1

GAME NAME Player Unknowns Battle Ground

	CPU	MOTHER BOARD	RAM	POWER SUPPLY	CPU COOLING	VIDEO CARD
Total 300.15	AMD FX-(	GA-970A-DS3P	Vengeance 8GB	CORSAIR CX450	CLP0556-B	GT 710 1GD3H LP
Total 302.15	AMD FX-(	GA-970A-DS3P	Vengeance 8GB	CORSAIR CX450	ROCC-16003	GT 710 1GD3H LP
Total 315.15	AMD FX-(	GA-970A-DS3P	Vengeance 8GB	CORSAIR CX450	RR-212E-20PK-R2	GT 710 1GD3H LP
Total 320.15	AMD FX-(	GA-970A-DS3P	Vengeance 8GB	ROSEWILL LEPTON 600	CLP0556-B	GT 710 1GD3H LP
Total 322.15						

(Fig.2)



Far Cry New Dawn

Game Name Far Cry New Dawn

	CPU	MotherBoard	RAM	PowerSupply	Cooling	VideoCard
Total 526.95	AMD RYZ	GA-970A-DS3P	Vengeance 8GB	CORSAIR CX450	CLP0556-B	GT 710 1GD3H LP
Total 528.95	AMD RYZ	GA-970A-DS3P	Vengeance 8GB	CORSAIR CX450	ROCC-16003	GT 710 1GD3H LP
Total 541.95	AMD RYZ	GA-970A-DS3P	Vengeance 8GB	CORSAIR CX450	RR-212E-20PK-R2	GT 710 1GD3H LP
Total 546.95	AMD RYZ	GA-970A-DS3P	Vengeance 8GB	ROSEWILL LEPTON 600	CLP0556-B	GT 710 1GD3H LP
Total 548.95						

(Fig.3)

Fig.2 and fig.3 show the reports after user used the database to find the correct components for the build. The reports show the name of all the components and the total price to build. Fig.3 is showing the software specific, for this case it was for the game called ‘Far Cry New Dawn’.

b. Reflection & Learning Goals

The development of database management solution made possible to think like an engineer and analyze the problem in many different ways. It allowed to think about how to create the interface, how to store data, how to link the tables and all the little pieces of processes to make this project successful. By learning how to utilize the tools that were involved in this project, this project contributed in more advanced courses in the later term of Applied Data Science program to explore many different aspects such as security, database architecture and how to cross use the tools in different fields.

To any Data Science students, this project allows students to sketch the blueprint of the database structure, design the interface for users and build the database to make the raw data can be useful tool. Learning goal is to think as an engineer, think as a user and think as a database administrator which can lead to advanced thinking and understand not only basis of database management but also advanced skills of database management.

3. IST 687 – Introduction to Data Science

    a. Project Description

Through studying of the Introduction to Data Science under Dr. Stanton, new data science techniques were introduced. How to clean the data to analyze and visualize the analysis were taught to solve the problems that involves the big data. For the final project, the Linear Regression Model, the Correlation Matrix, the Association Rule mining, the Support Vector Machine, the Apriori Rule Association, the Word Cloud and the visualization used to analyze the data, compare each airline satisfactory survey and give insights to how to increase the satisfactory of other airlines based on the results. R Studio is used for this project, and it was first introduction to the software and was very helpful to analyze such a big amount of data.

The data was given by the professor since the collection of data was not part of the course. However, the data cleaning process was the major component of this project since there are many rows and columns that may cause the faulty results on analysis. The names of columns also cleaned which contains spaces and dots (Eakins, 'IST687', 2019).

For the data analysis, it began with the summary of the data and visualize data in bar charts. Among the 14 airlines, 1 airline showed the highest satisfactory

survey result. The correlation matrix between the columns of the dataset taken from the survey which affects the satisfaction. The sublimation of colors on the extreme right in a bar format gives us the basic idea of the statistical values ranging from -1 to 1 along with the size of the dots in the matrix and the color (Eakins, 'IST687', 2019). The Linear Regression model showed the independent and dependent variables to understand the positive or negative impact on the satisfactory of the customers. The level of p-value (Fig.4) explains whether the variable has the significance to the model or not. Furthermore, the linear regression model provides the adjusted R-Squared value (Fig.5) to understand the quality of the model.

```
Residual standard error: 0.7175 on 193607 degrees of freedom
Multiple R-squared:  0.4481,    Adjusted R-squared:  0.446
F-statistic: 219.8 on 715 and 193607 DF,  p-value: < 2.2e-16
```

(Fig.4)

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    1.7764272  0.0108206 164.171  < 2e-16 ***
GenderMale                     0.0627233  0.0031281  20.052  < 2e-16 ***
Type.of.TravelMileage tickets -0.1883424  0.0057737 -32.621  < 2e-16 ***
Type.of.TravelPersonal Travel  0.2932543  0.0033956  86.362  < 2e-16 ***
ClassEco                      -0.0100541  0.0055682  -1.806 0.070981 .
ClassEco Plus                 -0.0246883  0.0071600  -3.448 0.000565 ***
Scheduled.Departure.HourHigh  -0.0083998  0.0042312  -1.985 0.047125 *
Scheduled.Departure.HourLow   -0.0299937  0.0042348  -7.083 1.42e-12 ***
Departure.Delay.in.MinutesHigh -0.0003173  0.0066891  -0.047 0.962164
Departure.Delay.in.MinutesLow -0.0049205  0.0061504  -0.800 0.423694
Arrival.Delay.in.MinutesHigh   0.0312066  0.0102470   3.045 0.002324 **
Arrival.Delay.in.MinutesLow    0.0074016  0.0066770   1.109 0.267638
Flight.cancelledYes           -0.1366458  0.0151053  -9.046  < 2e-16 ***
Flight.time.in.minutesHigh     0.0048322  0.0066630   0.725 0.468312
Flight.time.in.minutesLow     -0.0185412  0.0070850  -2.617 0.008873 **
Flight.DistanceHigh            0.0032167  0.0067459   0.477 0.633483
Flight.DistanceLow             0.0133225  0.0068991   1.931 0.053479 .
Arrival.Delay.greater.5.Minsyes 0.1581237  0.0086641  18.250  < 2e-16 ***
```

(Fig.5)

After determining the key drivers have the significance to the satisfaction, the Association Rule Mining was used to find the co-occurrences of variables. The goal of this analysis was to find a combination of variables that inevitably lead to high or low customer satisfaction (Fig.6). SVM is used to build a classification model with the customer satisfaction as the target variables by showing the confusion matrixes of the results (Fig.7)

```
> inspect(ruleset)
    lhs                                                    rhs                                      support confidence      lift count
[1] {CheapseatsAirlineDF.Flight.cancelled=No,
     CheapseatsAirlineDF.Arrival.Delay.greater.5.Mins=no} => {CheapseatsAirlineDF.Satisfaction=High} 0.3236991  0.5627790 1.1279952 12734
[2] {CheapseatsAirlineDF.Arrival.Delay.greater.5.Mins=no} => {CheapseatsAirlineDF.Satisfaction=High} 0.3278172  0.5577855 1.1179866 12896
[3] {CheapseatsAirlineDF.Flight.cancelled=No}             => {CheapseatsAirlineDF.Satisfaction=High} 0.4948016  0.5010812 1.0043325 19465
[4] {}                                                    => {CheapseatsAirlineDF.Satisfaction=High} 0.4989196  0.4989196 1.0000000 19627
[5] {CheapseatsAirlineDF.Class=Eco,
     CheapseatsAirlineDF.Flight.cancelled=No}             => {CheapseatsAirlineDF.Satisfaction=High} 0.3983324  0.4964359 0.9950218 15670
[6] {CheapseatsAirlineDF.Class=Eco}                       => {CheapseatsAirlineDF.Satisfaction=High} 0.4021709  0.4944372 0.9910156 15821
>
```

(fig.6)

```
testData.SatHuH      0     1
        Happy    10206   103
        unHappy   1426  1378
```

(Fig.7)

b. Reflection & Learning Goals

This project was the best exercise which provided the lesson that the cleaning procedure before analyzing the big data is the most important step in Data Science. Without the clean data, none of the analysis or modeling was useful since it will give faulty results. By using multiple analysis tools, it gave the opportunity to observe the satisfactory of airline customers and create the different models that lead to an insight to the bad satisfactory airlines. In the final

term of Applied Data Science program, these basic tools were used to clean and understand bigger and complex data and apply the complicated analysis.

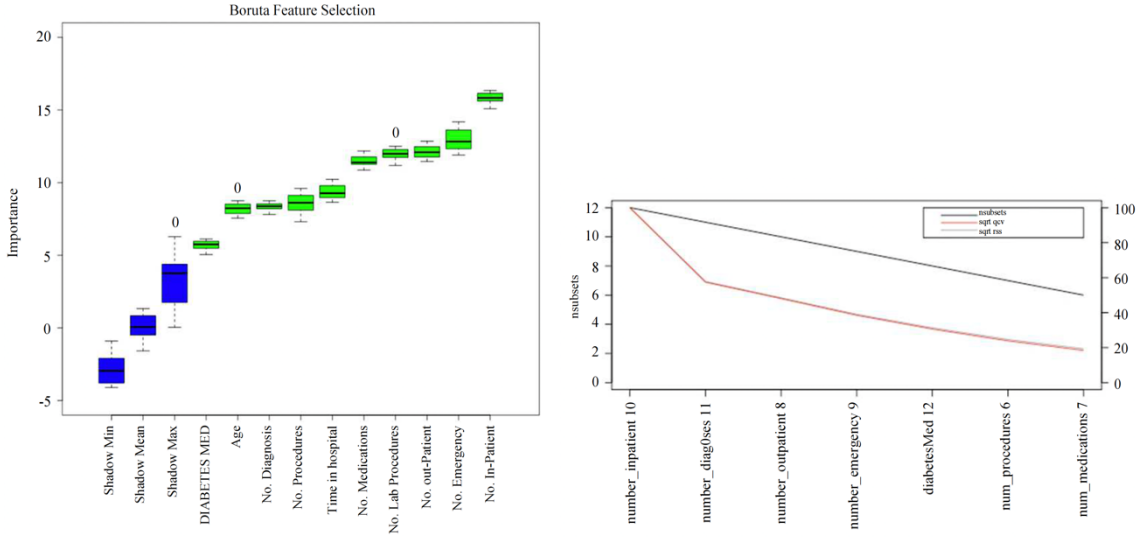
Understanding the data is incredibly important when it comes to Data Science. The data scientists can make all the cool analysis results and visualization to the customers when they have deeper understanding of what they are dealing with. As the data get complicated with texts, numbers and other types, clean and well modeled data will allow the data scientists to do complex analysis and provide better results to customers.

4. IST 707 – Data Analytics

    a. Project Description

The project for the Data Analytics course under the direction of Dr. Gates was more advanced project that involves with the mining of data, more advanced analysis tools such as regression, classification and clustering, and interpretation of results to make the story out of the numbers. In the Final presentation, there were four analysis techniques implemented to give insightful results about re-admission rate of hospital patients. The study provides with an efficient prediction model that can be deployed to a clinical scenario and help healthcare units to be prepared for the unavoidable re-admissions and provide alternative care to preventable re-admissions (Eakins, “IST707”, 2020).

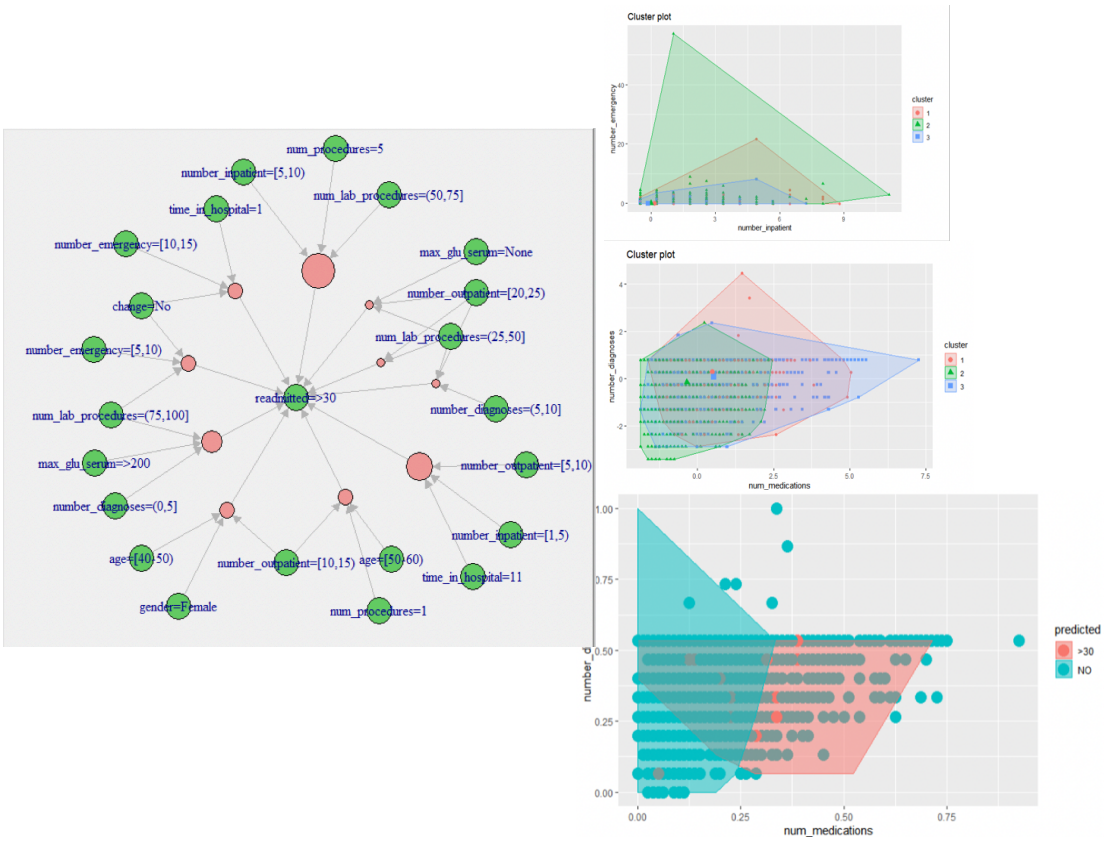
The data required cleaning and selecting features to eliminate unnecessary variables and focus on a few specific targets. To do so, multiple of visualization techniques used to show the significance of the variables (fig.8). The concern raised due to data size is termed as the minimal-optimal problem (Nilsson, 2007). This study employs Boruta algorithm and stepwise regression to determine the best features within the dataset (Fig.8).



(fig.8)

Once the features were selected, the project utilized all four techniques to see different outcomes of the analysis. However, Support vector machine and Random forest did not have good results. SVM generated an accuracy of about sixty-seven percent where are random forest was found to highly over-fitting (Due to the size of dataset). Association rules and data visualization (Fig.9) was found to be the two most useful methods to understand the underlying factors

causing hospital re-admissions (Eakins, “IST707”, 2020).



(fig.9)

The model predicted higher numbers for not readmitted with higher probabilities on number of diagnoses and number of medications. Those two variables are the significant variables to predict the result of patients. Fig.10 shows the classification matrix of the model.

	TestSet_labels		
readmission_Prediction	<30	>30	NO
<30	58	91	39
>30	72	294	168
NO	443	1425	2498

(Fig.10)

This study implements a predictive analytical approach to identify patients prone to readmission and thus, systematically reduce the number of avoidable re-admissions mainly caused by patient non-compliance to medication instruction or early discharge from hospital. The novelty of this method is to directly incorporate patients' history of re-admissions into modeling framework along with other demographic and clinical characteristics. This project also verifies the effective-ness of the proposed approach by validating training accuracy. Some contributions made in this paper are applying Boruta algorithm and stepwise variable selection and implementing genetic and greedy ensemble algorithm to optimize the predictive models (Eakins, "IST707", 2020).

b. Reflection & Learning Goals

Understandably, there are instances where techniques or tools sometimes cause unpredicted results as SVM and Random Forest shown from this project. It is always important to understand how to utilize other tools to analyze and predict the model when some of tools not showing the results as expected. Also, understanding how to use visualization in the big data project is essential to the data scientist. The visualization is a great tool to show the results of the analysis, but it also helps to pick the key features, or see where to focus on when it comes to large datasets.

This project successfully met all four learning goals where it required students to collect the data from outside of sources, analyze the data, find different strategies and decide which tools to use when there are obstacles in the way, and finally provide insightful conclusion to readers to implement the solutions to the problems which can be found in the real world.

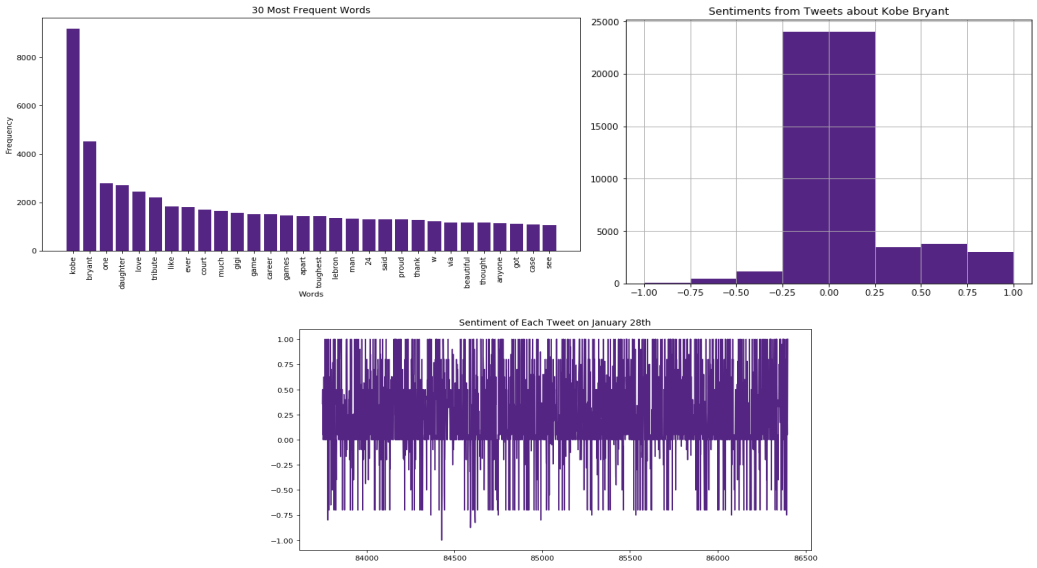
5. IST 736 – Text Mining

    a. Project Description

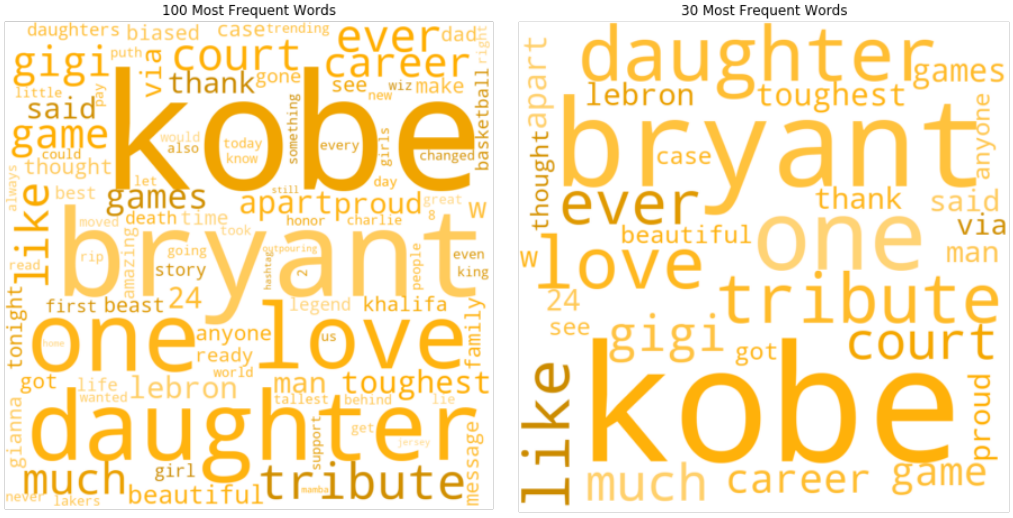
The project for IST 736 course was launched under direction of Dr. Gates, and Text mining from tweets, analyze the sentiments and text analysis techniques were introduced to understand how to treat the unstructured data from the scratch. In the final presentation, the tweets about recent tragedy of Kobe Bryant was collected to analyze public sentiments toward Kobe Bryan and his daughter, Gigi. With the history of Kobe Bryan's unpleasant incidents and the status as heroic player at Los Angeles Lakers, this study was to give insightful recommendation to commercial companies on whether they should feature Kobe Bryant for the commercial campaign.

This project required immediate data collection from Twitter since the limitation on how far the user can collect the data from the past. By using the Python's Tweepy package, 35804 tweets were collected from the public. Once the

data was saved as ‘csv’ file, the data was cleansed to have only raw text from the tweets since the emojis and other non-English characters can ruin the analysis. The first exploratory analysis implemented the statistical analysis and the Word Cloud to show that there are more positive sentiment tweets were found than the negative comments (Fig.11) and the topics that were mentioned the most among the Tweets (Fig.12).

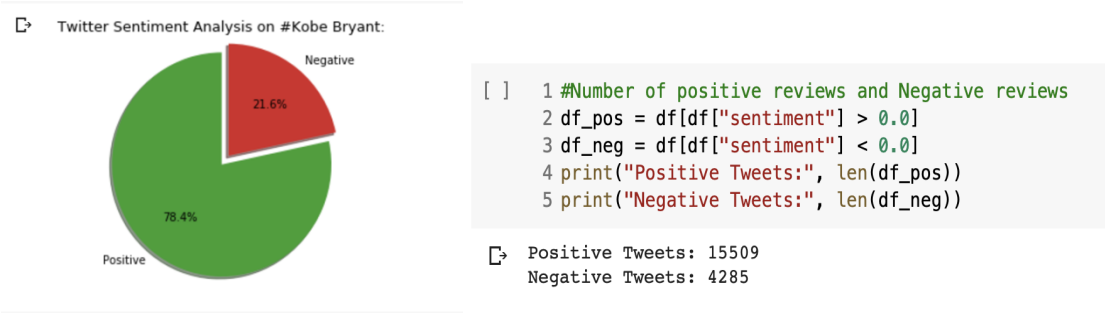


(Fig.11)

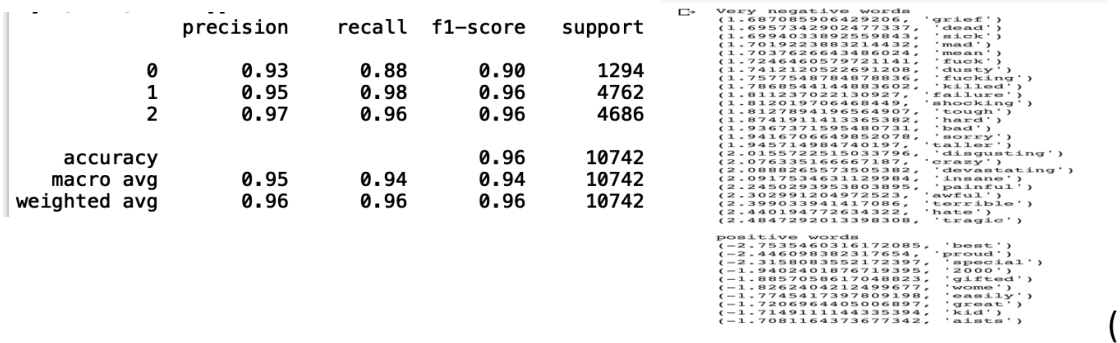


(Fig.12)

After the first exploratory analysis, the sentiment analysis (Fig.13) showed the statistics of sentiments of tweets with more accurate results, and the SVM (Fig.14) was used to predict based on training data, whether or not a tweet had a negative or positive or neutral sentiment. It also categorized and sorted the words that had an extremely negative sentiment as well as very positive sentiment. Lastly, in order to identify the possible location for campaigns around the country, this study has identified the tweets location with positive sentiments (Fig.15) (Eakins, "IST736", 2020).

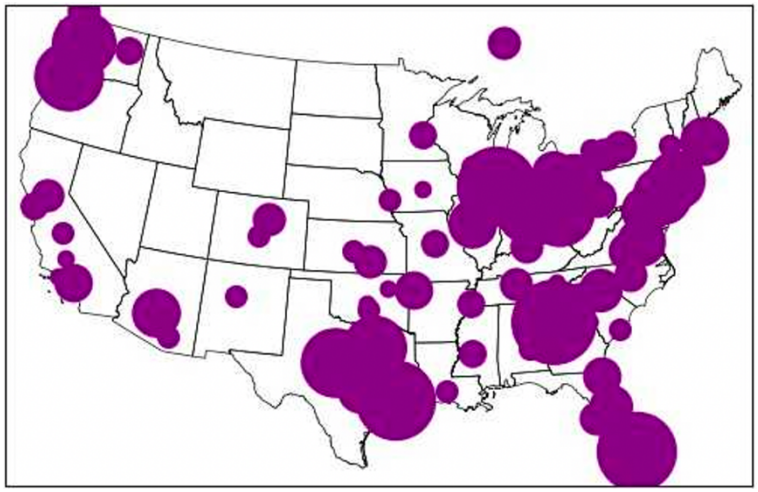


(Fig.13)



	precision	recall	f1-score	support
0	0.93	0.88	0.90	1294
1	0.95	0.98	0.96	4762
2	0.97	0.96	0.96	4686
accuracy			0.96	10742
macro avg	0.95	0.94	0.94	10742
weighted avg	0.96	0.96	0.96	10742

(Fig.14)

 (Fig.15)

b. Reflection & Learning Goals

In the beginning phase of this project, there were opportunities to challenge the knowledge to achieve the data that it required. From collecting the data by writing scripts in Python to analyzing the data by choosing proper tools and techniques, it reminded that how challenge a project can be if the instruction or guidelines were not existed. Also, it was great opportunity to learn various private policies and how data need to be handled if the data was collected from the public available source. One thing that brought the attention was, even if it is a public source, the data does not represent the whole population which could lead to false assumption or conclusion.

There are many types of data, such as numbers, pictures, texts and so on. But the text data can be analyzed in both quantity and quality. The limitation of the source is unlimited since it can be anything from the comments on yelp to the

novels. The recent introduction of personal assistants like, Siri or Amazon Alexa, they were developed from the small text mining to understand the Human language and provide the answers to improve day-to-day life.

This project was combination of all learning goals and added more challenges to the objectives; The data collection process required the automation to increase the productivity of the collections, the analysis of the text data increased the complexity of data cleaning and utilization of tools, the strategy and decision making to find which technique to use and what to focus, and finally, the insightful conclusion and communication skill to give better implementation to business were definitely exceed expectation of the learning goals.

6. Conclusion

School of Information of Studies at Syracuse University has four learning objectives for its students; Data collection, Data analysis, Strategy and decision and Implementation. This portfolio is the best example of the successful student who acquired all learning objectives throughout the courses by completing the challenging projects. The data collection was performed in many different ways such as writing a script to automate the process or download from a website, the database created, stored the data and managed to give tools for users, the collected data was analyzed by using data analysis and statistical tools and

techniques such as Regression, Support Vector Machine, Random Forest, Classification, Clustering, or Bernoulli. There were charts, graphs, and other visualization accompanied to show the analysis in a graphical way to represents the results from the analysis (Eakins, “IST659” “IST687” “IST707” “IST736”). Furthermore, the strategy and decision-making process were implemented along the way of the project to execute the project in the right direction and conclude the insightful implementation.

The presentations of the project helped to learn how to compose slides that easy to read for people who does not involved in the project. The communication skills were also developed to use easy terms and give deliverables in a non-technical manner so that others can understand what the problem was to solve and how to implement the solution (Eakins, “IST687” “IST707” “IST736”). Lastly, there were many other aspects learned throughout the projects, such as how to manage the timeline, how to be a team player, what to consider when collect the data, and etc.

From Applied Data Science program, the learning objectives challenged students to adapt new ways to view the issues around the world. By having perspective of data scientists and the methods to solve the problem, it gives the opportunity to orchestrate the skills that learned from the program; the data

collection, analysis, strategy and make decisions and implementation. The data scientists from Syracuse University will have proper skills and techniques to solve a variety of problems in the organization.

7. References

Eakins, J. Y. (2019) IST 659: Data Admin Concepts & Database Management. Retrieved from https://github.com/jyeakins/MSADS_Portfolio/tree/master/IST659

Eakins, J. Y. (2019) IST 687: Introduction to Data Science. Retrieved from https://github.com/jyeakins/MSADS_Portfolio/tree/master/IST687

Eakins, J. Y. (2020) IST 707: Data Analytics https://github.com/jyeakins/MSADS_Portfolio/tree/master/IST707

Eakins, J. Y. (2020) IST 736: Text Mining https://github.com/jyeakins/MSADS_Portfolio/tree/master/IST736

Nilsson, R., 2007. Consistent feature selection for pattern recognition in polynomial time. J. Mach. Learn. Res.