

MASTER OF SCIENCE APPLIED DATA SCIENCE

PROJECT PORTFOLIO MILESTONE

Debra A. Kernstock

SUID: 392592182

dakernst@syr.edu

June 2021

INTRODUCTION

The Applied Data Science program at Syracuse University's School of Information Studies teaches students the opportunity to collect, manage, analyze, and develop insight using data from various domains using different tools and techniques.

The Project Portfolio Milestone will show a sample size learned skillsets from the program:

- IST 659 – Database Administrative Concepts and Database Management
- IST 707 – Data Analytics
- IST 736 – Text Mining

**IST 659 – DATABASE ADMINISTRATIVE
CONCEPTS AND DATABASE
MANAGEMENT**

IST 659

IST 659 is an introductory course to database management systems. This course examines data structures, file organizations, concepts, and principles of database management systems (DBMS) as well as data analysis, database design, data modeling, database management, and database implementation.

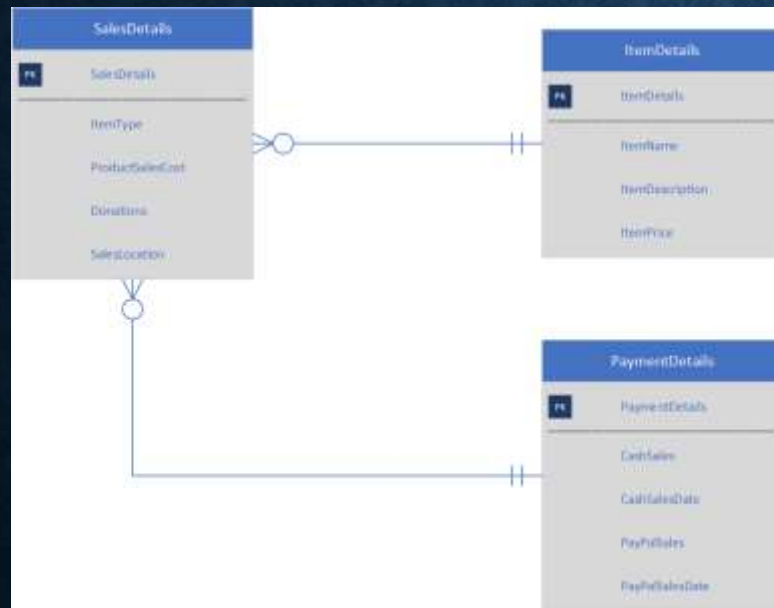
More specifically, it introduces hierarchical, network, and relational data models; entity-relationship modeling; basics of Structured Query Language (SQL); data normalization; and database design. Using Microsoft's Access and SQL Server DBMSs as implementation vehicles, this course provides hands-on experience in database design and implementation through assignments, lab exercises, and course projects. This course also introduces advanced database concepts such as transaction management and concurrency control, distributed databases, multitier client/server architectures, web-based database applications, data warehousing, and NoSQL.

IST 659 PROJECT DESCRIPTION

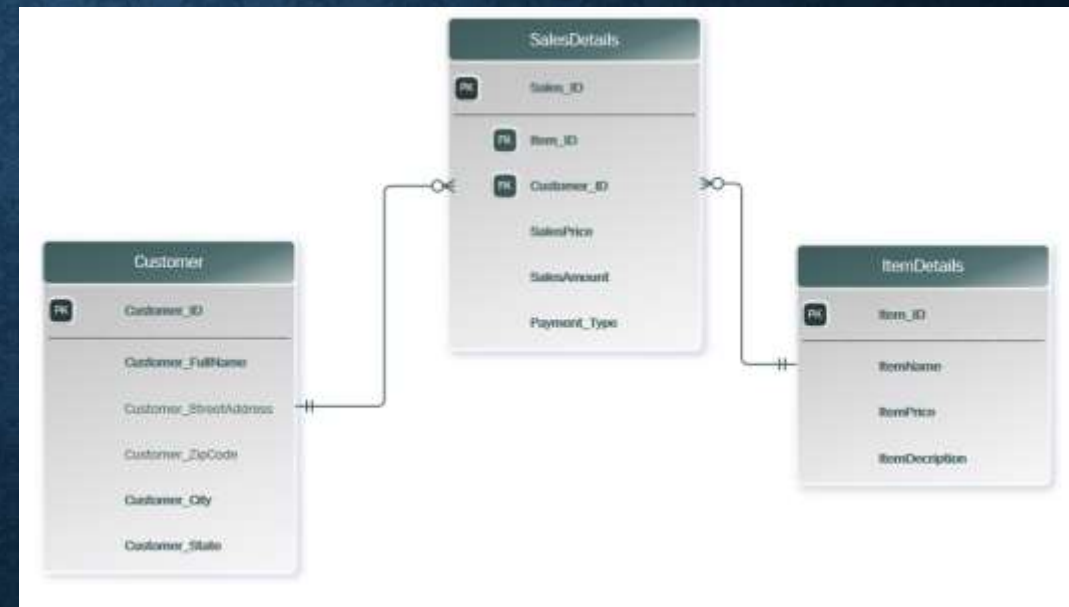
During the months of August, September, and October my son who is now 13 years old sells Boy Scouts popcorn. These sales are a way the scouts fundraise to earn their own way in Scouting. It provides him the opportunity to fund his entire year in Scouting. It provides Units the funding needed to execute a successful program year.

IST 659 PROJECT

Conceptual ERD



Normalized Logical Model



REFLECTIONS

Before starting this project, I had assumptions which caused me more time to fix my mistakes. Towards the end of the project, I learned that I need to focus more on the smaller errors I made and to be more vigilant of my ERD. It did change and I expected it change. Moreover, the next time I work on this database, I would want to build a better model. I had to make some changes that I had to go back and fix. This was a learning experience and I learned that I need to take my time. In my job, I will be working with the database and I will definitely be deliberate with my actions. Since this project was completed, I have worked on 2 very distinct databases which I was able to apply my knowledge and assist in managing multiple databases with a front end and a back end.

LEARNING GOALS

This project examined data structures, file organization, concepts, and principals of database management systems (DBMS) as well as data analysis, database design, data modeling, database management, and database implementations. More specifically, this project showed relational data models, entity-relationship modeling, basics of Structured Query Language (SQL); data normalization; and database design. By using Microsoft Access and SQL Server DBMSs as implementation vehicles, a database was created and maintained.

IST 707

This course will introduce popular data mining methods for extracting knowledge from data. The principles and theories of data mining methods will be discussed and will be related to the issues in applying data mining to problems. Students will also acquire hands-on experience using state-of-the-art software to develop data mining solutions to scientific and business problems. The focus of this course is in understanding of data and how to formulate data mining tasks in order to solve problems using the data.

The topics of the course will include the key tasks of data mining, including data preparation, concept description, association rule mining, classification, clustering, evaluation and analysis. Through the exploration of the concepts and techniques of data mining and practical exercises, students will develop skills that can be applied to business, science or other organizational problems.

IST 707 PROJECT

Define a problem on the dataset and describe it in terms of its real-world organizational or business application. The problem may use one or more of the types of data mining algorithms that we have studied this semester: Classification, Clustering and Association Rules, in an investigation of the solution to the problem. This investigation must include some aspects of experimental comparison: depending on the problem, you may choose to experiment with different types of algorithms, e.g. different types of classifiers, and some experiments with tuning parameters of the algorithms.

Alternatively, if your problem is suitable, you may use more than one of the algorithms (Clustering + Classification, e.g.). If there are a larger number of attributes, you can try some type of feature selection to reduce the number of attributes. You may use summary statistics and visualization techniques to help you explain your findings.

IST 707 PROJECT

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Age	Outcome	data.frame': 2000 obs. of 8 variables:											
2	138	62	35	0	33.6	47	1	\$ Pregnancies	: int	2	0	0	0	1	0	4	8	2	2
0	84	82	31	125	38.2	23	0	\$ Glucose	: int	138	84	145	135	139	1				
0	145	0	0	0	44.2	31	1	\$ BloodPressure:	int	62	82	0	68	62	78	72			
0	135	68	42	250	42.3	24	1	\$ SkinThickness:	int	35	31	0	42	41	32	17			
1	139	62	41	480	40.7	21	0	\$ Insulin	: int	0	125	0	250	480	265				
0	173	78	32	265	46.5	58	0	\$ BMI	: num	33.6	38.2	44.2	42.3						
4	99	72	17	0	25.6	28	0	\$ Age	: int	47	23	31	24	21	58	28			
								\$ Outcome	: int	1	0	1	1	0	0	0	0	0	0

IST 707 PROJECT

Model name	Algorithms	accuracy	Algorithm parameters
Model 1	Naïve Bayer's	0.763285	NB1<- naive_bayes(train_df\$label~., data=train_df)
Model 2.1	Decision tree	0.9130435	tree1 <- rpart(train_df\$label~., data = train_df, method="class")
Model 2.2	Decision tree	0.821256	tree2 <- rpart(label~ Glucose + Insulin+ BMI+Age, data=train_df, method="class", control=rpart.control(minsplit=60, cp=0.001))
Model 3.1	SVM polynomial	0.2898551	SVM_p <- svm(label~., data=train_df, kernel="polynomial", cost=.1, scale=FALSE)
Model 3.2	SVM Linear	0.8019324	SVM_l <- svm(label~., data=train_df, kernel="linear", cost=.1, scale=FALSE)
Model 3.3	SVM radial Kernel	0.7101449	SVM_r <- svm(label~., data=train_df, kernel="radial", cost=.1, scale=FALSE)
Model 4	Random forest	0.9758454D	rf<- randomForest(label~., data=train_df, ntree=100, proximity=TRUE)

Model 1 Naïve Bayer's

	Real results		
	0	1	
Predict results	0	118	21
	1	22	46

Model 2.1 Decision tree 1

	Real results		
	0	1	
Predict results	0	124	14
	1	16	53

Model 2.2 Decision tree 2

	Real results		
	0	1	
Predict results	0	122	14
	1	18	53

Model 3.1 SVM polynomial

	Real results		
	0	1	
Predict results	0	122	44
	1	18	13

Model 3.1 SVM Linear

	Real results		
	0	1	
Predict Results	0	121	23
	1	19	44

Model 3.1 SVM Radial

	Real results		
	0	1	
Predict results	0	140	67
	1	0	0

Model 4 Random Forest

	Real results		
	0	1	
Predict Results	0	549	12
	1	9	258

REFLECTIONS

Our models used Naïve Bayes, Decision Trees, SVM, and Random Forrest. Using the Support Vector Machines gave the highest accuracy with the linear kernel. However, it was the Random Forest algorithm that yielded the highest accuracy in predicting pre-diabetes based on data points. This data mining project is important because diabetes is in the top ten of leading causes of death.

LEARNING GOALS

The work on this project taught me how to document, analyze, and translate data mining needs into real life solutions. Also, there was a data story that developed from the data and patterns emerged that had validity and a high percent of accuracy. The first step to telling a data story is cleaning the data which included in checking for missing values, incorrect values, and other outliers. The next step was to create a training and testing datasets so that we could create data models using different algorithms.

IST 736

The main goal of this course is to increase student awareness of the power of large amounts of text data and computational methods to find patterns in large text corpora. This course is designed as a general introductory level course for all students who are interested in text mining.

Programming skill is preferred but not required in this class.

This course will introduce the concepts and methods of text mining technologies rooted from machine learning, natural language processing, and statistics. This course will also showcase the applications of text mining technologies in (1) information organization and access, (2) business intelligence, (3) social behavior analysis, and (4) digital humanities.

IST 736


The main goal of this course is to increase student awareness of the power of large amounts of text data and computational methods to find patterns in large text corpora. This course is designed as a general introductory level course for all students who are interested in text mining. Programming skill is preferred but not required in this class. This course will introduce the concepts and methods of text mining technologies rooted from machine learning, natural language processing, and statistics. This course will also showcase the applications of text mining technologies in (1) information organization and access, (2) business intelligence, (3) social behavior analysis, and (4) digital humanities.

IST 736 PROJECT DESCRIPTION

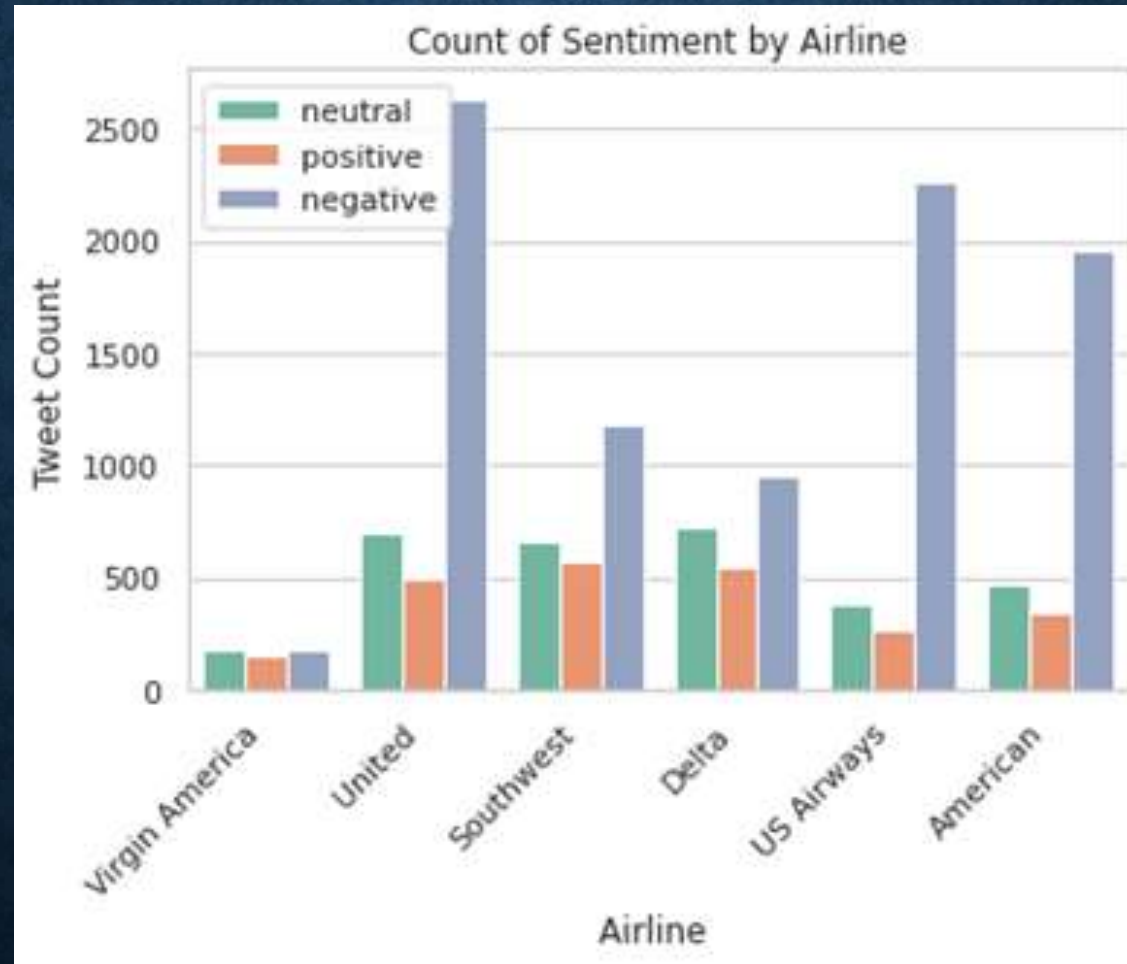
With more than 31MM monetizable daily active users in the US in Q1 2020, Twitter has become a vast medium for users to voice their opinions, and for brands or businesses to reach a broad audience and engage with customers. Customers sharing their sentiment on Twitter can play a significant part in a brand's reputation, especially airlines. The following tweets provide examples of customers who use Twitter to 1) voice their frustration and 2) call upon a brand to act.

The intent of this project was to analyze tweets from Twitter to produce a sentiment analysis dataset and describe it in terms of its real-world organizational or business application. The problem required the use of one or more of the types of data mining algorithms.

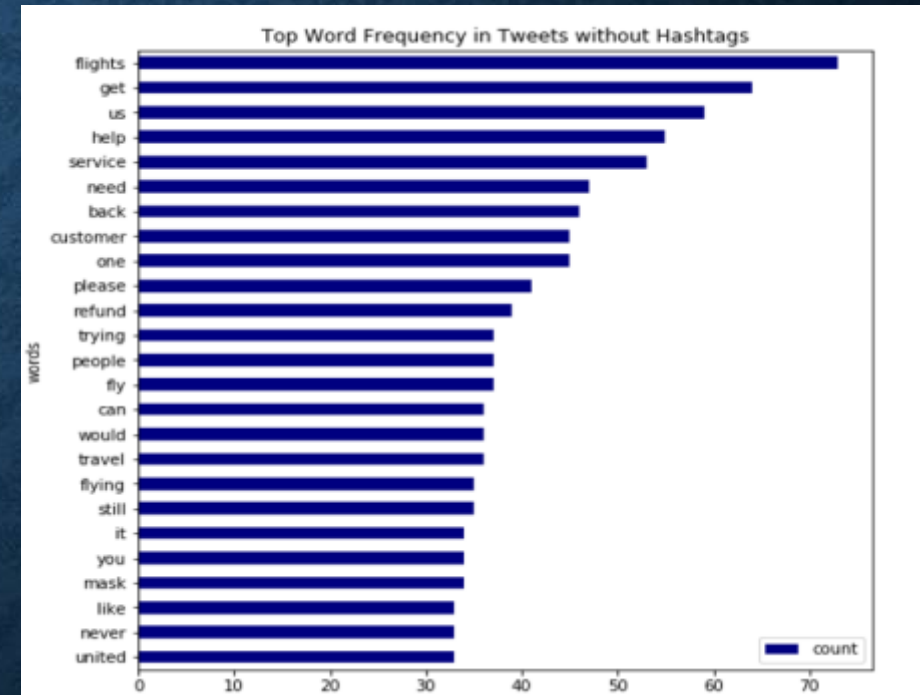
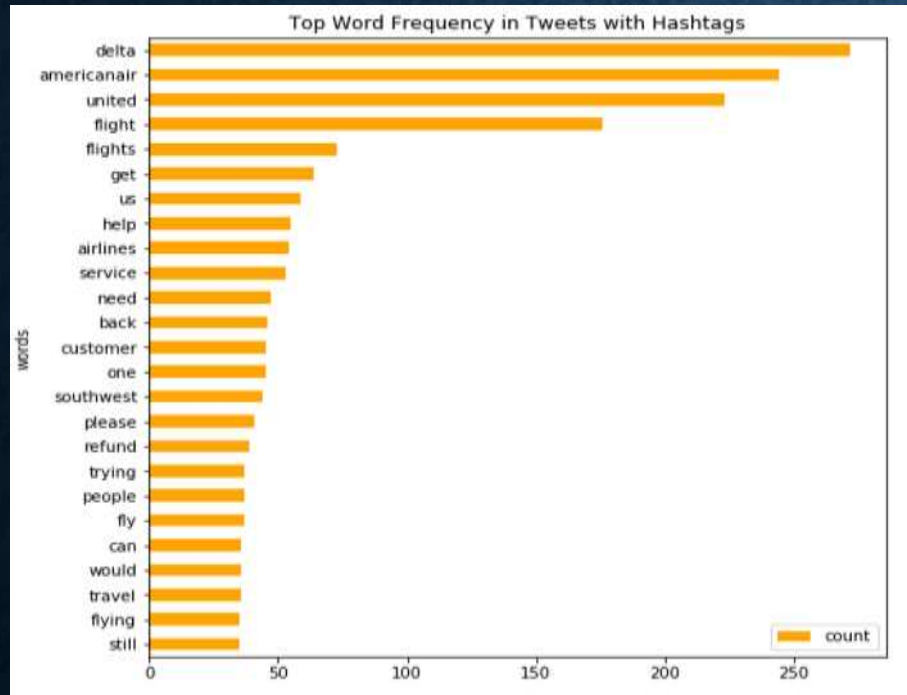
IST 736 PROJECT

	🔍 tweet_id	⚙️ airline_sentiment	⚙️ airline	⚙️ text
	 56758827... 57031060...	negative 63% neutral 21% Other (1) 16%	United 26% US Airways 20% Other (4) 54%	14427 unique values
1	570306133677760513	neutral	Virgin America	@VirginAmerica What @dhepburn said.
2	570301130888122368	positive	Virgin America	@VirginAmerica plus you've added commercials to the experience... tacky.
3	570301083672813571	neutral	Virgin America	@VirginAmerica I didn't today... Must mean I need to take another trip!
4	570301031407624196	negative	Virgin America	@VirginAmerica it's really aggressive to blast obnoxious "entertainment" in your guests' faces & they have little recourse

IST 736 PROJECT



IST 736 PROJECT



REFLECTIONS

This project focused on using text mining on real-world data and developing a dataset where sentiment analysis was used to answer business questions such as airlines that are receiving positive reviews versus airlines that receiving negative reviews and how they compare against each other. The project focused on text mining focuses on unstructured text data, which come in words. It converted text to numbers that still bear the meaning of text is an important topic in text mining.

LEARNING GOALS

I learned how to create a dataset and use algorithms to conduct sentiment analysis. Also, learned how to apply LDA to data. Once the model was created, three topics were created on data with hashtags on which customers were tweeting about.

REFERENCES

Hoffer, J. A, Ramesh, V., & Topi, H. (2016). Modern database management (12th ed.). New York, NY: Pearson.

Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of EMNLP 2002, 79-86. url: <https://arxiv.org/pdf/cs/0205070.pdf>

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar (2005) Introduction to Data Mining. (Free sample chapters available at authors' website <http://wwwusers.cs.umn.edu/~kumar/dmbook/index.php>)

Weiss, S. M., Indurkha, N., & Zhang, T. (2010). Fundamentals of predictive text mining. New York: Springer. ISBN: 978-1849962254

