

Damon Porter

PortfoLio

Contents

[Introduction 2](#_Toc21738946)

[Scripting for Data Analysis 2](#_Toc21738947)

[Project Overview: 2](#_Toc21738948)

[Learning Objectives: 3](#_Toc21738949)

[Data Analytics 4](#_Toc21738950)

[Project Overview: 4](#_Toc21738951)

[Learning Objectives: 6](#_Toc21738952)

[Natural Language Processing 7](#_Toc21738953)

[Project Overview: 7](#_Toc21738954)

[Learning Objectives: 9](#_Toc21738955)

[Conclusion 10](#_Toc21738956)

[Github Directory: 11](#_Toc21738957)

[References: 11](#_Toc21738958)

# Introduction

The field of Data Science aims to drive decision making by analyzing vast amounts of complex big data by combining a variety of fields such as machine learning, statistics, and data mining and analysis. Syracuse’s Applied Data Science program aims to prepare their students in these tenets of data science by providing courses directed towards effectively collecting and capturing the data, managing data, analyzing the data, and effectively communicating the findings of their research. These tenets are a part of the overall objective of the Applied Data Science program. These objectives include:

1. Describe a broad overview of the major practices’ 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 in practice.

While these skills were intertwined with all of the courses in the program, they were most evident through the course work in Scripting for Data Analysis, Data Analytics, and Natural Language Processing. Through the use of a variety of technologies, statistical concepts, and modeling techniques, these were exemplar courses in regards to the learning objective of the Applied Data Science Program.

# Scripting for Data Analysis

## Project Overview:

The objectives of the final project for the Scripting for Data Analysis Course was, utilizing Python, gather, preprocess(clean), and analyze a large dataset. To address the requirements of this project, a question was posed as to whether Topic Modeling techniques possessed the capability of finding commonality between lyrics of popular Spotify songs and grouping them together. “Topic modeling is a statistical modeling concept that allows analyst to view documents in abstract and create a topic based on these abstracted words in from the documents.” (Porter, IST 652, 2019)

Prior to creating the Topic Model, it was necessary to clean and preprocess the data. The initial data preprocessing included removing observations that lacked lyrics, such as instrumental tracks, removing non-English language lyrics, and removing songs that were duplicated. Next it was necessary to vectorize the text data utilizing two prominent text mining libraries: SKLearn and NLTK. Utilizing these packages in conjunction ensured that a clean corpus was created by “the removal of punctuation, removing unnecessary words, stemming words, and making all the words lower case.” (Porter, IST 652, 2019)

Once the data were cleaned and the corpora were created, the Gensim library, a package specifically utilized in Latent Dirichlet Allocation Topic Modeling, was loaded and run against the aforementioned corpora (Figure 1). The figure below displays 50 different songs grouped together by potential topics. The larger the circle, the more distinct the topic. In these results, song 1 and song 2 are well defined topics that more than likely won’t have many similarities. The songs that overlap with songs 1 and 2, are likely to belong to the same or a very similar topic.

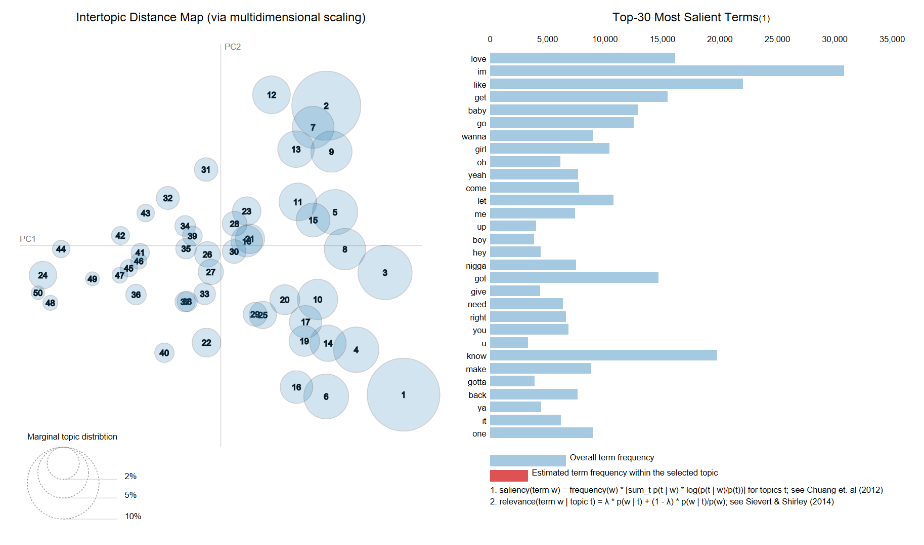
****

Figure : LDA Visualization

## Learning Objectives:

1. Describe a broad overview of the major practices’ areas in data science.
   1. This project allowed for the use exploratory text mining, text data preprocessing (vectorization, TDIDF, corpus creation, tokenization, etc.), Latent Dirichlet Allocation Topic Modeling, and exposure to Python
2. Collect and Organize data.
   1. One of the key objectives of this research assignment was to gather outside data to create a data script. The data from this research was acquired from Kaggle and can be found here: <https://www.kaggle.com/edalrami/19000-spotify-songs>
3. Identify patterns in data via visualization, statistical analysis, and data mining and develop alternative strategies based on the data
   1. During the exploratory phase of this research the initial models were found to contain a number of filler words that skewed the models, so that provided the insight that these words (la, oh, uh, hmm, etc..) didn’t have any intrinsic value, and it would prove beneficial to remove them from the analysis.
4. Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization
   1. Throughout this research, it was necessary to document and explain Topic Modeling, and how the Gensim library worked, but just as necessary to document the code behind the data preprocessing, Gensim library, and the associated visualizations.

# Data Analytics

## Project Overview:

While the assignment for the Data Scripting predominately focused on the mechanics of creating a script utilizing Python, the Data Analytics focused on the intricacies of data modeling, implementing and deploying models apropos to the problem at hand, and delivering data driven solutions. The objective for the final project was to utilize the course criteria to solve a real-world problem. This project was performed in unison with the Lyrics exploration of the Scripting for Data Analytics course, but focused on the components of what makes popular music.

The research conducted for this experiment included taking the components of the 19000 Spotify songs and finding commonality and determining if it possible to classify a hit song utilizing several different supervised learning algorithms. These predictors (figure 2) encompass key characteristics of any song.

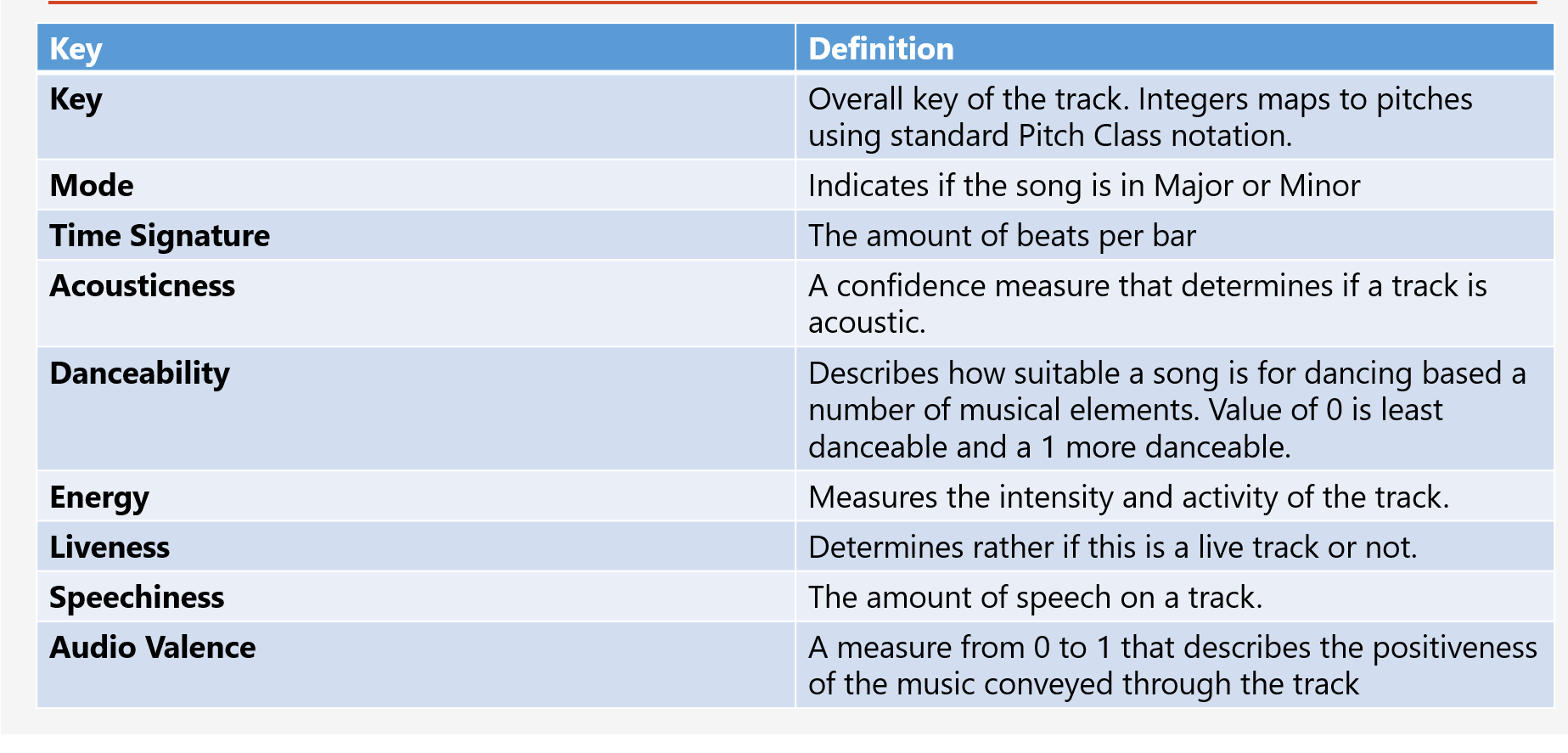


Figure : Predictor Variables

Prior to building the classifier with these predictors, it was necessary to perform extensive preprocessing on this data set. Some of the common preprocessing steps were taken (deduping, removing records with Nulls, and transforming the data utilizing logarithms to effectively compare numeric vales), but it was also necessary to derive a new field and subset the data. The data set was categorized by playlist and not genre, so it was necessary to derive genres from these playlists. This was done through analyzing the names of playlist and deducing the most likely genre name.

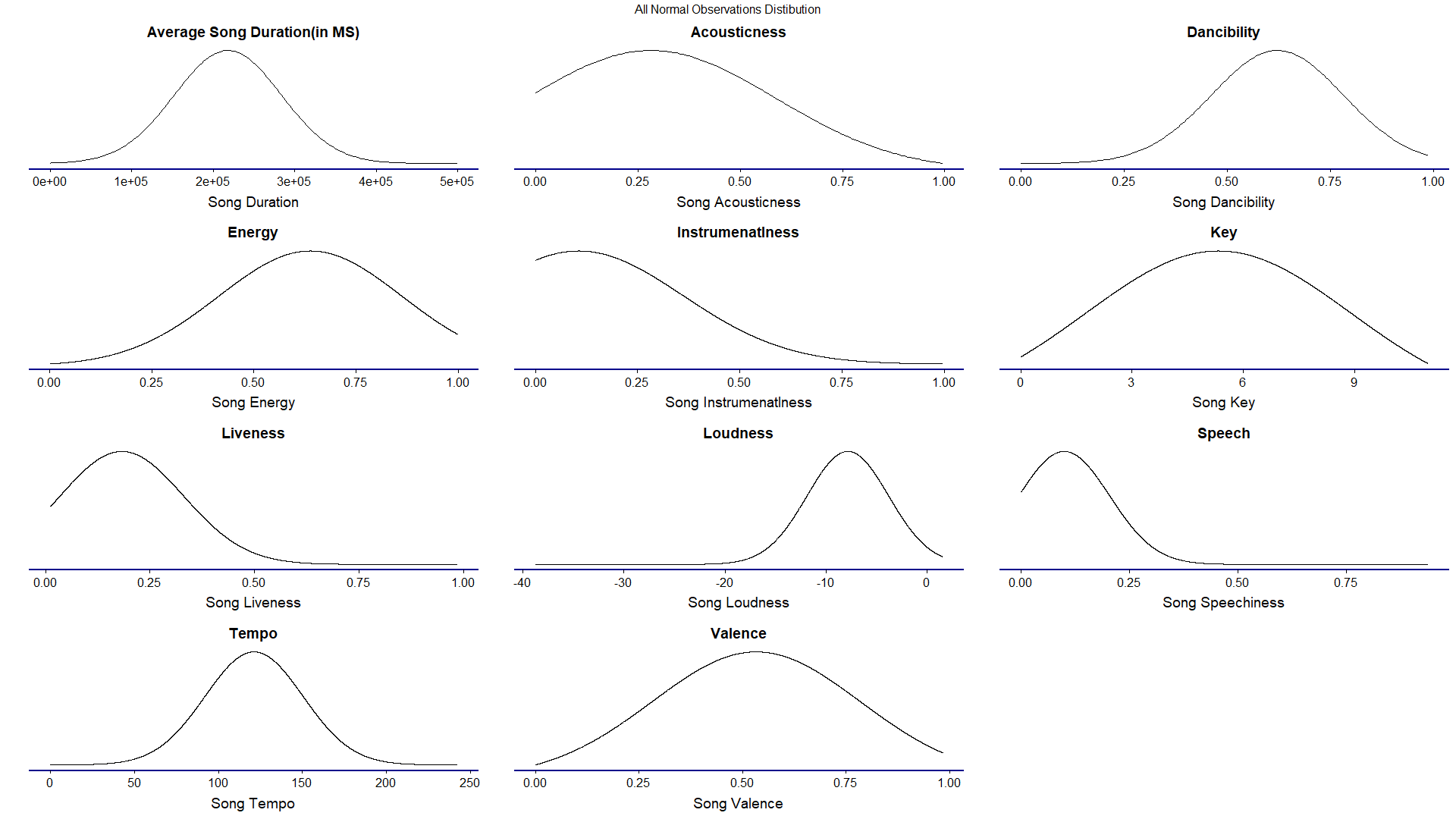


Figure : Distributions of the predictors

Following the preprocessing steps, the data was split into a training and test set to build the supervised machine learning models. The algorithms applied in this research were Support Vector Machines, Naïve Bayes, and Decision Trees classifiers. Each of these models were trained with the 60% of the data in the training set and tested against the test set by utilizing cross fold validations. The results of the classifiers were mixed. The Decision tree classifier accuracy for predicating the genre accurately was only 50% (Figure 4). The SVM, the most effective model in this analysis, was approximately 60% effective in classifying the genres (Figure 5).

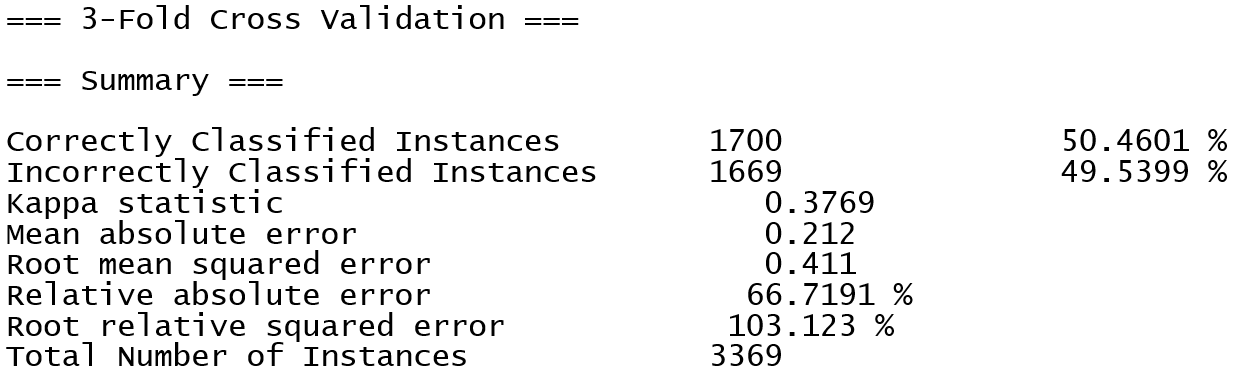


Figure : DECISION TREE RESULTS

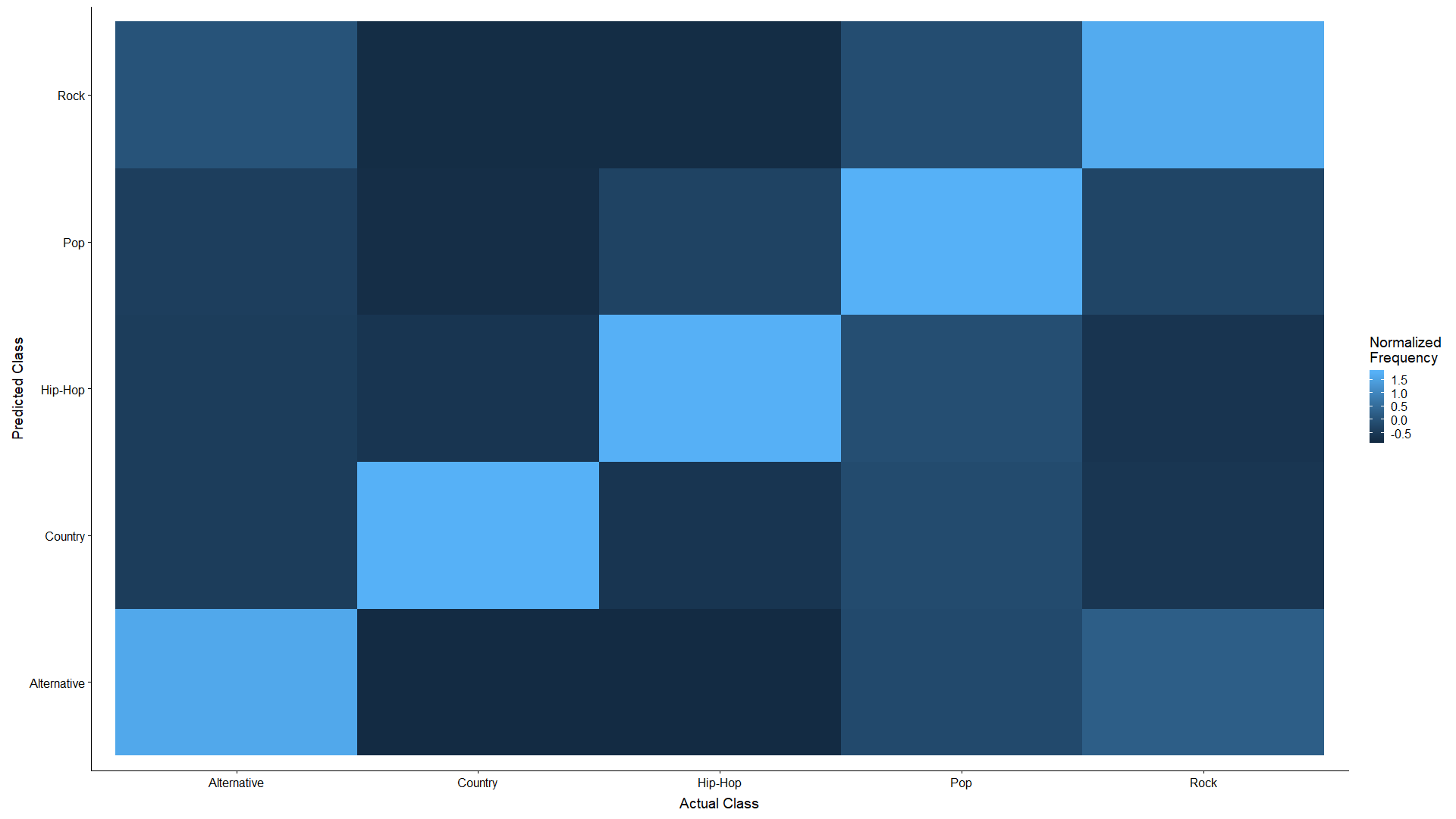


Figure : SVM HeatMap

While the models presented in the research were fairly effective in determining the correct genre in the cross -validation evaluation, they each had difficulty distinguishing between some genres that share similarities. “Correlation between two genres” (Porter, IST 707, 2019) could be the cause of misclassification and further analysis on lyrical content may be necessary to create a more effective classifier.

## Learning Objectives:

1. Describe a broad overview of the major practices’ areas in data science.
   1. The Data Analytics course focused on the use of R as an analytics’ tool to perform data preprocessing, perform expletory data analysis through the use of visualizations, and incorporated machine learning algorithms to do predictive analytics.
2. Collect and Organize data.
   1. While the 19000 Spotify Songs data set was the same for this project, the method of ensuring the data was suitable for the needs of this project were completely different than the text analysis. This reinforced the idea that data can provide a variety of insights depending on the problem at hand.
3. Identify patterns in data via visualization, statistical analysis, and data mining and develop alternative strategies based on the data
   1. To enhance the models developed during this research, Information Gain was used to determine which of the predictors provided the greatest insight into classifying genres. This could allow for more streamlined, informative models that only utilized the most effective predictors.
4. Develop alternative strategies based on the data.
   1. While exploring the data for this project, it was found that the outcome variable, genre, did not exist. Due to this circumstance, it was necessary to derive this column by analyzing the existing data and transforming it to meet the needs of the research project.
5. Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization
   1. While research is driven by successful findings, this project also provided insight into algorithms failing. This allowed the researcher to not only explain the successful experimentation, but the failures as well. Being able to do this shows a level of thoroughness and completeness.

# Natural Language Processing

## Project Overview:

The objective of the final project in the Natural Language Processing Course was text classification. Text classification has proven successful in determining whether an email is spam or not, if a customer review online is legitimate or was it fabricated, or determining the general sentiment of text data. These things are fairly easy to detect and classify due to the nature of how the words are used or the structure of an email determined to be spam. But can a machine successfully classify text data that even humans struggle classifying at a high percentage? The problem statement for this research was to build a sarcasm classifier. This would not only show how well a machine detects sarcasm, but it also combines aspects of both of the previous research topics.

Data collection for this projected could be summarized by the following excerpt from the research paper:

“The data for this research project was collected from the Kaggle dataset ‘Sarcasm on Reddit.” This dataset contains 1.3 million comments from Reddit. These comments were combined into a csv and balanced by taking an even amount of sarcastic, labeled as ‘1’, and non-sarcastic comments, labeled as ‘0’. To perform this analysis, a subset of data, 20,00 records evenly distributed. was utilized. This data was then converted into a corpus by utilizing the built in NLTK Corpus creator and read back into Python.” (Porter, IST 664,2019)

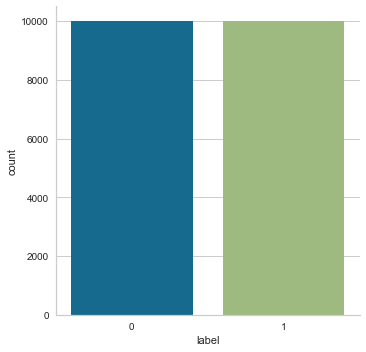
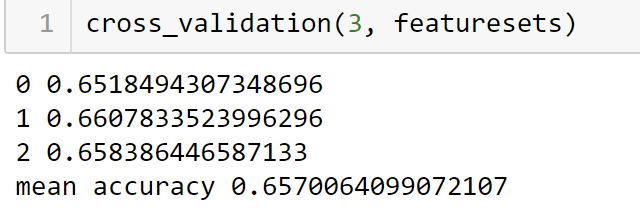


Figure : Sarcasm Data Distribution

Once the data was collected, it was necessary to create a corpus that viewed each individual post as a document. After the corpus was created, there were several different experiments performed with differing methods of preprocessing. Using several preprocessing procedures for feature extraction enabled the ability to gauge how responsive the classifier performed to the transformation of the text. The feature extraction included lower all words in corpus to ensure features wouldn’t appear multiple times in the frequency due to changes in spelling, stemming and lemmatizing the text in order to only account for the roots of the features, working with n-grams to observe how evaluating word phrases affected the classification, and removing stop words.

Utilizing a combination of the feature extraction techniques, the top 49000 features, by frequency, were utilized to train the classifier. Once the classifier was built with training data, hold out data was used to make predictions. The accuracy of the predictions all hovered between 59 and 65%. Rather solely focusing on the accuracy score of the classifier, cross fold validation was utilized to reperform the predictions on different subsets of the training set. Using this method of validation, the accuracy scores generally hovered around 66% accuracy.



For the final evaluation of the models created for this classification task, a confusion matrix was utilized to compare the results and provide Precision, Recall, and F1 of the classifier. These are all statistical measures of evaluation. The precision measures all the true positives and divides them by number of true positives plus false positives. The recall measures the number of true positives divided by the number of true positives plus false negatives. The F1 score takes both precision and recall into account to calculate the score.

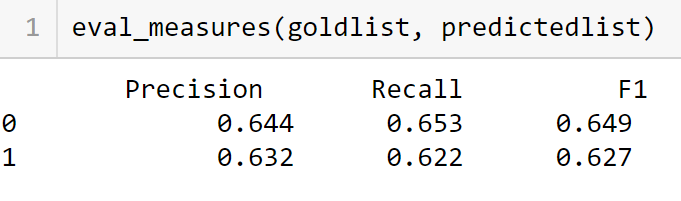


Figure : Cross Validation

## Learning Objectives:

1. Describe a broad overview of the major practices’ areas in data science.
   1. This research provided an opportunity to normalize, transform, and preprocess unstructured data. The ability to manipulate features and create corpora within the NLTK framework was a necessity to adequately complete coursework. This research also provided an opportunity to utilize a number of machine learning classifiers to make predictions.
2. Collect and Organize data.
   1. While the dataset for this project was obtained from Kaggle, there were a number preprocessing steps necessary to create the corpuses needed for this analysis.
3. Identify patterns in data via visualization, statistical analysis, and data mining and develop alternative strategies based on the data
   1. This project involved the use of creating Bi and Trigrams to gain additional insight. Yet, instead of using the raw n-gram numbers, PMI was utilized. PMI, or Pointwise Mutual Information, is a statistical measure takes into account the probability that a feature will exist together in a data rather than independently.
4. Develop alternative strategies based on the data.
   1. As the accuracy scores, and trifold validations, were consistent no matter how the features were extracted, a number of different tests were performed. A sentiment analysis gauged the polarity of the statements and other classifiers were used, but none provided an increase in accuracy.
5. Synthesize the ethical dimensions of data science in practice.
   1. As this project dealt with individual user names, a model was created to strip out none English words in a dictionary. This allowed the models to not utilize these usernames, but it also ensured a level of anonymity.

# Conclusion

With the learning objectives acting as guiding principles, the Applied Data Science Program at Syracuse ensures that each student understands the elements of data science. Each of these projects reinforced concepts such as gathering data and transforming it to a format that fits within the confines of that assignment, adding a visual component to aide in the consumption of exploratory data analysis, and not having preconceived notions about data that may bias the experiment and not leave room for adjustments. They also reinforce the ability to discuss findings with peers, instructors, or course leads by ensuring that we are effectively communicating our results in a report.

Each of the projects above provided additionally clarity on how to properly utilize data science techniques, such as preprocessing and transforming data, feature extraction, the proper way to deploy machine learning algorithms, creating adequate training and test data, and creating reproducible results.

The skillsets and traits embedded into each Applied Data Science course, will aid an aspiring data scientist to utilize the information taught in the course to the business sector. The constant reinforcement of these learning objectives creates data scientist who are resourceful, knowledgeable, fluid, and ethical which is invaluable in business no matter the sector.

# Github Directory:

<https://github.com/DPorter283/Portfolio>

# References:

Porter, Damon (2019) Damon Porter IST\_707\_Project2

<https://github.com/DPorter283/Portfolio/blob/master/Data%20Analytics%20Project%20Report-Damon%20Porter.zip>

Porter, Damon (2019) Damon\_Porter\_Project\_Report

<https://github.com/DPorter283/Portfolio/blob/master/Data%20Scripting%20Final%20Project%20Report-Damon%20Porter.zip>

Porter, Damon (2019) Damon\_Porter\_Final\_Project\_NLP

<https://github.com/DPorter283/Portfolio/blob/master/NLP-Final%20Project%20Classification%20of%20Text-Damon%20Porter.zip>