Syracuse University
Library Sciences

Masters of Science Applied Data Science

Portfolio Milestone

Daniel Caley

SUID: 352784643

https://github.com/dcaley5005/Data_Science

March 2022

Introduction

The Applied Data Science program at the School of Library Science at Syracuse University provides the foundational knowledge necessary to tackle descriptive, inferential, and predictive type problems.

Reports showcased in this Portfolio demonstrates the skills and intent of the program:

- Data Administration Concepts & Database Management (IST 659)
- Big Data Analytics (IST 718)
- Natural Language Processing (IST 664)

The Learning Objectives of the Program (Example of Bulleted List Slide)

1. Describe a broad overview of the major practice areas in data science.
2. Collect and organize data.
3. Identify patterns in data via visualizations, 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.

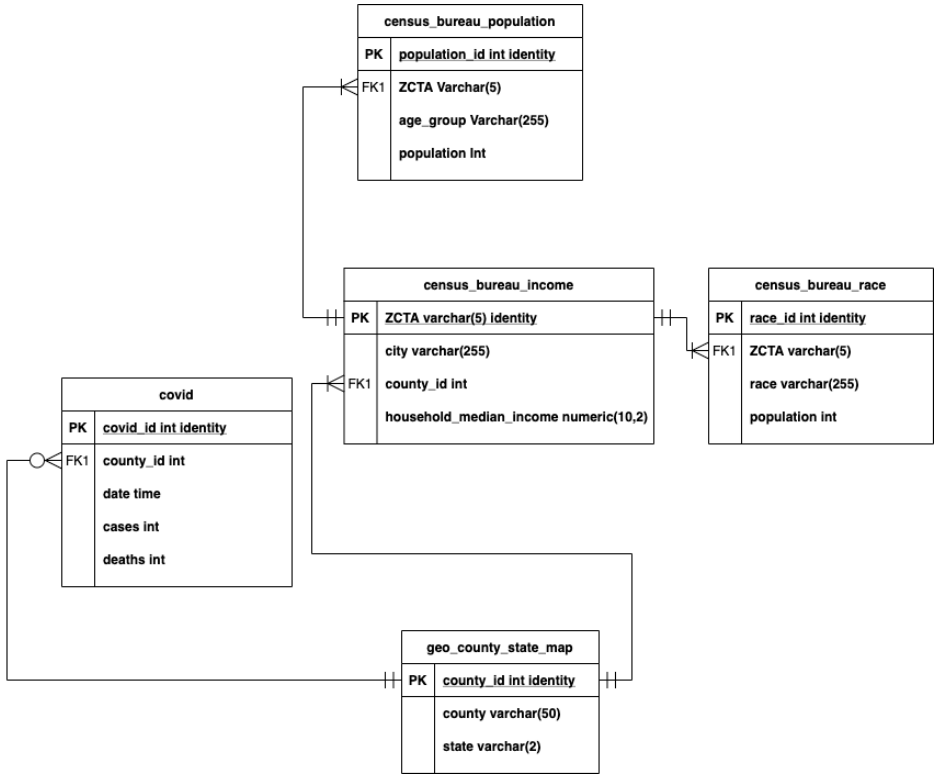
IST 659 Data Administration Concepts & Database Management

Custom Built Socioeconomic COVID SQL Data Base

IST 659: Database Administration

Introduction

- Database Administration under the guidance of Professor Chad Harper, a research focused database surrounding the socioeconomic impacts of COVID on the American Society.
- The database included the following:
  - Daily COVID Cases and Deaths by County.
  - Census Bureau Data at the ZCTA and County Level with the intention to aggregate up to the County:
    - Household Median Income
    - Population
    - Broken out by Race
  - A Geo County State Dimensional table in order to aggregate further up when needed.



IST 659: Database Administration

Modeling and Deployment

- A conceptual and Logical model were created to help organize how the data relates to each other.
- This is also known as an Entity Relationship Diagram and is useful to visualize the Database architecture.
- The database included the following:
  - Daily COVID Cases and Deaths by County.
  - Census Bureau Data at the ZCTA and County Level with the intention to aggregate up to the County:
    - Household Median Income
    - Population
    - Broken out by Race
  - A Geo County State Dimensional table in order to aggregate further up when needed.
- With the ERD in mind, SQL code was written to execute the overall objective of the database architecture. The Data was then loaded via SQL server.

IST 659: Database Administration

Sample of the Data

	zcta	county_id	household_median_income
1	35013	01009	0.00
2	35034	01007	39087.00
3	36003	01001	37000.00
4	36005	01005	49722.00
5	36480	01003	27461.00

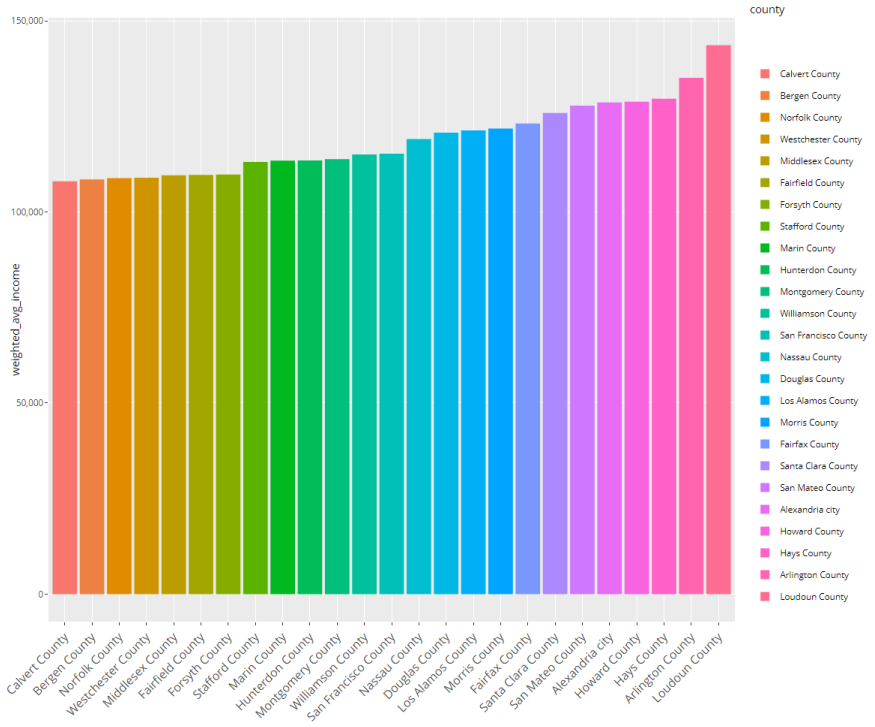
	population_id	zcta	age_group	population
1	350131	35013	Under 5 years	0
2	350341	35034	Under 5 years	547
3	360031	36003	Under 5 years	111
4	360051	36005	Under 5 years	57
5	364801	36480	Under 5 years	28

	race_id	zcta	race	population
1	35013105	35013	Other Race alone	0
2	35034100	35034	White alone	2942
3	36003101	36003	Black or African American alone	1102
4	36005102	36005	American Indian and Alaska Native alone	0
5	36480100	36480	White alone	1389

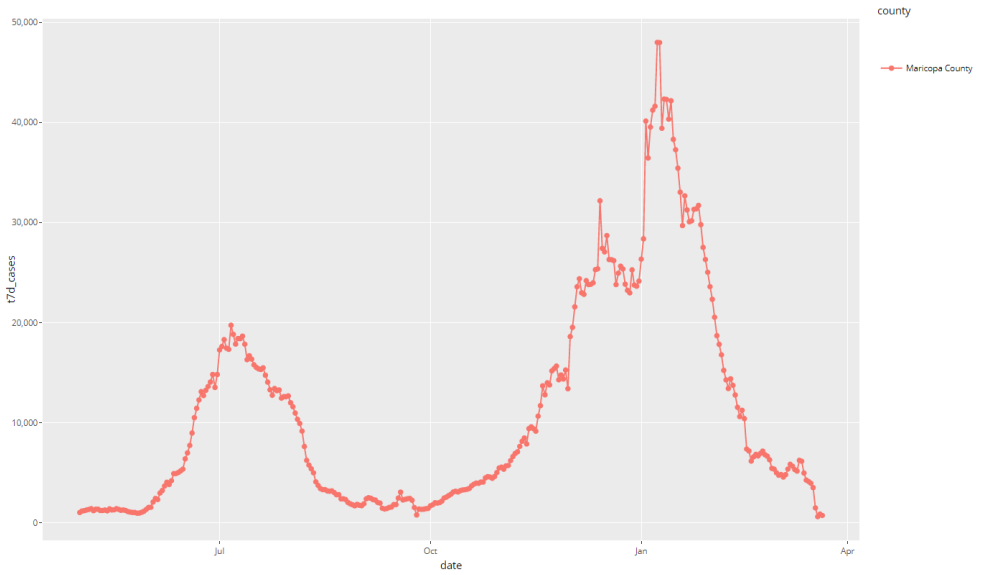
- On the right hand side is a direct look into the Database using SQL.
- The first output shows, by ZCTA the household median income, along with an ID for county in order to join up on COVID data.
- The second output shows Census data by population and broken out by age. A factor to identify how different areas were affected by COVID.
- The third output shows the population by Race. Another factor in identifying how different areas were affected by COVID.

IST 659: Database Administration

Reporting



- R was used on the reporting side to plug directly into the newly created database and extract information.
- Then Ggplot2, an R package, was used to visualize the data.



Reflection

- I have worked with databases now for 9 years and at first, I was under the impression that I would only get a little out of this class. In short, I was wrong.
- From Logical and Conceptual Models to Normal Form and then creating functions and procedures, I'm able to really approach database tasks and problems more efficiently and intelligently.
- For example, I would have skipped the modeling phase, dodged the normal form exercise and began grabbing data to put into tables prior to this class.
- I would have ended up spending an exorbitant amount of time fixing my table, making mistakes when doing analytics, and had to update multiple tables with the same data. Something I could have done better is understanding the COVID data.

IST 718 Big Data Analytics

Home Price Forecast
Using Python

IST 718: Big Data Analytics

Introduction

In the world of Big Data being able to extract insights is key to help an organization make well informed decisions.

In this case looking to a new market to open a business or buy a home can be challenging. A way to solve this problem is to look to housing data.

Specifically understanding historical returns, risk of an area, and by using forecasting techniques to distil where a business should be opened.

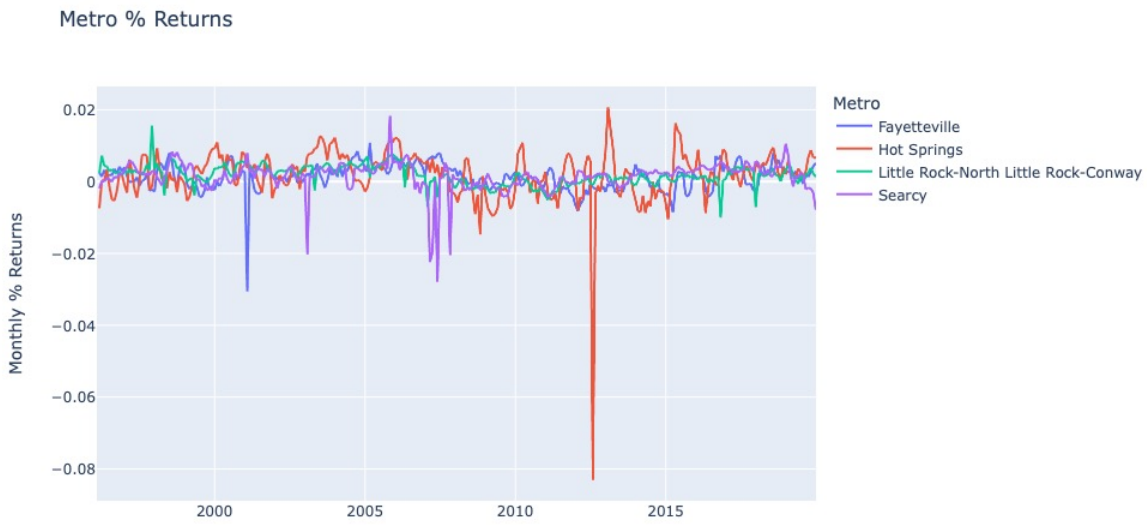
Big Data Analytics taught the OSEMi approach to help tackle any big data problem.

IST 718: Big Data Analytics

OSEMi

- **O**btain data and explain data structures and data elements.
- **S**crub data by applying scripting methods, to include debugging, for data manipulation in Python, R or other languages.
- **E**xplore data by analyzing using qualitative techniques including descriptive statistics, summarization, and visualizations.
- **M**odel relationships between data using the appropriate analytical methodologies matched to the information and the needs of clients and users.
- Interpret the data, model, analysis, and findings. Communicate the results in a meaningful way.
- Select an applicable analytical methodology for real problems in areas such as business, science, and engineering.

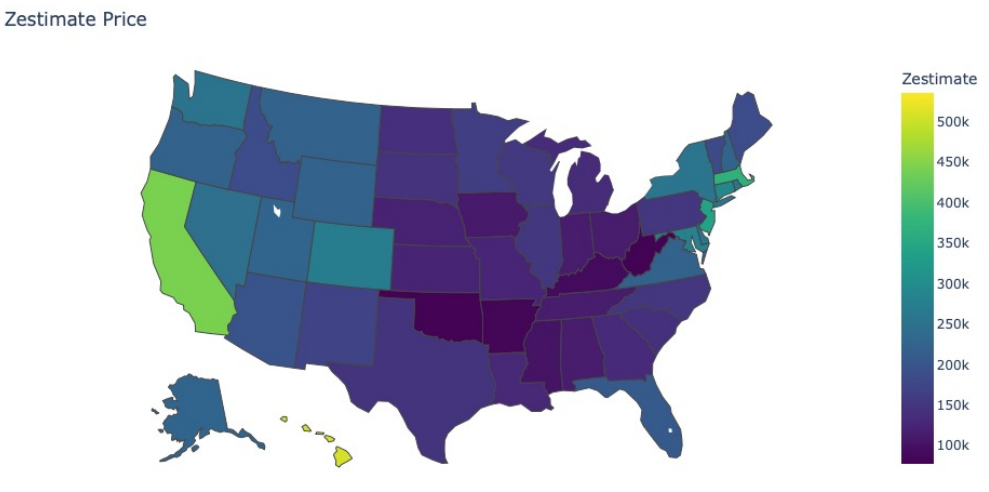
Sample of the Data

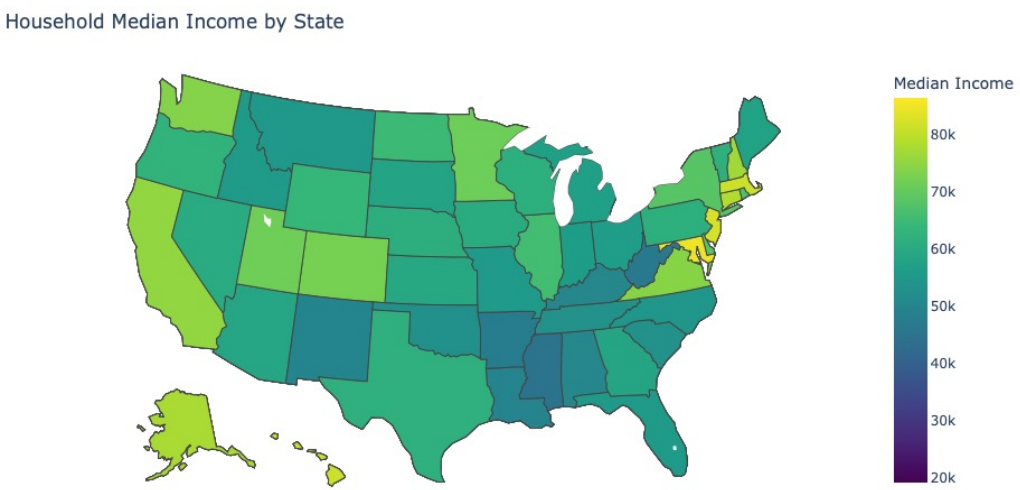


- **Obtain** Census and Single-Family Residential Home data.
- **Scrub** the data by creating a single dataframe of Census data and residential home data.
- **Explored** the data to identify any correlations or trends that can be used in the forecast
- **Modeled** by using Facebook prophet to The third output shows the population by Race. Another factor in identifying how different areas were affected by COVID.
- **Interpreted** the results to make actionable business decisions.

IST 718: Big Data Analytics

Reporting

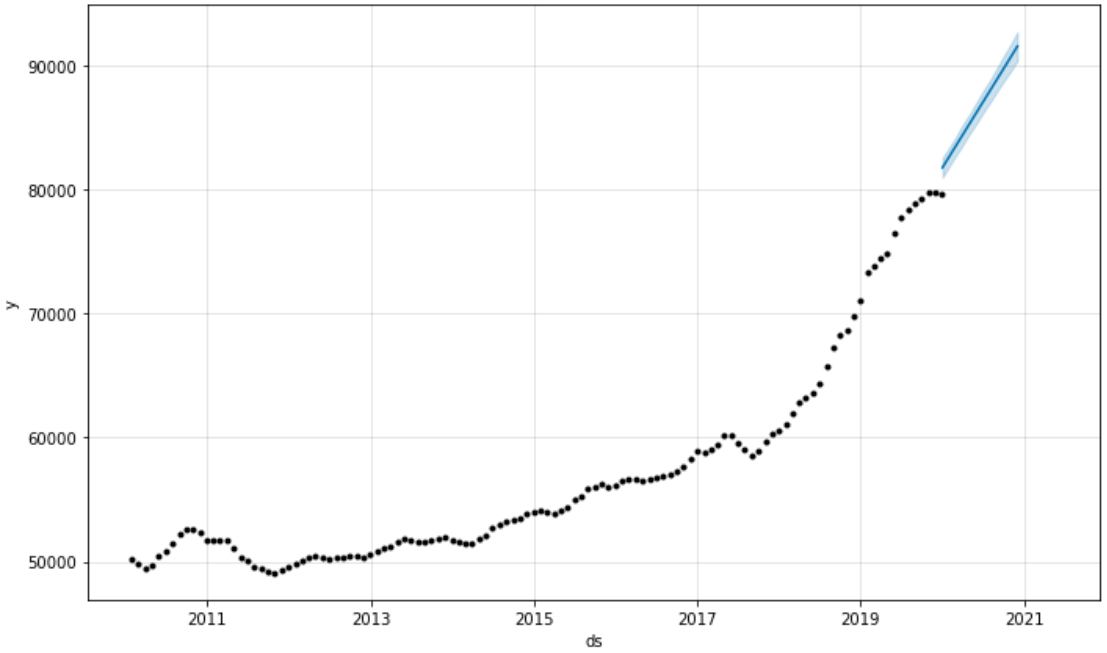




- Zestimate Price Against Household Median Income
- Python was used to graph across America.

Reflection

- My undergraduate was in Finance yet I did very little forecasting. Being able to wrangle the data and then predict how a particular market might performed was invigorating.
- The graph on the right shows Zipcode 71935 a market that appears to be moving in a positive direction.
- In combination of this graph I borrowed finance terms like Sharpe ratio to identify the best market to pick from a risk and reward perspective.



IST 664 Natural Language Processing

Predicting IMDB Ratings

IST 664: Natural Language Processing

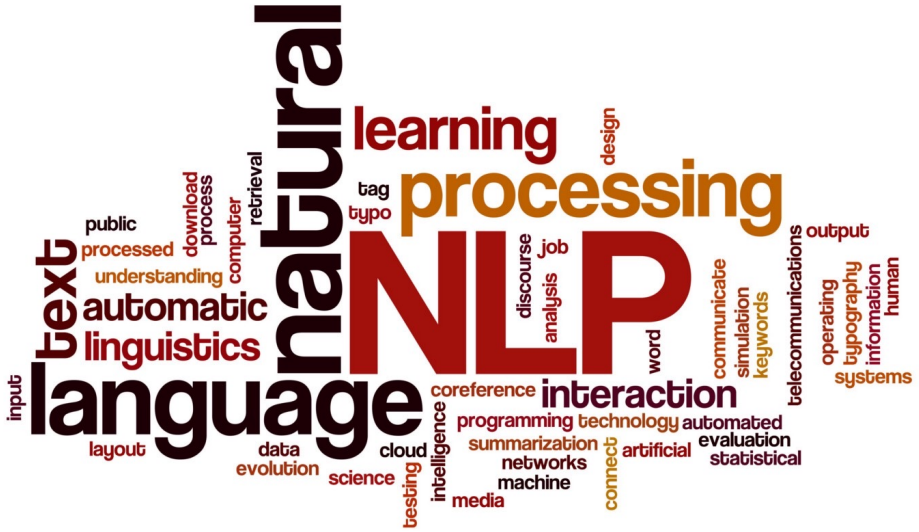
Introduction

The goal we set out today is to predict the rating of a list of comments using classification algorithms and different cleaning techniques.

The exhaustive list of items includes tokenizing, filtering, pre-processing the word list.

Then apply feature engineering techniques that span Unigrams, Bigrams, and all the way to subjectivity.

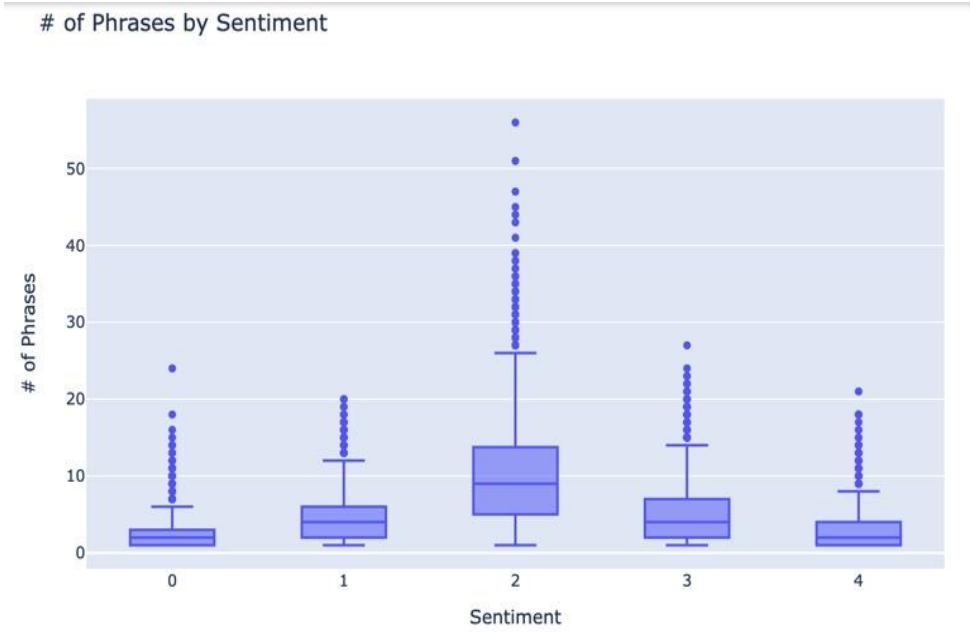
Then use multiple models like Bayes Classifier, Random Forest, and Support Vector Machine Classification. To round off the analysis, we used cross-validation to help prove the results will help in the prediction task.



Reviewing The Data

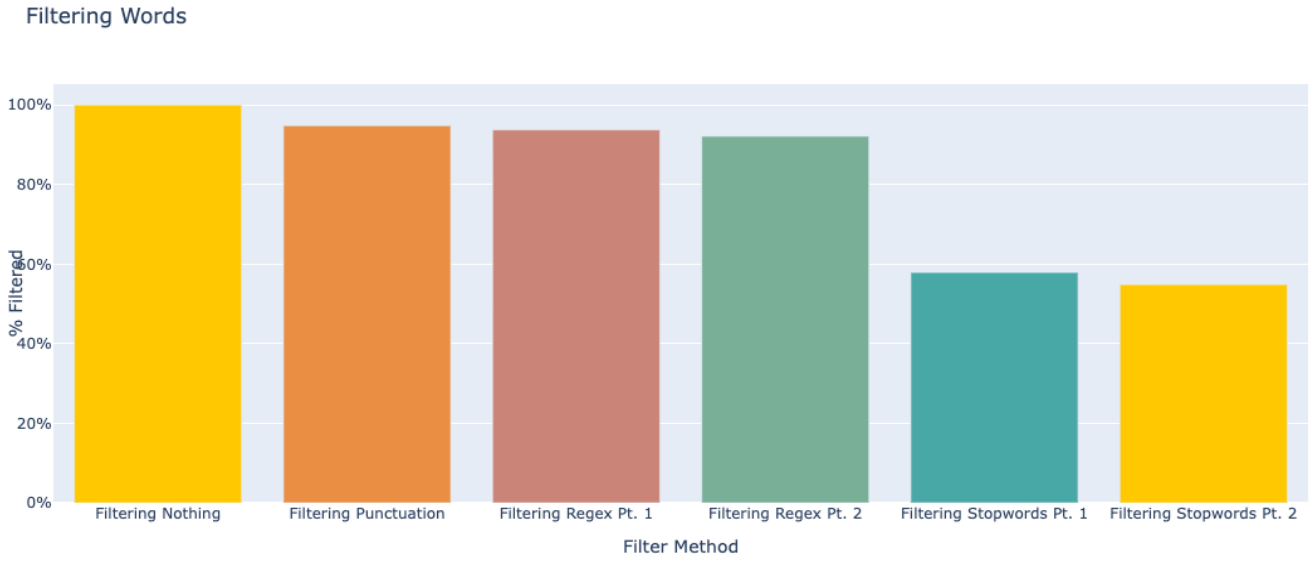
- The dataset we chose to review and analyze was the Kaggle movie review list. This dataset was produced for the Kaggle competition, which uses data from the sentiment analysis
- The data was originally taken from the Pang and Lee movie review corpus based on reviews from Rotten Tomatoes website.
- Before reviewing the original dataset we needed to import all the necessary packages needed for preprocessing and filtering. First we created training and test data frames using the given train.tsv and test.tsv files that were given to us.
- The total number of full sentences (also considered “unique_sentences”) were 8520, out of a total of 156,060 listed phrases (also considered “total_sentence”). That is approximately 18 times the amount of unique sentences.
- This confirmed that there was a lot of cleaning and filtering that needed to be done prior to creating a new model to run our experiments on.

Sample of the Data



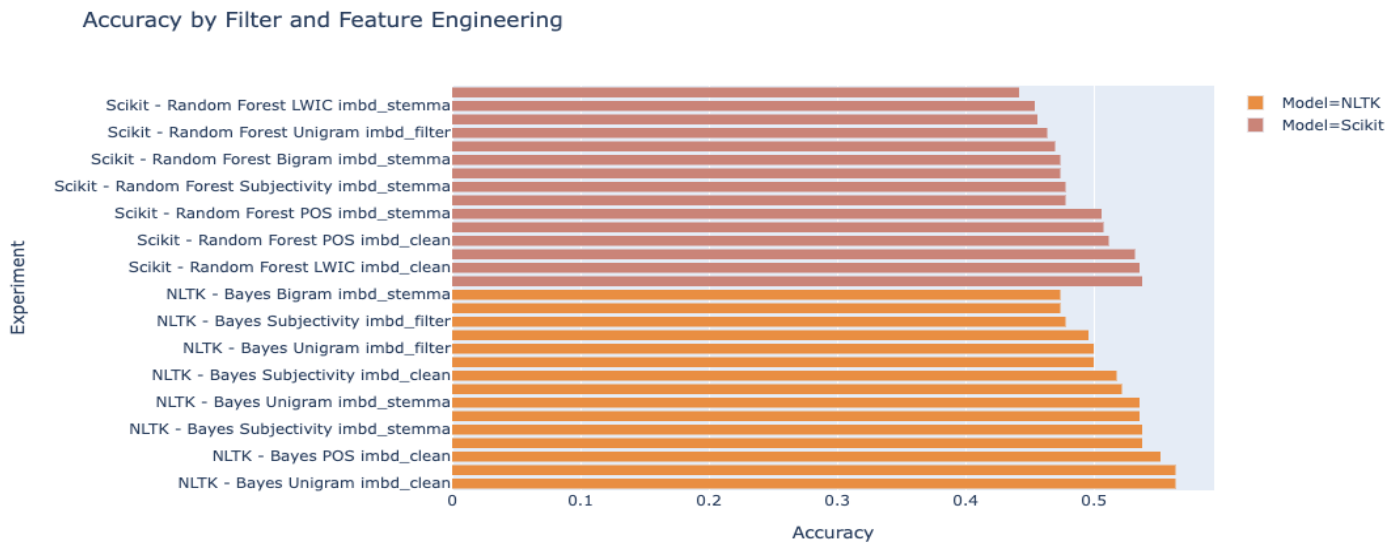
- compared the number of phrases to the sentiment scores that were assigned to them.
- As scores 1 and 4 were the least, the score of 2 had the largest upper and lower bounds with exceeding whiskers past the 50 count of assigned phrases as you can see below.
- Filtering and tokenizing the data was the next step.
- Tokenizing the data is the process of taking a large quantity of text and dividing into smaller parts.
- This is an essential piece in classifying the data.

Summary on Filtering



- The biggest gain in filtering came from removing stopwords.
- Each other method helped refine the population.

Results



- After building 3 filters, 5 feature engineering, and 2 models ran every permutation of possible experiments.
- This included 30 different experiments between Bayes NLTK and Random Forest Scikit learn.
- The results ranged from 56% to 44% with each new experiment incrementally improving from the other.
- The overall results are still poor with accuracy scores equating to nearly 50/50 shot in predicting the right rating based on comments.

Reflection

- Building this model was ground breaking. At work we have message transcripts and survey scores associated to them.
- Not all messages have this score as many users don’t fill them out.
- Being able to predict a score would help better serve customers and improve customer service.

Final Thoughts

Final Thoughts

Conclusion

- The path to predicting, using inferential statistics, and descriptive analytics are all exciting ventures to take on.
- Through filtering, transforming the data, applying multiple engineering techniques, and then modeling the featuresets, a conclusion can be seen. The power to predict comes at a great price.
- Throughout the process of building my knowledge the journey was hard, the exercises learned were invaluable, and most importantly the people will never be forgotten.