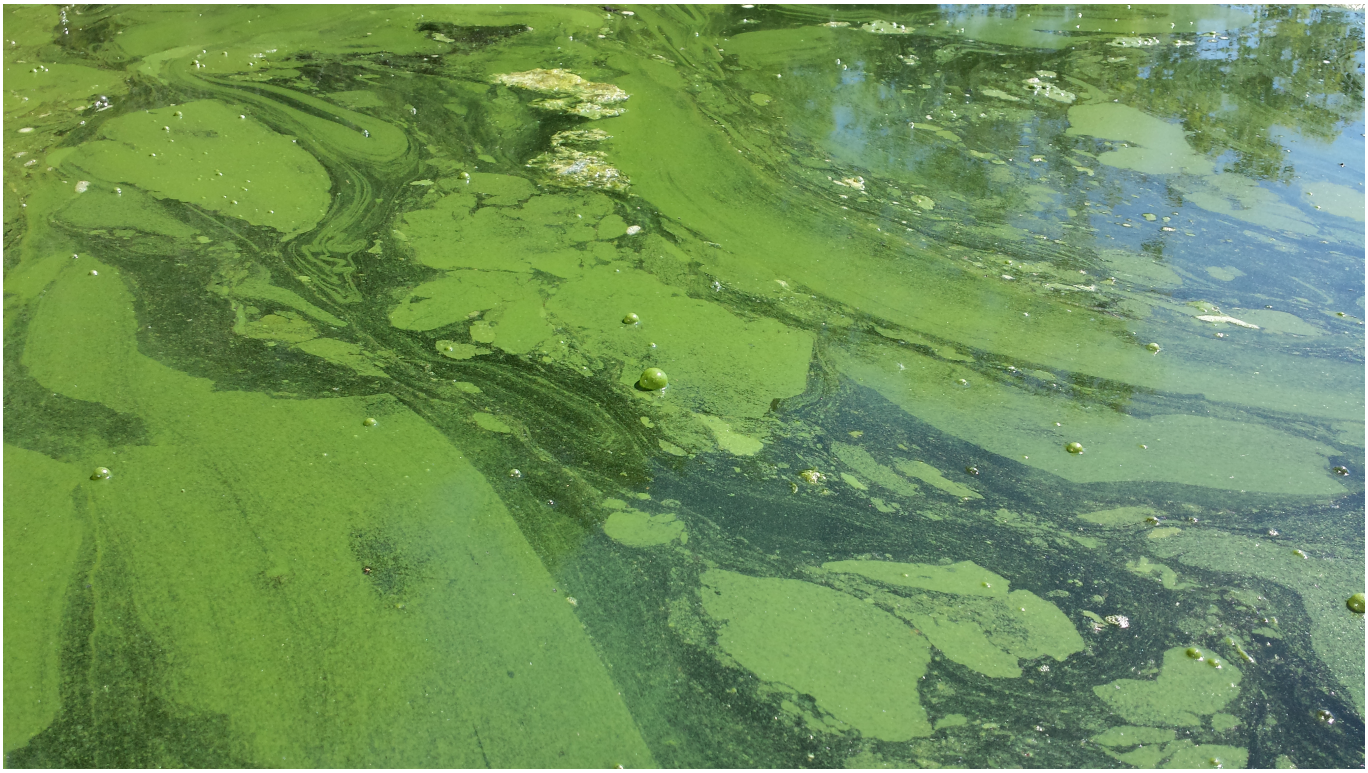
Texas Water Response: Predicting Priority with NLP Modeling



Overview

For my capstone, I explored three possible NLP models to find the most predictive parameters when trained on unique, text-based complaints submitted to Texas' Commission on Environmental Quality (TXEQ).

I found that high priority complaints featured 'oil' and 'discharge' more than any other priority, medium mainly featured 'sewage' and property disputes, while low priority complaints could focus on water pressure.

Another finding was the increase of complaints over the last 15 years, with complaints most commonly ranking as mid-level.

The final finding was that both the tuned multinomial and logisitic regression model had an average accuracy of 60% in my balanced dataset. Both models were most successful at identifying high priority responses, and struggled most with identifying mid-level priorities.

Business Problem

The state of Texas receives over 60 complaints about environmental quality a week. To best address these needs, I prepared a predictive model for the Texas Commission on Environmental Quality (TCEQ) that could help flag high priority complaints and lead to a quicker response rate than currently available.

Data Understanding

The data comes from TCEQ and is regularly updated. My project pulls from all data up to December 3, 2023. Results may vary with more recent entries.

Initially, there were 215,000 rows and 33 columns. By focusing on water complaints, and randomly balancing the data set to the least common priority (low), I ended with a dataset of 16,500 rows.

Modeling

The models included a baseline multinomial bayes model, a multinomial bayes model informed by GridSearch, and a multi-class logistic regression. I simplified the eight categories of complaints to three targets, low priority, medium priority, and high priority.

Recommendation

Accuracy for finding high priority complaints came up to 71% in the tuned multinomial model. This is high enough to consider automated flags on TF-IDF top words, such as 'oil' -- words that are truly unique to the top priority cluster.

Another recommendation would be having complainants to tag entries to assist in identifying common complaints, such as leaks, odors, and wastewater leakage.

My final recommendation would be to educate Texans' on preventative measures they can take to mitigate wastewater flooding. Solutions such as permeable pavements, vegetale swales, and retention pools are explored by Texas Living Waters.

Future Considerations

I would like to compare frequency of complaint sites to see if many complaints revolve around a few bodies of water.

I'm interested to study water samples from high priority incident sites before and after complaints are marked 'addressed', to confirm the sites are left in condition equal or better than their original condition.

I hope to explore other possible dataframe features that could have strong predictive power for necessary response.

Repository Structure

├── Images <- Images and Graphs used in this project

├── .gitignore <- Contains list of files to be ignored from GitHub

├── presentation.pdf <- Slide Presentation of the project

├── README.md <- Contains README file to be reviewed

└── predict_priority.ipynb <- Jupyter notebook of the project containing codes and analysis