Predicting Covid-19 Hospitalization using EHR and Biomedical Text

Mohammad Bakir, Sheng-Ting Lin, Eric Pan

Georgia Institute of Technology

Abstract

The EHR DREAM Challenge[1] is a competition supported by various research institutes in Washington state to develop machine-learning models for supporting clinical care of COVID-19 patients. One challenge involves predicting hospitalization of COVID-19 patients within 21 days of testing positive for COVID-19. Building more reliable models can help healthcare providers better predict hospitalization needs of patients testing positive for COVID-19. We conducted a retrospective study using electronic health records (EHR) in predicting hospitalization using Logistic Regression, Support Vector Machine (SVM), and Random Forest models with dimensionality reduction methods and data resampling. Our synthetic Data trained prediction model has an AUC performance of 0.66 while our model has an AUC performance of 0.7204 on the University of Washington real patient dataset. In addition, we used Natural Language Processing (NLP) on the CORD-19 dataset[2] by identifying terms related to COVID-19 and using the results to guide feature engineering and selection. A model trained on only NLP-curated features resulted in a best AUC model performance of 0.7082.

Introduction

On March 11, 2020, the WHO declared Coronavirus Disease 2019 (COVID-19) a pandemic[3]. As of October 2020, over 36 million confirmed cases and 1 million deaths have been reported worldwide[4]. Approximately 20-30% of patients with COVID-19 require hospitalization, and 5-12% of those hospitalized may require critical care in an intensive care unit (ICU)[5]. Due to the potentially rapid escalation in disease severity, preemptive and preventative care can dramatically improve outcomes. Several COVID-19 related studies center on predicting the mortality of COVID-19 patients and identifying biomarkers correlating to development of a severe COVID-19 case [6,7,8]. In this project, we present binary supervised machine learning classifiers for predicting the probability of patient hospitalization within 21 days of a positive COVID-19 result. The effort is supported by COVID-19 DREAM Challenge, which provides synthetic EHR data for local model training and provide model metrics on model submissions by running training and evaluation on real EHR data in their private servers. We also introduce an NLP analysis on the separate CORD-19 dataset using tokenized Unified Medical Language System (UMLS) terminology to inform the feature selection process.

Data Overview

The COVID-19 DREAM Challenge data used for this study comprises of de-identified electronic health record (EHR) data collected from hospitals within the University of Washington Medial System. The data contains 10 years of clinical records (2010-2020) from 9,500 patients who have at least one RT-PCR test for COVID-19, and of which ~800 tested positive for the disease. From these records, a larger synthetic dataset of ~1200 COVID-positive patients are generated to protect privacy with 107 of the synthetic patients hospitalized.

While the DREAM challenge database focuses on patient-specific outcome, the CORD-19 dataset is a large collection of over 345,000 medical research publications related to coronavirus strains in general[2]. This dataset was first released in March 2020 is continuously maintained by the Allen Institute for AI. This project focuses on analyzing two datasets separately: the dataset provided for the DREAM challenge (the "DREAM data") and the CORD-19 data.

The DREAM data follows version 5.3.1 of the OHDSI OMOP Common Data Model (CDM)[9]. The dataset is sourced and organized based on 9 different OMOP CDM table schemas. The OMOP CDM schema use a discrete list of medical concepts IDs which is maintained by OHDSI. These IDs can be searched on OHDSI-maintained database[10] and provide a standardized representation of medical concepts. A binary "gold standard" hospitalization label is provided for each person in the training dataset. Both training and evaluation sets are provided for the challenge with the results evaluated on real EHR data hosted across different clinical institutes.

Methods

The overall project workflow involves two separate components: ETL and model generation using the DREAM dataset, and an independent NLP analysis using the CORD-19 dataset.

ETL and Exploratory Data Analysis

The ETL pipeline involved loading all the available columns containing concept IDs and then aggregating the counts per concept ID per patient (Table 1). Each unique concept ID in the dataset represents a potential feature to use in the training. To perform initial data analysis and model training, all unique concept IDs per patient were aggregated across both test and train datasets and used to generate the unfiltered list of features.

In addition to the counts of each concept ID, numerical values were parsed-out separately and introduced as additional features when available (using custom concept IDs). These value-based features include average values from measurements and observations and person age. Abnormal measurement counts were also parsed-out and introduced as custom concept IDs.

In the final feature matrix, each column represents a concept ID used as a feature and each row represents a patient. If the feature is count-based, the value of each cell represents the number of table entries containing that concept ID for that patient (default: 0). Otherwise, if the feature is value-based, the value of each cell represents average of recorded measurement/observations (default: average across all recorded measurements).

In this iteration, we planned to boost model performance using primarily clinical concepts that are less susceptible to demographic biases (e.g. race, gender, location). The split of clinical vs. non-clinical concepts selected is shown below in Figure 1.

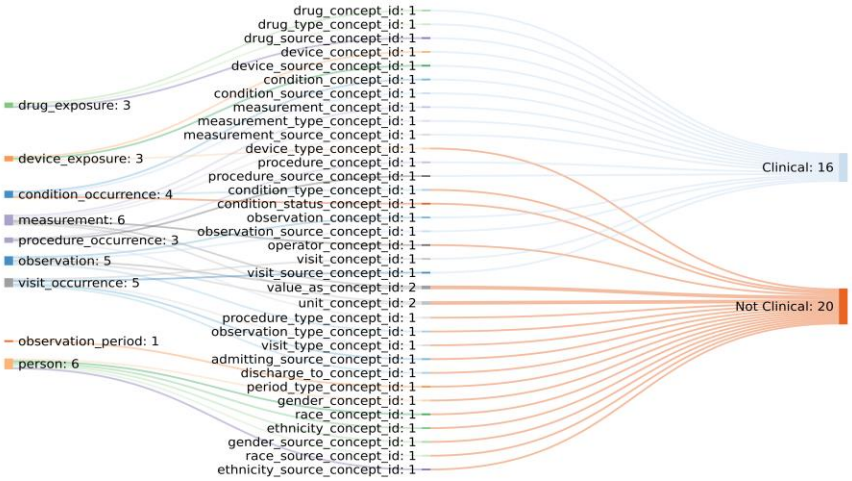

Figure 1. Manual clinical label assignment for each column containing concept IDs.

After filtering based on the "clinical" columns outlined above, there are a total of 2849 unique concept ID instances in the provided training and evaluation datasets. Given the wide breadth of medical conditions, over 90% of unique concept IDs appeared for less than 20% of the patient population. To address data sparsity, our initial analysis focused on identifying features that are reasonably present in the population and highly correlated with patient hospitalization. To identify promising concept IDs, we calculated the Pearson correlation coefficient of each feature to the training labels and ranked them based on the absolute value of the correlation.

Feature Engineering

Based on the results of the Pearson correlation, features in general do not have high correlation with training labels. Principle Component Analysis (PCA) was used to reduce features considered for modeling by preserving features that capture the majority of variance. Using PCA helped uncover the underlying feature signals hidden in our data, with approximately 1/3 of total feature set, representing 80% of variability in the dataset.

Due to an approximate 1:10 imbalanced ratio (Hospitalized vs Non-Hospitalized cases), the overall predictive performance of model was expected to be constrained in favor of the majority class. Using Synthetic Minority Oversampling Technique (SMOTE)[12] in the imblearn Python package, we oversampled the minority class to produce a balanced oversampled dataset with 1144 patients in each of the two classes. We also under-sampled the majority class to produce another balanced under-sampled dataset with 100 patients in each class, and measure the AUROC (Table 2) and AUPRC scores (Table 3).

Model Architecture and Metrics

In the modeling phase, binary classifiers were trained using the PCA extracted features to predict hospitalization. Post feature extraction and addressing class imbalance, the study population was split into an 80% training set and 20% test set. Logistic Regression, Support Vector Machine, and Random Forest classifiers were trained on the training set to predict hospitalization. Hyperparameter tuning was performed using gird search on 10-fold cross validation of the training set. The model performance of the best classifier was then evaluated on the test set.

The multiple machines learning models trained in the study were evaluated and compared using the Area under the Receiver Operating Characteristic Curve (AUROC) and Area under the precision-recall curve (AUPRC).

The AUROC curve summarizes the trade-off between the true positive rate and false positive rate for a predictive model using different probability thresholds. AUROC provides a single measure of the diagnostic ability of a binary classifier as its discrimination threshold is varied. We also use the AUPRC average precision score as the graph provide an average precision score, a useful performance metric for imbalanced data in a problem setting where importance is placed on finding positive examples. The two scores provided a full picture of model performance.

NLP on Biomedical Text

We used NLP tools to map each concept IDs from the DREAM dataset to a list of UMLS terms, and then find the total counts of each UMLS terms in research publications on COVID-19. Our goal was to identify the most relevant features (here: concept IDs) based on existing coronavirus research to guide our model training. We used spaCy[13], an NLP Python package, and scispaCy[14], a package containing pretrained models and pipeline tools to optimize the use of spaCy on scientific text. We ran the name-entity extraction on just the titles of the ~345k publications. Using scispaCy's pretrained models optimized for medical entities, we linked each paper title and concept name to the most likely candidate UMLS IDs.

We used *en_ner_bc5cdr_md*, a SciSpacy model trained on BC5CDR corpus to recognize disease and chemical terms. We added UMLS linker to the model's pipeline because it contains the largest number of concepts (~3 millions). To compensate for the imperfect recall, we set the model to return the top 5 candidates for any candidate with score greater than 0.7 as a suggested default threshold. The model generates a list of lists of UMLS IDs from each title and each concept name from the DREAM dataset. Finally, for each concept name, we count the number of times each of its 1~5 UMLS IDs appear in all the titles of COVID-19 research papers. We then aggregate all the counts for each concept name, and then sort the concepts.

We inspected the top five concepts with the highest aggregated counts and found that we cannot use these concepts to train the models on. They all contain the word "infection". Upon closer inspection, one of the UMLS IDs correspond to the entity "infection" appears in ~10k titles and ~79 concepts. This shows that if we simply aggregate the raw counts of the UMLs IDs for each entity, the concepts with the highest aggregated counts would be those with UMLs term that frequently but unspecific to each concept they appear in. Therefore, we used term frequency – inverse document frequency (TF-IDF) [17] to normalize for commonly occurring UMLS IDs:

$$tfidf(t,d,D) = tf(t,d) \cdot idf(t,D) = \log(f_{t,d}) \cdot \log\left(\frac{N}{n_t}\right)$$

where $f_{t,d}$ is the counts of the UMLS ID (t) in a document d, N is the total number of concepts, and n_t is the number of concepts containing t. We performed the same steps on the COVID-19 research abstracts. The top concepts in titles and abstracts are shown in Tables 6 and 7.

Model Deployment on DREAM Challenge Server

The environment setup and model training are built into a Docker image that adheres to the DREAM Challenge's format. The image is then uploaded to the DREAM Challenge's Docker Hub and can then be submitted to the University of Washington's sever to be run on real EHR data. We then would get notified the AUROC and AUPR scores of the model trained and evaluated on the real data, as well as minimal logs for any debugging purposes.

Experimental Results

Exploratory Data Analysis on Synthetic Dataset

Table 1 shows some example concept IDs, their correlation magnitudes, and information on data sparsity. The highest correlation magnitude was less than 10% for any given feature (0.09602). We used these results to guide feature selection for training the model.

Concept ID	Concept Name	Correlation Magnitude	% Populated in Provided Data	Average # of Instances Per Patient (when present)
201826	Type 2 diabetes mellitus	0.0875	53.00%	1.4510
77670	Chest pain	0.0875	26.46%	1.1964
436659	Iron deficiency anemia	0.0087	23.90%	1.1438
436962	Insomnia	0.0087	25.90%	1.1605
4196147	Peripheral oxygen saturation	0.0058	87.53%	4.9151

Table 1. Some example concept IDs with metadata after aggregating the counts in the ETL step. Correlation magnitudes were calculated using the Pearson Coefficient Correlation.

Model Performance on Synthetic Dataset

Tables 2 and 3 compare the performances of the trained models on the two balanced synthetic data test sets. The results show that our models had no discriminatory capacity to distinguish between positive negative classes. All three models used 464 features identified from PCA.

Model	Under-sampled		Oversampled	
	Train	Test	Train	Test
Logistic Regression	0.84	0.66	0.85	0.50
Support Vector Machine	0.76	0.62	0.75	0.51
Random Forest	0.79	0.55	0.77	0.52

Table 2: AUROC test scores of classification models based on filtered features from PCA

Model	Under-sampled		Oversampled	
	Train	Test	Train	Test
Logistic Regression	0.16	0.091	0.098	0.078
Support Vector Machine	0.13	0.072	0.082	0.072
Random Forest	0.19	0.076	0.073	0.069

Table 3: AUPRC test scores of classification models based on filtered features from PCA

Our best performing AUC and AUPRC model was the Logistic Regression model using features selected from PCA. Per the performance metric summary below, our model can identify non-hospitalized patients with a high degree of precision, and recall, with scores of .84 and .87 respectively. The model however lacked the ability to identify hospitalized patients with precision and recall test scores of 0.5. This showed that our models overfit the training data, impacting out of sample test performance.

Actual	Predicted Non-Hospitalized	Predicted Hospitalized
Non-Hospitalized	54	8
Hospitalized	10	8

Table 4: Confusion Matrix of the best performing model - Logistic Regression Model (AUROC = 0.66)

Class	Precision	Recall	F1-Score
Non-Hospitalized	0.84	0.87	0.86
Hospitalized	0.50	0.44	0.47

Table 5: Performance Metrics of the best Logistic Regression Model

Model Performance on Synthetic Dataset using NLP-Curated Features

From the ~345,000 titles, ~732,000 UMLS IDs were extracted but only ~29,000 IDs are unique. For the concept names, there are also about ~29,000 unique UMLS IDs extracted. Using PySpark, we calculated word count on the UMLS IDs from the publication titles, and then aggregated the counts for IDs belonging to the same concept. Tables 6 and 7 below show the top-weight concepts from paper titles and abstracts respectively.

Concept ID	Concept Name	Weights
19131544	Ascorbic Acid 50 MG / Calcium Sulfate 250 MG / Cholecalciferol 400...	741.71
40233612	Calcium ascorbate 60 MG / … Vitamin B 12 0.025 MG Oral Capsule ...	707.28
44783628	Pulmonary hypertension due to lung disease and/or hypoxia	563.74
19131481	Calcium ascorbate 60 MG / .... Vitamin B 12 0.01 MG Oral Table	546.86
40173507	0.5 ML Streptococcus pneumoniae serotype 1 capsular antigen diphtheria ...	507.69

Table 6: Top 5 concepts weighted with TF-IDF frequency using research titles.

Concept ID	Concept Name	Weight
19131544	Ascorbic Acid 50 MG / Calcium Sulfate 250 MG / Cholecalciferol 400 ..	58508
40233612	Calcium ascorbate 60 MG / … Vitamin B 12 0.025 MG Oral Capsule	58508
19131481	Calcium ascorbate 60 MG / .... Vitamin B 12 0.01 MG Oral Table	58489
44783628	Pulmonary hypertension due to lung disease and/or hypoxia	57753
4098740	Vitamin B12 deficiency anemia due to malabsorption with proteinuria	57439

Table 7: Top 5 concepts weighted with TF-IDF frequency using research abstracts.

We used the list of concept IDs identified from the titles and abstracts to generate the feature matrix with the results shown in Table 8 below. Unfortunately, the models performed poorly which suggests that using only generalized coronavirus knowledge may not be informative enough and future approaches require more specific literature.

Model	Biomedical Text Titles		Biomedical Text Abstracts	
	AUC	AUPRC	AUC	AUPRC
Logistic Regression	0.50	0.13	0.50	0.15
Support Vector Machine	0.51	0.14	0.52	0.13
Random Forest	0.53	0.12	0.55	0.16

Table 8: Performance of the models using only concept IDs identified from the CORD-19 analysis.

Model Performance on DREAM Challenge Dataset

ML model type	Features	AUROC	AUPR
*Logistic regression	Age, race, gender	0.6911	0.1346
Logistic regression	Concepts from abstracts	0.7082	0.1284
Logistic regression	Concepts from titles	0.5815	0.1653
Logistic regression	Concepts from correlation	0.7204	0.1295

Table 9: results reported from DREAM challenge platform. *: baseline model provided by DREAM Challenge

Table 9 above shows our model results using the DREAM challenge dataset. Unfortunately, we were unable to deploy our outlined approach in time for this project submission. However, we were able to deploy the baseline model provided by the DREAM challenge as a performance comparison for future submissions. As later outlined in the Discussion and Conclusion, we plan to continue our efforts after the conclusion of this course.

Discussion

Model Development

During the model development, we monitored the public leaderboards provided by the DREAM challenge to better understand the most predictive features in the dataset. The current best models rely on non-clinical concepts like a patient's race, gender, and ethnicity. While these are metrics we plan to incorporate in future iterations, the use of these features may introduce unethical biases which remains an open research issue for machine learning algorithms [15, 16]. The DREAM challenge is scored on AUC and AUPRC scores of trained models. The model choices we made revolved around the need to reduce overfitting of the trained classifier due to the complex, and feature rich dataset. The best performing model, Logistic Regression, overfit the training set due the noisy complex dataset. Although feature size was reduced using PCA, the high sparsity and similarity of clinical concepts made it difficult for model training.

Evaluating model using the synthetic data may also be unreliable, based on the model metrics on test data from synthetic data versus Dream Challenge's real EHR data. The DREAM Challenge website does not provide how they generated the synthetic data and how representative they are to the real world. The only way to properly assess a model is through a submission to DREAM Challenge to have a model evaluated against the patient EHR. However, submission is limited to once per day, and three times a week. Logs on failed submissions are not included in the feedback to protect patient privacy, making debugging and model development process. Therefore, we were not able to try out more models near the end of the project using the real EHR data. Based on the results we have, the features selected from using NLP and correlation scores have improved performance when evaluated with AUROC scores, but lower in AUPR metric, which is more useful in imbalanced dataset.

Concepts Obtained from NLP

Concepts obtained from NLP pipelines did not improve the model evaluation metrics much. The use of the TF-IDF reduces the repeating concepts retrieved initially. However, as seem in Tables 6 and 7, some top concepts are still closely related and are not intuitively linked to COVID-19. This suggest that we may need to explore other weighting schemes such as other TF-IDF variants, and perhaps try other vocabularies set such as MeSH terms. We would also need a more

automatic way to inspect the result from NLP instead of manually looking at the top k concepts to help us efficiently improve the NLP pipeline.

Project Challenges

The main challenges during implementation are focused in primary two areas: data complexity and deployment troubleshooting. In our case, the data complexity related to the high variety, high sparsity, and high similarity of many clinical concepts while the deployment troubleshooting is more difficult due to patient privacy concerns.

While the OMOP data schema is clearly documented and well structured, the challenge of generating features from the dataset is that the available clinical feature set is entirely dependent on the concept IDs found in the available training data. A concept ID that shows up for one patient is not guaranteed to show up for another – in fact for most concepts it is highly likely it does not appear in novel datasets. This is due to the nature of rare medical conditions and the very specific assignment of some of the concepts (e.g. for the same medication, different concept IDs are assigned for different dosages). Thus, while we were able to identify promising concept IDs to use as features, the given features were not guaranteed to be recorded for the final evaluation set. Finally, given the high similarity between some of the concepts, it was difficult for the NLP algorithm to effectively segment and identify distinctly useful features.

An additional challenge involved actual deployment to the DREAM platform to train on the real EHR data. Due to patient privacy concerns, the infrastructure error log must be accessed via support tickets which have slow response and iteration times. This made it difficult to perform more complex ETL operations and receive feedback on the availability of features needed in order to complete our deployment in a faster cadence. While prohibitive, it is ultimately a good thing that patient privacy is taken seriously, and the current infrastructure makes traditionally inaccessible data available for ML projects. This is an example for future citizen science projects on data with privacy regulation.

Conclusion and Future Optimization

In this paper we present a supervised machine learning approach that returns the hospitalization classification of a person with positive COVID-19 test result. We train our model using all concept IDs present in the data as feature and applied techniques including Pearson Correlation, PCA, and SMOTE to improve the consistency of the model. We also mined insights from the CORD-19 dataset and use PySpark to summarize the NLP result from ~345,000 articles.

In future iterations, we plan to improve this project in the future by compressing the feature set using approaches including multicolinarity analysis, manual concept grouping, and clustering analysis to algorithmically identify concept groups. While the Pearson Correlation was used to identify features with high correlation to the gold standard hospitalization label, the correlation information between features was not incorporated. By adding a pairwise comparison between features, similar correlation features can be identified and removed or grouped into a single feature to reduce the size of the feature set. On the NLP side, we can explore different entity linker to other knowledge bases, such as MeSH (Medical Subject Headings) and other models that can recognize of entities other than drugs and conditions. Ultimately, we can improve our model by further organizing our results and identifying a minimum size, maximum effectiveness

feature set. We can also incorporate custom features, such as gender, age, and race as the baseline model from DREAM Challenge. We also hope that our application of NLP in the feature selection process encourages novel and creative approaches to traditional feature selection.

References

1. SageBionetworks. COVID-19 DREAM Challenge [Internet]. Available from: https://www.synapse.org/#!Synapse:syn21849255/wiki/601865
2. Lu Wang L, Lo K, Chandrasekhar Y, Reas R, Yang J, Eide D, et al. CORD-19: The Covid-19 Open Research Dataset [Internet]. ArXiv. Cornell University; 2020 [cited 2020Nov15]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7251955/
3. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020 [Internet]. World Health Organization. World Health Organization; 2020 [cited 2020Oct10]. Available from: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020
4. WHO Coronavirus Disease (COVID-19) Dashboard [Internet]. [cited 2020Nov15]. Available from: https://covid19.who.int/
5. Lu Wang L, Lo K, Chandrasekhar Y, Reas R, Yang J, Eide D, et al. CORD-19: The Covid-19 Open Research Dataset [Internet]. ArXiv. Cornell University; 2020 [cited 2020Nov15]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7251955/
6. Yan L, Zhang H-T, Goncalves J, Xiao Y, Wang M, Guo Y, et al. An interpretable mortality prediction model for COVID-19 patients. Nature Machine Intelligence. 2020;2(5):283–8.
7. Tian W, Jiang W, Yao J, Nicholson CJ, Li RH, Sigurslid HH, et al. Predictors of mortality in hospitalized COVID-19 patients: A systematic review and meta-analysis. Journal of Medical Virology. 2020;92(10):1875–83.
8. Shang W, Dong J, Ren Y, Tian M, Li W, Hu J, et al. The value of clinical parameters in predicting the severity of COVID-19. Journal of Medical Virology. 2020;92(10):2188–92.
9. OMOP CDM v5.3.1 [Internet]. cdm531.utf8. [cited 2020Nov15]. Available from: https://ohdsi.github.io/CommonDataModel/cdm531.html
10. Athena. 2020 [cited 2020Nov15]. Available from: https://athena.ohdsi.org/
11. ACL 2020 Workshop NLP-COVID [Internet]. OpenReview. [cited 2020Nov15]. Available from: https://openreview.net/group?id=aclweb.org%2FACL%2F2020%2FWorkshop%2FNLP-COVID
12. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 2002;16:321–57.
13. Honnibal M, Montani I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2017.
14. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. Proceedings of the 18th BioNLP Workshop and Shared Task. 2019;
15. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Intern Med.* 2018;178(11):1544–1547. doi:10.1001/jamainternmed.2018.3763
16. Ledford H. Millions of black people affected by racial bias in health-care algorithms [Internet]. Nature News. Nature Publishing Group; 2019 [cited 2020Nov15]. Available from: https://www.nature.com/articles/d41586-019-03228-6
17. Zhang W, Yoshida T, Tang X. A comparative study of TF* IDF, LSI and multi-words for text classification. Expert Systems with Applications. 2011 Mar 1;38(3):2758-65.

Team Contributions

Team Repo: https://github.gatech.edu/epan9/BD4H-Team58-Submission

- The repo contains the commit history and separate branches indicating the primary split in workload.
- Main Shared Branch: master

Mohammad Bakir: Model Generation + Evaluation

- Individual branch: mohammad, deploy

Sheng-Tin Lin: NLP Analysis

- Individual branch: santina/nlp , deploy

Eric Pan: ETL Pipeline

- Individual branch: eric

All members contributed equally to the Final Report and Presentation