Capstone 2: Milestone Report

Predicting Hospital Readmission Using Unstructured Clinical Notes

Problem Statement

In order to improve patient outcomes through coordinated care and post-discharge planning, the Affordable Care Act established the Hospital Readmissions Reduction Program (HRRP). To incentivize upgraded patient care, the HRRP lowers Medicare payments to hospitals with too many readmissions. The excess readmission ratio (ERR) is measured for the following conditions to determine reimbursement payments to hospitals:

- Acute Myocardial Infarction (AMI)
- Chronic Obstructive Pulmonary Disease (COPD)
- Heart Failure (HF)
- Pneumonia
- Coronary Artery Bypass Graft (CABG) Surgery
- Elective Primary Total Hip Arthroplasty and/or Total Knee Arthroplasty (THA/TKA)

In addition to procedural updates and training, hospitals would also likely benefit from the ability to predict which patients are most at-risk for readmission in order to target additional care toward those patients. Rajkomar et al used deep neural networks to predict hospital readmissions with a ROCAUC score of 0.75-0.76. Manyam used survival analysis and deep learning with structured clinical data to predict hospital readmissions with a C-statistic of 0.712. For this project, we will use natural language processing and deep learning with unstructured clinical notes.

Data

The model will use data from the MIMIC-III database, which is comprised of over 58,000 hospital admissions for 38,645 adults and 7,875 neonates. To access this data, I first had to complete a series of training on human subject research ethics and agree not to share data from the database. Once access was acquired, I set up a PostgreSQL database in order to access the data on my local machine. Data about admissions date and type (emergency, elective, urgent, or newborn) was pulled from the ADMISSIONS table. Text data was pulled from the NOTEEVENTS table. In particular, discharge summary and nursing notes were used from the NOTEEVENTS table. All notes for a given patient and hospital visit were concatenated into a single list of notes.

The ADMISSIONS data was pulled into a Pandas dataframe. Three new features were created:

- `next_admission`: provides the datetime of the next hospital admission for each subject and each hospital admission
- `next_admission_type`: a categorical description of whether the hospital visit was elective, emergency, urgent, or newborn.
- `days_between_admit`: the number of days between admission dates. This will be used to determine the target variable.

Since we are looking for 30-day readmissions that were unplanned, the `next_admission` data was deleted for admissions defined as 'elective'.

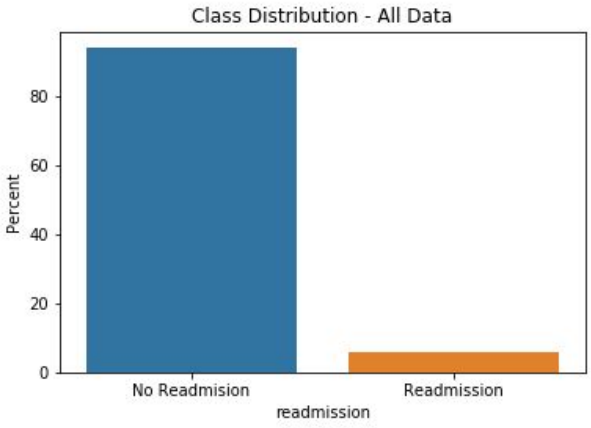
The NOTEVENTS data was pulled from the database into a Pandas dataframe. Since there are multiple nursing notes and a discharge summary for each hospital admission (HADM_ID), we must combine the notes into a single entry based on HADM_ID and patient (SUBJECT_ID). First, duplicate notes were dropped from the dataframe. Duplicate notes were determined by the exact same text and HADM_ID. The dataframe was then grouped by SUBJECT_ID and HADM_ID and all notes were concatenated into the TEXT column. This allows each row of the dataframe to represent one hospital visit.

Next, the admissions and notes dataframes were merged. The percent of rows missing text data were then calculated for each admission type. Since over half of the NEWBORN admissions were missing, all NEWBORN admissions were dropped from the dataframe. The HRRP does not consider neonatal readmissions, so this should not affect our model. The target variable (`readmission`) was created by marking a 1 for all rows where `days_between_admit` was less than 30 and 0 for all others.

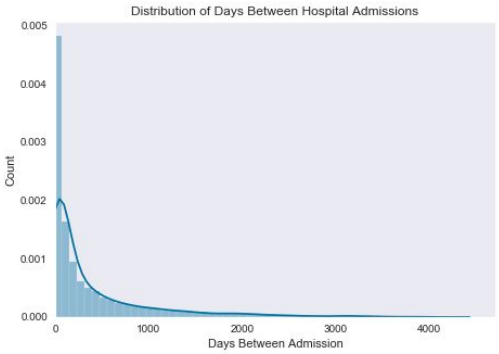
Admission Type	Percent Missing Text
NEWBORN	53.67
EMERGENCY	3.24
ELECTIVE	4.49
URGENT	4.12

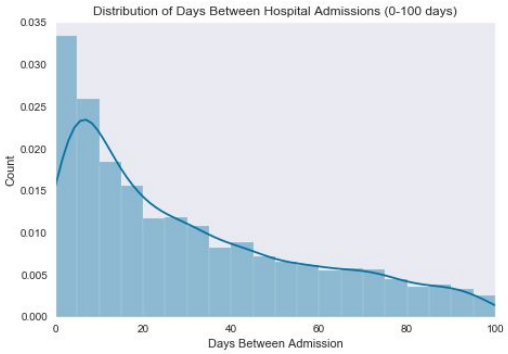
At this point, 30% of the dataset was put aside as a test set. This will not be used until we are ready to test the final model. The remaining 70% will be split into a training and validation set to be used in model building. The text data from both datasets was cleaned by filling empty text cells with an empty string, converting the list of notes for each hospital visit into a single string, and removing all instances of '\n' (otherwise the letter n becomes the most prominent feature since it demarks all line breaks). The datasets are then exported as separated train and test CSV files.

Exploratory Data Analysis

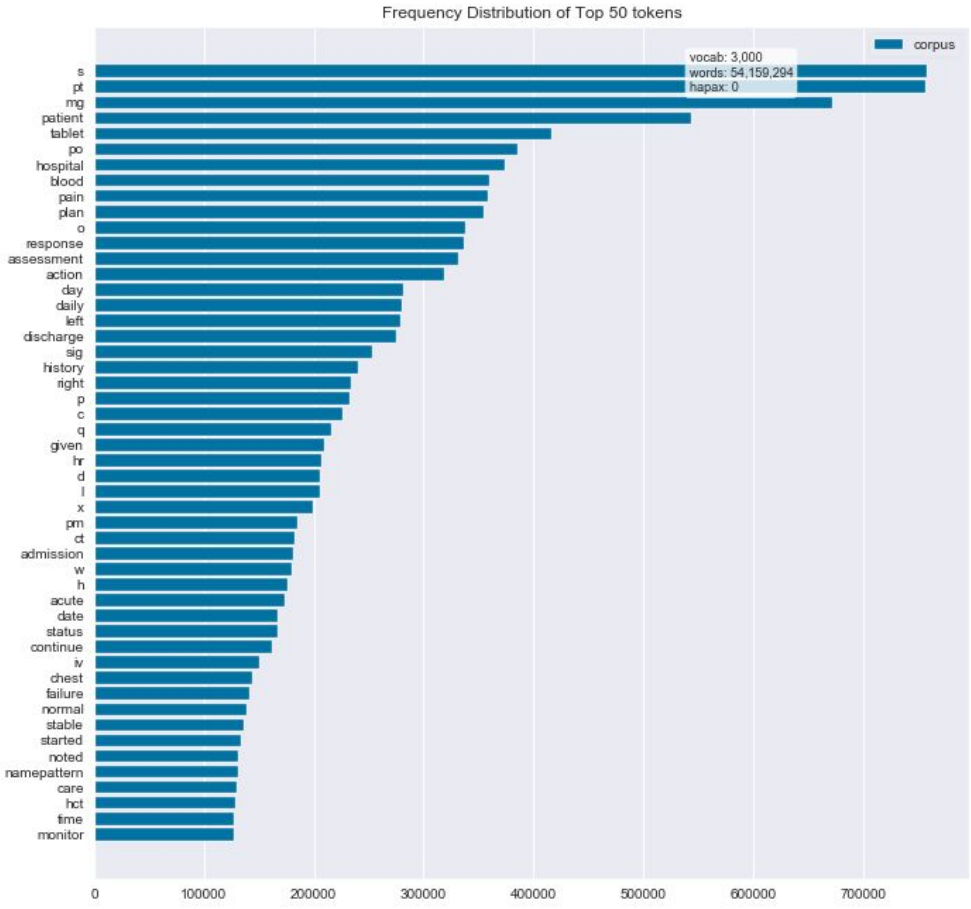


With only 5.88% of samples in the positive class (readmission), there is a fair amount of class imbalance. This will be accounted for as we move forward into training models. A histogram of the days between admissions indicates that most urgent and emergent readmissions occur within the first 100 days after discharge. If we zoom in to the first 100 days after discharge, we can see that the number of readmissions peak in the first 30 days and start to trail off after that. This helps to explain why it is so important to get past those 30 days after discharge. If a patient makes it more than 30 days without readmission, they may be less likely to have to return at all.

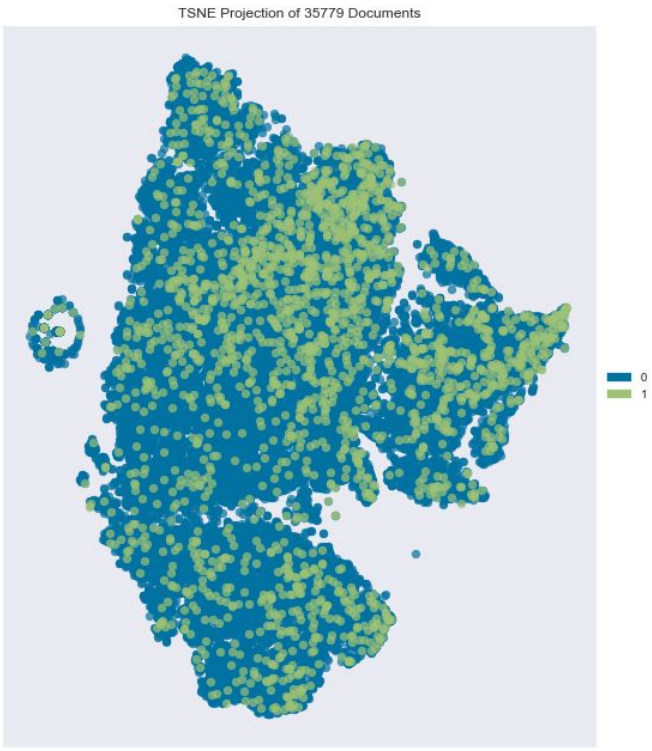




Using a quick count vectorizer, we can get a visual sense of the text data. First, a frequency distribution chart shows that the most common words, after dropping stop words, are common medical jargon (e.g. patient, blood, plan, assessment) and abbreviations (e.g. s = sign/symptom, pt = patient, p.o. = by mouth). We might be able to improve the model by removing frequent words that do not add much information, but for now we will keep them in.



To get a sense of whether the two classes can be divided using a clustering algorithm, a tSNE map colored by class is generated. While there do appear to be some clusters, they are decidely not divided by class. Instead, the two classes seem to share fairly even distributions. Therefore, clustering will likely not be a fruitful approach for this problem.



References

1. Hospital Readmission Reduction Program. (2018, December 04). Retrieved December 18, 2018, from https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/HRRP/Hospital-Readmission-Reduction-Program.html
2. Long, A. (2018, June 04). Introduction to Clinical Natural Language Processing: Predicting Hospital Readmission with... Retrieved December 12, 2018, from https://towardsdatascience.com/introduction-to-clinical-natural-language-processing-predicting-hospital-readmission-with-1736d52bc709
3. Manyam, R.B., et al. (2018). Deep Learning Approach for Predicting 30 Day Readmissions after Coronary Artery Bypass Graft Surgery. Machine Learning for Health (ML4H) Workshop, NeurIPS 2018. arXiv:1811.07216
4. MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). DOI: 10.1038/sdata.2016.35.
5. Rajkomar, A., et al. (2018). Scalable and accurate deep learning with electronic health records. Npj Digital Medicine, 1(1). doi:10.1038/s41746-018-0029-1