E. Chris Lynch

Using natural language processing on clinical notes to predict hospital readmission

Comparing bag-of-words and word embedding models

ABSTRACT

30-day hospital readmissions have been targeted as a key metric of patient care. In 2012, the Affordable Care Act initiated the Hospital Readmission Reduction Program (HRRP) to incentivize improved patient outcomes by financially penalizing hospitals with excessive readmission rates. According to the American Hospital Association, in the first five years of the HRRP, hospitals experienced $1.9 billion in penalties.1 This project uses 283,208 clinical notes (nursing and discharge summary) on 35,779 patient admissions from the MIMIC-III database.2 Various natural language processing models were assessed in their ability to predict all-cause 30-day readmission for all patients (excluding neonates), including word embeddings and pre-trained language models. A skip-gram word embedding model using Word2Vec with Latent Dirichlet Allocation (LDA) proved to be the best fit model based on a ROC-AUC of 0.7078. Future models should focus on improving recall without sacrificing ROC-AUC or precision.

All code available at github.com/TheeChris/springboard/tree/master/Capstone_2.

INTRODUCTION

The Problem

In order to improve patient outcomes through coordinated care and post-discharge planning, the Affordable Care Act established the Hospital Readmissions Reduction Program (HRRP). The HRRP incentivizes upgraded patient care by lowering Medicare payments to hospitals with too many unscheduled, 30-day 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)

For this project, we will use clinical notes and natural language processing to predict 30-day readmission. While the HRRP currently only takes the six diagnoses above into consideration, this project uses data from all non-neonatal patients since the HRRP will likely continue to expand its definition. The goal is to help the hospital predict which patients are most at-risk for readmission in order to target additional care toward those patients.

About the Data

This project uses anonymized patient data from the MIMIC-III database. MIMIC-III contains electronic health record data on over 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012.

After completing the required course in the ethical use of human data in research, the MIMIC-III database was accessed as 26 CSV files totaling 6.2GB. These files were loaded in a local PostgreSQL database for easy access and manipulation. All data for this project was pulled from the NOTEEVENTS and ADMISSIONS tables. 223,556 nursing notes and 59,652 discharge summary notes were pulled from the NOTEEVENTS table. 58,976 rows of admissions data were pulled from the ADMISSIONS table. The ADMISSIONS and NOTEEVENTS data were pulled into two separate pandas dataframes in order to be cleaned and processed.

Data Pre-Processing

First, three new features were created for the admissions dataframe:

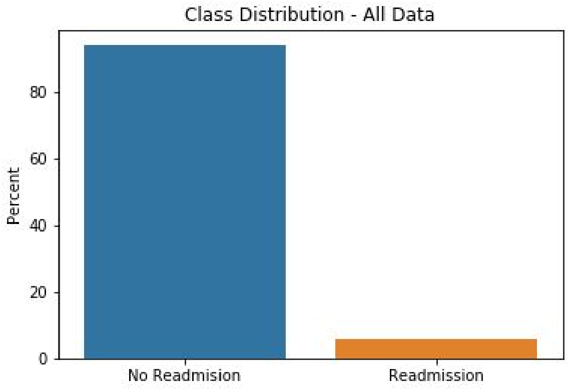
- `next_admission`: provides the DateTime of the next hospital admission for each subject and each hospital admission
- `next_admission_type`: a categorical description of whether the hospital visit was elective, emergency, urgent, or newborn.
- `days_between_admit`: the number of days between admission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

In order to measure unplanned 30-day readmissions, the `next_admission` and `next_admission_type` data were marked empty for admissions defined as 'elective'.

There were multiple nursing notes and a discharge summary for each hospital admission (HADM_ID). Therefore, the notes for each admission were combined into a single entry. First, 22,742 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 as a list. This allowed each row of the dataframe to represent one hospital visit.

Next, the admissions and notes dataframes were merged on SUBJECT_ID and HADM_ID. The percent of rows missing text data were then calculated for each admission type. Since over half of the NEWBORN admissions were missing (and we are not currently concerned with newborn readmission rates), all NEWBORN admissions were dropped from the dataframe. 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

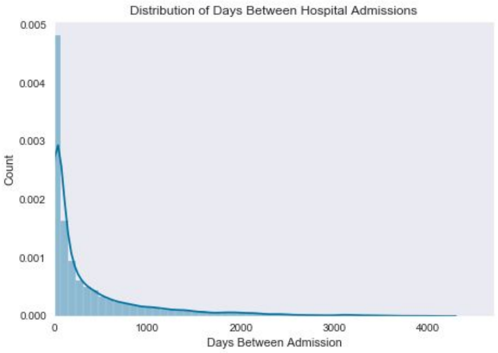


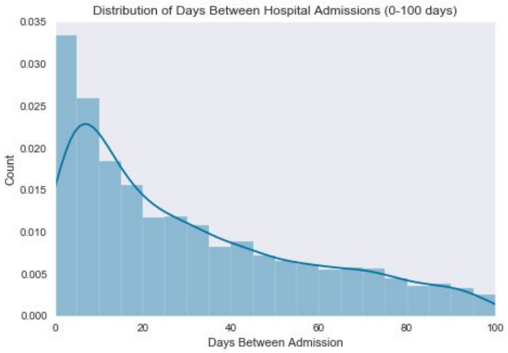
30% of the dataset was put aside as a test set, which will be used to evaluate the final model. The remaining 70% will be split into a training and validation set to be used in model building. It is obvious from the chart to the right that this is an imbalanced dataset with only 5.88% of the data being in the positive class. For this reason, the data was stratified when split so that the training dataset contained a similar proportion (6.02%).

The text data from both datasets were 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.

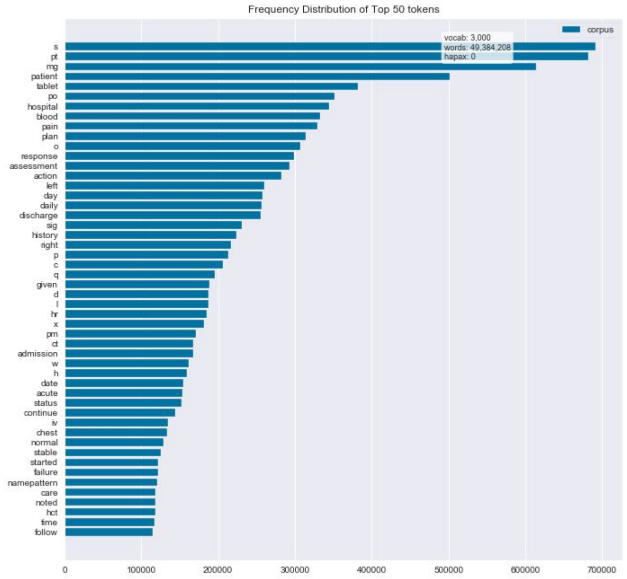
EXPLORATORY DATA ANALYSIS

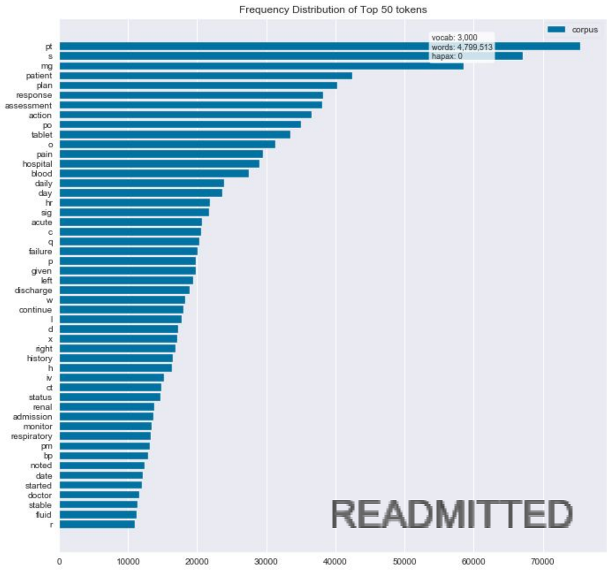
A histogram of the number of days between admissions indicates that most urgent and emergent readmissions occur within the first 100 days after discharge. If we zoom in to the first 100 days after discharge, we can see that the number of readmissions peak in the first 30 days and start to trail off after 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.

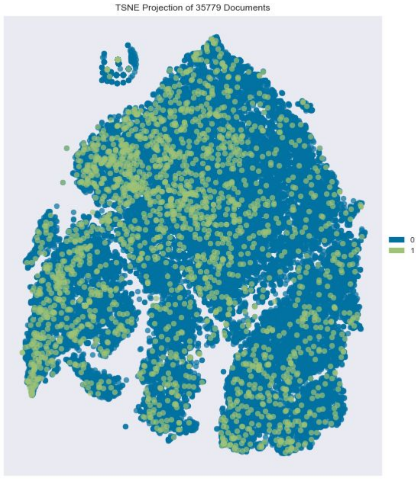

Distribution of Days Between Hospital Admissions


Distribution of Days Between Hospital Admissions (0-100 days)

The bar charts below give a sense of common words used in the clinical notes of patients who were and were not readmitted. First, the frequency distribution charts 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). In addition, the charts illustrate that while the order and counts may change, both sets of patients seem to share a similar lexicon. This indicates that the predictive ability of the model will likely be nuanced and difficult to explain without in-depth exploration.


Frequency Distribution of Top 50 tokens


Frequency Distribution of Top 50 tokens

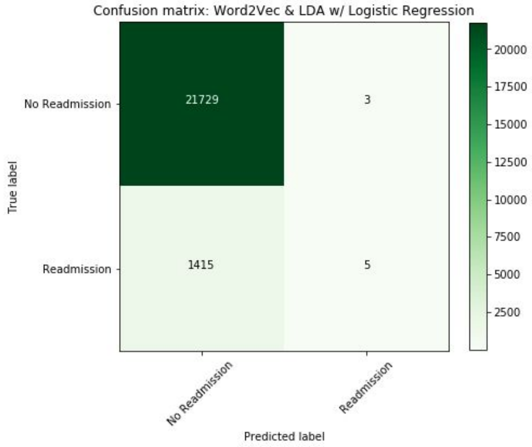

TSNE Projection of 35779 Documents

The t-SNE plot to the left shows that there do appear to be some clusters in the data, but they are decidedly not divided by class. Instead, the two classes seem to share fairly even distributions throughout much of the dataset. This is not surprising given the difficult nature of accurately predicting hospital readmission.

PREDICTIVE MODELING

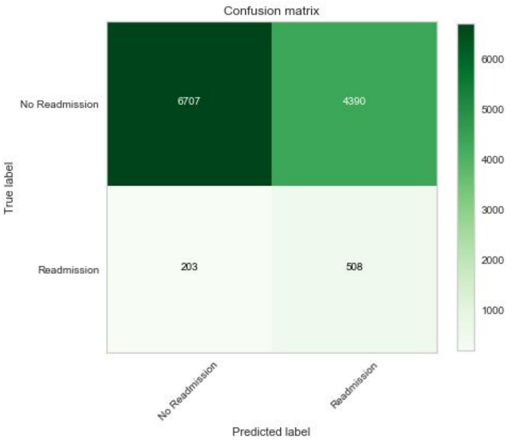
A traditional bag-of-words was the first approach used to predict the 30-day readmission status of patients. This provides a baseline of performance to compare against more computationally expensive neural network models. Receiving Operating Characteristic - Area Under the Curve (ROC-AUC) was used to measure the trade-off between sensitivity and specificity. Since readmissions only account for a small percentage of hospital admissions and the primary objective is to capture hospital readmission, recall and precision are computed to capture the model's ability to accurately predict readmission. To account for the imbalanced data set, the initial bag-of-words approach using logistic regression was tested with random under-sampling and over-sampling with SVM-SMOTE. Word2Vec models using random forests were trained but abandoned as they performed worse than logistic regression.

	ROC-AUC	Precision	Recall
Word2Vec & LDA w/ Logistic Regression	0.7078	0.6250	0.0035
Random Forest (under-sampling)	0.7076	0.1052	0.7145
Random Forest w/ TF-IDF (under-sampling)	0.7059	0.1093	0.6793
SVM (under-sampling)	0.6972	0.1096	0.6399
Logistic Regression (under-sampling)	0.6958	0.1143	0.5724
Word2Vec w/ Logistic Regression (under-sampling)	0.6924	0.6347	0.6620
Word2Vec & LDA w/ Logistic Regression (under-sampling)	0.6905	0.6387	0.6324
Logistic Regression (SVM-SMOTE)	0.6041	0.1558	0.1181



While Word2Vec with Latent Dirichlet Allocation (LDA) and no under-sampling had a better ROC-AUC score, the model's recall suffers greatly. This may be fixed by changing the threshold by which the model predicts readmission. If we can maintain this model's precision while improving recall, Word2Vec with LDA will likely prove to be the best fit model.

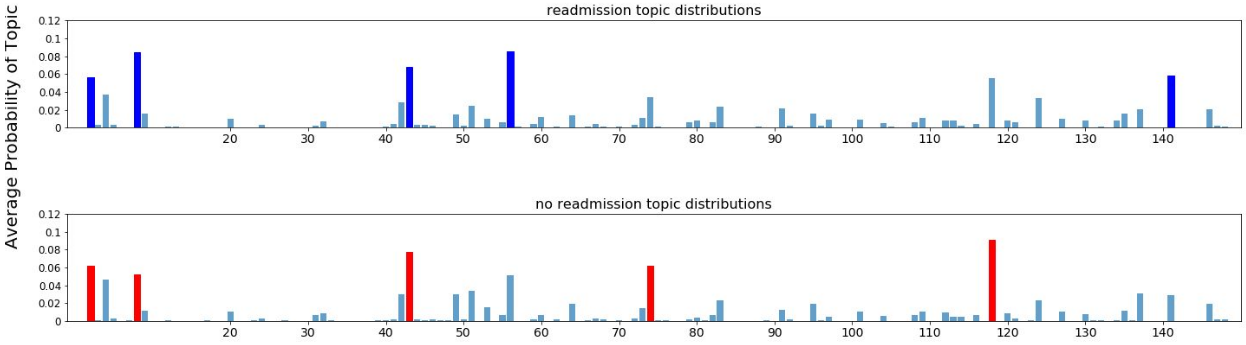
Random forests with random under-sampling and bag-of-words appear to produce the best fit model indicating that it can predict hospital readmission with 71.45% recall and an AUC of 0.7076, but this comes at the cost of a lot of false positives. Approximately 40% of patients who were not readmitted were falsely predicted to readmit. In a clinical setting, this may result in costly care that is not needed. For this reason, improving the model would need to focus on improving precision.



Feature Engineering

For the bag-of-words models, the clinical notes were cleaned by removing punctuation and digits, converting to lowercase, and tokenized. Word2Vec and LDA texts were also stemmed and lemmatized. For Word2Vec, we used a Skip-Gram model with a window of 6 words and 5 noise words to model the corpus into 200 features.

LDA has the benefit of grouping the text into topics. The clinical notes were modeled into 150 topics and their distributions are illustrated in the graph below. We can see from the graph that while the two groups share some of the top topics, there are differences that would likely help with modeling.



We can also look at the top words from the top topics of each group to see if any themes appear in the topics. For example, topics that are common in both (2, 8, 43) appear to be cardiac patients, patients at hospital1, and status updates.

Top Topics for Readmitted Patients	
Topic	Top Words
2	tablet, daili, sig, cardiac, ventricular
8	tablet, daili, sig, hospital1, pt
43	statu, unit, show, number, also
56	tablet, sig, daili, need, capsul
141	daili, tablet, cultur, sig, neg

Top Topics for Not Readmitted Patients	
Topic	Top Words
2	tablet, daili, sig, cardiac, ventricular
8	tablet, daili, sig, hospital1, pt
43	statu, unit, show, number, also
74	tablet, daili, sig, disp, refil
118	arteri, coronari, qd, postop, statu

Improving the Model

The only parameter of the Word2Vec model using LDA and logistic regression that was evaluated using grid search was the penalty term (l1 or l2). This model could be iterated upon through a more exhaustive grid search that looks at altering such hyperparameters as the window size, the learning rate, number of epochs, and downsampling threshold. As mentioned above, the threshold (or class weight) could be altered to improve the model's recall.

To improve the random forest model, two hyperparameters were modulated and validated against a test set of data: the size of the trees (n_estimators) and the number of branches (max_depth). Models were also tested using term frequency-inverse document frequency (TF-IDF), but this seems to slightly hurt the model performance. This may indicate that certain words that help define one or both of the classes are very frequent in the respective class and are, therefore, penalized by inverse document frequency. The best model used 500 estimators and a depth of 50. A more extensive grid search on the random forest hyperparameters, as well as additional feature engineering (perhaps adding LDA to the random forest model) may be able to improve predictive power or reduce the number of false positives.

While the discharge summary and nursing notes should convey the the majority of patient visit information (procedures, diagnosis, lab results, etc), it is also possible that adding these additional features explicitly may improve model performance.

MOVING FORWARD

The model could be productionized to give a real-time prediction of a patient's readmission risk. Examining feature importance through hypothesis testing could provide further insight into what factors need to be focused on when patients are marked as at risk for readmission. While we have seen promising ROC-AUC scores, it is advised to define acceptable precision and recall scores before a minimally-viable product is produced.

CONCLUSION

30-day readmission for all patients can be predicted using unstructured nursing and discharge summary notes. Prediction accuracy proved to be best with a word embedding model. This model compares well to previous, procedure-specific models such as the study from Manyam et al[3] (ROC-AUC: 0.712), but can likely be improved through additional feature engineering and hyperparameter tuning. Training a language model on a large set of clinical notes would also likely improve prediction results. The next step in production is to create a readmission flag for clinicians to help them better determine which patients may need additional care before discharge or directed post-discharge follow-up care. This can be A/B tested to see if it significantly improves patient outcomes.

REFERENCES

1. American Hospitals Association. AHA Fact Sheet: Hospital Readmissions Reduction Program. https://www.aha.org/other-resources/2016-01-18-aha-fact-sheet-hospital-readmissions-reduction-program. Last accessed January 29, 2019.

2. 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.

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