Using natural language processing on clinical notes to predict hospital readmission

Hospital Readmission Reduction

- targeted as a key metric of patient care
- 2012: Affordable Care Act initiated the Hospital Readmission Reduction Program
- Incentivize improved patient outcomes by financially penalizing hospitals with excessive readmission rates
- $1.9 billion in penalties in first 5 years (American Hospital Association)

About The Data

Data Collection

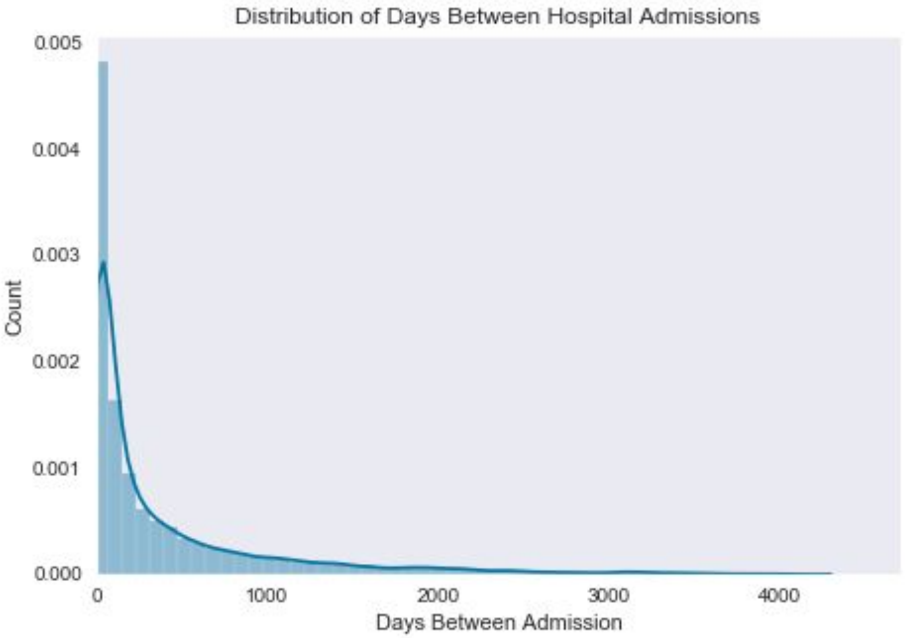
MIMIC-III version 1.4

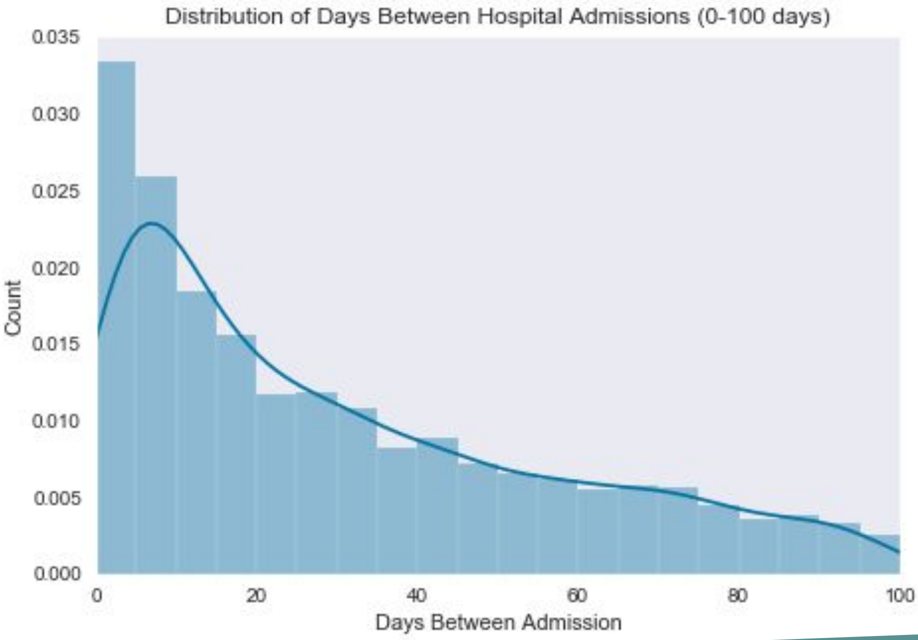
- over 58,000 hospital admissions from critical care units of the Beth Israel Deaconess Medical Center
- 38,645 adults and 7,875 neonates
- data spans June 2001 - October 2012
- Collected as 26 CSV files (6.2GB) and loaded in a PostgreSQL database
- Pulled nursing notes and discharge summary

Data Preprocessing

- Defined next admission date for each subject and admission
- Defined admission type (elective, emergency, urgent, newborn)
- Compute number of days between admissions
- Mark elective visits as empty for next admission date and type
- Combined all notes for each subject and admission into a single string
- Dropped all duplicate and newborn admissions
- Compute target variable using days between admissions
- Split 70% of data into training set, 30% into test set

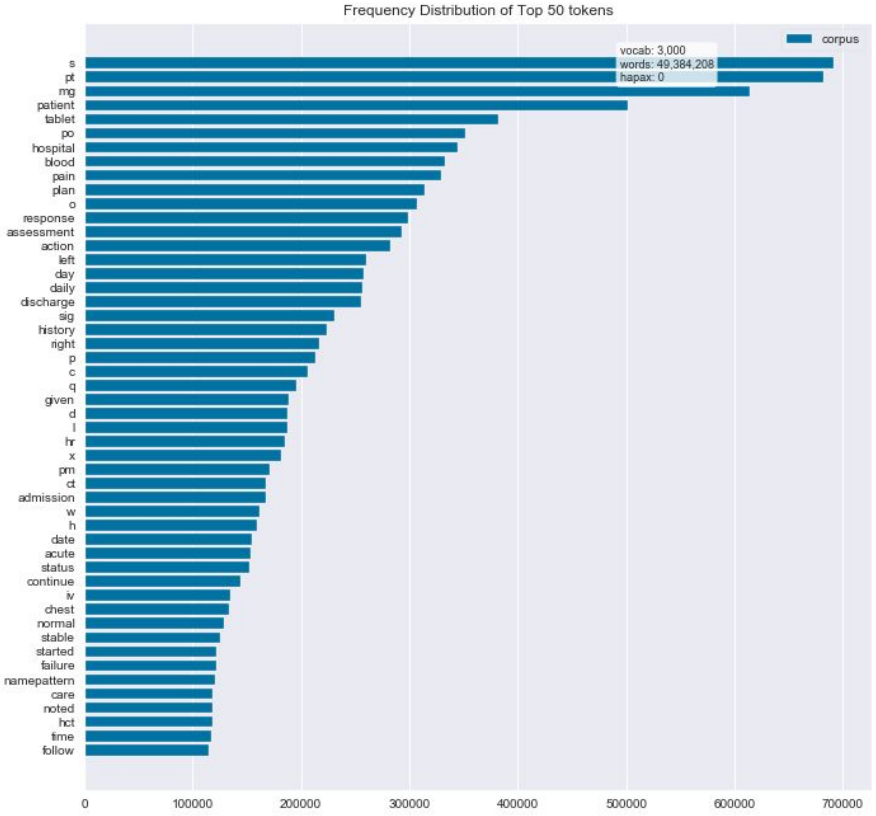
Exploratory Data Analysis



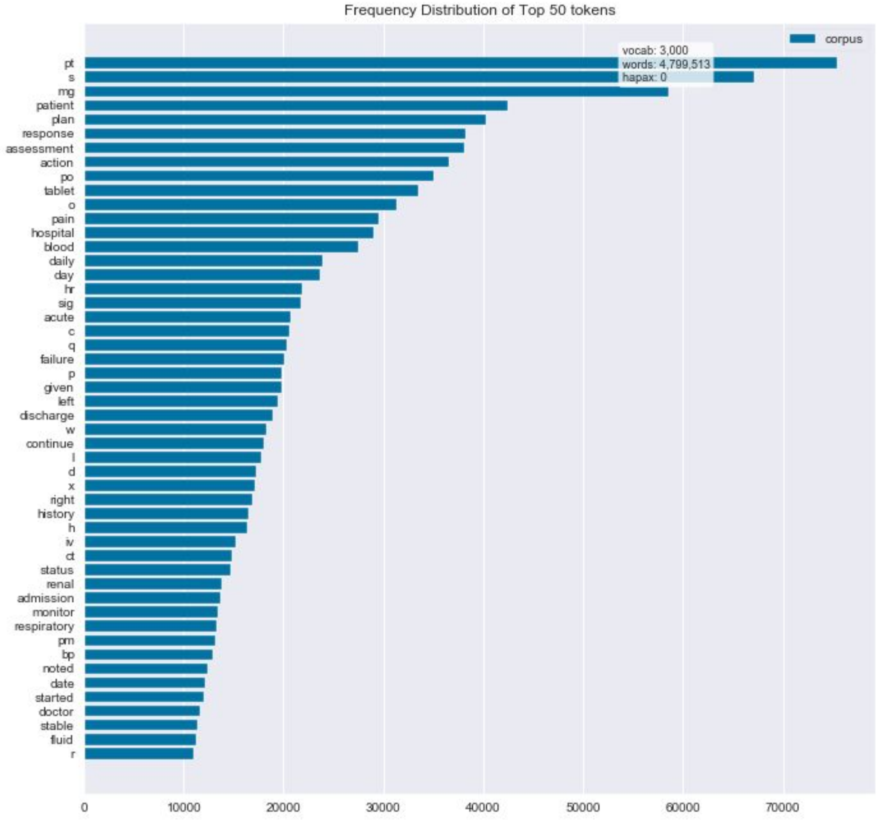


Readmissions peak early

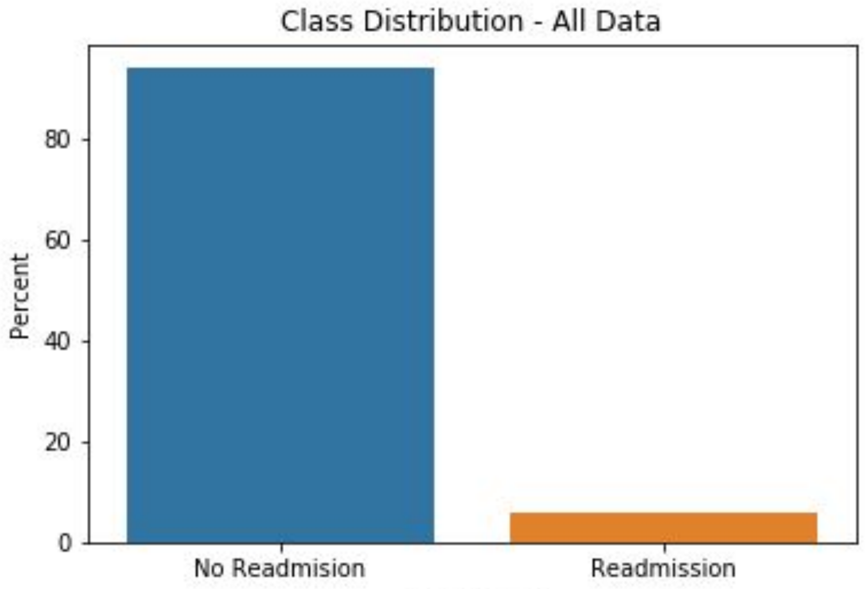
NOT READMITTED

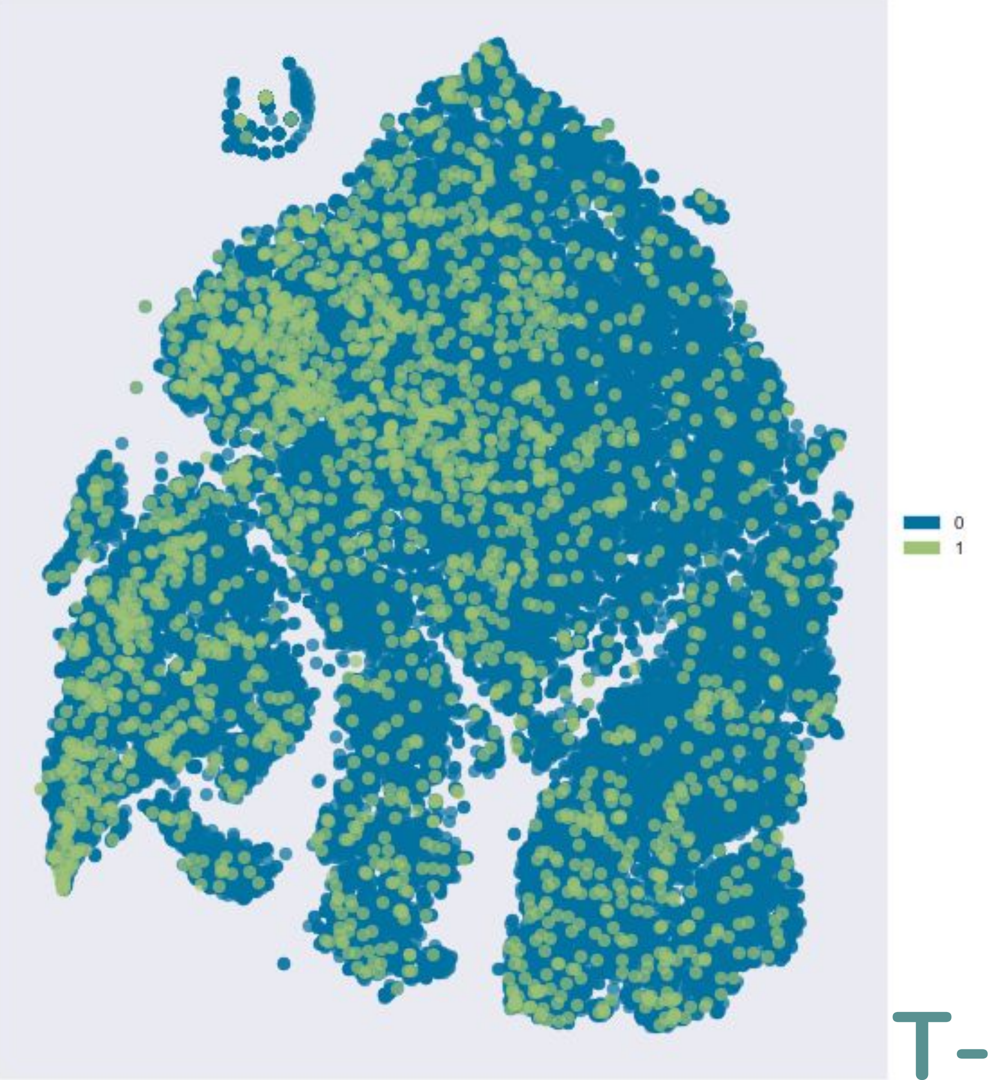


READMITTED



Imbalanced Dataset





T-SNE MAP

Feature Engineering

Bag-of-words

Removed punctuation and numbers

Lowercase

Tokenized

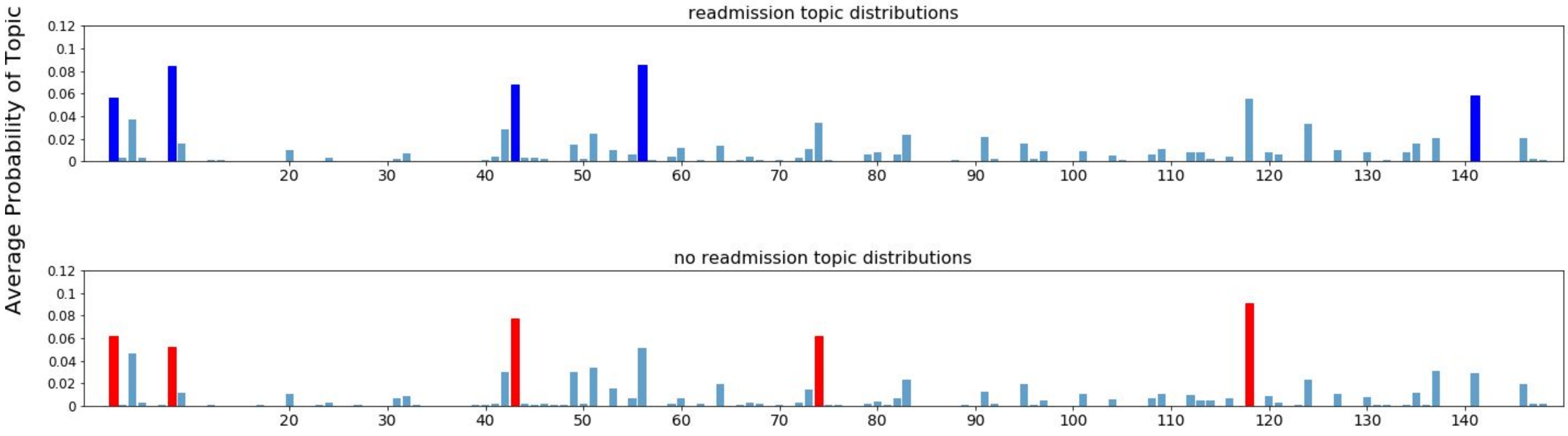
Word Embeddings

Cleaned and tokenized

Stemmed

Lemmatized

Latent Dirichlet Allocation

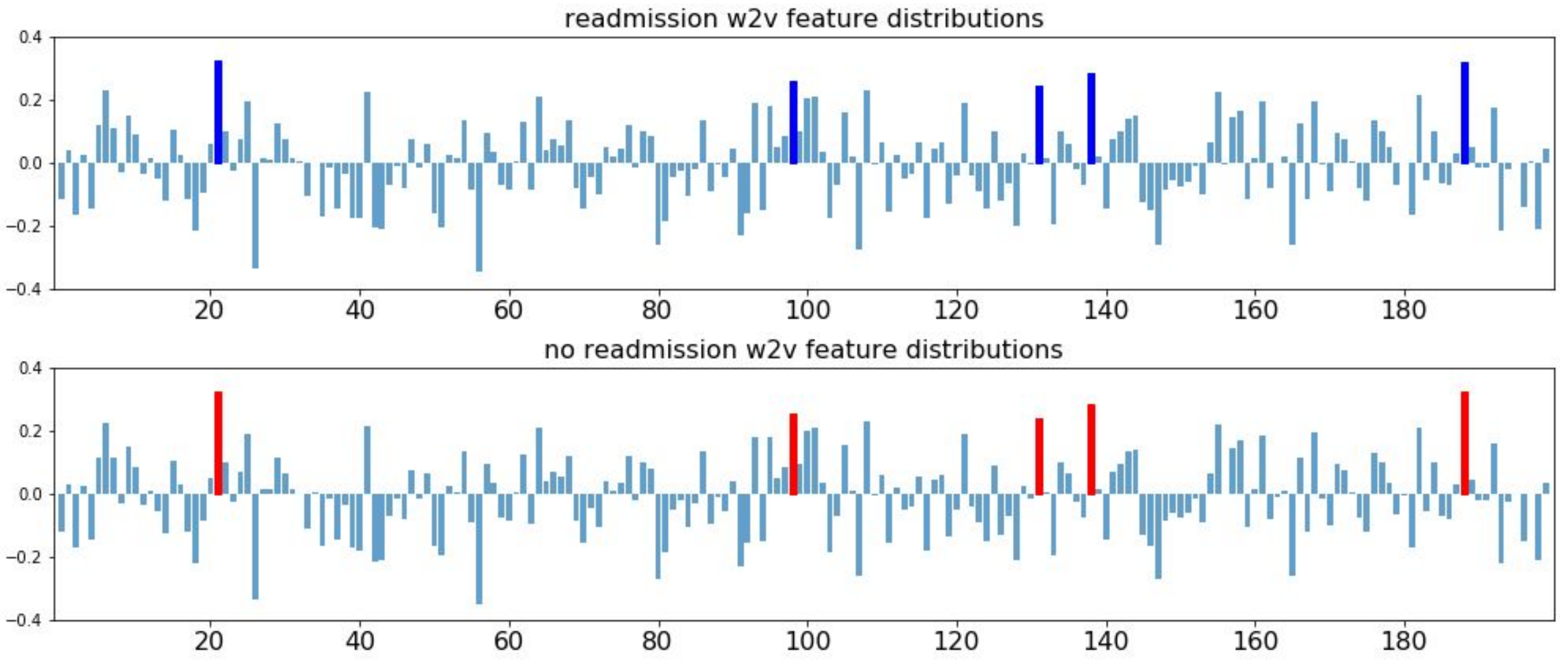


Top Topics for Readmitted Patients	
Topic	**Top Words**
2	tablet, daili, sig, cardiac, ventricular
8	tablet, daili, sig, hospital1, pt
43	statu, unit, show, number, also
56	tablet, sig, daili, need, capsul
141	daili, tablet, cultur, sig, neg

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

Defining LDA Topics By Examining the Top Words

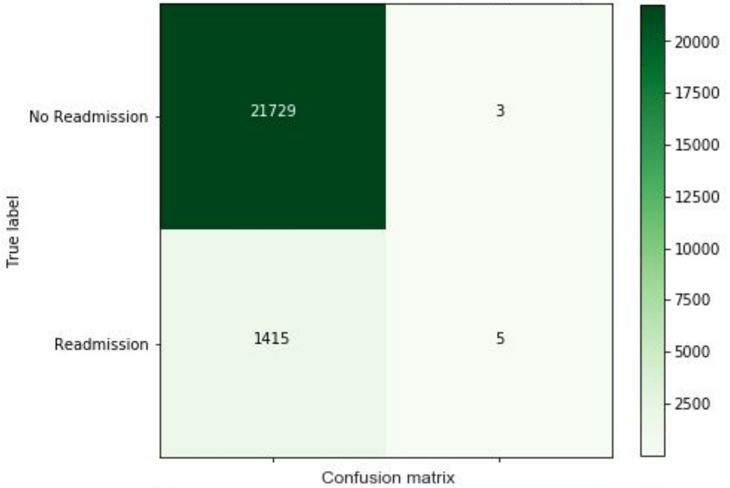
Word2Vec Feature Distribution



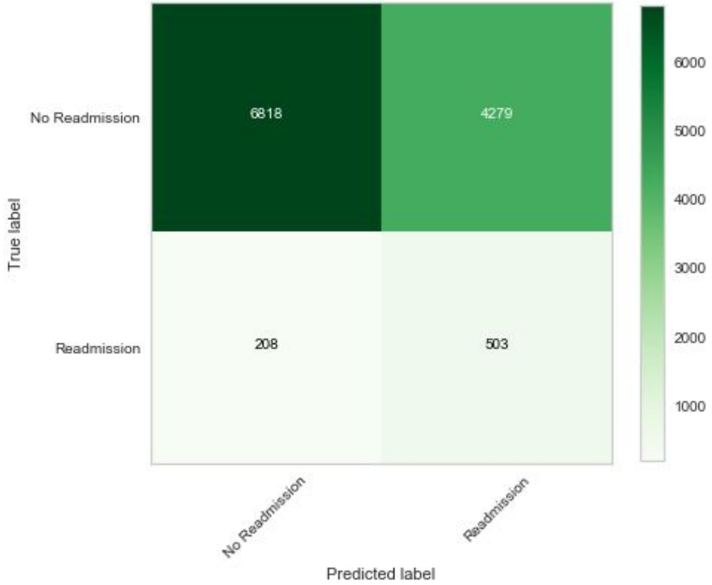
Predictive Modeling

MODEL PERFORMANCE

	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



Word2Vec with Latent Dirichlet Allocation and Logistic Regression



Bag-of-words with Random Forest

Moving Forward

Next steps in model improvement

Word2Vec

Grid search:

- window size
- learning rate
- number of epochs
- downsampling threshold

Alter threshold (or class weight)

Random Forest

More extensive grid search:

- Number of estimators
- Tree depth
- Leaf and node parameters
- Etc.

Balance class weight

Additional feature engineering:

- LDA
- Lab work, pharmacy, etc

Next steps in production

- EHR flag for readmission risk
- A/B test to determine effectiveness

Any Questions?

E. Chris Lynch

echrislynch@gmail.com

github.com/TheeChris

linkedin.com/in/echrislynch