Predicting Hospital Readmission

By Jerad Williams

Table of Contents

Introduction

Problem Statement

The goal of this project is to look into how we could help improve hospital resource utilization.

By using the information included in discharge summaries, we could predict hospital readmission, hence achieving:

- Improved budgeting and allocation of resources
- Reduction of waiting time
- Better care of patients coming in
- More precautionary measures of patients being discharged
- Overall better health of the hospital's patients

Results Summary

Output of the Project:

A Machine Learning model was built to predict a probability score for the patient being admitted once again within 30 days to the same hospital based on the discharge summaries

Accuracy of the Model:

The best tested model was XGBoost, yielding a 0.73 AUROC Score, which is just 0.02 less than the State of the Art method which has score of 0.75-0.76.

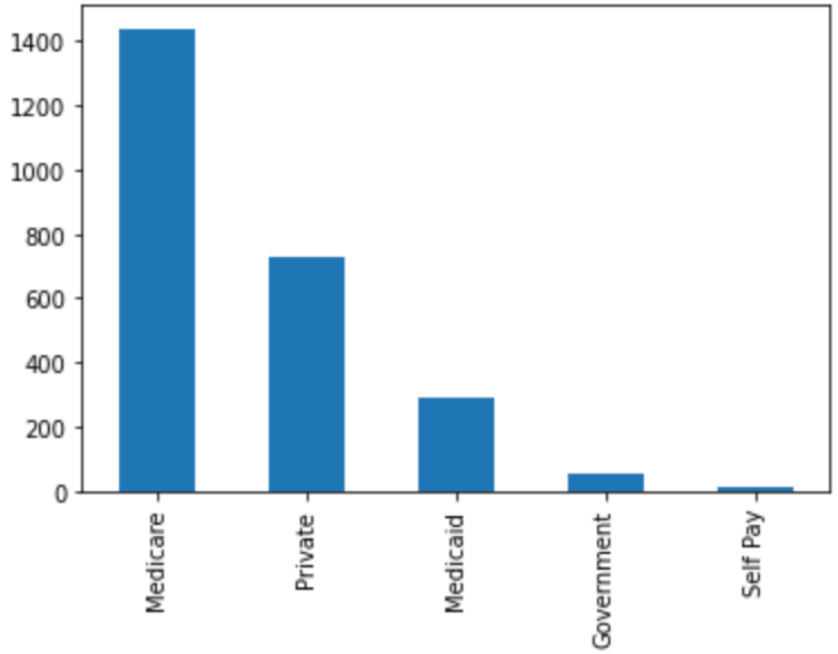
Data Used

- Real patients' data from the MIMIC-III database
- Info on 50k patients in Boston, MA between 2001 and 2012
- Encrypted sensitive information
- Data fields:
  - Diagnosis
  - Admission date and time
  - Number of days in hospital
  - Admission type
  - Type of insurance, marital status, religion and ethnicity
  - Etc.

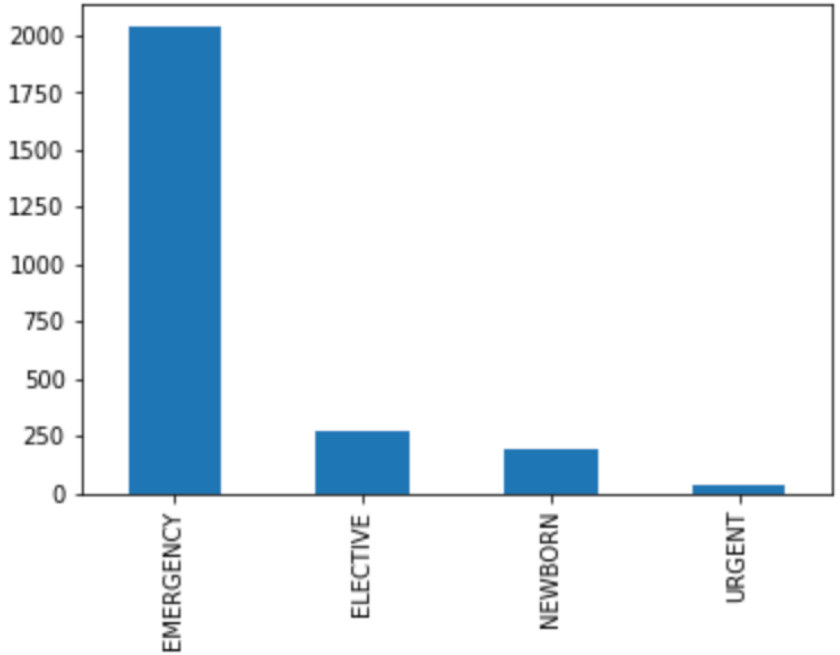
Data Exploration

Insurance & Admission Type of Readmitted Patients

Insurance Type

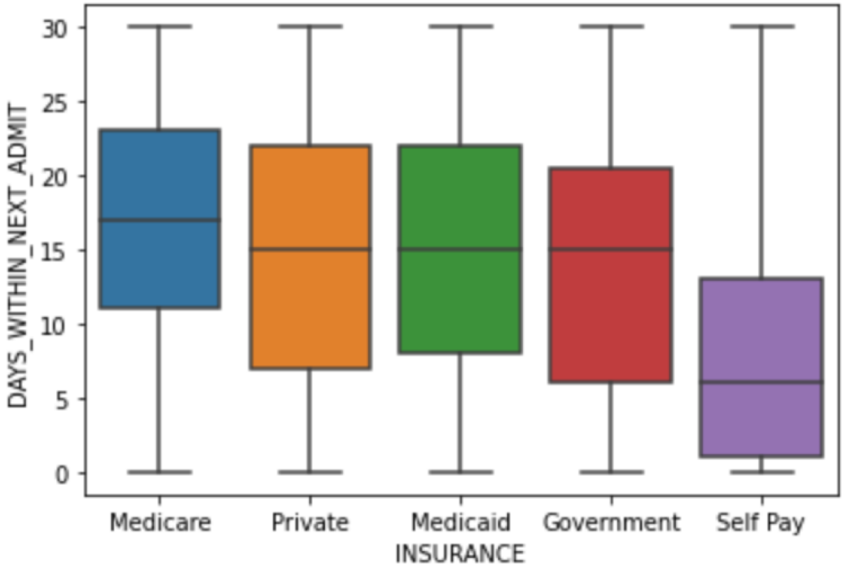


Admission Type

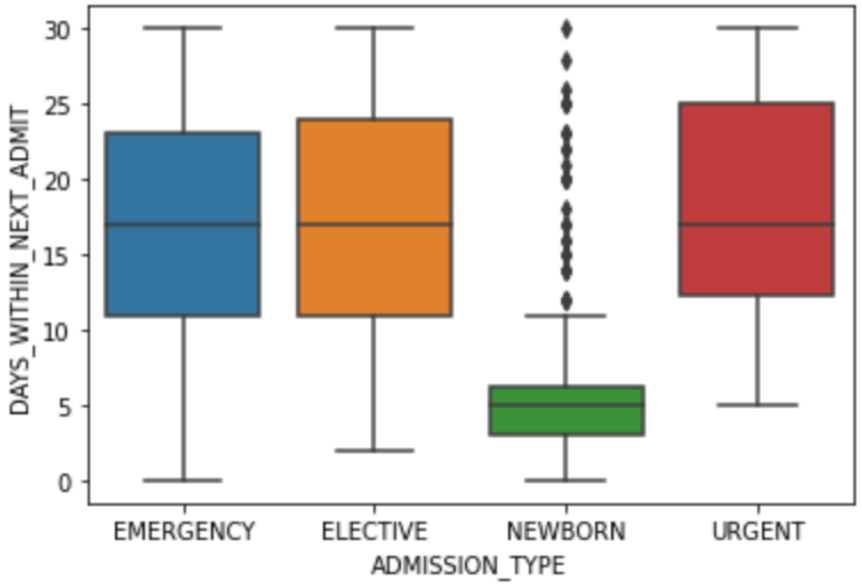


Insurance & Admission Type vs Days until Readmission

Insurance Type and Days until Next Admission

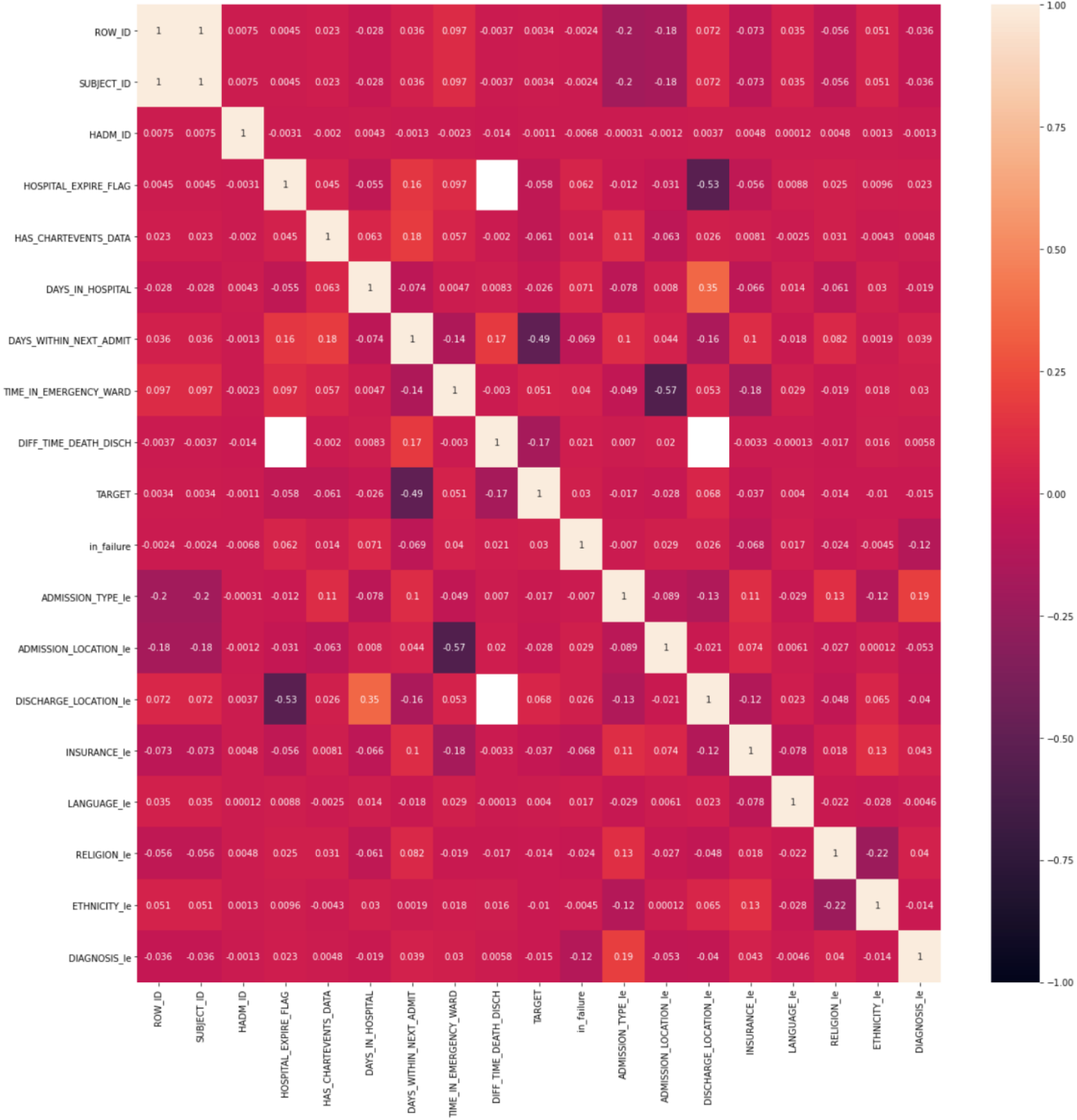


Admission Type and Days until Next Admission



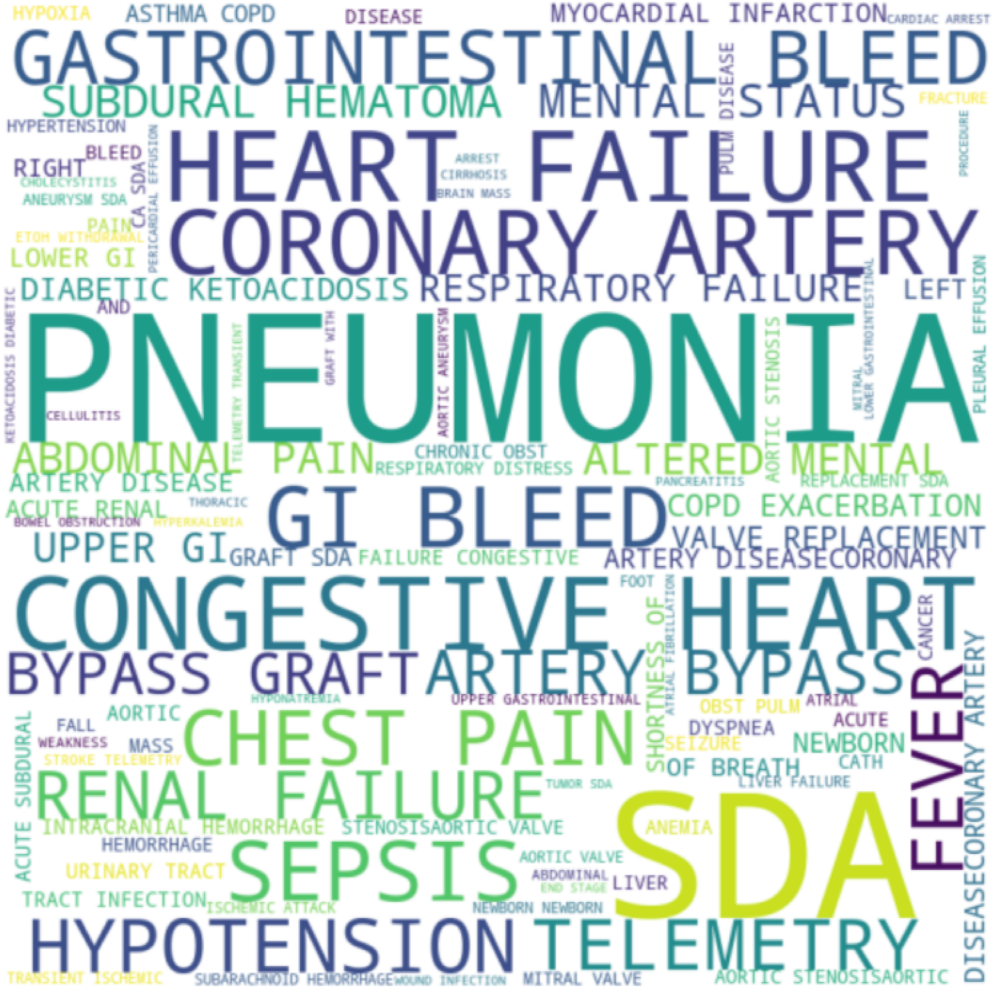
Correlation Heatmap

- Moderate negative correlation between time in emergency ward and admission location (-57%)
- Moderate negative correlation between discharge location and hospital expire flag (-53%)
- No correlation between non-medical variables (ethnicity, religion, language) and target

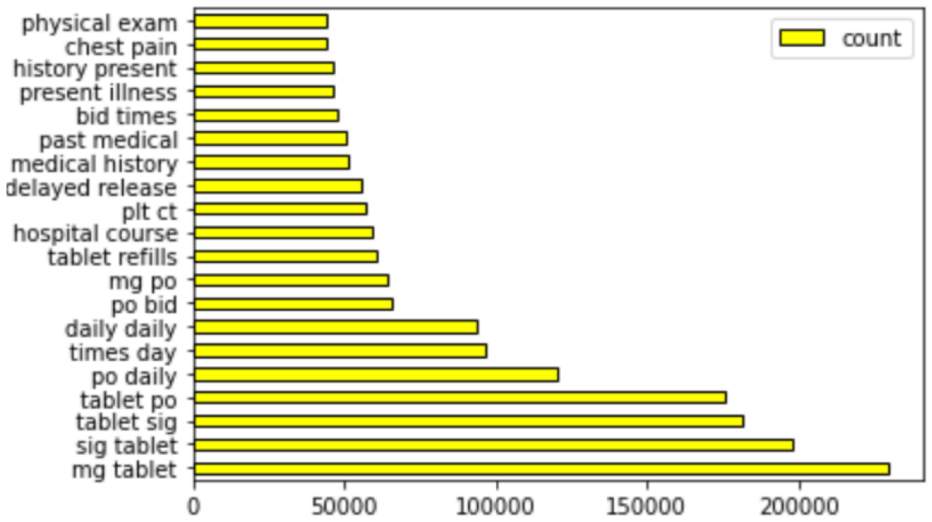


Predictive Model

Feature Engineering



- NLP techniques were used for analysis and feature engineering
- Feature engineering techniques include:
  - TfidfVectorizer
  - Word2Vectors
  - Keras Tokenizer



Predictive Techniques

- Tested models:
  - Logistic Regression
  - Bernoulli Naïve Bayes
  - Decision Tree
  - LSTM Based Deep Learning
  - **XGBoost**

- Training Data
  - Dataset split into training and testing data (80-20 split)
  - Models trained on all features or a subset of them

XGBoost Model - Results

- The XGBoost model trained with all features except the Diagnosis feature vector (Diagnosis is already included in detail in summary text) and with all the given dataset gave the best AUROC score of 0.73.

- The XGBoost model trained with only Doc2Vec embeddings with the downsampled data gave only 0.6690 AUROC Score.

- The XGBoost model was tried with different summary text max features for the TfIdf Feature vectors, the maximum we could achieve was with 5000 max features values that is 0.73 AUROC Score, which is just 0.02 less than the State of the Art method which has score of 0.75-0.76.

Limitations

- Low amount of data
  - <60k admissions for 50k patients
  - Target variable only for 4.45% of the admissions
- Representativeness of data
  - Outdated data (2001-2012)
  - Data relevant only to one hospital in Boston, MA

Thank you!