Discover Alpha

30-Day Hospital Readmission Predictive Model

By Santosh Patil & Angsuman Dutta

Highlights

- We compared over 10 different models for predicting hospital readmissions.
- Models prediction is over 80% accuracy across all hospitals, exceeds similar global benchmark models
- Data anomalies exist for some model influence variables, appropriate adjustments have been made to prep the data for the model
- Stacked Ensemble and Gradient Boost Machine model perform best in terms of Area Under Curve(AUC).
- Deep Learning techniques did not provide any significant lift to the prediction rate. They work well with vast volumes of data (> 1MM of records)

Abstract

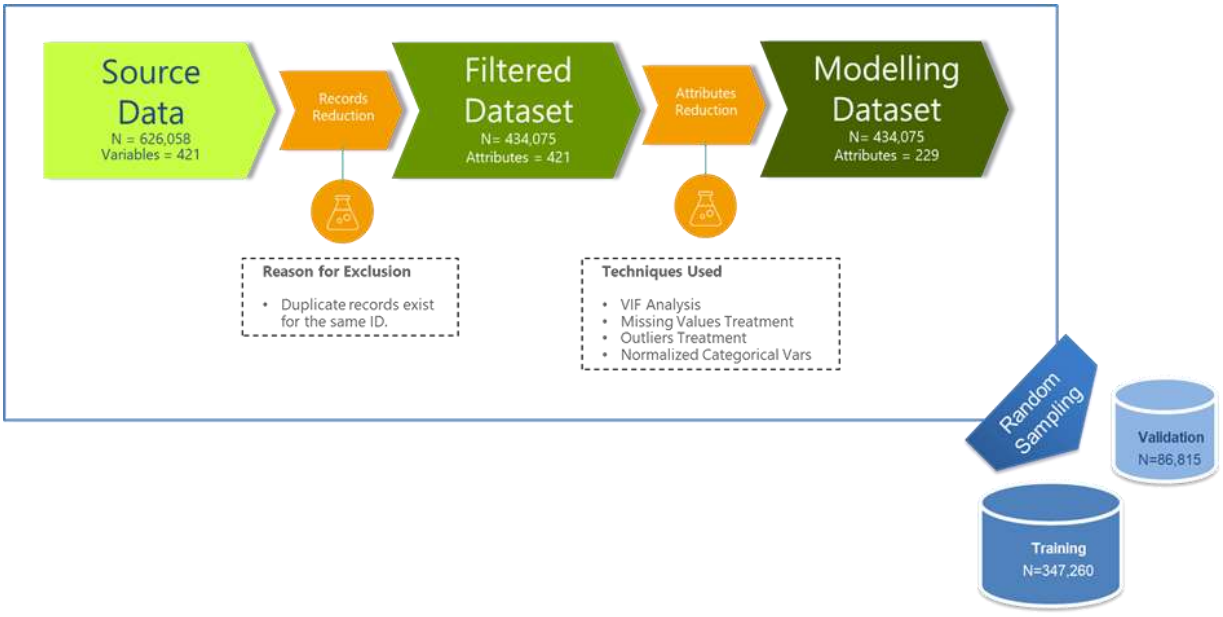
As healthcare industry moves towards a value-based model under the bundled payment program, hospitals are financially accountable for the quality and cost of an episode of care. The program also requires hospitals to coordinate care with the downstream care providers to mitigate financial risks. Risk sharing arrangements between hospitals and payers together with penalties imposed by the CMS are driving an interest in decreasing readmissions. There are number of published risk models predicting 30-day readmissions for different patient populations. In this work we describe the predictive models developed, some of which outperform the regression methods that are typically applied in the healthcare literature. As an added feature, the model also recommends the best discharge location for each prediction.

Data Summary

The dataset used is Hospital Readmissions dataset from over 200 hospitals in the United States. The data contains about 625,000 hospital visits for various procedures between 2010 and 2015. Each record consists of about 420 variables. We formalize the task of predicting early patient readmissions as a binary classification task. For each visit, we have background information on the patient's race, sex, age, and length of stay. Additionally, we also know the type of diagnosis codes, surgery times and certain risk conditions.

Data Processing and Preparation

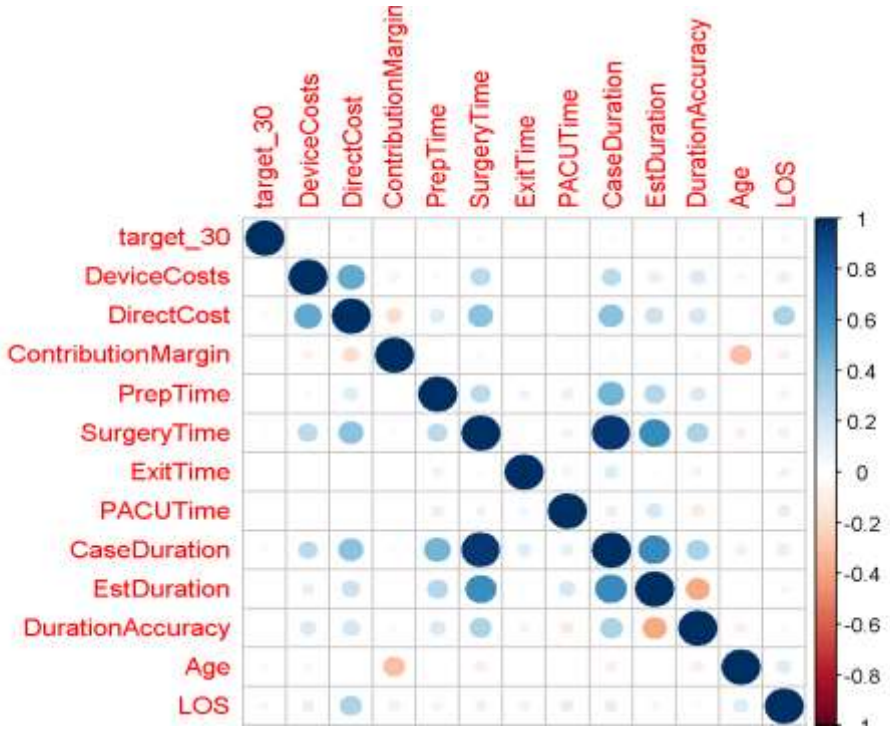
Different cleansing and filtering techniques were applied to the source dataset. Due to the highly imbalanced nature, several steps were taken so that the algorithms "generalize" learning the patterns in the dataset. We removed variables that are not predictive and descriptive and those that had very few observations. We also converted the text categorical variables into numbers before feeding them to the machine learning algorithms. We used several measures to determine the classification accuracy of our algorithms due to the highly imbalanced nature such as AUC, F1 score, precision, recall etc.



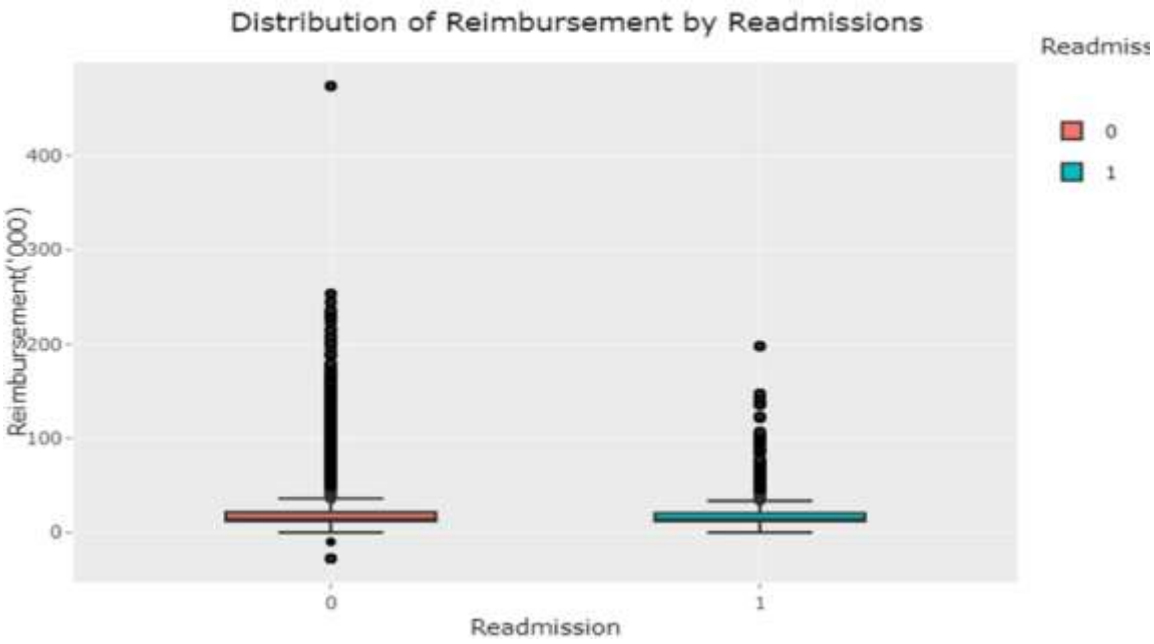
The filtered dataset for modelling consists of about 434,075 visits with 229 variables. The modeling dataset was randomly split into Training and Validation dataset. The Training dataset consists of 347,260 visits and is exclusively used to initially fit the model. Successively, the fitted model is used to predict the responses for the observations in a second set Validation Dataset consisting of 86,815 records.

Data Observations

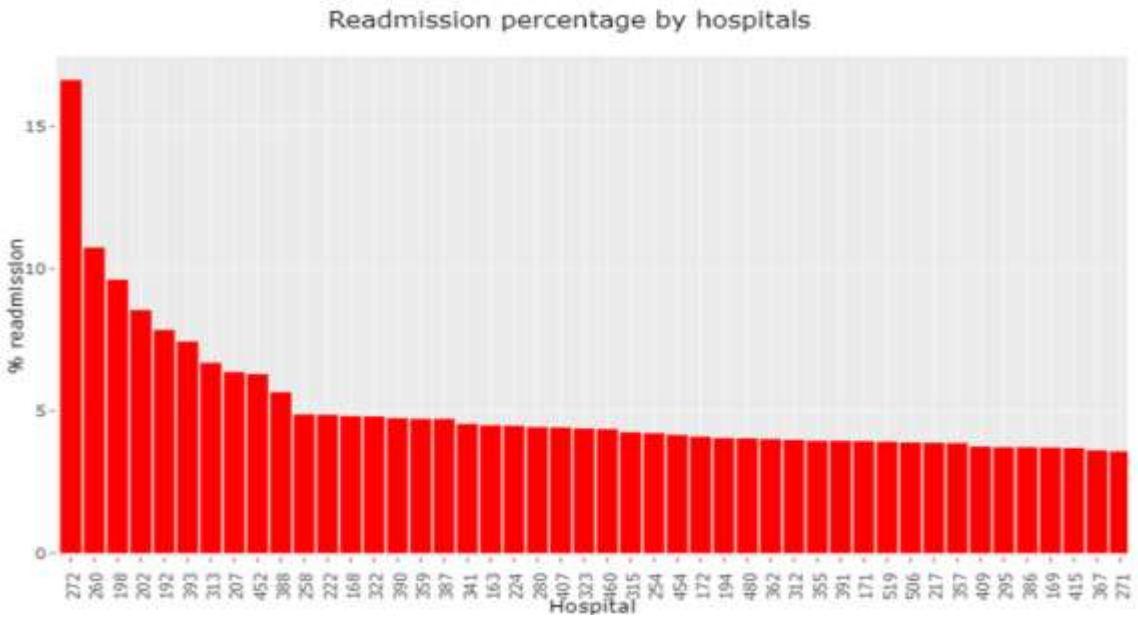
The correlation among the internal characteristics and readmissions does bring any significant relation to the fore. It's mainly because of missing values in the data Correlation Analysis



The reimbursement feature has some outliers when compared to readmissions. The mean is 1.8312 and some of them as high as 4.74

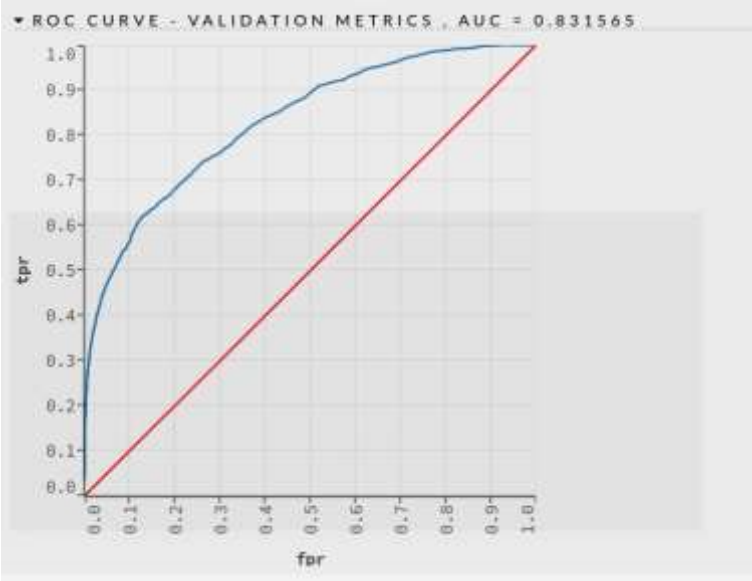


We also observed that certain hospitals have relatively higher number of readmission rates. As such we can observe that about 10 hospitals have over 5% readmission rates.



Model Results

We applied different modeling techniques to the training data set.The model with best lift has over .82 AUC.



The following table describes the confusion matrix with the best fit model and highest true positives predictions of readmissions

	Actual/Predicted	0	1	Error	Rate
CM	**0**	59283	3906	0.0618	3906 / 63189
	1	731	719	0.5041	731 / 1450
	Total	**60014**	**4625**	**0.0717**	**4637 / 64639**

Conclusions

The ultimate goal of this predictive model is to initiate interventions to lower the frequency of 30-day readmissions. Our meta-analysis across the universe of data

shows that the 30-day readmission rate across all specialties is about 2.35 percent.

Our model outcomes have a higher lift compared to other benchmark models as compared below.

Performance of our predictive models

	model_id	auc	logloss
0	StackedEnsemble_0_AutoML_20170915_222616	0.828648	0.087525
1	DRF_0_AutoML_20170915_222616	0.818696	0.091977
2	XRT_0_AutoML_20170915_222616	0.811438	0.093354
3	GBM_grid_0_AutoML_20170915_222616_model_3	0.787233	0.094752
4	GBM_grid_0_AutoML_20170915_222616_model_2	0.772581	0.097987
5	GBM_grid_0_AutoML_20170915_222616_model_1	0.767919	0.099260
6	GBM_grid_0_AutoML_20170915_222616_model_0	0.758105	0.101653
7	GBM_grid_0_AutoML_20170915_222616_model_4	0.707715	0.108982
8	GLM_grid_0_AutoML_20170915_222616_model_0	0.688505	0.107110

Performance of Benchmark Models*

Table 3. Results for predicting readmission across 280 DRGs

Method name	Mean (SE) AUC across all 280 DRGs	# DRGs where method had highest mean AUC
RF	0.684 (0.004)	80
PLR (α = 0.01)	0.683 (0.004)	27
PLR (α = 0.5)	0.682 (0.004)	15
PLR (α = 1)	0.681 (0.004)	22
SGD (Huber, α = 1)	0.672 (0.004)	39
SVM (linear)	0.671 (0.004)	47

Source: Journal of Biomedical Informatics, *Aug 2015