Case Study: Predicting Patient Readmission

September 5th, 2018

Table of Contents

Problem Statement

Chronic disease readmission is one of the leading cost drivers in US healthcare

By the Numbers

2017 Readmission Cost: **$310M** across top 10 heath systems

Growth: **9.5-15%** Expected CAGR from 2017-2022

Key Challenges

- Readmission penalties
- Expensive specialty treatments
- Resource allocation challenges

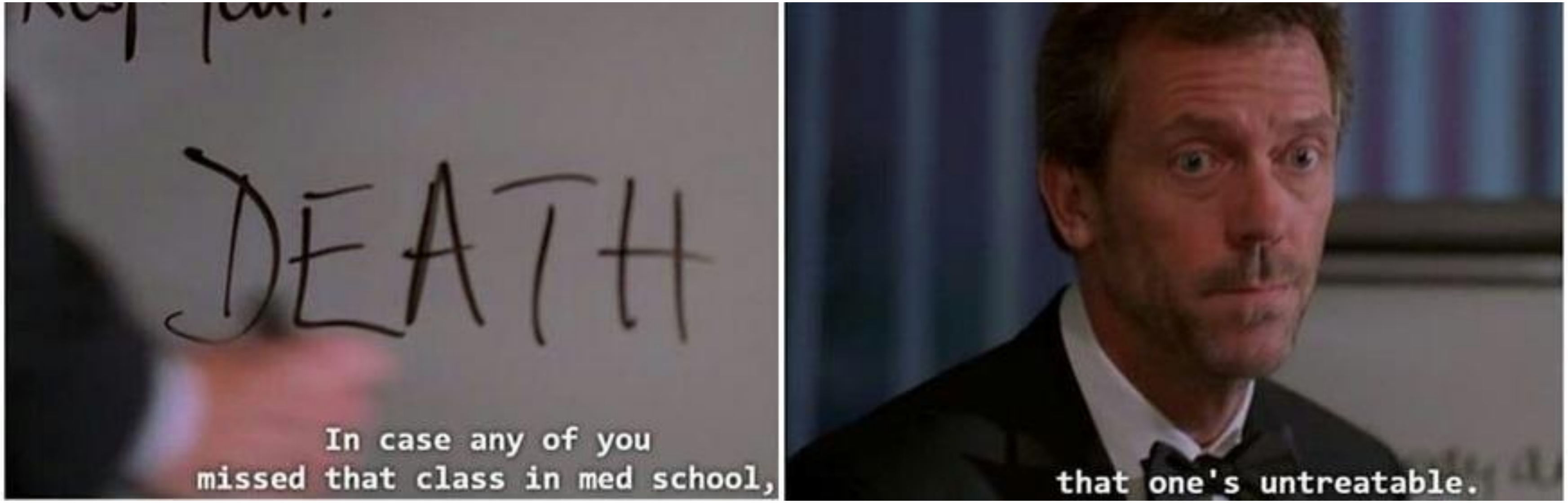
Top Readmission Cases

- Schizophrenia
- **Diabetes**
- Surgical complications

In this study, we analyzed a patient dataset from 2001-2011 that contains demographic and treatment information for 100,000 diabetes cases, and built a classification model to predict whether or not a given patient would be readmitted within 30 days of discharge

Problem Statement (Cont.)

Identifying high-risk readmission cases is key to reducing mortality rates related to chronic health complications



Dataset, Assumptions, and Metrics

Dataset

Data was pulled from the UCLA medical learning library

101,776 Rows x 50 Columns

Messy: Contained missing values, string and number pairs, ambiguous features

Assumptions

Results from the dataset should only be generalized for patients with diabetes

Assumed data was accurate for a given patient

Assumed that dataset was pulled from a random sample of diabetes patients

Key Metrics

Model was evaluated against a **null model that predicted that the patient would not be readmitted for all cases**, because that is currently the industry standard.

Goal: Outperform null model

Data

63768	1.15E+08	Caucasian	Male	[70-80)	?	1	1	7	5	?	?	73	0	12	0	0	0	428
12522	48330783	Caucasian	Female	[80-90)	?	2	1	4	13	?	?	68	2	28	0	0	0	398
15738	63555939	Caucasian	Female	[90-100)	?	3	3	4	12	?	InternalM	33	3	18	0	0	0	434
28236	89869032	AfricanAm	Female	[40-50)	?	1	1	7	9	?	?	47	2	17	0	0	0	250.7
36900	77391171	AfricanAm	Male	[60-70)	?	2	1	4	7	?	?	62	0	11	0	0	0	157
40926	85504905	Caucasian	Female	[40-50)	?	1	3	7	7	?	Family/Ge	60	0	15	0	1	0	428
42570	77586282	Caucasian	Male	[80-90)	?	1	6	7	10	?	Family/Ge	55	1	31	0	0	0	428
62256	49726791	AfricanAm	Female	[60-70)	?	3	1	2	1	?	?	49	5	2	0	0	0	518
73578	86328819	AfricanAm	Male	[60-70)	?	1	3	7	12	?	?	75	5	13	0	0	0	999
77076	92519352	Afric[illegible]	[illegible]	[50-60)	?	1	1	7	4	?	?	45	4	17	0	0	0	410
84222	1.09E+08	Caucasian	Female	[50-60)	?	1	1	7	3	?	Cardiology	29	0	11	0	0	0	682
89682	1.07E+08	AfricanAm	Male	[70-80)	?	1	1	7	5	?	?	35	5	23	0	0	0	402
148530	69422211	?	Male	[70-80)	?	3	6	2	6	?	?	42	2	23	0	0	0	737
150006	22864131	?	Female	[50-60)	?	2	1	4	2	?	?	66	1	19	0	0	0	410
150048	21239181	?	Male	[60-70)	?	2	1	4	2	?	?	36	2	11	0	0	0	572
182796	63000108	AfricanAm	Female	[70-80)	?	2	1	4	2	?	?	47	0	12	0	0	0	410
183930	1.07E+08	Caucasian	Female	[80-90)	?	2	6	1	11	?	?	42	2	19	0	[illegible]	[illegible]	V57
216156	62718876	AfricanAm	Female	[70-80)	?	3	1	2	3	?	?	19	4	18	0	0	0	189
221634	21861756	Other	Female	[50-60)	?	1	1	7	1	?	?	33	0	7	0	0	0	786
236316	40523301	Caucasian	Male	[80-90)	?	1	3	7	6	?	Cardiology	64	3	18	0	0	0	427
248916	1.15E+08	Caucasian	Female	[50-60)	?	1	1	1	2	?	Surgery-G	25	2	11	0	0	0	996
250872	41606064	Caucasian	Male	[20-30)	?	2	1	2	10	?	?	53	0	20	0	0	0	277

rosiglitazo	acarbose	miglitol	troglitazor	tolazamid	examide	citogliptor	insulin	glyburide-	glipizide-m	glimepirid	metformir	metformir	change	diabetesM	readmitted
No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	NO
No	No	No	No	No	No	No	Up	No	No	No	No	No	Ch	Yes	>30
No	No	No	No	No	No	No	No	No	No	No	No	No	No	Yes	NO
No	No	No	No	No	No	No	Up	No	No	No	No	No	Ch	Yes	NO
No	No	No	No	No	No	No	Steady	No	No	No	No	No	Ch	Yes	NO
No	No	No	No	No	No	No	Steady	No	No	No	No	No	No	Yes	>30
No	No	No	No	No	No	No	Steady	No	No	No	No	No	Ch	Yes	NO
No	No	No	No	No	No	No	No	No	No	No	No	No	No	Yes	>30
No	No	No	No	No	No	No	Steady	No	No	No	No	No	Ch	Yes	NO
Steady	No	No	No	No	No	No	Steady	No	No	No	No	No	Ch	Yes	NO
No	No	No	No	No	No	No	Steady	No	No	No	No	No	No	Yes	>30
No	No	No	No	No	No	No	Steady	No	No	No	No	No	Ch	Yes	<30
No	No	No	No	No	No	No	Down	No	No	No	No	No	Ch	Yes	<30
No	No	No	No	No	No	No	Steady	No	No	No	No	No	No	Yes	NO
No	No	No	No	No	No	No	Steady	No	No	No	[illegible]	No	No	Yes	>30
No	No	No	No	No	No	No	Up	No	No	No	No	No	Ch	Yes	NO
No	No	No	No	No	No	No	Steady	No	No	No	No	No	Ch	Yes	<30
No	No	No	No	No	No	No	No	No	No	No	No	No	No	Yes	NO
No	No	No	No	No	No	No	Steady	No	No	No	No	No	No	Yes	>30
No	No	No	No	No	No	No	Steady	No	No	No	No	No	Ch	Yes	NO
No	No	No	No	No	No	No	Down	No	No	No	No	No	Ch	Yes	NO
No	No	No	No	No	No	No	Steady	No	No	No	No	No	Ch	Yes	NO
No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	NO
No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	>30

Approach

Identify Problem

Identified patient readmission as the problem to address and looked for relevant data sources

Pull explore data

Pulled, cleaned, and dropped data as necessary

Model and Predict

Trained decision tree, random forest, and boosted models

Re-engineered data as necessary

Measure Performance and Report

Measured tree accuracy and reported

Train-Test-Split → Train Decision Tree → Train Random Forest → Train Ada Boost

- Train Decision Tree ← Apply Feature Selection
- Train Random Forest ← Apply Feature Selection
- Train Ada Boost ← Apply Feature Selection

Model and Solution

Three classification models were trained using the patient dataset, and measured against a null model performance

Decision Tree

No max depth

No min-leaf size

SelectFromModel feature-selector was applied to isolate important features

Accuracy: 47.1%

Random Forest Classifier

100 decision trees

No max depth

SelectFromModel feature-selector was applied to isolate important features

Accuracy: 57.6%

Ada Boost Classifier

100 decision trees

0.1 learning rate

SelectFromModel feature-selector was applied to isolate important features

Accuracy: 59.1%

Null Model Performance: 51%

GENERAL ASSEMBLY

Retrospective

Spend more time up-front cleaning data

- Features and rows were dropped due to missing data or lack of interpretability, which reduced robustness of the model
- More effort in cleaning features would have increased accuracy

Train models with small n_estimators at first

- Due to high number of features, Random Forest and Ada boost classifiers took a longtime to run (~25-40 minutes) and yielded marginal improvements run-over-run

Incorporate more robust classifiers into ensemble

- Outside of Adaboost, I would have liked to implement XG Boost and potentially a deep learning classifier

Questions?


-Ok, got my pulse
-Start counting
-Stop, how many?
-26
-OK. Either you suck at math,
or you are gonna die in 2 seconds
-You suck at maths
VIA 9GAG.COM