Leveraging Data Science to Understand and Predict Hospital Readmissions in Diabetes Patients

Katherine Brown, Joseph Bivens, Scott VandePolder, William Eberle

Motivation

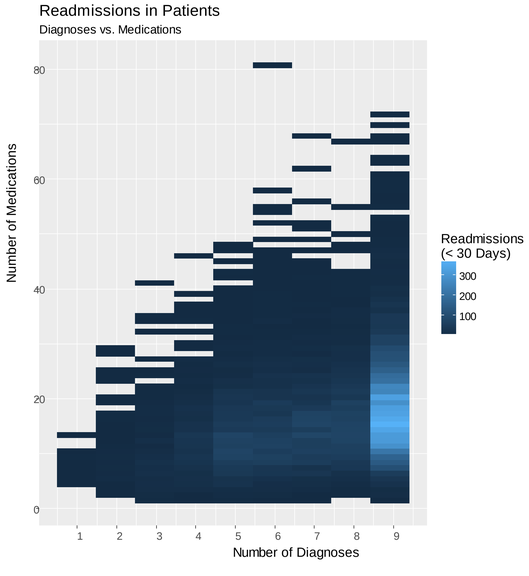
The fundamental goal in healthcare is simple: improve patient health. Reducing the rate of patient readmission provides enormous benefits to both patients and healthcare providers, but is difficult to achieve without identifying the underlying causes and recognizing which patients are at risk.

Goals

- Explore the University of California, Irvine data repository on hospital readmissions of diabetic patients.
- Identify patient characteristics that are positively or negatively correlated with readmission events.
- Develop predictive models that could provide decision support to healthcare professionals by recognizing patients at increased risk.

Exploratory Data Analysis

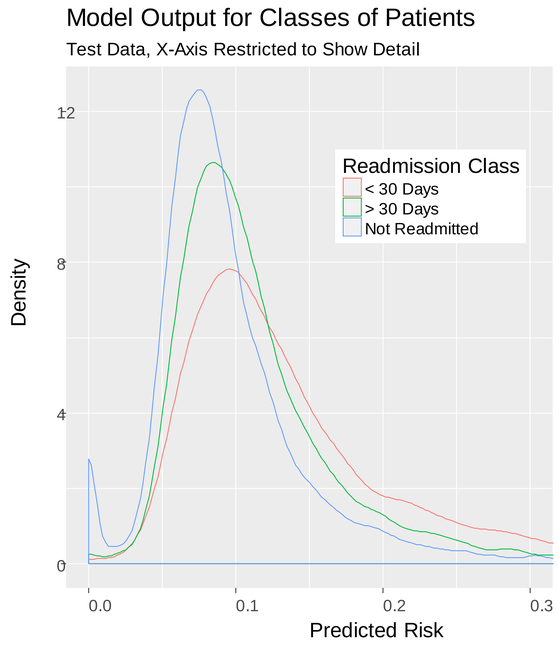
- We began by cleaning the data and creating exploratory visualizations.



- The stark contrast between the final two columns suggests that 9 may actually indicate “9 or more” diagnoses.

Logistic Regression

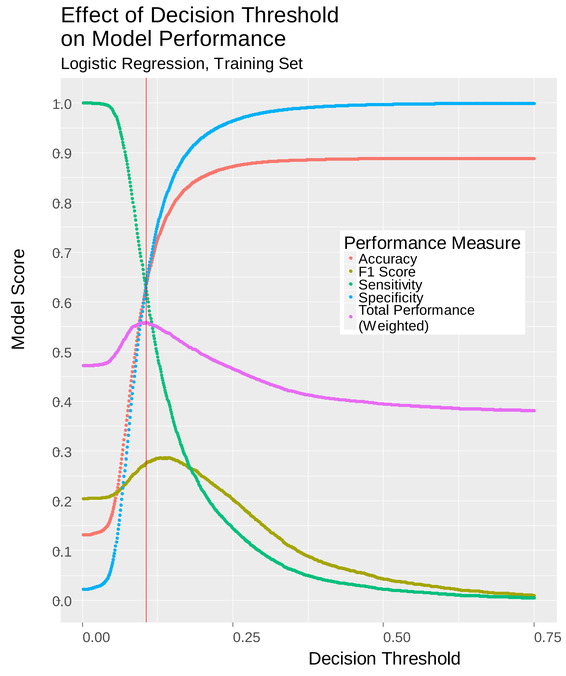
- Logistic Regression is a powerful and relatively simple supervised learning technique.
- Given a set of attributes (features) of a patient, the model assigns a score that indicates the predicted risk of readmission.



- Note that the distribution of Risk Scores tends to be higher for patients who eventually are readmitted.
- The model was trained on only two Readmission Classes (“< 30 Days” and “Other”). However, the prediction scores correctly show that the patients in the “> 30 Days” class have risk levels that fall between patients readmitted sooner and those not readmitted at all.
- This suggests that the generated scores are well-correlated with the true underlying risk and that the highest risk patients require readmission soonest.

Threshold Optimization

- These predicted risks are continuous on the interval [0,1]. We must select a threshold of risk that is significant enough to warrant concern.



- In particular, finding a balance between Sensitivity and Specificity is critical.
- Maximizing **sensitivity** ensures patients at elevated risk are reliably flagged by the model. This is essential in the healthcare domain, so it is given additional weight.
- Maximizing **specificity** ensures that false alarms are rare.
- A insensitive model will ignore patients at risk, but one that is too sensitive might be ignored by healthcare providers. The vertical line shows the chosen threshold.

	Prediction		
Reality	< 30	Other	**Total**
< 30	1,350	900	**2,250**
Other	6,449	11,655	**18,104**
Total	**7,799**	**12,555**	**20,354**

Confusion Matrix for Chosen Threshold

Results

Performance Measure	Model Score
Overall Accuracy	0.639
Balanced Accuracy	0.622
Sensitivity	0.600
Specificity	0.644
Precision	0.173
F1	0.269

Performance Evaluation of Best-Fitting Logistic Regression Model (20% Test Data)

- The Accuracy measures indicate that this model has potential as a diagnostic tool.
- The Precision is somewhat low as a consequence of emphasizing Sensitivity. Additional experimentation with feature selection may improve performance.

Conclusions and Future Work

Attribute	Value	Coefficient
Age	[70-80)	+1.62
Number of Diagnoses	--	+0.03
On Diabetes Meds	Yes	+0.15
Discharge Disposition	Hospice (Home)	-2.98
Primary Diagnosis	Musculoskeletal / Connective Tissue	+1.33

- The attribute / value combinations above were among those found to be statistically significant. A positive coefficient increases a patient’s risk score, and negative coefficients reduce the risk score.
- We will generate, optimize, and evaluate additional models using other algorithms, such as decision trees and SVMs.
- We will allow programmatic access to the models so they can be used in a future decision support application.
- We will continue to search for a better understanding of risk factors and how providers might compensate for them.

The team wishes to acknowledge Dr. William Eberle, TTU, for his guidance and UCI Machine Learning Repository for the data.