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.

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.

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.

Results

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

<table>
<thead>
<tr>
<th>Performance Measure</th>
<th>Model Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Accuracy</td>
<td>0.639</td>
</tr>
<tr>
<td>Balanced Accuracy</td>
<td>0.622</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.600</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.644</td>
</tr>
<tr>
<td>Precision</td>
<td>0.173</td>
</tr>
<tr>
<td>F1</td>
<td>0.269</td>
</tr>
</tbody>
</table>

Conclusions and Future Work
• 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.

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

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.

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