AI Quiz
AI Quiz
AI Quiz
Objectives

Learn some buzzwords

Why Bother?

How to build a predictive model

Examine real-world predictive models

Getting Buy-In from Clinicians
AI: Artificial Intelligence

General Artificial Intelligence

“Narrow” Artificial Intelligence
Machine Learning

Machine learning explores the study and construction of algorithms that can learn from and make predictions on data.

https://en.wikipedia.org/wiki/Machine_learning

Predictive analytics, or making predictions based on past data, is one of the artificial intelligence tasks that machine learning can solve.
I’m still confused…

**Artificial Intelligence** tries to replicate the capabilities of the human mind.

**Machine Learning** uses complex math to solve difficult problems.

**Predictive Analytics**, from the standpoint of healthcare or business, is one of the most important activities that is enabled by Machine Learning.

**Predictive Models** and **Risk Models** are the products of Predictive Analytics.
Why bother?
Classic Approaches

Mortality prediction
The Charlson Index was introduced in 1987 in the Journal of Chronic Disease as mortality risk score.

Readmission prediction
The LACE Index was introduced in the Canadian Medical Association Journal in 2010 to predict early death or unplanned readmission after discharge.
CONCLUSION: The LACE Index may not accurately predict unplanned readmissions within 30 days from hospital discharge in CHF patients. The LACE high risk index may have utility as a screening tool to predict high risk ED revisits after hospital discharge.

CONCLUSION: The LACE Index is a poor tool for predicting 30-day readmission in older UK inpatients. The absence of a simple predictive model may limit the benefit of readmission avoidance strategies.
Limitations

Most standard models are trained with data from a broad, general population.

Most standard models are based upon data elements that are available through billing or claims data.
Advantages of building models

Trained on data from your environment.

Trained on data from your patients.

Answers your specific questions.
When should I build a model?

Trying to differentiate outcomes for complex cohorts

Predict infrequent events

Prioritize attention of limited resources to very frequent events

Predict outcomes as the result of modified behaviors

Incorporate features unlikely to be available to “standard” models
  - Socio-economic data
  - Geo-location data
Let’s Try It
Let’s Build a Predictive Model
Steps to build a model

1. Determine event of interest.
2. Determine our population.
3. Decide upon “features.”
4. Build feature sets.
5. Run through various algorithms: Train and Test.
6. Select the best model.
Typical Workflow for Building a Predictive Model

1. Data Source
2. Gnarly SQL Query
3. Data Manipulation
4. Feature Set
5. Tools/Algorithms (SAS | Weka | R | Python)
6. Candidate Models
7. Evaluate & Select Best
## Features

<table>
<thead>
<tr>
<th>Delivery Date</th>
<th>Delivery Location</th>
<th>Humour Temperament</th>
<th>Blood Letting</th>
<th>Physician Type</th>
<th>Hand Washing</th>
<th>Died</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/1/1844</td>
<td>Clinic 1</td>
<td>Sanguine</td>
<td>Yes</td>
<td>Physician</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>1/1/1844</td>
<td>Clinic 1</td>
<td>Melancholy</td>
<td>No</td>
<td>Physician</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>1/1/1844</td>
<td>Clinic 1</td>
<td>Balanced</td>
<td>No</td>
<td>Physician</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>1/1/1844</td>
<td>Clinic 2</td>
<td>Choleric</td>
<td>No</td>
<td>Midwife</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>1/1/1844</td>
<td>Clinic 2</td>
<td>Phlegmatic</td>
<td>No</td>
<td>Midwife</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>
Training and Testing

Most records will be used to “train” or create the models.

The remaining records will be used to test, or determine the accuracy, of each model.
Developing a Predictive Model

Features (i.e. age, comorbidities, polypharmacy)

Definition: Simply put, a feature is an input to a machine learning model.

Algorithms (i.e. Lasso, Random Forest, k-means)

Definition: Algorithms are complex mathematical processes that discover the relationship between features (input) and the outcome being predicted.

Result:

- Handful of best (most predictive) features
- Best algorithm that computes the relationships between input features to generate prediction
- Performance report summarizing best ‘model’
Dr. Semmelweis’s Model

When Delivery Location = Clinic 1 and Hand Washing = No, women are 3 times more likely to die. Humours are not predictive, and blood letting correlates slightly with death.
Real World Models
Real World Use Case: COPD Readmissions

From nih.gov
COPD Readmission Challenge

Can we develop a model to help Pulmonary Navigators identify which COPD patients are most likely to experience an exacerbation that would lead to a readmission?
COPD Model Example

Total number of respiratory disease index admissions: 90,312
Total number of features: 29
Final number of features used: 19
COPD Model Example

Variable importance
Health Catalyst used the random forest's Gini impurity index along with area under the ROC curve (AUROC) maximization to determine which variables to incorporate into the final model. By removing variables below 0.1 on this index that did not have an impact on the AUROC (once removed), fourteen input variables were shown to account for the most significant impact on readmission prediction. The final variables are detailed below.

Variables considered

- Acute exacerbation
- Admitted through ED
- Age
- Arrive Date
- Barriers to medication
- BMI
- COPD stage
- County
- Facility
- Financial class
- Gender
- History of dementia
- History of depression
- History of diabetes
- History of heart failure
- History of psychoses
- History of renal disease
- History of vascular disease
- Living arrangement
- Palliative care
- Passed through ICU
- Powerplan utilization
- Primary care physician
- Prior inpatient admissions
- Resp. discharge medications
- Self-reported disability
- Smoking status
- Walking limitation

*Crossed-out variables were considered but didn’t increase the model’s accuracy.*
COPD Readmissions

Note: Data is from de-identified data set and in some places fabricated in order to show a reasonable representation of actual trends and observations from production data. All names, addresses, and other PHI are fabricated.
Likelihood of No Shows

Defining the Model

<table>
<thead>
<tr>
<th>The question</th>
<th>data used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Which patients are likely to no-show to their</td>
<td>Total number of appointments: ~10</td>
</tr>
<tr>
<td>scheduled appointments?</td>
<td>million</td>
</tr>
<tr>
<td></td>
<td>Total number of input variables: 30</td>
</tr>
<tr>
<td></td>
<td>Final number of input variables used: 14</td>
</tr>
</tbody>
</table>

Variable importance

Health Catalyst used the random forest Gini impurity index to determine which variables to incorporate into the final model. By removing those below 0.1 on this index, fourteen input variables were shown to account for the most significant impact on no-show prediction. The final variables are detailed below:

Variables considered

| Appointment Department                       | Days from Scheduled               |
| Region                                        | to Appointment                     |
| Appointment Duration                          | Days to Next Holiday               |
| Appointment Time                              | Distance in Miles (patient zip to  |
| Appointment Type                              | location zip)                      |
| Benefit Plan                                  | Ethnic Group                       |
| Cost Center                                   | Financial Class                    |
| Day of Week                                   | Gender                             |
|                                              | Language                           |
|                                              | Location-ID                        |
|                                              | Month                              |
|                                              | Operating Hours Group              |
|                                              | Patient Age                        |
|                                              | Race                               |
|                                              | Recency                            |
|                                              | Running                            |
|                                              | Cancellations                       |
|                                              | Running No Shows                   |
|                                              | Service Line                       |
|                                              | Week of Year                       |
|                                              | Provider-Specialty                 |
|                                              | Provider-Type                      |
|                                              | Race                               |
|                                              | Recency                            |
|                                              | Running                            |
|                                              | Cancellations                      |

*Crosset-out variables were considered but didn’t increase the model’s accuracy.

Choice of model

A Random Forest model was used to calculate the relative impact of the above variables in respect to the appointment status of no-show. The model was created using 200 trees. The performance was as follows:

- Model performance: AUROC (c-statistic): 0.88
- Example cut point of .065: True-positive rate (Sensitivity) 0.800; False-positive rate (1-Sensitivity): 0.218

Deploying the Model in the Patient Access Application

This model has been deployed into an output table named SAM.PracticeManagement.NoShowOutput using the above logic. A probability score for each appointment is calculated and appended to an output table each time on the same schedule as the Patient Access subject area mart (SAM) refresh. This output table joins to the Patient Access SAM using AppointmentID.

References


Likelihood of No Shows
### CLABSI

<table>
<thead>
<tr>
<th>The question</th>
<th>data used</th>
</tr>
</thead>
<tbody>
<tr>
<td>For patients with a central line, what is their risk of CLABSI over the course of the encounter?</td>
<td>Total number of central-line cases: 71,019</td>
</tr>
<tr>
<td></td>
<td>Total number of input variables: 23</td>
</tr>
<tr>
<td></td>
<td>Final number of input variables used: 16</td>
</tr>
</tbody>
</table>

### Variable importance

Two-sample t-tests of input variables were used against the CLABSI result label to determine which variables should be included in the final model. Sixteen input variables accounted for the most significant impact on CLABSI prediction. Including additional input variables beyond these sixteen did not materially improve the model accuracy for this data set. The final variables are detailed below.

### Variables considered

- **AgeInDays**
- **HistoryCLABSI**
- **HistoryHIV**
- **HistoryImmuneDeficiency**
- **HistoryLeukemia**
- **HistoryLymphoma**
- **HistoryNeutropenia**
- **LineDays**
- **LineDaysFemoral**
- **LineDaysInternal Jugular**
- **LineDaysMultilumen**
- **LineDaysNonTunneled**
- **LineDaysPercutaneous**
- **LineDaysPort**
- **LineDaysTunneled**
- **Parenteral Nutrition**
- **Pregnancy**
- **Sex**
- **DaysSinceAdmit**
- **DaysBeforePlacement**

*Crossed-out variables were considered but didn’t increase the model’s accuracy.*

### Choice of model

A Random Forest model was used to calculate the relative impact of the above variables in respect to the labeled outcome of patient infection. The model was created using the Gini criterion and 2,000 trees. The performance was as follows:

- Model performance: AUROC (c-statistic): 0.870
- Example cut point: True-positive rate (Sensitivity): 0.816; False-positive rate (1-Sensitivity): 0.169

### Deploying the Model in the CLABSI Application

This model has been deployed directly into the CLABSI subject area mart (SAM) using the above logic. A risk score for any patients that receive a central line is calculated and appended to an output table each time the SAM refreshes.

### Reference

Get Buy-In
Getting Buy-In from Clinicians

“My patients are sicker.”

“You have a FALSE POSITIVE rate of what?”
Tips for Getting Buy In from Clinicians

#1 Clinicians need to understand the model

If you cannot explain the algorithm, do not use it. Use a simpler algorithm that you can explain.
Tips for Getting Buy In from Clinicians

#2 Build a “model performance report”

Documentation for any interested stakeholder to learn about the model:

- Why was it created?
- What features were tried? Which were used?
- What algorithm was used?
- How accurate is the model?
Tips for Getting Buy In from Clinicians

#3 Provide details to end users
Tips for Getting Buy In from Clinicians

#4 It’s just a suggestion

“Suggestive Analytics” may be a better term than “Predictive Analytics” to demonstrate that we are not trying to replace human judgement.
Review

Useful vocabulary for discussing predictive analytics

Usefulness of custom predictive models

The steps to build a predictive model

Examples of how predictive analytics has been deployed in the wild

Tips for getting buy-in from clinicians
## You Need Smart People!

<table>
<thead>
<tr>
<th>Data Scientist</th>
<th>Data Architect (Engineer)</th>
<th>Machine Learning Engineer</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Formulates hypotheses about features driving a predictive model (with clinical input)</td>
<td>• Finds and provisions source data</td>
<td>• Develops software to automate machine learning workflow</td>
</tr>
<tr>
<td>• Tries various algorithms to determine best approach for prediction</td>
<td>• Leverages definitions in analytics environment</td>
<td>• Requires data science knowledge</td>
</tr>
<tr>
<td>• Assesses model output and accuracy and operationalizes the best approach</td>
<td>• <strong>Feature engineering</strong></td>
<td>• Requires knowledge of software engineering best practices</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• <strong>A rare find!</strong></td>
</tr>
</tbody>
</table>
Our open-source machine learning software product

Automates key tasks in developing models, or customizing existing models using local data

Makes deployment in an analytics environment easy and ‘production quality’
Machine Learning in Healthcare: Now for Everyone

Machine learning for healthcare just got a whole lot easier

The healthcare.ai packages are designed to streamline healthcare machine learning. They do this by including functionality specific to healthcare, as well as simplifying the workflow of creating and deploying models.

Learn more about machine learning via the healthcare.ai community by reading and subscribing to our weekly blogs, viewing our weekly healthcare live event broadcasts, and engaging our data science experts.

LEARN & ENGAGE WITH OUR EXPERTS

Getting Started in R and RStudio
Hosted by Levi Therken, Director of Data Science, Health Catalyst
Thursday, February 23, 2017 3:00pm - 6:00pm

WATCH LIVE

Machine Learning Broadcast Topic
TBD