Session 24

Deploying Predictive Analytics: A Practitioner’s Guide

Eric Just
Senior Vice President
Health Catalyst
Learning Objectives

• How and why machine learning and predictive analytics are increasingly influencing healthcare improvement

• Common mistakes with scaling predictive analytics

• Three key principles to enable scalable predictive analytics
Poll Question #1

How important are predictive analytics for the future of healthcare?

1) Not at all important
2) Low importance
3) Neutral
4) Moderately important
5) Extremely important
6) Unsure or not applicable
Predictive analytics is about using pattern recognition to predict future events but...

Predicting something is not good enough; you must have the data to act and intervene.
What Is Machine Learning?

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

Within the field of data analytics, machine learning is a method used to devise complex models and algorithms that lend themselves to prediction. In commercial use, this is known as **predictive analytics**.

https://en.wikipedia.org/wiki/Machine_learning
Predictive Analytics in Healthcare: “Classic” Approaches

A new method of classifying prognostic comorbidity in longitudinal studies: development and validation.
Charlson ME, Pompei P, Ales KL, Mackenzie CR.

Abstract
The objective of this study was to develop a prospectively applicable method for classifying comorbid conditions which might alter the risk of mortality for use in longitudinal studies. A weighted index that takes into account the number and the seriousness of comorbid disease was developed in a cohort of 559 medical patients. The 1-yr mortality rates for the different scores were: "0", 12% (181); "1-2", 26% (225); "3-4", 52% (71); and "greater than or equal to 5", 85% (82). The index was tested for its ability to predict risk of death from comorbid disease in the second cohort of 685 patients during a 10-yr follow-up. The percent of patients who died of comorbid disease for the different scores were: "0", 8% (588); "1", 25% (54); "2", 48% (25); "greater than or equal to 3", 59% (18). With each increased level of the comorbidity index, there were stepwise increases in the cumulative mortality attributable to comorbid disease (log rank ch 2 = 165; p less than 0.0001). In this longer follow-up, age was also a predictor of mortality (p less than 0.001). The new index performed similarly to a previous system devised by Kaplan and Feinstein. The method of classifying comorbidity provides a simple, readily applicable and valid method of estimating risk of death from comorbidity disease for use in longitudinal studies. Further work in larger populations is still required to refine the approach because the number of patients with any given condition in this study was relatively small.

Charlson Index, 1987

LACE Index, 2010
What Has Happened Since 2010?

**Using the LACE index to predict hospital readmissions in congestive heart failure patients.**
Wang H¹, Robinson RD, Johnson C, Zenerosa NR, Jayswal RD, Keithley J, Delaney KA.

**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.

**Predicting readmissions: poor performance of the LACE index in an older UK population.**
Cotter PE¹, Bhalla VK, Wallis SJ, Biram RW.

**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.
What Has Happened Since 2010?

ML – Based Predictive Analytics

Better Machine Learning Tools

Limitations on ‘Index’ models

Analytics more pervasive

Data availability
Poll Question #2

What was or would be the biggest barrier to implementing predictive analytics?

a) We do not have the people or skills
b) We do not have the right data or technical tools/infrastructure
c) We do not have executive support or budget
d) Past efforts have failed to show results
e) Other
f) Unsure or not applicable
Predictive analytics is easy (or at least easier!)

Organizations are struggling with making predictive analytics routine, pervasive, and actionable.
Typical ‘Current State’ for Predictive Analytics

Data Source → Gnarly SQL Query → Data Manipulation → Tools/Algorithms (SAS, Weka, R, Python) → Predictive Model → Deploy → ?
Three Key Recommendations for Scaling Predictive Analytics

**Fully leverage your analytics environment**

**Standardize tools and methods using production quality code**

**Deploy with a strategy for intervention**

Data Source

Predictive Model

Gnarly SQL Query

Data Manipulation

Tools/Algorithms

SAS | Weka | R | Python

Deploy with a strategy for intervention
Fully Leverage Your Analytics Environment
What is a Feature?

“In machine learning and pattern recognition, a feature is an individual measurable property of a phenomenon being observed. Choosing informative, discriminating and independent features is a crucial step for effective algorithms in pattern recognition, classification and regression.”

https://en.wikipedia.org/wiki/Feature_(machine_learning)
Leverage Your Analytics Environment

• A data warehouse provides access to raw data and pre-defined data like:
  ▪ Clinical registries
  ▪ Comorbidity models (i.e. Charlson Score)
  ▪ Readmissions
  ▪ Length of stay
  ▪ Other calculated fields

• Read-only access is not enough!
The Dangers of ‘Polypharmacy,’ the Ever-Mounting Pile of Pills

Paula Span
THE NEW OLD AGE  APRIL 22, 2016

Comorbidity-Polypharmacy Scoring Facilitates Outcome Prediction in Older Trauma Patients

David C. Evans, MD,* Charles H. Cook, MD,* † Jonathan M. Christy, MD, ‡ Claire V. Murphy, PharmD, § Anthony T. Gerlach, PharmD, ‡‡ Daniel Eiferman, MD,* David E. Lindsey, MD,* Melissa L. Whitmill, MD,* Thomas J. Papadimos, MD, MPH, † Paul R. Beery, II, MD,* Steven M. Steinberg, MD,* † and Stanislaw P. A. Stawicki, MD*

Article in Journal of the American Geriatrics Society · July 2012
DOI: 10.1111/j.1532-5415.2012.04075.x · Source: PubMed
## Polypharmacy Data Mart

<table>
<thead>
<tr>
<th>PatientID</th>
<th>MedicationID</th>
<th>StartDT</th>
<th>EndDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z006F600A51F</td>
<td>10994</td>
<td>7/27/2006</td>
<td>NULL</td>
</tr>
<tr>
<td>Z006F600A51F</td>
<td>41801</td>
<td>7/27/2006</td>
<td>NULL</td>
</tr>
<tr>
<td>Z006F600A51F</td>
<td>10994</td>
<td>7/27/2006</td>
<td>NULL</td>
</tr>
<tr>
<td>Z006F600A51F</td>
<td>41801</td>
<td>7/27/2006</td>
<td>NULL</td>
</tr>
<tr>
<td>Z006F600A51F</td>
<td>10994</td>
<td>7/27/2006</td>
<td>NULL</td>
</tr>
<tr>
<td>Z13148BF2583</td>
<td>4996</td>
<td>9/14/2005</td>
<td>11/15/2005</td>
</tr>
<tr>
<td>Z13148BF2583</td>
<td>11798</td>
<td>9/14/2005</td>
<td>11/15/2005</td>
</tr>
<tr>
<td>Z13148BF2583</td>
<td>15061</td>
<td>9/14/2005</td>
<td>11/15/2005</td>
</tr>
</tbody>
</table>

### Step 1:
**Clean up data**
- Missing end dates
- One time doses

### Step 2:
**Aggregate into polypharmacy count at each encounter**

<table>
<thead>
<tr>
<th>PatientEncounterID</th>
<th>PolypharmacyCNT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1048826</td>
<td>6</td>
</tr>
<tr>
<td>1048912</td>
<td>0</td>
</tr>
<tr>
<td>1048923</td>
<td>0</td>
</tr>
<tr>
<td>1048924</td>
<td>0</td>
</tr>
<tr>
<td>1048925</td>
<td>0</td>
</tr>
<tr>
<td>1048926</td>
<td>0</td>
</tr>
<tr>
<td>1049094</td>
<td>2</td>
</tr>
<tr>
<td>1049095</td>
<td>2</td>
</tr>
<tr>
<td>1049096</td>
<td>2</td>
</tr>
<tr>
<td>1049097</td>
<td>3</td>
</tr>
<tr>
<td>1049098</td>
<td>3</td>
</tr>
<tr>
<td>1049099</td>
<td>4</td>
</tr>
<tr>
<td>1049100</td>
<td>2</td>
</tr>
</tbody>
</table>
What Is Feature Engineering?

“Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy…”

Jason Brownlee in “Discover Feature Engineering, How to Engineer Features and How to Get Good at It”


“Much of the success of machine learning is actually success in engineering features that a learner can understand.”

Scott Locklin in “Neglected Machine Learning Ideas”

https://scottlocklin.wordpress.com/2014/07/22/neglected-machine-learning-ideas/
Other Examples of Feature Engineering

- Number of ER visits in the last year
- Line days
- Number and types of comorbid conditions
- Almost any input into a predictive model will be engineered in some way

The ability for data scientists to engineer features is critical to a successful predictive analytics/machine learning strategy.
Fully Leverage Your Analytics Environment

Most feature engineering should be done in the analytics environment (data warehouse).

Give data scientists enough access to data warehouse to promote efficient re-use of engineered features.

Use standard data warehouse ETL tools to operationalize engineered features.
Standardize Tools and Methods Using Production-Quality Code
Three Key Recommendations for Scaling Predictive Analytics

1. Fully leverage your analytics environment
2. Standardize tools and methods using production quality code
3. Deploy
You Need Lots of Smart People!

**Data Scientist**
- Formulates hypotheses about features driving a predictive model (with clinical input)
- Tries various models to determine best approach for prediction
- Assesses model output and accuracy and operationalizes the best approach

**Machine Learning Engineer**
- Develops software leveraged by data scientists to test and deploy their models
- Requires data science knowledge
- Requires knowledge of software engineering best practices
- **A rare find!**
Predictive Analytics Processes

Development Process

1. Build model experiment (~30-40 features)
2. Split data into train and test
3. Run multiple algorithms
4. Measure & report performance
5. Select best algorithm & important features (~10 features)
6. Store parameters

Running the Model

1. Load model parameters
2. Receive patient 'record' (~10 features)
3. Calculate prediction
4. Output prediction
Developing a Machine Learning Code Base

Why?

- Focus data scientist on model development – not writing code or reinventing the wheel
- Standardize methodologies so that best practices can be deployed
- Predictive models in production require production quality code
Developing a Machine Learning Code Base: Best Practices

Version control

• Documentation
• Continuous integration

Unit Testing

Also,

This is production code – it should be treated as such!
Developing a Machine Learning Code Base: Technology Choices

**R**
- Open source, deeply entrenched in healthcare
- More familiar to analysts/statisticians

**Python**
- Open source, newer approach (with lots of momentum)
- More familiar to developers

**AzureML**
- Cloud-based
- Easy to deploy

**Plenty of other choices**
Our Code Base Includes:

<table>
<thead>
<tr>
<th>Data Ingestion/Preparation</th>
<th>Model Development</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Load data from database and/or CSV</td>
<td>• Split test and train</td>
<td>• Model performance reports</td>
</tr>
<tr>
<td>• Impute missing values</td>
<td>• Feature selection</td>
<td>• Trend identification</td>
</tr>
<tr>
<td>• Remove ‘bad’ data (rows and columns)</td>
<td>• Run algorithms</td>
<td>• Risk adjusted comparisons</td>
</tr>
<tr>
<td>• Date/Time expansion (i.e. Day of week, week of year)</td>
<td>▪ Random Forest</td>
<td></td>
</tr>
<tr>
<td></td>
<td>▪ Lasso</td>
<td></td>
</tr>
<tr>
<td></td>
<td>▪ Mixed Models (Q4 2016)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>▪ k-means (Q4 2016)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Evaluate model performance</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Save model</td>
<td></td>
</tr>
</tbody>
</table>
Scaling People

- **Data Architects**
  - Great domain knowledge
  - Often looking for opportunities to advance career/skills

- **With the right tools…**
  - Data architects make great feature engineers
  - Data architects can easily get started in predictive analytics

“One awesome thing about the output from the R [package] you put together is the output aligns perfectly with creating Patient Stratification algorithms. ...The fact that I feel comfortable running this stuff speaks to how easy you have made it. Thanks again, Levi.”
Putting Predictive Models in Production
Modality #1: Extract, Transform, Load (ETL) Process

Deploy in this modality if:

- Prediction is not based on highly dynamic data
- Prediction is used in analytic application
- Intervention strategy ‘OK’ with some latency (up to 24 hours)
- Example: Readmission prediction
Modality #2: Web Services

Deploy in this modality if:

- Prediction is dynamic data
- Prediction is used in workflow application
- Intervention strategy not ‘OK’ with some latency (up to 24 hours)
- Example: Sepsis early detection
Scaling Predictive Analytics

- Fully leverage your analytics environment
- Standardized tools and methods using production quality code
- Deploy with a strategy for intervention

Data Source ➔ Data Manipulation ➔ Tools/Algorithms (SAS, Weka, R, Python) ➔ Predictive Model ➔ Deploy with a strategy for intervention

Standardized tools and methods using production quality code

#HASUMMIT16
Deploy With a Strategy for Intervention
Case Study: Central Line Associated Bloodstream Infection (CLABSI)

- Approximately 41,000 patients with central lines will end up with a bloodstream infection (CLABSI)
- One in four patients with a CLABSI will die
- CLABSI improvement team looking at compliance with evidence-based guidelines
- Retrospective analysis led to increased insight into problem areas and associated interventions
- Team wanted more pro-active notification of high-risk patients
- Developed predictive algorithm based on 16 features
MODEL PERFORMANCE REPORT
RISK OF CLABSI

SUMMARY

Predictive Question: For patients with a central line, what is their risk of CLABSI over the encounter?

- Total # of past cases: 70,218
- Total # of input variables: 23
- Final # of input variables used: 16

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.

<table>
<thead>
<tr>
<th>Variables Considered</th>
<th>ApnInDays</th>
<th>HistoryCLABSI</th>
<th>LineDays</th>
<th>LineDaysPort</th>
</tr>
</thead>
<tbody>
<tr>
<td>AgeAtAdmission</td>
<td>HistoryAge</td>
<td>LineDaysTotal</td>
<td>LineDaysTunneled</td>
<td>ParenteralNutrition</td>
</tr>
<tr>
<td>DayBeforePlacement</td>
<td>HistoryPneumoniaDeficiency</td>
<td>LineDaysTunneled</td>
<td>RouteBathingNonCompliant</td>
<td></td>
</tr>
<tr>
<td>DaysSinceAdmit</td>
<td>Historylymphoma</td>
<td>LineDaysPort</td>
<td>RoutineBathingNonCompliant</td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>HistoryNeutropenia</td>
<td>LineDaysTunneled</td>
<td>RoutineBathingNonCompliant</td>
<td></td>
</tr>
</tbody>
</table>

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 1000 trees. The performance was as follows:

- Model performance: AUROC: 0.871
  - Example cut point: True positive rate: 0.81 (Sensitivity), False positive rate: 0.16 (1-Sensitivity)

Deployment

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

References


POWERED BY DATA SCIENCE @ HEALTH CATALYST
What Does It Look Like?

Key features:

- Part of broader CLABSI effort
- Team in place was already responding to care gaps exposed in the application
- Workflow (daily huddle) was amenable to up to 24 hour latency
- High risk patients easy-to-identify
- Show risk factors in addition to the risk score – enables intervention
Models Built To Date:

**Built**

- Central line-associated bloodstream infection (CLABSI) – Clinical Decision Support
- Forecast IBNR claims/year-end expenditures – Financial Decision Support
- Congestive Heart Failure, Readmissions Risk – Clinical Decision Support
- Predictive Risk & Cost – Population Health and Accountable Care
- Patient Flight Path, Diabetes Future Risk – Clinical Decision Support
- Patient Flight Path, Diabetes Future Cost – Clinical Decision Support
- Patient Flight Path, Diabetes Top Treatments – Clinical Decision Support
- Patient Flight Path, Diabetes Next Likely Complications (Glaucoma) – Clinical Decision Support
- Patient Flight Path, Diabetes Next Likely Complications (Retinopathy) – Clinical Decision Support
- Patient Flight Path, Diabetes Next Likely Complications (ESRD) – Clinical Decision Support
- Plus several more… (Nephropathy, Cataracts, CHF, CAD, Ketoacidosis, Erectile Dysfunction, Foot Ulcers)

**In Development**

- Predictive appointment no shows – Operations and Performance Management
- Propensity to pay – Financial Decision Support
- Pre-surgical risk (Bowel) – Clinical Decision Support and client request
- Post-surgical risk (Hips and Knees) – Clinical Decision Support
- Patient Flight Path, Congestive Heart Failure (5-6 new flight path algorithms similar to Patient Flight Path, Diabetes below)
- Patient Flight Path, Coronary Artery Disease (5-6 new flight path algorithms similar to Patient Flight Path, Diabetes below)
- Geo-spatial health system service area definition, network referral/leakage
- INSIGHT socio-economic based risk – Clinical Decision Support and client request
- Native SQL/R predictive framework and standard package - Platform
- Feature selection, Parallel Models, Rank and Impact of Input Variables – Platform
- Predictive ETL batch load times – Platform

**Planned**

- Early detection of CLABSI, CAUTI, Clostridium difficile (c. diff) hospital infections – Clinical Decision Support
- Early detection of Sepsis/Septicemia (Blood Infection) – Clinical Decision Support
- Public data sets, benchmarks, “Catalyst Risk”, expected mortality, length of stay – CAFÉ collaboration
- Clusters of population risk (near term risk/cost) – Population Health and Accountable Care
Poll Question #3

What are or would be the top three most important data sources to your organization in making predictions? (select 3, if applicable)

a) Clinical EMR data
b) Claims data
c) Patient outcomes data
d) Financial data
e) Non-medical patient data (e.g. socioeconomic, behavioral)
f) Patient satisfaction data
g) Unsure or not applicable
Scaling Predictive Analytics

- Fully leverage your analytics environment
- Standardized tools and methods using production quality code
- Deploy with a strategy for intervention

Data Source → Predictive Model

Gnarly SQL Query → Data Manipulation → Tools/Algorithms (SAS | Weka | R | Python) → Deploy with a strategy for intervention

Standardized tools and methods using production quality code
What the Future Holds
Closed Loop Architecture

Clinical Workflow Engine

Web/Mobile Apps (SMART on FHIR)

EHR
Analytic insights drive:
- Alerting
- Ordering
- Refills
- Inbox
- Diagnosis
- Referrals
- Dynamic screen generation
- Suggestions
- Risk management

API (FHIR)

Algorithm Library

Data Warehouse
- Local patient data
- Clinical trials
- Social determinants data
- Patient reported outcomes
- Regional and National Data Sets
- Activity Based Costing
- Genomic (+ other ‘omics)
- Environmental
- Geospatial
- Device data (includes fitness)
- Patient engagement

Registry Definitions

Text Processing/NLP Algorithms

Clustering Algorithms (Patients Like This)

Readmission Predictors

High and rising risk predictors

CAFÉ 65M+ Patient Records

Integrated Data Repository

Clinicians need analytic insight delivered in their natural workflow

Patients need to be involved, too!

Patients need to be involved, too!
Lessons Learned

• Developing a predictive analytics capability that scales requires the right analytics environment, the right tools, and the right deployment strategy.

• The foundation of a good predictive model is good model input, or “features.” An analytics environment provides a good environment for “feature engineering,” which allows domain experts to manipulate data to create better model inputs.

• Operationalizing machine learning and predictive analytics requires production-ready software. Good engineering principles like version control and automated testing should be applied to software that is being used to generate and deploy predictive models.

• Deploying actionable predictive analytics should be done in the context of an intervention. The predictive value should be displayed with information that spurs intervention, like specific factors driving the prediction.
Analytic Insights

Questions & Answers
What You Learned…

Write down the key things you’ve learned related to each of the learning objectives after attending this session
Thank You