Digging Into the Modeling Lifecycle
About the Presenters

- **Greg Frankowiak**, FCAS, MAAA, CSPA, CPCU, MSM
  - Senior Consulting Actuary
  - Bloomington, Illinois

- **Radost Wenman**, FCAS, MAAA, CSPA
  - Consulting Actuary
  - San Francisco, CA
Agenda

• Overview of modeling lifecycle
• Data usage and quality
• Validation/quality checks
• Retention example
• Implementation considerations
• Change management
• Model monitoring
What insurance applications have you seen predictive modeling used for? Select all applicable.

A. Pricing
B. Underwriting
C. Claims
D. Marketing
E. None

Polling Question
Modeling Lifecycle

- Many uses of predictive modeling within insurance
- Opportunity to utilize data more fully to address business challenges
Modeling Lifecycle

The models may get the glory... but there is so much more to it!
Modeling Lifecycle

- Business Question
- Data
- Methods
- EDAs
- Model Building
- Model Validation
- Model Implementation
- Model Monitoring
Modeling Lifecycle

• Communicating and understanding the business question
• How will the results of the model be implemented
• Systems implementation considerations
• Documentation
Data Usage/Data Quality

• Data is the raw material that fuels analysis

• End goal is to change raw material into something useful—turning data into actionable insights

• Sometimes this gets lost—and the data is seen as the end goal

• 75% of the time is spent on data!
Data Usage/Data Quality

• Data quality for data stores

• Data quality during modeling process

• Consider using 3rd party data vs. agent or customer-entered data from a quality and consistency standpoint

• Metadata
Validation/Quality Checks

- Need a plan for validation along the way to ensure foundational building blocks are correct—data, code
- Consider having others evaluate the methodology
Validation/Quality Checks

• It’s right, but is it good?

• Various ways that models are evaluated
Model Validation

COMP Frequency Lift Chart - Validation Data

freq_act
freq_pred
# Modeling Applications for Insurance – Classification Case

<table>
<thead>
<tr>
<th>Pricing</th>
<th>Underwriting</th>
<th>Claims</th>
<th>Marketing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating plan development</td>
<td>Straight through processing</td>
<td>Fraud</td>
<td>Characteristics of shoppers/quoters</td>
</tr>
<tr>
<td>Vehicle classification</td>
<td>Selection/rejection</td>
<td>Claim settlement value</td>
<td>Likelihood of insureds responding to</td>
</tr>
<tr>
<td>Custom insurance scores</td>
<td>Target report ordering (MVR, CLUE)</td>
<td>Early warning indicator</td>
<td>marketing initiatives</td>
</tr>
<tr>
<td>Territory Definition</td>
<td>Action indicators</td>
<td>Close w/o pay</td>
<td>Advertising effectiveness</td>
</tr>
<tr>
<td>Homeowner by peril pricing</td>
<td>Vehicle inspection/re-inspection</td>
<td>Chance to reopen</td>
<td>Retention/Conversion Analysis</td>
</tr>
<tr>
<td>Expanded SDIP</td>
<td>Home inspection/re-inspection</td>
<td>Chance to close within a set timeframe</td>
<td></td>
</tr>
<tr>
<td>Tier plans/Scorecards</td>
<td></td>
<td>Chance for a large adjustment to reserves</td>
<td></td>
</tr>
<tr>
<td>Usage based insurance</td>
<td></td>
<td>Chance for litigation</td>
<td></td>
</tr>
<tr>
<td>Price optimization</td>
<td></td>
<td>Early warning indicator</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Claim assignment</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Service provider evaluation</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Salvage / Subrogation</td>
<td></td>
</tr>
</tbody>
</table>

- [Retention/Conversion Analysis](#)
Retention Example

Who are the customers most likely to churn?

Why would they churn?

What is the length of time that the customer will stay?
Retention Example: Two Types of Attrition Models

**Who?**
- Treats attrition as a binary response
- Predicts which customers will leave

**How Long?**
- Treats attrition as a lifetime response
- Predicts how long customers will stay
Retention Example - Dataset Description

- IBM sample dataset from telecommunication market
- Data contains fictitious customer data—available from IBM Watson Community
- Rows represent customers and columns are attributes
- Information includes:
  - Demographic info (age, gender, dependents)
  - Account info (tenure, contract term, payment and billing method)
  - Types of services (phone, internet, streaming TV, tech support)
  - Customer active status (churn)
Retention Example – R Code

> my_packages = c('caret', 'caTools', 'DataExplorer', 'e1071',
>                  'jtools', 'rpart', 'rpart.plot', 'summarytools')

> sapply(my_packages, require, character.only = TRUE)
Retention Example – R Code

> introduce(dataset)

```
rows       7043
columns    21
discrete_columns 17
continuous_columns  4
all_missing_columns   0
total_missing_values 11
complete_rows   7032
total_observations 147903
memory_usage     1169928
```
## Retention Example – R Code

```r
> view(dfSummary(dataset))
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Stats / Values</th>
<th>Freqs (% of Valid)</th>
<th>Graph</th>
<th>Valid</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>gender</td>
<td>1. Female</td>
<td>3488 (49.5%)</td>
<td></td>
<td>7043</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>2. Male</td>
<td>3555 (50.5%)</td>
<td></td>
<td>(100%)</td>
<td>(0%)</td>
</tr>
<tr>
<td>tenure</td>
<td>Mean (sd) : 32.4 (24.6)</td>
<td></td>
<td></td>
<td>7043</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>min &lt; med &lt; max: 0 &lt; 29 &lt; 72</td>
<td></td>
<td></td>
<td>(100%)</td>
<td>(0%)</td>
</tr>
<tr>
<td></td>
<td>IQR (CV) : 46 (0.8)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>73 distinct values</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TotalCharges</td>
<td>Mean (sd) : 2283.3 (2266.8)</td>
<td></td>
<td></td>
<td>7032</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>min &lt; med &lt; max: 18.8 &lt; 1397.5 &lt; 8684.8</td>
<td></td>
<td></td>
<td>(99.84%)</td>
<td>(0.16%)</td>
</tr>
<tr>
<td></td>
<td>IQR (CV) : 3393.3 (1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6530 distinct values</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Churn</td>
<td>1. No</td>
<td>5174 (73.5%)</td>
<td></td>
<td>7043</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>2. Yes</td>
<td>1869 (26.5%)</td>
<td></td>
<td>(100%)</td>
<td>(0%)</td>
</tr>
</tbody>
</table>
Retention Example – R Code

```r
> barplot(prop.table(table(Churn, gender),2), legend = c('Yes', 'No'),
       args.legend = list(title = 'Churn', x = 'topleft', cex = .9), horiz = TRUE,
       ylim = c(0.5, 8.7), xlim = c(0, 1.3), width = c(2, 2), space = 0.3)

> title('Gender', line = -7.5, adj = 0.35)
```
Retention Example – R Code

```r
> boxplot(tenure ~ Churn, width = c(1,1), xlim = c(0.5,5),
         legend = c('No', 'Yes'), frame = FALSE, horizontal = TRUE)

> title('Tenure', line = -10.5, adj = 0.5)
```
Retention Example – R Code

```R
> train_test = sample.split(Churn, SplitRatio = .70)
> tab = table(Churn, train_test)
> tab

<table>
<thead>
<tr>
<th></th>
<th>FALSE</th>
<th>TRUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>1549</td>
<td>3614</td>
</tr>
<tr>
<td>Yes</td>
<td>561</td>
<td>1308</td>
</tr>
</tbody>
</table>

> prop.table(tab, margin = 1)

<table>
<thead>
<tr>
<th></th>
<th>FALSE</th>
<th>TRUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>0.3000194</td>
<td>0.6999806</td>
</tr>
<tr>
<td>Yes</td>
<td>0.3001605</td>
<td>0.6998395</td>
</tr>
</tbody>
</table>

> trainSet = subset(dataset, train_test == TRUE)
> testSet = subset(dataset, train_test == FALSE)
```
Retention Example – R Code

```r
> LogMod = glm(Churn ~ ., family = binomial(link = 'logit'), data = trainSet)
> plot_summs(LogMod, scale = T)
```
Retention Example – R Code

> trainProb = predict(LogMod, newdata = trainSet, type = 'response')
> trainPred = factor(ifelse(trainProb >= 0.5, 'Yes', 'No'))
> trainActual = factor(ifelse(trainSet$Churn == 'Yes', 'Yes', 'No'))

> testProb = predict(LogMod, newdata = testSet, type = 'response')
> testPred = factor(ifelse(testProb >= 0.5, 'Yes', 'No'))
> testActual = factor(ifelse(testSet$Churn == 'Yes', 'Yes', 'No'))
Retention Example – R Code

> trainConf = confusionMatrix(data = trainPred, reference = trainActual)
> testConf = confusionMatrix(data = testPred, reference = testActual)
Retention Example – R Code

> fourfoldplot(trainConf$table)
> fourfoldplot(testConf$table)
Retention Example – R Code

> TreeMod = rpart(Churn ~ ., data = trainSet, cp = .005)
> rpart.plot(TreeMod, type = 4, clip.right.labs = FALSE, branch = .3,
Retention Example – R Code

> svmMod = svm(Churn ~ ., type = 'C-classification', kernel = 'radial', data = trainSet)

> trainProb = predict(svmMod, newdata = trainSet, probability = T)
> trainPred = factor(ifelse(attr(trainProb, 'probabilities')['Yes'] >= 0.5, 'Yes', 'No'))
> trainActual = factor(ifelse(trainSet$Churn == 'Yes', 'Yes', 'No'))

> testProb = predict(svmMod, newdata = testSet, probability = T)
> testPred = factor(ifelse(attr(testProb, 'probabilities')['Yes'] >= 0.5, 'Yes', 'No'))
> testActual = factor(ifelse(testSet$Churn == 'Yes', 'Yes', 'No'))

> trainConf = confusionMatrix(data = trainPred, reference = trainActual)
> testConf = confusionMatrix(data = testPred, reference = testActual)

SVM Accuracy on Hold-out Test Data = 0.782
Retention Example – R Code

10-fold Cross Validation

- Logistic: 0.7999
- SVM: 0.7887
- Decision Tree: 0.7006
## Retention Example – Further Steps for Model Improvement

<table>
<thead>
<tr>
<th>Data</th>
<th>Modeling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable correlations</td>
<td>Neural Networks, Genetic Algorithms, Bayesian</td>
</tr>
<tr>
<td>Data reduction techniques</td>
<td>Ensemble models - bagging, boosting, stacking</td>
</tr>
<tr>
<td>Clustering</td>
<td>Parameter tuning</td>
</tr>
<tr>
<td>Variable transformations</td>
<td>Text mining analytics</td>
</tr>
<tr>
<td>Derived variables</td>
<td>Social media analytics</td>
</tr>
<tr>
<td>Missing and outlying values</td>
<td>Variable selection</td>
</tr>
</tbody>
</table>

Variations in this dataset are evident and require further investigation.
Implementation Considerations

- Implementation considerations can:
  - inform methods or data used
  - complexity of approach

- Is data available in production?

- Can model be put into production or will that take additional work?

- Could data or methods be of concern to regulators or those using its outputs—underwriters, claims handlers, agents, etc.
Implementation Considerations

- Implementation includes the technical aspect—code reviews, test cases, etc.

- Consider pre-implementation and post-implementation testing

- Need data as real-time as possible:
  - Catch and resolve production errors quickly
  - Minimize damage (refunds, lost customers, etc.)
Change Management

• Why don’t they like my awesome model?!

• Change management is something that should be started early

• And not done after implementation
Model Monitoring

Two aspects of monitoring: correctness and business outcomes desired

- Business is trying to accomplish something, monitoring is to see if you are doing that
- Want to monitor inputs and outputs
Model Monitoring

- Suggest starting small and building up—focus on final outcome
- Setting tolerances can streamline the monitoring process
- Goal is to look at more without having a person do so
- Actions taken if out of tolerance
Questions
Join Us for the Next APEX Webinar

Thursday, October 17
2:00 p.m. ET
Registration is Open

An Update to Pinnacle’s Risk Retention Group Benchmarking Study

Rob Walling  Erich Brandt  Greg Fears
Final Notes

• We’d like your feedback and suggestions!
  • Please complete our survey

• For copies of this APEX presentation
  • Visit the Resource Knowledge Center at Pinnacleactuaries.com
Thank You for Your Time and Attention

Greg Frankowiak

gfrankowiak@pinnacleactuaries.com

Radost Wenman

rwenman@pinnacleactuaries.com