Introduction to Predictive Modeling

December 13, 2007
Introduction / Objective

1. What is Predictive Modeling?
2. Types of predictive models.
3. Applications – case studies.
Predictive Modeling: A Review of the Basics
Definition of Predictive Modeling

“Predictive modeling is a set of tools used to stratify a population according to its risk of nearly any outcome... ideally, patients are risk-stratified to identify opportunities for intervention before the occurrence of adverse outcomes that result in increased medical costs.”

“The year 1930, as a whole, should prove at least a fairly good year.”

-- Harvard Economic Service, December 1929
Why do it? Potential Use of Models

Medical Management Perspective

- Identifying individuals at very high risk of an event (death, LTC, disability, annuity surrender, etc.).

- Identify management opportunities and determine resource allocation/prioritization.
Identification – how?

• The art and science of predictive modeling!

• There are many different algorithms for identifying member conditions. THERE IS NO SINGLE AGREED FORMULA.

• Condition identification often requires careful balancing of sensitivity and specificity.
## Identification – example (Diabetes)

### Inpatient Hospital Claims – ICD-9 Claims Codes

<table>
<thead>
<tr>
<th>ICD-9-CM CODE</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>250.xx</td>
<td>Diabetes mellitus</td>
</tr>
<tr>
<td>357.2</td>
<td>Polyneuropathy in diabetes</td>
</tr>
<tr>
<td>362.0, 362.0x</td>
<td>Diabetic retinopathy</td>
</tr>
<tr>
<td>366.41</td>
<td>Diabetic cataract</td>
</tr>
<tr>
<td>648.00-648.04</td>
<td>Diabetes mellitus (as other current condition in mother classifiable elsewhere, but complicating pregnancy, childbirth or the puerperium.</td>
</tr>
</tbody>
</table>
# Diabetes – additional codes

<table>
<thead>
<tr>
<th>CODES</th>
<th>CODE TYPE</th>
<th>DESCRIPTION - ADDITIONAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIABETES;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G0108. G0109</td>
<td>HCPCS</td>
<td>Diabetic outpatient self-management training services, individual or group</td>
</tr>
<tr>
<td>J1815</td>
<td>HCPCS</td>
<td>Insulin injection, per 5 units</td>
</tr>
<tr>
<td>67227</td>
<td>CPT4</td>
<td>Destruction of extensive or progressive retinopathy, (e.g. diabetic retinopathy) one or more sessions, cryotherapy, diathermy</td>
</tr>
<tr>
<td>67228</td>
<td>CPT4</td>
<td>Destruction of extensive or progressive retinopathy, one or more sessions, photocoagulation (laser or xenon arc)</td>
</tr>
<tr>
<td>996.57</td>
<td>ICD-9-CM</td>
<td>Mechanical complications, due to insulin pump</td>
</tr>
<tr>
<td>V45.85</td>
<td>ICD-9-CM</td>
<td>Insulin pump status</td>
</tr>
<tr>
<td>V53.91</td>
<td>ICD-9-CM</td>
<td>Fitting/adjustment of insulin pump, insulin pump titration</td>
</tr>
<tr>
<td>V65.46</td>
<td>ICD-9-CM</td>
<td>Encounter for insulin pump training</td>
</tr>
</tbody>
</table>
Insulin or Oral Hypoglycemic Agents are often used to identify members. A simple example follows; for more detail, see the HEDIS code-set.

<table>
<thead>
<tr>
<th>Insulin</th>
</tr>
</thead>
<tbody>
<tr>
<td>2710* Insulin**</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Oral AntiDiabetics</th>
</tr>
</thead>
<tbody>
<tr>
<td>2720* Sulfonylureas**</td>
</tr>
<tr>
<td>2723* Antidiabetic - Amino Acid Derivatives**</td>
</tr>
<tr>
<td>2725* Biguanides**</td>
</tr>
<tr>
<td>2728* Meglitinide Analogues**</td>
</tr>
<tr>
<td>2730* Diabetic Other**</td>
</tr>
<tr>
<td>2740* Reductase Inhibitors**</td>
</tr>
<tr>
<td>2750* Alpha-Glucosidase Inhibitors**</td>
</tr>
<tr>
<td>2760* Insulin Sensitizing Agents**</td>
</tr>
<tr>
<td>2799* Antiadiabetic Combinations**</td>
</tr>
</tbody>
</table>
All people are not equally identifiable

Definition Examples:

Narrow: Hospital Inpatient (primary Dx); Face-to-face professional (no X-Ray; Lab)

Broad: Hospital I/P (any Dx); All professional

Rx: Narrow + Outpatient Prescription

Prevalence of 5 Chronic conditions

<table>
<thead>
<tr>
<th></th>
<th>Narrow</th>
<th>Broad</th>
<th>Rx</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medicare</td>
<td>24.4%</td>
<td>32.8%</td>
<td>30.8%</td>
</tr>
<tr>
<td>Commercial</td>
<td>4.7%</td>
<td>6.3%</td>
<td>6.6%</td>
</tr>
</tbody>
</table>
### False Positive Identification Incidence through Claims

**Medicare Advantage Population (with drug benefits)**

**Diabetes Example**

<table>
<thead>
<tr>
<th>Year 1</th>
<th>Narrow</th>
<th>+ Broad</th>
<th>+ Rx</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>75.9%</td>
<td></td>
<td>85.5%</td>
<td>92.6%</td>
</tr>
<tr>
<td>Year 2</td>
<td>24.1%</td>
<td>14.5%</td>
<td>7.4%</td>
<td></td>
</tr>
</tbody>
</table>

**TOTAL** 100.0% 100.0% 100.0% 100.0%
Prospective versus Retrospective Targeting

- Last Year's Members: 5%
- Last Year's Costs: 45%
- This Year's Costs: 18%

SOLUCIA, INC.
## Cost Stratification of a Large Population

<table>
<thead>
<tr>
<th>Population</th>
<th>0.0% - 0.5%</th>
<th>0.5% - 1.0%</th>
<th>Top 1%</th>
<th>Top 5%</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>67,665</td>
<td>67,665</td>
<td>135,330</td>
<td>676,842</td>
<td>13,537,618</td>
</tr>
<tr>
<td>Actual Cost</td>
<td>$3,204,433,934</td>
<td>$1,419,803,787</td>
<td>$4,624,237,721</td>
<td>$9,680,579,981</td>
<td>$21,973,586,008</td>
</tr>
<tr>
<td>PMPY Total Actual Cost</td>
<td>$47,357</td>
<td>$20,977</td>
<td>$34,170</td>
<td>$14,303</td>
<td>$1,623</td>
</tr>
<tr>
<td>Percentage of Total Cost</td>
<td>14.6%</td>
<td>6.5%</td>
<td>21.1%</td>
<td>44.1%</td>
<td>100%</td>
</tr>
</tbody>
</table>

**Patients with > $50,000 in Claims**

<table>
<thead>
<tr>
<th>Number of Patients</th>
<th>0.0% - 0.5%</th>
<th>0.5% - 1.0%</th>
<th>Top 1%</th>
<th>Top 5%</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Patients</td>
<td>19,370</td>
<td>5,249</td>
<td>24,619</td>
<td>32,496</td>
<td>35,150</td>
</tr>
<tr>
<td>Percentage of Total</td>
<td>55.1%</td>
<td>14.9%</td>
<td><strong>70.0%</strong></td>
<td>92.4%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>
Program Evaluation/ Reimbursement Perspective

- Predicting *what would have happened* absent a program.
- Predicting resource use in the “typical” population.
Example 1: Time Series

Predicted costs

Actual costs

Savings

$0

$400

$800

$1,200

$1,600

PDMPM

Jan-99 Jan-00 Jan-01 Jan-02 Jan-03 Jan-04

4 years of pre-program data
## Example 2: Normalized resources

<table>
<thead>
<tr>
<th>Member ID</th>
<th>Single Condition</th>
<th>RiskScoreID</th>
<th>PgmCode</th>
<th>NonDup Patient Count</th>
<th>Patient Count x Risk Score</th>
<th>Expected Claims Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>1080</td>
<td>CHF</td>
<td>39.8</td>
<td>200</td>
<td>1</td>
<td>39.774</td>
<td>$58,719</td>
</tr>
<tr>
<td>532</td>
<td>Cancer 1</td>
<td>174.2</td>
<td>100</td>
<td>1</td>
<td>174.189</td>
<td>$210,829</td>
</tr>
<tr>
<td>796</td>
<td>Cancer 2 + Chronic cond.</td>
<td>159.7</td>
<td>100</td>
<td>1</td>
<td>159.671</td>
<td>$1,289,469</td>
</tr>
<tr>
<td>531</td>
<td>Cancer 2 + No Chron. cond</td>
<td>135.3</td>
<td>100</td>
<td>1</td>
<td>135.289</td>
<td>$338,621</td>
</tr>
<tr>
<td>1221</td>
<td>Multiple Chron cond.</td>
<td>28.8</td>
<td>200</td>
<td>1</td>
<td>28.811</td>
<td>$34,660</td>
</tr>
<tr>
<td>710</td>
<td>Acute conds and Chron</td>
<td>110.9</td>
<td>100</td>
<td>1</td>
<td>110.87</td>
<td>$100,547</td>
</tr>
<tr>
<td>795</td>
<td>Acute conds and Chron</td>
<td>121.1</td>
<td>100</td>
<td>1</td>
<td>121.083</td>
<td>$148,107</td>
</tr>
<tr>
<td>882</td>
<td>Diabetes</td>
<td>25.7</td>
<td>200</td>
<td>1</td>
<td>25.684</td>
<td>$22,647</td>
</tr>
<tr>
<td>967</td>
<td>Cardiac</td>
<td>24.5</td>
<td>200</td>
<td>1</td>
<td>24.465</td>
<td>$1,308</td>
</tr>
<tr>
<td>881</td>
<td>Asthma</td>
<td>24.1</td>
<td>200</td>
<td>1</td>
<td>24.096</td>
<td>$15,776</td>
</tr>
</tbody>
</table>

$2,220,683
Actuarial, Underwriting and Profiling Perspectives

- Calculating renewal premium
- Profiling of provider
- Provider & health plan contracting
Types of Predictive Modeling Tools

- Risk Groupers
- Predictive Modeling Tools
- Statistical Models
- Artificial Intelligence
Uses of Risk Groupers

Risk Groupers can be used for these 3 purposes ... but best for actuarial, underwriting and profiling

- Actuarial, Underwriting and Profiling Perspectives
- Medical Management Perspective
- Program Evaluation Perspective
What are the different types of risk groupers?
# Selected Risk Groupers

<table>
<thead>
<tr>
<th>Company</th>
<th>Risk Grouper</th>
<th>Data Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>IHCIS/Ingenix</td>
<td>ERG</td>
<td>Age/Gender, ICD-9 NDC, Lab</td>
</tr>
<tr>
<td>UC San Diego</td>
<td>CDPS</td>
<td>Age/Gender, ICD -9 NDC</td>
</tr>
<tr>
<td>DxCG</td>
<td>DCG RxGroup</td>
<td>Age/Gender, ICD -9 Age/Gender, NDC</td>
</tr>
<tr>
<td>Symmetry/Ingenix</td>
<td>ERG PRG</td>
<td>ICD – 9, NDC NDC</td>
</tr>
<tr>
<td>Johns Hopkins</td>
<td>ACG</td>
<td>Age/Gender, ICD – 9</td>
</tr>
</tbody>
</table>
1. Similar performance among all leading risk groupers*.

2. Risk grouper modeling tools use *different algorithms* to group the source data.

3. Risk groupers use *relatively limited data* sources (e.g. DCG and Rx Group use ICD-9 and NDC codes but not lab results or HRA information)

4. Most Risk Grouper based Predictive Models combine also use statistical analysis.

* See New SOA study (Winkelman et al) published this year. Available from SOA.
Types of Predictive Modeling Tools

PM Tools

Risk Groupers

Statistical Models

Artificial Intelligence
Uses of Statistical Models

Statistical models can be used for all 3 uses

- Medical Management Perspective
- Actuarial, Underwriting and Profiling Perspectives
- Program Evaluation Perspective
What are the different types of statistical models?
Types of Statistical Models

- Logistic Regression
- ANOVA
- Time Series
- Non-linear Regression
- Trees
- Survival Analysis
- Linear Regression
## Multiple Regression Model Example

<table>
<thead>
<tr>
<th>Finding</th>
<th>Hierarchy</th>
<th>Coefficient</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetes</td>
<td>Low cost DM</td>
<td>0</td>
<td>Trumped by Hi cost</td>
</tr>
<tr>
<td>Diabetic nephropathy</td>
<td>Hi cost DM</td>
<td>2.455</td>
<td></td>
</tr>
<tr>
<td>Angina</td>
<td>Low cost CAD</td>
<td>0</td>
<td>Trumped by Hi cost</td>
</tr>
<tr>
<td>Migraines</td>
<td>Med cost headache</td>
<td>0.208</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Subtotal</strong></td>
<td>2.763</td>
</tr>
<tr>
<td>Age-related base</td>
<td></td>
<td>0.306</td>
<td></td>
</tr>
<tr>
<td>Gender-related base</td>
<td></td>
<td>-0.087</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Risk</strong></td>
<td>2.982</td>
</tr>
</tbody>
</table>
Time series modeling tools is another type of statistical modeling tool – it requires a lot of historical data.
Time Series analysis is to

a) Identify the pattern of observed time series data and

b) Forecast future values by extrapolating the identified pattern.
Example: Time Series

4 years of pre-program data

Predicted costs

Actual costs

Savings
1. Statistical models can be used for a number of actuarial applications: evaluation, premium calculation, provider profiling and resource allocation.

2. The predictive model is a critical component of successful medical management intervention programs - “impactability is key in medical management”.

3. Statistical models can use all available detailed data (e.g. lab results or HRA).
Types of Predictive Modeling Tools

- Risk Groupers
- Statistical Models
- Artificial Intelligence
What are the different types of artificial intelligence models?
Artificial Intelligence Models

- Neural Network
- Genetic Algorithms
- Nearest Neighbor Pairings
- Conjugate Gradient
- Rule Induction
- Principal Component Analysis
- Kohonen Network
- Simulated Annealing
- Fuzzy Logic
- Conjugate Gradient
- Simulated Annealing
Features of Neural Networks

Reality

- NN tracks complex relationships by resembling the human brain

Perception

- NN can accurately model complicated health care systems

Reality

- Performance equals standard statistical models
- Models overfit data
1. Good academic approach.
2. Few data limitations.
3. Performance comparable to other approaches.
4. Can be hard to understand the output of neural networks (black box).
In Summary

1. Leading predictive modeling tools have similar performance.

2. Selecting a predictive modeling tool should be based on your specific objectives - one size doesn’t fit all.

3. A good predictive model for medical management should be linked to the intervention (e.g. impactability).

4. “Mixed” models can increase the power of a single model.
PM is **NOT** always about *Cost Prediction*.....

.....it IS about resource allocation.

- Where/how should you allocate resources?
- Who is *intervenable* or *impactable*?
- What can you expect for outcomes?
- How can you manage the key drivers of the economic model for better outcomes?
Remember this chart?

<table>
<thead>
<tr>
<th></th>
<th>0.0% - 0.5%</th>
<th>0.5% - 1.0%</th>
<th>Top 1%</th>
<th>Top 5%</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>67,665</td>
<td>67,665</td>
<td>135,330</td>
<td>676,842</td>
<td>13,537,618</td>
</tr>
<tr>
<td>Actual Cost</td>
<td>$3,204,433,934</td>
<td>$1,419,803,787</td>
<td>$4,624,237,721</td>
<td>$9,680,579,981</td>
<td>$21,973,586,008</td>
</tr>
<tr>
<td>PMPY Total Actual Cost</td>
<td>$47,357</td>
<td>$20,977</td>
<td>$34,170</td>
<td>$14,303</td>
<td>$1,623</td>
</tr>
<tr>
<td>Percentage of Total Cost</td>
<td>14.6%</td>
<td>6.5%</td>
<td>21.1%</td>
<td>44.1%</td>
<td>100%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>0.0% - 0.5%</th>
<th>0.5% - 1.0%</th>
<th>Top 1%</th>
<th>Top 5%</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients with &gt; $50,000 in Claims</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Patients</td>
<td>19,370</td>
<td>5,249</td>
<td>24,619</td>
<td>32,496</td>
<td>35,150</td>
</tr>
<tr>
<td>Percentage of Total</td>
<td>55.1%</td>
<td>14.9%</td>
<td><strong>70.0%</strong></td>
<td>92.4%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>
Decreasing Cost / Decreasing Opportunity

Population Risk Ranking

Event frequency (percent)

Cumulative Total Population

0.2% 0.7% 1.3% 4% 14% 25%
Economic Model: Simple example

- 30,000 eligible members (ee/dep)
- 1,500 – 2,000 with chronic conditions
- 20% “high risk” – 300 to 400
- 60% are reachable and enroll: 180 - 240
- Admissions/high-risk member/year: 0.65
- “Change behavior” of 25% of these:
  - reduced admissions: 29 to 39 annually
  - cost: $8,000/admission
- Gross Savings: $232,000 to $312,000
  - $0.64 to $0.87 pmpm.
Key drivers of the economic model

- Prevalence within the population (numbers)
- Ability to Risk Rank the Population
- Data quality
- Reach/engage ability
- Cost/benefit of interventions
- Timeliness
- Resource productivity
- Random variability in outcomes
Understanding the Economics

DM Program Savings/Costs at different penetration levels

- Gross Savings
- Expenses
- Net Savings

Penetration (%) vs. Savings/Cost ($ millions)

2% 17% 32% 47% 62% 77% 92%

Savings/Cost ($ millions)

$- $1 $2 $3 $4

SOLUCIA, INC.
Modeling
What is a model?

- A model is a set of coefficients to be applied to production data in a live environment.

- With individual data, the result is often a predicted value or “score”. For example, the likelihood that an individual will purchase something, or will experience a high-risk event (surrender; claim, etc.).

- For underwriting, we can predict either cost or risk-score.
Practical Example of Model-Building
Available data for creating the score included the following

• Eligibility/demographics
• Rx claims
• Medical claims

For this project, several data mining techniques were considered: neural net, CHAID decision tree, and regression. The regression was chosen for the following reasons:

With proper data selection and transformation, the regression was very effective, more so than the tree.
1. Split the dataset randomly into halves

Put half of the claimants into an analysis dataset and half into a test dataset. This is to prevent over-fitting. The scoring will be constructed on the analysis dataset and tested on the test dataset. Diagnostic reports are run on each dataset and compared to each other to ensure that the compositions of the datasets are essentially similar. Reports are run on age, sex, cost, as well as disease and Rx markers.

Diagnostics
• In any data-mining project, the output is only as good as the input.
• Most of the time and resources in a data mining project are actually used for variable preparation and evaluation, rather than generation of the actual “recipe”.

2. Build and Transform independent variables
3. Build composite dependent variable

- A key step is the choice of dependent variable. What is the best choice?
- A likely candidate is total patient cost in the predictive period. But total cost has disadvantages:
  - It includes costs such as injury or maternity that are not generally predictable.
  - It includes costs that are steady and predictable, independent of health status (capitated expenses).
  - It may be affected by plan design or contracts.
- We generally predict total cost (allowed charges) net of random costs and capitated expenses.
- Predicted cost can be converted to a risk-factor.
3. Build and transform Independent Variables

The process below is applied to variables from the baseline data.

- Typical transforms include
- Truncating data ranges to minimize the effects of outliers.
- Converting values into binary flag variables.
- Altering the shape of the distribution with a log transform to compare orders of magnitude.
- Smoothing progression of independent variables
3. Build and transform Independent Variables

- A simple way to look at variables
- Convert to a discrete variable. Some variables such as number of prescriptions are already discrete. Real-valued variables, such as cost variables, can be grouped into ranges
- Each value or range should have a significant portion of the patients.
- Values or ranges should have an ascending or descending relationship with average value of the composite dependent variable.

![Graph showing the relationship between % Claimants and Avg of composite dependent variable.](image)
4. Select Independent Variables

- The following variables were most promising
- Age - Truncated at 15 and 80
- Baseline cost
- Number of comorbid condition truncated at 5
- MClass
  - Medical claims-only generalization of the comorbidity variable.
  - Composite variable that counts the number of distinct ICD9 ranges for which the claimant has medical claims.
  - Ranges are defined to separate general disease/condition categories.
- Number of prescriptions truncated at 10
4. Select Independent Variables (contd.)

- Scheduled drug prescriptions truncated at 5
- NClass
  - Rx-only generalization of the comorbidity variable.
  - Composite variable that counts the number of distinct categories distinct ICD9 ranges for which the claimant has claims.
  - Ranges are defined using GPI codes to separate general disease/condition categories.
- Ace inhibitor flag
- Anticoagulants flag
- Diuretics flag
- Number of corticosteroid drug prescriptions truncated at 2
- Neuroleptic drug flag
- Digoxin flag
5. Run Stepwise Linear Regression

An ordinary linear regression is simply a formula for determining a best-possible linear equation describing a dependent variable as a function of the independent variables. But this pre-supposes the selection of a best-possible set of independent variables. How is this best-possible set of independent variables chosen?

One method is a stepwise regression. This is an algorithm that determines both a set of variables and a regression. Variables are selected in order according to their contribution to incremental $R^2$. 
5. Run Stepwise Linear Regression (continued)

Stepwise Algorithm

1. Run a single-variable regression for each independent variable. Select the variable that results in the greatest value of $R^2$. This is “Variable 1”.

2. Run a two-variable regression for each remaining independent variable. In each regression, the other independent variable is Variable 1. Select the remaining variable that results in the greatest incremental value of $R^2$. This is “Variable 2.”

3. Run a three-variable regression for each remaining independent variable. In each regression, the other two independent variables are Variables 1 and 2. Select the remaining variable that results in the greatest incremental value of $R^2$. This is “Variable 3.”

......

n. Stop the process when the incremental value of $R^2$ is below some pre-defined threshold.
6. Results - Examples

- Stepwise linear regressions were run using the "promising" independent variables as inputs and the composite dependent variable as an output.
- Separate regressions were run for each patient sex.
- Sample Regressions

- Female
  - Scheduled drug prescription 358.1
  - NClass 414.5
  - MClass 157.5
  - Baseline cost 0.5
  - Diabetes Dx 1818.9
  - Intercept 18.5

Why are some variables selected while others are omitted? The stepwise algorithm favors variables that are relatively uncorrelated with previously-selected variables. The variables in the selections here are all relatively independent of each other.
6. Results - Examples

- Examples of application of the female model

**Female Regression Regression Formula**

\[(\text{Scheduled Drug} \times 358.1) + (\text{NClass} \times 414.5) + (\text{Cost} \times 0.5) + (\text{Diabetes} \times 1818.9) + (\text{MClass} \times 157.5) - 18.5\]

<table>
<thead>
<tr>
<th>Raw Value</th>
<th>Transformed Value</th>
<th>Predicted Value</th>
<th>Actual Value</th>
<th>Transform Function</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Claimant ID</strong></td>
<td><strong>Schedule Drugs</strong></td>
<td><strong>NClass</strong></td>
<td><strong>Cost</strong></td>
<td><strong>Diabetes</strong></td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>620</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Value Range</th>
<th>Transformed Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RV &lt; 2</td>
<td>1.0</td>
</tr>
<tr>
<td>2 &lt; RV &lt; 5</td>
<td>2.0</td>
</tr>
<tr>
<td>RV &gt; 5</td>
<td>3.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Value Range</th>
<th>Transformed Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RV &lt; 2</td>
<td>0.5</td>
</tr>
<tr>
<td>2 &lt; RV &lt; 7</td>
<td>3.0</td>
</tr>
<tr>
<td>RV &gt; 7</td>
<td>6.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Value Range</th>
<th>Transformed Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RV &lt; 2</td>
<td>2,000</td>
</tr>
<tr>
<td>2 &lt; RV &lt; 10</td>
<td>6,000</td>
</tr>
<tr>
<td>RV &gt; 10</td>
<td>10,000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Value Range</th>
<th>Transformed Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>1.0</td>
</tr>
<tr>
<td>No</td>
<td>0.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Value Range</th>
<th>Transformed Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RV &lt; 1</td>
<td>0.5</td>
</tr>
<tr>
<td>1 &lt; RV &lt; 7</td>
<td>2.0</td>
</tr>
<tr>
<td>RV &gt; 7</td>
<td>3.0</td>
</tr>
</tbody>
</table>
Model Modifications
Expanding and Changing the Model

Expanding definitions
Models for separate populations
Models for varying renewal years
Form of output
Trend
Evaluation
EVALUATION - Testing

Various statistics available for evaluation:
  R-squared
  Mean Absolute Prediction Error
  \[ \frac{(\text{Prediction} - \text{Actual})}{\text{Prediction}} \]

Compare to existing tools
Evaluate results and issues
Selected references

This is not an exhaustive bibliography. It is only a starting point for explorations.

- Shapiro, A.F. and Jain, L.C. (editors); *Intelligent and Other Computational Techniques in Insurance*; World Scientific Publishing Company; 2003.
- Dove, Henry G., Duncan, Ian, and Robb, Arthur; *A Prediction Model for Targeting Low-Cost, High-Risk Members of Managed Care Organizations*; The American Journal of Managed Care, Vol 9 No 5, 2003
- Berry, Michael J. A. and Linoff, Gordon; *Data Mining Techniques for Marketing, Sales and Customer Support*; John Wiley and Sons, Inc; 2004
- Montgomery, Douglas C., Peck, Elizabeth A., and Vining, G Geoffrey; *Introduction to Linear Regression Analysis*; John Wiley and Sons, Inc; 2001
- Kahneman, Daniel, Slovic, Paul, and Tversky (editors); *Judgment under uncertainty: Heuristics and Biases*; Cambridge University Press; 1982
Selected references (contd.)


Further Questions?

iduncan@soluciaconsulting.com

Solucia Inc.
220 Farmington Avenue
Farmington, CT 06032

860-676-8808

www.soluciaconsulting.com