Automating Business Modeling with the AutoModelR package

Derek McCrae Norton

InterContinental Hotels Group
Outline

1. Introduction
   - The Reasoning Behind AutoModelR
   - What is AutoModelR?

2. AutoModelR Specifics
   - Data Exploration and Reduction
   - Modeling / Model Assessment and Selection
Outline

1. Introduction
   - The Reasoning Behind AutoModelR
   - What is AutoModelR?

2. AutoModelR Specifics
   - Data Exploration and Reduction
   - Modeling / Model Assessment and Selection
Why AutoModelR?
... or should I keep my eyes open for the remainder of this talk?

- The modeling process can be lengthy.
- A standardized process can be defined.
- Many steps can be simple in implementation, if not in scope.
- A standardized report can aid in understanding.
Why AutoModelR?

... or should I keep my eyes open for the remainder of this talk?

- The modeling process can be lengthy.
- A standardized process can be defined.
- Many steps can be simple in implementation, if not in scope.
- A standardized report can aid in understanding.
Why AutoModelR?
... or should I keep my eyes open for the remainder of this talk?

- The modeling process can be lengthy.
- A standardized process can be defined.
- Many steps can be simple in implementation, if not in scope.
- A standardized report can aid in understanding.
Why AutoModelR?
... or should I keep my eyes open for the remainder of this talk?

- The modeling process can be lengthy.
- A standardized process can be defined.
- Many steps can be simple in implementation, if not in scope.
- A standardized report can aid in understanding.
Process is standardized in a broad sense.

Process is followed with or without automation.

When 40+ requests are sitting in queue, any savings of time is greatly appreciated.
- Process is standardized in a broad sense.
- Process is followed with or without automation.
- When 40+ requests are sitting in queue, any savings of time is greatly appreciated.
Process is standardized in a broad sense.

Process is followed with or without automation.

When 40+ requests are sitting in queue, any savings of time is greatly appreciated.
Process is standardized in a broad sense.

Process is followed with or without automation.

When 40+ requests are sitting in queue, any savings of time is greatly appreciated.
Outline

1 Introduction
   • The Reasoning Behind AutoModelR
   • What is AutoModelR?

2 AutoModelR Specifics
   • Data Exploration and Reduction
   • Modeling / Model Assesment and Selection
What is AutoModelR?

- Automation of three standardized tasks.
  1. Data Exploration and Reduction
  2. Modeling
  3. Model Assessment and Selection

- Most of the focus is on task 1.
- Tasks 2 and 3 are generally first steps to more involved work.
What is AutoModelR?

- Automation of three standardized tasks.
  1. Data Exploration and Reduction
  2. Modeling
  3. Model Assessment and Selection

  Most of the focus is on task 1.
  Tasks 2 and 3 are generally first steps to more involved work.
What is AutoModelR?

- Automation of three standardized tasks.
  1. Data Exploration and Reduction
  2. Modeling
  3. Model Assessment and Selection
- Most of the focus is on task 1.
- Tasks 2 and 3 are generally first steps to more involved work.
What is AutoModelR?

- Automation of three standardized tasks.
  1. Data Exploration and Reduction
  2. Modeling
  3. Model Assessment and Selection

- Most of the focus is on task 1.
- Tasks 2 and 3 are generally first steps to more involved work.
What is AutoModelR?

- Automation of three standardized tasks.
  1. Data Exploration and Reduction
  2. Modeling
  3. Model Assessment and Selection

- Most of the focus is on task 1.
- Tasks 2 and 3 are generally first steps to more involved work.
What is AutoModelR?

- Automation of three standardized tasks.
  1. Data Exploration and Reduction
  2. Modeling
  3. Model Assessment and Selection
- Most of the focus is on task 1.
- Tasks 2 and 3 are generally first steps to more involved work.
Outline

1. Introduction
   - The Reasoning Behind AutoModelR
   - What is AutoModelR?

2. AutoModelR Specifics
   - Data Exploration and Reduction
   - Modeling / Model Assessment and Selection
General Steps

- Drop variables with a high percentage of missing values.
- Create indicator variables for presence in variables with a medium percentage of missing values.
- Impute variables with a low percentage of missing values.
- Drop variables with zero variation.
- Add log transformed variables.
- Conduct basic filter variable selection based on relevance and redundancy.
General Steps

- Drop variables with a high percentage of missing values.
- Create indicator variables for presence in variables with a medium percentage of missing values.
- Impute variables with a low percentage of missing values.
- Drop variables with zero variation.
- Add log transformed variables.
- Conduct basic filter variable selection based on relevance and redundancy.
General Steps

- Drop variables with a high percentage of missing values.
- Create indicator variables for presence in variables with a medium percentage of missing values.
- Impute variables with a low percentage of missing values.
- Drop variables with zero variation.
- Add log transformed variables.
- Conduct basic filter variable selection based on relevance and redundancy.
General Steps

- Drop variables with a high percentage of missing values.
- Create indicator variables for presence in variables with a medium percentage of missing values.
- Impute variables with a low percentage of missing values.
- Drop variables with zero variation.
- Add log transformed variables.
- Conduct basic filter variable selection based on relevance and redundancy.
General Steps

- Drop variables with a high percentage of missing values.
- Create indicator variables for presence in variables with a medium percentage of missing values.
- Impute variables with a low percentage of missing values.
- Drop variables with zero variation.
- Add log transformed variables.
- Conduct basic filter variable selection based on relevance and redundancy.
General Steps

- Drop variables with a high percentage of missing values.
- Create indicator variables for presence in variables with a medium percentage of missing values.
- Impute variables with a low percentage of missing values.
- Drop variables with zero variation.
- Add log transformed variables.
- Conduct basic filter variable selection based on relevance and redundancy.
General Steps II

- Create summary tables.
- Create plots. (Thanks to Jim Porzak)

### IHG_TOTAL_SHARE

<table>
<thead>
<tr>
<th>MIN</th>
<th>MEDIAN</th>
<th>MAX</th>
<th>MEAN</th>
<th>STD DEV</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0125</td>
<td>0.4286</td>
<td>1</td>
<td>0.4798</td>
<td>0.3008</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>N</th>
<th>UNIQUE</th>
<th>MISSING</th>
<th>SKEWNESS</th>
<th>KURTOSIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>326</td>
<td>0</td>
<td>0.3837</td>
<td>-1.066</td>
</tr>
</tbody>
</table>

### MI_EMAIL_PREF_VAL_CD

<table>
<thead>
<tr>
<th>Label</th>
<th>Frequency</th>
<th>Barchart</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25</td>
<td>31.6</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>18.6</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>15.7</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>15.2</td>
</tr>
<tr>
<td>5</td>
<td>17</td>
<td>4.7</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>4.2</td>
</tr>
<tr>
<td>7</td>
<td>24</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>57</td>
<td>1.6</td>
</tr>
<tr>
<td>9</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>10 ** Others (23 Levels) **</td>
<td>3.4</td>
<td></td>
</tr>
</tbody>
</table>

This Space For Rent
Apply Inside
Outline

1. Introduction
   - The Reasoning Behind AutoModelR
   - What is AutoModelR?

2. AutoModelR Specifics
   - Data Exploration and Reduction
   - Modeling / Model Assessment and Selection
Modeling Goals

- As yet incomplete.
- Create a portfolio of models for particular types of dependent variables
  - Continuous
  - Binary Categorical
  - Multicategory Categorical
Modeling Goals

- As yet incomplete.
- Create a portfolio of models for particular types of dependent variables
  - Continuous
  - Binary Categorical
  - Multicategory Categorical
Modeling Goals

- As yet incomplete.
- Create a portfolio of models for particular types of dependent variables
  - Continuous
  - Binary Categorical
  - Multicategory Categorical
The initial goal of AutoModelR was to automate as much as possible of the Exploratory Data Analysis in the modeling process.

- This is achieved with a Sweave document that is intelligent enough to make choices similar to mine.

- The goal is now to expand this to include automated initial modeling and model comparison.

Outlook

- Continue to add initial models to be fit.
- Consider a rewrite specifically for time series data.
The initial goal of AutoModelR was to automate as much as possible of the Exploratory Data Analysis in the modeling process.

This is achieved with a Sweave document that is intelligent enough to make choices similar to mine.

The goal is now to expand this to include automated initial modeling and model comparison.

Outlook

- Continue to add initial models to be fit.
- Consider a rewrite specifically for time series data.
The initial goal of AutoModelR was to automate as much as possible of the Exploratory Data Analysis in the modeling process.

- This is achieved with a Sweave document that is intelligent enough to make choices similar to mine.

The goal is now to expand this to include automated initial modeling and model comparison.

Outlook
- Continue to add initial models to be fit.
- Consider a rewrite specifically for time series data.
1 Dataset Information

1.1 Independent Variable Assignments
The variables from the data frame have been moved to the following places:
- Numeric Variables: x.num
- Categorical Variables: x.cat
- Dummy Variables Created From Categorical Variables: x.dummy
- Variables With More Than 75% Missing Values: x.hmis
- Variables With Between 40% and 75% Missing Values: x.mmis

1.2 Data Demographics
Initial Data Summary:
- Total Variables: 459
- Numeric Variables: 351
- Categorical Variables: 108
- Total Observations: 1000

Data Summary After Data Cleaning and Variable Selection:
- Numeric Variables: 176
- Categorical Variables: 48
- Dummy Variables: 85

R version 2.7.0 (2008-04-22)
Platform: i386-pc-mingw32

1.2 Categorical Variable Summaries

Dataset Information EDA Report for Data Frame x.fact
1 Independent Variable Information
1.1 Numeric Variable Summaries

<table>
<thead>
<tr>
<th>Variable</th>
<th>MIN</th>
<th>MED</th>
<th>MAX</th>
<th>AVG</th>
<th>STD</th>
<th>OBS</th>
<th>UNIQUE</th>
<th>MISSING</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS_AGE35_44</td>
<td>0</td>
<td>17.5</td>
<td>999</td>
<td>46.8</td>
<td>168</td>
<td>1000</td>
<td>380</td>
<td>0</td>
</tr>
<tr>
<td>CS_FAMILIES</td>
<td>0</td>
<td>76.2</td>
<td>999</td>
<td>101</td>
<td>159</td>
<td>1000</td>
<td>608</td>
<td>0</td>
</tr>
<tr>
<td>CS_HISPANICS</td>
<td>0</td>
<td>2.89</td>
<td>999</td>
<td>36.7</td>
<td>170</td>
<td>1000</td>
<td>404</td>
<td>0</td>
</tr>
<tr>
<td>CS_UNIT10P</td>
<td>0</td>
<td>2</td>
<td>999</td>
<td>38.7</td>
<td>170</td>
<td>1000</td>
<td>437</td>
<td>0</td>
</tr>
<tr>
<td>CS_VAL500</td>
<td>0</td>
<td>12.8</td>
<td>999</td>
<td>52.3</td>
<td>168</td>
<td>1000</td>
<td>586</td>
<td>0</td>
</tr>
<tr>
<td>CS_WHITE</td>
<td>0</td>
<td>91.1</td>
<td>999</td>
<td>112</td>
<td>157</td>
<td>1000</td>
<td>528</td>
<td>0</td>
</tr>
<tr>
<td>MI_AMB_TENURE</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0.025</td>
<td>0.385</td>
<td>1000</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>MS_ENROLL_MONTH</td>
<td>1</td>
<td>7</td>
<td>12</td>
<td>6.61</td>
<td>3.13</td>
<td>1000</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>MS_IN_STAYS_PERC</td>
<td>0</td>
<td>0</td>
<td>0.667</td>
<td>0.000827</td>
<td>0.0215</td>
<td>1000</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>MS_MPT_PTS_EARNED</td>
<td>0</td>
<td>3.23e+03</td>
<td>3.67e+05</td>
<td>1.61e+04</td>
<td>3.56e+04</td>
<td>1000</td>
<td>733</td>
<td>0</td>
</tr>
<tr>
<td>SG_AIR_MEMO</td>
<td>0</td>
<td>0</td>
<td>1.1e+05</td>
<td>300</td>
<td>4.85e+03</td>
<td>1000</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>SG_CROWNE</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td>0.447</td>
<td>2.05</td>
<td>1000</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>SG_CROWNE_RESORT_C</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0.01</td>
<td>0.109</td>
<td>1000</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>SG_DIST0_PCT</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.0306</td>
<td>0.134</td>
<td>1000</td>
<td>45</td>
<td>0</td>
</tr>
<tr>
<td>SG_DOWNTOWN_C</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>0.06</td>
<td>0.582</td>
<td>1000</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>SG_EM_ES_COUNT</td>
<td>0</td>
<td>3</td>
<td>14</td>
<td>3.23</td>
<td>3.27</td>
<td>1000</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>SG_ENEWS_COUNT</td>
<td>0</td>
<td>4</td>
<td>22</td>
<td>6.11</td>
<td>6.76</td>
<td>1000</td>
<td>23</td>
<td>0</td>
</tr>
</tbody>
</table>

1.2 Categorical Variable Summaries

<table>
<thead>
<tr>
<th>Variable</th>
<th>OBS</th>
<th>UNIQUE</th>
<th>MISSING</th>
<th>CAT1</th>
<th>% CAT2</th>
<th># CAT 80%</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS_DIVISION</td>
<td>1000</td>
<td>10</td>
<td>0</td>
<td>20.300</td>
<td>18.500</td>
<td>6</td>
</tr>
<tr>
<td>CS_REGION</td>
<td>1000</td>
<td>5</td>
<td>0</td>
<td>36.500</td>
<td>27.600</td>
<td>3</td>
</tr>
<tr>
<td>MI_CCD_TYP</td>
<td>1000</td>
<td>6</td>
<td>0</td>
<td>98.300</td>
<td>0.800</td>
<td>1</td>
</tr>
<tr>
<td>MI_CURRENT_IND</td>
<td>1000</td>
<td>2</td>
<td>0</td>
<td>81.400</td>
<td>18.600</td>
<td>1</td>
</tr>
<tr>
<td>MI_EMAIL_PREF_VAL_CD</td>
<td>1000</td>
<td>32</td>
<td>0</td>
<td>31.600</td>
<td>18.600</td>
<td>4</td>
</tr>
<tr>
<td>MI_SEC_PHONE_IND</td>
<td>1000</td>
<td>6</td>
<td>0</td>
<td>90.700</td>
<td>7.500</td>
<td>1</td>
</tr>
<tr>
<td>MS_COPRT_AMEX</td>
<td>1000</td>
<td>7</td>
<td>0</td>
<td>99.700</td>
<td>0.200</td>
<td>1</td>
</tr>
<tr>
<td>MS_COPRT_VISA</td>
<td>1000</td>
<td>51</td>
<td>0</td>
<td>91.600</td>
<td>1.900</td>
<td>1</td>
</tr>
<tr>
<td>MS_DOMINANT_BRAND</td>
<td>1000</td>
<td>23</td>
<td>0</td>
<td>46.800</td>
<td>35.700</td>
<td>2</td>
</tr>
<tr>
<td>MS_ENRL_BRAND_CD</td>
<td>1000</td>
<td>18</td>
<td>0</td>
<td>65.200</td>
<td>17.600</td>
<td>2</td>
</tr>
<tr>
<td>MS_ENRL_CHAIN_CD</td>
<td>1000</td>
<td>7</td>
<td>0</td>
<td>65.200</td>
<td>20.100</td>
<td>2</td>
</tr>
<tr>
<td>MS_ENRL_MKT_RGN_DESC</td>
<td>1000</td>
<td>6</td>
<td>0</td>
<td>65.200</td>
<td>34.400</td>
<td>2</td>
</tr>
<tr>
<td>MS_ENRL_SUB_MKT_RGN_~1</td>
<td>1000</td>
<td>16</td>
<td>0</td>
<td>65.200</td>
<td>34.400</td>
<td>2</td>
</tr>
<tr>
<td>MS_ENRL_SUB_MKT_RGN_~2</td>
<td>1000</td>
<td>16</td>
<td>0</td>
<td>65.200</td>
<td>34.400</td>
<td>2</td>
</tr>
<tr>
<td>MS_EOFFERS</td>
<td>1000</td>
<td>2</td>
<td>0</td>
<td>63.300</td>
<td>36.700</td>
<td>2</td>
</tr>
<tr>
<td>MS_ESOURCE</td>
<td>1000</td>
<td>5</td>
<td>0</td>
<td>48.800</td>
<td>34.800</td>
<td>2</td>
</tr>
</tbody>
</table>