A Comparison of Feature Generation Techniques for Credit Risk Modelling

Dan Kellett
Director of Data Science
Feature Generation is a core part of modelling but is often overlooked for more ‘glamorous’ parts of the project.
A move to tradeline bureau data dramatically expanded the potential for new feature creation

<table>
<thead>
<tr>
<th>ACCT_ID</th>
<th>Ave balance</th>
<th>Worst status</th>
<th># Searches</th>
</tr>
</thead>
<tbody>
<tr>
<td>12345</td>
<td>500</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>38410</td>
<td>120</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>86500</td>
<td>1750</td>
<td>8</td>
<td>1</td>
</tr>
</tbody>
</table>

1 row per customer

1 row per customer, per bureau record
The expert driven approach uses innovation techniques to access an organisation’s bank of knowledge.
The brute force approach is about rapid progression through a feature extraction funnel

- Mechanistically created 1.4 million new variables from bureau data for risk prediction
- Just 15 hours processing time reduced this to 12 powerful new splitters
- The app-stage risk model contains 4 of these covariates as some of the top variables
The AI approach uses Deep Learning to build pictures of ‘typical’ risk profiles for direct comparison.

- Bureau-gram of low risk individual
- Bureau-gram of high risk individual