Validating and Understanding a Highly Nonlinear Machine Learning Model

Douglas W. Dwyer, Managing Director | Moody’s Analytics
Agenda

1. Prologue *Winning the Indy 500*
2. Modeling Approaches and Validation Principles
3. Why this approach will work
4. Implementation
5. Looking at Outliers
6. Combining Both Approaches
7. Model makes a million dollar difference
8. Conclusion
1 Prologue: Winning the Indy 500
Winning in Indy
Requires both carburetion and the right non-linear surface
Validating a Highly Nonlinear Model

Winning the Indy 500

Ever since the automobile was invented, people have been racing.

» Different types of tracks
» Different conditions

A mix of air and fuel is injected into the engine; optimizing mixture is key.

In 2018, Will Power and Team Penske had a winning time of just under 3 hours, less than 3 seconds ahead of the second place finisher.

Design innovations can be the difference between winning and losing, but such innovations require extensive testing prior to race day.
Bankers have always rated their borrowers

A bank that only lends to the safest borrowers will not have much Pre-Provision Net Revenue (PPNR).

A bank that lends to everyone may find that provisions exceed PPNR.

Being competitive requires differentiating between the good and the bad borrowers.

A modeling innovation may make a difference, but it must be carefully tested before “race time.”
Modeling Approaches
Validating a Highly Nonlinear Model

Modeling approaches

Let data speak \rightarrow\text{Impose some structure on data} \rightarrow\text{More structure imposed on data}

Unsupervised Learning
Natural Language Processing
Kernel Methods

Deep Neural Networks
Generalized Additive Models
Support Vector Machines
Nearest Neighbor
Random Forests
Boosting

Traditional Nonlinear Model
New Highly Nonlinear Model

Option pricing theory
Hierarchical Bayesian models
Bias correction methods
Structural Equation Modeling
Structural Estimation
Core validation principals

» Evaluation of conceptual soundness
» Ongoing monitoring
» Outcome analyses
» Effective Challenge
Recipe for validating a highly nonlinear PD model

A data table with these columns:
1) Default flag
2) Traditional non-linear PD
3) Highly non-linear PD
4) Set of “risk drivers”
5) Set of “cut variables”

And hundreds of thousands of observations (rows)

By Moody’s Analytics

- Data Preparation time: 20 years
- Cook time: 20 minutes

Procedure: Run a Model Comparison Report to:
1) Demonstrate that “highly non-linear model” has better discriminatory power
2) Compare the level of the PDs to each other and to actual
3) Examine cases where two models produce very different results.
4) Look at “higher order effects” and validate “final model”
Recipe for validating a highly nonlinear PD model

Ingredients

A data table with these columns:

1) A default flag
2) A PD from a “traditional non-linear model”
3) A PD from the “highly non-linear model”
4) A set of “risk drivers”
5) A set of “cut variables”

And hundreds of thousands of observations

ML PD model

Recipe for validating a highly nonlinear PD model

By Moody's Analytics

Data Preparation time: 20 years
Cook time: 20 minutes

Ingredients

A data table with these columns:

1) Default flag
2) Traditional non-linear PD
3) Highly non-linear PD
4) Set of “risk drivers”
5) Set of “cut variables”

And hundreds of thousands of observations (rows)

To garnish

Labels and Formats

Method

Procedure: Run a Model Comparison Report to:

1) Demonstrate that “highly non-linear model” has better discriminatory power
2) Compare the level of the PDs to each other and to actual
3) Examine cases where two models produce very different results.
4) Look at “higher order effects” and validate “final model”
Recipe for validating a highly nonlinear PD model

Method: Run a Model Comparison Report to:

1) Demonstrate that “highly nonlinear model” has better discriminatory power than the “traditional nonlinear model.”

2) Compare the level of the PDs to each other and to actual.

3) Examine cases where two models produce very different results.

4) Look at “higher order effects” and validate “final model.”

ML PD model

By Moody’s Analytics

Data Preparation time: 20 years
Cook time: 20 minutes

A data table with these columns:
- Default flag
- Traditional non-linear PD
- Highly non-linear PD
- Set of “risk drivers”
- Set of “cut variables”

And hundreds of thousands of observations (rows)

Ingredients

To garnish

Labels and Formats

Method

Procedure: Run a Model Comparison Report to:

1) Demonstrate that “highly non-linear model” has better discriminatory power

2) Compare the level of the PDs to each other and to actual

3) Examine cases where two models produce very different results.

4) Look at “higher order effects” and validate “final model”
Why this Approach Works
Why this approach works

» We have an objective and concrete outcome variable -- Default
» Well-established set of risk drivers
» Large “panel” dataset
» “Credit Cycle” is treated separately
A “Traditional Nonlinear Model”

Graphic displays interrelationship between three key risk drivers and the RiskCalc EDF as a staircase.

Upstairs is better liquidity, backward is better profit, and leftward is more sales growth. Red steps represent highest risk. Highest risk is where both liquidity and profitability and sales growth are in the lowest quartile, but low profit, low liquidity and rapid sales growth is risky as well.
### Highly nonlinear models: decision tree-based ensemble learners

**Random Forest:**

- Bootstrap aggregation of classification trees, where each tree casts a unit vote for the most popular class.
- Select $n_{\text{tree}}$, the number of trees to grow, and $m_{\text{try}}$, number of variables.
- For $i = 1$ to $n_{\text{tree}}$: Draw a bootstrap sample from the data. Grow a "random" tree, where, at each node, the best split is chosen from among $m_{\text{try}}$ randomly selected variables.
- Prediction of test data is done by majority vote from predictions from the ensemble of trees or by averaging the predictions from individual trees.

**Boosting:**

- The classification trees here are trained in a sequence. Each subsequent tree focuses on those cases incorrectly classified in the last round by giving more weight to those cases while constructing the next tree.
- Combine the classifiers by taking a weighted average of their outputs.
Machine Learning – without traditional transformation vs. traditional approach

Raw ratios + sector
Ratios imputed by median

• Boosting AR = 51%
• Probit AR = 38%*

Raw ratios + sector
Ratios imputed by median and windsorized (1% each)

• Boosting AR = 59.7%
• Probit AR = 48.9%

Percentiles of ratios + sector

• Boosting AR = 59.7%
• Probit AR = 53.5%

Full Transform

• Boosting AR = 59.7%
• GAM AR = 57.7%
Issues to watch

A linear model that includes “year-fixed effects” may have fit the development data better than a linear model without year-fixed effects, but not on future data. The same issue may be hidden within an ML model.

ML fits both economic relationships and data imperfections.

Important to cross-validate at the firm level.
4 Validating a Highly Nonlinear Model
Validating a highly nonlinear model

Start with the U.S. Credit Research Database:

1,234,854 observations of 250,141 firms with 17,288 defaults from 1994-2018

Acquire a “Traditional Nonlinear PD” – we use the RC4.0 FSO model structure re-estimated on this data.

Acquire a “Highly Nonlinear PD” – we use a Boosting Model estimated on the same data.

Risk drivers are the financial ratios and sector dummies used in RC4.0.

Cut variables are: year, size, and sector.
Validating a Highly Nonlinear Model

Provides back
Validating a Highly Nonlinear Model

Provides back
Validating a Highly Nonlinear Model

Provides back

Summary Statistics

Performance Stats “by Cut”

Cumulative Accuracy Profiles
Validating a Highly Nonlinear Model

Provides back

Calibration curves for each model, risk driver, and cut variable

Summary Statistics

Performance Stats “by Cut”

Cumulative Accuracy Profiles
Validating a Highly Nonlinear Model

Provides back

Calibration curves for each model, risk driver, and cut variable

Summary Statistics

Performance Stats “by Cut”

Cumulative Accuracy Profiles

Scatter Plots
Findings
Validating a Highly Nonlinear Model

Boosting model delivers better CAP plot

Figure 2: Cap Plot

- Alt. PD
- Bench PD
- Perfect Model

AR Alternative: 59.72%
AR Benchmark: 57.7%
Obs: 1,234,654
Defaults: 17,288
Performance differences are robust across cuts

<table>
<thead>
<tr>
<th>Grouping</th>
<th>Number of Obs</th>
<th>Number of Defaults</th>
<th>AR Alt. Model</th>
<th>AR Bench. Model</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business Products</td>
<td>134,598</td>
<td>1,977</td>
<td>59.70</td>
<td>59.07</td>
<td>0.64</td>
</tr>
<tr>
<td>Business Services</td>
<td>160,678</td>
<td>1,891</td>
<td>55.45</td>
<td>53.39</td>
<td>2.05</td>
</tr>
<tr>
<td>Consumer Products</td>
<td>62,014</td>
<td>1,011</td>
<td>60.34</td>
<td>59.10</td>
<td>1.24</td>
</tr>
<tr>
<td>Health Care</td>
<td>111,315</td>
<td>918</td>
<td>52.28</td>
<td>48.25</td>
<td>4.03</td>
</tr>
<tr>
<td>Agriculture</td>
<td>27,269</td>
<td>443</td>
<td>50.59</td>
<td>48.83</td>
<td>1.75</td>
</tr>
<tr>
<td>Communication</td>
<td>15,930</td>
<td>269</td>
<td>55.12</td>
<td>51.72</td>
<td>3.40</td>
</tr>
<tr>
<td>Construction</td>
<td>156,904</td>
<td>3,416</td>
<td>62.62</td>
<td>60.45</td>
<td>2.17</td>
</tr>
<tr>
<td>HiTech</td>
<td>35,620</td>
<td>398</td>
<td>63.08</td>
<td>60.37</td>
<td>2.72</td>
</tr>
<tr>
<td>Mining</td>
<td>25,455</td>
<td>549</td>
<td>60.22</td>
<td>54.63</td>
<td>5.58</td>
</tr>
<tr>
<td>Services</td>
<td>148,209</td>
<td>1,900</td>
<td>57.79</td>
<td>55.77</td>
<td>2.02</td>
</tr>
<tr>
<td>Trade</td>
<td>272,276</td>
<td>3,498</td>
<td>58.85</td>
<td>57.39</td>
<td>1.46</td>
</tr>
<tr>
<td>Transportation</td>
<td>49,040</td>
<td>623</td>
<td>53.67</td>
<td>50.44</td>
<td>3.23</td>
</tr>
<tr>
<td>Unassigned</td>
<td>21,184</td>
<td>325</td>
<td>58.96</td>
<td>52.55</td>
<td>6.41</td>
</tr>
<tr>
<td>Utilities</td>
<td>14,359</td>
<td>70</td>
<td>63.93</td>
<td>54.98</td>
<td>8.95</td>
</tr>
</tbody>
</table>
Performance differences are robust across cuts

**Groupings Based on Financial Statement Year**

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of Observations</th>
<th>Number of Defaults</th>
<th>AR Alt. Model</th>
<th>AR Bench. Model</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1994</td>
<td>3,803</td>
<td>28</td>
<td>62.06</td>
<td>52.87</td>
<td>9.19</td>
</tr>
<tr>
<td>1995</td>
<td>6,618</td>
<td>51</td>
<td>68.70</td>
<td>49.52</td>
<td>19.18</td>
</tr>
<tr>
<td>1996</td>
<td>13,680</td>
<td>50</td>
<td>63.59</td>
<td>53.47</td>
<td>10.12</td>
</tr>
<tr>
<td>1997</td>
<td>19,467</td>
<td>60</td>
<td>63.51</td>
<td>57.16</td>
<td>6.35</td>
</tr>
<tr>
<td>1998</td>
<td>33,732</td>
<td>209</td>
<td>49.74</td>
<td>47.22</td>
<td>2.53</td>
</tr>
<tr>
<td>1999</td>
<td>39,967</td>
<td>482</td>
<td>54.28</td>
<td>51.90</td>
<td>2.38</td>
</tr>
<tr>
<td>2000</td>
<td>44,935</td>
<td>505</td>
<td>58.43</td>
<td>55.34</td>
<td>3.09</td>
</tr>
<tr>
<td>2001</td>
<td>48,977</td>
<td>599</td>
<td>55.94</td>
<td>53.41</td>
<td>2.54</td>
</tr>
<tr>
<td>2002</td>
<td>54,152</td>
<td>561</td>
<td>63.09</td>
<td>60.66</td>
<td>2.43</td>
</tr>
<tr>
<td>2003</td>
<td>62,600</td>
<td>610</td>
<td>59.64</td>
<td>59.50</td>
<td>0.14</td>
</tr>
<tr>
<td>2004</td>
<td>68,518</td>
<td>685</td>
<td>55.55</td>
<td>54.53</td>
<td>1.02</td>
</tr>
<tr>
<td>2005</td>
<td>72,504</td>
<td>666</td>
<td>56.52</td>
<td>54.35</td>
<td>2.17</td>
</tr>
<tr>
<td>2006</td>
<td>72,680</td>
<td>1,146</td>
<td>60.54</td>
<td>58.22</td>
<td>2.32</td>
</tr>
<tr>
<td>2007</td>
<td>75,224</td>
<td>2,064</td>
<td>60.74</td>
<td>59.50</td>
<td>1.25</td>
</tr>
<tr>
<td>2008</td>
<td>77,579</td>
<td>2,528</td>
<td>54.48</td>
<td>53.55</td>
<td>0.92</td>
</tr>
<tr>
<td>2009</td>
<td>76,396</td>
<td>1,808</td>
<td>56.40</td>
<td>55.04</td>
<td>1.36</td>
</tr>
<tr>
<td>2010</td>
<td>75,136</td>
<td>1,285</td>
<td>58.89</td>
<td>56.12</td>
<td>2.77</td>
</tr>
<tr>
<td>2011</td>
<td>72,417</td>
<td>927</td>
<td>59.70</td>
<td>58.05</td>
<td>1.65</td>
</tr>
<tr>
<td>2012</td>
<td>63,746</td>
<td>610</td>
<td>61.11</td>
<td>59.41</td>
<td>1.70</td>
</tr>
<tr>
<td>2013</td>
<td>70,488</td>
<td>637</td>
<td>63.47</td>
<td>59.62</td>
<td>3.85</td>
</tr>
<tr>
<td>2014</td>
<td>67,027</td>
<td>722</td>
<td>64.77</td>
<td>61.41</td>
<td>3.36</td>
</tr>
<tr>
<td>2015</td>
<td>59,662</td>
<td>734</td>
<td>65.73</td>
<td>63.44</td>
<td>2.29</td>
</tr>
<tr>
<td>2016</td>
<td>47,066</td>
<td>309</td>
<td>66.36</td>
<td>63.37</td>
<td>2.99</td>
</tr>
<tr>
<td>2017</td>
<td>8,477</td>
<td>12</td>
<td>67.62</td>
<td>55.45</td>
<td>12.17</td>
</tr>
</tbody>
</table>
Performance differences are robust across cuts

Groupings Based on Firm Size (Total Assets)

<table>
<thead>
<tr>
<th>Size Range</th>
<th>Number of Obs.</th>
<th>Number of Defaults</th>
<th>AR Alt. Model</th>
<th>AR Bench. Model</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;500k</td>
<td>77,514</td>
<td>1,511</td>
<td>44.40</td>
<td>43.74</td>
<td>0.66</td>
</tr>
<tr>
<td>500k to 1MM</td>
<td>53,092</td>
<td>1,107</td>
<td>53.95</td>
<td>52.44</td>
<td>1.51</td>
</tr>
<tr>
<td>1MM to 2.5MM</td>
<td>133,801</td>
<td>2,462</td>
<td>54.95</td>
<td>53.78</td>
<td>1.16</td>
</tr>
<tr>
<td>2.5MM to 5MM</td>
<td>153,407</td>
<td>2,300</td>
<td>59.48</td>
<td>57.89</td>
<td>1.59</td>
</tr>
<tr>
<td>5MM to 10MM</td>
<td>180,521</td>
<td>2,342</td>
<td>61.08</td>
<td>59.05</td>
<td>2.03</td>
</tr>
<tr>
<td>10MM to 20MM</td>
<td>172,033</td>
<td>2,247</td>
<td>59.85</td>
<td>58.55</td>
<td>1.31</td>
</tr>
<tr>
<td>20MM to 40MM</td>
<td>131,143</td>
<td>1,629</td>
<td>59.20</td>
<td>56.96</td>
<td>2.24</td>
</tr>
<tr>
<td>40MM to 80MM</td>
<td>92,164</td>
<td>1,090</td>
<td>61.72</td>
<td>59.23</td>
<td>2.49</td>
</tr>
<tr>
<td>80MM to 160MM</td>
<td>64,154</td>
<td>779</td>
<td>61.49</td>
<td>59.59</td>
<td>1.91</td>
</tr>
<tr>
<td>160MM to 320MM</td>
<td>43,923</td>
<td>587</td>
<td>60.37</td>
<td>59.19</td>
<td>1.18</td>
</tr>
<tr>
<td>&gt;320MM</td>
<td>133,099</td>
<td>1,234</td>
<td>65.98</td>
<td>62.48</td>
<td>3.49</td>
</tr>
</tbody>
</table>
Highly nonlinear model fits industry differences better

Traditional Nonlinear Model underweights Current Liabilities to Sales in Construction
New approach fits nonlinear relationship between EBITDA to interest expense and default better.
Look at Outliers
Inspect cases where PDs are very different
For example…

- Black box nature of the models – lack of interpretability
- Hard to anticipate model behavior outside of the data range

Case 1: A non-defaulted company in the Trade industry

- EBITDA to Interest Expense: 4X
- Return on Assets: -17%
- Cash to Assets: 1%
- Debt to (Debt plus Equity): 77%
- Retained Earnings to Current Liability: -6X
- Total Assets: $500,000

- Boosting PD = 0.2% (A3)
- RiskCalc PD = 8.9% (Caa/C)

Case 2: A non-defaulted company in the Business Services industry

- EBITDA to Interest Expense: 40X
- Return on Assets: 212%
- Cash to Assets: 10.5%
- Debt to Debt plus Equity: 89%
- Retained Earnings to Current Liability: 357X
- Total Assets: $5.8MM

- Boosting PD = 13.7% (Caa/C)
- RiskCalc PD = 0.54% (Baa3)
Combining the Approaches
How to use a really good signal, except when it is bad

Suppose we have two measures of the same signal – e.g. the PD.

One signal is known to be much better most of the time, but it is sometimes faulty.
Bayesian approach lowers the weight on the second signal, as the gap between the two increases.
Combination improves power a bit

Highly Nonlinear Model vs. Traditional

Combination vs. Traditional

MOODY'S ANALYTICS
Validating a Highly Nonlinear Model
Combination tightens the differences between the two

Highly Nonlinear Model vs. Traditional

Combination vs. Traditional
Million Dollar Difference
Model difference is worth a million dollars

Assumptions

$10 billion portfolio

Loosening Credit Standards Causes

» More loans to good borrowers (PPNR* Increases)
» More loans to bad borrowers (Provisions Increases)

Net Income (NI) is the PPNR less Provisions

Central Default Tendency (CDT) of population is 2%

A bank that lends to everyone has no income:

» Spread = Central Default Tendency * LGD / (1 - Central Default Tendency)

*PPNR is Pre Provisions Net Revenue
Tighter lending standards result in low provisions, but low PPNR
Loose lending standards result in large PPNR, but large provisions as well.

**Impact of Lending Standards**
- Horizontal axis: Good Borrowers Lent to (%)
- Vertical axis: Defaulters Lent to (%)

**Impact on Revenue**
- Horizontal axis: Good Borrowers Lent to (%)
- Vertical axis: $mm

**Marginal Cost and Benefit**
- Horizontal axis: Good Borrowers Lent to (%)
- Vertical axis: Marginal Benefits and Costs ($mm)

**Graphs:**
- PPNR
- Provisions
- NI

**Legend:**
- Marginal Cost
- Marginal Benefit
Setting marginal cost to marginal benefit optimizes net income.
New model yields better cost curves worth $1.5mm

More accurate new model produces $1.454mm by turning away 2.8% more of the bad borrowers and only 1.4% more of the good borrowers.
New model yields better cost curves worth $1mm

More accurate, new model produces $1.454mm by turning away 2.8% more of “bad” borrowers and only 1.4% more of “good” borrowers.
Conclusion
Conclusion

» In validating a highly nonlinear model, a \textit{traditional nonlinear model} provides a useful reference.

» Using known risk drivers to predict default helps achieve a transparent model.

» Can use a weighted average of the two models to achieve even better results.

» Care must be taken to evaluate the performance across multiple dimensions.

» Modern software and hardware makes the process easier.

» Small differences in AR can be worth a lot of money.

» Do not need to sacrifice transparency.
Extra Credit: interaction assessment for each pair of risk drivers

Risk Driver 1

Risk Driver 2

Risk Driver 3

Risk Driver 4
Appendix: What is PPNR?

The term is found in the CCAR methodology documentation.

PPNR is net interest income plus noninterest income less noninterest expense.

and

Pretax Net Income is approximately PPNR less Provisions.

Source: *Dodd-Frank Act Stress Test 2016: Supervisory Stress Test Methodology and Results*