Boosting Credit Risk Models

Lessons learnt from more than two decades of research and consulting

Prof. Dr. Bart Baesens

Bart.Baesens@kuleuven.be
Prof. Dr. Bart Baesens

- Studied at KU Leuven (Belgium)
  - Business Engineer in Management Informatics, 1998
  - Ph.D. in Applied Economic Sciences, 2003
- Ph.D. title: Developing Intelligent Systems for Credit Scoring Using Machine Learning Techniques
- Lecturer at School of Management at University of Southampton (UK)
- Research: big data & analytics, credit risk, fraud analytics
- www.bluecourses.com
- www.bartbaesens.com
Books
BlueCourses

- **Basic Credit Risk Modeling for Basel/IFRS 9 using**
  In this course, students learn how to develop credit risk models in the context of the Basel and IFRS 9 guidelines.

- **Advanced Credit Risk Modeling for Basel/IFRS 9**
  In this course, students learn how to do advanced credit risk modeling.

- **Machine Learning Essentials**
  In this course, participants learn the essentials of machine learning.

- **Fraud Analytics**
  In this course, participants learn the essentials of fraud analytics.

- **Social Network Analytics**
  In this course, participants learn the essentials of social network analytics.

- **Recommender Systems**
  In this course, you will learn the essentials of recommender systems.

- **Customer Lifetime Value Modeling**
  In this course, you learn the essentials of Customer Lifetime Value Modeling.

- **Text Analytics**
  In this course, you learn the essentials of text analytics.

www.bluecourses.com
Overview

• Credit Risk Model Architecture
• Credit Risk Model Requirements
• Boosting Credit Risk Models
  – Alternative Data Sources
  – Feature Engineering
  – Deep Learning
  – Profit Driven Modeling
• Model Risk
Credit Risk Model Architecture

Level 0
- Internal Data
- External Data
- Expert Judgment

Level 1
- Model
- Scorecard
- Logistic Regression

Level 2
- Risk ratings definition
- PD calibration

Credit Risk Model Requirements

• Statistical performance

• Profitability

• Interpretability
• Operational efficiency
• Privacy and Ethics
• Ecological footprint
Boosting Credit Risk Models

- Alternative Data Sources
- Feature Engineering
- Deep Learning
- Profit Driven Modeling
Alternative Data Sources

• Call Detail Record (CDR) Data
• Google Street View
• Google Trends
• Social Media Data
• API/Web Scraping
• Open data
• Note: privacy/ethics always respected!
Alternative Data Sources: CDR Data


• Use Call Detail Record (CDR) data for credit scoring
• People tend to call those in their economic circle
• Inspect default influence for credit scoring
• Featurize CDR data and combine with other data
Alternative Data Sources: CDR Data

• Combining CDR with traditional data in credit scoring significantly increases AUC
• In terms of profitability, best model built with only calling behavior features
• Calling behavior features predictive both in terms of statistical and economic performance
• Closing thoughts
  – financial inclusion in developed countries: people joining financial market for first time
  – financial inclusion in developing countries: no historical financial data
  – privacy: check local privacy regulations and ethical guidelines
Alternative Data Sources: Google Street View

- Gebru et al., Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States, *PNAS*, 2017 (https://www.pnas.org/content/114/50/13108)
- Portray demographic and economic makeup of neighborhoods using Google Street View
- Socioeconomic attributes such as income, race, education, and voting patterns can be inferred from cars detected in Google Street View images using deep learning
- More foreign cars in neighbourhood, implies higher average income
  - German and Japanese cars (Lexus in particular) found in areas with high median household income
  - American cars made (Buicks, Oldsmobiles and Dodges) associated more with lower median incomes
- Add deep learning features to credit scoring models
- Privacy (e.g., GDPR)
Alternative Data Sources: Google Trends

• Provides search data and popular search terms across time, geography, etc.
• Can be used for nowcasting where aim is to forecast present or near future
• Allows to spot trends more quickly than traditional channels
• E.g., forecasting unemployment based upon Google searches with key terms jobs, unemployment (benefits), social security, etc.
• Useful for
  – PD/LGD/EAD calibration (e.g., PIT calibration in IFRS 9 setting)
  – anticipating economic downturns
  – stress testing
Feature Engineering

• Transforming data set variables into features to help credit risk modes achieve better performance in terms of
  – predictive power
  – interpretability

• Simple example
  – date of birth ➔ age

• Manual feature engineering versus automated feature engineering

• Best way to improve performance of credit risk models is by designing smart features!
Feature Engineering

Two classes in coordinate system

See https://www.kdnuggets.com/2018/12/feature-engineering-explained.html
Feature Engineering

See https://www.kdnuggets.com/2018/12/feature-engineering-explained.html
Feature Engineering

• Trend features
• Ratio features
• Transformations
  – Logarithmic transformation
  – Power transformation
  – Box-Cox transformation
  – Yeo-Johnson transformation
  – Principal Component Analysis
  – Percentile coding
  – Thermometer coding
Trend Features

• Trends summarize historical evolution

• Absolute trends: \( \frac{F_t - F_{t-x}}{x} \)

• Relative trends: \( \frac{F_t - F_{t-x}}{F_{t-x}} \)

• Useful for size variables (e.g., asset size, loan amounts) and ratios

• Beware with denominators equal to 0!

• Can put higher weight on recent values

• Extension: time series analysis!

Yeo Johnson Transformation

• Yeo Johnson (Yeo and Johnson, 2000)

\[ x \rightarrow f(x; \lambda) = \begin{cases} 
\frac{((1 + x)^{\lambda} - 1)/\lambda}{\log(x + 1)} & \text{for } \lambda \neq 0, x \geq 0 \\
\frac{-(1 - x)^{2 - \lambda} - 1}{2 - \lambda} & \text{for } \lambda \neq 2, x < 0 \\
-\log(-x + 1) & \text{for } \lambda = 2, x < 0 
\end{cases} \]

• Set \( \lambda \) by
  – experimentation
  – visual inspection
  – performance optimization (AUC, profit)
Yeo Johnson Transformation

Deep Learning


• Motivation
  – performance of classification algorithms for credit scoring extensively researched
  – deep learning little attention within credit scoring community
  – deep learning algorithms compared to performance of 2 ensemble methods (random forests, XGBoost) and 2 conventional methods (logistic regression, decision trees)
Deep Learning

<table>
<thead>
<tr>
<th>Data set</th>
<th>Cases</th>
<th>Covariates</th>
<th>Prior default rate</th>
<th>Nx2 cross-validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC</td>
<td>690</td>
<td>14</td>
<td>.445</td>
<td>10</td>
</tr>
<tr>
<td>GC</td>
<td>1,000</td>
<td>20</td>
<td>.300</td>
<td>10</td>
</tr>
<tr>
<td>Th02</td>
<td>1,225</td>
<td>14</td>
<td>.264</td>
<td>10</td>
</tr>
<tr>
<td>Bene1</td>
<td>3,123</td>
<td>28</td>
<td>.667</td>
<td>10</td>
</tr>
<tr>
<td>HMEQ</td>
<td>5,960</td>
<td>12</td>
<td>.199</td>
<td>5</td>
</tr>
<tr>
<td>Bene2</td>
<td>7,190</td>
<td>27</td>
<td>.300</td>
<td>5</td>
</tr>
<tr>
<td>UK</td>
<td>30,000</td>
<td>15</td>
<td>.040</td>
<td>5</td>
</tr>
<tr>
<td>TC</td>
<td>30,000</td>
<td>23</td>
<td>.221</td>
<td>5</td>
</tr>
<tr>
<td>Bene3</td>
<td>38,649</td>
<td>11</td>
<td>.002</td>
<td>5</td>
</tr>
<tr>
<td>GMC</td>
<td>150,000</td>
<td>10</td>
<td>.067</td>
<td>5</td>
</tr>
</tbody>
</table>
Deep Learning

- 2 ensemble methods considered overall best performing classifiers
  - XGBoost is best performing classifier on all performance measures considered except one where random forests is best
- Deep networks do not outperform shallower networks with 1 hidden layer
- Deep learning algorithms not appropriate for credit scoring and 2 ensemble methods considered should be preferred
Neural Networks in Credit Risk

- Not as final credit risk model because of black box property
- Benchmarking
- Detecting non-linear or interaction effects
- Use features/score from deep learning neural network as input to logistic regression model
- Combine neural network with linear/logistic regression in two stage model
- Start from property address, retrieve Google Street/Satellite view image, estimate value using deep learning (e.g. convolutional neural network) and relate to default (PD) or loss (LGD) risk
- Privacy + Ethics!
Profit Driven Modeling


- Key ideas
  - profit-based classification performance measure for credit scoring (as opposed to AUC, Brier score, etc.)
  - enables selection of most profitable credit scoring model and provides optimal cutoff point
  - results indicate that this outperforms traditional approaches in terms of accuracy and monetary value
Profit Driven Modeling

• EMP package
  – https://cran.r-project.org/web/packages/EMP/index.html

• ProfLogit
  – use EMP to estimate coefficients of logistic regression model
  – Stripling E., vanden Broucke S., Antonio K., Baesens B., Snoeck M., Profit
    Maximizing Logistic Model for Customer Churn Prediction Using Genetic

• ProfTree
  – use EMP to estimate decision tree
  – Höppner S., Stripling E., Baesens B., vanden Broucke S., Verdonck T., Profit Driven
Model Risk

- FICO (2021)
  - 65% of companies cannot explain how specific AI model decisions or predictions are made
  - 73% have struggled to get executive support for prioritizing AI ethics
  - Only 20% actively monitor their models in production
  - 30% of organizations report an increase in adversarial and other attacks against their model
Model Risk

• “Model risk is the risk of expected or unexpected loss resulting from the inadequate development or usage of analytical models across all business units and activities of the company.”
Model Risk

- Chapter 9: Managerial Risk
- Chapter 8: Security Risk
- Chapter 7: Operational Risk
- Chapter 6: Validation Risk
- Chapter 3: Data Risk
- Chapter 4: Specification Risk
- Chapter 5: Development Risk

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Deployment
- Evaluation
Conclusions

• Credit Risk Model Architecture
• Credit Risk Model Requirements
• Boosting Credit Risk Models
  – Alternative Data Sources
  – Feature Engineering
  – Deep Learning
  – Profit Driven Modeling
• Model Risk
BlueCourses

Basic Credit Risk Modeling for Basel/IFRS 9 using
In this course, students learn how to develop credit risk models in the context of the Basel and IFRS 9 guidelines

Advanced Credit Risk Modeling for Basel/IFRS 9
In this course, students learn how to do advanced credit risk modeling.

Machine Learning Essentials
In this course, participants learn the essentials of machine learning.

Fraud Analytics
In this course, participants learn the essentials of fraud analytics.

Social Network Analytics
In this course, participants learn the essentials of social network analytics.

Recommender Systems
In this course, you will learn the essentials of recommender systems.

Customer Lifetime Value Modeling
In this course, you learn the essentials of Customer Lifetime Value Modeling.

Text Analytics
In this course, you learn the essentials of text analytics.

www.bluecourses.com