Big Data, Alternative Data, and the Assessment of Credit Risk in the Consumer Lending Industry

Keith Shields, CAO, Magnify
Susan Arnot, Director – Decision Sciences, Magnify
Keith Shields
• 20+ years in financial services, developing and deploying statistical models to predict credit risk across multiple asset classes
• CAO Magnify: 8+ years of Analytic Services and Solutions
• 14 years at Ford Motor Credit, Director of Global Analytics
• 2+ years at Credit Acceptance Corp, VP Portfolio Management

Susan Arnot
• 20+ years in financial services, developing and deploying statistical models to predict credit risk across multiple asset classes
• Director of Decision Sciences at Magnify
• 10 years at Ford Motor Credit, Manager of Credit Risk Analytics, Latin America
• 4 years at Ally Bank, Manager of Credit Risk Modeling
Credit Risk Modeling
Credit Bureaus and Typical “Big Data” Lending Problems
Scorecards and Cutoff Scores
Alternative Data, “AltFi”, Marketplace Lending, and Fintech
Common predictive modeling conventions
Regulatory Pressure
The Demand for Creativity and Disruption
What Makes “Alternative Lending Analytics” Different from “Banking Analytics”?
The Continued Relevance of Logistic Regression
“It is not the model fitting techniques that need to change, as much as it is the treatment of samples, potential predictors, and model refits”.
Quantify the Risk of a Loan Applicant

- An Example:

  If Susan has a 9% PROBABILITY OF DEFAULT, then my expected revenue is:
  
  - $110 \times .91 + \$0 \times .09 = \$100.10
  
  - => 91% is a breakeven probability. I loaned $100 and I expect roughly that in return.

If I can’t accurately quantify Susan’s probability of default, I can’t have a lending business. At Magnify we build statistical models that accurately quantify Susan’s probability of default.

The only way I fail to make a profit is if Susan doesn’t pay me back. This is called DEFAULT.
The Credit Risk Model
Still one of the best and most pervasive applications of analytics...

- Definition and Function
  - A credit risk model is, usually, a linear statistical model that uses combination of credit, contract, and personal attributes to predict the likelihood that a loan applicant will default (fail to pay back the loan).
  - “Attributes” = “Characteristics” = Independent variables
  - Credit Score
  - Loan-to-Value Ratio (LTV)…insert diatribe on 2008 mortgage disaster
  - Payment-to-Income Ratio (PTI)
  - Revolving Utilization
  - Recent Credit Inquiries

- Credit risk models are used in almost every lending decision today
  - PD’s, Basel III, SR 11-7
  - Is lending ever done without models? The Big Short…
The “FICO Score”…

- 1989: Bill Fair, Earl Isaac…”FICO”
- Credit Scoring System: Credit Risk Models are converted to “scorecards” by creating a point system whereby the parameter estimates of the credit risk model are multiplied by the possible values of the independent variables to create points.
- Example: $Y = b_0 + b_1 x_1$. Suppose $x_1$ can take on a value of 1, 2, or 3. Scorecard:

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$b_1 \times 1 \times 100$</td>
</tr>
<tr>
<td>2</td>
<td>$b_1 \times 2 \times 100$</td>
</tr>
<tr>
<td>3</td>
<td>$b_1 \times 3 \times 100$</td>
</tr>
</tbody>
</table>

Note: Typically we transform $X_1$ first

- Why was this a “Big Data” problem? Why is it still?
  - 300 million consumer bureaus, each of them roughly 32,000 bytes wide, most of those bytes heavily codified data.
This is what a credit bureau file looks like...

<table>
<thead>
<tr>
<th>TRADES</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUBNAME</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>ACCOUNT#</td>
</tr>
<tr>
<td>ECOA COLLATRL/LOANTYPE</td>
</tr>
<tr>
<td>DPT ED/NAVI</td>
</tr>
<tr>
<td>E00000000000000511</td>
</tr>
<tr>
<td>I</td>
</tr>
<tr>
<td>NAVIENT</td>
</tr>
<tr>
<td>99572734200000001</td>
</tr>
<tr>
<td>I</td>
</tr>
<tr>
<td>NELNET</td>
</tr>
<tr>
<td>50000005400000009</td>
</tr>
<tr>
<td>I</td>
</tr>
<tr>
<td>NELNET</td>
</tr>
<tr>
<td>50000005400000009</td>
</tr>
<tr>
<td>I</td>
</tr>
<tr>
<td>CAPITAL ONE</td>
</tr>
<tr>
<td>5178600000238</td>
</tr>
<tr>
<td>I</td>
</tr>
</tbody>
</table>

And this is what we need to build a model:

- Amount of open mortgage payment
- Number of satisfactory revolving bank trades
- Number of open auto trades paid as agreed
- Worst present status on an open trade opened within the last 24 months
- Total amount of collection items
- Revolving utilization %
- And so on...how many more?

At Ford Credit we experimented with over 1,000 of these, and we had over 10 million aged loans to build models on.
Suppose we make $1,000 profit on “good” loans and suffer a $10,000 loss on “bad” loans:

<table>
<thead>
<tr>
<th>Credit Score</th>
<th>% of Applicants</th>
<th>Good Rate</th>
<th>Loss (Bad) Rate</th>
<th>Expected Net Profit</th>
<th>Cumulative Approval Rate</th>
<th>Cumulative Loss Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>40%</td>
<td>99.5%</td>
<td>0.5%</td>
<td>$945</td>
<td>40%</td>
<td>0.5%</td>
</tr>
<tr>
<td>400</td>
<td>20%</td>
<td>96.0%</td>
<td>4.0%</td>
<td>$560</td>
<td>60%</td>
<td>1.7%</td>
</tr>
<tr>
<td>300</td>
<td>15%</td>
<td>92.0%</td>
<td>8.0%</td>
<td>$120</td>
<td>75%</td>
<td>2.9%</td>
</tr>
<tr>
<td>200</td>
<td>15%</td>
<td>87.0%</td>
<td>13.0%</td>
<td>-$430</td>
<td>90%</td>
<td>4.6%</td>
</tr>
<tr>
<td>100</td>
<td>10%</td>
<td>75.0%</td>
<td>25.0%</td>
<td>-$1,750</td>
<td>100%</td>
<td>6.7%</td>
</tr>
</tbody>
</table>

Some simple math on each credit tranche leads us to very useful conclusions:

- Cutoff score = 300. We lose money on borrowers who score below 300.
- We will decline roughly 25% of our applicants.
- Our loss rate will be 2.9% (as opposed to 6.7% if we approved everyone).
- Our profit per contract is $677 (as opposed to $268 if we approved everything).
Alternative lenders look for opportunities to “expand the lending box” through the use of nontraditional data sources (i.e. not the credit bureau).

What’s an alternative lender?
- The simplest definition is that an alternative lender is any lender that isn’t a bank (or captive, in my opinion).

Historically the customers of alternative lenders are people, or small businesses, who are unable to secure a bank loan. This is still largely the case, but not nearly to the extent it once was.

Recently many start-up alternative lenders have distinguished themselves from banks based on heavier use of technology, removal of traditional loan structures, and tapping into overlooked sources of collateral and non-traditional funding sources:
- Technology: Fintech platforms
- Removal of traditional repayment plans: Merchant Cash Advance, Payday Loans, and Uber
- Overlooked sources of collateral: merchant invoices
Alternating Data That Predicts Credit Risk

- Yelp reviews...merchant risk
- Address changes, evictions (Lexis Nexis)... subprime auto risk
- Student loans, tweets, and the use of the word “broke”
- Facebook likes
- LendDo and phone records
- Yodlee and bank statements
- Payday loan databases, e.g. Factor Trust

Perceived improvements in modeling techniques don’t make near as much impact as the integration of the right alternative data attributes. But what are the right alternative data attributes? That’s another Big Data problem.
Is there a place for us (statisticians, data scientists, analysts)?

- In a big way yes:
  - Huge expenditures on alternative data (Yodlee, Clarity, LendDo, Yelp)
  - 100% technology-enabled, model-driven, loan adjudication and pricing
  - Progressive underwriting

- But there are also minefields:
  - Small samples and the “overfitting contest”
  - Biased samples
  - Relentless pressure to grow (before regulation) and strong appetite from secondary markets. *The Big Short?*
Logistic regression offers the right construct for estimating risk...

Let $PD = \text{"Probability of Default"}$. Let $Y=1$ be default and $Y=0$ be paid as agreed

$PD$ is related to attributes $X_k$’s by taking a weighted sum of the attributes called “$xbeta$” where $xbeta = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n$ and substituting $xbeta$ into the logistic function

- $PD = P(Y=1|X_k) = \frac{\exp(xbeta)}{1+\exp(xbeta)} \ldots \text{the logistic function}$
- We need to find the $\beta_k$’s which best connect the $X_k$’s to the $Y$’s
- If $Y=1$, want $PD$ big  
  If $Y=0$, want $1-PD$ big
- Done by maximizing the likelihood function “$L$” as a function of $\beta_k$’s
  - Max $L(\beta | X) = \prod PD^Y (1 - PD)^{1-Y}$
Logistic regression models are at the root of most credit scores

- Per Wiki: Although logistic (or non-linear) probability modelling is still the most popular means by which to develop scorecards, various other methods offer powerful alternatives, including MARS, CART, CHAID, and random forests.

- Machine learning has become really popular lately, particularly among alternative lenders. The overfitting contest continues (next slide)…
Comparing a Logistic Regression Model with SVM Model

- SVM performs a little better as a “loan renewal model”, but performance degraded by an alarming amount from development to validation.

- Why loan renewal? Why not credit risk? Developing an actual credit risk model with machine learning is really not an option. The ECOA requires adverse action reasons, which are difficult to extract from “black box” machine learning models.
Building predictive models has become a canned process for many banks – why?
- Reg B: Empirically-derived, statistically sound
- ECOA
- OCC 11-7: Supervisory Guidelines on Model Risk Management
- CFPB and lenders’ lack of transparency
- Credit Ratings and Secondary Markets
- Banks shrinking desire to do retail banking. The alternative? Commercial lines-of-credit for alternative lenders
- Millennials, millennials, millennials...
Big Data can't be just a buzzword...the search for 4-leaf clovers should be continuous and disciplined.

- Build models to understand, not always to predict

- Unstructured data has to be structured in an intelligent way

- Personally-verifiable, or volunteered, data

- Technology devoted to collecting unstructured data and structuring it
Questions

- Thanks for your time and attention.
- Keith Shields: kshields@magnifyas.com
- Susan Arnot: sarnot@magnifyas.com