Approximations for Rare Events Forecasting Under Monotonic Constraints

Matthew Ginley

R/Finance 2018
Figure 1: A Rare Events Metaphor
Distress Risk as a Monotonic Rare Events Problem

Approximation of Rare Events Datasets Using Kernel Density Estimators

Preliminary Results and Closing Remarks
Distress Risk as a Monotonic Rare Events Problem
Corporate Bankruptcy or Distress Risk Modeling with Financial Ratios

- Classics: (Altman 1968), (Ohlson 1980), (Lo 1986). **Demonstrate efficacy with well known ratios and linear methods.** Debt To Equity, Earnings Yield, etc.
- More recent: (Shumway 2001), (Beaver, McNichols, and Rhie 2004), (Chava and Jarrow 2004). **Relationship is robust over time.** Other improvements, but still primitive, linear methods.
- Model target is binary indicator (discrete, observable event) of Ch. 7 or 11 Bankruptcy Filing
- Model forecasts are probabilities (continuous, latent quantity), interpreted as a measure of distress risk.
Corporate Bankruptcy as a Rare Events Data Problem

- Universe of US Listed Equity, 1965 - 2014
- Monthly observations of all firms, roughly 2.4 million firm-months
- Only 1,052 bankruptcy filings, sample mean bankruptcy rate is about 0.05%
- Literature suggested 3-8 predictor ratios and some market observables such as capitalization. This is open problem, not our focus.
Rare Events Data Problems

Binary classification problem with a severe imbalance in the observed class proportions (majority/minority or class 0/1) (King and Zeng 2001)

- 50/50: heads/tails coin flip (even)
- 60/40: NFL home/away team winning (balanced)
- 71/29: 538 final election win probability Clinton (imbalanced)
- 90/10: (severe imbalance)
- 99/01: (rare events)
- 99.95/0.05: monthly, corporate non-bankrupt vs. bankruptcy (extreme rare events)
- 100.0/0.00: Cleveland Browns not/winning Super Bowl (determinism)
Figure 2: Bankruptcy Data Example (Firm-Month Obs. from 2014)
Figure 3: Bankruptcy Data Example, with Jitter (Firm-Month Obs. from 2014)
Figure 4: Bivariate Scatter Plot of Non-Bankrupt and Bankruptcy Classes (Firm-Month Obs. from 2014)
Monotonicity

- SIMPLE: If firm A has greater leverage than firm B, firm A must have greater distress risk
- Technically, there is Monotonic Non-Decreasing vs. Monotonic Non-Increasing
- Linear models are a subset of all monotonic models. Monotonicity can be thought of as a “relaxation” of linearity. Very useful!
- Monotonically constrained non-parametric models are easy to conceptualize but estimation is difficult in practice
The relationship is non-linear but monotonic, and we have lots of data!

But fully non-parametric models can take a while, and technically right now this doesn’t exist.

Big Data, Rare Events is kind of the “worst of both worlds”. Very challenging setting for machine learning

We can approximate the information contained in the data and then exploit the presence of monotonicity to minimize what we might have lost.
Approximation of Rare Events Datasets Using Kernel Density Estimators
Naive Bayes

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

$$\mathbb{P}[Y = 1|X = x] = \frac{f(x|Y = 1)\mathbb{P}[Y = 1]}{f(x)}$$

$$\mathbb{P}[Y = 1|X = x] = \frac{\phi(x|\mu_1, \sigma_1^2 1') \left( \frac{n_1}{n} \right)}{\phi(x|\mu_1, \sigma_1^2 1') \left( \frac{n_1}{n} \right) + \phi(x|\mu_0, \sigma_0^2 1') \left( \frac{n_0}{n} \right)}$$
Kernel Smoother, or Naive Bayes with Non-parametric Densities

Same as before . . .

\[ P[Y = 1|X = x] = \frac{\phi(x|\mu_1, \sigma^2_1 1')(\frac{n_1}{n})}{\phi(x|\mu_1, \sigma^2_1 1')(\frac{n_1}{n}) + \phi(x|\mu_0, \sigma^2_0 1')(\frac{n_0}{n})} \]

But change from Gaussian to Non-parametric densities . . .

\[ \phi(x|\mu_1, \Sigma_1) \to \frac{1}{n_1} \sum_{\{i:y_i=1\}} \phi(x|x(i), h_1^2 1') \]
\[ \phi(x|\mu_0, \Sigma_0) \to \frac{1}{n_0} \sum_{\{i:y_i=0\}} \phi(x|x(i), h_0^2 1') \]
Figure 5: KDE of 3 points, Gaussian Kernel, bandwidth = 1
Figure 6: Bankruptcy Data Example, KDE (Firm-Month Obs. from 2014)
Figure 7: Bankruptcy Data Example, KDE (Firm-Month Obs. from 2014)
Figure 8: Bankruptcy Data Example, Combo KDE + Monotonically Constrained Non-parametric Regression (Firm-Month Obs. from 2014)
Re-imagine the KDE as Sampling with Noise

- Nearly identical to the ROSE (Random Over-Sampling Examples), but applied to classification instead of regression (Lunardon, Menardi, and Torelli 2014)
- But with a tweak: Don’t repeatedly sample, just use the noise assumption (KDE) to generate proportions at minority class points only and follow with a **monotonically constrained** regression
- Repeated sampling with noise and model fitting, and likewise subsetting and ensembling, won’t work due to ultra low minority class proportion. Rare Events . . . signal is extremely faint.
Figure 9: Needle In A Haystack, Needle Locations
Figure 10: Needle In A Haystack, Needle To Hay Proportions
How can this approximation work for training a regression model? Monotonicity!

- Given this property, we are indifferent to the nature of the data with respect to how it evidences the true odds ratio. More class 1 points, fewer class 0 points in any region, vice versa.
- With monotonicity, we have additional understanding of the nature of the relationship. We can interpolate. Counter-example: without monotonicity, the relationship between class 1 points would be impossible to infer (using our approach).
- By virtue of rare events, relationship could be assumed to be 0 everywhere. We only know for sure where it is greater than 0 at observed minority class points. Let's maximize what we know (signal) at those points and the interpolate.
1. Collect data and construct binary target indicator, predictors, etc.
2. Select a KDE bandwidth
3. Using the full training dataset, calculate sample proportions, but at minority class points only
4. Use sample proportions (only) to train a monotonically constrained regression (with link fun)
5. Generate final forecasts of minority class probability for the test set from the regression (and invert link fun)
Bandwidth Specifies the Degree of Noise in the Dataset

(Scott 1992) Multivariate Scott Rule bandwidth estimator:
\[ \hat{h}_{SR} = \hat{\sigma} n^{-1/(d+4)} \]

▶ Theoretical guarantees to minimize error. Driven by sample variance and dataset dimensions \((n \times d)\)
▶ Good results using some multiple of this rule (and without class differentiation)
▶ Selection is then an open-ended, but very manageable 1-dim problem
▶ We validate on the degree of monotonicity exhibited by the transformed dataset, but 1-dim grid search based on training error works
Our Pick for Monotonically Constrained Regression

Bayesian Additive Regression Trees (BART)

- Original (Chipman, George, and McCulloch 2010)
- Monotonically constrained (Chipman et al. 2016)
- Impractical with full dataset (and classification not implemented yet)
- BUT, nice synergy with our approximation. estimation works on spacings between data points. Think “sparse BART”
Figure 12: Approximation As A Shortcut
Positive Takeaways

▶ effectively capture non-linear dynamics
▶ shines with out of sample extrapolation in the extremes corresponding with high risk
▶ procedural, statistical efficiencies
Figure 13: Bankruptcy Data Example, Proposed Approximation vs. Logistic Regression (Firm-Month Obs. from 2014)
3 Sources for Degradation

1. Decreasing severity of class imbalance
2. Increasing number of non-monotonic predictors
3. Increasing number of predictors
Preliminary Results and Closing Remarks
Evaluation of Rare Events Is Challenging

- Topic of on-going research within specialties like Meteorology (Ferro and Stephenson 2011)
- Traditional metrics like MSE and 0-1 Loss are worthless. Constant 0 forecast performs well
- Area Under the Receiver Operating Characteristic Curve (AUC or AUROC) is latest standard (Fawcett 2006), but not helpful here because of absolute lack of bankruptcy events.
Figure 14: Test Set Results, Various Methods (July 2007 - December 2014)
Donald J. Trump
@realDonaldTrump

Moron Matt just isn't getting it done with his bankruptcy work. Years down the drain and still no success! Sad!

Figure 15: What A Bad Day at the Office . . .
Probability Calibration

- Can be thought of as a visual depiction of the long run batting average or hit rate for a classification model.
- Easy sports example using NFL games ("Evaluating logistic regression beyond classification accuracy: The case of NFL matchup prediction" 2016)
Figure 16: Probability Calibration Plot, Hypothetical, Perfect Model
Figure 17: Probability Calibration Plot, Linear Model (Live Backtesting July 2007 - December 2014)
Figure 18: Probability Calibration Plot, Monotonic Model (Live Backtesting July 2007 - December 2014)
Final Comments

- Monotonic Non-parametric is powerful class of models. Best of both worlds. Unconstrained, fully non-parametric often results in “Garbage Out”.
- BART and monotonically constrained BART are great and have readily available software.
- KDEs have been around forever but very useful. This approximation can be used by itself for EDA and validating other models, and transforming classification to regression with small data.
- Tukey: “Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.”
Thank You

R/Finance committee and sponsors, Michael Weylandt and other conference regulars!

Special Thanks to Dr. Chava (Georgia Tech) for bankruptcy data, and Dr. McCulloch (ASU) for feedback and distribution of BART software.

I am actively seeking work, am always looking to learn more, and can notify you upon publication of paper and code (currently in submission).

mginley@gmail.com


