When Production Machine Learning Fails

John Urbanik
DataEngConf
10/31/17

predata

OR:

When initially promising seeming supervised learning models don't quite make it to production, or fail shortly after being productionized, why?

How can we avoid these failure modes?

Media Coverage of AI/ML Failure

03/28/2016 07:54 am ET

The Future Of Crime-Fighting Or The Future Of Racial Profiling?: Inside The Effects Of Predictive Policing

The idea of PredPol is that if officers focus their attention on an area that's slightly more likely to see a crime committed than other places, they will reduce the amount of crime in that location.

Alexis C. Madrigal Fusion

LUCAS JACKSON / REUTERS

Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter

Attempt to engage millennials with artificial intelligence backfires hours after launch, with TayTweets account citing Hitler and supporting Donald Trump

Tay uses a combination of artificial intelligence and editorial written by a team including improvisional comedians. Photograph: Twitter

Microsoft's attempt at engaging millennials with artificial intelligence has backfired hours into its launch, with waggish Twitter users teaching its chatbot how to be racist.

The company launched a verified Twitter account for "Tay" - billed as its "AI fam from the internet that's got zero chill" - early on Wednesday.

Tesla driver killed in crash with Autopilot active, NHTSA investigating

294

by Jordan Golson | @jlgolson | Jun 30, 2016, 4:42pm EDT

SHARE TWEET LINKEDIN

A Tesla Model S with the Autopilot system activated was involved in a fatal crash, the first known fatality in a Tesla where Autopilot was active. The company revealed the crash in a blog post posted today and says it informed the National Highway Transportation Safety Administration (NHTSA) of the incident, which is now investigating.

predata

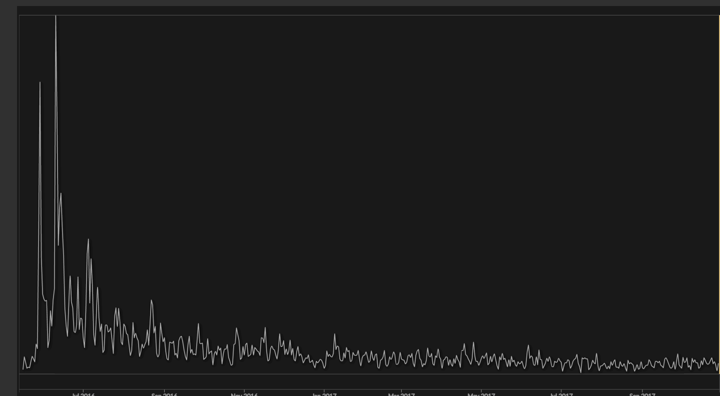
A Framework

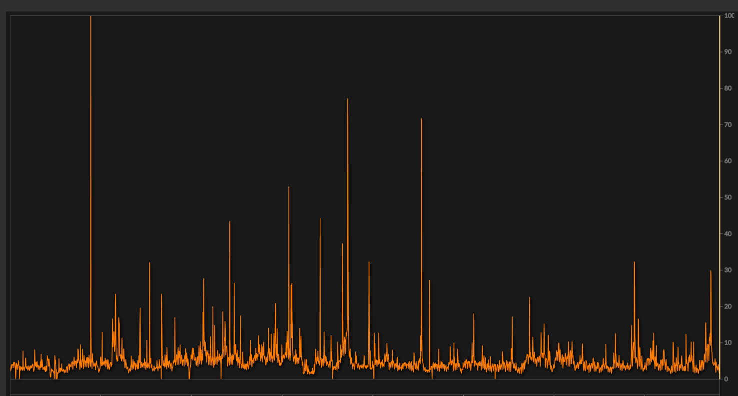
1. A survey of some less discussed failure modes
2. Techniques for detecting and/or solving them

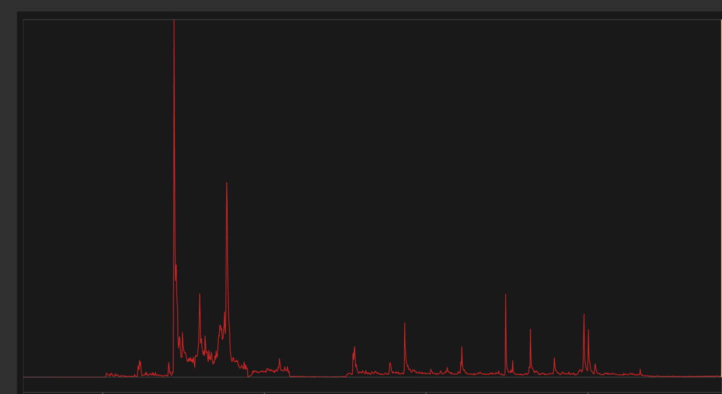
- Class Imbalance
- Time based effects
  - latent time dependence
  - concept drift
  - Non-stationarity
  - Structural breaks
- Business applicability
  - Dataset availability,
  - Look-ahead bias
  - Metrics and loss functions

Predata Data

Our data exhibits all sorts of non-stationarity, is extreme value distributed, have many structural breaks. Our prediction targets are heavily imbalanced and exhibit multiple modes of concept drift.







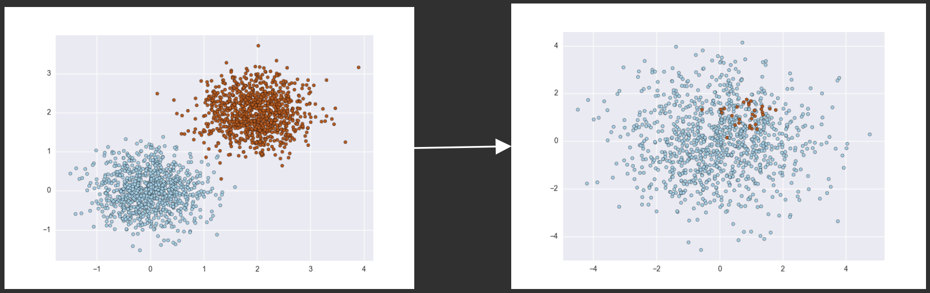
Things Not Covered

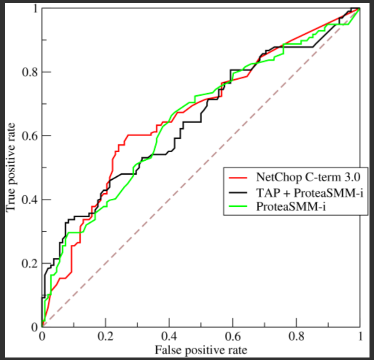
- Conventional overfitting
- Interpretability
  - Most commonly raised obstacle, often used to help with model selection
- Lack of data
  - In some cases this is solvable with money or time
  - Also see Claudia's talk titled "All The Data and Still Not Enough"
- Dirty, noisy, missing, or mislabeled data
  - Refer to Sanjay's talk yesterday
- Problems without 'straightforward' solutions (i.e. censored data, unsupervised learning and RL)

Class Imbalance

- Classical examples: cancer detection, credit card fraud
- Predata examples: terrorist incidents, large scale civil protests

- MSE / Accuracy derived metrics don't work well
- ROC, Cohen's Kappa, macro-averaged recall better, but not the end all





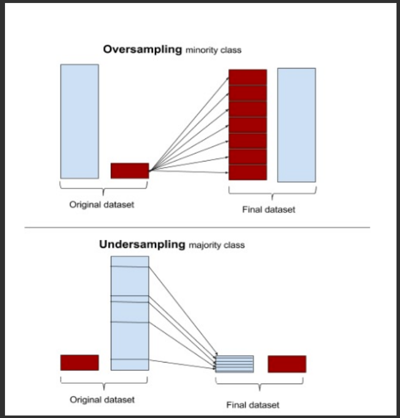
$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

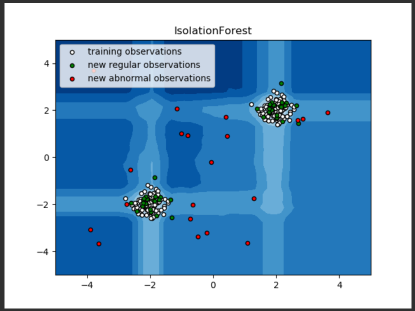
predata

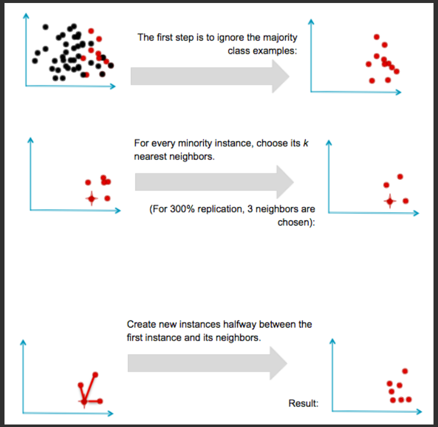
Class Imbalance (cont’d)

1. Oversampling, undersampling
2. Adjust class / sample weights
3. Frame as anomaly detection problem (only in two class case)
4. SMOTE and derivatives - ADASYN and other variants

Check out [imbalanced-learn]()







https://svds.com/learning-imbalanced-classes/

predata

Latent Time Dependence

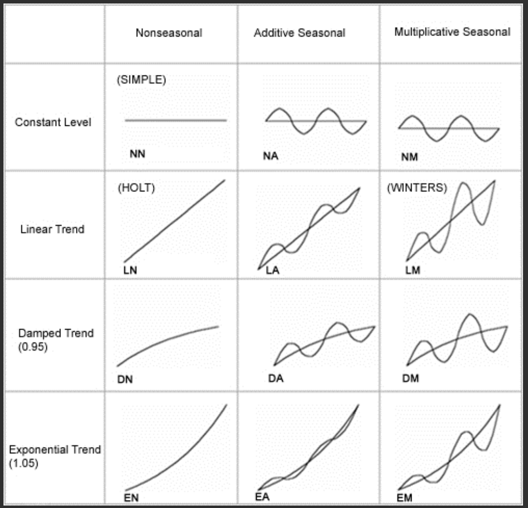
- Don't JUST use K-Fold cross validation
  - Also use a set of time oriented test/train splits
  - Some time series splits are ‘lucky’ or ‘easy,’ especially in the presence of concept drift and class imbalance
- Plot performance metrics via a sliding window over time in holdout

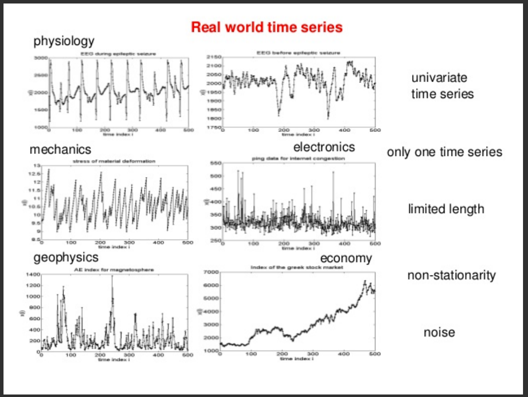
https://svds.com/learning-imbalanced-classes/

Non-stationarity

- Seasonality / weak stationarity
  - seasonal adjustment
  - feature engineering
- Trend stationary
  - Growth (exponential or additive)
  - KPSS test
  - Model the trend, remove it
  - Rolling z-score
- Difference stationary
  - ADF unit root test
  - Use differencing to remove
  - Beware fractional integration - long memory (GPH test)

$$
\begin{aligned}
(1-B)^d &= \sum_{k=0}^{\infty} \binom{d}{k} (-B)^k \\
&= \sum_{k=0}^{\infty} \frac{\prod_{a=0}^{k-1}(d-a)\,(-B)^k}{k!} \\
&= 1 - dB + \frac{d(d-1)}{2!}B^2 - \cdots .
\end{aligned}
$$

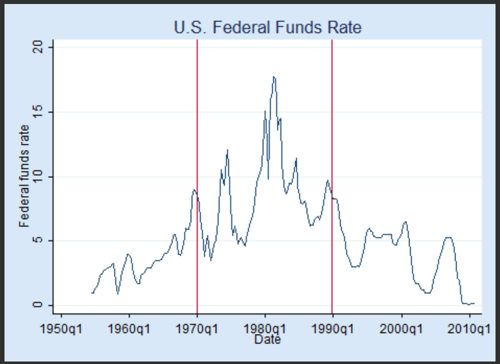




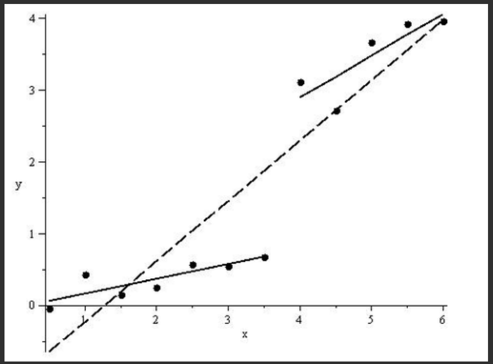
http://www.simafore.com/blog/bid/205420/Time-series-forecasting-understanding-trend-and-seasonality

Structural Breaks

- Unexpected shift, often caused by exogenous events
- Change detection is a very active area of research
  - Chow test for single change-point
  - Multiple breaks require tests like sup-Wald/LM/MZ
  - These make assumptions like homoskedasticity
- Mitigate by using just recent data



https://www.stata.com/features/overview/structural-breaks/



https://en.wikipedia.org/wiki/Structural_break#/media/File:Chow_test_example.png

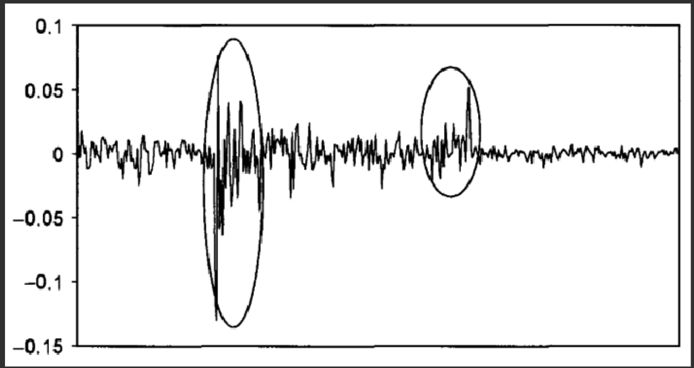
Concept Drift

Changing relationship between independent and dependent variables
OR
Changing class balance / Mutating nature of classes

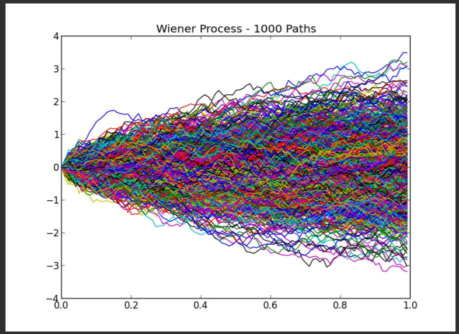
- Active and passive solutions:
  - Active rely on change detection tests / online change detection
  - Passive solutions continuously update the model
  - There is active research in ensembling based on time based performance
    - Predata is particularly interested in resurfacing old successful classifiers after some transient change / exogenous shock

Other Time Series Effects

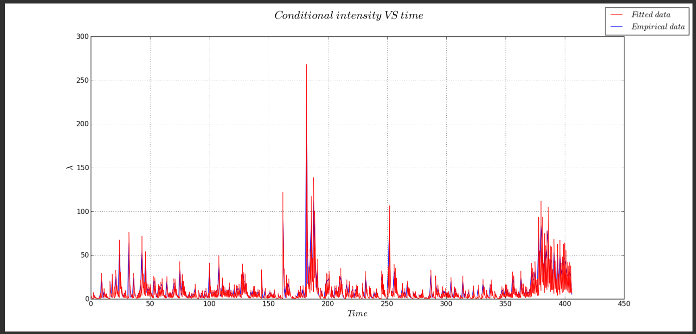
- Volatility clustering
- Poisson/Cox/Hawkes processes
- Random walks / Wiener processes



Volatility Clustering Phenomenon of Financial Time Series Source: Alexander, C. (2001)



https://github.com/matthewfieger/wiener_process



https://stackoverflow.com/questions/24785518/how-to-compute-residuals-of-a-point-process-in-python

predata

Look-Ahead Bias and Time Delays

- Make sure that you have guarantees (or mitigation strategies) if you have data availability failures
  - Ensemble models with different delays
  - Surface data outages to data consumers
- Feature engineering done now might not have been intuitive in the past. If there is concept drift, how can we be sure that performance will continue.
  - Look at performance over time in live test
  - Automated feature engineering / feature selection
  - Use judgement; use features that seem like they would be stable across time (little concept drift) or features that would likely be discovered in real time

Loss Functions and Metrics

- How does you business value Type I/II errors?
- Time series prediction specific:
  - Is an early prediction useful?
  - Should a late prediction be penalized fully?
  - How do we weight samples based on their importance?
- How do you translate business concerns to the optimization / modeling layer
  - Writing custom loss functions
  - AutoGrad, PGM like Edward
  - Genetic algorithms

Questions?

John Urbanik
jurbanik@predata.com
@johnurbanik

predata