University of Nottingham
UK | CHINA | MALAYSIA

Validate AI

Metamorphic Exploration for Machine Learning Validation and Model Selection

Dr. Anthony Bellotti*,
Zhihao Ying*,
Dr. Joe Breeden & Prof. David Towey***

*** School of Computer Science, University of Nottingham Ningbo China**
****Deep Future Analytics LLC**

University of Nottingham Ningbo China (UNNC)

- University of Nottingham Ningbo China (UNNC) was the first Sino-foreign university to open its doors in China in 2004.
- At UNNC we have over 9,000 students and more than 7% are from China's Hong Kong, Macao, Taiwan regions and other overseas countries.
- We have over 900 members of staff, split across academic and professional services.
- Our staff and students come from about 70 countries and regions around the world.
- Teaching and research university.

University of Nottingham
UK | CHINA | MALAYSIA

Introduction to Validate AI

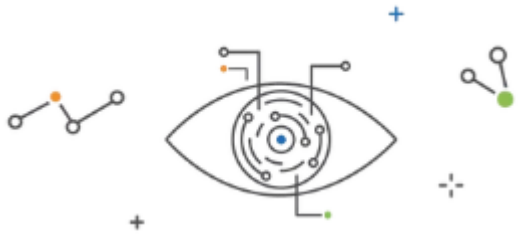
Validate AI

- Validate AI (VAI) is an independent community interest company (UK) that strives to be the 'go to' organisation dedicated exclusively to improving how AI systems are validated to build trust. We advocate a cross-sectoral collaborative approach between academia, government, industry, and charity sectors to maximise impact.

- Our primary objectives are to validate AI and promote:
  1. Assurance and practitioner-centric standards.
  2. Innovation in how we validate AI.
  3. Convening of the community to AI validation best practice.

Current programme of roadshows starting with

Toward Trustworthy AI: Operationalising AI Assurance, Bayes Centre, University of Edinburgh, 5th September



Web site: validateai.org

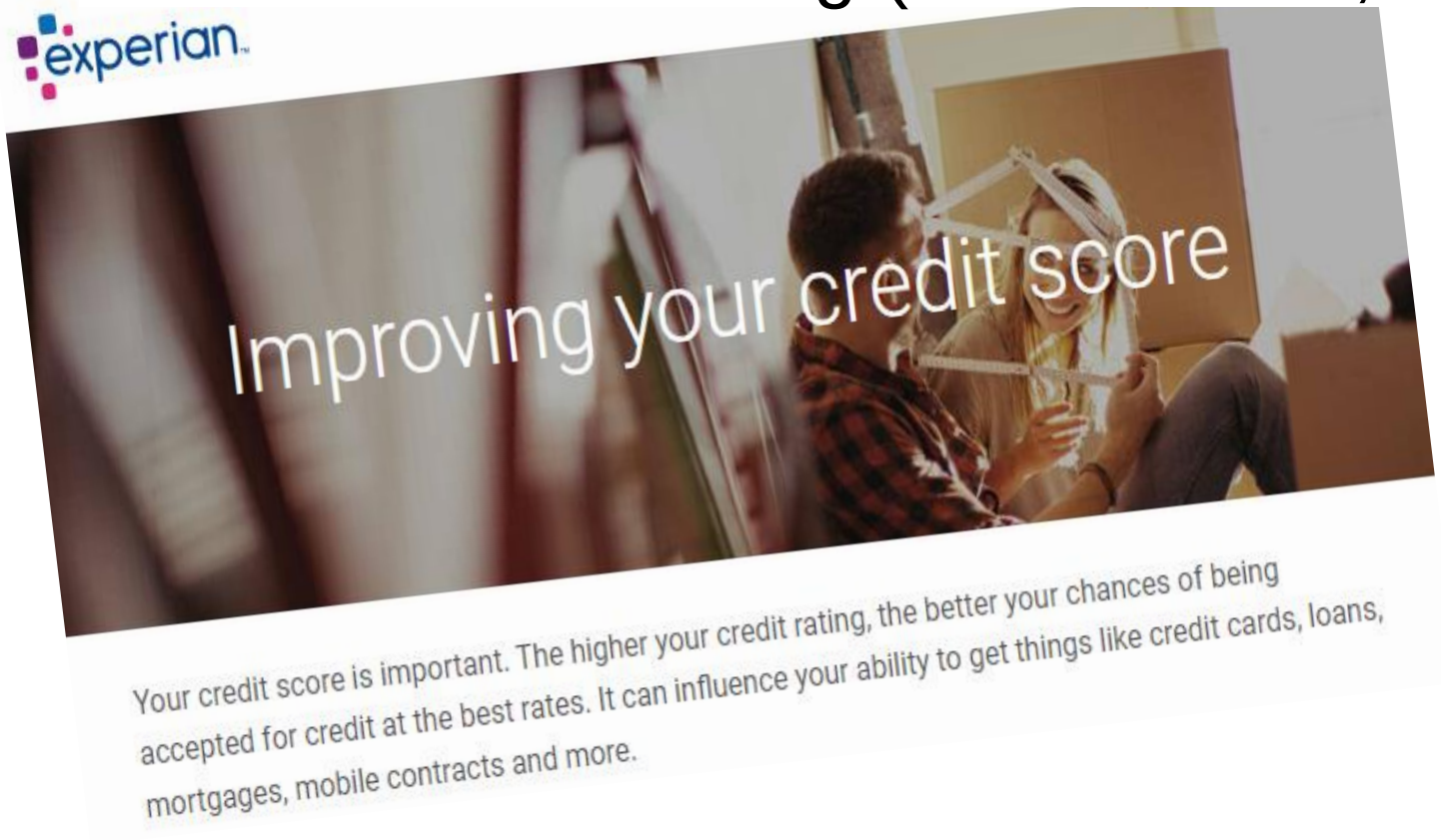
Business Expectations

- We expect our credit scoring models to match business expectations or intuition.
- Example 1. Increase in bureau credit score should lead to decrease in probability of default (PD) (everything else being equal).
- Example 2. Decrease in Loan-to-Value (LTV) should lead to lower interest rate (everything else being equal).

Business Expectations

Implications for customers:

- Acceptance or rejection of loan application.
- Price of borrowing (interest rate, fees).



www.experian.co.uk/consumer/guides/improve-credit-score.html

Which?

"If you have a low or 'bad' credit score, you're more likely to be turned down when you apply to borrow money, or offered less favourable interest rates, in which case you should take steps to improve your score."

Using Tables or Linear Models

- Fairly easy to spot relationship between predictors and PD based on table scores or direction of model coefficients.

Example: LTV band table

LTV	PD	Interest Rate
60%	0.007	5.9%
75%	0.014	6.1%
80%	0.018	6.3%
95%	0.024	6.6%

Example: Model of default:
higher value means higher PD

Feature	Model coefficient	Example	Score
Bureau credit score	+0.01	623	+6.23
LTV	+0.02	80	+1.6
Debt-to-Income	+0.05	20	+1.0
Remortgage (0/1)	-1.2	1	-1.2
		TOTAL	7.63

Using Tables or Linear Models

- Fairly easy to spot relationship between predictors and PD based on table scores or direction of coefficient.

Example: LTV band table

LTV	PD	Interest Rate
60%	0.007	5.9%
75%	0.014	6.1%
80%	0.018	6.3%
95%	0.024	6.6%

Example: Model of default:
higher value means higher PD

Feature	Model coefficient	Example	Score
Bureau credit score	-0.01	623	-6.23
LTV	+0.02	80	+1.6
Debt-to-Income	+0.05	20	+1.0
Remortgage (0/1)	-1.2	1	-1.2
		TOTAL	-4.83

Direction of model coefficient as selection criteria

- For linear models trained on historic data, the direction of coefficients can be used as a model selection criteria.

- If direction of effect of a feature does not match business expectations, then reject model.

See e.g. Joseph L. Breeden and Nikolay Dobrinov. *Quantifying model selection risk in macroeconomic sensitivity models*.

Journal of Risk Model Validation, 16(3), 2022.

Rise of Machine Learning in Credit Scoring

- Machine Learning (ML) is increasingly being used in credit scoring:-
  1. Improved performance / accuracy;
  2. Expanded market by use of new data sources;
  3. Automation: operational efficiency.

- Machine learning in UK financial services (Bank of England 2022 survey):-
  - The number of UK financial services surveyed firms that use machine learning (ML) continues to increase.
  - Overall, 72% of firms that responded to the survey reported using or developing ML applications.
  - These applications are becoming increasingly widespread across more business areas.
  - ML applications are now more advanced and increasingly embedded in day-to-day operations.
  - 79% of ML applications are in the latter stages of development, ie either deployed across a considerable share of business areas and/or critical to some business areas.

https://www.bankofengland.co.uk/report/2022/machine-learning-in-uk-financial-services

Challenges of Machine Learning in Credit Scoring

- Bias and discrimination.
- Data quality and Data security.
- Robustness *to changes and crises over time.*
- Explainability. *Being able to explain lending decisions.*
- Consistency *with business expectations.*

Why are these especially a problem for machine learning algorithms?

Challenges of Machine Learning in Credit Scoring

- Bias and discrimination.
- Data quality and Data security.
- Robustness *to changes and crises over time.*
- Explainability. *Being able to explain lending decisions.*
- Consistency *with business expectations.*

Why are these especially a problem for machine learning algorithms?

Complexity!

There are no simple coefficients that link features to model output.

Black Box

Metamorphic Testing

- How to test a ML model's consistency with business expectations?

- Use metamorphic testing:-

  - Construct a Hypothetical Metamorphic Relation (HMR).
    1. Construct a source test case (STC).
    2. Construct a follow-up test case (FTC).
    3. Form a metamorphic relation (MR) between STC and FTC which reflects expected behaviour of the system.
    4. Execute the test with multiple (STC,FTC) pairs and record violations of MR.

See e.g. Zhi Quan Zhou, Liqun Sun, Tsong Yueh Chen, and Dave Towey. Metamorphic relations for enhancing system understanding and use. *IEEE Transactions on Software Engineering*, 46(10):1120–1154, 2020.

Metamorphic Testing: Credit Scoring Example

- Hypothetical metamorphic relation (HMR1): Increasing bureau credit score (everything else the same) should give

$$PD_{STC} > PD_{FTC}$$

- MR procedure:
  1. STC is an example from test set (historic credit data);
  2. FTC is same as STC but with bureau credit score increased by 10%.
  3. Use the model to output PD for STC and FTC, PD_{STC} and PD_{FTC} respectively.
  4. Test HMR1: $PD_{STC} > PD_{FTC}$

- Repeat the test across all test examples, and compute proportion of examples where HMR1 is violated.

Metamorphic Testing: Composite Example

- Hypothetical metamorphic relation (HMR5): Increasing bureau credit score and decreasing Debt-to-income (DTI) (everything else the same) should give

$$PD_{STC} > PD_{FTC}$$

- MR procedure:
  1. STC is an example from test set (historic credit data);
  2. FTC is same as STC but with bureau credit score increased by 10% *and* DTI decreased by 10%.
  3. Use the model to output PD for STC and FTC, PD_{STC} and PD_{FTC} respectively.
  4. Test HMR5: $PD_{STC} > PD_{FTC}$

- Repeat the test across all test examples, and compute proportion of examples where HMR1 is violated.

Experiments with US mortgage data

- Use Freddie Mac loan-level data for US mortgages.

  `https://www.freddiemac.com/research/datasets/sf_loanlevel_dataset.page`

- Default defined as 90-days or more repayment delinquency, within 3 years of origination.
- Originating in 2016 (43,000 loans with 1,360 defaults) and 2018 (44,672 clients with 1,467 default).
- Build models using neural networks (NN), gradient boosting decision trees (GB) and random forest (RF). Common ML methods in the literature – *which is worst or best?*
- Measure AUC (area under the ROC curve) on an independent test set (25%) for model fit criterion.
- Measure percentage violations of HMRs on the independent test set.

Experimental Results: Neural Networks

- NN size = number of neurons in each layer * number of layers
- HMR 1 (credit score), HMR 2 (LTV), HMR 3 (credit score & DTI)

Year	HMR	2*1	3*1	5*1	2*2	5*2	2*3	5*3
2016	1	0.29%	0.18%	0.81%	0.64%	3.47%	0.08%	6.98%
	2	0.15%	0.87%	1.68%	0.56%	4.13%	0.51%	11.61%
	5	0.16%	0.12%	0.36%	0.3%	17.95%	0.10%	4.52%
AUC		0.718	**0.728**	0.717	0.693	0.703	0.673	0.699
2018	1	0.08%	0.08%	0.08%	0.17%	0.95%	0.27%	2.71%
	2	0.39%	1.67%	3.37%	0.49%	9.18%	0.91%	12.49 %
	5	0.03%	0.05%	0%	0.13%	0.63%	0.25%	1.78%
AUC		0.755	**0.759**	0.748	0.733	0.740	0.682	0.731

Experimental Results: Gradient Boosting Decision Trees

- GBDT size = number of trees in ensemble
- HMR 1 (credit score), HMR 2 (LTV), HMR 3 (credit score & DTI)

Year	HMR	1	10	50	80	100	500	1000
2016	1	0%	0.04%	21.19%	39.57%	38.91%	35.09%	23.67%
	2	0%	0.49%	0.75%	1.46%	1.56%	10.05%	17.37%
	5	0%	0.04%	12.15%	29.80%	28.86%	28.31%	23.35%
AUC		0.680	0.712	0.729	**0.731**	0.730	0.707	0.679
2018	1	0%	0.18%	0.40%	0.69%	1.2%	1.29%	2.89%
	2	0%	1.64%	1.1%	2.56%	1.67%	16.11%	23.44%
	5	0%	0.25%	0.92%	0.96%	0.98%	1.29%	3.04%
AUC		0.707	0.733	0.755	**0.759**	**0.759**	0.749	0.734

Experimental Results: Random Forest

- RF size = number of trees in ensemble
- HMR 1 (credit score), HMR 2 (LTV), HMR 3 (credit score & DTI)

Year	HMR	1	50	100	500	1000	5000	10000
2016	1	3.21%	42.45%	48.84%%	54.72%	55.50%	56.32%	56.4%
	2	1.27%	18.89%	23.04%	29.66%	32.41%	34.49%	33.75%
	5	3.47%	40.29%	45.42%	49.83%	50.4%	51.27%	51.31%
AUC		0.517	0.660	0.677	0.691	0.695	**0.696**	**0.696**
2018	1	1.49%	16.2%	18.77%	26.24%	27.83%	30.46%	30.87%
	2	1.60%	19.94%	24.04%	30.23%	31.93%	34.32%	34.66%
	5	2.05%	20.83%	24.09%	27.80%	28.5%	29.6%	29.68%
AUC		0.527	0.693	0.707	0.718	0.719	**0.721**	**0.721**

Conclusions and Implications

- Selecting best model on model fit (test AUC) does not typically yield good results for business expectations.
- Introduce Metamorphic Testing (MT) as a key part of the test process;
- Think carefully about Hypothetical Metamorphic Relations (HMR)
- Introduce MT as a model selection criterion;
- Use MT for monitoring after deployment.
  - Note that final outcome (default/non-default) is not needed, so results of metamorphic testing are immediate, given a new example.

University of Nottingham
UK | CHINA | MALAYSIA

Thank you!

Dr. Anthony Bellotti
School of Computer Science
University of Nottingham Ningbo
China