The Importance of Being Calibrated: A Study on Probability Calibration and Interval Estimation for Binary Classification

Elisa Terzini, Sezer Mezgil, Sandro Khizanishvili

Statistical Learning
Sapienza University of Rome
Sep 2025

Agenda

Introduction & Motivation: Can We Trust a Model's Confidence?

Question:

A model predicts a 20% chance of positive class. What does this actually mean?

For a **well-calibrated model**, over many predictions, a model that gives a 20% probability score should be correct **approximately** 20% of the time for similar predictions

But:

Most powerful modern algorithms (e.g., SVM, GBM, Random Forests, Logistic Regression ...) are naturally "scoring classifiers". They excel at discrimination (ranking instances) but are often **poorly calibrated**.

Predicted Probability $P(Y=1 \mid X=x)$	True Outcome	Observed Frequency
0.2	0	$1/5 = 0.2$
…	…	…
…	…	…
0.2	0	$1/5 = 0.2$
…	…	
0.2	1	$1/5 = 0.2$
0.2	0	$1/5 = 0.2$
…	…	…
…	…	…
0.2	0	$1/5 = 0.2$

Agenda

Discrimination vs Calibration:

Two models that are equally accurate (70% correct) show different levels of confidence in their predictions. **Model A** uses ***well-calibrated*** probabilities while **Model B** only uses extreme probabilities.

Formal Definition :

A random variable $\boldsymbol{P}$ taking values in $[0, 1]$ is ***well-calibrated*** for a random variable $\boldsymbol{Y}$ taking values in $\{0, 1\}$ if

$$\mathbb{E}(\boldsymbol{Y} \mid \boldsymbol{P}) \approx \boldsymbol{P}$$

$\boldsymbol{P}$ is the prediction made by a probabilistic predictor for $\boldsymbol{Y}$

		Model A		Model B	
	LABEL	PREDICTION	PREDICTION PROBABILITY	PREDICTION	PREDICTION PROBABILITY
1	NO	NO	0.3	NO	0.0
2	NO	NO	0.3	NO	0.0
3	NO	NO	0.3	NO	0.0
4	NO	YES	0.7	YES	1.0
5	NO	YES	0.7	YES	1.0
6	YES	NO	0.3	NO	0.0
7	YES	YES	0.7	YES	1.0
8	YES	YES	0.7	YES	1.0
9	YES	YES	0.7	YES	1.0
10	YES	YES	0.7	YES	1.0
		Accuracy 0.7		Accuracy 0.7	

Agenda

- Introduction & Motivation: Can We Trust a Model's Confidence?
- Accuracy is Not Enough: Discrimination vs. Calibration
- Calibration Methods: Platt, Isotonic, and Venn-Abers
- Our Dataset & Experimental Setup
- Credit Risk Data Preparation
- Results on Credit Risk Data: Accuracy vs. Calibration
- Simulated Data Generation
- Results on Simulated Data
- Probability Intervals
- Conclusion & Key Takeaways
- References

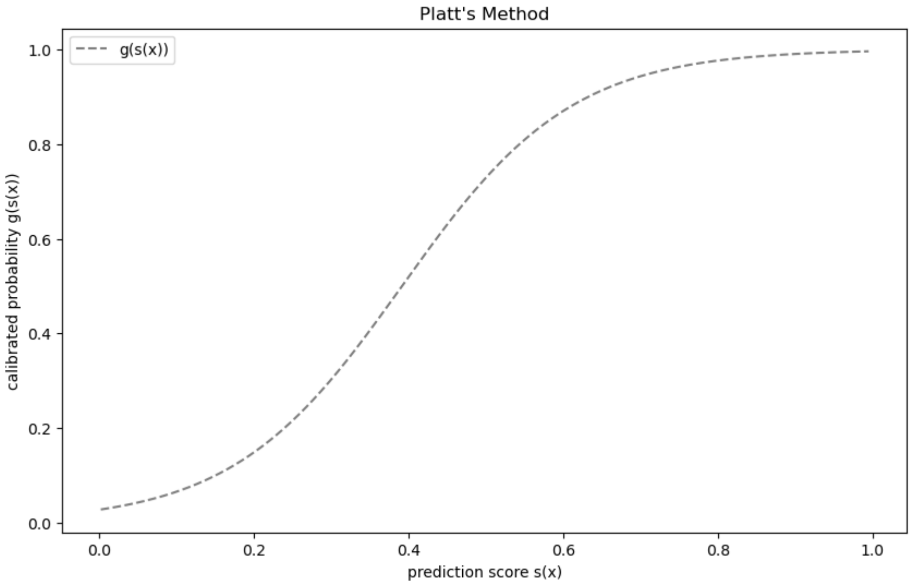
Calibration Methods: Platt’s method

- Platt’s method uses sigmoid:

$$g(s) \coloneqq \frac{1}{1+\exp(As+B)}$$

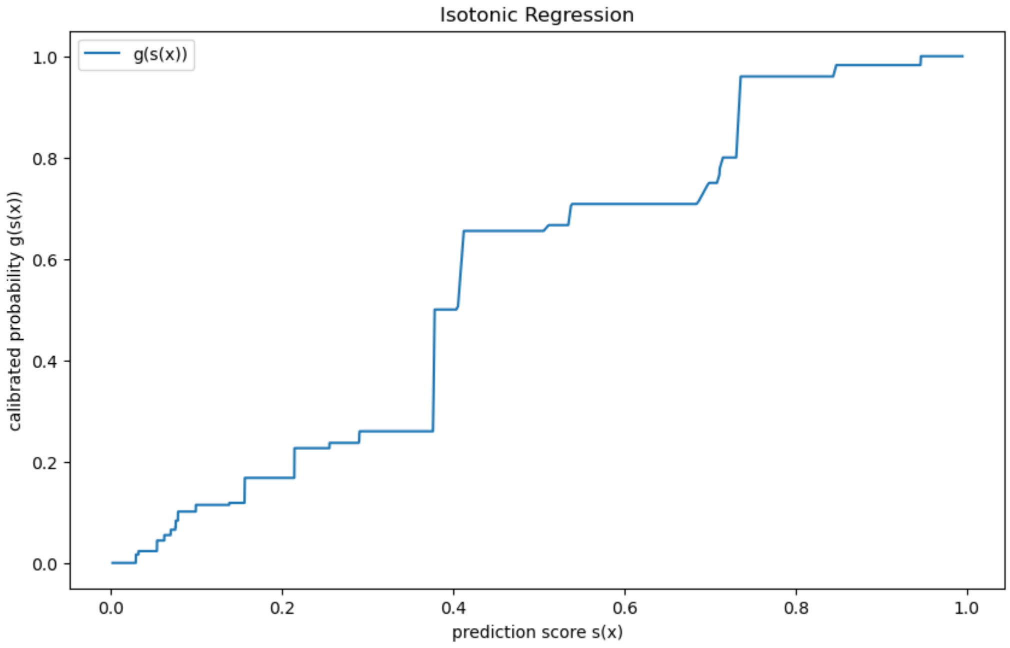
Where A<0 and B are parameters that are determined via MLE.

- **Pros:** Simple, fast, less prone to overfitting on small datasets.
- **Cons:** Makes a strong assumption that the distortion in scores follows a sigmoid shape.



Calibration Methods: Isotonic Regression

- A **non-parametric approach** that fits a flexible, step-wise function to the model's scores.
- Fit an **isotonic regression model** (a.k.a. a non-decreasing step function) to the pairs $(s(x_i), y_i)$ using **PAVA**.
- **Pros**: Highly flexible. Can learn any non-decreasing calibration pattern. Often more accurate than Platt when we have enough data.
- **Cons**: Prone to overfitting on small datasets. Requires more data to reliably learn the function.



Calibration Methods: Inductive Venn-Abers predictors (IVAPs)

- IVAPs are a specific class of **Venn predictors** that utilize **isotonic regression** to calibrate probabilistic predictions.

- **Output**: Unlike traditional probabilistic predictors that issue a single probability, IVAPs produce **multi-probabilistic predictions** $[p_0, p_1]$.

- **Theoretical Validity**: Either p_0 or p_1 is guaranteed to be **perfectly calibrated** under the i.i.d. assumption.

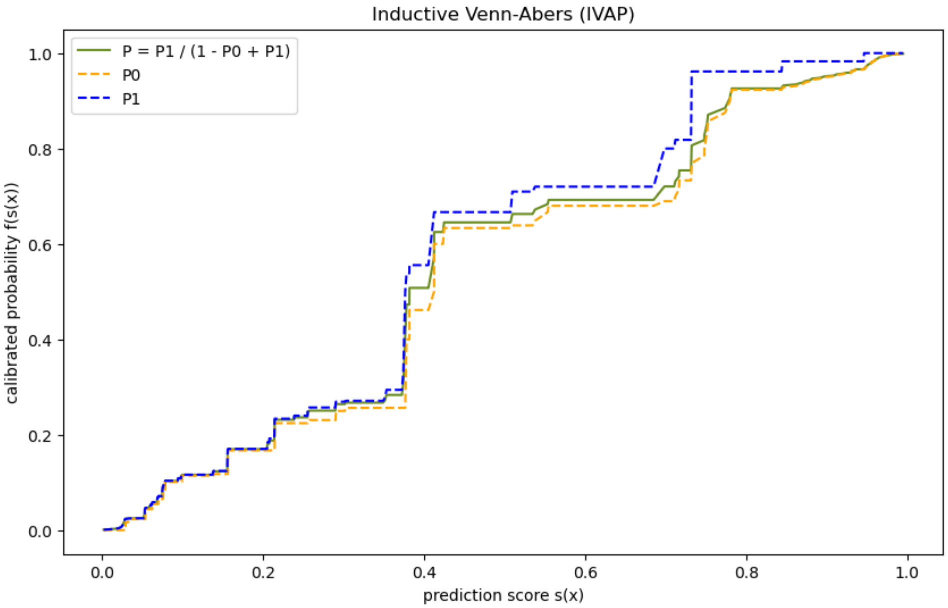
Algorithm:

1. Divide the training set of size l into two subsets, the **proper training set** of size m and the **calibration set** of size k, so that $l = m + k$.

2. Train the scoring algorithm on the proper training set

3. Find the scores $s_1, \dots, s_k$ of the calibration objects $x_1, \dots, x_k$.

4. When a new test object x arrives, compute its score s. Fit isotonic regression to $(s_1, y_1), \dots, (s_k, y_k), (s, 0)$ obtaining a function f_0. Fit isotonic regression to $(s_1, y_1), \dots, (s_k, y_k), (s, 1)$ obtaining a function f_1. The multiprobability prediction for the label y of x is the pair $(p_0, p_1) := (f_0(s), f_1(s))$.



Using *log loss* corresponding minimax probabilistic prediction p is:

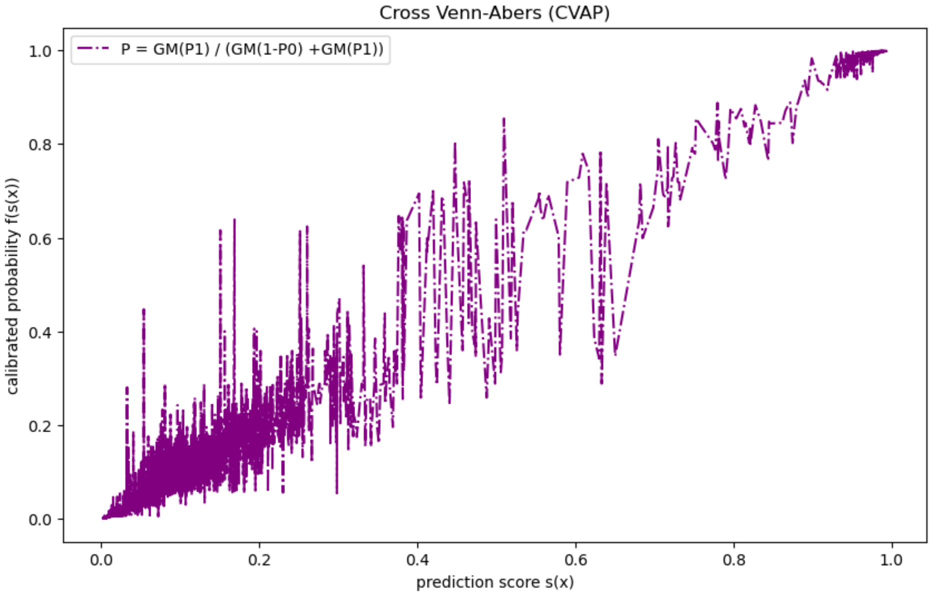
$$p = \frac{p_1}{1 - p_0 + p_1}$$

Calibration Methods: Cross Venn-Abers predictors (CVAPs)

- A CVAP is just a combination of K IVAPs, where K is the parameter of the algorithm.

Algorithm :

1. Split the training set T into K folds $T_1, \ldots, T_k$
2. for $k \in \{1, \ldots, K\}$
   $$(p_0^k, p_1^k) \coloneqq \mathrm{IVAP}(T \setminus T_k, T_k, x)$$
3. return $\frac{GM(p_1)}{GM(1-p_0)+GM(p_1)}$



Agenda

Our Dataset & Experimental Setup

Data

Credit Risk Data

- **Source**: Kaggle
- **Objective**: Binary classification task to predict customer default (Default vs. No Default).
- **Size**: 32,461 instances
- **Features**: 11 features (e.g., income, loan amount, credit history)
- **Class Balance**: 78% No Default (0); 22% Default (1)

Simulated Data

- **Source**: Beta distribution
- **Objective**: To study calibration in an idealized setting with known ground truth, we simulated 'model scores' from two distinct **Beta** distributions
- **Size**: 1,000 / 10,000 / 50,000
- **ClassBalance**: 5%, 10%, 20%, 30%, 40%, 50%

Model	Calibration Methods Applied	Interval Estimation	Discrimination Evaluation Metrics	Calibration Evaluation Metrics
Logistic Regression	• Platt's Method • Isotonic Regression • Venn-Abers	• Venn-Abers • 500 samples Bootstrap (2.5% , 97.5%)	• ROC AUC • KS statistics • Log Loss	• Log Loss • Brier Score • ECE (Expected Calibration Error)
XGBoost		• Venn-Abers		

Agenda

Credit Risk Data Preparation

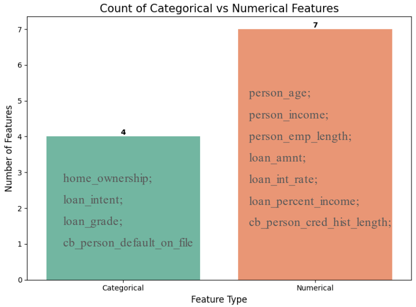
1. **EDA:** Before any data preprocessing we explored our dataset.

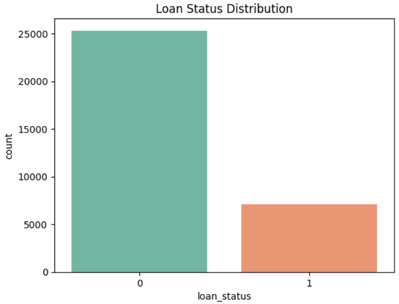
2. **Dataset Split** The dataset was first divided into three sets: training, validation, and test.

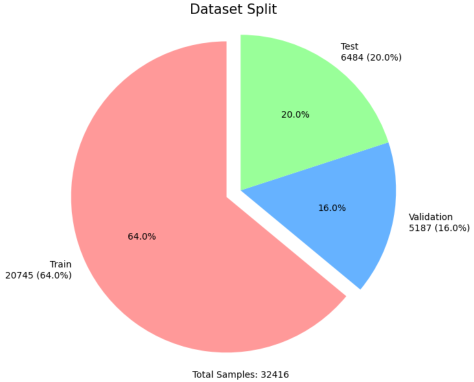
3. **Outlier Capping**: We capped outlier values in the features to reduce their impact on model performance.

4. **Missing Value Imputation**: Missing values in the dataset were handled by applying appropriate imputation methods to ensure data completeness.

5. **Target Encoding for Categorical Features** Categorical features were encoded using target encoding, where the mean of the target variable for each category was used to represent the categorical feature.



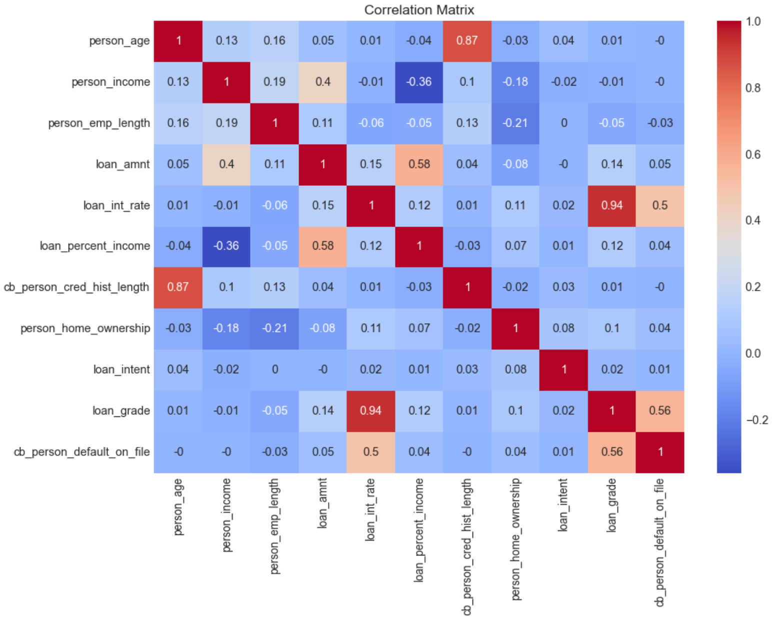




Credit Risk Data Preparation (Feature Selection)

1. **Correlation Analysis** We calculated the correlations between features and identified highly correlated ones. Features with high correlation were dropped based on their relevance and interpretability.

2. **Stability Check with PSI** To ensure feature stability, we used **Population Stability Index (PSI)** to compare feature distributions across the training and validation sets. This helped us identify features that might cause data drift or instability.



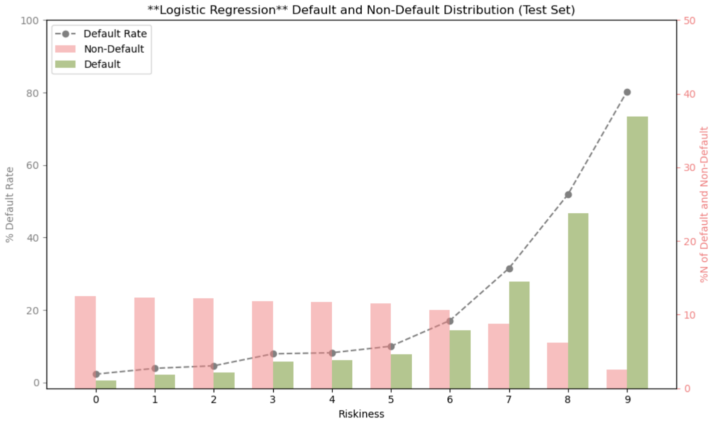
1. **Correlation exclusion:** ‘`person_age`’ , ‘`loan_int_rate`’
2. **PSI exclusion:** All features are stable!

Agenda

Results on Credit Risk Data: Accuracy

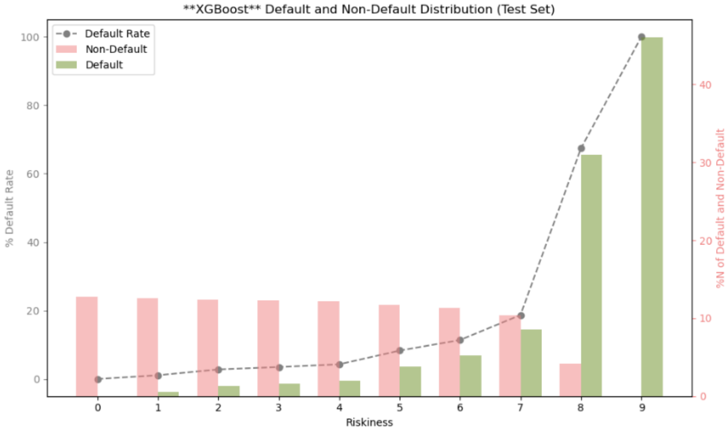
Logistic Regression

	Train	Valid	Test
ROC AUC	0.86	0.87	0.86
KS statistics	0.56	0.58	0.58
Log Loss	0.359	0.345	0.357

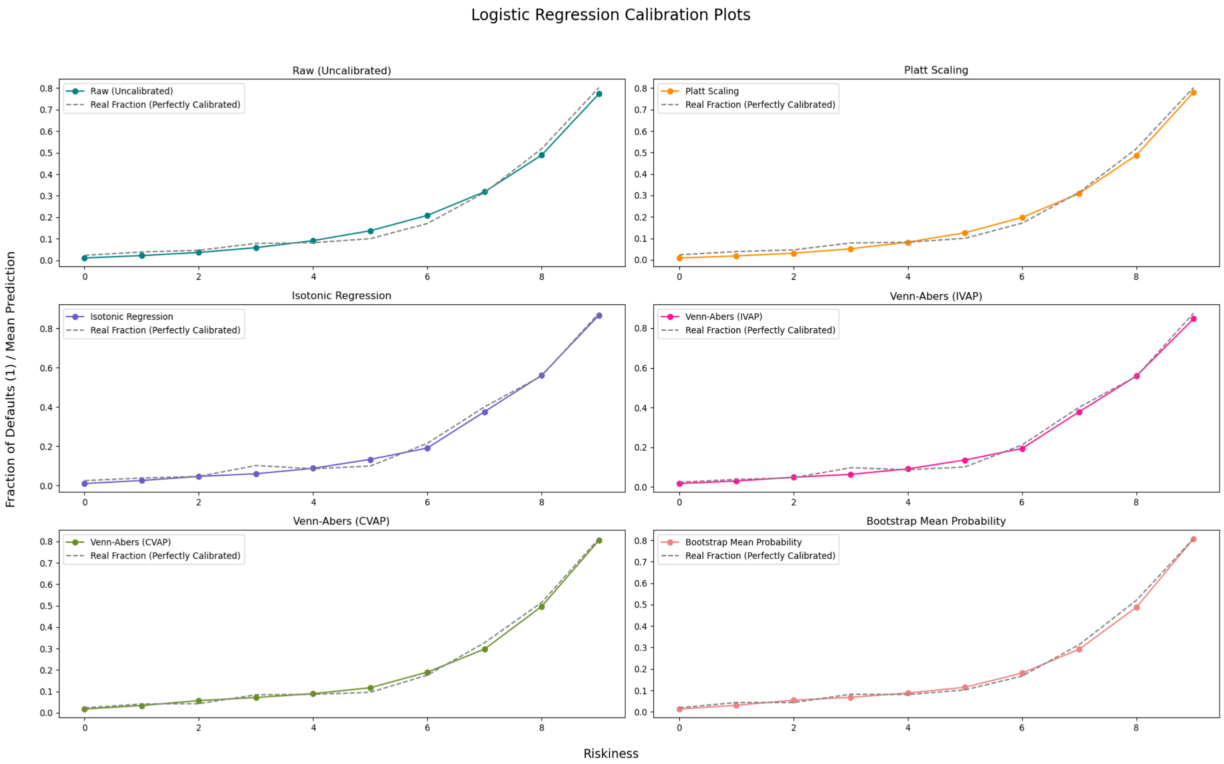


XGBoost

	Train	Valid	Test
ROC AUC	0.96	0.94	0.94
KS statistics	0.77	0.74	0.73
Log Loss	0.187	0.194	0.203



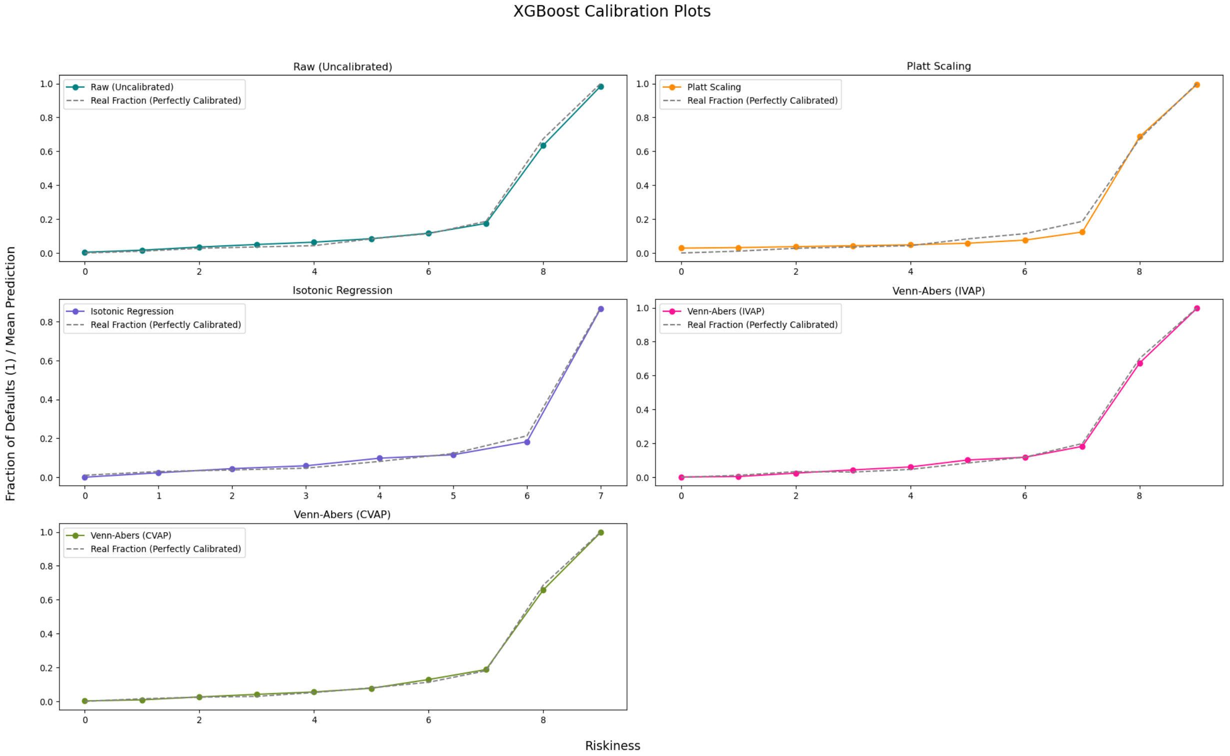
Results on Credit Risk Data: Calibration (Logistic Regression)



Valid/ CalibSet			
	ECE	Log Loss	Brier Score
Raw	0.017	0.345	0.106
Platt	0.016	0.344	0.105
Isotonic	0	0.337	0.104
IVAP	0.005	0.339	0.104
CVAP	0.012	0.343	0.105
Bootstrap	0.009	0.341	0.105

Test Set			
	ECE	Log Loss	Brier Score
Raw	0.022	0.357	0.109
Platt	0.02	0.357	0.109
Isotonic	0.015	0.385	0.109
IVAP	0.016	0.356	0.109
CVAP	0.013	0.354	0.108
Bootstrap	0.013	0.359	0.108

Results on Credit Risk Data: Calibration (XGBoost)



Valid/ CalibSet			
	ECE	Log Loss	Brier Score
Raw	0.015	0.194	0.053
Platt	0.023	0.198	0.053
Isotonic	0	0.183	0.052
IVAP	0.004	0.186	0.052
CVAP	0.019	0.172	0.048

Test Set			
	ECE	Log Loss	Brier Score
Raw	0.013	0.203	0.056
Platt	0.018	0.207	0.057
Isotonic	0.010	0.267	0.056
IVAP	0.010	0.201	0.056
CVAP	0.008	0.1990	0.056

Agenda

Simulated Data Generation

Data Generation Process

For Class 0 :

- 70% of samples: $\boldsymbol{Beta}(\boldsymbol{\alpha} = \mathbf{2}, \boldsymbol{\beta} = \mathbf{8})$ (Accurate, low-confidence predictions)
- 30% of samples: $\boldsymbol{Beta}(\boldsymbol{\alpha} = \mathbf{7}, \boldsymbol{\beta} = \mathbf{3})$ (Overconfident errors)

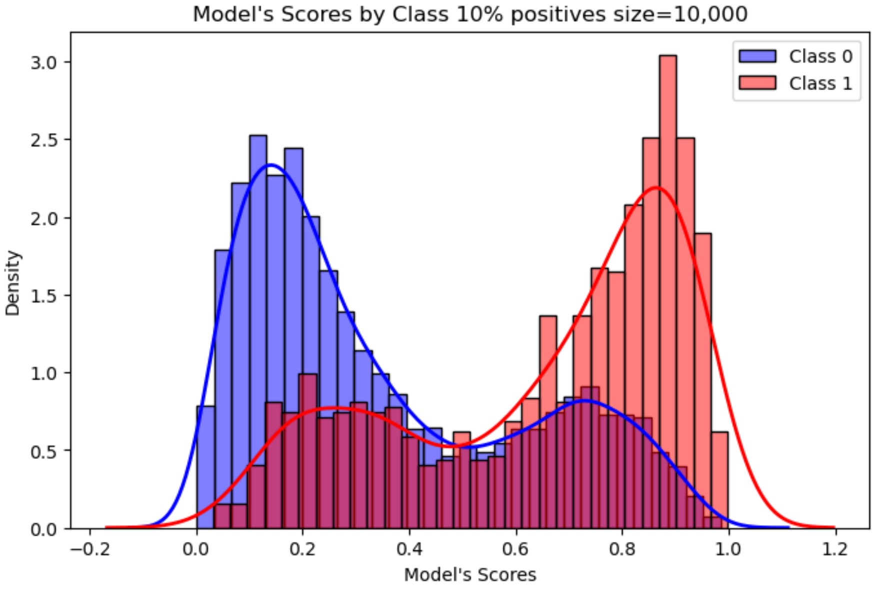
For Class 1 :

- 70% of samples: $\boldsymbol{Beta}(\boldsymbol{\alpha} = \mathbf{8}, \boldsymbol{\beta} = \mathbf{2})$ (Accurate, high-confidence predictions).
- 30% of samples: $\boldsymbol{Beta}(\boldsymbol{\alpha} = \mathbf{3}, \boldsymbol{\beta} = \mathbf{7})$ (Underconfident errors)

Simulation Scenarios

- **Dataset Size:** 1,000 / 10,000 / 50,000;
- **Class Imbalance:** 5%, 10%, 20%, 30%, 40%, 50% positive class prevalence

This resulted in a grid of **3 x 6 = 18** distinct simulation scenarios, allowing us to draw strong conclusions about the performance and reliability of **Platt Scaling**, **Isotonic Regression**, and **Venn-Abers** under a wide range of conditions.



Agenda

Results on Simulated Data

ECE				
Size = 1,000				
	RW	PS	ISO	IVAP
5%	0.309	0.015	0.015	0.015
10%	0.302	0.042	0.050	0.048
20%	0.232	0.052	0.035	0.052
30%	0.140	0.037	0.032	0.025
40%	0.083	0.059	0.070	0.066
50%	0.090	0.067	0.073	0.050

ECE				
Size = 10,000				
	RW	PS	ISO	IVAP
5%	0.312	0.008	0.008	0.008
10%	0.277	0.013	0.011	0.009
20%	0.218	0.021	0.013	0.015
30%	0.143	0.021	0.024	0.023
40%	0.075	0.016	0.012	0.011
50%	0.063	0.031	0.020	0.019

ECE				
Size = 50,000				
	RW	PS	ISO	IVAP
5%	0.314	0.005	0.004	0.004
10%	0.277	0.014	0.006	0.006
20%	0.207	0.014	0.008	0.008
30%	0.139	0.017	0.011	0.012
40%	0.073	0.026	0.013	0.012
50%	0.053	0.015	0.012	0.012

Log Loss				
Size = 1,000				
	RW	PS	ISO	IVAP
5%	0.599	0.175	0.279	0.176
10%	0.609	0.275	0.283	0.281
20%	0.566	0.398	0.401	0.405
30%	0.573	0.497	0.720	0.495
40%	0.555	0.533	0.637	0.529
50%	0.581	0.566	0.563	0.562

Log Loss				
Size = 10,000				
	RW	PS	ISO	IVAP
5%	0.562	0.176	0.176	0.176
10%	0.561	0.274	0.274	0.274
20%	0.560	0.395	0.395	0.395
30%	0.551	0.478	0.490	0.479
40%	0.574	0.547	0.549	0.549
50%	0.568	0.558	0.555	0.555

Log Loss				
Size = 50,000				
	RW	PS	ISO	IVAP
5%	0.564	0.172	0.173	0.171
10%	0.554	0.268	0.267	0.267
20%	0.557	0.406	0.405	0.405
30%	0.558	0.489	0.493	0.488
40%	0.559	0.535	0.537	0.533
50%	0.559	0.550	0.552	0.550

Results on Simulated Data

Brier Score				
Size = 1,000				
	RW	PS	ISO	IVAP
5%	0.190	0.043	0.043	0.043
10%	0.206	0.079	0.082	0.081
20%	0.190	0.124	0.125	0.126
30%	0.194	0.165	0.166	0.164
40%	0.188	0.178	0.177	0.177
50%	0.199	0.193	0.193	0.192

Brier Score				
Size = 10,000				
	RW	PS	ISO	IVAP
5%	0.191	0.044	0.044	0.045
10%	0.190	0.078	0.079	0.079
20%	0.190	0.124	0.125	0.125
30%	0.184	0.157	0.157	0.157
40%	0.195	0.184	0.185	0.185
50%	0.192	0.189	0.187	0.187

Brier Score				
Size = 50,000				
	RW	PS	ISO	IVAP
5%	0.191	0.044	0.044	0.044
10%	0.187	0.077	0.077	0.077
20%	0.188	0.128	0.128	0.128
30%	0.189	0.161	0.161	0.161
40%	0.189	0.179	0.179	0.179
50%	0.188	0.185	0.185	0.185

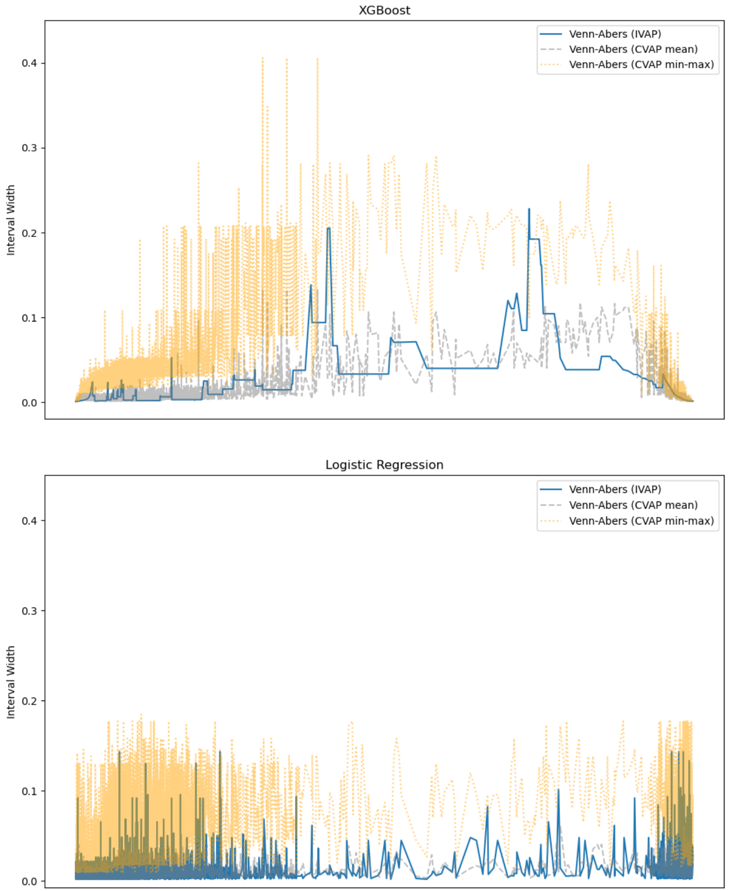
Overall Performance

- *Platt's Scaling*: ECE – **4/18**; Log Loss – **9/18**; Brier Score – **13/18**;
- *Isotonic Regression*: ECE – **12/18**; Log Loss – **6/18**; Brier Score – **11/18**;
- *Inductive Venn-Abers*: ECE – **12/18**; Log Loss – **13/18**; Brier Score – **13/18**;

Agenda

Probability Intervals

Model	Method	Average Interval Width
XGBoost	IVAP	0.0067
XGBoost	CVAP (Mean)	0.0069
Logistic Reg	IVAP	0.0080
Logistic Reg	CVAP (Mean)	0.0081
XGBoost	CVAP (Min, Max)	0.0333
Logistic Reg	CVAP (Min, Max)	0.0514
Logistic Reg (Bootstrap)	(2.5%, 97.5%)	0.0627



Agenda

Conclusion & Key Takeaways

1. **Calibration is a Crucial, Separate Property from Accuracy:** a model can be highly accurate yet poorly calibrated, making its probability scores misleading and untrustworthy for real-world risk assessment. Explicitly measuring and improving calibration is essential for any application relying on probabilistic predictions.

2. **The Best Calibration Method is Context-Dependent:**
   - **Platt Scaling:** A good, fast default for simple models like Logistic Regression.
   - **Isotonic Regression:** Powerful but can overfit on small datasets. Excellent for larger, well-behaved datasets.
   - **Venn-Abers Predictors:** Provide robust, distribution-free calibration guarantees and are highly competitive, especially on complex models like XGBoost. They offer a unique advantage: inherently valid probability intervals.

3. **For Precise Uncertainty Quantification, Venn-Abers is Superior:** Our results demonstrate that Venn-Abers predictors (IVAP and CVAP mean) generate prediction intervals that are significantly tighter than traditional bootstrap methods, while maintaining validity. This makes them ideal for applications requiring precise uncertainty estimates.

Agenda

References

- Dataset - https://www.kaggle.com/datasets/laotse/credit-risk-dataset
- Vovk, Vladimir, Ivan Petej. "Venn-Abers predictors". Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence (2014) (arxiv version https://arxiv.org/pdf/1211.0025)
- Vovk, Vladimir, Ivan Petej and Valentina Fedorova. "Large-scale probabilistic predictors with and without guarantees of validity." Advances in Neural Information Processing Systems 28 (2015) (arxiv version https://arxiv.org/pdf/1511.00213)
- Model Calibration, Explained- https://towardsdatascience.com/model-calibration-explained-a-visual-guide-with-code-examples-for-beginners-55f368bafe72/
- Venn-Abers implementation - https://github.com/ip200/venn-abers

Thank you for the attention!

Sep 2025