Beyond the Match
Enhancing Product Matching with Model Calibration

Wayfair Advisors: Rajesh Bawa, Yuxin Chen
MIT Advisors: Prof. Georgia Perakis, Leann Pearl Geetha Thayaparan

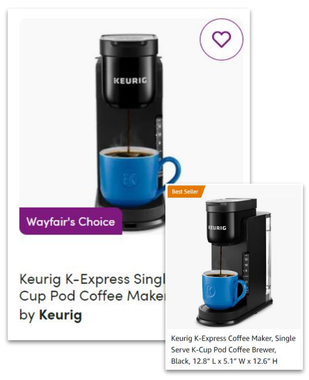

Claire Guan **Sophie Long**

Problem Statement

Context: Wayfair is a leading online retailer in the home goods industry with **25M** products across **1600+** categories. They own many classification models where the predicted probability needs to be accurate calibrated. A key use case is the product matching model which identifies matches between Wayfair's products and competitors' products to optimize pricing strategy.

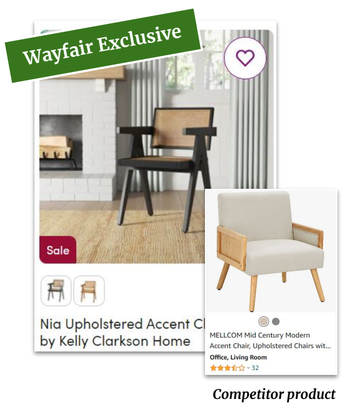
The main objective of this project is to **develop a generalizable model calibration framework** for better robustness & efficiency.

- Enhance the **credibility** of model predictions
- Alleviate frequent model retraining challenges, which can take up to **3** months
- Opportunities to support **22+** production models driving **$50M** annual revenue



Competitor product

Score: 96%



Competitor product

Score: 23%

Dataset

We compiled our data by extracting from various internal data tables, including matching pairs pool, model results, and agents' review data using Google Cloud Platform.

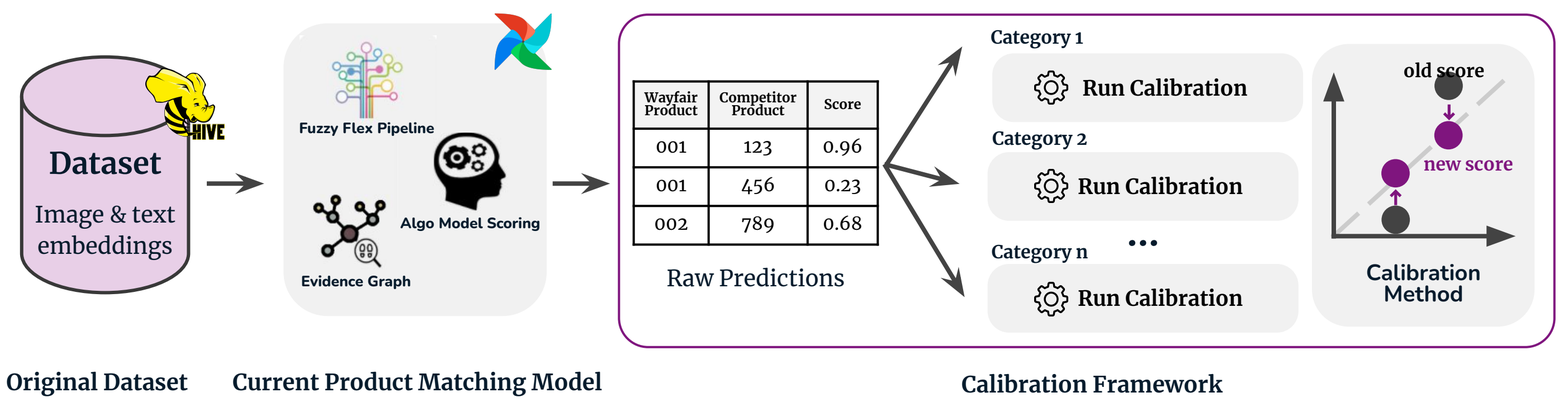
Product ID	Competitor Product ID	Category	Model prediction	Groundtruth
001	123	Clocks	0.96	Match
001	456	Clocks	0.23	No Match
002	789	Beds	0.68	Match

➡ Evaluated by offshore agents to provide the ground truth label

25M Wayfair Products
400M Competitor Products
1600+ Product Categories

Methodology

The focus is to develop a generic model calibration framework to improve the reliability and accuracy of predictions from different machine learning models. Hence, we conducted extensive research on various calibration methods, and implemented **four** efficient calibration methods for the binary classification scenarios. In addition, we designed **5 classifier metrics** and **6 calibration metrics** to evaluate these models, where **Accuracy**, **AUC**, and **Binary-ECE** are the most important.



Original Dataset **Current Product Matching Model** **Calibration Framework** **End User Scenario**

Calibration Methods
(Multiclass methods are detailed in report)

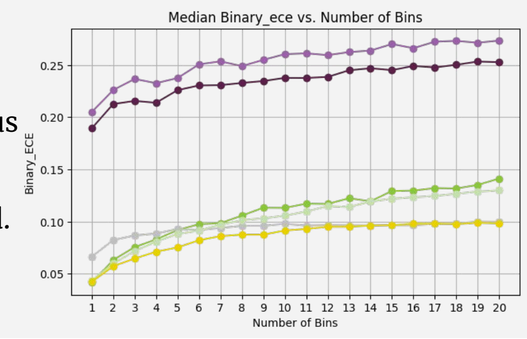
Experimentation

1 Determining data granularity

We experimented with two training variants: (a) training separate models for each product category, (b) training a single model using the entire dataset. Testing on granular category data and the entire dataset showed that calibration using each product category data was the most effective approach.

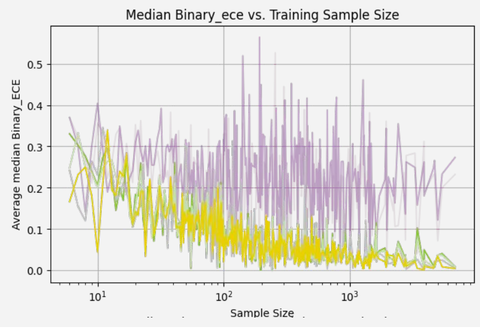
2 Determining optimal bin size

As calibration metrics are sensitive to the number of bins used, we tested various bin sizes (1 to 20) and observed that as bin size increases, errors increased. The optimal bin size is 7 where an elbow occurs.

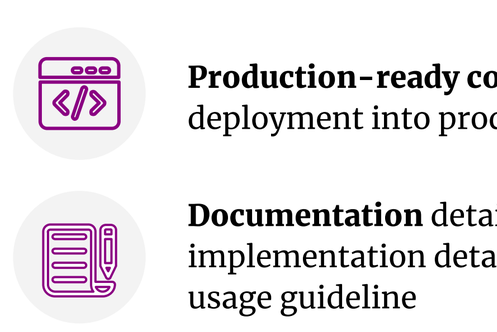


3 Determining minimum sample size

We assessed the relationship between training sample size and Binary ECE. Larger sample size leads to reduced errors. The min sample size is 100 for isotonic calibrator.



Results



Methods	Accuracy	AUC	Binary ECE
Uncalibrated	0.50	0.78	0.33
Current Calibration	0.68	0.76	0.25
Sigmoid Calibration	0.76	0.76	0.16
Beta Calibration	0.76	0.77	0.15
BBQ Calibration	0.70	0.50	0.15
Isotonic Calibration	**0.76**	**0.75**	**0.13**

Model trained with per-category data (7 bins). The reliability diagram (left) compares predicted probabilities (x) against observed frequencies (y) where diagonal line represents perfect calibration.

- ➔ The isotonic calibrator consistently achieves the best calibration performance, reducing **49%** of the Median Expected Calibration Error (Binary ECE) while maintaining **0.75** AUC compared to current baseline.
- ➔ The different calibration models have consistent performance across multiple calibration metrics, making it easier to choose the best model.

Recommendations

We provide the following recommendations to stakeholders

- Calibrate match scores separately for each product category for better performance.
- For categories with limited training samples (<100), use a calibration model trained on the entire dataset.
- Use the Binary ECE threshold of 0.1 to achieve well-calibration with match rate of 80%.

Other Applications

Prevent fraudulent transactions to safeguard Wayfair's operations and customers' shopping environment

Automate B2B customer identification to optimize business potential and revenue growth

Select the lead image of products to enhance customer engagement and visual appeal

Predict customer intent through text and audio transcriptions to provide guided solutions to agents

Deliverables

Methodology presentation outlining the key approaches used in the project

Production-ready codebase for easy deployment into production pipeline

Documentation detailing the implementation details and usage guideline

Impact

Calibration Error Reduced by
49%
compared to baseline

Approximate expected lift of
$15M
in annual revenue

Opportunities to support over
22
production models

Next Steps

- **Multi-class Calibration:** Explore new use cases and extend to multi-class models for broader applicability.
- **Real-time calibration**: Investigate methods to maintain accuracy amid dynamic data changes.
- **Optimize feature engineering:** Enhance model performance by selecting relevant features from the original raw dataset utilized for training ML models.