**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 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 the focus is to develop a generic model calibration framework to improve the reliability and accuracy of predictions from different machine learning models.

Wayfair’s products and competitors’ products to optimize pricing strategy.

**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 them for different machine learning models.

**Experimenation**

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 increase. 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**

- **Reliability Diagram: Model Comparison**
  - 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**
  - 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.

**Impact**

- **Calibration Error Reduced by 49% compared to baseline**
  - **Approximate lift of $15M in annual revenue**
  - **Opportunities to support over 22 production models**

**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

**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.