Model Evaluation for Product Performance

Evaluate your AI models to ensure they meet product quality and deliver tangible business value. Learn metrics, experimentation and analytics techniques to make data-driven decisions.

July 30, 2025

Evaluation Metrics

Measure model performance using multiple dimensions.

Metric	What it measures
Accuracy	Overall proportion of correct predictions
Precision	True positives among all predicted positives
Recall	True positives among all actual positives
F1 Score	Harmonic mean of precision and recall

Accuracy
How many predictions were correct?

Precision
Focuses on relevance of positive predictions.

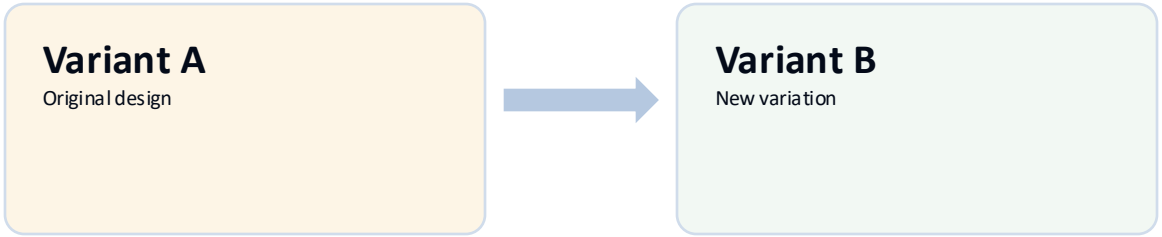
Recall
Focuses on completeness of positive predictions.

F1 Score
Balances precision and recall.

[1] [2]

A/B Testing

Compare multiple variations to learn what drives engagement.

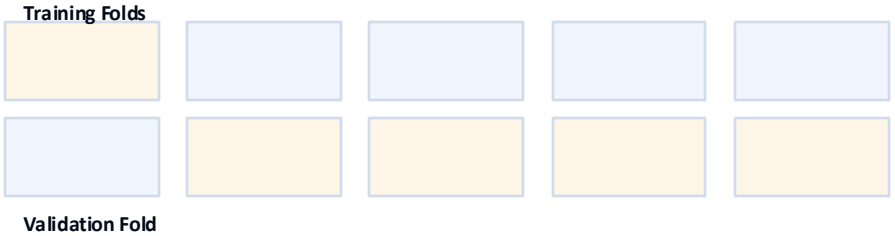


- Split real users into random groups – each sees a different variant.
- Variants differ in one design element (e.g., button text).
- Collect metrics like conversion or click-through rate to determine winners.
- Choose the winning design only if results are statistically significant.

[3] [4] [5]

Cross-Validation

Robustly estimate model performance and avoid overfitting.

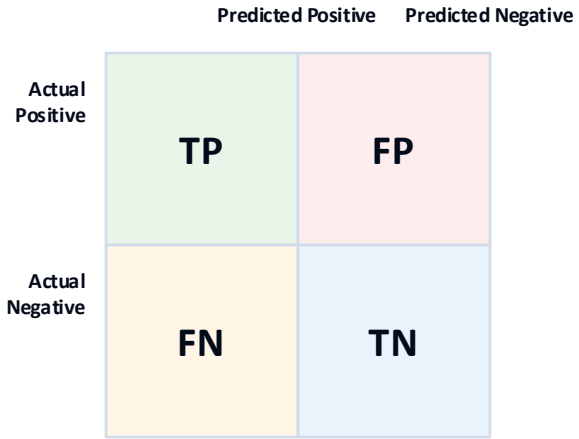


- Split data into k equal folds; iteratively train on k-1 folds and validate on the remaining fold.
- Average results across folds for a more accurate performance estimate.
- Use scikit-learn's cross_val_score or cross_validate to automate evaluation with multiple metrics.

[6] [7] [8]

Confusion Matrices

Visualise errors to fine-tune your classifier.



- TP: model correctly predicted a positive case.
- FP: model predicted positive but it was actually negative.
- FN: model predicted negative but it was actually positive.
- TN: model correctly predicted a negative case.

[9] [10]

Product Analytics & Tools

Track real user behaviour and measure feature impact.

Scikit-learn

Open-source library built on NumPy and SciPy.
Provides cross_val_score, cross_validate and a suite of metrics (accuracy, precision, recall, F1).

Mixpanel

Product analytics platform for tracking in-app behaviour.
Event tracking, segmentation, funnels and retention analysis help you connect user actions to model impact.

[11] [12] [13] [14] [15]

Defining Success Metrics

Align model performance with business goals.

Why it matters

Success metrics quantify how well a product meets its goals and user needs.
Monitoring metrics drives data-driven decisions, optimises user experience and ensures ROI aligns with product goals.

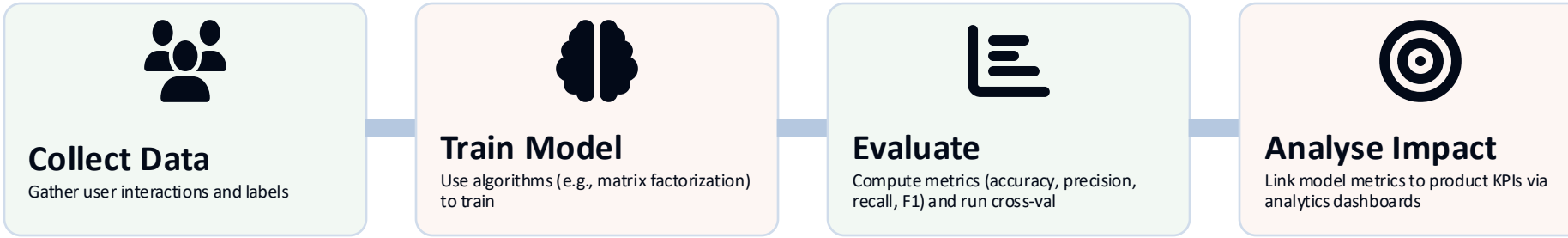
How to define

- Identify business goals using frameworks like OKR or SMART.
- Select North Star, HEART or AAARRR metrics to track across the user journey.
- Leverage no-code analytics tools and custom dashboards for visualisation.
- Use acquisition, activation, adoption, retention, referral and revenue metrics to evaluate business impact.

[16] [17]

Let’s Build!

Apply evaluation techniques to a recommender system.



We’ll implement and evaluate a recommendation model using scikit-learn and analyse its effect on product engagement with Mixpanel.
Collect data → Train model → Evaluate metrics → Analyse business impact.

[18] [19] [20]