CASE STUDY ·PRODUCT MANAGEMENT

Replacing a Third-Party Face Match Service with an In-House AI Solution

A 3-month initiative to migrate a production-critical KYC face verification system from AWS Rekognition to a custom in-house model — reducing costs by **80%**, improving recall, and establishing full ownership of a core identity verification capability.

EXECUTIVE SUMMARY

The Challenge & The Outcome

Our digital onboarding platform relied on AWS Rekognition to verify that a customer's selfie matched the face on their government-issued ID. While the solution worked reliably, it introduced recurring infrastructure costs, limited our ability to improve the model, and created vendor dependency for a critical part of the onboarding journey.

80%

Cost Reduction

Reduction in face matching infrastructure costs after migrating to an in-house solution

+2%

Recall Improvement

Improvement in verification recall — more genuine users successfully completing onboarding

<1s

End-to-End Latency

Total verification latency maintained within internal performance SLA

3

Months to Production

Full cross-functional delivery from discovery to global rollout

The project successfully reduced face matching costs by 80%, improved verification performance, and enabled the organization to fully own and continuously improve its face verification capabilities.

THE PROBLEM

Why the Existing Solution Wasn't Enough

Face Match is one of the most critical steps during digital KYC. Every onboarding journey depends on correctly verifying that the person taking the selfie is the same individual shown on the identity document. The AWS Rekognition-based solution presented several business challenges that limited our ability to scale and innovate.

Business Constraints

- High recurring inference costs at scale
- Limited control over model improvements
- Vendor dependency for a core verification capability
- Slower response to emerging fraud patterns
- Limited flexibility to customize verification logic

Why It Mattered

Since face verification directly impacts onboarding success rates and fraud prevention, replacing the solution required maintaining production-grade reliability while reducing cost. The goal was to build an in-house solution that delivered better performance, lower operational expense, and full ownership of a critical capability.

DISCOVERY

Understanding Real Production Failures

Rather than immediately building a new model, the first step was understanding *why* the production system failed. We extracted historical predictions from AWS Rekognition and manually audited thousands of verification pairs. This analysis revealed recurring failure patterns that became the foundation for improving both the training dataset and the overall verification pipeline.

Age Gaps

Large differences between ID photo age and current selfie appearance caused mismatches.

Image Quality

Blurry, low-quality, or poorly lit ID images reduced match confidence significantly.

Rotated Documents

Rotated document images broke the detection pipeline before comparison could occur.

Similar Faces

Visually similar individuals were incorrectly classified as matches, increasing false positives.

Background Faces

Portraits or secondary faces on ID cards were detected instead of the primary subject.

Improper Crops

Inconsistent face cropping removed critical facial context needed for accurate matching.

PRODUCT STRATEGY

Data Quality Over Model Architecture

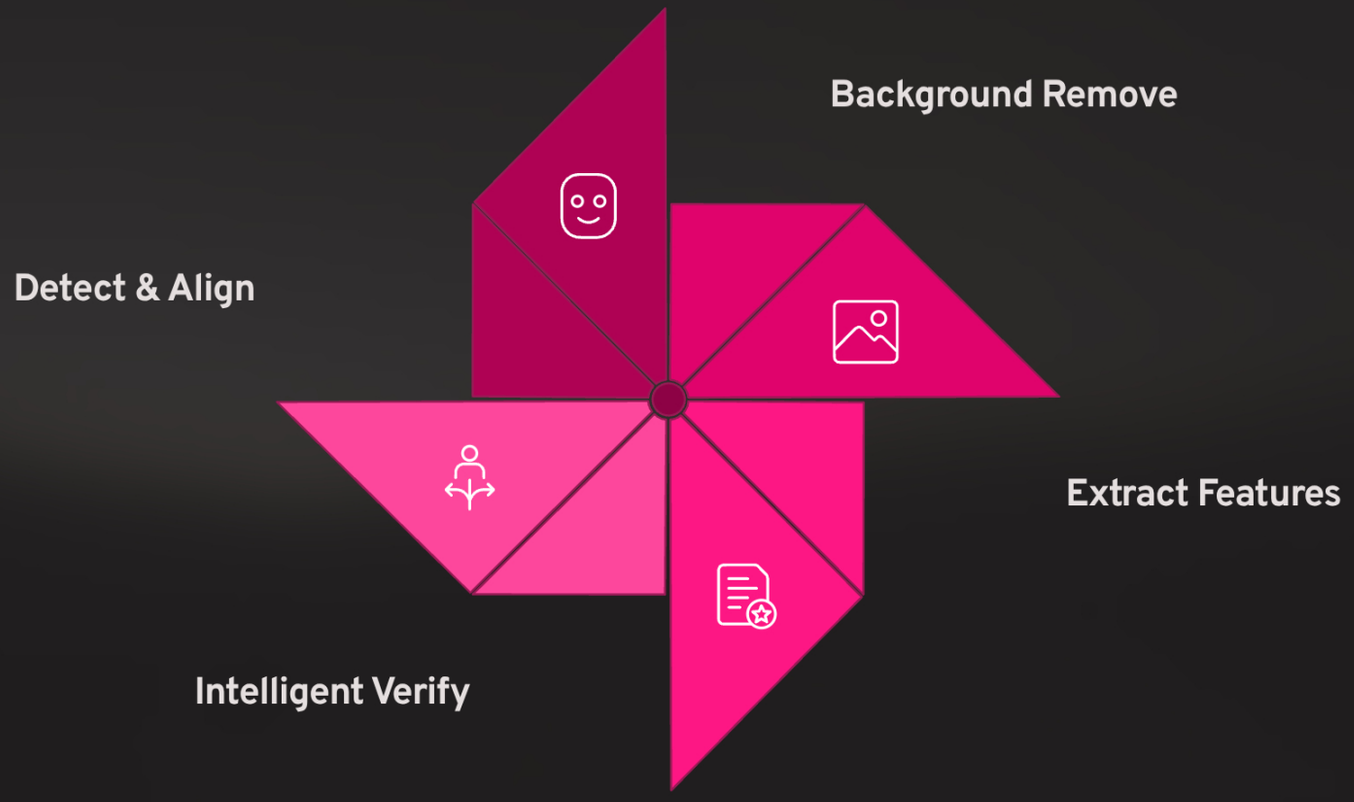
One of the most consequential decisions in this project was prioritizing **training data quality** over model architecture changes. Three complementary data strategies were deployed to create a production-representative dataset.

1	2	3
Production Data Collection	**Hard Negative Generation**	**Similar Face Dataset**
Face pairs were extracted from recorded Video KYC sessions — including selfie vs. ID image (genuine matches) and agent image vs. customer ID (genuine non-matches) — creating realistic, production-like training data.	Non-match pairs were generated by swapping identities across different users. These "hard negatives" closely resembled production traffic and helped the model learn to reject incorrect matches, reducing false positives.	Difficult non-match examples containing visually similar individuals were intentionally created. Training on these cases improved the model's ability to distinguish between people with similar facial characteristics.

SOLUTION DESIGN

A Hybrid Verification Pipeline

Instead of depending solely on facial similarity scores, we designed a **four-stage hybrid verification pipeline** that combined facial embeddings with image quality signals to make more robust verification decisions across real-world onboarding scenarios.



The final verification stage considered both facial embeddings and image quality signals — blur, brightness, and head pose — rather than relying on a simple similarity threshold. This multi-signal approach significantly improved robustness across challenging real-world onboarding scenarios including poor lighting, partial occlusion, and document quality variation.

ITERATION

Continuous Improvement Through Production Feedback

The project followed an iterative development cycle rather than a single model release. Each production failure informed the next model version, creating a self-reinforcing feedback loop between live traffic, dataset improvements, and model retraining.

The Iteration Loop

01

Production Analysis

Identify failure patterns from live traffic

02

Dataset Improvement

Augment training data to cover edge cases

03

Model Training & Validation

Retrain and evaluate against holdout sets

04

Controlled Rollout & Monitoring

Deploy gradually and track production metrics

Key Iterative Improvements

- **Orientation Detection:** An orientation detection model automatically corrected rotated document images before verification
- **Improved Face Cropping:** Increased crop margins preserved more facial context, improving accuracy on partially visible faces
- **Diverse Training Data:** Expanded dataset to include hijabs, turbans, beards, laminated cards, blur variations, lighting differences, and age gaps

EVALUATION & TRADE-OFFS

Model Performance & The Latency Decision

Offline evaluation showed that the in-house solution achieved higher recall while maintaining production-grade precision. The marginal decrease in precision was an acceptable trade-off given the meaningful improvement in genuine user verification success rates.

Metric	AWS Rekognition	In-House Model
Precision	99.75%	99.57%
Recall	83.05%	**85.66%**

The Latency Trade-off

The enhanced pipeline introduced additional preprocessing — background removal, image quality analysis, and metadata extraction — increasing average API response time by approximately **200 milliseconds**.

Why We Accepted It

- End-to-end verification latency remained **under one second**, within our internal performance SLA
- The latency increase was outweighed by significant improvements in model performance and operational cost
- Internal testing confirmed no noticeable impact on the overall onboarding experience

ROLLOUT

Phased Deployment to Minimize Production Risk

Given that face verification is a critical onboarding capability, we adopted a phased rollout approach to minimize production risk. Continuous monitoring and rollback mechanisms remained in place throughout deployment.

1

Phase 1 — Pilot

Enabled for a small set of enterprise customers. Monitored verification success rate, precision, recall, API latency, error rates, and production stability.

2

Phase 2 — Validation

Production metrics compared against the legacy AWS solution. Confirmed latency within SLA, improved verification performance, and maintained system stability.

3

Phase 3 — Global Rollout

After meeting predefined success criteria, the solution was gradually expanded to all customers with continuous monitoring and rollback mechanisms in place.

All three phases completed without disrupting customer onboarding. The migration was fully successful with zero production incidents attributed to the new system.

RESULTS & LEARNINGS

Business Impact & Key Takeaways

The project successfully achieved its primary business objectives across financial, product, and operational dimensions — while establishing a repeatable model improvement loop for the future.

Financial Impact

80% Cost Reduction

Migrated from a third-party service to an in-house solution, eliminating recurring inference costs at scale

Product Impact

- Improved verification recall while maintaining production-grade precision
- Better handling of age variation, blur, lighting differences, and rotated documents
- Reduced dependency on external vendors
- Faster ability to iterate based on production feedback

Operational Impact

- Successfully migrated a production-critical service without disrupting onboarding
- Established a repeatable feedback loop between production failures, dataset improvements, and model retraining

Key Learnings

Data Quality Drives Outcomes

The biggest performance gains came from improving training data, not changing model architectures.

Balance Three Metrics

Effective AI product management requires balancing accuracy, cost, and user experience — not optimizing a single metric.

Accept Thoughtful Trade-offs

The 200 ms latency increase was accepted because it stayed within SLA and delivered substantial long-term business value.

Validate in Production

Controlled experiments and phased rollouts de-risked a critical migration and built stakeholder confidence.