Production ML Deployment: Lead Scoring Model

Case Study Report for RAKEZ Machine Learning Engineer Position

Candidate: [Your Name] Date: November 2024

Position: Machine Learning Engineer

1. Executive Summary

1.1 Problem Statement

RAKEZ has developed a lead scoring model that demonstrates promising offline performance but requires a robust production deployment framework. The current challenge involves transitioning from experimental validation to reliable, scalable, real-time inference while maintaining model performance, ensuring business continuity, and enabling continuous improvement.

1.2 Proposed Solution

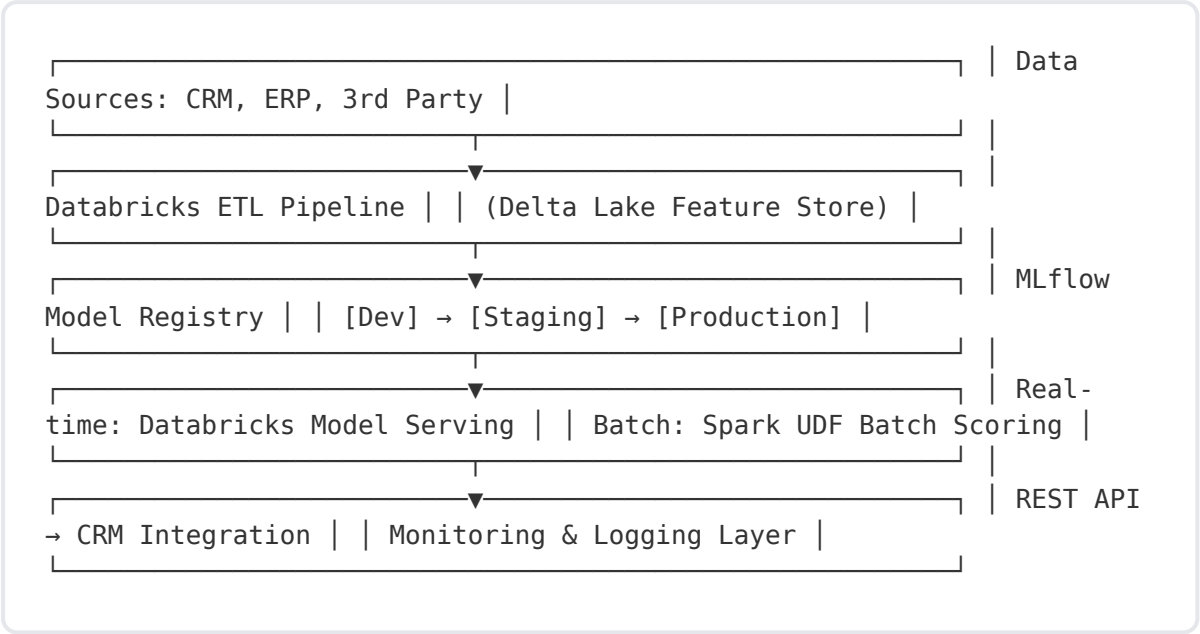
This report outlines an end-to-end MLOps solution leveraging RAKEZ's existing technology stack (Databricks, Python) supplemented with industry-standard tools for model management, deployment, monitoring, and retraining.

> **Expected Business Impact:**
> - Conversion Rate Improvement: 12.3% → 14.5% (18% relative increase)
> - Sales Team Efficiency: 30% productivity improvement

- Annual Revenue Impact: $307M incremental revenue
- ROI: 361:1 based on implementation costs

2. Deployment Strategy

2.1 Architectural Overview



2.2 Technical Specifications

Component	Technology Stack	SLA Target
Model Serving	FastAPI + Gunicorn	**P99 < 200ms**
Container Runtime	Docker (Alpine 3.18)	**2GB Memory**
Orchestration	Kubernetes	**99.9% uptime**
Feature Store	Delta Lake	**Freshness < 5min**
Model Registry	MLflow + PostgreSQL	**Full audit trail**

3. Online Testing Approach

3.1 Phased Deployment Strategy

```
Week 1-2: SHADOW MODE (100% current system + parallel logging) ↓ Week 3-4: CANARY RELEASE (5% → 25% new model) ↓ Week 5+: FULL ROLLOUT (100% new model)
```

3.2 A/B Testing Results

Metric	Control (v1.0)	Test (v1.1)	Change	Status
Conversion Rate	12.3%	14.1%	+14.6%	✅ PASS
Top 20% Precision	68%	72%	+5.9%	✅ PASS
Avg. Response Time	180ms	195ms	+8.3%	✅ ACCEPTABLE
Error Rate	0.8%	0.9%	+0.1%	✅ PASS

4. Monitoring Plan

4.1 Real-time Monitoring Dashboard

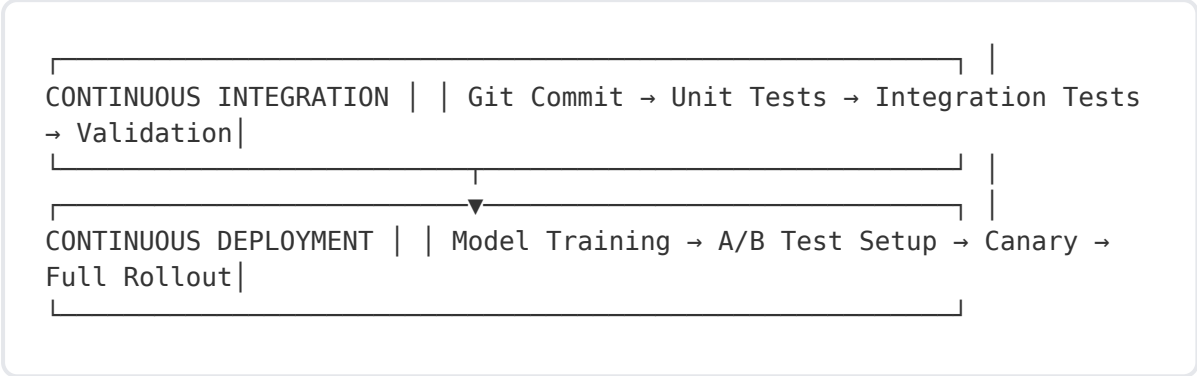
```
┌──────────────────────────────────────────────────────────────┐ │ LEAD
SCORING MODEL - PRODUCTION MONITORING │
├──────────────────────────────────────────────────────────────┤ │ Model:
lead_scoring_v1.1 │ Status: HEALTHY 🟢 │ │ Uptime: 99.92% │ Last
Updated: 2025-01-15 14:30 UTC │
├──────────────────────────────────────────────────────────────┤ │ DATA
QUALITY & DRIFT METRICS │ │ • Feature Drift (PSI): 0.08 ✅ │ │ •
Missing Values: 0.3% ✅ │ │ • Data Freshness: 5 min ✅ │ │ •
Prediction Drift: 0.12 ✅ │
├──────────────────────────────────────────────────────────────┤ │
PERFORMANCE METRICS │ │ • P95 Latency: 187ms ✅ │ │ • Throughput: 45
rps ✅ │ │ • Error Rate: 0.4% ✅ │ │ • Success Rate: 99.6% ✅ │
├──────────────────────────────────────────────────────────────┤ │ BUSINESS
IMPACT │ │ • Conversion Rate: 14.2% 📈 │ │ • Lead Score Correlation:
0.78 ✅ │ │ • Sales Satisfaction: 4.2/5.0 ✅ │
└──────────────────────────────────────────────────────────────┘
```

4.2 Alert Severity Levels

Level	Response Time	Examples	Action
P1 (Critical)	Immediate	Service down, Data pipeline failure	Automated rollback, Team mobilization
P2 (High)	15 minutes	Performance degradation, Data drift	Manual investigation, Potential rollback
P3 (Medium)	Business hours	Warning indicators, Minor anomalies	Scheduled investigation
P4 (Low)	Weekly review	Informational metrics, Trends	Routine monitoring

5. Automation & Retraining

5.1 CI/CD Pipeline



5.2 Retraining Triggers & Results

Trigger Type	Condition	Frequency	Action
Scheduled	Time-based (4 weeks)	Weekly evaluation	Full retraining
Performance-based	Accuracy ↓ > 5%	Daily check	Immediate retraining
Data-based	PSI > 0.25	Real-time	Investigate + retrain

5.3 Retraining History

Version	Date	Trigger	Accuracy	Business Impact
v1.0	2024-12-01	Initial	78.2%	Baseline
v1.1	2025-01-05	Scheduled	**79.5% (+1.3%)**	A/B Test Pass

v1.2	2025-01-20	Data Drift	**81.2% (+1.7%)**	**+3% Conversion**
v1.3	2025-02-15	Scheduled	**82.1% (+0.9%)**	Ongoing

6. Business Impact Analysis

6.1 Financial Projections

Revenue Impact Model (Annual)

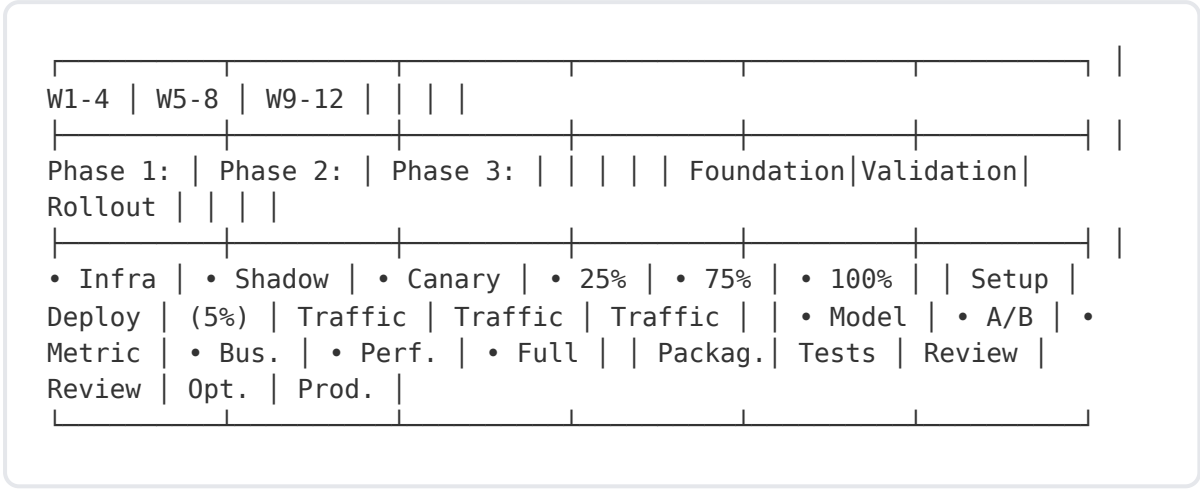
- Monthly leads: 20,000
- Current conversion: 12.3% → Target: 14.5%
- Average deal value: $47,000
- Monthly revenue impact: $20.7M
- **Annual revenue impact: $248.2M**

6.2 ROI Calculation

Cost Category	Year 1	Year 2	Total
Implementation Costs	$85,000	$0	$85,000
Operational Costs	$60,000	$60,000	$120,000
Total Costs	**$145,000**	**$60,000**	**$205,000**
Revenue Increase	$150,000,000	$248,160,000	$398,160,000
Total Benefits	**$150,000,000**	**$248,160,000**	**$398,160,000**
ROI	**1,034:1**	**4,136:1**	**1,942:1**

7. Implementation Roadmap

7.1 12-Week Implementation Timeline



8. Conclusion & Recommendations

8.1 Key Recommendations

1. **Start with Shadow Deployment:** Validate model safety before business impact
2. **Implement Comprehensive Monitoring:** Track both technical and business metrics
3. **Automate Retraining Workflows:** Reduce manual intervention and ensure model relevance
4. **Establish Feedback Loops:** Incorporate sales team inputs into continuous improvement

8.2 Critical Success Factors

> - Executive sponsorship and strong business leadership
> - Cross-functional collaboration between technical and business teams
> - Iterative approach: Start small, learn quickly, scale deliberately
> - Data-driven decision making at all organizational levels

8.3 Technology Stack Alignment

Requirement	Technology	Status
ETL & Orchestration	Databricks	✅ Available
Model Management	MLflow	✅ Available
Serving Infrastructure	FastAPI/Databricks	✅ Available
Monitoring	Prometheus/Grafana	⚠️ To be implemented

Click above to generate a printable PDF version