# RAKEZ Lead Scoring Model - Deployment & Monitoring ## 10-Slide Presentation

## Slide 1: Title Slide **RAKEZ Lead Scoring Model** **End-to-End ML Engineering Solution** **Deployment, Monitoring & Operations**

## Slide 2: Problem Summary ### Business Challenge - **Need**: Prioritize leads to maximize sales team efficiency - **Current State**: Manual lead qualification, inconsistent prioritization - **Goal**: Automated lead scoring with real-time predictions ### Technical Requirements - Real-time scoring API (< 200ms latency) - Continuous monitoring and drift detection - Automated retraining when performance degrades - Seamless CRM integration - Production-grade reliability (99.9% uptime) ### Success Metrics - **Conversion Rate**: Increase by 15% - **Sales Efficiency**: Reduce time-to-contact by 30% - **Model Performance**: Maintain AUC > 0.85

## Slide 3: End-to-End Architecture ### System Components

■ **Architecture Diagram**

*See ARCHITECTURE_DIAGRAMS.md for visual version*

**■ Full Architecture Diagrams**: See `06_docs/ARCHITECTURE_DIAGRAMS.md` for complete visual documentation. ### Key Technologies - **Databricks**: Data processing and batch inference - **MLflow**: Model versioning and registry - **FastAPI**: Real-time scoring API - **Plotly Dash**: Monitoring dashboard - **Delta Lake**: Data storage and versioning

## Slide 4: Deployment Plan ### Deployment Strategy **Phase 1: Model Registry Setup** - Register initial model to MLflow - Set up Production and Staging stages - Configure model metadata and tags **Phase 2: Batch Inference** - Deploy Databricks job for batch scoring - Schedule daily runs at 2 AM - Update CRM with lead scores **Phase 3: Real-time API** - Deploy FastAPI service - Load model from MLflow registry - Enable shadow model for testing **Phase 4: Monitoring** - Deploy monitoring jobs (drift detection, metrics) - Set up Streamlit dashboard - Configure alerting (Slack, Email) ### Deployment Methods - **Canary Deployment**: 10% → 50% → 100% traffic - **Shadow Deployment**: Parallel evaluation of new models - **Rollback Mechanism**: One-click reversion to previous version ### Timeline - Week 1: Infrastructure setup - Week 2: Batch inference deployment - Week 3: Real-time API deployment - Week 4: Monitoring and alerting setup

## Slide 5: Online Testing Strategy ### A/B Testing Framework **Traffic Splitting** - Production Model: 90% traffic - New Model (Staging): 10% traffic - Compare performance metrics **Evaluation Metrics** - Conversion rate by model - Revenue per lead - Sales team feedback - Model performance (AUC, Precision, Recall) **Decision Criteria** - New model must show: - +2% AUC improvement OR - +5% business KPI improvement - No increase in error rate - No latency degradation ### Shadow Model Deployment **Silent Evaluation** - Deploy new model alongside production - Route all traffic to both models - Compare predictions without affecting users - Monitor for 1-2 weeks before promotion **Benefits** - Zero-risk evaluation - Real-world performance data - Gradual confidence building ### Testing Schedule - **Week 1-2**: Shadow deployment - **Week 3**: A/B test (10% traffic) - **Week 4**: Full promotion if successful

## Slide 6: Monitoring & Alerting ### Key Metrics **Performance Metrics** - **Latency**: P50, P95, P99 (Target: P95 < 200ms) - **Throughput**: Requests per second (Target: > 10 req/s) - **Error Rate**: HTTP errors (Target: < 1%) **Business Metrics** - **Conversion Rate**: Leads → Customers - **Revenue per Lead**: Average revenue - **Score Distribution**: Lead score buckets **Model Metrics** - **AUC**: Model discrimination (Target: > 0.85) - **Precision/Recall**: Classification performance - **Calibration**: Prediction probability accuracy ### Alerting System **Alert Levels** - **Info**: Logged to dashboard - **Warning**: Slack notification (#ml-alerts) - **Critical**: Email + Slack + PagerDuty **Alert Thresholds** - Latency P95 > 500ms: Critical - Error rate > 1%: Critical - PSI > 0.5: Critical drift - Conversion rate drop > 10%: Warning ### Monitoring Dashboard - Real-time metrics visualization - Drift detection charts - Alert log viewer - Model performance trends

## Slide 7: Drift Detection ### Drift Types **Data Drift** - Changes in feature distributions - New categories in categorical features - Missing value pattern changes **Concept Drift** - Changes in feature-target relationship - Model performance degradation - Business environment changes ### Detection Methods **PSI (Population Stability Index)** - Measures distribution shift - Thresholds: - PSI < 0.1: No change - PSI 0.1-0.25: Moderate change (Warning) - PSI > 0.25: Significant change (Critical) **KL Divergence** - Measures distribution difference - Threshold: KL > 0.1 **Statistical Tests** - Kolmogorov-Smirnov test (continuous) - Chi-square test (categorical) - P-value < 0.05 indicates drift ### Drift Response **Automatic Actions** - Log drift metrics to Delta table - Send alerts to monitoring team - Trigger retraining pipeline if PSI > 0.5 **Monitoring Schedule** - Real-time: Feature statistics - Hourly: Distribution checks - Daily: PSI calculation - Weekly: Comprehensive drift report

## Slide 8: CI/CD Workflow ### Pipeline Stages **1. Code Quality** - Linting (Flake8) - Code formatting (Black) - Unit tests (Pytest) **2. Validation** - Notebook syntax validation - Model validation checks - Integration tests **3. Deployment** - **Staging**: Auto-deploy on develop branch - **Production**: Manual approval + canary deployment **4. Model Registry** - Register new models to MLflow - Promote from Staging to Production - Archive previous versions ### Canary Deployment Process **Phase 1: 10% Traffic (1 day)** - Monitor latency, error rate, conversion - Automatic rollback on failure **Phase 2: 50% Traffic (2 days)** - Continue monitoring - Validate business metrics **Phase 3: 100% Traffic** - Full production deployment - Intensive monitoring for 48 hours ### Rollback Mechanism - Automatic: Error rate > 2%, Latency > 50% degradation - Manual: One-click rollback to previous version - Maintains model registry history

## Slide 9: Retraining Strategy ### Retraining Triggers **Automatic Triggers** - Data drift detected (PSI > 0.25) - Model performance degradation (AUC drop > 5%) - Scheduled retraining (weekly/monthly) **Manual Triggers** - Business requirement changes - New feature availability - Admin-initiated retraining ### Retraining Workflow **1. Data Collection** - Latest 6 months of labeled data - Minimum 10,000 records - Time-based train/test split **2. Model Training** - Hyperparameter optimization (Optuna, 50-100 trials) - Time series cross-validation - XGBoost/LightGBM models **3. Model Evaluation** - Compare with production model - Must show improvement: - +2% AUC OR - +5% business KPI - Shadow testing for 1-2 weeks **4. Model Promotion** - Register to MLflow Staging - Manual review and approval - Canary deployment - Promote to Production ### Retraining Schedule - **First 3 months**: Weekly retraining - **After 3 months**: Monthly retraining - **Drift-triggered**: As needed

## Slide 10: Sales Team Complaint Investigation ### Scenario: "Model scores are wrong - high-scored leads aren't converting" ### Investigation Workflow **Step 1: Immediate Response (Within 1 hour)** - Check model health metrics - Verify API is functioning correctly - Review recent model changes - Check for data quality issues **Step 2: Data Analysis (Within 4 hours)** - Analyze conversion rates by score bucket - Compare current vs historical performance - Check for data drift (PSI, KL divergence) - Review feature distributions **Step 3: Model Performance Review (Within 24 hours)** - Evaluate model metrics (AUC, Precision, Recall) - Compare production vs shadow model - Review calibration plots - Check for concept drift **Step 4: Root Cause Analysis** - **If Data Drift**: Investigate data source changes - **If Concept Drift**: Business environment may have changed - **If Model Issue**: Review training data and features - **If Integration Issue**: Check CRM sync and data pipeline **Step 5: Resolution** - **Data Issue**: Fix data pipeline, retrain model - **Model Issue**: Retrain with updated data/features - **Business Change**: Update model to reflect new patterns - **Integration Issue**: Fix CRM sync mechanism **Step 6: Communication** - Document findings and resolution - Update sales team with explanation - Implement preventive measures - Schedule follow-up review ### Tools & Dashboards - Streamlit dashboard for metrics - MLflow UI for model comparison - Drift detection reports

## Slide 11: Model Explainability & Fairness ### Model Explainability **SHAP (SHapley Additive exPlanations)** - Feature importance for each prediction - Local and global explanations - Explainability API endpoint (`/explain-prediction`) **LIME (Local Interpretable Model-agnostic Explanations)** - Local model explanations - Feature contribution analysis - Human-readable explanations ### Bias Detection & Fairness **Fairness Metrics** - **Demographic Parity**: Equal selection rates across groups - **Equalized Odds**: Equal true/false positive rates - **Selection Rate**: Fair selection across sensitive attributes **Bias Detection** - Automatic bias detection - Fairness threshold monitoring - Alert on bias violations ### Transparency Benefits - ■ Regulatory compliance (explainable AI requirements) - ■ Trust and stakeholder confidence - ■ Bias detection and mitigation - ■ Model interpretability

## Slide 12: Enhanced Auditability & Governance ### Comprehensive Audit Logging **Audit Trail Components** - All API calls logged with user, timestamp, IP - Model deployment/rollback events - Data access tracking - Failed action logging **Compliance Features** - Immutable audit logs (7-year retention) - Compliance reporting - Regulatory audit support - Complete action traceability ### ML Governance Framework **Model Approval Workflow** - Multi-stage approval process - Risk-based approval levels - Compliance validation - Documentation requirements **Risk Assessment** - Automated risk scoring - Risk factor analysis (data quality, performance, bias, stability, security) - Risk-based approval requirements - Risk mitigation planning **Governance Benefits** - ■ Controlled model deployment - ■ Risk management - ■ Regulatory compliance - ■ Accountability and transparency

## Slide 13: Disaster Recovery & Business Continuity ### Disaster Recovery Plan **Recovery Objectives** - **RTO (Recovery Time Objective)**: 4 hours - **RPO (Recovery Point Objective)**: 1 hour **Backup Strategy** - Model registry: Daily backups - Data: Daily incremental, weekly full - Configuration: On every change - Retention: 7 years (compliance) ### Failover Mechanisms **API Failover** - Primary/Secondary regions - Automatic failover on health check failure - Load balancer routing **Model Failover** - Fallback to previous model version - Automatic rollback on errors - Performance-based failover **Data Failover** - Cross-region replication - Delta Lake failover - Data integrity validation ### Business Continuity - ■ 99.9% uptime target - ■ Automated failover - ■ Regular DR drills - ■ Complete recovery procedures - Conversion rate analysis ### Prevention - Daily monitoring of conversion rates - Weekly model performance reviews - Proactive drift detection - Regular stakeholder communication

## Appendix: Key Metrics Dashboard ### Real-time Metrics - API Latency: 145ms (P95) - Throughput: 25 req/s - Error Rate: 0.2% - Conversion Rate: 12.5% ### Model Performance - AUC: 0.87 - Precision: 0.82 - Recall: 0.75 - F1 Score: 0.78 ### Drift Status - Overall PSI: 0.15 (Normal) - No critical drift detected - All features within thresholds

**End of Presentation**