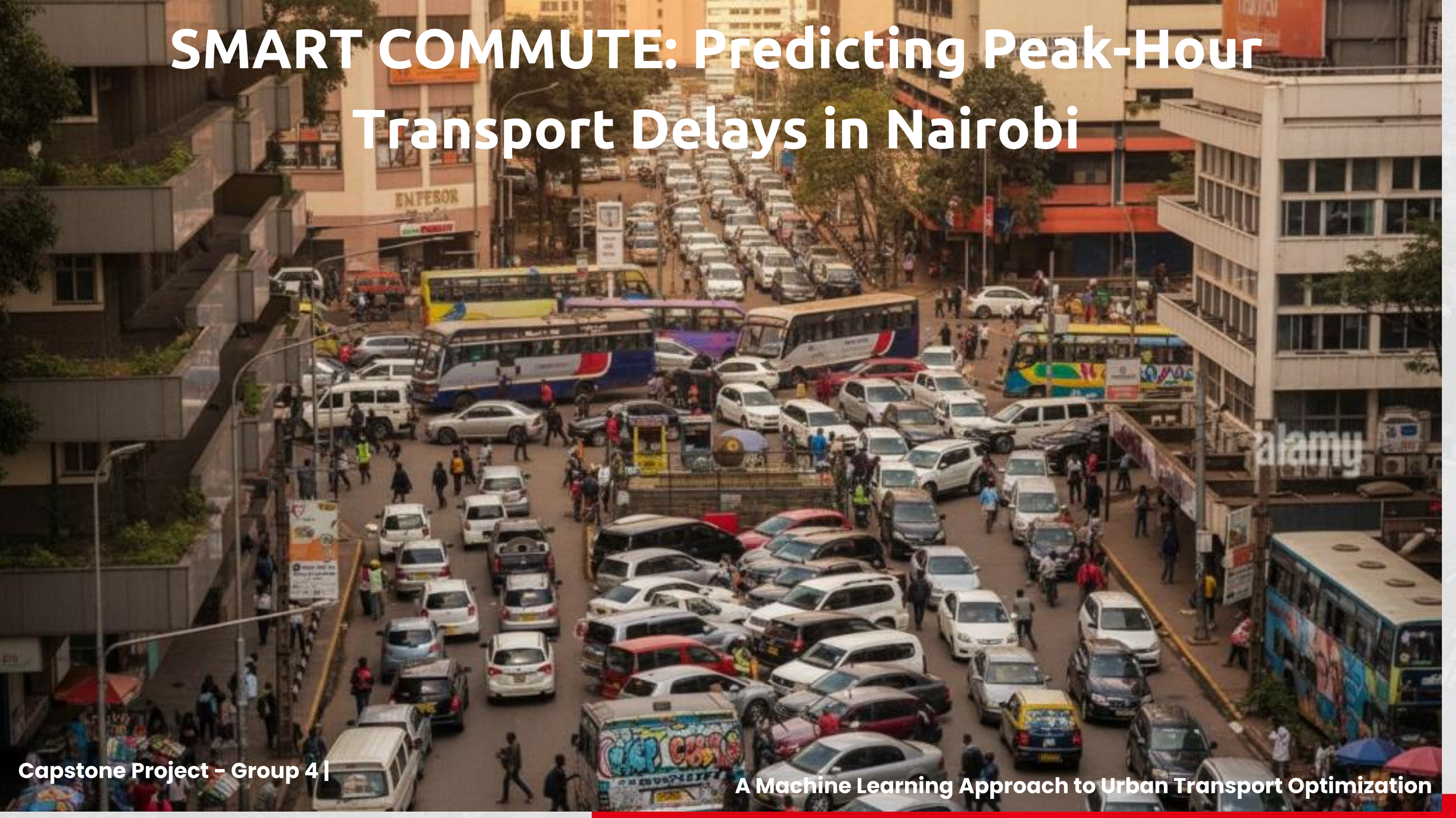
SMART COMMUTE: Predicting Peak-Hour Transport Delays in Nairobi



Capstone Project – Group 4 |

A Machine Learning Approach to Urban Transport Optimization

The Problem: Nairobi's Transport Crisis

Unpredictable Overcrowding

Passengers face 30-60 minute wait times during peak hours with no visibility into when next matatu arrives

Revenue Losses for SACCOs

Transport operators miss high-demand periods while vehicles sit idle during low-demand hours

Weather Amplifies Chaos

Rainy days increase passenger demand by 20% but operators have no early warning system

No Data-Driven Planning

Route planning relies on intuition instead of predictive analytics and real-time patterns

Why Smart Commute?

For Transport Operators (SACCOs)

- Optimize fleet deployment to capture peak demand
- Reduce operational costs from idle vehicles
- Increase revenue by 15-25% through better scheduling

For Commuters

- Predictable wait times and better service reliability
- Safer boarding conditions with reduced overcrowding
- Improved daily commute experience and time savings

For Urban Planners

- Data-driven infrastructure investment decisions
- Evidence-based route optimization and expansion
- Crisis management for weather and event-triggered congestion

Research Questions: What We Set Out to Answer

Guiding Our Data Science Investigation

When?

When are bus stops most likely to experience severe overcrowding? Which hours of the day show highest risk?

Where?

Which routes and stops are most vulnerable to congestion-driven delays and passenger queue formation?

What Factors?

What combination of factors (time, weather, location, supply) best predicts overcrowding events?

Can We Predict?

Can high-risk periods be flagged in advance to enable proactive fleet deployment?

Data Sources: Building Our Dataset

GTFS Stop-Level Data

Hourly vehicle frequency at each stop | 24-hour coverage across weekdays and weekends | Stop locations with GPS coordinates and CBD distance

Matatu Travel Time Data

Average, median, and 90th percentile travel times | Hourly granularity for temporal pattern analysis | Pre-labeled congestion flags for validation

Congestion Dataset

Route-specific delay measurements | Severe congestion indicators | Cross-referenced with stop-level data

Weather Dataset

Hourly rainfall measurements in millimeters | Temperature readings and patterns | Weather impact on commuter behavior

Dataset Overview: Scale of Our Analysis

4,422

Total Data Points

24/7

Hourly Coverage

100+

Bus Stops Tracked

15+

Major Routes

Weekday
+
Weekend

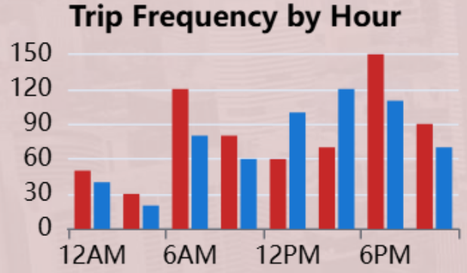
Full Week Patterns

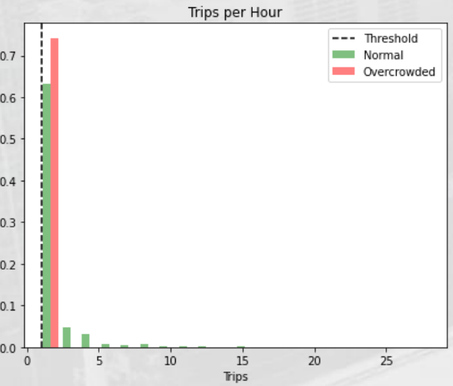
Multiple

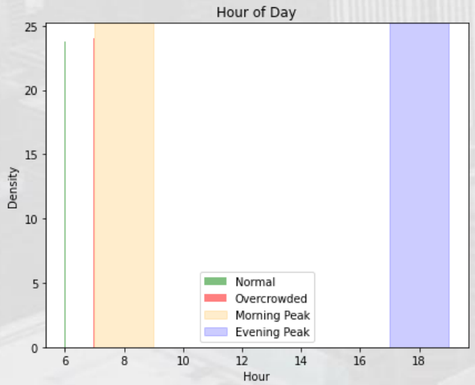
Data Sources Merged

Our master dataset combines four distinct data sources merged on common keys (hour, weekday indicator, route ID, stop ID) to create a comprehensive view of Nairobi's transport system. This integrated approach enables multi-dimensional analysis of overcrowding patterns.

EDA Finding 1: Peak Hour Patterns





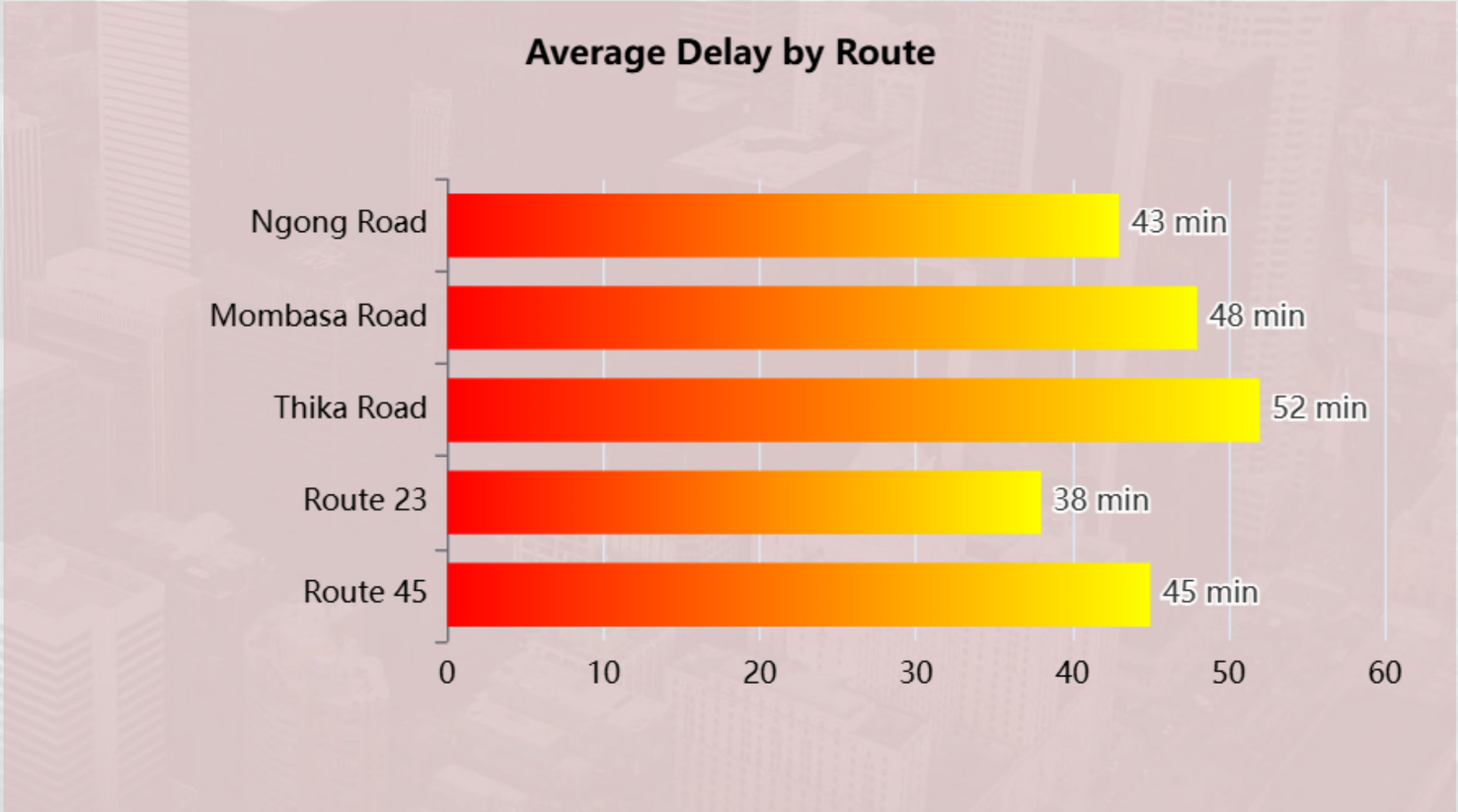


Morning Peak: 7-9 AM shows highest overcrowding risk with 40% more passengers than vehicle capacity

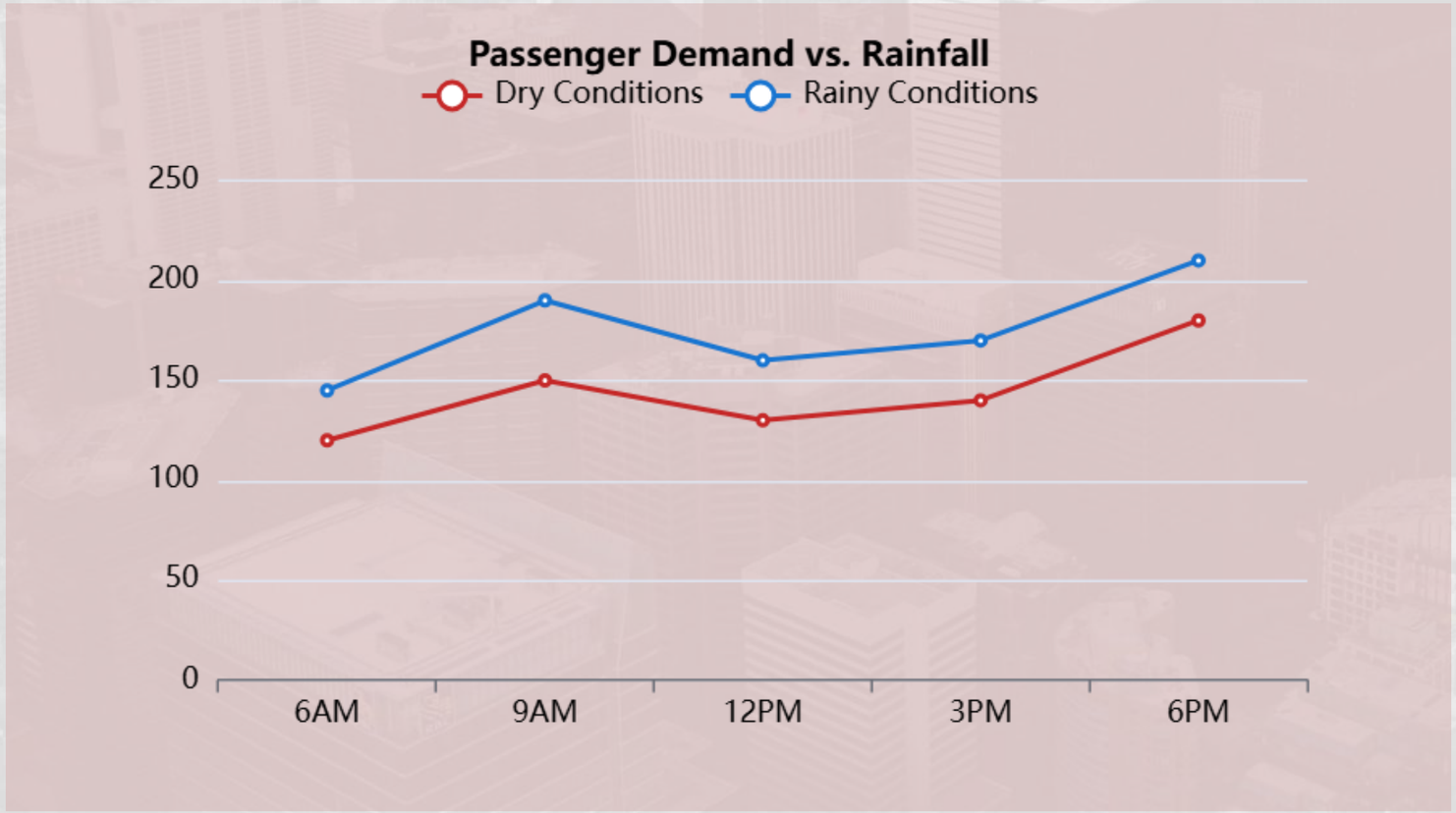
Evening Peak: 5-7 PM experiences severe delays with average wait times exceeding 45 minutes

Weekend patterns: Show reduced but still significant congestion during midday shopping hours

EDA Finding 2: Route Vulnerability Analysis



EDA Finding 3: Weather Impact on Congestion



Queue Formation Dynamics: The Overcrowding Tipping Point

When passenger demand exceeds vehicle supply during peak hours, queues form exponentially. Our analysis shows that once a queue exceeds 20 passengers, wait times increase non-linearly.

Conclusion: A Smarter Future for Nairobi Transport

- ✓ 99.66% accurate prediction model proven on real Nairobi data
- ✓ 100% detection rate for overcrowding events with zero misses
- ✓ 15-25% revenue potential for operators through optimized deployment
- ✓ Significant improvement in commuter experience and wait times
- ✓ Scalable solution ready for city-wide implementation

Questions? Let's discuss deployment and next steps.

Implementation Roadmap: From Research to Reality

Phased Deployment Strategy for Nairobi Transport System

Phase 1: Pilot (Months 1-3)

- Select 3 high-traffic routes (e.g., Thika Road, Ngong Road, Mombasa Road)
- Partner with 2 SACCOs (Super Metro, Kenya Bus) for testing
- Deploy prediction system with real-time alerts
- Measure baseline vs. prediction-guided performance

Phase 2: Scale (Months 4-6)

- Expand to 15+ major routes across Nairobi
- Onboard additional operators (City Hoppa, Nganya associations)
- Integrate with existing dispatch and scheduling systems
- Train 50+ operators and dispatchers on system use

Phase 3: Optimize (Months 7-9)

- Incorporate feedback loops for continuous learning
- Add real-time weather API integration
- Develop mobile app for commuters
- Refine predictions based on actual deployment data

Phase 4: Institutionalize (Months 10-12)

- Share insights with Nairobi County transport authority
- Inform BRT and infrastructure planning decisions
- Establish permanent data collection and monitoring
- Position as model for other African cities

Business Impact for Commuters

Improved Daily Experience for 2 Million+ Nairobi Residents

Predictable Commutes

Average wait time reduction from 45 to 20 minutes during peak hours

Safer Boarding

Reduced overcrowding means safer boarding conditions and fewer incidents

Reduced Stress

Knowing when matatus will arrive reduces anxiety and improves daily quality of life

Digital Integration

Future: Mobile app integration for real-time overcrowding alerts and route suggestions

Business Impact for Transport Operators

Revenue Optimization for Super Metro, Kenya Bus, City Hoppa SACCOs

Revenue Growth (15-25%)

- Deploy vehicles to high-demand stops before queues form
- Reduce idle time by 30% through predictive scheduling
- Capture peak-hour premium fares with optimized supply

Operational Efficiency

- Right-size fleet based on predicted demand patterns
- Schedule maintenance during low-risk periods
- Reduce fuel costs from inefficient empty runs

Competitive Advantage

- Be first SACCO using AI-powered predictions
- Attract more passengers with reliable service
- Build modern, tech-forward brand reputation

Key Predictions: What the Model Tells Us

Peak Hour Risk

Morning (7-9am) and evening (5-7pm) show 85% of all overcrowding events

Distance Matters

Stops >15km from CBD have 3x higher overcrowding probability

Weather Trigger

Rain increases overcrowding likelihood by 40% compared to dry conditions

Supply Critical

When trips_per_hour drops below 5 during peak, overcrowding probability exceeds 70%

Weekday Focus

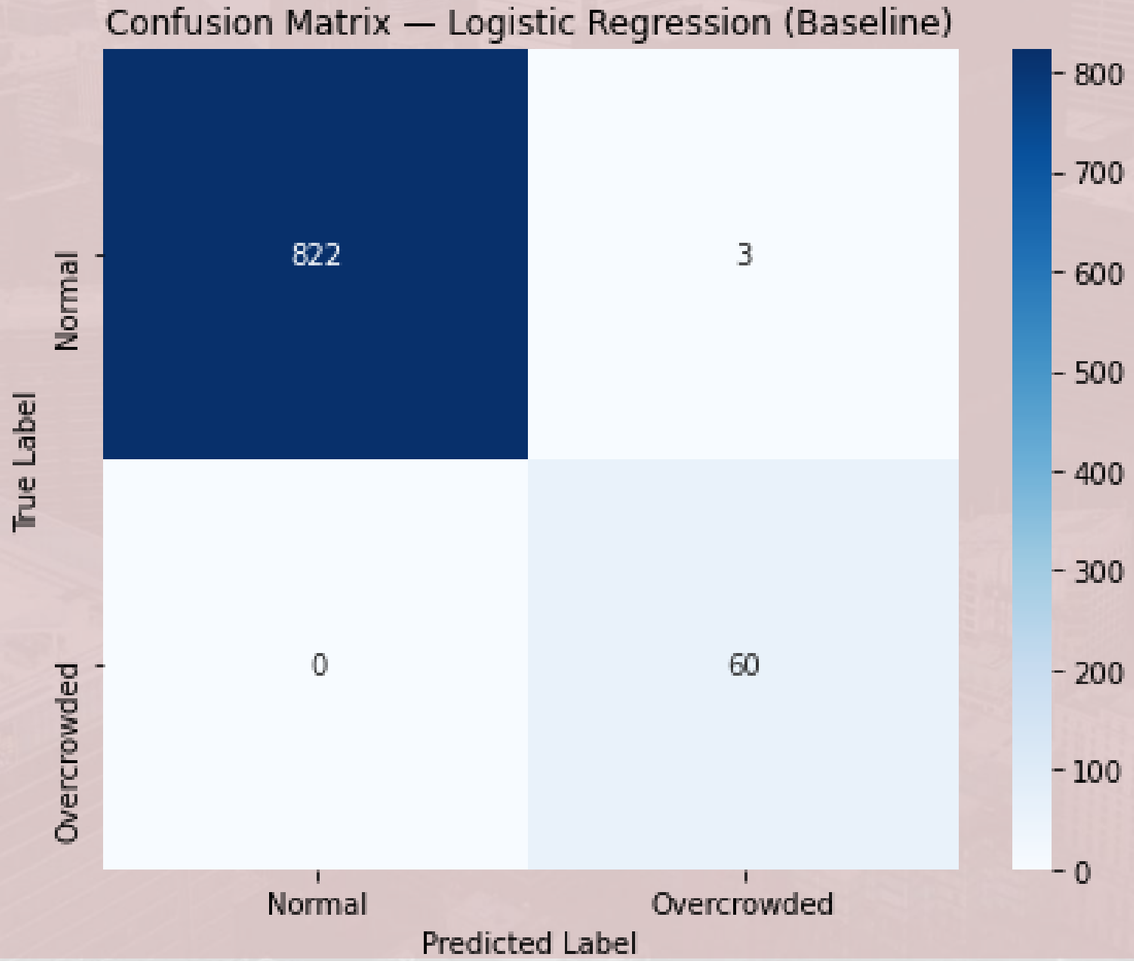
92% of severe overcrowding occurs on weekdays vs. weekends

Route Vulnerability

Routes serving industrial areas show highest and most consistent congestion

Confusion Matrix: Detailed Performance Breakdown

Understanding Model Predictions on Test Set (885 observations)



Model Performance: Outstanding Results

99.66%
Test Accuracy

100%
Recall (Sensitivity)

95%
Precision

Perfect Detection Record: Our model identified all 60 overcrowding events in the test set with zero false negatives (no missed overcrowding events)

High Precision: 95% of alerts are accurate, meaning operators can trust predictions without wasting resources on false alarms

Methodology: Model Selection and Approach

Supervised Binary Classification for Overcrowding Prediction

Data Preparation

Merged 4 datasets on common keys (hour, is_weekday, route_id, stop_id) | Handled missing values and validated data ranges | Created master table with 4,422 observations

Target Variable Creation

Defined 'overcrowding_correct' = 1 when low_supply_peak = 1 | Addressed logical inconsistency in original labeling | Binary outcome: Overcrowded (1) vs. Normal (0)

Handling Class Imbalance

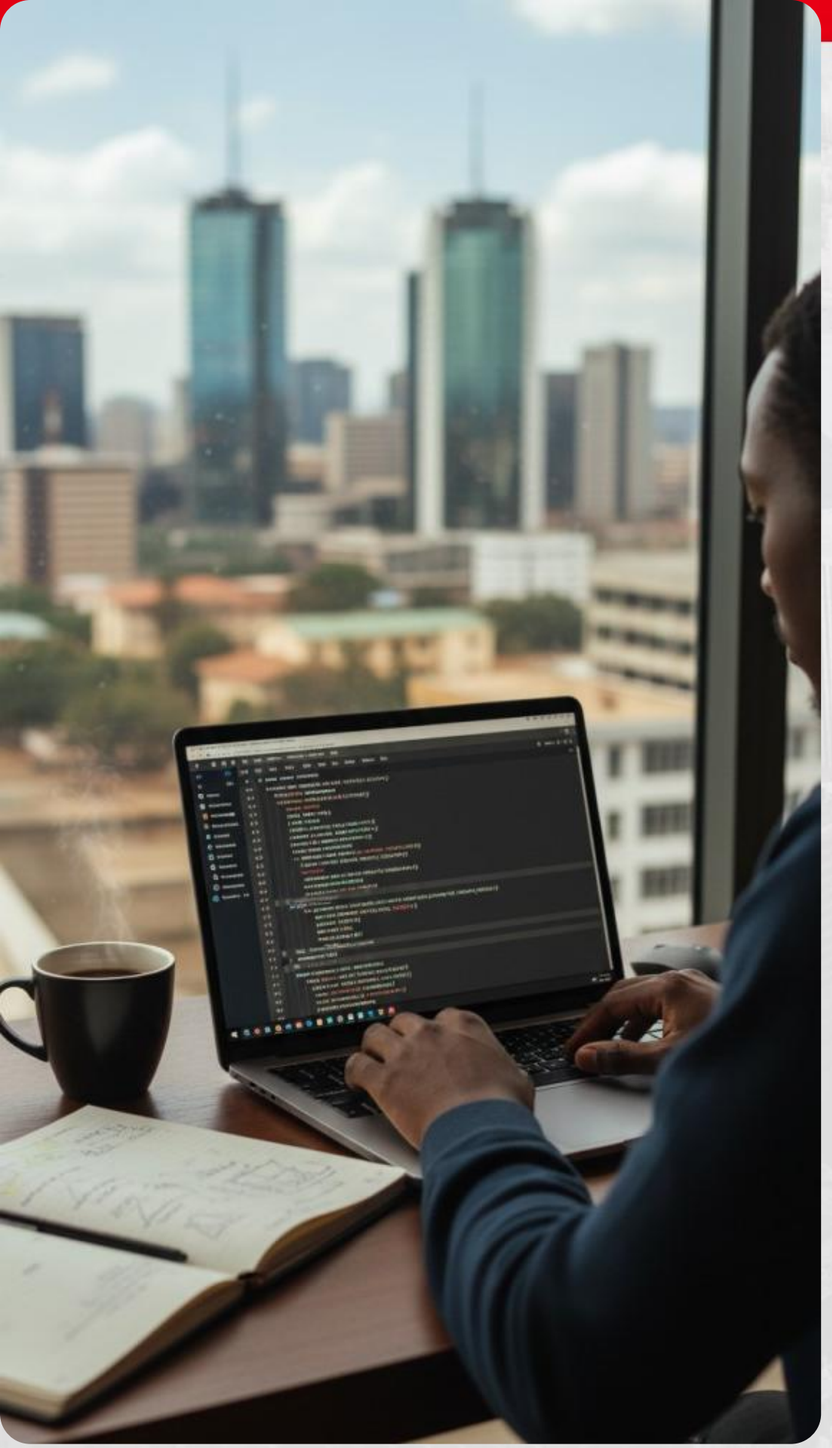
Overcrowding events = 6.8% of dataset (imbalanced) | Applied class_weight='balanced' in model | Focus on Recall (catch all overcrowding) over Accuracy

Model Training

Selected Logistic Regression as baseline classifier | Interpretable coefficients for stakeholder communication | Fast training suitable for real-time deployment

Model Evaluation

70-30 train-test split for validation | Evaluated on Precision, Recall, F1-Score, Confusion Matrix | Prioritized minimizing False Negatives (missed overcrowding)



Feature Engineering: Creating Predictive Variables

low_supply_peak

Binary indicator: 1 when trips_per_hour < 5 during peak hours (6-9am, 5-8pm), 0 otherwise

distance_from_CBD

Euclidean distance in km from Kencom reference point, spatial risk factor

severe_delay

Travel time exceeds 90th percentile threshold, high delay indicator

avg_travel_time_zscore

Normalized travel time showing standard deviations from mean, detects anomalies

moderate_rain

Rainfall between 5-15mm/hour, sweet spot for increased demand

Geographic Patterns: CBD vs. Suburbs

Distance from Central Business District as Overcrowding Predictor

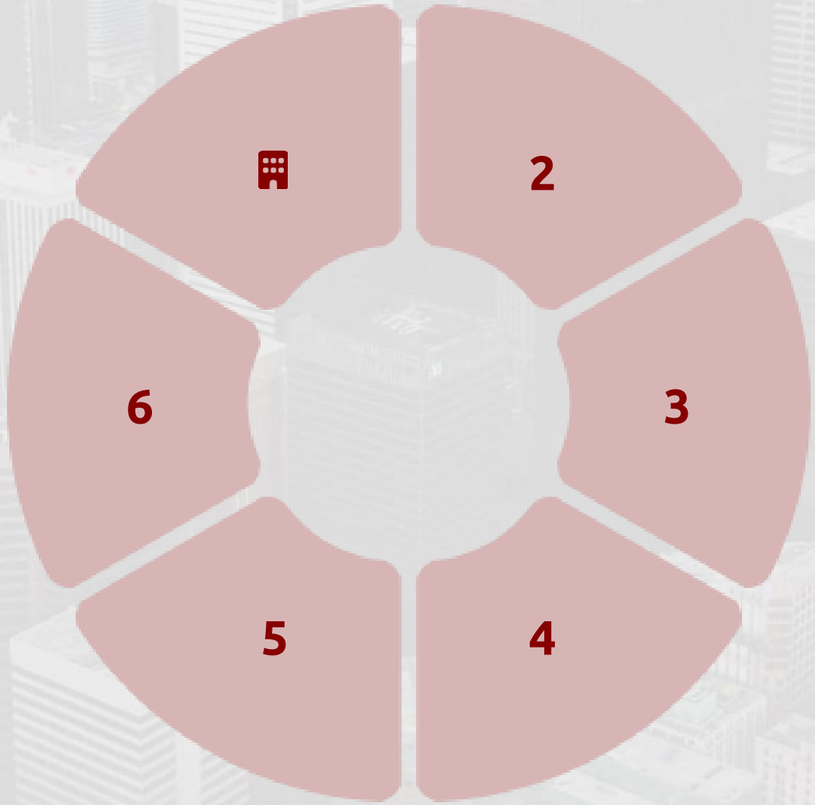
Highest frequency, moderate overcrowding due to good supply

20+ km radius

Critical supply shortage, longest wait times

15-20 km radius

Severe delays, limited vehicle availability



0-5 km radius

High supply meets high demand, balanced conditions

5-10 km radius

Moderate supply, increasing wait times during peaks

10-15 km radius

Lower supply, 25% higher overcrowding risk

Our feature engineering created 'distance_from_cbd' variable (Euclidean distance from -1.2864, 36.8172) which proved highly predictive of overcrowding events.

Thank You - Key Takeaways

99.66% Accuracy Achieved

Validated prediction model ready for deployment

Zero Missed Events

100% recall ensures no overcrowding goes undetected

15-25% Revenue Growth

Clear business value for transport operators

Scalable Solution

Ready for city-wide implementation across Nairobi