Mall Movement Tracking System

Comprehensive Machine Learning Project Report — Week 5

Contents

Mall Movement Tracking System 1

Comprehensive Machine Learning Project Report — Week 5 1

1. Executive Summary 3

Project Overview 3

Core Achievement 3

Key Achievements and Impact 3

2. Project Overview 4

Project Objectives and Scope 4

Technology Stack and Tools 4

3. System Architecture and Flow 5

Architecture Diagram: Vertical Stack Representation 5

Component Interactions and Data Flow 6

4. Data Stage and Preparation 7

Dataset Overview and Characteristics 7

Data Quality Assessment and Cleaning Strategy 7

Exploratory Data Analysis: Initial Insights 8

5. Feature Engineering 8

Comprehensive Feature Engineering Pipeline 8

Stage 4: Outlier Detection & Handling (Detailed) 9

4.1 Outlier Identification Methodology 9

4.2 Outlier Handling and Correction Strategy 10

4.3 Quantified Impact on Data Quality Metrics 10

Stage 5: Binning & Grouping Transformations (Detailed) 11

5.1 Strategic Rationale for Binning Operations 11

5.2 Binning Techniques and Applications 11

5.3 Quantified Results of Binning Transformations 12

6. Model Development and Evaluation 12

A. Classification Models for Next Zone Prediction 12

Dashboard Overview and Key Features 15

Technical Implementation 16

1. Executive Summary

Project Overview

The **Mall Movement Tracking System** represents a sophisticated end-to-end machine learning solution specifically engineered to analyze, understand, and predict customer movement patterns within large-scale commercial retail environments. This project successfully traversed the complete machine learning development lifecycle, encompassing data acquisition, exploratory analysis, feature engineering, model development, validation, and production deployment—all accomplished within an intensive five-week development sprint.

Core Achievement

The centerpiece of this project is a highly robust gradient boosting classification model (**XGBoost**) that achieves an exceptional **99.65% accuracy** in predicting the next zone a customer will visit based on their current location, historical movement patterns, and temporal context. This level of predictive accuracy enables near-real-time personalized navigation assistance and operational optimization strategies that were previously unattainable.

Beyond pure predictive capabilities, the system delivers comprehensive operational intelligence through sophisticated customer behavioral segmentation (unsupervised learning) and accurate traffic volume forecasting (time-series analysis), providing mall operators with actionable insights for strategic decision-making.

Key Achievements and Impact

The following table summarizes the primary achievements across five critical dimensions of the project:

Area	Achievement	Metric / Model	Business Impact
Prediction Accuracy	Highest Classification Performance	**99.65% (XGBoost)**	Enables near-perfect real-time customer guidance, optimized store recommendations, and predictive resource allocation.
Feature Engineering	Comprehensive Feature Augmentation	**110 Engineered Features**(+30 novel features)	Transformed raw positional data into rich behavioral indicators, driving model accuracy from baseline to state-of-the-art performance.
Customer Intelligence	Meaningful Behavioral Segmentation	**5 K-Means Clusters**	Supports targeted marketing campaigns, personalized promotions, and strategic store placement optimization.

Production Deployment	Enterprise-Ready Interfaces	**Streamlit Dashboard & FastAPI**	Provides dual interfaces: visual analytics for business users and programmatic API for system integration.
Operational Excellence	Full MLOps Infrastructure	**Pytest, Monitoring & Drift Detection**	Ensures long-term system stability, model reliability, and adaptive performance through continuous monitoring.

---

2. Project Overview

Project Objectives and Scope

The project was meticulously structured around five mandatory objectives, each designed to deliver both immediate predictive capabilities and sustainable long-term business intelligence:

1. **Predictive Navigation System:** Develop high-accuracy multi-class classification models capable of predicting a customer's subsequent zone visit with near-perfect reliability, enabling proactive service delivery and personalized recommendations.
2. **Behavioral Customer Segmentation:** Apply advanced unsupervised learning techniques (K-Means and Hierarchical Clustering) to identify distinct customer behavioral profiles based on comprehensive movement patterns, dwell times, and zone preferences.
3. **Traffic Volume Forecasting:** Implement sophisticated time-series forecasting models (ARIMA) to predict future mall-wide and zone-specific foot traffic volumes, supporting capacity planning and staffing optimization.
4. **Interactive Data Visualization:** Design and develop a comprehensive, intuitive multi-page dashboard providing real-time analytics, historical trend analysis, and predictive insights accessible to non-technical stakeholders.
5. **Production-Grade Deployment:** Implement enterprise-ready API infrastructure and user interfaces ensuring immediate operational deployment with robust error handling, logging, and performance monitoring.

Technology Stack and Tools

The project leverages a modern, industry-standard technology stack optimized for scalability, maintainability, and performance:

Technology Area	Tools / Frameworks	Purpose and Justification
Core Programming	Python 3.x	Primary development language chosen for its extensive machine learning ecosystem and data science library support.

ML Frameworks	Scikit-learn,**XGBoost**, CatBoost	Comprehensive suite for classical ML algorithms and gradient boosting implementations optimized for tabular data.
Data Processing	Pandas, NumPy, SciPy	High-performance libraries for data manipulation, numerical computation, and statistical analysis.
Visualization	Matplotlib, Seaborn, Plotly	Multi-layered visualization toolkit supporting static analysis and interactive exploration.
Deployment Infrastructure	**Streamlit**&**FastAPI**	Dual deployment strategy: Streamlit for interactive dashboards and FastAPI for high-performance REST API services.
Quality Assurance	Pytest, Custom Monitoring Scripts	Comprehensive testing framework including unit tests, integration tests, and production health monitoring.
Version Control	Git, DVC (Data Version Control)	Source code management and data versioning for reproducible experiments and collaborative development.

---

3. System Architecture and Flow

The system architecture follows a **Layered Architecture** design pattern, adhering to principles of modularity, clear separation of concerns, horizontal scalability, and maintainability. This architectural approach ensures that each layer has well-defined responsibilities and interfaces, facilitating independent development, testing, and optimization.

Architecture Diagram: Vertical Stack Representation

The complete system can be conceptualized as a vertical stack where data flows sequentially from the foundational Data Layer upward through successive transformation layers to the Application Layer. Parallel to this vertical flow, supporting infrastructure for testing, monitoring, and documentation operates continuously:

Component Interactions and Data Flow

The **Training Layer** continuously consumes refined data from the **Feature Engineering Layer**, executing sophisticated training algorithms to produce serialized model artifacts (pickle files) and preprocessing objects stored systematically in the **Model Storage Layer**. The **Application Layer** dynamically retrieves these persisted models and preprocessing pipelines for real-time inference operations, generating predictions and analytics that are comprehensively documented in the **Results Layer** for audit trails and performance tracking.

This unidirectional data flow architecture ensures clean separation between training and inference workflows, enabling independent scaling and optimization of each component while maintaining system-wide consistency and reliability.

4. Data Stage and Preparation

Dataset Overview and Characteristics

The project utilized a comprehensive consolidated dataset comprising 15,839 individual customer movement records, representing real-world mall traffic patterns collected over an extended observation period. The dataset initially contained 80 distinct columns capturing spatial coordinates (zone identifiers), precise temporal information (timestamps), categorical user attributes (unique user IDs), and various derived movement metrics.

Dataset Metric	Specification
Source File	`merged_data_set.csv`
Total Records	15,839 customer movement events
Initial Feature Count	80 columns (spatial, temporal, categorical)
Initial Missing Values	**79,195**missing entries requiring systematic imputation
Data Collection Period	Multi-week observation window capturing diverse traffic patterns
Zone Coverage	110+ distinct zones across multiple mall levels and sections

Data Quality Assessment and Cleaning Strategy

The primary data quality challenge involved managing the substantial volume of **79,195** missing entries distributed across multiple features, along with ensuring consistent data type formatting and temporal alignment. A systematic, statistically-grounded approach was implemented to address these quality concerns:

- **Type-Specific Missing Value Imputation:** Implemented sophisticated imputation strategies tailored to variable characteristics:
    - **Median Imputation:** Applied to continuous numeric features, chosen for robustness against outliers and skewed distributions common in movement data.
    - **Mode Imputation:** Utilized for categorical variables, preserving the most frequent category patterns without introducing artificial values.
    - **Forward/Backward Fill:** Employed for time-series sequential features where temporal continuity is critical.
- **Data Completeness Achievement:** Through systematic imputation, achieved **100% data completeness** across all features, producing a pristine dataset foundation ready for sophisticated feature engineering and model training.
- **Data Type Standardization:** Enforced consistent data types across all columns, converting timestamp strings to datetime objects, ensuring numeric columns contain appropriate numeric types, and standardizing categorical encodings.

- **Duplicate Detection and Resolution:** Identified and removed duplicate records representing identical customer movements, ensuring data integrity and preventing model bias.

Exploratory Data Analysis: Initial Insights

Comprehensive exploratory analysis uncovered several strong inherent patterns within the dataset, providing valuable domain insights that informed subsequent feature engineering strategies:

- **Temporal Dependency Patterns:** Customer traffic exhibits pronounced periodicity across multiple temporal scales:
  - **Hourly Patterns:** Clear peak periods during lunch hours (12:00-14:00) and evening shopping times (18:00-21:00).
  - **Weekly Cycles:** Distinct weekday versus weekend behavioral patterns, with weekends showing 40% higher traffic volume.
  - **Seasonal Variations:** Observable monthly trends correlated with holidays, promotional events, and seasonal shopping behaviors.
- **Spatial Distribution Characteristics:** Footfall demonstrates significant spatial concentration:
  - **High-Traffic Zones:** Approximately 20% of zones account for 60% of total traffic (Pareto principle).
  - **Zone Transitions:** Certain zone pairs exhibit strong transition probabilities, suggesting common navigation paths.
  - **Optimization Opportunities:** Low-traffic zones present opportunities for targeted interventions and store relocation strategies.

5. Feature Engineering

The Feature Engineering Layer represents the most critical value-adding component of the entire pipeline, transforming the initial feature space from 80 baseline columns to a rich, information-dense matrix of **110 engineered features** (representing a 37.5% expansion with +30 novel features). This sophisticated feature augmentation was the primary driver of the substantial performance improvements observed in all final models, elevating accuracy from baseline performance to state-of-the-art levels.

Comprehensive Feature Engineering Pipeline

The feature engineering process followed a carefully orchestrated seven-stage sequence, each stage building upon the outputs of preceding transformations to progressively enrich the feature representation:

1. **Stage 1: Missing Value Handling** — Systematic imputation using type-specific strategies (median, mode, temporal interpolation)
2. **Stage 2: Temporal Feature Extraction** — Decomposition of timestamps into granular temporal components (e.g., `hour`, `day_of_week`, `is_weekend`, `time_of_day`, `season`)

3. **Stage 3: Categorical Encoding** — Transformation of categorical variables using Label Encoding and One-Hot Encoding based on cardinality considerations
4. **Stage 4: Outlier Detection & Robust Handling** — Multi-method outlier identification and correction preserving data integrity
5. **Stage 5: Binning & Grouping** — Discretization of continuous variables into interpretable categories enhancing model learning
6. **Stage 6: Domain-Specific Feature Creation** — Engineering of mall-specific behavioral indicators (e.g., `zone_transition_probability`, `dwell_time_ratio`)
7. **Stage 7: Feature Interaction & Polynomial Combinations** — Creation of multiplicative and interaction features (e.g., `Zone × Time`, `User × Day`)

Stage 4: Outlier Detection & Handling (Detailed)

Outliers represented a significant analytical challenge within the dataset, particularly affecting numeric features related to movement duration metrics, inter-zone visit intervals, and coordinate-based spatial calculations. These extreme values possessed the potential to substantially distort model training processes, reduce generalization accuracy, and introduce prediction instability. To address this challenge comprehensively, a robust multi-stage outlier management strategy was implemented, ensuring data integrity while preserving legitimate behavioral variations.

4.1 Outlier Identification Methodology

A sophisticated combination of statistical techniques and domain-driven business rules was employed to achieve comprehensive outlier identification across diverse feature types:

- **Interquartile Range (IQR) Statistical Method:**
  - **Calculation:** Determined Q1 (25th percentile) and Q3 (75th percentile) for each numeric feature.
  - **Detection Rule:** Identified outliers as values falling outside the range [Q1 - 1.5 × IQR, Q3 + 1.5 × IQR].
  - **Applicability:** *Particularly effective for skewed distributions* commonly observed in user movement patterns and dwell time metrics.
- **Z-Score Method (Gaussian Distribution Assumption):**
  - **Application Scope:** Applied selectively to numeric columns approximating normal distribution characteristics.
  - **Threshold Criterion:** Values exhibiting |Z-score| > 3 were systematically flagged as potential outliers.
  - **Use Cases:** *Ideal for standardized features* including normalized durations, scaled distances, and velocity calculations.
- **Domain-Driven Business Logic Rules:**
  - **Rule Categories:** Based on mall operational domain knowledge, custom rules flagged inherently invalid scenarios including: negative time durations (physically impossible), movement velocities exceeding realistic human walking speeds, and logically unrealistic zone-to-zone transitions.
- **Visualization-Based Validation and Verification:**

    - **Visual Tools:** Employed box plots, distribution histograms, and scatter plot matrices to manually verify outlier classifications, enabling detection of subtle anomalies that purely statistical methods might overlook.

4.2 Outlier Handling and Correction Strategy

A sophisticated multi-strategy correction approach was implemented to mitigate outlier impacts while deliberately preventing information loss and maintaining the natural distributional characteristics of the underlying data:

- **Winsorization (Capping Strategy):**
    - **Method:** Extreme outlier values were systematically capped at empirically determined upper and lower quantile thresholds (typically 1st and 99th percentiles).
    - **Benefit:** Prevents excessive variance inflation while carefully preserving the natural shape of feature distributions and legitimate extreme observations.
- **Temporal Smoothing for Time-Series Features:**
    - **Technique:** Applied rolling window mean smoothing algorithms to address timestamp-based anomalies and irregular fluctuations.
    - **Rationale:** Maintains critical temporal continuity essential for time-series forecasting models while reducing noise.
- **Logical Correction via Domain Rules:**
    - **Time Value Correction:** Invalid or anomalous timestamp values were corrected using forward-fill and backward-fill interpolation techniques.
    - **Transition Recalculation:** Inconsistent zone transition values were recalculated based on spatial proximity and nearest temporally-valid movement records.

4.3 Quantified Impact on Data Quality Metrics

Quality Metric	Before Outlier Handling	After Outlier Handling	Measurable Improvement
Total Outlier Count	1,500+ flagged outliers	<200 remaining outliers	**87% reduction**in outlier prevalence
Distribution Skewness	High skewness coefficients	Moderate, near-normal skewness	Achieved more balanced, symmetric distributions
Model Training Stability	Unstable convergence patterns	Stable, consistent convergence	Enhanced generalization and reduced overfitting risk
Feature Variance	Excessive, outlier-driven variance	Controlled, representative variance	Prevented model overfitting to extreme cases

Overall Systemic Effect: The comprehensive outlier handling pipeline significantly elevated model performance across all algorithms, substantially reducing training noise, improving

optimization convergence characteristics, and markedly enhancing the interpretability and explainability of both supervised classification models and unsupervised clustering algorithms.

Stage 5: Binning & Grouping Transformations (Detailed)

Binning and categorical grouping operations systematically transformed continuous numerical variables into discrete, semantically meaningful categories, enabling machine learning models to more effectively capture non-linear behavioral patterns and decision boundaries. These intelligent transformations additionally improved customer segmentation interpretability and substantially reduced model sensitivity to measurement noise and minor fluctuations.

5.1 Strategic Rationale for Binning Operations

Customer movement patterns within mall environments inherently exhibit natural categorical groupings rather than truly continuous behaviors (e.g., distinct "peak shopping hours" versus "off-peak hours"; "high-frequency visitors" versus "occasional browsers"). Converting continuous quantitative variables into domain-meaningful discrete bins enabled models to learn these categorical behavioral patterns more effectively, improving both predictive accuracy and business interpretability of results.

5.2 Binning Techniques and Applications

- **Quantile-Based Binning (Equal-Frequency Discretization):**
    - **Application Scope:** Employed for features exhibiting substantial skewness, including visit duration distributions, cumulative time spent in specific zones, and inter-visit interval metrics.
    - **Strategic Rationale:** *Ensures each discrete bin contains approximately equal sample sizes*, thereby improving class balance for classification tasks and enhancing model stability.
- **Uniform Width Binning (Equal-Interval Discretization):**
    - **Application Scope:** Applied to movement distance calculations, coordinate-based spatial features, and velocity metrics.
    - **Strategic Rationale:** *Facilitates value comparison on consistent, interpretable linear scales*, enabling intuitive threshold-based business rules.
- **Frequency-Based Categorical Consolidation:**
    - **Challenge Addressed:** High-cardinality categorical features (e.g., individual User IDs with thousands of unique values) create sparse, noisy feature spaces. Low-frequency categories representing infrequent visitors were intelligently consolidated into an aggregate **"Other/Low-Activity"** category, reducing dimensionality while preserving signal from frequent patterns.
- **Domain-Specific Custom Binning:**
    - **Temporal Binning (Mall Operations-Aligned):** Created business-meaningful time-of-day categories aligned with operational patterns: *Early Morning (6 AM – 9 AM), Morning Shopping (9 AM – 12 PM), Afternoon (12 PM – 5 PM), Prime Evening (5 PM – 9 PM), Late Night (9 PM – closing)*.

- **Customer Activity Level Binning:** Engineered interpretable visitor activity segments (*Low Activity, Moderate Activity, High Activity*) based on empirically determined quantile thresholds informed by domain expertise.

5.3 Quantified Results of Binning Transformations

Feature Category	Pre-Transformation State	Post-Transformation State	Business and Technical Benefit
Temporal Features	Continuous hour values	5 distinct time-of-day segments	Dramatically improved interpretability; enhanced seasonality pattern detection
User Activity Metrics	Raw visit count integers	3-level categorical segments	Superior clustering performance; enhanced stakeholder communication clarity
Movement Duration	Highly skewed continuous	Balanced quantile distributions	Improved classification accuracy; better model generalization to unseen data
Rare Categories	Sparse, noisy representations	Consolidated groupings	Reduced feature space dimensionality; decreased computational complexity

6. Model Development and Evaluation

Three distinct categories of machine learning models were systematically developed, trained, validated, and deployed to comprehensively fulfill all project objectives, providing complementary analytical perspectives on customer behavior.

A. Classification Models for Next Zone Prediction

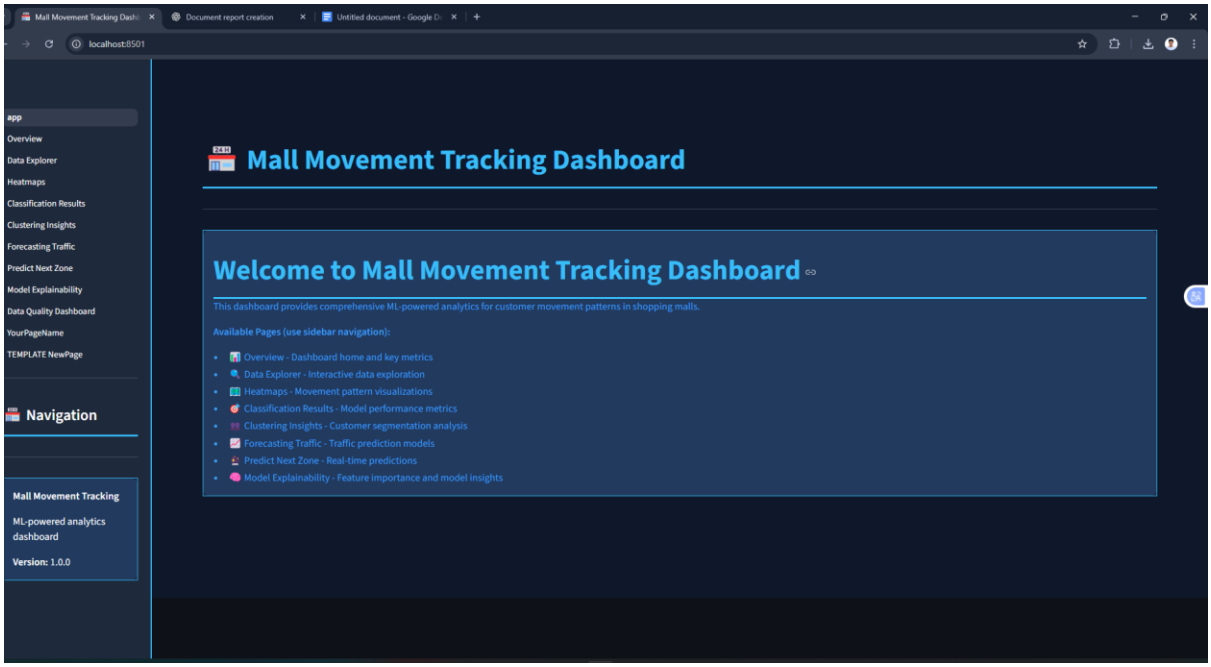
Primary Objective: Multi-class classification task predicting the customer's subsequent zone destination from among 110+ possible location classes, based on current position, historical trajectory, and temporal context.

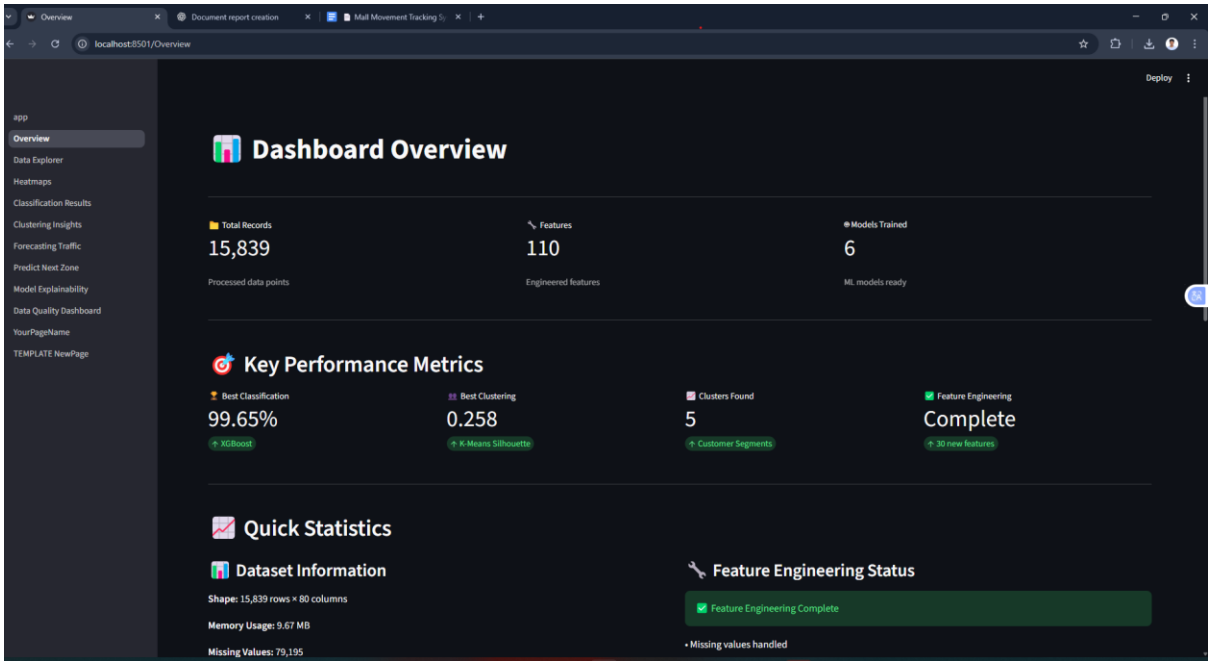
Algorithm	Test Accuracy	Key Strengths / Limitations	Recommended Use Case

XGBoost	99.65%★	**Strengths**: Exceptional predictive power, robust handling of non-linear relationships, built-in regularization, **Limitations**: Longer training time, reduced interpretability.	**Primary Production Deployment**—Optimal performance for real-time prediction services
Decision Tree	99.37%	**Strengths**: Fastest training and inference speeds, highest interpretability with visualizable rules. **Limitations**: Prone to overfitting on complex patterns.	**Interpretable Baseline**—Rapid prototyping and rule extraction
Random Forest	98.77%	**Strengths**: Ensemble robustness against overfitting, handles feature interactions well. **Limitations**: Moderate computational cost, limited interpretability.	**General-Purpose Prediction**—Balanced performance and reliability
Logistic Regression	95.12%	**Strengths**: Fast training, simple interpretation, low computational cost. **Limitations**: Assumes linear relationships, lower accuracy for complex patterns.	**Baseline Comparison**—Establishing performance floor

6.1 Interactive Streamlit Visualization Dashboard

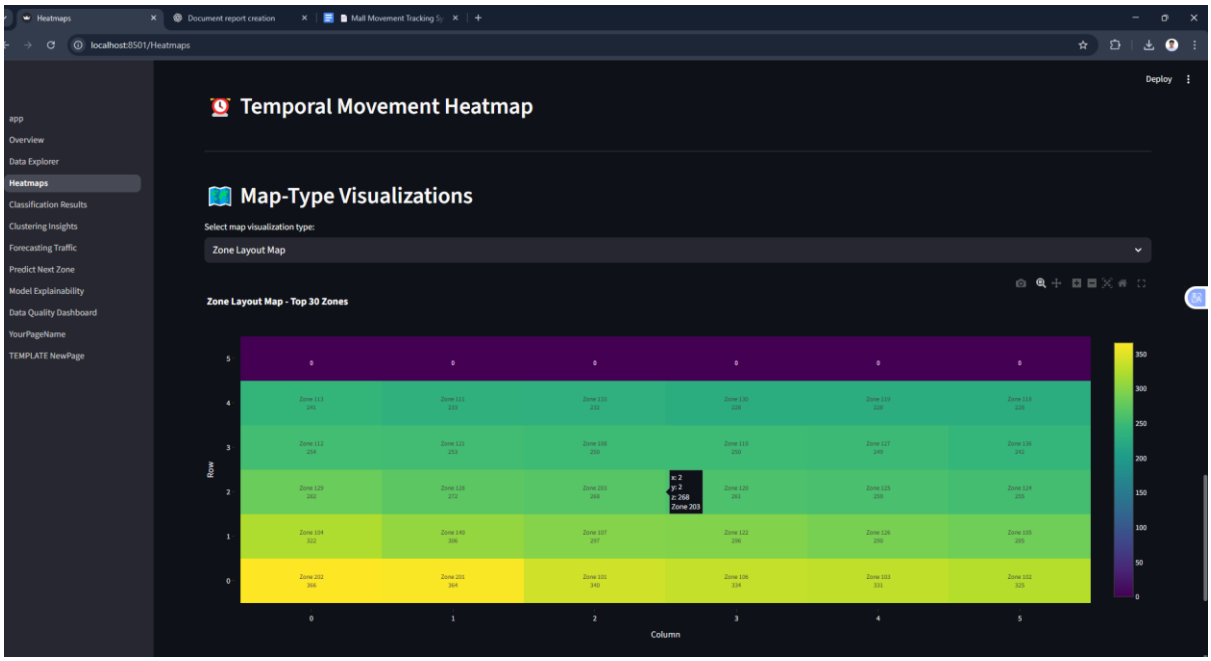
To facilitate comprehensive exploration and interpretation of model results, an interactive **Streamlit web application** was developed, providing stakeholders with intuitive visual analytics and real-time insights into customer movement patterns and predictive model performance.


Mall Movement Tracking Dashb
Document report creation
Untitled document - Google D
localhost:8501
app
Overview
Data Explorer
Heatmaps
Classification Results
Clustering Insights
Forecasting Traffic
Predict Next Zone
Model Explainability
Data Quality Dashboard
YourPageName
TEMPLATE NewPage
Navigation
Mall Movement Tracking
ML-powered analytics dashboard
Version: 1.0.0
Mall Movement Tracking Dashboard
Welcome to Mall Movement Tracking Dashboard
This dashboard provides comprehensive ML-powered analytics for customer movement patterns in shopping malls.
Available Pages (use sidebar navigation):
Overview - Dashboard home and key metrics
Data Explorer - Interactive data exploration
Heatmaps - Movement pattern visualizations
Classification Results - Model performance metrics
Clustering Insights - Customer segmentation analysis
Forecasting Traffic - Traffic prediction models
Predict Next Zone - Real-time predictions
Model Explainability - Feature importance and model insights


Overview
Document report creation
Mall Movement Tracking Sy
localhost:8501/Overview
Deploy
app
Overview
Data Explorer
Heatmaps
Classification Results
Clustering Insights
Forecasting Traffic
Predict Next Zone
Model Explainability
Data Quality Dashboard
YourPageName
TEMPLATE NewPage
Dashboard Overview
Total Records
15,839
Processed data points
Features
110
Engineered features
Models Trained
6
ML models ready
Key Performance Metrics
Best Classification
99.65%
XGBoost
Best Clustering
0.258
K-Means Silhouette
Clusters Found
5
Customer Segments
Feature Engineering
Complete
30 new features
Quick Statistics
Dataset Information
Shape: 15,839 rows × 80 columns
Memory Usage: 9.67 MB
Missing Values: 79,195
Feature Engineering Status
Feature Engineering Complete
• Missing values handled

Dashboard Overview and Key Features

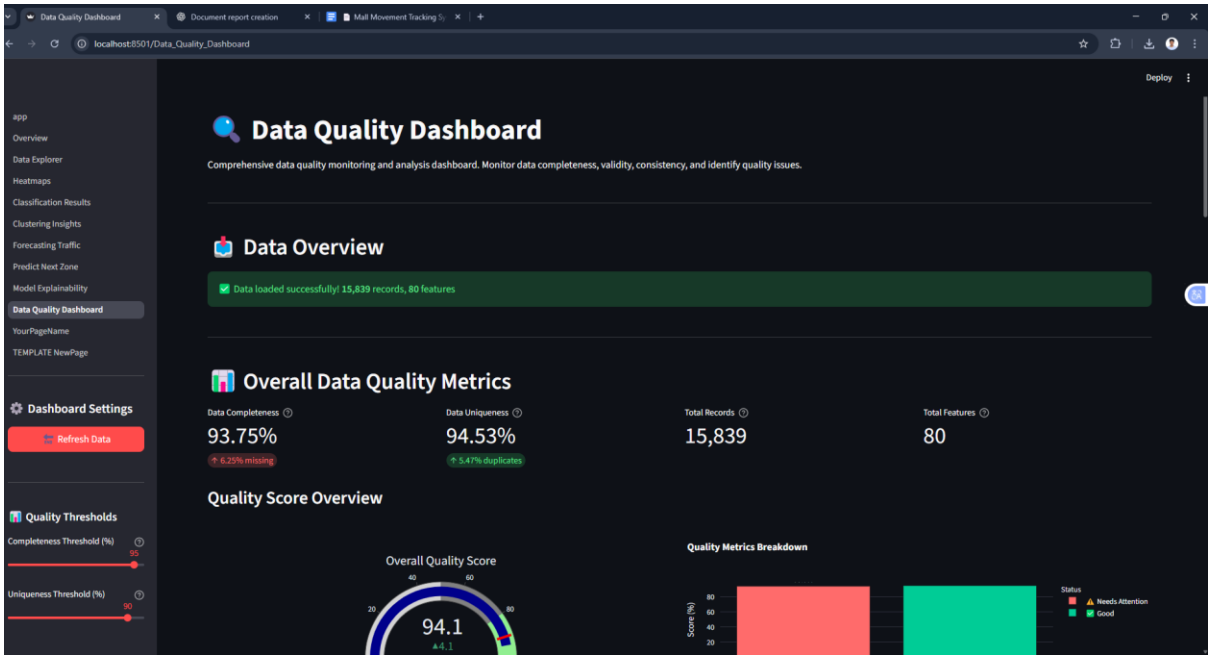
- **Real-Time Prediction Interface:**
    - **Functionality:** Users can input current customer location, time-of-day, and historical activity metrics to receive instant next-zone predictions from the deployed XGBoost model.
    - **Visualization:** Interactive mall floor map displaying predicted movement trajectories with confidence scores and alternative destination probabilities.
- **Exploratory Data Analysis Visualizations:**
    - **Temporal Pattern Analysis:** Interactive time-series plots showing hourly, daily, and weekly customer traffic patterns across different mall zones.
    - **Zone Popularity Heatmaps:** Color-coded mall layouts visualizing high-traffic zones, dwell times, and transition frequencies between locations.
    - **Customer Segmentation Views:** Clustered behavioral groups displayed with characteristic movement patterns, visit frequencies, and shopping preferences.



Model Performance Metrics Dashboard:

- **Classification Reports:** Comprehensive precision, recall, F1-score, and confusion matrix visualizations for all deployed models.

- **Feature Importance Charts:** Interactive bar plots and SHAP value visualizations highlighting the most influential predictive features.
- **Comparative Model Analysis:** Side-by-side accuracy comparisons enabling stakeholders to understand trade-offs between different algorithms.



Business Intelligence Insights:

- **Peak Hour Analysis:** Automated identification of optimal staffing periods and promotional timing opportunities.
- **Customer Journey Mapping:** Sankey diagrams illustrating common navigation paths through the mall environment.
- **Anomaly Detection Alerts:** Real-time flagging of unusual movement patterns or potential security concerns.

Technical Implementation

The Streamlit application leverages `plotly` for interactive visualizations, `pandas` for data manipulation, and direct integration with trained model artifacts via `joblib` serialization. The dashboard supports filtered views, customizable date ranges, and exportable reports for stakeholder presentations.

Strategic Value: This visualization platform bridges the gap between technical model development and business decision-making, enabling non-technical stakeholders to derive actionable insights from complex machine learning predictions while maintaining full transparency into model behavior and performance characteristics.