Forecasting TTC Subway Delays (2022–2024)

Overview

This project investigates subway delay patterns across Toronto's TTC system using three years of historical data. By applying machine learning models, your team uncovered high-impact delay causes, forecasted delay durations, and offered actionable insights for transit operators, planners, and riders.

Business Objective

- **Improve Service Reliability:** Predict long and short delays to inform operations.
- **Optimize Resource Allocation:** Identify peak delay windows for better staffing.
- **Empower Riders and Planners:** Deliver transparent, data-driven communication.

Dataset Summary

- **Source:** Open Data Toronto – TTC Subway Delays
- **Volume:** 69,071 cleaned entries from 2022–2024
- **Focus:** Controllable subway delays (Lines 1–3)
-

Methodology

Data Processing

- Cleaned and standardized station names
- Dropped columns with excessive missing values
- Derived temporal features (hour, day, min gap)

Feature Engineering

- Encoded categorical variables
- Created severity classes for classification
- Log-transformed delay duration for regression

Machine Learning Models

Classification (Delay Severity)

- Models Used: **Random Forest**, **Tuned XGBoost**
- Classes:
  - **Class 0**: No Delay
  - **Class 1**: Short Delay (
  - **Class 2**: Long Delay (>10 min)
- **Best Performance:** XGBoost

- **Overall Accuracy:** 98%
- **Class 2 F1 Score:** 0.98 → reliably detects long delays

Regression (Delay Duration)

- Target: `Delay` (log-transformed)
- Best Model: **XGBoost Regressor**

- Strong predictive power for 2–5 minute delays
- Top features: `Station`, `Delay Code`, `Min Gap`, `Hour of Day`

Key Visuals & Insights

- Most delays occur during **rush hours** (6–8AM, 3–6PM)
- **Operator-related issues** (e.g., "No Operator") drove longer delays
- **Top delay codes**: SUDP, MUPAA, SUO
- **Top stations**: Bloor-Yonge, St. George, Union
- Even **zero-minute delays** showed predictive signal for classification

Economic Impact Estimation

- **Monthly Delay Minutes:** ~5,903
- **Estimated Cost (at $50/min):** ~$295,000+/month
- **Annual Projection:** ~$3.5M CAD

Team & Contributions

Member	Role
Valerie Poon	Planning, documentation, final README
Sahil Modi	EDA and station analysis
Saad Khan	Data sourcing, GitHub setup
Sneha Gupta	Feature engineering, modeling
Sucharitha Sundararaman	Regression pipeline, SHAP analysis
Faiz Shaikh	Code review, solution pitch

Next Steps

- Enhance visual labeling and clarity
- Expand dataset for deeper historical trends
- Build a dashboard for real-time delay awareness
- Investigate why **Union Station** sees fewer delays despite high traffic

Tools & Libraries

- Python (pandas, NumPy, scikit-learn)
- XGBoost, SHAP

- Jupyter Notebooks
- matplotlib & seaborn for visualization