Predicting NYC Rideshare Prices: Using Subway Delays, Ridership, and Weather Conditions

IST 707 – Applied Machine Learning

Team Members: Dawryn Rosario, Marko Masnikosa, Rianne Parker

Overview

Problem:

- NYC rideshare prices fluctuate due to various factors, but external transportation disruptions (e.g., subway delays) are often overlooked.
- Current surge pricing models lack transparency and may not account for alternative transit options.
- Machine learning can analyze historical data to **predict price fluctuations** based on external factors.

Why ML?

- Traditional statistical analysis cannot dynamically capture complex relationships between subway delays, ridership, weather, and rideshare prices.
- ML models can learn these patterns and provide **predictive insights** for riders and service providers.

Stakeholders

Major Stakeholders:

- **Rideshare Companies (Uber, Lyft,etc.)** → Optimize surge pricing models.
- **NYC Residents** → Plan trips efficiently, avoiding costly fares.
- **City Transportation Authorities** → Understand demand shifts between public transit and rideshares.

Minor Stakeholders:

- **Tourists & Business Travelers** → Need accurate fare estimates.
- **Commuters Affected by Subway Delays** → Require alternative transport options.

The Gap

Current Approaches & Limitations:

- **Existing ML models** consider weather and traffic but rarely integrate **public transit data.**
- **Surge pricing algorithms** are proprietary, offering **no transparency** into price fluctuations.
- **Static fare estimations** don’t account for real-time subway conditions.

Why Our Approach?

- Combines **multiple data sources (subway, weather, rideshare)** for improved predictions.
- Provides insights for **riders and authorities**, enhancing trip planning.

Data Sources & Quality

Datasets:

- **MTA Subway Delays** (Gov Data)
- **MTA Ridership** (Gov Data)
- **Rideshare Price Data (Uber-NYC)** (Kaggle)
- **Weather Data (NY – Central Park)** (EPA Data)

Why These Datasets?

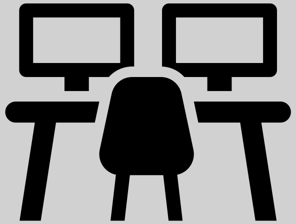
- **High-quality, publicly available, historical trends** from reliable sources.
- Captures **key external factors affecting price fluctuations.**

** Data Sources*

Machine Learning Techniques

Approach:

1. **Data Preprocessing**
   1. Handle missing values, normalize features, encode categorical data.
2. **Feature Engineering**
   1. Identify key predictors (e.g., delays, ridership, temperature).
3. **Modeling Techniques**
   1. **Regression models (Linear, XGBoost, Random Forest)** for fare prediction.
   2. **Classification models** for surge pricing prediction.
4. **Evaluation Metrics**
   1. **Regression:** RMSE, MAE
   2. **Classification:** Accuracy, F1-score



Most Effort Required In:

1. **Feature selection & engineering** to integrate different datasets effectively.
2. **Hyperparameter tuning** for model optimization.

Risk & Mitigation Strategies

RISK	MITIGATION STRATEGY
Incomplete Or Missing Data	Use imputation techniques, remove anomalies
Model Underperformance	Experiment with various ML algorithms & tuning
Data Integration Challenges	Standardize & preprocess all datasets
Stakeholder Needs Change	Adjust model features based on new insights
Computational Limitations	Use cloud-based training if required