Predicting School Bus Breakdowns and Delays in NYC

1. Project Team-B

- Team B Members: Nasim Aalemi, Shobha Panthi, and James Gilmore
- Title: Predicting School Bus Breakdowns and Delays in NYC

2. Project Overview

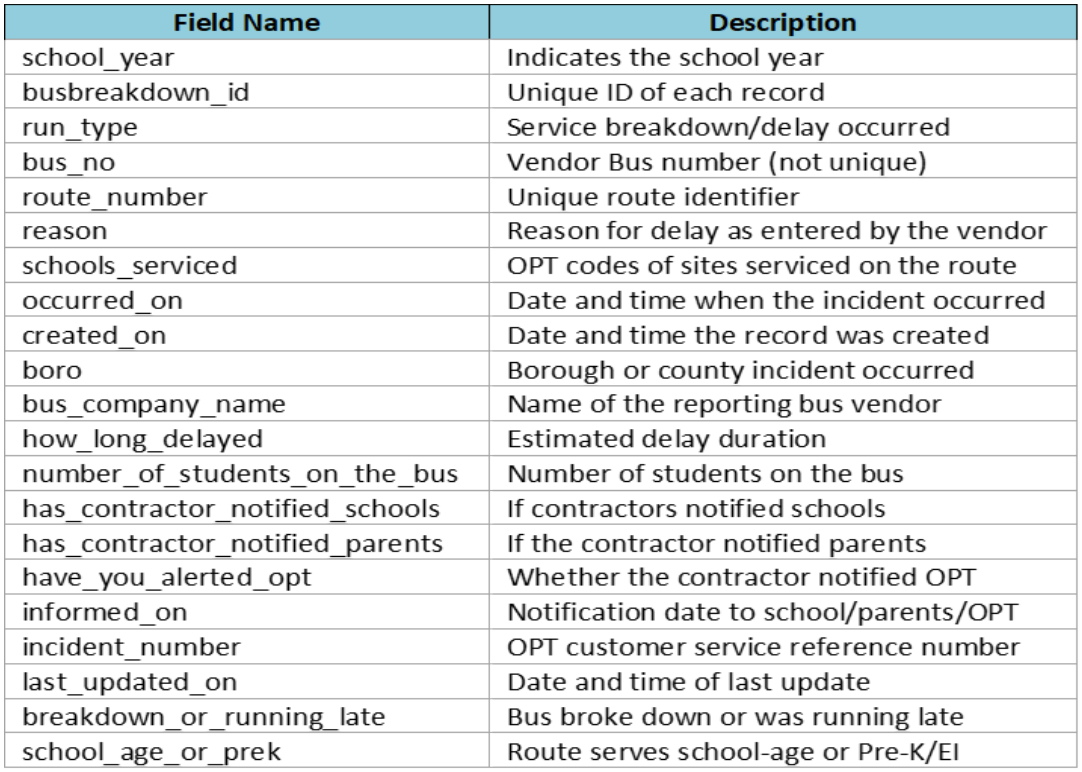
This capstone project aims to analyze and model school bus breakdowns and delays in New York City using a large-scale dataset containing 747,173 incident records. In the number one public school system in the United States, a reliable school transportation system is essential for students. Travel breakdowns affect student learning, family logistics, and long term academic success. Our project will combine exploratory data analysis, machine learning, and interactive dashboards to uncover patterns, predict outcomes, and offer actionable insights. We are specifically aiming to answer what operational factors contribute most to school bus delays and breakdowns in NYC, and can we accurately predict future incidents based on route, contractor, and historical patterns.

3. Data Source

- Primary Dataset and Metadata Link: https://data.cityofnewyork.us/Transportation/Bus-Breakdown-and-Delays/ez4e-fazm/about_data

4. Dataset Summary

The dataset covers several academic years and includes: Incident Information, Operational Details, Contextual Data, Response Behavior. Detailed information provided in below table.



Field Name	Description
school_year	Indicates the school year
busbreakdown_id	Unique ID of each record
run_type	Service breakdown/delay occurred
bus_no	Vendor Bus number (not unique)
route_number	Unique route identifier
reason	Reason for delay as entered by the vendor
schools_serviced	OPT codes of sites serviced on the route
occurred_on	Date and time when the incident occurred
created_on	Date and time the record was created
boro	Borough or county incident occurred
bus_company_name	Name of the reporting bus vendor
how_long_delayed	Estimated delay duration
number_of_students_on_the_bus	Number of students on the bus
has_contractor_notified_schools	If contractors notified schools
has_contractor_notified_parents	If the contractor notified parents
have_you_alerted_opt	Whether the contractor notified OPT
informed_on	Notification date to school/parents/OPT
incident_number	OPT customer service reference number
last_updated_on	Date and time of last update
breakdown_or_running_late	Bus broke down or was running late
school_age_or_prek	Route serves school-age or Pre-K/EI

Targets for Prediction:

- Classification: Breakdown or Running Late
- Regression: How Long Delayed

The dataset is large enough to support deep learning, ensemble modeling, and time-based analysis, with ample records for training and validation.

Modeling Approach and Analysis

Exploratory Data Analysis (EDA)

- Identify common reasons for delays
- Analyze temporal trends (monthly, seasonal, yearly)
- Examine delay patterns by borough, bus company, and route
- Explore relationships between response behavior and delay outcomes

Classification Model

- Goal: Predict if a bus incident will result in a breakdown or running late
- Input: Borough, company, run type, reason, and student count.
- Model: Logistic Regression, Random Forest, XGBoost
- Metric: Accuracy, F1-Score, Precision, Recall

Regression Model

- Goal: Predict delay duration in minutes, Input: All contextual and operational features
- Model: Linear Regression, Random Forest Regressor, Gradient Boosting
- Metric: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R-squared (R^2)

5. Frontend Implementation

Power BI Dashboard or Visual Analytics

- Map View: Delay density by city or borough, Charts: Delay reason frequency, Average delay by company and route, YoY trends, Filters/Slicers: School year, company, borough, and delay reason, KPIs: % of breakdowns, Avg delay duration by route or company

Streamlit Web App or Interactive Prediction Tool

- User Inputs: Borough, route, bus company, and reason
- Model Outputs: Classification: Breakdown or Running Late
- Regression: Estimated delay in minutes

6. Conclusion

This project will demonstrate how large-scale operational data can be leveraged to improve urban public services. Through combining data science, predictive modeling, and intuitive interfaces, this capstone aims to help NYC’s Department of Education and transit partners make more informed, proactive decisions about school bus logistics.