Project Report - Group 5

Port Authority of New York and NJ Data Analytics Project

BANL 6900-01 - Business Analytics Capstone

A. Understanding of the Business Goals:

The Port Authority of New York and New Jersey manages a complex transportation network of bridges, tunnels, and the Port Authority Bus Terminal that serves millions of commuters annually. To improve traffic flow, reduce congestion, and optimize toll operations, the corporation has outlined five central questions that guide this project.

We aim to -

1. Find the key factors that influence traffic at Port Authority facilities.
2. Measure toll violations by time and facility to guide enforcement.
3. Identify the busiest times of the year, month, and week and what causes them.
4. See how congestion patterns changed after the 2025 toll adjustments.
5. Forecast traffic beyond 2025 to support future planning.

B. Tools we plan to use -

Python, Azure AutoML, and Looker Studio / Power BI

- **Python -** will serve as the primary tool for data preparation, cleaning, joining datasets, and machine learning. Libraries such as `pandas and numpy` will manage and transform the datasets, while `matplotlib` and `seaborn` will support visualization and `scikit-learn, statsmodels, xgboost,` and `prophet` will support regression, classification, and forecasting tasks.

- **Azure AutoML -** will be applied for the forecasting task. By automatically testing multiple algorithms and hyperparameter settings, AutoML ensures that the most accurate and reliable forecasting model is selected.

- **Looker Studio / Power BI** - will be the final visualization platform for this project.

- **SQL –** Might use for extracting, cleaning, and preprocessing and joining but priority to do with python (Optional).

C. Algorithms and Models planned to use for each goal:

Goal	Model / Algorithm	Rationale for Choice
1. Identify Top Usage Factors	Multiple Linear Regression with Ridge regularization	We will use Multiple Linear Regression with Ridge regularization to identify the top five factors influencing facility usage. This approach provides clear insights while handling predictors like weather and seasonality.
2. Toll Violations	Logistic Regression and XGBoost Classifier	Logistic Regression will serve as the baseline for predicting toll violations, while XGBoost will enhance accuracy and handle class imbalance.
3. Busiest Times - Time Series	SARIMAX, Prophet	SARIMAX and Prophet will capture seasonal and holiday effects for time series forecasting.
4. Pricing & Congestion Shifts	Panel Regression and Difference-in-Differences (DiD)	Panel regression and difference-in-differences will be used to evaluate whether the 2025 pricing changes caused traffic to shift between facilities.
5. Forecasting Beyond 2025	Azure AutoML	Azure AutoML will automate model testing and selection, ensuring the most accurate and the best model.

D. Rationale for Choosing These Tools and Models:

- **Goal 1 - Ridge Regression** was chosen because it clearly shows how different factors affect traffic while handling correlations between variables.
- **Goal 2 - Logistic Regression and XGBoost** provide a good balance: logistic regression is simple and easy to explain, while XGBoost improves accuracy and captures complex patterns.
- **Goal 3 - SARIMAX and Prophet** are effective time series models that capture seasonality, holidays, and long-term patterns in traffic.
- **Goal 4 - Panel Regression with DiD** helps us measure the true impact of toll pricing changes by comparing before and after periods across facilities.
- **Goal 5 - AutoML** automates model testing and selects the best one, making long-term forecasting both accurate and scalable.