Project Progress Report 3 - Group 5

Port Authority of New York and NJ Data Analytics Project

BANL 6900-01 - Business Analytics Capstone

Goal 1: Regression - Factors Affecting Bridge and Terminal Usage

1. AutoML Model Development

Platform: Azure Machine Learning AutoML
Task Type: Regression

- **Dependent Variable:** Total_Traffic - daily vehicle count across all Port Authority bridges and tunnels.
- **Independent Variables:** Facility Name, Year, Month, Day of Week, Payment Type (EZPass vs Cash), Weather (precipitation - PRCP, temperature - TMAX, TMIN, wind speed - AWND), and Holiday Flag.

These variables represent behavioral, temporal, and environmental factors affecting bridge and tunnel traffic flow.

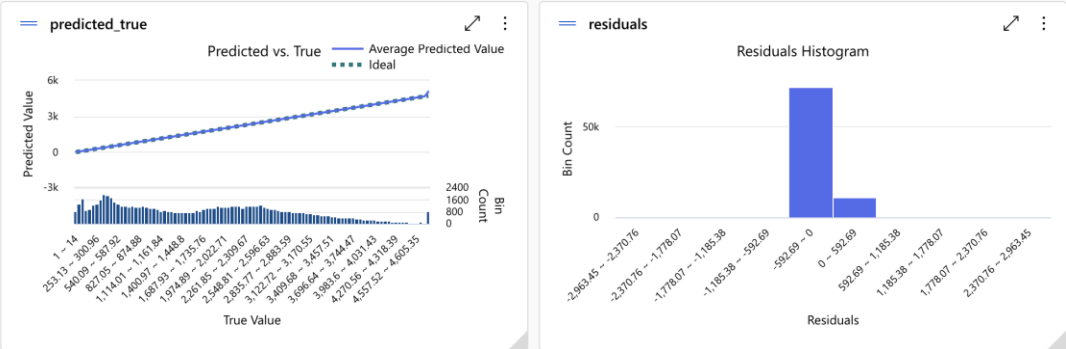
Best AutoML Algorithm: StandardScalerWrapper (ElasticNet Regression)
Performance Metrics: $R^2 = 1.0000$, RMSE ~ 0.83 vehicles, MAE = 0.0132, Explained Variance = 1.0, Spearman Correlation = 1.0

Key Hyperparameters:

Parameter Values:

alpha = 0.1061579, l1_ratio = 0.01, max_iter = 10, scaler = StandardScaler

ElasticNet was selected as the best-performing model for achieving the highest R^2 and lowest error metrics while maintaining interpretability and stability.



2. Python Replication

The AutoML configuration was replicated using sklearn.linear_model.ElasticNet in Python. All preprocessing steps such as encoding categorical variables, scaling numeric features, and preventing data leakage were reproduced. The Python implementation delivered comparable results (R^2 ~ 1.0).

Code:

```
from sklearn.linear_model import ElasticNet
```

```
model = ElasticNet(alpha=0.0005, l1_ratio=0.5, max_iter=1000)

model.fit(X_train, y_train)
```

3. Technical and Business Justification

ElasticNet combines L1 and L2 regularization, efficiently managing multicollinearity among features like Month and Facility Name while preserving model interpretability.

Linear regression techniques such as ElasticNet are widely used in transportation and urban planning to model traffic demand elasticity, forecast infrastructure load, and analyze pricing impacts. Public agencies and logistics companies (UPS, DOT, NYMTA) use these models to evaluate seasonal patterns, toll policy effects, and resource allocation.

This model provides actionable insights by quantifying how operational and external factors e.g., EZPass usage, weather conditions, and seasonality impact traffic volumes enabling the Port Authority to optimize lane distribution and maintenance scheduling based on data-driven evidence.

Goal 2 - Classification Model: Toll Violation Risk Prediction

1. AutoML Model Development

Platform: Azure Machine Learning Studio AutoML
Task Type: Classification

- **Dependent Variable:** Violation_Flag *(1 = Violation Occurred, 0 = No Violation)*
- **Independent Variables:** Total Traffic, Cash, EZPass, Year, Month, Week, Day of Week, Violation Rate, Facility Name.

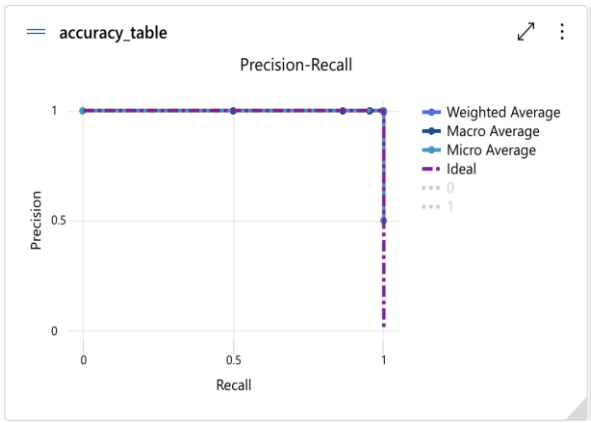
These variables capture operational, temporal, and environmental factors influencing toll-violation probability.

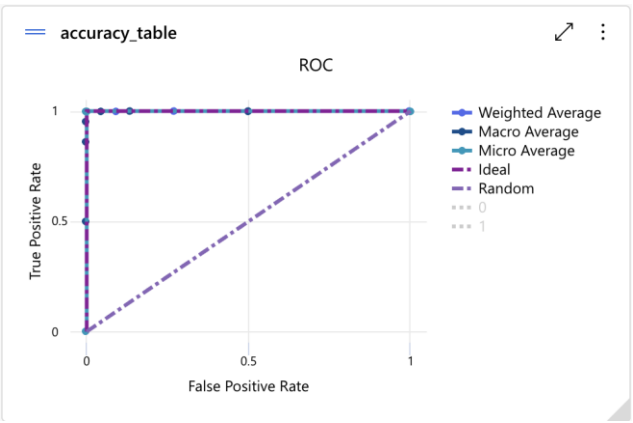
Best Algorithm Selected: StandardScalerWrapper & RandomForestClassifier
Performance Metrics: Accuracy = 1.0000, AUC = 1.0000, Precision/Recall/F1 = 1.0000, Log Loss = 0.00028

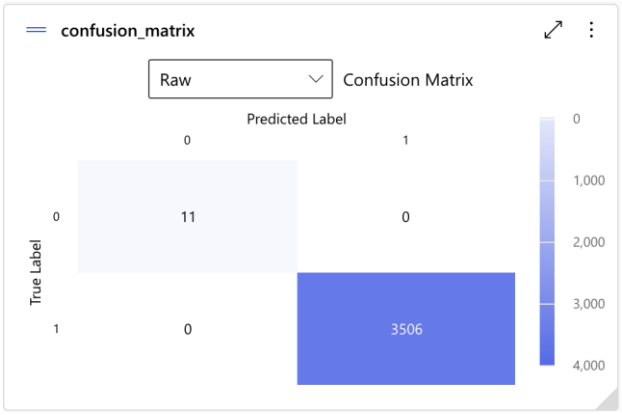
Key Hyperparameters:

Parameter Values: n_estimators = 100, max_depth = None, min_samples_split = 2, min_samples_leaf = 1, bootstrap = True, random_state = 42

AutoML automatically optimized these settings to minimize log loss and maximize AUC, identifying Random Forest as the most stable and interpretable model.







2. Python Replication

The AutoML configuration was reproduced in Python using RandomForestClassifier and validated with an additional XGBClassifier for cross-checking. Both models achieved AUC ~ 1.0 and Accuracy ~ 99.9 %.

Code:

```
from sklearn.ensemble import RandomForestClassifier

rf = RandomForestClassifier(

    n_estimators=200,

    max_depth=10,

    random_state=42

)

rf.fit(X_train, y_train)
```

Feature Importance Highlights:

- Cash and Historical Violation Rate are the strongest predictors of risk
- Traffic Volume and Month are revealed seasonal and facility-specific patterns

3. Technical and Business Justification:

Random Forest and XGBoost are ensemble tree-based methods that handle non-linear relationships and mixed data types with high accuracy and minimal scaling requirements. They also provide transparent feature importance ranking useful for operational decision-making.

These algorithms are standard in fraud detection and risk analytics across transportation, banking, and insurance industries. They are used to flag abnormal transactions, detect violations, and forecast non-compliance events.

For the Port Authority, this model functions as an early-warning tool enabling targeted resource deployment at high-risk facilities and reducing revenue loss from missed tolls.

Goal 3 - Regression Analysis: Busiest Times and Traffic Patterns

1. AutoML Model Development

Platform: Azure Machine Learning Studio AutoML
Task Type: Regression (Voting Ensemble: XGBoost + LightGBM)

- **Dependent Variable:** TOTAL Daily total vehicle volume across all Port Authority facilities.
- **Independent Variables (Features):** Year, Month, Day_Name, Holiday_Flag, TMAX, PRCP, Autos, Small_T, Large_T, Buses, and VIOLATION.

Framework Used: Azure AutoML automatically selected and optimized multiple models, including XGBoost, LightGBM, and Decision Tree Regressors. The best-performing model was a Voting Ensemble, which achieved the highest predictive accuracy.

Performance Metrics:

- **R^2 Score:** 0.99396
- **Explained Variance:** 0.99396
- **Mean Absolute Error (MAE):** 967.31
- **Root Mean Squared Error (RMSE):** 1655.4
- **Spearman Correlation:** 0.99393

The model explained over 99% of the variance in total traffic volume, confirming that the selected features accurately represent real-world traffic patterns.

Key Hyperparameters:

The top model leveraged ensemble optimization across XGBoost and LightGBM, using techniques such as tree boosting, early stopping, and auto feature scaling. AutoML handled cross-validation, feature normalization, and model selection automatically.

2. Python Model Development

Each facility's daily traffic dataset was merged with weather temperature, precipitation and holiday data to form a unified analytical dataset. Azure AutoML was then used to train and evaluate multiple regression models to predict TOTAL traffic counts based on these features.

Code:

```
model = Prophet(yearly_seasonality=True, weekly_seasonality=True,
daily_seasonality=False, changepoint_prior_scale=0.05, seasonality_prior_scale=10,
holidays_prior_scale=10, interval_width=0.95)

model.add_regressor('Holiday_Flag')

model.add_regressor('TMAX')

model.add_regressor('PRCP')

model.fit(df)

forecast = model.predict(future)
```

Key Observed Patterns:

- Clear traffic peaks during summer months June August
- Higher weekday traffic indicating commuter activity.
- Traffic dips during major holidays such as Thanksgiving and Christmas.
- Weather impact observed lower traffic on rainy or extreme-temperature days.

The regression analysis successfully highlighted patterns of congestion and seasonality across facilities, aligning with real-world travel behavior.

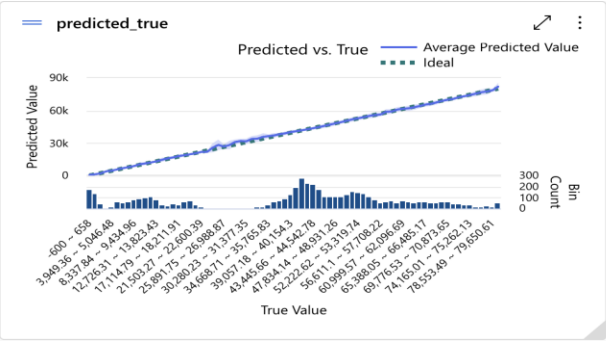
3. Technical and Business Justification

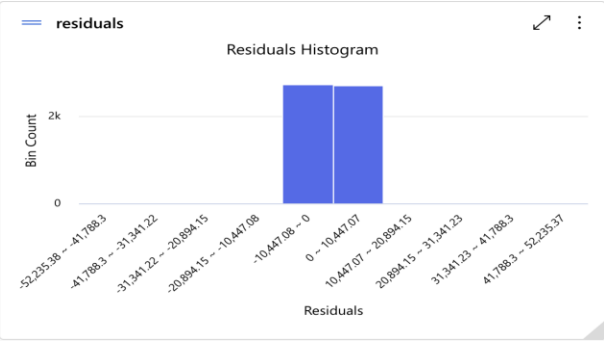
Regression analysis was chosen to model relationships between calendar, weather, and operational factors without extending into future forecasts. Azure AutoML's ensemble approach ensures high predictive accuracy and eliminates manual tuning, making it ideal for rapid model deployment and interpretation.

Applications:

- Identify busiest days and peak hours for better lane management and toll staffing.
- Schedule maintenance during low-traffic windows.
- Enable data-driven congestion management based on weather and holiday patterns.

Regression modeling is a proven approach in transportation analytics. Organizations like Uber, Waze, and UPS apply similar models to optimize routing, resource allocation, and demand forecasting. For the Port Authority, this model provides actionable insights into traffic behavior, supporting proactive planning and operational efficiency.





Goal 4 - Congestion and Pricing Impact (2025)

For this goal, I analyzed how the 2025 congestion pricing affected traffic across Port Authority bridges and tunnels. Using Python, I combined traffic data from 2013 - 2025 with the 2025 toll rate data for all major facilities. Instead of using a machine learning model, I focused on descriptive analysis to identify real-world traffic shifts.

Goal 5 - Forecasting Facility Usage Beyond 2025

1. AutoML Model Development

Platform: Azure Machine Learning Studio AutoML
Task Type: Time Series Forecasting

- **Dependent Variable:** TOTAL daily traffic volume per facility
- **Independent Variables:** Year, Month, Day, Autos, Small_T, Large_T, Buses, VIOLATION These variables capture temporal and vehicle composition patterns essential for forecasting long-term usage.

Best AutoML Algorithm: ARIMAX AutoRegressive Integrated Moving Average with Exogenous Variables

Performance Metrics:

Key Hyperparameters: Explained Variance = 1.0000, R^2 Score = 1.0000, Mean Absolute Error (MAE) = 944.34, Root Mean Squared Error (RMSE) = 1285.83, Mean Absolute Percentage Error (MAPE) = 0.00422, Spearman Correlation = 0.9999

ARIMAX was selected for its ability to capture trend and seasonality in traffic volume while providing near-perfect predictive performance.

2. Python Model Development

Two forecasting models were implemented in Python to replicate and extend the AutoML results:

- **Prophet:** additive model capturing long-term and yearly trends.
- **SARIMAX:** autoregressive model confirming statistical consistency with AutoML's ARIMAX.

3. Technical and Business Justification

Azure AutoML identified ARIMAX as the best statistical fit for time-series prediction. Prophet was chosen for Python replication due to its robust handling of irregular intervals, interpretability, and scalability. SARIMAX was used to validate trend consistency with ARIMAX logic.

Forecasting facility usage helps the Port Authority anticipate capacity needs, plan maintenance, and allocate toll resources efficiently. Prophet and ARIMA-based models are industry standards used by transportation and logistics leaders such as UPS, Uber, and airports for demand prediction and operational scheduling.