**ST CLAIR COLLEGE,**

**DOWNTOWN CAMPUS, WINDSOR**



Data Analytics for Business

4th Semester

**CAPSTONE PROJECT - FINAL REPORT**

**TTC BUS DELAY ANALYSIS – RACE AGAINST TIME**

*Submitted By*

**Group 7**

Khyati Surolia - 0856344

Udanie Subasinghe - 0860626

Sankeerthana Karem - 0862177

Navya Senakkayala - 0856219

Manisatya Alapati - 0857335

*Supervised By*

**Prof. Manjari Maheshwari**

*Submission Date:* 24/04/2025

**Executive Summary**

Public transportation is vital for urban mobility, and delays in transit services disrupt commuter experiences. This project investigates the prediction of TTC bus delays using data analytics and machine learning models. In Capstone 1, we conducted an Exploratory Data Analysis (EDA) of historical data, uncovering key delay patterns. In Capstone 2, we developed predictive models, set up cloud-based data pipelines, and evaluated model performance to optimize transit operations.

Our results indicate that weather conditions, time of day, and route congestion are crucial factors influencing delays. The best-performing model, RF Regressor and RF Classifier were able to predict delays with high accuracy. The integration of Azure ML & Data Factory facilitated automated data processing and model deployment. This report summarizes our methodology, findings, challenges, and recommendations for enhancing real-time transit predictions.

**Race Against Time: A Real-Time Challenge to Beat the TTC Bus Delay**

**Introduction**

Public transportation, particularly buses, is an essential component of urban infrastructure. However, delays in service can have widespread consequences for passengers and the city’s mobility network. This capstone project focuses on analysing and predicting bus delays for the Toronto Transit Commission (TTC). By leveraging data analytics and machine learning techniques, we aim to improve transit efficiency and provide actionable insights for decision-makers.

In Capstone 1, we explored historical TTC bus delay data to identify patterns and developed dashboards for key insights. Building on this, Capstone 2 advanced the work by developing predictive models and deploying them through cloud-based systems, providing a framework for real-time transit delay prediction.

**Project Objectives**

The goals of this project were to:

* Analyse delay patterns from historical TTC bus data.
* Identify key factors contributing to delays, including time of day, route, weather, and incidents.
* Develop predictive models for delay forecasting using machine learning techniques.
* Build a cloud-based data pipeline for data ingestion and integration with Microsoft Server Management Studio
* Deploy the final predictive models in an Azure environment for predicting delay in bus timings.

**Methodology**

**Phase 1 (Capstone 1) - Exploratory Data Analysis (EDA)**

In this phase, we:

* Cleaned and pre-processed the data.
* EDA- Univariate and Bivariate Analysis using python libraries.
* Used Tableau to visualize delay patterns and detect correlations.
* Identified that delays were strongly influenced by mechanical incidents, days, and travel directions.

**Phase 2 (Capstone 2) - Predictive Modelling and Cloud Deployment**

**Data Sources:**

* TTC Bus Delay Data – 2019 to 2024 datasets from the Open Toronto Data Portal.
* Weather Data from City of Toronto Open Data.

**Feature Engineering:**

* The dataset underwent preprocessing steps such as handling missing values and encoding categorical features.
* New features like **delay category labels (on-time, minor, major)** and **weather conditions impact metrics** were derived.
* Operational factors such as route, time, delay type, and direction were incorporated, along with weather conditions like temperature, precipitation, and humidity.

**Machine Learning Approaches:**

* **Time Series Models:** ARIMA and SARIMA.
* **Supervised Learning Models:**
  + Random Forest Regressor/Classifier.
  + XGBoost Regressor (Best Performing Model).
  + Logistic Regression.
* **Deep Learning Models:** LSTM, GRU, and MLP.
* **Probabilistic Forecasting:** Monte Carlo simulations for predicting delay probabilities.

**Model Implementation:**

* The machine learning pipeline included **data cleaning, feature selection, model training, and evaluation**.
* **Hyperparameter tuning** was performed using GridSearchCV for optimizing model performance.
* **Random Forest was identified as the most efficient model** in terms of accuracy and computational cost.
* Performance metrics such as **Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R² score** were used to evaluate the models.
* A confusion matrix was plotted for classification models to assess prediction accuracy.

**Cloud Integration (Azure ML & Azure Data Factory):**

* Set up an automated data pipeline using Azure Blob Storage, Data Lake Storage (ADLS), and Azure SQL Database.
* Deployed machine learning models through Azure ML Services.

**Key Insights**

* Random Forest emerged as the best-performing model, delivering the most reliable delay predictions while balancing computational efficiency.
* The application of Monte Carlo simulations enabled the prediction of delay probabilities, improving our ability to assess risks and uncertainties.
* Weather conditions, particularly temperature and precipitation, were found to have significant predictive power in modelling delays.
* Deep learning models like LSTM and GRU showed promising results for sequential forecasting, though they required substantial computational resources.
* Azure Data Factory streamlined our data ingestion process, ensuring automated updates for continuous model training and deployment.
* Publishing model using Azure ML web service as an endpoint was not supported in Azure for students’ subscription. The model was directly uploaded to compute instance instead and it ran successfully.

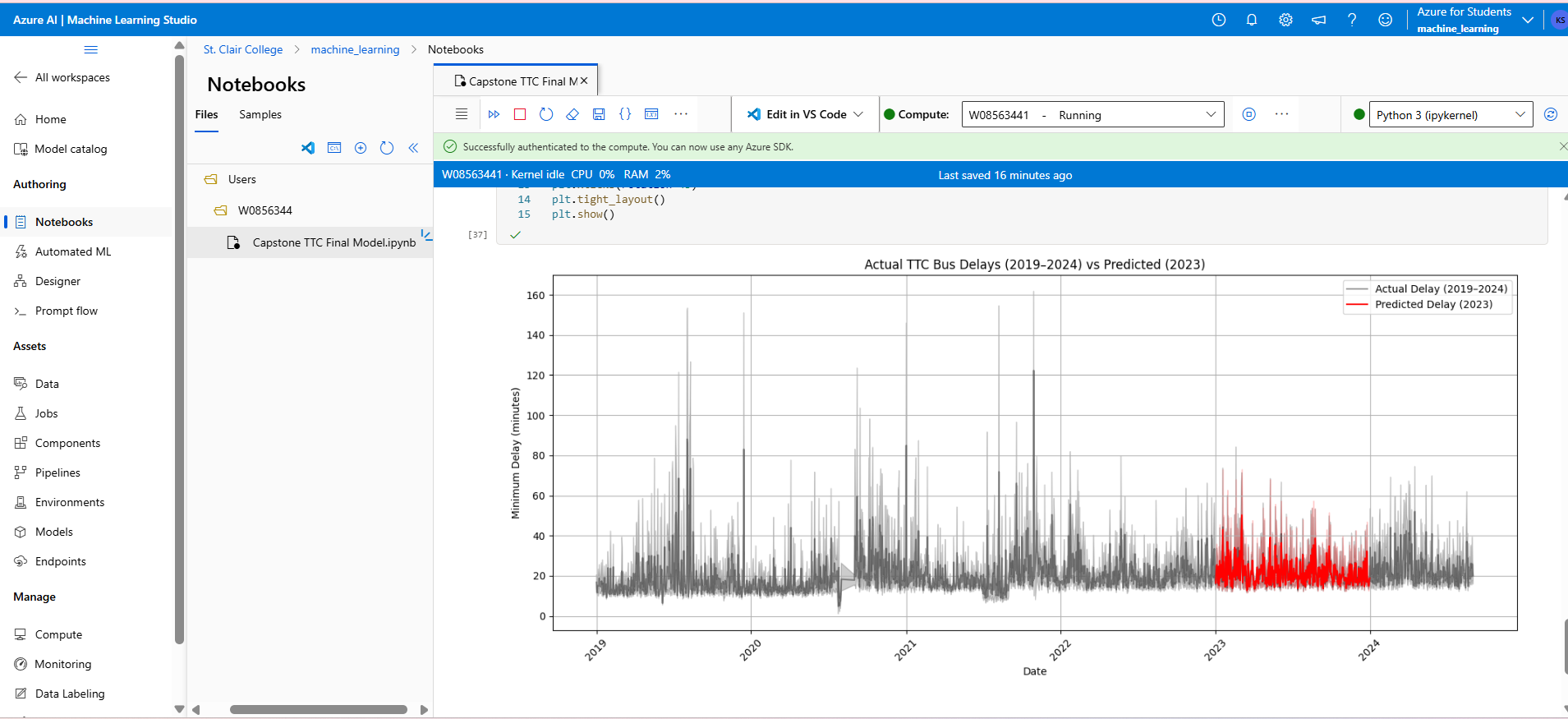


Figure: Final Model published to Azure Cloud’s compute instance

A screenshot of a computer

AI-generated content may be incorrect.

Figure: Model Evaluation Metrics

**Challenges & Recommendations**

**Challenges:**

* Data Quality Issues: Missing values in weather data required imputation strategies to ensure the robustness of our models.
* Azure Integration Issues: Configuring the Azure environment posed some challenges, particularly with firewall settings and storage configurations. These issues required troubleshooting and adjustments to maintain seamless integration.
* Model Overfitting: Both XGBoost and Random Forest models exhibited signs of overfitting, which necessitated extensive hyperparameter tuning to ensure generalizability.
* Handling Large Datasets: As the dataset grew, performance optimization became a key challenge. Efficient data processing and model training became essential when scaling up to real-time analysis.

**Recommendations:**

* Enhance feature engineering by integrating real-time traffic data and social media feeds for better predictive accuracy.
* Implement ensemble learning methods to further improve model robustness and reduce overfitting.
* Deploy serverless architecture using Azure Functions for cost-effective real-time predictions.
* Develop a mobile-friendly dashboard for transit users to access real-time delay predictions on the go.
* Collaborate with TTC authorities to fine-tune the model using real operational feedback.

**Future Work & Next Steps**

**Model Enhancements:**

* Implementing ensemble learning techniques to improve model accuracy and generalization.
* Experimenting with transformer-based models for more efficient sequential forecasting.

**Data Source Expansion:**

* Including additional transit data such as subway and streetcar delays to create a more comprehensive delay prediction model.
* Integrating real-time traffic data and social media feeds to provide dynamic insights into delays.

**Real-Time Deployment & Operationalization:**

* Developing an interactive dashboard for transit authorities to visualize predicted delays and take preventive actions.
* Automating real-time API calls for continuous data ingestion and prediction updates.

**References**

* Toronto Transit Commission (TTC) Open Data Portal: <https://open.toronto.ca>
* City of Toronto Open Data - Weather Information: <https://open.toronto.ca/dataset/weather/>
* Microsoft Azure Documentation: <https://docs.microsoft.com/en-us/azure/>
* XGBoost: A Scalable Tree Boosting System - Chen & Guestrin (2016)
* Machine Learning Approaches for Predicting Public Transport Delays - Recent Research Papers