1. Business Summary

- The goal was to develop forecasting models to predict how many ferry tickets will be redeemed and sold daily for the Toronto Island Park service, helping to avoid overcrowding or underutilization of equipment and staff.
- Two forecasting models - LightGBM and LSTM - were developed using 10 years of historical data. The data was cleaned by filling missing days and removing anomalies to ensure reliable predictions. Features were engineered to capture patterns like recent sales trends and special days that influence ferry usage.
- The models were tested across multiple time periods to confirm consistent accuracy. Prediction errors for redemptions and sales were up to 60% lower than those of the baseline model (Figure 1), helping reduce unexpected surges or drops in demand. Deploying these models will enable early forecasting, improving preparation during peak times and enhancing customer satisfaction.
- Feature importance analysis provided actionable insights for marketing strategies and resource planning - including potential price adjustments on less busy days.
- Regularly updating the models with new data will help prevent data drift and maintain performance.

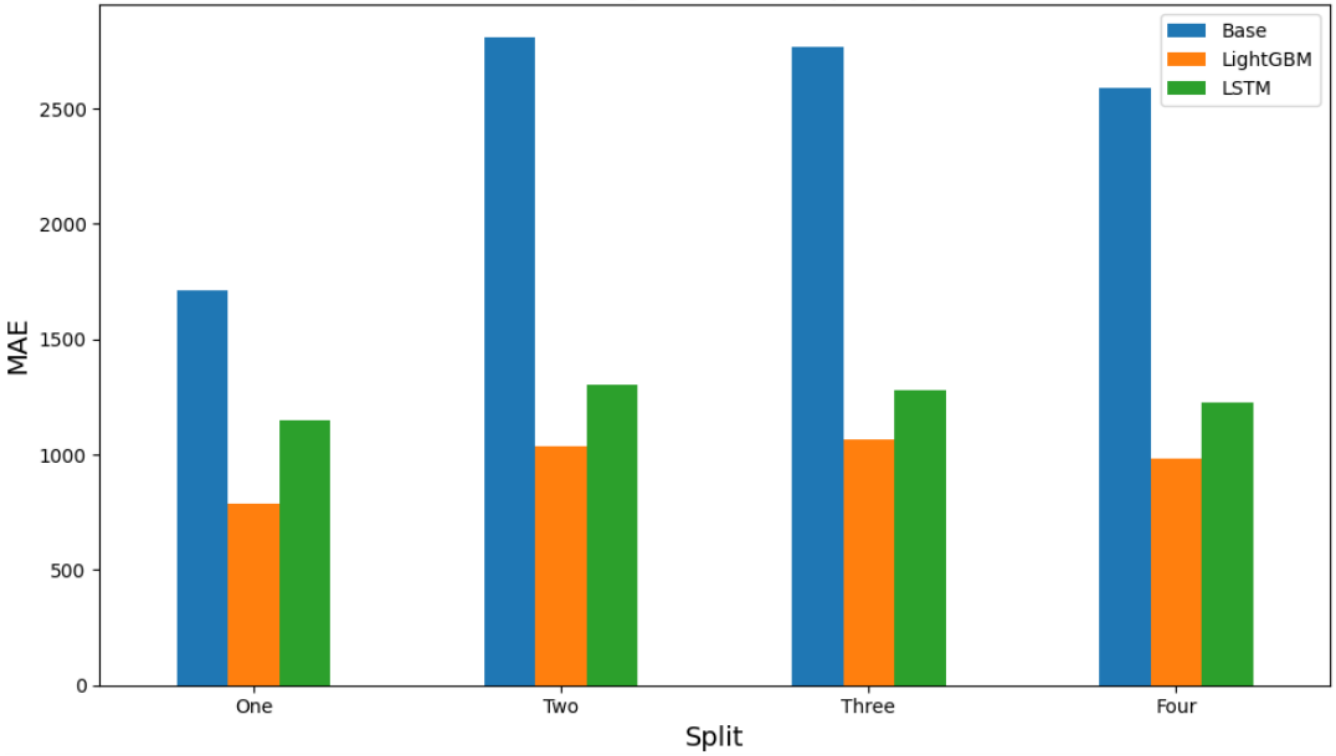

Figure 1: Comparison of mean absolute error between the baseline model, lighGBM and LSTM.

2. Technical Summary – Redemption Model

- The primary objective of this task was to develop a redemption forecasting model to predict daily ferry ticket redemptions for the Toronto Island Park service. Two advanced models were implemented and benchmarked against a seasonal baseline: LightGBM (LGB), a tree-based ensemble model, and Long Short-Term Memory (LSTM), a deep learning model for sequence data. Both were integrated into a modular and reproducible pipeline built entirely with open-source tools.
- The pipeline incorporated structured classes for data loading, exploratory analysis, outlier detection, missing value imputation, and feature engineering. The dataset was timestamp-indexed and aggregated at the daily level. Missing calendar days were imputed using historical averages for the corresponding day across different years. Outliers were detected using STL decomposition of the time series, followed by Isolation Forest with a contamination threshold of 0.001. Feature engineering introduced lag features, rolling averages, and key temporal indicators such as day-of-week, holidays, and seasonal patterns. Redundant or potentially leakage-inducing columns were excluded prior to model training.
- LightGBM was trained using time-series cross-validation with four splits and a test window of 365 days per fold. The model was configured with a learning rate of 0.03, a maximum depth of 4, 20 leaves, and L1/L2 regularization. Early stopping (50 rounds) was used to prevent overfitting. The LSTM model used a sequence length of 15, two hidden layers, dropout of 0.2, and was trained over 50 epochs.
- Both models effectively captured the underlying seasonal and trend patterns in redemptions. LightGBM achieved the most consistent and accurate results, reducing MAE by approximately 60% over the baseline model (Figure 2). LSTM also outperformed the baseline but was not fully fine-tuned due to time constraints.
- Future improvements may include hyperparameter optimization for the LSTM model, exploration of hybrid architectures, and incorporation of external factors such as special events or weather conditions to enhance forecast accuracy.

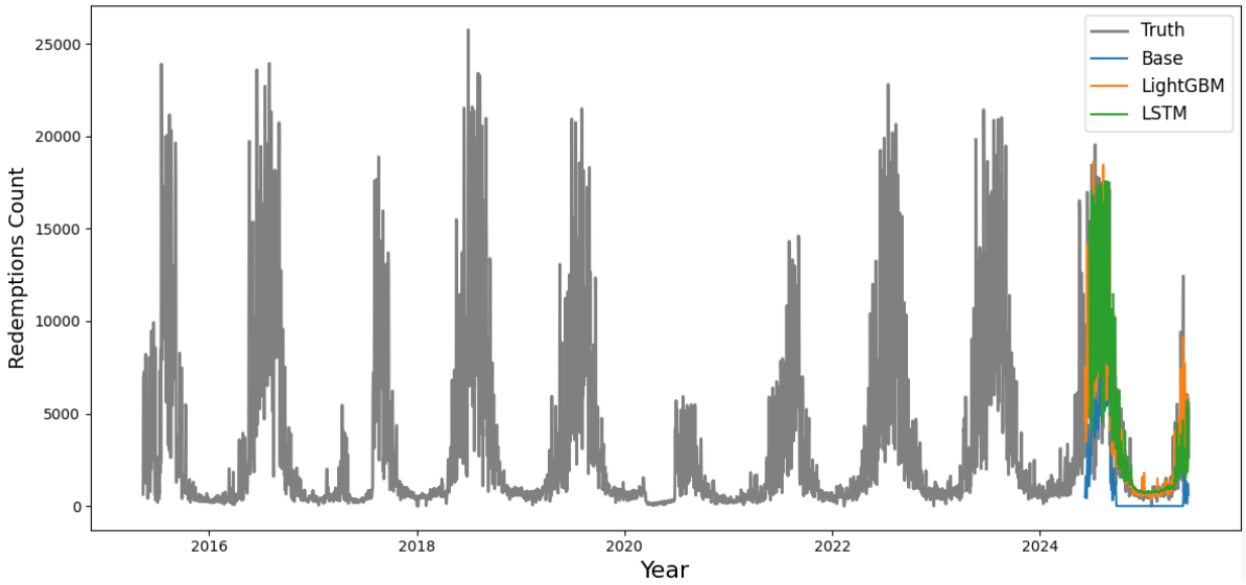

Figure 2: Comparison of ground truth with predictions from the baseline model, LightGBM, and LSTM.

3. Technical Summary – Sales Model

An additional forecasting model was developed to predict daily ferry ticket sales for the Toronto Island Park service. This model reused the same class-based architecture and modular pipeline as the redemption model, including components for data preprocessing, missing value imputation, outlier detection, and feature engineering. A LightGBM model was trained using time-series cross-validation and evaluated using mean absolute error (MAE). The model employed a low learning rate (0.03) and early stopping (50 rounds) to ensure stable convergence. Regularization (lambda_l1 = 0.5, lambda_l2 = 0.5) and a shallow tree structure (max_depth = 4) helped mitigate overfitting. Validation MAE closely tracked training MAE, indicating good generalization performance (Figure 3). Predicted sales aligned well with actual values over time, effectively capturing seasonal and trend components (Figure 4). SHAP (SHapley Additive exPlanations) was applied to interpret model predictions, revealing that key features included recent redemption lags, rolling averages, and indicators for weekends, seasons, and holidays. Future work may explore incorporating LSTM architectures similar to the one used for the redemption model.

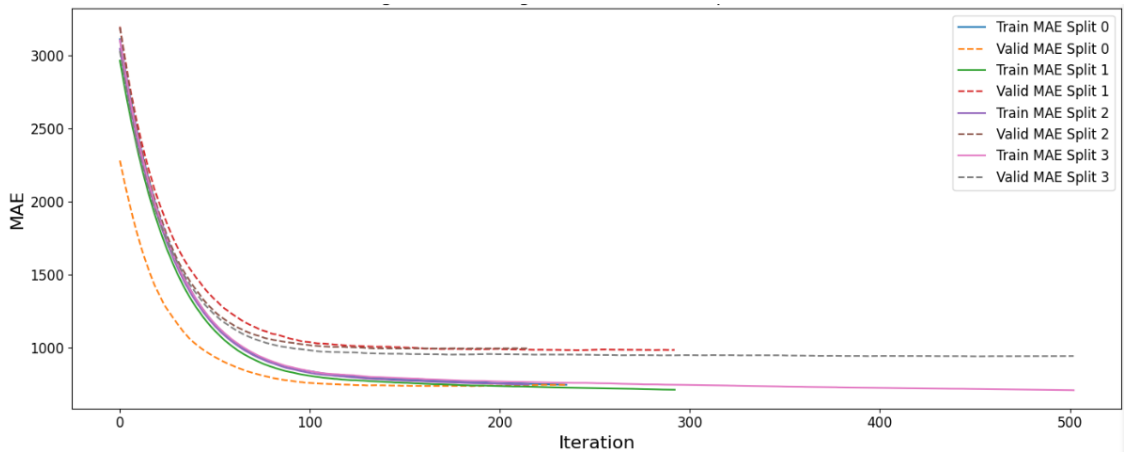

Figure 3: Comparison of training and validation mean absolute error (MAE) over iterations.

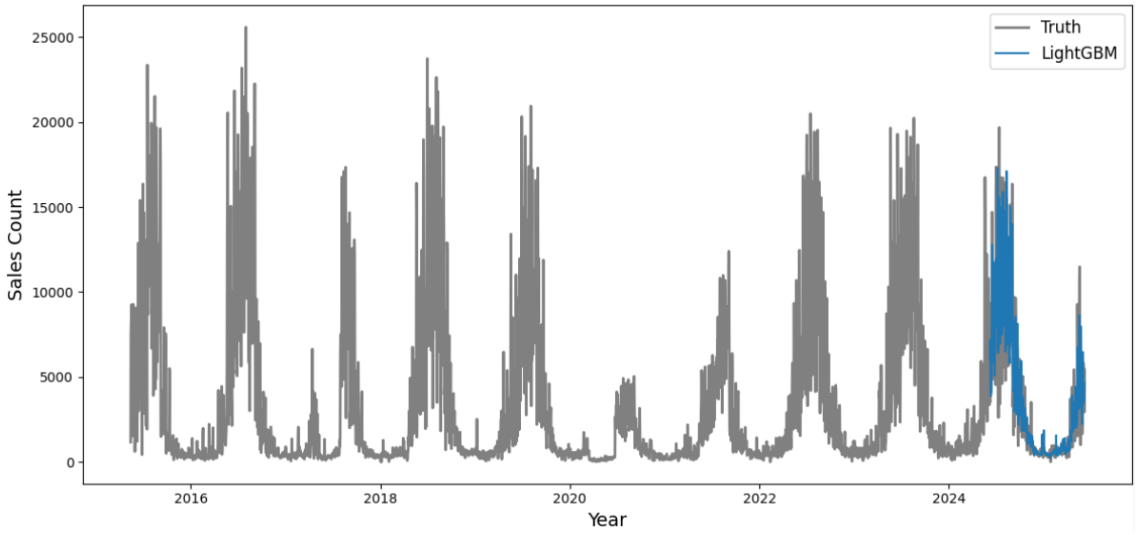

Figure 4: Comparison of ground truth with predictions from the LightGBM.

4. Others

Model Assumptions

The models assume stable seasonal patterns, accurate historical data, and that missing days were operational. It is also assumed that selected features capture key drivers and that residual errors are independent without significant autocorrelation.

AI Usage Disclosure

ChatGPT was utilized to assist with adding docstrings to the code, debugging, and formatting the report.