THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

RapidRoute Logistics: Traffic Prediction Model Report

Prepared By:
Bilal Sajid Channa

Date:
November 25, 2025

1. Introduction

The objective of this project was to develop a robust machine learning model capable of predicting hourly interstate traffic volume with high accuracy. RapidRoute Logistics faces significant operational inefficiencies due to unpredictable congestion on I-40 and I-75. Accurate forecasts are critical for optimizing delivery schedules, minimizing fuel waste, and improving customer satisfaction.

This report outlines the analytical approach, model architecture, and key findings derived from historical traffic data. It culminates in the recommendation of a specific predictive solution which balances aggressive pattern recognition with conservative stability measures to minimize risk.

2. Executive Summary

To accomplish the objective of this project, I developed and rigorously tested multiple predictive models, ranging from baseline regressions to advanced gradient boosting ensembles. The final selected model, the *Golden Blend*, achieved a predictive error **(RMSE) of ~256.7 cars per hour**, significantly outperforming the client's target threshold of **< 300 cars**.

Key Findings

- **Traffic is highly cyclical:** The single strongest predictor is the interaction between the hour of the day and the day of the week.
- **Weather acts as a suppressor:** While clear skies correlate with high volume, severe weather (snow, thunderstorms) acts as a "brake" on traffic, significantly lowering volume and increasing unpredictability.
- **Holidays disrupt patterns:** The analysis of dataset was not conclusive for this variable since we did not have enough data points to measure its true effect on traffic.

- **Ensembling reduces risk:** By combining a high-performance model with a stability-focused model, we reduced the risk of "overfitting" to noise, ensuring reliable predictions even during unseen validation checks to gauge model performance.

3. Model Architecture & Workflow

The solution is built on a **Stacked Ensemble** architecture. Instead of relying on a single algorithm, we combined two variations of **Histogram Gradient Boosting Regressors (HGBR)**, an algorithm known for its speed and efficacy to handle non-linear relationships.

The Two Components

The "Performance" Engine (70% Weight):

- **Focus:** Maximizing accuracy on "normal" days.
- **Optimization:** Tuned using Bayesian Optimization (Optuna) to find the perfect balance of learning rate (0.08) and tree depth (22).
- **Role:** Provides the baseline trend, capturing complex interactions like "Friday Evening Rush Hour."

The "Stability" Engine (30% Weight):

- **Focus:** Safety and robustness.
- **Configuration:** Uses a more conservative learning rate and explicitly includes detailed weather categorization (weather_final).
- **Role:** Acts as a "check." If the Performance engine predicts high traffic during a blizzard because it's a Monday, this engine pulls the prediction down to a realistic level.

The Workflow:

1. **Data Cleaning:** Raw data was processed to correct temperature scales (Kelvin to Celsius) and engineer time features (hour, dayofweek).
2. **Feature Engineering:** Created interaction terms (weather_final) and "Holiday Flags" to help the model distinguish exceptional days from routine ones.

3. **Validation:** Models were tested using 5-Fold Cross-Validation to ensure they weren't just memorizing the past but could generalize to the future.
4. **Ensembling:** The final predictions are a weighted average:

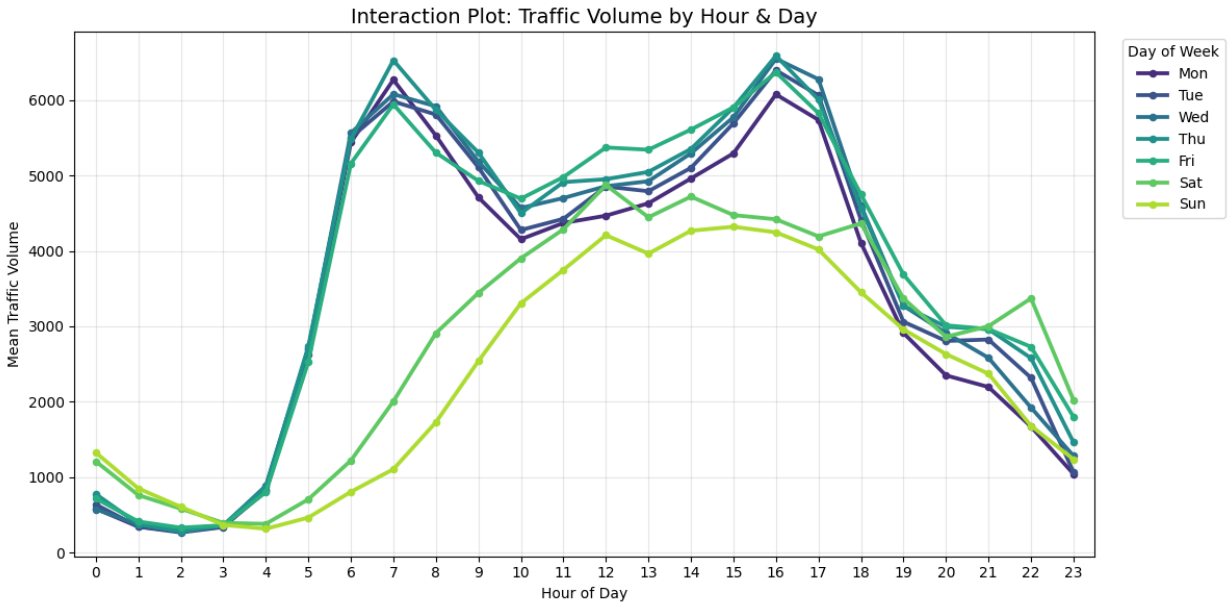
$$\textit{Prediction} = 0.7 \times (\textit{Performance Model}) + 0.3 \times (\textit{Stability Model})$$

4. Key Relationships & Data Insights

A. The Rhythm of the Road (Time & Day)

Traffic volume follows a distinct "heartbeat" that varies by day type. Weekdays (Mon-Fri), display a "Double Hump" pattern. The sharp peaks at 7:00–9:00 AM and 4:00–6:00 PM represent commuter traffic. On the other hand, weekends (Sat-Sun) display a "Single Bell Curve." Traffic rises slowly to a midday peak (12:00–2:00 PM) and tapers off, reflecting leisure travel.

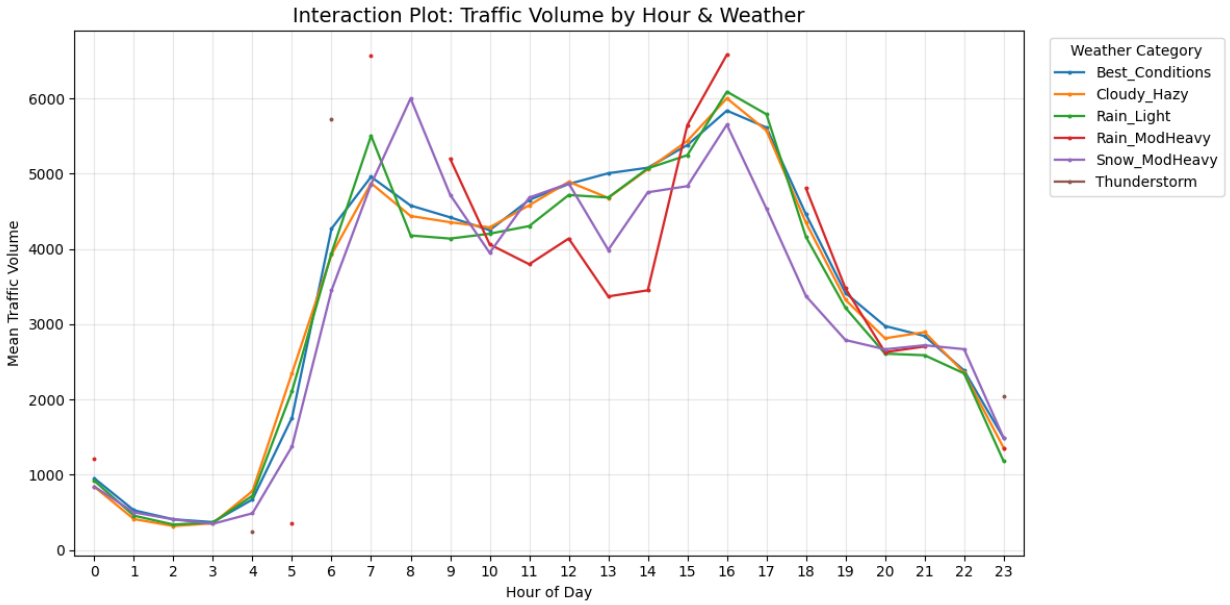
The model uses this interaction as its primary signal. It essentially learns 168 distinct hourly profiles (24 hours × 7 days) to set the baseline prediction.



B. The Impact of Weather

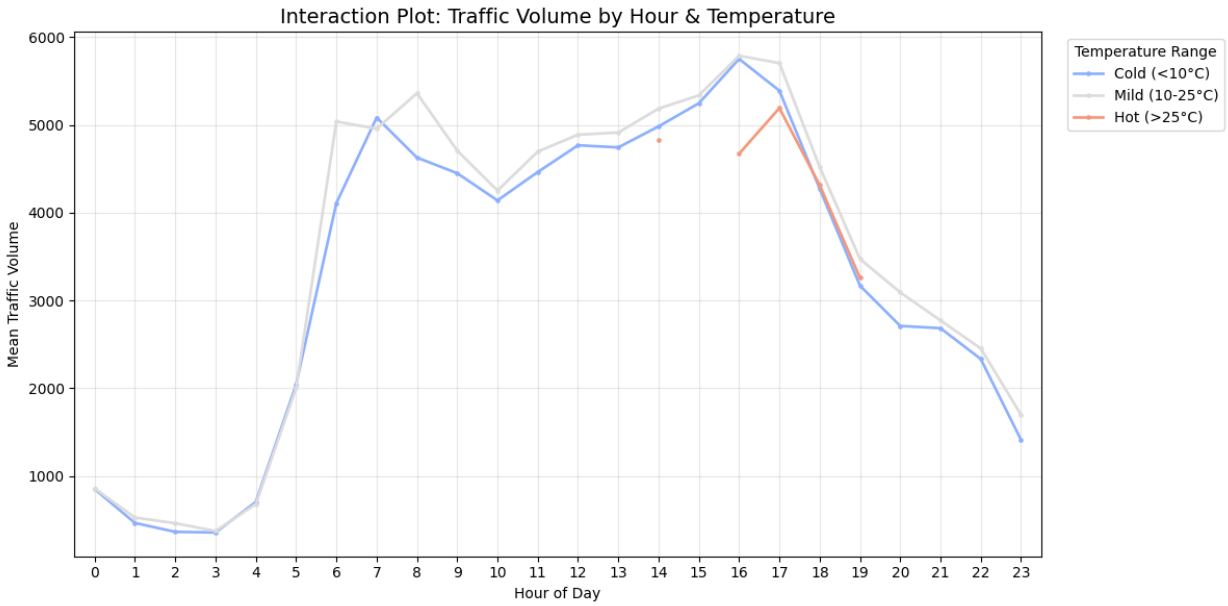
Weather conditions act as a hierarchy of constraints. "Best Conditions" (Clear/Cloudy) allow for maximum road capacity while "Rain" and "Mist" cause a slight reduction in volume but maintain the standard curve. The most significant drop in volume can be attributed to “Snow”. However, one must be cautious with the last observation since the dataset did not contain enough entries for us to say this with significant confidence.

The model uses weather as a penalty factor. It learns that 5:00 PM on a rainy Friday is not the same as 5:00 PM on a clear Friday, adjusting the volume downward accordingly.



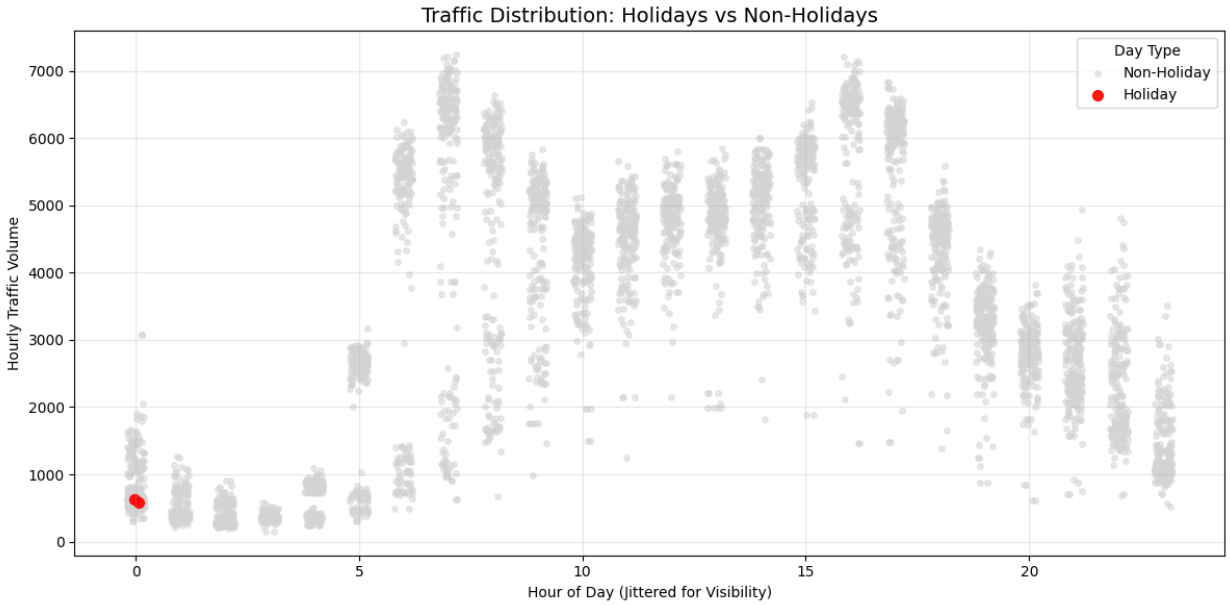
C. The Impact of Temperature

The relationship is non-linear and threshold-based. Traffic volume is robust and effectively identical across mild and hot temperatures. However, below a certain threshold (< 10°C), traffic volume consistently dips. The model does not use temperature given there is not enough significant difference for the model to tell apart noise from actual signals present in the data.



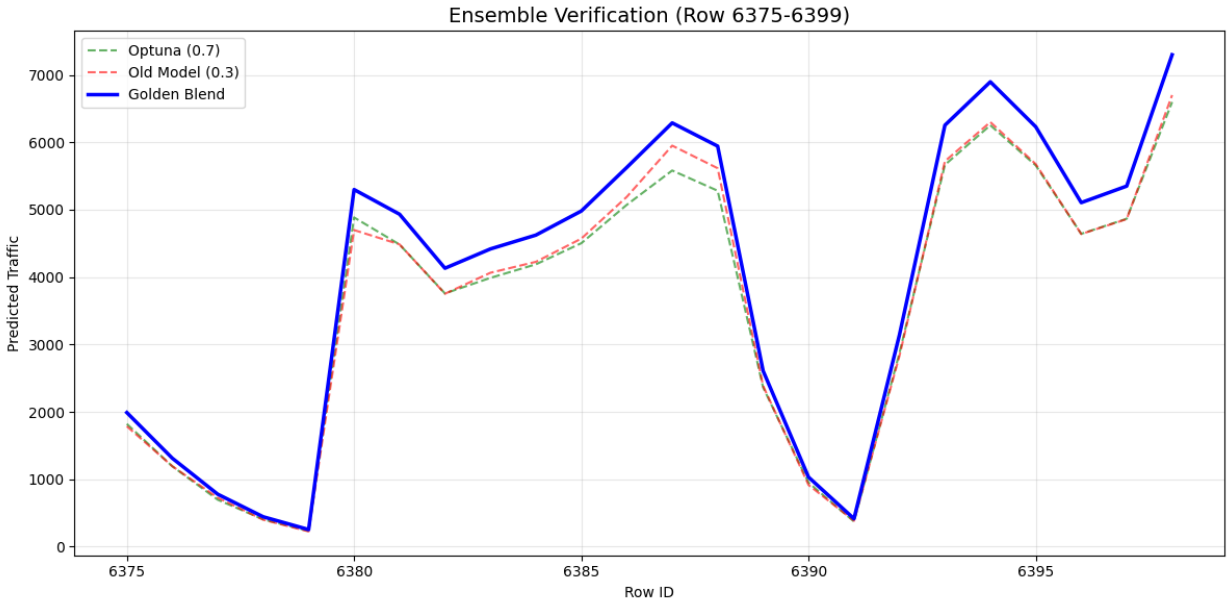
D. The Holiday Effect

Holidays act as "Pattern Disruptors." The scatter plot reveals that traffic on holidays (Red dots) is consistently lower than on non-holidays (Gray dots) for the same hours. However, owing to absence of datapoints, this feature was not useful.



5. Visualizing the "Golden Blend" Strategy

This visualization demonstrates the power of our ensemble approach during a period of high model disagreement (like a weather event). *Green Line (Performance Model)* predicts high volume based on historical averages. It is aggressive and accurate for normal patterns. The *Red Line (Stability Model)* predicts a sharp drop. It reacts strongly to a specific condition (e.g., snow). Ultimately, *Blue Line (Golden Blend)* is the final prediction, which navigates between them. By averaging the two, this "Model Stitching" is able to achieve **a low RMSE of 256.71**.



Appendix: Generalization Error (Model Zoo)

The following table summarizes the performance of the models developed during the hackathon phase. Lower RMSE indicates better performance.

Model Name	Algorithm	RMSE (CV)	Stability (Std Dev)	Role
Golden Blend	***Ensemble***	***~256.71****	***< 0.015***	***Final Model***
Lean Optuna	HGBR	258.56	0.011	Component A
Lean Clean	HGBR	260.23	0.010	Component B
Linear Baseline	Linear Regression	~1850.4	> 15.0	Discarded
N/A	Random Forest	423.04	8.773	Discarded
N/A	XGBoost	~300	18.496	Discard