Camera Traffic Counts: Final Project

By: John Trelford, Lucas Chiang, Travis Welsh, Brian Pham, and Ritesh Penumatsa
(Group 11)

Dataset Overview

Camera Traffic Counts

- 15-minute interval traffic by intersection in Austin
- Collected by GRIDSMART optical traffic detectors

Visual Crossing Weather Data

- Includes Austin climate conditions by day

Subsetted Data

How?

- Filter to only rows where year = 2019
- 14,717,624 rows in dataset where year = 2019
- Shape (rows, columns) = (14717624, 19)

Why?

- Prevents interference from the COVID-19 pandemic period
- The full Camera Traffic Counts dataset was 82.1 million rows
    - Overwhelms RAM
    - Computationally Intensive

Explanation Task:
What factors drive congestion at intersections?

Overall Goal

- We aim to identify the key features driving traffic volume to help people know under what conditions they should expect more or less traffic.
- This will allow people to more efficiently plan their excursions on the road so that they waste less time.

Goal

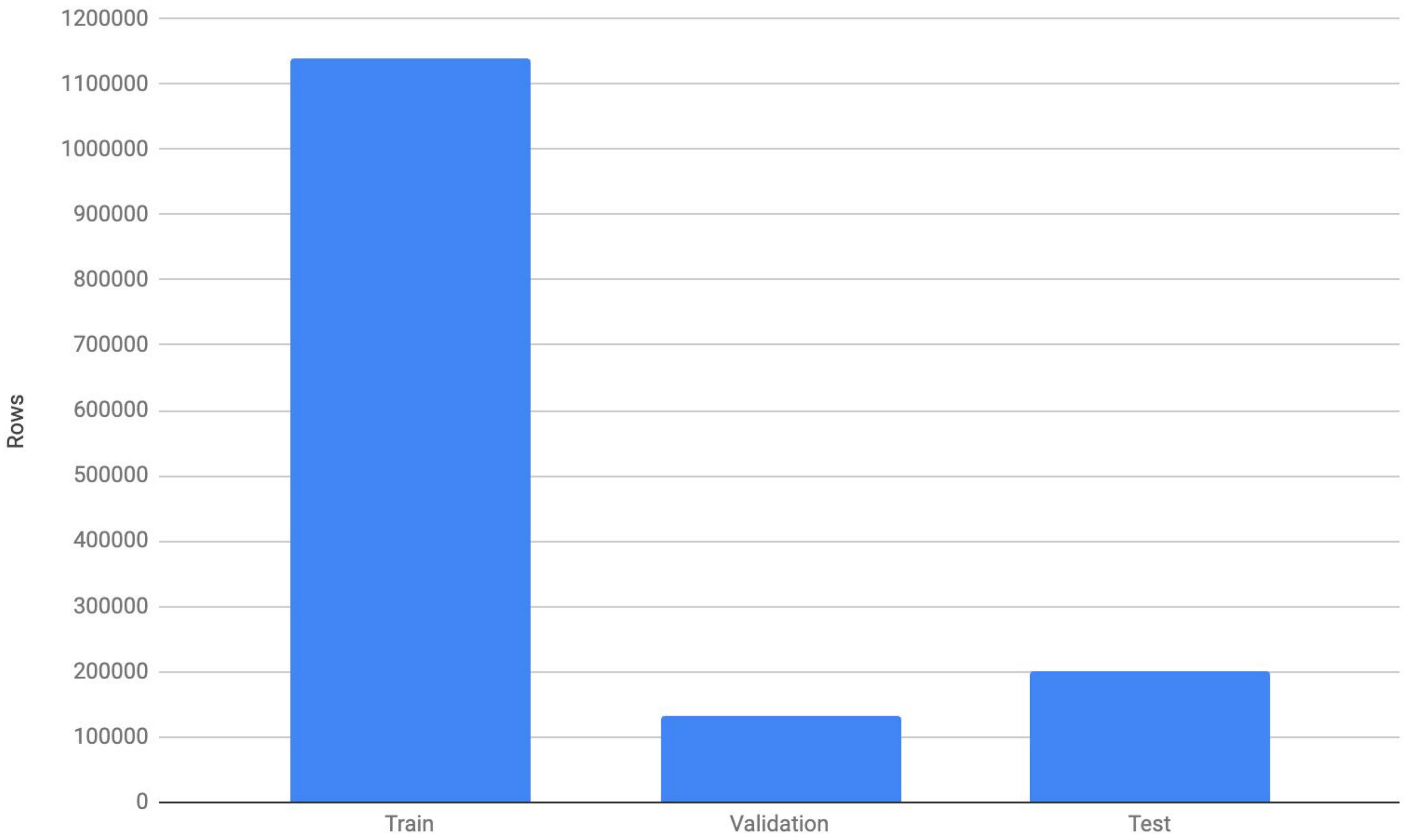
Our goal is causal understanding to inform individuals’ driving decisions by identifying factors that can reduce wasted time on the road.

Data & Choice of Features

- Target Variable: Volume
- Choice of Features:
  - Day of the Week - Travel patterns vary throughout the week
    - One-hot encoded to properly weight the categorical variable
  - Holiday - Travel patterns deviate around major holidays
  - Precipitation - Severe weather affects travel patterns
  - Heavy Vehicle Present - Larger vehicles affect turns
  - Time - Travel patterns vary throughout the day (ex: morning and afternoon rushes)
  - Lagged Volumes - Captures short-term dynamics
  - Visibility - Low visibility makes driving less safe; some drivers would be more cautious
  - Temperature - May be more/less likely to drive in drastic temperatures

Model Setup

- Splitting the dataset
  - Training set: January-September
  - Validation set: October-November
    - k-fold cross-validation is inappropriate because it breaks the time order
  - Testing set: December
  - Avoids temporal leakage so that only past data predicts future observations
    - Ensures an unbiased estimate of generalization because it simulates the real world scenario of predicting traffic volume without knowledge of the future.


Rows
1200000
1100000
1000000
900000
800000
700000
600000
500000
400000
300000
200000
100000
0
Train
Validation
Test

Justification of Feature Set

01 **Nature of Outcome**

Volume is a real-valued outcome

02 **Data Availability**

Data is available for every observation

03 **Data Integrity**

There is no data leakage

04 **Causal Structure**

All predictors are causally before volume

05 **Predictive Signal**

All features are expected to have a meaningful relationship with volume

Limitations

Avoidance of Temporal Leakage: By strictly partitioning the data chronologically, we ensure that the model has **no knowledge of the future** (temporal leakage), guaranteeing that evaluation metrics are **less biased estimates** of our model's performance.

Real-World Applicability: If the model performs well on the December test set, it provides confidence that the important features in the model reflect what factors are associated with changes in traffic volume.

Bias and Noise:

01

Seasonality

Seasonality dictates travel and commuter volume

02

Demographic change

Long-term population increases could dramatically impact traffic volume

03

Holiday Impact

The holiday period potentially impacts traffic patterns

04

Anomalies

Natural disasters, road closures, and pandemics can disrupt usual trends

If we used traffic data for more years, we could account for seasonality, demographic change, and holiday impact. Anomalies would likely require real-time adjustment to deal with. For us, some error is inevitable due to the *novelty* of the December test environment and only using 2019 data for training.

Model Choice

Model Class 1: Linear Regression

- Appropriate for temporal data with smoothly changing trends
- Easy to interpret the coefficients
- Low variance
- Low computational cost
- Avoids overfitting (few features yet millions of samples)
- Can optimize coefficients by L2 regularization

Model Class 2: Random Forest

- Able to model complex relationships between variables
- More resistant to outliers in the data
- Adapts to non-linear relationships

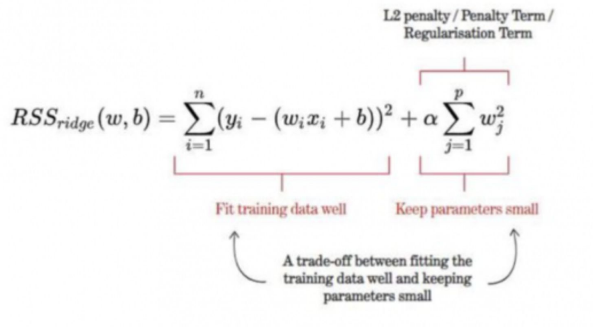
Model Choice cont.

Linear Regression Limitation

- Coefficients become unstable and unreliable
- Problem caused by multicollinearity (highly correlated features)
- Our one-hot encoding creates this issue

Ridge Regression Solution

- It is Linear Regression plus an L2 regularization term
- L2 penalty shrinks coefficient magnitudes toward zero
- Helps in preventing overfitting
- This shrinking stabilizes the coefficient estimates
- Result: Reliable coefficients for interpretation



$$RSS_{ridge}(w,b) = \underbrace{\sum_{i=1}^{n}(y_i - (w_i x_i + b))^2}_{\text{Fit training data well}} + \alpha \overbrace{\underbrace{\sum_{j=1}^{p} w_j^2}_{\text{Keep parameters small}}}^{\text{L2 penalty / Penalty Term / Regularisation Term}}$$

A trade-off between fitting the training data well and keeping parameters small

Hyperparameter Tuning

Ridge Regression

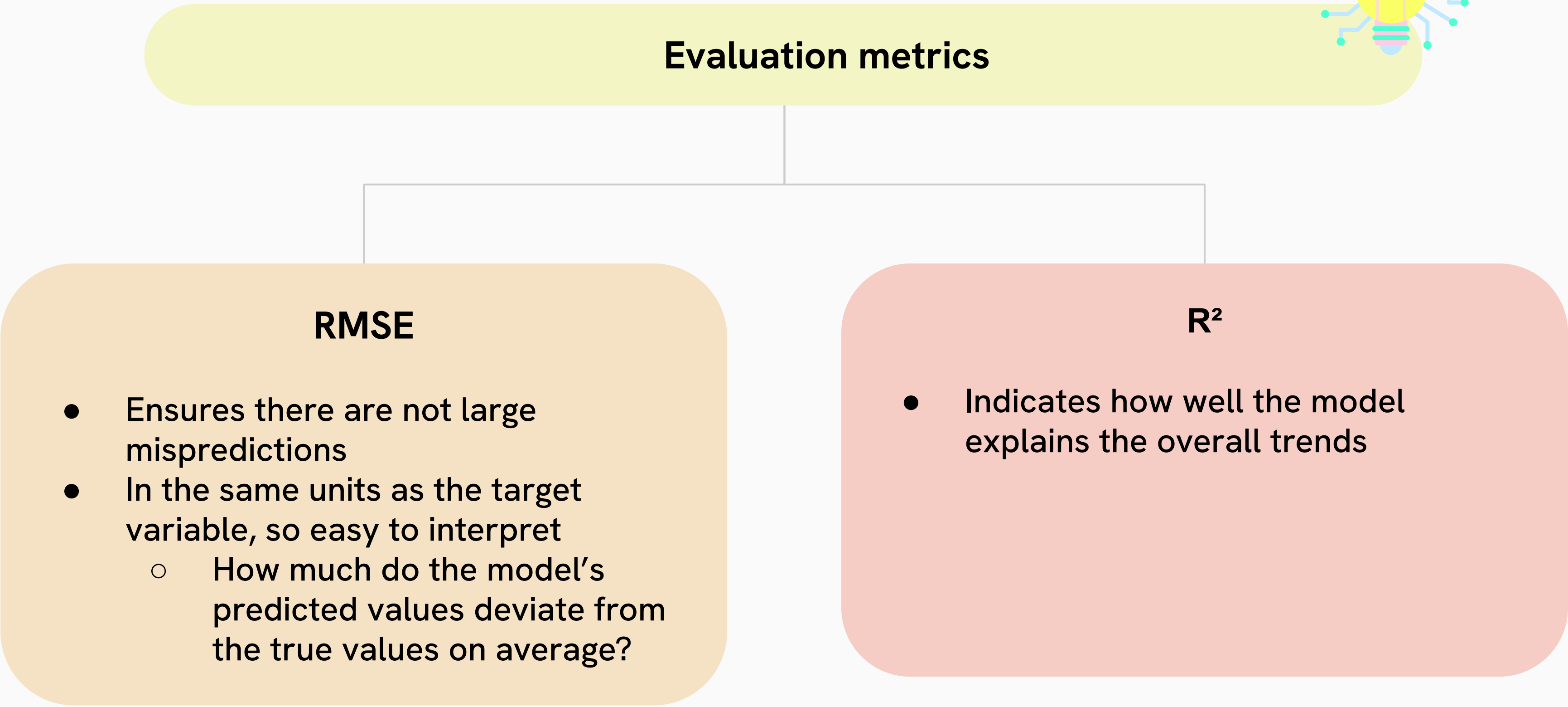
- ɑ, the strength parameter of the ridge regression penalty was tuned (.001 was selected from values: 0.001, 0.01, 0.1, .2, .3, .5)
- Search method: Minimizing RMSE with a list of parameters on the validation set
  - Penalizes large errors more heavily
  - Want to protect against large mispredictions

Random Forest

- N_estimators, max_depth, min_samples_leaf, and min_samples_split were tuned
- The search method was done on a predefined set of values and yielded the following results:
  - N_estimators = 50
  - Max_depth = 9
  - Min_samples_leaf = 2
  - Min_samples_split = 5

Note: We specifically didn't use GridSearchCV or RandomizedSearchCV to avoid temporal leakage hence why most of the parameters were manually tuned on the validation set (October - November).

Evaluation Metrics



Evaluation metrics

RMSE

- Ensures there are not large mispredictions
- In the same units as the target variable, so easy to interpret
  - How much do the model's predicted values deviate from the true values on average?

R^2

- Indicates how well the model explains the overall trends

Ridge Regression Model

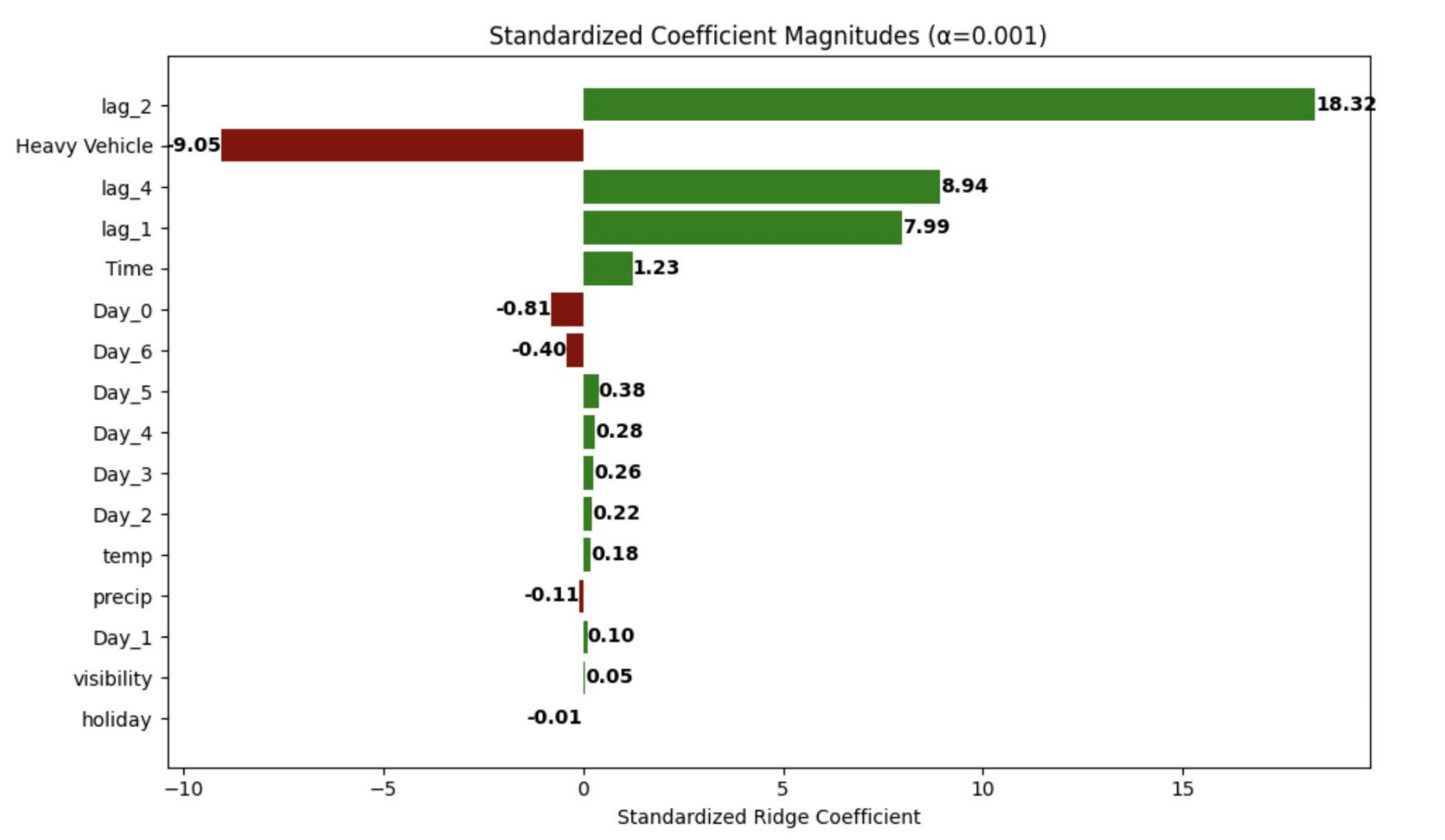
Evaluation and Results

Statistically Significant Features

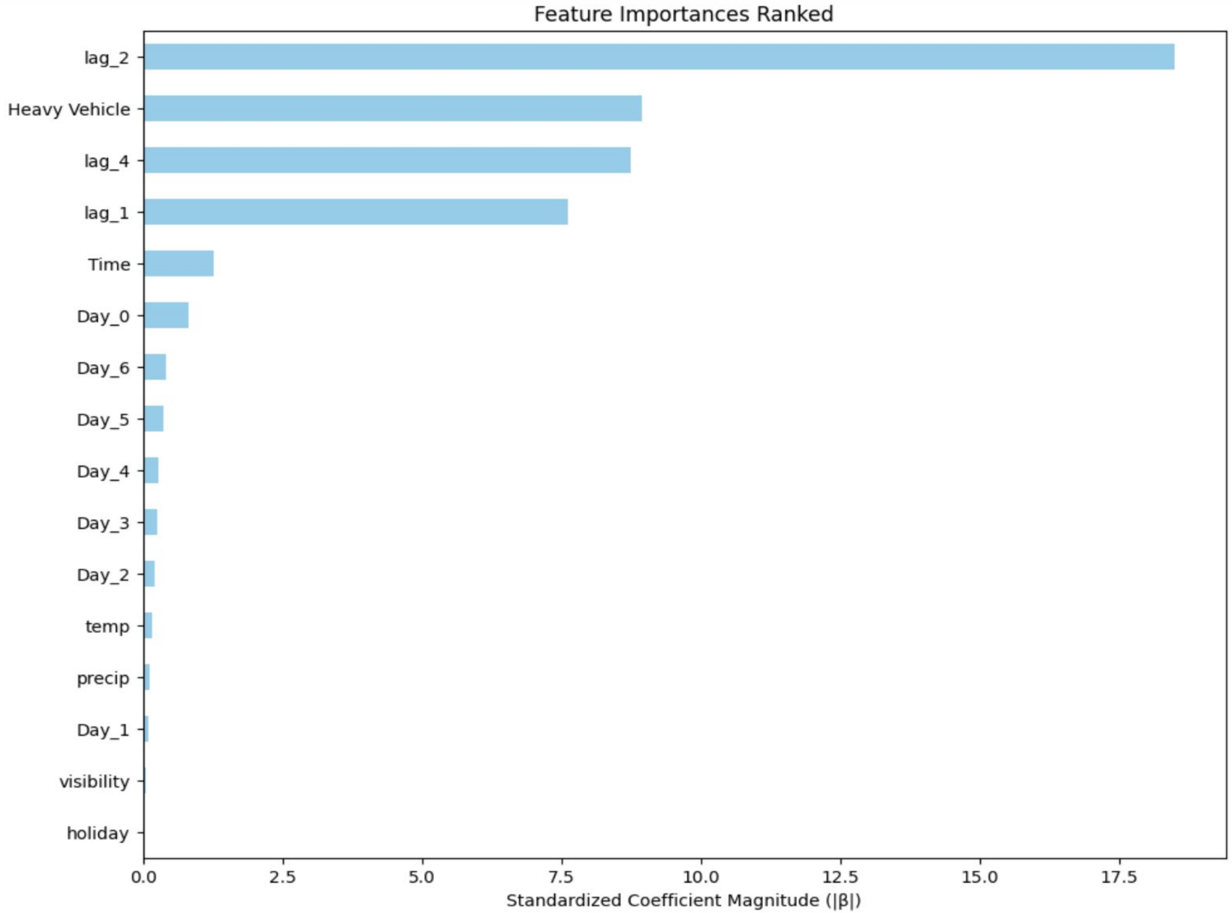
Feature	OLS Coefficient	P-Value
Precipitation	-2.499510e+08	9.752530e-01
Time (of day)	2.015019e-01	0.000000e+00
lag_1	1.177328e-01	0.000000e+00
lag_2	2.859718e-01	0.000000e+00
lag_4	1.353272e-01	0.000000e+00
Visibility	3.929092e-02	1.129897e-02
Temperature	1.240402e-02	2.797525e-21
Heavy Vehicle	-1.800484e+01	0.000000e+00

- We first fit a normal linear regression model using OLS and found the regression summary to see which features in normal linear regression are statistically significant.
- The features on the left were statistically significant in the model, meaning that there is evidence of a relationship between these features and volume
- The dummy variables for days of the week were not statistically significant, likely because the **Time (of day)** already accounted for that variation in this run of the model (multicollinearity)

Ridge Regression Results



Ridge Regression Feature Importances



- The five most important features:
  - Lag_2
  - Heavy Vehicle Presence
  - Lag_4
  - Lag_1
  - Time

Evaluation of Ridge Regression

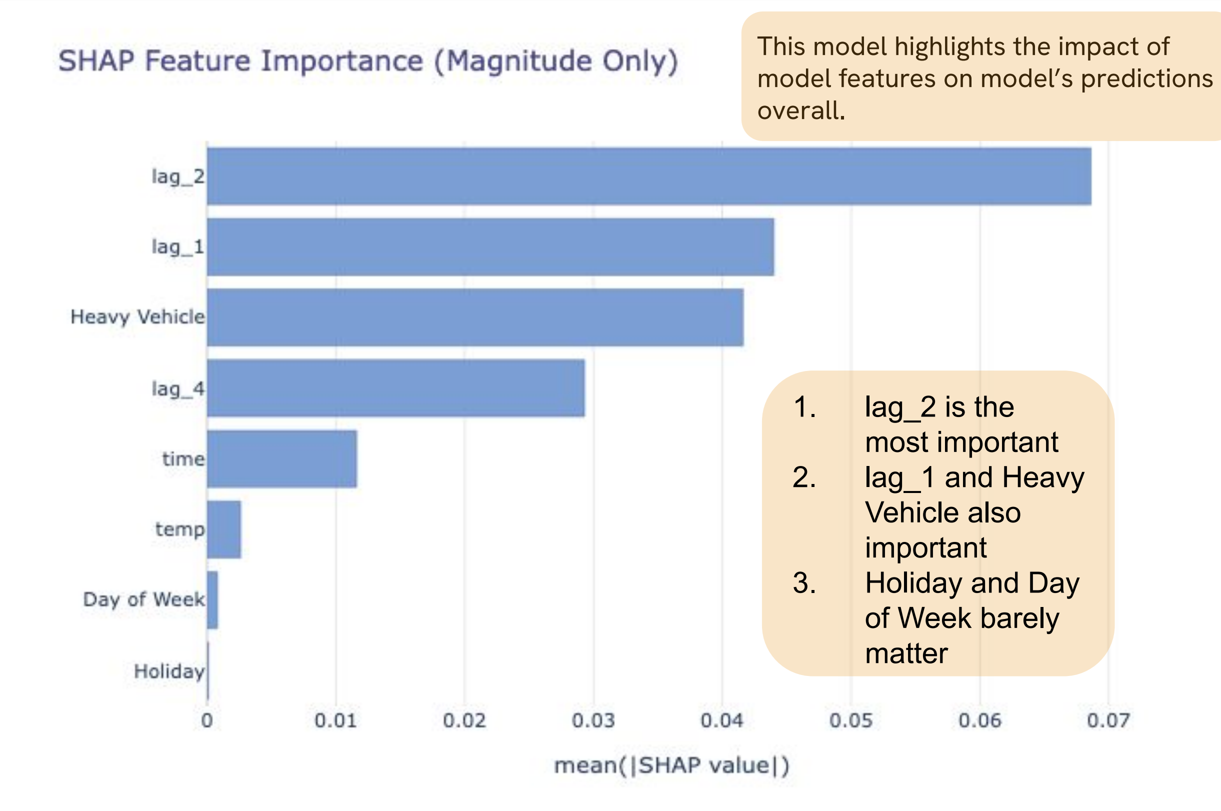
R^2: 0.2283659
RMSE: 50.51293

The R-Squared for ridge regression is **.23**, which means that the model explains around **23%** of the variation in traffic volume. The Root Mean Squared Error is **50.513**, meaning the model's prediction error on average is about **50** units of traffic volume.

Random Forest Regression

Evaluation and Results

Random Forest Feature Importance



This model highlights the impact of model features on model's predictions overall.

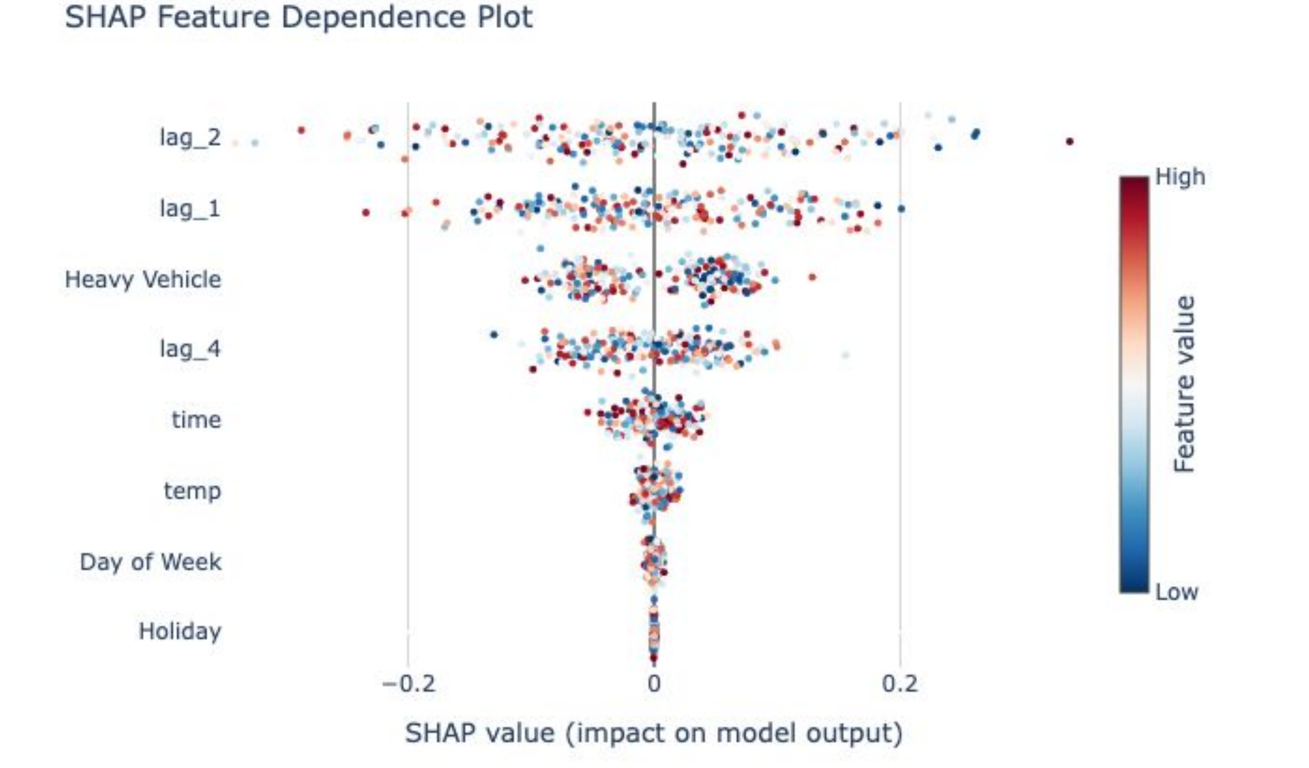
1. lag_2 is the most important
2. lag_1 and Heavy Vehicle also important
3. Holiday and Day of Week barely matter

The five most important features:

- Lagged Volumes
- Vehicle Type
- Time
- Weather
- Day of Week

It averages the **absolute SHAP values** across all samples. Larger bar = feature has a bigger impact on predictions.

Random Forest Feature Importance



This scatter plot visualizes the impact of a single feature value on a model's prediction.

01

Direction of Impact

Whether a feature pushes predictions **up or down** from average model prediction. **Positive** SHAP values push it up and **Negative** SHAP values push predictions down.

02

Feature interactions

The spread of **red/blue points** shows how the model reacts to different values of that feature.

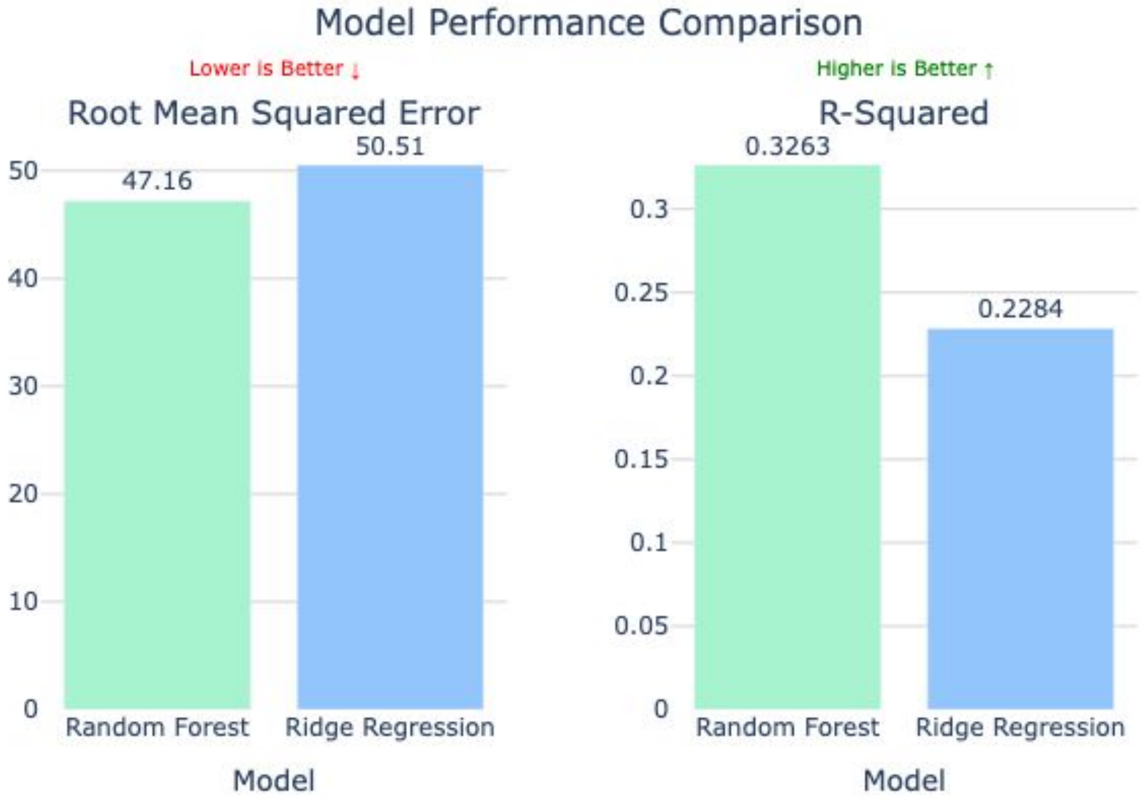
Features with **wide** spreads (lags, heavy vehicles) have a big influence on predictions.
Features with **narrow** spreads (temp, holidays) barely move the prediction at all.

Evaluation of Random Forest Regression

R^2: 0.3263
RMSE: 47.1627

The R-Squared for random forest regression is **.33**, which means that the model explains around **33%** of the variation in traffic volume. The Root Mean Squared Error is **47.163**, meaning the model's typical prediction error is on average about **47** units of traffic volume.

Model Comparison


Model Performance Comparison
Lower is Better ↓
Root Mean Squared Error
47.16
50.51
50
40
30
20
10
0
Random Forest
Ridge Regression
Model
Higher is Better ↑
R-Squared
0.3263
0.2284
0.3
0.25
0.2
0.15
0.1
0.05
0
Random Forest
Ridge Regression
Model

Energy Considerations

Model	Total Training Fits	Avg. Training Time per Fit (s)	Energy Consumed (kWh)	CO_2 Emitted (kg CO_2 eq)
Ridge Regression	1	0.5262	0.000007	0.000003
Random Forest	1	57.1712	0.000794	0.000318

Explanation and Interpretation

Our primary objectives are maximizing precision and minimizing large errors. Because our model does not need to be ran very frequently, we are not too concerned with the speed and energy trade-offs. Therefore, Random Forest (RF) is the superior predictive model for our problem.

- RF Performance: Random Forest achieved a significantly greater R^2 and lower Root Mean Squared Error (RMSE) compared to Ridge Regression. The focus on minimizing RMSE aligns directly with our goal to protect against large mispredictions of traffic volume, which is critical for understanding the features most related to volume.
- Trade-off Justification: While Ridge Regression was dramatically more energy- and time-efficient, these are not too important for practical applications of our model, which will not be ran frequently.
- The Ridge Role: Although not the optimal model, the Ridge model was still crucial for its ability to produce stable, interpretable coefficients for analysis, which was its dedicated role in our project.

Interpretation

- The models suggest that the majority of traffic congestion is driven by how busy the traffic intersection has been in the past, specifically the past 15-30 minutes (denoted by lag_1 and lag_2 and lag_4 to an extent). This tells us that traffic queues tend to “snowball” or stay busy when busy.
    - People should expect when traffic is busy, it will stay busy, so they could wait for less congested times to run their errands.
- Another factor is the presence of heavy vehicles, which is correlated with less congestion.
    - People could choose to take alternate paths that heavy vehicles are likely to take.
- Time of day affects congestion. For example, later times are correlated with higher congestion.
    - People can plan their outings to be earlier in the day.
- Ridge regression indicates that volume is lower on weekends than on other days of the week

Thank You!
Any questions?