Traffic Speed Analysis at Austin Intersections

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

Introduction

Vehicle speed sits at the pinnacle of road safety. Most drivers want to move efficiently, sacrificing safety for speed, which increases the overall danger for others on the road, especially in cities like Austin. In this analysis, we try to answer the following question: “What factors are associated with average speeds at intersections?” We hope to identify factors that can serve as guidance for inexperienced drivers who may not be as comfortable driving higher speeds as others.

Question

“What factors are associated with higher average speeds at intersections?”

Our goal is to identify interpretable and actionable insights that can help inexperienced drivers make driving decisions to feel safer on the road.

Data

The primary dataset consists of 15-minute interval traffic counts by intersection across the city of Austin, Texas. The data was collected via GRIDSMART optical detectors installed at select

intersections. Each row of the data corresponds to a single 15-minute window and includes the following features:

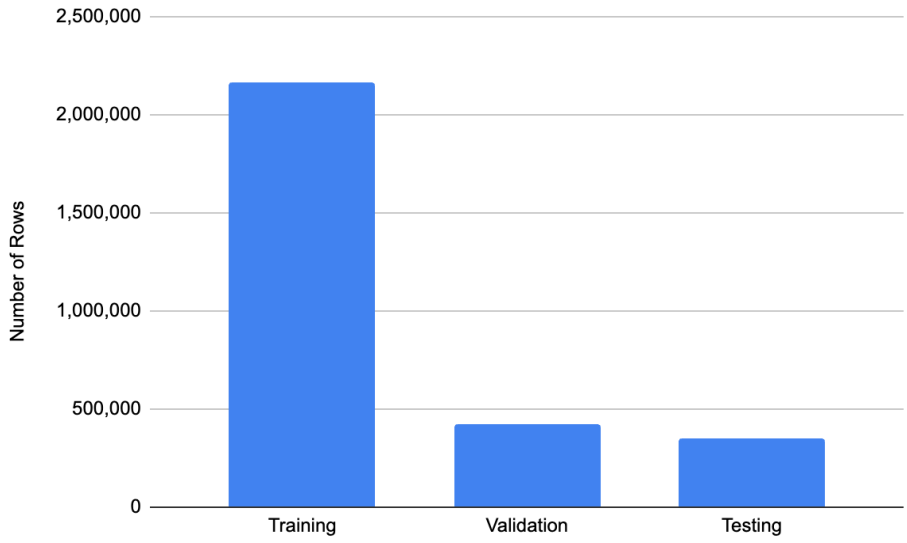
Feature	Unit of Measure	Justification
Wind Speed	Miles Per Hour	Higher wind speed can worsen vehicle stability. Drivers may drive more slowly as a precaution.
Precipitation	Inches	Precipitation can worsen traction on roads, causing people to drive more slowly.
Heavy Vehicle Presence	True/False	People may drive abnormally in the presence of a heavy vehicle.
Visibility	Miles	When visibility is worse, people are likely to drive more slowly due to not being able to see as far ahead.
Temperature	Degrees Fahrenheit	Temperature can impact engine performance, affecting speed.
Holiday	True/False	People may be in more of a rush during holidays.
Lagged Volumes	Number of Cars	Recent congestion in an intersection would likely affect driving speeds due to vehicles being close together.
Lagged Speeds	Miles Per Hour	Average speeds in an intersection may be sticky and not change much over short time intervals.
Snow	Inches	Snow can worsen friction, and people may drive more slowly as a result.
Day of Week	Categorical listing of Day of Week	People may be more likely to drive higher/lower speeds on

		certain days of the week.

Given this data and our question, the target variable for this problem is average speed.

Problem Setup

To balance feasibility and interpretability, we subsetted the dataset to only rows from the year 2019. This was to ensure that the data did not conflict with the effects of COVID-19 during and after the pandemic. It was also more feasible for computational purposes since the original dataset contained over 82 million rows. Lastly, the single-year window reduces confounding of long-term demographic changes. After this filter, the dataset contained a little over 14 million rows with 19 columns. Afterwards, we conducted an 80-10-10 split: the training set contains the months January through September, the validation set consists of October and November, and the testing set includes December. The data was split in this manner to avoid temporal leakage, ensuring that only past data is used to predict future observations. It also guarantees that there is an unbiased estimate of generalization because it simulates the real-world scenario of predicting traffic volume without knowledge of the future. It is also worth noting that we did not use k-fold cross-validation because it would have broken the time order and caused data leakage.



Justification of Feature Selection

The features were selected based on strict inclusion criteria to ensure rigor. Firstly, all independent variables are causally before speed. Second, all variables are available for every observation, meaning that there are no missing values, which exhibits complete data integrity. Third, the study design fully prohibits data leakage. Finally, each feature chosen has a predictive relevance that can be justified with general physical, behavioral, or environmental reasoning.

Limitations

This study is subject to several limitations due to the restricted temporal scope. First, the limitation of a single year of data prohibits the capture of multi-year seasonal effects and also excludes any influence of a longitudinal demographic shift. Furthermore, the limitation of using only a single year exposes the dataset to many irregularities, such as holidays, weather anomalies, and various other unexpected events, which may distort trends heavily.

Model Classes

Ridge Regression

Ridge Regression is well-suited for analyzing temporal data with smooth trends, offering high interpretability through clear coefficient analysis and low computational cost. Its L2 regularization is particularly effective for high-dimensional datasets, preventing overfitting while optimizing feature weights to maintain low variance.

Random Forest

The Random Forest model excels at capturing complex, non-linear relationships within data and is notably robust against outliers.

Hyperparameter Tuning

Ridge Regression

The parameter α, the strength parameter of the ridge regression penalty, was tuned. To avoid temporal leakage, we avoided the usage of grid search cross-validation and randomized search cross-validation techniques. Instead, α was chosen from the following set of α values: [0.001, 0.01, 0.1, 0.2, 0.3, 0.5, 1.0, 2.0, 5.0, 10.0]

The α value that yielded the smallest root mean squared error (RMSE) on the validation set was chosen. Ridge regression fitted on α = .001 resulted in the smallest validation RMSE (11.257503), hence it was chosen as the value on the test set.

Random Forest

We tuned only **n_estimators** and **max_depth** to balance performance with computational efficiency. These two hyperparameters are the most influential in controlling Random Forest complexity, and limiting the search space helped keep training time manageable given the large dataset and chronological split. Because our primary goal was interpretability rather than maximizing predictive accuracy, an expanded hyperparameter search was unnecessary.

The best parameters selected were: {'model__max_depth': 10, 'model__n_estimators': 100}.

Model Choices

Ridge Regression

Statistically Significant Features (P < .01) from Standard Ordinary Least Squares

Feature	OLS Coefficient	P-Value
Snow	5.325e-14	7.608e-201
Day 0 (Sunday)	2.713e+00	1.660e-79
Day 1 (Monday)	2.146e+00	2.804e-50
Day 2 (Tuesday)	2.141e+00	3.653e-54
Day 3 (Wednesday)	2.1204e+00	3.297e-53
Day 4 (Thursday)	1.877e+00	3.483e-41
Day 5 (Friday)	1.868e+00	7.037e-42
Day 6 (Saturday)	2.410e+00	1.180e-67
Lag 2 (Avg. Speed at T - 2)	2.131e-02	1.105e-107
Lag 4 (Avg. Speed at T - 4)	4.860e-03	2.139e-08
Lag_1_Volume (Volume at T - 1)	5.242e-02	4.796e-128
Lag_2_Volume (Volume at T - 2)	2.6165e-02	1.174e-13
Lag_4_Volume (Volume at T - 4)	1.429e-02	4.447e-05
Heavy Vehicle Presence	1.341e+00	5.138e-34
Time	-2.002e-01	8.920e-132

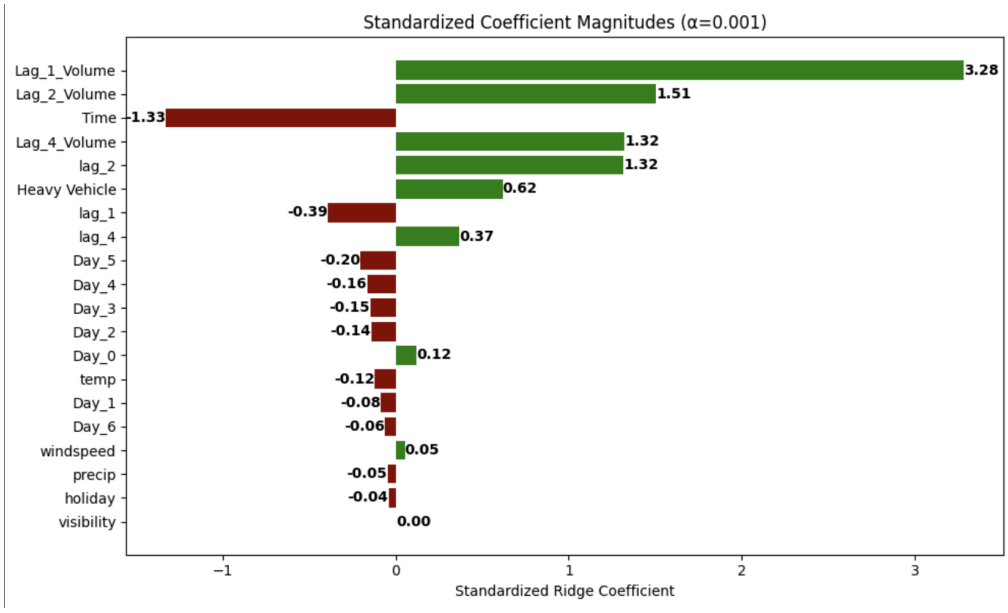
Interpretation

The above chart is a summary of coefficients and their associated p-values resulting from an ordinary least squares linear regression model without the ridge regression penalty factor. Note that the chart only displays those that were significant at the 1% significance level, and the OLS model was run on a random sample of 50,000 observations due to computational limitations. Statistical significance here suggests that there is evidence of relationships between the above features with the target variable, average speed.

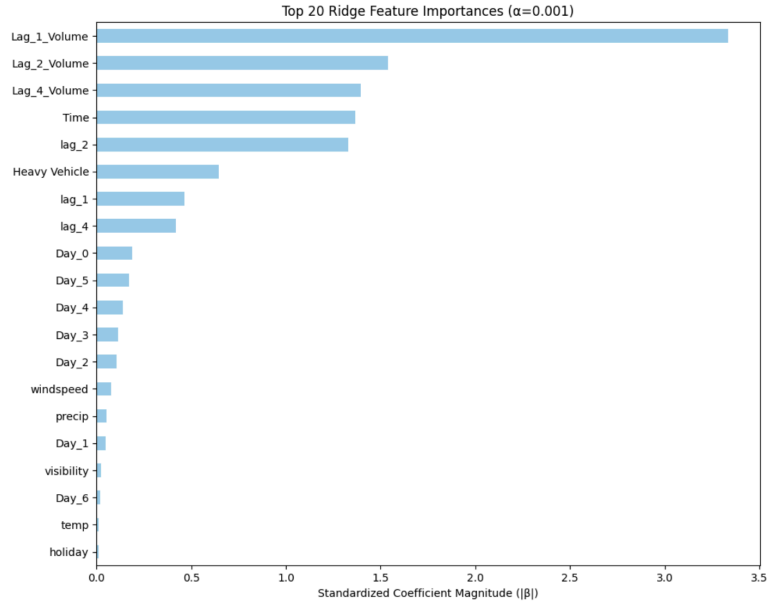
Based on the results, there is convincing evidence that lag average speed is associated with the average speed. The positive coefficient suggests that high average speeds are persistent, as are low average speeds. This indicates a pattern where average speed remains consistently high or consistently low for periods of time. Additionally, the positive coefficient of lag volume indicates that higher volumes of vehicles tend to yield higher average speeds at intersections.

Additional significant features include time of day, which has a positive coefficient indicating rising speeds throughout the day, and day of the week, reflecting distinct daily patterns. Counterintuitively, heavy vehicle presence is associated with higher average speeds. A reasonable explanation might be that cars attempt to overtake heavy vehicles, leading to higher average speeds. Finally, snow also seems to be statistically significant. However, it is worth noting that the distribution of snow is highly skewed due to Austin not having much snow throughout the year, which may have led to biased results in statistical significance for this feature.

Coefficient Magnitudes from Ridge Regression



Feature Importances (Absolute Values of Coefficient Magnitudes) from Ridge Regression



Interpretation

The chart above displays the standardized coefficients from a ridge regression model ($\alpha = 0.001$) fitted on a random testing sample of 50,000 observations. The key drivers identified align with the feature importance analysis from the Random Forest model, confirming the significance of

temporal factors: lagged volumes, lagged speeds, time of day, and day of the week. The presence of heavy vehicles also emerges as a notable predictor. Conversely, weather-related features demonstrate the least influence on the model.

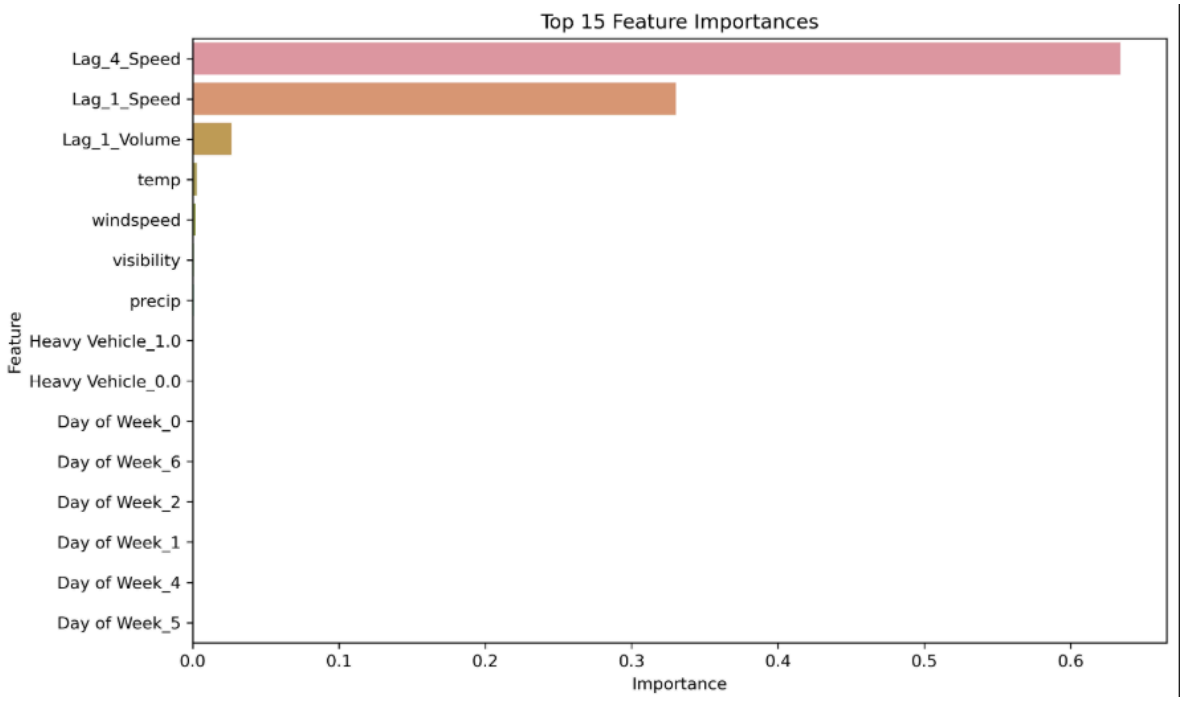
Overall Takeaway

Our analysis identifies recent history and temporal factors—specifically lagged volume, lagged speed, and time of day—as the most influential factors of average speed, with weather-related factors showing the weakest influence. The results show a strong pattern of persistence: higher lagged speeds and volumes predict higher current speeds, with their coefficients being mostly positive and statistically significant. The presence of heavy vehicles is also positively associated with speed. However, the model's explanatory power is limited ($R^2 = 0.20$), and a counterintuitive result—higher volumes predicting higher speeds—suggests the potential influence of unobserved confounders. Consequently, these findings should be interpreted with caution.

Random Forest

Feature	Feature Importance
Lag_4_Speed	**0.6340**
Lag_1_Speed	**0.3303**
Lag_1_Volume	**0.0266**
temp	**0.0029**

windspeed	**0.0019**
visibility	**0.0009**
precip	**0.0009**
Heavy Vehicle_1.0	**0.0007**
Heavy Vehicle_0.0	**0.0006**
Day of Week_0	**0.0002**



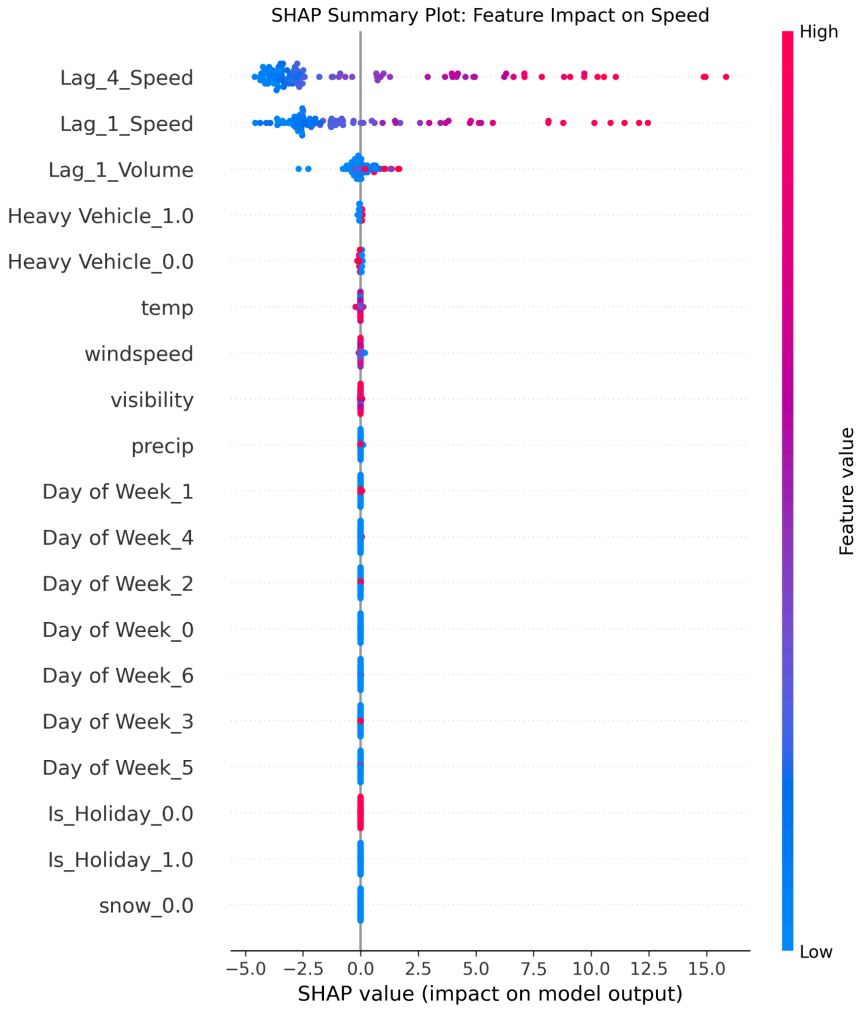
Random Forest Feature Importance

The Random Forest model shows a very clear hierarchy in which features matter most for predicting average traffic speed:

- Lag_4_Speed (speed one hour prior) is by far the strongest predictor.
- Lag_1_Speed (speed 15 minutes prior) is the next most influential feature.
- Lag_1_Volume also contributes, though much less.
- Weather variables (temperature, visibility, wind speed, precipitation) have very little importance.
- Categorical features like Heavy Vehicle type, Day of Week, Holidays, and Snow have minimal effect.

Interpretation

Traffic speed is highly autocorrelated. The best way to estimate current speed is simply to look at what speeds were in the recent past. Weather and calendar effects play a role, but are not strong enough to significantly shift speeds in short 15-minute windows. This is consistent with real-world traffic patterns.



What does the SHAP Summary Plot mean?

The SHAP summary plot illustrates how each feature influences individual predictions in the Random Forest model. Each point represents a single observation, with its position along the x-axis indicating whether that feature pushed the predicted speed higher (right) or lower (left). The color scale reflects the feature value (red = high, blue = low), allowing us to see how different feature magnitudes affect the model's output.

Random Forest SHAP Summary Plot

- Lag_4_Speed and Lag_1_Speed again appear at the top, confirming they provide the majority of predictive power.
- SHAP values show both positive and negative effects, meaning high lagged speeds push predictions upward, while low lagged speeds pull them downward.
- Lag_1_Volume contributes meaningfully—higher volume tends to reduce predicted speed.
- Weather variables show tightly centered SHAP values near zero, suggesting they have low influence.
- Day-of-week and holiday indicators have almost no effect, aligning perfectly with feature importance rankings.

Interpretation

The SHAP results suggest the model's predictions are driven almost entirely by recent traffic patterns, which aligns with our goal of understanding what factors influence speeds at intersections. Lagged speeds and lagged volume consistently have the strongest influence, meaning current speed is most dependent on how fast and how congested the intersection was shortly beforehand. Weather, day-of-week, and holiday effects contribute very little, indicating that broader environmental or temporal factors do not meaningfully shift speeds within short 15-minute intervals. Overall, SHAP analysis confirms that the model relies on the most relevant signals—recent traffic flow—to explain variation in vehicle speeds.

Overall Takeaway

Short-term traffic speed is overwhelmingly driven by its recent history. Lagged speeds are by far the strongest predictors, with lagged volume also playing a meaningful but smaller role. In contrast, weather conditions, calendar effects, and other categorical factors contribute very little and do not meaningfully shift speeds within 15-minute intervals. This pattern aligns with realistic traffic behavior in Austin, where immediate traffic flow indicates how fast vehicles move through intersections—not broader environmental or temporal factors.

Results

What are our evaluation metrics and loss function?

R^2 and RMSE were used as our evaluation metrics, and we used MSE as our loss function because they are well-suited for continuous prediction problems like traffic speed.

R^2 helps us understand how much variation in speed the model can explain, which aligns with our goal of identifying which factors drive speed differences across intersections. RMSE provides an interpretable measure of prediction error in miles per hour, making it easy to assess how far off the model is in a real-world context.

We selected MSE as the loss because it penalizes larger errors more heavily, which is important in prioritizing safety where large speed-prediction mistakes are more devastating. Taken together, these metrics allow us to evaluate both explanatory power and reliability in predicting average speeds.

Ridge Regression

Evaluation Metrics

Validation RMSE	Validation R^2	Test RMSE	Test R^2
11.2575	**0.2099**	**11.5371**	**0.2019**

The Ridge Regression model achieved an R-squared value of .20, indicating that it can explain around 20% of the variations in average vehicle speeds across Austin intersections. Although we achieved insight into the influential features, there is much variability not explained by the model, so the results should be interpreted carefully. The RMSE of 11.26 suggests that the model's predictions are typically off by about 11 mph.

Random Forest

Evaluation Metrics

Validation RMSE	Validation R^2	Test RMSE	Test R^2
9.7053	**0.4184**	**9.9871**	**0.4037**

The Random Forest model achieved an R-squared value of 0.40, indicating that it could explain approximately 40% of the variations in average vehicle speeds across Austin intersections. The RMSE of 9.99 suggests that the model's predictions are typically off by roughly 10 mph.

How do these results relate to our problem?

In the context of our problem, this indicates that while the model captures some meaningful patterns in how intersection characteristics relate to speed, there is still substantial variability driven by factors outside the dataset—such as driver behavior, other weather conditions, or unmeasured roadway features.

Discussion

Several studies support the importance of environmental and traffic factors in determining vehicle speeds. What factors contribute to average vehicle speeds has been an important question nearly as long as traffic has existed. In 1962, Rowan and Keese analyzed the influence of some similar features, such as visibility, on driver speeds. However, they also explored factors that we did not, such as the curvature of streets, urban developments around the streets, and proximity to radar detectors. Incorporating these into our model would likely improve its strength by adjusting for these confounders. The goal of their report was to inform public policy, particularly regarding the establishment of speed limits. Our report's goal is a smaller scale but more useful for ordinary individuals rather than policymakers.

Alomari et al. focus on utilizing the results for the creation of better traffic management strategies, while our research focuses on helping inexperienced drivers target calmer driving conditions (lower speeds). Their data is from streets where drivers are able to reach a free-flow state (ideal, uninterrupted driving conditions), while our data is specifically from vehicles at intersections. Both of these are important aspects of driving and traffic congestion, so the results should be able to align and build off of each other. Alomari et al. analyze important features that

we did not use, such as road roughness/quality, while we analyze features such as weather that they did not include. There is some overlap between our projects, primarily regarding the use of speed and volume variables. They found that volume has a negligible, positive impact on speed. However, we found that volume had a small but non-negligible impact on speed. This is likely due to Alomari et al. focusing on free-flow speed while our aim was the average speed at intersections. They found that road roughness/quality decreases speed (worse quality leads to lower speed). This would likely add to the strength of our model if we were to include it in a future version of the report.

Conclusion

This analysis identifies factors associated with average vehicle speeds at Austin intersections. There is a central finding between both models: average speed at Austin intersections is strongly determined by recent traffic conditions. Both Ridge and Random Forest models identify lagged speeds and volumes to be the strongest predictors. In contrast, weather variables, day of the week, holidays, and the presence of heavy vehicles demonstrate minimal to negligible influence within the 15-minute time intervals studied. This highlights that driver behavior is primarily reactive to the immediate flow rather than broader environmental or temporal factors.

The models' limited explanatory power (R^2 of 0.20 for Ridge Regression and 0.40 for Random Forest) underscores a critical caveat: a significant portion of speed variability remains unexplained, suggesting the influence of unobserved factors like individual driver behavior, road structure, or real-time traffic controls. Therefore, while the analysis successfully identifies the dominant signal of recent driving history, it also reveals the complexity of traffic systems. For

practical application, this means guidance for drivers or planners must emphasize proactive decision-making, as one cannot only rely on environmental conditions, day of the week, holidays, or the behavior of surrounding traffic to gauge speed level.

Real-World Applications for Inexperienced Drivers

Our research originated to aid inexperienced drivers. An important note is that weather conditions (rain, visibility) had relatively little influence on speeds, implying that traffic does not naturally slow down during poor conditions. Thus, there is little reason for inexperienced drivers concerned about their safety to drive under more difficult conditions because the slight reductions in the speeds of drivers likely will not outweigh the difficulties posed by precipitation and low visibility. Due to the predictive power of lagged volumes and speeds, inexperienced drivers should know that the group behavior of traffic tends to set a pace that is unlikely to change quickly. If a road is highly congested or full of cars driving at higher speeds, then inexperienced drivers should avoid it. The ridge regression found that the presence of a heavy vehicle was associated with higher traffic speeds. We suggest that inexperienced drivers try to drive on roads where heavy vehicles are less likely to be encountered.

Acknowledgment

Contribution Scores (0-100):

Ritesh	Lucas	Brian	John	Travis
100	100	100	100	100

For this project, we used generative AI tools—primarily ChatGPT and Gemini—to help with brainstorming analysis approaches, clarifying conceptual questions, and improving the clarity of written explanations. We used it to refine code structure, troubleshoot issues such as reducing dimensionality, improving training efficiency, and simplifying the model when needed, and strengthen the readability of the written report. All modeling decisions, interpretation, and implementation were ultimately guided by our judgment, but AI tools helped speed up problem-solving and improve communication throughout the project.

References

Alomari, Ahmad H., et al. "Traffic Speed Prediction Techniques in Urban Environments." Heliyon, vol. 8, no. 12, Dec. 2022, e11847. Cell Press, https://doi.org/10.1016/j.heliyon.2022.e11847.

Rowan, Neilon J., and Charles J. Keese. *A Study of Factors Influencing Traffic Speeds*. Texas Transportation Institute, Jan. 1962, library.ctr.utexas.edu/hostedpdfs/tti/rp-17-04.pdf.