Final

Connor Jacobs & Madi Tansy

Introduction

This project focused on predicting the most and least congested state routes in California for 2025 to identify opportunities for traffic mitigation and infrastructure planning. We employed two models: a simple linear regression model to capture basic trends and a neural network to uncover complex patterns in traffic data. This work is significant as it aims to provide actionable insights for reducing congestion, improving travel efficiency, and minimizing environmental impact.

Analysis

The data we used came from Caltrans and included things like traffic counts, freeway locations, and how busy roads are during rush hour. First, we cleaned and organized the data to make sure everything was ready to use. Then, we made maps and charts to see where traffic is worst and when it happens. The linear regression model helped us understand simple connections, like how location affects traffic. The neural network dug deeper, finding hidden patterns that are harder to spot. These models helped us predict traffic and offer ideas to improve driving and road planning.

Methods

Linear Regression Model
The linear regression model was designed as a simple and interpretable way to analyze traffic patterns. It used key features like geographic location (longitude and latitude), peak-hour traffic counts, and route information to predict the average daily traffic volume. After splitting the data into training and testing sets, we trained the model using the training data to understand

how these factors relate to traffic volumes. Linear regression provided a straightforward baseline, showing how individual features impact traffic without introducing unnecessary complexity. Its predictions were then compared to the true values to evaluate performance using metrics like mean squared error and R^2 score.

Neural Network Model

The neural network model was developed to uncover more complex relationships in the data that the linear regression model might miss. Its architecture consisted of an input layer for the standardized features, followed by two dense hidden layers with 64 and 32 nodes, each using ReLU activation to process non-linear patterns in the data. The final output layer provided traffic volume predictions. The model was trained using the Adam optimizer and mean squared error as the loss function. By leveraging its multiple layers, the neural network learned deeper patterns in the data, such as how combinations of features interact. The performance was evaluated on the test set, and results were visualized to compare predictions to the actual traffic volumes.

Results

Table 1: Results of Basic and Advanced Models

	Linear Regression	**Neural Network**
MSE	244132863.03	257423739.42
R2	0.95	0.95
MAPE	74.37%	63.13%

Linear Regression

The linear regression model did a pretty good job overall. It had a mean squared error (MSE) of about 244 million and an R^2 score of 0.95, meaning it explained 95% of the differences in traffic volumes. That's solid, but its Mean Absolute Percentage Error (MAPE) was pretty high at 74.37%, which shows it sometimes struggled with individual predictions, especially when traffic volumes were either really high or really low. Looking at the scatter plot of actual vs. predicted values, most of the predictions lined up well, but there were definitely a few points where the model either overestimated or underestimated traffic volumes.

Neural Network

The neural network model was a bit more complex, but its performance was similar. It had an MSE of about 250 million and the same R^2 score of 0.95, which means it captured trends just as well as the linear regression model. However, its MAPE was slightly better at 63.13%, meaning it was more accurate at predicting specific values. Like with linear regression, the scatter plot showed that while most predictions were close, there were some clear misses, especially in more extreme cases.

Comparison

Both models performed similarly when looking at overall trends in Figure 1, but the neural network model was slightly better at making precise predictions for individual data points. What's interesting is that both models picked up on patterns that make sense if you've spent any time driving in Southern California. For example, the most congested freeways included routes like the 405, 91, and 605, which anyone from this area knows can feel like parking lots during rush hour. On the other hand, the least congested routes in Figure 2, like Route 211 or Route 140, are in the middle of nowhere. Most of us haven't even heard of these freeways because they're so far from major urban areas, which really validates the results.

The bar charts and heatmap also helped bring these findings to life. Seeing the 405 at the top of the "most congested" list in Figure 3 isn't surprising—it's infamous for its traffic jams. Meanwhile, the heatmap in Figure 4 clearly shows heavy traffic clustered around major urban hubs like Los Angeles and Orange County, while quieter routes are spread out in less populated areas. All of this backs up what we already know about traffic in California: it's brutal in the city, but out in the middle of nowhere, the roads are wide open. These results not only make sense but also show that the models captured real-world patterns accurately.

Table 1: Prediction Comparison of Model

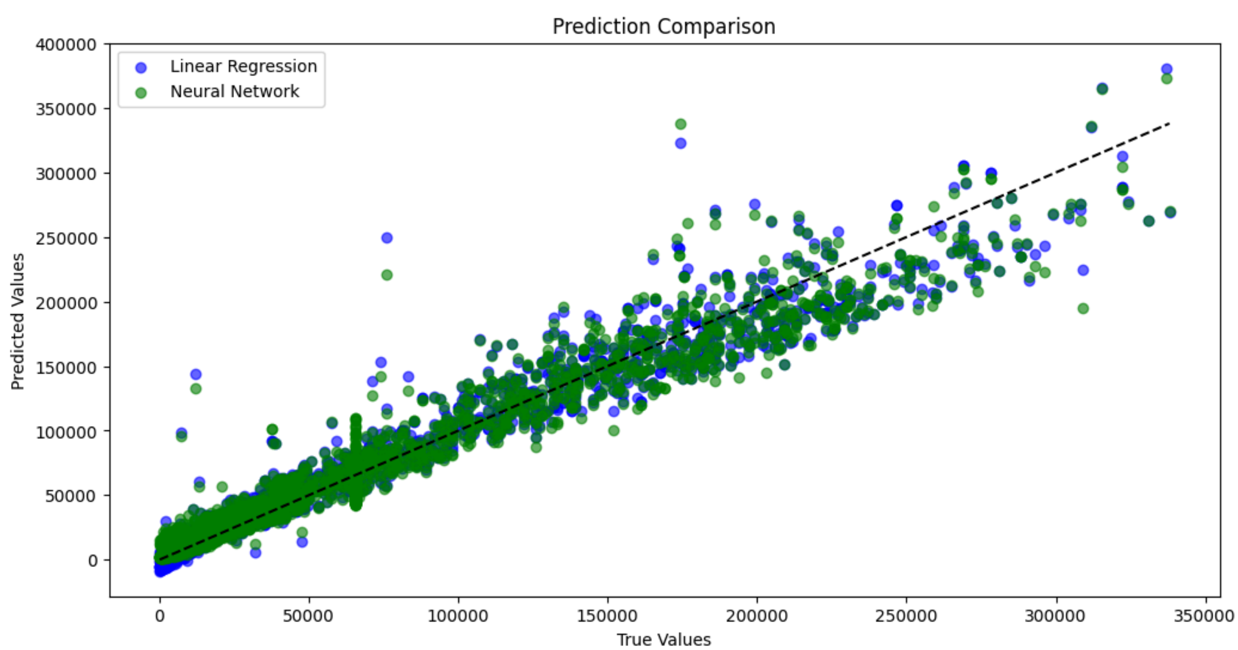


Table 2: Top 10 (Predicted) Least Congested Freeways

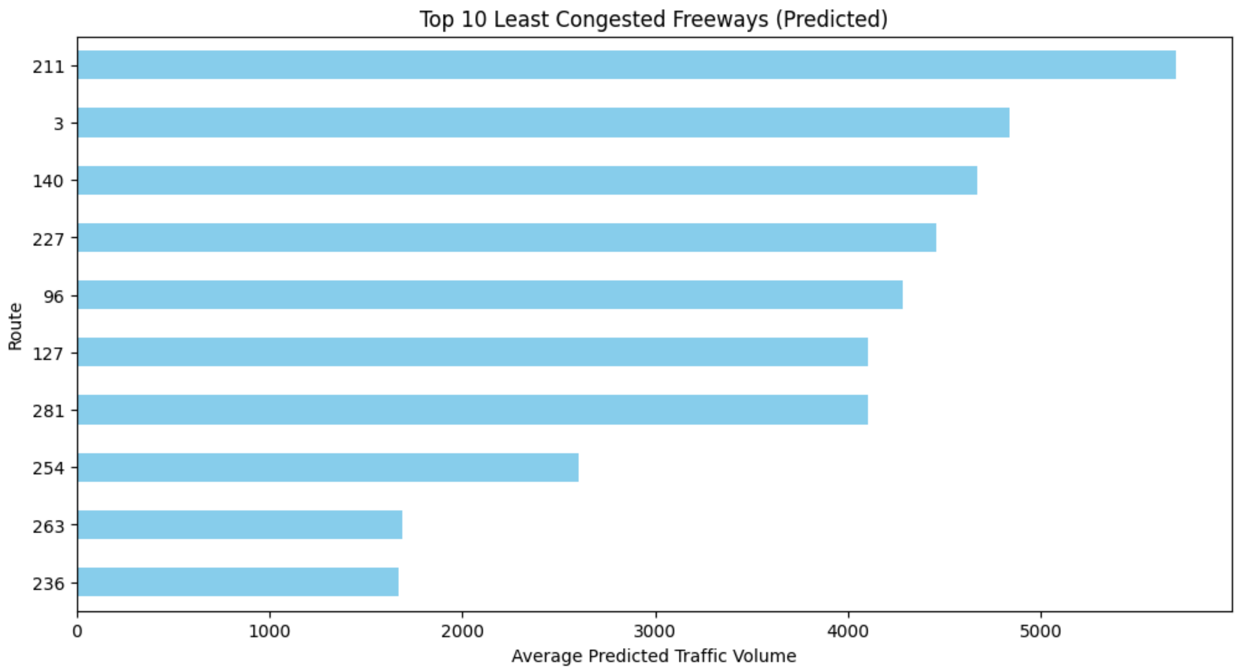


Table 3: Top 10 (Predicted) Most Congested Freeways

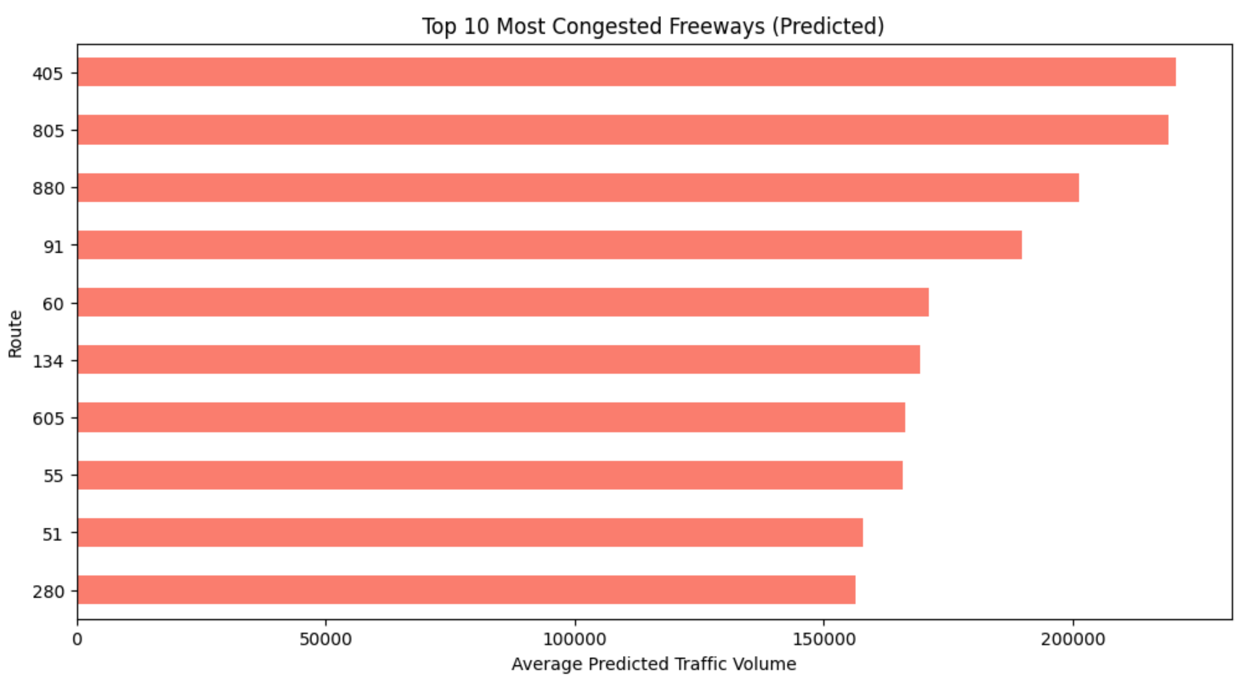
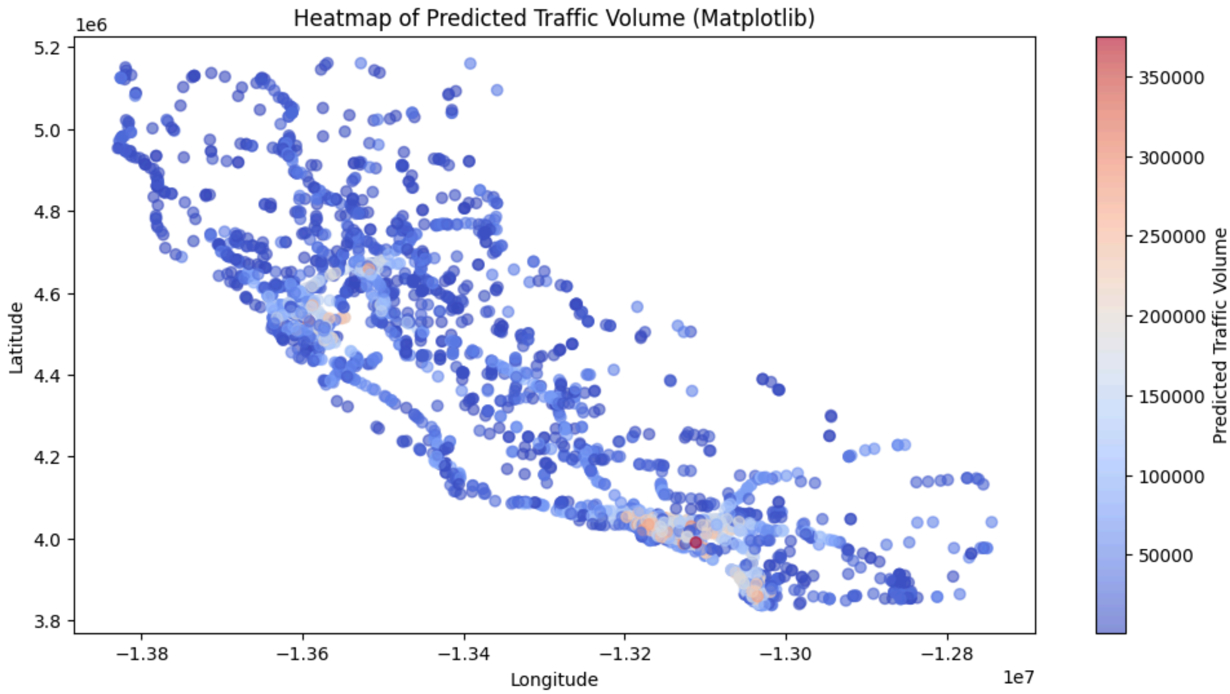


Table 4: Heatmap of Predicted Traffic Volume Throughout California



Reflection

This project highlighted the differences between using a simple model such as linear regression, and a more advanced model, such as a neural network for predicting traffic patterns. Both models performed reasonably well, but the neural network slightly outperformed linear regression. The neural network's ability to model non-linear patterns made it better suited for this task, though linear regression remained useful for its simplicity and interpretability. Overall, the project demonstrated how using both simple and complex models can help provide a clearer picture of traffic trends.

One of the main challenges during this project was choosing the best advanced model to use. Initially, we selected an SVM model, which had MAPE scores over 500%, so we decided to change to a NN, which was significantly better. Another difficulty was fine-tuning the training process—adjusting hyperparameters like learning rate and batch size took some trial and error to prevent overfitting and improve performance. Despite these challenges, the project showed how powerful neural networks can be when applied to real-world problems such as this one.