CS5300 Final Project Report - Phase 4

Scott Sanchez

December 2, 2024

Introduction

In this phase, the goal was to evaluate the importance of input features by iteratively removing them and measuring the resulting validation accuracy. The aim was to identify non-informative or redundant features, streamline the dataset, and ultimately improve the performance of the final model by reducing complexity. The findings from this process allow for better generalization while maintaining or even improving predictive performance. This phase builds upon the earlier analyses, where we observed which features contributed the most to prediction accuracy.

Methodology

Step 1: Individual Feature Importance

The first part of this analysis involved training models using only one feature at a time. For each feature, a simple neural network was trained, and the validation accuracy was recorded. This step provided insights into the relative importance of each feature.

Each model was defined as follows:

- Architecture: 16 neurons in the first hidden layer and 8 neurons in the second hidden layer.
- Activation: ReLU for the hidden layers and softmax for the output layer.
- Loss function: Categorical crossentropy.
- Optimizer: Adam.

Validation accuracies for each feature were plotted as a bar chart (see Figure 1), revealing which features contributed the most to model performance.

Step 2: Iterative Feature Removal

Building on the insights from Step 1, features were iteratively removed in order of increasing importance. At each iteration, a model was trained using the remaining features, and the validation accuracy was recorded. This process continued until all features were removed, allowing us to observe the effect of feature removal on validation accuracy.

The final subset of features was determined by identifying the point where validation accuracy began to significantly decline. This reduced feature set was then used to train a final model, which was compared to the full-feature model for performance.

Results

Feature Importance

Figure 1 shows the validation accuracy of models trained with individual features. Interestingly, `Start_Lat` emerged as the most important feature, achieving the highest validation accuracy when

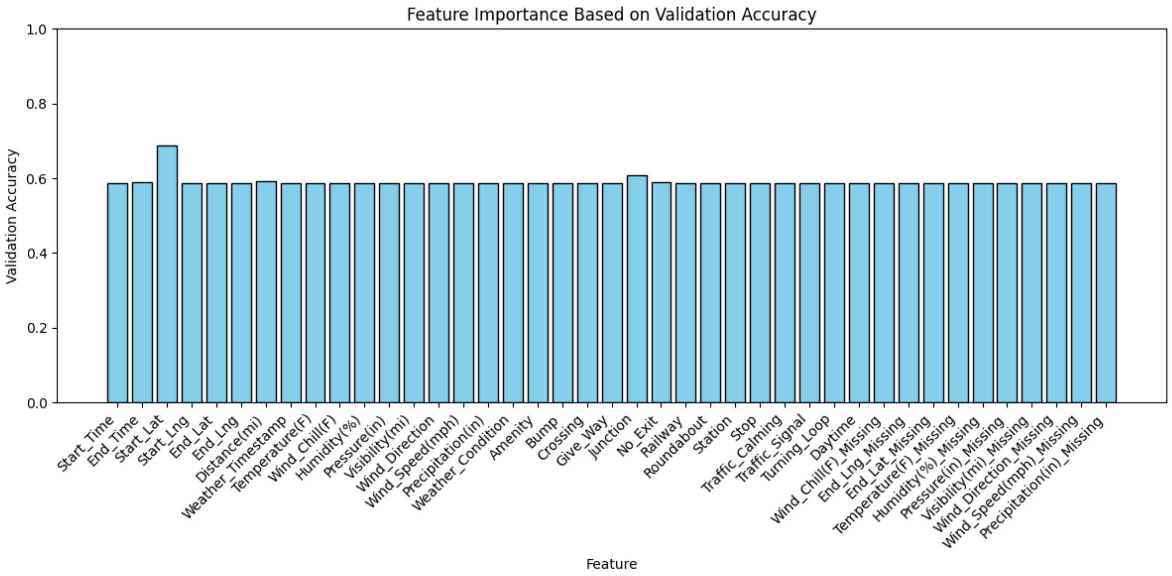

Figure 1: Validation Accuracy for Models Trained with Individual Features

used as the sole input. Other factors, when considered alone, had only minimal predictive power when considered alone.

Iterative Feature Removal

Figure 2 illustrates the validation accuracy at each step of the iterative feature removal process. As expected, removing features (in the order of importance identified in Step 1) led, generally, to a decrease in accuracy. In short, the overall trend-line through Figure 2 is a gradual downwards one. Interestingly, the model trained with only features ranked strictly higher than `Daytime` achieved a validation accuracy that surpassed the full-feature model.

Final Model Performance

The final model was trained using the most important features as identified during iterative removal. This reduced feature model achieved a validation accuracy of 70.3%, compared to 68.8% for the full-feature model. Table 1 summarizes the performance of the full-feature model and the reduced feature model.

Model	Validation Accuracy	Number of Features
Full-Feature Model	68.8%	35
Reduced-Feature Model	70.3%	12

Table 1: Comparison of Full-Feature and Reduced-Feature Models

Discussion

This exercise directly shows how important feature selection is in machine learning. Removing redundant or non-informative features not only simplified the model but also improved validation accuracy. This improvement suggests that the removed features introduced noise or complexity, which hindered the model's ability to generalize.

The reduced feature model is not only more interpretable but also more computationally efficient. Features like `Start_Lat`, `End_Time`, and `Start_Time` proved to be the most critical for predicting crash severity. This makes sense; location and time are, logically, the biggest components of accident risk.

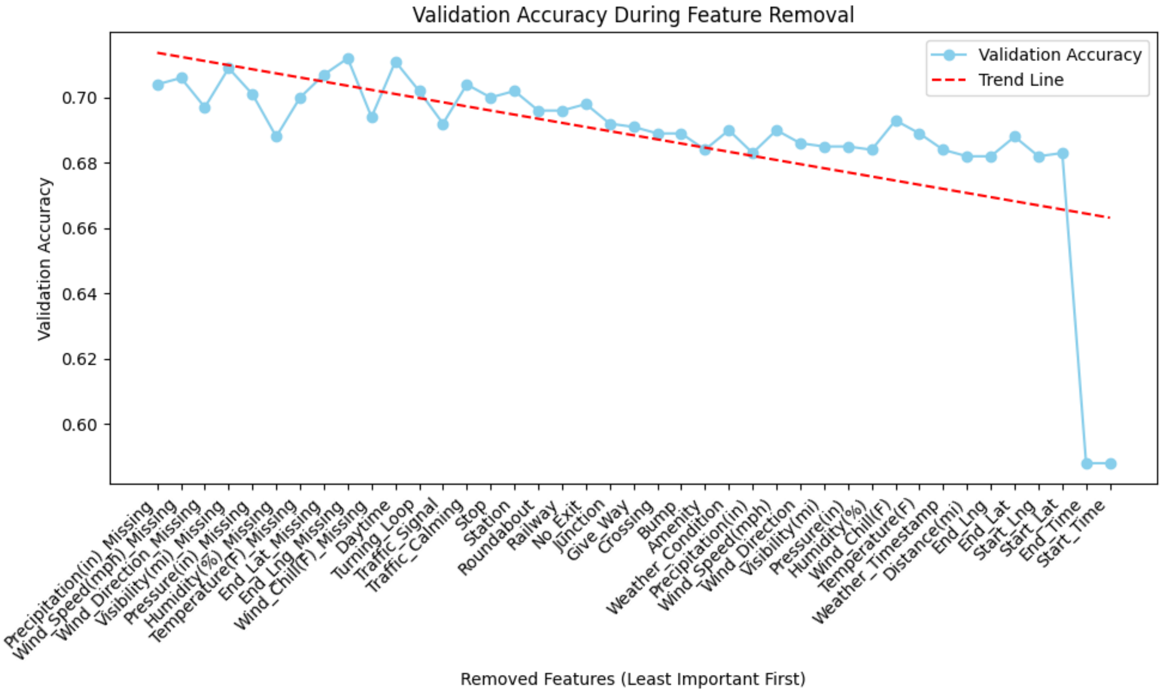

Figure 2: Validation Accuracy During Iterative Feature Removal

Conclusion

This phase displayed the value of iterative feature removal for improving model performance and interpretability. By identifying and removing less critical features, a reduced feature model outperformed the full-feature model. Future work could explore advanced feature selection techniques or consider interactions between features for optimization.