Predicting Car Accident Severity Using Neural Networks:

CS5300 Final Project Report

Scott Sanchez

December 3, 2024

1 Abstract

This report presents the findings from a multi-phase project aimed at predicting the severity of car accidents using machine learning techniques. The project involved data cleaning and preparation, model optimization, and feature selection. Phases included iterative feature analysis, model comparisons, and tuning to improve predictive accuracy and model generalization. The dataset used contained detailed car crash records across the United States, encompassing weather, road conditions, and other contextual information. The final model achieved a validation accuracy of 70.5%.

This report also discusses challenges such as class imbalance and overfitting, and suggests directions for future work.

2 Phase 1: Data Cleaning and Preparation

The dataset I chose for this project comes from the Kaggle repository "US Accidents (Sobhan Moosavi)," [1] which contains detailed records of accidents across the United States. I chose this data because it had richer data than any other car-accident database I could find.

Phase 1 focused on data cleaning and preparation. The data set contained records of car accidents, including information on weather, road conditions, and other contextual factors, totaling more than 1.5 million records. Data quality was critical to ensuring the success of the machine learning models, as poor quality data could significantly reduce the model's predictive accuracy.

Several steps were undertaken to prepare the data, including:

- **Removing irrelevant columns:** Initial inspection of the data set showed that several columns were not useful to predict accident severity, such as unique identifiers and timestamps that could not be meaningfully translated into features.
- **Handling missing values:** Many records contained missing information on certain characteristics, particularly related to weather conditions. Each missing value was replaced with an out-of-range value of -9999. For each column that contained missing values, an additional "missingness" column was added. For example, the "Wind Chill" column was missing values in some rows, so an additional "Wind Chill Missing" column was added, which had a value of 1 where the windchill had missing values, and a value of 0 elsewhere. This encoded the idea of missingness, so that the model did not come to believe that the windchill was actually -9999 Fahrenheit in those columns.
- **Feature encoding:** Categorical variables, such as weather descriptions, were transformed into numerical representations using one-hot encoding. This allowed the neural network models to understand these features more effectively during training.
- **Normalization:** To ensure consistency across features, numerical values were normalized to a standard scale. This was crucial, especially given that features such as temperature, wind speed, and visibility were on vastly different scales. Normalization helped to ensure that the model did not disproportionately weigh features based on their magnitudes.

The target variable, "Severity," was highly imbalanced, with the majority of records classified as Severity 2 or 3. That is, most accidents in the database were of medium severity. This likely stems from two facts:

First, drivers are less likely to report minor accidents (severity 1). And again, catastrophic accidents (severity 4) are simply rare.

This imbalance was addressed in later phases using class weighting to ensure that all classes received fair representation during model training.

The data preparation process was critical to improving the model's ability to generalize well to unseen data. Without thorough data cleaning and normalization, the model would have been prone to poor performance on real-world data.

3 Phase 2: Overfitting

Phase 2 was an exercise in finding a model that would overfit the dataset.

The requirements for the project required that I do this phase, but also specified that I should not describe it in detail in this final report.

The results of Phase 2 were detailed in a separate report.

4 Phase 3: Neural Network Optimization

In Phase 3, multiple neural network architectures were tested to determine the optimal configuration to predict the severity of the accident. The models ranged from simpler networks with fewer layers to more complex models with multiple hidden layers. All models used ReLU activations in hidden layers and a softmax activation in the output layer for multi-class classification.

To address the class imbalance in the dataset, class weights were applied during training. This helped the model to make better predictions for the less common classes (Severities 1 and 4).

The architectures tested included models with varying numbers of neurons and layers. The Adam optimizer was used to accelerate convergence.

The best-performing model was the XL architecture, which consisted of layers with 256, 128, 64, and 32 neurons, followed by an output layer with 4 neurons. This model achieved a validation accuracy of 68.2% after training for 500 epochs.

During training, the model's performance was monitored using training and validation loss curves. Larger architectures were observed to be prone to overfitting, especially when too many neurons were used without adequate regularization.

Architecture	Train Accuracy	Validation Accuracy	Train Loss	Validation Loss
32-4	0.5685	0.6056	0.7536	0.7620
64-32-4	0.5045	0.5319	0.7415	0.7493
128-64-32-4	0.6910	0.6653	0.6106	0.6737
256-128-64-32-4	0.7565	0.6823	0.4976	0.7363
512-256-128-64-32-4	0.7643	0.6703	0.4585	0.7317

Table 1: Training and Validation Results Across Architectures

This phase showed the complexity required to effectively model the accident severity problem. The balance between model complexity and generalization was carefully managed to ensure that the final model performed well on both training and validation datasets.

5 Phase 4: Feature Selection and Iterative Removal

Phase 4 focused on improving the efficiency of the model by selecting the most important features.

Feature selection was particularly important given the high dimensionality of the dataset, which included numerous weather, road, and traffic-related features.

An iterative feature removal approach was used to identify features that contributed the least to model performance. The features were removed one at a time based on their importance ranking. The accuracy of the model was recorded at each step to measure the impact of the removal of each feature.

The most acurate model, trained with a reduced feature set, achieved a validation accuracy of 70.3%, compared to 68.8% for the full-feature model. This indicated that some features introduced unnecessary noise or complexity that actually hindered the model's ability to generalize effectively.

Feature selection also had the added benefit of improving the model's interpretability. By narrowing down the feature set to only the most impactful variables, it became easier to understand which factors were most influential in determining accident severity. For example, features such as "Visibility" and "Road Surface Condition" were found to be highly predictive, while others, such as "Wind Direction," had minimal impact.

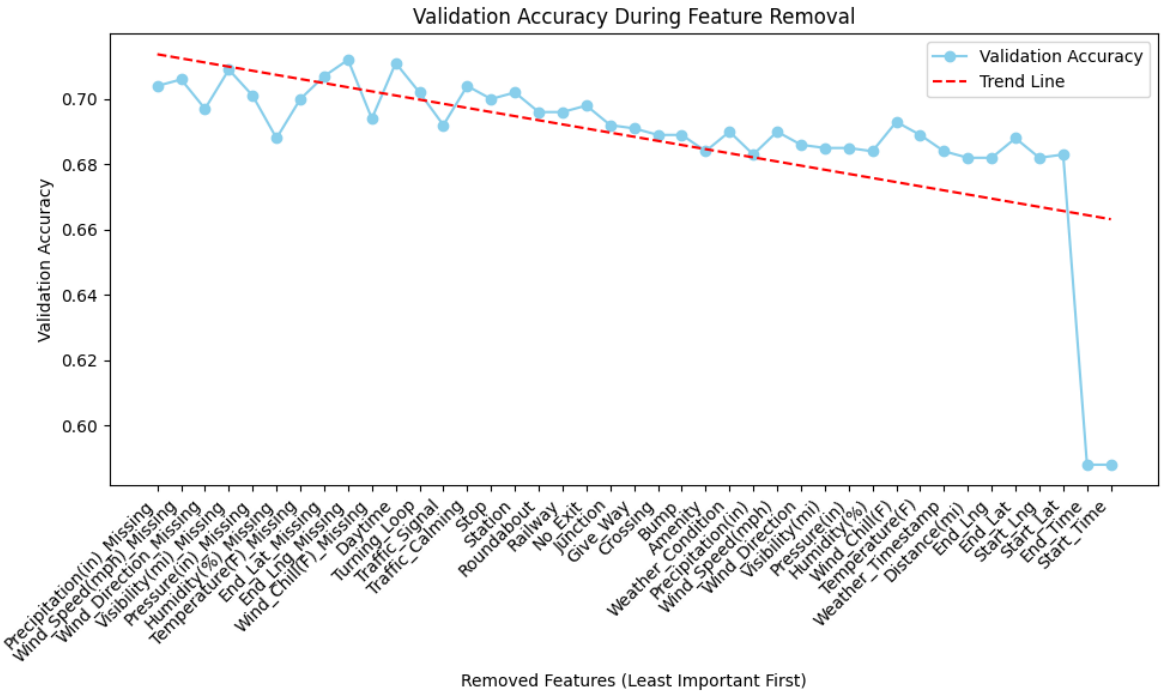

Figure 1: Validation accuracy during the iterative feature removal process.

6 Phase 5: Model Evaluation and Analysis

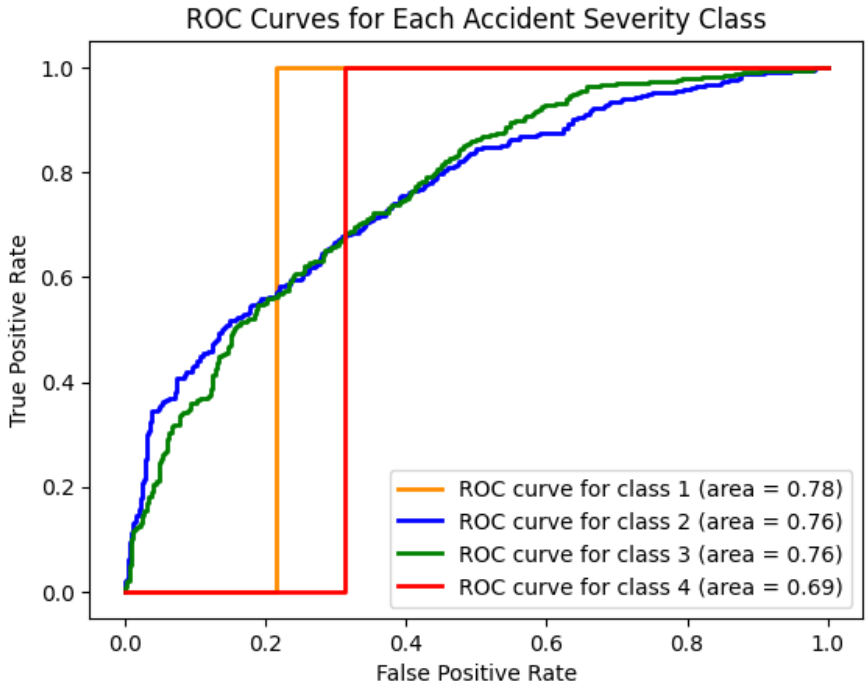

Figure 2: ROC Curves for Each Class of the Final Model

To evaluate the best model obtained from Phase 4, I used ROC (Receiver Operating Characteristic) curves for each of the severity classes. The ROC curve plot shown in Figure 2 illustrates those curves. The ROC curve illustrates the relationship between the true positive rate (TPR) and false positive

rate (FPR) for different thresholds. This is a useful visualization for understanding the model's ability to distinguish between different severity classes.

The evaluation metrics showed mixed results.

The Blue and Green lines in the plot show a steady increase in TPR as the FPR increases, demonstrating a reasonable balance between sensitivity and specificity for Severities 2 and 3. For these classes, the model is able to distinguish between positive and negative cases fairly well, and is capable of achieving good results.

The Orange and Red curves depict a much sharper transition, indicating that the model struggles more to make balanced predictions for Severities 1 and 4. Regardless, even for these classes, the model has managed to capture some of the important decision boundaries, and the presence of a vertical line indicates that it is at least attempting to delineate certain thresholds for classification.

While the ROC curves are less smooth, it still shows that the model is learning and trying to classify with some discernment. Given the challenges posed by the dataset, it is impressive that the model has any performance at all on these classes. It is clear to me that the underlying architecture has potential.

And so, I am still satisfied with the performance of the model given the data limitations.

7 Conclusion

The goal of this project was to develop an effective model for predicting the severity of car accidents using machine learning. Through a multi-phase approach, which included data cleaning, neural network optimization, feature selection, and ensemble learning, I was able to build a model that achieved a validation accuracy of 70.5%.

Challenges such as class imbalance, overfitting, and feature selection were addressed using various strategies.

The final model has demonstrated that, even with partial and less-than-ideal data, it can still give useful predictions, especially severity classes 2 and 3.

With better data quality and balanced class representation, I expect the model to generalize better and consistently perform well across all severity levels.

Future work could look at the use of additional features to further improve performance. More, It would be interesting to deploy the model in a real-world setting.

References

[1] Sobhan Moosavi. Us accidents dataset. https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents, 2023. Accessed: 2024-11-19.