Traffic Accident Severity Analysis and Prediction: Milestone Report

Authors: Tikki Cui (ttcui@vt.edu), Kshitiz Dhakal (kshitiz@vt.edu)

1. Introduction

Our project aims to analyze factors affecting traffic accidents and predict accident severity using machine learning techniques. Traffic accidents impact both lives and infrastructure significantly, and understanding their causes and severity can help inform preventative measures. We are analyzing a U.S. traffic accident dataset to uncover trends and build predictive models that could potentially improve safety outcomes.

2. Dataset Overview

The dataset used is sourced from Kaggle and includes approximately 500,000 records spanning February 2016 to March 2023, covering 49 states (excluding Hawaii). It contains 45 initial features related to accident details, such as time, location, weather, and road conditions. The raw data is sourced from several credible sources, such as the US Department of Transportation, traffic cameras, and traffic sensors. Overall the dataset provides a thorough and detailed overview of each accident and their surrounding circumstances.

One key feature is *Severity*. This feature categorizes how severe an accident is on a discrete scale ranging from 1 to 4. This feature is key because it serves as our target variable. We can use this feature for predictive models that take in accident details and attempt to classify how severe it was. This also gives us the chance to analyze any correlations between features and the severity. This gives us insights into what factors can influence accident severity.

3. Data Prepocessing

The first thing we did was to get rid of unnecessary features. There were 2 features, `ID` and `Source` that were purely metadata. Since these features make no meaningful contributions to our analysis, we dropped these features from the dataset.

Next, we used One-Hot Encoding to convert some categorical features into binary features so our machine learning models could work with the data. The 2 initial features we one-hot encoded were `State` and `Weather_Condition`. Although there are plenty of other categorical features, we don't currently plan on using them for any machine learning models. Therefore it's unnecessary and redundant to one-hot encode them. If future analysis requires it, we can always go back and encode the features we need.

Next, we fixed inconsistencies in the timestamp data. Some timestamps contained milliseconds while others did. For the ones that did, they were mostly all 0's. Since this level of precision is so small, we decided it was unnecessary to keep it. By removing the milliseconds from the timestamps, it ensured a consistent datetime format for all timestamp data. This will help ensure consistency and accuracy down the road when we starting training our models.

A feature we added is `Is_Weekend`. This feature indicates whether the accident occurred on a weekday or weekend. If an accident spans multiple days, we only account for the start time of the accident. This feature can be useful for analyzing frequency of accidents on

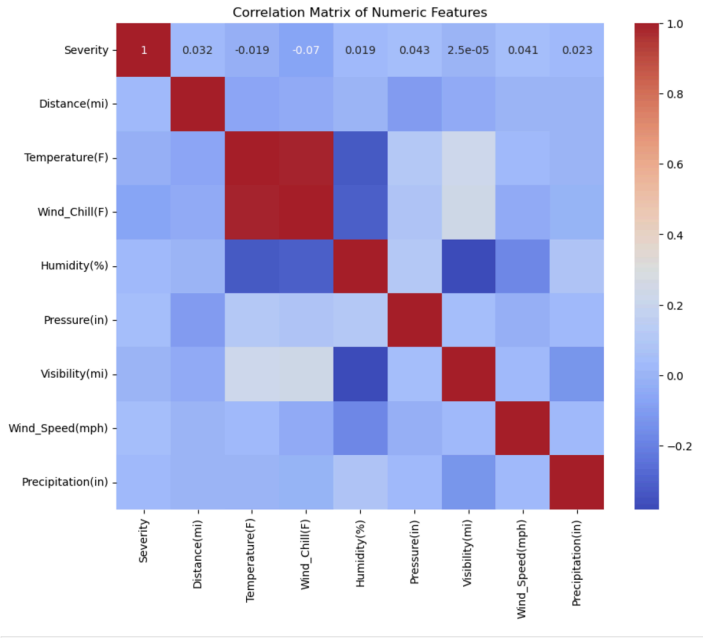
weekedays versus weekends. This may also help our models to gain another level of understanding as to what is causing these accidents.

We also created a feature that tracked the elapsed time for the accident. This was merely a way of encapsulating how long an accident lasted based on the start and end time. Although this information is technically present from the start and end, we think that capturing this data in an explicit feature will be useful for the models, especially for predicting the accident severity.

After all this preprocessing, we finished with 205 features to work with.

4. Initial Findings on dataset:

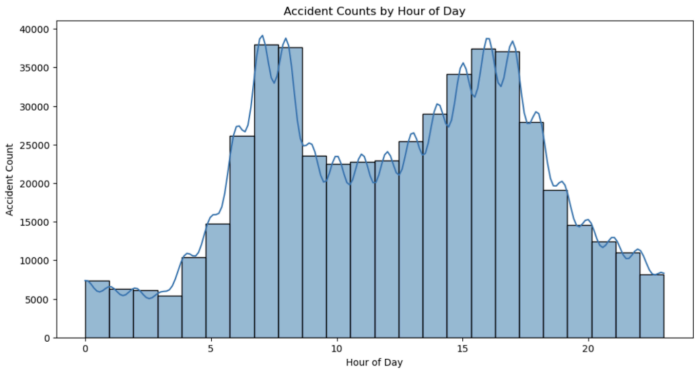
- **Correlation Matrix:** We generated a correlation matrix to identify any significant correlations among features. This analysis is guiding our feature selection for machine learning models. The **correlation heatmap** revealed that certain features, like visibility and precipitation, had moderate correlations with accident severity. This suggests that weather conditions could be meaningful predictors in the severity model.



- In addition to the correlation matrix, we also observed key patterns regarding the timing and frequency of accidents, as represented in the histplot and countplot:

    1. **Accident Timing**:

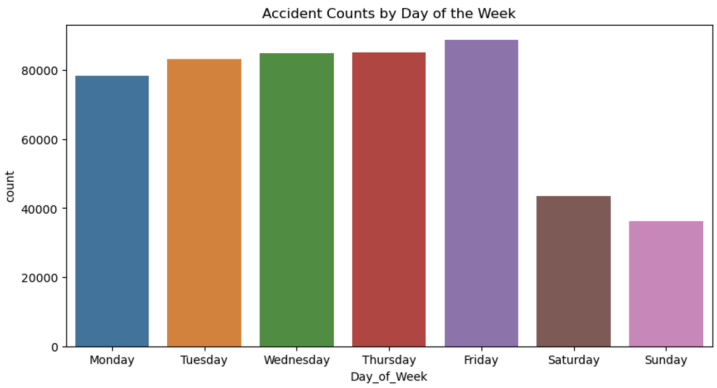
    - The histplot reveals that the highest number of accidents occur around **7-8 AM**, which coincides with the beginning of office hours, and **4-5 PM**, which is typically the end of office hours. These times correspond to peak traffic hours when people are commuting to and from work, making the roads more congested and increasing the likelihood of accidents.



- This observation suggests that traffic management systems could focus on these time periods to enhance safety, possibly through increased monitoring, deployment of traffic officers, or improved signaling.

2. **Accident Frequency by Day of the Week**:

- The countplot shows that **Friday** experiences the highest number of accidents, followed by Thursday, Wednesday, Tuesday, and Monday. This trend may be attributed to several factors, such as the start of the weekend, increased social activities, or people being more fatigued towards the end of the workweek.



- The least number of accidents occur on Saturdays and Sundays, which might indicate fewer vehicles on the road during the weekends or different traffic patterns, such as reduced commuting to work. People may also be more cautious when traveling for leisure, further contributing to the decrease in accidents during these days.

5. Machine Learning Models

We have also begun testing several supervised learning models, specifically focusing on multiclass classification to predict accident severity (appendix fig 1-8).

<u>Key Observations</u>

- Accuracy across all the models is consistently high, at 85%, indicating that the model has learned to classify most of the instances correctly, despite the imbalances in class distributions. However, the detailed evaluation metrics (precision, recall, and F1-score) for certain classes suggest that there are areas where the model's performance can be improved.

- Class Imbalance is evident, as the majority of predictions fall under class 2 (with recall and precision of 1.00 for this class), while the other classes (1, 3, and 4) have poor precision and recall (mostly 0.00). This suggests that the model predominantly predicts class 2, leading to a significant performance bias toward this class.

- Precision and Recall for Class 2: For most features, the precision and recall for class 2 are excellent (close to 1.00). This indicates that the model is very good at predicting class 2 instances, but this performance comes at the cost of other classes.

- Poor Performance for Classes 1, 3, and 4: In most cases, the precision, recall, and F1-score for classes 1, 3, and 4 are zero. These classes are likely underrepresented in the dataset, or the model may be unable to distinguish these classes due to their features being similar to class 2. This class imbalance can be problematic in applications where the classification of these minority classes is important.

Specific Feature Insights

- Wind Speed shows slightly better precision for class 3 compared to other features, with a precision of 1.00. However, recall for class 3 is 0.00, indicating that although class 3 predictions are precise when made, they are extremely rare, and the model fails to identify most of them.

- Macro Average Metrics: The macro average precision (around 0.21–0.24), recall (around 0.25), and F1-score (around 0.23) are quite low, showing that the model's performance is not consistent across all classes. The macro average treats all classes equally, so these metrics suggest that the model struggles with balancing the minority and majority classes.

- Weighted Average: The weighted averages of precision (0.73), recall (0.85), and F1-score (0.79) indicate that the model's performance is better when weighted by class distribution. This highlights the dominance of class 2 in the dataset, but also suggests that for a balanced or fair classification task, improvements are needed for classes 1, 3, and 4.

Challenges

Several challenges have arisen in the project so far:

- Data Imbalance: Severity classes are not equally represented, with lower severity classes being more common. We may need to apply techniques like SMOTE or class weighting to address this imbalance.

- Feature Engineering Complexity: Engineering new features, particularly from time and location data, required careful handling to avoid introducing bias.

- Computational Resources: Training complex models with large feature sets is computationally intensive. We are exploring ways to optimize model complexity without sacrificing performance.

6. Next Steps

Our immediate next steps are:

1. **Preprocess Missing Data:** Currently, our dataset contains missing data and we haven't done anything to preprocess it. We plan to investigate each feature that contains missing

data and come up with a strategy to rectify it. For each feature, we will try to find a way to impute the missing data, only dropping the feature when all else fails. The goal is to find imputation techniques that best preserves accuracy without over complex solutions.

2. **Additional Feature Engineering:** As we continue to dive into the dataset, we can engineer more features to help improve model accuracy. These features would not be presenting existing data in a different format, like one-hot encoding. Rather, these would be brand new features that provide meaningful insights into the problem.

3. **Optimize Model Performance:** Experiment with class balancing techniques and additional hyperparameter tuning to improve model performance across all severity classes.

4. **Explore Additional Models:** We plan to test alternative machine learning models, including Random Forest and Gradient Boosting, to compare with our initial neural network results.

5. **Further Visualization:** Create more detailed visualizations to understand the role of weather and time in accident severity. This will provide insights that can guide feature selection for final models.

6. **Clustering and PCA Analysis:** To understand potential patterns or clusters within the dataset, we will apply KMeans clustering and PCA for dimensionality reduction. This will allow us to visually analyze groupings and may help with feature selection.

7. Conclusion and Expected Outcomes

Through this milestone, we have established a foundational understanding of the dataset and identified preliminary patterns that may influence traffic accident severity. By improving our models and deepening our analysis, we hope to develop a reliable framework for predicting accident severity. These insights could inform strategies for risk mitigation, such as implementing weather-specific warnings or adjusting road conditions in high-risk areas.

8. Appendix

ML Model training to classify severity levels from different features
1. Humidity and Severity

Classification Report of Humidity and Severity:

	precision	recall	f1-score	support
1	0.00	0.00	0.00	792
2	0.85	1.00	0.92	57966
3	0.00	0.00	0.00	7517
4	0.00	0.00	0.00	1628
accuracy			0.85	67903
macro avg	0.21	0.25	0.23	67903
weighted avg	0.73	0.85	0.79	67903

2. Visibility and Severity

```
Classification Report of Visibility and Severity:
              precision    recall  f1-score   support

           1       0.00      0.00      0.00       792
           2       0.85      1.00      0.92     57966
           3       0.00      0.00      0.00      7517
           4       0.00      0.00      0.00      1628

    accuracy                           0.85     67903
   macro avg       0.21      0.25      0.23     67903
weighted avg       0.73      0.85      0.79     67903
```

3. Distance and Severity

```
Classification Report of Distance and Severity:
              precision    recall  f1-score   support

           1       0.00      0.00      0.00       792
           2       0.85      1.00      0.92     57966
           3       0.00      0.00      0.00      7517
           4       0.11      0.00      0.00      1628

    accuracy                           0.85     67903
   macro avg       0.24      0.25      0.23     67903
weighted avg       0.73      0.85      0.79     67903
```

4. Temperature and Severity

```
Classification Report of Temperature and Severity:
              precision    recall  f1-score   support

           1       0.00      0.00      0.00       792
           2       0.85      1.00      0.92     57966
           3       0.00      0.00      0.00      7517
           4       0.00      0.00      0.00      1628

    accuracy                           0.85     67903
   macro avg       0.21      0.25      0.23     67903
weighted avg       0.73      0.85      0.79     67903
```

5. Wind chill and Severity

```
Classification Report of Wind_Chill and Severity:
              precision    recall  f1-score   support

           1       0.00      0.00      0.00       792
           2       0.85      1.00      0.92     57966
           3       0.00      0.00      0.00      7517
           4       0.00      0.00      0.00      1628

    accuracy                           0.85     67903
   macro avg       0.21      0.25      0.23     67903
weighted avg       0.73      0.85      0.79     67903
```

6. Pressure and Severity

```
Classification Report of Pressure and Severity:
              precision    recall  f1-score   support

           1       0.00      0.00      0.00       792
           2       0.85      1.00      0.92     57966
           3       0.00      0.00      0.00      7517
           4       0.00      0.00      0.00      1628

    accuracy                           0.85     67903
   macro avg       0.21      0.25      0.23     67903
weighted avg       0.73      0.85      0.79     67903
```

7. Wind speed and Severity

```
Classification Report of Wind_Speed and Severity:
              precision    recall  f1-score   support

           1       0.00      0.00      0.00       792
           2       0.85      1.00      0.92     57966
           3       1.00      0.00      0.00      7517
           4       0.00      0.00      0.00      1628

    accuracy                           0.85     67903
   macro avg       0.46      0.25      0.23     67903
weighted avg       0.84      0.85      0.79     67903
```

8. Precipitation and Severity

```
Classification Report of Precipitation and Severity:
              precision    recall  f1-score   support

           1       0.00      0.00      0.00       792
           2       0.85      1.00      0.92     57966
           3       0.00      0.00      0.00      7517
           4       0.00      0.00      0.00      1628

    accuracy                           0.85     67903
   macro avg       0.21      0.25      0.23     67903
weighted avg       0.73      0.85      0.79     67903
```