Factors Influencing Traffic Accident Severity

Maryam Saamin

Abstract. This study investigates factors influencing traffic accident severity using a dataset spanning February 2016 to March 2023, covering 49 states in the USA. With 7.7 million accident records and 46 features, including weather conditions, time of day, points of interest, and traffic-related attributes, the analysis aims to understand patterns and correlations. Machine learning models, including Random Forests, XGBoost, Decision Trees, and MLPClassifier, are used to predict accident severity. Results indicate an accuracy range of 75.7% to 80.4%, with significant variations in precision, recall, and F1-score across severity levels. Insights from this study can inform strategies for accident prevention and mitigation.

Keywords. US-Accident, Severity of Accident, ML Prediction Accident

1. Introduction

This project involves analyzing a comprehensive dataset of car accidents covering 49 states in the USA from February 2016 to March 2023. With around 7.7 million records and 46 columns, the dataset offers rich insights into various factors affecting traffic accidents.

2. Question and objectives

Question:

What factors contribute to traffic accident severity?

- Investigate the impact of various features such as weather conditions, time of day, points of interest (POI), and traffic-related attributes on accident severity.

- Explore correlations and patterns among these factors.

Objectives:

1. Data Preprocessing:
    - Relevant data on traffic accidents, including attributes such as weather conditions, time, POI, and traffic-related variables. (Region, Accident and incident, etc.)
    - Clean and preprocess the data to handle missing values, outliers.
2. Feature Engineering:
    - Create meaningful features from raw data. For example:
    - Extract time-related features (hour of the day, day of the week).
    - Encode categorical features (e.g., weather conditions, POI, etc.) using one-hot encoding or embeddings.
3. Exploratory Data Analysis:
    - Visualize the relationships between features and accident severity.
    - Identify any trends, anomalies, or patterns.
4. Model Building and Evaluation:
    - Develop predictive models using ML techniques (e.g., random forests, gradient boosting, decision tree and neural network model).
    - Split the dataset into training and validation sets.
    - Evaluate model performance using appropriate metrics (accuracy, precision, recall, F1-score).
5. Interpretability and Insights:
    - Interpret model predictions to understand the importance of different features.
    - Identify which factors contribute most to severe accidents.

3. Key Data Points and Columns

1. Location and Time:

- Start_Time: The timestamp when the accident was reported.
- End_Time: The timestamp when the accident was cleared or resolved.
- Start_Lat, Start_Lng: Latitude and longitude coordinates of the accident location.
- End_Lat, End_Lng: Latitude and longitude coordinates of the accident's end location (if applicable).
- Distance(mi): The distance of the accident from a reference point (e.g., intersection, landmark).

2. Description and Context:

- Description: A natural language description of the accident.
- Weather_Timestamp: The timestamp when weather conditions were recorded.
- Weather_Condition: Describes the weather conditions at the time of the accident.
- Timezone: The time zone of the accident location.

3. Severity and Impact:

- Severity: The severity level of the accident (e.g., minor, moderate, severe).
- Amenity, Bump, Crossing, etc.: Binary indicators for the presence of certain amenities or road features near the accident location.

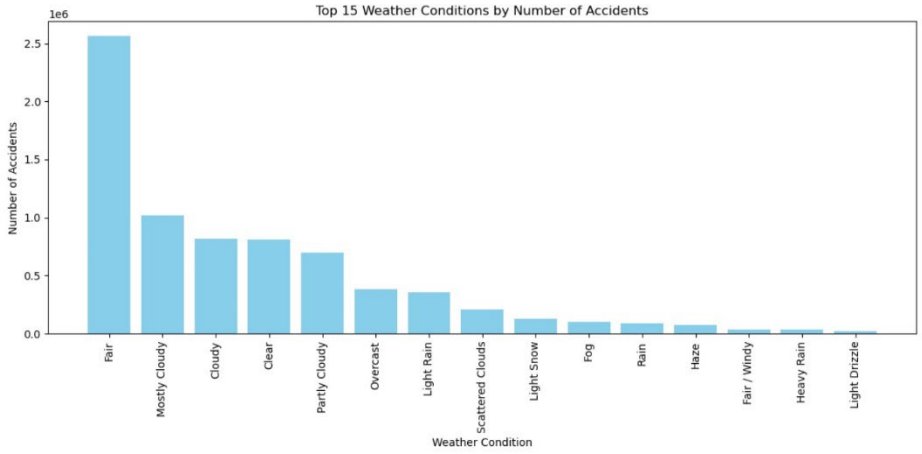
4. Points of Interest (POI):

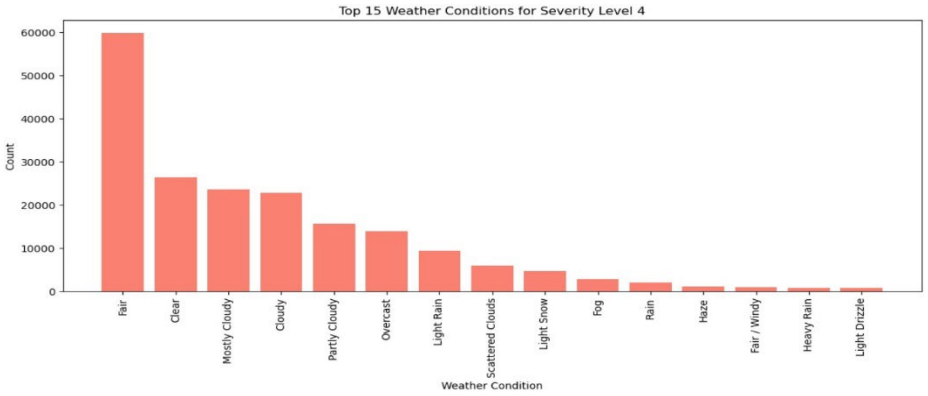
- Railway, Station, Traffic_Signal, etc.: Binary indicators for the presence of specific points of interest near the accident location.

4. Data Visualization

Weather condition:

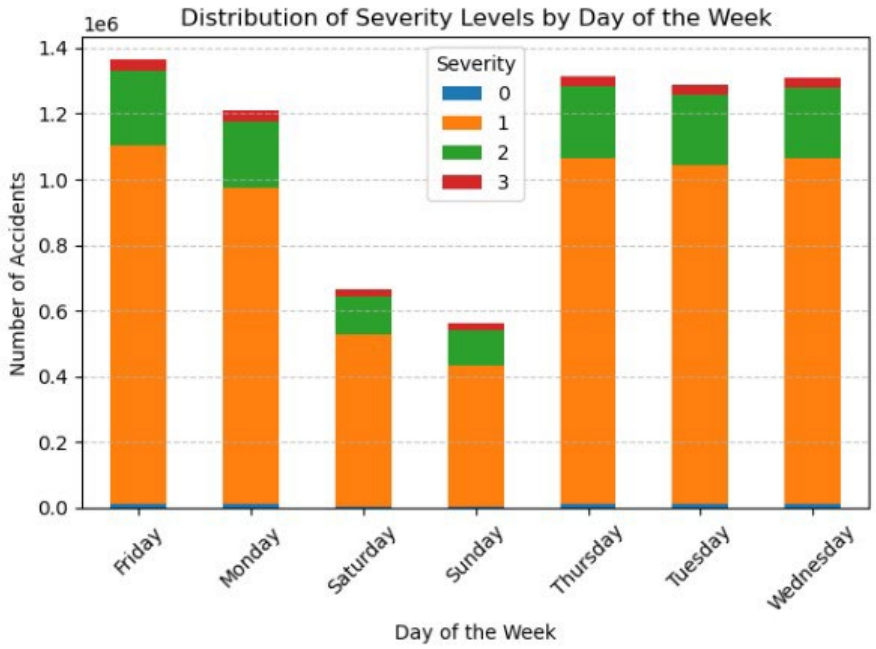
Out of the 144 different weather types, fair weather is the most common among the top 15 conditions for both total accidents and those marked as severity level 4. This shows that even in good weather, a lot of accidents happen, so it's important to figure out why.





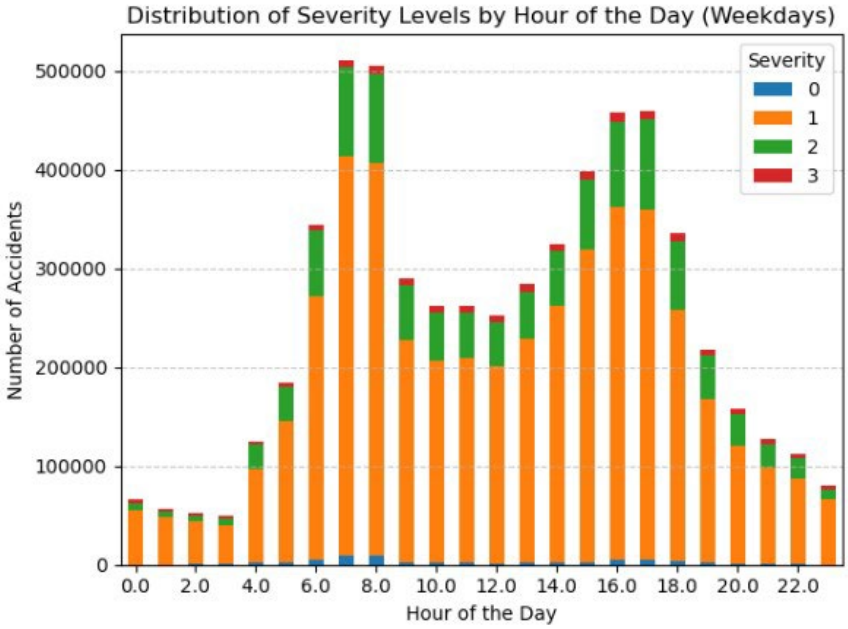
Accidents in the week:

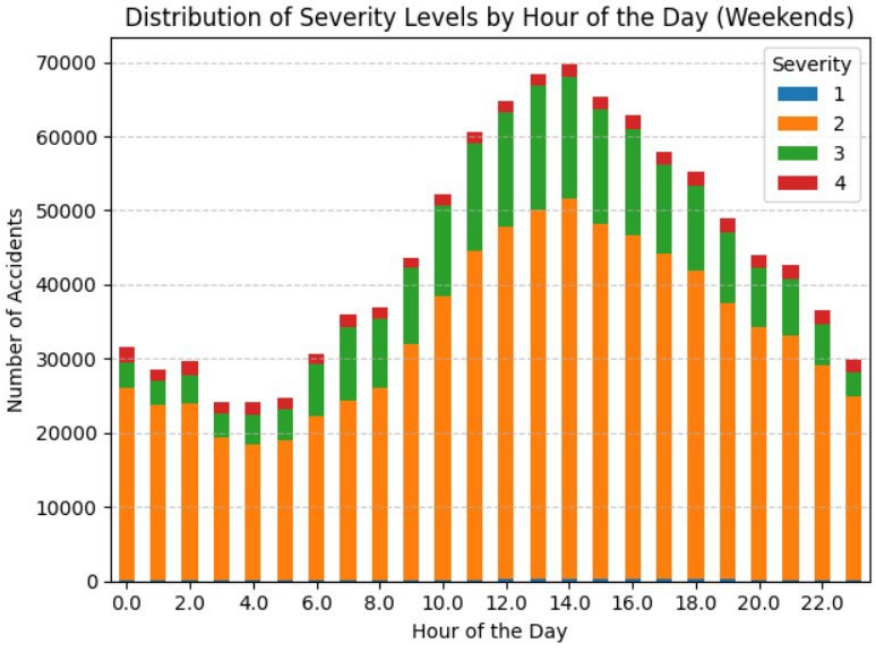
In the distribution of severity levels by day of the week, Friday consistently records the highest number of accidents.



Peak Accident Hours on Weekdays and Weekends:

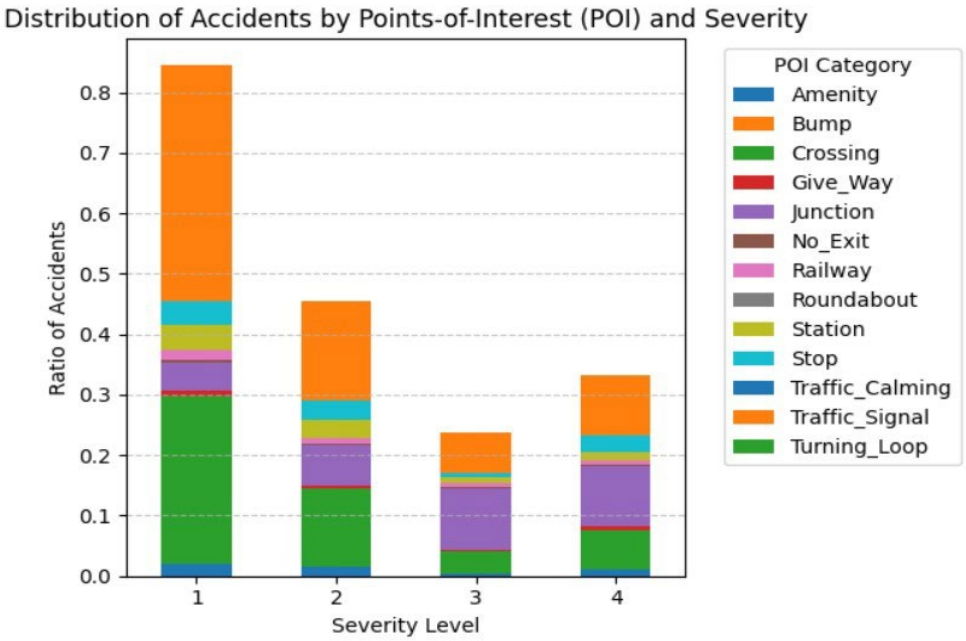
During weekdays, the highest number of accidents occurs between 7 to 8 AM and then again between 4 to 5 PM. However, during weekends, there's a notable increase in accidents between 12 to 2 PM, followed by a decrease afterwards.

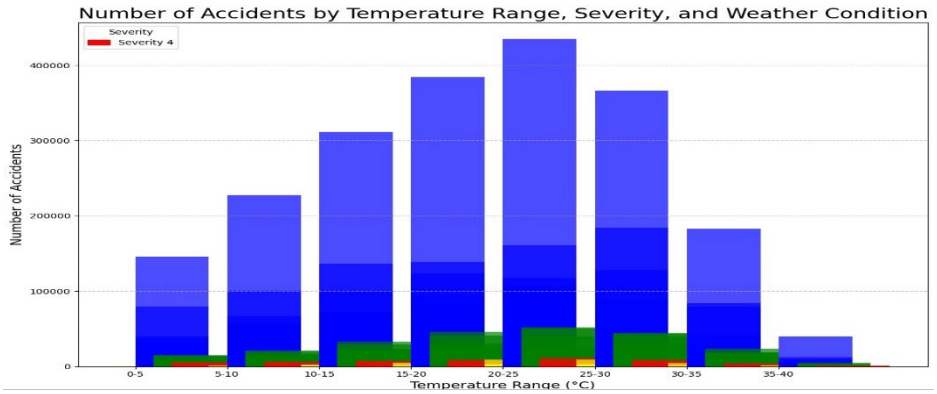

Distribution of Severity Levels by Hour of the Day (Weekdays)
Severity
0
1
2
3
Number of Accidents
500000
400000
300000
200000
100000
0
0.0
2.0
4.0
6.0
8.0
10.0
12.0
14.0
16.0
18.0
20.0
22.0
Hour of the Day


Distribution of Severity Levels by Hour of the Day (Weekends)
Severity
1
2
3
4
Number of Accidents
70000
60000
50000
40000
30000
20000
10000
0
0.0
2.0
4.0
6.0
8.0
10.0
12.0
14.0
16.0
18.0
20.0
22.0
Hour of the Day

Analysis of Accident Distribution by Point of Interest, Severity, and Weather Temperature:

This analysis looks at how accidents are spread out across different places, how severe they are, and what the weather is like. More accidents happen when the temperature is between 25°C and 30°C. Also, it was noticed that where accidents happen and how severe they are varied. Two images are included to help understand these findings better.





5. Models' Accuracy and insights

XGBoost Model:

XGBoost model shows 80.40% accuracy and an AUC score of 0.848. While it excels in classifying label '1', precision and recall for other classes are lower, suggesting room for improvement, particularly in balancing class performance. Fine-tuning may enhance overall effectiveness.

Table 1.

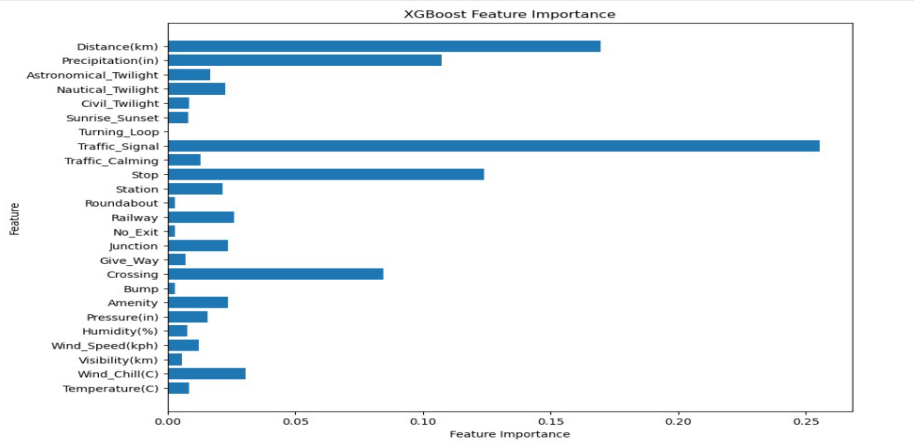
XGBoost Accuracy: 0.8040032891693554

Classification Report:

	precision	recall	f1-score	support
0	0.57	0.02	0.05	13509
1	0.82	0.97	0.89	1230523
2	0.56	0.17	0.26	260525
3	0.47	0.02	0.04	41122
accuracy			0.80	1545679
macro avg	0.60	0.30	0.31	1545679
weighted avg	0.76	0.80	0.75	1545679

AUC Score: 0.8484855836431987

Figure 1.



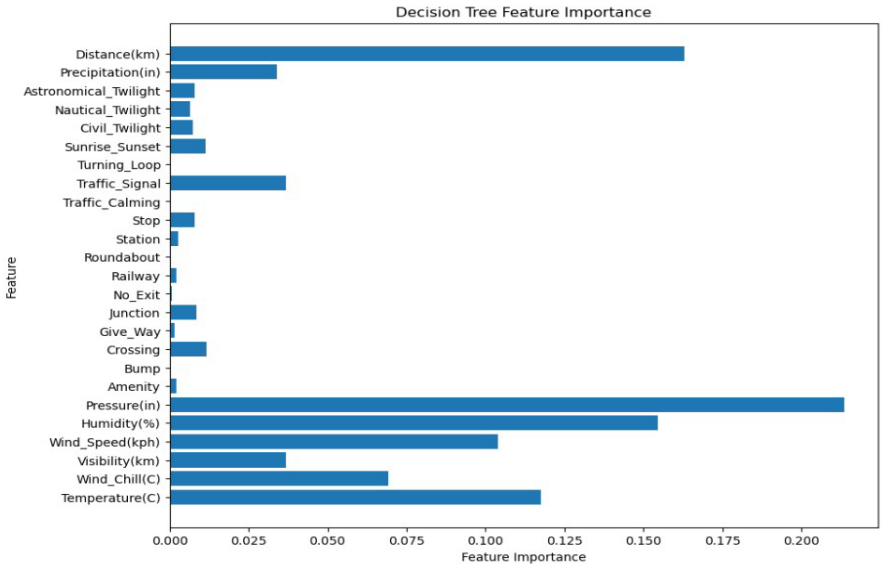
Decision Tree Model:

The Decision Tree model reached 75.73% accuracy and an AUC score of 0.649. It performs fairly well in classifying some labels, but there's room for improvement, especially in distinguishing between different classes. Fine-tuning could help boost its overall effectiveness.

Table 2.

```
Accuracy: 0.7572917792115956
Decision Tree Classification Report:
              precision    recall  f1-score   support

           1       0.23      0.29      0.26     13509
           2       0.85      0.86      0.85   1230523
           3       0.42      0.40      0.41    260525
           4       0.22      0.23      0.23     41122

    accuracy                           0.76   1545679
   macro avg       0.43      0.44      0.44   1545679
weighted avg       0.76      0.76      0.76   1545679

AUC Score: 0.6491496915094546
```

Figure 2.



Random Forest Model:

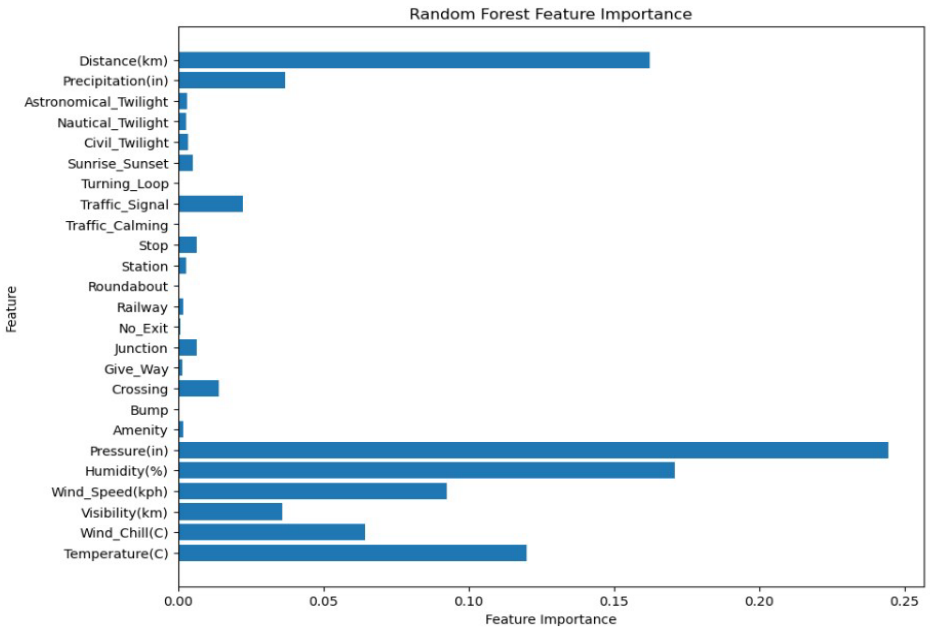
The Random Forest model scored 80.35% accuracy and an AUC of 0.830. While it's strong in classifying label '2', performance varies for other labels, suggesting room for improvement. Adjustments could help better balance its performance across different classes.

Table 3.

```
Accuracy: 0.803535533574565

Classification Report:
              precision    recall  f1-score   support

           1       0.52      0.22      0.31     13509
           2       0.85      0.93      0.88   1230523
           3       0.52      0.35      0.42    260525
           4       0.44      0.19      0.26     41122
```

```
    accuracy                           0.80   1545679
   macro avg       0.58      0.42      0.47   1545679
weighted avg       0.78      0.80      0.78   1545679

AUC: 0.830043757008013
```

Figure 3.



MLP Classifier:

Accuracy: 0.8005594952121365

The MLP Classifier attained an accuracy of 80.06%, indicating its capability to make correct predictions across the dataset. However, further analysis with additional metrics would provide a clearer understanding of its performance compared to other models.

6. Conclusion

In conclusion, weather conditions significantly influence the severity of accidents, with most incidents occurring during fair weather. Moreover, the timing of accidents varies between weekdays and weekends, with Friday being the day with the highest accident rate. Factors such as humidity, pressure, temperature, wind speed, wind chill, visibility, and precipitation also play crucial roles in determining accident severity. Understanding these patterns is vital for implementing targeted public safety measures aimed at reducing accident rates and mitigating their impacts, particularly during peak accident periods and weather conditions.

References

[1] Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. "A Countrywide Traffic Accident Dataset.", arXiv preprint arXiv:1906.05409 (2019)

[2] Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.

[3] https://smoosavi.org/datasets/us_accidents

[4] https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents