Analysis of US Traffic Accidents

CIS 5450

Contributors: Yue Wang, Raafae Zaki, Aria Xingni Shi

Penn Engineering
University of Pennsylvania

Objective and Value Proposition

- **Objective**: Analyze US traffic accident severity using various factors (time, weather, location).
- **Motivation**: Importance of predicting accident severity for public safety and traffic management.

Dataset Overview

Source: U.S. Traffic Accidents (2016–2023).

- **Scope**: 49 states, 500,000 sampled accidents from a dataset of 7.7M records.
- **Collection**: Real-time APIs gathering data from cameras, sensors, and law enforcement.
- **Example**: Display a snippet of the dataset。

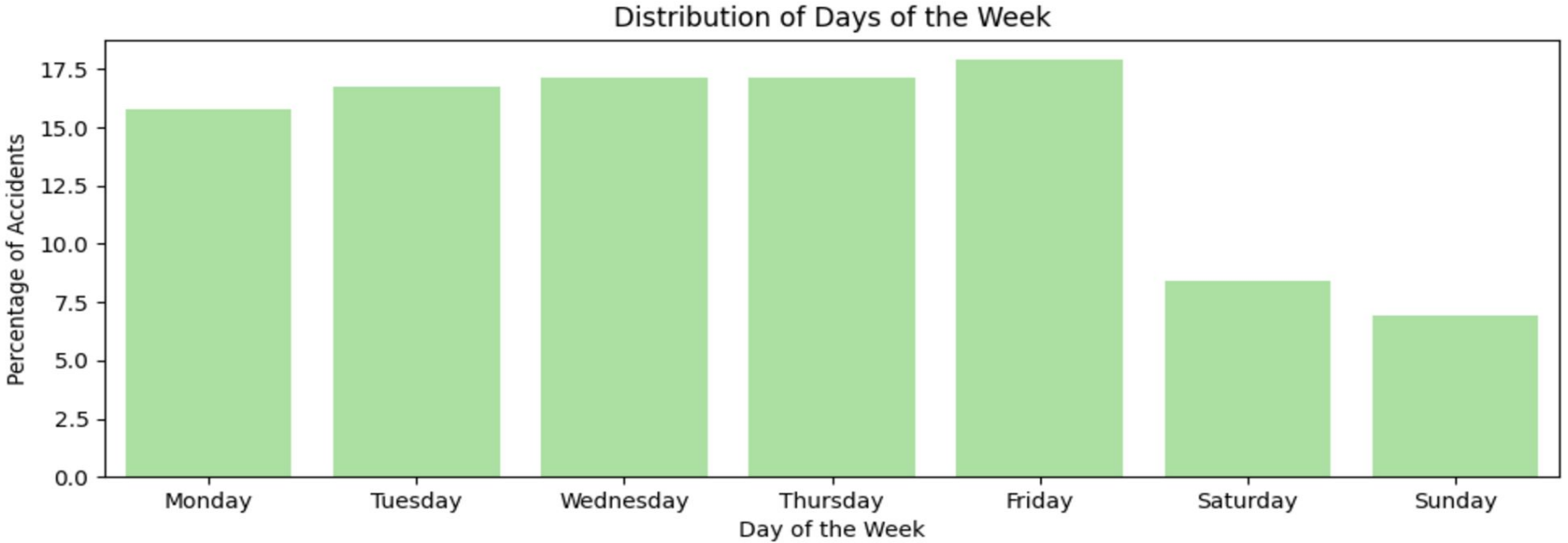
	ID	Source	Severity	Start_Time	End_Time	Start_Lat	Start_Lng	End_Lat	End_Lng	Distance(mi)	...	Roundabout	Station	Stop	Traffic_Calming	Traffic_Signal	Turning_Loop
0	A–2047758	Source2	2	2019-06-12 10:10:56	2019-06-12 10:55:58	30.641211	-91.153481	NaN	NaN	0.000	...	False	False	False	False	True	False
1	A–4694324	Source1	2	2022-12-03 23:37:14.000000000	2022-12-04 01:56:53.000000000	38.990562	-77.399070	38.990037	-77.398282	0.056	...	False	False	False	False	False	False
2	A–5006183	Source1	2	2022-08-20 13:13:00.000000000	2022-08-20 15:22:45.000000000	34.661189	-120.492822	34.661189	-120.492442	0.022	...	False	False	False	False	True	False
3	A–4237356	Source1	2	2022-02-21 17:43:04	2022-02-21 19:43:23	43.680592	-92.993317	43.680574	-92.972223	1.054	...	False	False	False	False	False	False
4	A–6690583	Source1	2	2020-12-04 01:46:00	2020-12-04 04:13:09	35.395484	-118.985176	35.395476	-118.985995	0.046	...	False	False	False	False	False	False

5 rows × 46 columns

Major Findings From EDA

Accident distribution for day of week

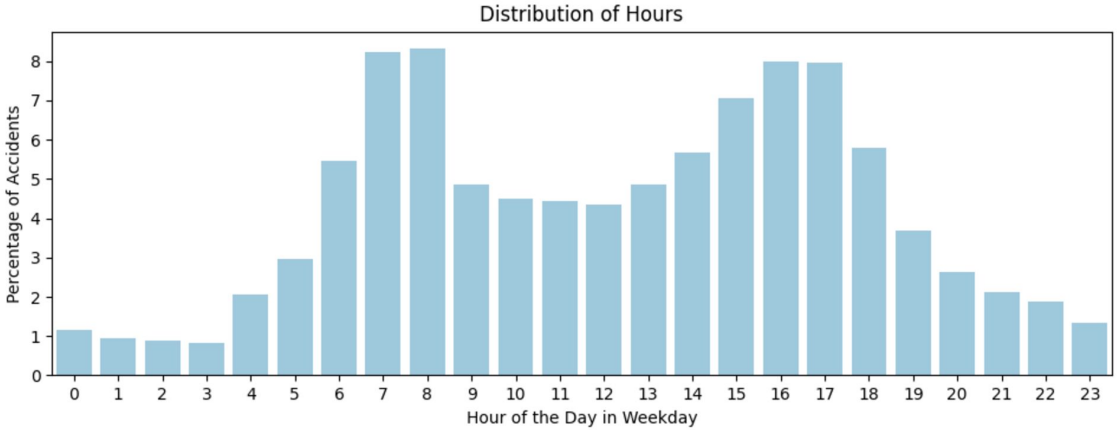
- More accidents occur from Monday to Friday compared to Saturday and Sunday.



Major Findings From EDA

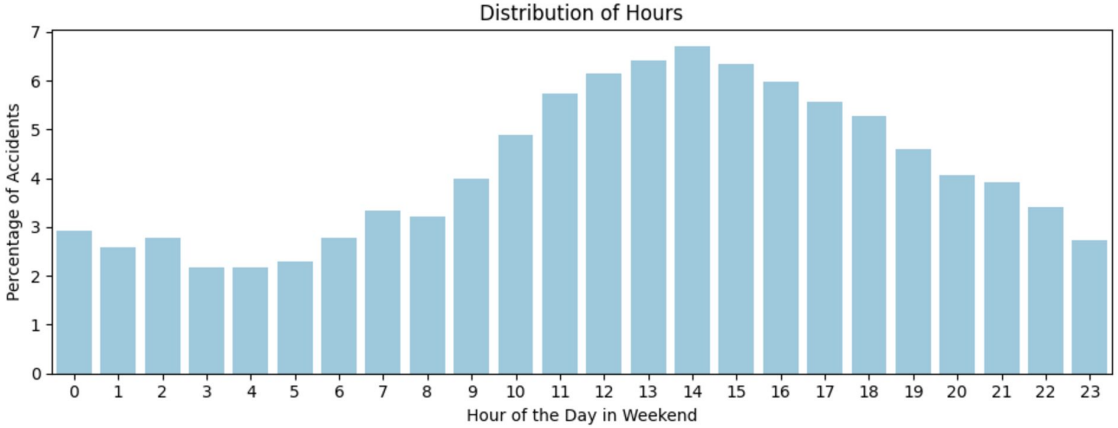
Accident distribution for hour of weekday

- The majority of accidents happen during the morning peak (7-8 AM) and the afternoon peak (3-6 PM)



Accident distribution for hour of weekend

- The majority of accidents happen during the afternoon peak (12-4 PM)

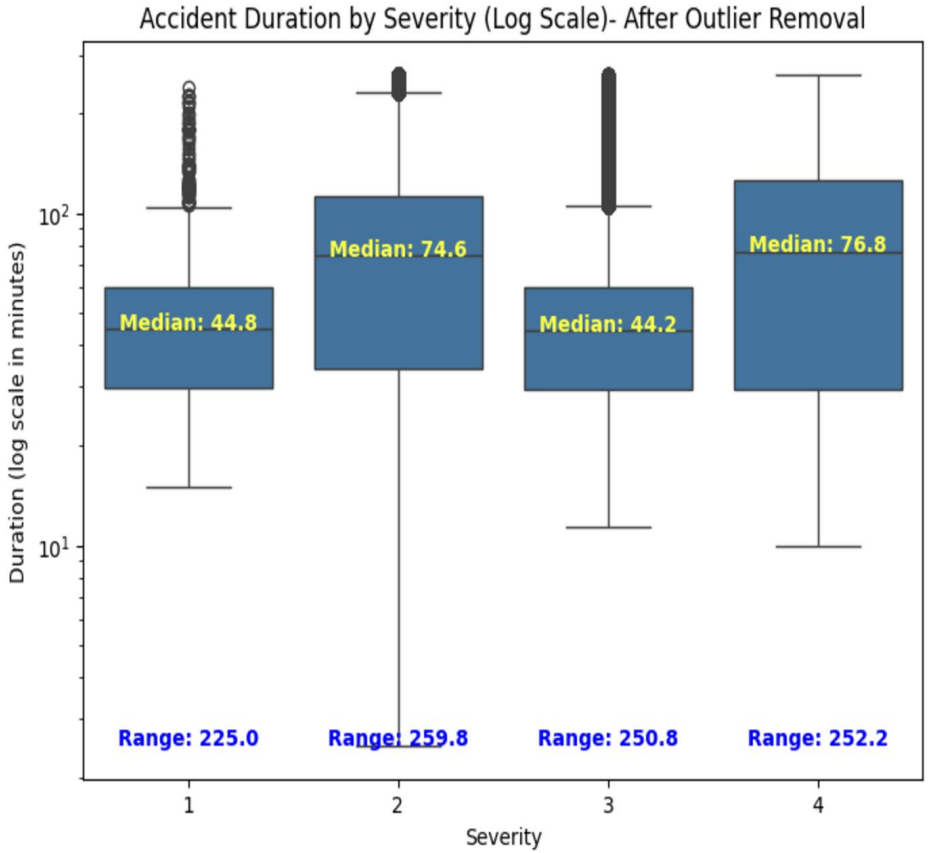


Penn Engineering

Major Findings From EDA

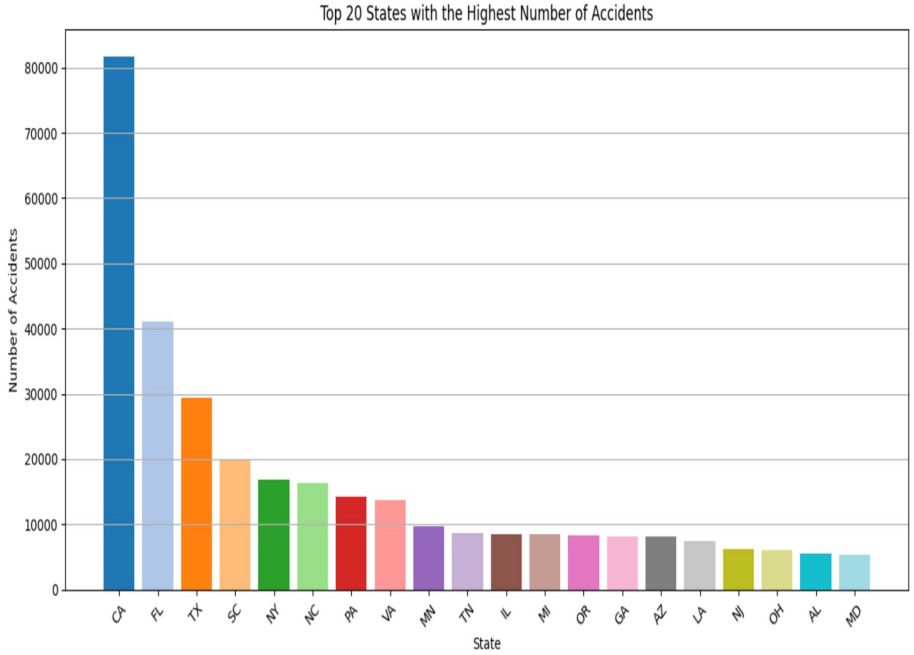
Accident distribution for duration

- The graph displays accident durations on a logarithmic scale (y-axis) across four severity levels (1-4 on x-axis)
- Duration is measured in minutes, using a log scale from 10^1 to 10^2 (10 to 100 minutes)
- Similar range values
- Similar median values for 1 and 3
- Similar median values for 2 and 4

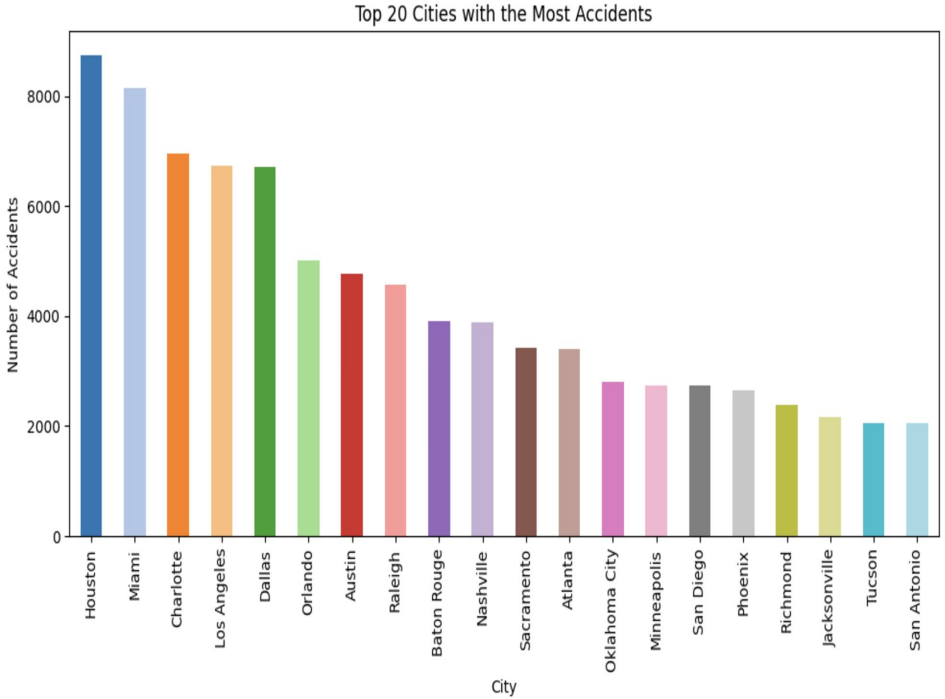


Major Findings From EDA

Top 20 states with the most accidents



Top 20 cities with the most accidents

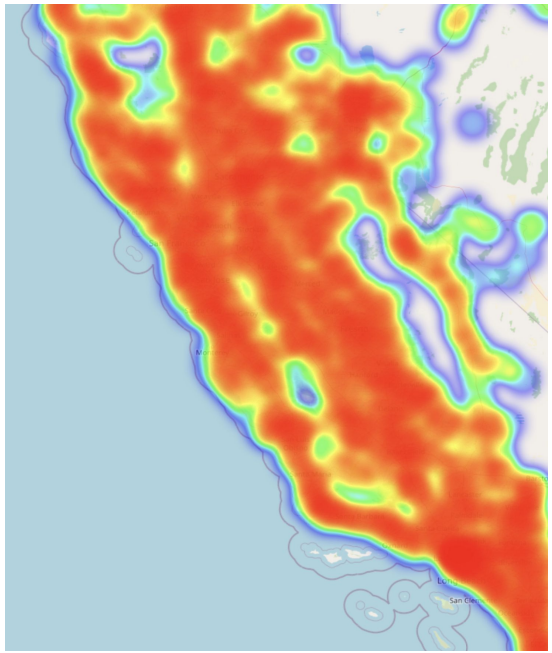


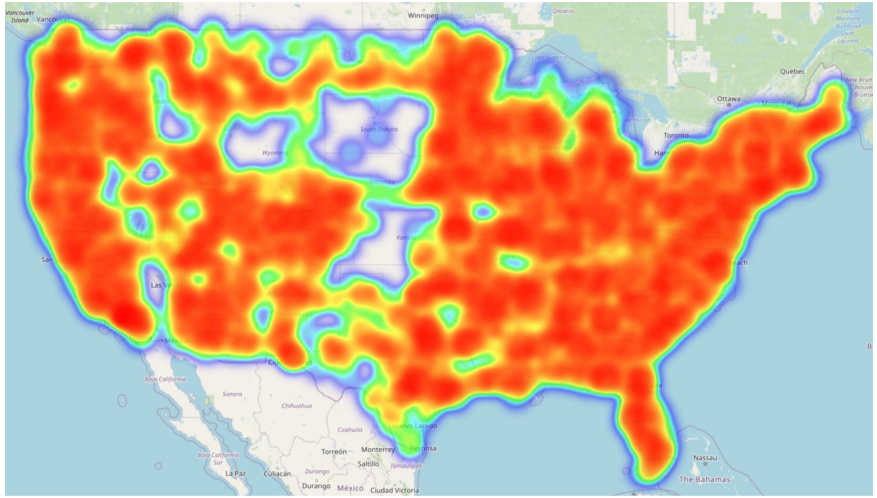
Penn Engineering

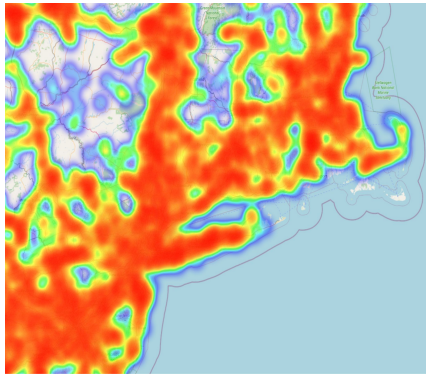
Major Findings From EDA

Visualize the location of accidents on the map

- **East coast** **Whole map** **West coast**







Modeling

Penn Engineering

Overview

Objective

Predict **severity** of car accidents in the US based on multiple factors to help inform decision-making and resource allocation for emergency services and safety regulations.

Method

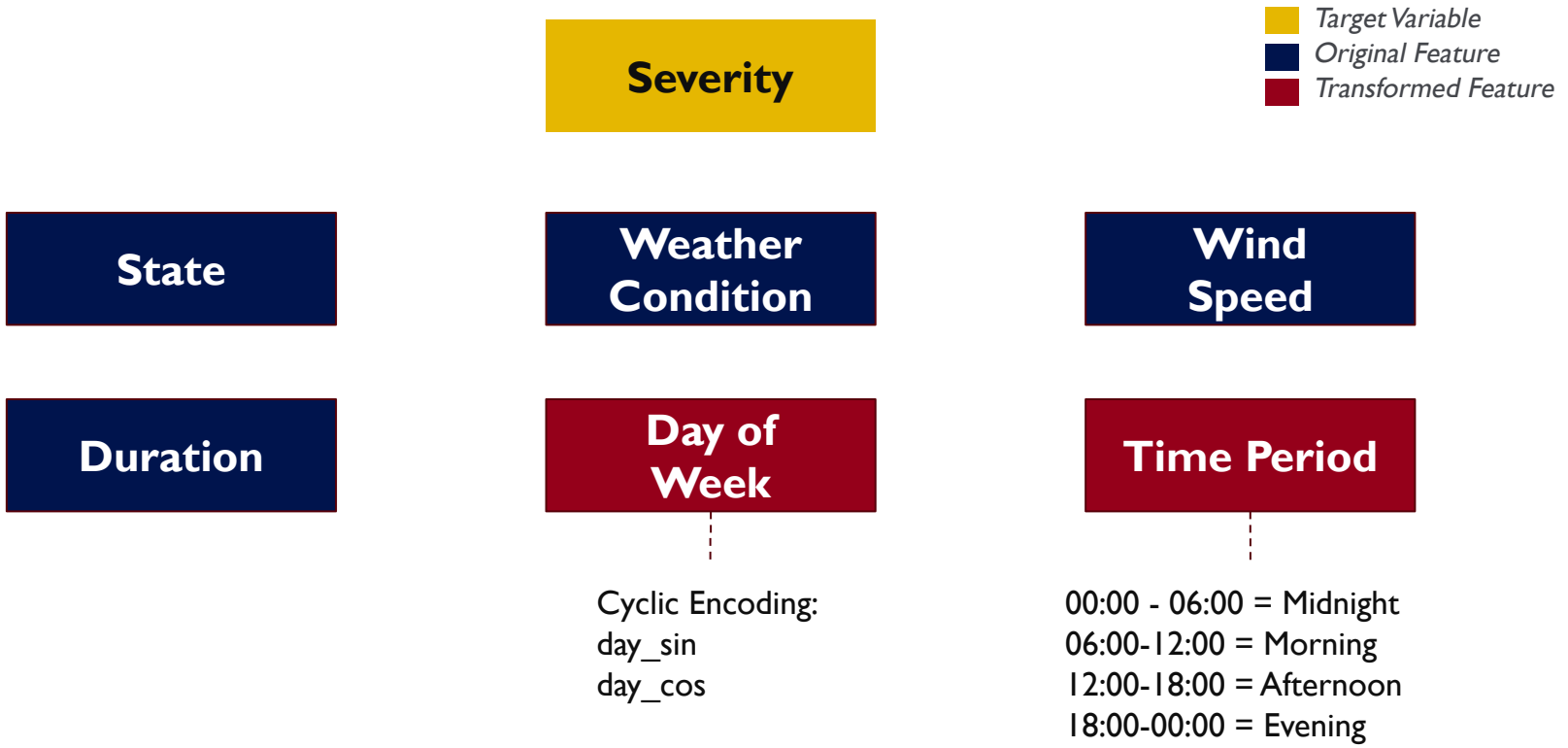
Target variable of severity score is a set of discrete values, ranging from 1-4. Therefore, we must use **classification** for our predictions.

Metric

Evaluate model performance using **recall** in order to minimize the false negative rate, reducing the chance of under-predicting accident severity which could be very costly.

Note: accuracy is unreliable due to high class imbalance, while precision is not as important since the cost of false negatives exceeds the cost of false positives in this use case.

Feature Engineering & Selection



Data Pre-processing

1. **Encode**: apply One-Hot Encoding to categorical features (nominal variables), cyclic encoding to *Day of Week*, and manually shift *Severity* (target variable) from 1-4 to 0-3 for appropriate model fitting

2. **Calculate Class-Imbalance**: found ~71x more observations in the majority class compared to minority class, requiring stratification

3. **Train–Test Split**: 80/20 split with stratification and shuffling of the dataset

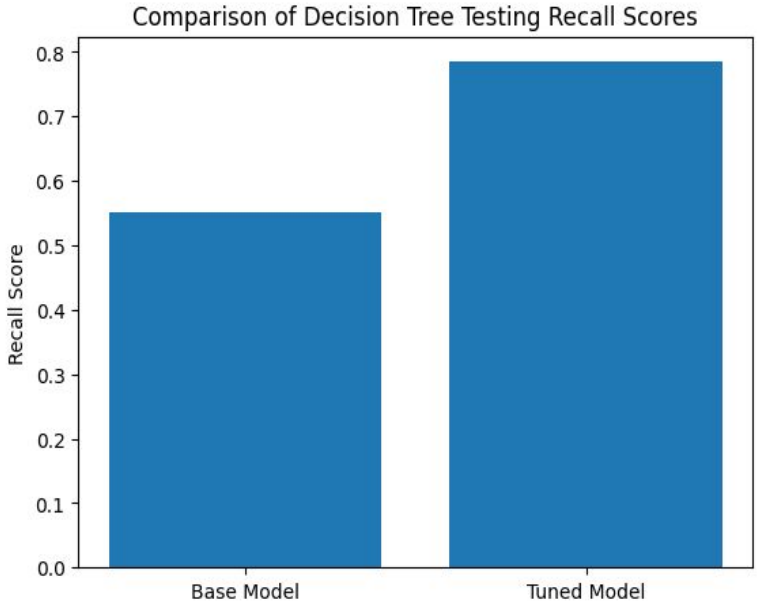
4. **Scale**: apply Standard Scaler to numerical features to reduce bias

Modeling Approach

Steps	Process
Setup	Define model with initial guess parameters and fit to training data
Evaluation	Predict on training and testing data, and calculate recall score
Tuning	Perform randomized search over parameter distribution with 5-fold cross-validation and extract the most optimal hyperparameters
Comparison	Compare testing recall between base model (using initial hyperparameters) and tuned model (using optimal hyperparameters)

Decision Tree Classifier *(Baseline)*

- **Setup:**
  a. Start with initial hyperparameters
  b. Establish "balanced" weighting to offset class imbalance
- **Evaluation:** base vs. tuned
  a. Training Recall = 62% → a. Training Recall = 82%
  b. Testing Recall = 55% → b. Testing Recall = 78%
- **Tuning:**
  a. Find most optimal hyperparameters across a randomized grid
  b. Employ 5-fold cross-validation for robustness
  c. Maximize the "recall" metric

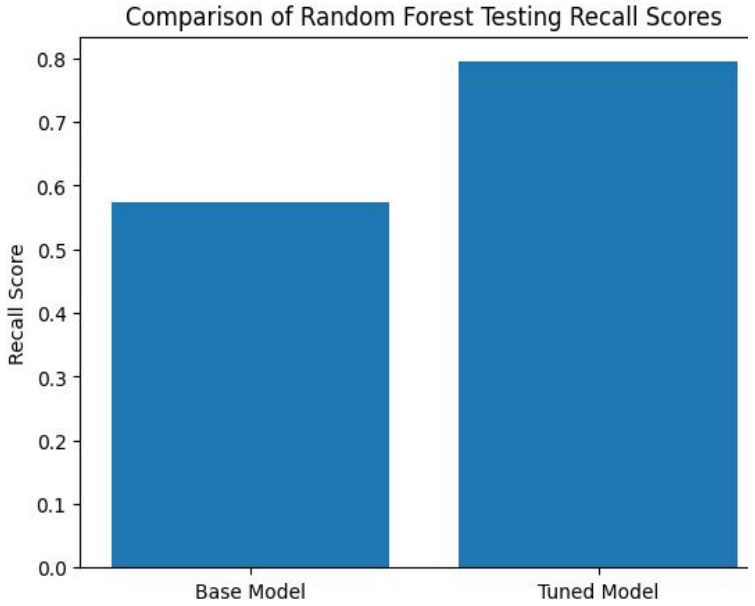


Result:

42% improvement in testing recall score by using the most optimal hyperparameters. The high performance of this simple model may be due to the relative low-complexity and straightforward nature of the relationships between our input features and *Severity.*

Random Forest Classifier

- **Setup:**
  a. Start with initial hyperparameters
  b. Establish "balanced" weighting to offset class imbalance
- **Evaluation:** base vs. tuned
  a. Training Recall = 64% → a. Training Recall = 82%
  b. Testing Recall = 58% → b. Testing Recall = 79%
- **Tuning:**
  a. Find most optimal hyperparameters across a randomized grid
  b. Employ 5-fold cross-validation for robustness
  c. Maximize the "recall" metric
  d. Increase number of decision trees
  e. No bootstrapping

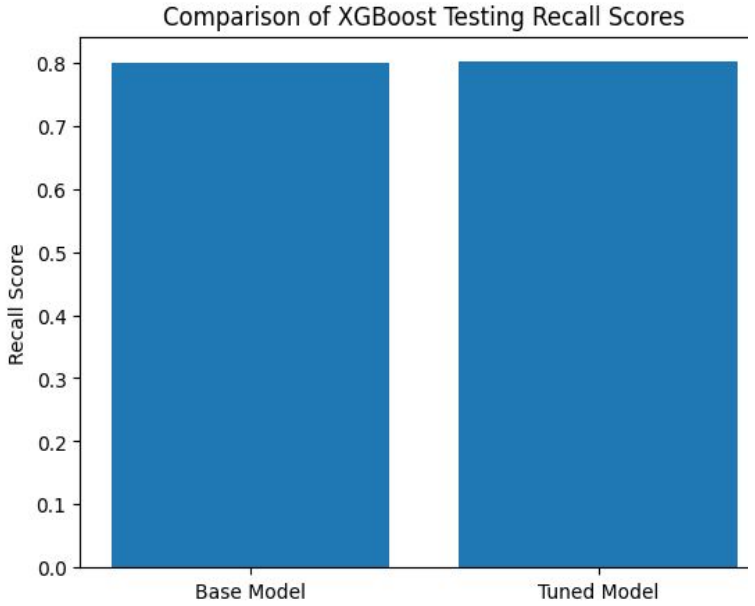


Result:

38% improvement in testing recall score by using the most optimal hyperparameters. Since it's recall score is close to the tuned decision tree, this suggests that the data is not as complex, making the model underutilized and it's advantages not as pronounced.

XGBoost Classifier

- **Setup:**
  a. Start with initial hyperparameters
  b. Establish weighting based off our calculated class imbalance ratio
- **Evaluation:** base vs. tuned
  a. Training Recall = 80.3% → a. Training Recall = 80.7%
  b. Testing Recall = 79.9% → b. Testing Recall = 80.1%
- **Tuning:**
  a. Find most optimal hyperparameters across a randomized grid
  b. Employ 5-fold cross-validation for robustness
  c. Maximize the "recall" metric
  d. Increase number of decision trees
  e. No bootstrapping



Result:

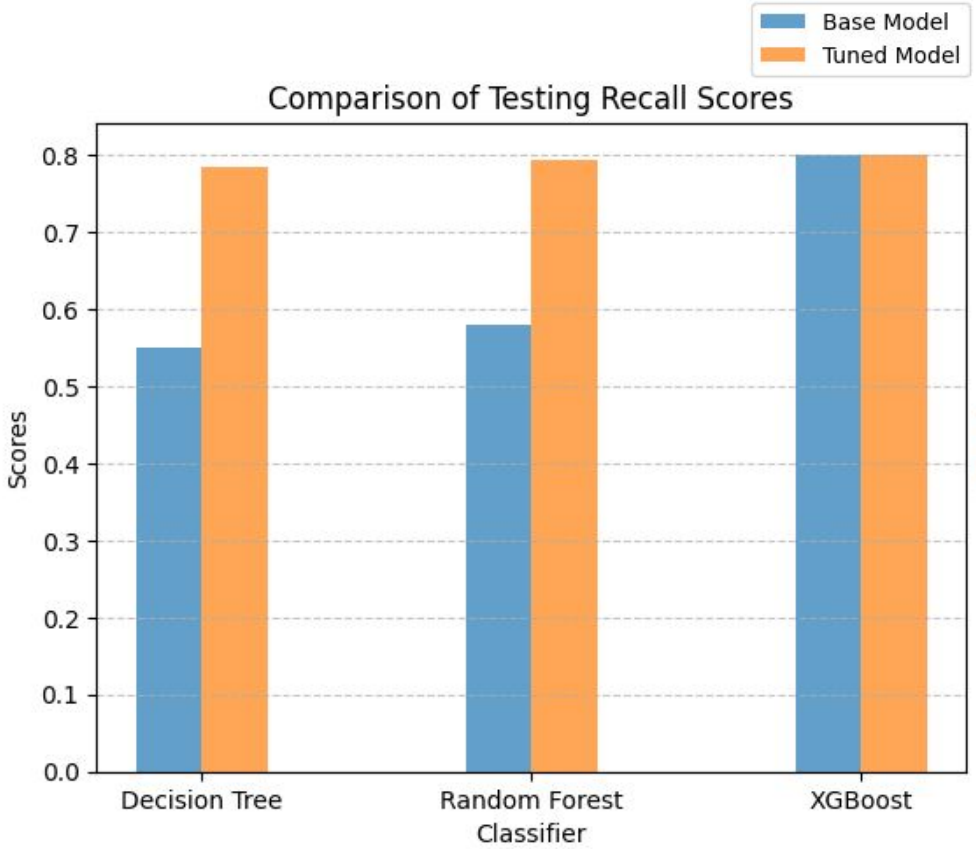
Only a **0.2%** improvement in testing recall score after tuning, and untuned performance is very similar to the previous two tuned models. Therefore, tuning may not be necessary in practical scenarios.

Model Comparison

Best-Performing Model: Tuned XGBoost **(80.1%)**

Most-Improved After Tuning: Decision Tree **(42%)**

Discussion: The robustness of the XGBoost Classifier is most likely due to its iterative boosting approach which is inherently effective at improving recall for imbalanced data. Given the simplistic nature of our data and how well the tuned decision tree generalized to the testing data, we may not need more complex ensemble methods to predict *Severity*. Alternatively, using XGBoost without tuning could also be sufficient to predict *Severity*, as tuning this model led to a negligible performance increase.



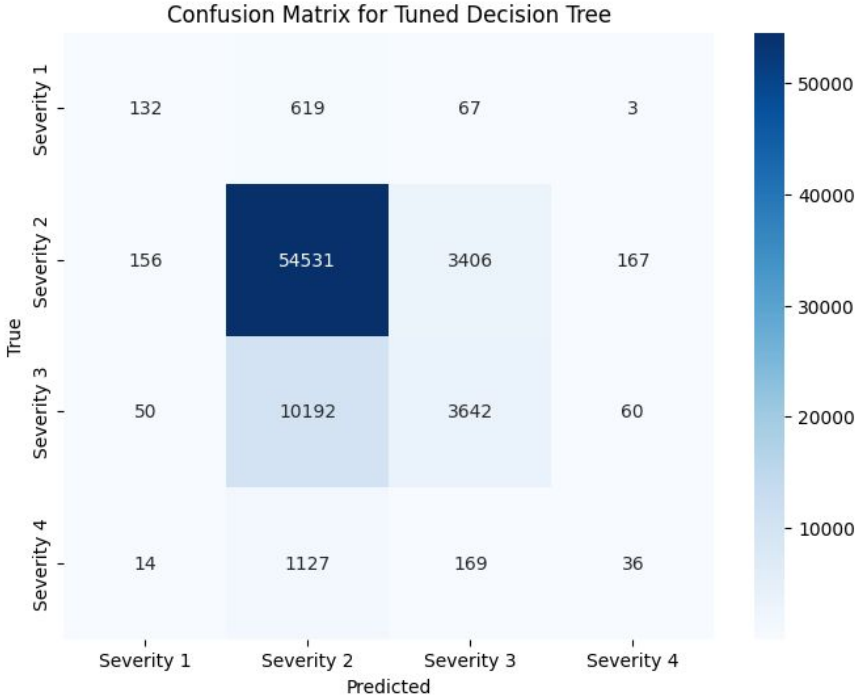
Final Takeaway: *Trading Power for Ease of Use*

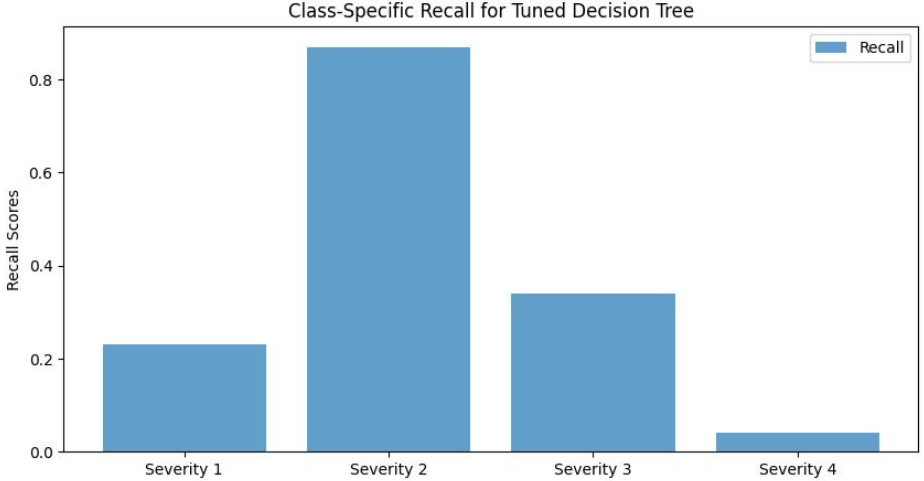
In order to best balance performance with efficiency, we recommend using either the tuned decision tree or the untuned XGBoost in more practical scenarios. This ensures high predictive power while minimizing unnecessary compute time and resources that come with tuning complex ensemble methods.

Hypothesis Testing

Penn Engineering

Performance Analysis: *Tuned Decision Tree by Severity Class*





- Strong performance for Severity 2 (majority)
- Poor performance for Severity 4 (minority)

Final Thoughts

Penn Engineering

Primary Insights

EDA

- The prevalence of accidents in average weather conditions (35.1% under fair weather) implies that extreme weather events might not be the primary drivers of accidents.
- Coastal areas see higher accident rates compared to inland regions, suggesting geographical location is a vital factor for safety measures.
- Accidents cluster near traffic signals, crossings, and junctions, emphasizing the role of traffic design in risk mitigation strategies.

Modeling

- Extensive tuning or complex ensemble approaches are not needed for predicting severity in this case.
- This allows for easy deployment of this model on simpler devices to maximize the velocity of predictions.

Hypothesis Testing

- Model excels at identifying the majority class (Severity 2) with high recall, demonstrating its ability to effectively handle the most frequently occurring accidents.
- This ensures that the majority of cases are well-represented and accurately predicted.

Challenges and Future Work

Challenges

- High class imbalance restricts predictive capabilities for more severe accidents due to their being less data for collisions with severity of 3 or 4
- Data contains simple relationships between the various factors, which is great for efficiency and usage of simpler models, but we cannot fully leverage more advanced approaches to maximize performance

Future Work

- Collect more data for higher-severity accidents to offset class imbalance.
- Identify additional factors that could contribute to accidents that were not in the original dataset e.g. traffic congestion, typical work hours, socioeconomic conditions of certain locations, etc. for more complex relationships between our features.