Driving Towards Safety: A Data-Driven Approach to Predicting Accident Severity

Executive Summary

Traffic accidents are a significant cause of economic and human loss globally. This project focuses on predicting the severity of accidents in the U.S. using the publicly available **US Accidents Dataset**. By leveraging historical data and machine learning techniques, this project aims to:

1. Identify the key factors influencing accident severity.

2. Build a robust predictive model to classify accidents into severity levels.

3. Provide actionable insights to mitigate high-severity accidents.

The tuned XGBoost model achieved an accuracy of **76%** and demonstrated the importance of features like traffic signals, weather conditions, and road junctions in predicting accident severity. The findings can aid stakeholders in optimizing road safety measures.

Problem Statement

The goal of this project is to predict the **severity of accidents** using historical data. Accident severity is categorized into four levels:

- **Severity 0:** Minimal impact.
- **Severity 1:** Minor traffic disruption.
- **Severity 2:** Moderate traffic disruption.
- **Severity 3:** Severe traffic impact.

Given the large-scale US Accidents Dataset, the specific objectives are:

1. Develop a machine learning model to classify accidents by severity.

2. Identify critical environmental and infrastructural factors influencing severity.

3. Provide actionable insights for reducing severe accidents.

Dataset Overview

The **US Accidents Dataset** covers accidents across 49 states in the U.S. from February 2016 to March 2023. It includes approximately **500,000 records** with 46 features such as:

- **Environmental Factors:** Weather conditions, visibility, temperature.
- **Infrastructural Factors:** Traffic signals, junctions, crossings.
- **Location Data:** Latitude, longitude, city, state.

Key Features Selected:

For this project, the following features were used:

1. Weather Condition: Categorical (e.g., Clear, Rain, Fog).

2. Temperature (°F): Continuous.

3. Humidity (%): Continuous.

4. Pressure (in): Continuous.

5. Visibility (mi): Continuous.

6. Junction: Binary (presence/absence).

7. Traffic Signal: Binary (presence/absence).

8. Crossing: Binary (presence/absence).

Preprocessing Steps:

1. Handled missing values by imputation or dropping columns with excessive nulls.

2. Encoded categorical features using **Label Encoding**.

3. Balanced the dataset using **SMOTE (Synthetic Minority Oversampling Technique)** to handle class imbalance.

4. Split the data into training (70%) and testing (30%) sets.

Model Development

Initial Model:

A **Random Forest Classifier** was used as the baseline. However, it struggled with imbalanced classes and failed to generalize well for Severity 2 and 3.

Final Model:

A **tuned XGBoost Classifier** was implemented with the following benefits:

1. Handles imbalanced datasets effectively.

2. Provides interpretable feature importance.

Hyperparameter Tuning:

- **Learning Rate:** Adjusted to control the step size (best: 0.2).

- **Max Depth:** Optimized for tree complexity (best: 7).
- **Number of Estimators:** Optimized to 200.

Evaluation Metrics:

The model was evaluated using:

1. **Accuracy:** Overall correctness of predictions.

2. **Precision and Recall:** Class-specific performance.

3. **F1-Score:** Balance between precision and recall.

4. **Confusion Matrix:** Visual representation of misclassifications.

Final tuned model accuracy: **76%**

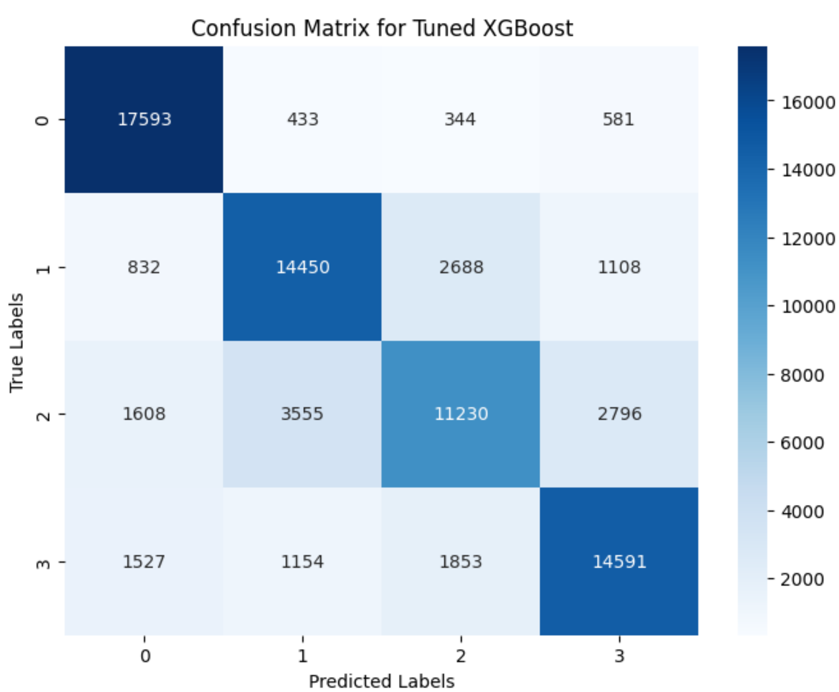
Results

Classification Metrics:

- **Accuracy:** 76%
- **Macro Average F1-Score:** 75%

Confusion Matrix:

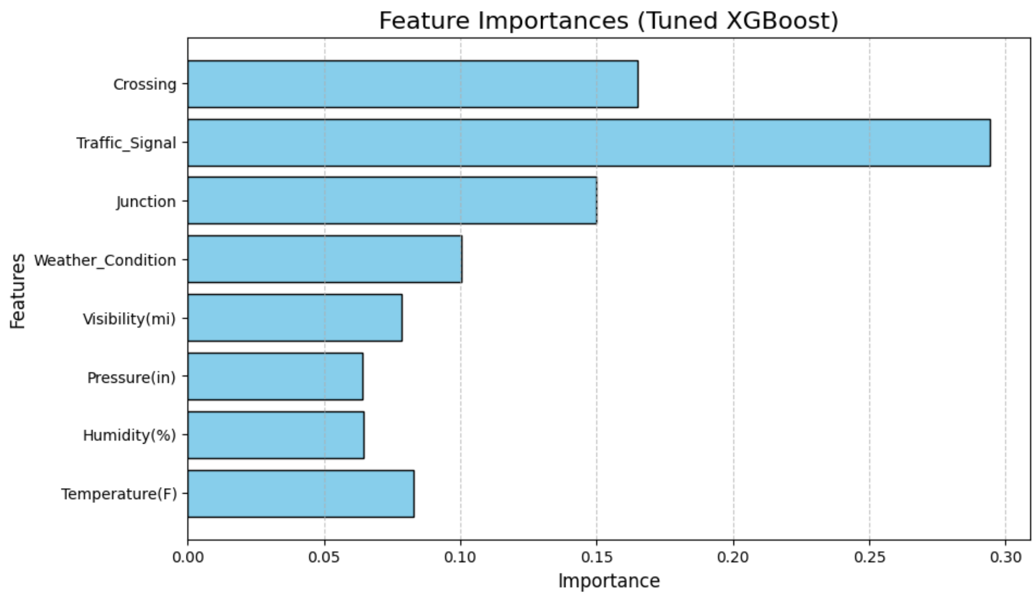
The confusion matrix highlights the distribution of predictions across classes. Most errors occur between adjacent severity levels (e.g., Severity 2 misclassified as Severity 1 or 3).



Feature Importance:

Key factors influencing accident severity include:

1. **Traffic Signal:** Most impactful feature.
2. **Junction:** Significant influence on severity levels.
3. **Weather Condition:** Determines road safety.
4. **Visibility:** Poor visibility leads to severe accidents.



Insights and Recommendations

Key Findings:

1. **Environmental Impact:** Adverse weather (e.g., fog, rain) significantly increases severity.
2. **Road Infrastructure:** Accidents at intersections and traffic signals are more severe.
3. **Visibility:** Poor visibility leads to disproportionately higher severity levels.

Recommendations:

1. Install **adaptive traffic signals** to reduce accidents at intersections.
2. Enhance **road lighting** and reflective signage to mitigate visibility-related accidents.
3. Develop **public awareness campaigns** to improve driver behavior in adverse weather.

Conclusion

This project successfully built a machine learning model to predict accident severity with an accuracy of **76%**. It identified key factors influencing severity, providing actionable insights for road safety improvements.

Future Work:

1. Incorporate real-time weather and traffic data for live prediction.
2. Explore additional features like vehicle type and driver behavior.
3. Expand the dataset to include more recent data for better generalization.

This project demonstrates the power of data science in solving real-world challenges and contributes to creating safer roads.

Dataset Link: https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents/data