PREDICTING ACCIDENT SEVERITY USING US ACCIDENTS DATASET

Dataset source - US Accidents (2016 - 2023) (kaggle.com)

TEAM MEMBERS

1. Mukul Rayana
2. Jayasurya
3. Srutileka S
4. Sai Vikram Karna

Introduction

Traffic accidents are a major public safety concern, and knowing the elements that contribute to accident severity can help mitigate their impact. The purpose of this research is to apply machine learning to anticipate the severity of traffic accidents based on variables such as weather, road conditions, time, and location. The model will forecast accident severity on a 1–4 scale, with one indicating certain minor accidents and the remaining four which represent serious accidents.

This project's dataset is the US incidents Dataset (2016-2023), which includes extensive information about incidents such as weather, location, timing, and road conditions. This rich dataset allows us to investigate which factors contribute the most to accident severity and provides recommendations for enhancing road safety.

Dataset Overview

The **US Accidents Dataset** consists of over 7 million traffic accident records across the United States. It includes 46 features describing each accident, such as:

- **Time**: When the accident occurred (start time, end time).
- **Location**: Latitude, longitude, city, county, state.
- **Weather**: Temperature, visibility, precipitation, weather conditions.
- **Traffic Conditions**: Traffic signals, road junctions, and road crossings.
- **Severity**: The target variable, ranging from 1 to 4, where 1 is the least severe and 4 is the most severe.

The primary objective is to predict the Severity of an accident based on these features.

Project Workflow

This project will follow a structured workflow that includes data preparation, exploratory data analysis (EDA), feature engineering, model building, and model evaluation.

Each stage is described below.

Data Preparation

1. **Handling Missing Data**

- **Why**: Missing data can bias the model and reduce performance. Therefore, we handled missing values by filling or removing them.
- **To-do**:
    - For **numerical features** (e.g., ‘Temperature(F)’, ‘Visibility(mi))’, we filled missing values with the **median** to maintain the data's distribution.
    - For **categorical features** like ‘Weather_Condition’, we filled missing values with the **most frequent category**.
    - Dropped irrelevant or highly missing features, such as ‘Wind_Chill(F)’, which had too many missing values.

2. **Handling Duplicates and Outliers**

- **Why**: Duplicates can lead to biased models, and outliers can cause models to overfit.
- **To-do**:
    - Checked for and removed duplicates based on accident ID, time, and location.
    - Use **Interquartile Range (IQR)** to detect and remove outliers in fields like Distance(mi) and Temperature(F).

3. **Date-Time Conversion**

- **Why**: Time is a crucial factor in understanding accident patterns.
- **To-do**:
    - Convert the ‘Start_Time’ and ‘End_Time’ columns to **datetime format** and extract some additional features such as **Hour of the Day**, **Day of the Week**, and **Month** for further analysis.

Exploratory Data Analysis (EDA)

The purpose of EDA is to explore the dataset and uncover patterns or relationships between the features and the target variable (accident severity).

1. **Distribution of Accident Severity**

- **Why**: To understand how severity levels (1-4) are distributed across the dataset.
- **To-do**:
  - Plot the **distribution of accident severity** to see if the dataset is balanced or imbalanced across severity levels.

2. **Time-Based Analysis**

- **Why**: Time factors, such as rush hours and weekends, can significantly affect accident patterns.
- **To-do**:
  - Create plots to analyze the number of accidents by **Hour of Day**, **Day of Week**, and **Month**.
  - Identified that accidents peak during rush hours and specific seasons.

3. **Weather and Accident Severity**

- **Why**: Weather conditions have a strong influence on traffic safety and accident severity.
- **To-do**:
  - Use **boxplots** to analyze the relationship between accident severity and weather factors like Temperature(F) and Visibility(mi).

Feature Engineering

Feature engineering is crucial for improving the model's predictive performance by creating new, meaningful features from the existing data.

1. Time-Based Features

- **Why**: Time-related factors, like rush hour and weekends, are highly predictive of accidents.
- **To-do**:
  - Create a new feature called **Rush Hour** to flag accidents occurring during peak traffic times (7-9 AM, 4-6 PM).
  - Create another **Weekend** feature to flag accidents that occurred on weekends.

2. Weather-Based Features

- **Why**: Poor weather conditions often result in more severe accidents.
- **To-do**:
  - Created a **Bad Weather** feature that identifies accidents occurring in poor weather conditions, such as rain, snow, or fog.

Model Building

Intending to use machine learning algorithms to predict accident severity based on the engineered features.

1. Base Random Forest

- **Why**: Random Forest is a robust ensemble model that is less prone to overfitting and handles complex datasets well.
- **To-do**:
  - To train a **Random Forest** classifier on the training data and evaluated its performance using **accuracy**, **precision**, **recall**, and **F1-score**.

2. **Tuned Random Forest**

- **Why**: Tuning hyperparameters can significantly improve model performance.
- **To-do**:
  - Learn and make use the **GridSearchCV** to tune hyperparameters like n_estimators, max_depth, and min_samples_split.
  - Retrained the model with the best parameters and evaluated its performance.

3. **XGBoost**

- **Why**: XGBoost is a powerful gradient-boosting algorithm that often outperforms traditional models, especially on structured datasets.
- **To-do** :
  - Understand the working and try out the **XGBoost** model on the data and compare its performance with the Random Forest models.

Model Evaluation and Comparison

Finally, the performance metrics and comparison of multiple models using key metrics:

- **Accuracy**: Overall percentage of correct predictions.
- **Precision**: How many of the predicted positive cases (e.g., high severity) were actually positive.
- **Recall**: How many of the actual positive cases were correctly predicted.
- **F1-score**: The harmonic mean of precision and recall.

Results Comparison

We have a plan of the **Base Random Forest**, **Tuned Random Forest**, and **XGBoost** models using a bar chart to visualize their performance across these metrics.

This plan might improvise depending on the work flow

Conclusion and Insights

By predicting accident severity, we can provide valuable insights into the factors contributing to more severe accidents. Key findings include:

- **Time of Day**: Accidents are more frequent and severe during rush hours and late at night.
- **Weather Conditions**: Poor weather conditions, such as rain or fog, significantly increase accident severity.
- **Seasonal Trends**: Accidents tend to increase in frequency during certain months, possibly due to holidays or weather changes.

Additional Work (project 2)

- **Model Deployment**: The model could be deployed in real-time applications to predict accident severity based on live traffic and weather data.
- **Geospatial Analysis**: Further geospatial analysis could provide more granular insights into accident-prone areas, allowing city planners to take preventive measures.
- **More Advanced Models**: Further improvement could be made by trying other advanced models such as **LightGBM** or using **deep learning** approaches.