Accident Severity
Impact of an Accident on Traffic Congestion
Project Proposal

Pedro Silva, Ramyad Raadi
uc2023235452@student.uc.pt, uc2023205634@student.uc.pt

University of Coimbra, Faculty of Science and Technologies
Bachelor of Data Science and Engineering, ML Course

february 2025

1 Introduction

Reducing traffic accidents is an important public safety challenge. Road accidents not only pose a significant risk to human life but also contribute to severe traffic congestion, leading to increased travel time, fuel consumption, economic losses, etc. Understanding how accidents may impact traffic congestion can help city planners, traffic managers and the police to develop effective strategies to mitigate delays, improve accident response and improve road safety, in general. Therefore, the main goal of this study is to investigate how accidents can affect traffic congestion.

2 Materials

2.1 Data Description

This project considers a countrywide car accident dataset that covers 49 states of the USA. This dataset can be found in https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents/data and is distributed by the Creative Commons Attribution-Noncommercial-ShareAlike license (CC BY-NC-SA 4.0) [1].

[1] https://creativecommons.org/licenses/by-nc-sa/4.0/

The dataset contains 7.7 million accidents and 47 features, recorded by multiple API's that provide streaming traffic incidents data, from February 2016 to March 2023. For the purposes of this project, we will use a sample of this dataset, provided in the same web page, with 500 000 accidents. The features, and it's description, are in the figure below and you can find more information about the dataset here.[2]

#	Attribute	Description	Nullable
1	ID	This is a unique identifier of the accident record.	No
2	Severity	Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay).	No
3	Start_Time	Shows start time of the accident in local time zone.	No
4	End_Time	Shows end time of the accident in local time zone. End time here refers to when the impact of accident on traffic flow was dismissed.	No
5	Start_Lat	Shows latitude in GPS coordinate of the start point.	No
6	Start_Lng	Shows longitude in GPS coordinate of the start point.	No
7	End_Lat	Shows latitude in GPS coordinate of the end point.	Yes
8	End_Lng	Shows longitude in GPS coordinate of the end point.	Yes
9	Distance(mi)	The length of the road extent affected by the accident.	No
10	Description	Shows natural language description of the accident.	No
11	Number	Shows the street number in address field.	Yes
12	Street	Shows the street name in address field.	Yes
13	Side	Shows the relative side of the street (Right/Left) in address field.	Yes
14	City	Shows the city in address field.	Yes
15	County	Shows the county in address field.	Yes
16	State	Shows the state in address field.	Yes
17	Zipcode	Shows the zipcode in address field.	Yes
18	Country	Shows the country in address field.	Yes
19	Timezone	Shows timezone based on the location of the accident (eastern, central, etc.).	Yes
20	Airport_Code	Denotes an airport-based weather station which is the closest one to location of the accident.	Yes
21	Weather_Timestamp	Shows the time-stamp of weather observation record (in local time).	Yes
22	Temperature(F)	Shows the temperature (in Fahrenheit).	Yes

[2] `https://smoosavi.org/datasets/us_accidents`

23	Wind_Chill(F)	Shows the wind chill (in Fahrenheit).	Yes
24	Humidity(%)	Shows the humidity (in percentage).	Yes
25	Pressure(in)	Shows the air pressure (in inches).	Yes
26	Visibility(mi)	Shows visibility (in miles).	Yes
27	Wind_Direction	Shows wind direction.	Yes
28	Wind_Speed(mph)	Shows wind speed (in miles per hour).	Yes
29	Precipitation(in)	Shows precipitation amount in inches, if there is any.	Yes
30	Weather_Condition	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)	Yes
31	Amenity	A POI annotation which indicates presence of amenity in a nearby location.	No
32	Bump	A POI annotation which indicates presence of speed bump or hump in a nearby location.	No
33	Crossing	A POI annotation which indicates presence of crossing in a nearby location.	No
34	Give_Way	A POI annotation which indicates presence of give way in a nearby location.	No
35	Junction	A POI annotation which indicates presence of junction in a nearby location.	No
36	No_Exit	A POI annotation which indicates presence of no exit in a nearby location.	No
37	Railway	A POI annotation which indicates presence of railway in a nearby location.	No
38	Roundabout	A POI annotation which indicates presence of roundabout in a nearby location.	No
39	Station	A POI annotation which indicates presence of station in a nearby location.	No
40	Stop	A POI annotation which indicates presence of stop in a nearby location.	No
41	Traffic_Calming	A POI annotation which indicates presence of traffic calming in a nearby location.	No
42	Traffic_Signal	A POI annotation which indicates presence of traffic signal in a nearby loction.	No
43	Turning_Loop	A POI annotation which indicates presence of turning loop in a nearby location.	No
44	Sunrise_Sunset	Shows the period of day (i.e. day or night) based on sunrise/sunset.	Yes
45	Civil_Twilight	Shows the period of day (i.e. day or night) based on civil twilight.	Yes
46	Nautical_Twilight	Shows the period of day (i.e. day or night) based on nautical twilight.	Yes
47	Astronomical_Twilight	Shows the period of day (i.e. day or night) based on astronomical twilight.	Yes

Figure 1: Feature table and it's description. Source: `https://smoosavi.org/datasets/us_accidents`

3 Challenges

We analyze briefly the context of the features and the dataset and we find some main challenges that we need to face before applying machine learning models:

1. "Severity" is the target feature. The original feature has 4 classes, but we are looking to change it to only 2 (Severe or not severe).

2. Unbalanced dataset. The majority of the accidents are not severe. For this, we are looking to use SMOTE to balance the classes.

3. There are a lot of features - we need to analyze better the complex

features and try to reduce the quantity of them.

4. There are missing data for certain days. We need to understand how we will manage this issue.

4 Acknowledgments

We would like to mention that AI [ChatGPT] was used to correct some syntax errors and to make text more cohesive and coherent.

5 References

1. Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. "A Countrywide Traffic Accident Dataset.", 2019.
Check `https://arxiv.org/abs/1906.05409`.

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.
Check `https://arxiv.org/abs/1909.09638`.