CS5300 Final Project Report - Phase 1

Scott Sanchez

December 2, 2024

1 Introduction

For this project, I decided to focus on predicting the severity of car accidents, an issue of real-world significance with clear applications in the insurance industry. Specifically, I aim to use machine learning to develop a model that can predict the severity of crashes based on a variety of contextual and environmental factors.

1.1 Why This Project?

The end goal of this kind of AI would be to enable dynamic pricing strategies for auto insurance, where companies can assess risk in real time and adjust premiums accordingly. Such an AI, if attached to cars on the road, could discourage behaviors or conditions that produce serious accidents. For example, a driver frequently exposed to high-risk conditions might see higher premiums.

2 Dataset Source

The dataset I chose for this project comes from the Kaggle repository "US Accidents (Sobhan Moosavi)," [1] which contains detailed records of accidents across the United States. I chose this data because it had richer data than any other car-accident database I could find.

For Phase 1, the focus was on ensuring that the data was clean, well-structured, and ready for use in a machine learning pipeline.

The dataset contains records of car crashes with information about weather, road conditions, and other factors. For practical reasons, I sampled 5,000 rows to work with a manageable subset. The target variable here is **Severity**. Because, given some set of real-world conditions, we want to predict how severe an accident is likely to be.

The data's "Severity" column classifies crashes into four levels of severity. This makes the task a multi-class classification problem, with the model predicting one of four severity levels based on input features.

The dataset met all the criteria for this project: it contains over 1,000 rows, has more than three features, and is structured as tabular data rather than time-series or textual data. This ensures that the data is suitable for training a neural network.

3 Data Cleaning & Preparation

Preparing the dataset for machine learning involved several steps. I started by removing irrelevant and redundant columns. For instance, the ID column, while unique to each crash, didn't provide any meaningful information for modeling. Similarly, textual columns like Description and Street were excluded, for two reasons:

1. The project description specified that we should avoid word-based data, as this isn't meant to be an NLP project, and
2. The location data is already given by latitude and logitude.

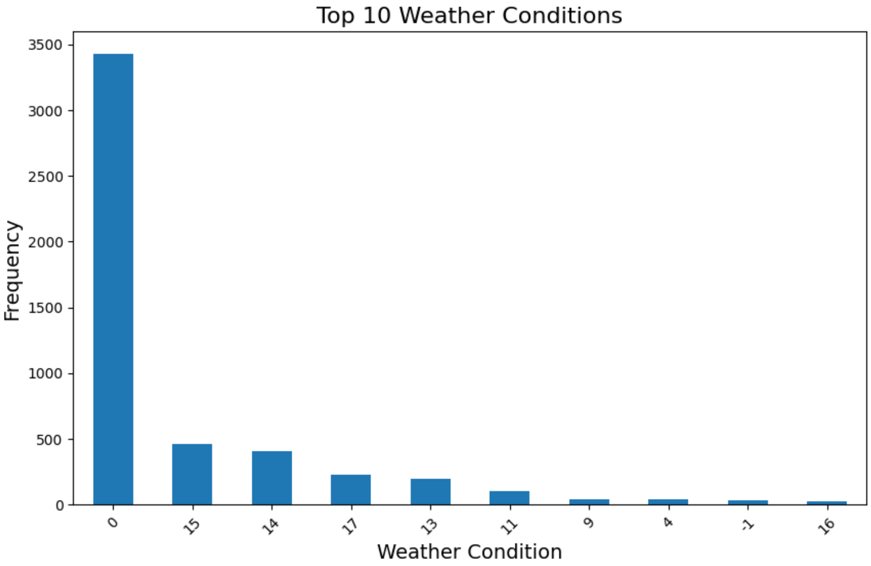

Figure 1: Distribution of weather, showing the skew towards accidents occurring in clear weather.

Several features had missing data, such as ‘Wind_Chill(F)‘ and ‘Precipitation(in)‘. Missing values were handled by substituting an out-of-range placeholder (‘-9999‘) and adding missingness indicator columns to retain information about their absence. While this approach preserves data integrity, the prevalence of missing values could still introduce noise into the model.

For categorical features, I transformed ‘Wind_Direction‘ into numerical values by mapping directions like N, NE, and S to angles. I also converted ‘Weather_Condition‘ into integer codes to make it usable for machine learning.

I normalized all numerical columns. This step was critical because features like Temperature(F) and Visibility(mi) had vastly different ranges, which could skew the model's training process. Finally, I converted all timestamp columns, such as ‘Start_Time‘ and ‘End_Time‘, into Unix epoch format to simplify handling time-based data.

4 Data Analysis

4.1 Target Variable Distribution: Severity

The target variable, ‘Severity‘, is heavily imbalanced

- Severity 2: 58.76
- Severity 3: 41.06
- Severity 4: 0.10
- Severity 1: 0.08

I believe that the imbalance stems from the fact that lower severity crashes (Severity 1) are significantly underreported, because drivers perceive them as negligible. Conversely, severe accidents (Severity 4) are rare, because catastrophic crashes are uncommon.

4.2 Weather Condition

Analysis of the ‘Weather_Condition‘ variable revealed that by far the most frequent condition is clear weather. This finding might initially seem counterintuitive but I suspect the following: drivers may be more cautious in adverse weather, while clear conditions might lead to overconfidence and riskier behavior.

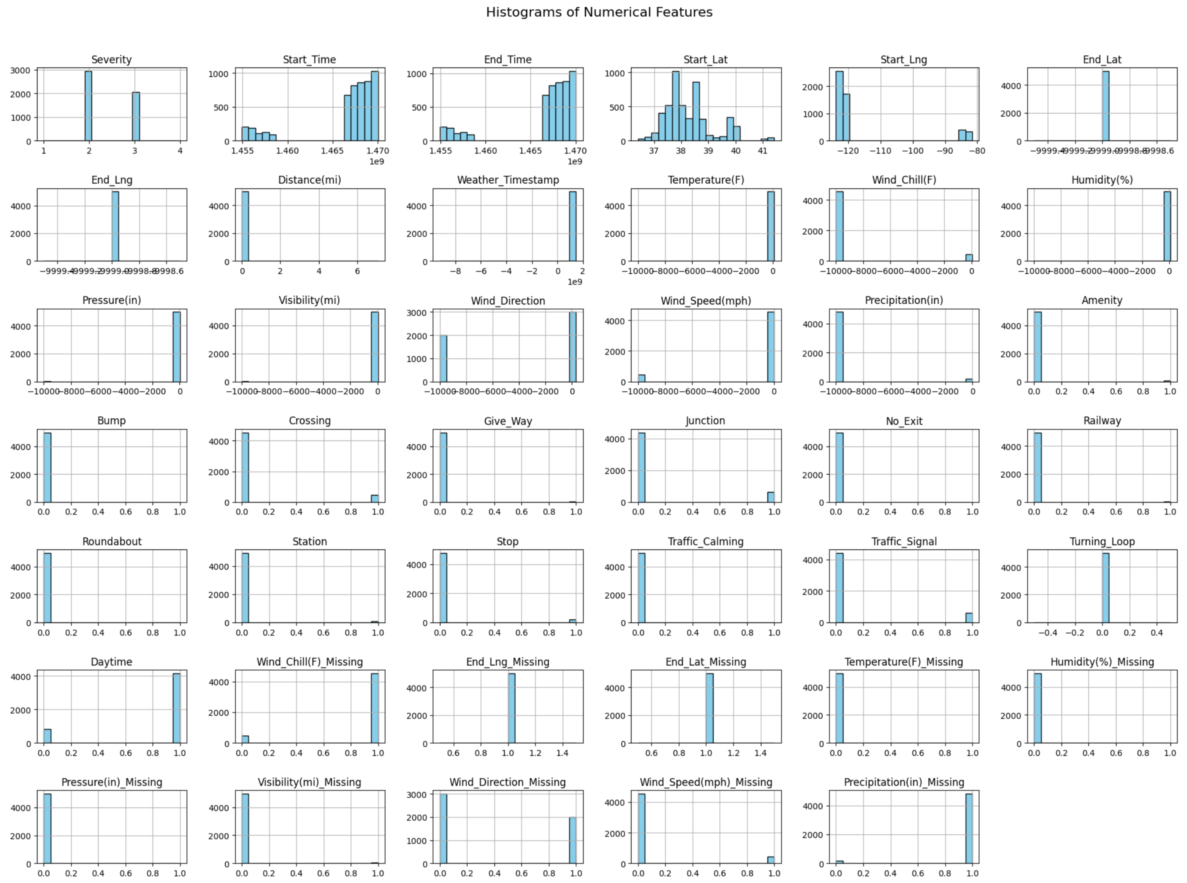

Figure 2: A chart of all columns, after cleaning, graphed as histograms.

4.3 Geospatial Features

Features such as ‘Start_Lat‘, ‘Start_Lng‘, and ‘End_Lng‘ show clustering around certain geographic regions, suggesting the data may be biased toward specific locations. This indicates that people live in cities.

4.4 Environmental Conditions

Features like ‘Visibility(mi)‘ are skewed toward high values, indicating most accidents occur under clear visibility ‘Wind Speed (mph)‘ and ‘Precipitation(in)‘ have distributions clustered at low values, with extreme outliers - again, likely due to people feeling over-confident in fair conditions. -

4.5 Binary Features

Features such as ‘Amenity‘, ‘Crossing‘, and ‘Traffic_Signal‘ are highly imbalanced, with the vast majority being ‘0‘. This suggests that most accidents do not involve these specific conditions, though they could still be significant in certain contexts.

References

[1] Sobhan Moosavi. Us accidents dataset. https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents, 2023. Accessed: 2024-11-19.