Quantifying Risk of Collisions in Autonomous Vehicles via Logistic Regression

Ariyan Sajid

October 20th 2025

Abstract

0.1. Objective

The goal of this project is to create a statistical model that evaluates the driving behavior of autonomous vehicles (AVs). More importantly, we wish to identify factors that most strongly predict collision events. Using real-world trajectory data, we will quantify how speed, time-to-collision (TTC), road curvature, and environmental context affect crash probability. The study applies principles of statistical inference, regression, and extreme-value analysis to evaluate AV safety performance.

0.2. Methodology

The **Open Motion** Dataset from Waymo provides 100K+ real-world driving scenes with rich trajectory information for multiple agents. We can examine how AV reacts to moving pedestrains, cyclists, and fellow vehicles. Each scenario includes the AV's position, velocity, acceleration, lane offset, and a binary collision flag (1 = crash, 0 = no crash). The dataset is larger than 1TB, so we will process it using Princeton University's Adroit cluster. Python and TensorFlow can be used to extract key motion features. After this processing, we can import the files into RStudio for R-only statistical analysis.

0.3. Statistical Analysis

1. **Descriptive Statistics**: Summarize braking, acceleration, and TTC distributions across different road environments (urban vs. highway).

2. **Distribution Fitting**: Fit various distributions (e.g., normal, lognormal, extreme value) to braking distance and TTC data.

3. **Hypothesis Testing**: Test for mean differences in driving metrics under different environments.

4. **Logistic Regression**: Develop a model that classifies whether a vehicle will experience an accident. Coefficients from the model can reveal which variables significantly increase crash risk (e.g, weather, time of day, speed, road type, etc.).

5. **Extreme Value Analysis**: Estimate the probability of near-miss (very low TTC) events.

0.4. Expected Outcome

Through our logistic regression model, we aim to statistically characterize AV safety behavior and quantify how driving and roadway conditions influence the likelihood of collision. The results will highlight critical predictors of AV crashes and demonstrate how advanced statistical tools can be applied to large-scale mobility datasets. With these insights, AV manufacturers can pinpoint areas of improvement to improve the safety of AVs.

Contents