SWITRS California Collision Dataset: Analysis of Collision Cause and Severity, and Potential Effects of AI Collision Avoidance Systems

Group 5

- Nolan Ollada
- Nathan Palmer
- Micah Simmerman

Our Github Link

https://github.com/Micah614/Group5_Data_Mining_Project/tree/main

Description

Through analysis of a large database of CA traffic collision records, we hope to discover relationships between collision-contributive factors (i.e., “lead measures”) and lag measures of interest, including: collision severity, personal injuries, fatalities, regional collision frequency, etc. Our goal is to cluster the data points into classes based on similar/shared attributes with the intent of comparing these findings with modern research on AI-based collision detection and/or self-driving vehicle systems to gain a better understanding of the potential advantages of AI in certain sections of the driving population.

Prior Work

- Analysis of traffic injuries and fatalities in the U.S. since 1966
  https://www.kaggle.com/code/silversurf/us-traffic-getting-to-the-bottom-of-it
- Dataset of Tesla deaths including autopilot usage since 2013
  https://www.kaggle.com/datasets/thedevastator/tesla-accident-fatalities-analysis-and-statistic?select=Tesla+Deaths+-+Sudden+Acceleration.csv

SWITRS California Collision Database (Kaggle)

- SWITRS California Collision Database
  - Gathered from California Highway Patrol (CHP) using the Statewide Integrated Traffic Records System (SWITRS). Records from January 1 2001 to mid-December 2020 compiled on Kaggle by user Alex Gude
  - https://www.kaggle.com/datasets/alexgude/california-traffic-collision-data-from-switrs
  - SWITRS is a large SQLite database file containing 3 main tables with dozens of attribute columns, and almost 10 million data points.

Each team member has a copy of the SQLite database downloaded

Proposed work

Data Cleaning:

- Explore data attributes and remove “unnecessary” columns using Principal Component Analysis (PCA), correlation analysis, or similar to determine the most predictive factors of collisions in the SWITRS database.
- Create routines to clean, smooth, aggregate, and normalize the attributes of interest. Examples include transforming date stamps for granular aggregation and smoothing a to reduce noise or compensate for missing record entries.
- Build and store cleaned SWITRS data in exploratory data cubes (an option).

Areas of research:

- Create a ranked list of collision-contributing factors based on support from the parent data set.
- Uncover unexpected collision classes by mining the database for frequent patterns.
- Measure the impact socioeconomic factors (e.g., “Median county income vs collision count and severity”, “Accident severity vs sex and age”,) on collision probability, severity, etc.
- Produce a model to predict collision severity, given the primary cause(s) of the collision
- Determine what, if any relationship exists between local (i.e., county) legislation and specific measures of interest, such as the frequency or severity of collisions in that jurisdiction.
- Determine which collision classes contain the greatest cross-section of AI preventable collision causes.
- Apply a theoretical success rate based on existing AI literature and estimate the potential for safety improvement in these targeted classes.

Evaluation:

"The truth will set you free. But not until it is finished with you."
-David Foster Wallace

Evaluation Tools *(Evaluating the quality of our work):*

- Verification of code correctness using test input, and/or creating a test harness.
- Using pair programming and peer reviews to detect discrepancies in algorithms, strategies, and code.

Desired Project Outcomes *(Evaluating the quality of our efforts):*

- A thorough PCA and/or correlation analysis will be vital for uncovering the most predictive attributes in the data set. As a result, this section should be organized, comprehensive, well-tested, and well-documented.
- Data cleaning, to the extent possible, should be performed at the beginning of the analysis and should ideally occur within a single, designated section of the codebase for the sake of correctness testing and/or troubleshooting.
- Data cubes may present an efficient way to aggregate and store warehouse data. If data cubes are used, they should be verified, well-organized, and documented.
- Analysis with respect to collision factors, demographic indicators, temporal trends, and geospatial proximity should be attempted, using whatever method appears most suitable unless the analysis is deemed unnecessary in light of a recent finding (for example)..
- Finally, the project should produce meaningful and interesting knowledge about 1.) categories of collisions clusters, and 2.) the likelihood of improving collision estimates by one or more lag measures of interest (i.e. fatalities, collision frequency, collision severity, etc.) by eliminating lead measure of interest (e.g. "using a cell phone while driving", "moderating speed under adverse weather conditions", "following too closely", "distractions caused by passengers", etc.)

Sources Cited

- Dataset Source: https://www.kaggle.com/datasets/alexgude/california-traffic-collision-data-from-switrs
- Data Dictionary: https://tims.berkeley.edu/help/SWITRS.php