

Group 5 Project Proposal

“Separating Driver Classes to Estimate the Safety Benefits of AI-Corrected and Autonomous Vehicles”

James “Micah”
Simmerman

Computer Science
Post-Baccalaureate
Program (CSPB),
University of Colorado at
Boulder,
jasi9001@colorado.edu

Nathan Palmer

Computer Science
Post-Baccalaureate
Program (CSPB),
University of Colorado at
Boulder,
napa8745@colorado.edu

Nolan Ollada

Computer Science
Post-Baccalaureate
Program (CSPB),
University of Colorado at
Boulder,
nool0624@colorado.edu

1 Problem Statement and Motivation

This project will investigate data mining techniques for efficient isolation of principle classes hidden inside massive datasets (i.e., those containing millions of data points or more). This project will identify a series of robust and sequential data mining techniques that can be used to extract, clean, normalize, and prepare the data for causal analysis. Our investigation will use Principal Component Analysis (PCA) and apriori-based Data Cube techniques to observe temporal and regional trends in automotive safety. We will use the insight derived from these Data Cubes to identify and predict the principal driver classes composing the SWITRS California Collision Data Set using Decision Tree

Classification (DTC), Frequent Pattern Mining (FPM), and Bayesian Belief Networks (BBN). The project will conclude with a k-fold cross validation to estimate the accuracy and usefulness of each model.

Artificial intelligent (AI) driving systems are a relatively new field of automotive research, based on extending the “awareness” of automotive computing systems with integrated onboard sensors (e.g., radar, lidar, infrared, etc.). The information relayed from these devices alert the AI driving system to the present driving conditions. This data, collected by the automotive control unit (ACU), enables the system to make split-second driving decisions to preserve the life and safety of pedestrians and automotive passengers. The AI-integrated automotive

market is positioned for explosive capital growth, with the current market value estimated at \$2.2 billion in 2022, and expected to climb to a whopping \$7.7 billion by 2027, according to an analytical report from marketsandmarkets.com[9]. Companies such as Argo AI, Tractable, BMW, and Intel are emerging as key players in a private automotive sub-sector market that is still anyone's game.

While the prospect of owning (and selling) AI-driven automobiles is an alluring prospect, the subject of AI vehicles has raised numerous ethical concerns about the implications of autonomous vehicles on traffic and pedestrian safety laws. In particular, the question of "*who should be held responsible in the event of an accident*" -[nasscom community](https://nasscom.com)[8], has come up again and again. Hence, there is still much to understand about where these AI technologies can be applied to the greatest benefit of the public and the passengers they transport.

2 Literature Survey

Our investigation into prior work done in this field yielded a few interesting analyses that we plan to reference and use to help guide some of our work. One resource we found is called "US Traffic - Getting to the Bottom of it", by Kaggle user "silversurf" (kaggle.com/silversurf)[6]. This analysis uses a data set called US Traffic and Fatality Records, containing data from Fatal car crashes in the U.S. from 2015-2016. The analysis is performed in a python notebook, and analyzes several data attributes collected from fatal car accidents. A set of findings are

presented which show which attributes are most likely to contribute to fatal car accidents. The attributes consist of things like weather conditions, lighting, types of roadways and intersections, vehicle types, and impairment level, among many more attributes.

Another resource we found is called "Tesla Deaths" ([Tesla Deaths data set](https://tesladeaths.com))[7]. This resource is a dataset consisting of accidents involving Tesla vehicles that include the death of a driver, occupant, cyclist, or pedestrian, as well as whether Tesla Autopilot was in use at the time of the accident. We plan to reference this dataset during our analysis on the potential advantages or disadvantages of Artificial intelligent (AI) driving systems.

3 Proposed Work

The data preprocessing section of this project will focus heavily on dimensionality reduction, as well as normalization and abstraction of numerical attributes. Given the extent of the dataset in question, some form of numerosity reduction could also be warranted e.g., if certain attribute outcomes can safely be regarded as uninteresting. Definitions for the attributes in the SWITRS database are available from online resources ([Attribute Definitions](https://www.switrs.com/attribute-definitions))[5] that should resolve any potential issues presented by the *entity identification problem*.

Our team currently anticipates using the following data preparation techniques:

- **Max-min normalization** can be employed in circumstances where

the relative weight of an attribute's values must be standardized to fit under a multi-attribute dependent model.

- ***Percentile-based normalization***, whenever outliers are of particular interest to the model.
- ***Categorical abstractions of numeric attributes*** to create abstract classes representing numerical attribute ranges (e.g., age, income, etc.) that are more meaningful when represented as a range of values than as specific values themselves. Creating abstract attributes in this manner also enhances the support needed for the (e.g.,) apriori-based data pruning techniques that will be used in later sections of the analysis.
- ***Bin-based averaging and smoothing*** is useful whenever noisy and/or missing data are creating gaps in an otherwise consistent model.
- ***Principal Component Analysis (or PCA)*** will be performed on the preprocessed data to identify the leading factors of collisions across the California SWITRS dataset.

4 Data set

Our project will analyze data from California Highway Patrol (CHP) using the Statewide Integrated Traffic Records System (SWITRS). The database we will use consists of almost 10 million data points, which represent every traffic collision in the state of California from January 1st 2001 to

mid-December 2020. This database is compiled by Kaggle user Alex Gude (kaggle.com/alexgude)[2] through several requests made to CHP for data over the past 20 years. The data come in the form of an SQLite file, and categorizes the data through dozens of attributes. These attributes consist of factors such as crash location, crash severity, weather and lighting conditions, road and intersection types, vehicle types, and victim information. ([Link to data set](#))[2]

5 Evaluation Methods

The following sections present the proposed evaluation methods for this project.

Using Data Cube Technologies to Assess Regional and Temporal Traffic Trends

We will use Data Cube technologies to refine the analysis by locating interesting sections of time-scaled and regionally-separated data. This step will assist us in determining what abstractions of time and/or regionality may reveal the most sensitive patterns and trends within the dataset. Derived attributes may also be defined at this stage.

Decision Tree Classifiers for Class Identification

Our group plans to build a decision tree classifier based on the principal components uncovered during PCA, and the abstractions discovered in our data cubes. Multiple python libraries (such as those available with Scikit) are also available to assist with the generation of DTCs.

Classification and Rule Generation using Bayesian Belief Networks

Traffic accidents are frequently described as a cross-section of probabilistic events. Bayes Theorem can therefore be used to assess the probability of a traffic collision outcome, based on prior knowledge of conditions that might be related to the event. Given that there are over 100 attributes in the central table of the SWITRS database (which resembles a star-schema), this dataset offers a wide range of probabilistic outcomes that can be evaluated under the Bayesian Belief Network model to derive conclusions about hidden driver classes.

The strengths of BBNs include their ability to incorporate incremental data without the need for recomputation, as well as the relative simplicity of BBN computation compared with other methods.

The relative computational ease of the BBN technique should enable an extensive analysis of the attribute combinations beyond that of what another model might be capable of. BBN can therefore be seen as an exploratory tool in this project. BBN results are also relatively intuitive and easy to verify by other computational means.

BBNs also provide ideal metrics for evaluating model confidence, accuracy, and precision; as well as for determining the relative support of pre-abstracted classes at different granularity levels. The varying support of the subgroups uncovered during the creation of the Bayesian Belief Network

may assist in the designation of newly discovered or newly-derived classes.

Isolation of Rules Based on Frequent Pattern Mining

Collisions may also be described in terms of frequent patterns. For example, many fatal collisions involving teenagers can be said to contain the frequent pattern: {driving_late, no_seatbelt, alcohol_impaired}. Among the FP classification techniques currently available, perhaps none is more effective than the Frequency Pattern Tree (or “FP Tree”, for short).

The FP-Tree algorithm works by reorganizing the dataset (or sample partition) into a compact tree structure, which is subsequently mined in search of frequent patterns based on available support and user provided threshold value. Python documentation sources and numerous online tutorials exist that can assist with these efforts.

Cross-Method Accuracy Summary and Method Comparisons

After the investigation is complete, we will conclude our analysis with a cross-comparison of the methods employed in each section. We will use the training accuracy estimates from each section to derive conclusions about the driver classes presented within the SWITRS dataset. We will also use this information as a basis for discussion about which AI features would have the most profound safety impact for drivers of each category.

6 Tools

In our upcoming project, we have chosen Python as our primary programming tool. This selection is based on Python's extensive functionality and flexibility.

Key among Python's offerings are several libraries that we'll be utilizing extensively. NumPy, a high-performance library for mathematical computations, will allow us to efficiently handle the numerical aspects of our work, especially where large datasets are involved.

On the other hand, Pandas, a sophisticated data manipulation library, will be instrumental in managing, analyzing, and transforming our data, ensuring that we can handle complex operations with ease and accuracy.

In addition, we will be employing Matplotlib, a dynamic plotting library. This tool will enable us to create comprehensive

two-dimensional graphics, enhancing our ability to visualize and interpret complex data, trends, and insights.

Another invaluable tool in our Python toolbox will be scikits-learn. This library is known for providing classic machine learning algorithms that offer simple and efficient solutions to learning problems. With scikits-learn, we can build robust machine learning models, enabling us to unearth the hidden patterns in our data and make more precise predictions. Thus, combined with the computational capabilities of NumPy, the data manipulation strength of Pandas, and the visualization prowess of Matplotlib, scikits-learn will help us to achieve a more comprehensive and effective analysis of our project.

7 Milestones

We plan to make weekly progress on our project in order to meet the deadlines specific for each part of the project

1. Data cleaning and preprocessing completed by Monday, July 24th 2023.
2. Initial Exploratory Data Analysis (EDA) and Feature Selection by July 24th 2023
3. Part 3: Project Progress Report completed by Monday, July 24th 2023
4. Exploratory Data Analysis and Feature Selection completed by July 31st 2023

5. Initial development of accident prediction models, model selection and training by July 31st 2023
6. Accident prediction model completed by August 7th 2023
7. Part 4: Project Final Report completed by August 14th 2023

7.1 Milestones Completed 7/24/2023

Data preprocessing:

1. Implemented function to read in and enumerate the tables and attributes in 16GB SWITRS.sqlite file.

Data cleaning:

1. Utilized data dictionaries to translate attributes into human readable format.
2. Identified and removed useless attributes in the data set. Many attributes include data types which are only used by CHP for record keeping purposes and won't be useful in regression models.
3. Identified and removed unint

Principal Component Analysis (PCA)

REFERENCES

- [1] Mubarak Almunairi, Kashif Muneer, and AqeelUrRehman. 2022. Vehicles Auto Collision Detection & Avoidance Protocol. IJCSNS International Journal of Computer Science and Network Security, VOL.22 No.3, March 2022.

http://paper.ijcsns.org/07_book/202203/20220315.pdf

- [2] Alex Gude. 2021. California Traffic Collision Data from SWITRS. Retrieved from <https://www.kaggle.com/datasets/alexgude/california-traffic-collision-data-from-switrs>
- [3] Hovannes Kuhandjian. 2022. AI-based Pedestrian Detection and Avoidance at Night Using an IR Camera, Radar, and a Video Camera. CSU Transportation Consortium. Project 2127. https://scholarworks.sjsu.edu/mti_publications/430/
- [4] Hazem H. Refai and Fadi Basma. 2009. Collision Avoidance System at Intersections FINAL REPORT - FHWA-OK-09-06. Electrical and Computer Engineering Department. <https://www.odot.org/hqdiv/p-r-div/spr-rip/library/reports/fhwa-ok0906.pdf>
- [5] github.com/agude. 2021. SWITRS-to-SQLite. SWITRS data dictionary retrieved from https://github.com/agude/SWITRS-to-SQLite/blob/master/switrs_to_sqlite/value_maps.py.
- [6] kaggle.com/silversurf. 2018. US Traffic - Getting to the Bottom of it. kaggle.com data analysis. Retrieved from <https://www.kaggle.com/code/silversurf/us-traffic-getting-to-the-bottom-of-it/notebook>.
- [7] Tesla Deaths. 2023. Retrieved from <https://www.tesladeaths.com/>.

- [8] Brijesh Saluja. 2023. “What are the Ethical and Safety Concerns with Autonomous Vehicles?”. <https://community.nasscom.in/communities/digital-transformation/what-are-ethical-and-safety-concerns-autonomous-vehicles>.
- [9] August, 2023. “Automatic Artificial Intelligence Market Report”. marketsandmarkets.com; Report Code: SE 5533. <https://www.marketsandmarkets.com/Market-Reports/automotive-artificial-intelligence-market-248804391.html>