Rakesh Gururaj
Sriramm Muthayala Sudhakar

Analysis of Road Accidents Based on Demography and Accident Forecasting

I. PROBLEM STATEMENT:

Road traffic safety is a measure that is often not given the due care that is needed. The knowledge of road safety should be taught starting from school. As per the reports by world health organization, there are almost 1.25 million deaths a year across the globe because of the negligence on road while driving leading to traffic accidents. With such a high fatality rate, accidents are predominant in developing countries than in undeveloped countries. The United Nation and rest of the world is seriously concerned about reducing the traffic accidents and decreasing the fatality rate.

Prevention of accidents could possibly be a complex research problem to address since it involves multiple parameters(internal such as vehicle age etc. and external parameters such as weather and road conditions). There is a huge scope for making meaningful analysis of the crash data and predicting/forecasting the accident based on the demographics such as location, vehicle type, driver age, weather, geo location.

Current advancements in the field of machine learning gives us ways to process multi-dimensional data, build a model that forecasts accidents based on the data provided to the model. The key factor to consider is, how accurate is our model and how much room for error we have in our predictions.

II. METHODOLOGY

We plan to build a model will focus on the following entities

- Investigation on what causes accidents
- Key features that has a correlation with the accident
- How do we forecast the accidents?
- What benefit do we obtain from this research?
- Visualizations derived from data
- Understand pattern in road accidents

Rakesh Gururaj
Sriramm Muthayala Sudhakar

The path that we might follow to reap the benefit out of our model is as follows

- Data processing
    - Cleaning
    - Feature selection
    - Preprocessing
    - Visualizations of processed data (to get meaningful insights)
    - Data ingestion pipeline
- Model Exploration
    - Using various ML techniques (Tentative; subject to change)
        - Random forest classifier
        - Regression
        - K-Nearest Neighbors
        - AdaBoost classifier
        - Time series analysis
    - Training and testing of dataset
        - K-Fold cross validation
    - Comparison with Baseline (need to explore on other projects built over same dataset)
- Model Enhancement
    - Hyperparameter tuning
    - Modification to the dataset
    - Using other classifiers for accuracy improvement
- Testing and evaluation
    - Accuracy
    - Precision
    - Recall
    - ROC AUC score
    - Sensitivity
    - Specificity
- Documentation

III. Dataset

The data is fetched from [1] . However, the dataset is assembled using the data that is published by UK government on accidents across the country annually.

Size: 1.2 GB

Format: CSV (We might fetch and convert it into JSON file for easier processing)

Time period: Approximately between 2004 to 2016

Data available:

- Accident Information – Each accident with a primary key on Accident Index column
- Vehicle Information – Each vehicle involved in accident and their relevant attributes along with driver attributes having a mapping with accident information with Accident Index key

IV. Benefits

1. Reducing the impact of accidents by finding out key feature contributing to the accidents such as Driver's age.
2. Identify areas of concern for accidents
3. Might help insurance companies in deciding the price for the driver / for region as such

V. Team Members

1. Rakesh Gururaj, 007500597
2. Sriramm Muthayala Sudhakar,

References

[1] Thanasis. “UK Road Safety: Traffic Accidents and Vehicles” Kaggle.com. https://www.kaggle.com/tsiaras/uk-road-safety-accidents-and-vehicles (accessed Sep. 26,2020).

[2] Global Status Report on Road Safety 2013. 2013, WHO, Geneva