CS 68 Machine Learning for Business with Python

Stanford Continuing Studies
Lecturer: Cathal J Flanagan

Prediction of Frequency and Severity of Car Accidents in California

By Denis Azarov (denis.azarov@icloud.com)

02/15/2023

ABSTRACT

Using datasets of car accidents and weather in California, prediction factors for car crashes were analyzed. The linear regression predicted the number of accidents per day in Los Angeles to 1.6% using weather data. Random Forest and XGB models yielded similar results of about 90% precision/recall in predicting the factors that determined the severity of car accidents in California.

GOALS

Developing predictive models to identify factors which have an impact on the frequency and severity of car collisions. These models can (i) provide information that can save human lives, (ii) lower insurance costs and (iii) enhance the safety of autonomous vehicles when incorporated into their systems.

DATA SOURCES

Countrywide car accident dataset was sourced via Kaggle. (To access the data a registration on Kaggle is required). The accident data is collected from February 2016 to Dec 2021, using multiple APIs that provide streaming traffic incident (or event) data. For the prediction models and regression, a subset of data for California and Los Angeles were used to avoid computational challenges. The dataset was also reduced to the timeframe from 2016 to 2019 to exclude possible COVID-19 effects.

To avoid endogeneity of the sample, the accident data was enriched with weather data from the LA weather station to include the days when no accidents were reported. The data from this weather station includes a rich set of weather parameters. To access the weather data an API request to the NCEI was generated.

KEY TAKEAWAYS

Regression to describe the number of accidents in Los Angeles

- Weather explains the number of car accidents in LA to a limited degree. Fog, precipitation, temperature, and haze were among the parameters that had a statistically significant effect on the frequency of accidents. Interestingly, temperature and fog showed a negative effect meaning a reduction in the number of accidents. Apart from limitations of the data described below, this could be attributed to the fact that less LA drivers are on the roads in foggy conditions. Smoke, haze, and precipitation increase the frequency of car accidents as one would expect.
- Limitations of the analysis - limited sources for weather data (only one weather station was used), limited timeframe, and unknown completeness of the car accidents data set.

ML models to predict factors affecting severity of car accidents in California

- Random Forest and XGBoost provided similar results with the level of precision/recall of ~90%.
- Among other important factors are daytime, weather conditions (such as clouds and wind), city, road attributes (stop, junction, traffic signal) and availability of the daylight.
- It appears logical that location and time are important factors that affect Severity, since the Severity in the dataset used is determined by the impact on traffic (i.e., long delays).