Analysis of Traffic Fatalities

The Problem

Can we target the roads on which fatal accidents are most likely to occur? How often is weather a factor (cloudy, foggy, raining, snowing, icy roads, etc.)? Are they more likely to occur at a particular time of day and/or year or on a particular type of road (for example, interstate or rural)? On the other hand, are human factors such as impairment due to drugs, alcohol or even distraction greater factors than the foregoing environmental factors?

The first question is therefore whether the chosen environmental factors have a greater or lesser impact than human impairment. If human impairment is the greater factor, is it more likely to be due to drugs/alcohol or to some other factor? Finally, regardless of the role of human impairment, when (time of day/season/weather conditions) and where (types of roads and general regions of a community) are traffic fatalities most likely to occur?

The Client

This topic would be of interest to law enforcement agencies, as they can make the best choices in deploying resources in those areas and at those times preemptively (i.e., putting speed traps or DUI checkpoints in those areas and shortly before those times when traffic fatalities are likely to occur in an attempt to encourage safe driving behavior) and to ensure that resources are available to quickly respond at those times. (It would also be of interest to auto insurance companies of individuals that live in those types of areas in setting their rates.) Additionally, if human factors are more to blame, law enforcement may be able to plan the appropriate educational outreach programs to focus more on issues such as driving while impaired or distracted.

The Data

The dataset will be taken from the 2015 Traffic Fatalities provided by NHTSA available here: https://www.kaggle.com/nhtsa/2015-traffic-fatalities. This consists of 17 related csv files which contain in common unique case identifiers. The accompanying NHTSA documentation defines the meanings of the codes and abbreviations used throughout each dataset.

The Approach

1. Data wrangling and cleaning—First I will determine which among the 17 datasets are the most useful. After choosing the most important/useful datasets, each will be cleaned, with some variables converted into categorical factors as needed, and then ultimately united into a single dataset.
2. Exploratory Data Analysis—Here, I will analyze relationships among the variables in the dataset using both statistical functions and graphical analyses.

It should then become apparent whether human or environmental factors play a greater role in traffic fatalities.

3. Machine Learning—After performing an analysis of the key variables, I will then build a model using the appropriate variables in order to predict whether a serious fatality (an accident involving more than one fatality) is likely to occur. Approximately 80% of the data will be used in training the model, while the remaining 20% will be used to test the model's accuracy.
4. Final Report—I will then provide a detailed report describing the procedure I used and the findings I obtained therefrom, including appropriate data visualizations along with a slide deck providing a high level summary of the report.

Deliverables

The deliverables will consist of a detailed report of my procedural approach, my findings, including appropriate data visualizations where needed, a slide deck containing a high level summary of report, and the corresponding Python code used in analyzing the data. All of the foregoing will be submitted and published on GitHub.