**Applied Data Science Capstone Project**

**Background:**

Automobile accidents present significant economic costs and, in some cases, bodily harm or even death. At the same time, automobiles are critical forms of transportation for many, and accidents may appear an unavoidable risk. While complete elimination of these risks is impossible, the purpose of this analysis is to determine the conditions that are most likely to cause accidents so that these risks can be mitigated. Further, some accidents are more severe than others. This analysis will analyse the conditions that contribute to increased accident severity.

**Data:**

The dataset consists of 194,673 observations and includes 37 predictor variables. The dependent variable is a discrete measure of accident severity, ranging from 0 (“unknown”) to 3 (“fatality”). The predictor variables indicate things like address, type of collision, number of persons involved, number of pedestrians involved, number of vehicles involved, whether or not the accident occurred in an intersection, driver inattention, whether the driver was under the influence of drugs or alcohol, weather conditions, road conditions, and light conditions.

**Methodology:**

The approach is to fit a logistic regression to the data to estimate whether or not an accident results in an injury using the available covariates. The data will be split into training (80%) and testing (20%) samples to assess the potential for overfitting. The analysis will uses several specifications to determine the best fit, noting that high correlations between predictive variables creates non-invertability problems with some specifications. The final specification includes as covariates dummy variables indicating whether the accident occurred in an intersection, whether road conditions were poor, whether the accident happened at an angle, head on, during a left turn or other, whether a parked car was involved, whether a pedestrian was involved, whether a cyclist was involved, whether a vehicle was rear-ended, whether it involved a right-hand turn, and whether it involved a sideswipe. The specification also includes ordinal variables indicating the number of vehicles involved in the accident, and the number of people involved.

**Results:**

The model performed reasonably well in predicting whether or not an accident would involve an injury. The model’s Jaccard similarity score is 0.76, and the test accuracy based on the confusion matrix is 74.5%.

Based on the model, the variables with the largest positive contribution to the likelihood of an injury were the number of vehicles, the number of people, the presence of a cyclist, and the presence of a pedestrian.

By examining marginal effects, in other words, examining the effect of each variable while holding the remaining variables at their sample means, the involvement of pedestrians and cyclists are the two most important variables. The involvement of cyclists or pedestrians increase the probability of an injury by 23% and 26%, respectively.

**Discussion:**

The most important conclusion from this analysis is that accidents involving pedestrians and cyclists are significantly more likely to result in injuries than those that do not. Therefore, if the aim is to devise policies that can reduce the numbers of injuries on the roads, policies must be adopted that provide greater protections to cyclists and pedestrians.

**Conclusion:**

This analysis used data on traffic accidents to assess the likelihood of injuries. The analysis included data on weather conditions, whether accidents occurred in an intersection, and the type of accident, but the largest contributors to the likelihood of an injury is the involvement of pedestrians and cyclists. The recommended policy to reduce road-based injuries is therefore to adopt policies that further protect pedestrians and cyclists.