**Factors Influencing Car Accident Severity**

**Introduction:**

Like most things in life, driving a vehicle carries a certain level of risk which depends on a wide range of various factors. In 2018, there were approximately 33,654 fatal car accidents within the US that resulted in the deaths of 36,560 individuals. If there were a reliable way to predict the safety levels of various driving conditions, then it could result in a significant reduction in car accidents and the resulting fatalities. Clearly, this is a pressing issue considering accidents are one of the leading causes of death and injury in the US.

This project aims to determine what the most influential factors are that contribute to the severity of a motor vehicle accident, such as weather conditions, road conditions, light conditions, etc. and then create a model that will be able to predict the severity of an accident based on these factors and determine the level of risk.

**Data:**

The dataset being used for this project is sourced from the SDOT Traffic Management Division, Traffic Records Group based in Seattle and includes all traffic collisions records dated from 2004 to the present. This data includes detailed information about each collision that took place in Seattle that can be useful for this project, such as the weather conditions, road conditions, light conditions, whether a driver was under the influence or inattentive at the time of the collision, whether speeding was a factor of the collision, the type of location/address the collision occurred at (such as an alley, intersection or block), and the severity of the collision (given as a numerical code). In the model being built for this project, the severity code will be the dependent variable that is to be predicted based on the aforementioned collision attributes (which will be the independent variables).

In order to deal with the data in a more consistent and reliable way, data cleaning was a necessary task. The severity code column was adjusted to reflect a binary pattern of categorical variables, 0 being less severe (property damage only) and 1 being more severe (physical injuries occurred). The various independent variables used in this analysis also had to be converted to a more organized system of variables – address types were converted to 0, 1, and 2 for block, intersection, and alley (respectively); inattention index was converted to 0 and 1 for no and yes (respectively); under the influence index was converted to 0 and 1 for no and yes (respectively); weather index was converted to 0, 1, 2, and 3 for clear, cloudy/overcast, fog/windy, and rain/snow conditions (respectively); road conditions index was converted to 0, 1, and 2 for dry, mixed, and wet conditions (respectively); light conditions index was converted to 0, 1, and 2 for light visibility, medium visibility, and dark visibility (respectively); and lastly, speeding index was converted to 0 and 1 for no and yes (respectively). Incomplete and missing data fields were normalized by filling the unknown values within the dataset.

**Methodology:**

After taking into consideration the distribution of the feature set, it was determined that it was not fully balanced and so SMOTE was used in order to balance the data. Exploratory analysis was done and a pie chart was created to compare the distribution of accidents depending on address type. The machine learning models chosen for analysis were decision tree analysis, logistic regression, and support vector machine. Decision tree analysis analyzes the dataset by breaking it down by feature into smaller partitions (or nodes) and develops a decision tree that demonstrates how the model reaches a particular decision. Logistic regression analyzes the data through a statistical model that is predicated on using a logistic function. Lastly, the support vector machine algorithm bases its modeling system on creating a hyperplane in a dimensional field that depends on the number of feature sets. This hyperplane is then able to specifically classify the given data points.

**Results:**

After running the various models and visualizing their classification reports and confusion matrices, it became apparent that they all had similar f1 accuracy scores with the logistic regression and decision tree models performing slightly superiorly.

**Discussion:**

The f1 accuracy scores for each of the models were quite similar to one another – logistic regression had a score of .5955, decision tree had a score of .5947, and the SVM model had a score of .5758. From these results, I would have to say that all models would be relevant and should not be ruled out when it comes to analyzing and evaluating the data.

**Conclusion:**

In conclusion, these models were about equally effective when it came to their overall accuracy. However, future analysis on this project should include different machine learning methods in order to determine if there is an even more effective model type for this kind of data analysis. Additionally, with the bulk of accidents occurring in the “block” neighborhood type, this may be an indicator for what areas to specifically avoid during particular driving conditions.