Data Understanding

Regarding the data, it was extracted from the Seattle City Council database as I mentioned before in a csv format. A close look to at it shows that it needs a thorough data cleaning. In fact, there are many columns I don't need for the machine learning model including empty colunms. I will get rid of missing data as well.

Only 2 severity codes (**1, 2**) are available with severitycode 1 cases been almost 2 times seviritycode 2. This is not surprising as vehicle collision prediction has always suffered from data imbalance issue. And, knowing that when dealing with severe data imbalance, most machine learning algorithms do not perform well. I will downsample severitycode 1 to match the severity code 2 samples size.

Moreover, most of the features are of type object, unfortunately, Sklearn Decision Trees which I am planning to use, do not handle categorical variables. I will then convert these features to numerical values using pandas.get_dummies(). Since I have more than two categories, I may create a new problem in the way. As I keep assigning different integers to different categories, it may create a confusion. This strategy might as well defeat its own purpose. So instead of having one column with n number of categories, I will use n number of columns with only 1s and 0s to represent whether the category occurs or not. To accomplish this task, I will import OneHotEncoder library. Then, I will visualize these categorical variables by using boxplots to determine the relationship between the different attributes and the severity code. Futhermore, I will perform group with multiple variables and look at the correlation of these variables with the severity code wandering about the dependency of the severity code on these variables.

Finally, I will build a supervised machine learning algorithm that will accurately predict road accidents severity. As I mentioned before, road accidents prediction is defined as a classification problem or a regression problem. Most of the studies that performed classification only reported the accuracy metric which is not well suited for problems with data imbalance such as road accident prediction [3]. The studies that performed regression used different definitions for the risk of accidents, which makes comparisons difficult. Therefore, in order to build an accurate model, I will build several prediction models using various machine learning algorithms such as K Nearest Neighbor(KNN), Decision Tree, Support Vector Machine and Logistic Regression.

K-Nearest Neighbor (KNN) which will help me predict the severity code of an outcome by finding the most similar to data point within k distance.

Decision Tree which will give me a layout of all possible outcomes to fully analyse the concequences of a decision. It context, the decision tree observes all possible outcomes of different weather, road and light conditions.

Support Vector Machine which is used for classification, regression and outliers detection**.**

Logistic Regression which is perect for binary data such as the present dataset containing only two severity codes.

References

1. M. M. Chong, A. Abraham, and M. Paprzycki, "Traffic accident analysis using ma-

chine learning paradigms,” Informatica, vol. 29, pp. 89–98, 05 2005.

2. J. Abellan, G. L ´ opez, and J. de O ´ na, “Analysis of traffic accident ˜ severity using decision rules via decision trees,” Expert Systems with Applications, vol. 40, no. 15, pp. 6047 – 6054, 2013.
3. H. He and E. A. Garcia, “Learning from imbalanced data,” IEEE Transactions on Knowledge & Data Engineering (TKDE), no. 9, pp. 1263–1284, 2008.