**Introduction**

In an effort to reduce the frequency of car collisions in a community, an algorithm could be developed to predict the severity of an accident given the current weather, road, visibility conditions and speed.

One way to unveil the truth behind these variables lets create some models to explore the potential Severity of collisions, and may help for future campaigns for a safety drive from local government.

Therefore, this document may be helpful for local community and any stakeholder in the car industry or insurance.

**Data**

The data has being collected by the Seattle Police Department and Accident Traffic Records Department from 2004 to present. Consists of 37 independent variables and 194,673 rows. The dependent variable, “SEVERITYCODE”, contains numbers that correspond to different levels of severity caused by an accident from 0 to 4.

Severity codes are as follows:

0 Little to no Probability (Clear Conditions) 1 Very Low Probability — Chance or Property Damage 2 Low Probability — Chance of Injury 3 Mild Probability — Chance of Serious Injury 4 High Probability — Chance of Fatality Furthermore, because there are null values in some records, the data needs to be preprocessed before any further processing.

And as reminder the independent variables selected for the models are weather, road, visibility conditions and speed.

**Methodology**

Exploratory data analysis was performed on the relevant categorical variables: address type, collision type, person count, pedestrian count, cyclist count, vehicle count, junction type, SDOT type, under the influence, weather, road conditions, light conditions, lane key, crosswalk key, and if a parked car was hit. The amount of categories quite significant and best way to proceed tis to have a common scale and transform them into matrices. Also, there are null values, so those will be eliminated, due the data set size will not suffer a bias effect. Nevertheless, since We are taking into account the speed (try to kill a myth about speed limits are “safer”, down the road); its Null Values will be filled as 1 and if it has a value like “Y” as yes, will be replaced as 2.

**Results & Evaluation**

Unexpectedly, the data shows that most accidents occur during the day with normal drivers and conditions. Most of the cases involved property damage, with no fatal losses.

What is quite interesting , most vehicle accidents occur during the best driving conditions, weather clear, daylight, roads are dry and no speeding(looks like speeding after all does not involve in an accident, so it is “user problem”).

**Discussion**

At the beginning there is categorical data and it has to transformed into a common scale with numerical data type, specifically as float type.

After solving that issue lets not forget the — imbalanced data with the target variable(SEVERITY); so it was resample the observations and end up with 58188 values each code as 1 or 2 according to severity of the damage.

The models to be used are:

K-Nearest Neighbor, Decision Tree, SVM and Logistic Regression.

Evaluation metrics used to test the accuracy of our models were jaccard index, f-1 score and logloss for logistic regression. Choosing different k, max depth and hyperamater C values helped to improve our accuracy to be the best possible.

According to data no serious injuries or fatalities have happened yet, so basically property damage is involved in accidents. For further studies, may be useful to obtain more data about pedestrian right of way, inattentive drivers, to have more knowledge about the problem. Since there are null values, more surveys could help.

It was determined that most accidents occur during normal weather and road conditions. However, further data is needed to analyze this trend. It may be that these types of days constitute the highest number of days in the year. Therefore, further data on the weather needs to be analyzed.

**Conclusion**

According to data, weather conditions have a certain impact as variable to conduct on accident, also if occurs will be more likely to result on property damage (class 1) or injury (class 2) and not fatal losses.

So, as data showed that most vehicle accidents occur during good conditions with somehow property damages what is left is to create campaigns to enhance “awareness” of driving and… DO NOT DRINK & DRIVE!