SEATTLE COLLISION DATA ANALYSIS

1. INTRODUCTION AND BUSINESS PROBLEM

In recent years many municipalities around the world gained access to large sets of traffic / road accident data that is still relatively underutilized. This project showcases how insights gained from road accident data can be used for the benefit of various stakeholders.

The key objective of the report is to predict the severity of the road accident based on the various data points such as weather and road condition, accident coordinates and etc. Analysis is based on the data provided by Seattle municipality and hence would only be relevant to Seattle though the model can still be extended (with relevant modifications) to any other city.

Predictive model can be utilized by individuals who are wondering how severe the road accident can be in a given area with specific road conditions and etc.

Predictions can also be leveraged by municipalities in city planning, e.g. improving road infrastructure in the problematic areas (with high proportion of severe accidents), increasing medical service coverage in such areas and etc.

2. DATA DESCRIPTION

As mentioned previously we will be leveraging data provided by Seattle municipality on road accident / collisions. Data contains a large number of features (37 without the label column). The features contain various info on road and weather conditions, coordinates of the accident, accident code provided by the authorities and etc. The label column is “SEVERITYCODE” which is populated with 1 (meaning property damage only) and 2 (meaning injury). The total number of entries is 194,673

The dataset is mildly imbalanced – ~30% of positive and ~70% of negative results. Hence, we would be required to balance the training sets (would be covered in more details later).

We will use data provided to build a predictive model to predict SEVERITYCODE. We would first clean data and remove redundant features. We would then try out different machine learning methods to build the most accurate / appropriate predictive model.

After first stab at the data cleaning we have decided to remove 19 redundant / non-relevant features. Below you can find reasons for removal:

- Duplication: “SEVERITYCODE.1", “SEVERITYDESK”, "INCDATE"
- Columns with too many missing data entries: "INTKEY", "INATTENTIONIND","EXCEPTRSNCODE","EXCEPTRSNDESC", "SPEEDING", "PEDROWNOTGRNT", “SDOTCOLNUM"
- Columns that have useless info (e.g. ID): "REPORTNO", "INCKEY", "COLDETKEY", "OBJECTID"
- Columns with categorical info that have too many unique values (impractical for one-hot encoding): "ST_COLDESC","SDOT_COLDESC", 'SEGLANEKEY', 'CROSSWALKKEY', "LOCATION",

In the end we are left with the following features: 'X', 'Y', 'STATUS', 'ADDRTYPE', 'COLLISIONTYPE', 'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'INCDTTM', 'JUNCTIONTYPE',

'SDOT_COLCODE', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND', 'ST_COLCODE', 'HITPARKEDCAR'

We have also removed rows with Nan losing 14,606 rows (~7.5%) which shouldn't affect our results much as this hasn't impacted the distribution between positive and negative results.