Predicting collision vehicles in Montreal

Introduction

Business Problem¶

A system to reduce accidents will be favoured by car users, insurance companies, road maintenance, cities/municipalities/governments and, if relevant, habitants in the area. It would reduce the costs caused by accidents, as well as improve the traffic by reducing congestion caused by accidents. Predict the possibility of getting to a car accident and its severity given the current driver and driving conditions in order to reduce damages in a real-life scenario.

Data

Data Source:

- Data is already provided through a CSV file called Data-Collisions.csv which contains 194673 rows.

Feature Selection:

- The next step is to remove irrelevant columns: COLDETKEY, REPORTNO, STATUS, INTKEY, LOCATION, EXCEPTRSNCODE, EXCEPTRSNDESC, JUNCTIONTYPE, SDOT_COLCODE, SDOT_COLDESC, INATTENTIONIND, UNDERINFL, PEDROWNOTGRNT, SDOTCOLNUM, SPEEDING, ST_COLCODE, ST_COLDESC, SEGLANEKEY, CROSSWALKKEY, HITPARKEDCAR, COLLISIONTYPE, ADDRTYPE, SEVERITYDESC,
- Then we have to identify the relevant columns by looking for trends and patterns in this case: WEATHER, ROADCOND, LIGHTCOND, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, INCDATE, INCDTTM.

Data Cleaning:

- We eliminate duplicate columns: SEVERITYCODE.
- Dealing with missing data by removing the rows.
- Balancing in order to dealing with inbalanced data.
- Normalising, balancing or transforming data in order to be able to work with different models in the next step. This could include for example changing strings that describe weather to numeric input.