Capstone Project - Car accident severity

Business Problem

Most practically useful task is to classify locations according to their level of danger. Traffic authorities could use this information to change road signs, road markings and traffic lights. However, the original sample lacks information on traffic: the number of cars per unit of time. Without this information, such an analysis is useless. Since only two cars a day can appear in a very dangerous place. And both have an accident. And in a safe place, thousands of cars can pass a day, and only two of them will have an accident.

Thus, we will deal with a less interesting problem. We will classify the level of danger of an accident according to the available factors: 1) Only property damage 2) An accident with human injuries.

Data

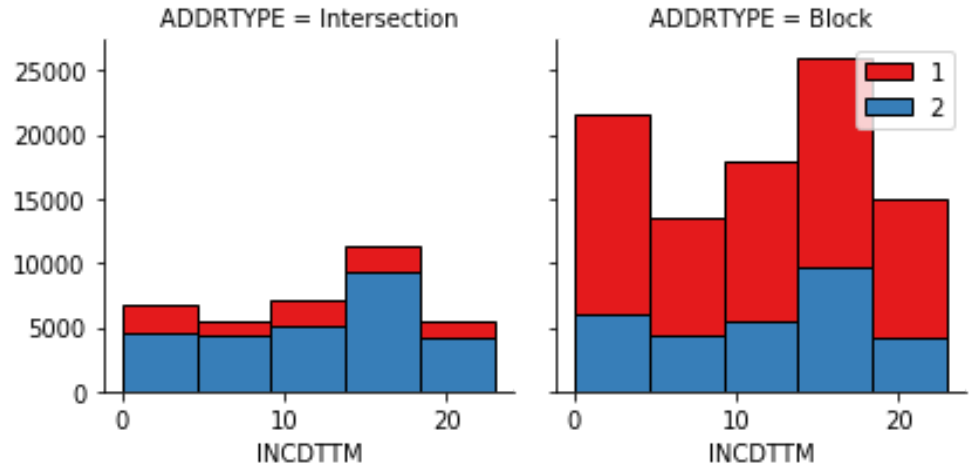
To analyze car accident severity, the following features were selected from the sample:

LOCATION - Location of the collision.
ADDRTYPE - Collision address type.
COLLISIONTYPE - Collision type.
PERSONCOUNT - The total number of people involved in the collision.
PEDCOUNT - The number of pedestrians involved in the collision.
PEDCYLCOUNT - The number of bicycles involved in the collision.
VEHCOUNT - The number of vehicles involved in the collision.
INCDTTM - The date and time of the incident.
JUNCTIONTYPE - Category of junction at which collision took place
INATTENTIONIND - Whether or not collision was due to inattention.
UNDERINFL - Whether or not a driver involved was under the influence of drugs or alcohol.
WEATHER - A description of the weather conditions during the time of the collision.
ROADCOND - The condition of the road during the collision.
LIGHTCOND - The light conditions during the collision.
PEDROWNOTGRNT - Whether or not the pedestrian right of way was not granted.
SPEEDING - Whether or not speeding was a factor in the collision.
HITPARKEDCAR - Whether or not the collision involved hitting a parked car.

The classification will be carried out into two classes: 1) Property Damage Only Collision 2) Injury Collision. The dependent variable SEVERITYCODE corresponds to these classes.

In the variables INATTENTIONIND, PEDROWNOTGRNT, SPEEDING, missing values have been replaced with 'N'. In variables 'ADDRTYPE', 'COLLISIONTYPE', 'JUNCTIONTYPE', 'WEATHER', 'ROADCOND', 'LIGHTCOND' missing values are replaced with 'Unknown'. Lines with empty values of the LOCATION variable and other variables have been removed.

On the basis of the INCDTTM variable, we formed the time of the accident feature. In theory, in the morning, at night, or at any other time, an accident can affect its severity.



It should be noted that, based on the laws of physics, the severity of an accident is primarily affected by kinetic moment. Kinetic moment is equal to the product of speed and mass. Unfortunately, the sample lacks data on the mass of cars and their speed in an accident. Therefore, the classification results without these data will have no practical value.

The categorical variables were converted to numeric using the binary encoding method.

In total, we got 187,223 examples with 52 features.