IBM Data Science Capstone project

Modelling Seattle transportation collision using Classification Machine Learning

Long Phan – Climate Change Data Analyst, SNV Vietnam

August 2020

tuanlongphannguyen@gmail.com

1. Introduction/ Business Problem

A description of the problem and a discussion of the background

In this capstone project, I will choose to work on a transportation problem in Seattle by analyzing the accident characteristic. Transportation accident is one of the most headache issue in any big cities. This problem, fortunately, thanks to the power of data, can be addressed and mitigated. Thus, the goal of the project is to help predict future accidents base on key variables, thus finding the ways to reduce collision rate.

2. Data information

A description of the data and how it will be used to solve the problem

From the SDOT Traffic Management Division, Traffic Records Group, a 72MB dataset is collected. It has 194673 weekly-updated collision records from 2004 to present, in 37 columns. The key variable to be use as predicted criteria (y) is Severity code, which has two unique values 1 and 2, representing for prop damage and injury collision severity, respectively. The dataset, with selected variables, will be modelling using four supervised classification machine learning techniques: K-Nearest Neighbor, Decision Tree, Logistic Regression, and Support Vector Machine. The four model results will be evaluated with three evaluation methods: F1-score, Jaccard index, and log loss. From that, the highest-accuracy model will be selected to predict and identify driven factors.