Capstone Project – Car Accident Severity

Wei-Che Chung

1. Introduction

The purpose of this study is trying to explore the possible reasons and conditions that would cause more severe collisions in the Seattle. By finding the correlation between severity of collision level and corresponding condition, such like weather, location…etc., building prediction model to help government department setting up strategy to decrease highly dangerous collision happens.

2. Data Information

i. Data Description

The data used in this study is publicly available by Seattle Police Department, which includes records of all types of collisions, locations and more details (total 37 attributes) from 2004 to present.

Data available:
https://dataseattlecitygis.opendata.arcgis.com/datasets/collisions/data

ii. Data Understanding

In this data, totally we have 194673 observations. The severity of collision is the target variable for the study, which is assigned to five levels: fatality, serious injury, injury, prop damage and unknown, and used number 5 to 0 to represent these five level. However, in this data, we only have two severity level: injury (2) and prop damage (1).

For the attributes used in this study, we removed some irrelevant attributes, then processed the feature selection and picked up 15 attributes (as below table shows) which we have more correlation with our study.

Categorical Attribute	Numerical Attribute
ADDRTYPE: COLLISIONTYPE JUNCTIONTYPE INATTENTIONIND UNDERINFL WEATHER ROADCOND LIGHTCOND PEDROWNOTGRNT SPEEDING HITPARKEDCAR	PERSONCOUNT PEDCOUNT PEDCYLCOUNT VEHCOUNT

The detail information of attributes and data can be found on below links.

Data Set information:
https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf

iii. Data Preprocessing

Since our dataset includes different types of data, such like text, string, or number. We need to transform these data so that we could apply analysis algorithms on them. Also, we have some attributes with different missing value condition. In order to keep our dataset as completed as possible, here we chose to interpolate the missing with meaningful values.

3. Methodology

We separated the dataset into training set (80%) and testing set (20%). In order to applying these machine learning algorithms, we also transformed the categorical attributes into some dummy representation.

In the study, we mainly considered four machine learning algorithms to build our prediction model:

i. K-Nearest Neighbor (KNN)

ii. Decision Tree

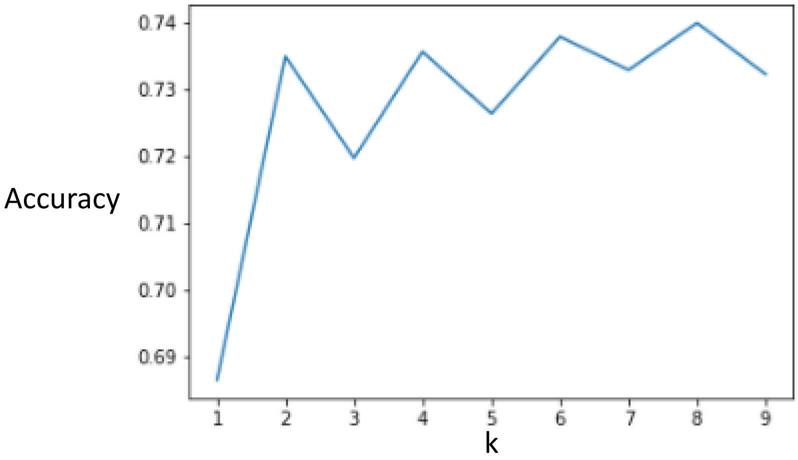
iii. Support Vector Machine (SVM)

iv. Logistic Regression

4. Result and Evaluation

v. K-Nearest Neighbor (KNN)

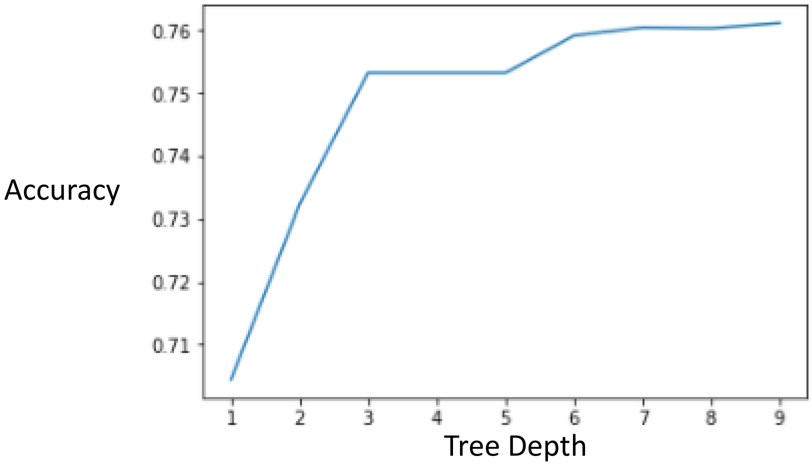
For KNN, we set k distances from 1 to 9 to see the accuracy changes.



We could see that when k = 8 gives us the best accuracy about 74%

vi. Decision Tree

For decision tree, we set tree depth from 1 to 9 to see the accuracy changes.



We could see that when tree depth = 9 gives us the best accuracy about 76%

vii. Support Vector Machine (SVM)

SVM gives us accuracy about 76%

viii. Logistic Regression

Since our target variable is binary, so we could try logistic regression, and at same time, it could predict the probability of each severity happened. For logistic regression, it gives us accuracy about 76%

5. Discussion

While there’re no missing in target variable, but other attributes have different missing conditions. Therefore, how to keep/remove data would be a problem before applying machine learning algorithm.

6. Conclusion

Based on our models, we could use historical data to predict potential severity of car accident. The model could help government department setting up strategy to decrease highly dangerous collision happens at specific weather, road type … etc.