Car Accident in Seattle

Coursera IBM Data Sciences Capstone Project

Introduction

- Car accident in Seattle has a different severity level.
- This study is to create a model to predict the severity level using data science.
- There are 37 attribute and 194,673 rows in the raw data.

Data Preprocessing

Data Cleaning

- Remove no meaning and duplicated data from the data set (a key data, description data)
- data contain too many null value which is easily lead to incorrect prediction (more than 50% of the data)
- After removal, there are 14 attribute left.

Data to be use to predict severity

Data	Meaning
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
JUNCTIONTYPE	Category of junction at which collision took place
SDOT_COLCODE	A code given to the collision by SDOT
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
ST_COLCODE	A code provided by the state that describes the collision
HITPARKEDCAR	Whether or not the collision involved hitting a parked car.

Handle Null Value

Data field	Method
ST_COLCODE	Replace blank value (‘ ‘) as NaN and replace NaN with maximum frequency data
JUNCTIONTYPE	Replace ‘Unknown’ value as NaN and replace NaN with maximum frequency data
UNDERINFL	Replace ‘Y’ as 1 and ‘N’ as 0 and replace NaN with maximum frequency data
HITPARKEDCAR	Replace ‘Y’ as 1 and ‘N’ as 0
ADDRTYPE	Replace NaN with maximum frequency data
COLLISIONTYPE	Replace NaN with maximum frequency data
WEATHER	Replace NaN with maximum frequency data
ROADCOND	Replace NaN with maximum frequency data
LIGHTCOND	Replace NaN with maximum frequency data

Change Data Type and Normalization

- Change ST_COLCODE, UNDERINFL, and HITPARKEDCAR to integer.
- Change category data to integer
- Normalize data

Split Testing and Training Data

- X = 'ADDRTYPE', 'COLLISIONTYPE', 'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'JUNCTIONTYPE', 'SDOT_COLCODE', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND', 'ST_COLCODE', 'HITPARKEDCAR'
- y = ‘SEVERITYCODE’

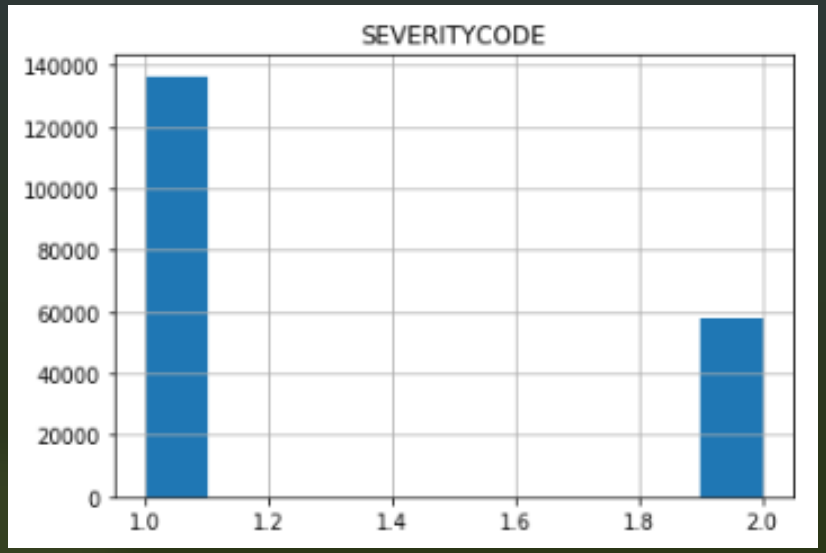
Data	Percentage of all data	Amount of data
Training	70%	136,271
Testing	30%	58,402

Data Exploration

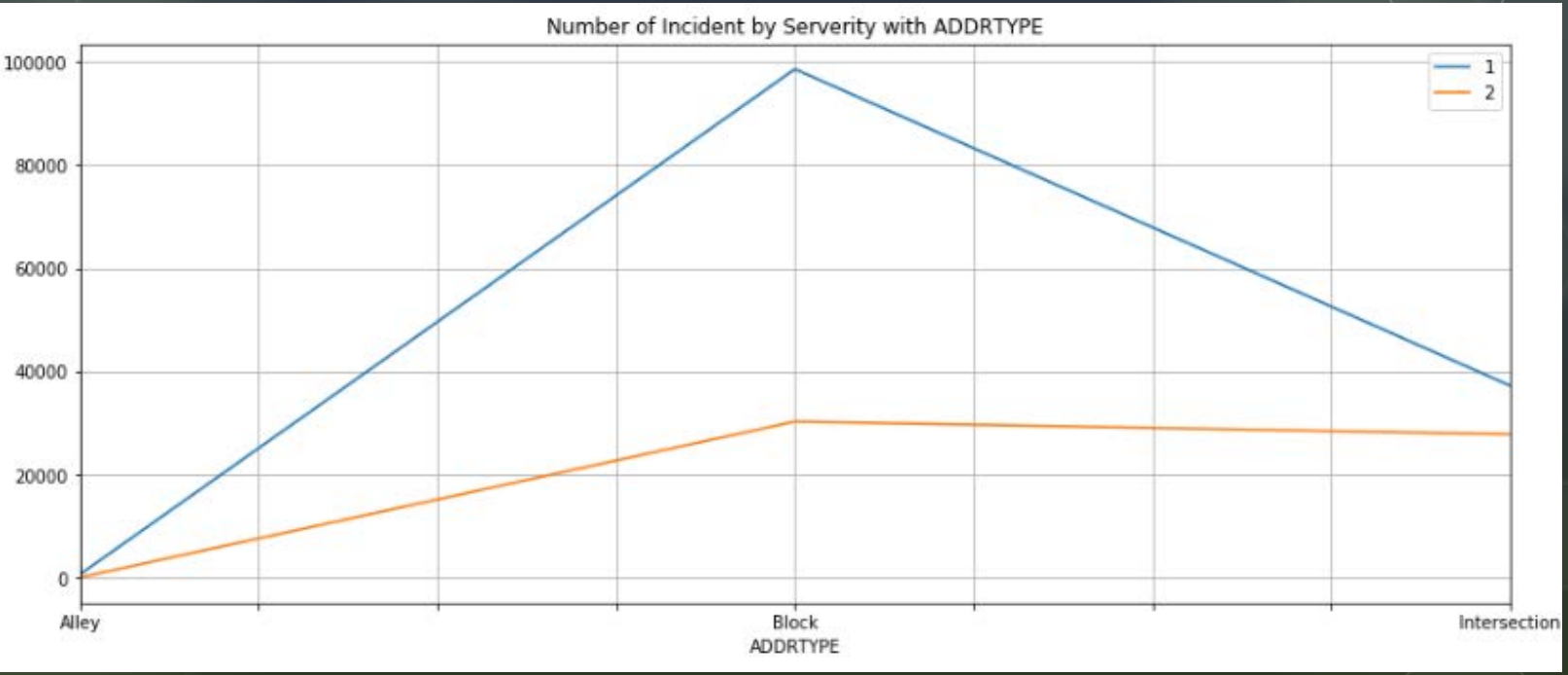
Data to Predict

- Severity Code

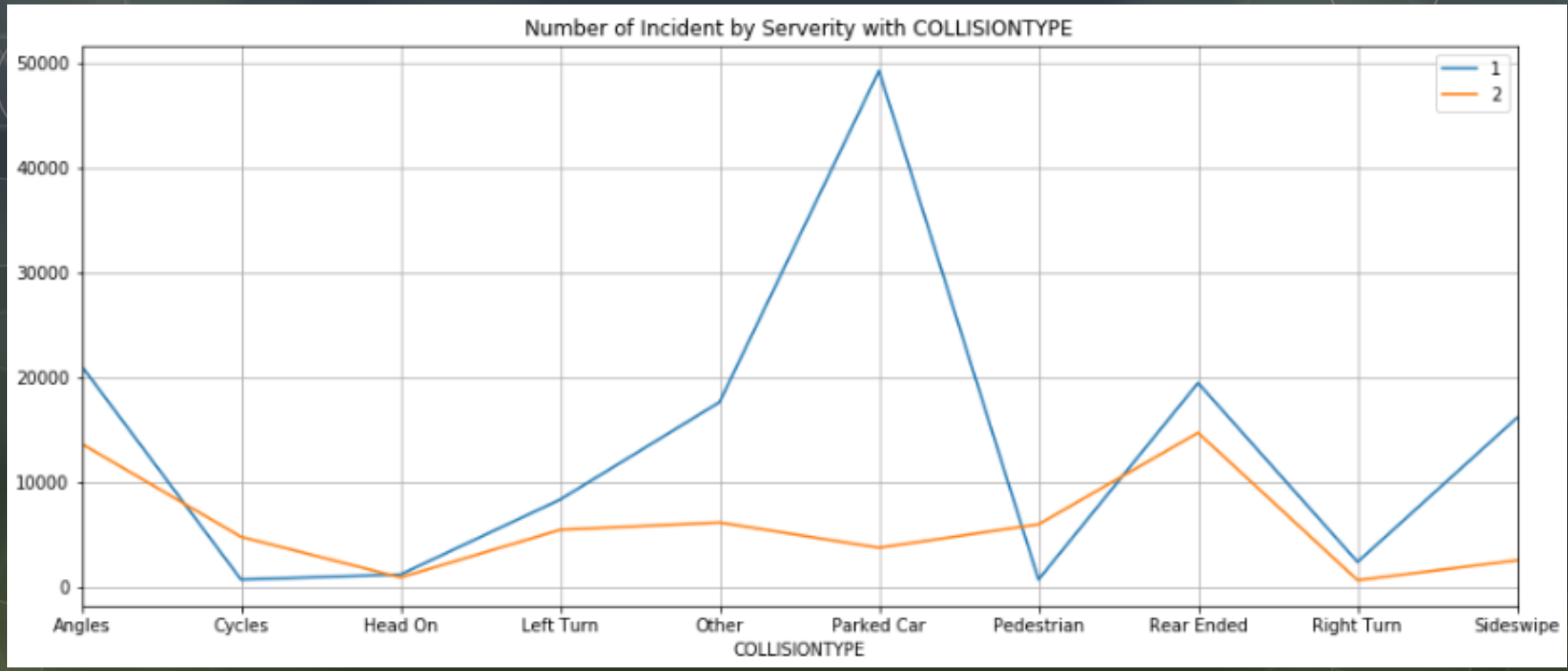
Severity Code	Description	Amount of data
1	Prop damage	136,485
2	Injury	58,188



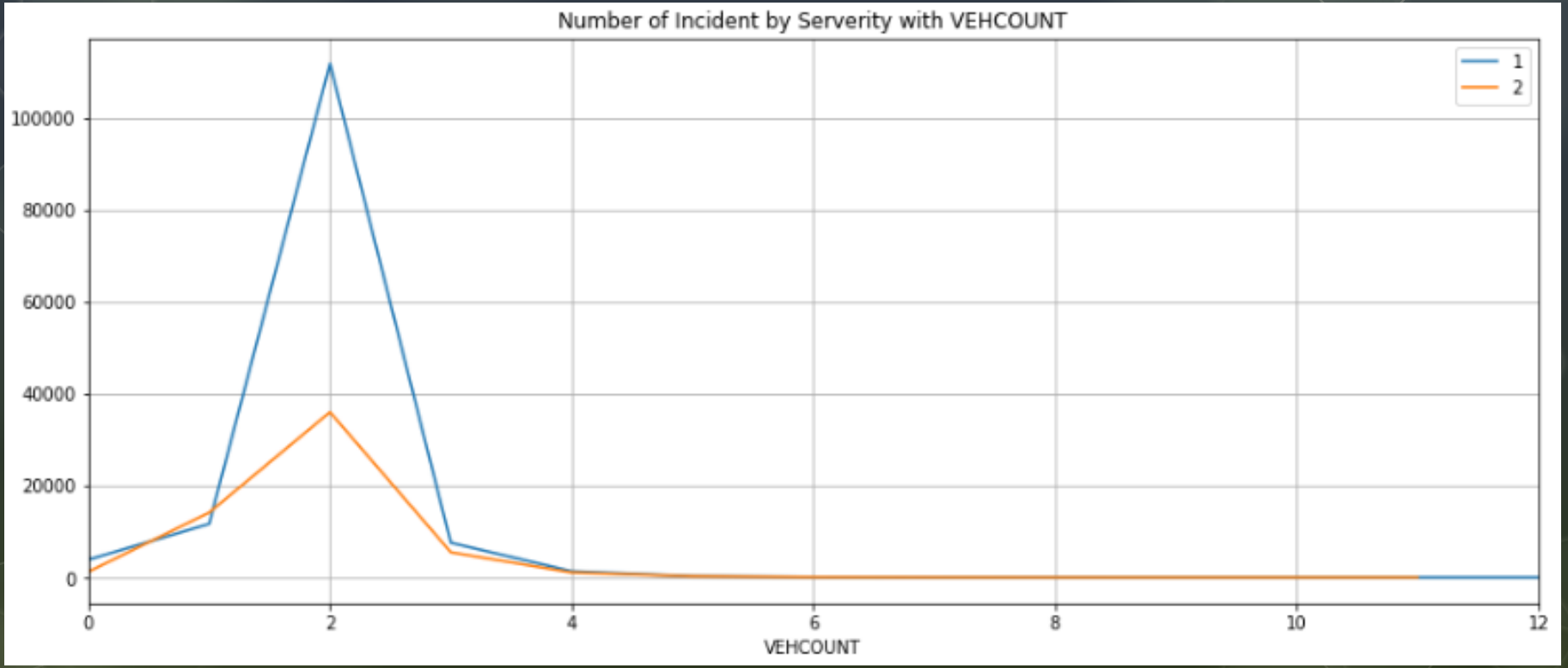
ADDRTYPE group by severity code



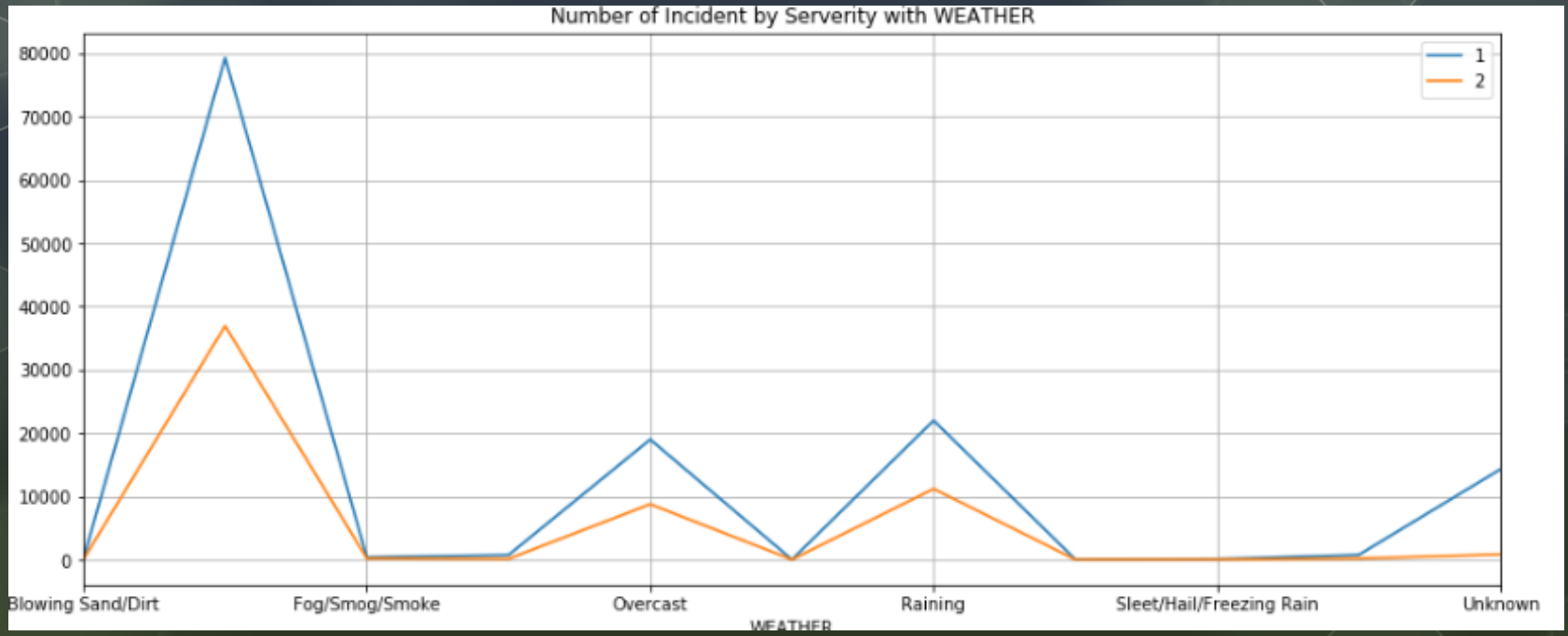
COLLISIONTYPE group by severity code



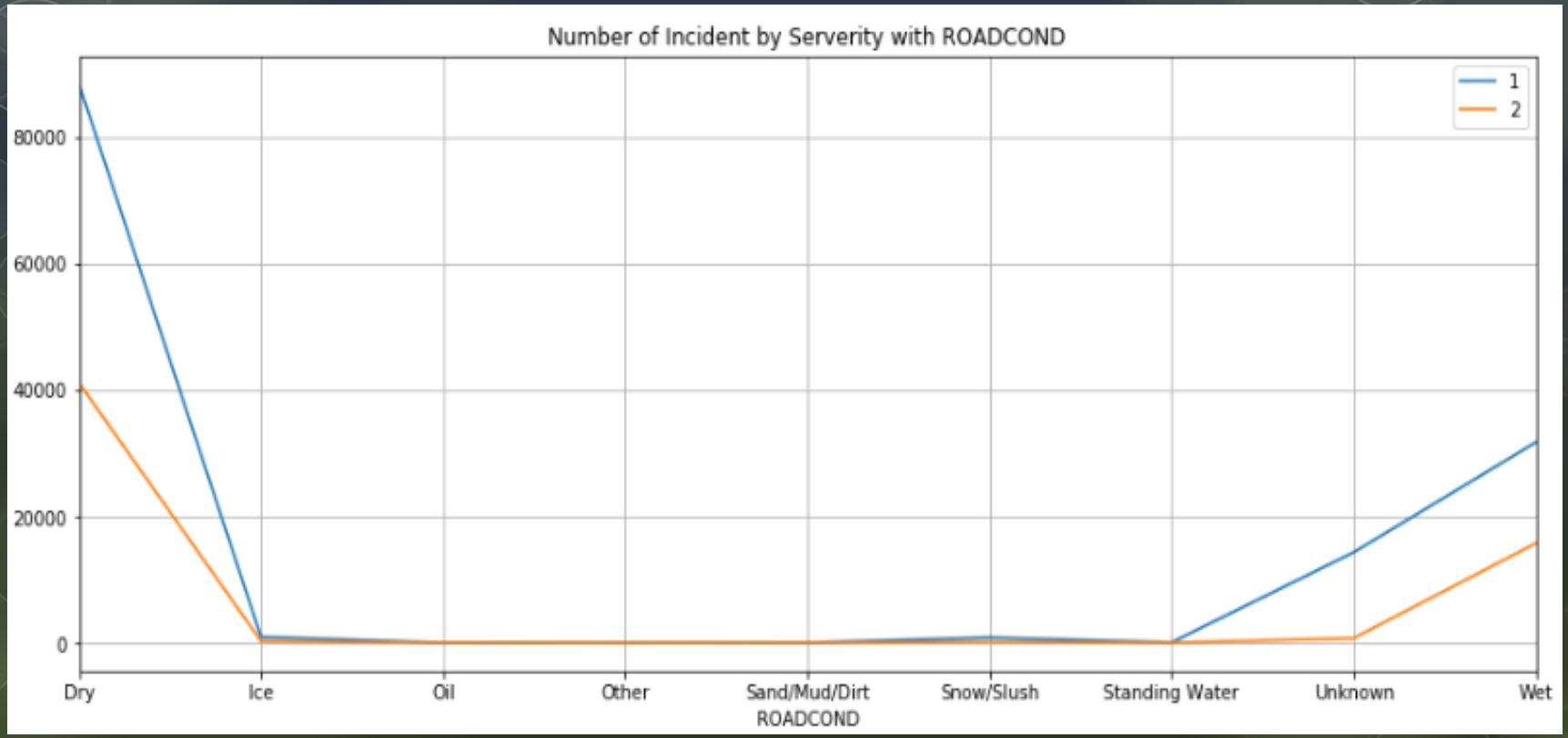
VEHCOUNT group by severity code



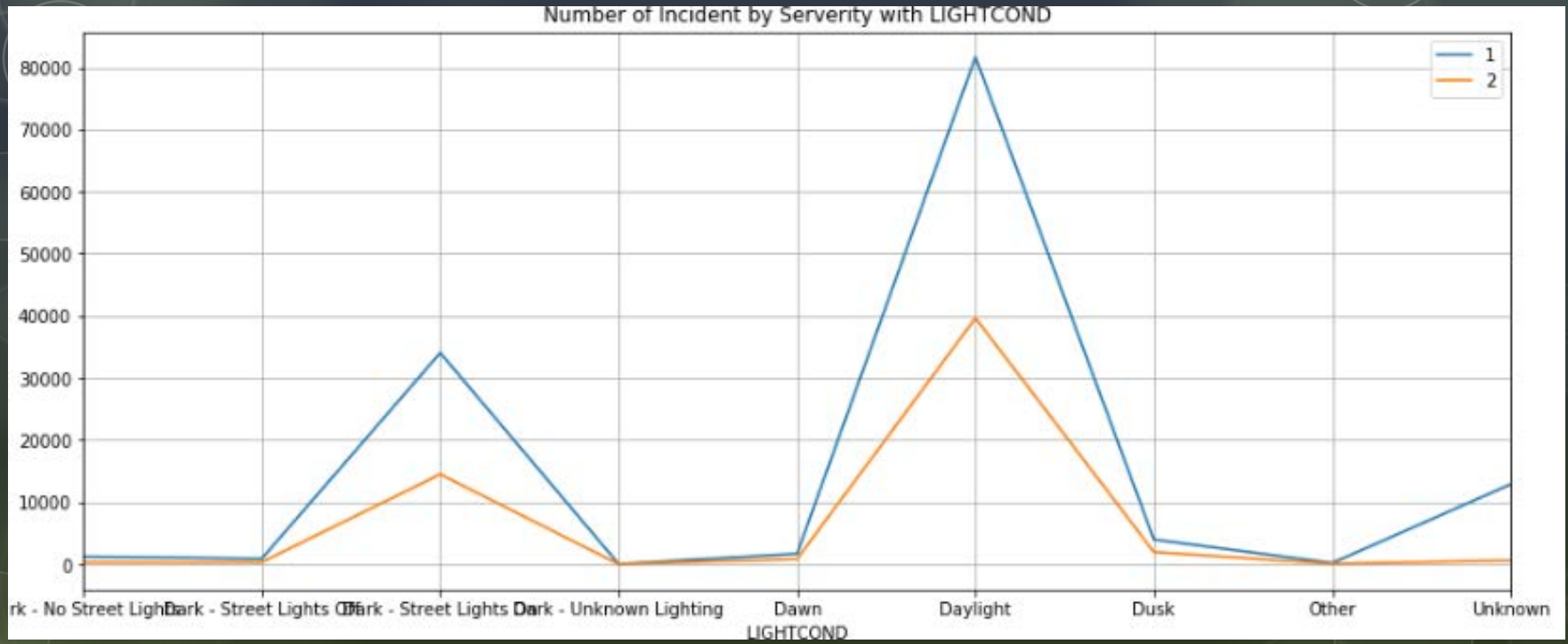
WEATHER group by severity code



ROADCOND group by severity code



LIGHTCOND group by severity code



Data Analysis

Pearson Correlation

Data	Pearson Correlation with SEVERITYCODE
SEVERITYCODE	1
PERSONCOUNT	0.130949
PEDCOUNT	0.246338
PEDCYLCOUNT	0.214218
VEHCOUNT	-0.054686
SDOT_COLCODE	0.188905
UNDERINFL	0.044377
ST_COLCODE	-0.165233
HITPARKEDCAR	-0.101498

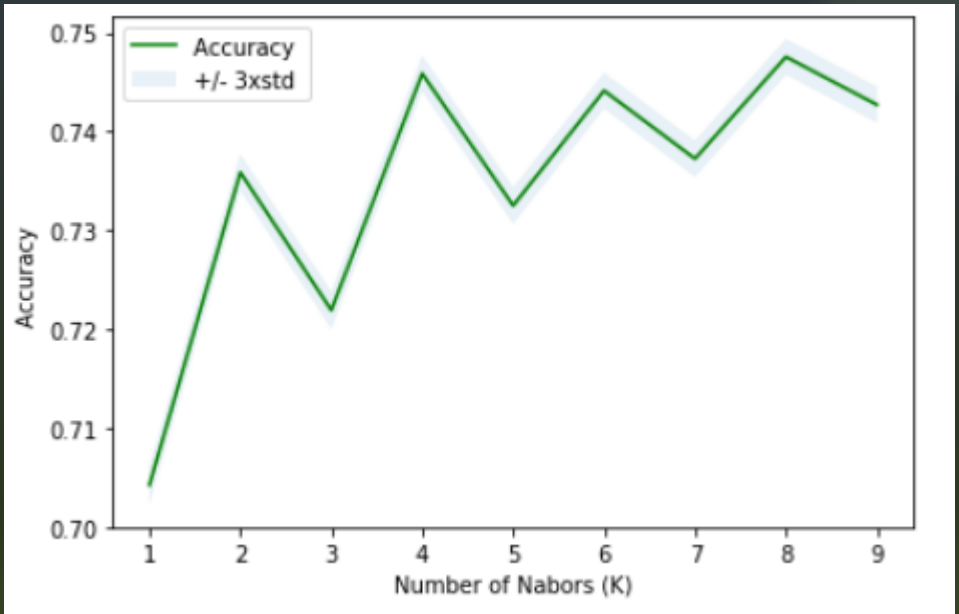
- There is no linear correlation to SEVERITYCODE.

Decision Tree

- Max depth = 5
- Accuracy (the fraction of correctly classified samples) = 0.753758432930379 1

K-Nearest Neighbor (KNN)

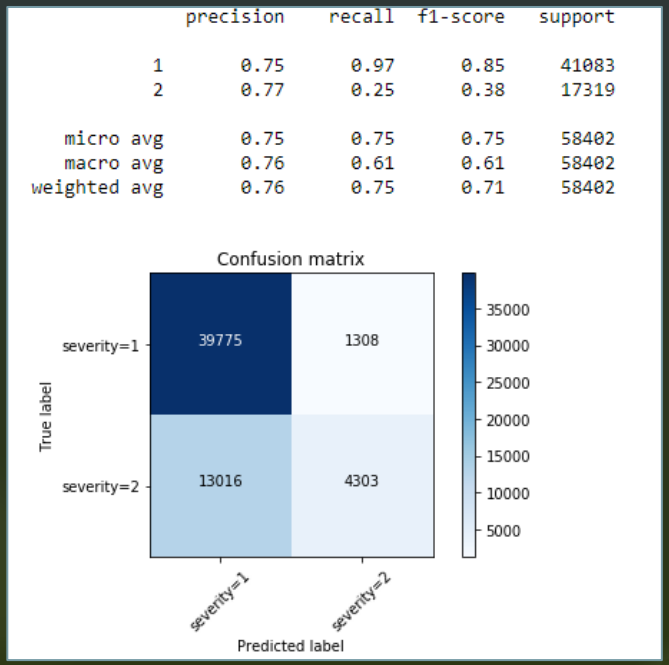
- From experiments, K is 1 to 10, the best K is 8.
- Accuracy (the fraction of correctly classified samples) = 0.7475771377692545



Logistic Regression

- Use Solver Liblinear
- Accuracy
  - Jaccard similarity score = 0.7547344269031883
  - Confusion matrix (F1 average) = 0.717368310349650 1

	precision	recall	f1-score	support
1	0.75	0.97	0.85	41083
2	0.77	0.25	0.38	17319
micro avg	0.75	0.75	0.75	58402
macro avg	0.76	0.61	0.61	58402
weighted avg	0.76	0.75	0.71	58402

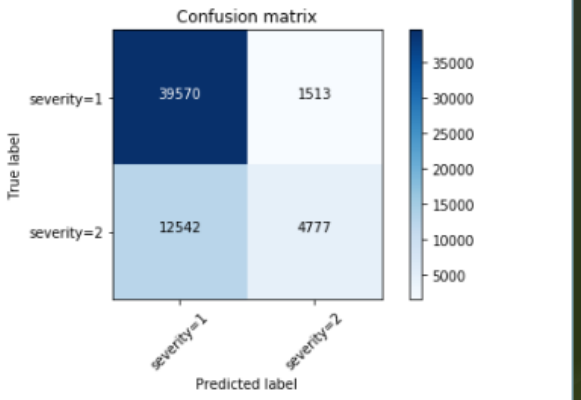


Support Vector Machine (SVM)

- Kernel ‘rbf’
- Accuracy
  - The fraction of correctly classified samples = 0.7593404335467964
  - Confusion matrix (F1 average) = 0.7173683103496501

	precision	recall	f1-score	support
1	0.76	0.96	0.85	41083
2	0.76	0.28	0.40	17319
micro avg	0.76	0.76	0.76	58402
macro avg	0.76	0.62	0.63	58402
weighted avg	0.76	0.76	0.72	58402

Confusion matrix, without normalization
[[39570 1513]
 [12542 4777]]



Summary

- Decision tree algorithm, K-Nearest Neighbor algorithm (KNN), Logistic Regression, and Support Vector Machine algorithm (SVM), shows similar ability to predict the severity level.
- Decision tree and SVM has accuracy score 76%, where KNN and Logistic Regression have accuracy score 75%.