CAR ACCIDENT SEVERITY PREDICTION

October 18, 2020

INTRODUCTION/BUSINESS PROBLEM

Car accidents are a serious problem. They result in personal injury, property damage and death. On average, six million car accidents occur each year in the United States alone. Of these accidents, three million people are injured often permanently. Also, more than ninety people die every day from car accidents.

The injuries, deaths and property damage have untold personal and economic ramifications such as lost wages and productivity, increased insurance premiums and medical costs that reverberate throughout society and impact us all. To gain some valuable insight into the factors related to this issue, this analysis will focus on one specific geographic area—Seattle, Washington. Attributes such as weather, road condition, light conditions, speeding, inattention and under the influence will be used to predict car accident severity.

Being able to accurately predict car accident severity will benefit both direct and indirect stakeholders. For example, if it is determined that low lighting conditions facilitate car accidents, perhaps additional streetlamps could be installed to increase driver visibility thus reducing accidents. Drivers, public safety officials, insurance companies, emergency room personnel, etc. would all benefit from these insights.

DATA

This project will work with data recorded by the Seattle DOT Traffic Management Division, Traffic Records group in the Data-Collisons.csv file which was downloaded from the Coursera website. The dataset has observations of traffic accidents from 2004 to the present day and is updated weekly. At the present time, there are approximately 194,673 observations in this data set with 38 different attributes. Non-essential attributes will be dropped from the dataset. A detailed description of the attributes can be found on the metadata.pdf in this project's GITHUB repository.

The target label for this project is SEVERITYCODE which has the following codes (according to the metadata.pdf):

SEVERITYCODE

CODE	DESCRIPTION
3	Fatality
2b	Serious Injury

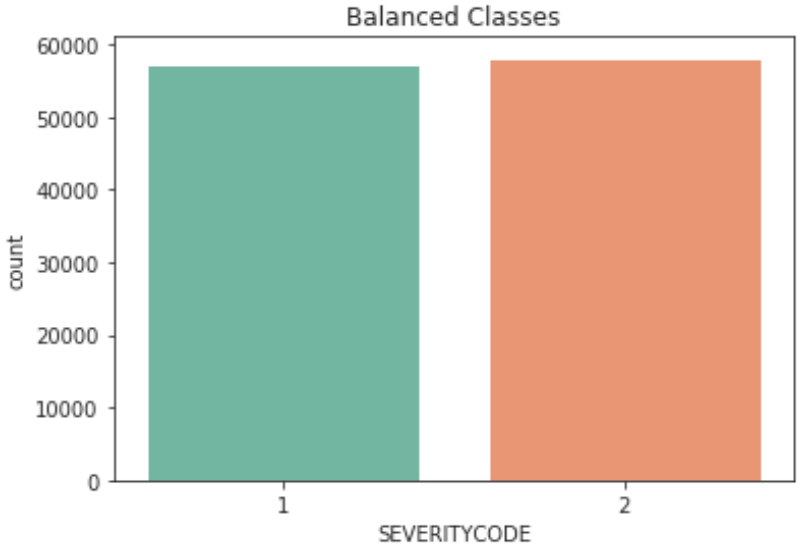
2	Injury
1	Property damage
0	Unknown

In the dataset, however, only severity codes 1 (Property Damage) and 2 (Injury) are used. There are 136,485 (70%) instances where severity code equals 1 and 58,188 (30%) instances where severity code equals 2.

METHODOLOGY

The dataset is imbalanced and required resampling (under sampling) so that the dataset is balanced.

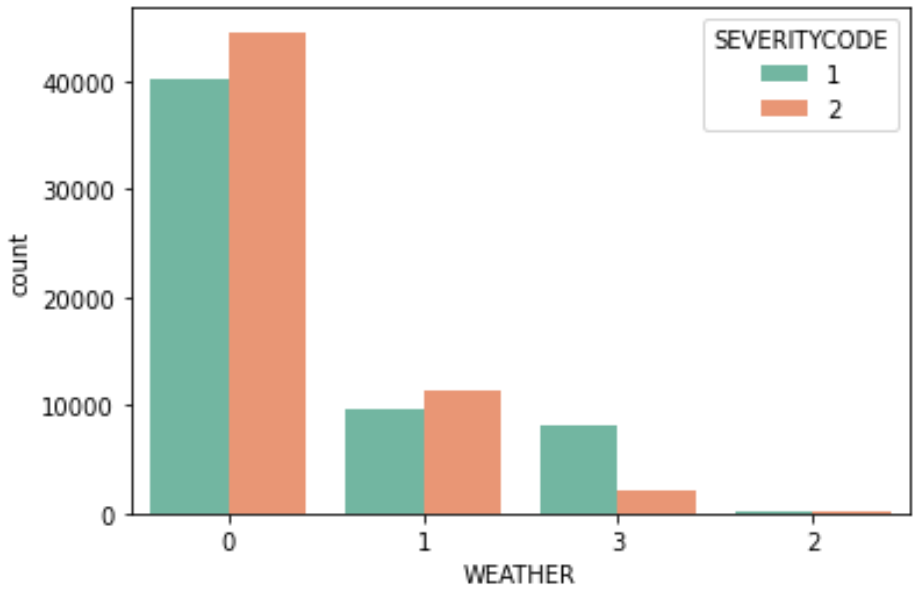
The graph below represents the dataset after resampling.



Upon initial exploratory analysis, the following is a list of important features in the dataset that will be used in the data model(s). Since many of the features contain classification information, they will have to be converted to numerical or binary format prior to running the model(s). Fields that contained null values and NaN entries were also cleaned up as well. Statistical analysis was used to determine if there were any apparent patterns and trends. The results were also visualized in bar graphs to show relationships between variables to the target.

WEATHER – There are 11 different types of weather condition classifications in the dataset: Clear, Raining, Overcast, Unknown, Snowing, Other, Fog/Smog/Smoke, Sleet/Hail/Freezing Rain, Blowing Sand/Dirt, Severe Crosswind. They will be grouped into four numerical codes as follows:

CODE	GROUP	DESCRIPTION IN DATASET
0	Dry	Clear, Partly Cloudy, Overcast
1	Wet	Raining, Snowing, Sleet/Hail/Freezing Rain
2	Unusual	Blowing Sand/Dirt, Severe Crosswind, Fog/Smog/Smoke
3	Other	Other, Unknown

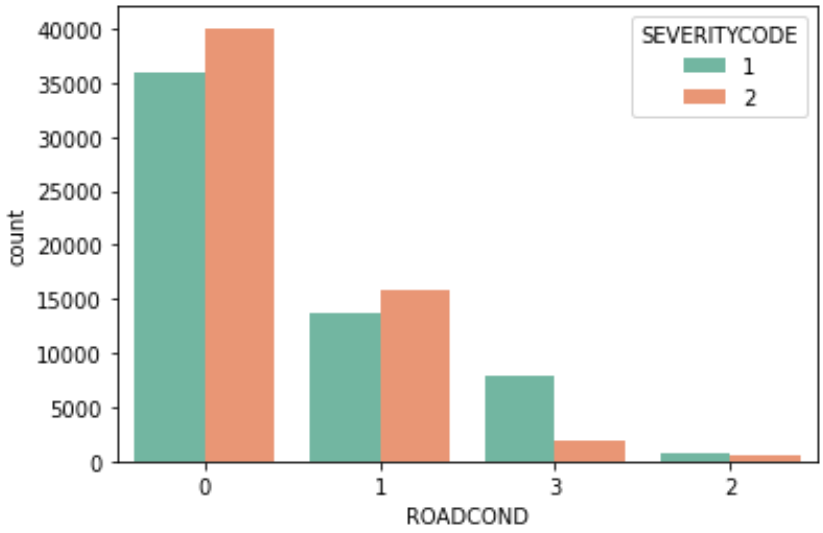


72.8% of accidents occur when weather is Dry (Clear, Partly Cloudy, Overcast)

ROADCOND – There are 9 different types of road condition classifications in the dataset: Dry, Wet, Unknow, Ice, Snow/Slush, Other, Standing Water, Sand/Mud/Dirt, Oil. They will be grouped into four numerical codes as follows:

CODE	GROUP	DESCRIPTION IN DATASET
0	Good	Dry
1	Fair	Wet, Sand/Mud/Dirt, Oil

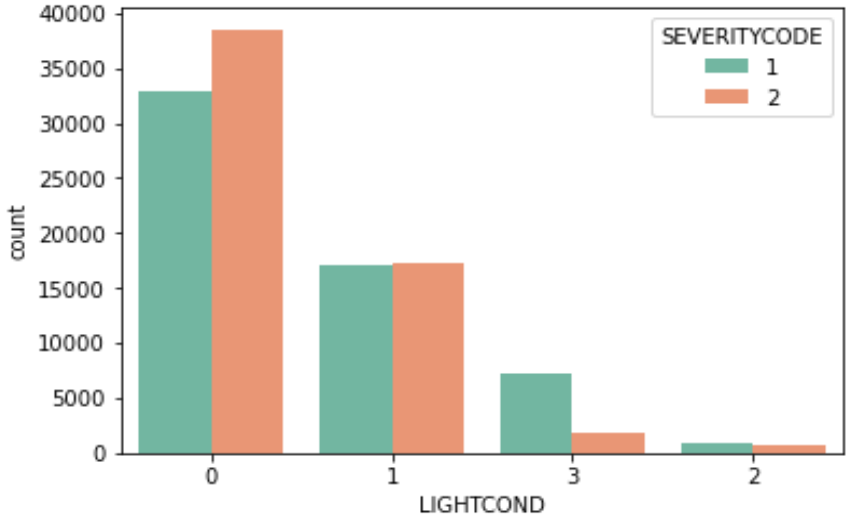
2	Poor	Ice, Standing Water, Snow/Slush
3	Other	Other, Unknown



65% of car accidents occur when road conditions are Good(Dry).

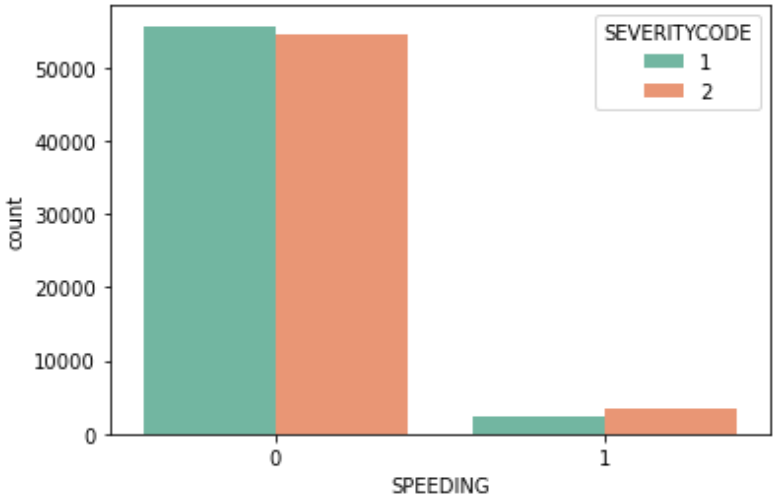
LIGHTCOND – There are 9 different types of light condition classifications in the dataset: Daylight, Dark – Street Lights On, Unknown, Dusk, Dawn, Dark – No Street Lights, Dark – Street Lights Off, Other, Dark – Unknown Lighting. They will be grouped into four numerical codes as follows:

CODE	GROUP	DESCRIPTION IN DATASET
0	Light	Daylight
1	Partial Light	Dusk, Dawn, Dark – Street Lights On
2	Dark	Dark – No Street Lights, Dark – Street Lights Off, Dark – Unknown Lighting
3	Other	Other, Unknown



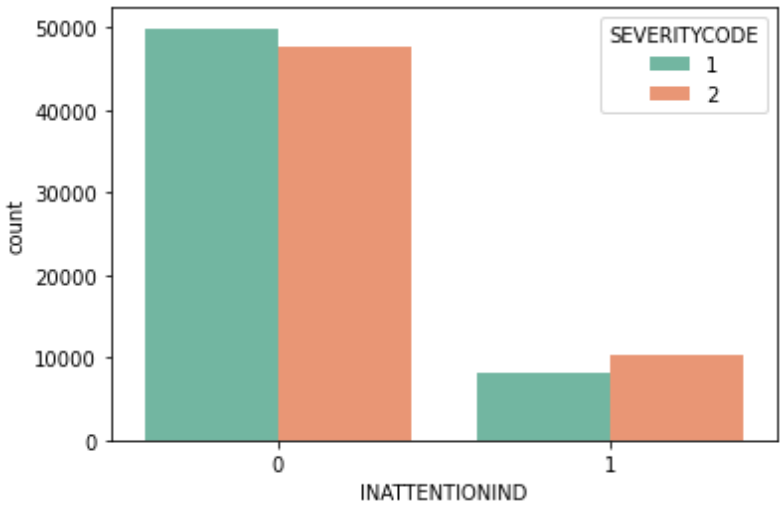
61.4% of accidents occur when it is Light(Daylight) outdoors.

SPEEDING – The speeding indicator denotes whether a driver was speeding by “Y”. This field will be converted to binary format. Fields containing a “Y” will be converted to a 1. Blanks or NaN will be converted to 0.



5.13% of car accidents involved speeding. Of these accidents, 49% resulted in property damage (Severity Code 1) while 59% resulted in injury (Severity Code 2).

INATTENTIONIND – The inattention indicator denotes if the driver is inattentive with a Y. This field will be converted to binary format. Blanks or NaN will be converted to 0.

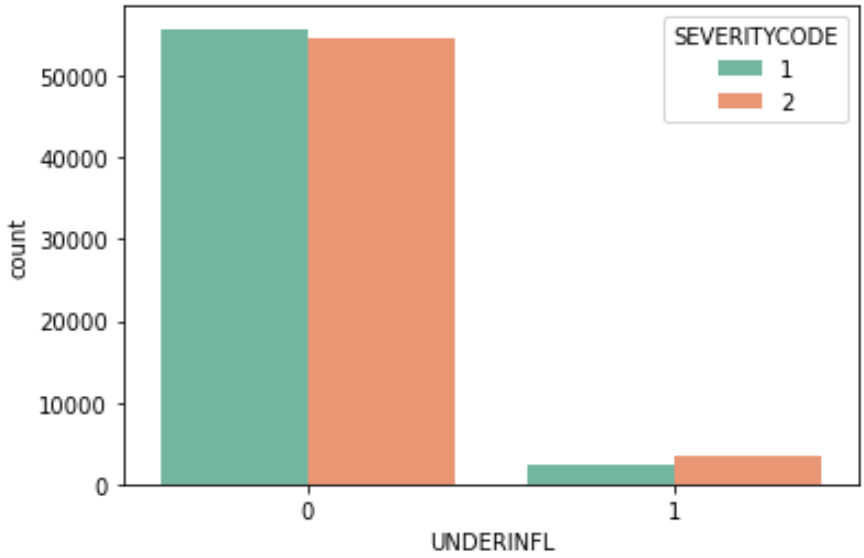


16% of car accidents involved drivers who were inattentive. Of these accidents, 44.3% resulted in property damage (Severity Code 1) while 55.7% resulted in injury (Severity Code 2).

UNDERINFL – The under the influence indicator denotes if the driver is under the influence with either a Y or a 1. N, 0 and NaN denote not under the influence. This field will be converted to binary format.

Example of current values in balanced dataset:

N 60232
0 47463
Y 3340
1 2648
Name: UNDERINFL, dtype: int64



5.15% of drivers involved in car accidents have a positive under the Influence indicator. Of these accidents, 40.5% resulted in property damage (Severity Code 1) while 59.5% Resulted in injury (Severity Code 2).

KEY OBSERVATION FROM EXPOLRATORY ANALYSIS

Based upon this analysis, controllable driver behavior which includes speeding, inattention or under the influence while driving results in a higher incidence of injury.

Once data was cleaned up, classification datatypes were changed to integers for use in the model(s). The following is an example of the Feature set.

Target(Y) **Feature Set(X)**:

SEVERITYCODE	WEATHER	ROADCOND	LIGHTCOND	SPEEDING	INATTENTIONIND	UNDERINFL
2	2	2	2	1	0	0
1	0	0	0	0	1	0

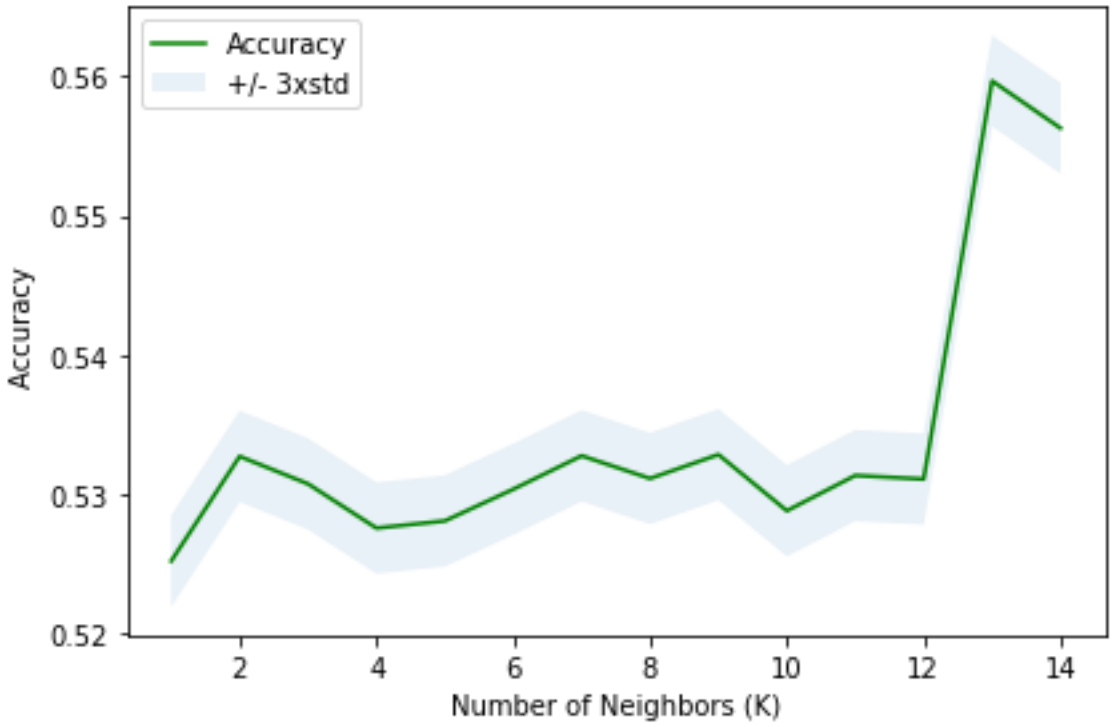
The feature set was normalized for the models.
X=preprocessing.StandardScaler().fit(X).transform(X)

Both the K Nearest Neighbors and Decision Trees were used for machine learning. These two methods were chosen for their predictive abilities. Since the objective of this program is to predict the severity of an accident based on variables in the feature set.

RESULTS

K Nearest Neighbors (KNN) - The training set accounted for 80% of the dataset while test set accounted for 20%. Initially, a k of 6 was used in the KNeighborsClassifier. This resulted in very similar accuracies between training set and test set, .5311 and .5303 respectively.

To determine if another k would be a better fit than the initial k=6, program code was created to test ks from (1 to 15). It determined that the Best K was 13 and resulted in an accuracy of .5597.



DECISION TREE – The training set accounted for 70% of the dataset while the test set accounted for 30% with a maximum depth of 6. This resulted in a Decision Tree Accuracy of .5617.

DISCUSSION/CONCLUSION

Several key observations were made with this dataset. One was very surprising. Most accidents occurred when the weather was dry, road conditions were “good” and it was light outside.

Speeding accounted for 5.13% of accidents where 49% of speeding accidents resulted in property damage (Severity Code 1) and 59% resulted in injury (Severity Code 2). Inattention accounted for 16% of accidents where 44.3% resulted in property damage (Severity Code 1) and 55.7% resulted in injuries (Severity Code 2). Drivers who were under the influence accounted for 5.5% of total accidents with 40.5% of those resulting in property damage (Severity Code 1) and 59.5% of those resulting in injury (Severity Code 2).

Speeding, driver inattention and drivers who were driving under the influence at the time of accident resulted in a higher percentage of injuries. Driver inattention was almost three times higher than that of drivers who were speeding or driving under the influence. Inattention could be caused by a variety of factors including device usage while driving.

One recommendation is that public safety officials can do an aggressive campaign to warn the public about the dangers of distracted driving. There can also be steeper fines for drivers caught texting while they are operating a vehicle. Activist groups have already done an excellent job about informing the public about the dangers of operating a vehicle while under the influence. Also, the legal and financial consequences for DUI and speeding has resulted in a decrease in these types of driver behaviors. More awareness will need to be promoted about the dangers of inattentive driving as well to reduce the number of accidents related to this behavior.

K Nearest Neighbors (KNN) and Decision Tree modules used with the Feature set did an adequate job at predicting accident severity. Of the two models, the Decision Tree performed the best with a .5617 accuracy.

Source

http://www.driverknowledge.com