Car accident severity report: A Capstone Project

6 minute read

by Aliah H.

Introduction / Business Problem

According to the Washington State Department of Transportation 2015 Annual Collision Summary, a crash occured every 4.5 minutes. Devastatingly, a person died in a crash every 16 hours and a person was injured in a crash every 11 minutes . There were a total of 117,053 collisions with 499 of them resulted in fatality and 1752 of them with serious injuries [1]. In 2018 alone, there were 10, 249 police reported collisions on Seattle street [2].

Meanwhile, the US Department of Transportation's National Highway Traffic Safety Administration (NHTSA) motor vehicle crashes imposed USD 836 billion in economic cost and societal harm on the country in 2010. And for the 2017 financial year, NHTSA requests USD 1.181 billion to effectively continue its mission of ensuring safer drivers, safer cars, and safer roads [3].

In this project, we will try to predict the severity of an accident based on road and weather conditions. Specifically, this report will be targeted to stakeholders interested in developing an app or system that could alert and warn drivers about the potential risk they are facing when driving.

Data

The dataset used for this project is based on reported car collisions in Seattle, Washington from 2004 to 2020. The dataset lists the severity of each car accidents along with the time and conditions under which each accident occurred. There are 38 attributes with 194673 entries included in the raw dataset. There is also a metadata document provided along with the dataset that describes each attribute.

Methodology

Machine learning method

In this project, two classifier models were built using Decision Tree and K Nearest Neighbour. Both models are built, analysed and visualised using Python libraries sklearn, pandas, numpy and matplotlib.

Exploratory data analysis

To get to know the data, the shape, the name and the datatype of the attributes were identified using pandas library. Then, the dataset was checked for missing values and the target label was checked for count balance.

Data preparation

To prepare the dataset for modelling, attributes identified with more than 50% missing values were dropped. Further, only entries with missing values were dropped for the remaining attributes. The minority class for the target label was then upsampled to balance the dataset. The identified categorical attributes were then converted into numerical using <html>LabelEncoder</html>.

Modelling, Prediction and Analysis The balanced dataset was split into training and testing datasets. Both models were built using the train dataset and predictions were carried out using the test dataset. Both models were then analysed for accuracy by calculating the accuracy and F1 score.

Result

Exploratory Data Analysis

For this step, I did some simple exploration to identify missing values and the datatypes of the attributes. The raw data has 38 attributes and 19,4673 entries. There were 6 attributes with more than 50% missing values in the dataset. These attributes were dropped from the dataset. For the remaining attributes, only entries with missing values were ddropped. Majority of the remaining attributes are numerical whereas 7 attributes were found numerical. A summary is shown below:

	Count
# of Entries	194673
# of Attributes	38
# of columns with >50% missing values	6
# of Categorical attributes	7

Dropping columns

There are 38 attributes in the data. In order to choose which features to use for modelling later on, we first identify the columns that would not be useful for the analysis. This is done by identifying the missing values.

- **PEDROWNOTGRNT** : Whether or not the pedestrian right of way was not granted. (Y/N)
- **EXCEPTRSNDESC** : No description provided
- **SPEEDING** : Whether or not speeding was a factor in the collision. (Y/N)
- **INATTENTIONIND** : Whether or not collision was due to inattention. (Y/N)
- **INTKEY** : Key that corresponds to the intersection associated with a collision
- **EXCEPTRSNCODE** : No description provided
- **SDOTCOLNUM**: A number given to the collision by SDOT.

These top 6 attributes has more than 50% missing values, therefore would not be useful if included in the data for modelling. Also preliminarily, four more additional attributes identified (‘REPORTN0’, ‘SDOTCOLNUM’,’X’, ’Y’) to would not be contributing to the analysis, thus dropped. These attributes were the report ID and the coordinates for the accident.

Additionally, columns that are not contributing or too complex are also removed. For example, the attribute 'LOCATION' is a categorical data with more than 10,000 different entries and 'SEVERITYDESC' attribute is redundant.

Data Preparation

Target label

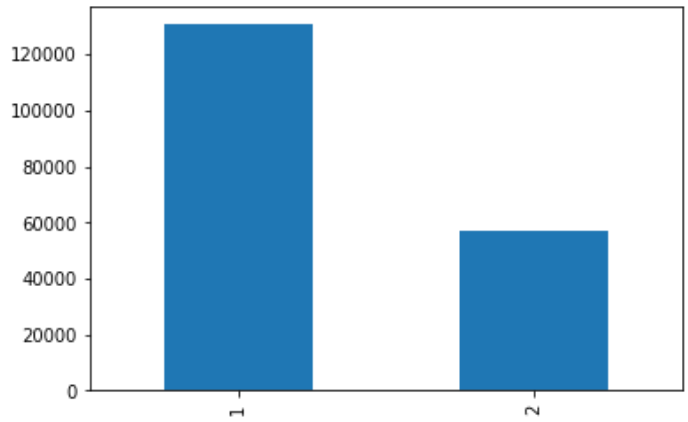
The target label is 'SEVERITYCODE' and the description is listed in 'SEVERITYDESC'.

The 'SEVERITYCODE' attribute lists the code that corresponds to the severity of the collision:

- 3 - fatality
- 2b - serious injury
- 2 - injury
- 1 - prop damage
- 0 - unknown

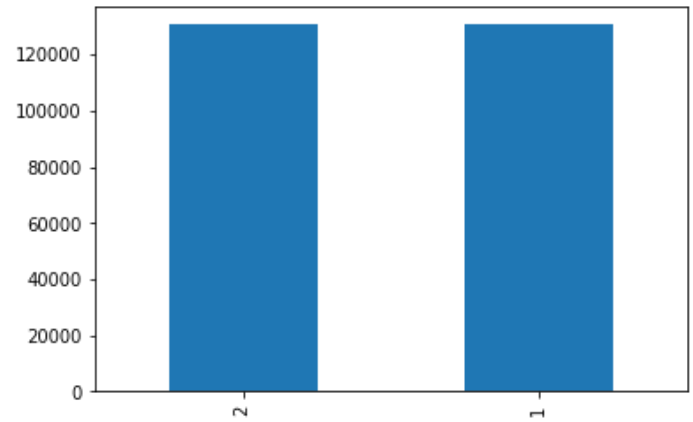
	SEVERITYCODE
1	130634
2	56870

The data is not balanced.



Balancing the dataset by upsampling minority class

The data was balanced by upsampling the minority class 2. This increased the category to 1300634 entries.



Setting up the dataset

Some preprocessing to generate feature set, X. All categorical data from the upsampled dataframe was converted into numerical using LabelEncoder. For 'UNDERINFL' attribute, the values were standardized using a replace method. The dataset was split into training and testing dataset with a 70:30 ratio.

'UNDERINFL' was categorised into 4 different values that was redundant.

	UNDERINFL
N	138460
0	109098
Y	7609
1	6101

'N' and 'Y' was replaced to '0' and '1' respectively.

	UNDERINFL
0	247558
1	13710

Attributes with categorical datatype is shown in the table below:

	ADDRTYPE	COLLISIONTYPE	UNDERINFL	WEATHER	ROADCOND	LIGHTCOND	ST_COLCODE
1	Block	Sideswipe	0	Raining	Wet	Dark - Street Lights On	11
2	Block	Parked Car	0	Overcast	Dry	Daylight	32
3	Block	Other	0	Clear	Dry	Daylight	23
5	Intersection	Angles	0	Clear	Dry	Daylight	10
6	Intersection	Angles	0	Raining	Wet	Daylight	10

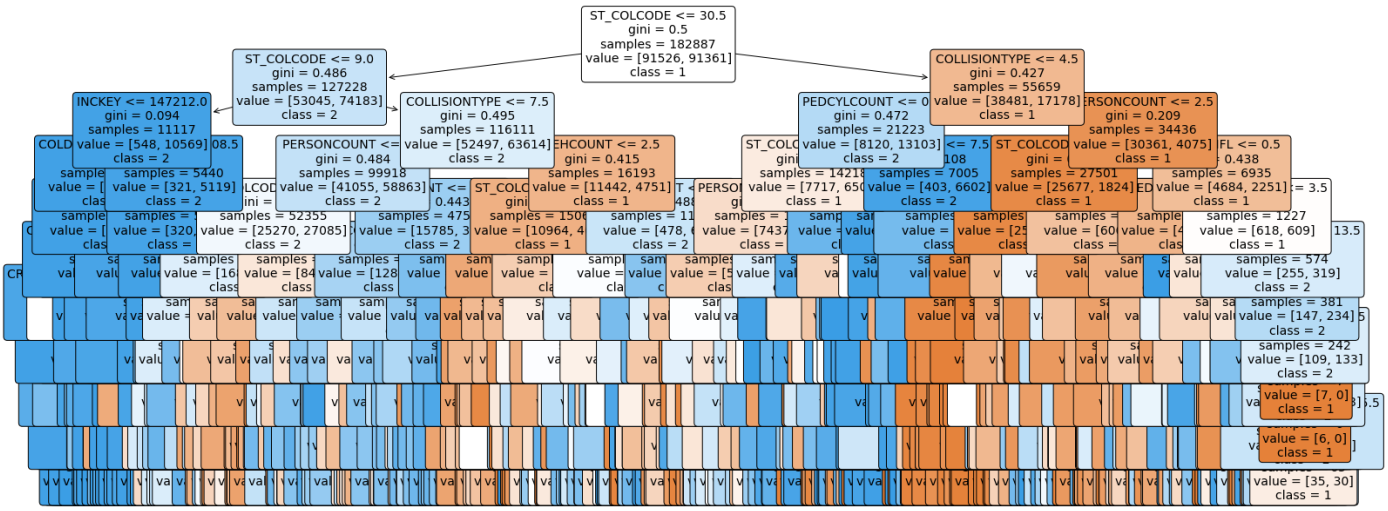
Decision Tree Classifier

Modelling

A decision tree was built using the DecisionTreeClassifier function from sklearn's library. The maximum depth for tree was set at 10 and the tree was trained using the training dataset.

Prediction

The trained model was then used to predict collision severity using the testing dataset.



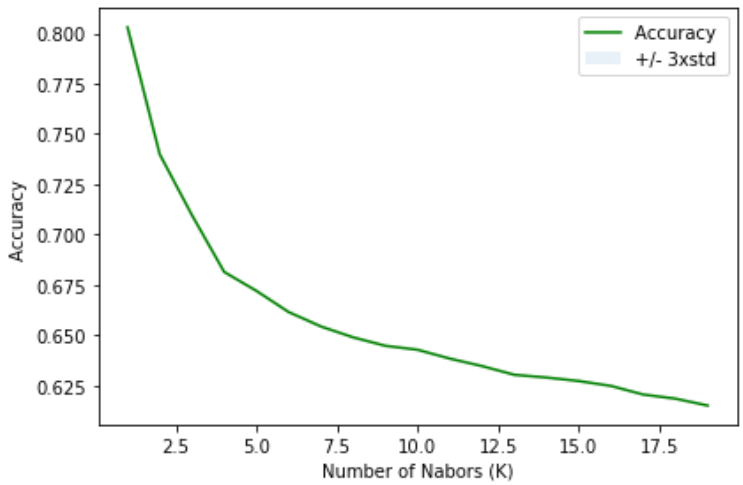
Evaluation

The accuracy score for the tree was 0.73, which was good. Top 3 factors predicting severity is the number of pedestrian, number of cyclists and number of individuals likely to involve in the collision.

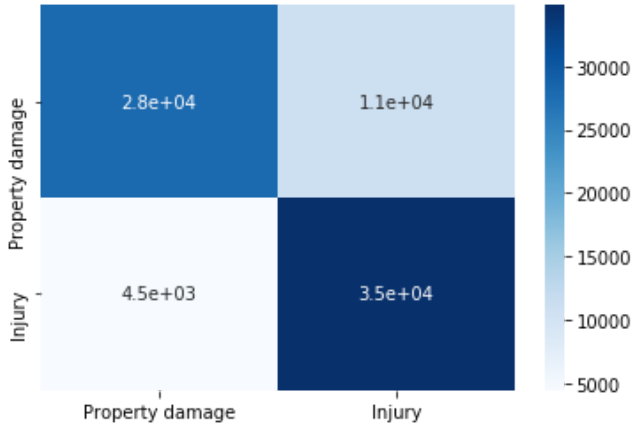
KNN Classifier

Modelling

A KNN model was built using the KNeighborsClassifier function from sklearn's library. The accuracy of the model was tested against a range of number of k to identify the best k. The best accuracy was 0.8 with k=1.



Evaluation



The accuracy score for the best KNN model was 0.80 and the F1 score was 0.78 and 0.82 for predicting 'Property Damage only' and 'Injury' collisions.

The model is optimised at k=1, at which the model correctly predicts accident severity code 1 and 2- 86% and 76% of the time, respectively. The F1 scores of the two accident outcomes are 0.78 and 0.82.

Discussion and summary

The scores for the KNN model are higher than of the Decision Tree demonstrating that for this dataset, the KNN model performs better.

References

1. Washington State Department of Transportation 2015 Annual Collision Summary. https://wsdot.wa.gov/mapsdata/crash/pdf/2015_Annual_Collision_Summary.pdf

2. Seattle Department of Transportation 2019 Traffic Report. https://www.seattle.gov/Documents/Departments/SDOT/VisionZero/2019_Traffic_Report.pdf

3. National Highway Traffic Safety Administration (NHTSA) Budget Estimates (Fiscal Year 2017). https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/fy2017-nhtsa_cbj_final_02_2016.pdf

Tags: Classifier (https://aliahhawari.github.io/tags/#classifier) Decision Tree (https://aliahhawari.github.io/tags/#decision-tree) KNN (https://aliahhawari.github.io/tags/#knn) Machine Learning (https://aliahhawari.github.io/tags/#machine-learning)

Categories: machinelearning (https://aliahhawari.github.io/categories/#machinelearning)

Updated: October 10, 2020