Predicting Severity Codes of Car Collision Accidents

Jungin Han

October 30, 2020

1. Introduction

1. Background

Car collisions have been huge problems to our society, sometimes leading to serious casualties. Thus it is important to analyze the previously obtained data and predict before they happen. Through our capstone project, our primary goal is to build an appropriate machine learning model and predict the severity codes, which is one of the main parameters describing the severity of accidents.

Our target variable "severity code" has five possible values as follows:

- 0 : unknown
- 1 : property damage
- 2 : injury
- 2b : serious injury
- 3 : fatality

2. Business Understanding

Classifying the severity of accidents using severity codes would lead to a big decrease in casualties and damages of accidents in future as people regarding this problem can use the data to improve and reform environments such as road conditions for reducing the total property and human damages.

2. Data

1. Data Acquisition

The provided example dataset, the data of all collisions in Seattle from 2004 to present, will be used for this project. This data have 35 attributes in total including SEVERITYCODE. As we do not need all the attributes, some attributes that look irrelevant to our modeling will be deleted. I chose to drop some attributes with descriptions and key codes and to have 15 attributes for further data preparation. Following is the first five row of our data.

	SEVERITYCODE	ADDRTYPE	COLLISIONTYPE	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	INCDATE	INATTENTIONIND	UNDERINFL	WEATHER
0	2	Intersection	Angles	2	0	0	2	2013/03/27 00:00:00+00	NaN	N	Overcast
1	1	Block	Sideswipe	2	0	0	2	2006/12/20 00:00:00+00	NaN	0	Raining
2	1	Block	Parked Car	4	0	0	3	2004/11/18 00:00:00+00	NaN	0	Overcast
3	1	Block	Other	3	0	0	3	2013/03/29 00:00:00+00	NaN	N	Clear
4	2	Intersection	Angles	2	0	0	2	2004/01/28 00:00:00+00	NaN	0	Raining

2. Data Cleaning

First I resampled the data because we have only two classes of the target variable and these classes have unbalanced number of data set. Downsampling creates a random subset to have same size of data set for each class.

Next I dropped values that have occurrences less than 50 times regarding as outliers and replaced the missing values as 'Unknown' values.

3. Feature Selection

By exploring the cleaned data, the following six attributes are chosen to be used: ADDRTYPE, WEATHER, ROADCOND, LIGHTCOND, PERSONCOUNT, and PEDCOUNT. We set the feature data set X having those six attributes and the target variable y, SEVERITYCODE.

I converted the categorical values into the numerical one using the label encoder and normalized the data as numerical values in our feature data set shouldn't be regarded as weights on each attribute.

Finally I split the normalized data into the train set and the test set. I chose to have 20 % of the total data as the test set and the rest as the train set.

3. Methodology

We will use three classification algorithms for our machine learning modeling, k-nearest neighbors, decision trees, and logistic regression. Each modeling process consists of two steps - finding the parameter which gives the highest accuracy and training the model with the obtained parameter.

1. K-Nearest Neighbors

When the k value is provided, it classifies and groups the data within the k nearest neighbor points.

2. Decision Trees

The decision tree model explore all possible outcomes depending on the variables in the feature dataset.

3. Logistic Regression

Logistic regression algorithm is a kind of linear regression and useful for predicting a class from independent variables. As there are only two classes in the target variable, we can use the logistic regression model.

4. Results

1. Evaluation

We will evaluate three parameters, f1 score, jaccard score, and log loss(log loss is only for logistic regression model). From the models we trained in the previous section, we can obtain these parameters as follows:

	KNN	Decision Tree	Logistic Regression
F1 score	0.643094	0.651396	0.633458
Jaccard score	0.643094	0.651396	0.633415
Log Loss	NA	NA	0.635921

Among those models, decision tree algorithm reached the best accuracy with 0.651396.

2. Discussion

From the evaluation, we can see that all three models give the accuracy between 0.63 and 0.66 and there are several things to consider to improve this machine learning modeling.

First, the way I have dealt with the missing values and chosen appropriate attributes would affect the results. I tried another way dropping all missing values in the feature data set other than converting them into ‘Unknown’ values and it gave the differences in the accuracy about 10 %.

Also, we need to take account into the characteristics and usages of algorithms and this will answer why the decision tree algorithm performs better than other algorithms in this project.

5. Conclusion

To predict the severity codes for car collision accidents, I picked three classification algorithms - k-nearest neighbors, decision tree, and logistic regression - for machine learning modeling. From the results, we can see that all three models yielded the accuracy over 0.6 and among them the decision tree performed best with the accuracy of 0.652413.