Car Crash Severity Capstone Project

Desiana Nurchalifah
2020

1. Introduction

With the growth of Artificial Intelligence (AI), especially in autonomous driving, building a system that learns from real life data to ensure safety and operational efficiency of transport systems is a growing interest. A trend where humans act only as supervisor of the automation system was predicted to happen within 10 years by the city governments (Boston Consulting Group, 2015). However, although the automation of a product is a key to the next generation of our economy, the safety of fully automated vehicles is an intricate issue (Luettel, 2012). With high economic impact in the transportation industry, safety in automated driving has become an imperative part as technology driver that applies also for other sectors of application domains.

As it comes with automation, the opening question will be: *How to deploy autonomous driving systems that are safe enough to leave humans out of the system?*. Safety in vehicles does not only mean it obeys the traffic laws, but there are also other factors, such as road hazards, flooding, and weather. Hence, solidifying in finding the severity of possible vehicle crashes to make computer-based automotive systems safe is the biggest challenge for current practice.

Safety introduced is not only in observation of traffic compliance violations, such as driver intoxications, or in a sense, human error, but also environment conditions, such as weather, road conditions, location of accidents, etc. Combination of these variables took part in alerting city infrastructures or business values for upcoming predictions.

The target of this observation is city infrastructure workers, such as the police department, paramedic, government, car insurance company within the Seattle area, and AI companies to integrate the learning and simulation problem of real-life occurrence.

2. Data

Upon further research for collision data, I obtained similar data from Seattle GeoData. The data is an open data from the Seattle government. Data is collected from the year 2004 until present time. The data includes all types of collisions.

The data consists of 40 independent variables as potential features and 221,144 collection of accidents. The dependent variable, **SEVERITYCODE**, contains numbers that correspond to different levels of severity caused by an accident from 0 to 3.

Severity codes are as follows:

- 3: Fatality — High Probability
- 2b: Serious Injury — Mild Probability
- 2: Injury — Low Probability
- 1: Property Damage — Very Low Probability
- 0: Unknown — Little to No Probability

On a side note, these codes are encoded into the range [0-4] with 4 as the high probability of a severe car crash. From observation in the data it is obtained:

```
In [8]: df_seattle['SEVERITYCODE'].value_counts()
Out[8]: 1     137414
        2      58665
        0      21619
        2b      3096
        3        349
        Name: SEVERITYCODE, dtype: int64
```

Labels are uneven, hence stratification, SMOTE (upsampling) and evaluation metrics such as: F1-measure, precision, recall, and Logloss are used to measure performance. Confusion matrix is also used to analyze better performance.

3. Feature Selection

Data needs to be preprocessed, such as removal/fill in null values and encoding of features. Possible features obtained from the environment are as follows:

- Weather
- Road
- Light
- Junction Type

To observe prominency in automation sector, we observe human factors that contributes the severity such as:

- Intoxication
- Attention
- Number of persons included

As the severity of collisions could affect the situation in the neighborhood, that leads to potential traffic, factors below are also observed:

- Collision description
- Collision state
- Hit of cars

As data is noisy, we need to preprocess the data first. Converting objects into values and operation to both NaN and empty values. In this project, NaN is removed from the target and all features associated with the particular NaN target as it does not help with classification.

Otherwise, NaNs in other features are marked as -1. The reason for this is to observe the whole possibility of features to be observed.

4. Exploratory Data Analysis

Feature are observed among variables using correlation matrix as follows:

	WEATHER	ROADCOND	LIGHTCOND	SDOT_COLCODE	PERSONCOUNT	JUNCTIONTYPE	INATTENTIONIND	UNDERINFL	ST_COLCODE	HITPARKEDCAR	SEVERITYCODE
WEATHER	1	0.754008	0.402468	-0.216594	-0.266384	0.0469144	-0.135614	-0.437789	-0.163403	0.153363	-0.358968
ROADCOND	0.754008	1	0.398593	-0.201116	-0.254065	0.0370012	-0.141003	-0.407452	-0.126886	0.147789	-0.343946
LIGHTCOND	0.402468	0.398593	1	-0.10913	-0.180661	-0.000788514	-0.107549	-0.231688	-0.0451975	0.115355	-0.232244
SDOT_COLCODE	-0.216594	-0.201116	-0.10913	1	0.00832317	-0.179434	0.055654	0.260193	0.332624	-0.158463	0.311602
PERSONCOUNT	-0.266384	-0.254065	-0.180661	0.00832317	1	-0.0323729	0.125228	0.383241	0.0520425	-0.11906	0.370575
JUNCTIONTYPE	0.0469144	0.0370012	-0.000788514	-0.179434	-0.0323729	1	-0.0139191	-0.0794896	-0.0981703	-0.00951596	-0.0306037
INATTENTIONIND	-0.135614	-0.141003	-0.107549	0.055654	0.125228	-0.0139191	1	0.111212	0.033378	-0.0215269	0.107642
UNDERINFL	-0.437789	-0.407452	-0.231688	0.260193	0.383241	-0.0794896	0.111212	1	0.462324	-0.175261	0.518683
ST_COLCODE	-0.163403	-0.126886	-0.0451975	0.332624	0.0520425	-0.0981703	0.033378	0.462324	1	-0.0219944	0.179909
HITPARKEDCAR	0.153363	0.147789	0.115355	-0.158463	-0.11906	-0.00951596	-0.0215269	-0.175261	-0.0219944	1	-0.201678
SEVERITYCODE	-0.358968	-0.343946	-0.232244	0.311602	0.370575	-0.0306037	0.107642	0.518683	0.179909	-0.201678	1

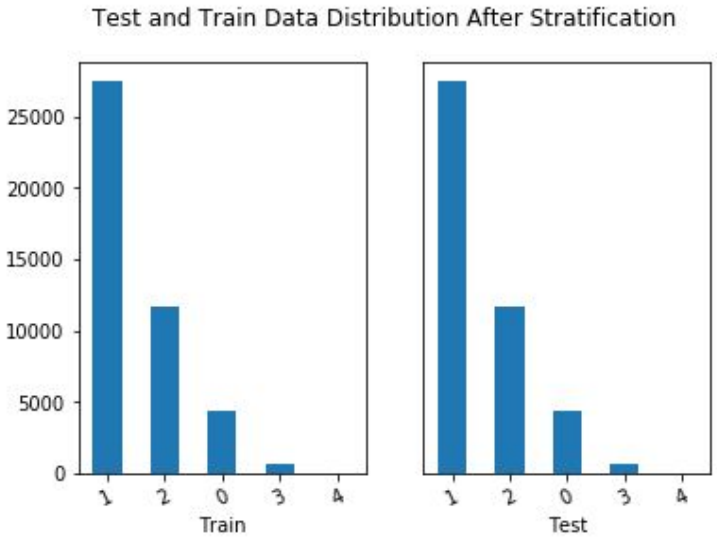
From all the features included, surprisingly the type of the junction is not that prominent with the correlation to the severity of a crash. Among 10 features included, the type of junction is the lowest prominent, followed by attention of the driver to the road. On the other hand, intoxication to the driver has the biggest correlation to the severity of a crash. If this is reported to the stakeholders with the intention of developing Safe Autonomous Driving, this could act as a base to strengthen the business value of autonomous vehicles.

5. Building Models

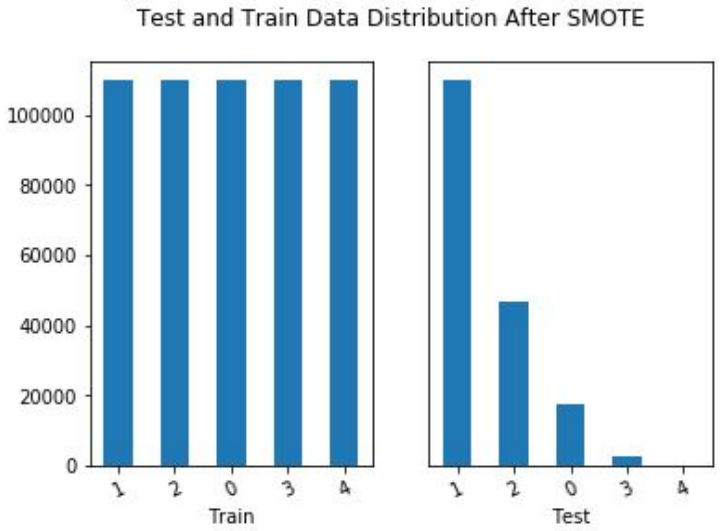
In this project we will normalize the data. Data standardization or normalization give data zero mean and unit variance. In the first section, data has been preprocessed and features are extracted from the data according to the perspective of the stakeholders. We will then build our machine learning models based on the features. As data is imbalanced, as checked in the first section, we will stratify the data such that division of classes for training and testing encompass all the classes evenly. Here three machine learning models are compared:

- K-Nearest Neighbor
- Decision Tree
- Logistic Regression

Below is the result of train and test data stratification to even out distributions:

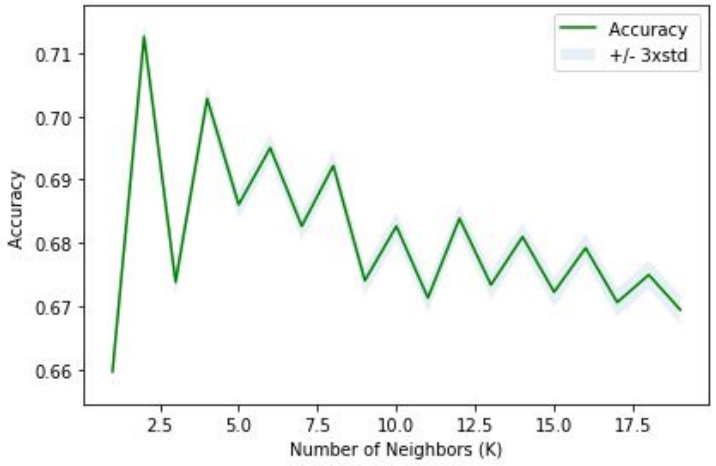


After making train and test dataset evenly distributed each class, we apply SMOTE to upsample the training data in order to get a higher level of fairness among classification. Below is the distribution after SMOTE:



In building KNN model, below is the result:

```
Best accuracy for KNN:  0.712607054888934  k:  2
```



While Decision Tree and Logistic Regression construction are respectively:

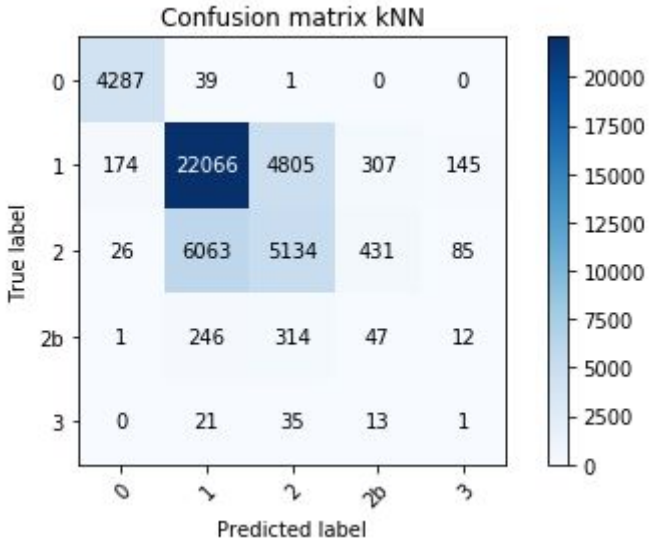
```
Best accuracy for Decision Tree:  0.6552098162836418  depth:  5
Logistic Regression's best acc:  0.5150611257993808  Regularization Values:  1
```

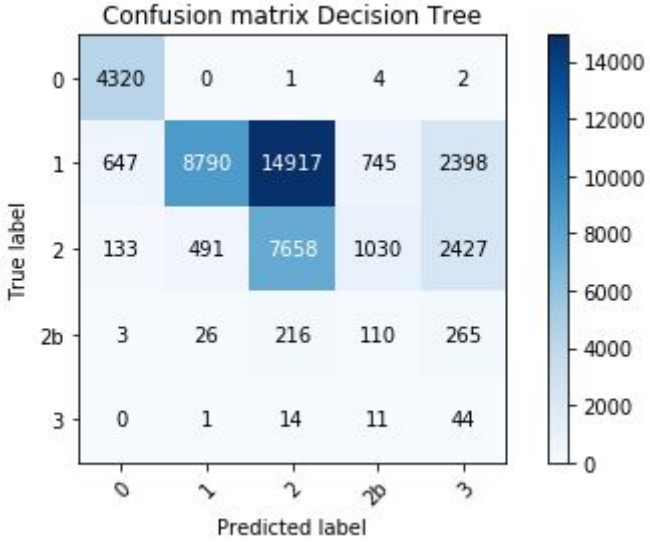
All models are previously searched in the space of [1,20] for k, [1,15] for depth and [0.001,0.01,0.1,1,10,100] for regression in logistic regression.

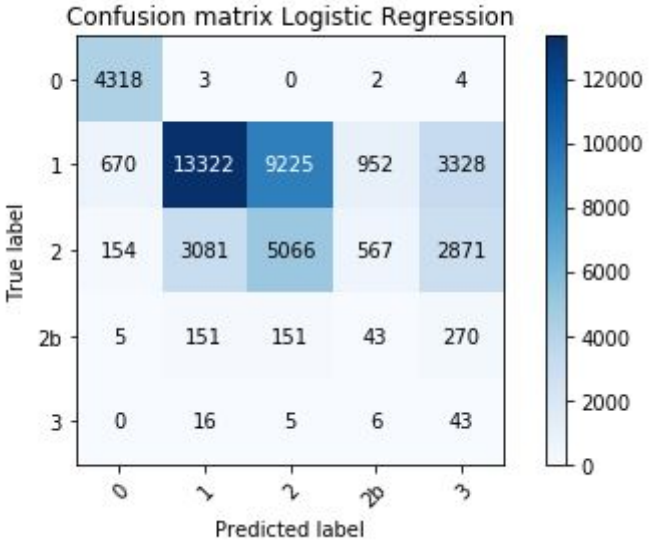
6. Results

	Algorithm	F1-score	LogLoss	Precision	Recall
0	KNN	0.460101	NA	0.458644	0.464136
1	Decision Tree	0.388289	NA	0.438632	0.55528
2	Logistic Regression	0.391077	1.12381	0.405484	0.519521

In the results, it shows that among three machine learning methods, KNN excels other methods with only a small difference in recall. Although KNN provides the best performance evaluation, parameter tuning in KNN is computationally exhaustive. In this experiment it is not proven the higher K is, the more probability that the model overfits the data, this could happen in defiance to the data being imbalanced. Improvement could be done by applying cross-validation to the test data, as the real test data would also contain imbalanced distribution. In using decision trees, the maximum depth is 5 for the best accuracy performance on training data. Hence this value will be used for testing. For logistic regression, the regularization factor that built maximum performance is when C=1. Confusion matrices of each model are given as follows:







Among all the ML methods, kNN is the safest approach as prediction using either Decision Tree or Logistic Regression has higher value in mis-classifying high probability of severe accident (type 3) with low probability accidents (type 1 and type 2). The highest total of true label which classified correctly is also achieved by kNN, followed by Logistic Regression and Decision Tree respectively.

7. Conclusion

Based on Seattle GeoData for car collisions, among 40 attributes which have a potential value in predicting car crash severity, human intoxication factor is still the highest feature importance that takes part in severity of car accidents. Among environment conditions, weather is imperative to contribute in car collisions, while location of the collision, such as junctions, are exceptionally insignificant. To predict whether collision will affect traffic, with the state of property damage to severity that leads to fatal injuries of the accident, it could be classified effectively using kNN, which had performed better than Decision Tree or Logistic Regression in comparison using 4 evaluation metrics. In the future, it is advisable to use PCA to obtain the number of features without having to hand-craft feature parameters. Approach can also be extended using Stratified K-fold.