Predicting Severity Codes of Car Collision Accidents

How we can predict the severity of car collision accidents Using machine learning algorithms

October 30, 2020

Background / Business Understanding

- Car collisions have been huge problems to our society thus it is important to analyze the historical data and predict before they happen
- Our primary goal is to **build an appropriate machine learning model and predict the severity code**, which is categorized as follows:
  - 0 : Unknown
  - 1 : Property Damage
  - 2 : Injury
  - 2b : Serious Injury
  - 3 : Fatality

Data Acquisition

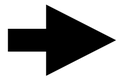
- The data of all collisions in Seattle from 2004 to present has been used
- The raw data have 35 attributes in total including SEVERITYCODE and I chose to have only 15 attributes for further data preparation
- Following is the first five rows of the data:

	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

Data Preparation

- Balance the data by downsampling
  - As the classes of SEVERITYCODE has different size we need resampling:

```
1    136485
2     58188
Name: SEVERITYCODE, dtype: int64
```



```
2    58188
1    58188
Name: SEVERITYCODE, dtype: int64
```

- Fill up the missing values and convert data types
- Create the feature dataset X and target y
  - X : ADDRTYPE, WEATHER, ROADCOND, LIGHTCOND, PERSONCOUNT, PEDCOUNT
  - y : SEVERITYCODE
- Normalize the feature dataset and split the data into train/test set

Methodology / Evaluation

- Three classification algorithms - **K-Nearest Neighbors, Decision Tree, and Logistic Regression** - has been used
- To compare the accuracy of each model, I have calculated **f1 score and jaccard score** (and log loss only for logistic regression regression)
- All three models give the accuracy between 0.63 and 0.66 and among those models the decision tree algorithm reached the best accuracy

	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

Conclusion

- All three models yielded the accuracy over 0.6 and among them **the decision tree model performed best with the accuracy of 0.652413**
- There are several things to consider to improve this modeling:
  - How to choose appropriate attributes and deal with the missing values
  - The characteristics and usages of algorithms