PREDICTING SEVERITY OF COLLISIONS

By Nguyễn Tường Quang

September, 2020

---o0o---

Table of contents:

1. Introduction

Traffic accident is something no one wants, but also a part of modern life. But, can traffic accident be predictable?

Actually, traffic accidents do not happen naturally, but have causes. If an accident can be predicted, from the initially assumed causes, we can adjust everything related to reduce the accident. Specifically, this report will be targeted to stakeholders to reduce crashes.

And the goal of this project is to forecast the severity of traffic accedents.

Under the aspect of data science, we should must collect data related to vehicle accidents in history, as much as possible, and then analyze the data to build a suitable predictive model.

If it is possible to build a good predictive model, that is, it is possible to establish a link between the input causes and the severity of the accident, we will be able to predict the accidents in future.

2. Analytic approach

With the goal is to forecast the severity of collisions, this is clearly the classification of damage caused by a traffic accident.The levels of damage can be classified into several levels, such as: property damage, human injured, or death. In this project, we only divided into two levels: level 1- only property damage that does not injure (prop damage), level 2- that hurts people (injury).

From that purpose, we can choose familiar classification algorithms such as: Naive Bayes, Logistic regression, Decision tree, K nearest neighbors, super vector machine, ... But for the purpose of only dividing into 2 levels, the binary type, we prefer to choose: Logistic regression and Decision tree.

3. Data

3.1 Data requirements

The prerequisite must be real data, recording many aspects related to accidents, such as location, date and time, road conditions, weather conditions, lighting conditions, number of people involved in an accident, number of vehicles and type of vehicle crashed ... And, ultimately and most importantly, the consequences of the accident. The aftermath of an accident must be fully recorded on the damage to property as well as human life. This is the target of classification in this project.

The next condition is that the data must be large enough, preferably recorded for many consecutive years in the same region or city, country ...

The third condition is that data must be recorded in a structured and clear way.

3.2 Data sourses

There are many datasets that meet the above requirements. For this project, we used the City of Seattle dataset, recorded from 2004 to 4/2020. This dataset is shared by IBM Data Science course.

3.3 Data descriptions

The dataset consists of nearly two hundred thousand rows, with 38 columns. In which, the first column is about the severity of the collision, which is the target for classification in the model. Consists of two serious levels: 1-prop damage and 2-injury.

The remaining 37 columns, in addition to the separate record columns of the state, are the remaining data that can be used as inputs to build the model. These columns are entirely categorical variables. Include:

- Type of collision

- Total number of people involved in a collision

- Number of pedestrians involved in a collision

- Number of cyclists involved in a collision

- Number of vehicles involved in a collision

- Traffic configuration where collision occurred (junctiontype)

- Weather conditions at the time of the collision

- Road conditions at the time of the collision

- Light conditions at the time of the collision

- The date and time of the collision

Which features to choose as input for the predictive model will depend on the relationship between that feature and the target (i.e the severity of the collision). This will be achieved after exploring data analysis and preprocessing step.

4. Data analysis: exploring and preprocessing

4.1 Target

The target is the severitycode column in the data. It includes two classes: 1- property damage, 2- human injury. Encoded as 1 and 2.

- Class 1: about 70%
- Class 2: about 30%

4.2 Number of persons, pedestrians, bicycles and vehicles involved in the collision

- Mean number of persons involved in a collision:
  Class 1: 2.329348
  Class 2: 2.714357

- Mean number of pedestrians involved in a collision:
  Class 1: 0.005268
  Class 2: 0.111896

- Mean number of bicycles involved in a collision:
  Class 1: 0.004975
  Class 2: 0.083316

- Mean number of vehicles involved in a collision:
  Class 1: 1.943312
  Class 2: 1.867928

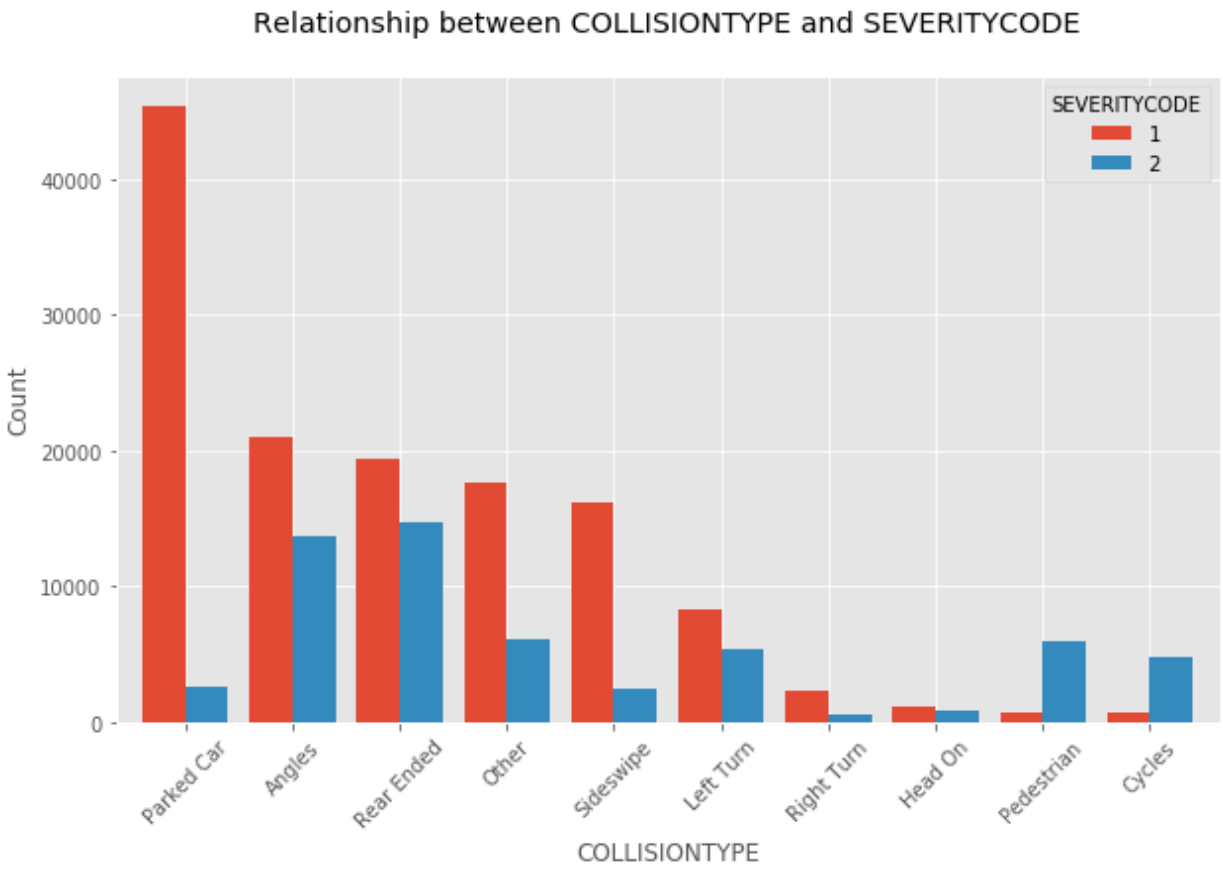
Discussion:

Number of pedestrians and bicycles are too small to use for predicting.

4.3 Relationship between type of collision with severity of collision

There are many types of collisions that are classified. Looking at the plot below can see the impact of each type on the severity. The collisions that occur most with parked vehicles, followed by at corners and hitting the back of the vehicle. Parking arrangements are also a problem in modern cities.

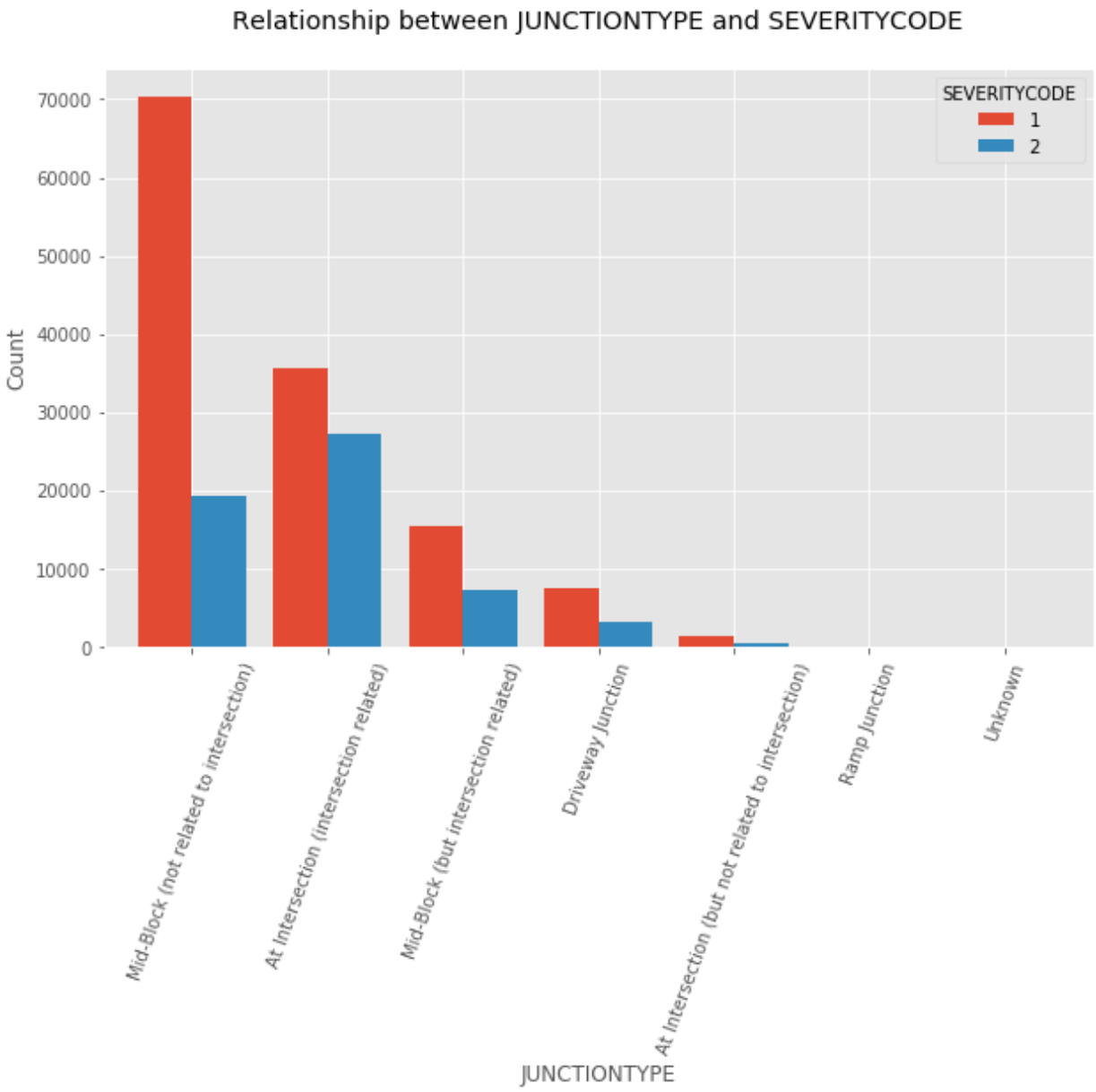
Figure 1:



4.4 Relationship between JUNCTIONTYPE with severity of collision

Looking at the plot we can see that the traffic configuration greatly affects the collisions. Building blocks, intersections, all have the potential to contribute to collisions.

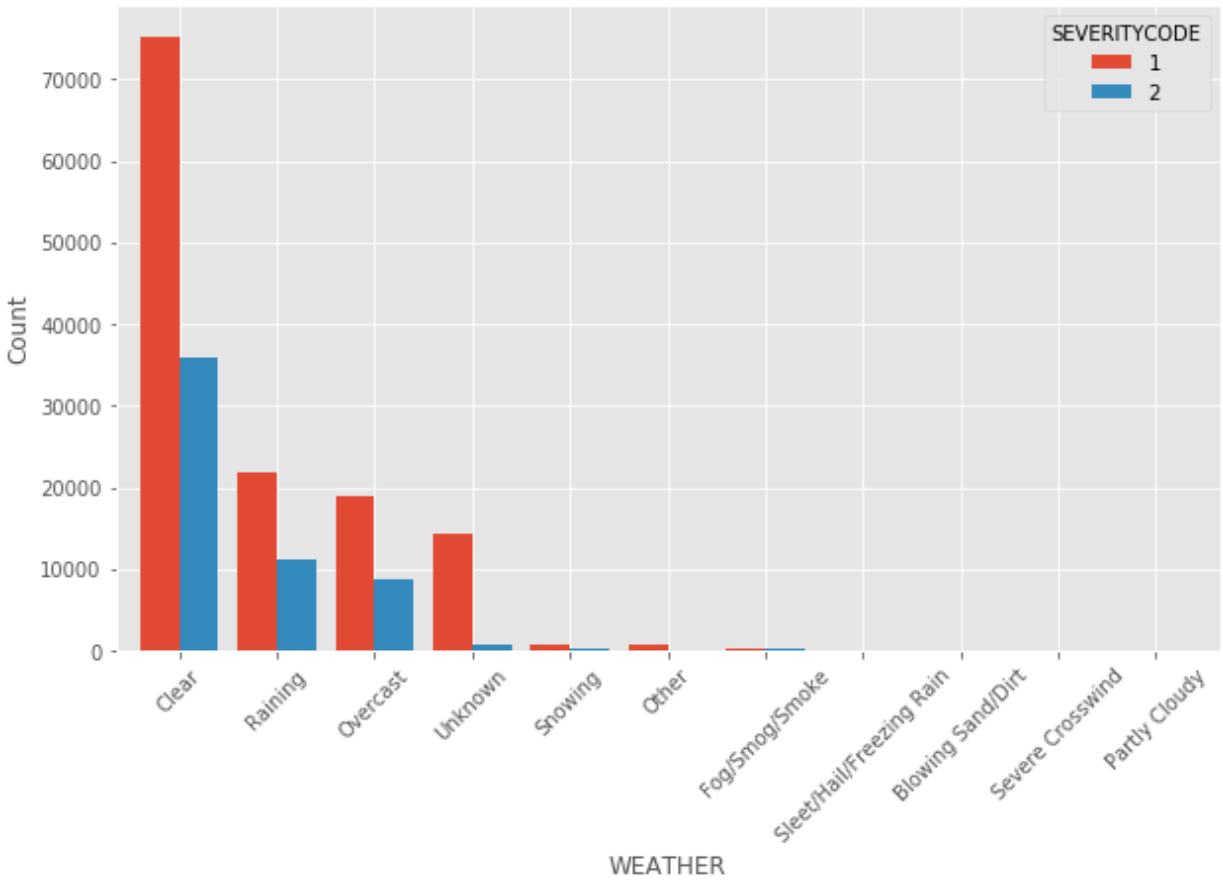
Figure 2:



4.5 Relationship between WEATHER with severity of collision

Intuitively can also predict the weather will affect the cause of the collision. Amazingly, the clear weather causes the most accidents, three times more than it rains. While it is snowing or foggy, it causes very few traffic accidents.

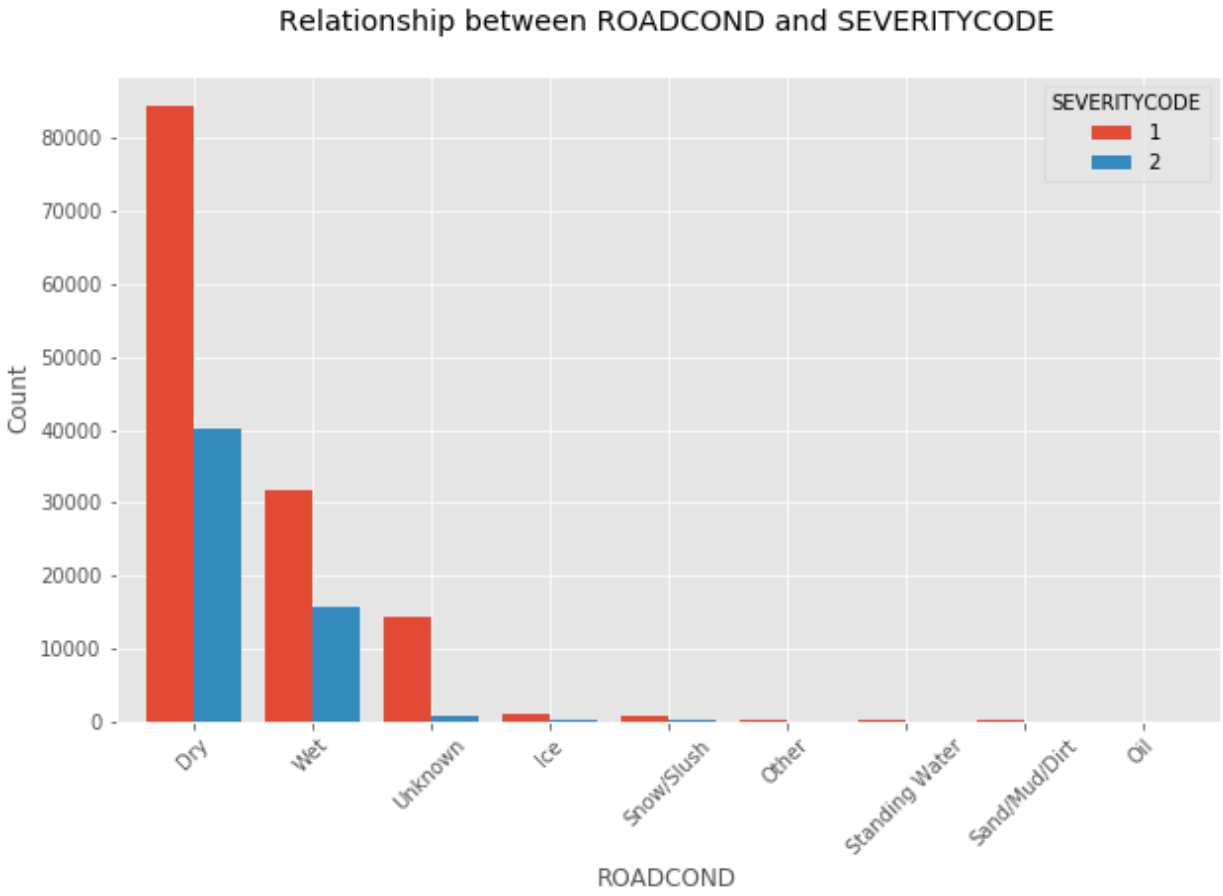
Figure 3:
Relationship between WEATHER and SEVERITYCODE



4.6 Relationship between the condition of the road with severity of collision

Road conditions, of course, have a decisive influence on traffic. But surprisingly when dry roads cause the most traffic accidents, almost 3 times when wet roads. While the road is icy or snowy, it causes very little accidents.

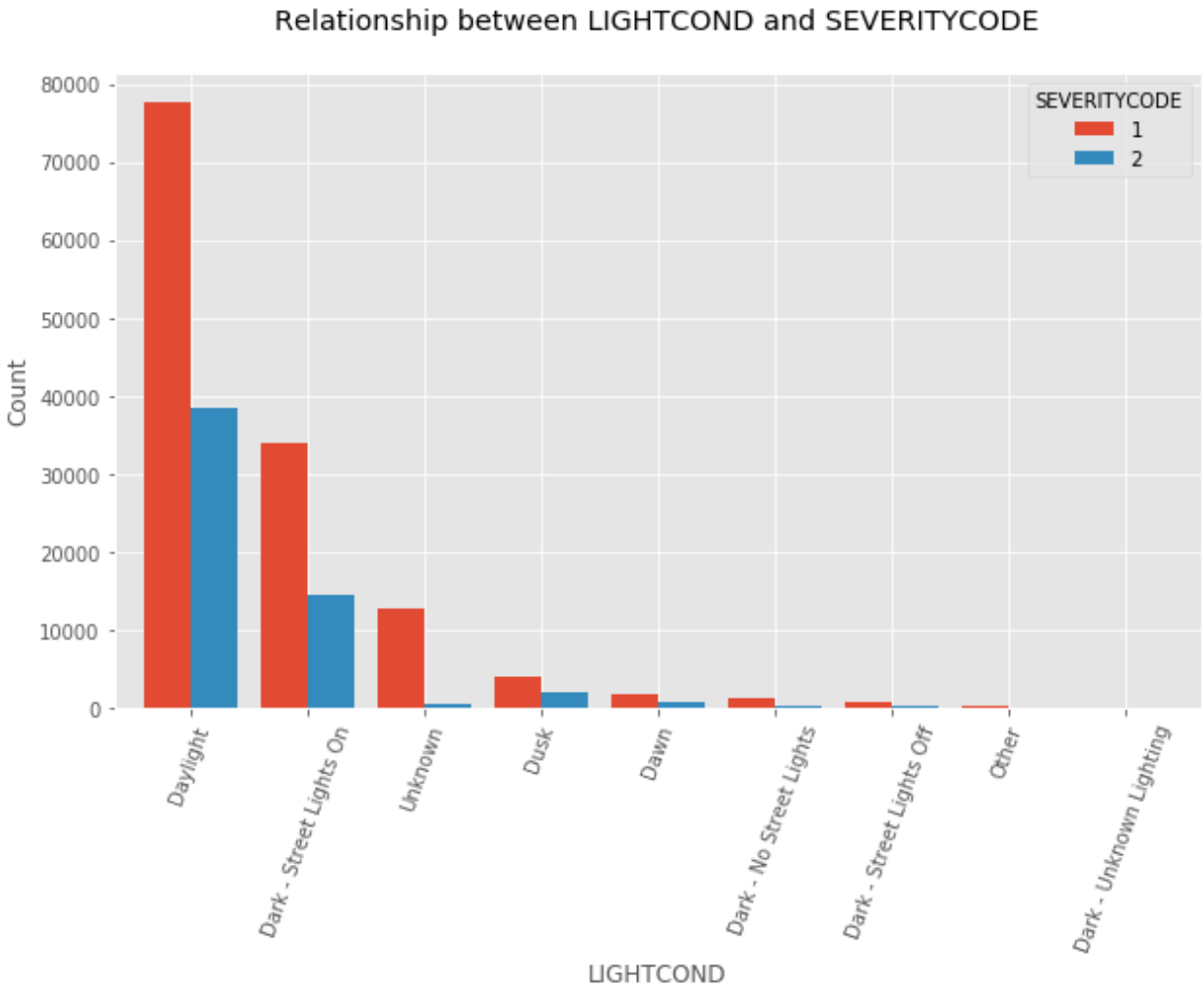
Figure 4:



4.7 Relationship between the light conditions with severity of collision

Most accidents happen during the day, twice as much at night. This is probably related to daytime traffic rather than night.

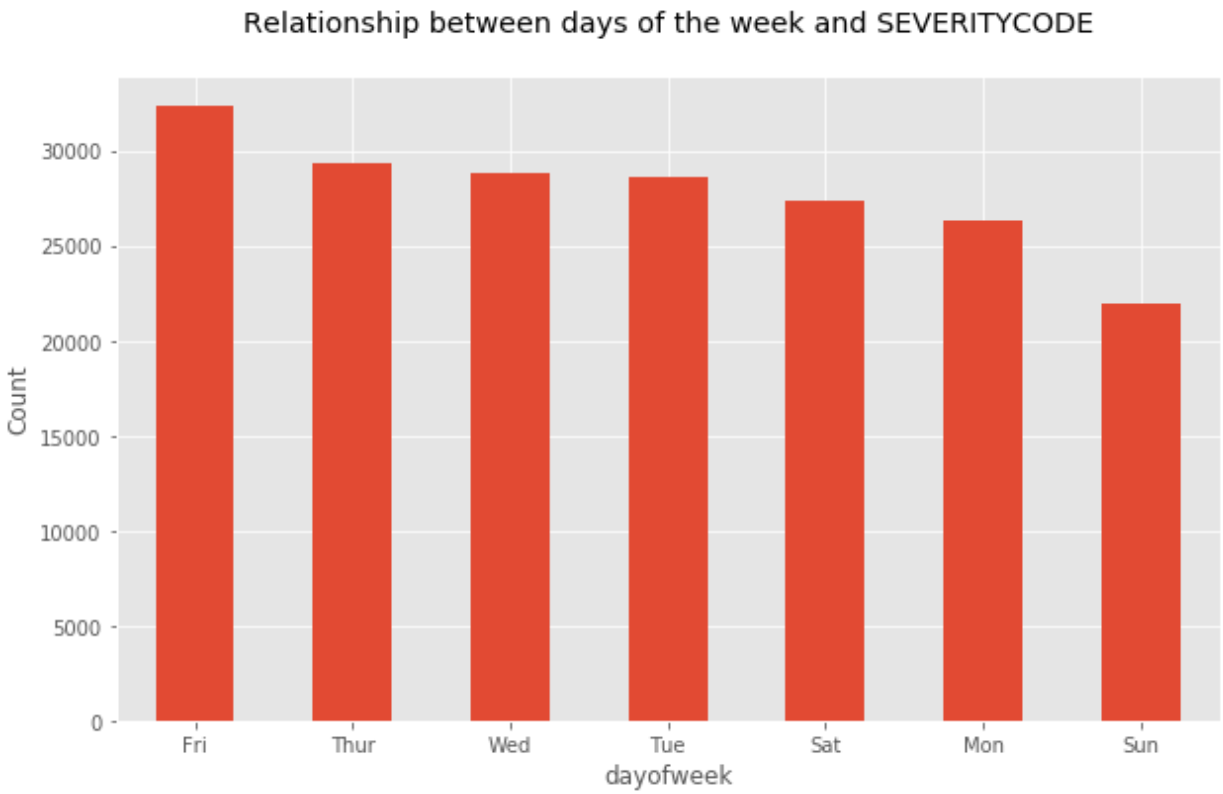
Figure 5:



4.8 The date in collisions, related to the severity

Modern societies move by day of the week. Usually we go to work Monday through Friday, rest on Saturday and Sunday. Does this have an effect on the severity of the collisions?
Here we survey by day of the week.
From the plot we can see, collisions happen most on Fridays, and at least on Sundays.

Figure 6:



4.9 The time problem in the dataset

The time data of the collisions are very important in building predictive models. This dataset has a timed column but is incomplete. Thus, to use time as input, we have to filter and remove rows that have no time. But this can cause a lot of data loss, affecting the prediction results.

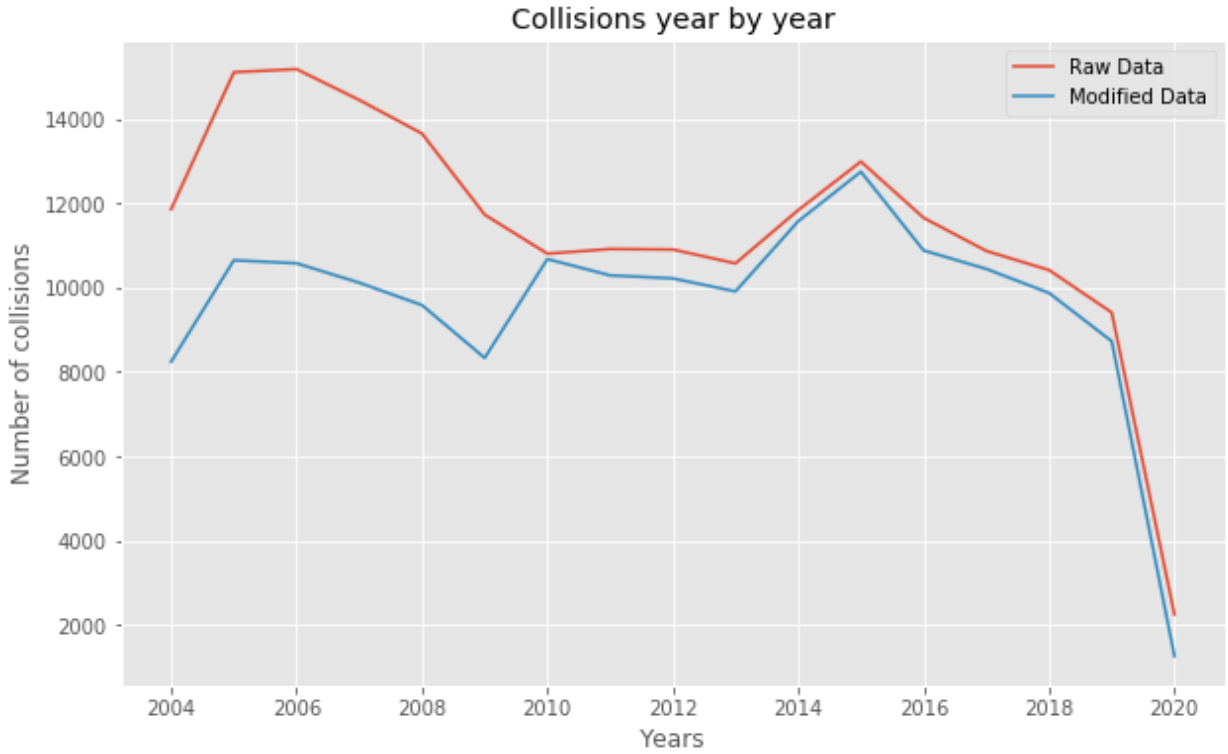
Below we will explore the collisions year by year in the case of keeping the raw data and only the timed rows.

Look at the plot below, from 2004 to 2019:

The red line is the original data, tends to decrease year by year. While the blue line, the timed data, tends to flat. It means that filtering and removing data have falsified reality.

Because of this bias, we decided not to use the hour data as an input to build a predictive model.

Figure 7:



4.10 Select features and preparing data for predicting model

- Select features: SEVERITYCODE, PERSONCOUNT, VEHCOUNT, COLLISIONTYPE, JUNCTIONTYPE, WEATHER, ROADCOND, LIGHTCOND and dayofweek

- The features are removed all missing values.
- Except SEVERITYCODE (target), PERSONCOUNT and VEHCOUNT, all the rest got one - hot encoding.
- The data is splited into 2 dataset: y-SEVERITYCODE, X-all the rest features.
- X is normalized by StandardScaler.

5. Modeling & Evaluation model

Two algorithms are selected are: Decision Tree and Logistic Regression.

The data is splited in two parts: trainset & testset with rate 0.25 (25%) for testset.

Evaluation model by jacard, f1-score and logloss. Deeper evaluation by confusion matrix:

- Results:

No	Algorithms	Jacard	F1-score	Logloss
1	Decision Tree	0.748772	0.709571	NA
2	Logistic Regression	0.747549	0.706120	0.495706

- Results of confusion matrix:

Figure 8: confusion matrix for Decision Tree

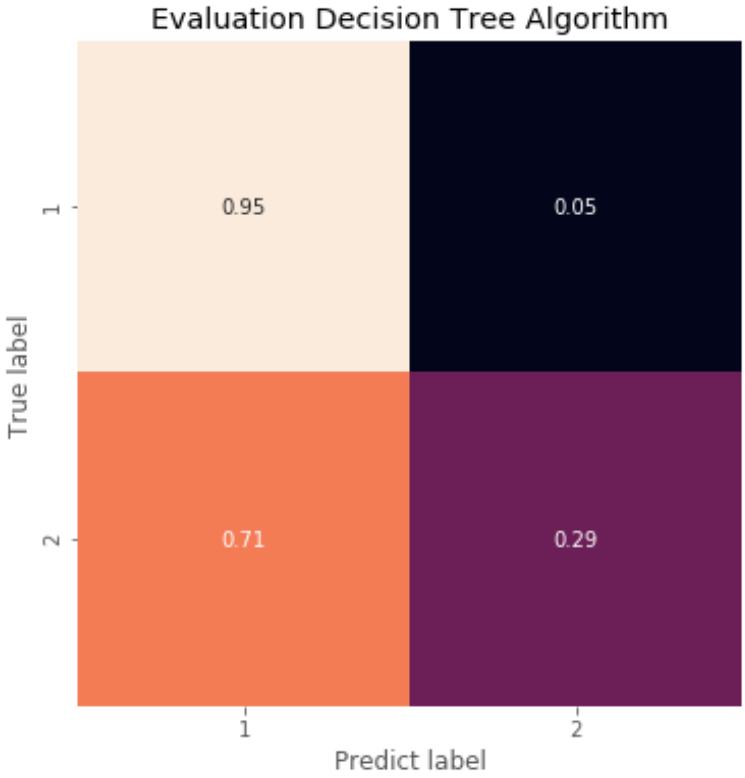
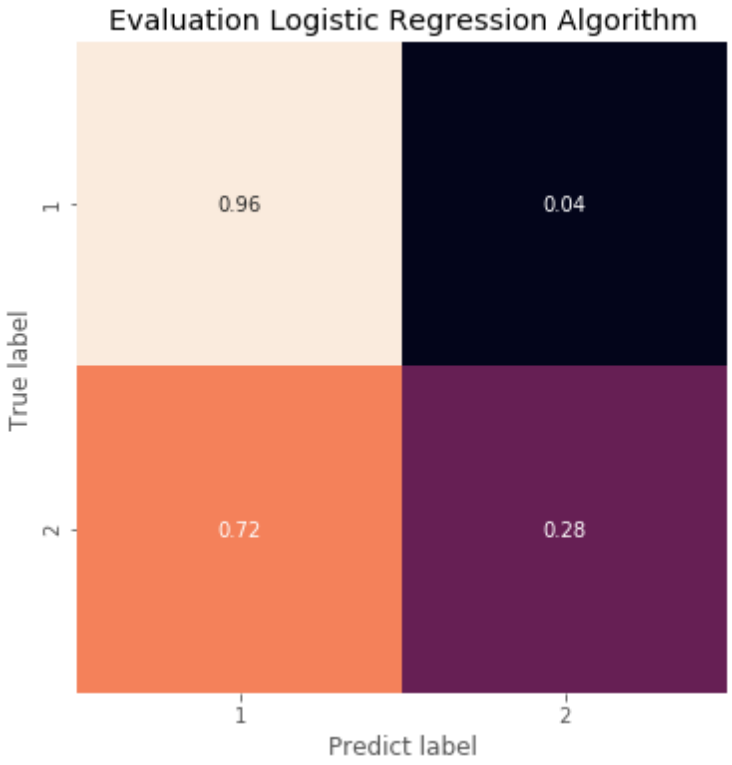


Figure 9: confusion matrix for Logistic Regression



6. Discussion

The results of the two algorithms are approximately. This shows the reliability of the predictive model. With an overall accuracy of nearly 75% is acceptable. However, the difference in accuracy in the two classes is very high.

With class 1 (prop damage), both methods are very high results, up to 95% and 96%. However, the opposite results in class 2 (injury), the results predicted by both methods are very low (only 29% and 28%).

This restriction is due to the input data. Presumably, the input data lacked the features by which traffic accident severity could be more clearly distinguished. For example, the speed of a vehicle before the time of a collision. The vehicle speed may be related to the speed specified in the traffic system. This data is available in the traffic monitoring system but has not been updated. Obviously, the higher the speed, the higher the risk of human injury.

7. Conclusion

With the above results, it can be seen as a source of reference for stakeholders. However, in future, if the input data can be improved, the accuracy will definitely be higher. And of course, forecasts are also more useful.

---o0o---