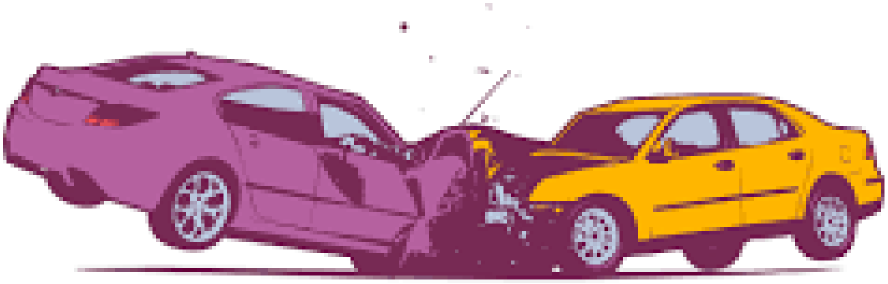
IBM – Coursera

IBM Data Science Professional Certificate

Capstone project – week2 report

SUBJECT: CAR ACCIDENT SEVERITY



Trung Kien NGUYEN – August 2020

1. Introduction

The IBM Data Science Professional certification course on Coursera concludes with a Capstone Project. This project is about using data science tool-set on a real-life problem: the traffic collision. I present here the summary of my project. The analysis was performed in Python. The details analysis is specified in the Jupyter notebook on Github.

2. Business problem

For this project, I choose a hypothetical business problem.

I'm driving to another city to visit some friends. It is rainy and windy, and on the way, I come across a terrible traffic jam on the other side of the highway. Long lines of cars barely moving and the police shutting down the highway. It is a terrible accident and they must be in critical condition for all of this to be happening. Hence, it will great if there is something in place that could warn you. By giving the weather information and the road conditions, we need to know the possibility of getting into a car accident and how severe it would be.

The assumption behind the analysis is that we can use supervised machine learning to predict the severity based on historical traffic collision data, whose severity was considered by the police. With this model, we can develop an mobile application to warn the driver in real-time.

3. Data

To perform this analysis, we need the following data:

- Accident location
- Road conditions
- Weather condition
- Junction
- Car speeding
- Number of people involved
- Light conditions
- Number of vehicles involved in
- The current date and current time

These information can be obtained from Seattle Department of Transportation (SDOT). SDOT has an open data platform which can be found in "https://data.seattle.gov/". In this platform, they update their information about collisions weekly. We can find all information we need in this dataset. The attribute information details can be found in "https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf".

4. Methodology

4.1. Exploratory data

The data set contains 38 columns:

- SEVERITYCODE: A code that corresponds to the severity of the collision.
- ***X: The longitude***
- ***Y: The latitude***
- OBJECTID: ESRI unique identifier
- INCKEY: A unique key for the incident
- COLDETKEY: Secondary key for the incident
- REPORTNO: Report number
- STATUS: Match or unmatch
- ADDRTYPE: Collision address type: Alley, Block, Intersection
- INTKEY: Key that corresponds to the intersection associated with a collision
- LOCATION: Description of the general location of the collision.
- EXCEPTRSNCODE: Enough or not enough information
- EXCEPTRSNDESC: Enough or not enough information
- SEVERITYCODE: *Duplicated column*
- SEVERITYDESC: A detailed description of the severity of the collision
- COLLISIONTYPE: Collision type
- ***PERSONCOUNT: The total number of people involved in the collision***
- PEDCOUNT: The number of pedestrians involved in the collision. This is entered by the state.
- PEDCYLCOUNT: The number of bicycles involved in the collision. This is entered by the state.
- ***VEHCOUNT: The number of vehicles involved in the collision. This is entered by the state.***
- ***INCDATE: The date of the incident.***
- ***INCDTTM: The date and time of the incident.***
- ***JUNCTIONTYPE: Category of junction at which collision took place***
- SDOT_COLCODE: A code given to the collision by SDOT.
- SDOT_COLDESC: A description of the collision corresponding to the collision code.
- INATTENTIONIND: Whether or not collision was due to inattention. (Y/N)
- UNDERINFL: Whether or not a driver involved was under the influence of drugs or alcohol.

- ***WEATHER: A description of the weather conditions during the time of the collision.***
- ***ROADCOND: The condition of the road during the collision.***
- ***LIGHTCOND: The condition of the road during the collision.***
- PEDROWNOTGRNT: Whether or not the pedestrian right of way was not granted. (Y/N)
- SDOTCOLNUM: A number given to the collision by SDOT.
- ***SPEEDING: Whether or not speeding was a factor in the collision. (Y/N)***
- ST_COLCODE: A code provided by the state that describes the collision.
- ST_COLDESC: A description that corresponds to the state’s coding designation.
- SEGLANEKEY: A key for the lane segment in which the collision occurred.
- CROSSWALKKEY: A key for the crosswalk at which the collision occurred.
- HITPARKEDCAR: Whether or not the collision involved hitting a parked car. (Y/N)

Note that our purpose is predicting the severity if an accident occurs to warn the driver. Hence, we can’t use the consequence columns such as SEGLANEKEY, CROSSWALKKEY, HITPARKEDCAR. Etc. We choose the bold columns to perform the prediction.

There are a lot of missing data such as in column "X", "Y", "ADDRTYPE", "INTKEY", "LOCATION", "EXCEPTRSNCODE", "EXCEPTRSNDESC", "COLLISIONTYPE", etc.

Based on our purpose, the problem is a classification problem. We will use the historical data to predict the label (severity) of incoming data. At the end of this section, we just see the data and data type. We also choose the necessary data and see if there is missing data. In the next section, we will do the data visualization and statistics.

4.2. Data visualization and statistics

The data set is imbalanced with 70% data with “1” label and 30% data with “2” label. It contains only 2 type of severity: "1" (prop damage) and "2" (injury). It will limit the prediction because the classification can not perform with the label which doesn't exist in data set such as "3" (fatality), "2b" (serious injury) and "0" (unknown).

Before do the data visualization, we need to work with missing data. The location data (longitude and latitude) is an importance information, hence, we drop all row with missing location data. The “SPEEDING” column is set to “Y” if the accident relates to speed problem, so, we replace all missing data to “0” and “Y” to 1.

For other columns, there are about 5012 rows missing data, which is very small if we compare with data set (194673 rows), we can drop all these rows.

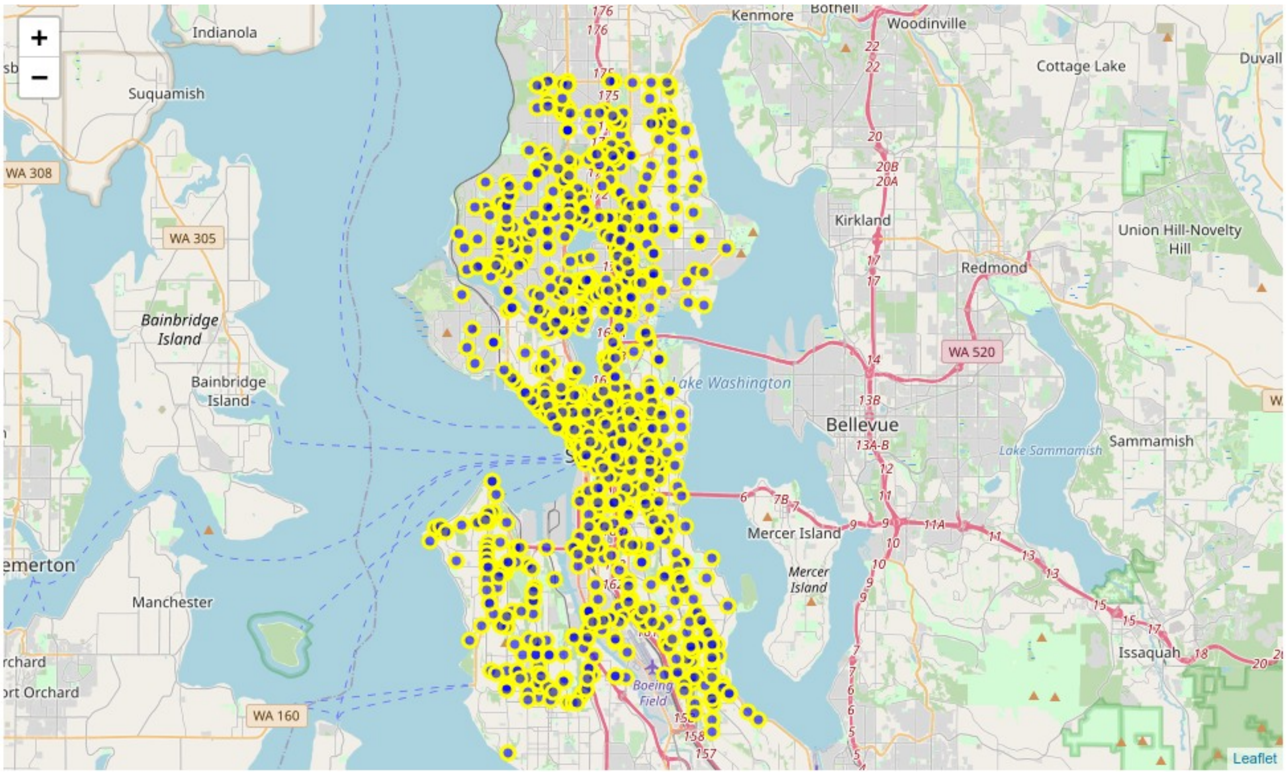


Figure 1. The first 1000 accidents location.

The data set is large to load all the data into map, so, I choose first 1000 points to display on the map (Figure 1). It seems that the Seattle Center have higher number of accidents.

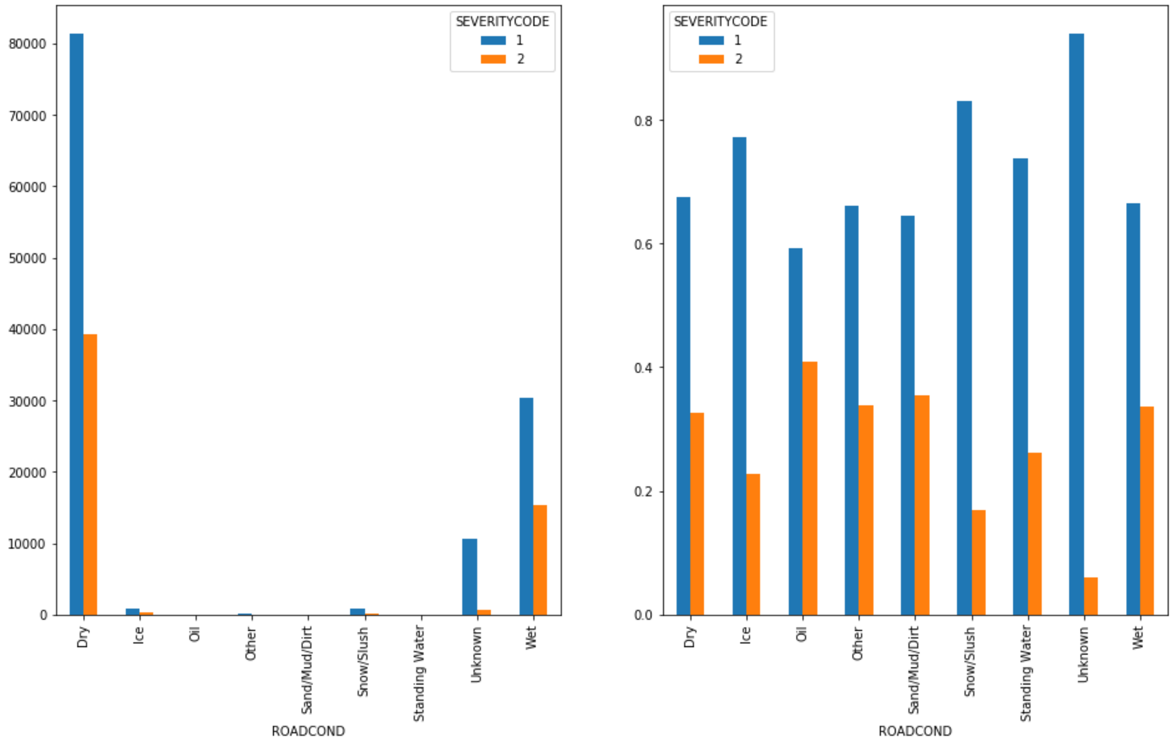


Figure 2. The distribution of severity based on road condition.

According the figure 2, we can conclude that almost data in the road condition set (Dry, Wet, Unknown) and when an accident occurs, the probability of severity "1" is more than the probability severity "2". But we can not say the impact of road condition on the probability.

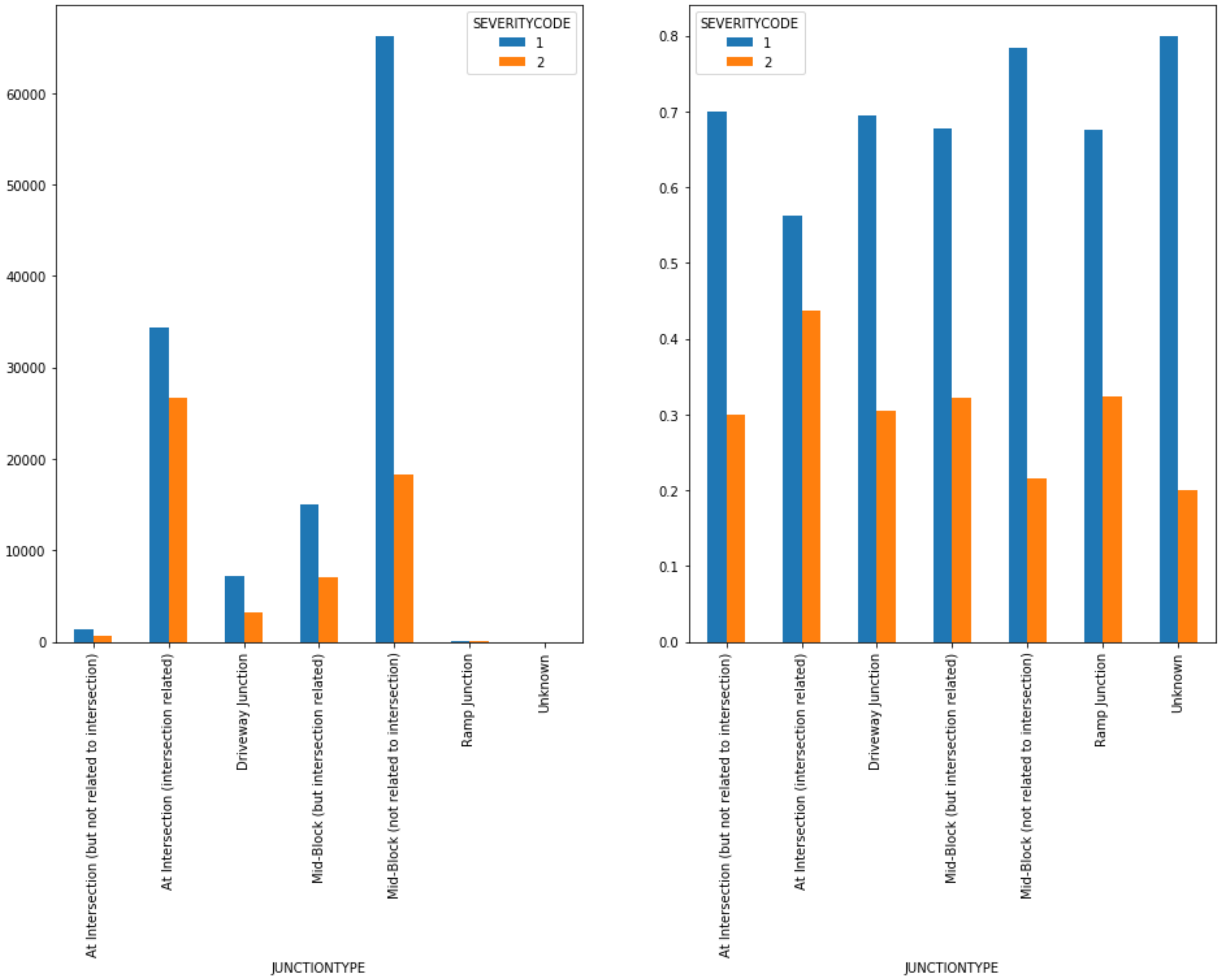


Figure 3. The distribution of severity based on junction.

According to figure 3, almost junction type data is in the set (At intersection, driveway junction, mid-block). When the accident occurs at intersection (intersection related), so, the higher probability the accident is more dangerous.

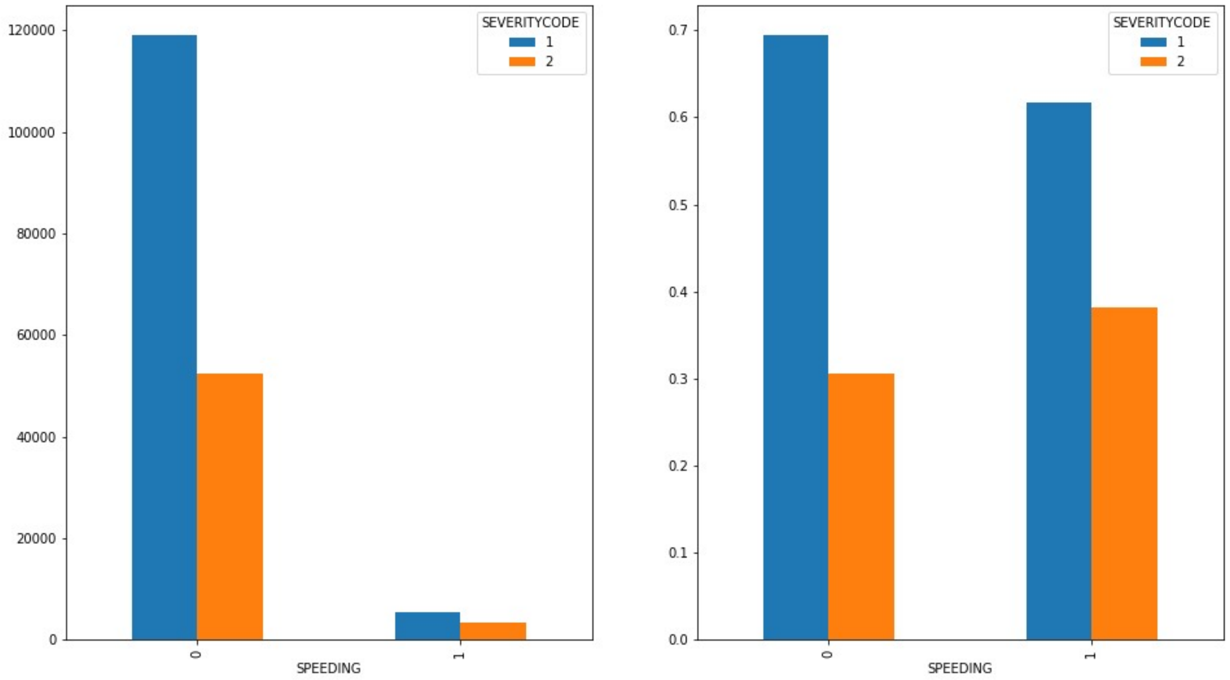


Figure 4. The distribution of severity based on speed.

According to figure 4, almost accidents does not relates to speed, but when the accident concerns about speed, it will have higher severity.

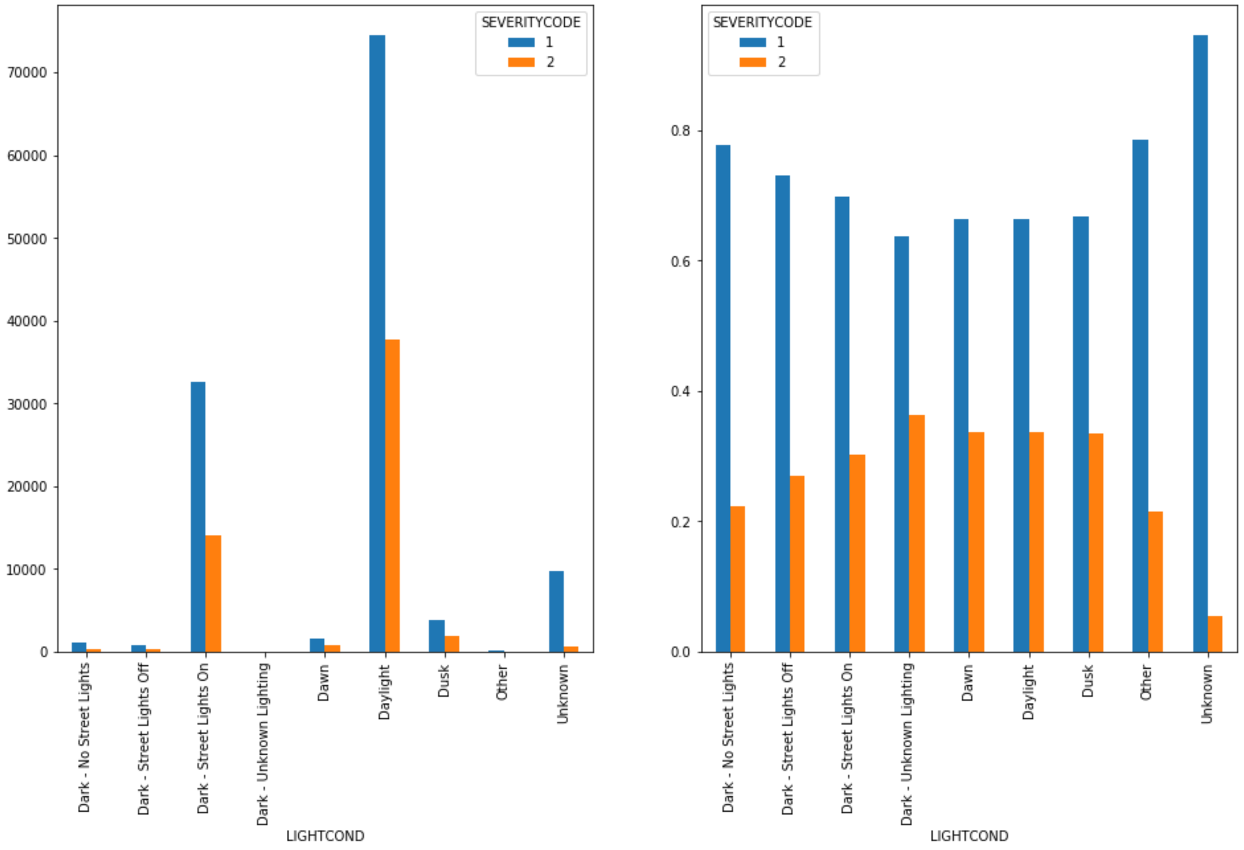


Figure 5. The distribution of severity based on light condition.

We can see that almost the accidents occur when the light is normal (Street lights On and Daylight) in figure 5.

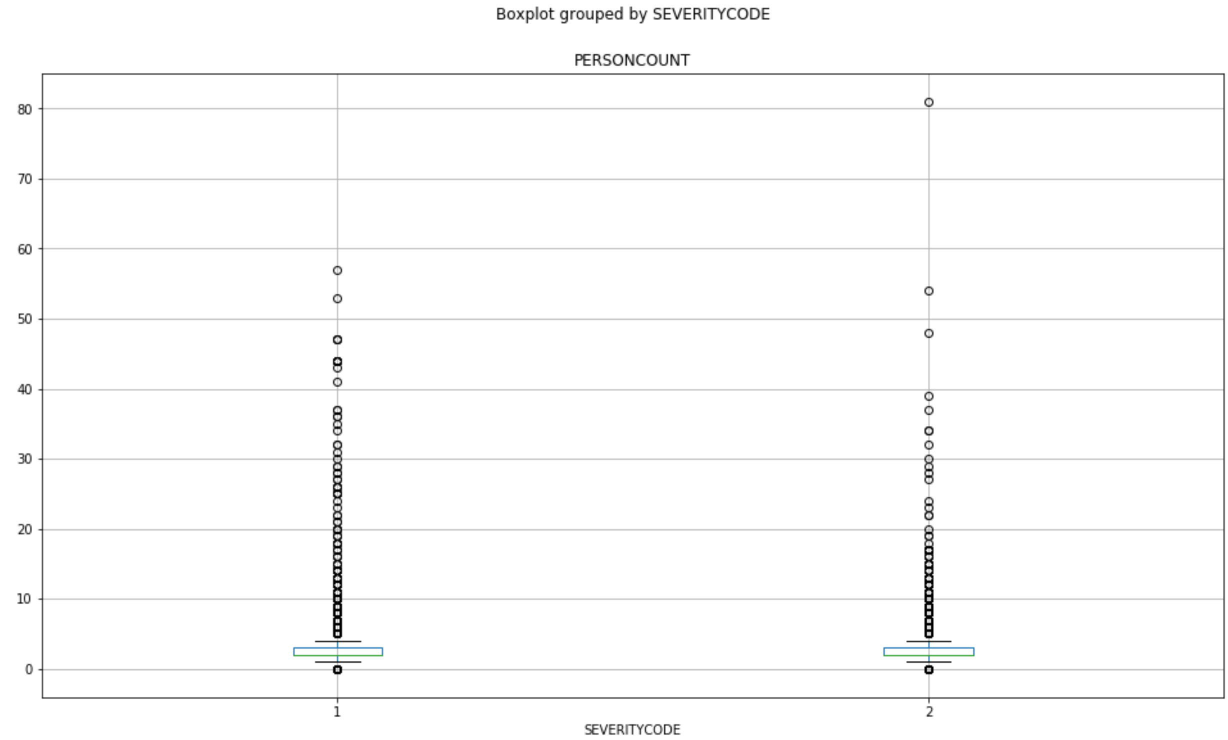


Figure 6. The distribution of severity based on number of person.

Related to figure 6, we can see that the average number of person in accident type is nearly the same. We also observe the same result with the number of vehicle.

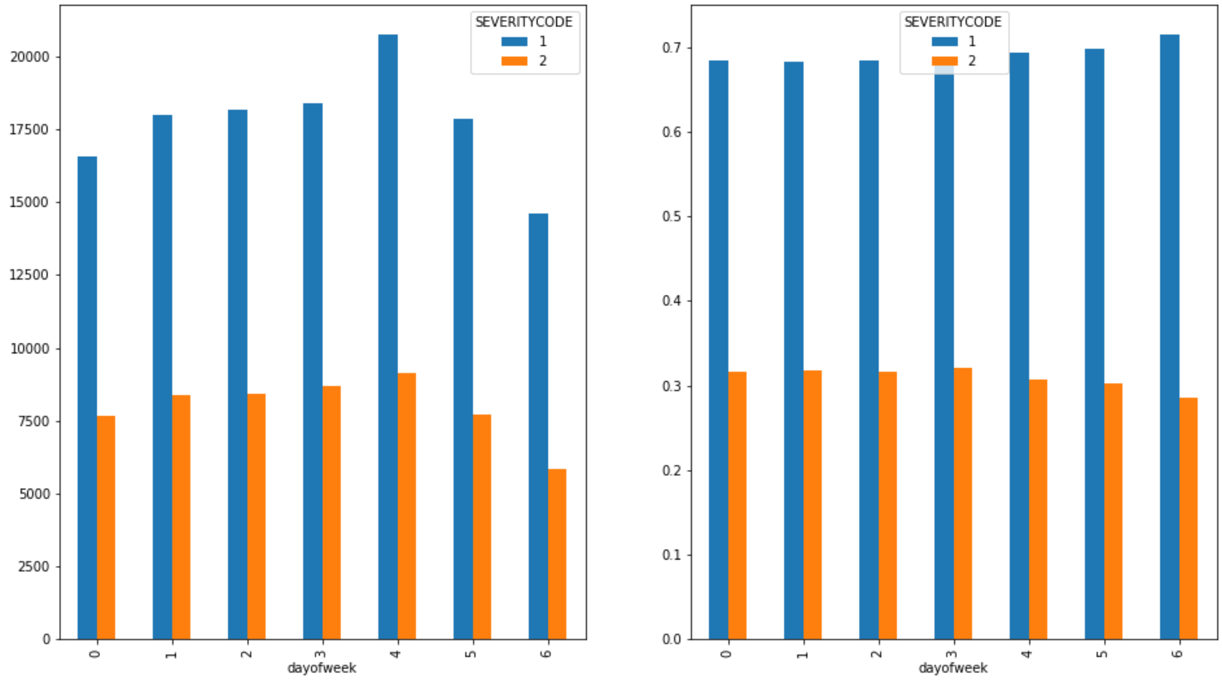


Figure 7. The distribution of severity based on day of week.

The number of accidents is greatest at 4th day (Friday) and at least in 6th day (Sunday) and the proportion of severity does not depend on the day.

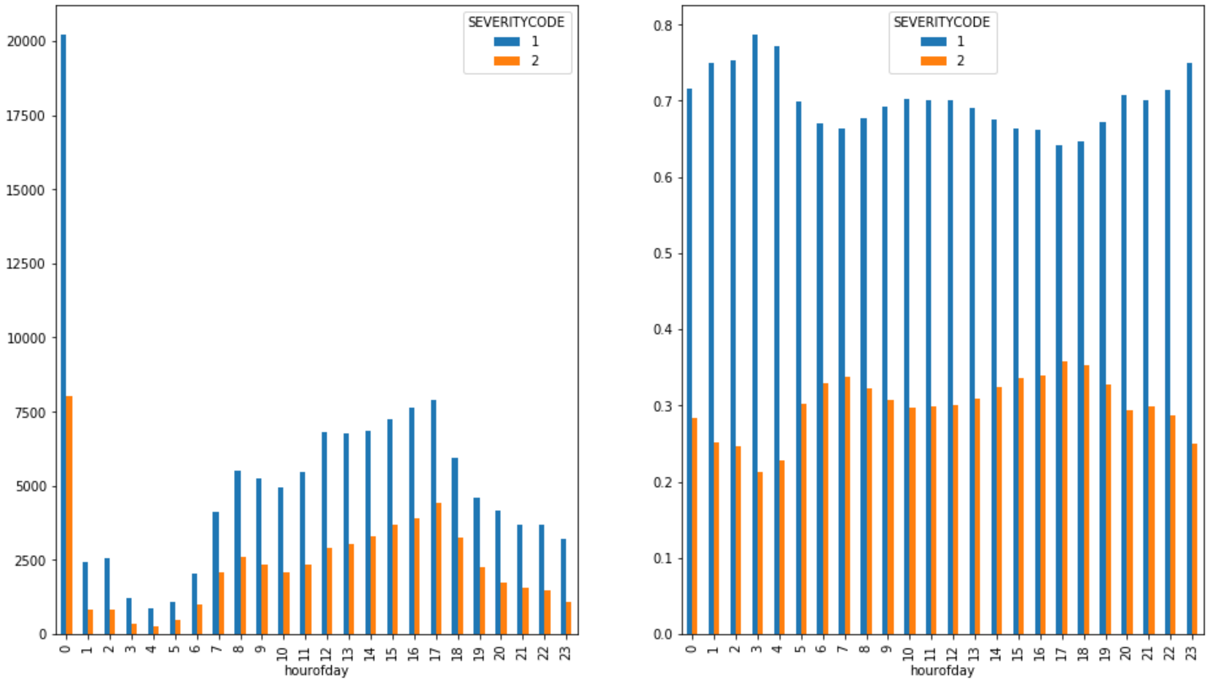


Figure 8. The distribution of severity based on hour of day.

In figure 8, there are a lot of accidents in mid-night. Related to the day analysis, it seems that people in Seattle has a lot of activities at Friday night and it causes accidents when they come home at 0h.

4.3. Data preparation

After clean the data by working with missing data, we need to encode the categorical data such as “ROADCOND”, “WEATHER”, “JUNCTIONTYPE”, “LIGHTCOND”.

The data is imbalanced, so, we can balance the data set by using under sampler, we remove 68422 rows to balance the data with each number of severity is 55822 rows. After under sampling data, we split 80% data to train and 20% data to test.

Because the problem is a classification problem. Hence, we choose 3 algorithms to apply this data set: K Nearest Neighbor (KNN), Decision Tree and Logistic Regression. SVM can’t be considered because the data set is large (more than 10000 rows)[1].

5. Result

	F1-Score	Jaccard score
KNN	0.64	0.47
Decision Tree	0.67	0.51
Logistic Regression	0.62	0.47

[1] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

The table shows algorithm analysis after under sampling data. The most appropriate model for this problem is decision tree with highest score.

6. Discussion

Our main aim was to predict the severity of the accident. It was very difficult to handle this large-sized data. After the data analysis, the data is not highly correlated and the data is highly imbalanced. With the imbalanced data set, it is easy to reach more than 70% accuracy (naive prediction – predicting all the accidents as severity “1”), but it was of no use. With the 0.67 in f1-score, the model is quite good, but if we can reach the score more than 0.7, it will be better and can be applied to the reality. We don’t choose the oversampling because there are a lot of categorical data.

7. Conclusion

In conclusion, most of the algorithms are biased towards most frequent class. However, pre-processing and corresponding imbalanced data techniques will give better results. Based on the current known condition of weather, light, road condition, speed, current time, number people in a street, number vehicle in a street, accident severity can be predicted. But there is no one feature that influences the accident severity.

Reference

The data information. *SDOT*. Retrieved from https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf