Predicting car accident severity

1. Business Problem

   In this project, we will try to predict the possibility of a car driver involved in a car accident under certain weather and road condition. This project will also try to give the possible severity of the accident. Specifically, this project will focus on Seattle City.

2. Data describe

   The data set is the collisions records of Seattle City. All collisions provided by SPD and recorded by Traffic Records. This includes all types of collisions. Timeframe: 2004 to Present.

   (1) Features selection

   Based on our problem, the data used in the project will include:

   the severity of the accident;

   the weather;

   road condition;

   the light condition;

   speeding or not;

   crosswalk or not.

   (2) Data cleaning

   First, replace the missing data in column "speeding" with 0

   Then, replace the non-zero values in column "crosswalkkey" as 1

   The following step is removing all rows that including missing data

   The last step is encoder the weather, road condition and light condition

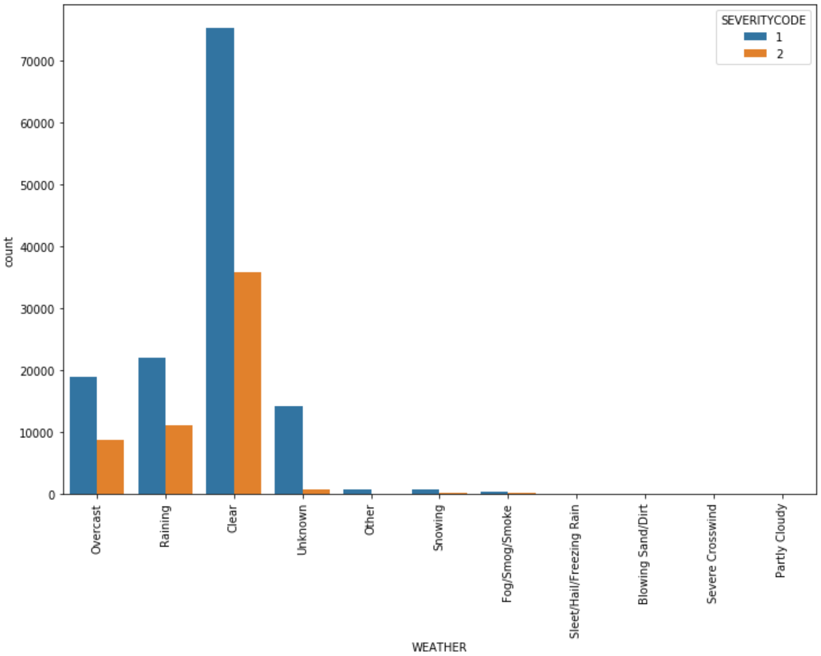
3. Methodology

   (1) First using bar graph to explore the relationship of the 5 features and the severity

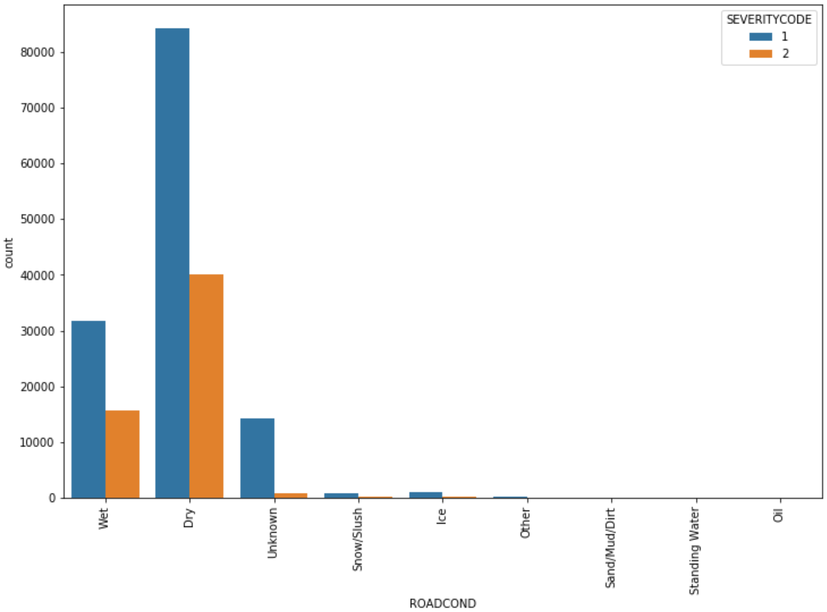
(2) Based on the business problem, we need predict the possibility of severity of the car accident, we choose Logistic Regression model.

4. Results

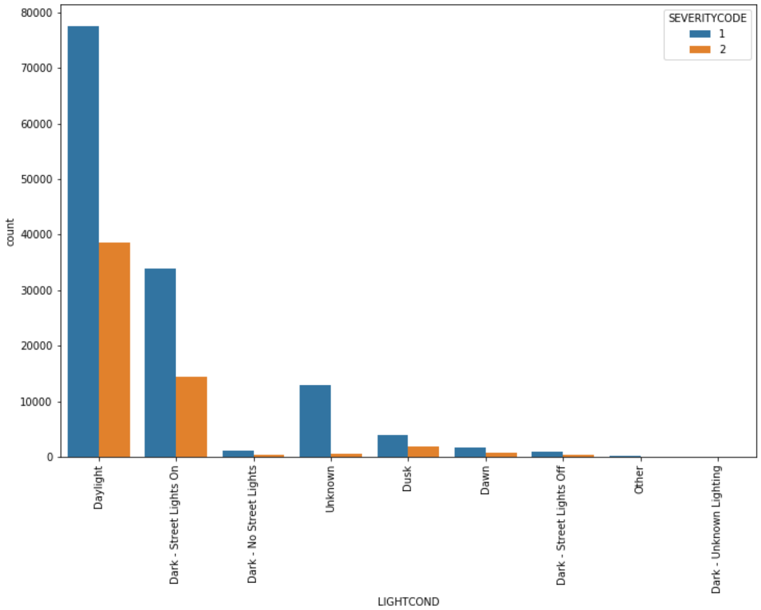
(1) Relationship of weather and severity:



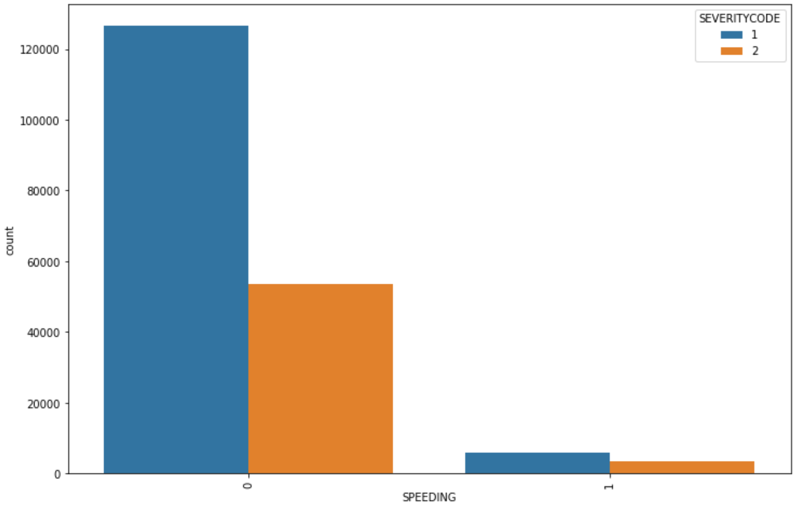
(2) Relationship of road condition and severity:



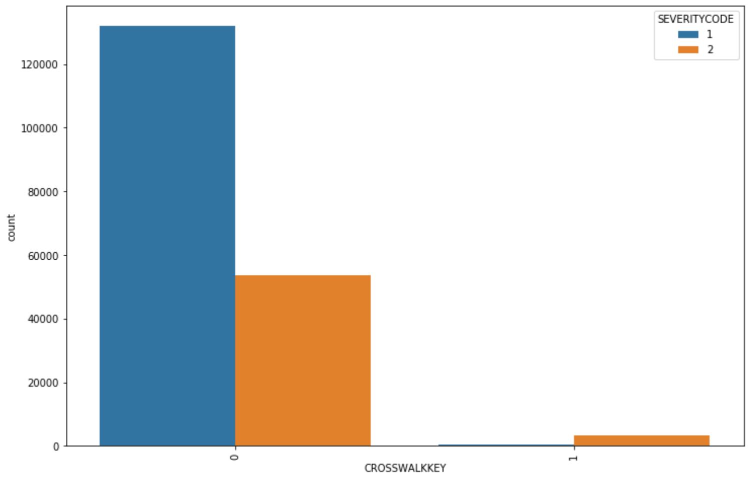
(3) Relationship of light condition and severity:



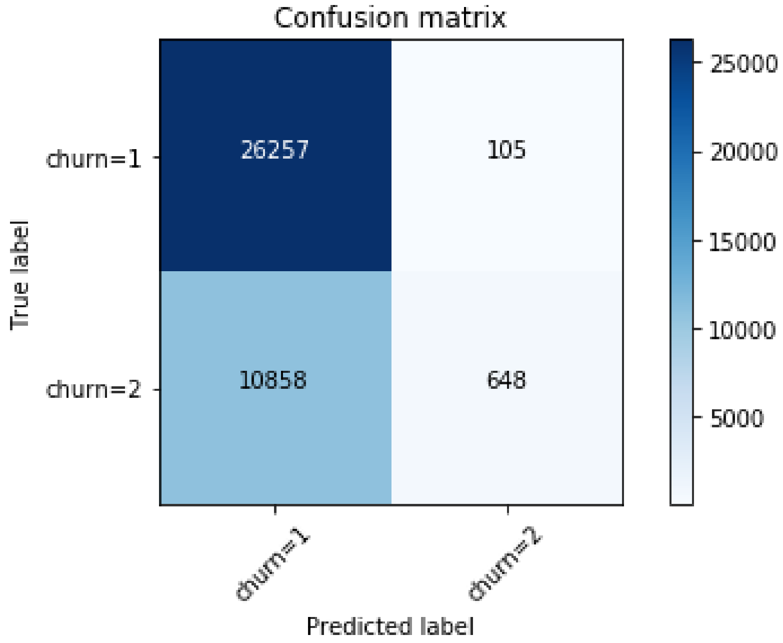
(4) Relationship of speeding and severity:



(5) Relationship of cross walk and severity:



(6) The Confusion matrix of model:



5. Discussion

   (1) The features have obviously effect with the severity;

   (2) The jaccard index of the model is about 0.71, it pretty good, but the Confusion matrix shows that there are lots of severity = 2 cases wrongly categorized to 1, the reason is that the data is unbalance labeled, some technologies need to use to deal this problem.

6. Conclusion

This case provide a possible method to predict the severity of car accident under certain weather and road conditions. The limitation of this case are:

1. There dataset is only including one city, we need more data to get better results;
2. The dataset is unbalance labeled, we need more technologies to deal with this.