Report of Final Assignment

Carisa Li

November 18, 2020

1. Introduction

1.1 Background

Car accidents are a kind of accidents which can be from vital to insignificant. Thus, how to determine a car accident's severity as soon as possible is important, since it is helpful to take countermeasures. As a result, I use the dataset which is about the collisions and try to predict the severity of a car accident.

1.2 Problem

Can we predict a car accident's severity?

1.3 Interest

This model can help traffic police and hospital to decide next action or be prepared. For example, traffic police can evacuate traffic to let ambulances arrive sooner and save other drivers' time, while hospital can be well-prepared when occur vital car accident. What's more, when several car accidents happen in the same time, this model can help these people to allocate resources. Therefore, the injured can get better treatment and these car accidents can have less impact on others.

2. Data

2.1 Source

This data is provided by Seattle Department of Transportation. I download it through this link: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

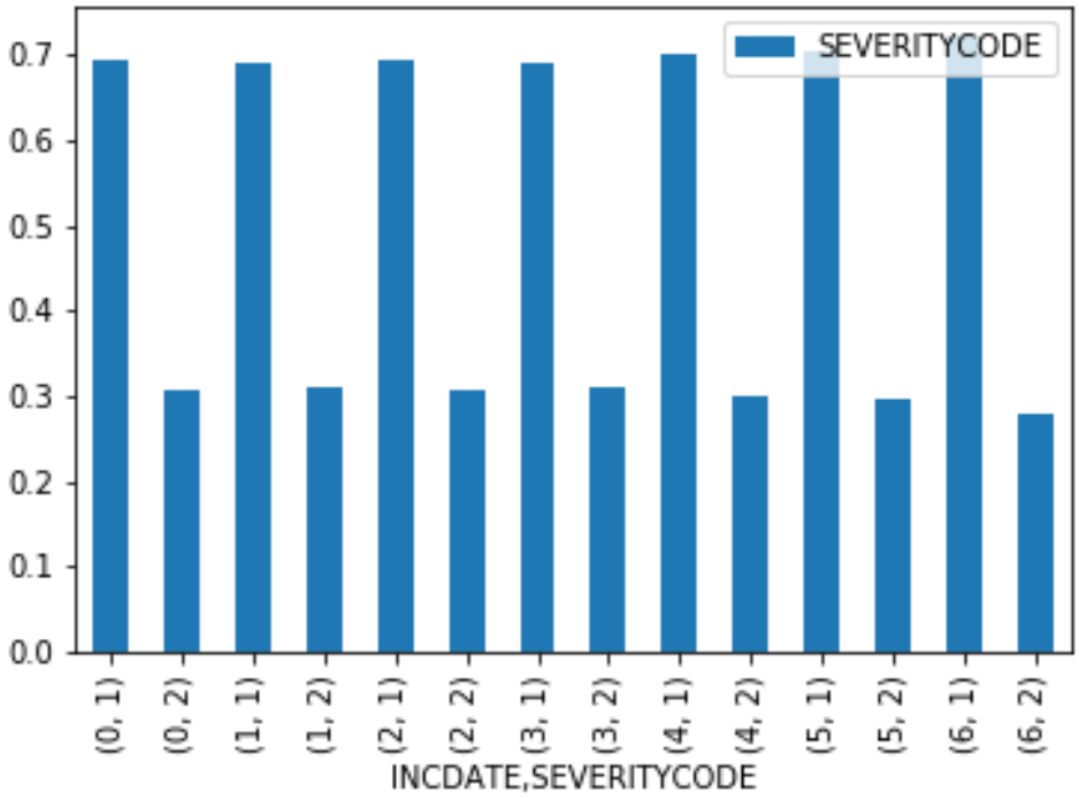
2.2 Content

This dataset is about all types of collisions from 2004 to present in Seattle. It originally has 194672 rows and 38 columns, including serial number, time, place and other details.

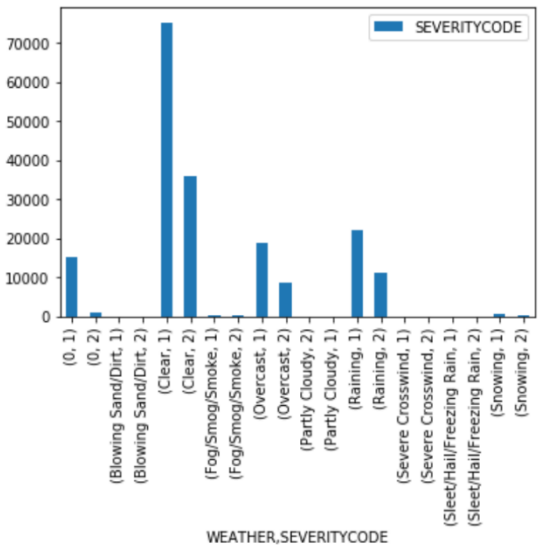
3. Methodology

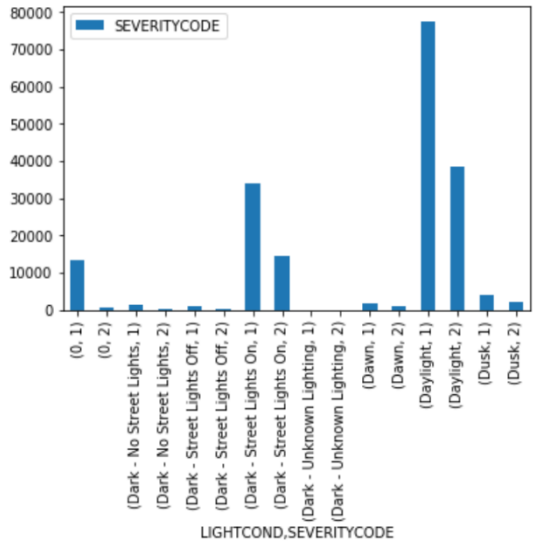
3.1 Data analysis and Visualization

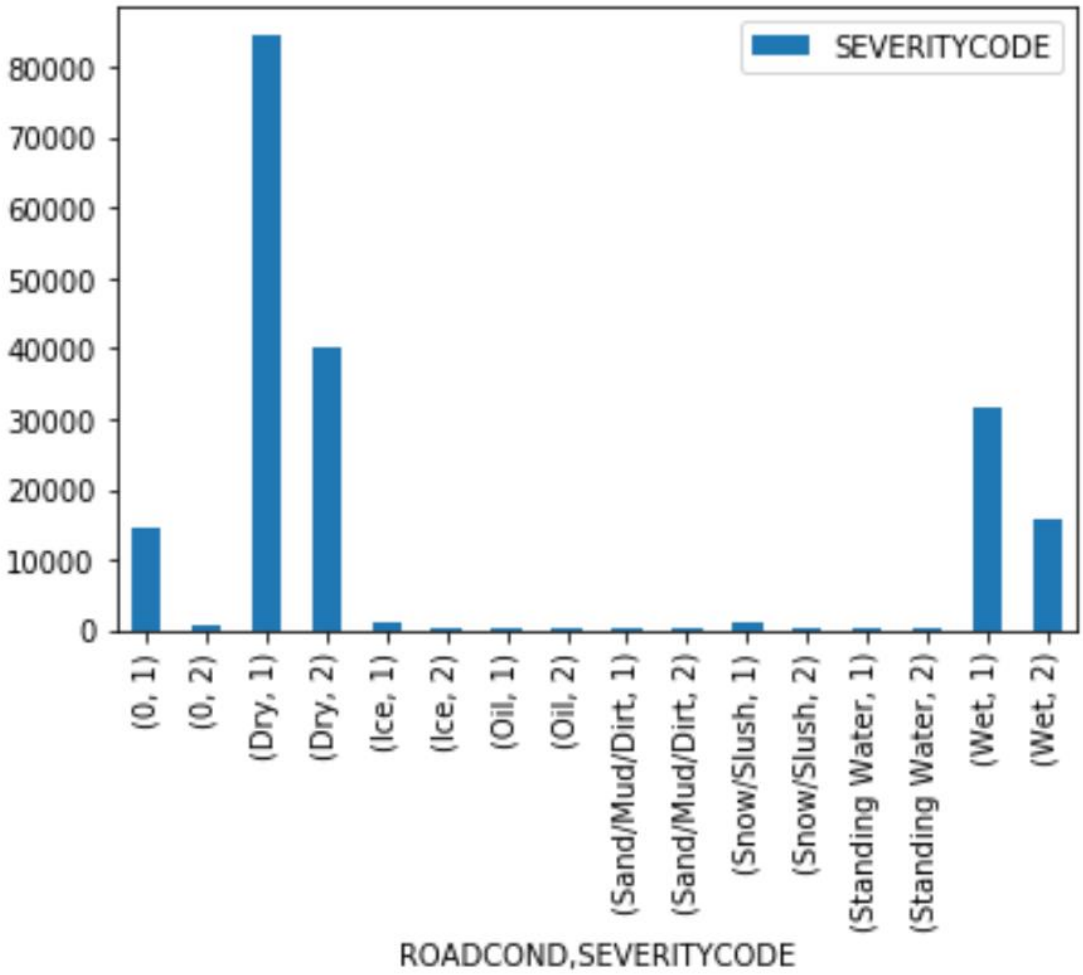
Through the bar charts of several factors and severity, I found the below things which are helpful for me to do pre-processing.



Above picture shows that the severity ratio of 1 to 2 is roughly 7:3 no matter what day of week it is. Thus, I think day of the week has a little effect on the severity of a car accident.







These pictures show that there are some conditions are too rare, so I combine them into “0” column, which means “NaN”, “Other”, “Unknown”, to simplify the models.

3.2 Pre-processing

First, I dropped out all the columns about serial number, time and place. Second, I combined some options to simplify my model. Third, I replaced those yes/no question with 1/0, and did one hot encoding on the columns of multiple options. Fourth, I did type casting to make sure all columns are int or float. Last, I standardized my data. Finally, dataset (X) has 189787 rows and 37 columns.

3.3 Modeling and Evaluating

I use 70% of data to train, while the other is used to evaluate.

Since the data is too large, K Nearest Neighbor and Support Vector Machine are not suitable. Therefore, I used Decision Tree (with max_depth from 4 to 9) and Logistic Regression to train model.

The below pictures show the comparison of different models.

4	0.7963697277829946
5	0.7963221169195861
6	0.7967959928418549
7	0.7998976942425119
8	0.8002969008090273
9	0.7988271432605216

The accuracy of Decision Tree with different max_depth.

	Algorithm	Jaccard	F1-score
0	Decision Tree	0.757838	0.847295
1	LogisticRegression	0.757363	0.845823

The accuracy comparison of Decision Tree with max_depth=8 and Logistic Regression

4. Results

I get a model having at least 75% correct rate. People can roughly predict the severity of a car accident through this model.

5. Discussion

We get this model by using weather, environment, details of collision, and so on. Thus, we should train staff member to ask suitable questions, get the necessary information, and predict severity through the model when receiving a report call. Additionally, for this model should be retrained on the regular basis, people should record data in detail to get a better model.

6. Conclusion

In this report, I train a model to predict the severity of car accident. After training staff member to use this model, police and hospital can allocate resources, decide next action and be prepared sooner.