# Car Collisions Severity Prediction Project

Renhuan He

September 29, 2020

## 1. Introduction

According to the annual United States road crash statistics by ASIRT, more than 38,000 people die every year in crashes on U.S. roadways. The U.S. traffic fatality rate is 12.4 deaths per 100,000 inhabitants. It is evident that with the increasing number of vehicles on urban and suburban roads, the cases of vehicle accidents are also increasing.

Say you are driving to another city for work or to visit some friends. It is rainy and windy, and on the way, you come across a terrible traffic jam on the other side of the highway. Long lines of cars barely moving. As you keep driving, police car start appearing from afar shutting down the highway. Oh, it is an accident and there's a helicopter transporting the ones involved in the crash to the nearest hospital. They must be in critical condition for all of this to be happening.

Now, wouldn't it be great if there is something in place that could warn you, given the weather and the road conditions about the possibility of you getting into a car accident and how severe it would be, so that you would drive more carefully or even change your travel if you are able to.

How to predict the severity of an accident? It would be useful for drivers and police officers if we could predict the severity of an accident based on environmental factors, human factors so as to avoid an accident or decrease severity of an future accident.

## 2. Data

### 2.1 Data Cleaning & Feature Selection

Data is provided by the Seattle Department of Transportation (SDOT) on vehicle collisions along with its severity starting from 2004 to 2020. The dataset consists of 38 columns having different kinds of data including collision severity, road conditions, number of people involved, location of collision, weather etc. This project is going to use this dataset to explore and predict the severity of an accident considering factors including weather, road conditions, location, light conditions, time and day of the week, etc. Based on definition of our problem, factors that will influence our decision are:

- weather conditions
- road conditions
- light conditions
- location type
- time of the day
- day of the week
- number of people involved
- number of vehicles involved
- car speeding

There are 38 columns and 194673 rows in the dataset. According factors we choose for the problem and , we could drop following redundant columns and drop missing values. And then we start to check missing values in the dataset and found some problems. First, For columns: INATTENTIONIND and SPEEDING, there are too many missing values, so we decide to remove these two columns. Secondly, the column UNDERINFL has four different values in dataset but supposed to only have two types of values. Fortunately,

the unexpected values have similar meaning, so we modified the values and encode the column to 0 and 1 to prepare for analysis.

The third problem is that in columns ADDRTYPE, WEATHER, ROADCOND, LIGHTCOND still a few null values existed. To avoid losing more valid data and also consider that these columns are all about environment conditions that might influence severity of car collisions, so we decided to add another value- "others" to fill those NAs. And the way of filling in NAs is also reasonable because there is always exceptions in the car collisions and their environment conditions.

After solving those problems in data, we can start to analysis the dataset for more insights.

# 3. Methodology

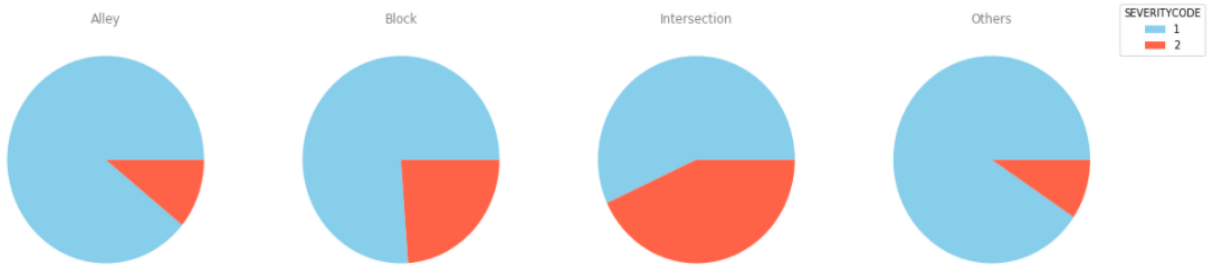
In order to predict the severity of an accident, we have selected several factors. Now first we need to do some exploratory data analysis to figure out the relationship between severity of an accident and each factor. Then we could transform data into proper format to do further analysis.

After exploratory data analysis and data transformation, we could start to use classification models in Machine Learning including KNN, SVM, Logistics Regression and Decision tree. Since there is only two classes of severity of an accident as the target variable, classification models are the most suitable and straightforward choice.

### 3.1 Exploratory Data Analysis

A. Relationship between accident severity and ADDRTYPE

There are three main types of location: Alley, Block and Intersection. It is reasonable that different location may have different influence on the severity of collisions. From the pie chart below, we can find that there are significantly highest percentage of level 2 severity of accidents happen in Intersection and block also has second highest percentage of level 2 severity.




B. Relationship between accident severity and status

| STATUS | Matched | Unmatched |
|---|---|---|
| SEVERITYCODE | | |
| 1 | 69.88 | 100.0 |
| 2 | 30.12 | 0.0 |

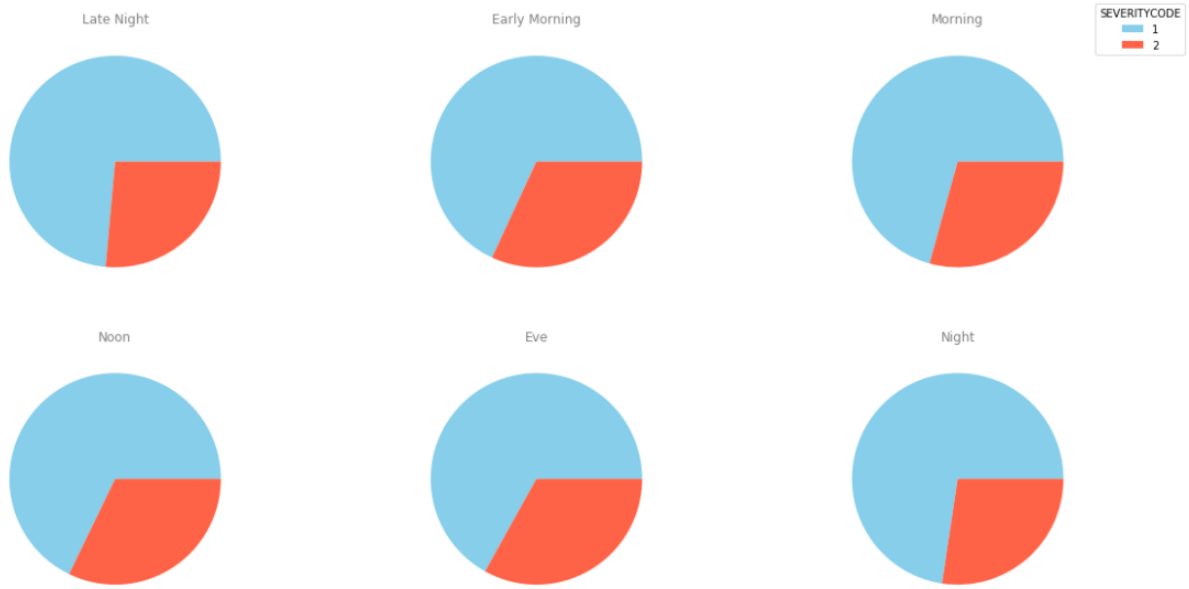
From the table, we can clearly tell that Matched status has much greater effect on severity of accident. Because there is no level 2 severity in unmatched status.

C. Relationship between accident severity and people/vehicles involved

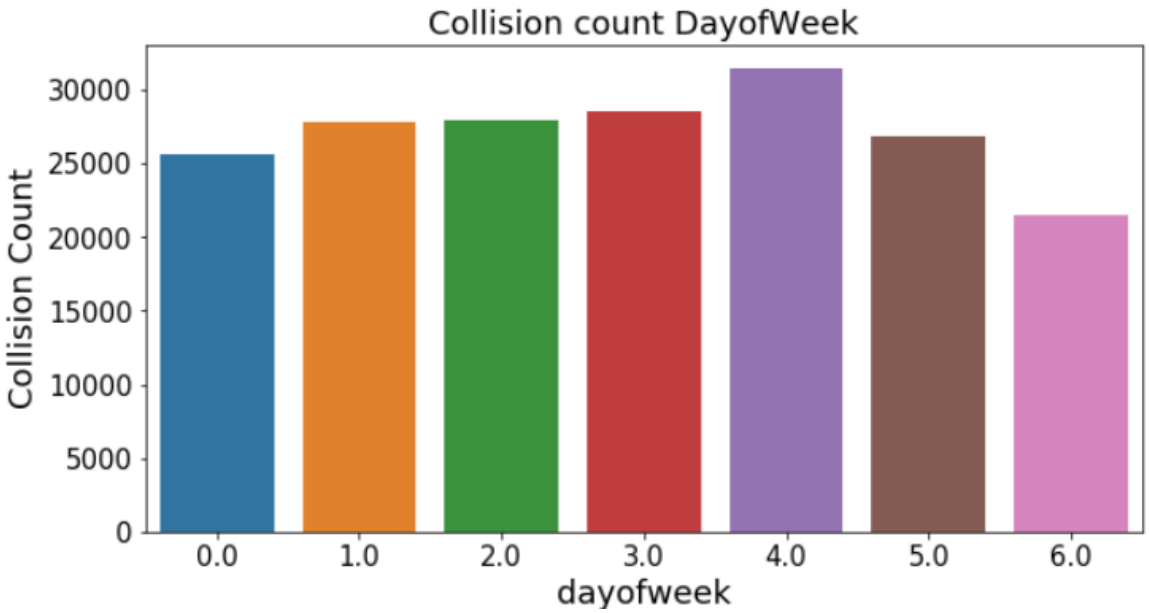
From the heatmap below we can find that the relation between people involved in collisions and collision severity is stronger than other three factors. This is quite reasonable because more people involved more people might get hurt during car collisions.


D. Relationship between accident severity and day/time

To explore the severity of collisions and time of the day, we used the INCTIME column and divided accident time into six categories: Late Night, Early Morning, Morning, Noon, Eve, Night. Then we plot the percentage of two level of severity in pie charts according the time period. Then we find that in the Early Morning, Noon and Evening have slightly higher percentage of level 2 collision severity.

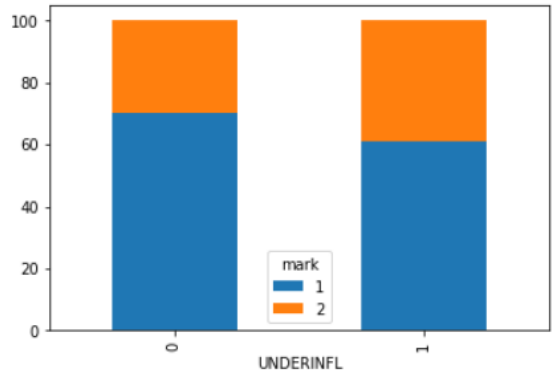


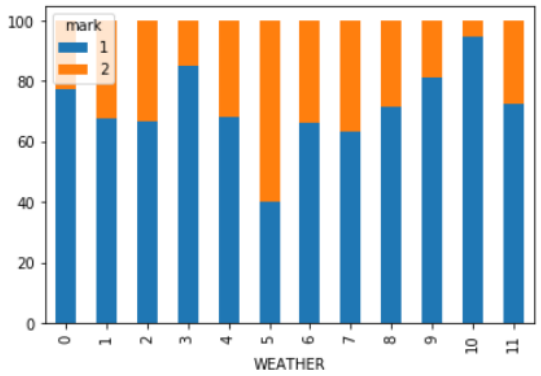
Then we used the INCTIME column and listed the day of week for every collision and created a count plot as follow. We found that more collisions happened in Friday but there is no apparent difference between each day of week in the percentage of level 2 collision severity.

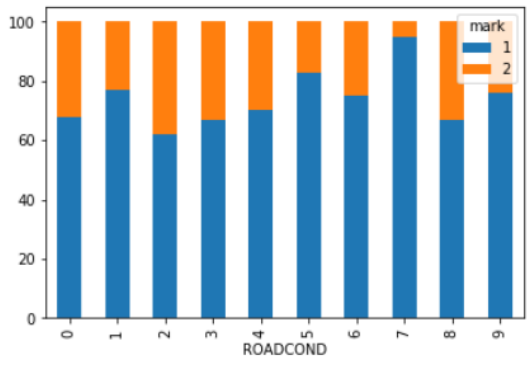


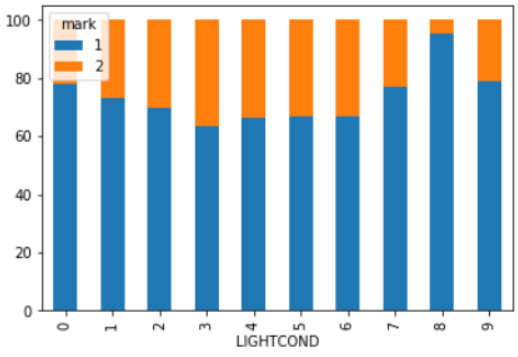

E. Relationship between accident severity and under influence, weather, road condition, light condition

Similarly, we compared influence of different categories of under influence, weather, road condition, light condition on severity.



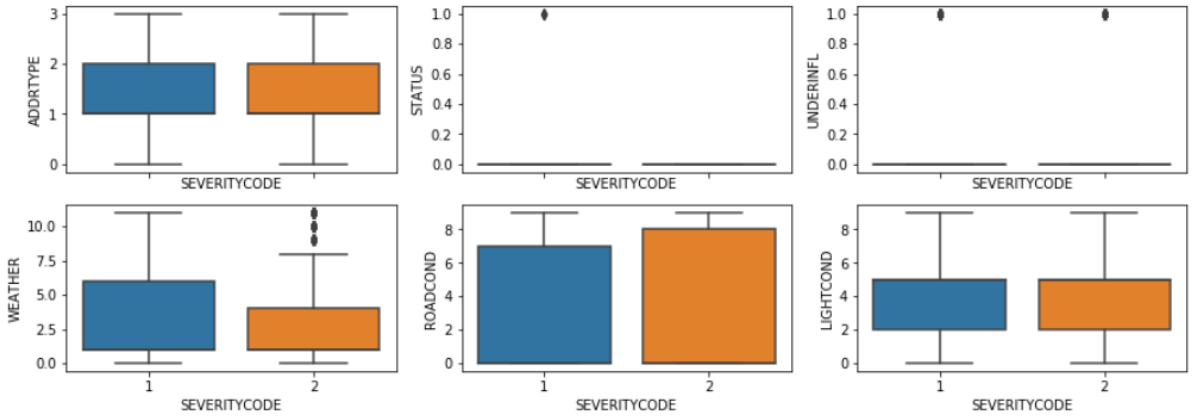



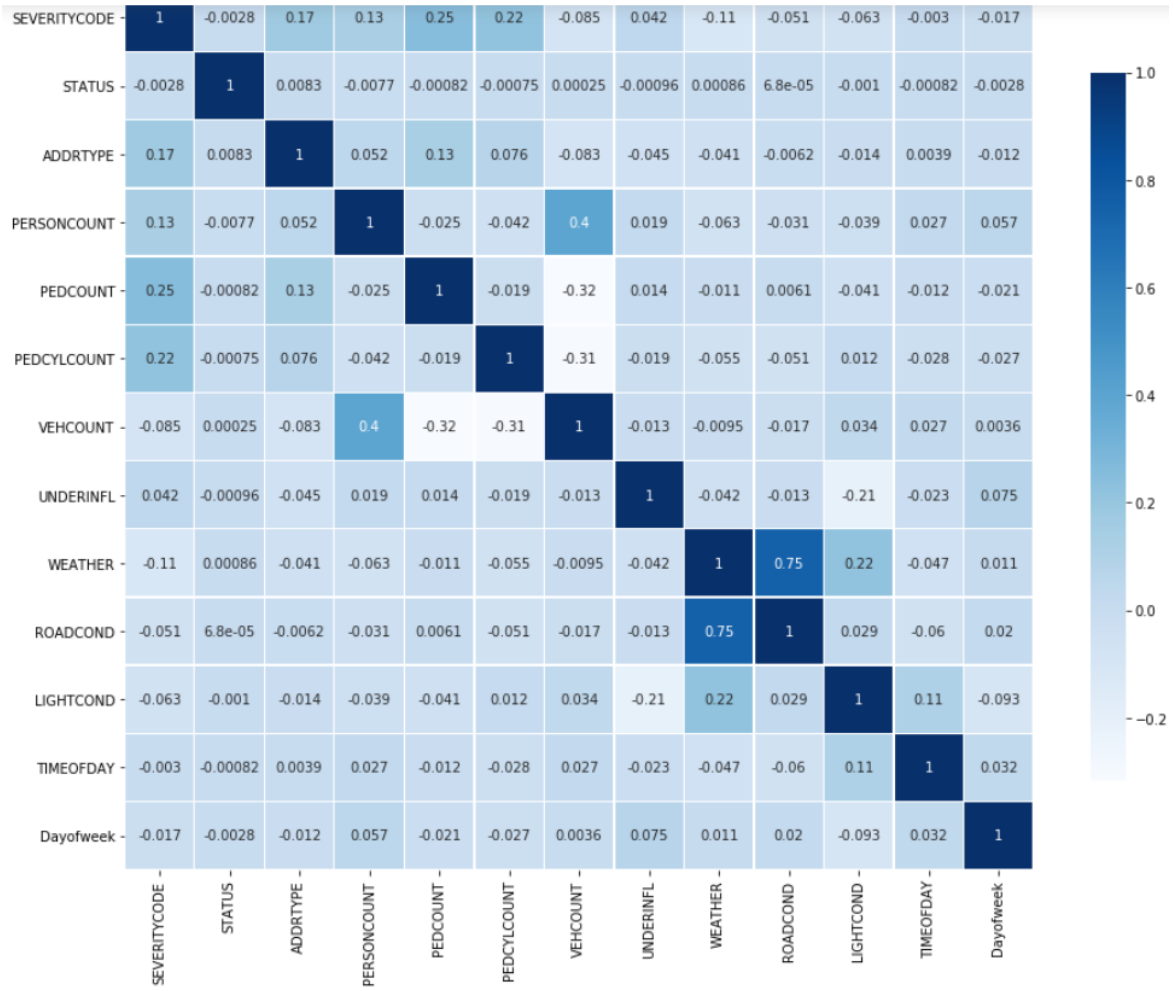



F. Encoding the data

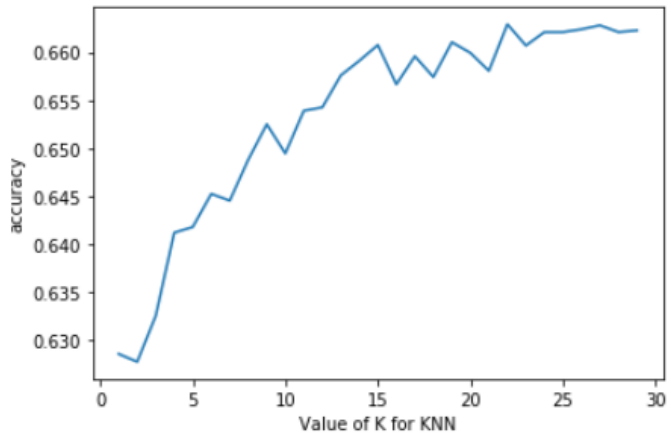
For the convenience of quantitative analysis, we need to transform all categorical factors into numerical values. So we simply encode each category into numbers and then we could use boxplot to check the distributions of values. Also we used heatmap for clearly show the relation between each two variables and found that two group of variables have significant positive relation. One is between vehicle involved and person involved, another is between weather and road condition. Weather and road condition have a strong relation -0.75 which can explained that the weather caused most changes in road condition.



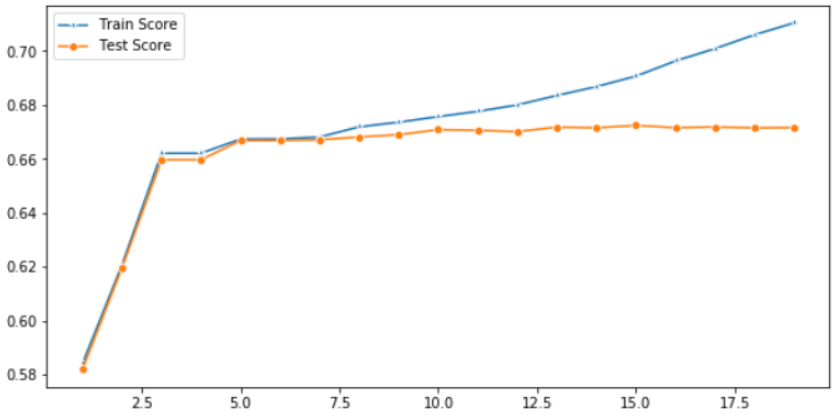



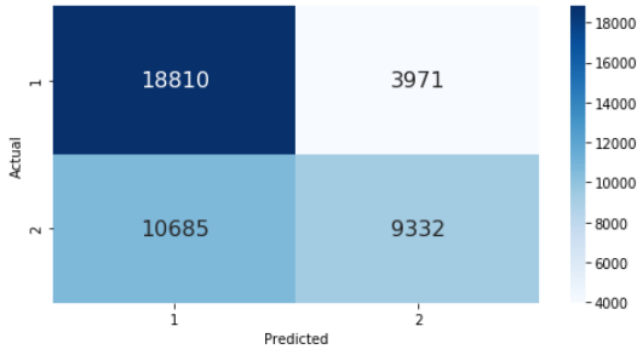
### 3.2 Predictive Modeling & Evaluation

In this project, we used four classification models(KNN, SVM, Logistics Regression and Decision tree) to predict the severity. These four model are most common classification models that would be most efficient for binary classification prediction. For knn model, we used a loop test to find the best k value and the best k is 22.




For svm model, we chose linear kernel and 0.1 as C value. The result is slightly better than knn but it took much longer computing time. For decision tree model, we used loop test again to find the most propriate max depth value. From the line chart below we find out that when max depth equals 5, the model would be more effective. For logistic regression model, we still chose linear kernel and 0.1 as C value. From the confusion matrix heat map below we can find that there are a great amount of false negative prediction.



# 4. Result & Discussion

In KNN model, the accuracy on test data is 0.6554 and the best k value is 22 according to our loop k value test. In SVM model, the accuracy is 0,6668 which is slightly better than knn but the computation time is extremely long which means svm require much more computation. In logistic regression model, the accuracy is 0.6575. It is again slightly better than svm model but much less computation needed. In decision tree model, the accuracy on test data is 0.6668 which is highest result among all four models. So, for this project, decision tree might be the most effective and suitable model. One reason might be that in this dataset more than half of variables are categorical factors at beginning, so decision tree could better work on it.

| Algorithm | accuracy | TureNegative | FalsePositive | FalseNegative | TruePositive |
|---|---|---|---|---|---|
| **KNN** | 0.655404 | 17250 | 5531 | 9217 | 10800 |
| **Decision Tree** | 0.666830 | 18914 | 3867 | 10392 | 9625 |
| **SVM** | 0.656666 | 19370 | 3411 | 11283 | 8734 |
| **Logistic Regression** | 0.657554 | 18810 | 3971 | 10685 | 9332 |

# 5. Conclusion

This project uses four most common and simple classification models to train and predict based on the data set. During the four models, decision tree model has the best performance and relatively less computation time. But there might be models more useful on this dataset. And the accuracy of the models could be improved with small changes on hyper parameter and data transformation. Lastly, there could be more data and more other important factors left without recording. Also, if we could use the data of exact location, we could predict the severity according each location that might more accurate.