**Predict Car Accidents Severity**

Liuqi Qian

August 2020

**Introduction**

As the most commonly used transports, automobiles play an important role in daily life. People drive cars to work, study, travel, and even move to a new house. With the widespread use of automobiles, the possibility of traffic accidents increases inevitably. Besides, the environmental factors including weather, location light, and more will also affect the probability of accidents.

However, sometimes emergency departments are hard to evaluate severity when they received the call from on-site which may cause irreparable loss and even people's death. Based on the situation, predict the severity is one of the crucial points.

To help the people who are involved in the accidents, this project will utilize certain know conditions to predict the severity of the accidents and reduce the risks of accidents happen by taking actions.

**Data Source**

To address the problem, the weekly collision from 2004 to present in Seattle is going to be used in this case. The original dataset is a csv file and can be found [here](https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv). Also, there is a [metedata form](https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf" \t "_blank) to give the basic information about the dataset.

**Data Description and Cleaning**

**A screenshot of a cell phone

Description automatically generated**There are total 194,673 rows and 38 columns in the original dataset.

Right is the list of all the columns (features):

In this project, the first column "SEVERITYCODE" is the target variable that will be predicted by other fields. According to the metadata file, there are a total of 5 codes that correspond to the severity. However, there are only two types recorded in the dataset: code 1 and code 2. There is 136,485 number of code 1 and it stands for "prop damage". 58,188 observations represent "injury" as code 2.

After checking the NA number in each column, there are seven fields that have more than 40% of the missing values including "INTKEY", "EXCEPTRSNCODE", "EXCEPTRSNDESC", "INATTENTIONIND", "PEDROWNOTGRNT", "SDOTCOLNUM", SPEEDING". These columns will be dropped in this project.

A screenshot of a cell phone

Description automatically generated

Also, following columns will be kept or dropped based on their reasons:

* The "WEATHER" column is related to the "ROADCOND" so only "ROADCOND" will be kept
* "X", "Y" columns are describing the geographic point of the "LOCATION" so only "LOCATION" will be kept
* "INCDTTM" includes the information in the "INCDATE" and only "INCDTTM" will be kept
* "PEDCOUNT" and "PEDCYLCOUNT" are related to "COLLISIONTYPE" and only "COLLISIONTYPE" will be kept
* "ST\_COLCODE" includes the information in the "HITPARKEDCAR" so "HITPARKEDCAR" will be dropped
* "COLLISTIONTYPE" includes the information about "STOD\_COLCODE", "STOD\_COLDESC", "ST\_COLCODE" as well as "ST\_CODESC", and only "COLLISTIONTYPE" will be kept
* "JUNCTIONTYPE" is similar to "ADDTYPE" and only "ADDTYPE" will be kept
* "STATUS" is meaningless in this project and will be dropped
* "LOCATION" is too massive in this project and will be dropped

After dropping the columns, there are only 9 columns left. Following is the screenshot of top 5 rows:

A screenshot of a cell phone

Description automatically generated

Since there are still many NA values and there are total 194673 records, the rows which contains NA values will be dropped. And finally there are 187,609 rows with 9 columns.

**Convert Data Type**

Since the "INCDTTM" is meaningless but the month and the hour time makes senses, we will extract the month and hour value from the "INCDTTM" column. And create two new columns, “MONTH” and “HOUR”. Followings are the bar chart shows the person count bar plot that group by month and hour.

A screenshot of a cell phone

Description automatically generated

Also, all the string type columns are be converted into integer type including “ADDRTYPE”, “COLLISIONTYPE”, “UNDERINFL”, “ROADCOND”, “LIGHTCOND”.

A screenshot of a cell phone

Description automatically generated

Following is the severity code bar plot in the dataset.

A screenshot of a social media post

Description automatically generated

**Predict Model**

First try the Linear Regression

A screenshot of a social media post

Description automatically generated

After testing, it seems like a non-linear regression model.

Hence, decision tree is used for the future predicting.

A screenshot of a cell phone

Description automatically generated

**Conclusion**

It is useful to use the machine learning to predict the severity of car accidents. Most of the algorithms are biased towards to most frequent class. Proper preprocessing of the data will give optimal results.