Car Accident Severity Prediction Model

Introduction

Car accidents could have severe consequences on our daily lives. By understanding the main factors of the severity of car accidents, we drivers could better plan and manage driving habit to avoid or reduce unnecessary risks. This will also benefit the insurance and the entire society.

The goal of the analysis is intended to identify relative factors and build a quantitative model to predict the severity of the car accidents. The data were collected for the accidents occurred from 2004 to 2020 in Seattle area. The target variable is the severity, which has two levels:

- Property Damage Only Collision
- Injury Collision

As the target variable represents the severity of the accidents with two levels, the classification model would be appropriate for this problem. Applicable models include:

- K-nearest neighbor
- Support Vector Machine
- Decision Tree
- Random Forest
- Logistic Regression

Data

The dataset, provided by Coursera, consists of car accidents data collected from Seattle area from 2004 to 2020. Various attributes of each accidents represent the road condition, weather, location, vehicles and persons involved. Obviously certain attributes relate more closely to the severity of the accidents while others play a lesser role.

Data Quality and Missing Data

The dataset consists of 194673 records with 38 fields. The target variable is SEVERITYCODE with the following distribution:

SEVERITYCODE	SEVERITYDESC	Record Count
1	Property Damage Only Collision	136,485
2	Injury Collision	58,188

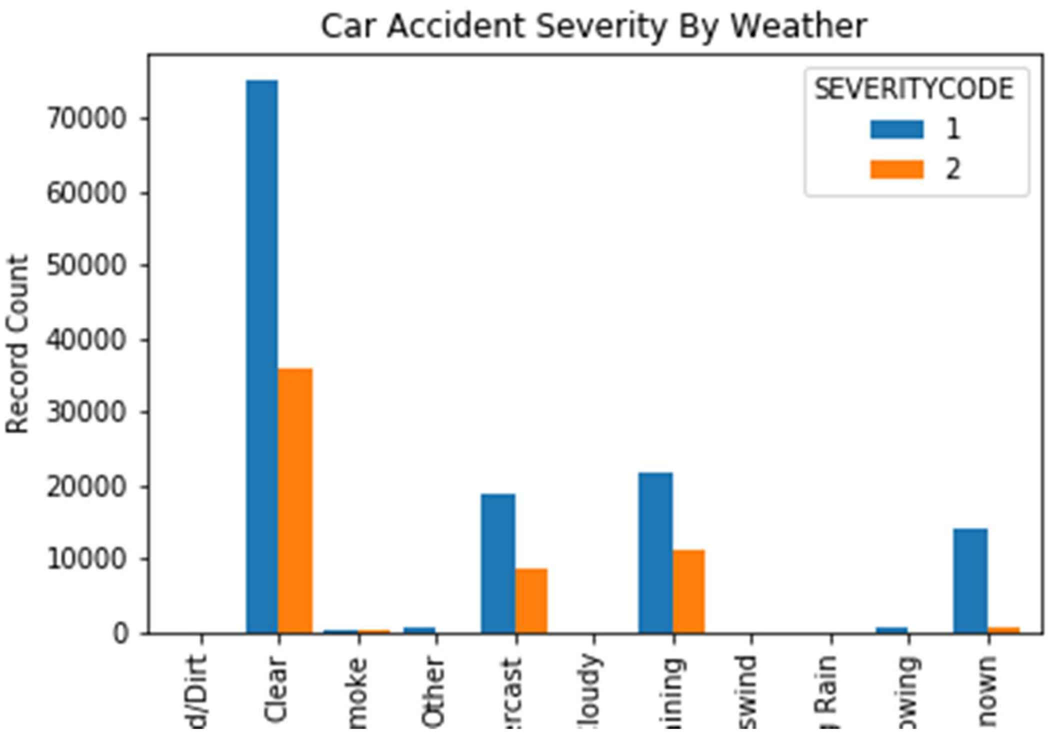
According to meta data, certain fields represent the identification of the incidence, which do not have direct relationship to the car accidents, such as 'X', 'Y', 'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO' and 'STATUS'. X and Y are longitude and latitude of the location. Additionally, a few fields have significant number of missing values, and they will be excluded from the modeling exercise.

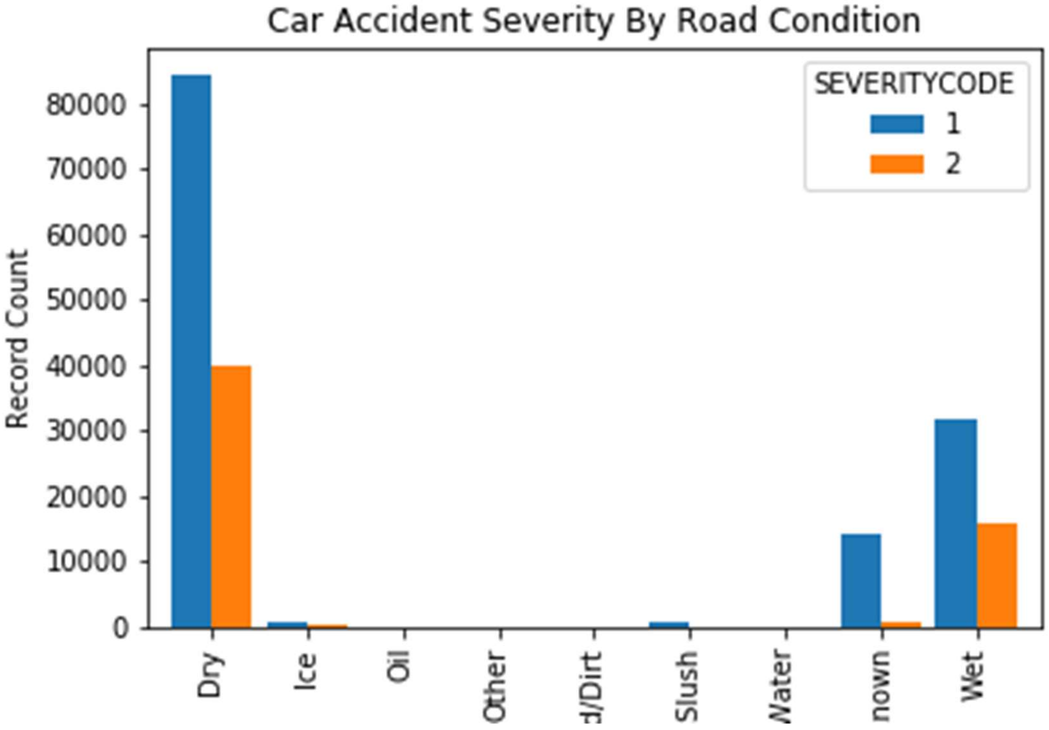
```
Field Name       Percent of Missing Value
-------------------------------------------
INTKEY               66.57%
EXCEPTRSNCODE        56.43%
EXCEPTRSNDESC        97.10%
INATTENTIONIND       84.69%
PEDROWNOTGRNT        97.60%
SDOTCOLNUM           40.96%
SPEEDING             95.21%
```

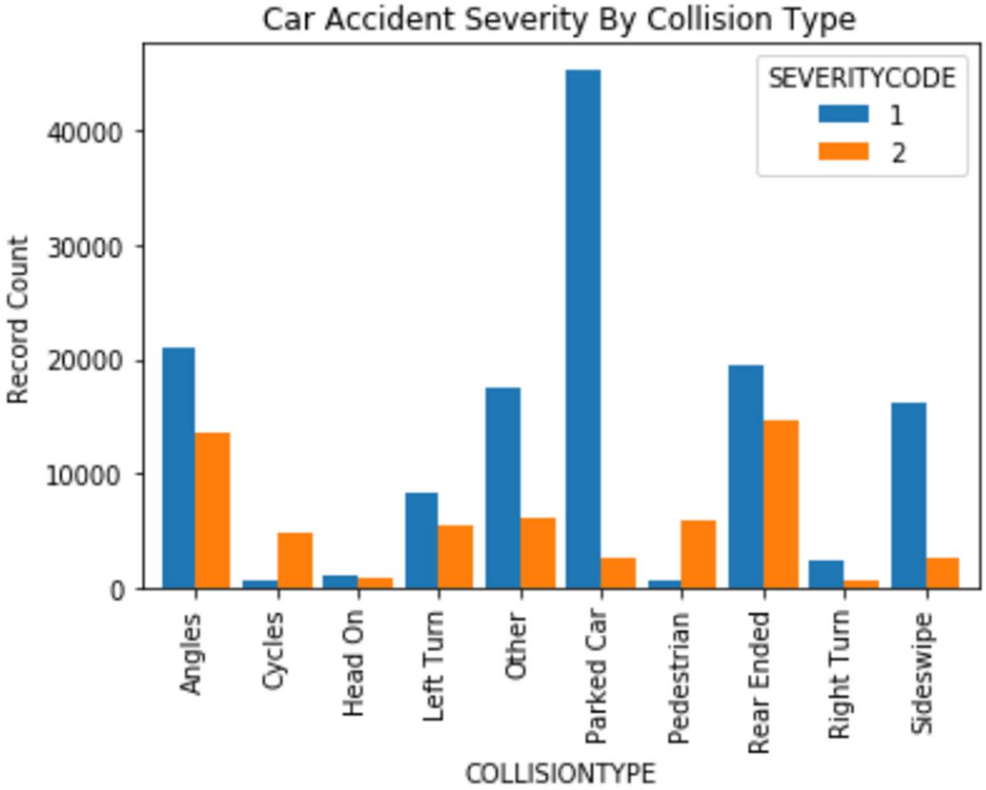
Exploratory analysis and visualization

The data exploration will help identify the variables and uncover relations between independent and target variables.

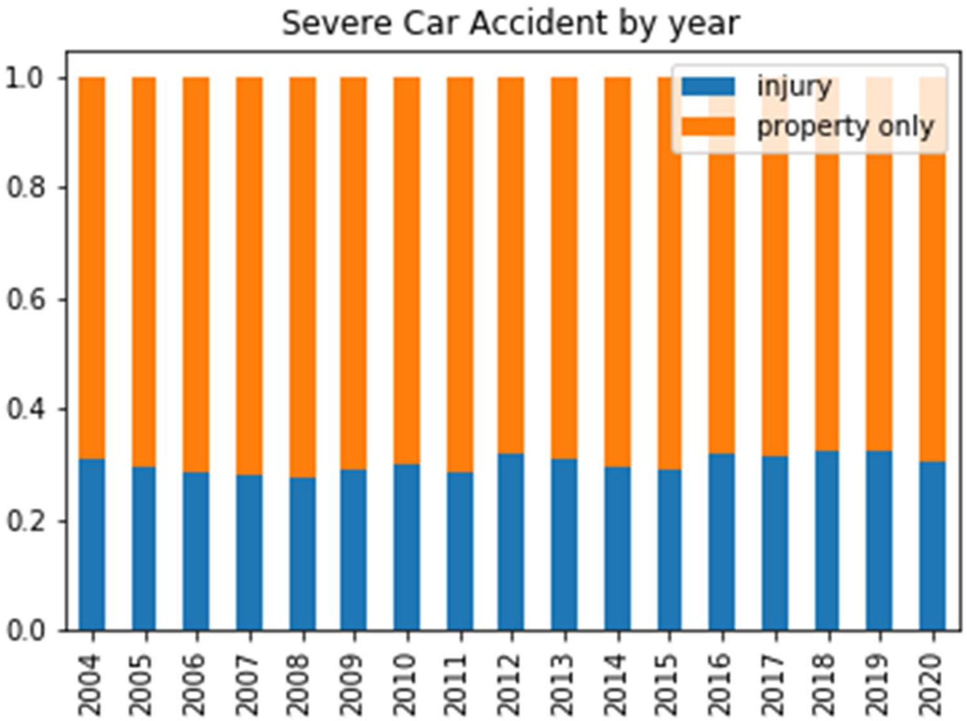
The “WEATHER”, “ROADCOND”, “LIGHTCOND” are categorical variables, and they have some predictive power in relations to the severity of car accidents.



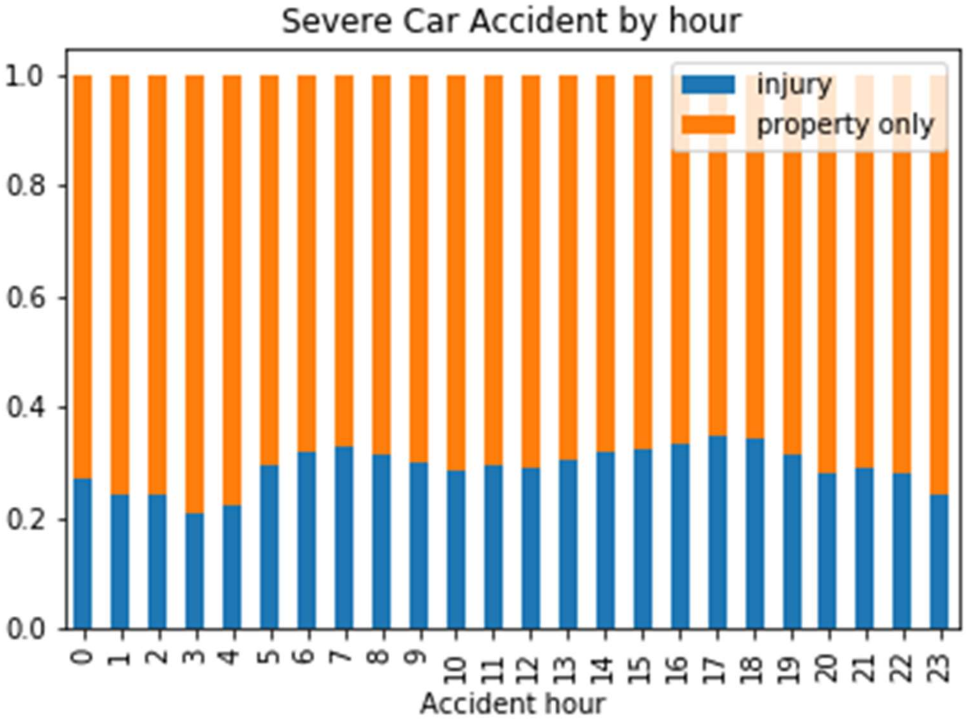

Car Accident Severity By Road Condition
SEVERITYCODE
1
2
Record Count
80000
70000
60000
50000
40000
30000
20000
10000
0
Dry
Ice
Oil
Other
d/Dirt
Slush
Vater
nown
Wet


Car Accident Severity By Collision Type
SEVERITYCODE
1
2
Record Count
40000
30000
20000
10000
0
Angles
Cycles
Head On
Left Turn
Other
Parked Car
Pedestrian
Rear Ended
Right Turn
Sideswipe
COLLISIONTYPE

There is no distinct long-term trend in terms of accident severity, as shown below:



However, the time of day does play a role in determining the severity of the accidents, where morning and evening traffic hours tend to observe more injuries.



Data Conversion for Modeling

Categorical variables are converted to on-hot coding for modeling purpose.

New features were extracted from ‘INCDTTM’ field to create new variables representing traffic hours.

Selected features include: 'COLLISIONTYPE', 'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND'.