Applied Data Science Capstone

Predicting Car Accidents by road and light conditions

Introduction - Background

- For the final capstone project in the IBM certificate course, we want to analyze the accident “severity” in terms of human fatality, traffic delay, property damage, or any other type of accident bad impact. The data was collected by Seattle SPOT Traffic Management Division and provided by Coursera via a link. This dataset is updated weekly and is from 2004 to present. It contains information such as severity code, address type, location, collision type, weather, road condition, speeding, among others.
- The target audiences of this study can be car companies, insurance companies and car owners, and especially in the transportation department. And it means to figure out the reason for collisions and help to reduce accidents in the future.

Introduction - Problem

- As there are more and more vehicles around the world, the number of car accidents, especially collisions, have increased in these years. It can be problems for all stakeholders and cause some serious troubles. This project aims to find out what type of Collision type has the highest severity level in the past, and whether road condition and light condition have impacts on these accidents.

Introduction - Limitations

- Although there are plenty of characteristics that might affect car accidents, deal to the time and efforts, this project will only focus on road condition and light condition and try to figure out how these attributes affect severity level.
- We will use data science power to generate Promising results and help stakeholders reduce accidents and optimize cars or roads condition.

Data Acquisition and Cleaning - Data Sources

- We will use Data-Collisions from IBM capstone course and predict a model of car accidents. There are 194,673 observations and 38 variables in this data set.

	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE
0	2	-122.323148	47.703140	1	1307	1307	3502005	Matched	Intersection
1	1	-122.347294	47.647172	2	52200	52200	2607959	Matched	Block
2	1	-122.334540	47.607871	3	26700	26700	1482393	Matched	Block
3	1	-122.334803	47.604803	4	1144	1144	3503937	Matched	Block
4	2	-122.306426	47.545739	5	17700	17700	1807429	Matched	Intersection

Data Acquisition and Cleaning - Data Cleaning

- Because this project only focuses on severity level, collision type, road condition and light condition, this data set will be extracted and transferred to a new data set. And we will also delete all non-value.

	SEVERITYCODE	COLLISIONTYPE	ROADCOND	LIGHTCOND
0	2	Angles	Wet	Daylight
1	1	Sideswipe	Wet	Dark - Street Lights On
2	1	Parked Car	Dry	Daylight
3	1	Other	Dry	Daylight
4	2	Angles	Wet	Daylight

Data Acquisition and Cleaning - Data Cleaning

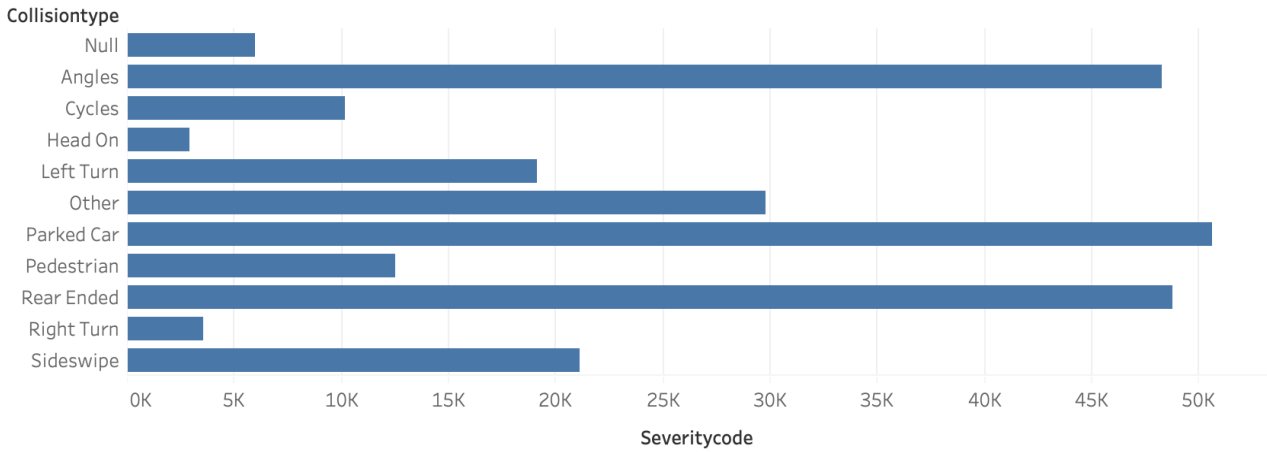
- Besides, we need to transform our categorical data to numerical data in order to make a decision tree. The collision type, road condition and light condition will be transform to numerical data by using ‘preprocessing’ method from ‘sklearn’.

```
array([[2, 'Angles', 'Wet', 'Daylight'],
       [1, 'Sideswipe', 'Wet', 'Dark - Street Lights On'],
       [1, 'Parked Car', 'Dry', 'Daylight'],
       [1, 'Other', 'Dry', 'Daylight'],
       [2, 'Angles', 'Wet', 'Daylight']], dtype=object)
```

```
array([[2, 0, 9, 5],
       [1, 10, 9, 2],
       [1, 6, 0, 5],
       [1, 5, 0, 5],
       [2, 0, 9, 5]], dtype=object)
```

Exploratory Data Analysis

Collision Types and Severity Level:



- As can be seen in this picture, parked car has the highest number of total severity codes, which means parked car is the most common collision types. And rear ended, angles are the following types.

Exploratory Data Analysis

Road Condition and Light Condition:

- We first divide our data into train set, with 70% of the data, and test set, with 30% of the data. The train set has 132595 rows of data, and test set has 56827 rows. And then we define our decision tree which has four level of max depths. Then we train our data with train set and predict our data. The prediction set will be compared with test set to evaluate the accuracy of this model. In this model, our decision tree accuracy is 1.0 which is really high and can predict accurate results.

Results and Discussion

- This project and analysis are quite helpful for the Seattle transportation department. Results show that parked car is the most common type of car accidents and we can use decision tree to predict severity level based on light condition and road condition.
- This could give car companies and insurance companies some ideas to prevent car accidents. For example, car companies should pay more attention on car parking and insurance companies should take road and light conditions in consideration when dealing with a car accident.

Conclusion

- Purpose of this project is to find out which car accidents types has the worst impact and what condition (road and light) could lead to higher severity level. This means to help stakeholders prevent car accidents in the future.
- Final decision based on collision database and could be useful for other stakeholders.