Capstone Project - Car accident severity in Seattle for Fleet Navigation Optimization

Xiaoyini “Ara” Zhang

1. Introduction

Problem: Risk prediction of impact of different factors to traffic accident
Audience: Fleet Services
Solution Provider: Navigation System Software Provider

Traffic accidents have been one of the major threats to fleet services, as fleet vehicles are facing more risks in long-distance commutes, and higher pressure to meet the delivery deadlines. Hereby, when experiencing severe weather and traffic accidents, fleet services are under high risks or may even cause accidents if not well operated. To help fleet companies and their drivers without compromising the high-efficiency that they are trying hard to meet, a navigation company is working on creating a prediction model of how likely different factors including weather, location, driver situation, vehicle types and more.

2. About Data

2.1 Data Source and Overview

The data is provided by **Seattle Police Station** which shows traffic collisions from **2004 to present on a weekly basis**.

The data contains **37 independent variables and 194,673 entries**, and records the situation of the collisions including **weather condition, location, type of collision, vehicles involved, road and light conditions and more**. We will be choosing the impact factors regarding collision situations that will be most applicable for fleet services.

2.2 Selected Variables

By looking at the data, we found that there are multiple parameters recorded for a single accident, including **collision ids and keys, collision data, severity code defined by SPD, collision status and type, location, pedestrian and vehicles involved, weather, road and light condition, speed, and more**.

SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	...	ROADCOND

LIGHTCOND	PEDROWNOTGRNT	SDOTCOLNUM	SPEEDING	ST_COLCODE	ST_COLDESC	SEGLANEKEY	CROSSWALKKEY	HITPARKEDCAR

In this case, the **severity code('SEVERITYCODE')** is the target value we want to predict, which means the level of severity - the more severe it is, the more likely the fleet company should adjust the delivery time window and notify the driver in advance. In this case:

- 3 = fatality
- 2b = serious injury
- 2 = injury
- 1 = prop damage
- 0 = unknown

Then we will choose the independent variables that are critical to predict the collision severity. Among all the factors, we have to pick the ones that **cause the collision** but **not the result of collision**. For example, collision IDs and keys which are for the Police's purpose to keep record should be removed, as well as how many pedestrians and vehicles involved - such details reflecting severity are already part of the target value we want to predict.

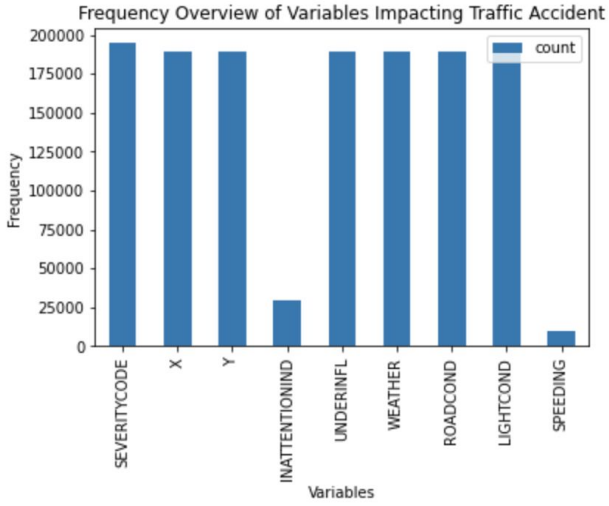
Hereby, we picked the following parameters shown below, and will continue to do exploratory analysis and data cleaning.

SEVERITYCODE	X	Y	INATTENTIONIND	UNDERINFL	WEATHER	ROADCOND	LIGHTCOND	SPEEDING
2	-122.323148	47.703140	NaN	N	Overcast	Wet	Daylight	NaN
1	-122.347294	47.647172	NaN	0	Raining	Wet	Dark - Street Lights On	NaN
1	-122.334540	47.607871	NaN	0	Overcast	Dry	Daylight	NaN
1	-122.334803	47.604803	NaN	N	Clear	Dry	Daylight	NaN
2	-122.306426	47.545739	NaN	0	Raining	Wet	Daylight	NaN

3. Methodology

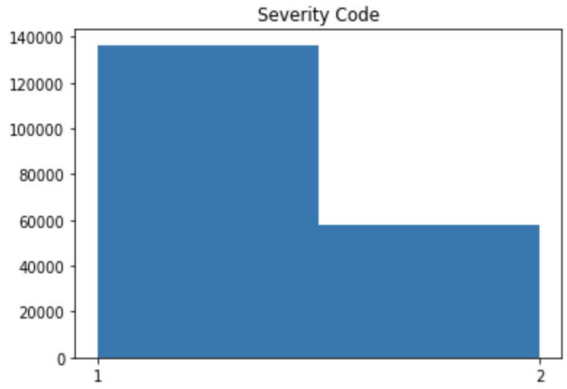
3.1 Data Cleaning and Exploratory Analysis

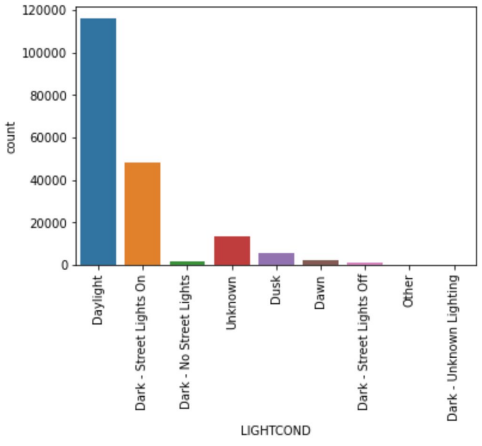
Based on the current given variables, we will further explore the entries to clean up missing variables and check if there're enough valid entries. Among all impacting factors, weather condition, road condition, light condition, and drug influences seem to have more valid data points and variables have enough data to create the model.

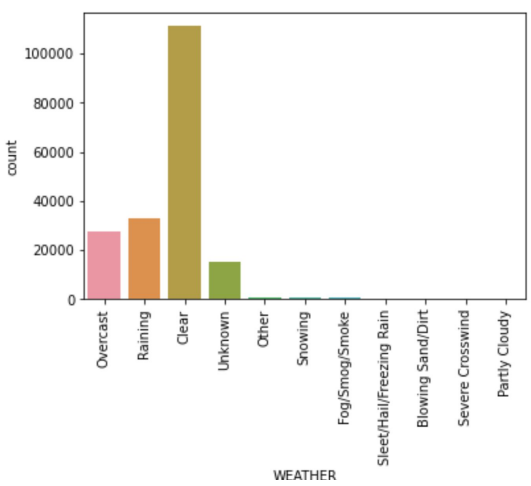


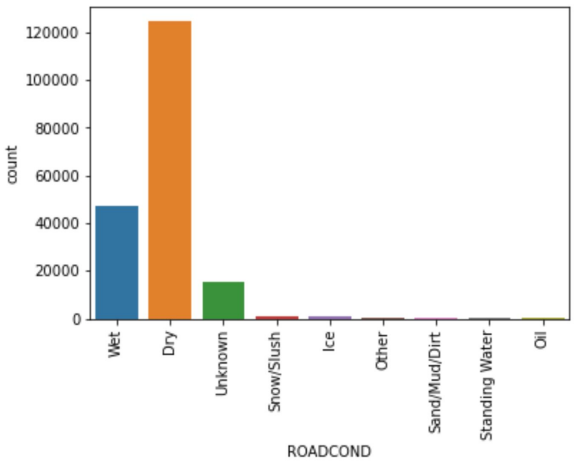
However if we look into the data, we will find the **values very imbalanced.**

In terms of the severity level of the accidents, the current entries contain 1 and 2, which indicates **property damage** and **injury** respectively with much more level 1 severity than level 2. Same thing for other independent variables, data is skewed to mild or non severe conditions, such as daylight, clear weather, or dry road, which and thus will cause the model in the end to predict more level 1 severity accidents (property damage) than level 2.

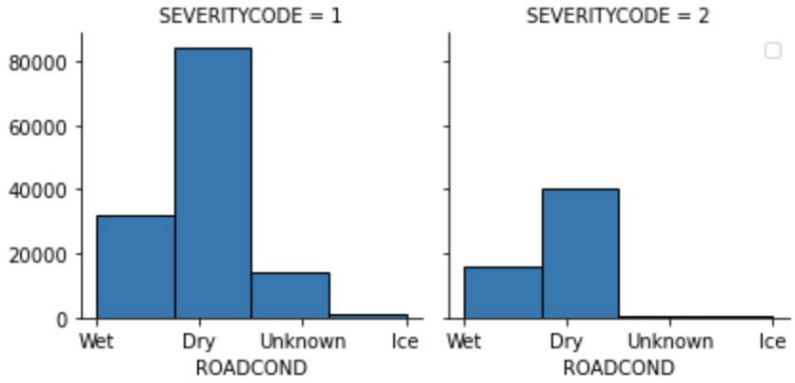








Furthermore, taking severity level within road condition as an example, road condition is also skewed to property damage with prominently high 'Dry' condition impacting the distribution.



On the other hand, in order to eliminate correlation within the variables which may add noise to the final training, we found that weather condition is largely correlated with road condition and devices to leave **"Weather"** only as it is more predictable and can be found and sourced on weather and climate datasets.

	SEVERITYCODE	INATTENTIONIND	UNDERINFL	WEATHER	~~ROADCOND~~	LIGHTCOND	SPEEDING
SEVERITYCODE	1.000000	0.046378	0.044377	0.036330	~~0.023395~~	-0.003625	0.038938
INATTENTIONIND	0.046378	1.000000	-0.025953	-0.007797	~~-0.020218~~	-0.038663	-0.048805
UNDERINFL	0.044377	-0.025953	1.000000	0.017130	~~0.016764~~	0.238641	0.092356
WEATHER	0.036330	-0.007797	0.017130	1.000000	~~0.729870~~	0.185692	0.124187
~~ROADCOND~~	~~0.023395~~	~~-0.020218~~	~~0.016764~~	~~0.729870~~	~~1.000000~~	~~0.186130~~	~~0.163104~~

LIGHTCOND	-0.003625	-0.038663	0.238641	0.185692	~~0.186130~~	1.000000	0.098468
SPEEDING	0.038938	-0.048805	0.092356	0.124187	~~0.163104~~	0.098468	1.000000

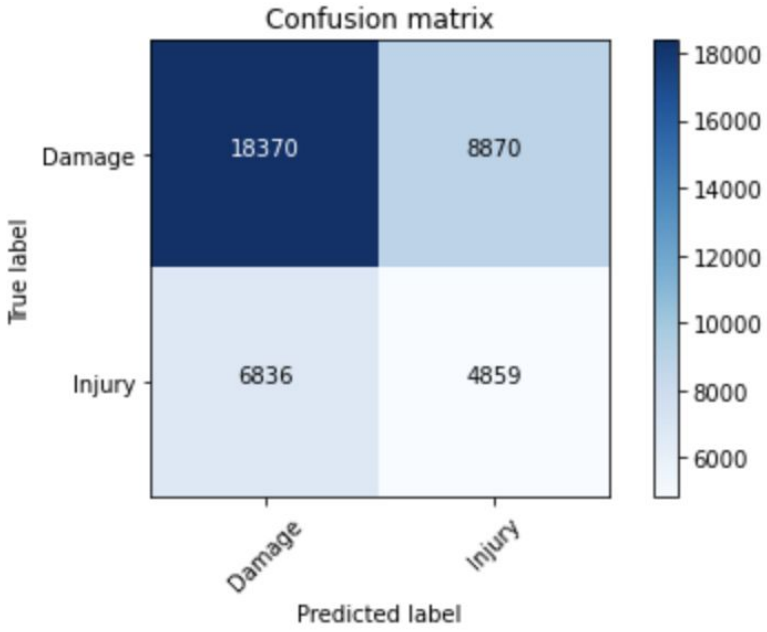
3.2 Machine Learning Methods

Based on the findings above, we will first balance the data and then predict the model. To do that, we categorize different conditions to numerical values and balance the data using SMOTE function.

- Inattention:
  - 1 = Yes
  - 0 = No
- Influence by Drug:
  - 1 = Yes
  - 0 = No
- Weather:
  - 4 = Severe Crosswind, Blowing Sand/Dir
  - 3 = Sleet/Hail/Freezing Rain, Snowing
  - 2 = Raining, Fog
  - 1 = Partly Cloudy, Overcast,
  - 0 = Other Unknown Clear, and missing values
- Light Condition:
  - 3 = Dark No Street Lights, Dark - Street Lights Off
  - 2 = Dark Unkown Lighting, Dark, Street Lights On
  - 1 = Dawn, Dusk
  - 0 = Daylight, Unknown, Others and missing value
- Speeding
  - 1 = Yes
  - 0 = No

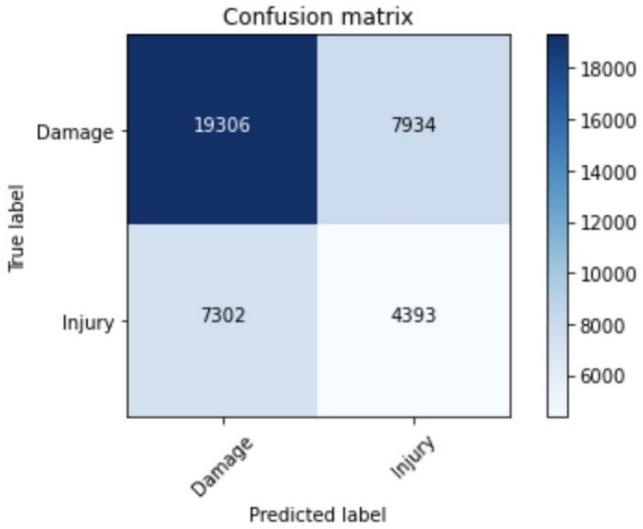
Decision Tree: Accuracy Score: 60%

	Precision	Recall	f1 Score	Support	Jaccard score
Property Damage	0.67	0.73	0.70	25206	0.5391
Injury	0.42	0.35	0.38	13729	



Logistic Regression: Accuracy Score: 59%

	Precision	Recall	f1 Score	Support	Jaccard score
Property Damage	0.73	0.71	0.72	27240	0.5589
Injury	0.36	0.38	0.37	11695	



Based on the result, we found that Decision Tree and Logistic Regression have similar overall accuracy rates, while **Decision Tree** has better balance between the two severity levels.

4. Results

Based on the selected decision tree model, we can then provide probability for accidents based on weather, road and light level before the fleet driver is out for delivery. Based on the predicted value, they will adjust route or delivery time accordingly.

For example, if the driver does not lose attention(0), did not take any drug(0), and not speeding, but the it was snowing that day (3) and road condition is icy(3) and there is dawn time, the model suggested that the severity level is 2 (injury), meaning that proceeding the route under such condition may cause damage to the driver or encounter accident caused by other cars. In this way, the navigation system should notify the driver and customers the potential delay in delivery and adjust the delivery time window accordingly.

5. Discussion

5.1 Supporting Data to Complete the Application

In the real world, simply traffic data will not be the sole source to determine drivers' activity. To apply the traffic accident model to truly benefit fleet services, more data regarding:

- Operational cost
- Historical drivers' data
- Drivers' experience
- And more will be further needed.

For example, another indicator that can be quickly taken into consideration is the experience level of the driver. If the driver is more experienced going on a certain route under poor weather, he/she may be able to quickly find route or others.

6. Conclusion

Based on the analysis and results above, we suggest that majorly focusing on weather and light conditions, and have detailed records on driver's condition on tiredness, healthy level as input of inattention condition and drug condition. Specifically:

- for weather such as **severe Crosswind, Blowing Sand/Dir and freezing rain even snowing**, with **limited lighting in the evening**, there is very likely hood of severe accidents that may cause injury, the navigation system should push notification to fleet drivers and customers for an extended delivery timeframe.
- In order to optimize the more input should be added including
    - Driver's driving record and experience level at the company and route, and

- Operational cost structure of extended deadline

To further determine how much an extended deadline should be further pushed that will have minimum impact on the fleet service' revenue model.