Prediction of car accidents severity

Jorge Méndez Astudillo

1. Introduction

There are many factors causing car accidents such as weather conditions, human factors and human error, and technical problems. Examples of weather conditions are rain and snow. For instance, a lot or rain decreases visibility and makes the road slippery therefore, extra care is needed when driving in this condition. Also, snow makes the driving conditions difficult. Human factors such as distractions, driving under the influence of alcohol and drugs, tiredness or medical conditions can be the cause of accidents. Furthermore, human error can cause accidents. Finally, technical problems of the car such as blown tires can be a cause of accidents.

1.1 Objective

The objective of this project is to build a model to predict car accident severity based on weather conditions. This will allow drivers to decide their best route and the level of alertness needed to drive safely. A data-driven model will be constructed and presented in this report.

1.2 Dataset:

The data has been collected in Seattle, Washington from 2004 to present. The Dataset is provided by Seattle Department of Transportation (SDOT) Traffic Management Division. It contains data from collisions registered in Seattle. The dataset includes the location of the accident (Latitude and Longitude), severity of the accident, collision address type, type of collision, whether pedestrians were involved, number of vehicles involved in the collision, number of injuries, number of fatalities, whether the collision was due to inattention, whether the driver was under the influence of drugs or alcohol, weather conditions during the time of collision as well as road and light conditions during the accident. This dataset has been chosen because of the climate in Seattle includes cold wet winters and warm dry summers, there are rarely temperatures below 0 ºC. Therefore, snow is not a common occurrence.

2. Data exploration

The downloaded data is loaded into a Pandas data frame. The dataframe contains 194673 rows × 38 columns. A new dataframe containing relevant information for this study has been constructed which contains the following columns:

Table 1. Variables to be used in the analysis.

Column name	Description	Values
'SEVERITYCODE'	A code that indicates the severity of the collision	• 3—fatality • 2b—serious injury • 2—injury • prop damage • 0—unknown
'Y'	Latitude	Latitude in deg
'X'	Longitude	Longitude in deg

'ADDRTYPE'	Collision address type	Three values possible: Alley Block Intersection
'SEVERITYDESC'	A detailed description of the severity of the collision	Text describing what happened
'COLLISIONTYPE'	Collision type	Text describing the type of collision
'INCDATE'	Date of the incident	
'SDTO_COLCODE'	State collision code	Numerical code described in metadata
'INNATENTIONIND'	Whether the accident is due to inattention	(Y/N) data will be changed to y=1, N=0 and missing data NaN
'UNDERINFL'	Whether the driver was under the influence of alcohol or drugs	(Y/N/1/0) data will be uniformed Y=1, N=0 and NaN for missing data
'WEATHER'	Description of weather conditions	Values: Blowing Sand/Dirt Clear Fog/Smog/Smoke Other Overcast Partly Cloudy Raining Severe Crosswind Sleet/Hail/Freezing Rain Snowing Unknown
'ROADCOND'	Description of the conditions of the road	Dry Ice Oil Other Sand/Mud/Dirt Snow/Slush Standing water Unknown Wet
'LIGHTCOND'	Description of light conditions.	Dark-No Street lights Dark-Street lights off Dark-Street lights on Dark-Unknown lighting Dawn Daylight Dusk Other Unknown

In all cases, the values coded as ‘UNKNOWN’ or ‘NAN’ will be removed from the data frame. Also, all of the values will be hot-encoded for further analysis.

- The severity of the collision will be related to weather and road conditions. Furthermore, the relation between light conditions and severity of condition will be studied.
- Latitude and Longitude data will be used to map severe collisions and study if there are areas of the city where these kinds of collisions happen often.
- Examples of information that ca be extracted from the data is the amount of high severity incidents ('SEVERITYCODE'=2) happened during rain ('WEATHER'=6). Also, we can study what weather conditions cause the road to be wet using columns 'WEATHER' and 'ROADCOND'.

3. Methodology

The collisions have been mapped and has been no pattern found. A new dataframe containing the columns that will be used in the prediction is made.

This new dataframe is called df_usefull and contains the following data:

SEVERITYCODE, Y, X; ADDRTYPE, SEVERITYDESC, COLLISIONTYPE, INCDATE, SDOT_COLCODE, INATTENTIONIND, UNDERINFL, WEATHER; ROADCOND and LIGHTCOND.

The relations between all variables in the new dataset are studied first by grouping all data in terms of the severity code and normalize the number of occurrences of weather conditions, road conditions and light conditions.

The weather when severe accidents happened is reported in the next graph.

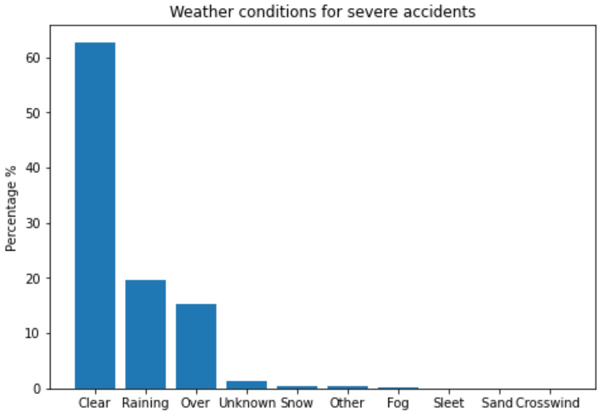

Figure 1. percentage of accidents with injuries during different weather conditions.

According to **Figure 1**, 60% of the accidents with severity code 2, namely with injuries, happened during a clear day. 20% happened during rain and 15% happened when it was overcast. All the other weather conditions did not contribute to accidents with injuries. The road conditions for severe accidents have been reported as:

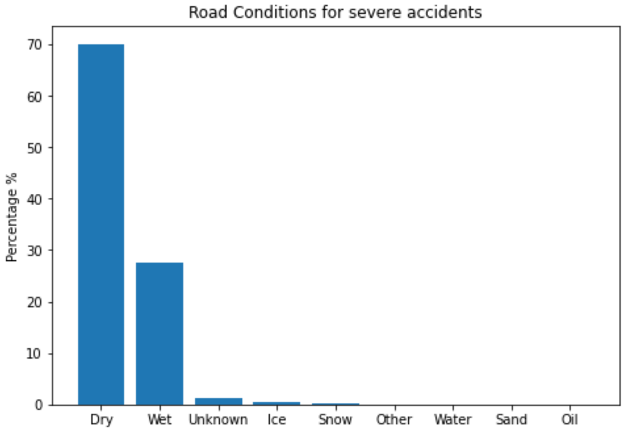

Figure 2. Road conditions for accidents with injuries.

According to **Figure 2,** 70% of severe accidents happened when the road was dry. Around 30% happened when the road was wet and about a 1% happened when there was ice on the road.

Rain causes the road to be wet, that seems to be the major cause of severe accidents. However, most of the accidents happened when the road was dry and the day was clear therefore, the weather did not play a major role in those accidents.

In order to better understand the relation between the variables, it is necessary to see the effect of distractions (INATTENTIONIND) and driving under the influence of drugs or alcohol (UNDERINFL).

Table 2. Relation between inattention and severity of accidents

INATTENTIONIND	SEVERITYCODE	PERCENTAGE
Y	1	65.11%
	2	34.88%

According to **Table 2**, 65% of accidents with low severity were caused because of a driver's distraction. 35% of accidents caused because of driver's distraction had injuries.

Table 3. Relation between underinfl and severity of accidents

UNDERINFL	SEVERITYCODE	PERCENTAGE
0	1	71.8%
	2	28.2%
1	1	59.4%
	2	40.6%
N	1	69.2%
	2	30.8%
Y	1	62.2%
	2	37.8%

It can be seen from **Table 3**, that only 38% of accidents where the driver was under the influence of alcohol or drugs were severe while 62% were not severe.

According to **Figure 1** and **2** and **Table 2** and **3**, severe accidents happened mostly during clear days, when the road was wet. Accidents with a high severity indicator were not caused by inattention of the driver nor was the driver under the influence of drugs or alcohol. Therefore, it is necessary to see the relation between INATTENTIONIND and weather and the relation between UNDERINFL and weather.

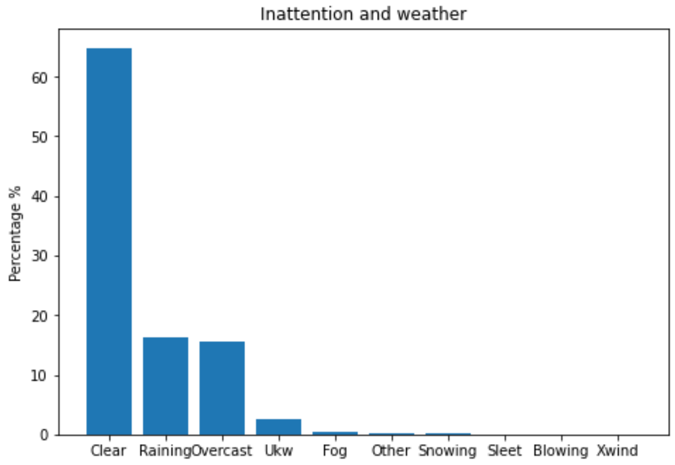

Figure 3. Percentage of accidents due to inattention during different weather conditions.

More than 60% of accidents due to driver inattention happened during clear days. The top three weather conditions when accidents due to inattention happened are: Clear, Raining and Overcast.

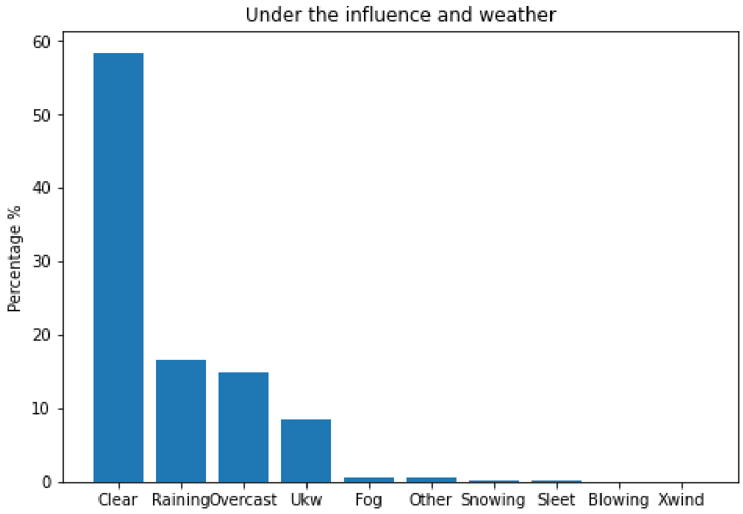

Figure 4. Percentage of accidents that happened during different weather conditions when the driver was under the influence of alcohol or drugs.

Almost 60% of accidents happening when the driver was under the influence of drugs or alcohol happened in clear days. The top weather conditions when drivers under the influence were involved in accidents are: Clear days, Raining and Overcast.

4. Data Analysis

In order to build a model to predict the severity of accidents with different weather and road conditions all categorical values were converted to numerical values. Then, the features 'INATTENTIONIND', 'UNDERINFL', 'LIGHTCOND', 'ROADCOND' and WEATHER have been selected.

Y is the severity code which is the variable we want to predict.

After selecting features, data is normalized using sklearn's preprocessing library.
Finally, the classification is done using the k-means model.

k-means model implementation

In order to implement *k*-means algorithm, a test and a train set are defined using the function train_test_split with 20% of the features used for testing. At the beginning, *k* is set to 3, then, the best *k* is calculated. The best *k* found is k=*6*. Therefore, the *k*-means algorithm is once again implemented with k=6.

A Decision Tree (DT) and a Support Vector Machine (SVM) were also implemented to compare the efficiency of the predictions obtained with *k*-means. DT and SVM are other classification models that could be used for the predition. However, I preferred to use k-means.

5. Validation

The Jaccard Similarity Index and the *f1-score* have been used as validation parameters. Both indicators are calculated by comparing the predicted value with values from test set. The validation has been implemented using sklearn.metrics. The results are shown in **Table 4**.

Table 4. validation of models used for classification

Algorithm	Jaccard score	F1-score
K-means	0.69	0.60
Decision Tree	0.70	0.58
Support Vector Machine	0.70	0.58

A good estimation would yield a high Jaccard Score and high F1-score, the closer to 1 the better.

6. Conclusion

The *k-mean*s model implemented to predict the severity of car crash depending on weather, road conditions, light conditions and whether the driver was distracted or under the influence of drugs or alcohol yielded a good precision as shown in **Table 4**. Although at the beginning it seemed that weather played an important role in causing accidents, it turned out that most serious accident happened when the day was clear, and the road was dry. Therefore, other variables had to be used in the prediction.