Predicting Road Accidents: An analysis of the main factors causing car accidents in Seattle, USA

Luis Terán

1. Introduction

a. Background: Motor vehicle accidents continue to be one of the leading causes of accidental deaths and injuries in the United States. They are responsible for billions in property damage and other economic losses each year. More than 38,000 people die every year in crashes on U.S. roadways. The U.S. traffic fatality rate is 12.4 deaths per 100,000 inhabitants. An additional 4.4 million are injured seriously enough to require medical attention. Road crashes are the leading cause of death in the U.S. for people aged 1-54 (ASIRT, 2020).

b. Problem: In order to identify the cause of the problem, the aim of this project is to identify those factors that influence the most on cars accidents and have a quantitatively estimate of the significance the relationship between the factors and the road accidents.

c. Interest: Even though this is a sample data from Seattle, this behavior patterns in car accidents can be related to other states or even countries with less amount of data of this problem. So this can be used as a feature reference for prevention of car accidents in all other places.

2. Data

a. Acquisition: The original dataset was obtained from the Seattle's government page at:
https://data.seattle.gov/Land-Base/Collisions/9kas-rb8d
The dataset if available for public access. It was created in April 8, 2020 and last update register is from August 27, 2020. Further information of the dataset is available in:
https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf

b. Feature selection: Before selecting the features for predicting it was necessary to state our main prediction objective. We want to know the probability and magnitude of an accident given some characteristics, in this way, the **"VEHCOUNT", "INJURIES", "SERIOUS INJURIES", "FATALITIES"** variable could be helpful for measuring in some way the severity of accident. These columns represent:
- VEHCOUNT: Number of vehicles involved in an accident
- INJURIES: Number of injured persons
- SERIOUS INJURIES: Number of people with serious injuries
- FATALITIES: Fatalities occurred

Those values were added together with different weights to create a new variable that according to the values, was classified into a 4 category variable:
1. Type 1 (0 - 4): No incident or a few car crashes with no injuries or fatalities.

2. Type 2 (5 - 9): One injured person and multiple car crashes.
3. Type 3 (10 - 20): Multiple Injuries or a serious injured person, with multiple car crashes.
4. Type 4 (>20): Sever accident. A lot of injured persons, more than one serious injured person or a fatality occurred

The complete dataset was split into two different dataframes. The first one was for feature selection, those variables that can contribute to an accident were selected and joined to this dataframe. The data description is shown as follows:

Attribute	Data type	Description
INCDATE	Date	Date of the accident
UNDERINFL	Boolean Values: (1: Yes, 0: No)	Whether or not a driver involved was under the influence of drugs or alcohol
WEATHER	Int Values: 1: Clear 2: Raining 3: Overcast 4: Snowing 5: Fog/Smog/Smoke 6: Sleet/Hail/Freezing Rain 7: Blowing Sand/Dirt 8: Severe Crosswind 9: Partly Cloudy 10: Other/Unknown	A description of the weather conditions during the time of the collision.
ROADCOND	Int Values: 1: Dry 2: Wet 3: Ice 4: Snow/Slush 5: Standing water 6: Sand/Mud/Dirt 7: Oil 8: Other/Unknown	The condition of the road during the collisions
LIGHTCOND	Int Values: 1: Daylight 2: Dark – Street Lights On 3: Unknown 4: Dusk	The light conditions during the collisions

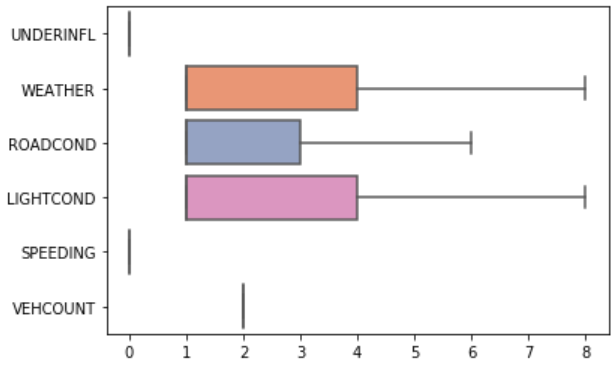
	5: Dawn 6: Dark – No Street Lights 7: Dark – Street Lights On 8: Dark – Unknown Lighting 9: Other	
SPEEDING	Int (1: Yes, 0: No)	Whether or not speeding was a factor in the collisions

On the other hand, there were selected other variables for additional information. The aim of this dataset is for getting a better understanding of the data and the problem.

c. **Data wrangling:** Since the amount of null observations in the data set was not significant for most of the variables, the null observations were deleted without taking into account the "SPEEDING" column. Then for every variable in the dataset, the data was transformed in order to get only integer values:
UNDERINFL: Transformed combined Y/N and 1/0 observations into only 1/0.
WEATHER: Transformed categorical data into integers.
ROADCOND: Transformed categorical data into integers.
LIGHTCOND: Transformed categorical data into integers.
SPEEDING: It was assumed that the missing values refer to the opposite of filled values.

3. Exploratory data analysis:

a. **General description:** Once we have a tidy dataset, we can check how does the dataset variate. We can see that we expect that the most frequent variable are those with lower values, those values were defined un purpose like that for the most common weather, light and road conditions.
Also, we clearly see that some columns have almost for every case the same value. Those columns are:
- Under Influence of alcohol or drugs: Almost all the cases correspond to 0 that means no influence at all.
- Speeding: Almost all the cases correspond to 0, that means no exceed in the speed limit.
- Vehcount: Almost all the cases only imply two cars in the accident.
For these columns there are more values, since there are just a few, those values are marked as outliers and are not shown in this plot.



b. **Severity:** One of the most important data columns in the dataset is the "Severity" column. From the plot below we can see the most of the values range between 0 and 12. Being the 50% of the data between 2 and 6 in the severity scale. This means almost all the values are in the type 1 and type 2 in the categorical scale of severity, this is telling that for most of the cases there are no injuries or just a few of them with no serious injuries or fatalities.
c. **Influence of drugs or alcohol:** From the analysis from the influence of drugs or alcohol, as expected, we detected that in case of accident, drunk or on drug people driving have a 4% extra of probability of having serious injuries and 1% more for death rate according to the observations in the dataset.
d. **Weather**: From the Weather analysis, we can conclude that the most dangerous weather for driving is "Severe Crosswind" (8) because from all the cases registered in this weather, 11.11% of cases ended up in fatalities. The least dangerous weather is "Partly cloudy" with only minor injuries in car accidents.
e. **Road conditions**: In general, it seems that the road conditions are not very related to fatalities since we see very small proportion related to it. From the dataset of Seattle, the most sever accidents happened in "standing water" roads. On the other, surprisingly most of the cases occurred on dry roads.
f. **Light conditions**: In the case of light conditions, the observed data for all cases was approximately the same. From all the features listed this is the least significant to the prediction.
g. **Speeding**: Finally, for the speed exceeding cases, just as the influence of alcohol or drugs, the percentage of serious injuries and fatalities grows 1% and 4% respectively.

4. **Predictive Modeling:** Before starting with the data modeling, a preprocessing of the data is needed. For this stage, the x and y matrix are declared for the data split. We've chosen to use the 85% of the data for training and 15% for testing. The data was standardized for algorithm's convergence purposes. Different algorithms were applied with different set of parameters for finding the best fit of the data:
    a. **K - Nearest Neighbor classifier:** The KNN algorithm was applied for different k neighbors from 1 to 10. The best accuracy was reached with 5 nearest neighbors with an accuracy of 0.6844.
    b. **Decision Tree classifier:** The decision tree classifier was applied for different depths between 1 and 10. For all the values the f-1 score and Jaccard index were approximately the same, the variation between them is not significant so we can choose any value of depth for the best model with and accuracy of 0.6867.
    c. **SVM classifier:** The SVM classifier was applied for different kernel functions (sigmoid, poly, rbf, linear). The f-1 score showed that the "Sigmoid" function reached a better performance with an accuracy of 0.571.
    d. **Logistic regression:** For the logistic regression a grid search of parameters was performed for different values of alpha (1,0.1,0.01,0.001,0.0001) and solvers ('liblinear', 'newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'). The value of alpha = 1 and the 'libinear' solver were the best fit of the data reaching a f-1 score of 0.5593.
5. **Conclusions:**

From all the models above, the Decision Tree classifier was the most accurate classifier, also this algorithm is chosen for its interpretability, also is the fastest from the shown here. Since all the depth levels have almost the same accuracy, the depth chosen level is for. From the plot bellow we can estimate what are the most important factors in the classification category, hence the most significant variables in the severity of the accident. The first 4 factors in order of significance are:

1. Road conditions
2. Light conditions
3. Influence of alcohol or drugs
4. Weather

Then, two probability predictions were performed for two completely different scenarios. One optimistic scenario and one pessimistic scenario. These probabilities represent that, in case of an accident, what is the probability of getting into the different types of severity crashes. The input is created from answering the following survey:

Finally, two cases are predicted as an example of the interpretability of the model of probabilities of the accident severity and how it is built. From the input, one person is able to estimate an expected probability for driving depending on the conditions of the environment for that particular case. The corresponding input is built from the answers of the following questions.

a.- ¿Are you drunk or under the influence of drugs?

1. Yes
2. No

b.- What's the weather like?

1. Clear
2. Raining
3. Overcast
4. Snowing
5. Fog/Smog/Smoke
6. Sleet/Hail/Freezing Rain
7. Blowing Sand/Dirt
8. Severe Crosswind
9. Partly Cloudy
10. Other/Unknown

c.- How would you define the road conditions?

1. Dry
2. Wet
3. Ice
4. Snow/Slush

5. Standing water
6. Sand/Mud/Dirt
7. Oil
8. Other/Unknown

d. What are the light conditions?

1. Daylight
2. Dark – Street Lights On
3. Unknown
4. Dusk
5. Dawn
6. Dark – No Street Lights
7. Dark – Street Lights On
8. Dark – Unknown Lighting
9. Other

e.- Are you in a hurry? Is it possible that you exceed the speed limit during your travel?

1. Yes
2. No

The results from the optimistic scenario and the pessimistic scenario are shown below:

Optimistic Scenario	Pessimistic scenario
Input: • No alcohol or drug influence • Partly cloudy weather • Dry road conditions • Daylight • No speed limit exceeded.	Input: • Alcohol or drug influence • Fog/Smog/Smoke • Standing water road conditions • Dark - street lights off • Speed limit exceeded.
Output prediction: • Type 1= 66.79%, • Type 2= 24.21% • Type 3=7.96% • Type 4=1.03%	Output prediction: • Type 1= 45.67%, • Type 2= 31.18% • Type 3=15.84% • Type 4=7.28%

6. Discussion

a. Even though the prediction model gives reasonable results, the accuracy can be improved by changing different forms of measuring the severity of an accident since it's really hard to find a way of predicting the severity of a road accident.

b. A better prediction can be made by getting more data. Also getting more data of no incident drives.

c. Using a different scale for the severity of a possible accident, different to the created here, could help for the accuracy of the predictions and a better understanding of the data.

7. References

- ASIRT, 2020 (Association for Safe International Road Travel): Annual United States Road Crash Statistics https://www.asirt.org/safe-travel/road-safety-facts/