Classification of Traffic Accident Severity

1. Introduction

Road safety is a primary concern for drivers since traffic accidents can be fatal. Without data, drivers rely on their intuitive perception of the environment to determine the level of cautiousness when they drive, which can be inaccurate at times. For example, drivers may underestimate the danger of speeding on a wet road even after the rain has stopped or overestimate the risks of driving with light fog.

Therefore, it is important to educate drivers with correct road safety knowledge that is backed-up by empirical evidence, to boost driving safety and confidence. A multi-class model that classifies accidents of various degrees of severity according to different road conditions can inform drivers of the relevant risks so that they can adjust their level of cautiousness when driving.

2. Data

2.1. Overview

A multi-class model can be used to classify the various degrees of severity. The features used to classify would be variables such as road condition, light condition, weather, speeding (or not), and the lane the driver is in. For example, consider a driver who is driving under 'wet', 'dark, streetlights on', 'not speeding', and on a lane. A trained model, when if an accident does occur, can classify the severity of the accident. This would provide the driver with the 'worst-case scenario', rather than a probabilistic estimate of an accident occurring. This can still have the effect of inducing an appropriate level of cautiousness in the driver.

2.2. Feature selection

The variables available in the data set is first inspected. The direction of collisions are deemed to be irrelevant in this investigation for two reasons. The first is that there are too many too account for and it might lead to non-convergence of my models. Secondly, it does not serve to warn drivers before an accident happens i.e. angle of

collision is determined near the time of impact, and by then, drivers cannot react to information of severity. The features I will be using for classification can be split into two broad categories: human and environment. The human conditions are: Speeding, whether accident due to inattention, and whether driver under substance influence. The environmental conditions are: Light conditions, road conditions, and weather conditions. These were selected primarily due to their intuitive relevance in determining the severity of an accident. For example, it is clear that if the driver was speeding, it is likelier for the accident (if it happens) to be more sever. Another example is that wet roads reduce the frictional index, which increases the force of collision.

2.3. Data Pre-processing

It is observed that the human conditions are binary and have "nan" values. Both 'Yes' and 'No' labels are required in such a case, and thus, the "nan" values have been replaced by the opposite value. For example, the Speeding condition only has 'Yes' values, so the 'nan' values have been replaced with 'No'. Next, the number of "Yes' values were tallied to ensure that there is enough variation in the data. As for the environmental conditions, which have many categories each, the one-hot encoding technique was used to generate dummies for each category. Similarly, the number of positives were tallied to ensure there is enough variation. Lastly, the remaining 'nan' values were dropped, as their presence in non-binary conditions cannot be attributed to either a positive or negative case with certainty.

3. Methodology

3.1. Exploratory Analysis

To first find out whether severity of accident indeed do vary between human and environmental condition, visualizations are relied on. The data is first grouped by the condition, and the proportions of the levels of severity are computed for each category of the condition. The differences in proportion can then tell whether severity is varying in conditions. Take the 'WEATHER' condition for example. There are many categories of weather, and in the bar plot below, we can see that there is some variation in proportions.

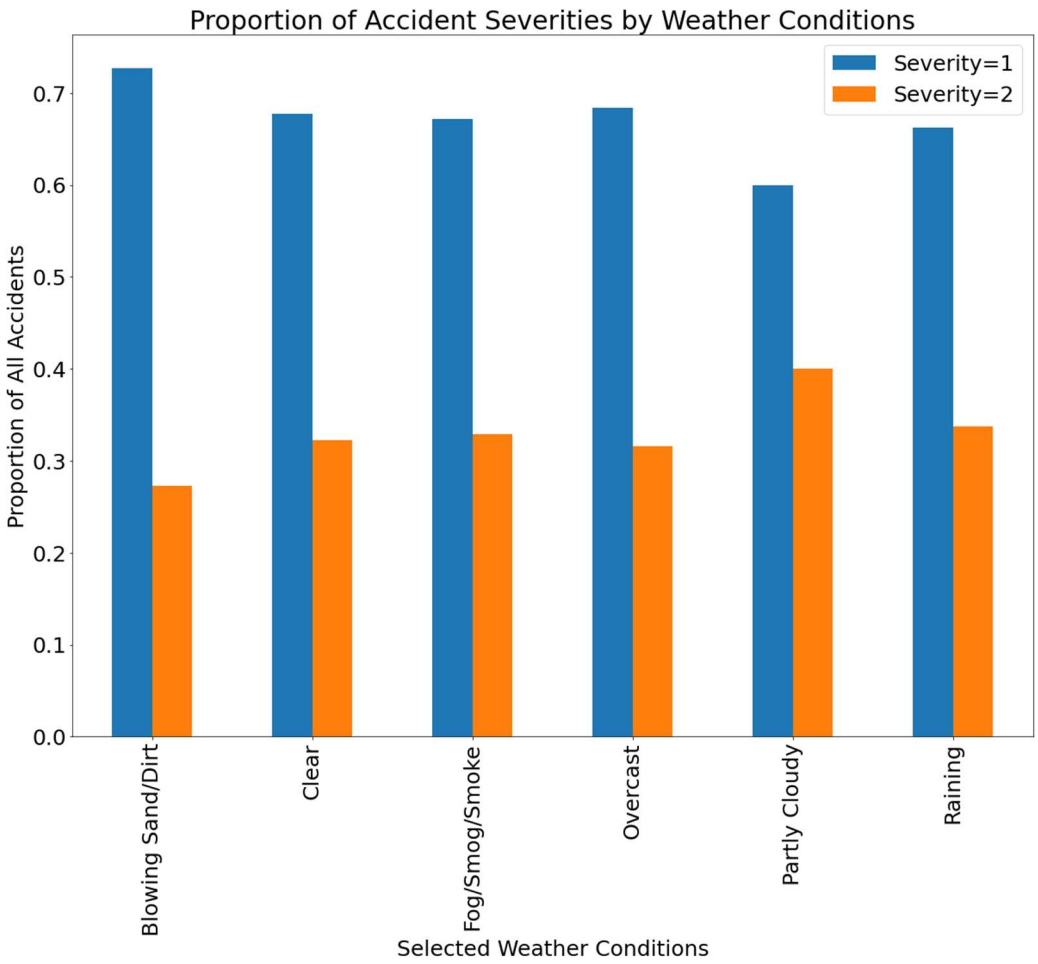

Fig.1 Bar plot of proportion of severity by categories of the Weather condition.

We can see that of all the accidents that have occurred, if the weather had been 'Partly Cloudy', the proportion of the more severe accident (level 2) is *higher* than when the weather had been "Blowing Sand/Dirt". This could be due to the fact that drivers are actually more careful when there are more extreme weather conditions, and more careless when the weather is good. We can also tell the degree of the impact in general from this plot. For example, having 'Fog/Smoke/Smog' has a lesser impact on severity than 'Partly Cloudly'. This again supports the hypothesis that drivers are more cautious under more extreme conditions. This is evidence that some environmental conditions do have an impact on severity. In the plot below, we can also see evidence of impact by the human condition.

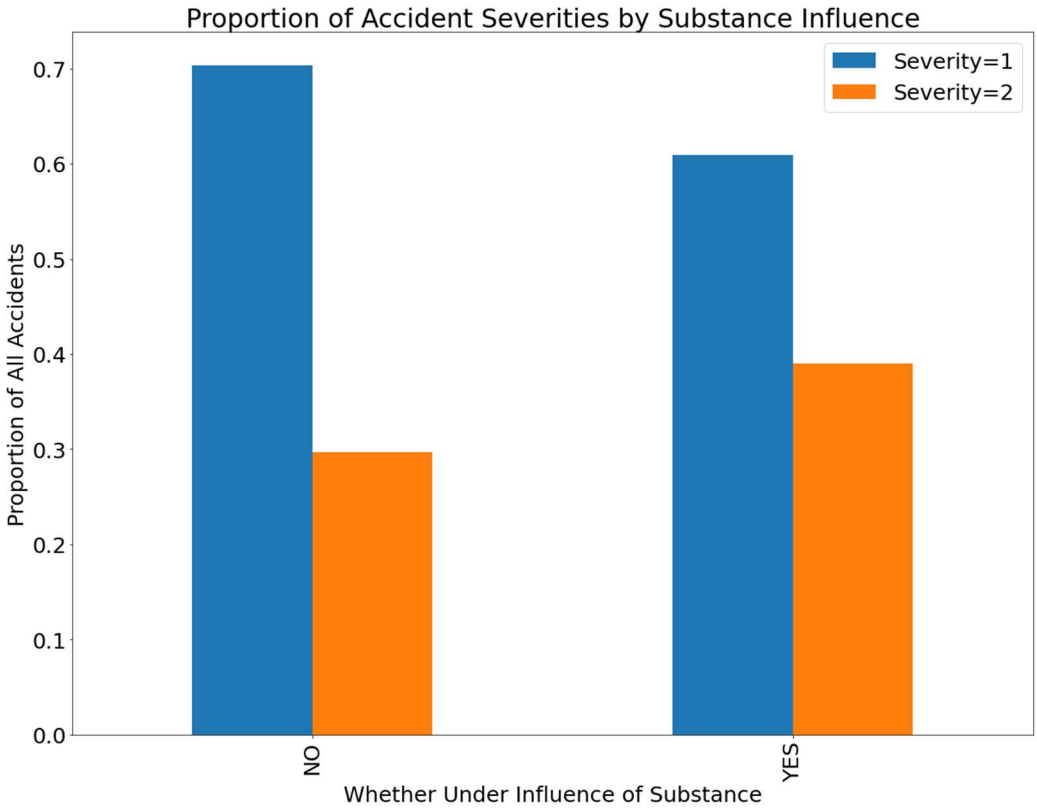

Fig.2 Bar plot of proportion of severity by categories of the human condition of substance-influence.

We can clearly see that of all accidents, the proportion of level 2 severity is *higher* if the driver had been under the influence of substances. This is evidence that the human condition can impact severity of traffic accidents.

3.2. Description of Machine Learning Approach

Having shown some initial evidence of impacts of the human and environmental condition on severity, we now need a more quantitative method for determining such relationship. As mentioned in the introduction, A multi-class model that classifies accidents of various degrees of severity according to different road conditions can inform drivers of the relevant risks so that they can adjust their level of cautiousness when driving. I utilize the Decision Tree and Logistic Regression methods. I first determine the best parameters for each method, then compare between the two to select the final method that will be used to generate results.

Selection is based on the accuracy score and F1 score criteria. These scores are generated by evaluating the models against a test set. To meet this requirement, the data was initially split into a training and testing set, where the testing set comprise of 20% of the entire processed dataset.

3.3. Decision Tree

The "maximum depth" parameter is tested iteratively to determine the 'best' Decision Tree model. The range of the parameter that is tested is 1 to 10. For each value of the parameter, the model performance against the test set in terms of accuracy score and F1 score is recorded.

	MaxDepth	AccScore	F1Score
5	6	0.700380	0.578062
7	8	0.700275	0.578107
6	7	0.700248	0.578095
8	9	0.700248	0.578191
4	5	0.700195	0.576851
0	1	0.700169	0.576692
1	2	0.700169	0.576692
2	3	0.700169	0.576692
3	4	0.700143	0.576679
9	10	0.699958	0.578242

Table 1. Performance metrics of the Decision Tree model against testing set for each value of the "Max Depth" parameter, ordered by descending accuracy score.

We can see that the scores are not significantly different. Thus, we should be indifferent to the value of the Maximum Depth parameter. The value of 5 is arbitrarily chosen to compete against the Logistic Regression model.

3.4. Logistic Regression

The "solver" parameter is tested iteratively to determine the 'best' Logistic Regression model. The range of the parameter includes "newton-cg", "lbfgs",

"liblinear", "sag", and "saga". For each value of the parameter, the model performance against the test set in terms of accuracy score and F1 score is recorded. In addition, the log-loss is also recorded.

	Solver	AccScore	F1Score	LogLoss
0	newton-cg	0.700169	0.576741	0.587845
3	sag	0.700169	0.576741	0.587846
4	saga	0.700169	0.576741	0.587846
1	lbfgs	0.700169	0.576741	0.587846
2	liblinear	0.700169	0.576741	0.587892

Table 2. Performance metrics of the Logistic Regression model against testing set for each value of the "solver" parameter, ordered by descending accuracy score.

Similar to the Decision Tree model, we can see that the scores are not significantly different. Thus, we should be indifferent to the value the type of solver used. The default solver is arbitrarily chosen to compete against the Decision Tree model.

3.5. Final Model Selection

The two methods are compared and selected based on the maximum score criterion.

	Algorithm	F1-score	Accuracy
0	Decision Tree	0.576851	0.700195
1	LogisticRegression	0.576741	0.700169

Table 3. Comparison of performance metrics between the two approaches.

Once again, the scores do not differ significantly. I chose the Logistic Regression model since it has the capability to predict probabilities of a severity instead of merely categorizing them. This can be justified as probabilities give a clearer picture to drivers by telling them how likely they will encounter a severe accident.

4. Results

To demonstrate what the model tells us about the conditions and severity of accident, and to show how the model can provide real-time information to drivers, artificial datasets were fed to the model to make predictions. For each of the three human conditions, the influence of drugs while driving is used to demonstrate the results. Scenarios are constructed by changing a few selected categories of the environmental conditions. The dark-wet-raining combination can be seen as the worst environment combination, and daylight-dry-clear is the best case. Only a few combinations in-between the best and worst-case are compared.

Human Condition: Drug Influence **'Yes' vs. 'No'**	**Environmental Condition: Light** **'Dark' vs. 'Daylight'**	**Environmental Condition: Road** **'Wet' vs. 'Dry'**	**Environmental Condition: Weather** **'Raining' vs. 'Clear'**	**Classification of Severity** **'Level 1' vs. 'Level 2'**
No	Dark	Wet	Raining	1
No	Daylight	Wet	Raining	1
No	Dark	Dry	Clear	1
No	Dark	Wet	Clear	1
No	Daylight	Dry	Clear	1
Yes	Dark	Wet	Raining	2
Yes	Daylight	Wet	Raining	2
Yes	Dark	Dry	Clear	2
Yes	Dark	Wet	Clear	2
Yes	Daylight	Dry	Clear	2

Table 4. Predicted classification of severity by the model under selected combinations of human and environmental factors.

Notice that regardless of the combination of environmental factors, the determining factor of severity is the human condition of drug influence while driving. Otherwise, we should see variation in severity categorization when the combination changes.

5. Discussion

5.1. Discussion of Results

As mentioned above, the determining factor of seems to the drug influence while driving. This is an important insight as it provides empirical evidence against the use of substances such as drugs or alcohol and driving. So if the driver values his life and the life of others, he should refrain from substance use when driving. On the other side of the coin, the results provide empirical evidence that environmental factors do not necessarily increase the potential severity of accidents when they happen.

5.2. Discussion of Limitations

There might exist potential confounders to be able to make claims about the flip-side of the results. That is, as mentioned earlier, drivers may have a natural response to the environment around them i.e. being more careful when road is wet or when it is dark. Since 'cautiousness' is not an input into the model, we may falsely conclude that they do not impact severity of accidents as seen in the results.

The other concern is that only two methods were compared in determining the best model. The K-nearest-neighbors and Support Vector Machine algorithms could be better candidates for example. However, they were excluded as they did not converge in multiple attempts and under different parameters.

6. Conclusion

The model provides empirical evidence that driving while under substance abuse increases the severity of an accident if it happens. Thus, it advises against the use of substance while driving. On the other hand, the results show that environmental factors such as weather, road, and lighting conditions do not impact severity of accident. However, this result could have been biased by not accounting for the driver's natural cautiousness under different environmental conditions. Lastly, the model can be deployed in a real-time setting where it warns a driver of his risks given inputs from sensors about his driving environments.