Predicting severe accidents using machine learning approaches

1.Introduction

According to the CDC, road accidents are the third top reason responsible for human mortality, so much so that 170 thousand people died of accidents in the US during the year 2017. The WHO statistics also show that road accidents are one of the top 10 reasons responsible for human mortalities. In fact, according to WHO, the counts of road accident-related mortalities have risen up to 2 million people in 2016 (ranked 8th cause of death) from the year 2000's 1.5 million people (ranked 10th cause of death). Many companies and organizations are trying to reduce the fatal injuries of road accidents. These include: urban planning organizations, ministry of health, Police departments, manufacturers of self-driving and non-self-driving cars (i.e. BMW, Tesla, etc.)

As a Data scientist, I believe a model that predicts road accident severity based on the given appropriate attributes could improve and/or possibly revolutionize these companies/organizations' actions. Below are examples of how such a predictive model could help these companies/organizations.

Urban planning organizations: such an organization could check the safety and quality of the urban roads by giving the road's attributes to the model and receive predictions. The roads whose predictions returned high severity of injuries could be prioritized for rethinking and repair.

Car manufacturing company: These companies could use the cameras and other tools implemented in them to measure features of the road, weather, etc. and feed the data to the model and receive predictions. If predictions are indicative of imminent severe accidents, the car could do multiple things.

For instance, a self-driving car would reduce its speed, increase the confidence threshold for a lane-changing decision. It could also notify the driver about the danger. As mentioned above an accident severity predicting model could help different industries.

2.Data

2.1 Data summary

Appropriate data for making a predictive model has requirements. For example, it needs to be labeled, and it needs to have proper features.

The dataset I'm using here is a collision dataset provided by SPD and recorded by Traffic Records. This dataset keeps track of the road accident in Seattle from 2004 till the present. This dataset, which is updated weekly, has almost 200 thousand rows and 38 columns. Our target/dependent variable is one of those columns and is called SEVERITY. Concretely each sample/row will have 37 independent variables and include the weather situation, the quality/quantity of the vehicles, and pedestrians involved. However, not all of these attributes will be used for training the model.

2.2 Data cleaning

As mentioned before, some of the features are totally unnecessary for our project. For instance, the intersection unique key (denoted INTKEY in the Dataset) could be helpful if we were using this model to predict collision severity in Seattle but not if we are using it to predict collision severity in New York.

The same goes for coordinates of the area where the accidents happened. Many other features were removed too, a list of which can be found in the project report.

Some features required moderate and some others require complex processing of the data. In this project, I dropped the severity Description (denoted SEVERITYDESC) and the similar columns because these columns contain human's natural language text, and cannot be translated into computer language(i.e. vectors of numbers) without the NLP machine learning technique called sentiment analysis, which is too complex of analysis and was out of scope for my project.

Examples of features that required low to moderate modification included the Speeding and inattentionID columns. These columns, all of which were of Yes/No type, had multiple representations for either Y/N. In some columns, Y was represented with the number "1", and N was represented by "0" or "numpy.nan".

Finally, the samples/rows that had NaN values were removed since guessing the NaN values using available techniques would simply lower the dataset quality as we already had a big

enough dataset. Now we can have a look at the final dataset that will be fed to our model.

OUNT	PEDCYLCOUNT	VEHCOUNT	UNDERINFL	PEDROWNOTGRNT	SPEEDING	HITPARKEDCAR	Alley	Block	...	Wet	Dark - No Street Lights	Dark - Street Lights Off	Dark - Street Lights On	Dark - Unknown Lighting	Dawn	Daylight	Dusk	Other
0	0	2	0	0	0	0	0	0	...	1	0	0	0	0	0	1	0	0
0	0	2	0	0	0	0	0	1	...	1	0	0	1	0	0	0	0	0
0	0	3	0	0	0	0	0	1	...	0	0	0	0	0	0	1	0	0
0	0	3	0	0	0	0	0	1	...	0	0	0	0	0	0	1	0	0
0	0	2	0	0	0	0	0	0	...	1	0	0	0	0	0	1	0	0

3. Methodology

3.1 Feature selection

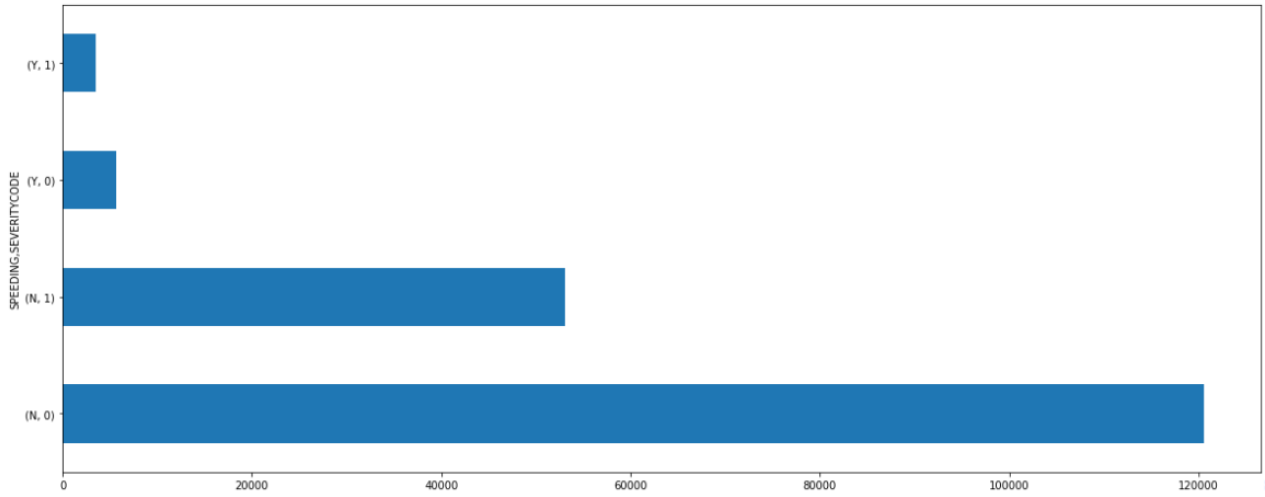
At this point, I needed to analyze each and every one of the features that existed in the final "clean" dataset in order to see if each feature was eligible for predicting accident severity. I came up with the formula "severe accident count / non-severe accident count". This formula could be used to find out if different subtypes of a categorical variable had different impacts on the accident severity. In other words, if all of the different subtypes a categorical variable had almost the same impact on the accident severity, there would be no benefit in categorizing the variable. The same formula could be used for numerical variables but instead of calculating the ratio of

severe to non-severe for each subtype, I divided the numerical values into bins and then checked if different bins made different impacts on the accident severity.

Below, are a couple of visualizations that demonstrate the intuition behind the formula.

$$Severity\ likeliness \approx \frac{count\ of\ severe\ accidents}{count\ of\ all\ accidents}$$

The equation above is the formula we mentioned before. If different subtypes/bins of a variable output different values of "Severity Likeliness", it is eligible for being used by the model.

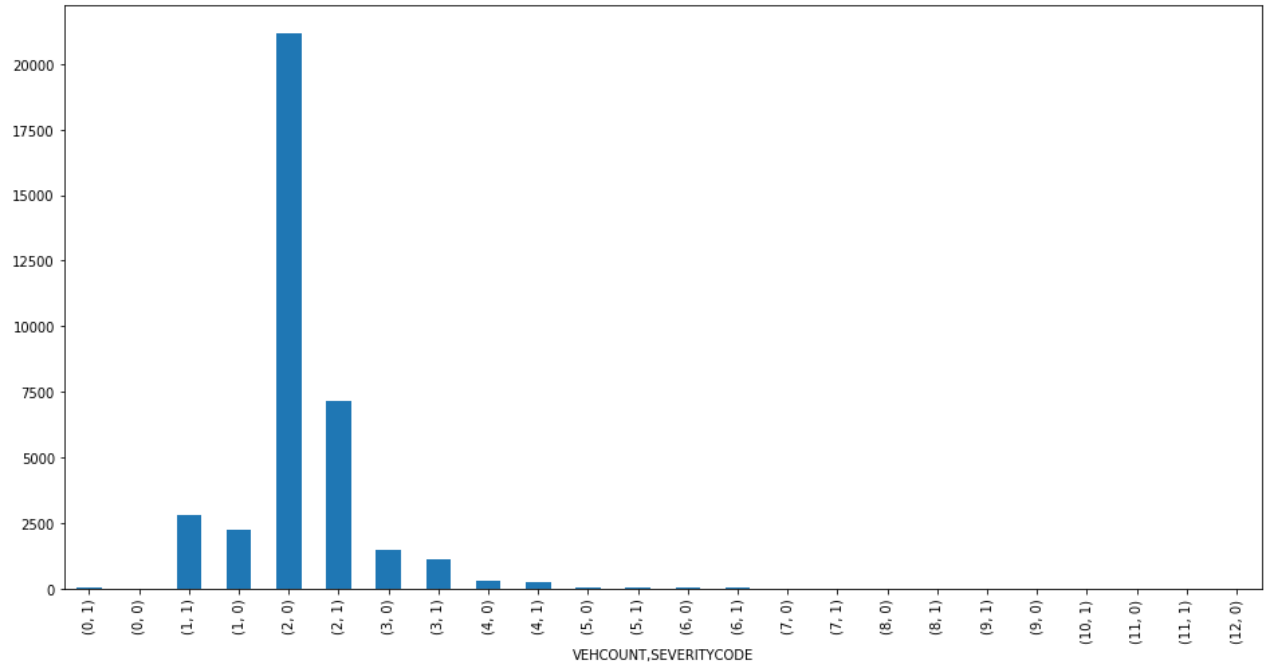


SPEEDING	SEVERITYCODE	
N	0	120569
	1	53133
Y	0	5701
	1	3492

The image above shows counts of accidents for each subtype of the variable "SPEED".

As we can see, the count of severe cases while the driver was not speeding is less than half of the minor ones. On the other hand, when the driver was speeding, the count of severe accidents is more than half of non-severe ones. This means that speeding increases the likelihood of having a severe accident in case one happens. Notice how comparing counts of only severe accidents when speeding to when not speeding is not a good

way to evaluate features as it gives us wrong insight.



In the plot above, we can see that different values of "VEHCOUNT" lead to very different severe/non-severe ratios. Notice how the length of bar pairs for values 2 and 3 of "VEHCOUNT", significantly differ.

3.2 Modelling

After the feature selection process, we can choose a model to fit our data. As I said in the introduction, the predictive model

can be used by car manufacturers. If a product is meant to be used by customers we need to ensure that the quality of our product is good enough. In our case, we need to alarm the driver only when our model is quite sure about its prediction. In order to be able to access the confidence score, we need a model that provides us with probability values. Concretely, our model of choice for this project is Logistic Regression.

I used GridCV to find the best combination of parameters for our model training process. In addition, I trained the model with different datasets, the results of which are listed below in the "Results" section.

4.Results

I used both balanced and unbalanced versions of the dataset and it turned out that although the model trained with the balanced dataset had lower test accuracy compared to the one trained with the unbalanced dataset, it was able to reach twice the f1 score of the model trained with unbalanced data which means that overall, the model trained with the balanced data set performed much better. Moreover, we do not see any signs of overfitting. The table below shows a detailed description of the results I got.

click to scroll output; double click to hide

balance	f1 test	f1 train	acc test
balanced	0.73	0.7296	0.70
unbalanced	0.43	0.4280	0.75

Notice how despite the large dataset, the unbalancedness caused the model to do a less than mediocre job at predicting accident severity.

5.Discussion

While my model was able to learn a great deal of information from the dataset, I believe there's a lot of ways in which the dataset can be improved. For example, I believe that if the data was more specific in terms of technical info, and less specific in terms of physical location info, my model could've been able to learn much more and thus perform much better. I'll name some examples. If there was a column that kept the Absolute amount of speed rather than holding a Yes/No, it could help a great

deal. Another improvement that can be made is that instead of submitting the accident locations based on their coordinates or their unique ID, the data collector and the Police could cooperate and gather more info about each intersection (i.e it's dimensions, asphalt quality etc.).

With all that said, I have a few suggestions for the people/organizations that want to create a predictor model using this dataset and then deploy it for the city of Seattle. Those people could use Seattle specific features like unique IDs of intersections and incident descriptions dictionary of Washington state to make a model that performs better in Seattle.

Ultimately, I believe that any researcher with proper equipment could create ANNs that can solve these problems much better than simple models like Logistic regression can. Unfortunately, using the so-called techniques was out of scope for our project.

6.Conclusion

Although the model we made was not able to perform as well as we expected, it learned a great deal about how features correlate with accident severity. I think that using the predictions of my model (those predictions that were made with high levels of confidence) could help the drivers of self-

driving vehicles to drive more safely and thus the rate of severe road accidents could be significantly reduced.

7.Refrences

Blogpost:
https://www.linkedin.com/pulse/new-ai-can-predict-accident-severity-mehregan-karbasi/