Predicting Severe Accidents in Seattle City

Batuhan Omer

September 2020

1.Introduction

Every year, roughly 1.3 million people die in car accidents worldwide – an average of 3,287 deaths per day. Road traffic crashes cause up to 50 million injuries globally each year. On the other hand, the use of smart cars is increasing each year. Wouldn't it be great if your car would warn you, given the weather and the road conditions about the possibility of you getting into a car accident and how severe it would be, so that you would drive more carefully or even change your travel if you are able to?

We tried to develop a machine learning model in an effort to reduce the amount of severe car accidents. The model predicts the severity of car accidents, given weather, road and lightning conditions. If the severity that is predicted by our model is high, than it should warn the driver to drive more carefully.

2.Dataset

We used a dataset for car accidents that was recored at Seattle City. You can download the Dataset by Clicking here. You can also find the Metadata by Clicking here. This dataset had many features that we did not need. Unfortunately, some good features, like if the driver was inattention-ed, had too many missing values. Therefore we dropped everything we did not need and we could not use.

	SEVERITYCODE	WEATHER	ROADCOND	LIGHTCOND
0	2	Overcast	Wet	Daylight
1	1	Raining	Wet	Dark - Street Lights On
2	1	Overcast	Dry	Daylight
3	1	Clear	Dry	Daylight
4	2	Raining	Wet	Daylight

We ended using only 4 features, the severity code, weather condition, road condition and light condition. The severity of each accident was encoded as followed:

1. Property Damage Only Collision
2. Injury Collision

The unique weather condition values in the provided dataset were:

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

The unique road conditions in our data are:

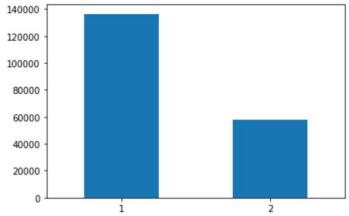
1. Dry
2. Wet
3. Standing Water
4. Sand/Mud/Dirt
5. Snow/Slush
6. Ice
7. Oil
8. Unknown
9. Other
10. NaN

Finally the unique lightning conditions were:

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

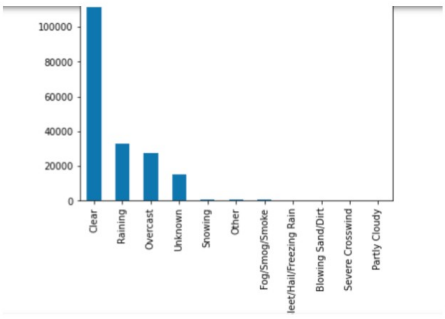
3.Visualization of the Data

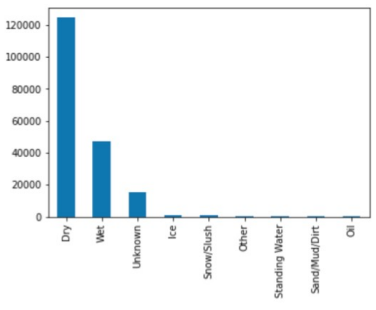
Our next step was to get a better understanding of our data, so we visualized it.

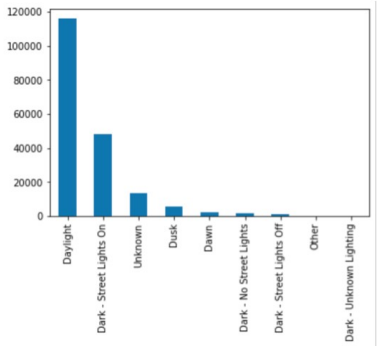


In this figure we plotted the severity code. As seen clearly our dataset was not balanced so we had to balance it later on. As our target value was the severity code

Next we plotted our three independent variables, the weather, road, and lightning condition.







Surprisingly, most of the accidents happened on good conditions. We could assume that most people lose their attention or drive carelessly when the conditions are good. On the other hand when the condition are bad most of the drivers drive carefully. In addition, we also visualized these conditions (the figures are not included here) for the severe accidents and the results were similar.

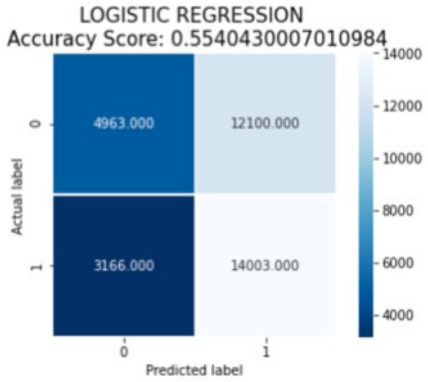
4.Building the Models

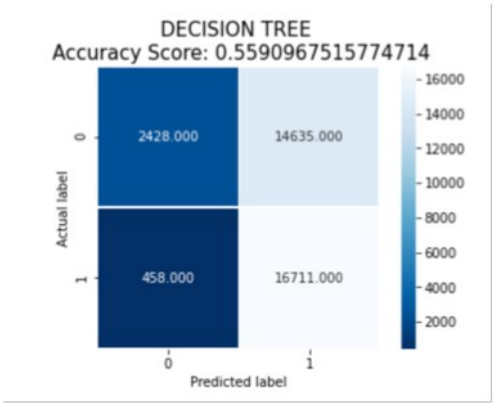
Before we started building our models we had to preprocess our dataset and make it ready for use. First we had to decide what we were going to to with the values NaN, Other and Unknown. We decided to drop all rows that included this three values. The reason was that there were not many values that were NaN and Other. Also most of the unknown values were associated with non severe accidents and as our data set was imbalanced we did not find a reason not to drop it. Just to note, we could have have merged the values unknown and others and have filled all missing values.

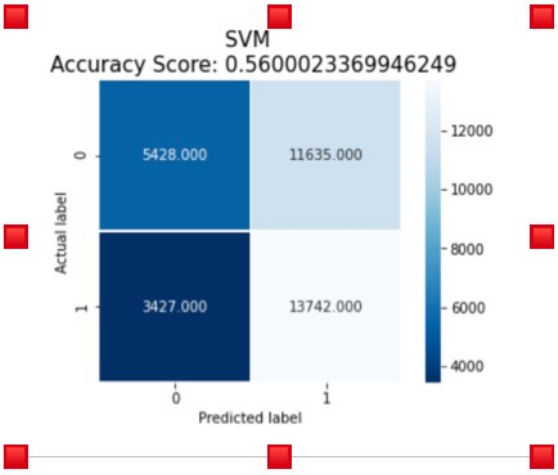
Next we perfectly balanced our target variable and also encoded all independent variables. The encoding follows the lists in the data section of this report. Finally our dataset was ready to be used.

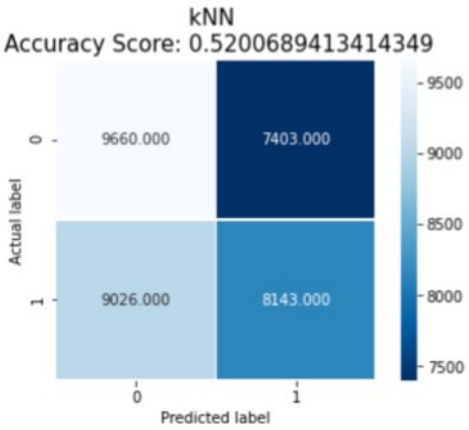
	SEVERITYCODE	WEATHER	ROADCOND	LIGHTCOND
0	1	1	1	1
1	1	3	1	1
2	1	1	1	1
3	1	1	1	1
4	1	10	8	1

Our next step was to create a training set and a testing set with a 70% split. Than we build 4 models, a linear regression model, a decision tree model, a SVM model, and kNN model with the best k. Bellow we can see the heat maps and the accuracies for each model. Unfortunately, the results were disappointing, that indicates, we have to go back and do a better data processing and feature engineering.









The best scoring model was the SVM with and accuracy of 56%

5.Conclusion

We tried to build a model that would predict the severity of an accident given the weather, road and lighting conditions. We used a data set of the accidents that occurred in Seattle city. By visualizing our data we learned that most of the accidents happen on ideal conditions. We would advise locals to warn drivers to be more carefully when the conditions are good. We build four models, a logistic regression, kNN, decision tree and SVM model. Our models performed really poor and should not deployed as the accuracy is not satisfying. By going back to feature engineering and data preprocessing we could improve our models accuracy in the future.