Seattle's collisions and severity recognition using machine learning algorithms

JUAN SEBASTIAN ALVARADO GALEANO

ACM Reference Format:
Juan Sebastian Alvarado Galeano. 2018. Seattle's collisions and severity recognition using machine learning algorithms. *ACM Trans. Graph.* 37, 4, Article 111 (August 2018), 6 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Car collisions and its influence on pedestrians in a problem which has always focused the attention of goverments in an continuous effort to preserve life. In general, collision might occur for several reasons which include distractions, influence of alcohol, the road and light conditions, the date among others which can result in various types of injuries. In order to classify all incidents, Seattle has been using a coding scheme which classifies these events as a fatality (3), a serious injury (2b), an injury (2), a prop damage (1) or even unknown (0). The latter coding scheme help to find patterns and important factors to take into account for future prevention of accidents implemented in upcoming laws as well as a way of testing actual regulations.

The present work shows a Machine Learning approach to collision severity clasification via a supervised algorithm as a function of the relevant involved factors such as fatalities, property damage, injuries etc. to identify relevant collision factors and optimize the statistical data recognition to give statistical results more efficiently.

Author's address: Juan Sebastian Alvarado Galeano, jsalvaradog@outlook.com.

0730-0301/2018/8-ART111 $15.00
https://doi.org/10.1145/1122445.1122456

2 DATA

A Seattle collisions data base recording all incidents from 2003 to present was obtained from Kaggle, which provides a suitable set for training the model prior to predict the severity of new incidents. The data set includes fatalities, collision type, number of involved persons their injuries and seriousness, the number of vehicles (and their speed), pedestrians, bicycles and parked cars involved, possible causes such as inattention, alcohol influence, whether or not the pedestrian right of way was not granted and environmental factors such as weather, road and light conditions and date among others. These features have been selected since I consider they are the main factors that helps to understand collisions and their severity, mostly influenced by the human affectation. However, it is worth to mention that the used data base includes more information which is not considered in this work.

First, the date information was splitted into three columns in order to verify the incidents per year, month and day. Then, data in come columns such as "UNDERINFL" has to be formatted in order to have the same convention for *yes*(Y) and *no*(N). Besides, numerical attributes have been re-scaled in such a way that represents the percentage over the total cases. In some graphs the sum of all values might not equal one, this happens because the difference is related to the unreported data.

Numerical values ensure the success of the machine learning algorithm since at first sight one

can see that a decision tree is suitable from the following table.

SEVERITYCODE	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	INJURIES	SERIOUSINJURIES	FATALITIES
0	0.008366	0.000000	0.000000	0.000000	0.000000	0.000000	0.0
1	0.649969	0.086023	0.113542	0.699043	0.000000	0.000000	0.0
2	0.323344	0.780138	0.809879	0.286929	0.943309	0.000000	0.0
2b	0.016202	0.116042	0.071772	0.012712	0.053206	0.969263	0.0
3	0.002115	0.017798	0.004641	0.001313	0.003486	0.030737	1.0

Nevertheless, it is important to find out if the algorithms succeeds with the environmental data, in this way predictions and decisions can be made before accidents. Since we have non-numerical attributes I take advantage of the *get_dummies*. KNN, decision tree, SVM and Logistic regression are going to be used for testing the 20% of the data.

To sum up, the considered features are:

- Numerical: SEVERITYCODE, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, INJURIES, SERIOUSINJURIES, FATALITIES, INATTENTIONIND, UNDERINFL, PEDROWNOTGRNT, SPEEDING, HITPARKEDCAR, YEAR, MONTH, DAY
- Categorical: SEVERITYCODE,ADDRTYPE, COLLISIONTYPE, WEATHER, ROADCOND, LIGHTCOND

for more information about their meaning the metadata is available in https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf

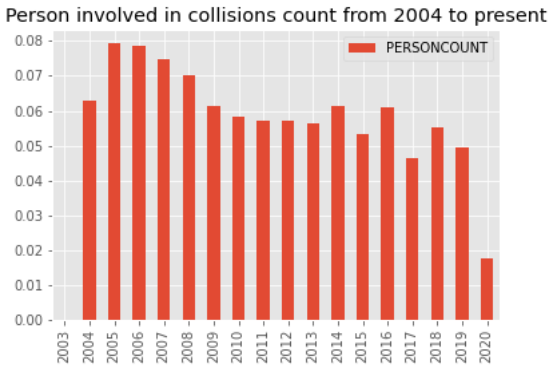
3 EXPLORATORY DATA ANALYSIS

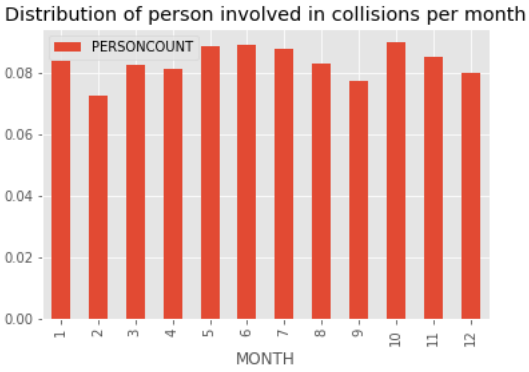
Let's recall the collision severity code scheme:

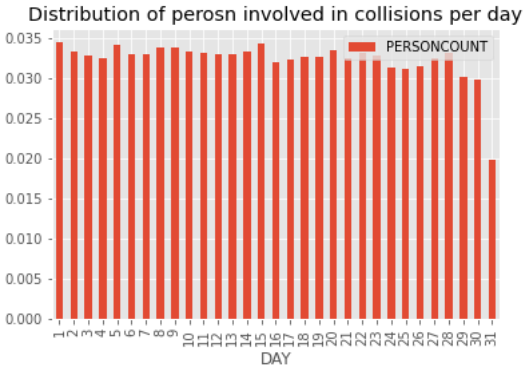
Code	Meaning
0	Unknown
1	Prop damage
2	Injury
2b	Serious injury
3	Fatality

since we have explicit *Fatalities*, *Serious Injuries*, *injuries* and *Vehicle count* attributes that make the classification quite obvious, let's see how the severity code behaves with all other attributes:

3.1 Relation with date



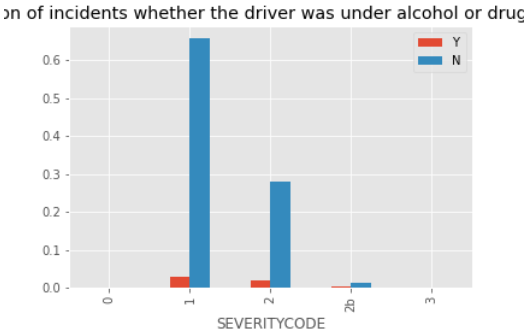


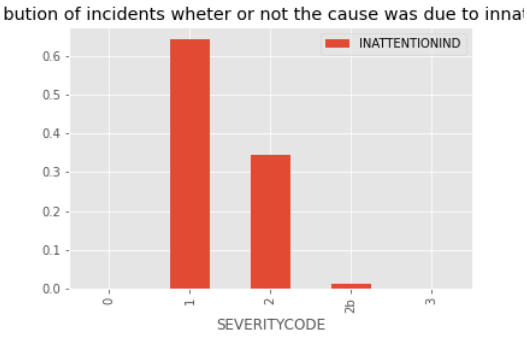


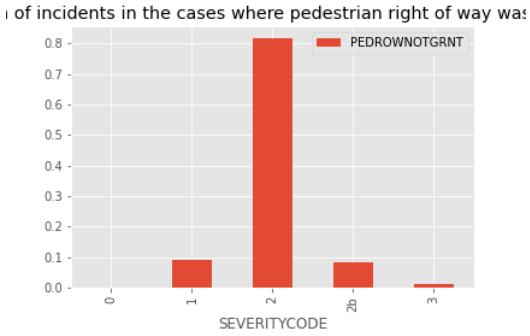
Although results not show any abrupt behavior, it is interesting to notice that during all months and days the number of incidents is approximately the

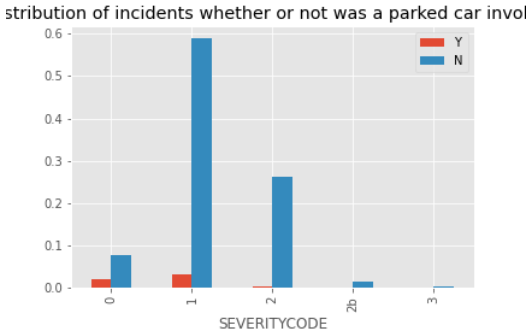
same. Nevertheless, we can observe that February has the least number of collisions but it is probably caused because the fewer days it has, which is not the case of September which shows a smaller percentage as well but the cause is not so obvious, probably due to a social context. However, we can see that implemented guidelines by the government have been effective in reducing the number of incidents over the years.

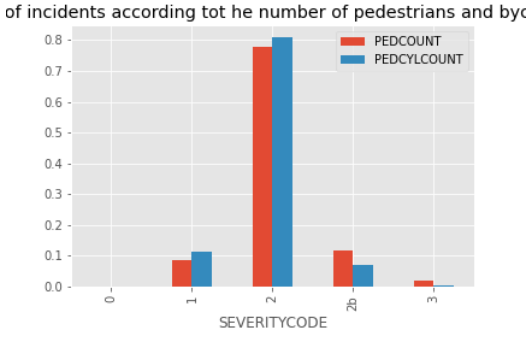
3.2 Relation with driver and pedestrian





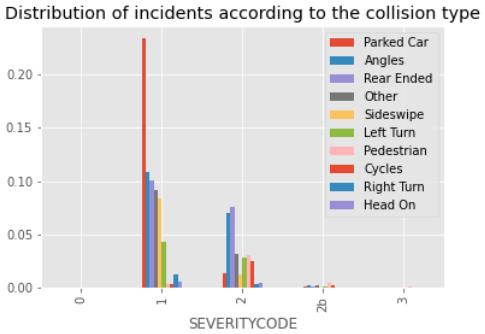


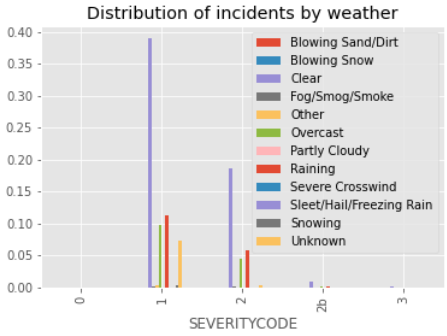


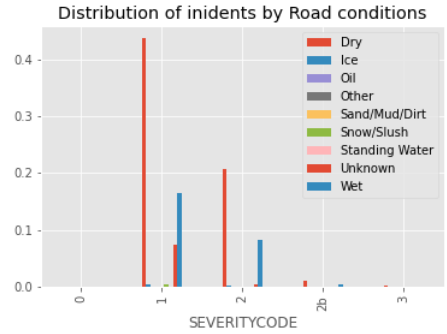


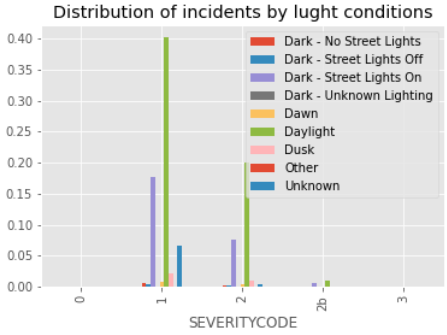
The general feature is that neither of the previous attributes was decisive for *fatalities* nor *unknown* although it is impressive the small fatalities percentage in every case. Secondly, we can see that in general, it is more probable to generate property damage rather than injuries expect when the pedestrian right of way is not granted, in which case there is a high probability of getting injuries but not serious ones. Likewise, injuries are highly probably when bicycles are involved; yet, not mainly serious.

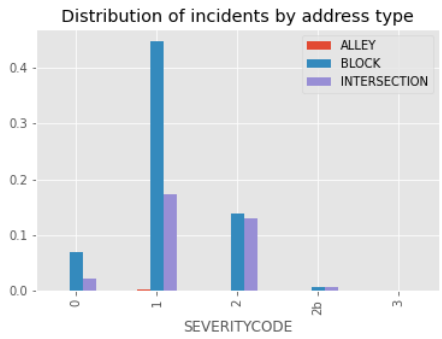
3.3 Relation with environment









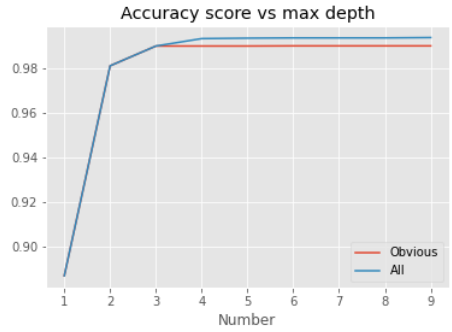


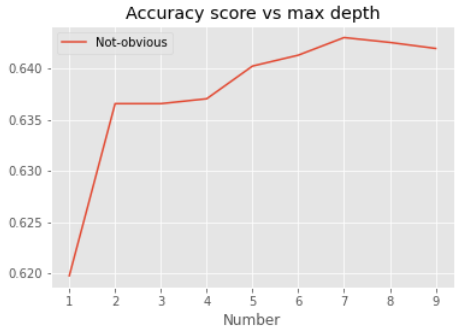
Again, we can see that none of there attributes are decisive to identify type-3 or type-0 collisions. However, property damage and injuries are mainly caused under clear conditions, with a dry road, with daylight and in blocks. Moreover, property damage is mostly to a parked car which shows that the main cause of collisions if the driver imprudence so better road education has to be implemented as well as more stringent guidelines in some zones.

4 PREDICTING MODELING

4.1 Using Numerical and yes/no attributes

In this case we are using the attributes PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, INJURIES, SERIOUSINJURIES, FATALITIES, INATTENTIONIND, UNDERINFL, PEDROWNOTGRNT, SPEEDING, HITPARKEDCAR, YEAR, MONTH and DAY. Among these, the "Obvious" attributes for classification are: PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, INJURIES, SERIOUSINJURIES and FATALITIES while its complement are the "Not-obvious" ones. The data contain NaN values in their entries which might be caused by not reported information or a null/no value which is not registered but can be interpreted by default. I build the predictive model by implementing a decision-tree algorithm because all other methods take a long computation time, so the proposed method was suitable as expected. a 20% test set was considered with a random seed 5. The resulting accuracy score for this data is given by:

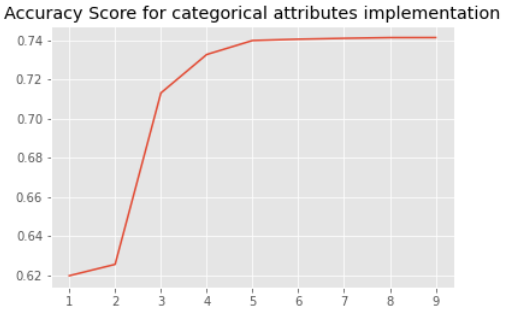




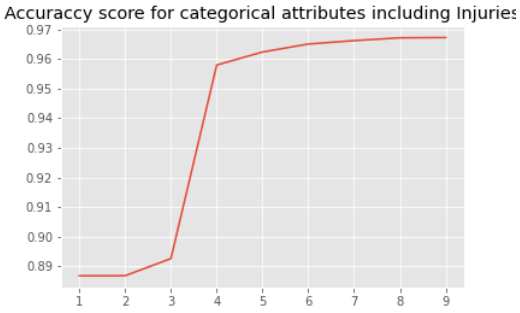
we can see that the optimal number of categories lies between 5 and 7, although 5 was a priori expected. However, when the obvious data is removed a 75% accuracy is achieved which is a surprising result. This help to identify possible risk in cities by giving a probability for an accident to end in fatalities injuries etc. However, the not-obvious data slightly increases the accuracy of the overall result, but in fact can be neglected and the classification can be done just with the obvious data

4.2 Using non-numerical attributes

the get_dummies method was used with this data set in order to implement a numerical scheme. After training the model it is found that:



Again, a surprising ≈ 75% result is achieved in this case, which also helps to identify risks and so prevention programs can be implemented for citizens. However, from the exploratory data analysis we learned that most attributes help to identify type 1 and 2 severity so one might think that by including a FATALITIES attribute the accuracy would improve, but the results show that the improvement is small. In contrast and out of the blue, the only feature that highly increase accuracy is INJURIES increasing close to the numerical attributes results. In this case there is no maximum in the accuracy but a depth of 7 is acceptable as well.



5 CONCLUSIONS

All in all, Seattle collisions data shows that September is the month with least number of incidents while on other months or days there is no patter and incidents tend to be the same. Although a decrease over the years have been seen. When modeling with numerical attributes one finds a high accuracy above the 99% while on categorical attributes

accuracy surprisingly tends to 75% but can be increased to almost 97% if INJURIES are registered. In this way, it has been shown that the decision tree algorithm was successful while all other methods took high computation times where a suitable maximum depth can be chosen between 5 and 7. This model implementation provides a way of identify potential risks prior to prevention programs and gives an efficient and optimal way of classification depending on just a few attributes.