Predict severity of an accident

Marcelo Quiroga

September, 2020

1. Introduction

Traffic accidents are a recurring problem in every corner of the world. Perhaps the biggest problem are the injured people because the material damage can be compensated but the human lives are irreplaceable, even if life is not at risk, the fact of being seriously injured remains a very big problem.

The first problem is the fatality of the injured, the second problem is the possibility of being seriously injured. Another problem no less important are the costs in repairing damages, costs that insurers assume, costs that governments assume and a long list of affected.

Governments in each country should analyze these types of studies to issue laws that try to reduce the risk of traffic accidents, the frequent driver who goes to their work every day should know the prevention recommendations that come from these reports because they aim to take care of their lives and the lives of their families.

2. Data

The data set used is based on Seattle city records comprising from 2004 to 2020, there are more than 194 thousand records consisting of different features totaling 37. These variables include data external to the driver and mobility, such as: weather status, date and time, road conditions, light conditions, among others. Also included are variables own or internal to the driver: alcohol, speed, driver responsibility.

The objective will be to predict the severity of accidents using the variable 'severity', the possible values for this variable are two: injury and prop damage; this represents a binary

analysis. Within the dataset you can see 30% of lesions and the remaining 70% prop damage. This creates the 'imbalance issue' problem and for this reason the relevant recommendations will be used[1][2].

Recommendations from other similar reports to address this analysis will be taken as an additional reference[3].

2.1. Data Wrangling

The dataset has 37 features and the following columns are used: 'INATTENTIONIND', 'UNDERINFL', 'SPEEDING', 'LIGHTCOND', 'ROADCOND', 'WEATHER', 'SEVERITYCODE'.
The data receives binary treatment where the value one (1) represents if the feature influences the severity of the accident and the opposite case (does not influence the severity) is given by the value zero (0).

- The values for the case that the driver is distracted (INATTENTIONIND) are 1 when the driver is distracted.
- The values if the driver is under the influence of alcohol (UNDERINFL) are 1, otherwise they are 0.
- If the driver drives with high speed (SPEEDING) they are 1 and otherwise 0.
- The different LIGHTCOND values are standardized to 'Daylight' = 0 and 'Darkness, Unknown, Streetlight' = 1.
- The values for ROADCOND are 'Dry' = 0 and 'Wet, Ice, Snow, Oil, others' = 1.
- The WEATHER values are 'Clear' = 0 and 'Rainy, Snowing, others' = 1
- For the WEATHER, ROADCOND and LIGTHCOND columns all null values are assigned one (1).

2.2. Dealing with data imbalance

Unbalanced data is a characteristic of this type of dataset where there are very few records of accidents with injuries and many more records with accidents

[1] How to Handle Imbalanced Classes in Machine Learning - https://elitedatascience.com/imbalanced-classes
[2] Bagging and Random Forest for Imbalanced Classification - https://machinelearningmastery.com/bagging-and-random-forest-for-imbalanced-classification/
[3] High-Resolution Road Vehicle Collision Prediction for the City of Montreal - https://arxiv.org/pdf/1905.08770.pdf

without injuries. To address these types of problems, data subsampling is used to balance the two classes of current data.

The first step is to reduce the 136485 records of SEVERITYCODE with value 1 to equal 58188 records, this coincides with case 2 (injuries) for SEVERITYCODE. For this, random sampling provided by the sklearn 'resample' library is used. This allows you to create a balanced dataframe.

The machine learning models that are used have the parameter 'class-weight', which receives two values: 'balanced' or 'balanced_subsample'. In this way, the model can be fed with the complete dataset that includes the imbalanced data.

3. Methodology

3.1. Divide the features

Two possible causes are being considered for traffic accidents:

- Internal causes, are the responsibility of the driver and are represented in the columns: 'INATTENTIONIND', 'UNDERINFL', 'SPEEDING'.
- External causes, these are causes external to the driver and are represented in the columns: 'LIGHTCOND', 'ROADCOND', 'WEATHER'.

This consideration allows dividing the dataset into two separate parts, to which an appropriate Machine Learning model is applied and the results for each of them are observed. Finally, these two parts are rejoined to apply the same algorithms and evaluate the results generated.

3.2. Modeling

The Machine Learning algorithms: Logistic Regression and Decision Trees are recommended for working with binary data, additionally the Random Forest

Classifier (RFC) algorithm is an extension of Decision Tree and is considered because it has already been used in other similar studies[4].

The metrics for evaluation are Jaccard index, to compare the similarity between the set of real data and the data set predicted by the model, and Log Loss, to predict the probability of the values obtained by the model.

4. Results and Discussions

In general, the two models used (Logistic Regression, Random Forest Classifier) receive modest evaluations, both of which have no more than 63% accuracy in their evaluations with the Jaccard index. If we evaluate the probability that the model correctly classifies as prop damage or injurie, an improvement is observed that reaches up to 68% and 69%.

The RFC model additionally provides important data that is the feature importance which is very useful for making recommendations.
The results of the model for the internal causes that are attributed to the speed of the car, inattention while driving and the influence of alcohol are:

Class_weight	LR		RFC	
	Jaccard	LogLoss	Jaccard	LogLoss
default	0.5349	0.6900	0.5349	0.6900
balanced	0.6343	0.6900	0.6343	0.6895
subsample	NA	NA	0.6343	0.6900

The external causes referring to the light conditions, the state of the road, the weather; present the following evaluations:

[4] High-Resolution Road Vehicle Collision Prediction for the City of Montreal - https://arxiv.org/pdf/1905.08770.pdf

Class_weight	LR		RFC	
	Jaccard	LogLoss	Jaccard	LogLoss
default	0.5391	0.7443	0.5412	0.6874
balanced	0.6343	0.6900	0.4959	0.6879
subsample	NA	NA	0.4959	0.6876

When modeling the complete dataset the precision drops to 47% but the probability of classification remains at 68%

Class_weight	RFC	
	Jaccard	LogLoss
default	0.5513	0.6864
balanced	0.4741	0.6842
subsample	0.47.35	0.6843

4.1. Feature importance

When looking at internal causes, the RFC provides additional information such as feature importance, so using subsampling, the evaluations are as follows:

Feature	Importance
Inattentionind	0.44
Alcohol influence	0.34
Speeding	0.22

Observing at external causes, the feature importance is:

Feature	Importance
Light condition	0.59
Road condition	0.21
Weather	0.20

Finally observing the six features, the values change with respect to the two previous tables but still provide important information:

Feature	Importance
Inattentionind	0.11
Alcohol influence	0.16
Speeding	0.10
Light condition	0.40
Road condition	0.17
Weather	0.06

5. Conclusions

- When you go out to drive avoid drinking this can reduce injuries in traffic accidents by up to 34%. Respect the speed limits as you can reduce accidents by up to 22%, and above all do not get distracted while driving because 44% of accidents are caused by being distracted.

- It is possible that on the road the weather is not good and that the track conditions are not favorable, this is out of your control but the most important thing is the light conditions, if you have a lot of difficulty in vision it is better to postpone the trip, or otherwise always use the high beams to have better visibility.