Predicting Severity of Car Accidents

Ali Alfares
09-10-2020

1- Introduction:

1-1 Background:

A traffic collision, also called a motor vehicle collision, car accident, or car crash, occurs when a vehicle collides with another vehicle, pedestrian, animal, road debris, or other stationary obstruction, such as a tree, pole or building. Traffic collisions often result in injury, disability, death, and property damage as well as financial costs to both society and the individuals involved.

1-2 Problem:

Every year the lives of approximately 1.35 million people are cut short as a result of a road traffic crash. Between 20 and 50 million more people suffer non-fatal injuries, with many incurring a disability as a result of their injury. Road traffic injuries cause considerable economic losses to individuals, their families, and to nations as a whole. These losses arise from the cost of treatment as well as lost productivity for those killed or disabled by their injuries, and for family members who need to take time off work or school to care for the injured. Road traffic crashes cost most countries 3% of their gross domestic product.

Causes of collisions could be rash driving, excessive speed of road users, violation of the regulations of traffic, sudden bursting of tires, insufficient lightings, failure of breaks, skidding of road surfaces, ruts, and potholes. "Severity code" is usually used to categorize the rank-wise priority at high crash locations, it goes up with more injuries and/or fatalities. Although it might be pretty straight forward to assume causes of collisions, but it gets very difficult to predict causes of high or low severity code associated with those collisions because there are a lot of factors impacting the possible severity code of accidents, mainly consisting of 2 parts (physical & behavioral).

1-3 Interest:

Understanding large datasets of records regarding those incidents is probably the keystone in the pursuit of having a safe driving experience, It's a best interest of governments, technology corporations and car manufacturers. The potential of this study is off limits, it can be directed towards many projects yielding many advances in cities governed by AI, projects such as:

1. A system that can assist drivers to take certain paths towards destination in order to reduce chances of accidents.
2. Advances in machine learning algorithms for automated AI cars with no chances of collisions.

The question here is "Can we build a model that can tell if it's dangerous or not to drive in a certain road?" and if so .. "Can this model determine to what extent it's dangerous to drive in a certain road?"

2- Data Acquisition & Cleaning:

2-1 Data Source:

Data used is "**Example Dataset**" provided by the IBM Data Science course on Coursera.org , Because of the following reasons:

1. It's rich with features which might be extremely useful for future research.
2. It's real life data recorded by Seattle state SPD & SDOT.
3. It has special records where behavior of drivers plays a role of making the collision (drugs or alcohol).
4. Physical features of collisions are recorded so it's highly valuable for machine learning projects for automated AI cars.

Dataset download link

2-2 Feature Selection:

Dataset has 38 features, 194673 rows representing records of collisions which has taken place in Seattle since 2004 until present. This report will focus on the factors impacting the severity_code, as mentioned before those factors are physical & behavioral, each consists of specific categorical features. Studying the physical circumstances during driving means to keep track of (weather conditions, road conditions & light conditions). Studying the behavioral impact of the driver during driving would imply considering the use of (drugs/alcohol) and/or being (inattentive). Selection of those features theoretically speaking is going to help reduce noise of the cumulative effect between different factors (physical & behavioral) which would help the model understand specific patterns in the complex reality of ours. Figure 1 shows what features will be taken into consideration in this report.

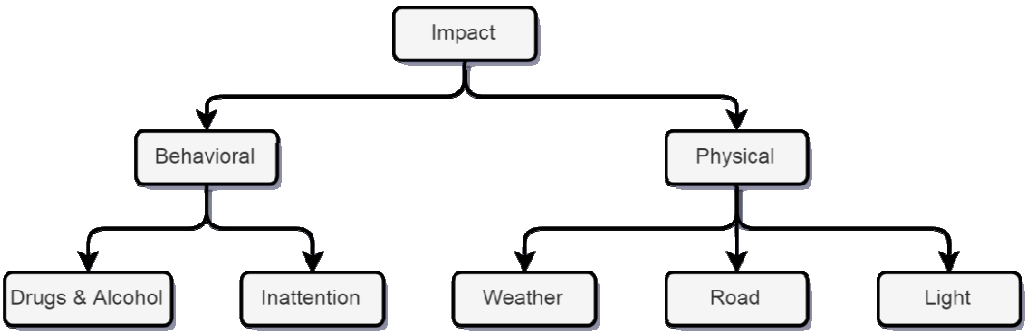

Figure 1 Features impacting severity_code

2-3 Data Cleaning:

"Garbage in, garbage out" This is a well-known principle in data science. It means "Your insights can only be as good as the data used to derive them". Thus it's extremely important to avoid using bad data. The idea behind feature selection in this report is to keep one factor fixed as the other can be free in providing patterns of impacting severity of accidents. Measures has been taken accordingly depending on "Metadata.pdf" file attached:

Stage 1: "Reducing human behavior impact"

1. Dropping data where drivers are drunk.
2. Dropping data where drivers are inattentive.
3. Dropping NAN in Weather & Light & Road Conditions.
4. Dropping "Unknown" + "Other" records in Weather & Light condition & Road Condition.

Rows left in the dataset after those measures were 133408 which is the cost of getting the best data to model.

Stage 2: "Evaluating human behavior impact"

1. Reloading dataset & dropping data where drivers are not drunk and attentive.
2. Dropping NAN in Weather & Light & Road Conditions.
3. Dropping "Unknown" + "Other" records in Weather & Light condition & Road Condition.

Rows left in the dataset after those measures were divided into 2 groups "focused" and "not focused".

3- Methodology:

3-1 Approach:

In order to model a categorical value of the target variable using values of categorical features, It's required to use classification. Building classification models would need a train-test split operation. The best classification model will be selected from different classifier models built using Scikit-learn on JupyterNotebook based on the best accuracy metric. Further observations on the most important features identified by the best classifier will also be shown.

3-2 Prediction Method:

Classification is a supervised machine learning approach, in which the algorithm learns from the data input provided to it and then uses this learning to classify new observations.

- Used classification models are:

1. K-Nearest Neighbor.
2. Decision Tree.
3. Support Vector Machine.
4. Random Forest Classifier.
5. Logistic Regression.

- Accuracy metrics of those models were (Jaccard, F1_score & LogLoss).

Figure 2 shows the process of prediction using different models.

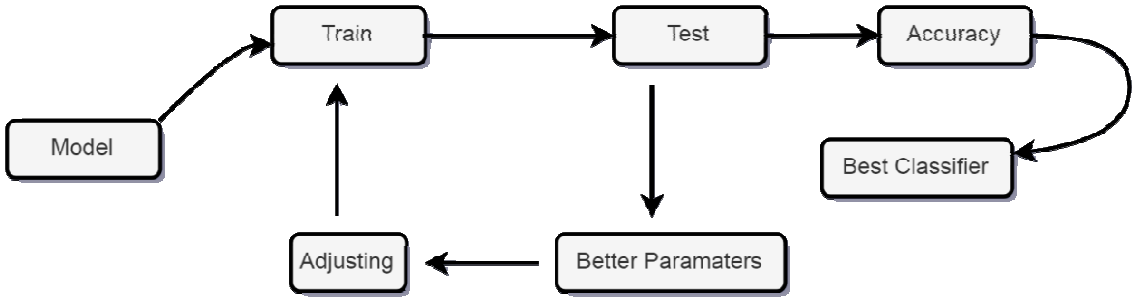

Figure 2 The process of prediction

- Train-test split produced 106726 training records and 26682 testing records. Before we start the predictive modeling, It's necessary to change data types for the model to be able to deal with. This concept is called one-hot encoding, which is essentially to change categorical data features into binary data features of each value involved within each feature.
- Data features before starting predictive modeling were:

WEATHER_Blowing Sand/Dirt	float64
WEATHER_Clear	float64
WEATHER_Fog/Smog/Smoke	float64
WEATHER_Overcast	float64
WEATHER_Partly Cloudy	float64

WEATHER_Raining	float64
WEATHER_Severe Crosswind	float64
WEATHER_Sleet/Hail/Freezing Rain	float64
WEATHER_Snowing	float64
ROADCOND_Dry	float64
ROADCOND_Ice	float64
ROADCOND_Oil	float64
ROADCOND_Sand/Mud/Dirt	float64
ROADCOND_Snow/Slush	float64
ROADCOND_Standing Water	float64
ROADCOND_Wet	float64
LIGHTCOND_Dark - No Street Lights	float64
LIGHTCOND_Dark - Street Lights Off	float64
LIGHTCOND_Dark - Street Lights On	float64
LIGHTCOND_Dark - Unknown Lighting	float64
LIGHTCOND_Dawn	float64
LIGHTCOND_Daylight	float64
LIGHTCOND_Dusk	float64

dtype: object

- Target features is "SEVERITYCODE" A code that corresponds to the severity of the collision: • 3—fatality • 2b—serious injury • 2—injury • 1—prop damage • 0—unknown.
- Classification models and their parameters are described below:

```
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
           metric_params=None, n_jobs=None, n_neighbors=3, p=2,
           weights='uniform')
DecisionTreeClassifier(class_weight=None, criterion='entropy', max_depth=4,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=None,
            splitter='best')
SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
            decision_function_shape='ovr', degree=3, gamma='scale', kernel='rbf',
            max_iter=-1, probability=False, random_state=None, shrinking=True,
            tol=0.001, verbose=False)
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
            max_depth=None, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=100, n_jobs=None,
            oob_score=False, random_state=None, verbose=0,
            warm_start=False)
LogisticRegression(C=0.01, class_weight=None, dual=False, fit_intercept=True,
            intercept_scaling=1, max_iter=100, multi_class='warn',
            n_jobs=None, penalty='l2', random_state=None, solver='lbfgs',
            tol=0.0001, verbose=0, warm_start=False)
```

Figure 3 Classification models' description

4- Results & Discussion:

4-1 Accuracy Metrics:

Letting the model learn took a lot of time nearly 45 minutes due to limited resources and the tremendous amount of rows to train/test on. The accuracy metrics were calculated and are shown in figure 4.

	Algorithm	Jaccard	F1-score	LogLoss
0	KNN	0.685631	0.559609	NA
1	Decision Tree	0.686043	0.558295	NA
2	SVM	0.685968	0.558398	NA
3	RandomForest	0.685893	0.558431	NA
4	Logistic Regression	0.686081	0.558383	0.620453

Figure 4 The accuracy metrics

Accuracy metrics of all models were very convergent with 68.5% for Jaccard, 55.8% for F1-score. This shows how difficult and complicated it's to predict severity code based on general physical circumstances of roads, all models are having convergent accuracy metrics means we need to input more features designed specifically to calculate severity code.

The best classifier can be any of these models but numerically we have to go with "Logistic Regression", as It's 68.6% accurate. Since the method ("feature_importances_") is not included in "Logistic Regression" model , we have to go with the next model which is "Decision Tree" to calculate importance of features.

4-2 Importance of Features:

We used Seaborn to visualize importance of features thus we could discuss how some features may appear more frequently than others, This may lead to better adjustments of models because it's more clear what to measure and how. Figure 5 visualizes importance of features.

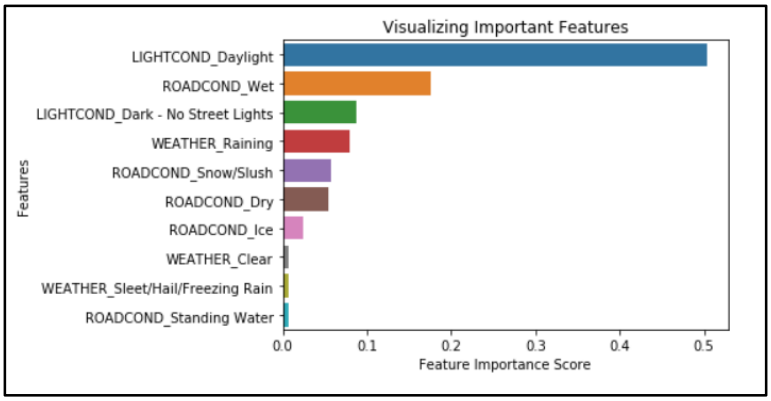

Figure 5 Importance of features

1. LIGHTCOND_Daylight: Although "Daylight" was one of the most frequent appearing value of light condition among the dataset. It may not be considered as a cause of accidents but a major cause of a more severe accidents. We can explain this using velocity of vehicles, when drivers drive during daylight they tend to increase their velocity and that's as a result exactly why it's more severe to drive in daylight and to have an accident.
2. ROADCOND_Wet, LIGHTCOND_Dark - No Street Lights & WEATHER_Raining : Those values are showing as often as they are partly major causes of accidents, they may be considered to be causes of more severe accidents as well.

4-3 Evaluating human behavior impact:

To evaluate human behavior impact on severity of accidents, we studied how frequent same physical circumstances are appearing when drivers are "focused" and "not focused". The 2 groups are representing the condition of attention the driver has that's through the use of Alcohol-drugs in addition to stating that they were inattentive for any other reason. The focused group was the same group we studied to extract patterns of physical circumstances leading to more severe accidents. The not focused group was the rest of the dataset. Both combined were representing some sort of normal distribution of circumstances. Table 1 shows how big is the impact of being focused or not focused on frequency of accidents' features, keeping in mind that all rows in dataset were about accidents and not possible accidents.

Table 1 Frequency of top 2 values of each feature

Value	Feature	Focused %	Not focused %	Entire dataset %
Top 2 values for weather				
Clear	Weather	**63.4**	**63.8**	**62.1**
Raining	Weather	**19.6**	**20**	**18.6**
Top 2 values for light				
Daylight	Light	**67.5**	**72.3**	**64**
Dark *street lights on*	Light	**26**	**21**	**26.7**
Top 2 values for road				
Dry	Road	**70.8**	**70**	**69.4**
Wet	Road	**27.7**	**29**	**26.4**
Total Records		**133408**	**8980**	**175216**

From table 1 we can see that the top 2 values of each feature is about 2-3 % different in frequency percentage overall between groups. By that we can conclude that being attentive or not is not correlated to specific physical circumstances. Thus it doesn't play a major role regarding severity of accidents.

5- Recommendations :

5-1 Data Features:

Data features designed specifically to predict an accurate value for the SEVERITY_CODE, which could be calculated the changes each second then let the model behave accordingly:

1. Passenger count of vehicle(s) involved (the more.. higher the risk).

2. Location of vehicle (the more frequent a location in logs appears.. higher the risk).

3. Location of vehicle regarding street with high traffic at the same time of driving.

4. Type of transportation (higher code with bigger vehicles).

Those parameters have potential of having more accurate prediction of SEVERITY_CODE. GPS systems could advise/suggest specific measures to avoid possible accidents with higher SEVERITY_CODE, measures such as changing driving route or changing velocity which could save lots of lives and lead to a safer more advanced driving experience. Models

should be trained and tested thoroughly over those newly created data parameters, then calculate their accuracy metrics thus determine if its efficient to run the driving-assistant software.

6- Conclusion :

This report concludes that the best classification model is highly depending on data's quality because the dataset was not built specifically to have severity code as a target feature. As all models were very convergent. the best classifier was "Logistic Regression" which had 68.6% Jaccard accuracy metric, It was trained and tested on 133408 records. We found that "Daylight" is a very important feature regarding severity of accidents because of drivers' tendency to increase velocity in daylight. We recommended working harder on designing better data features which would help provide better accuracy in order to use it for model implementation in future projects.

THE END