Prediction of the severity of an accident : The case of Seattle

Final Report presented by :

Marwan Gebran

Introduction :

In this report I am trying to predict the severity of potential accident in a specific region. In fact, the severity of the accident depends on several parameters such as the time of day, the weather, the location of the second car, the type of the car, the number of person inside the car, etc. All these factors affect the severity of the accident. To do so, I will try first to detect the most important factors that play a role in determining the severity of the accident and then use them with the existing data to built a model that can predict for specific conditions (features), the probability of the severity of an accident if it occurs.

To reach the desired objective of this project, I need first to pinpoint the attributes that affect the most and/or are somehow correlated with the targeted label. First, a correlation table will be used to confirm the importance of these parameters. Second, a supervised machine learning algorithm will be used for classification purposes. The choice of the Classification algorithm is based on the fact that the dataset contains only two possible possible output for the targeted parameter.

The choice of the Machine Learning algorithm will depend on different evaluation parameters that will be discussed in this report.

Data collection:

The data used in this work is related to the severity of an accident in SEATTLE based on some features such as the light condition, the weather condition, the number of person in the car,…

The file that is used was suggested by the Coursera team and is downloadable online. It is “Data-Collisions.csv”, containing 194673 rows, each one is related to a specific accident. For each accident, 38 columns exist to describe the details of condition of this accident. As an example of a column, we find the following:

ADDRTYPE	Collision address type: • Alley • Block

	• Intersection
SEVERITYCODE	A code that corresponds to the severity of the collision.
WEATHER	A description of the weather conditions during the time of the collision.
LIGHTCOND	The light conditions during the collision.
ROADCOND	The condition of the road during the collision.
SPEEDING	Whether or not speeding was a factor in the collision. (Y/N)

Many other features exist but they will show no importance to the current project. In fact, the purpose of this work is to predict the Severity of the accident so the choice of the important features is very crucial for an accurate prediction.

Data reduction and preparation:

The first step pf the work is to understand the target label, SEVERITYCODE. By analyzing this column using classical python/pandas routine that you can find in the attached notebook, it was clear that only tow possible output exist, 1 and 2.
The output 1 states that the accident had leaded to a property damage only whereas the output 2 states that the accident lead to an injury. Also by plotting the number of accidents as a function of the output, we found that there was a bias in the number of damage with respect to the number of injuries (Fig.1).

The number of property damage is almost two times larger than the one of injuries (136485 Vs 58188). Although this is a good indicator for safety, this could lead to an unbalanced dataset that will affect the training process of our model. In that case our model will be more biased to damage than to injury.

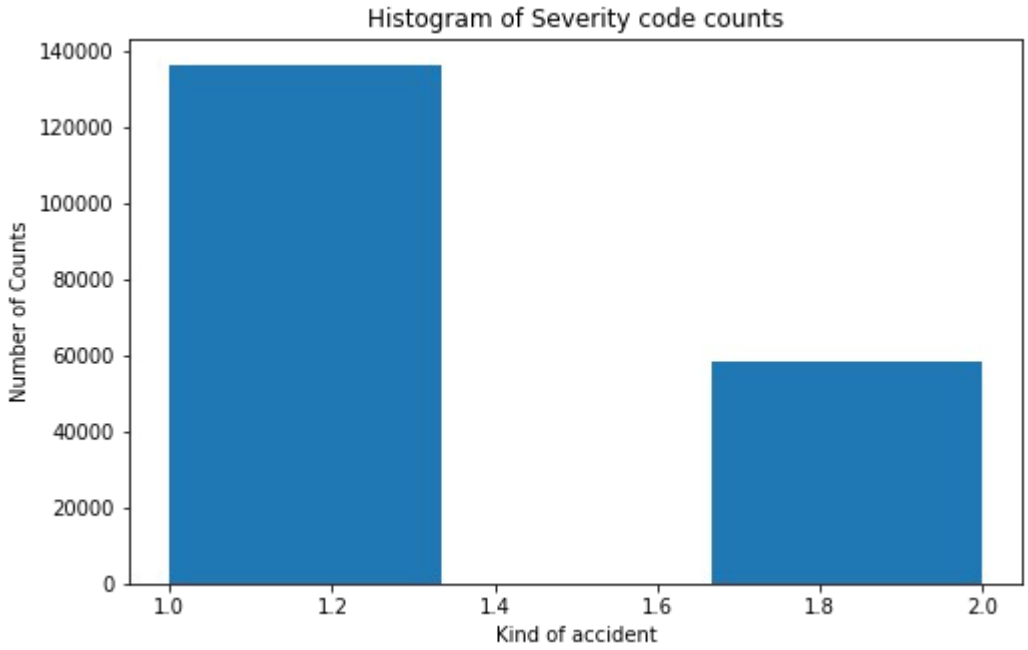


Fig. 1

To remedy this issue, I have selected randomly 58188 property damage rows and added them to the 58188 injury ones and creating a new dataset that is balanced between the two cases.

On the other hand, and for practical reasons, it is always better to have numerical values instead of categorical ones when trying to model a machine learning algorithm. That is why, most of the categorical features were transformed to numerical ones. As an example, the weather condition possibilities are displayed in the following table with the new assigned numerical values.

Wether Condition	New assignment (numerical)
Clear	0
Raining	1
Overcast	2
Unknown	3
Snowing	4
Other	5
Fog/Smog/Smoke	6
Sleet/Hail/Freezing	7
Rain	8
Blowing Sand/Dirt	9
Severe Crosswind	10

This step was done for different parameters such as the road condition, light condition and speeding. In order to select the features that affect the most the target SEVERITYCODE label, a correlation table was performed. This correlation table, that is detailed in the notebook, shows that not all parameters have an effect on the output and that some parameters are correlated but are not really suitable for prediction. The features that were chosen in this work are the PERSONCOUNT, the WEATHER, the ROADCOND, the LIGHTCOND, and the SPEEDING. The parameters were standardized to avoid biases in the model training.

Models testing and evaluation :

The prediction type that is considered in this problem is a classical clasification scheme. For that resason a supervised learning model should be built and learned on the balanced dataset. Four models will be tested, the K-Nearest Neighbor (KNN) search, the Decision Tree (DT), the Support Vector Machine (SVM), and the Logistic regression (LR).

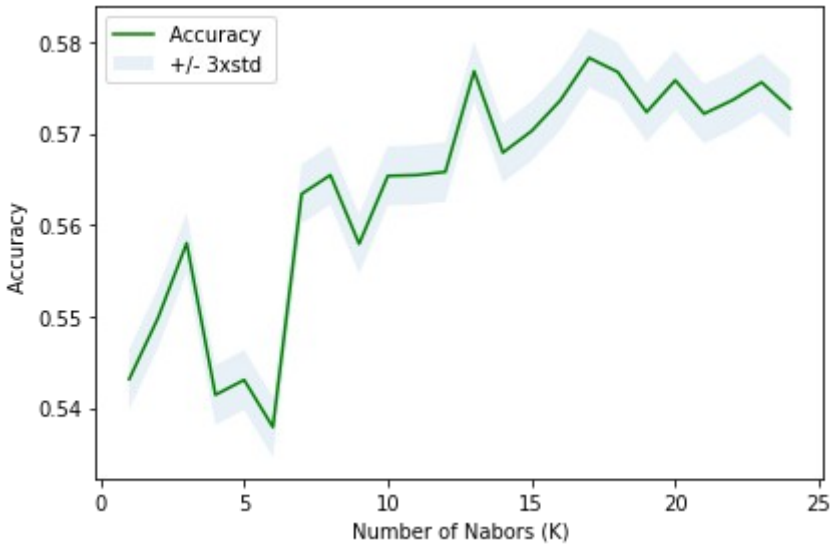
The training dataset will be divided 2 sets, a training and a test one. The training will consist of 80 % of the total while the test one will consist of the remaining 20 %.

Data	percentage	Number of rows
Training	80 %	93100
Test	20 %	23276

Once the training is done the 4 algorithm, the evaluation was done using 3 different indices, the F1 score, the Jaccard score and the log loss. The first two indices should be the closest to 1 for a good match whereas the LogLoss should be close to 0 for an accurate model.
The testings and the technical details are displayed in the attached notebook.

Results :

The first model to be tested was the KNN one. It requires to find the most optimal number of nearest neighbor to apply the search. To find that, an iteration was done on this number and the accuracy of the test was found in each step. The best accuracy was corresponded to the number of neighbor and this number was selected in the next steps.
It is found that using the Training dataset, the best value for the neighbors count is k=17.



Once k has been selected, the three accuracy indices were calculated on the Test set.

For the Decision Tree, a depth of 6 was chosen and the same data was used for training and testing. It was clear that the DT model was extremly fast in executing the training and testing.

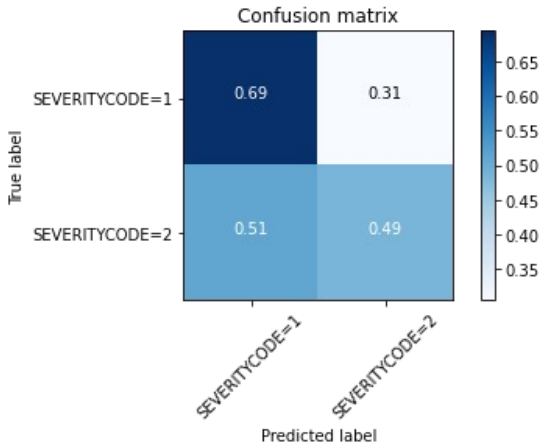
The SVM model took the longuest time to compute. Using the same data for testing and training, it took 100 times more time than the DT algorithm. The SVM was calculated using a Radial Basis Function as a Kernel.

Finally, the Logistic regression was tested using Liblinear solver. This solver is recommended when dealing with high dimension dataset.

After the evaluating the accuracy parameters for the 4 algorithm (following Table). The choice was made on the Decision Tree because it had the most optimal parameters for accuracy and the less time consuming procedure among the 4 of them.

Algorithm	F1 scord	Jaccard index	LogLoss
KNN	0.578234552445627	0.4068994683421943	*
Decision Tree	0.5846108113676355	0.4605344934742076	*
SVM	0.5857180406481002	0.45786213559129396	*
LR	0.5816742365301966	0.38250844002802725	0.674550673182746

The DT model was able to delivre the best accuracy parameters. The accuracy that was found could be represented by a simple confusion matrix as the one below. In this matrix, True poistive, True negative, False positive and False negative are depicted in percentages (eg. 69 % or True positive).



Conclusion

Using the SEATTLE dataset for car accident, a Decision Tree model was developped in order to predict the Severity of the accident based on some features such as the light condition, the road condition, and the weather condition. The Decision Tree model was used because it was the most accurate one among the remained supervised classification scheme. Once the model was selected, the training will be done using the whole (Balanced) dataset and applied later on the new data. It is clear that some extra information could lead to a better prediction such as the value of the speed at the moment of the accident.