Predicting Traffic Accident Severity

Suresh Dudi

October 2020

1. Introduction

1.1 Background

Every year car accidents cause hundreds of thousands of deaths worldwide. According to a research conducted by the World Health Organization (WHO) there were 1.35 million road traffic deaths globally in 2016, with millions more sustaining serious injuries and living with long-term adverse health consequences. Globally, road traffic crashes are a leading cause of death among young people, and the main cause of death among those aged 15–29 years. Road traffic injuries are currently estimated to be the eighth leading cause of death across all age groups globally, and are predicted to become the seventh leading cause of death by 2030[1]. Leveraging the tools and all the information nowadays available, an extensive analysis to predict traffic accidents and its severity would make a difference to the death toll. Analyzing a significant range of factors, including weather conditions, locality, type of road and lighting among others, an accurate prediction of the severity of the accidents can be performed. Thus, trends that commonly lead to severe traffic incidents can help indentifying the highly severe accidents. This kind of information could be used by emergency services, to send the exact required staff and equipment to the place of the accident, leaving more resources available for accidents occurring simultaneously. Moreover, this severe accident situation can be warned to nearby hospitals which can have all the equipment ready for a severe intervention in advance. Consequently, road safety should be a prior interest for governments, local authorities and private companies investing in technologies that can help reduce accidents and improve overall driver safety.

1.2 Problem

Data that might contribute to determining the likeliness of a potential accident occurring might include information on previous accidents such as road conditions, weather conditions, exact time and place of the accident, type of vehicles involved in the accident, information on the users involved in the accident and off course the severity of the accident. This projects aims to forecast the severity of accidents with previous information that could be given by a witness informing the emergency services

1.3 Interest

Governments should be highly interested in accurate predictions of the severity of an accident, in order to reduce the time of arrival and to make a more efficient use of the resources, and thus save a significant amount of people each year. Others interested could be private companies investing in technologies aiming to improve road safeness.

2. Data

2.1 Data source

The data can be found in the following Kaggle data set click here.

2.2 Feature Selection

The data is divided in 5 different data sets, consisting of all the recorded accidents in France from 2005 to 2016. The characteristics data set contains information on the time, place, and type of collision, weather and lighting conditions and type of intersection where it occurred. The places data set has the road specifics such as the gradient, shape and category of the road, the traffic regime, surface conditions and infrastructure. On the user data set it can be found the place occupied by the users of the vehicle, information on the users involved in the accident, reason of traveling, severity of the accident, the use of safety equipment and information on the pedestrians. The vehicle data set contains the flow and type of vehicle, and the holiday one labels the accidents occurring in a holiday. All five data sets share the accident identifications number. An initial analysis of the data was performed for the selection of the most relevant features for this specific problem, reducing the size of the dataset and avoiding redundancy, click here. With this process the number of features was reduced from 54 to 28.

2.3 Data Cleaning

- From the summary of the data we see that the data types are coherent with their respective values, with the only exception of the date, and that some features have missing values.
- More than half of the values for the coordenates are missig, as well as roughly a 10% of the data regarding the road_num and more than a 50% of the remaining samples are a 0. Thus, to keep the amount of samples the mentioned features will be dropped.
- Few values are missing in some features such as the atmospheric conditions or road category.
- Missing values and outliers will be filled with the label for *Other cases* category if possible. If not the most frequent value of the feature will be applyed.

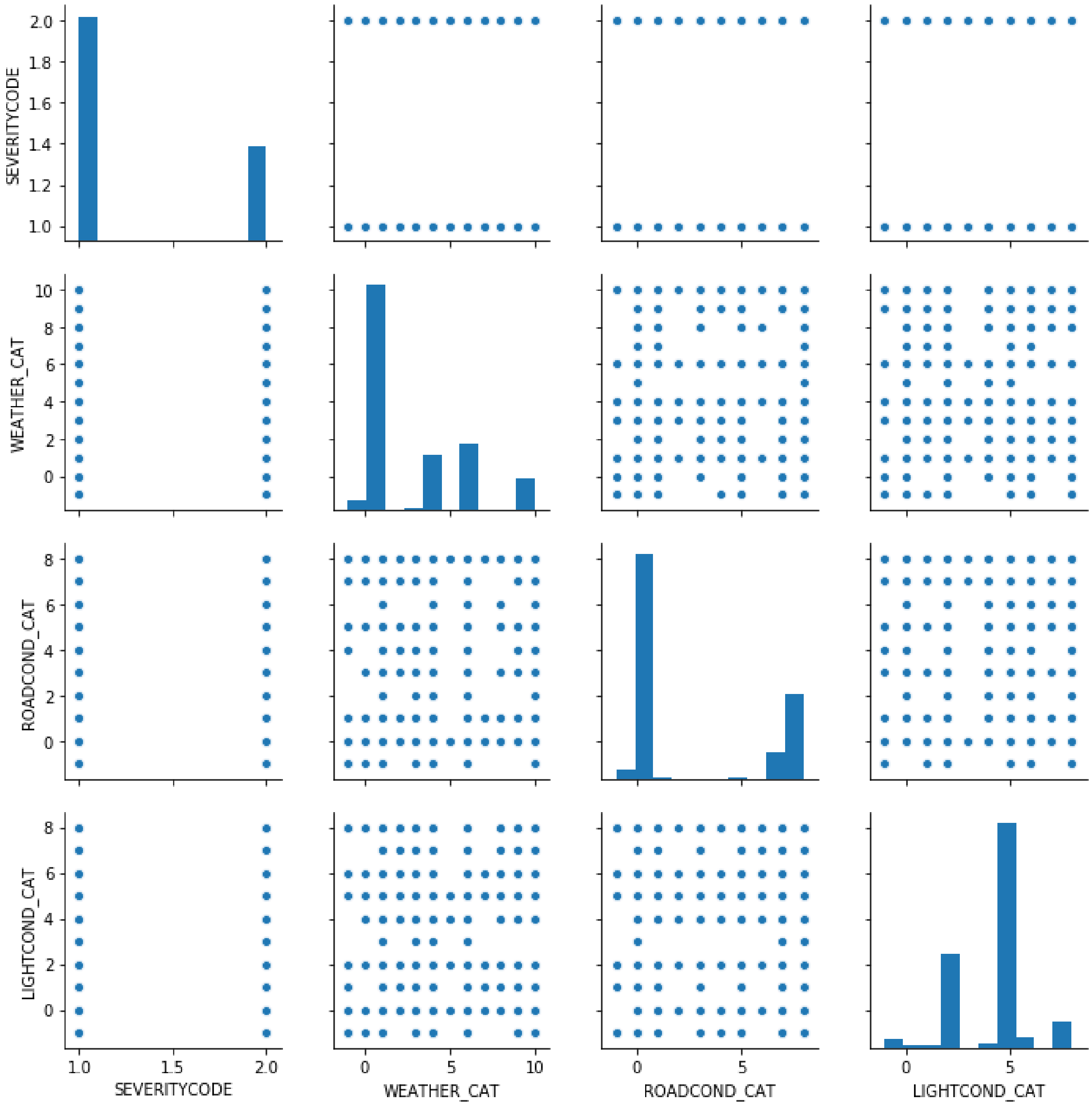
3. Exploratory Data Analysis:

From the summary of the data we see that the data types are coherent with their respective values, with the only exception of the date, and that some features have missing values.

- More than half of the values for the coordenates are missig, as well as roughly a 10% of the data regarding the road_num and more than a 50% of the remaining samples are a 0. Thus, to keep the amount of samples the mentioned features will be dropped.

- Few values are missing in some features such as the atmospheric conditions or road category.

Missing values and outliers will be filled with the label for *Other cases* category if possible. If not the most frequent value of the feature will be applied.


SEVERITYCODE
WEATHER_CAT
ROADCOND_CAT
LIGHTCOND_CAT
1.0
1.2
1.4
1.6
1.8
2.0
0
2
4
6
8
10
1.5
5

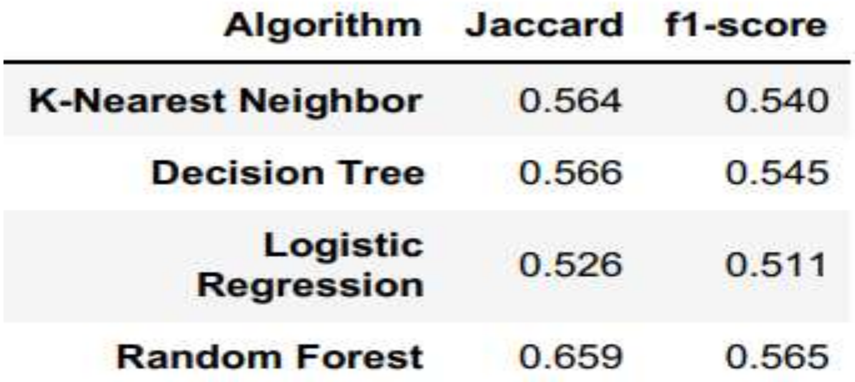
4. Predictive Modeling

Different classification algorithms have been tuned and built for the prediction of the level of accident severity. These algorithms provided a supervised learning approach predicting with certain accuracy and computational time. These two properties have been compared in order to determine the best suited algorithm for his specific problem. Firstly, the 116,376 rows where split 70/30 between the training and test sets.Then the data was standardized giving zero mean and unit variance to all features. Four different approaches were used:

- Decision Tree, Random Forest
- Logistic Regression
- K-Nearest Neighbor
- Supervised Vector Machine

  - K-Nearest Neighbor (KNN):
  - KNN will help us predict the severity code of an outcome by finding the most similar to data point within k distance.
  - Decision Tree:
  - A decision tree model gives us a layout of all possible outcomes so we can fully analyze the consequences of a decision. It context, the decision tree observes all possible outcomes of different weather conditions.
  - Logistic Regression:
  - Because our dataset only provides us with two severity code outcomes, our model will only predict one of those two classes. This makes our data binary, which is perfect to use with logistic regression.
  - Random Forest
  - To go a bit further I will develop a Random Forest model. A random forest fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The number of decision trees is specified with the n-estimators parameter.
    An upside of this algorithm is its feature importance attribute, it returns the impurity based feature importance.

5. Results



Algorithm	Jaccard	f1-score
K-Nearest Neighbor	0.564	0.540
Decision Tree	0.566	0.545
Logistic Regression	0.526	0.511
Random Forest	0.659	0.565

With no doubt the *Random Forest* is the best model, in the same time as the logistic regression it improves the accuracy from 0.56 to 0.65 and the f1-score from 0.511 to 0.57.

6. Discussion

- In the beginning of this notebook, we had categorical data that was of type 'object'. This is not a data type that we could have fed through an algorithm, so label encoding was used to created new classes that were of type int8; a numerical data type.
- After solving that issue we were presented with another - imbalanced data. As mentioned earlier, class 1 was nearly three times larger than class 2. The solution to this was down sampling the majority class with sklearn's resample tool. We downsampled to match the minority class exactly with 58188 values each.
- Once we analyzed and cleaned the data, it was then fed through three ML models; K-Nearest Neighbor, Decision Tree and Logistic Regression. Although the first two are ideal for this project, logistic regression made the most sense because of its binary nature.
- Evaluation metrics used to test the accuracy of our models were jaccard index, f-1 score and logloss for logistic regression. Choosing different k, max depth and hyperamater C values helped to improve our accuracy to be the best possible.

7. Conclusion

- In this study, I analyzed the relationship between severity of an accident and some characteristics which describe the situation that involved the accident. I identified the road condition, climate, lightning among the most important features that affect to the gravity of the accident. I built and compared 4 different classification models to predict whether an accident would have a high or low severity. These models can have multiple applications in real life such as, Severity of a accident can be predicted in real time by using above data when an accident is reported and from there measures can be taken quickly.