PREDICTING THE SEVERITY OF CAR ACCIDENTS IN CATALONIA

Marc Bara

Coursera Capstone Project

Applied Data Science

INTRODUCTION

- Road accidents are a major world economic and social problem of loss of lives and properties in many countries around the world. Key reports indicate the number of fatalities from road accidents per year of about 1.3 million and 50 million injuries [1] or an average of 3000 deaths/day and 30,000 injuries/day.

- We will develop a machine learning system, operated by local/regional administration, that could enable decision making in near-real time. Depending on local road conditions, traffic data, weather conditions, type of roads, day of the week, etc., the system will try to predict the severity of the accident; severity meaning possible fatalities involved.

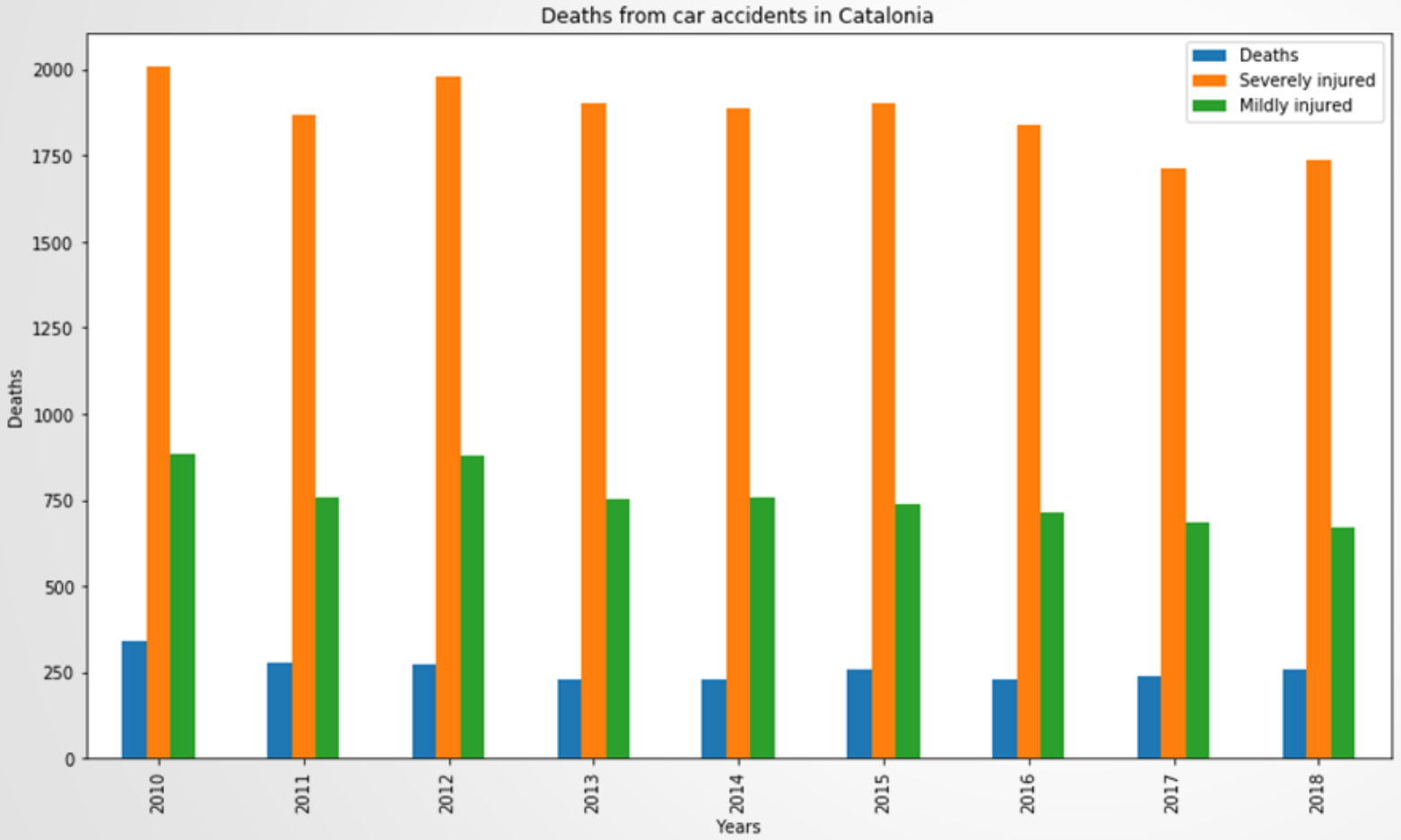
DATA SOURCE

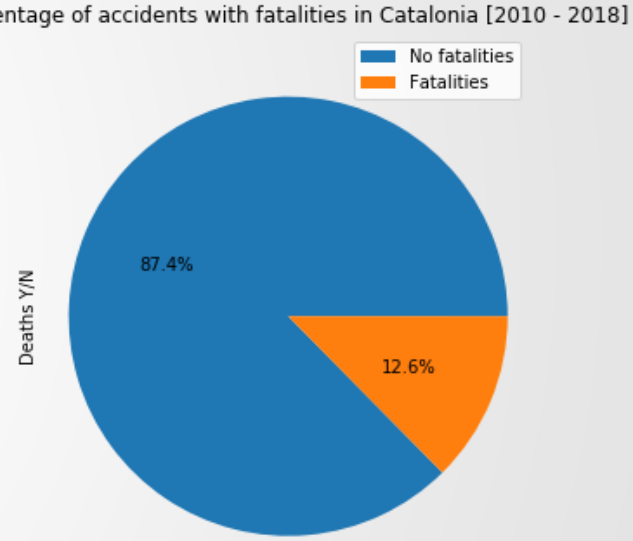
- Given the problem to tackle, we will be using relevant data with all possible descriptors that help our algorithm to succeed. In particular, we will focus on the area of Catalonia, in Spain, including Barcelona and its region.
- The data set to be used for algorithm training and testing is here:

https://analisi.transparenciacatalunya.cat/Transport/Accidents-de-tr-nsit-amb-morts-o-ferits-greus-a-Ca/rmgc-ncpb

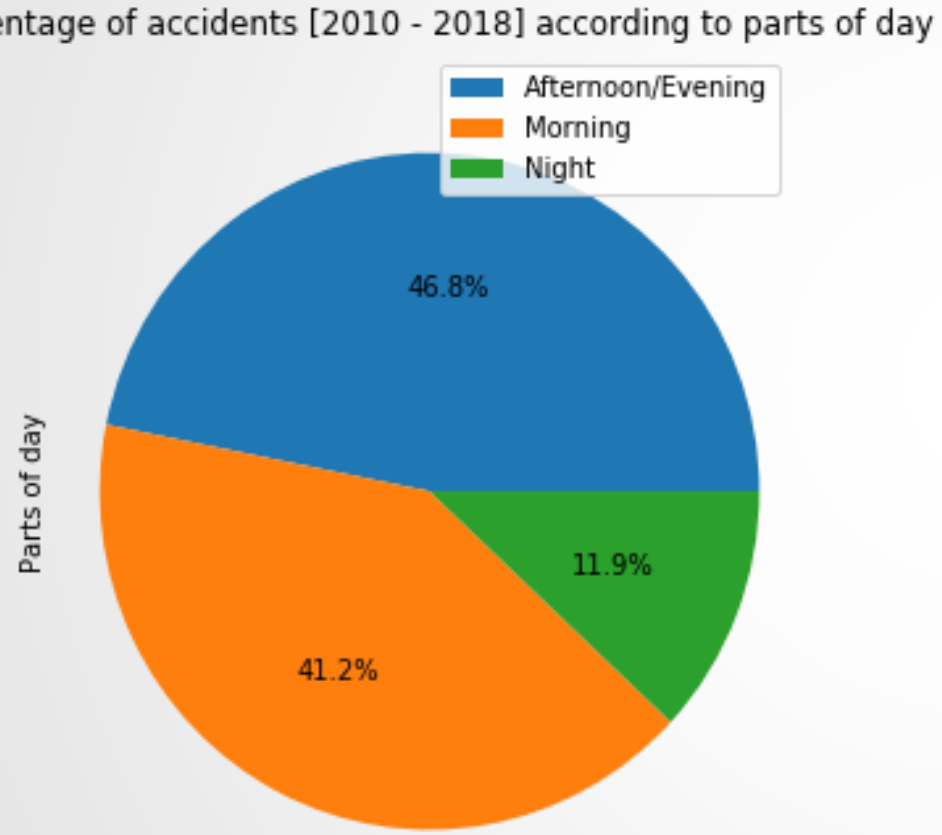
- It consists of a structured data set with information on traffic accidents with deaths or serious injuries that have occurred in Catalonia since 2010. Among the data fields we find road, light and weather conditions when the accident occurred, number of fatalities or injured people, type of road, time/day of the week, etc. Quantitively speaking, it records more than 16,000 accidents from 2010 to 2018, and for each of them we find 58 fields.

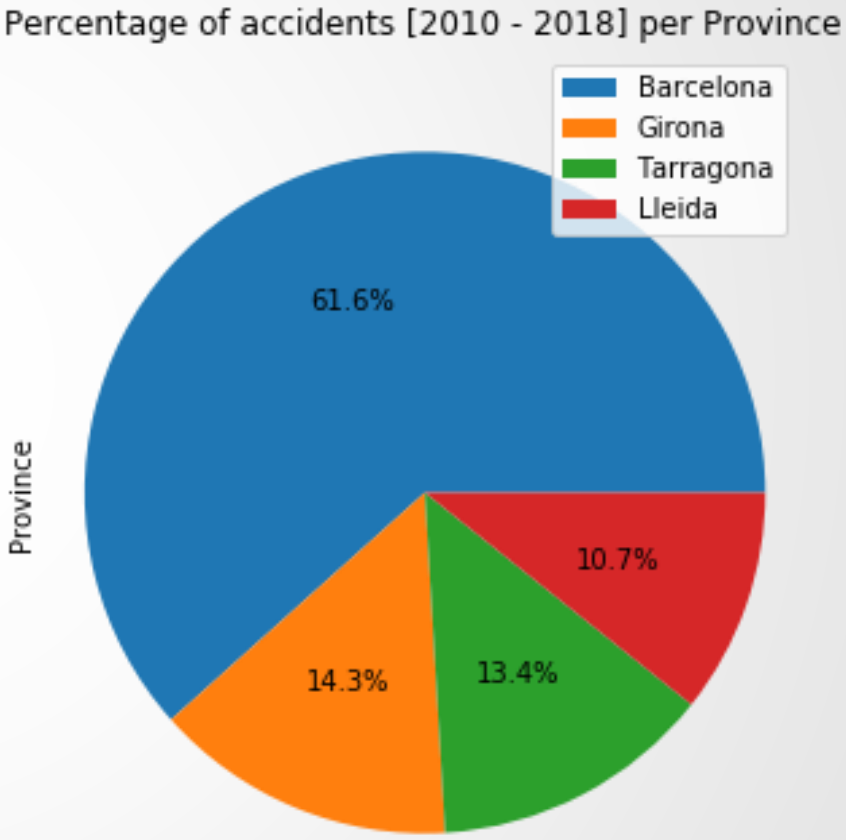
DATA VISUALIZATION



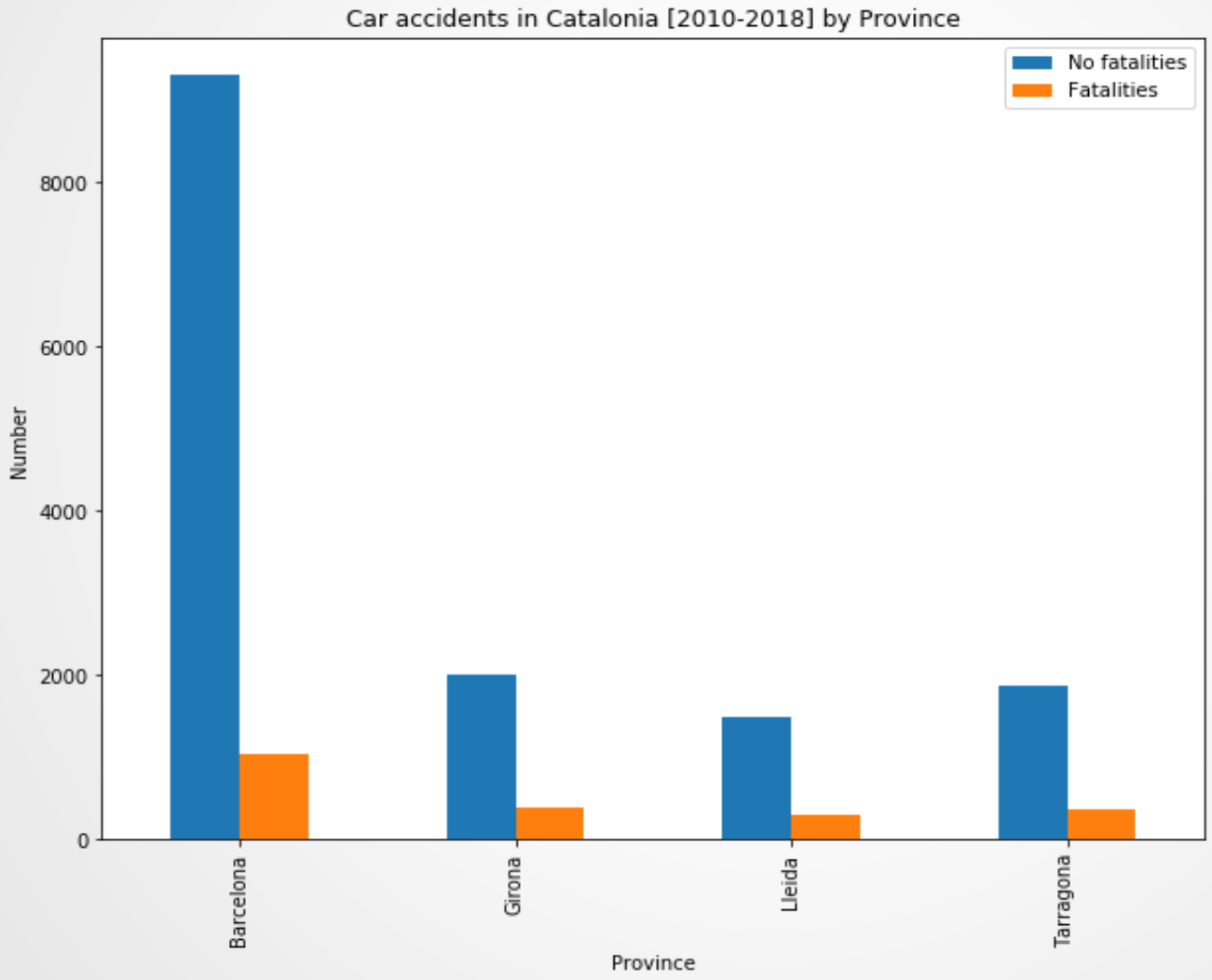


DATA VISUALIZATION





DATA VISUALIZATION

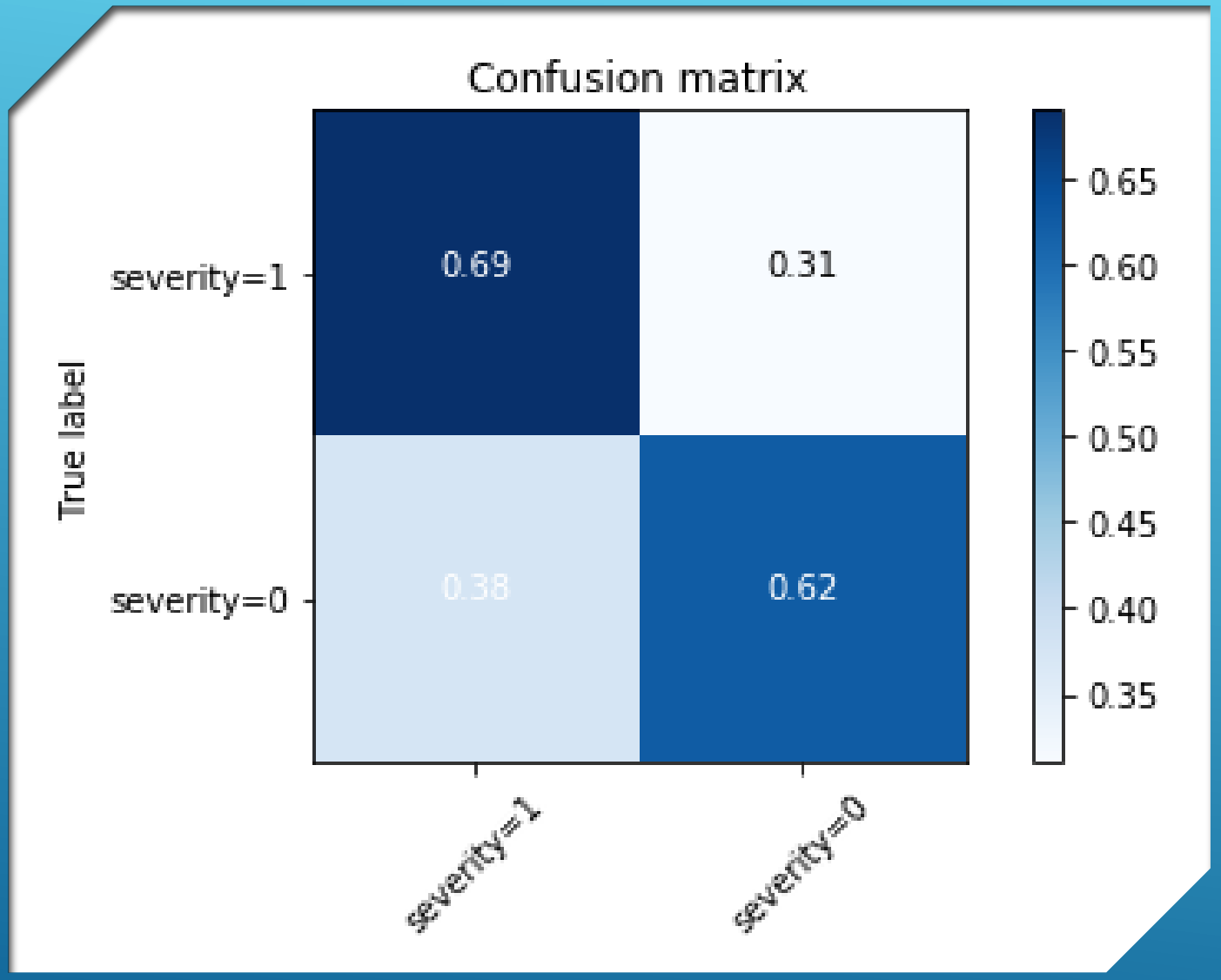


MODEL: LOGISTIC REGRESSION

- Our decision has been to apply a Logistic regression model, which is very appropriate to determine a 0/1 type of prediction. We want to simplify the severity of the accident in terms of fatality occurred (1) or not (0).
- We create target variable Y that it's 1 if there are any number of fatalities, 0 if none:

```
Y = (df_data_1['F_MORTS'] > 0).astype(int)
```

- Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression is estimating the parameters of a logistic model (a form of binary regression).

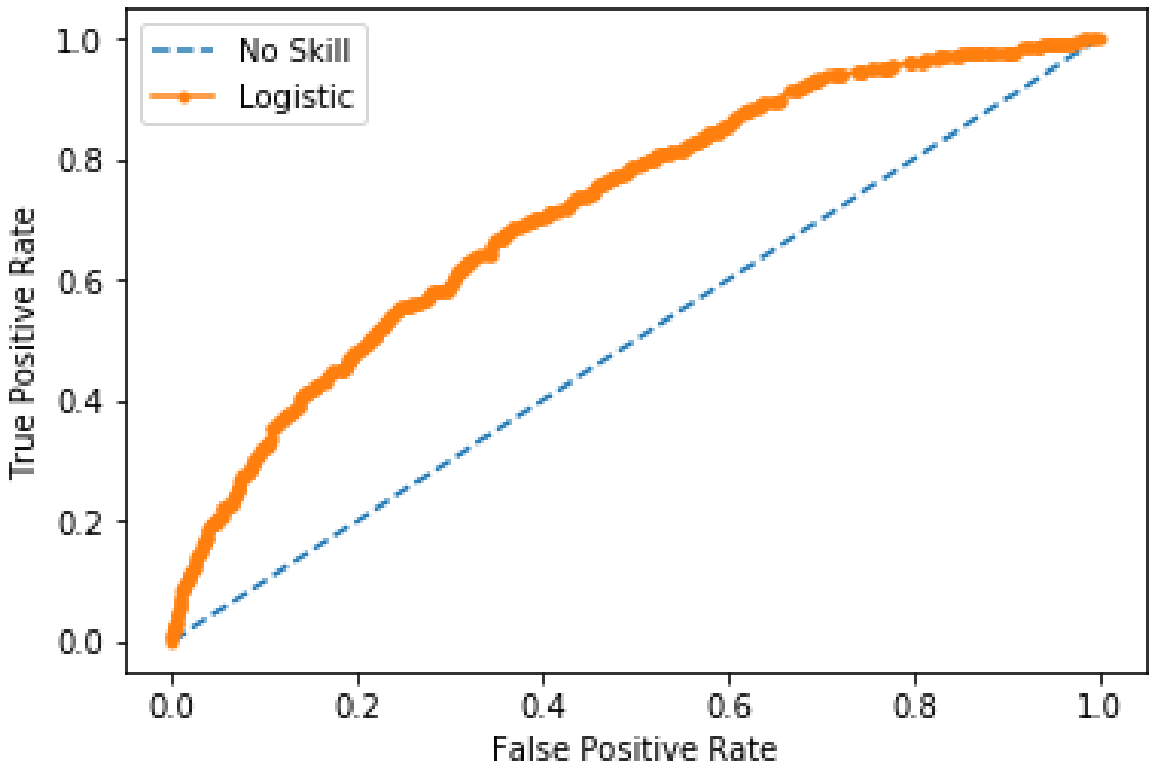


MODEL RESULTS AND EVALUATION

CONFUSION MATRIX

This is a specific table layout that allows visualization of the performance of an algorithm. Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class (or vice versa). The name stems from the fact that it makes it easy to see if the system is confusing two classes (i.e. commonly mislabeling one as another).

In our case we find that the true positives are 69%, and the true negatives 62%.



ROC CURVE AND AUC

►ROC Curves summarize the trade-off between the true positive rate and false positive rate for a predictive model using different probability thresholds.

►A no-skill classifier is one that cannot discriminate between the classes and would predict a random class or a constant class in all cases. A model with no skill is represented at the point (0.5, 0.5). A model with no skill at each threshold is represented by a diagonal line from the bottom left of the plot to the top right and has an AUC of 0.5. In our case, our predictor is much better, and reaches a value AUC of 0.71.

CONCLUSIONS AND FUTURE DIRECTIONS

In this study, we have analyzed the relationship between fatalities in accidents and their associated conditions at the moment of occurrence, like time of day, day of week, weather/light/wind conditions, etc.

We have built a logistic regression model that, given certain conditions at every moment, predicts the probability of having fatalities in case of an accident, i.e., its severity as defined in this problem.. These models can be very useful in helping authorities in taking real-time decision to close roads, or give recommendations to drives via some panels or other technology. Also, it will be possible to integrate these algorithms in the vehicle for assisting drivers in route selection and overall decision making.

We believe that, given the available data, the algorithm gives notable results, and to improve it further we may need to incorporate other data sets which will enrich the features with more accurate weather conditions, etc. Multiple data sets (combined in a synchronous way) should be necessary to improve the algorithm performance.