Predicting life-threatening car crashes in the UK using Machine Learning

Final project report

Fragkiskos Malliaros

Flora Attyasse
CentraleSupelec
Paris, France
flora.attyasse@student-cs.fr

Vadim Benichou
CentraleSupelec
Paris, France
vadim.benichou@student-cs.fr

David Kiskovski
Centrale Supelec
Paris, France
david.kiskovski@student-cs.fr

Marc Terzikhan
CentraleSupelec
Paris, France
mark.terzikhan@student-cs.fr

Abstract

The characteristics of car crashes that occurred in 2017 in the UK were analysed and models using various machine learning methods predicting their severity were written. This was done in order to get a better grasp on what causes deadly accidents, which could help in turn prevent them. Three datasets containing information on the accident, the driver and the casualties were treated, cleaned, modified and merged to accomplish this task. Features ranging from engine capacity of cars, types of roads, type of junctions, age of driver and time of crash were used. A Ridge feature selection was applied on the final data in order to select the best features for the model. Parameters such as the driver being a man, the casualty being elder, the accident occurring on a Friday or weather conditions including rain were seen to have the highest probability of causing a fatal accident. The predictions had a 78% accuracy for the gradient boosting method, and an AUC score of 0.53.

1. INTRODUCTION

Car accidents are responsible for more than 1.25 Million deaths worldwide each year [1]. They are the leading cause of death for people aged years 15 to 29.

Although the issue is especially major in impoverished countries, a lot is still to be done in the more developed ones, such as the UK. In the UK alone, close to 2000 deaths were recorded in 2017.

The World Health Organization adopted a new agenda for 2030to cut fatalities related to car accidents by half. This effort is thought to be attainable partly thanks to the development of poor countries, but also as a result of the new tools machine learning is putting at our disposal.

The objective of this project was to understand influencing factors of car crashes in the UK. Correlations had to be found between different parameters involved. The aim in turn was to infer and predict car fatalities, in order to reduce them as much as possible. Indeed, the observations from the project could be used for applications in:

-safe route planning,

-emergency vehicle allocation,

-roadway design,

-accident prevention

-where to place additional signage (to warn for dangerous curves for example)

-implementation of additional crucial functionalities to the future of the automobile industry: autonomous vehicles.

This year at the Esri DevSummit in Palm Springs, it was demonstrated how data could be used to predict accident probability. An iOS application was built by the APL (Applications Proto-type Lab) proposing the path using the

safest route available,allowing drivers to route around accident prone areas, and to ride in the safest roads[2].

2. THEORY

Multiple models and feature selection techniques were used in this model. In order to better understand them, the relevant theory is explained here for each one used.

1. Decision Tree

Decision trees runs simple tests conditioned on a choice over all attributes , hence creating a tree. The leaves are assigned a majority vote and the process is repeated for all branches until the tree incorporates all features[3].

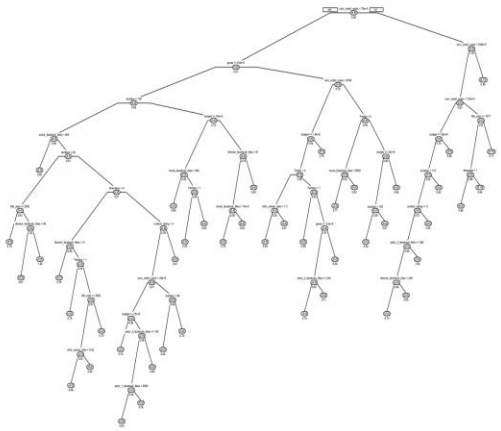


Figure 1: Example of decision tree

2. K-Nearest Neighbor

Uses distance function to compute a nearest neighbor algorithm. To do this, all training examples are store and for a point x, assign the label of the training example closest to it. This can be used in classification using the majority vote, hence predicting the class of the most frequent label among the k neighbors.[3]

3. Gradient Boosting

Gradient boosting is an ensemble method which is typically applied in classification. It is constituted by a set of classifiers whose decisions are combined using a majority rule. The boosting is a sequential method, iteratively reweighting the training examples, as shown in figure 2.

4. Ridge

Ridge is a feature selection technique used to get rid of all features the do not impact the model in a big way and only keep the ones that are important. It does so by shrinking coefficients to 0 if its related feature has negligible impact on the model. Its mathematical formulation is shown on figure 4

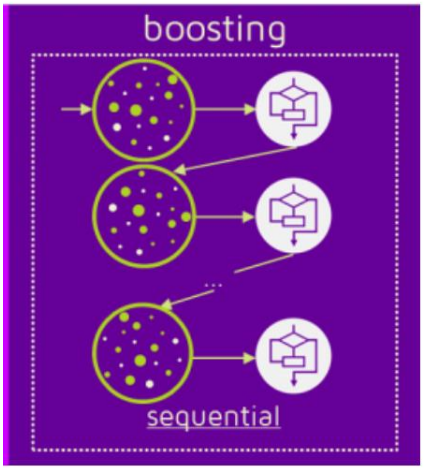


Figure 2: Working mechanism of boosting ensemble method [3]

$$J_{\mathrm{RR}}(\boldsymbol{\theta}) = J(\boldsymbol{\theta}) + \lambda||\boldsymbol{\theta}||_2^2$$
$$= \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{\theta}^T\mathbf{x}^{(i)} - y^{(i)})^2 + \lambda\sum_{k=1}^{K}\theta_k^2$$

prefers parameters close to zero

Hyperparameter λ: how much should we trade-off accuracy versus complexity

Model complexity

Figure 4: Ridge process

3. LITERATURE REVIEW

The domain of traffic accidents is highly studied nowadays thanks to machine learning.

In 2004, Miao Chong, Ajith Abraham and Marcin Paprzycki worked on Traffic Accident Data Mining Using Machine Learning Paradigms[4], having for aim to develop a prediction model that automatically classifies the type of injury severity of various traffic accidents.

For this, four machine learning models were used: support vector machines, neural networks trained using hybrid learning approaches, decision trees and finally a concurrent hybrid model involving decision trees and neural networks. The data used came from the National Automotive Sampling System (NASS) and the General Estimates System (GES) providing representative probability samples

from the annual estimated 6.4 million accident reports in the United States.

The initial dataset for the study contained traffic accident records from 1995 to 2000, a total number of 417,670 cases and as features : labels of year, month, region, primary sampling unit, the number describing the police jurisdiction, case number, person number, vehicle number, vehicle make and model; inputs of drivers' age, gender, alcohol usage, restraint system, eject, vehicle body type, vehicle age, vehicle role, initial point of impact, manner of collision, rollover, roadway surface condition, light condition, travel speed, speed limit and the output injury severity.

They defined 5 injury classes : no injury, possible injury, non-incapacitating injury, incapacitating injury, and fatal injury.

The best results were obtained through the last hybrid model.

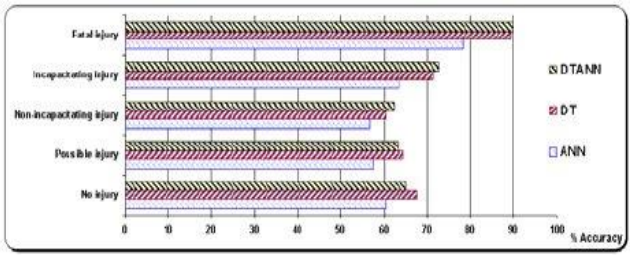


Figure 5: DT : Decision Tree – ANN : Neural Networks – DTANN : the hybrid combination.

More recently, S. Vasavi presented results from a research study on road accident data of major national highways that pass through Krishna district for the year 2013, entitled Extracting Hidden Patterns Within Road Accident Data Using Machine Learning Techniques[5].

The datasets came from police stations and full of noise, missing values, and inconsistencies. Missing values were replaced with NULL.

A clustering approach was used using K-medoids, and expectation maximization algorithms which were then analyzed to discover hidden patterns using a priori algorithm. Results showed that the selected machine learning techniques are able to extract hidden patterns from the data. Eight clusters were created:

• which accidents happen because of low and high traffic

• time of accident cluster in which accidents happen (morning, afternoon, evening, and night)

• age of the drivers

• accident occurred month

• weather condition at the time of accident

• lightening condition at the time of accident

• type of accident (rash driving, overlooking, vehicle, skidding…)

• speed limit of vehicles at the time of accident

F-measure was used for cluster analysis for performing node-based analysis.

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

F-measure = (1 + alpha) / ((1/Precision) + (alpha/Recall)), *with alpha = 1

Table 6 Performance measures for K-medoids and EM algorithm

Dataset	Precision		Recall		F-measure	
	EM	K-medoids	EM	K-medoids	EM	K-medoids
1504 tuples	0.5	0.8	0.4	0.6	0.45	0.69

4. DATA PROCESSING

A. Data Structure

The raw dataset used comes from the UK government data website and is composed of 3 datasets:

- an Accident set of nearly 130,000 instances with 32 features
- a Casualties set of 240,000 instances with 16 features
- a Vehicle set of 171,000 instances with 23 features.

For a hence total of 69 features, each row corresponding to an accident.

Examples of features	Description
Road Type	Roundabout, Single or dual carriageway
Accident Severity	Slight, Serious, Fatal
Number of Casualties	numbers
Day of Week	Monday to Sunday

B. Data Reduction

Since the dataset contained a lot of variables and observations, the data preparation performed constituted a major part of the project. The first inspection of the dataset was performed in order to get rid of any: redundant features (for example, multiple chronological and spatial components), as well as features too specific for the scope of this project. Indeed, the dataset contained detailed information on the circumstances surrounding the accident, such as the presence of police forces at the scene, or the specific type of pedestrian crossing that could have been

present. The location variable of Easting and Norting OSGR were hence removed as longitude and latitude were already present. Thus, the project focused only on key explanatory variables of accidents.

C. Handling Missing Values

After data reduction the dataset nevertheless still contained several features with a high percentage of missing values.

Information is originally present in the data in two different ways: discretized with integers or labels with codes.

Missing values were marked on the entierty of the data set as -1 values. Hence, all -1 in the 3 datasets were substituted by Nas. There were also missing values represented by other numbers in the data, so each feature was investigated individually to make sure that all elements are relevant.

First, in order not to retain too many variables and impute too many random values, a decision was made to remove all features with a missing values percentage of above 50%.

1. Accident Data:

For some features possessing a very low amount of missing values (less than 10) rows associated were simply deleted as they were defined as missing completely at random (MCAR) and insignificant for data analysis. This was the case for the Urban_Rural_road.

The same was performed on the Longitude, Latitude and Time.

Accident_Index	0
Longitude	29
Latitude	29
Accident_Severity	0
Number_of_Vehicles	0
Number_of_Casualties	0
Date	0
Day_of_Week	0
Time	3
1st_Road_Class	0
Road_Type	0
Speed_limit	0
Junction_Detail	609
Junction_Control	56296
Light_Conditions	1
Weather_Conditions	1
Road_Surface_Conditions	1937
Urban_or_Rural_Area	0

Figure 7: Accident missing values

The feature Junction_Control was observed to hold a large amount of missing values, which was counterintuitive as it was so closely linked with Junction_Detail (possessing 609 missing values). Upon investigation, it was discovered that the value '0' for Junction_Detail representing no junction involved in the accident was displayed as a missing value in Junction_Control. Hence the missing values in Junction_Control could be replaced to 0s when the corresponding value in Junction_Detail was also 0. The Junction_Control feature was then left with just over 100 missing values, hence the corresponding rows were dropped. The remaining Nas for Junction_Detail were found using a decision tree which was trained on every feature in the dataset apart from the time, the index and Road_Surface_conditions as they would not impact the result, being uncorrelated. The accuracy on this decision tree was computed using the training data with no missing values in that column, and was 79.34%.

The Weather_Conditions feature seemed to only contain 1 missing value, however in this case it had another value than -1 representing them. After finding all missing values associated with this feature, it was observed that it possessed just as many as the "Road_Surface_Conditions" feature (1937). This makes sense as the two features are very strongly correlated, as raining conditions would induce a wet road, just as snow an icy one. Both features are nonetheless kept, as weather conditions indeed impact the condition of roads, but also driver visibility which could be an important factor inducing accidents. The weather missing values were then estimated using another decision tree trained on the month of the year (which is a feature that was added based on the column data) as well as the location (longitude and latitude) where the accident took place. This was thought to be a good idea as meteorological patterns are heavily impacted by the location and the time of year. The accuracy of this tree was 81,94%. A Knn method could also have been used to find these values as a distance parameter was used, but the accuracy obtained was already satisfactory.

The weather conditions then helped find the missing values of the road surface feature by applying another decision tree using it in its training data.

Finally, the only feature with missing values remaining was the 'Light_conditions' column. This contained two types of Nas and "darkness but no lighting information". This was dealt with by using the fact that the "Time" feature would be able to determine that this parameters would come up in night time, and probably on a rural road, hence applying a decision tree would yield good results. Its accuracy was 83%. Indeed all decision trees with accuracies above 75% were thought smart to keep.

Finally, the '9' value of the 'Road_type' feature also constituted missing values. A good correlation between this column and those for 'Junction_Detail' and 'Urban_Rural_road' was found hence the missing values

were filled using a group by method, by taking the mode of each parameter already existing to fill the missing values.

2. Casualty Data:

The missing values for the casualty dataset were computed and are seen in figure 8.

```
(Accident_Index              0
 Casualty_Class              0
 Sex_of_Casualty            27
 Age_Band_of_Casualty     2214
 Casualty_Severity           0
 Casualty_Type               3
```

Figure 8: Casualty missing values

Some are present for "sex" which would be hard to estimate and are not very important as there are so few, hence the associated rows were dropped. The same was applied when dealing with the three missing values in the Casualty_type. Finally, although "Age_band" contains more Nas relatively, they do only represent 1.2% of the total dataset, and the decision to drop their rows was taken.

It may appear as a lot of rows were dropped arbitrarily, however the datasets dealt with in this project were very large, and they had to be merged in the end, hence getting rid of some rows that did not provide an optimal amount of information was not important in the scale of whole work being done.

3. Vehicle Data

```
Accident_Index              0
Vehicle_Type              293
Vehicle_Manoeuvre        3957
1st_Point_of_Impact      3961
Sex_of_Driver               3
Age_Band_of_Driver      27774
Engine_Capacity_(CC)    53381
```

Figure 9: Vehicles missing values

For vehicles, the sum of missing data is observed on figure 9. The features "1st_point_of_impact" and "Vehicle Maneuver" contained the same missing values, hence it was impossible to predict them according to one another. The rows for their missing values were henceforth dropped. Vehicle maneuver is then left with around 1000 missing values, which were filled using the mode according to the 1st point of impact feature. This made sense as some poinst of impact would without a doubt originate from specific vehicle maneuvers.

The many missing values of "Age_Band_of_Driver" were also chosen to be dropped. As stated previously, the amount of Data allows for flexibility when dealing with missing values. Very few Vehicle_type NAs remained hence were dropped.

Finally, the engine capacity feature missing values were dealt with in the following way: for rows with "Vehicle_type" values representing vehicles without motors, the column was attributed a "0". For the remainder, the mode of the engine capacity of specific vehicle types were inputted to the missing values of each of these vehicle types. The mode is chosen again here as we expect the data to be skewed.

The feature of the sex of driver has its value "3" also representing missing values. These were attempted to bed filled using a decision tree, however its accuracy was below 70%, hence it was chosen to keep the missing values for now and attempt to fill them again once the whole data was merged together.

This paid off as an accuracy for the decision tree trained with the merged data was 79.62%.

D. Data Engineering

1. Accident Data

A feature representing the month of the accident was added as it was a more valuable generalization of the time frame as the actual date was.

The accident severity was also transformed into a binary component, with slight damage being a 0 and fatal and serious damage a 1. This was done as the aim of this project is to be able to predict when life threatening accidents could take place, hence "serious" also constituted concerns. "None serious" accidents constituted more than four times more observations than its counterpart. This data could have been normalized but was not due to reasons that will be discussed later in this report.

2. Casualty Data

The "Casualty_type" feature was simplified by regrouping some of its components together. Indeed it originally constituted all possible types of vehicles involved in the crash. However, all similar vehicles were instead put together, signifying all types of cars, mini trucks, trucks,

motorcycles, pedestrians, bicycles etc… This was done to avoid any potential overfitting when running models, as some of the original vehicles were very rare.

The age band feature was also modified as it was decided that the bands were too narrow. Hence the new division incorporated infants, children, young adults (16-35), adults, and seniors over the age of 65. This was done to get more generality in the model.

The feature "Casualty_severity" was discussed in length. It was however determined that using it in the final model would not make sense, as the severity of the casualties condition would be directly correlated to the severity of the overall accident, which is what this project tries to predict. Hence this feature was finally dropped.

3. Vehicles Data

The same bins were implemented as in the previous dataset for the age of the driver, as well as the types of vehicles that was being driven.

It was then thought to be a good idea to check for the engine capacity of each vehicle according to its vehicle type, in order to determine if it has a powerful motor and hence can reach high speeds. This was done according to vehicle type as a motorcycle with 300cc is considered powerful as opposed to a car with the same capacity. The maximum speed that a car could reach could impact the odds of its driver to survive an accident with it.

The 3 datasets were then merged and the final dataset had the size of 288635 rows and 29 columns. The merging was possible thanks to the column "Accident index".

5. DATA VISUALIZATION

Visualization was a first necessary step to gain primary insights by highlighting the information contained or missing, especially with such a high-dimension, in the dataset.
The aim here was hence to figure out inferring features but also look for outliers and missing values regarding accidents in the UK.

How did accidents compare? It was quickly noticed that scatter plots of a feature against another were only relevant to observe outliers since plotting similar discrete values variables creates overlapping.

Time Analysis

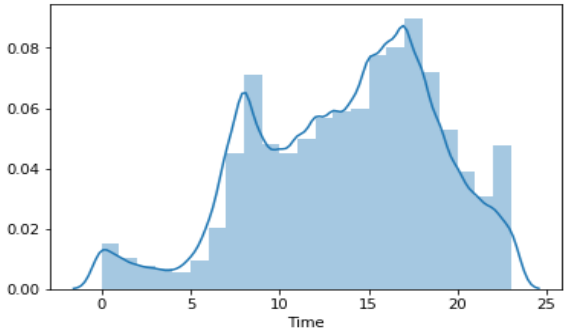


Figure 10 Time distribution of accidents 24h frame

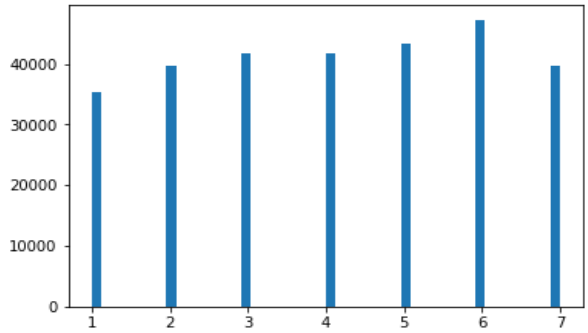


Fig11: Day of week accidents distribution, (1) being Sunday

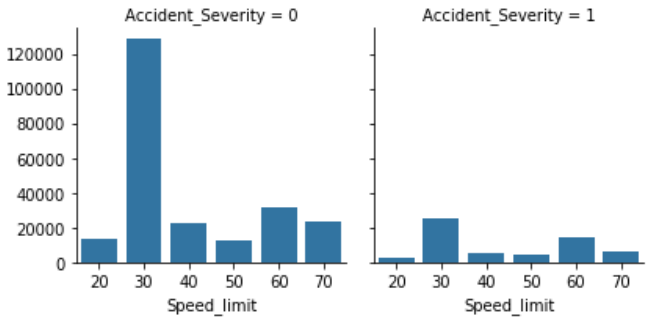


Figure 12: Roads' speed limits according to the severity of the accident

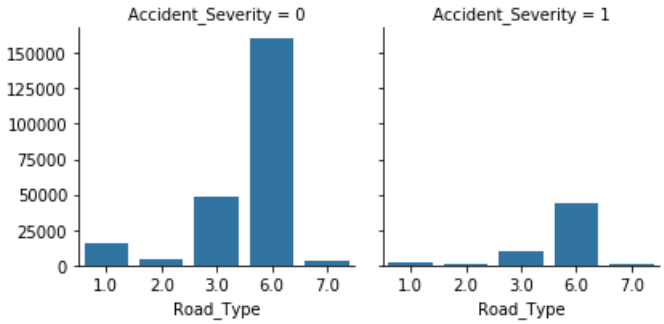


Figure 13: severity regarding road types

Severity regarding road types

It was noticed that accidents arose uniformly over the week; a slight peak for Friday could be noticed on figure 11.
Figure 10 revealed two interesting peaks: peak of accidents at 8 am in the morning and main one between 3pm to 6pm in the afternoon.
This probably corresponded to the main parts of the day cars are taken.

The UK car speed limits is the following: 30mph in urban areas, 60mph on main single-carriageway road, 70mph on dual carriageways and motorways. Those main road types hence explained the distribution shown on figure 15 as they were largely the most used across the country.

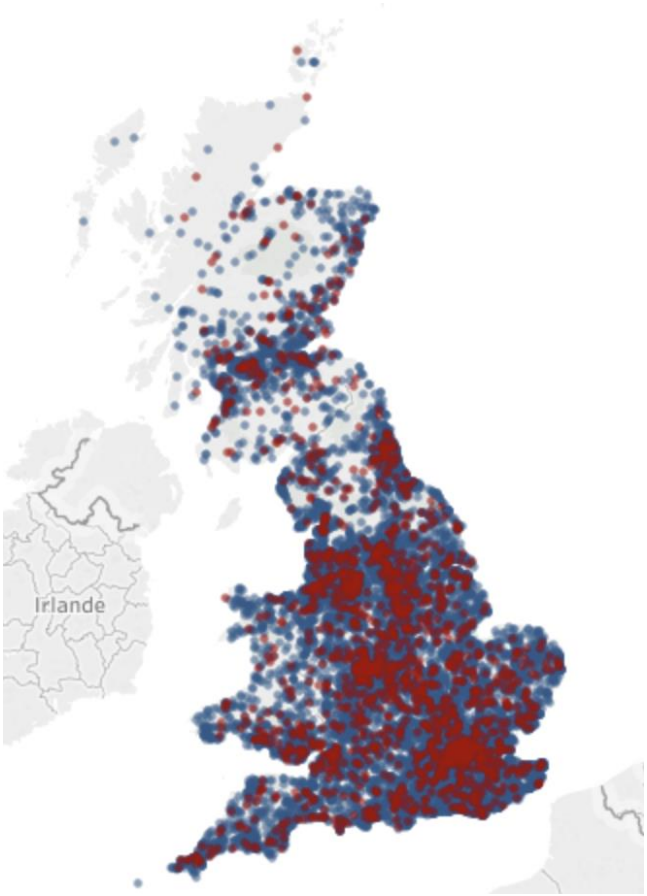


Figure 14 UK Accident Severity Map

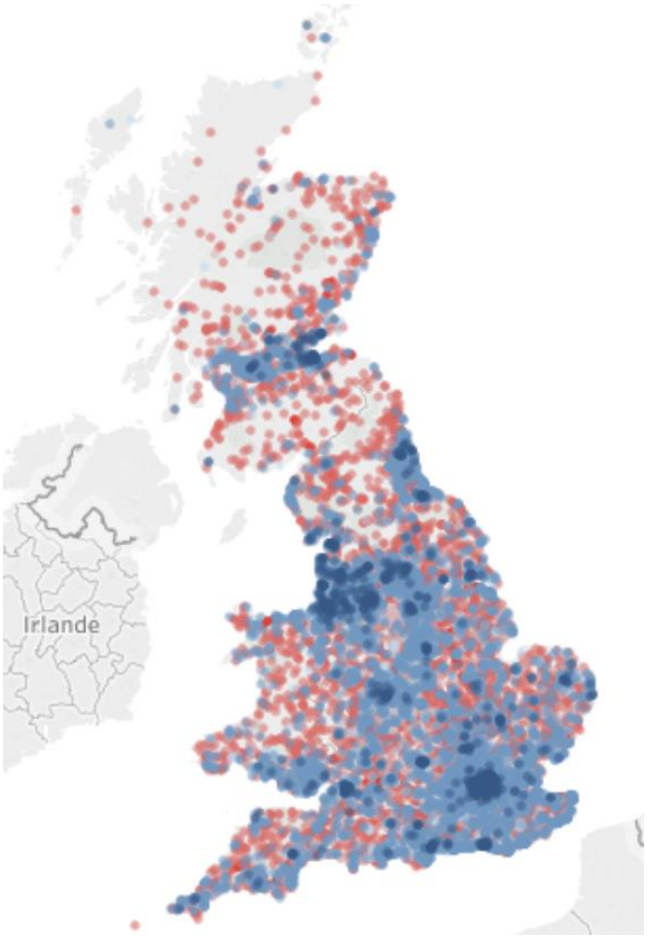


Figure 15 Roads' Speed Limit of accidents Map

6. MODELLING

Ridge Regression was used as features selection to estimate the weight impact of all our variables on accident severity.

In The end, only variables with significant coefficients were kept and the others dropped in the test set.

The final data set was reduced to 9 explicative variables composed of:

- Day of week
- 1st road class
- Sex of the driver
- 1st point of impact
- Age of casualty
- Casualty class
- Month
- Weather condition
- Road surface condition

It was observed that men had a tendency to cause more serious accidents relative to their female counterparts.
This analysis also showed that casualties aged over 65 years old were the most likely to pass away. Those kinds of feature cannot really be reduce with some precaution, because it is impossible to disallow certain groups of individual to drive on roads, hence no measures can be taken to limit them.

However, it also appears that most of the dangerous accident are on Friday, and a big proportion of them are in July. For those variables some prevention measures can be put in place or accentuate during this period in order to reduce accident or accident gravity.
Several models were used after the cleaning process:

- Logistic Regression
- Decision Tree
- Random Forest
- K-Nearest Neighbors
- Gradient Boosting

Models were tested on a Test Set accounting for 20% of the dataset. Another 20% were given for a Validation Set.

Choice of evaluation metric

As seen through the data visualization, the distribution of the accidents was highly skewed. This account for the fact that there were way more instances being slight accident than fatal ones.
The idea of normalizing the instances arose during the project but the distribution was considered to be nonetheless quite realistic. The slight/fatal ratio seems to be indeed observed and decision was hence taken not to decrease the slight accidents' weights.

Choice was made toward the accuracy metric. Accuracy fits normally better for nearly uniformly distributed classes; but in the trade-off between potentially smaller metric value versus a realistic point of view, the second option was retained.

	Model	Validation	Test
4	Gradient Boosting	79.61	78.22
1	Regression	79.61	78.15
0	KNN	74.71	73.95
3	Random Forest	74.36	73.41
2	Decision Tree	72.91	71.84

Figure 216: Accuracies for models

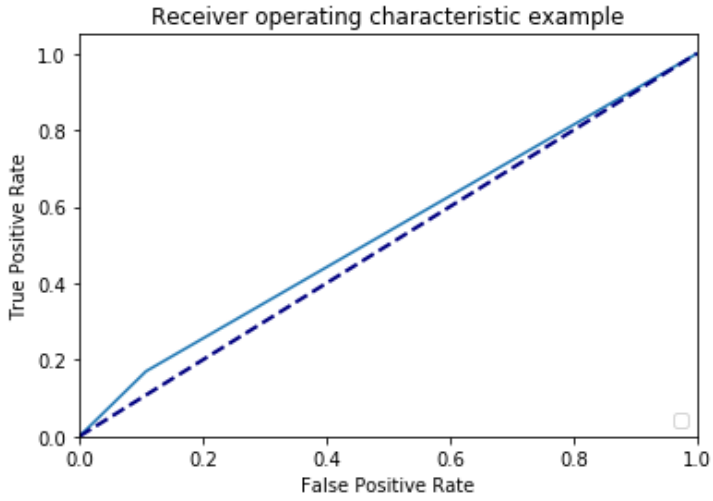
The highest accuracy reached was through the Gradient Boosting model and the Regression with respectively 78.22% and 78.15%
The smallest value obtained was with the Decision tree model with a value of 71.84%.

Next step to take was to compute Confusion matrices for each model: predicted vs true class label and computing the ROC Curve and the AUC score.

Gradient Boosting Confusion matrix

Predicted class \ True class	Positive(1)	Negative(0)
Positive(1)	44977	135
Negative(0)	12438	177

ROC CURVE obtained on Gradient Boosting model:



Computation of the Area Under the Curve
ROC_AUC : 0.5055191818528977

The modelling was also performed without using the Ridge, but the obtained results were inferior to when Ridge was used.

7.CONCLUSION

Modeling provided descent results regarding our purpose of identifying influencing factors. A prediction accuracy of 78% was obtained.
However, further steps could be taken regarding the features in order to improve the accuracy.

An interesting way to deepen the models would be to add extra relevant features regarding car accidents. It is famous that lack of concentration, emotional states as a strong anger are ones. These would probably reveal to be relevant factors. Finally, the autonomous car which is to be delivered in a handful of years will definitely transform the driving experience and reduce drastically the number of accidents. The freedom of driving will be the main trade-off and an important societal issue in the years ahead.

REFERENCES:

[1] Number of road traffic deaths ,Global Health Observatory (2013). WorldHealth Organization.

[2] Using Machine Learning to Predict Car Accident Risk, Daniel Wilson (2018). GeoAI.Availableathttps://medium.com/geoai/using-machine-learning-to-predict-car-accident-risk-4d92c91a7d57

[3] Machine learning course, 2018 CentraleSupelec, Fragkiskos Malliaros

[4] Traffic Accident Analysis Using Machine Learning Paradigms, Miao Chong, Ajith Abraham and Marcin Paprzycki. (2004). Fermi National Accelerator Laboratory, University of Chicago.

[5] Extracting Hidden Patterns Within Road Accident Data Using Machine Learning Techniques ,S. Vasavi, 2013