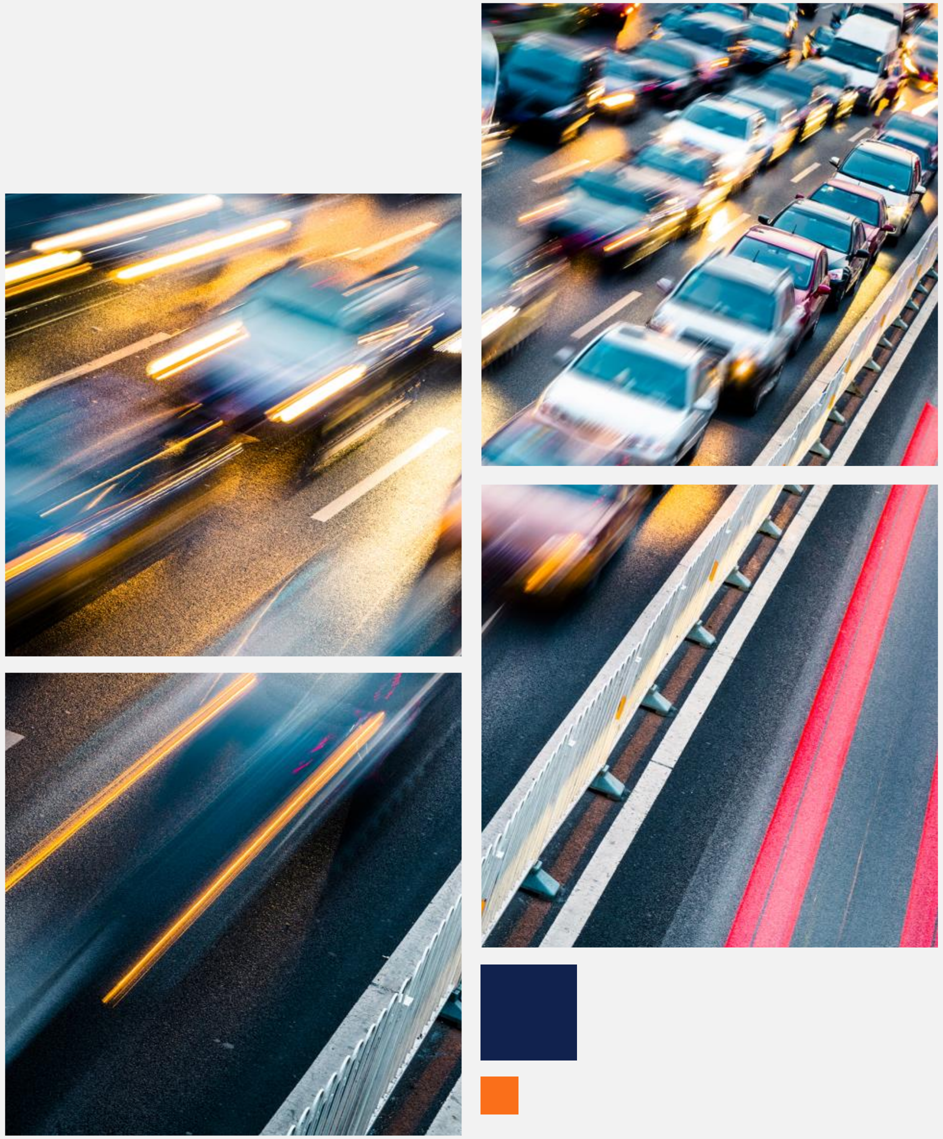
UK Road Safety Data Analysis: Understanding Accidents & Predicting Casualties

November 2022

Kyle Hincz

Introduction

- Road accidents are a traumatic experience for everyone involved and frequently end in death or serious injury.
- One of the priorities of the UK Government and the police force is to actively monitor and prevent traffic accidents.
- Despite this, in 2021 the UK reported over 100k vehicle accidents with over 128k casualties. Of those, 1.6k persons were killed and 25k were killed or seriously injured (KSI).
- This project will analyze all vehicle accidents as reported by the UK Department for Transport. The aim of the project is to understand the nature of these accidents and their underlying causes.
- The project applies machine learning and statistical methods to extract useful insights and predict serious casualties of these incidents.



The Dataset

The dataset has been obtained from the UK government website: Road Safety Data - data.gov.uk

It provides road safety data about the circumstances of personal injury road accidents in the UK from 1979 to 2021. The data contains only non-sensitive fields that can be made public. There are 3 main tables:

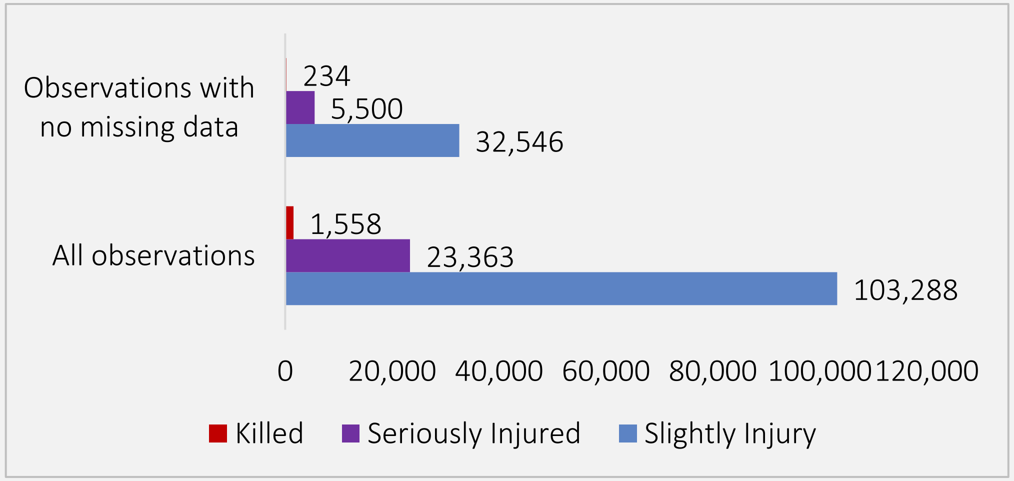
Table	Data Description	# of records in 2021	# of variables
Accident	List of accidents with date, time, location, road parameters, light conditions, number of casualties and vehicles, speed limit, weather conditions.	101,088	36
Casualty	Age & sex of casualty, whether pedestrian/driver/passenger, severity of injury	128,210	19
Vehicle	Driver age, sex, vehicle age, engine capacity, car make & model, which object was hit, journey purpose, vehicle maneuver, vehicle type	186,444	28

All three tables were joined so that each casualty record in the final data frame has corresponding vehicle and accident variables.

Data Pre-processing

Dealing with missing data

- Any missing data in the dataset was encoded as “-1”
- Comparison of balanced accuracy of models with and without incomplete records was performed
- Removing incomplete records resulted in better model performance so these observations were removed for the final analysis



Variable Encoding

- Original dataset obtained was clean with majority of variables being categorical
- Text variables such as car make/model and geographical coordinates were removed to facilitate analysis
- Continuous variables such as engine capacity and age of car/casualty were coded into deciles
- Finally, all variables were one-hot-encoded into columns containing only zeros and ones. This resulted in 378 columns

Feature Selection

Because of the large number of variables, feature selection was performed.
Importance weightings of features were obtained using the Ridge Regression with 5 fold cross validation. This technique deals well with correlated variables and tries to establish variables that have exactly zero effect. 100 most important variables resulting in KSI were selected and summarised:

Vehicle variables (58/100)

- Type: trucks, motorcycles, agricultural vehicles, buses or coaches
- Hit object off carriageway
- Submerged in water
- Not near junction
- Skidded or overturned
- Age of driver 11-15 yo

Casualty (33/100)

- Truck occupant
- Motorcycle driver/passenger
- Bus/van occupant
- Over 75 or under 10 yo
- Pedestrian
- Drivers significantly less likely to die than passengers

Accident (9/100)

- Extreme weather conditions (flood, snow, winds)
- Road conditions (oil spillage, defective road markings, signs or surface)
- Relatively low importance of rain or day/night

Methodology

Training and Testing split

- Data was split as follows:
  - 80% training data
  - 20% testing data
- Firstly, multiple models were tested with their base settings using 5 fold cross validation
- Top 3 models were selected for further hyperparameter tuning
- Final model was trained and tested using 10 fold cross validation

Choice of response variable and scoring metric

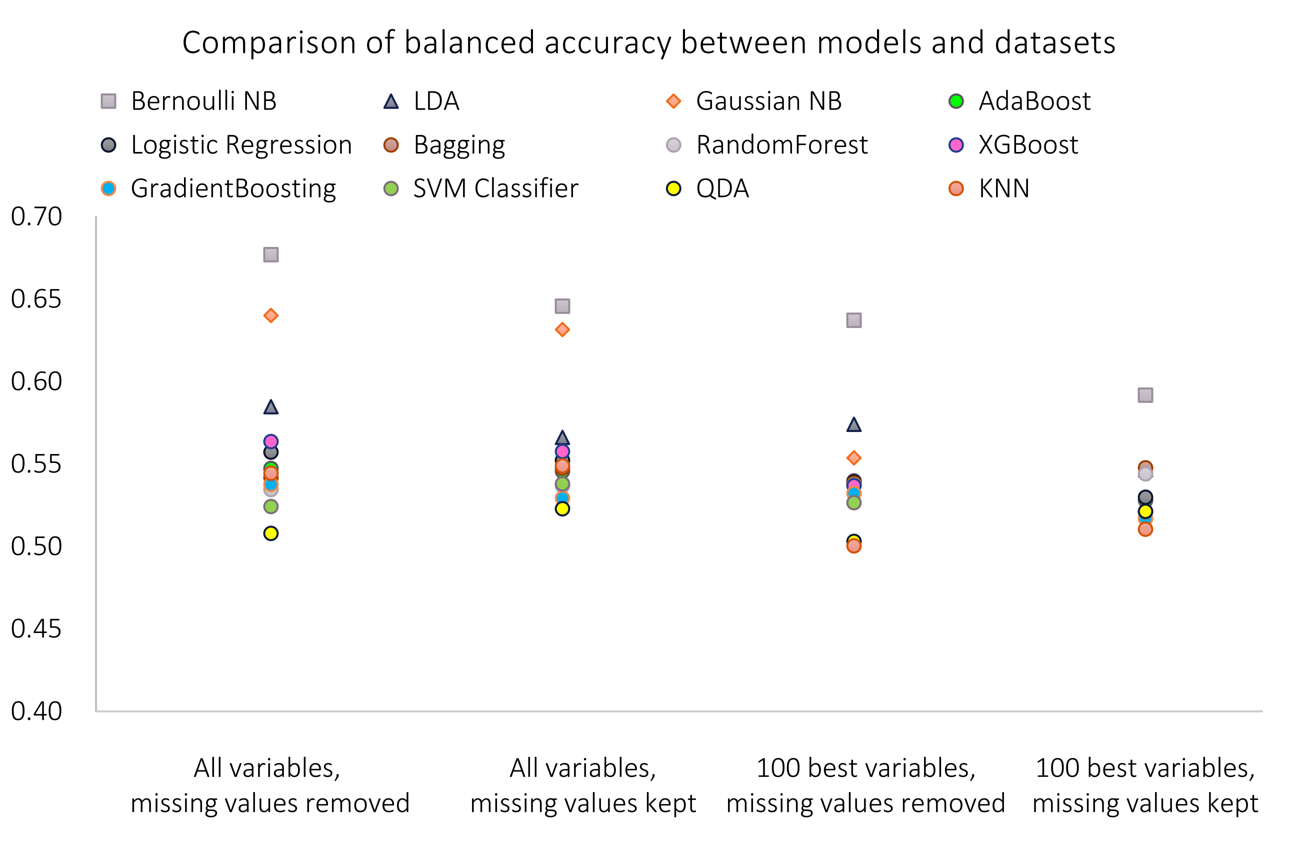
- Response variable: casualty being Killed or Seriously injured (KSI)
- Multiple scoring metrics have been evaluated (balanced accuracy, precision, recall, F1 score). The dataset does not have one scoring metric that accurately captures all its features.
- There are many other variables not captured in the dataset i.e. blood alcohol level, driver mental state, vehicle safety specification) that may influence casualty severity.
- Priority for the model is to predict KSI accurately but as the imbalance of “Not Serious” injury to KSI is approximately 6:1, balanced accuracy will be used.
- Balanced accuracy is the arithmetic mean of specificity and sensitivity and is more robust than standard accuracy.

Methods & Results

Findings

- Tested 12 different models
- Models performed best on full dataset with missing values (rows with -1 in any field) removed
- Bernoulli Naïve Bayes had best balanced accuracy in all cases as well as one of the best F1 scores
- Other models tended to be biased towards the non-KSI class or had either good precision or recall, but not both

Graph of model performance in predicting KSI



Analysis of Results

Naïve Bayes: Explanatory power

- Naïve Bayes determines probabilities of each class associated with the features and predicts class with the highest probability. The model does not offer a standard way of evaluating feature importances.
- Permutation importance has been performed on all features of the test dataset. This measures the decrease in feature importance if a feature is randomly shuffled. Top 10 features are listed on the right.
- Pedestrians are most likely to get seriously injured, especially if they are located in traffic. Motorcyclists and cyclists are also at risk. Drivers are less likely to be injured than passengers.

Most important features

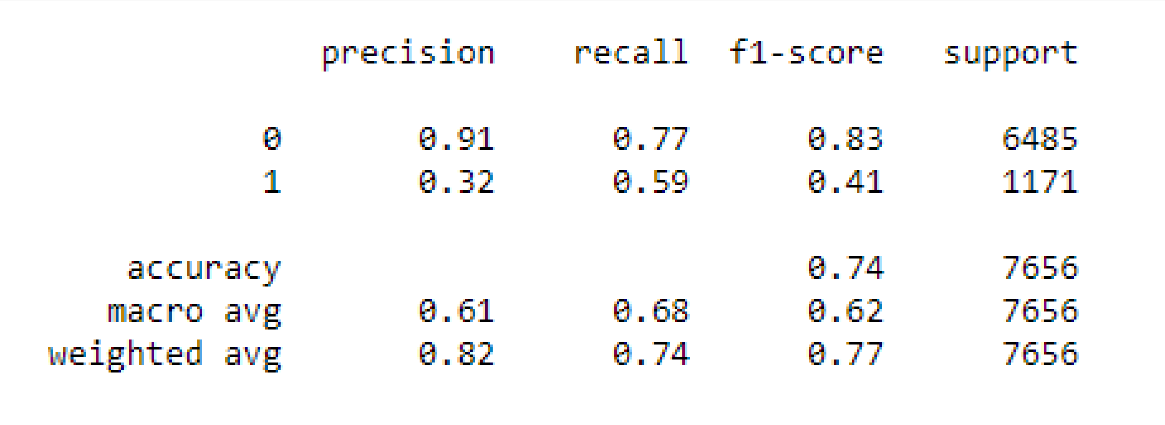
- Pedestrians (casualty class)
- Motorcycles (vehicle class)
- Occupants of motorcycles
- Pedestrians crossing carriageway
- Pedestrians crossing from driver’s offside
- Pedestrians crossing on junction
- Pedestrians walking in carriageway
- Cyclists
- Other vehicle occupants
- Drivers or riders (casualty class)

Results: Final Model

Naïve Bayes: Final Model

Final classification model using the python sklearn Bernoulli NB classifier with optimized parameter alpha of 1.24479 achieved balanced accuracy of 0.68.

Accuracy could be further improved by including more relevant features that the government omits from the dataset to protect identities of casualties.

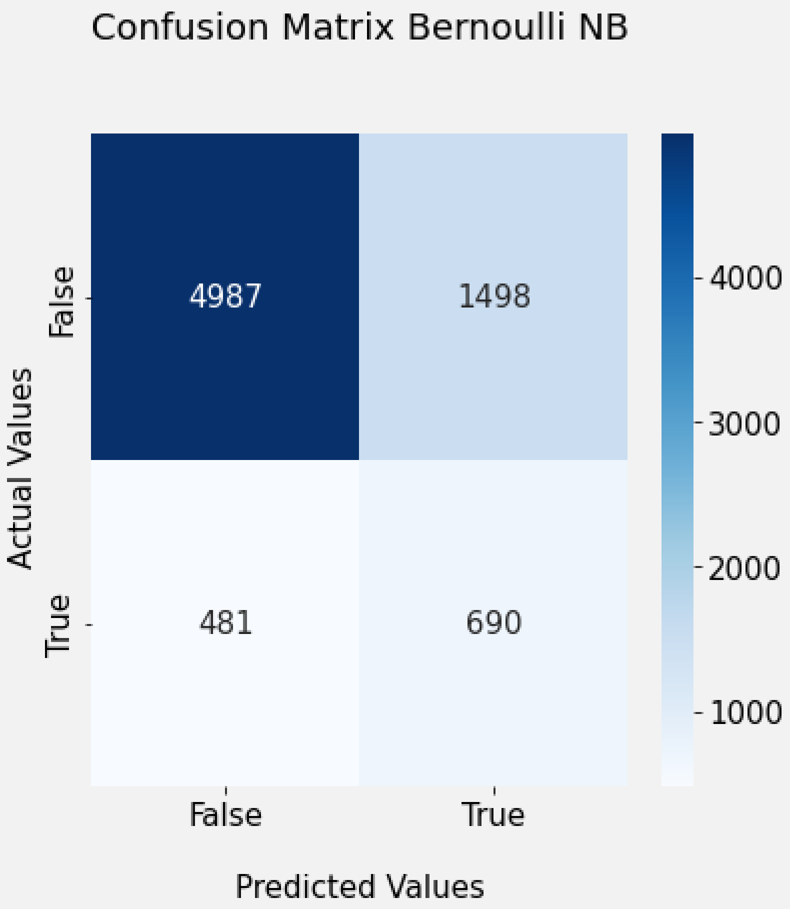


```
              precision    recall  f1-score   support

           0       0.91      0.77      0.83      6485
           1       0.32      0.59      0.41      1171

    accuracy                           0.74      7656
   macro avg       0.61      0.68      0.62      7656
weighted avg       0.82      0.74      0.77      7656
```

Confusion Matrix



Conclusions and Future Directions

1

Pedestrians, cyclists and motorcycle riders are most at risk

Governments should prioritize safety of these road users.

2

Drivers are at lower risk than passengers

With more vulnerable casualties at either end of the age spectrum

3

Several contributing factors to road casualties

Poor road markings and extreme weather conditions to name a few

4

Dataset could be expanded to improve model accuracy

Crucial factors like speed, drink driving or distracted driving were not included, and could improve model performance

5

Feature engineering could be employed

New features such as proximity or driver to home location could offer useful insight

6

Extend the time period of analysis

The dataset goes back over 40 years. Expanding the training set beyond 2021 could offer better performance but could impact computation time.

Thank you

The End