UK Road Safety Data Analysis:
Understanding Accidents & Predicting Casualties

Kyle Hincz

27 November 2022

***Abstract*—Governments aim to prevent road accidents and resulting casualties, especially those ending in death or serious injury. This paper tests various classification algorithms on the 2021 UK Road Accidents data to understand causes and predict whether an accident results in a fatality or injury. The method achieving best predictive power is Binomial Naïve Bayes and the author attempts to enhance the explanatory power of this model. Results indicate that the UK government should focus on improving safety of pedestrians, cyclists, and motorcycle riders as those are the road users most at risk of death or serious injury.**

1 INTRODUCTION

Road accidents are a traumatic experience for everyone involved and frequently end in death or serious injury (Gądek-Hawlena, 2020). As well as being a public health issue, traffic accidents carry a great economic cost (Moyer et al., 2017). One of the priorities of the UK government and the police force is to actively monitor and prevent traffic accidents (Department for Transport, 2019). 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) (Department for Transport, 2022b).

This paper analyses all vehicle accidents as reported by the UK Department for Transport in 2021. The aim of this research is to understand the nature of these accidents and their underlying causes. The paper applies machine learning and statistical methods to extract useful insights and predict deaths and serious casualties of these incidents.

2 PROBLEM STATEMENT & DATASET

The dataset has been obtained from the UK government website (Department for Transport, 2022b). 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

Table 1: Summary of data tables

For this analysis, all three tables were joined so that each casualty record in the final data frame has corresponding vehicle and accident variables.

The problem statement is as follows: can a fatality or serious injury be predicted based on the description of the accident and other factors present in the dataset? If the fatality can be predicted with good confidence, there is a possibility of preventing the loss of human life. Some research in the field has been done with application of logistic regression to Italian accident data (Eboli et al., 2020) but comparing results across countries remains difficult due to different reporting styles of each country's police force.

The main challenge of the UK dataset will be its class imbalance. In 2021 there were ca 1,500 casualties out of 101k, equating to a fatality rate of just under 1,5%.

Figure 1 presents key variables in the dataset visually. We can see that most casualties were male (62%) with slight injuries (79%). Most casualties happened in an urban environment (64%). Proportionally more men were injured as drivers (60k injured male drivers vs 26k injured female drivers) but more women were injured as passengers (10k injured male passengers vs 14k injured female passengers). This could partially be explained by the fact that only 35% of registered cars in the UK were registered to a female keeper (Driver and Vehicle Licensing Agency, 2022).

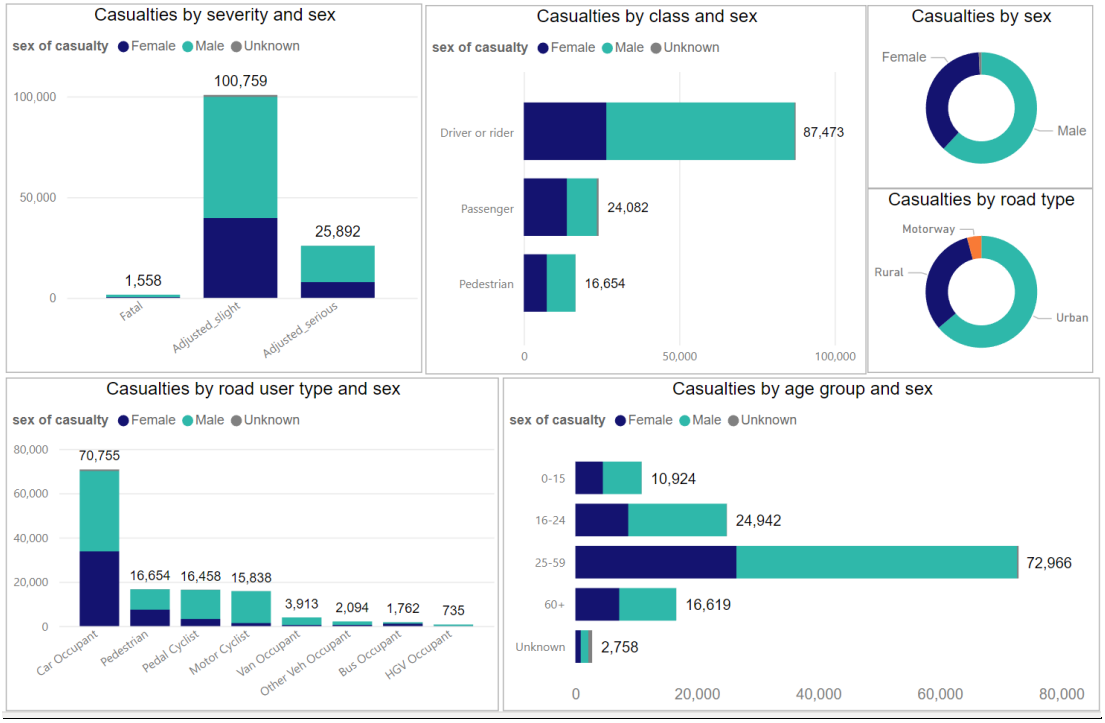

Figure 1: Breakdown of UK 2021 road accident casualties by severity, sex, class, age, road user type and road type (prepared by Department for Transport, 2022a)

The dataset has already been cleaned and is in a format that is conducive to analysis. Most categorical variables have been encoded into integers. For instance, the y variable has been encoded into 3 categories: 1: fatal, 2: serious, 3: slight. Not all variables are populated for every accident – the dataset contains field value of -1 where the data is missing.

One other challenge of the dataset is that not all parameters of the accident have been made public in order to protect the accident participants. Things like alcohol level of the driver have been removed and they could possibly have high predictive power.

3 PROPOSED METHODOLOGY

3.1 Data Preprocessing

As mentioned, the source dataset was clean, with any missing data already encoded as "-1" so no data cleaning was required. Despite this, the treatment of the missing records had to be considered as missing records constitute a significant proportion of all records and their inclusion or exclusion could impact model performance. For the distribution of records with missing data in the dataset, see Figure 2 below.

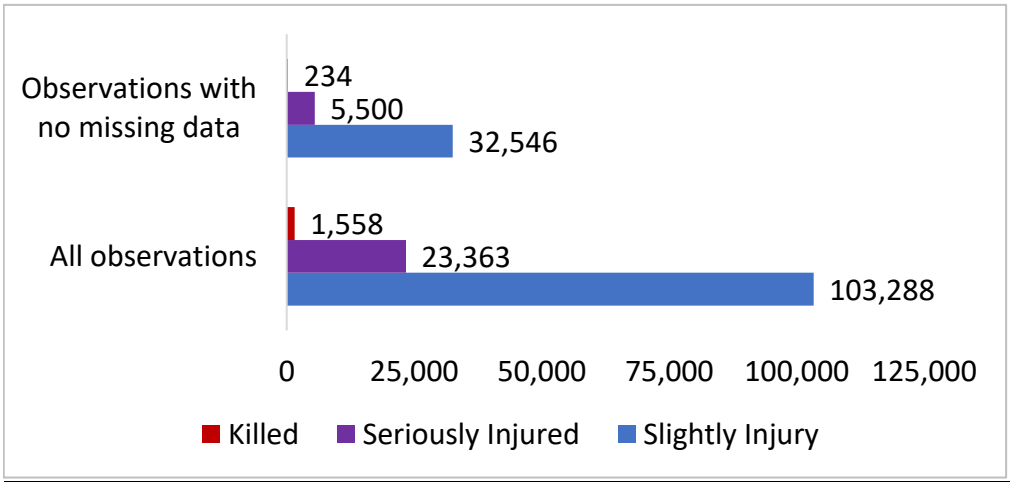

Figure 2: Distribution of complete observations and observations with missing data by casualty severity.

An initial comparison of balanced accuracy of various models with and without incomplete records was performed. Removing incomplete records from training and testing data resulted in better model performance so these observations were removed before the final analysis

3.2 Variable encoding

Majority of variables in the dataset were 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.

3.3 Response variable and evaluation metric choice

Two potential response variables were considered for this dataset: Only killed casualties and killed or seriously injured casualties (KSI).

Multiple scoring metrics have been evaluated (balanced accuracy, precision, recall, F1 score). It transpired that the dataset does not have one scoring metric that accurately captures all its features. Additionally, 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. Also, casualties' death is heavily influenced by the level of medical response received after the accident. Lives are frequently saved by fast and appropriate medical response. The dataset does not contain any medical response variables, therefore KSI is considered to be a more appropriate response variable. The priority of our model would be to predict KSI accurately but as the imbalance of "Not Serious" injury to KSI is approximately 6:1, balanced accuracy will be used.

If standard accuracy was used, the model could predict all records to be not KSI, resulting in high accuracy. Balanced accuracy avoids this as it is the arithmetic mean of specificity and sensitivity and is more robust than standard accuracy. Balanced accuracy ensures predictive power of the model is spread between correctly predicting negative and positive outcomes.

3.4 Data split and cross validation

The entire dataset was randomly split 80:20 into training and testing. Cross validation was used on the dataset. Cross validation is the process of spitting the dataset into chunks and training and testing the models on these chunks to ensure greater robustness of the result. For deriving the initial testing errors, 5-fold cross validation was used. For fine tuning the parameters I used the experimental halving cross validation algorithm in scikit-learn as the ordinary cross validation was very time consuming. I hyper tuned using the cross-validation grid and 10 folds for a more convincing result.

3.5 Feature Selection

Because of the large number of variables, feature selection was performed. Using fewer features in a model can sometimes increase its accuracy and reduce variance. 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)	Casualty (33/100)	Accident (9/100)
▪ 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	▪ Truck occupant ▪ Motorcycle driver/passenger ▪ Bus/van occupant ▪ Over 75 or under 10 yo ▪ Pedestrian ▪ Drivers significantly less likely to die than passengers	▪ Extreme weather conditions (flood, snow, winds) ▪ Road conditions (oil spillage, defective road markings, signs or surface) ▪ Relatively low importance of rain or day/night

Table 2: Summary of 100 most important variables identified using ridge regression.

3.6 Machine learning techniques used

The problem has been reduced to a supervised classification problem. Using python, several ensemble methods were used and compared to baseline methods.

3.6.1 *Parallel Ensemble methods*

Random Forest: combines and aggregates many decision trees that each vote for the class based on the underlying splits in the data.

Bagging: or bootstrap aggregating; uses random sampling with replacement from training data. It generates multiple weak learners from this data and aggregates them improving the prediction by reducing variance and overfit. Computationally expensive.

3.6.2 *Sequential Ensemble methods:*

AdaBoost or Adaptive Boosting: uses one level decision trees as weak learners putting weight on difficult to classify points.

Gradient Boosting: sequentially adds predictors that improve on ones already included in the model. Gradient descent is used to select and correct the predictors, improving accuracy.

XGBoost or Extreme Gradient Boosting: uses weights assigned to variables to construct decision trees. If a variable is predicted wrong, the weight of it is increased and sent to a second decision tree.

3.6.3 *Baseline methods:*

- Logistic regression,
- KNN – K Nearest Neighbor,
- Linear Discriminant Analysis
- Quadratic Discriminant Analysis
- Gaussian Naïve Bayes Classifier
- Bernoulli Naïve Bayes Classifier
- SVM Classifier

4 ANALYSIS AND RESULTS

Overall, 12 different models were tested ad compared. Models performed best on full dataset with missing values removed. Bernoulli Naïve Bayes had best balanced accuracy (0.68) 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 Figure 3 presents the results with averaged cross validated balanced accuracy.

Top three performing models: Bernoulli NB, LDA and Gaussian NB were hyper tuned. For final hyper tuned parameters see Appendix 1. For full results and cross validated accuracy scores on the testing data see Appendix 2.

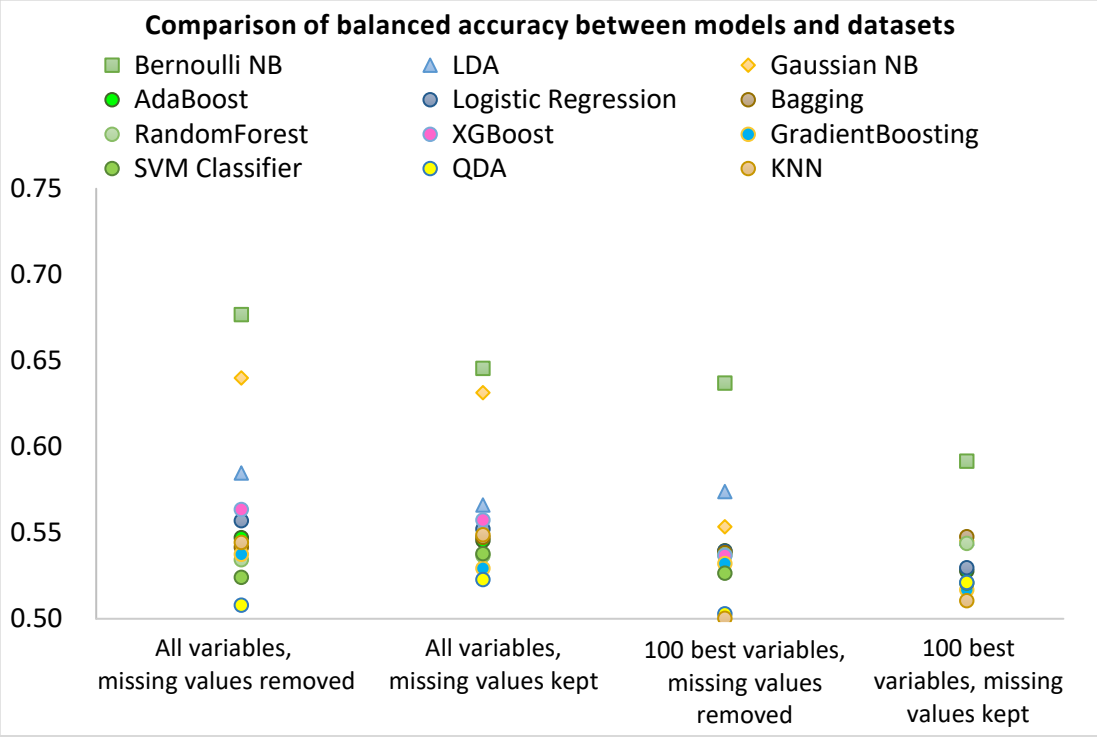

Figure 3: Comparison of balanced accuracy between models and datasets

4.1 Final model: Explanatory power

Naïve Bayes determines probabilities of each class associated with the features and predicts class with the highest probability. Unfortunately, this model does not offer a standard way of evaluating feature importances. Therefore, despite highest predictive power, the model offers little in terms of explainability of results.

In order to understand the output of the model permutation importance has been performed on all features of the dataset. This technique measures the decrease in feature importance if a feature is randomly shuffled. Top 10 features are as follows:

- 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)

In summary, 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.

4.2 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.

Confusion matrix of the model in Figure below indicates that the models correctly predicted 4,987 non-KSI accidents and 690 KSI accidents. It incorrectly classified 481 KSI accidents as non-KSI and 1,498 non-KSI accidents as KSI.

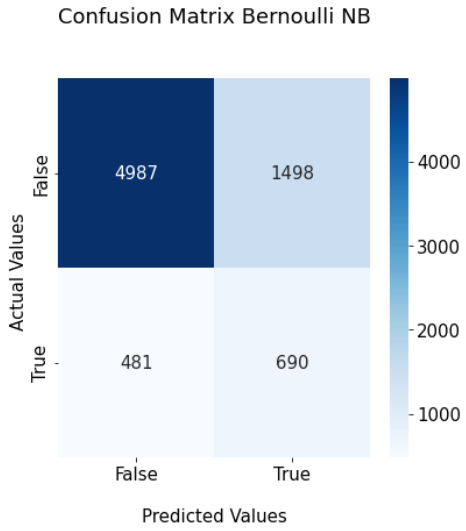

Figure 4: Final model confusion matrix

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

5 CONCLUSIONS

Applying the outputs of the model in the context of the problem, several important conclusions can be made.

- Pedestrians, cyclists, and motorcycle riders are most at risk. Governments should prioritize safety of these road users.
- Drivers are at lower risk than passengers with more vulnerable casualties at either end of the age spectrum (0-10 and 60+ years old)
- There are several contributing factors to road casualties. Poor road markings and extreme weather conditions to name a few.
- It is difficult to fully separate all factors as there are multiple variables in each incident. Each accident is a unique combination of thousands of variables that contribute to its outcome.

5.1 Future work

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.

Feature engineering could be employed. New features such as proximity or driver to home location could offer useful insight.

It would be possible to extend the time period of analysis. The dataset goes back over 40 years. Expanding the training set beyond 2021 could offer better performance but would impact computation time.

5.2 Lessons learned

Selecting the evaluation metric was harder than expected. No one metric was perfect to evaluate the models and I spent considerable time researching the pros and cons of each method. The imbalance of the dataset was responsible for a significant proportion of this difficulty, but the dataset also had several important variables missing or obscured.

Most significant variables obtained through ridge regression did not match the most significant variables identified through permutation importance of the Naïve Bayes model. While there was some overlap, I expected more alignment. This highlights

the fact that different models look at different features and it may sometimes be preferred to use a more explainable model in the analysis. Interpreting the results of Naïve Bayes was challenging.

Hyperparameter tuning took up a lot of time and it was impossible to run the full hyperparameter grid for models with many input parameters. Understanding the mathematical foundations of the model being used and its hyperparameters is crucial to a successful data mining project. One can mindlessly optimize but, without knowing the underlying formulas, it's very hard to interpret the final parameters.

6 APPENDIX

6.1 Appendix 1 – cross validation final model parameters

('Gauss Naive Bayes', GaussianNB(), {'var_smoothing': 4.32876128108305 26e-08})

('Bernoulli Naive Bayes', BernoulliNB(), {'alpha': 1.2447871461879063})

('LDA', LinearDiscriminantAnalysis(), {'shrinkage': 0.0, 'solver': 'lsqr'}

Resulting improvements in balanced accuracy were marginal.

6.2 Appendix 2 – Final results with their balanced accuracy score

	All variables, no missing	**All variables, including missing data**	**100 best variables, no missing**	**100 best variables, including missing data**
Bernoulli NB (tuned)	**0.68**	**0.65**	**0.64**	**0.59**
Gaussian NB (tuned)	0.64	0.63	0.55	0.55
LDA (tuned)	0.58	0.57	0.57	0.55
AdaBoost	0.55	0.55	0.54	0.53
Logistic Regression	0.56	0.55	0.54	0.53
Bagging	0.54	0.55	0.54	0.55
RandomForest	0.53	0.54	0.54	0.54
XGBoost	0.56	0.56	0.54	0.55
GradientBoosting	0.54	0.53	0.53	0.52
SVM Classifier	0.52	0.54	0.53	0.52
QDA	0.51	0.52	0.50	0.52
KNN (k=3, CV)	0.54	0.55	0.50	0.51

7 REFERENCES

Department for Transport. (2019). *Road safety statement 2019: a lifetime of road safety*. GOV.UK. https://www.gov.uk/government/publications/road-safety-statement-2019-a-lifetime-of-road-safety

Department for Transport. (2022a). *Reported road casualty statistics in Great Britain: interactive dashboard, from 2017*. Maps.dft.gov.uk. https://maps.dft.gov.uk/road-casualties/index.html

Department for Transport. (2022b, February 24). *Road Safety Data*. Www.data.gov.uk. https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

Driver and Vehicle Licensing Agency. (2022). *Vehicle licensing statistics: 2021*. GOV.UK. https://www.gov.uk/government/statistics/vehicle-licensing-statistics-2021/vehicle-licensing-statistics-2021

Eboli, L., Forciniti, C., & Mazzulla, G. (2020). Factors influencing accident severity: an analysis by road accident type. *Transportation Research Procedia*, *47*, 449–456. https://doi.org/10.1016/j.trpro.2020.03.120

Gądek-Hawlena, T. (2020). Economic Consequences of Road Accidents for TSL Companies. *Communications in Computer and Information Science*, 343–353. https://doi.org/10.1007/978-3-030-59270-7_26

Moyer, J. D., Eshbaugh, M., & Rettig, J. (2017). Cost analysis of global road traffic death prevention: Forecasts to 2050. *Development Policy Review*, *35*(6), 745–757. https://doi.org/10.1111/dpr.12263