Analysing Road Accidents Severity and Making Predictions Using Machine Learning Techniques

Introduction

Thousands of people are killed or injured on our roads daily. Despite the various developments in road safety and stricter regulations, predicting the severity of accidents remain an important challenge to reduce fatalities. Around 1.2 million lives are lost worldwide by road traffic accidents annually and millions more face life-altering injuries [1]. Individuals of all ages, commuting to work or school, or embarking on a long journey can fall victim of road fatalities, very often leaving behind families and communities devastated. According to the World Health Organization (WHO), injuries from road accidents are the leading cause of death for children and young adults between the age of 5 to 29 [2], showing the urgent need for effective interventions. Previous studies have shown that road category and the number of vehicles in an accident determine severity. This study focuses on leveraging machine learning techniques to predict severity and identify the most significant factors, using road accidents statistics from the UK.

The goal is to enhance road safety strategies and reduce the burden of traffic accidents by performing data preprocessing, exploratory data analysis, feature engineering and implementation of machine learning algorithms. The results are intended to highlight the most significant factors affecting severity of road accidents and identify potential data-driven models to improve road safety.

Analytical Questions and Data

Understanding the circumstances under which accidents happen, and the locations are very powerful information that can be used to take action to avoid them. The research aim is to present a comprehensive analysis and a powerful predictive model that can help reveal key factors affecting accident and how these results can be further used to define dynamic hotspots. We expect to answer the following research questions:

- How do environmental (e.g. road types, road conditions) and temporal factors (e.g., time of day, weather, and seasonality) influence the likelihood of severe accidents?
- What are the hotspots zones (accident blackspots) for the different accidents' severity?
- Can we accurately predict the severity of road accidents using machine learning models by implementing environmental, temporal, and road-related features?
- What are the most significant factors leading to these accidents?

The dataset used in this study is from Kaggle [3]. It covers a wide date range of events from 2021 to 2022. The dataset primarily focuses on UK Road Accidents which are collected and published by the Department of Transport. There are 307,955 instances and 21 features available for the analysis. The variables included in the dataset are time of accident, number of vehicles, severity, location, road conditions, weather conditions and many more. Given the features in the dataset, regression, classification or clustering can be applied depending on what is to be identified.

Data Processing

Preparing the dataset for analyses and modelling is an important step to get accurate information. The raw data needs to be clean as the results from the analysis depends on data consistency, accuracy and missing data points should be resolved.

The first step is to identify missing values and duplications. Missing values in the variable time have been removed. From the variable date; year and month have been extracted. Missing values in other variables have been replaced by using data available in other variables; for instance, datapoints missing the road type have been replaced according to speed limit. Some grammatical mistakes have also been amended to ensure consistency. Dimensionality has been reduced by grouping datapoint into categories for some variables. Outliers have been adjusted by doing further research. Interquartile range and transformations were considered but may not reflect the reality.

In the dataset, three types of accident severity are observed:

1. Slight: A road crash where medical attention was not necessary.
2. Serious: A road crash where medical attention was needed.
3. Fatal: A road crash that resulted in death.

Feature engineering is also an important step to any analysis to add further insights. A variable considering UK rush hour between 7am to 10am and 4pm to 7pm. We have also created a new feature for visibility condition where the value one would indicate that there was light and zero suggest darkness. The types of vehicles have been grouped into seven categories. We have not dropped features that we deem not useful for the analysis or modelling. After these changes, we have 307,619 rows and 24 variables which can be seen in Table 1.

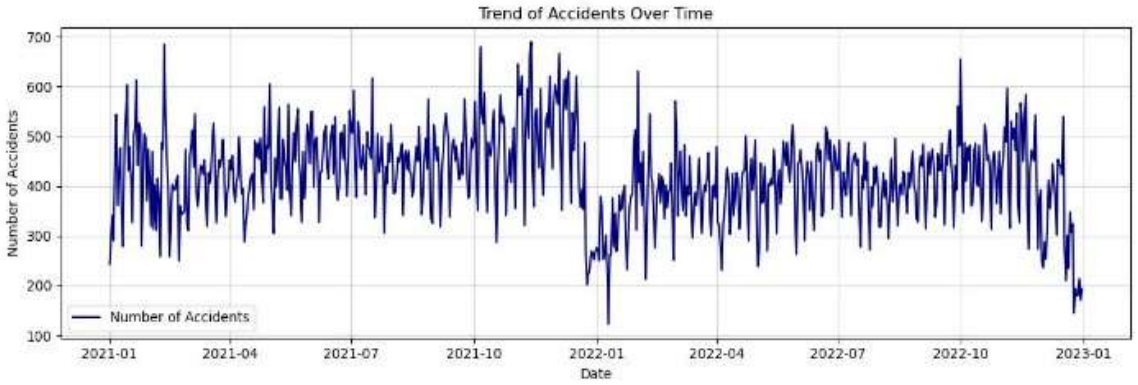
Table 1: List of variables

Accident Details	Location and Environment	Road and Traffic Details	Vehicles and Casualties	Others
Accident Index	Latitude	Junction control	Number of vehicles	Police Force
Accident Date	Longitude	Junction detail	Number of casualties	
Day of the Week	Urban or Rural Area	Road type	Vehicle Type	
Accident Severity	Local Authority (District)	Carriageway Hazards	Vehicle	
Time	Weather conditions	Speed Limit		
Month	Road Surface conditions			
Year	Light Conditions			
Hour	Visibility			
Rush hour				

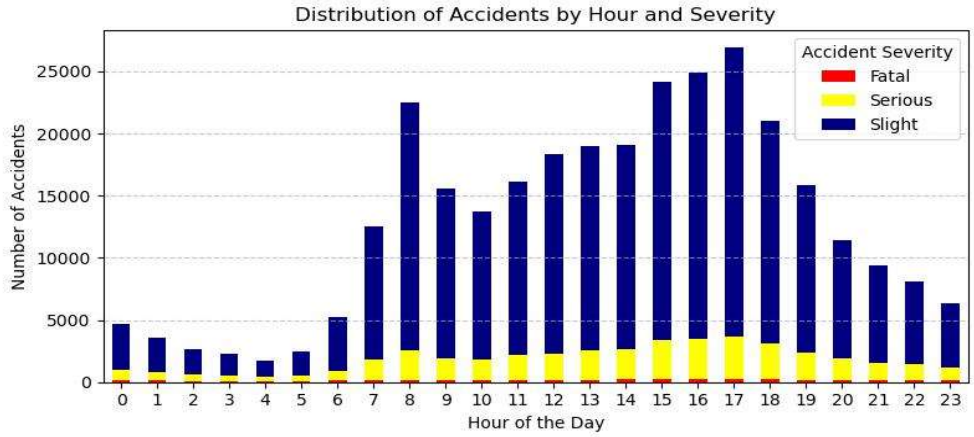
Road Accident Data Analysis and Statistics

Table 2: Descriptive Statistics

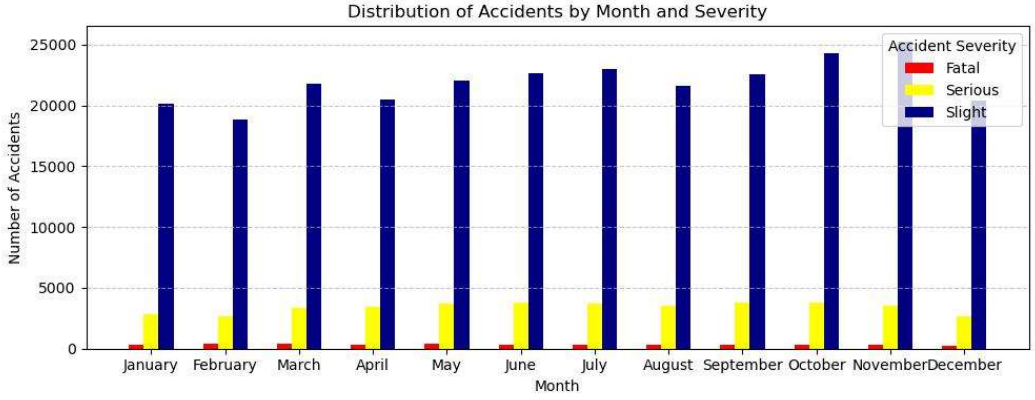
Variables	Values	Count	Slight (%)	Serious (%)	Fatal (%)
Year	2021	163370	85.09	13.44	1.47
	2022	144249	85.92	13.01	1.07
Month	1 (January)	23352	86.39	12.30	1.31
	2 (February)	21853	86.01	12.26	1.72
	3 (March)	25501	85.46	13.09	1.45
	4 (April)	24199	84.46	14.25	1.29
	5 (May)	26153	84.27	14.20	1.54
	6 (June)	26734	84.81	14.04	1.16
	7 (July)	26929	85.23	13.61	1.16
	8 (August)	25477	84.88	13.71	1.41
	9 (September)	26721	84.51	14.27	1.22
	10 (October)	28334	85.61	13.24	1.14
	11 (November)	29058	86.86	12.10	1.02
	12 (December)	23308	87.46	11.44	1.09
Day	Monday	43869	86.43	12.44	1.13
	Tuesday	46328	86.49	12.41	1.10
	Wednesday	46330	86.58	12.37	1.05
	Thursday	45595	85.97	12.92	1.10
	Friday	50464	86.07	12.70	1.22
	Saturday	41521	83.62	14.67	1.71
	Sunday	33512	82.06	16.06	1.88
Junction Control	Authorised person	460	90.43	9.13	0.43
	Auto traffic signal	32317	88.91	10.16	0.93
	Data missing or out of range	97901	82.73	15.32	1.95
	Give way or uncontrolled	149881	87.02	12.16	0.82
	Not at junction or within 20m	25364	82.31	15.73	1.96
	Stop sign	1683	89.01	10.34	0.65
Road Type	Dual carriageway	45450	86.47	11.72	1.81
	One way street	6185	87.16	12.01	0.82
	Roundabout	20895	91.10	8.53	0.37
	Single carriageway	231844	84.66	14.05	1.29
	Slip road	3232	91.06	8.51	0.43
Urban or Rural	Urban	109260	82.10	15.75	2.14
	Rural	198346	87.34	11.85	0.81
Speed Limit	10	3	100	0	0
	15	2	100	0	0
	20	2899	86.51	13.00	0.48
	30	199846	87.19	12.04	0.77
	40	25625	85.72	12.85	1.43
	50	10188	89.47	14.36	2.16
	60	46754	78.35	18.85	2.80
	70	22289	85.63	12.12	2.24



(a)



(b)



(c)

Figure 1: Accidents severity by (a)Day (2021-2022) (b)Hour (c)Month

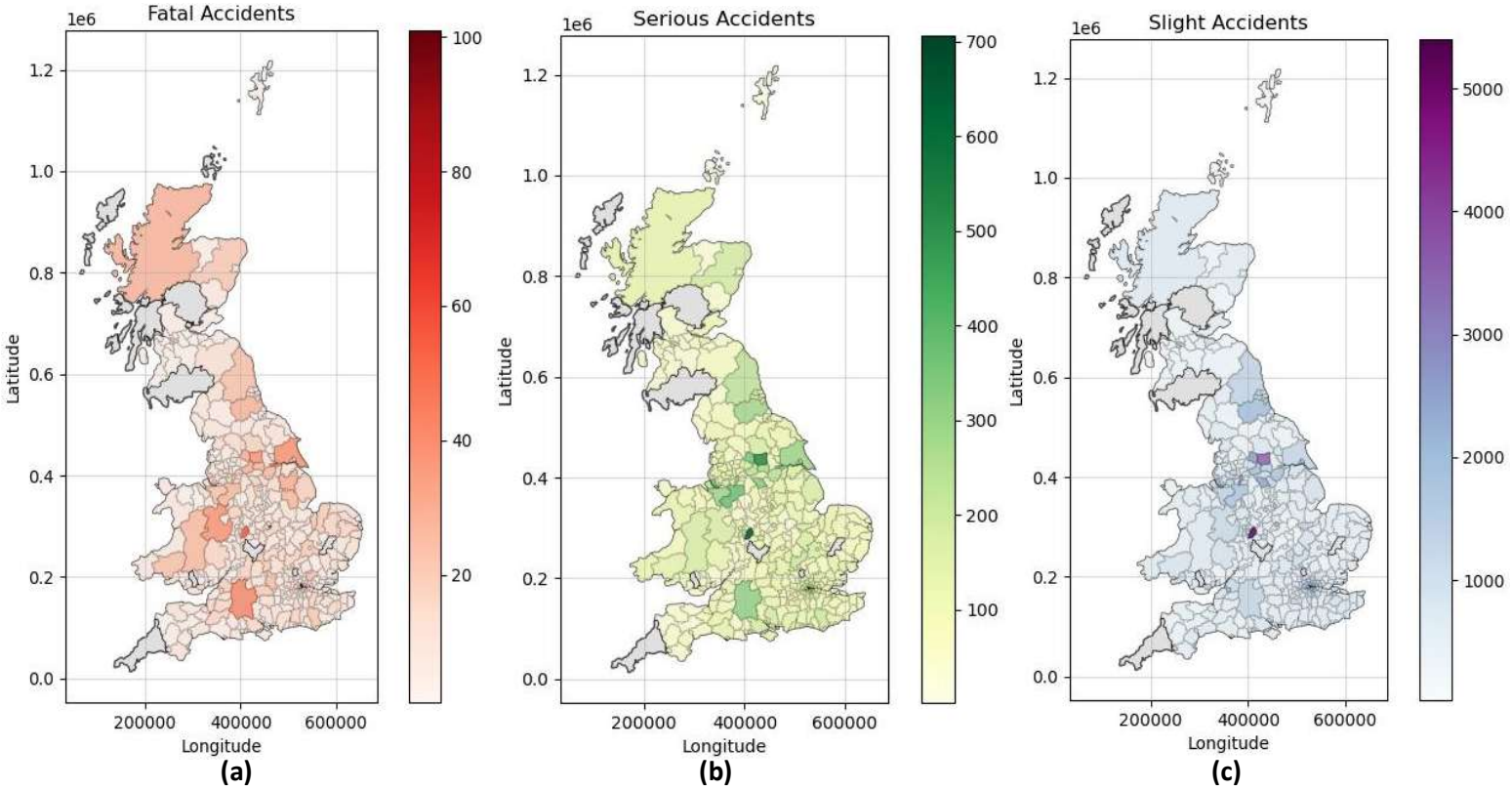


Figure 2: Spatial representation of Accidents severity by (a)Fatal (b)Serious (c)Slight

Descriptive Statistics

Table 4 provides the descriptive statistics of the variables in terms of counts and percentages for dependent variables. More accidents happened in 2021. November experienced the highest count. The percentage of fatal accidents is higher during winter months. Accidents are more frequent on weekdays. A high percentage of slight accidents occur at uncontrolled junctions. Fatal accidents are highest on dual carriageway. Lastly, higher-speed roads (70 mph) have a higher proportion of fatal accidents (2.24%).

Temporal Analysis

From Figure 1(a), a slight decrease can be seen in accidents in 2022. More accidents can be seen during winter/holiday periods. In diagram (b), slight accidents are the most frequent however severe and fatal accidents increase during early morning and late hours. Severe accidents are more likely to happen during peak hours. In (c), we can observe seasonality trends. Winter months can again be seen as high-risk season probably due to road and weather conditions. Fatal and serious accidents do not occur very often.

Spatial Analysis

The spatial comparison between the three severity levels can be observed in figure 2. Fatal accidents are concentrated along highways and major urban centres like London or Birmingham. Serious and slight accidents occur more broadly but clusters can still be seen in high density regions**.** Serious accidents are more common in Wales and North England and slight accidents occur more frequently across the UK.

Environmental Analysis

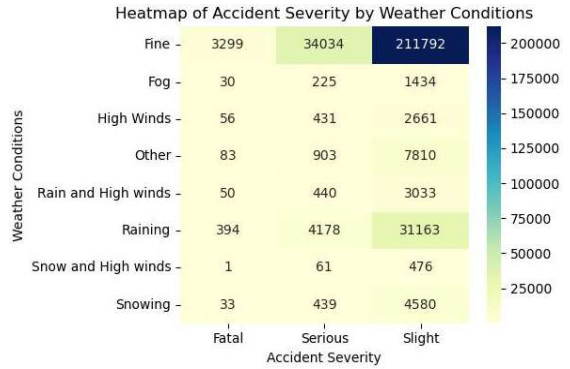


Figure 3: Heatmap of Accident Severity and Weather conditions

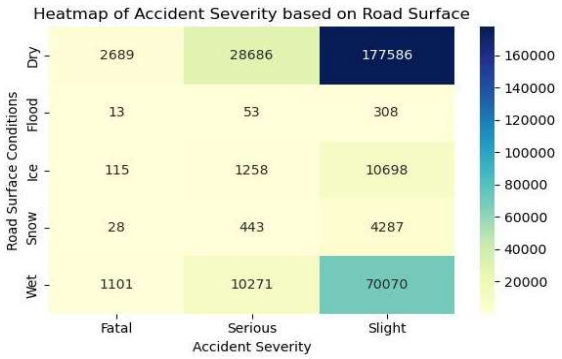


Figure 4: Heatmap of Accident Severity and Road Surface conditions

Figure 3 identifies weather conditions affecting number of accidents and severity. Fine weather conditions contribute to the highest number of accidents, most of which are of the lowest severity. Adverse weather conditions such as rain increase the risk of serious and fatal accidents. Figure 4 shows that most accidents occur on dry roads most probably due to less careful drivers. Wet and icy conditions increase the likelihood of serious and fatal accidents. Fine weather and dry road dominate number of accidents but with low severity.

Methodology

Road accident prediction is performed as a multiclass classification problem since accident severity is categorised into three levels: slight, serious and fatal. The following ML algorithms are considered:

1. *Multiclass Logistic Regression*

Logistic regression statistical technique used to predict binary outcome using one or more input variables [4]. In our case, since we are predicting three outcomes, it is a 'one vs all' logistic regression.

2. *Random Forest (RF)*

An ensemble technique where predictions from many decision trees are averaged to boost prediction stability and accuracy.

3. *Light Gradient Boosting (LGBM)*

LGBM is a decision-tree based algorithm used for classification and regression tasks. Compared to XGBoost and CatBoost, LGBM provides faster training and lower memory usage suitable for large datasets [5].

The machine learning algorithms have been selected based on previous similar studies. The features in Table 3used in the models have been selected based on domain knowledge as well as their performance in past articles. Road type was a significant predictor in Ali S. Al-Ghamdi's article about road accident prevention using logistic regression.

Table 3: Features

Features	Description
Number of casualties	Number of casualties in accidents.
Number of Vehicles	Number of vehicles involved in road accidents
Speed Limit (mph)	10, 15, 20, 30, 40, 50, 60, 70
Road Type	The different road types
Urban or Rural	Location of the accident
Rush Hour	Peak traffic hours
Visibility	Lighting condition
Road Conditions	Dry, Wet or Slippery
Weather	Good, Moderate or Adverse
Vehicle	The type of vehicle

Table 4: Performance metrics

Models	Imbalance Data				Balanced Data (After SMOTE)			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Logistic Regression	0.855	0.759	0.855	0.788	0.534	0.788	0.534	0.626
Random Forest	0.853	0.767	0.853	0.789	0.579	0.791	0.579	0.659
L-GBM	0.855	0.780	0.855	0.787	0.597	0.795	0.597	0.673
L-GBM (Grid Search)	0.854	0.774	0.854	0.789	0.591	0.793	0.591	0.667

80% of the dataset is used for training and 20% for testing. A stratified 10-fold cross validation is applied on the training for better generalization and the models are evaluated using the following metrics: Accuracy, Precision, Recall and F1-Score. Numerical variables are standardised, and categorical variables have been one-hot encoded. The target variable is label encoded. To mitigate overfitting, accuracy is compared between cross-validation and testing. The algorithms are run on the imbalanced data before applying SMOTE. Based on initial results, the best performing model is further optimised, for instance through grid search analysis.

Experimental Results

The result for prediction is presented in Table 4. The models performed well on the imbalance data and similar performance metrics are observed. Logistic regression and L-GBM performed best in terms of Accuracy and Recall while RF has a slightly higher precision (0.767). Cross-validation accuracy and test accuracy are similar suggesting no overfitting.

After SMOTE, L-GBM is the best performing model with accuracy and recall of 0.597 and F1-score of 0.673. RF follows closely with slightly lower metrics, but the results are better than logistic regression. Accuracy has been reduced for all models and slight overfitting is observed.

L-GBM is the best model overall. To further optimise this model, grid search is implemented and the optimised hyperparameters are seen below but these have not yielded better results.

Table 5: L-GBM Grid Search Parameters	
Learning Rate	0.1
Max Depth	7
Number of Estimators	500
Number of leaves	50

Findings and Reflections

We have been able to identify how environmental factors influence accident severity. Good weather conditions and dry surfaces lead to higher frequency of accidents but with less severe outcomes. Serious accidents are often under adverse weather and hazardous road surfaces. We have also identified that accident trends remain stable over time with minor fluctuations due to seasonal or behavioural patterns. Rush hours lead to increased non-serious accident occurrences. Monthly variations shows that summer holidays and winter influence accident frequency and severity.

It has also been observed that urbanized areas like central England and part of Scotland with high traffic density are hotspots for fatal accidents. Serious accidents are present across the UK. Slight accidents dominate in densely populated urban areas and are linked to vehicle congestion.

To predict accident severity, logistic regression, random forest and light gradient boosting have been the machine algorithms of choice. Logistic regression underperformed both before using SMOTE and after compared to the other models. In a study by the University of Cyprus in 2023, a 90% accuracy was observed however the target variable was binary which may be more suitable for predictions using logistic regression [6].

Random forest was expected to show some improvements when underrepresented classes (Serious and Fatal) were synthetically oversampled however the RF model being robust skewed distributions performed better on the imbalanced data. After SMOTE, an increase in precision is seen which suggests a balanced performance across all classes but a lower accuracy indicates the model inability to generalize effectively.

L-GBM is found to be the best model for accident severity prediction. Before SMOTE, an accuracy of 85.5% is observed proving its suitability to handle real-world data where severity is unevenly distributed. However, after SMOTE, accuracy drops reducing the model's ability to generalize to the test data. L-GBM maintains more competitive precision and F1-scores compared to logistic regression and random forest. L-GBM proves to be an excellent choice but the decision to apply SMOTE or not should be considered; figure 5 may help with this decision.

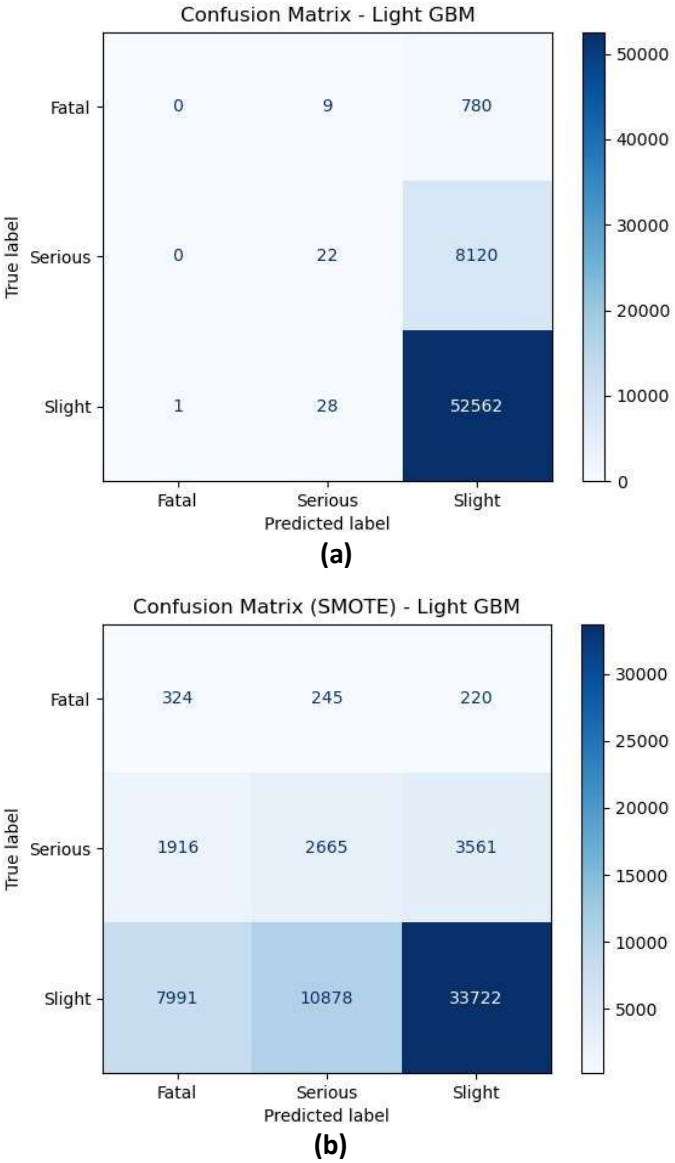


Figure 5: Confusion Matrix: (a) before (b) after SMOTE

In Figure 5(a), the model fails to predict fatal accidents; they are misclassified. For serious accidents, around 99% are misclassified. The model without SMOTE predicts slight accidents correctly by a high margin since the data is skewed towards this severity. With SMOTE, the model improves its predictive capabilities for minority classes: 324 fatal accidents and 2665 serious accidents are correctly predicted. The choice of implementing SMOTE would depend on whether minority classes predictions are prioritised, or accuracy is the main goal. SMOTE did not overwhelmingly improved the model as observed in previous studies.

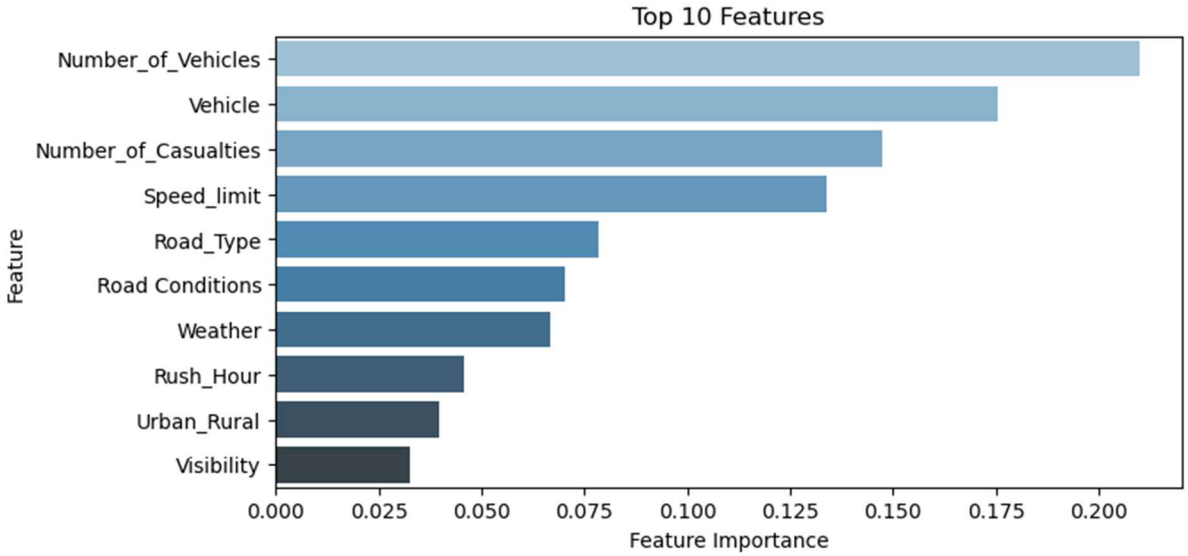


Figure 6: Feature Analysis

To identify the main factors for accident severity, feature importance through random forest node purity index has been used. SHAP has been considered since we can see the impact of the features for each class (in appendix).

From figure 6, the factors are arranged in descending order. The top 3 factors are number of vehicles, vehicle type and number of casualties. Road type which was an important determinant in previous study done in other countries, does not seem to be a very important factor here in the UK. Visibility has the least impact on the target variable.

Conclusion

Throughout the project, we have seen the performance of machine learning models in predicting accident severity. Random Forest and L-GBM performed well but interpretability from Logistic Regression may be more important in situations where policy decisions are important. Imbalanced dataset was a challenge as models tend to favour majority classes but techniques like SMOTE can help at the cost of lower accuracy and less robust models. Hence, accuracy as the only measure of performance initially masked the challenges of imbalance data; other metrics like F1-Score, AUC and confusion matrix should be considered. A final issue from the dataset was the broad categorization of multiple features which could have led to potential bias during encoding process.

Future work

Neural networks or ensemble methods and other features such as age and gender may help yield more accurate models. Hyperparameter tuning should be explored further. Use SHAP to identify relevant factors from models.

REFERENCES

[1] Kremer, W. (2024, May 19). *More than a million people die on roads every year. Meet the man determined to prevent them*. BBC Home - Breaking News, World News, US News, Sports, Business, Innovation, Climate, Culture, Travel, Video & Audio. https://www.bbc.com/future/article/20240517-vision-zero-how-europe-cut-the-number-of-people-dying-on-its-roads#:~:text=An%20estimated%201.2%20million%20lives,suffer%20often%20life-changing%20injuries.

[2] *Road safety*. (n.d.). World Health Organization (WHO). https://www.who.int/health-topics/road-safety#tab=tab_1

[3] Zrirak, M. (n.d.). *Project Excel Road Accident Dataset*. Kaggle. Retrieved [18.11.24], from

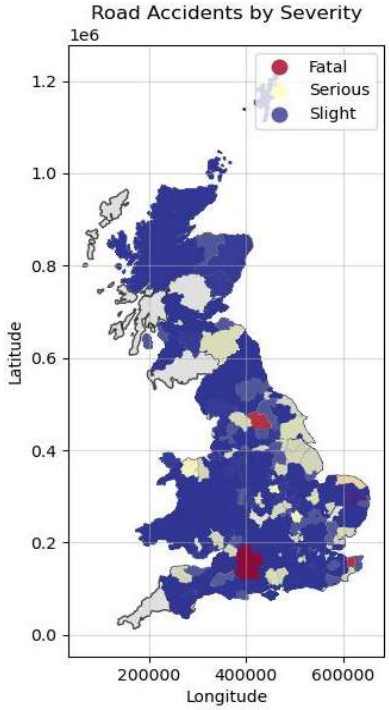
https://www.kaggle.com/datasets/mohamedzrirak/project-excel-road-accident

[4] Mendez, G., Buskirk, T. D., Lohr, S., & Haag, S. (2008). Factors Associated With Persistence in Science and Engineering Majors: An Exploratory Study Using Classification Trees and Random Forests. *Journal of Engineering Education*, *97*(1), 57–70. https://doi.org/10.1002/j.2168-9830.2008.tb00954.x

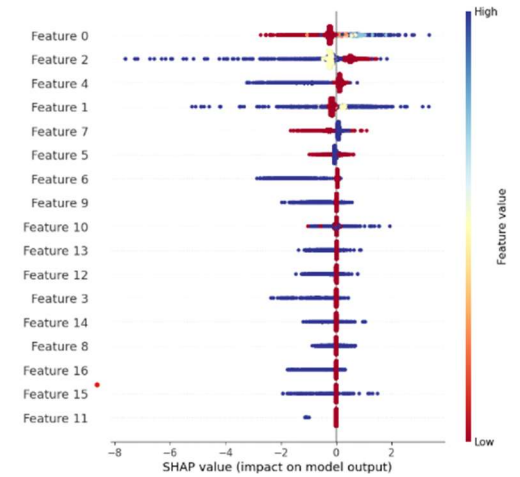
[5] Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2020). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-020-09896-5

[6] Obasi, I. C., & Benson, C. (2023). Evaluating the effectiveness of machine learning techniques in forecasting the severity of traffic accidents. *Heliyon*, *9*(8), Article e18812. https://doi.org/10.1016/j.heliyon.2023.e18812

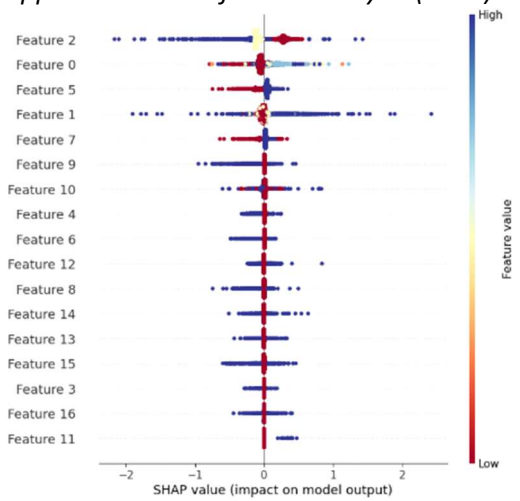
Section	Word Count
Introduction	213
Analytical question and data	225
Analysis	963
Findings, reflections and further work	677



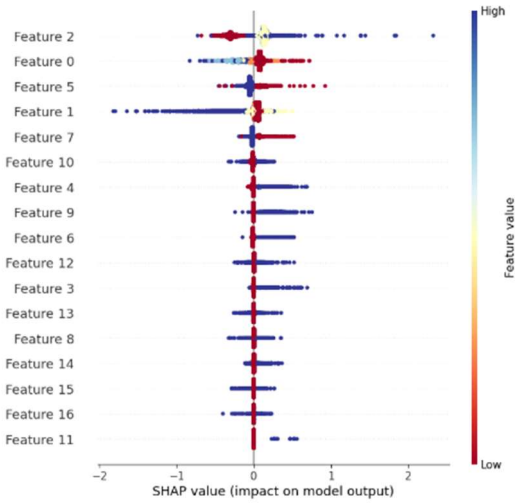
Appendix 1: Accident severity across UK



Appendix 2: SHAP feature analysis (Fatal)



Appendix 3: SHAP feature analysis (Serious)



Appendix 4: SHAP feature analysis (Slight)