Road Accident Severity project

Team members: Nadezhda Videneeva

GitHub link: https://github.com/VNVid/AccidentSeverityProject

1. Introduction

Road traffic accidents are a major public health concern globally, leading to significant loss of life, injury, and economic costs. Beyond the tragic human toll, these accidents impose a substantial financial burden on individuals, families, and nations. This makes the study of road accident severity not only a matter of public interest but also of urgent necessity.

The core scientific question driving this project is: How can we enhance road safety and possibly decrease the number of accidents? This question encompasses several key objectives:

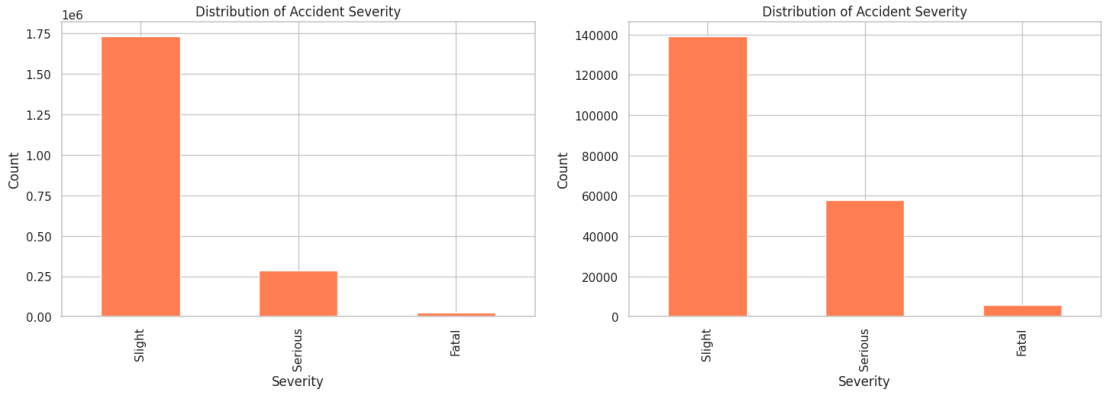
1. **Accident Severity Prediction**: Developing an ML model that can accurately predict the severity of an accident based on various factors such as road conditions, weather, traffic volume, and driver behaviour.
2. **Risk Factor Analysis**: Identifying the most significant factors contributing to high-severity accidents, which can inform policy decisions and targeted interventions.
3. **Preventive Strategy Formulation**: Utilising the insights gained from ML models and EDA to propose effective strategies for accident prevention and road safety enhancement.

2. Data collection and preprocessing

Data description. There are several publicly available datasets describing road accidents. For this project, I have utilised the datasets provided by the Department of Transport of the UK government. The UK government collects and publishes [2] on an annual basis detailed information about all registered traffic accidents across the country. I took the data from the years 2005-2017 from Kaggle [1]. The data does not contain any sensitive personal information about the drivers, ensuring privacy compliance. This consideration ensures that the study maintains ethical integrity.

The dataset includes two main files: one for accident information and another for vehicle information. The data encompasses various attributes such as vehicle manoeuvres, road type, weather conditions, age of vehicle, engine capacity, drivers' age, purpose of journey, etc, in total, 57 features. The target variable is accident severity categorised into 3 types: slight, serious, and fatal. In this project I will try to analyse how severity could be decreased.

Loading and subsampling data. Initially, the datasets were loaded to understand their structure and contents, and joined into one dataset on the accident index. Given the extensive size of the dataset (> 2 million samples), a subsampling strategy was employed to reduce their volume while maintaining a representative sample. This involved analysing the distribution of accident severity (Fig. 1, a) and implementing weighted sampling to retain a sufficient representation across all severity levels. Slight accidents were chosen with 0,25 probability, whereas serious and fatal with probability of 0,75. The final dataset includes 10% of the initial one, counting approx. 200.000 samples. As we can see from the accident severity distribution (Fig. 1, b), the data is still imbalanced.



(a) Accident severity distribution of initial dataset (b) Accident severity distribution of subsample

Fig. 1

Primary feature dropping. The next step was an examination of the features with further removal of several ones. Firstly, categorical features with a high number of unique values are likely to be uninformative. Therefore, 'Accident_Index', 'LSOA_of_Accident_Location', 'model' were dropped. Secondly, features with a large proportion of missing values were eliminated. My analysis revealed various terminologies for missing data, such as 'Data missing or out of range', 'Unknown', or 'Unallocated', which were standardised to numpy.NaN. Features with over 100,000 missing values (which is more than a half of samples), including Carriageway_Hazards, Special_Conditions_at_Site, Hit_Object_in_Carriageway, Hit_Object_off_Carriageway, Journey_Purpose_of_Driver, Skidding_and_Overturning, were deemed uninformative and removed. Lastly, features were scrutinised for their relevance and utility. Those specific to the UK context, like the type of police force which registered the accident or the district

where it happened, or not directly related to accident characteristics, such as latitude or longitude, were excluded, being not helpful in developing a preventive strategy. Here 'Unnamed: 0', 'Did_Police_Officer_Attend_Scene_of_Accident', 'Latitude', 'Local_Authority_(District)', 'Local_Authority_(Highway)', 'Location_Easting_OSGR', 'Location_Northing_OSGR', 'Longitude', 'Year_x', 'InScotland', 'Year_y', 'Date', 'Time', 'make', '2nd_Road_Class','1st_Road_Number', '2nd_Road_Number', 'Driver_IMD_Decile', '1st_Road_Class', 'Police_Force' were deleted. This step was crucial to streamline the dataset and focus on the most relevant variables.

Dealing with missing values. Remaining missing values were addressed by applying different imputation strategies for numerical and categorical data. Numerical columns with missing values were imputed with the median value, while the most frequent value was used for categorical columns. This approach ensured the retention of data integrity and robustness in subsequent analyses.

Categorical feature analysis. The distributions of categorical features were closely analysed. Categories with insufficient representation were either merged with similar groups or removed as outliers to reduce noise in the data. This step was instrumental in refining the dataset for more accurate model training and analysis and reducing dataset size after further one-hot encoding transformation. As a result, the following modifications were made (distributions on Fig.2-10):

- Samples where 'Junction_Control' feature had value 'Authorised person' were dropped.

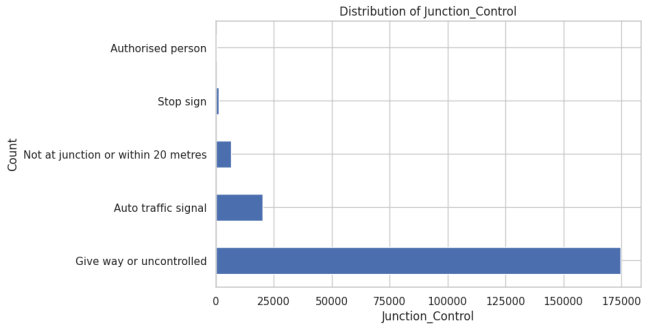


Fig. 2

- There are few mini-roundabouts (feature ‘Junction_Detail’), so they were replaced with roundabouts.

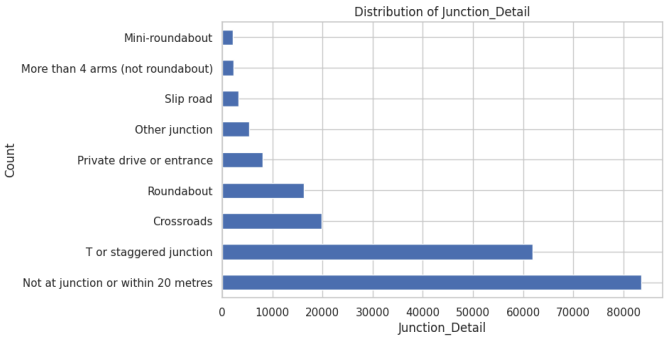


Fig. 3

- Lights unlit and no lightning are basically the same when analysing accidents, so these categories were merged together.

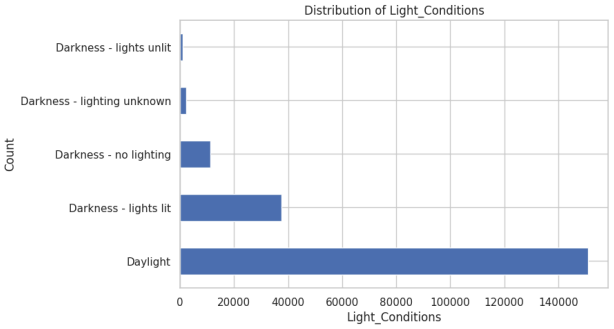


Fig. 4

- Concerning road surface conditions, 'Flood over 3cm. deep' value is an outlier, therefore, it was added to the group 'Wet or damp'. 'Snow' and 'Frost or ice' types were also merged together.

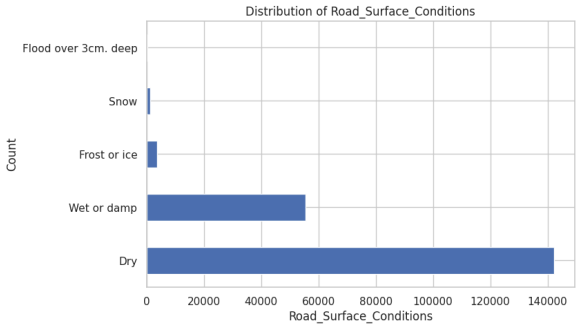


Fig. 5

- All age bands of drivers lower than 15 years were replaced by the new value '<=15'.

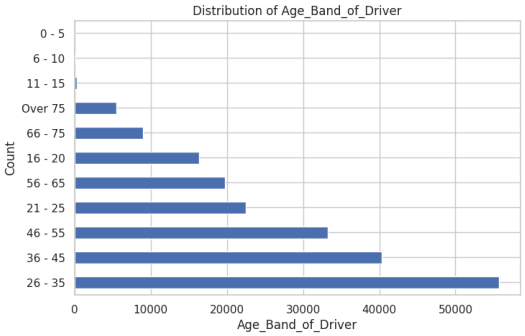


Fig. 6

- There are many poorly represented types of propulsion codes. All electric fuels were merged together. Samples with 'Gas/Bi-fuel', 'Petrol/Gas (LPG)', 'Steam', 'Gas', 'Gas Diesel' and 'New fuel technology' values were removed.

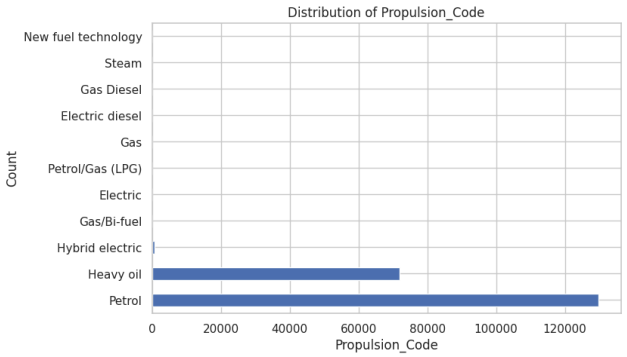


Fig. 7

- 'Single trailer', 'Other tow', 'Double or multiple trailer' and 'Caravan' values of 'Towing_and_Articulation' feature were replaced with new 'Trailer, caravan or other tow' category.

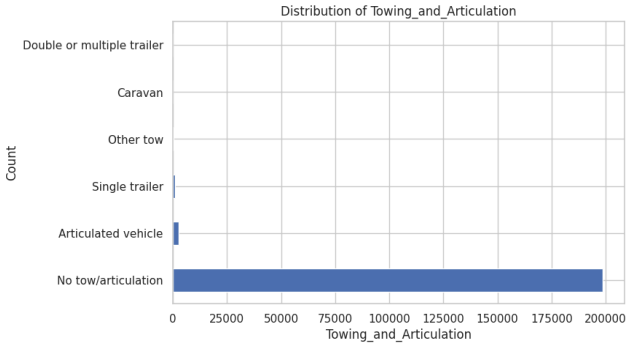


Fig. 8

- 'Offside - crossed central reservation' and 'Offside on to centrl res + rebounded' categories of 'Vehicle_Leaving_Carriageway' feature were joined to 'Offside and rebounded' type.

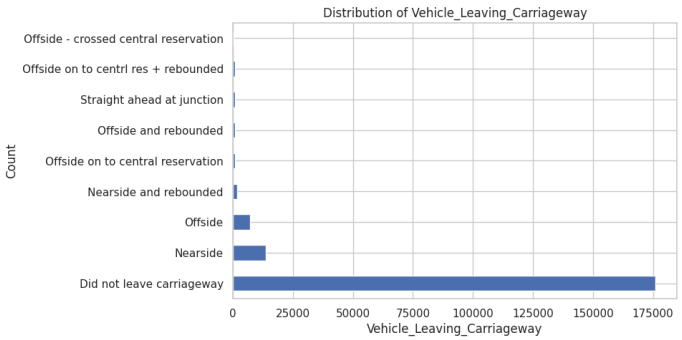


Fig. 9

- 'Was_Vehicle_Left_Hand_Drive' feature turned out to be uninformative, so it was deleted.

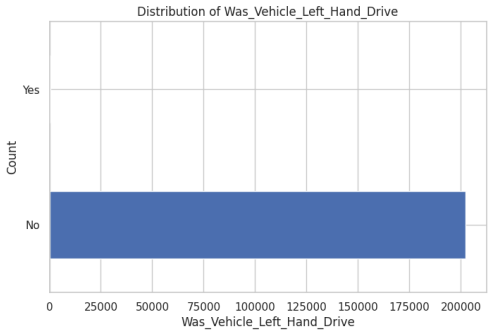


Fig. 10

Feature transformation. To prepare the categorical features for machine learning algorithms, one-hot encoding was employed, as the features are not ordinal.

Ethical considerations. Throughout the data collection and preprocessing stages, ethical considerations, particularly regarding data privacy, were observed. The chosen dataset does not contain sensitive personal information, aligning with ethical standards in data handling.

3. Models to predict accident severity

From a machine learning perspective, the task is to classify road accidents into three categories (slight, serious, fatal) based on various features. This multiclass classification problem requires models capable of handling imbalanced datasets and making predictions based on a mixture of categorical and numerical data. Besides, the models should be capable of measuring the relative importance of each feature on the prediction, as this is crucial for further risk factor analysis. Thus, Logistic Regression, Random Forest and Gradient Boosting on Decision Trees with CatBoost were used.

The dataset was divided into training, validation, and testing sets in a 70:15:15 ratio. This split was strategically chosen to ensure a substantial amount of data for training the models while providing adequate and separate datasets for validation and testing purposes.

Logistic Regression is an interpretable model and could be extended to multiclass classification with cross-entropy loss. However, it assumes linear relationships between a target variable and the log odds of the dependent variables, which may not be reflected in the dataset. To accommodate the model's sensitivity to feature scales, data was standardised using MinMaxScaler to the [0, 1] range. When training, L2-regularisation was employed. The hyperparameter tuning, executed via

GridSearchCV, was performed to find the best combination of the following parameters:

- class weights: ['balanced', None]. The “balanced” mode, which adjusts weights inversely proportional to class frequencies in the input data, was used due to imbalance of severity distribution.
- regularisation strength (C): [0.1, 1, 10]
- tolerance (tol): [1e-3, 1e-4, 1e-5]

To perform hyperparameter tuning training and validation splits were concatenated, as this step included 5-fold cross-validation within GridSearchCV, which ensured unbiased evaluation and selection of hyperparameters. The final model was trained with the identified optimal parameters on the training split.

As an ensemble method, Random Forest is robust against overfitting and effective in handling non-linear relationships. As with Logistic Regression, hyperparameter tuning of Random Forest Classifier was performed:

- class weigh': ['balanced', None],
- number of estimators: [1, 10, 100, 500],
- criterion: ['gini', 'entropy', 'log_loss'],
- Minimum number of samples in leaves: [1, 10, 100].

The final model was then trained with the identified optimal parameters on the training split.

CatBoost is renowned for delivering high accuracy and speed, particularly in datasets with complex feature interactions. Some studies demonstrate the effectiveness of CatBoost compared to other gradient boosting algorithms like XGBoost[3]. As CatBoost is known for efficient handling of categorical features, the model was trained on the training split of the processed dataset where categorical features were not yet transformed with one-hot encoding. It was trained using the GPU for faster computation. The model used a validation set while training to prevent overfitting and select the best iteration.

Justified by the need to equally represent all classes in the model's performance evaluation, particularly crucial in imbalanced datasets, average f1-score was chosen as the main optimality metric. The results of models’ performance on the test set could be found on Fig.11 below. CatBoost has achieved a slightly higher performance according to almost all mentioned metrics, so it will be used further for risk factor analysis.

	precision	recall	f1-score	support
Fatal	0.15	0.00	0.01	880
Serious	0.55	0.18	0.27	8764
Slight	0.72	0.95	0.82	20758
accuracy			0.70	30402
macro avg	0.47	0.38	0.36	30402
weighted avg	0.65	0.70	0.64	30402

(a) Logistic Regression

	precision	recall	f1-score	support
Fatal	0.57	0.01	0.03	880
Serious	0.54	0.23	0.32	8764
Slight	0.73	0.93	0.82	20758
accuracy			0.70	30402
macro avg	0.61	0.39	0.39	30402
weighted avg	0.67	0.70	0.65	30402

(b) Random Forest

	precision	recall	f1-score	support
Fatal	0.48	0.03	0.06	880
Serious	0.55	0.25	0.34	8764
Slight	0.73	0.93	0.82	20758
accuracy			0.71	30402
macro avg	0.59	0.40	0.41	30402
weighted avg	0.67	0.71	0.66	30402

(c) CatBoost

Fig. 11. Classification reports on test set.

4. Analysis and proposals

Using the CatBoost model, as the best of tested ones, the feature importances were calculated and visualised. The top 15 features, according to their importance, were plotted in a descending order (Fig.12). Let's take a closer look at some of them.

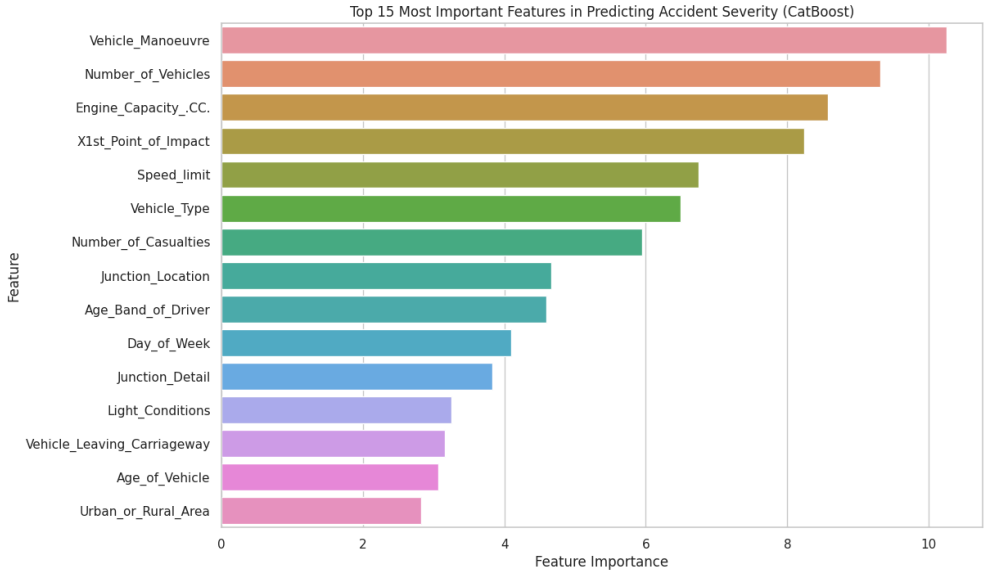


Fig. 12

The distribution analysis of vehicle manoeuvres showed a varied number of accidents associated with different manoeuvres. A bar graph (Fig.13) shows that going ahead another car is more frequently involved in accidents. A detailed crosstab (Fig.14) analysis provides a two-dimensional view of how different manoeuvres contribute to accident severities, visualised using heatmaps. These heatmaps are particularly useful in identifying specific manoeuvres that were more often associated with serious or fatal accidents. The most dangerous manoeuvres are different types of going ahead or overtaking other vehicles. Introducing restrictions on such manoeuvres on dangerous parts of roads may enhance road safety.

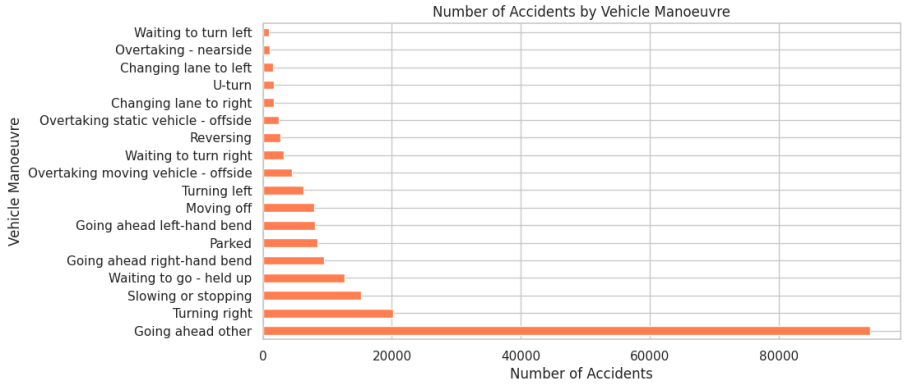


Fig. 13

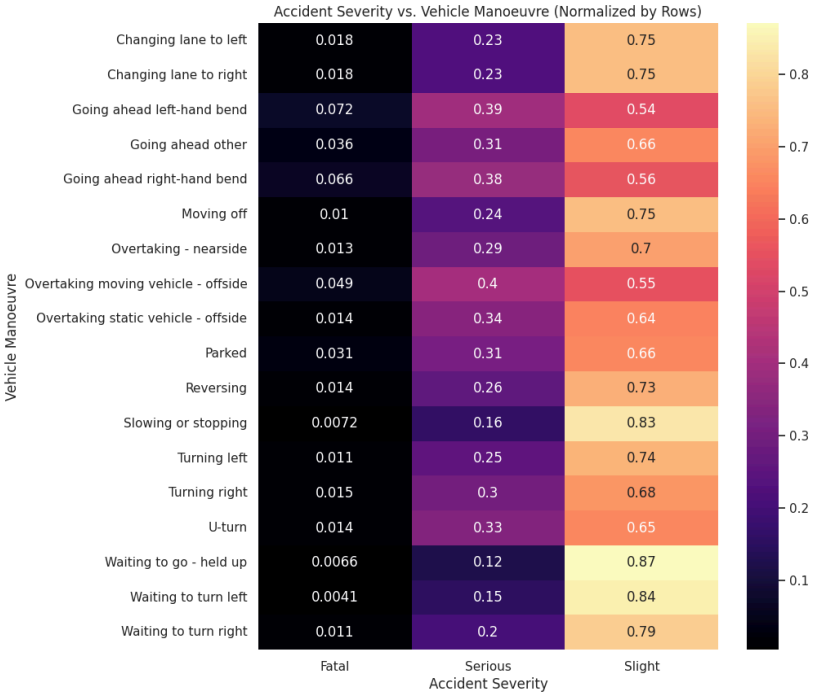


Fig. 14

The histogram on Fig. 15 now displays the number of accidents for each day of the week. Friday is the day with the highest frequency of accidents, which could be due to weekend journeys. Some traffic limitations could be introduced to decrease the number of accidents.

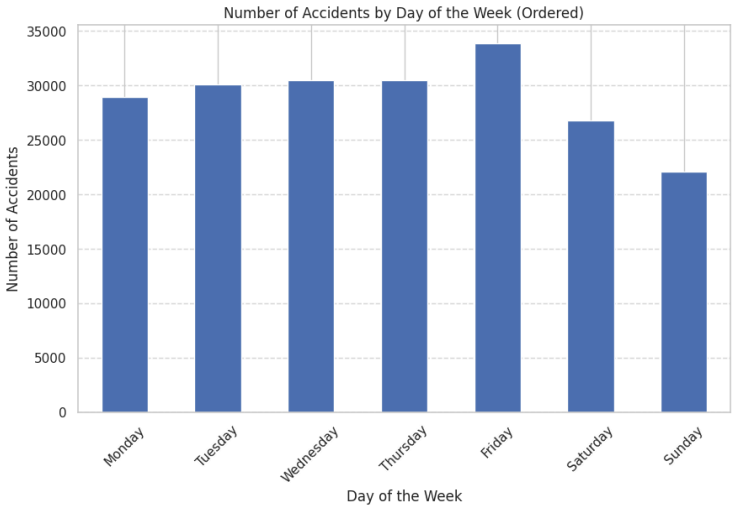


Fig. 15

The next histogram (Fig.16) illustrates the density of accidents by severity and age band of the driver. Each age band includes three bars, each representing a different level of accident severity: 'Slight' (green), 'Serious' (orange), and 'Fatal' (red). The height of each bar indicates the density of accidents of that severity within the respective age group. The plot reveals the fact that elderly people are more likely to get into a more serious accident. More strict health checks for drivers may be proposed.

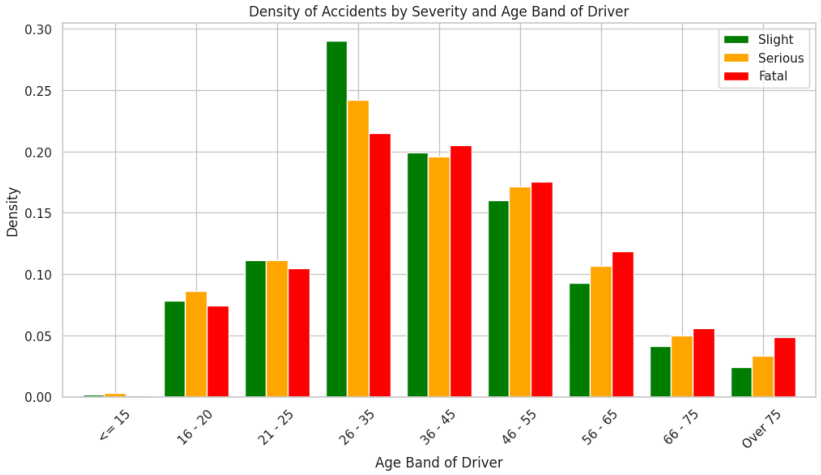


Fig. 16

The crosstab on Fig.17 provides a view of how different light conditions contribute to accident severities. Unlit areas result in a higher probability of serious or fatal accidents. An experiment was conducted by modifying the 'Light_Conditions' feature in the dataset, changing 'Darkness - no lighting' to 'Darkness - lights lit'. The results indicated a decrease in the percentage of fatal and serious accidents, suggesting that improving lighting conditions, especially in unlit areas, could significantly enhance road safety.

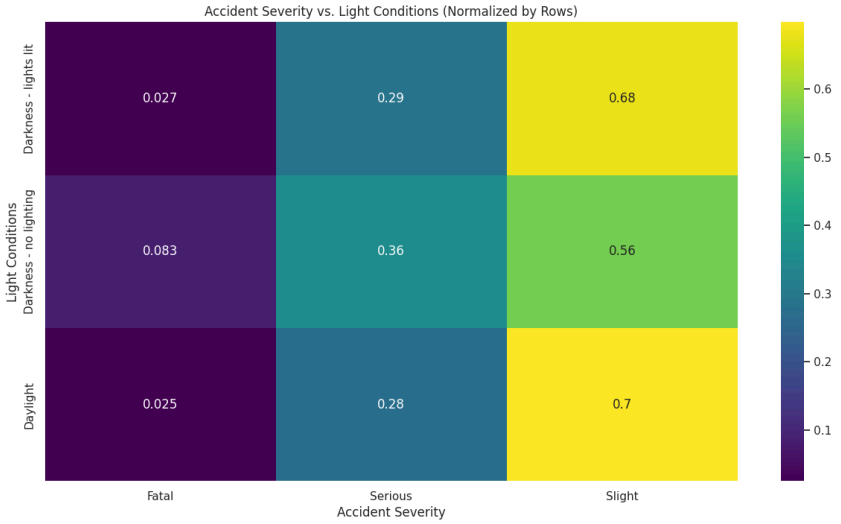


Fig. 17

The project explored the impact of reducing speed limits in areas near junctions, which are often sites of high-risk(Fig. 18). By changing higher speed limits (60, 70) to a lower limit (30) in the dataset and analyzing the results, it was observed that the percentages of serious and fatal accidents decreased. This implies that implementing lower speed limits in dangerous zones, particularly near junctions, can be a potent strategy in reducing accident severity.

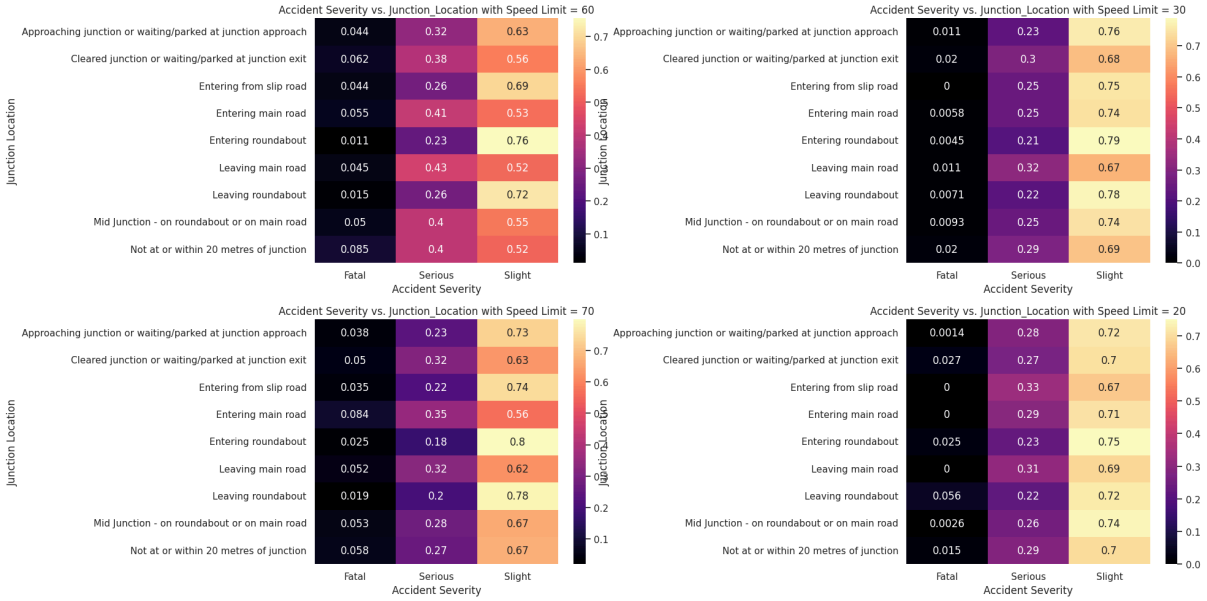


Fig. 18

5. Conclusion

Based on the mentioned analysis, several preventive strategies can be proposed:

1. Enhanced Illumination: The analysis showed a notable decrease in severe accidents when 'Darkness - no lighting' conditions were hypothetically changed to 'Darkness - lights lit'. This suggests that improving lighting conditions, especially in unlit areas, could significantly reduce the risk of severe accidents.
2. Speed Limit Regulation: Modifying speed limits in high-risk areas, particularly near junctions, demonstrated a decrease in severe accidents. Implementing lower speed limits in these areas could be an effective strategy to enhance road safety.
3. Focused Education on Vehicle Manoeuvres: Given the high impact of certain vehicle manoeuvres on accident severity, targeted educational campaigns and stricter regulations for high-risk manoeuvres could be beneficial.
4. Age Consideration: The age band analysis of drivers involved in accidents suggests the need for tailored driving education and stricter licensing processes for specific age groups, particularly elderly drivers.

The implementation of the project could be found on GitHub.

6. References

[1] https://www.kaggle.com/datasets/tsiaras/uk-road-safety-accidents-and-vehicles

[2] https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

[3] Ibrahim, Abdullahi & Raheem, Ridwan & Muhammed, Muhammed & Abdulaziz, Rabiat & Ganiyu, Saheed. (2020). Comparison of the CatBoost Classifier with other

Machine Learning Methods. International Journal of Advanced Computer Science and Applications. 11. 11.