Prediction severity of car accidents in the UK

School of Mathematics, Computer Science and Engineering
Department of Computer Science
City, University of London

Abstract— Nowadays, the majority of people use the roads for commuting to work or errands. The number of accidents increases according to the number of drivers on the road. Accidents are inevitable and unpredictable nobody wishes them to occur. However, if we learn from the prior incidents. It can remind us to avoid similar mistakes and commute more consciously. In summary, this paper use over 1 million observation of road accidents in the UK with several feature parameters including location and time of accident, light and weather conditions, road types, sex, and age of driver and more. Using several techniques of data analysis and machine learning to predict the severity of accidents and identify the main factors affecting the severity of an accident. This paper will use two datasets provided by the British police on the UK government website regarding car accidents that occurred between 2017 and 2021.

Keywords — accidents, road, machine learning, severity, data analysis

I. Introduction

Accidents happen regularly every day, many incidents certainly involve the police. The record of the accident around the United Kingdom was compiled by the British Police Government website. Whether that accident is slight or fatal they collect everything related to that incident such as the area of the accident, the condition around that area, and vehicle information including people who were involved with that accident and experienced injury or death. Accidents are beyond human control, but there contain some interesting insights that allow us to learn and avoid making the same mistakes. if we analyse from the previous records that at any time and in what area and in what condition, there is a high chance of accidents will occur. Avoiding that route and making a better commute choice improves safety when commuting.

Across the world suffer from financial and social expenses because of traffic accidents[1], which is why lowering accident rates is desired. Understanding the factors impacting the likelihood of accidents is crucial. To solve this problem, we apply machine learning approaches to analyse the level of severity of the accident using the observing data and identify the relationships between the severity of the accident and correlated factors. The process of analysing data will visualize important insights to choose the proper features for the training model. This paper will consider several models that will attempt to predict the severity of accident outcomes based on an array of predictor variables. In addition, we will use the result from the methods mentioned above to assist and answer the research questions. These answers should be adapted with the end goal of delivering real-world suggestions for people who use the road for commuting to remind and raise the level of the journey to be more secure.

II. Analytical question

The UK government website provides an interesting dataset of road accidents last five years. Therefore, it is a good opportunity to leverage this data. The analysis process has been conducted to gain better insight from the dataset and identify the relationships between the severity of the accident and correlated factors. Analytic questions define as delivering real-world suggestions for people who use the road to commute whether what kind of driver, what time, or what kind of road conditions are the most likely to have accidents. Especially, identify the main factor contributing to the level of accidents to remind people to beware when they are on the journey.

The below question is what we are going to focus on answering and the scope of our analysis.

- Does the pandemic incident reflex the trend of accidents for the last five years?
- What age groups or sex are most likely to be involved in accidents?
- What are the factors that affect the severity of car accidents? Identify the main factors correlating with the severity of car accidents.
- Do other factors like road surface, weather condition affect fatal road accidents?

III. Data (Materials)

In this paper, there are two primary sources that we going to use for analysing and training our model. The data provides detailed road safety information about the circumstances of personal injury road accidents between 2017 and 2021 by the British police on the UK government website[2].There are 2 CSV files that have been chosen to carry on in this project, the first and primary one is Accidents.csv and the other is vehicles.csv.

1) Accidents.csv: This file contains information on the year, day of the week, time, location of accidents, weather conditions, road types etc. In addition, every line in this file represents a unique traffic accident identified by the accident_index column

2) Vehicles.csv: This file contains information on vehicles involved in collisions such as types of vehicles, age of vehicle also including sex and age of the driver.

The two files as mentioned can be linked through the unique traffic accident identifier. In addition, the target variable that we will focus on in this project is severity in the accident file. It is a categories variable that represents the severity of each accident by numbers 1, 2, and 3 which mean fatal, serious and slight respectively.

The data from the British government contain many features with various types of data such as continuous features like age, latitude, longitude, engine capacity and categorical feature like the severity of the accident, road types, sex etc. However, the majority of the dataset is organized in categorical defined by integer numbers. It is necessary to limit the number of features to be easier to interpret when analysing and training the model and only select those features that were important and relevant to the answer research question.

IV. Analysis

The objective is to leverage and gain better insight from the dataset using visualization and machine learning approaches to answer the analytic questions.

A. Data Selection & Preparation

I. Handling with Missing Data & Imputation

Merge two CSV by the unique traffic accident identifier or known as accidents index columns. The combination of the two datasets has 1035534 million rows and 63 columns.

In this particular dataset, there are two types of missing values '-1' and 'Nan' Start with dropping a few columns that contain Nan values since these columns are not useful for our analysis. Each row in some columns contains -1 value which is mean unknown data removing it directly from this dataset we will lose numerous of data. Hence, to solve this problem we impute mode in the categories column that contains -1 instead and impute mean for continuous data such as age columns. After we impute mode and mean into some missing features, we decided to drop the row that contains -1 since the row that contains -1 now is fairly small now. We can remove it without losing too much data. In the process of removing missing values in the dataset, we lose our data only 19,283 rows instead of 339,116 rows if we remove it directly.

II. Initial Feature Selection

From domain knowledge, we manually picked a set of features that we assumed to be appropriate for our study. Domain knowledge is an essential part to prevent us from proceeding blindly and reducing features that are not relevant to our analysis. As a result, we selected 15 features that we believe are consistent with our assumptions and will help us answer our research questions. The detail and description in each feature have been explained in the table below in figure (1).

Variable	*Description*
accident_severity	*The quality or state of being severe in that accident (1: fatal,2: serious,3: slight)*
number_of_vehicles	*Number of vehicles involved in that case*
number_of_casualties	*Number of casualties involved in that case*
day_of_week	*The day of the week the accident happened. (0-6)*
road_type	*Type of road during the accidents*
speed_limit	*The maximum speed at which a vehicle may legally travel on a particular stretch of road.*
light_conditions	*The brightness of the area during the incident*
weather_conditions	*The weather in the area during the incident*
road_surface_conditions	*The surface of the road during the accidents is slippery, wet, or dry*
vehicle_type	*Type of vehicle involved in the accidents*
sex_of_driver	*The gender of drivers who involved in accidents*
age_band_of_driver	*The range of ages that determines*
age_of_vehicle	*Age of vehicles involved in the accident*
Did_police_officer_attend	*Did the police get involved in that incident or not?*
engine_capacity_cc	*The sum of all the capacities of all the cylinders taken together (cc.)*

Figure 1 – Features description

B. Feature Engineering & Data Deriviation

I. Feature Engineering

In order to maximize observation and analysis, we need to derive and transform our data to be easier to interpret. By adjusting columns or inserting additional features. First, we change date columns from object type into DateTime to make it simple to plot graphs corresponding to the time to compare the frequency of accidents in each year.

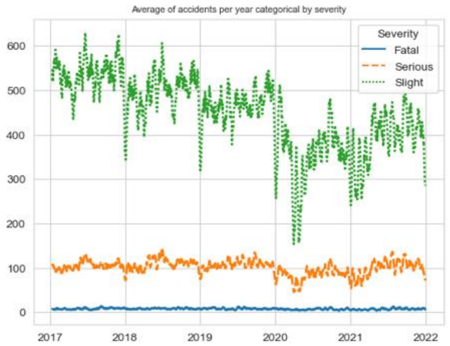


Figure 2 – Severity of accident trend last five years

Second, create new features call “city” and “area” from latitude and longitude columns using an existing python library called reverse geocode introduced by Richard Penman3 to return the city name and area. In order to limit our data when training the model in the first place to save computation time.

Most of the categorical variables in the data are stored as a numeric code, there are some quantitative variables in this dataset as we mentioned. We apply Discretize/bin technique to continue variables to categorize them based on the values of the variable. Change all the features into categories to make them more robust for comparison and contrast before the training model.

II. Important feature & Feature Selection

Multiple Correspondence Analysis (MCA)

Due to our data after discretizing is all categorial, we apply the MCA technique for analysing and studying the association between two or more nominal categories variables[6]. Which can reduce our features into two dimensional for visualizing and observing the distance between categories.

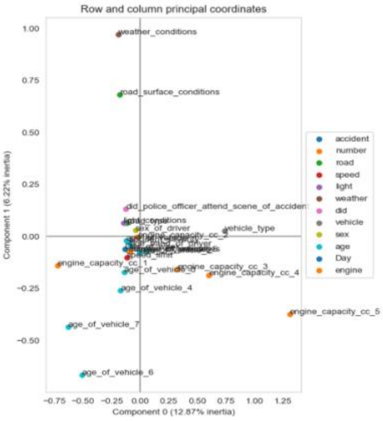


Figure 3 – MCA first two components scatter plot

Scatter Plot's first two components which capture our data around 19.09% allow us to examine the structure of our loadings. From figure (3), weather condition and road surface condition are quite distant from other features, assuming that these features are quite unrelated to other features.

Pearson's Chi-Square Test

The purpose of applying Pearson's Chi-Square method is to gain better insight which features are related to our prediction target for selecting to training our model. In order to, reducing features to make our model easy to interpret and less computing time.

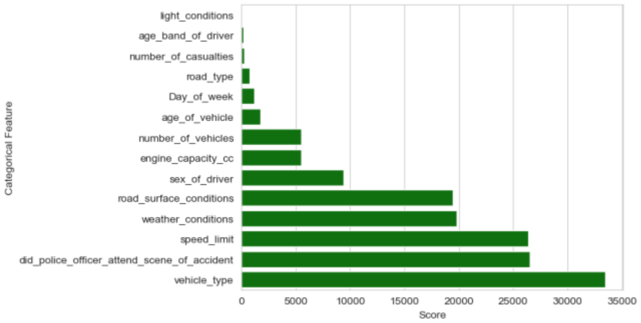


Figure 4 – Pearson's Chi-Square test score important feature

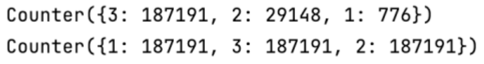
The way to interpret the figure (4) above is that categorical features with the highest values for the chi-squared score indicate higher relevance and importance[4] in predicting severity of accident that is vehicle types, police attendance and speed limit etc. On the other hand, light condition is not an important to choose for building our model.

C. Construction & Optimisation modeling

Machine learning approaches are applied to help us answer the research method and identify the relationship between independent and dependent variables.

I. Imbalanced classification: Border line SMOTE

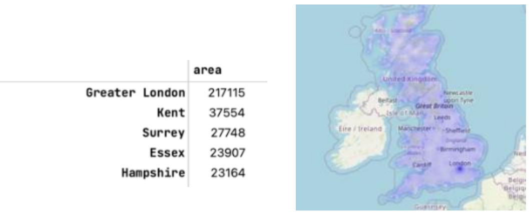
A problem with this dataset is the prediction class in the target variable is imbalanced causing our model inaccuracy and poor performance due to minority classes that are misclassified. Oversampling minority classes and providing more resolution by Synthetic Minority Oversampling Technique known as SMOTE will be required before building the model.



```
Counter({3: 187191, 2: 29148, 1: 776})
Counter({1: 187191, 3: 187191, 2: 187191})
```

Figure 5 – Target class before and after applying border line SMOTE

II. Scoping the dataset



	area
Greater London	217115
Kent	37554
Surrey	27748
Essex	23907
Hampshire	23164

Figure 6– The frequency heatmap area of accident around the UK

Due to working on the large scale of data, we decide to scope our model from the entire country into each area like the greater London area that should consider because of having a higher frequency rate of accident. Different cities have different behavior and features such as speed limits, road types, the number of police etc. Training separation in each area may make the model more accurate and probably more effective for working on the small project to save computing time which is enough to answer our research question.

III. Algorithm & Optimisation

The initial data exploration and analyses allowed for gain insight from the data and describing the relationship between variables like Multiple Correspondence Analysis and chi-square were also used to identify relationships at the beginning. However, to identify the main factors correlating with the severity of car accidents we need to implement a machine learning method based on their ability and popularity for classification task data.

- Random forest (RF) is a powerful and most used supervised learning method used for classification and regression tasks. Random forest is based on multiple decision trees and uses their majority vote for classification and average for regression.

For hyperparameter tuning, we apply grid search cross-validation to run through all different hyperparameter and find the best combination of optimal hyperparameters and selects the best value for the hyperparameters[7] in each model base on a scoring matrix we used accuracy to compare in this research. To find the most accurate predictive model in this analysis.

D. *Validation of Results*

A confusion matrix has been used to evaluate the performance for each model along with the using precision and recall to measure the success rate when classes are imbalanced[5]. Moreover, another metric is F1 score which is a measure of a test's accuracy. F1 is more precise and appropriate to deal with imbalanced classes in this study rather than investigate only accuracy scores.

V. RESULT AND DISSCUSION

A. *Findings from Analysis*

I. *Overview of accident last five years*

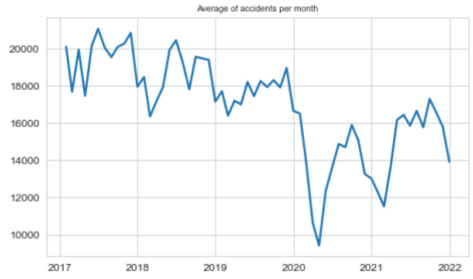


Figure 7 – Accident trends in the last five years

This graph from figure (7) can answer our first analytic question. According to World Health Organization (WHO), the first outbreak of the pandemic was identified in Wuhan in December 2019[8]. Corresponding to figure (7) the trend of accidents decreases dramatically at that time when the pandemic started. We can presume that Covid affects our daily life, fewer people are more likely to drive, resulting in fewer accidents.

II. *The distribution of age and gender reflex to the severity of the accident*

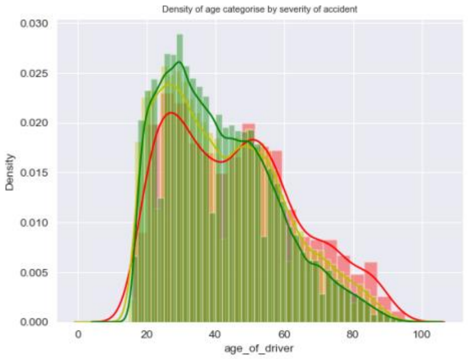


Figure 8 – Distribution age of driver categories by severity of accident

From figure (8) the colour represents the severity of the accident red: fatal, yellow: serious, green: slight, people in the age range of 20-40 are more likely to be involved in an accident than other age groups. On the other hand, the more people get older, the higher probability that accidents can be serious or fatal accident compared to other ages.

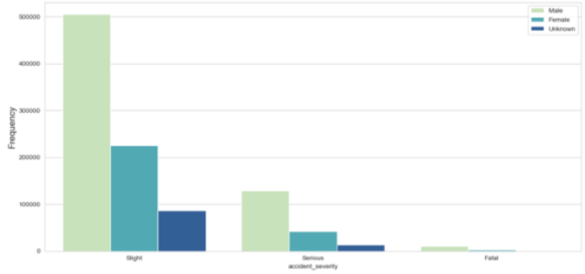


Figure 9 – Proportion of gender in accidents

Figure (9) above show that the proportion of male involved in an accident twice in every severity compared with female and unknown groups which are not surprising due to men tend to get responsibility for driving more than women, emotions and self-control etc. may also play a part to contribute the severity of the accident.

B. *Modelling Outcomes*

To answer our research question 3 and 4 we apply machine learning method to find features that affect the severity of the accident.

I. *Borderline SMOTE result*

We cannot assume from unreliable model having a poor performance of our model when predict the minority class. To solve this problem border line SMOTE has been applied to oversample the examples in the minority class.

Model	**Score**			
	Accuracy	***precision***	***Recall***	***F1 Score***
RF	0.860	0.764	0.860	0.797
	After using Border Line SMOTE			
RF	0.848	0.851	0.848	0.846

Figure 9 – Result comparison after applying borderline SMOTE

Although, the model has decreased the accuracy when performing on the training set but overall, our new model after applying SMOTE perform well equally to predict all three classes and correct predict of the model when working with imbalanced datasets.

II. *Random forest Result*: *Comparison between model with weather and road surface condition and without*

Score		
With weather & road condition		**Without weather & road condition**
RF	***Model***	***RF***
0.808	***Accuracy***	0.776
0.807	***precision***	0.774
0.808	***recall***	0.776
0.804	***F1-Score***	0.771

Figure 10 – Result comparison with and without interested features

Obviously, the absence of weather and road surface condition affect our model performance decrease only 3% overall. These features were given less weight than other features in decision-making for the RF model. As a result, we can conclude that these two features still influence model accuracy but not as much as other features, which corresponds to the results of MCA in the previous data analysis step showing the distance between other features and these two features.

III. Important feature in model

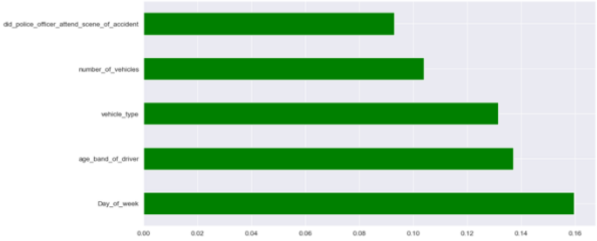


Figure 11 – Top 5 Important features in RF model

Vehicle type and police attendance still plat the part that contributes affect the severity of accident but surprisingly the Day of the week that we derive from the date column and age of the driver are two features that affect the severity in the random forest model as we assume in figure (8).

C. Reflections

The outcome of the model shows that removing weather and road surface condition does slightly affect our model. These features still contribute to level of severity of the accident. However, driving date and driver's age range play the important role in this analysis which is consistent with our general knowledge. If older people have an accident, the severity of the accident increases accordingly. Likewise, if we drive on a busy day in traffic, we probably have a chance to be involved in an accident than in the days when nobody rarely uses cars to travel around. The more accidents occur, the higher chance that the severity of the accident also increases. This evidence from the modelling process can answer our research question number four and three respectively.

D. Further Work

Due to the time limitations that make this study focus on the small aspect of this dataset. We can do further analysis by identifying the trend and effective factor that causes an accident in each city rather than specifically on a big scale like an entire country. Consider using another machine learning as a potential method. One would strongly suggest that using Bayesian optimization for tuning parameters saves computing time than using grid search.

VI. CONCLUSION

The analysis conducted in this paper is a starting point for further research. The time constraints make this study focus on a few aspects of this dataset. However, we achieve the project goals to answer the research questions and identify the features that significantly affect the severity of accidents.

A similar study should be carried out with more emphasis on the impact of features in each city or different regioun[9]. Because different areas tend to have different identities that indicate the level of an accident. Consider employing different analyses of model and algorithm tuning as an enhanced method. A precise model is required to analyze the factor and provide insight to remind people who travel on the road to be more careful and conscious every time they drive.

VII. WORD COUNT

Section	Expected	Actual
Abstract	150	147
Introduction	300	293
Analytical questions and data	300	186
Data (Materials)	300	283
Analysis	1000	1012
Findings, reflections and further work	600	612

REFERENCES

[1] Retallack AE., Ostendorf B. (2019) Current Understanding of the Effects of Congestion on Traffic Accidents. Int J Environ Res Public Health. 2019 Sep 13;16(18):3400. doi: 10.3390/ijerph16183400.

[2] D. for T. UK Government. (2017) “Road Safety Data” [Online]. Available: https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data (Accessed: November 11, 2022).

[3] Thampiman, A. Thampiman/Reverse-Geocoder: A fast, offline reverse geocoder in Python, GitHub. Available at: https://github.com/thampiman/reverse-geocoder (Accessed: December 7, 2022).

[4] Gebeyaw, M. (no date) Selecting categorical features in customer attrition prediction using Python, DataScience+. Available at: https://datascienceplus.com/selecting-categorical-features-in-customer-attrition-prediction-using-python/ (Accessed: December 2, 2022).

[5] Huilgol, P. (2022) Precision vs recall: Precision and recall machine learning, AnalyticsVidhya. Available at: https://www.analyticsvidhya.com/blog/2020/09/precision-recall-machine-learning/ (Accessed: December 5, 2022).

[6] Gilula, Z., Haberman, S.J. and van der Heijden, P.G.M. (2001) “Multivariate analysis: Discrete variables (correspondence models),” International Encyclopedia of the Social & Behavioral Sciences, pp. 10218–10221. Available at: https://doi.org/10.1016/b0-08-043076-7/00477-0.

[7] Shah, R. (2022) GRIDSEARCHCV: Tune hyperparameters with GRIDSEARCHCV, Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/06/tune-hyperparameters-with-gridsearchcv/ (Accessed: December 2, 2022).

[8] Agency, U.K.H.S. (2022) Covid-19: Background information, GOV.UK. GOV.UK. Available at: https://www.gov.uk/government/publications/wuhan-novel-coronavirus-background-information#:~:text=On%2031%20December%202019%2C%20the,City%2C%20Hubei%20Province%2C%20China. (Accessed: December 10, 2022).

[9] M. F. Labib, A. S. Rifat, M. M. Hossain, A. K. Das and F. Nawrine, "Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning in Bangladesh," 2019 7th International Conference on Smart Computing & Communications (ICSCC), 2019, pp. 1-5, doi: 10.1109/ICSCC.2019.8843640.