Analysis of Road Accidents in UK 1979-2015 Using Big Data Approach.

V.L.B. De Mel
Faculty of Information Technology,
University of Moratuwa
Sri Lanka.

U.D.D. Gunarathne
Faculty of Information Technology,
University of Moratuwa
Sri Lanka.

W.K.A.J. Wijethunga
Faculty of Information Technology,
University of Moratuwa
Sri Lanka.

Abstract—

Road accidents are a challenging issue in the world. There is a huge number of lives are deprived and a lot of properties are damaged due to road accidents. So, analyzing road accidents, identifying factors and conditions, and implementing prevention models for road accidents are emerging research areas. In this research area, a major problem is the huge amounts of data required to analyze and interpret the proper solution. This paper presents a descriptive analysis of road accidents in Great Britain from 1979 to 2015. based on a big data analyzing approach. This analysis identifies the correlation between road accidents and driver details such as driver's age and driver's gender. It also includes an analysis of how environmental factors are affecting road accidents. This project uses SPARK which is an open-source unified analytics engine for large-scale data processing. Spark provides an interface for programming entire clusters with implicit data parallelism and fault tolerance. Logistic Regression is used to predict accident severity based on environmental conditions and predict accident severity based on environmental conditions.

Keywords— Big Data Analysis, big data, Apache Spark, Accidents Analysis, Logistic Regression, Random Forest Classifier.

I. INTRODUCTION

Road accidents have become a crucial social issue since the invention of the automobile. There have been a lot of road accidents all over the world, plenty of life and property loss has occurred many problems. According to world health organization statistics, approximately 1.35 million people die each year as a result of road accidents, Road Traffic crashes cost most countries 3% of their gross domestic product, and road traffic injuries are the leading cause of death for children and young adults aged 5-29 years.

There are many reasons behind those accidents. Some major factors such as over speeding, drunken driving, distractions to drivers, red light jumping and avoiding safety gears like seat belts and helmets, and unsafe road infrastructure and vehicles are mainly affected by that. In this project, the goal is to analyze causes for road accidents, finding interest patterns, and finding solutions for avoiding or reducing road accidents using big data analytics techniques

II. RELATED WORK

For the Analysis of Road Accidents, should consider a substantial amount of past data and various attributes. there are several road accidents analyses but those approaches are not good enough to handle huge amounts of data in a fast manner. This project addresses that problem by using SPARK which is a fast and general engine for large-scale data processing.

III. METHODOLOGY

The dataset was collected from the Kaggle. Then processed the data cleaning process the ensuring data is correct, consistent, and usable. In the data cleaning, Firstly, removed duplicated, unknown, and missing observations from the dataset.

Analyzed data based on various attributes using various techniques. For the prediction, here used a statistical machine learning algorithm called Logistic Regression and Random Forest Classifier. By using Random Forest Classifier and Logistic Regression predicted Accident Severity based on the environment as well as Casualty Severity based on the vehicle.

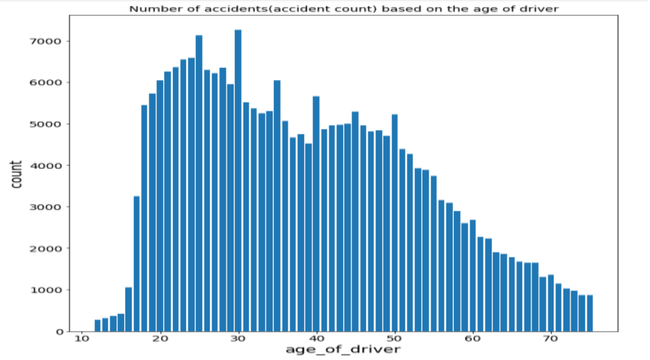
IV. DATASET

The dataset was taken from the Kaggle platform. All the data variables are coded rather than containing textual strings. This dataset includes Road Accidents in the UK between 1979 and 2015 and has 70 features/columns and about 250K rows.

V. ANALYSIS

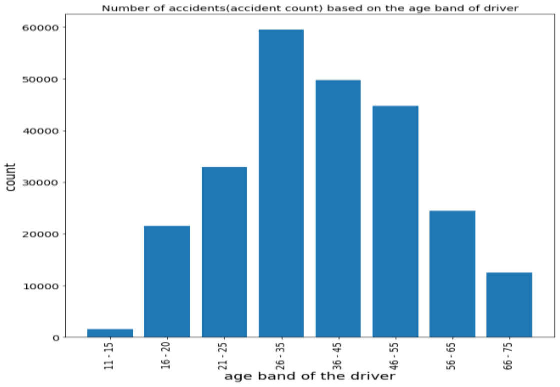
A variety of analyses can be done using the available data.

A. *Analysis of the accident based on Age of the Drivers.*



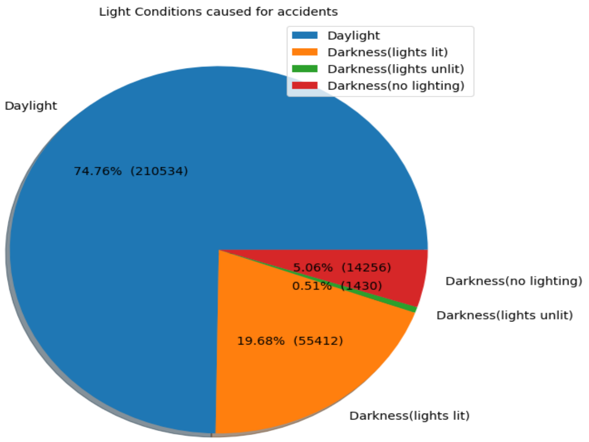
Histogram illustrates the number of accidents based on the age of the driver. It shows below 17 years old people also causing some driving accidents. But there was a law which said that people below 17 years old can't take driving licenses. So, those people cannot be taken as a driver and ignore them for further analysis. This graph is right-skewed. So, it can be said that most young drivers are causing road accidents. According to this graph, drivers of age 30 caused 7262 of the highest number of accidents.

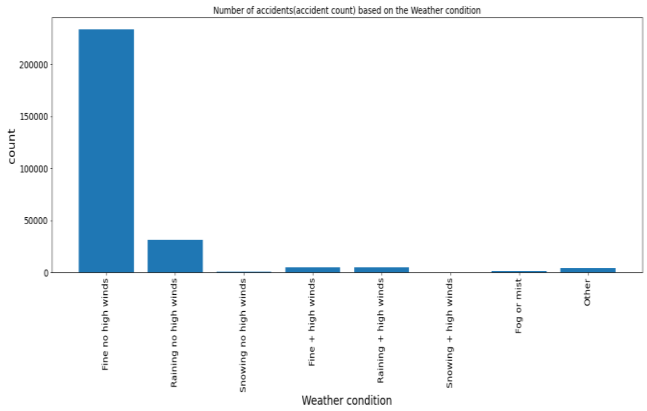
B. *Analysis of the accident based on the Age band of the Drivers.*

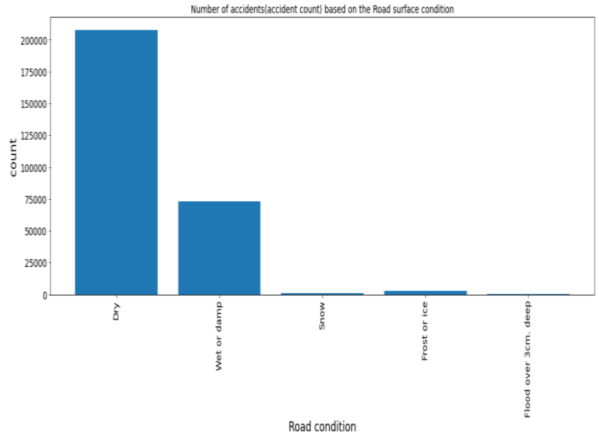


Histogram illustrates the number of accidents based on the age band of the driver. According to this histogram, drivers between 26-35 years old cause the most accidents. Age between 11 and 15 peoples have the least number of accidents, but that cannot take because age below 17 is not allowed for driving. Age between 66 and 75 people has the next least number of accidents. But it also cannot be taken as fact because generally, people of that age are not driving much, therefore, the ratio of the number of drivers is very low.

C. *Analysis of Accidents Based on Road and Environmental Conditions.*







In this work, Environment conditions analyze three ways. The first way is to analyze accidents based on the light conditions of the environment. According to the pie chart, 74.76% of accidents occur in daylight. It is more than 2/3 of total accidents. Darkness condition further can be divided into three parts such as 'lights lit', 'lights unlit' and 'no lighting'. In that 'darkness light lit' condition the highest number of accidents. This shows light conditions cannot be taken as a reason for an accident. The second way is to analyze environmental conditions using weather. Almost every accident happens in 'fine or no high' wind conditions. According to the histogram, more than 200000 accidents happen in 'fine or no high wind' weather. Lastly, analysis was done using the surface conditions of the road. There are more than 200000 accidents that happen in dry conditions. But there are also a considerable number of accidents that happen in 'water or damp' conditions which is more than 75000 accidents.

D. *Time series analysis*.

Here are some of the time series analysis

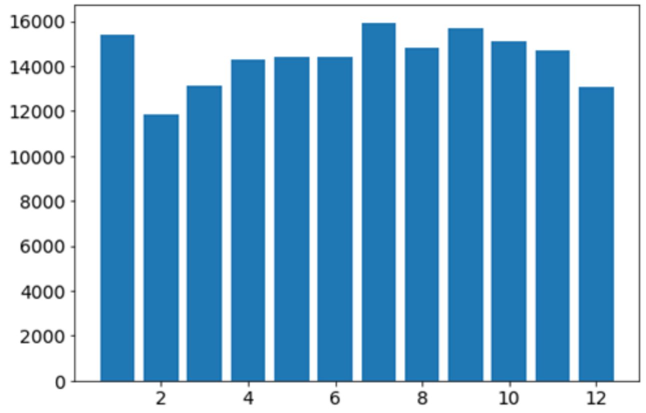


Figure 1 Monthly Accidents Histogram

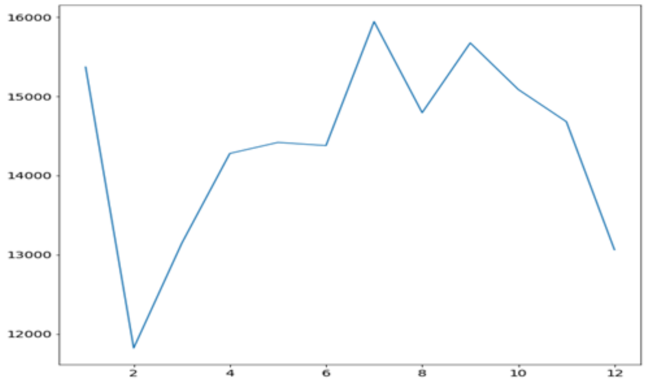


Figure 2 Monthly Accidents Line Graph

This work also has a time series analysis. Time series analysis is done using three ways. They are monthly analysis, weekly analysis, and daily analysis. They can’t extract much information using monthly histogram charts. By observing that, there are on average 14000 accidents happening per month. In July there are the highest number of accidents occurred and in February month there are least number of accidents occurred.

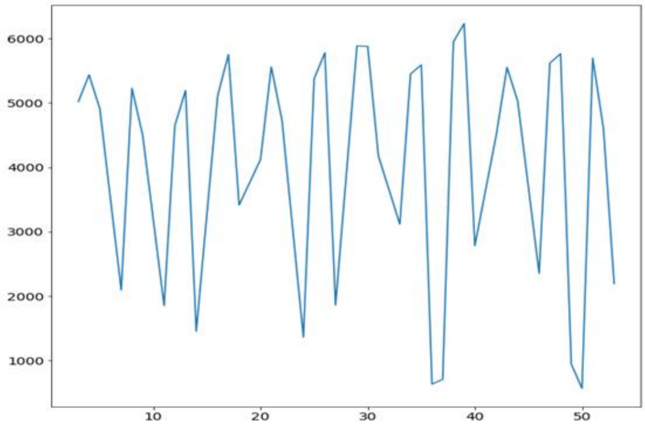


Figure 3 Weekly Accidents Line Graph

Analyzing the weekly accidents line graph, it shows some approximate zig-zag pattern. The zig-zag pattern means it shows that accidents are increasing and then decreasing, the same pattern repeats. This work cannot identify the direct reason for this situation. It is further work and an interesting area to analyze the reason for occurring that pattern in weekly accidents line graphs

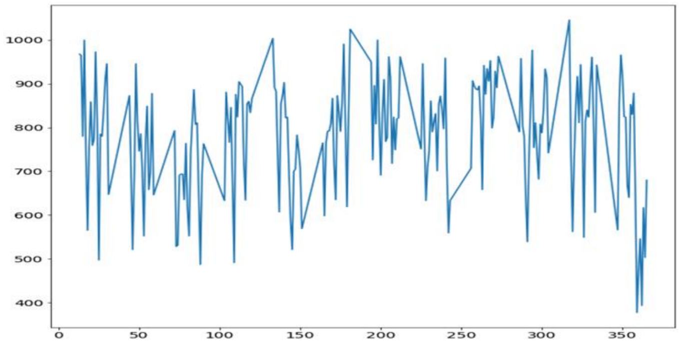


Figure 4 Daily Accidents Line Graph

This work includes daily base time series analysis also but can't do proper analysis daily because of imbalances in data. The daily Accident line graph shows how accidents are occurring daily. Unable to find useful information or patterns from daily basis time-series analyzing.

E. *Predicting Accident Severity Based on Environmental Conditions.*

This prediction model uses a statistical machine learning algorithm called Logistic Regression to predict Accident Severity based on environment. And for the training of this model used weather_conditions, special_conditions_at_site, road_surface_conditions, and light_conditions features. By Using this Logistic Regression algorithm able to formulate a model with an accuracy of 0.7801.

E.1 Random forest Classifier.

Random forest is a flexible, easy-to-use machine learning algorithm that provides exceptional results most of the time even without hyper-tuning. As mentioned in the Decision tree, the major limitation of the decision tree algorithm is overfitting. It appears as if the tree has memorized the data. Random Forest prevents this problem: It is a version of ensemble learning. Ensemble learning refers to using multiple algorithms or the same algorithm multiple times. Random forest is a team of Decision trees. And greater the number of these decision trees in Random Forest, the better the generalization [2].

F. *Predicting Casualty Severity Based on Vehicle Conditions.*

This prediction model uses a statistical machine learning algorithm called Random Forest Classifier to predict Casualty Severity based on the vehicle. And for the training of this model used engine_capacity_(cc), vehicle_manoeuvre, age_of_vehicle, and vehicle_type features. By Using this Random Forest Classifier able to formulate a model with an accuracy of 0.8225.

F.1 Logistic Regression.

Regression is an analysis or supervised learning task which is used for modeling and predicting the association between the dependent variable and independent variable. Regression can be classified as linear regression and logistic regression. In logistic regression, the dependent variable should be discrete whereas in linear regression the dependent variable should be continuous. Logistic regression is also popularly known as logit model which uses the logistic function. Predictions in the logit model can be mapped between 0 and 1 [3].

VI. FUTURE WORK

This project uses Random Forest and Logistic Regression for predictions. For further can test accuracy and reliability with other algorithms and techniques for the model prediction. For increasing accuracy can use a balanced dataset since the current dataset is not much balanced.

VII. CONCLUSION

The analysis implies that there is no considerable relation between Accident Severity with Engine Capacity of the Vehicle, Road Surface Conditions, Weather Conditions, and Light Conditions.

The analysis shows high accidents from the middle-aged drivers, but the reason behind that is the ratio of middle-aged drivers is high compared to other aged drivers.

VIII. REFERENCES

1. " Road Accidents Incidence", Kaggle.com, 2021. [Online]. Available: https://www.kaggle.com/akshay4/road-accidents-incidence
2. S. Grampurohit and C. Sagarnal, "Disease Prediction using Machine Learning Algorithms," 2020 International Conference for Emerging Technology (INCET), 2020, pp. 1-7, doi: 10.1109/INCET49848.2020.9154130.
3. S. Saxena and S. N. Prasad, "Machine learning based sensitivity analysis for the applications in the prediction and detection of cancer disease," 2019 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER), 2019, pp. 1-4, doi: 10.1109/DISCOVER47552.2019.9008083.