Road Traffic Accidents Injury Data Analytics

Mohamed K Nour[1]
College of Computer and Information Systems
Umm Al-Qura University

Atif Naseer[2]
Science and Technology Unit
Umm Al-Qura University

Basem Alkazemi[3]
College of Computer and Information Systems
Umm Al-Qura University

Muhammad Abid Jamil[4]
College of Computer and Information Systems
Umm Al-Qura University

***Abstract*—Road safety researchers working on road accident data have witnessed success in road traffic accidents analysis through the application data analytic techniques, though, little progress was made into the prediction of road injury. This paper applies advanced data analytics methods to predict injury severity levels and evaluates their performance. The study uses predictive modelling techniques to identify risk and key factors that contributes to accident severity. The study uses publicly available data from UK department of transport that covers the period from 2005 to 2019. The paper presents an approach which is general enough so that can be applied to different data sets from other countries. The results identified that tree based techniques such as XGBoost outperform regression based ones, such as ANN. In addition to the paper, identifies interesting relationships and acknowledged issues related to quality of data.**

***Keywords*—Traffic Accidents Analytics (RTA); data mining; machine learning; XGBOOST**

I. Introduction

Road Traffic Accident (RTA) is an unexpected event that unintentionally occurs on the road which involves vehicle and/or other road users that causes casualty or loss of property. Over 90% the world's fatalities on roads occur in low and middle income countries which account for only 48% of world's registered vehicles [1]. The financial loss, which is about US$518 billion, is more than the development assistance allocated for these countries. While developed rich nations have stable or declining road traffic death rates through co-ordinated correcting efforts from various sectors, developing countries are still losing 1–3% of their gross national product (GNP) due to the endemic of traffic casualties. World Health Organization (WHO) fears, unless immediate action is taken, road crash will rise to the fifth leading cause of death by 2030, resulting in an estimated 2.4 million fatalities per year [1].

Thus, measures to reduce crashes based on in-depth understanding of the underlying causes are of great interest for researchers.The 21st century has been seeing a rapid growth of road motorisation due to rapid increase of population, massive urbanisation, and increased mobility of the modern society, risks of road traffic fatality (RTF) may also become higher and RTA can also be assumed as a "modern epidemic". This paper presents an analytic framework to predict accident severity for road traffic accidents [1]. Past research on road traffic accidents analysis had mainly relied on statistical methods such as linear and Poisson regression. This paper presents an analytic framework to predict accident severity for road traffic accidents. In particular, the paper addresses issues related to data preprocessing and preparation such as data aggregation, transformation, feature engineering and imbalanced data. In addition, the paper aims to apply machine learning models to enable more accurate predictions. Hence, the compares the performance of several machine learning algorithms in predicting the accident injury severity. In particular, the paper applies logistic regression, support vector machines, decision trees, random forest , XGBoost and artificial neural network models. The rest of this paper is organized as follows: Section II introduces some previous works. Section III shows the methodology used in this work, Section IV describes the data management and the patterns of traffic accident data. Section V shows the results and analysis of the all the approached used in this work. Section VI gives the conclusions and future works.

II. Literature Review

Mehdizadeh et al. [2] presented a comprehensive review on data analytic methods in road safety. Analytics models can be grouped into two categories: predictive or explanatory models that attempt to understand and quantify crash risk and (b) optimization techniques that focus on minimizing crash risk through route/path-selection and rest-break scheduling. Their work presented a publicly available data sources and descriptive analytic techniques (data summarization, visualization, and dimension reduction) that can be used to achieve safer-routing and provide code to facilitate data collection/exploration by practitioners/researchers. The paper also reviewed the statistical and machine learning models used for crash risk modelling. Hu et al. [3] categorized the optimization and prescriptive analytic models that focus on minimizing crash risk. Ziakopoulos et al. [4] critically reviewed the existing literature on different spatial approaches that include dimension of space in its various aspects in their analyses for road safety. Moosavi et al. [5] identified weaknesses with road traffic accidents research which include: small-scale datasets, dependency on extensive set of data, and being not applicable for real time purposes. The work proposed a data collection technique with a deep- neural-network model called Deep Accident Prediction(DAP); The results showed significant improvements to predict rare accident events. Zagorodnikh et al.[6] developed an information system that displays the accidents concentration on electronic terrain map automatically mode for Russian RTA to help simplifying the RTA analysis.

Kononen et al. [7] analysed the severity of accidents occurred in United States using logistic regression model. They reported performance 40% and 98%, for sensitivity and specificity respectively. Also, they identified the most important predictors for injury level are: change in velocity, seat belt use, and crash direction.

The Artificial Neural Networks (ANNs) are one of the data mining tools and non-parametric techniques in which researchers have analysed the severity of accidents and injuries among those involved in such crashes. Delen et al. [8] applied a ANNs to model the relationships between injury severity levels and crash related factors. They used US crash data with 16 attributes. The work identified four factors that influence the injury level: seat belt, alcohol or drug use, age, gender, and vehicle.

Naseer et al. [9] introduces a deep learning based traffic accident analysis method. They highlighted deep learning techniques to build prediction and classification models from the road accident data.

Sharma et al. [10] applied support vector machines with different Gaussian kernel functions for crash to extract important features related to accident occurrence. The paper compared neural network with support vector machines. The paper reported that SVMs are superior on accuracy. However, the SVMs method has the same disadvantages of ANN in traffic accident severity predication as mentioned earlier

Meng et al. [11] used XGBoost to predict accidents using road traffic accident data from multiple sources. They used historical data along with weather and traffic data. Schlogl et al. [12] performed multiple experiments to proof that XGBoost performs better as compared to several machine learning algorithms.

Ma et al. [13] proposed the XGBoost based framework which analysed the relationship between collision, time and environmental and spatial factors and fatality rate. Results show that the proposed method has the best modelling performance compared with other machine learning algorithms. The paper identified eight factors that have impact on traffic fatality.

Cuenca et al. [14] compared the performance of Naive Bayes, Deep Learning and Gradient Boosting to predict the severity of injury for Spanish road accidents. Their work reported that Deep Learning outperform other methods

III. Methodology

The methodology adopted in this paper is shown in Fig. 1. The first step for data analytics process is data collection which is regarded as the primary building block for successful data analysis project. There are many data sources like sensors, visual data through cameras, and IoT and mobile devices which captures data in different formats and need to be stored realtime or offline. In addition, data collected from different authorities related to traffic volume, accident details and demographic information. The storage can be on the local servers or on cloud. The key of data management pyramid is data preprocessing. The data acquired from the storage locations cannot be used as it is. It requires preprocessing before performing any analysis. The acquired data may include missing information that needed to rectify as well as many information needed to be removed due to duplication. The preprocessing may involve the data transformation as it helps in data normalization, attribute selection, discretization, and hierarchy generation. Data reduction maybe required on the large scale of data as the analysis of a huge amount of data is harder. Data reduction increases the efficiency of storage and the analysis cost. Data Analysis use multiple machine learning algorithms to get the insight of data. The data analysis is very crucial for any organization as it provides the detailed information about the data and is helpful in certain decision making and predictions about the business. Data can be presented in various forms depending on the type of data being used. The data can be shown into organized tables, charts, or graphs. Data presentation is very important for business users as it provides the results from the analysis of data in a visual format.

One of the most important tasks for road risk analysis and modelling is to predict accident severity level. This paper looks at building predictive model for accident severity level and investigate the process of constructing a classification model to predict accident severity level, in particular the study:

- Presents the data management framework. This is followed by discussion on how the data was prepared prior to modelling. This includes pre-processing and data cleansing. This section is presented in the Data Management Framework section
- Identifies gaps in the RTAs predictive modelling techniques. This section gives brief background on each technique used in this paper and presents prior work in road traffic accident prediction together with with data requirements and recorded performance results. This topic is presented in the Data Analysis section
- Build prediction models. This section begins by stating performance metrics then followed by data used. Then compares classifiers in particular; logistic regression, support vector machines, neural networks, decision trees, random forest and Extreme gradient boosting tree (XGboost) . In this section, the unbalanced class distribution is investigated to see their impact on injury severity during accidents. This is presented in the results section

IV. Data Management

A. Data Collection

The data comprises publicly available data from UK government which spans the period of 2005 to 2019 [15]. Although the UK department of transport provide data from 1979, it was reported the data collected from 2005 onwards are more accurate and contains less missing data. The records shows information on road traffic collisions that involve personal injury occurring on public roads which have been reported to the police. Data is collected by the authorities at scene of an accident or, in some cases, reported by a member of the public at a police station, then processed and passed on to the authorities. Data includes, 2 million unique collisions, with x, y space coordinates available. Data related to traffic flow is and information about all UK network roads and local authorities are also available separately. The dataset contains a single entry for each accident with 33 attributes (features). The attributes can be grouped by geography, accident-focused, weather, time . Data Related to vehicles involved with the accidents is stored in separate file with 16 attributes. Data related to casualty involved with the accident is stored in 23 fields file. The relation between these three files is one to many, i.e one accident row can contain many casualty rows and many vehicle rows with the accident index is the linking field.

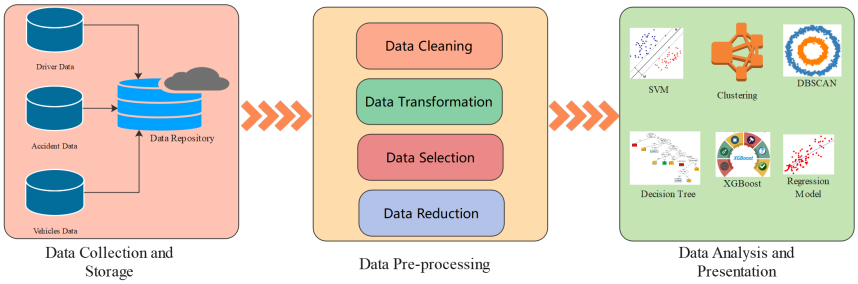

Fig. 1. Proposed Methodology

The severity in any accident is the most important feature to analyze the injury pattern. According to the WSDOT [16], the severity level during the accidents are measured using the KABCO scale, which uses the parameters: fatal (K), incapacitating-injury (A), non-incapacitating injury (B), minor injury (C), and property damage only (PDO or O)).

In this work, we divided the accidents into three categories i.e. Fatal (where the death occurs within 30 days of accident), Serious injury (where the person requires hospital treatment) and Slight injury (where the person not required any medical treatment).

In this project the data is stored in relation database. In future work however, the relational data first will be denormalized then transformed into Hadoop key-value records

B. Data Preprocessing

Different data preprocessing and cleaning methods were applied to the data. Data pre preprocessing involve many tasks and techniques. This include: dealing with missing value, outlier or anomaly detection, feature selection. A key stage in the data analytic is the selection of data. Data needs to be of good quality and clean.

Data quality considerations include accuracy, completeness and consistency [17]. In addition data volume is important as well. Data should be large enough to be of value in predictive modelling. It must be split into training, test and validation subset in order to evaluate the model. The following data preprocessing steps were applied to the data in order to make the data ready for analysis and machine learning algorithms:

- Most machine learning methods require the data to be in either binary or numeric format. However, in real life data sources include category attributes such road type, casualty class. All category attributes will be converted into numeric values.
- Numeric values differ in ranges. To avoid bias to large numeric values all numeric values will be normalized to values between 0 and 1
- Records with missing values will be removed.
- Date field will create several relate fields such as month, year and week.
- Determine less quality attributes. Attributes with more than 70% missing values will be removed.
- Calculate the correlation between severity level and all other attributes. Attributes with high correlation values will be removed as well as attributes with very low correlation values.
- Create fields related to easting and northing to create zones for accidents instead of specific location. The threshold value for zone level is $1km^2$.

One of the issues that faces building analytic models for crash severity, is the imbalance of data [18] where the occurrence fatality which is infrequent or rare event compared to no or minor injury accidents. Due to the extreme imbalance of accident data most algorithms will not produce good predictive models and perform poorly will likely misclassify the fatal accidents as it is not prevalent in the dataset [19]. For imbalanced data sets such as traffic accidents data, sampling techniques can help improve classifier accuracy [19]. Two sampling techniques; undersampling and oversampling techniques will be discussed below.

Under sampling is used to adjust the class distribution of a dataset in favour of the minority class. With undersampling, the majority class is reduced or under sampled [17] and randomly eliminates data from the majority class until both classes match. Oversampling is a technique used in data mining to adjust the class distribution of a dataset in favour of the majority class [18]. Oversampling, on the other hand, the minority class increased or over sampled until the size meets that of the majority class. However, these techniques require specialised skill and it can take a significant time-frame to identify the best sample.

In this study, we need to apply feature selection task on the dataset. The dataset is preprocessed as specified in Tables I, II and III. The Table I shows the features with respect to accidents, Table II shows all features of vehicles, while the Table III highlights the features with respect to casualty with their type. The tables also shows the preprocessing on the features so that some features excludes from the list while some of them adjusted with scale.

TABLE I. ACCIDENTS FEATURES

Variable Name	Type	Preprocessing
Accident Index	Link field	EXECLUDE (unique in accidents)
Police Force	Number from 1-98	0 london 1 otherwise
Accident Severity	1 Fatal 2 Serious 3 Slight	EXECLUDE
Number of Vehicles	Numeric	SCALE
Number of Casualties	Numeric	SCALE
Date (DD/MM/YYYY)	DATE	Split into Moth, Year and Week, weekend Weekday
Day of Week	1 TO 7	SCALE
Time (HH:MM)	TIME	Split into Rush hours and Non Rush Hours
Location Easting OSGR (Null if not known)	Numeric	Remove Last two dists and scale
Location Northing OSGR (Null if not known)	Numeric	Remove Last two dists and scale
Longitude (Null if not known)	Numeric	INCLUDE
Latitude (Null if not known)	Numeric	INCLUDE
Local Authority (District)	1 to 941	exclude
Local Authority (Highway Authority - ONS code)	208 Items	exclude
1st Road Class	1 to 6	0 motoryway, 1 othewise
1st Road Number	Numeric	**exclude**
Road Type	1 to 12	1 motoryway
Speed limit	Numeric	SCALE
Junction Detail	0 TO 9	0 no juntion, 1 otherwise
Junction Control	0 TO 4	0 no junction control, 1 otherwise
2nd Road Class	0 TO 6	0 motoryway, 1 othewise
2nd Road Number	Numeric	exclude
Pedestrian Crossing-Human Control	0 TO 2	0 no pedestrain crossing, 1 otherwise
Pedestrian Crossing-Physical Facilities	0 TO 8	0 no pedestrain crossing, 1 otherwise
Light Conditions	1 TO 7	0 daylight 1 otherwise
Weather Conditions	1 to 9	0 good conditions, 1 otherwise
Road Surface Conditions	1 to 7	0 dry, 1 otherwise
Special Conditions at Site	0 to 7	0 no special conditions, 1 otherwise
Carriageway Hazards	0 to 7	0 no hazard, 1 otherwise
Urban or Rural Area	1 to 3	0 urban, 1 otherwise
Did Police Officer Attend Scene of Accident	1 TO 3	0 attend, 1 otherwise

TABLE II. VEHICLE FEATURES

Variable Name	Type	Preprocessing
Accident Index	Link field	EXECLUDE (unique accident, one or more vehicle)
Vehicle Reference	Link field	EXECLUDE (unique vehicle & one or more casualty)
Vehicle Type	1 TO 113	0 car 1 otherwise
Towing and Articulation	1 TO 5	0 no towing, 1 otherwise
Vehicle Manoeuvre	1 TO 18	0 reversing,1 otherwise
Vehicle Location-Restricted Lane	0 TO 10	0 on main c way, 1 otherwise
Junction Location	0 TO 8	0 no juntion, 1 otherwise
Skidding and Overturning	0 TO 5	0 no skidding, 1 otherwise
Hit Object in Carriageway	0 TO 12	0 no object, 1 otherwise
Vehicle Leaving Carriageway	0 TO 8	0 not leaving, 1 otherwise
Hit Object off Carriageway	0 TO 11	0 not object off c way, 1 otherwise
1st Point of Impact	0 TO 4	0 no impact, 1 otherwise
Was Vehicle Left Hand Drive	1 TO 2	0 right hand, 1 otherwise
Journey Purpose of Driver	1 TO 6	0 work, 1 otherwise
Sex of Driver	1 TO 3	0 male, 1 otherwise
Age Band of Driver	1 TO 11	
Engine Capacity	Numeric	
Vehicle Propulsion Code	1 TO 10	0 Petrol, 1 otherwise
Age of Vehicle (manufacture)	Numeric	
Driver IMD Decile	0 TO 10	0 deprived 1 othewise
Driver Home Area Type	1 TO 3	0 deprived 1 othewise

TABLE III. CASUALTY FEATURES

Variable Name	Type	Preprocessing
Accident Index	Link field	EXECLUDE (unique accident, one or more casualty)
Vehicle Reference	Link field	EXECLUDE (unique in casualty table)
Casualty Reference	Link field	EXECLUDE (unique in vehicle and one or more in casualty)
Casualty Class	1 TO 3	0 driver, 1 otherwise
Sex of Casualty	1 TO 2	0 male, 1 otherwise
Age Band of Casualty	Numeric	SCALE
Casualty Severity	1 TO 3	TARGET VALBLE (0 fatal , 1 otherwise)
Pedestrian Location	1 TO 10	0 crossing,1 otherwise
Pedestrian Movement	1 TO 9	0 crossing,1 otherwise
Car Passenger	0 TO 2	0 passenger, 1 otherwise
Bus or Coach Passenger	0 TO 4	0 bus or coach passenger, 1 otherwise
Pedestrian Road Maintenance Worker (From 2011)	0 TO 2	0 road worker, 1 otherwise
Casualty Type	0 TO 113	0 pedestrain 1 otherwise
Casualty IMD Decile	0 TO 10	0 deprived 1 othewise
Casualty Home Area Type	1 TO 3	0 urban 1 otherwise

C. Data Analysis

Methods for traffic accident prediction can be broadly classified into three categories, namely statistical models, machine learning and analytics approaches, and simulation-based methods [17]. In this research we will concentrate on machine learning approaches.

Machine learning is a broad concept, which include supervised learning and unsupervised techniques. Supervised learning techniques include:artificial neural networks and its variations (Deep Learning, self-organised map), support vector machine (SVM), decision trees, Bayesian inference. Unsupervised learning include: association rules and clustering techniques.

Unsupervised learning involves searching for previously unknown patterns or groupings. Usually these techniques work without a prior target variable. Clustering and association rules fall under this group of techniques. Supervised learning, on the other hand, involves classification, prediction and estimation techniques that contain a target variable. Classification is a machine learning technique that assigns a class to an instance, i.e. automatically assigning traffic accident to one predefined class of severity. Prediction is similar to classification but involve assigning a continuous value to an instance.

Supervised learning methods usually use two sets: a training and test set. Training data is used for learning the model and requires a primary group of labelled traffic accident. Test set is used to measure the efficiency of the learned model and includes labelled traffic accident instances, which do not participate in learning classifiers.

This paper focuses on applying classification methods to classify accident severity. Five techniques will be applied and compared: (1) logistic regression models, (2) deep neural networks, (3) support vector machines, (4) decision trees, (5) extreme gradient boosting.

1) Logistic Regression: Regression models have become an integral component of any data analysis concerned with the relationship between a response variable and one or more explanatory variables. Logistic regression is a maximum-likelihood method that has been used in hundreds of studies of crash outcome.Traditionally, statistical regression models are developed in highway safety studies to associate crash frequency with the most significant variables. The logistic regression is a special case of the generalized linear model (GLM), which generalizes the ordinary linear regression by allowing the linear model to be related with a response variable that follows the exponential family via an appropriate link function. Logistic regression can be binomial or multinomial. The binomial logistic regression model has the following form:

$$p(y|x, w) = Ber(y|sigm(w^T x))$$

where w and x are extended vectors, i.e.,

$$w = (b, w1, w2, ..., w_D), x = (1, x1, x2, ..., x_D).$$

2) Artificial Neural Networks: Artificial Neural Networks (ANN) was build to imitate how the human brain works. It is formed by creating a network of small processing units called Neurons. Each neuron is very primitive, but the network can achieve complex tasks such as pattern recognition, image classification, and detection, natural language processing, etc. Mathematically ANN can be looked like a type of regression system that predicts and estimates new values from historical records. ANN is able to estimate any non-linear functions provided enough datasets were supplied for training the ANN. The architecture of ANN is built with three layers:

1) Input Layer: This layer receives the feature for the model.
2) Hidden Layer: This layer consists of one or more layers that identify the depth of the ANN. Each layer is connected through nodes with weighted edges. The performance of the model depends greatly on the hidden layers and their connectivity with input and output layers.
3) content...

3) Support Vector Machines: Support Vector Machines (SVMs) have been introduced as a new and novel machine learning technique according to the statistical learning theory. SVMs are used for classification and regression problems. Structural Risk Minimization (SRM) applied by SVM can be superior to Empirical Risk Minimization (ERM) since SRM minimizes the generalization error.

The primal form of SVM for classification is:

$$H : y = f(x) = sign(wx + b)$$

For regression the SVM is represented as:

$$H : y = f(x) = w^T x + b$$

4) Decision Trees: Decision trees are powerful data mining methods that can be used for classification and prediction.Decision trees represent rules, which are easy to interpret. There are a multiple methods used in creating decision trees for example: Iterative Dichotomiser 3 (ID3) and C4.5. Decision

trees are supervised learning methods. The decision trees working mechanism is to divide the data data into training and testing sets randomly.

Decision trees have high variance because the model yields to different results. Namely, bagging and boosting. In Bagging techniques, many decision trees build in parallel, form the base learners of bagging technique.The sampled data is input to the learners for training.

In boosting techniques, the trees are build sequentially with fewer splits. Such small trees, which are not very deep, are highly interpretable. The validation techniques like k-fold helps in finding the optimal parameters which helps in finding the optimal depth of the tree. Also, it is very important to carefully stop the boosting criteria to avoid over-fitting.

5) eXtreme Gradient Boosting (XGBoost): XGBoost, a scalable machine learning system for tree boosting which is proofed very popular in machine learning competitions such as kaggle and kdnuggests Most winning teams either utilize or supplement their solution with XGBoost . This success can be mainly attributed to the scalabitliy feature that is inherit inside the algorithm. Scalabitliy is due to the optimized learning algorithm to work with sparse data,parrallisim and its utilitisation of mutlithreading [20]. XGBoost is a boosting algorithm which uses gradient descent optimization technique with a regulized learning objective function. It has the following features:

1) Regularization: XGBoost prevents overfitting by using L1 and L2 regularization.
2) Weighted quantile sketch: Finding the split points is core task of most decision tree algorithms. Their performance affected if the data is weighted. XGBoost handles weighted data through a distributed weighted quantile sketch algorithm.
3) Block structure for parallel learning: XGBoost utilizes multiple cores on the CPU using a block structure which part of its design. Data is sorted and stored in in-memory units or blocks which enables the reuse of data by iterations. This also useful for split finding and column sub-sampling tasks.
4) Handling sparse data: Data can become sparse for many reasons such as missing values or one-hot encoding. XGBoost split finding algorithm can handle different types of sparsity patterns in the data.
5) Cache awareness: In XGBoost, non-continuous memory access is required to get the gradient statistics by row index. Hence, XGBoost has been designed to make optimal use of hardware. This is done by allocating internal buffers in each thread, where the gradient statistics can be stored.
6) Out-of-core computing: This feature optimizes the available disk space and maximizes its usage when handling huge datasets that do not fit into memory.

D. Model Evaluation

Evaluation is a key stage in the data analytics that assesses the predictive capability of the model and identify the model which performs best [17]. Several techniques normally used to evaluate classification models such as the confusion matrix, receiver operator curve (ROC) and the area under the curve (AUC). A confusion matrix shows the correct classifications true positives (TP) and true negatives (TN) in addition to incorrect classification false positives (FP), and false negatives (FN) [21]. The accuracy is calculated from the confusion matrix which gives the precision (percentage of data correctly classified) and recall (percentage of data which are correctly labelled) values. The equations from 1-6 shows the performance matrices formulas.

$$TPR = \frac{TP}{TP + FN} \tag{1}$$

$$FPR = \frac{FP}{FP + TN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F - measure = \frac{2 * Recall * Precision}{recall + precision} \tag{6}$$

V. Results and Analysis

Using python, jupyter notebook and Scikit learn, pandas and matplot data science libraries we have developed a workflow for processing the dataset and generate the corresponding accident severity prediction models. It is composed of a number of nodes, namely:

1) Dataset: contains the pre-processed data for the experiment
2) Explore Data: is an optional node to help in data exploration and viewing some statistics about the data before modelling.
3) Model: contains the algorithms that will be used for model generation.
4) Apply: where the model is applied to the predictors to generate the required results
5) Predictors: sample dataset for testing the prediction.
6) Prediction: the resulted table after applying the model on the predictors.

The dataset we have used is the UK traffic accidents that occurred between 2005-2019 obtained from UK department of transport. The data comes in three different files:accidents, casualties and vehicles in Tables I, II and III. Accidents_ index field joins the three tables in a one to many relationship, with one accident record corresponds to one or more casualties and vehicle records. Casualty table has a field called Vehicle Reference links a particular casualty record with vehicle and driver information record with the same accident_ index field. The original dataset contains about 2M accidents, 2M casualties and 5M Vehicles. The combined table resulted in records 3M records.

Data was explored using bar charts and histograms to look for trends and patterns in the data. Examples of such graphs are shown in Fig. 2, 3 and 4.

The following observations can be noted from the data:

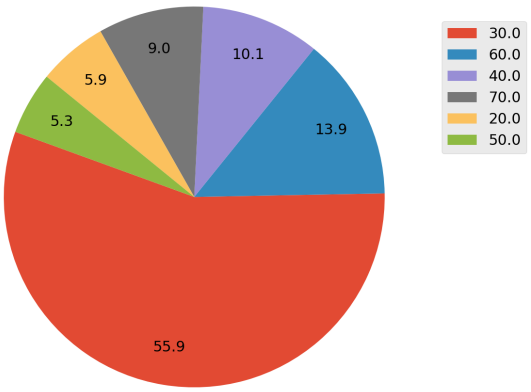

Fig. 2. Accidents per Speed Zone

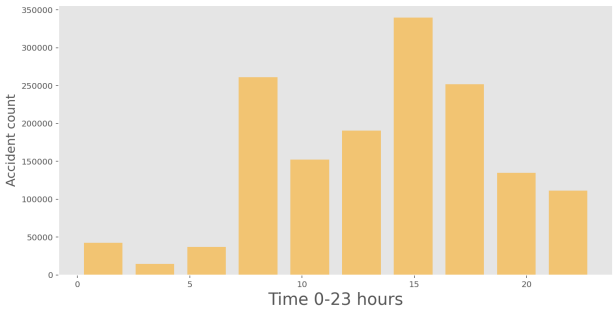

Fig. 3. Accident Time of Day

1) Accident severity (more than 90% records with slight severity).
2) Most crashes involve less than five cars with median 2 and mean 1.8.
3) Number of casualty range between 1 to 93 with mean 1.3 and median 1.
4) Accidents spread throughout the week with slight increase of accidents in Thursdays.
5) Accidents spread throughout the day with slight increase at school return time during working days.
6) First road class 3 roads has maximum number of accidents and single carriage ways roads and speed limit 30 MPH.
7) Uncontrolled junctions has more accidents than other types of accident.
8) Most accidents occur with fine weather conditions with Dry road surface.

The date then cleaned from incomplete records. All records with empty cells or value -1 were considered missing and removed. Then a histogram diagram was created for each column and non widespread columns were removed.

1) $Bus\_or\_Coach\_Passenger$
2) $Towing\_and\_Articulation$
3) $Vehicle\_LocationRestricted\_Lane$
4) $st\_Point\_of_Impact$
5) $Was\_Vehicle\_Left\_Hand\_Drive?$
6) $st\_Road\_Class, st\_Road\_Number$
7) nd_Road_Class

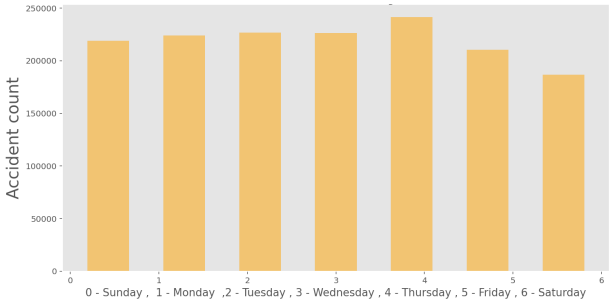

Fig. 4. Accidents Day of the Week

8) nd_Road_Number
9) $Pedestrian\_Crossing$
10) $Human\_Control$
11) $Special\_Conditions\_at_Site$

The resulting number of attributes used for model building is 48 features. Extra preprocessing was implemented on category type attributes by encoding with 0 and 1 values. Numeric and ordinal fields were scaled to remove bias due to large values. Out of 3M records of accidents only 29K were fatalities and 190K serious injuries. Serious and fatal records are grouped into one class and slight injury was the second class. This reduces the class imbalance together with under sampling method will enables the models achieve better results.

The selected models for this experiment are: logistic regression, decision trees, random forest, neural networks and XGBoost. Hyper parameter tuning was applied the methods. The dataset was partitioned two parts, 70% for training and 30% as test data. The metrics used for evaluating the algorithms were: balanced accuracy which is usually used with imbalanced data set. The balanced accuracy results is shown Table IV and ROC curves are shown in Fig. 5.

TABLE IV. BALANCED ACCURACY RESULTS

Method	balanced accuracy
Logistic Regression	66.26
Decision Trees	69.42
Support Vector Machines	53.22
Neural Networks	67.23
Random Forest	73.82
XGBoost	74.40

XGBoost and Random forest has shown better performance than logistic regression and, support vector machine and neural networks. This can be attributed to the nature of the modelling task and the data used. Most attributes have category values where decision trees based methods are reported to outperform regression based methods. Although with high number of dimensions decision trees based methods tend to affect its performance, in this data these methods continue to outperform linear and non linear classifiers. One downside is performance, however, as the data size increases time increased to obtain the results compared to logistic regression. In addition, further investigations needed to be undertaken to compare the performance with rule based methods which are report to perform well with categorical data.

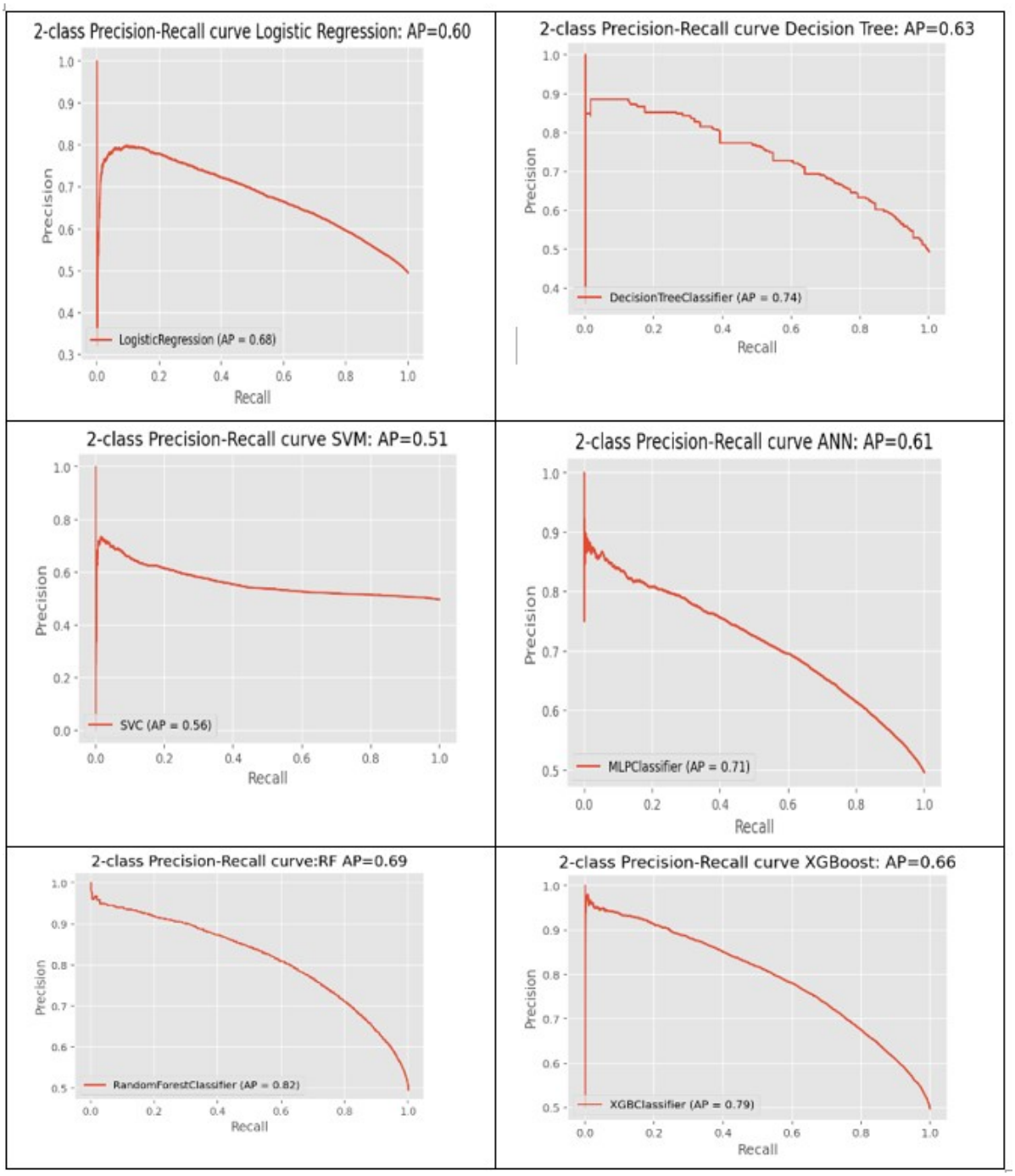

Fig. 5. ROC Curves for LR, SVM, ANN, DT, RF, XGBoost

The feature importance figure can be shown in Fig. 6. The figure shows the top 20 attributes that has effect on severity level. The top attribute is the casualty type which specify whether the casualty was a pedestrian or a passenger. This is followed by vehicle and area attributes. Although 66% of accidents were in 30 miles speed limit , it appear from the feature importance table that speed limit has less effect on the injury type. This insight can help raod traffic authorities to prioritise measures to reduce injury levels.

VI. Conclusion

This paper presented a data analytic framework in which UK traffic accidents data was analysed to established a model for predicting injury severity. The paper used publicly available data from 2005 to 2019 to build prediction models for injury severity level. The paper has combined all attributes from three data sources to analyse 63 attributes and it relation with accident severity. The paper highlighted issues related to data quality and imbalanced data and applied techniques to tackle these issues. The paper compared performance between

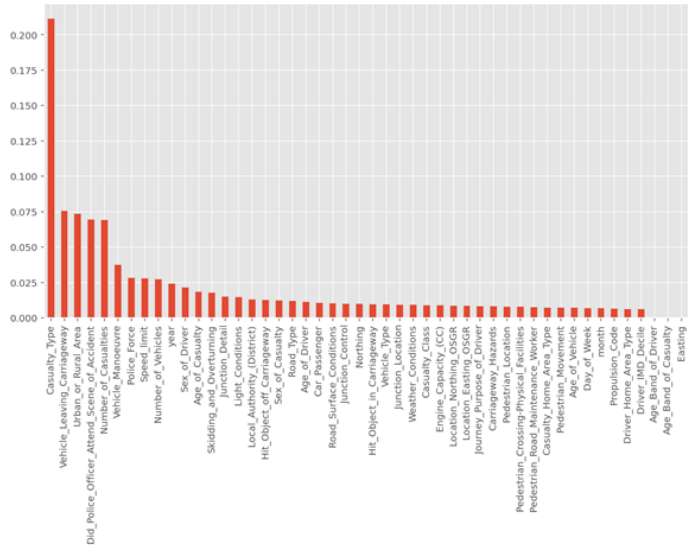

Fig. 6. Feature Importance

different machine learning techniques, XGBoost algorithm was shown to outperform other techniques with higher accuracy rate even with imbalanced data. Further work is suggested to use parallel processing libraries and compare the performance of rule based techniques and decision based techniques.

ACKNOWLEDGMENT

This work is supported by grant number 17-COM-1-01-0007,Deanship of Scientific Research (DSR) of Umm al Qura University, Kingdom of Saudi Arabia. The authors would like to express their gratitude for the support and generous contribution towards pursuing research in this area

REFERENCES

[1] World Health Organization (WHO), *A Road Safety Technical Package*, 2017. [Online]. Available: http://iris.paho.org/xmlui/bitstream/handle/123456789/34980/9789275320013-por.pdf?sequence=1{\&}isAllowed=y

[2] A. Mehdizadeh, M. Cai, Q. Hu, M. A. A. Yazdi, N. Mohabbati-Kalejahi, A. Vinel, S. E. Rigdon, K. C. Davis, and F. M. Megahed, "A review of data analytic applications in road traffic safety. Part 1: Descriptive and predictive modeling," *Sensors (Switzerland)*, vol. 20, no. 4, pp. 1–24, 2020.

[3] Q. Hu, M. Cai, N. Mohabbati-Kalejahi, A. Mehdizadeh, M. A. A. Yazdi, A. Vinel, S. E. Rigdon, K. C. Davis, and F. M. Megahed, "A review of data analytic applications in road traffic safety. Part 2: Prescriptive modeling," *Sensors (Switzerland)*, vol. 20, no. 4, pp. 1–19, 2020.

[4] A. Ziakopoulos and G. Yannis, "A review of spatial approaches in road safety," *Accid. Anal. Prev.*, vol. 135, no. July, p. 105323, 2020. [Online]. Available: https://doi.org/10.1016/j.aap.2019.105323

[5] S. Moosavi, M. H. Samavatian, A. Nandi, S. Parthasarathy, and R. Ramnath, "Short and long-term pattern discovery over large-scale geo-spatiotemporal data," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 2905–2913, 2019.

[6] N. Zagorodnikh, A. Novikov, and A. Yastrebkov, "Algorithm and software for identifying accident-prone road sections," *Transp. Res. Procedia*, vol. 36, pp. 817–825, 2018. [Online]. Available: https://doi.org/10.1016/j.trpro.2018.12.074

[7] D. W. Kononen, C. A. Flannagan, and S. C. Wang, "Identification and validation of a logistic regression model for predicting serious injuries associated with motor vehicle crashes," *Accid. Anal. Prev.*, vol. 43, no. 1, pp. 112–122, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.aap.2010.07.018

[8] D. Delen, R. Sharda, and M. Bessonov, "Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks," *Accid. Anal. Prev.*, vol. 38, no. 3, pp. 434–444, 2006.

[9] A. Naseer, M. K. Nour, and B. Y. Alkazemi, "Towards deep learning based traffic accident analysis," in *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)*, 2020, pp. 0817–0820.

[10] B. Sharma, V. K. Katiyar, and K. Kumar, "Traf fi c Accident Prediction Model Using Support Vector Machines with Gaussian Kernel Á Accident characteristics Á Data mining Á," *Adv. Intell. Syst. Comput.*, vol. 437, pp. 1–10, 2016.

[11] H. Meng, X. Wang, and X. Wang, "Expressway crash prediction based on traffic big data," *ACM Int. Conf. Proceeding Ser.*, pp. 11–16, 2018.

[12] M. Schlögl, R. Stütz, G. Laaha, and M. Melcher, "A comparison of statistical learning methods for deriving determining factors of accident occurrence from an imbalanced high resolution dataset," *Accid. Anal. Prev.*, vol. 127, no. January, pp. 134–149, 2019. [Online]. Available: https://doi.org/10.1016/j.aap.2019.02.008

[13] J. Ma, Y. Ding, J. C. Cheng, Y. Tan, V. J. Gan, and J. Zhang, "Analyzing the Leading Causes of Traffic Fatalities Using XGBoost and Grid-Based Analysis: A City Management Perspective," *IEEE Access*, vol. 7, pp. 148 059–148 072, 2019.

[14] L. G. Cuenca, E. Puertas, N. Aliane, and J. F. Andres, "Traffic Accidents Classification and Injury Severity Prediction," in *2018 3rd IEEE Int. Conf. Intell. Transp. Eng. ICITE 2018*, 2018, pp. 52–57.

[15] Department for Transport, "Road Traffic Statisticsn guidance," pp. 1–13, 2014. [Online]. Available: http://data.dft.gov.uk/gb-traffic-matrix/all-traffic-data-metadata.pdf

[16] B. Burdett, "Improving Accuracy of KABCO Injury Severity Assessment by Law Enforcement," *Univ. Wisconsin-Madison*, 2014.

[17] J. Han, M. Kamber, and J. Pei. (2012) Data mining concepts and techniques, third edition. Waltham, Mass. [Online]. Available: {http://www.amazon.de/Data-Mining-Concepts-Techniques-Management/dp/0123814790/ref=tmm_hrd_title_0?ie=UTF8\&qid=1366039033\&sr=1-1}

[18] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*, 2018.

[19] L. Gautheron, A. Habrard, E. Morvant, and M. Sebban, "Metric learning from imbalanced data," *Proc. - Int. Conf. Tools with Artif. Intell. ICTAI*, vol. 2019-Novem, no. 9, pp. 923–930, 2019.

[20] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 13-17-August-2016, pp. 785–794, 2016.

[21] P. Hájek, M. Holeňa, and J. Rauch, "The GUHA method and its meaning for data mining," *J. Comput. Syst. Sci.*, vol. 76, no. 1, pp. 34–48, 2010.