Fatality Detection:UK Traffic Accident Prediction

Amitha Raghava Raju
NC State University
araghav4@ncsu.edu

Prasanth Yadla
NC State University
pyadla2@ncsu.edu

Ian Menezes
NC State University
ivmeneze@ncsu.edu

Sankalp Gaharwar
NC State University
ssgaharw@ncsu.edu

1 INTRODUCTION AND BACKGROUND

1.1 Problem Statement

According to the World Health Organisation, more than 1.25 million people die each year as a result of road accidents. In 2018, England experienced the highest number of road fatalities, accounting for 85% of road deaths in the UK and subsequently resulted in nearly 2.2 million British Pounds being spent in prevention costs. Moreover, road accidents are usually affected by multiple factors, either in tandem or in isolation. For instance, on an average 22% of all road accidents in UK, are caused by adverse weather conditions. Additionally, less obvious factors such as the road infrastructure and the ambient lighting conditions also end up contributing to road accidents. The factors may be patently quantifiable such as the Speed limit and the Time of the day or they may be subjective factors such as human behavior which flows from the metrics.
We believe that this problem merits extensive data analysis to be able to focus on the factors which affect fatality rates of accidents to the maximum extent, so that risks due to these factors can be mitigated. Our objective in this project is stated as follows:

- Build a predictive machine learning model to detect road accident severity by considering various factors associated with the accident such as weather conditions, time of day, accident location, mode of travel of the affected person, speed limit, junctions types, road type, road infrastructure needs, ambient lighting and other human factors.
- We aim to be able to establish and verify hypothesis about multiple factors, which may not immediately stand out as contributing-factors for fatal accident.

We are confident that we will be able to identify and isolate patterns within the data set and hence, come up with suggestions for reducing the number of fatal accidents.

1.2 Related Work

Ehsaei et al.[2] propose a methodology for data mining in the domain of road accidents. They manage to establish the fact that road accidents are governed by a multitude of factors such as speed limit, time of the day and human behaviour.
Michilaki et al.[3] suggest the incorporation of latent factors such as weather conditions, ambient lighting, road infrastructure etc. in analysing the contributing factors such as frequency and severity of accidents.
Beshah et al.[4] use an adaptive regression tree modelling approach for building decision support systems for handling road traffic accident analysis through classification.
Zhang et al.[10] employ deep learning in detecting the traffic accident from social media data.
Dong et al.[11] propose a deep learning model to explore the complex interactions among roadways, traffic, environmental elements, and traffic crashes. The proposed model includes two modules, an unsupervised feature learning module to identify functional network between the explanatory variables and the feature representations and a supervised fine tuning module to perform traffic crash prediction.
A similar study is performed by Yuan et al.[12] who gauge severity in urban expressway crashes through the use of Machine Learning models.

2 METHOD

2.1 Approach

In the following section, we propose various pre-processing tasks and machine-learning models we aim to build to successful predict accident fatality rate and provide useful recommendations.

- Exploratory Data Analysis
- Data Pre-processing
- Predictive Machine Learning Models

2.1.1 Exploratory Data Analysis: In this section, initial investigation on the dataset is performed to discover patterns.

- **Categorical Context Mapping**: The primary data files consist of numerical data for all attributes. An additional context data file is provided to interpret the numerical data for each attribute in the data files.
- **Uncover underlying structure**: Extensive data analysis is conducted with the help of summary statistics,

graphical representation and spatial detection to identify auto-correlation factors, detect outliers and anomalies.
- **Determine optimal factors**: On performing data analysis we will be able to identify those attributes in the dataset that can be play an important role in predicting accident severity.
- **Test underlying assumptions:** Here we introduce novelty by perform time series forecast using ARIMA model to test the hypothesis that "Accident Fatality rates decline over the years".

2.1.2 Data Pre-processing: It is an integral step as the quality of data and the useful information that can be derived from it directly affects the ability of the model to learn.

- Handling Null Values
- Standardization
- Dataset Sampling
- Handling Categorical Variables
- Principal Component Analysis

2.1.3 Machine Learning Models: We explored the following Machine Learning techniques as a part of the project:

- Logistic Regression
- Random Forest
- XG Boost Classifiers
- K-Nearest Neighbours
- Artificial Neural Network

Additionally, we followed some common guidelines while implementing each of the the ML models. These are:

- **Split Dataset:** The dataset is split into 67% Training and 33% Test dataset. Additionally stratify parameter is introduced in order to ensure appropriate distribute of the target variable in the test set.
- **Hyper-parameter Tuning:** Various Machine Learning models are built and for each model, hyper-parameter tuning and k-fold cross validation is conducted in order to obtain optimum parameters.
- **Predictive Performance**: The model with best set of parameters is built and tested for performance on the test-set.

2.2 Rationale

The Data-set that we were working with was inherently challenging, since the relative frequencies of data objects corresponding to the different severity classes was not balanced. A major problem was that the presence of different number of data points corresponding to each of the severity classes 1, 2 and 3. By using a conventional random sampling approach, we risked wrongly predicting the associated Accident severity classes corresponding to the correlated factors, due to the limitation that the correct Severity class group may not be adequately represented in the training data for our Machine Learning models. Thus, to counter the inherent imbalance in the training data set and ensure fair representation to data points of all class types, we utilised a Stratified Sampling approach for our data.

When it came to choosing our Machine Learning models, we went for models that mirrored the constraints entailed by our data set and ensured better prediction performance. For instance, we chose the Ridge Regression over a conventional Linear Regression approach since we were initially unsure about the degree of influence that each of the multiple independent variables had on our dependent variable(Accident Severity). By opting for a Ridge Regression approach, we were able to ensure that the less influential independent variables were penalized much more heavily than the ones that influenced the accident severity more prominently. This immediately improved the prediction accuracy for our model. The data within our data set was highly correlated and this aspect was efficiently handled by our choice of Ridge Regression as our model. Thus, we were able to overcome the challenges that a conventional Linear Regression model would have faced on our data set.

Along similar lines, we ended up choosing the XGBoost model over Gradient Boosting classifiers. XGBoost did a much better job of reducing bias, as opposed to GBM as the degree of over-fitting was massively reduced due to the more regularized model formulation for XGBoost. This was particularly useful for avoiding any undue reliance on any singular attribute of the data.

In an identical manner, Random Forest ended up performing much better as a Bagging algorithm as opposed to a conventional Decision Tree model. While Decision Trees were increasingly susceptible to over-fitting, the Random Forest classifiers employed a collective average and hence proved to be immune to over fitting. Consequently, their results ended up being much more accurate on our Test Data-Sets

Building an exceptionally accurate prediction model on our data-set proved to be challenging, owing to the the sheer volume of data points and the large number of independent variables that we were trying to correlate to find the resulting Accident severity. To efficiently handle such a complicated input data-set, we found ANNs to be an extremely efficient ML tool.They allowed us to analyze the various input variables better through effective weight-allocation, which mirrored individual attributes' influence on the severity of accidents. By training an ANN across multiple layers across a set of epochs, we were able to obtain an unprecedented degree of accuracy on our test data-set. Other metrics such as the Precision and Recall also ended up improving by a significant margin as opposed to the other conventional Machine Learning models that we used on our data-set.

3 EXPERIMENT

3.1 Dataset(s)

The dataset used in this project is retrieved from Kaggle UK Car Accidents 2005-2015. Alternatively the dataset can be extracted from the UK Department for Transport. The dataset consists of 1.8 million data-points collected from 2005 until 2015 and provided in three separate files, with the primary key Accident Index.

- **Accidents.csv** - is the primary reference file and consists of 1651142 records with 32 attributes.
- **Casualties.csv** - consists of 2216720 records with 15 attributes.
- **Vehicles.csv** - consists 3004425 records with 22 attributes.
- **Context Folder** - consists of .csv files that provide an interpretation of the numerical data in the original files

The attributes can be essentially categorized as

- Seriousness of the incident: Accident Severity, fatalities.
- Road Type and conditions: Type of carriageway, slip road, Automatic signal, Wet/Dry/Snow, rural/urban, poorly lit.
- Vehicle Description: Model, vehicle manoeuvre at the time of accident, driver details.
- Casualty Description: Pedestrian location and movement, car/bus passenger, age demographic.
- Spatial and Temporal Variables: Geographic co-ordinates, recorded time of the incident

The target variable is Accident Severity where **1=fatal, 2=serious, 3=slight**.

3.2 Hypotheses

Hypothesis: Has the UK road accident severity declined over the years?
The objective of the hypothesis is to identify if over the years, essential mandates are put in place to make roads safer to drive, less prone to accidents and reduce the causality count. We aim to test the hypothesis by performing time series forecasting on the dataset. Moreover, from the time series forecast we can identify seasonal patterns, for instance the number of casualties could be low at the beginning of each year. Additionally, the trend plot represents general forecast model over the years.

3.3 Experimental Design

Data Preprocessing

Before performing any analysis, we treated our dataset suitably to ensure that our results do not suffer due to bias/shortcomings

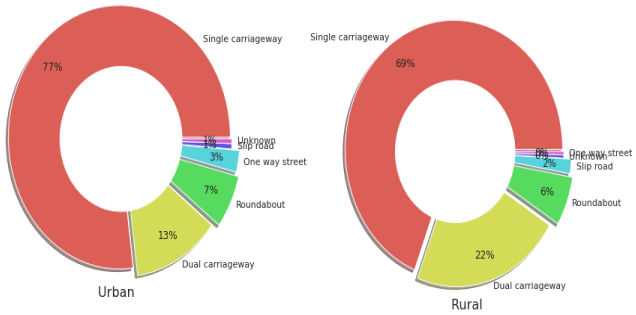

Figure 1: Urban and Rural Traffic Casualties by Road Type

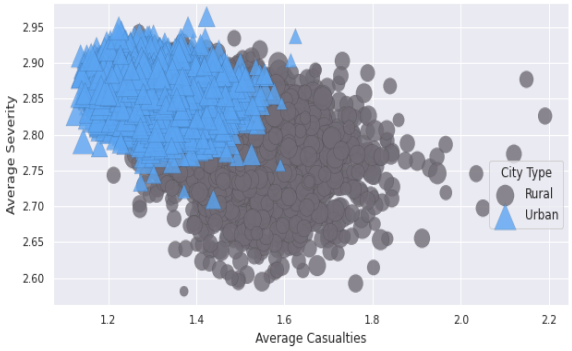

Figure 2: Average Severity vs. Average Casualty

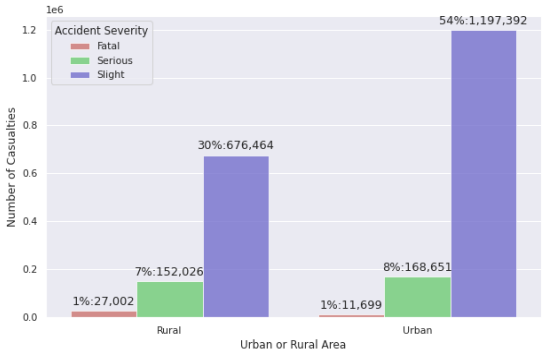

Figure 3: Total Traffic Accidents by Area

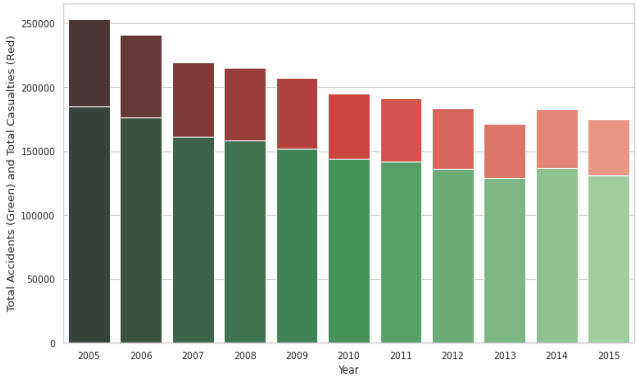

Figure 4: Total Casualty vs Accident Variation over year

that are inherently present within the data. We took the following measures to pre-process the data:

- **Handling Null Values:**: As the dataset is large (1.8 million records), we removed records with null values.
- **Standardization** Since the dataset has attributes of different scale it is essential to standardize the data such that the mean of the values is 0 and the standard deviation is 1.
- **Dataset Sampling**: Since the dataset is large, working on the complete dataset can be expensive considering the memory and time constraints. We performed stratified data sampling and ensured the sample retained the properties of the original dataset. The ratio of the target categories in original dataset was 1:10:68 and reduced to 1:4:16 in the sample.
- **Handling Categorical Variables:** The dataset consists of nominal and ordinal attributes. Additionally, the target variable severity is one-hot encoded when fed into the Artificial Neural Network Model.
- **Principal Component Analysis:** This technique is employed for feature extraction and ensuring the variables are independent of each other. Since the original dataset consists of 30 attributes, the number of planes occupied by the data increases thus adding more sparsity to the data which is difficult to model and visualize.

Data Visualization

Subsequently, we performed visualization experiments on the various data attributes. The rationale behind this was that we aimed to correlate as many disparate and seemingly unrelated attributes as we could as a part of our data analysis tasks. "A picture tells a thousand words", and we were confident that our visualization experiments would help us to gauge patterns and form hypothesis in the most accurate manner.

Temporal Data Analysis

We sought to correlate Accident severity with Temporal attributes of the data such as the Day,Date,Month, Year etc. By looking into Time-oriented data, we were able to identify and isolate patterns and answer questions such as: What time of the day is a severe accident most likely to happen?

We used measures such as AutoRegressive Integrated Moving Average (ARIMA) for tasks such as time- series analysis on our data. Performing Temporal Data Mining on our data set was an essential part of our quest to identify subtle patterns that helped us validate multiple hypotheses.

Spatial Data Design

Two pairs of co-ordinates - Geographic (latitude and longitude) and Projected (Easting and Northing), were available for identifying the location of every accident as part of the data-set. While Projected co-ordinates are more accurate, it's complexity and lack of being widespread fails to have them as inputs to all spatial libraries we were experimenting with. Also, the large data-set forbid us from using certain libraries. Geographic co-ordinates were this utilized.

Among various libraries, GoogleMapPlotter failed to plot all data-points. At any given instant, a maximum of only 29000 data-points could be plotted which doesn't even cover 10% of our data-set. We eventually resorted to MatPlotlib in Python for obtaining a complete heatmap with all the data-points. A base map of UK (.shp file) is needed for the plotted points to fall within the boundary limitations of the nation. This was provided by the website iGIS Map.

To investigate the effectiveness of the police, we relied on the R library - stats19, as it provided a demarcation of the territory controlled by various police departments in Great Britain.

Decision Tree Rule Generation

Given our mammoth-sized data-set and extensive list of attributes, the accident data-set was considered and trimmed, both in terms of width and depth. Nominal attributes such as police force, road types were dropped in favor of ratio attributes - casualties, vehicles involved etc., as the rules were based on data provided to the decision tree classifier and it wouldn't take any string-based nominal data. The entire data-set generated too many rules comprehend so equal-sized samples of our target variable were taken for this experiment.

Neural Network Architecture

Since this is a multi-class classification problem, we have used a simple baseline feed forward neural network with 1 hidden layer with 10 neurons and 1 output layer with 3 neurons to train. We have evaluated using several types of optimizers like Stochastic Gradient Descent, Adam and RMSProp and found that Adam optimizer with learning rate 0.001 and beta_1 0.9 performs better compared to others. Our choice of activation functions for the hidden layer is ReLU and the choice for output layer is softmax. Our loss function is categorical cross entropy.

Random Forest

We used sklearn's RandomForestClassifier with it's default parameters of 100 as number of trees and gini index to train the data.

K nearest neighbours

We have built a K-Nearest Neighbor Model and performed hyperparameter tuning.We considered leaf sizes of 30, 50 and 70 and n-neighbors to be 3, 5 and 7.

Logistic Regression

We build a Logistic Regression model and performed hyper-parameter tuning.We considered the L1 and L2 penalty factor.Additionally, C was set between 1.0 to 2.5 and max_iter was between 100 to 140,

XG Boost

Our XGBoost classifier uses learning rate of 0.1 with estimators as 140 and min child weight as 1, gamma as 0, subsample as 0.8, and colsample_bytree 0.8 and seed as 27 and scale_pos_weight as 1. Additionally, the model was tested for a range of estimators between 100 to 160.

4 RESULTS

4.1 Results

4.1.1 Exploratory Data Analysis:

4.1.1.1 Urban vs Rural Traffic Accident Casualties. The dataset was segregated into rural and urban area type, on performing analysis and from Figure 3, it can be inferred that 64% of casualties occur in Urban area, of which 54% were less severe. From Figure 1, it can be seen that majority of traffic accidents occurred on a single carriage for both urban (78%) and rural (69%) areas. Additionally, dual carriage road type also show high potential for traffic accident occurrence. Moreover, from Figure 2, it can be concluded that one is more likely to get involved in an accident in urban setting that in a rural region, but if an accident occurs in a rural setting it is more probable to be severe.

4.1.1.2 Impact of Weather Condition on Accident Severity. It can be inferred from Figure 16 that 80% of the total accidents occur when the weather is fine with no influence of winds. Since, most accidents occur during pleasant weather conditions, we can conclude that weather conditions do not significantly contribute to traffic accidents.

4.1.1.3 Impact of Lighting Conditions on Accident Severity. Figure 17, depicts that 73% of all the accidents occur during daylight and about 19% of all accidents occur at night when lights are lit. It is clear that infrastructure needs are in place and current lighting conditions do not contribute to Traffic accidents.

4.1.1.4 Spatial Representation of the Accidents. In Figure 13, we've plotted every reported accident to give a better idea of which locations in the UK are more prone to accidents. These are plotted along with their severity, and unsurprisingly some of the populous cities of London, Birmingham and Manchester have more reported cases of Fatal accidents (shown in yellow).

4.1.1.5 Number of Fatalities per Police Force. Given how most of the reported accidents are in England, we obtained the

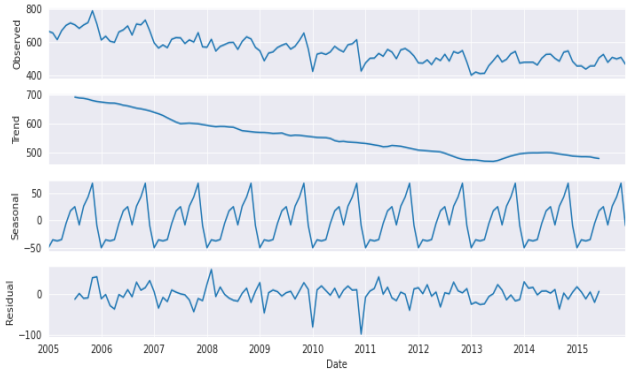

Figure 5: Decomposition of time series

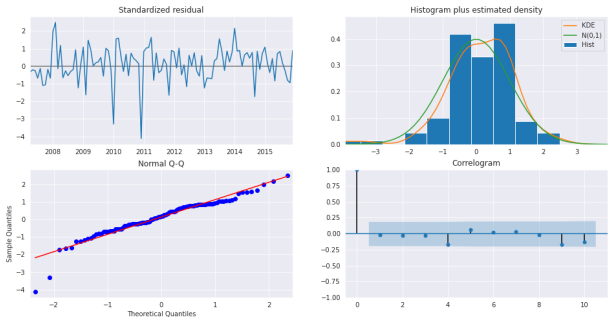

Figure 6: Model Diagnostics to identify Anomalies

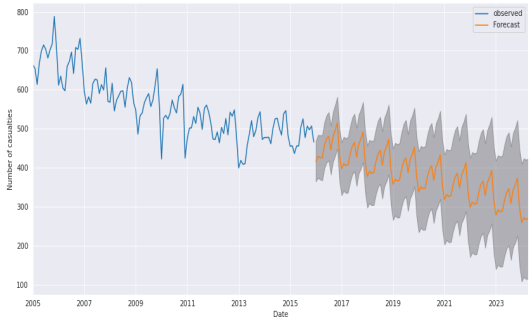

Figure 7: Produce and Visualize Forecast

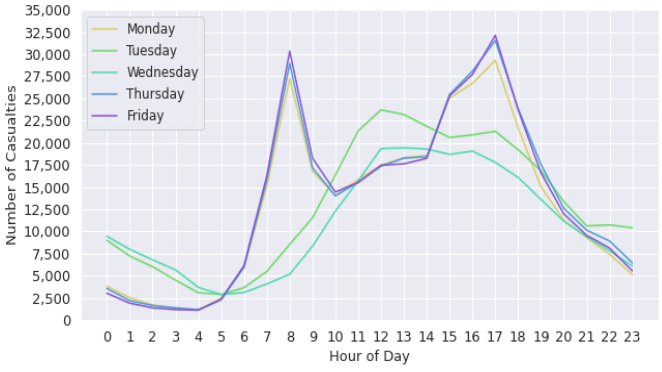

Figure 8: Weekday Traffic Accidents by Time

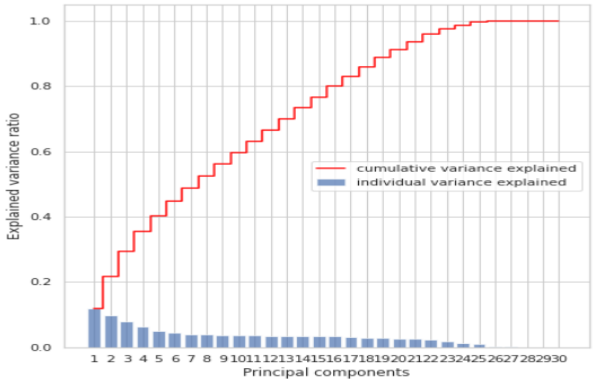

Figure 9: Principal Component Analysis

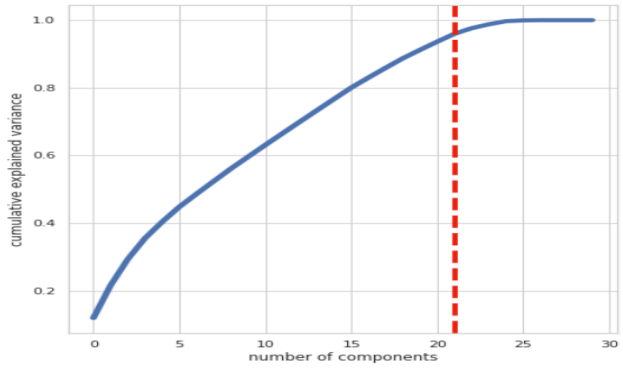

Figure 10: Cumulative Explained Variance - PCA

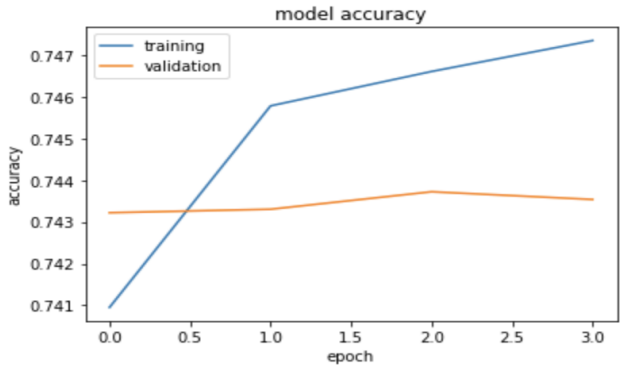

Figure 11: Accuracy Curve - Neural Network

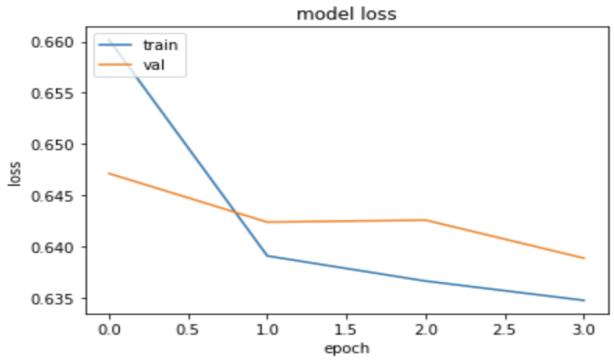

Figure 12: Loss Curve for Neural Network

territorial boundary that each English police force is responsible for. With this we were able to plot the territory-wise distribution of the accidents to view the effectiveness of each police force. Figure 14 plots only Fatal accidents with the region around London reporting the most cases.

4.1.1.6 Impact on Pedestrians at the Accident sites. More often than not, the casualties involve pedestrians walking around at the time of the accidents. Cyclists too are the major fatalities due to the cycle's lower momentum. Figure 15 shows these collateral casualties along with the passengers and their death counts across the nation for a range of vehicle speed-limits reported at the time of the accident. It can be noticed that passengers are most likely to die during a high-speed accident. Whereas for cyclists and especially pedestrians walking around, vehicles with an average speed of 30 proves fatal for them.

4.1.1.7 Accident Severity proportional to Age of the Driver. Figure 18 shows that drivers in the age-band 25 - 45 years are more likely to meet an accident, possibly since a large number of drivers belong to this age group. Additionally, there is more variation among age groups of drivers who encounter accidents that are less severe.

4.1.1.8 Accident Casualties by Days of Week and Time. The dataset is segregated into weekdays and from Figure 8 and 19 we infer that the number of casualties are typically high during commute hours which correspond to 8AM and 5PM respectively. We also plot the variation over the days of the week in Figure 8.

4.1.1.9 Accident vs Casualties co-relation. We can conclude that the total number of accidents and casualties are linearly co-related. They also decrease over time.

4.1.2 Time Series Analysis:

The Autoregressive Integrated Moving Average, ARIMA method is applied for Time series analysis. ARIMA models are denoted with notation ARIMA(p, d, q) that correspond to seasonality, trend and noise in the data. The "grid search" method is applied to obtain optimal parameters.
For the given dataset, the lowest AIC score of 1008.95 was obtained for parameters ARIMA(1, 1, 1)x(1, 1, 1, 12). The model diagnostics presented in Figure 5, suggests that the model residuals are near normal distribution.
The Root Mean Squared Error of the forecast is 18.41.The model diagnostics is presented in Figure 6 and time series forecast is shown in Figure 7.

4.1.3 Machine Learning Models:

The summary of results generated by applying various machine learning models on the dataset is presented in this section:

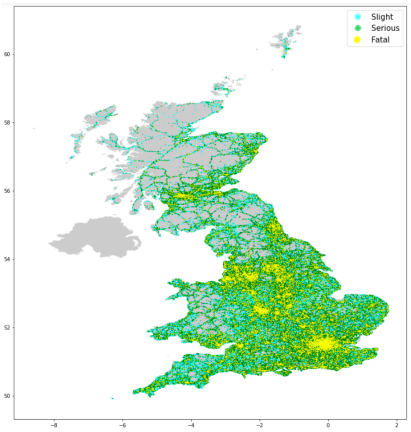

Figure 13: Accidents reported and their Severity

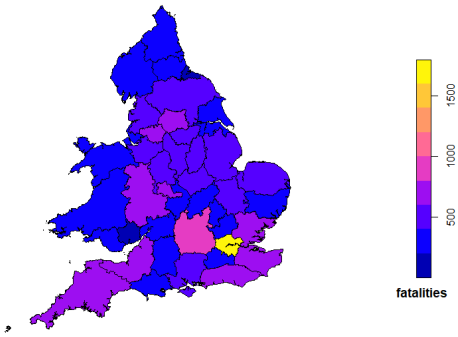

Figure 14: Fatalities per Territorial Police Force

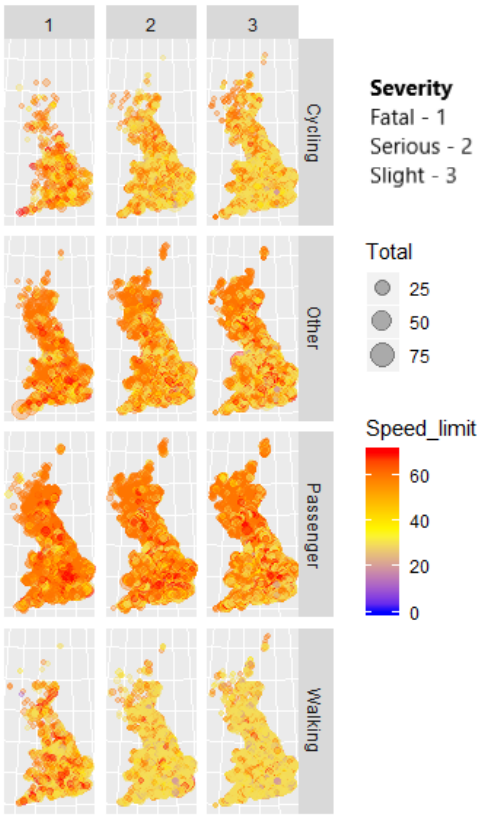

Figure 15: Impact on Pedestrians and Passengers

4.1.3.1 Logistic Regression: We implemented Logistic regression with hyper-parameter tuning on the input features. The best parameter set : C: 2.5, dual : False, max_iter:140, penalty='l2', tol=0.01, solver = 'saga'.

Confusion Matrix :-

		Predicted		
		Fatal	*Serious*	*Slight*
Actual	*Fatal*	92	59	6600
	Serious	90	85	27000
	Slight	60	99	110000

A training accuracy of 76.19% was achieved and a test accuracy of 76.20% was achieved. The classification report is presented in the below table.

Class	Precision	Recall	F1-score	Support
1	0.38	0.01	0.03	6788
2	0.35	0.00	0.01	27151
3	0.76	1.00	0.87	108605

4.1.3.2 K-Nearest Neighbor Classifier: While trying to predict the severity of the accident, the K-nearest neighbors classifier produced correct labels with intermediate accuracy. The results obtained are presented in the confusion matrix below.

Confusion Matrix :-

		Predicted		
		Fatal	*Serious*	*Slight*
Actual	*Fatal*	690	1300	4800
	Serious	930	4000	22000
	Slight	1800	9100	98000

A training accuracy of 79.01% was achieved and a test accuracy of 71.86% was achieved. The classification report is presented in the below table.

Class	Precision	Recall	F1-score	Support
1	0.20	0.01	0.13	6788
2	0.28	0.15	0.19	27151
3	0.78	0.90	0.84	108605

4.1.3.3 Neural Networks: We used a baseline feed-forward neural network model with 1 input layer, 1 hidden layer and 1 output layer. The input layer has 43 normalized features with 3 one-hot-encoded outputs representing severity.

We used around 4 epochs to train the neural network and achieved around 75% accuracy in the test dataset which was split before hand. The model accuracy curve for training data and validation data is shown in Figure. The model loss curve for training data and validation data is shown in Figure.

Confusion Matrix :-

		Predicted		
		Fatal	*Serious*	*Slight*
Actual	*Fatal*	1461	759	9065
	Serious	793	1170	25462
	Slight	621	1035	106679

Class	Precision	Recall	F1-score	Support
1	0.51	0.13	0.21	11285
2	0.39	0.04	0.08	27425
3	0.76	0.98	0.86	108335

A training accuracy of 76% was achieved and a test accuracy of 75% was achieved using Adam optimizer with learning rate = 0.001, beta_1=0.9, beta_2=0.999.

4.1.3.4 Random Forest. The criterion to train the model for Random Forest Classifier was 'gini',hence, the splitting decision implemented in the project was based on the Gini Index. The classifier was able to predict the accident severity reasonably well. The confusion matrix obtained for Random Forest Classifier is presented below.

Confusion Matrix :-

		Predicted		
		Fatal	*Serious*	*Slight*
Actual	*Fatal*	160	440	6200
	Serious	120	850	26000
	Slight	110	1200	110000

A training accuracy of 100% was achieved and a test accuracy of 76.02% was achieved. The classification report is presented in the below table.

Class	Precision	Recall	F1-score	Support
1	0.41	0.02	0.04	6788
2	0.35	0.03	0.06	27151
3	0.77	0.99	0.86	108605

4.1.3.5 XGBoost Classifier: The XGBoost classifier predicted the accident severity with intermediate results. The following are the set of parameter feed into the model for the prediction on the test set. Best Parameter set: learning_rate =0.1, n_estimators=140, max_depth=5, min_child_weight=1, gamma=0, subsample=0.8, colsample_bytree=0.8, nthread=4, scale_pos_weight=1. The confusion matrix obtained is presented below:

Confusion Matrix :-

		Predicted		
		Fatal	*Serious*	*Slight*
Actual	*Fatal*	130	200	6500
	Serious	110	230	27000
	Slight	66	220	110000

A training accuracy of 76.40% was achieved and a test accuracy of 76.24% was achieved. The classification report is presented in the below table.

Class	Precision	Recall	F1-score	Support
1	0.42	0.02	0.04	6788
2	0.36	0.01	0.02	27151
3	0.77	1.00	0.87	108605

4.2 Discussion

Based on the accuracy metrics, we feel that Random Forest performs the best on the data, with a training accuracy of 100% and with a test accuracy of 76%. The other models evaluated all have a training-test accuracy of around 75%.The default value of trees is 100. This makes sense, considering it's a form of ensemble classifier.KNN performs the least when test accuracy is concerned with 71.8%. On performing PCA, we visually interpreting the scree plot, that the elbow cutoff should be at 21 principal components which capture the maximum variance. Based on previous work on using AI techniques for UK, there have been studies on using adaptive regression trees and rule mining classification. Our work presents a novel comparative study of all models and presents metrics to compare performance.Also, to the best of our knowledge, no prior work has been done on evaluating Neural Networks efficacy on the traffic fatality data. Some of the important features we used are Sex of the driver, Age of the driver, Light conditions, Weather conditions, Road Surface conditions, Journey Purpose of Driver, Engine Capacity.

Additionally, the time series analysis is able to forecast a decline in the number of causalities with time. As the forecast proceeds into the future, the confidence interval increases indicating uncertainty with time. Hence, our test hypothesis is satisfied.

Another novel technique we used was extracting rules from our decision tree classifiers. Based on the sample selected, we were able to print them as if-else statements of which an exhaustive list can be seen in the Appendix. Some of them include: For speed-limits upto 35, involving 2 vehicles and at most 2 casualties, the police were at the site in 30% of cases. At least 7 casualties were reported with high severity in all cases where the speed limit is higher than 35. The code is available on Github here [13].

5 CONCLUSION

Based on the time-series data analysis, we can conclude that the number of severe fatalities decrease with time. On evaluating several machine learning models like Logistic Regression, K-Nearest Neighbors, Artificial Neural Networks, Random Forests and XGBoost, we found that Random Forests performs the best when we train it on our chosen dataset.

PROJECT SCHEDULE AND MEETING PLAN

Date	Time	Topic
02/09/2020	6 - 7 P.M	Feature Selection, K-Nearest Neighbours, Logistic Regression (L1, L2 Penalty)
02/12/2020	6 - 7 P.M	Decision Tree, Random Forest, Custom Tree Rules
02/16/2020	6 - 7 P.M	Gradient Boosting Classifier, XG Boost, Evaluation Metrics
02/20/2020	6 - 7 P.M	Neural Network hyper-parameter tuning, Evaluation Metrics
02/21/2020	6 - 7 P.M	Final Report Write-up Division
02/23/2020	6 - 7 P.M	Final Report and Presentation

APPENDIX

Rules Extracted from Decision Trees

Following is a snippet of if-else rules our method was able to generate.

Based on different attributes, it returns the count of records with severity = Fatal, Serious and Slight, in this order.

```
if ( Number of Vehicles <= 1.5 ) {
if ( Number of Casualties <= 1.5 ) {
if ( Speed Limit <= 25.0 ) {
return [[59. 79. 32.]]
} else {
if ( Police Attended Scene of Accident <= 0.0 ) {
return [[1. 0. 0.]]
} else {
return [[3091. 2136. 1231.]]
}
}
} else {
if ( Number of Casualties <= 2.5 ) {
if ( Speed Limit <= 25.0 ) {
return [[14. 3. 2.]]
} else {
return [[602. 233. 108.]]
}
} else {
if ( Police Attended Scene of Accident <= 0.0 ) {
return [[0. 0. 1.]]
} else {
if ( Number of Casualties <= 7.5 ) {
if ( Number of Casualties <= 3.5 ) {
if ( Speed Limit <= 25.0 ) {
return [[3. 1. 0.]]
} else {
return [[145. 44. 13.]]
}
} else {
if ( Number of Casualties <= 6.5 ) {
```

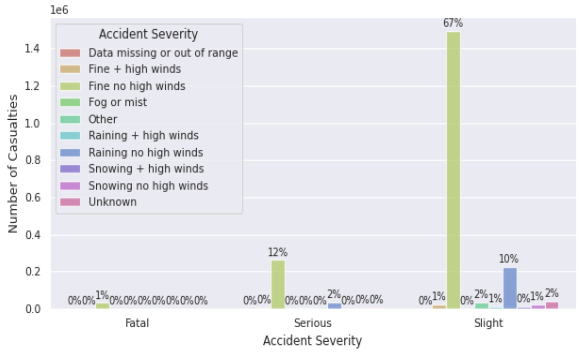

Figure 16: Total Accidents proportional to Weather

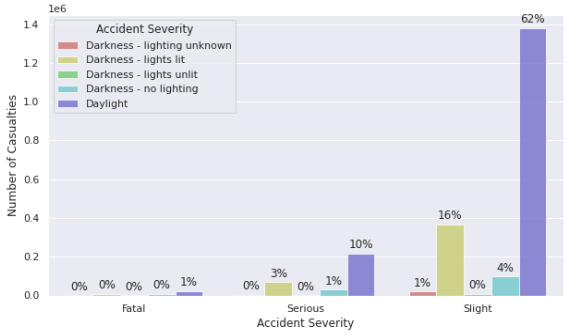

Figure 17: Impact of Lighting on Accidents

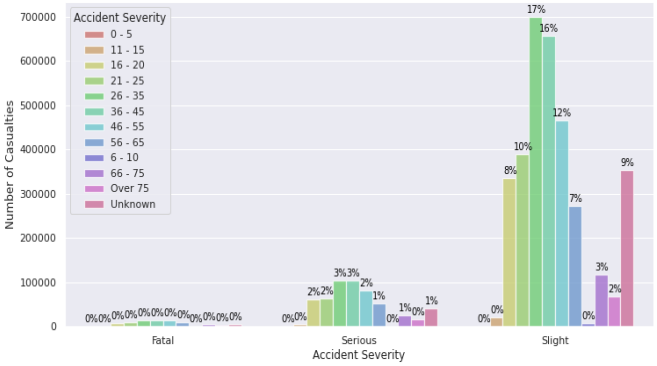

Figure 18: Accidents proportional to Driver Age-Band

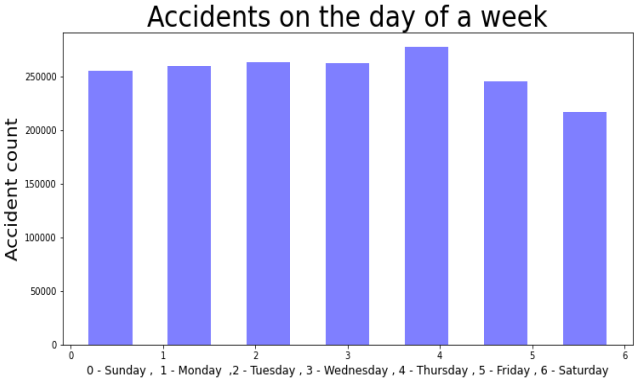

Figure 19: Accidents by Days of Week

```
if ( Number of Casualties <= 5.5 ) {
if ( Speed Limit <= 25.0 ) {
return [[2. 0. 0.]]
} else {
if ( Number of Casualties <= 4.5 ) {
return [[64. 12. 3.]]
} else {
return [[38. 5. 2.]]
}
}
} else {
return [[8. 1. 0.]]
}
} else {
if ( Speed Limit <= 25.0 ) {
return [[1. 0. 0.]]
} else {
return [[3. 2. 0.]]
}
}
}
} else {
if ( Number of Casualties <= 15.0 ) {
if ( Number of Casualties <= 10.5 ) {
if ( Number of Casualties <= 8.5 ) {
return [[0. 1. 0.]]
} else {
if ( Number of Casualties <= 9.5 ) {
return [[1. 0. 0.]]
} else {
return [[1. 1. 0.]]
}
}
} else {
return [[0. 2. 1.]]
}
} else {
return [[1. 0. 0.]]
}
}
}
}
}
```

REFERENCES

[1] UK car accidents 2005-2015,"https://www.kaggle.com/silicon99/dft-accident-data".

[2] A. Ehsaei and H. Evdorides, "Temporal variation of road accident datacaused by road infrastructure".

[3] P. Michalaki, M. Quddus, D. Pitfield, and A. Huetson, "Exploring thefactors affecting motorway accident severity in england using the gen-eralised ordered logistic regression model,"Journal of Safety Research,vol. 55, 11 2015.

[4] T. Beshah, A. Abraham, and C. Grosan, "Rule mining and classificationof road traffic accidents using adaptive regression trees,"InternationalJournal of Simulation: Systems, Science and Technology, vol. 6, 09 2005.

[5] Jean-leah Njoroge PhD, "Significance of Exploratory Data Analysis (EDA)", Nov 13, 2017, Retrieved from "http://www.jeannjoroge.com/significance-of-exploratory-data-anaysis/"

[6] Li. Susan, "An End-to-End Project on Time Series Analysis and Forecasting with Python", Jul 8, 2018, Retrieved from "https://towardsdatascience.com/an-end-to-end-project-on-time-series-analysis-and-forecasting-with-python-4835e6bf050b".

[7] VanderPlas Jake, " Python Data Science Handbook", Retrieved from "https://jakevdp.github.io/PythonDataScienceHandbook/03.11-working-with-time-series.html".

[8] Stewart Ryan,"GeoPandas 101: Plot any data with a latitude and longitude on a map", Oct 31, 2018, Retrieved from "https://towardsdatascience.com/geopandas-101-plot-any-data-with-a-latitude-and-longitude-on-a-map-98e01944b972".

[9] Gras, Robin Mashayekhi, Morteza. (2016). Rule Extraction from Decision Trees Ensembles: New Algorithms based on Heuristic Search and Sparse Group Lasso Methods. International Journal of Information Technology and Decision Making. 10.1142/S0219622017500055.

[10] Zhang, Z., He, Q., Gao, J., Ni, M. (2018). A deep learning approach for detecting traffic accidents from social media data. Transportation research part C: emerging technologies, 86, 580-596.

[11] Dong, C., Shao, C., Li, J., Xiong, Z. (2018). An improved deep learning model for traffic crash prediction. Journal of Advanced Transportation, 2018.

[12] Yuan, Q., Xu, X., Zhao, J., Zeng, Q. (2020). Investigation of injury severity in urban expressway crashes: A case study from Beijing. PLoS one, 15(1), e0227869.

[13] Github Link - Code Repository - "https://github.ncsu.edu/araghav4/ALDA_Fatality_Detection".