JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

BENEFITS OF USING SMOTE ALONG WITH NEURAL NETWORKS FOR AUTOMOBILE FATAL INJURY CLASSIFICATION MODELS

[1]Krishnakanth Gunta, [2]Satya Sai Baba Mudigonda, [3]Phani Krishna Kandala, [4]Rohan Yashraj Gupta, [5]Pallav Kumar Baruah

[1]M.Sc. Data Science and Computing, [2] Adjunct faculty, [3]Visiting faculty, [4]Visiting faculty, [5]Professor
[1]Department of Mathematics and Computer Science,
[1]Sri Sathya Sai Institute of Higher Learning, Puttaparthi, India

Abstract: Automobile crashes are a worldwide problem that result in deaths and injuries, in addition to some direct and indirect expenses. Predicting the risk of accidents can be useful for an insurance company to better estimate claims costs from such accidents. The majority of this research looks at the effects of other vehicles, the environment, and weather on crashes. The data for this study is taken from the Waka Kotahi Crash Analysis System (CAS), which tracks all road accidents recorded by the New Zealand Police. The data is preprocessed by handling missing values and tuning the features post which descriptive analysis is performed. Additionally, to handle data imbalance of fatal injuries, the synthetic minority oversampling technique (SMOTE) has been performed. An artificial neural network (ANN) model is then used to classify a crash as a fatal injury or not. The Artificial neural network model is then applied to both the unbalanced & balanced datasets. Their predictive accuracy, precision, recall, F1 score, and area under the receiver operator characteristic (AUC-ROC) have been compared. We thereby demonstrate that the artificial neural network model with balanced data gives the best accuracy (80%), precision (81%), recall (81%), f1-score (81%), and AUC-ROC (.80). Since artificial neural networks are black-box models, local interpretable model agnostic explanations (LIME) is used to check the model interpretability.

Keywords:
Automobile crashes, Fatal injury, Insurance, Data imbalance, Crash Analysis System, Artificial neural network, synthetic minority oversampling technique, local interpretable model agnostic explanations.

I. INTRODUCTION

Fatal injury classification for crashes is essential for automobile and property insurance to estimate the claims costs. "A fatal injury is one that results in death. It is distinguished from a disability in accident and disability insurance policies, which includes those injuries that prevent the insured from doing his or her regular job but do not result in his or her death."[1] The number of fatal crashes occurring on a daily basis is increasing, posing a serious public health challenge for all authorities involved in preventing them. To avoid this public health disaster, increased awareness, rigorous adherence to traffic laws, and scientific engineering measures are required. According to the World health organization "Every year, the lives of approximately 1.3 million people are cut short as a result of a road traffic crash. Every day, almost 3,700 people are killed globally in crashes involving cars, buses, motorcycles, bicycles, trucks, or pedestrians. More than half of those killed are pedestrians, motorcyclists, or cyclists. From a young age, males are more likely to be involved in road traffic crashes than females. Crash injuries are estimated to be the eighth leading cause of death globally for all age groups and the leading cause of death for children and young people 5–29 years of age. About three-quarters (73%) of all road traffic deaths occur among young males under the age of 25 years who are almost 3 times as likely to be killed in a road traffic crash as young females"[2-4]. Figure-1 shows the deaths across different regions due to crashes for the year 2019. It can be observed that Africa is having a high number of fatalities.

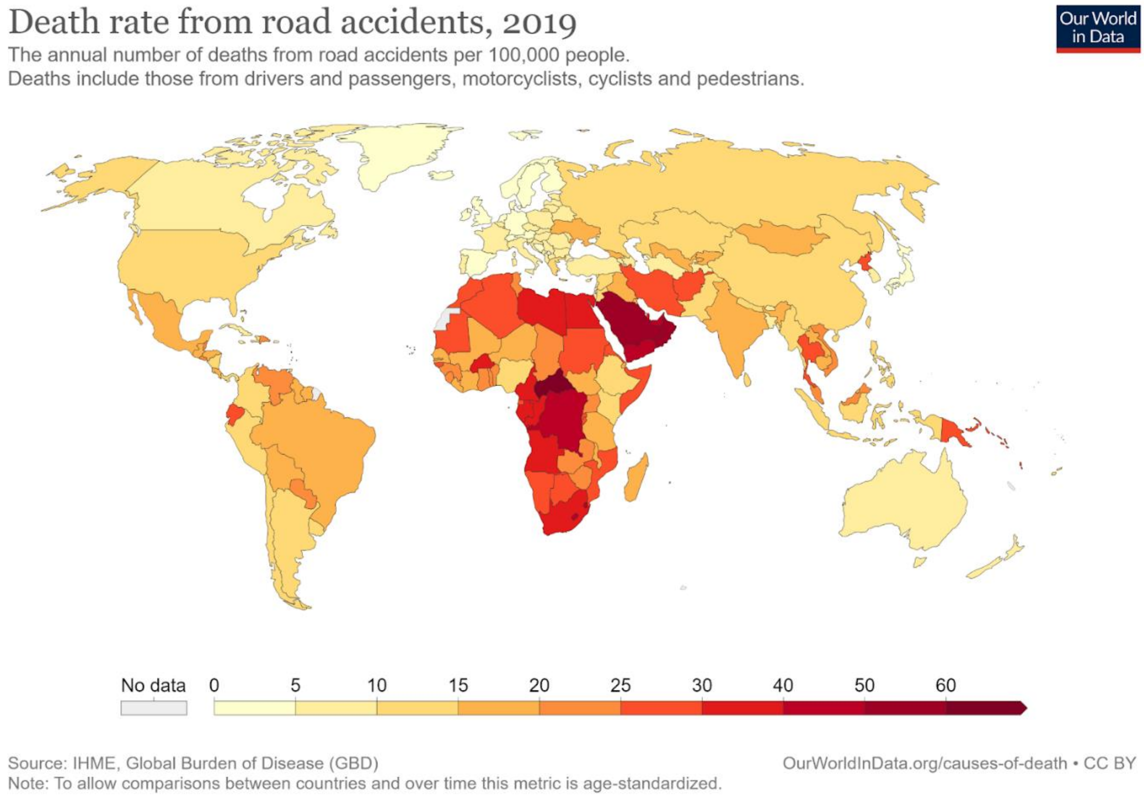


Figure-1: Deaths on the roads across different regions. Image Source [5]

In this study, we are trying to classify records as fatal injury or not, which is critical for the insurance industry to calculate claim costs and premiums. First, the data is analyzed and some insights are driven to reveal the hidden patterns in the data. second, using machine learning techniques, a model for classifying a crash is developed. The data used for this study is from Waka Kotahi Crash Analysis System. We have only 9.1 percent of fatal injury records out of all data points, making the data highly skewed. To handle the imbalance problem Synthetic minority oversampling technique (SMOTE) is used. SMOTE is a "statistical technique for increasing the number of cases in your dataset in a balanced way. The component works by generating new instances from existing minority cases that you supply as input. This implementation of SMOTE does not change the number of majority cases"[6-8]. Finally, the artificial neural network model with unbalanced data and balanced data is compared. Because artificial neural networks are black-box models, we employed local interpretable model agnostic explanations (LIME)[9-10] to validate the model.

This study is represented into nine sections: section 1 is an introduction to the study, section 2 is a literature review, section 3 is the methodology, section 4 is a data description, section 5 is data wrangling, section 6 contains the classification models and results, section 7 is understanding the classification model using LIME, section 8 is the conclusion of the study, and section 9 is the future scope.

II. LITERATURE REVIEW

The insurance domain and mathematical theory were obtained from Arthur Charpentier's "Computational Actuarial Science with R" [11]. The goal of this book is to provide a general overview of computational elements of actuarial science in the R environment. The study of the book's sample dataset adds clarity, and the coding is in R. "Introduction to Linear Regression Analysis by Douglas Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining" [12] was the second most important source of mathematical theory. This book provides a general understanding of regression analysis and the development of statistical models. It also served as a valuable go-to resource when additional knowledge on specific topics was required to move forward with the project work. "Hands-On Machine Learning with Scikit-Learn and TensorFlow"[13] is a book about machine learning. The key source for machine learning is this book. It analyses the machine learning landscape, particularly neural nets, uses the TensorFlow library to create and train neural nets, and provides methodologies for training and scaling deep neural nets.

In the paper "Classification and association rule mining of road collisions for analyzing the fatal severity, a case study"[14] written by Momeni Kho, S., and Pahlavani, P. used three non-parametric classification models based on road collision data to predict fatality severity. The prediction performance of the non-parametric models was evaluated and compared. The risk maps given by each classifier were displayed in order to assess the severity of fatalities across the research area. "Predictive Modeling for Motor Insurance Claims Using Artificial Neural Networks"[15] written by Zuriahati Mohd Yunos, Aida Ali, Siti Mariyam Shamsuddin, Noriszura Ismail, Roselina Salleh, the authors discussed the expected claim frequency and severity that are employed in predictive modeling for vehicle insurance Claims. Two types of claims were considered: third-party property damage and own-damage claims. They created the predictive model using data sets from 2001 to 2003. The backpropagation neural network model is proposed as a technique to model the problem in this research. In the paper "Generalized Linear models in vehicle insurance"[16] written by Silvie kafkova, the author discussed how a GLM is used to analyze a portfolio of automobile insurance data. The key benefit of the method described in this article is that rigid preconditions do not constrain the GLMs. The work provided a classification analysis approach that handles the selection of predictor factors based on a large real-world sample of data from 57410 cars. Based on the AIC comparison, they selected the model with the best estimate of annual claim frequency.

This literature review shows the foundations of fatal classification and claims cost estimation. This work is a comparative study of fatal classification using unbalanced and balanced data and an analysis of the outcome.

III. METHODOLOGY

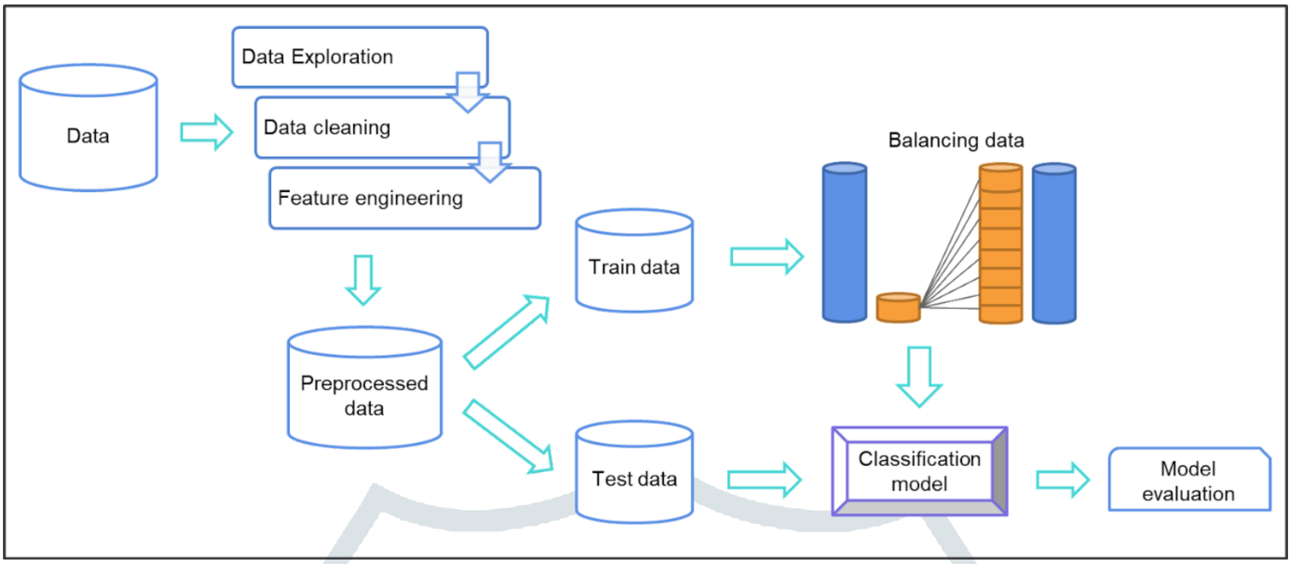


Figure-2: Overview of study

The proposed methodology for fatal classification has shown in figure-2. It has two major parts. First, Data is explored and cleaned, and then feature engineering is performed, resulting in preprocessed data. second, the preprocessed data is divided into two sets: a train set for modeling and a test set for evaluating the model. On the train set, the Synthetic minority oversampling technique is applied to balance the data. The balanced train set is then sent into the artificial neural network. The test set is delivered to the model to evaluate its performance.

IV. DATASET DESCRIPTION

The data for this study came from the Waka Kotahi Crash Analysis System (CAS), which tracks all road accidents recorded by the New Zealand Police. CAS covers crashes on New Zealand roads or anywhere the general public has legal access to a motor vehicle. "CAS combines three key sources of road safety data: crash reports, crash diagrams (since 1996), and roading data, which includes road categorization and traffic patterns. The crash data collection is based on fatal, injury, and non-injury crashes reported to Land Transport NZ by the police"[17]. There are 72 attributes and 758757 records in the dataset. 51 of the 72 attributes are numerical, while the remaining 21 are categorical. We have 9.1 percent (6966) fatal records out of 758757 records or data points, making the data highly skewed. The data is having 80% missing data in 6 attributes, 60% missing data in 23 attributes, and less than 1% missing data in the rest of the attributes. The data dictionary is available on the CAS website[18]. Table 1 contains the description of attributes that are used in the model.

Attribute	Description
Bus	Derived variable to indicate how many buses were involved in the crash.
Bridge	Derived variable to indicate how many times a bridge, tunnel, abutments, and handrails were struck in the crash.
CarStationWagon	Derived variable to indicate how many cars or station wagons were involved in the crash.
Crash severity	This is the severity of a crash. The worst injury sustained in the crash at the time of entry is used to determine this. The following values are possible: fatal, serious, minor, and non-injury.
Fatal count	The total number of fatalities associated with this crash.
Fence	Derived variable to indicate how many times a 'fence' was struck in the crash.
Flat hill	Whether the road is flat or sloped. Possible values include 'Flat or 'Hill'.
House or building	Derived variable to indicate how many times a house, garages, or sheds were struck in the crash
Kerb	Derived variable to indicate how many times a kerb was struck in the crash, that contributed directly to the crash.
Minor injury Count	The total number of minor injuries associated with this crash.

Motorcycle	Derived variable to indicate how many motorcycles were involved in the crash.
Moped	Derived variable to indicate how many mopeds were involved in the crash.
Light	The light at the time & place of the crash. Possible values: 'Bright Sun', 'Overcast', 'Twilight, 'Dark' or 'Unknown'.
Other object	Derived variable to indicate how many times an object was struck in a crash. This variable includes stockpiled materials, garbage cans, fallen poles, and fallen trees, among other things.
OtherVehicleType	Derived variable to indicate how many other vehicles were involved in the crash.
Parked vehicle	Derived variable to indicate how many times a parked or unattended vehicle was struck in the crash.
Post or pole	Derived variable to indicate how many times a post or pole was struck in the crash.
SchoolBus	Derived variable to indicate how many school buses were involved in the crash.
Serious injury count	The total number of Serious injuries associated with this crash.
Slop or food	Derived variable to indicate how many times landslips, washouts, or floods were objects struck in the crash.
Street light	At the time of the crash, the street lighting. Possible values are 'On,' 'Off,' 'None,' or 'Unknown.'
Suv	Derived variable to indicate how many SUVs were involved in the crash.
Taxi	The number of taxis involved in the crash was derived as a variable.
Tree	Derived variable to indicate how many times trees or other growing items were struck in the crash.
Truck	The number of trucks involved in the crash was derived as a variable.
unknownVehicleType	When the vehicle type is unknown, this variable is used to indicate how many vehicles were involved in the crash.
Urban	Derived variable using the 'speed limit' variable. Possible values are 'Urban' or 'Open Road'.
Vehicle	Derived variable to indicate how many times a stationary attended vehicle was struck in the crash. This includes broken down vehicles, workmen's vehicles, taxis, and buses.
water river	The number of times a body of water (such as rivers, lakes, or swamps) was struck in the crash.

Table 1: Data description

V. DATA WRANGLING

5.1. Data Exploration

Data exploration, also known as exploratory data analysis (EDA), is the process through which people examine and comprehend their data using statistical and graphical tools. This step aids in the identification of trends and problems in the dataset, as well as the selection of the model or algorithm to utilize in later steps.

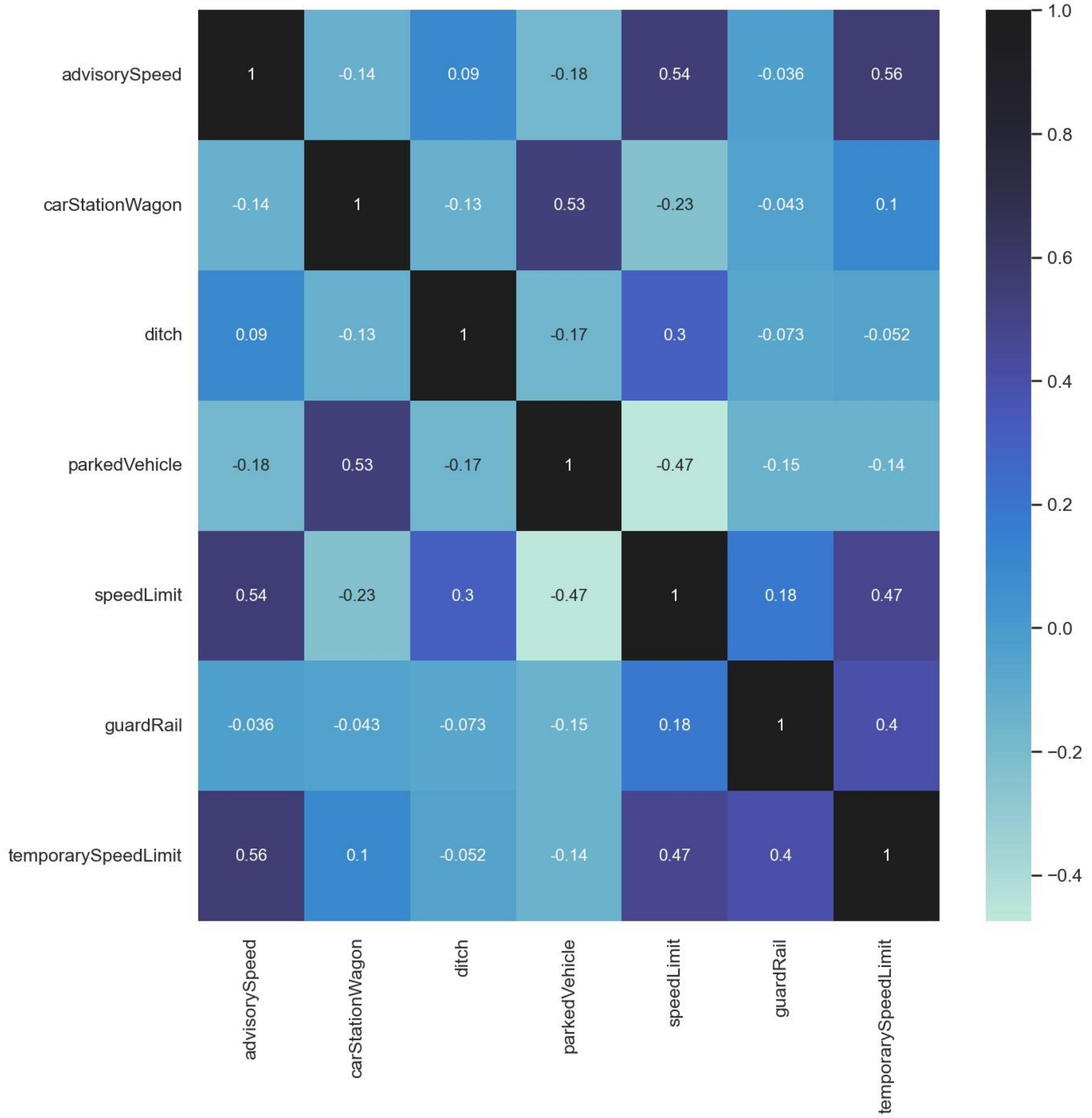


Figure-3: Correlation plot

Figure-3, is the correlation plot of numerical attributes. “Correlation means association - more precisely it is a measure of the extent to which two variables are related. There are three possible results of a correlational study: a positive correlation in which both variables move in the same direction, a negative correlation in which an increase in one variable is associated with a decrease in the other, and no correlation when there is no relationship between two variables”[19]. most of the variables in the data are having low correlation and few variables that are correlated are shown in figure-3.

5.2 Handling missing data

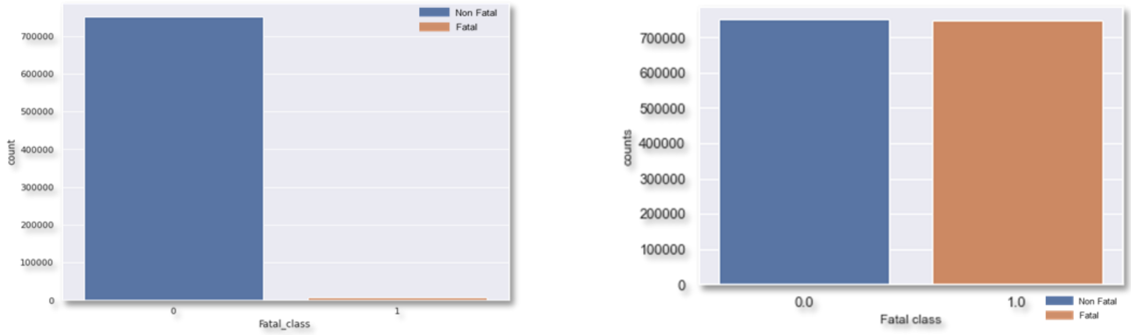
The null values for the majority of attributes are filled with zeros in numerical variables and with none in categorical variables. A function was constructed to fill in a few attributes, such as serious injury count, minor injury count, and fatal count, utilizing the crash severity attribute. Due to the fact that 90% of the data was missing, some of the attributes are removed including junction, advisory speed, etc., Additionally, Unique IDs are removed.

5.3 Feature engineering

A new categorical feature namely fatal class is derived from the fatal count. The fatal class is derived using the following logic if the fatal count is zero which means there is no death, then fatal else non-fatal.

5.4 Handling imbalance

To balance the data, the “Synthetic Minority Oversampling Technique” (SMOTE) [2-4] is used. The fatal classes were generated using the SMOTE approach in such a way that the final fatal to non-fatal ratio was 50:50 (approx.). This is a critical stage in this model since it ensures that there is enough fatal data for the model to train on. Figure-4 depicts this visually



Before smote **After smote**

Figure-4: fatal and non-fatal classes after and before Smote

VI. Fatal injury classification

The workflow for fatal injury classification is depicted in figure-5, the artificial neural network model is applied to unbalanced data (case1) and balanced data (case2). The results are then compared to decide whether balancing data with SMOTE is making an impact or not.

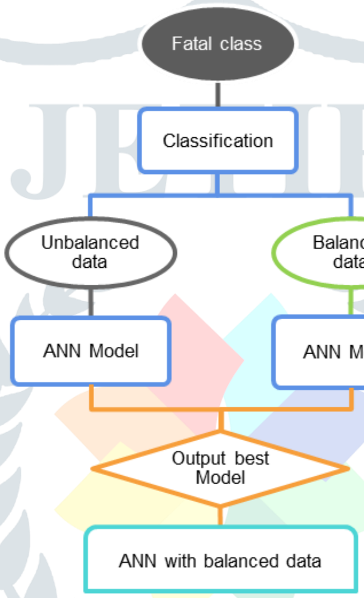


Figure-5: workflow for fatal classification

6.1 Model Parameters

ANN [13][20] with five layers, including hidden and output layers, is used to model this classification problem. At the input layer, data is normalized. RELU is employed as an activation function for all of the layers except for the output layer which employs a sigmoid activation function because it is a binary classification problem. The Binary cross-entropy function is utilized as a loss function for the network. For evaluation of the model, F1 score, accuracy, recall, and precision metrics are used. To find the best threshold AUC-ROC curve is helpful. The dropout strategy is used to avoid overfitting. Table 2 shows all the hyper parameters that are used in the modeling stage.

Parameters	values
Number of layers	5
Batch size	64
Initial learning rate	0.0001
Epochs	70
loss	Binary cross entropy
Optimizer	Adam
Activation functions	RELU and sigmoid

Metrics	Accuracy, F1-score, Recall, precision
Overfitting reduction technique	Dropouts

Table 2: Model parameters

6.2 Model summary

Number of inputs to the network is 31, and the output layer throws a probability that says whether the observation belongs to non-fatal or fatal. After the input layer, the first hidden layer has 512 neurons, the second hidden layer has 256 neurons, the third hidden layer has 128 neurons, the fourth hidden layer has 64 neurons, and the final hidden layer has 32 neurons. Finally, this model has 191,040 parameters, 190977 of which are trainable and 63 of which are not.

6.3 Model loss

Case 1 Case 2

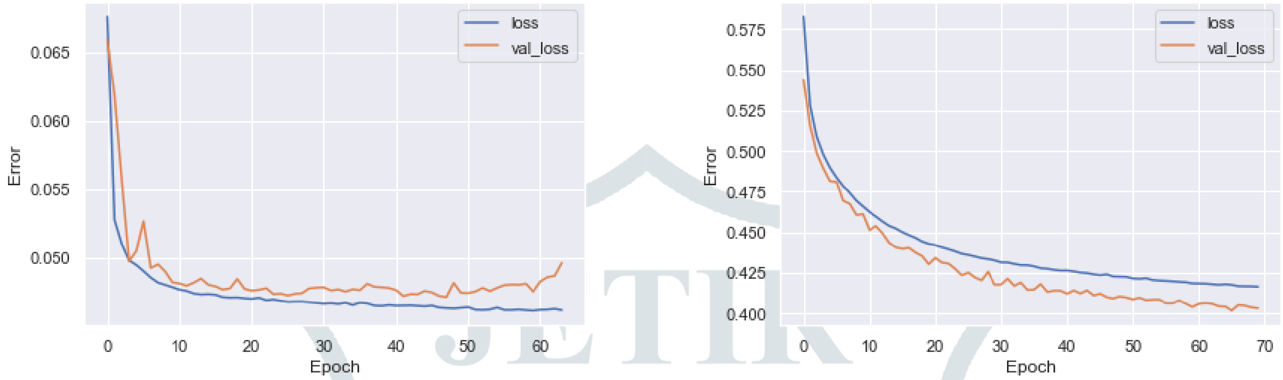
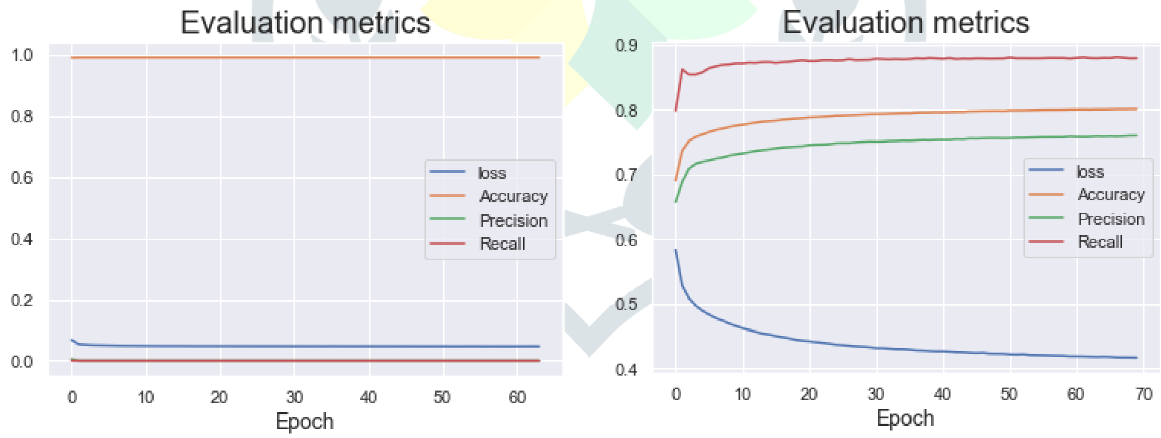


Figure-6: loss vs epoch

The model Figure-6 depicts the loss over the period of epochs for case 1 and case 2. It can be observed that in both cases validation loss and training loss are coming down as the training increases. The loss is converging to 0.4165 after 70 epochs for case 2 whereas the loss for case 1 is 0.046 after 70 epochs. It can be observed that loss is less in case 1 but the predictions are more accurate in case 2.

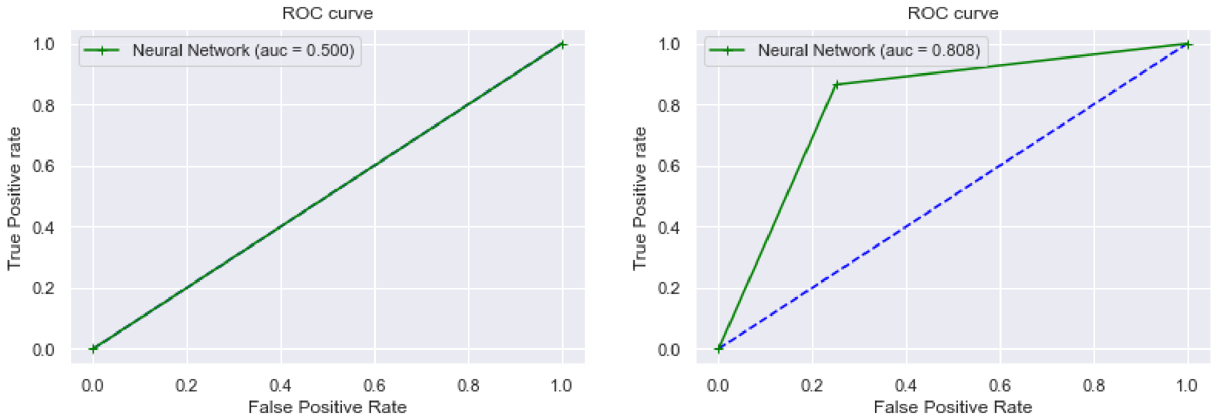
6.4 Model Results



Case 1 Case 2

Figure-7: Evaluation metrics

Figure-7 depicts the evaluation metrics over the period of 70 epochs. As the number of epochs increases, it is observed that accuracy, recall, and precision increase while loss decreases. It is very evident that in case 1 the accuracy is very high and other metrics values are very low in fact zero. Whereas in case 2, it can be observed that accuracy, precision, and recall is much better. The accuracy is high in case 1 because the non-fatal class (0) has more data points than the fatal class (1) which is a data imbalance problem. But in case 2 the problem of imbalance is solved by the synthetic minority oversampling technique (SMOTE).



Case 1 Case 2

Figure-8: Auc-Roc curve

Further, the auc-roc curve is used to determine the best threshold and best model among available models. The Auc-Roc curve is depicted in figure-8. The auc values for case1 and case 2 are 0.50 and 0.80 respectively. It is very evident here that case 2 is giving the best results compared to case 1 because the AUC value of case 2 is far better than case 1. Then the f1 score is examined to evaluate which model performed best. The classification report produced the following outcomes:

	Precision	Recall	F1-score
Non-fatal (0)	0.99	1.00	1.00
Fatal (1)	0	0	0

Table 3: Results of case 1

Table 3 shows the results of the model with unbalanced data. In this case for non-fatal class (0), the precision, recall, and f1 score are 0.99, 1.00, and 1.00, respectively. For fatal class (1), precision, recall, and f1 score are all 0. The overall accuracy is 0.99, which is really high. However, looking at the fatal injury class metrics, it's clear that the model is not operating at all; it's merely categorizing the observation as non-fatal injury owing to an imbalance problem. To declare the model is excellent, one needs the fatal and non-fatal injury classes f1 scores to be close. However, in this case, it is clear that they are quite far apart; the f1 score for non-fatal classes is 1, while the f1 score for fatal classes is 0. All of this leads us to the conclusion that models with unbalanced data are inefficient.

	Precision	Recall	F1-score
Non-fatal (0)	0.85	0.75	0.80
Fatal (1)	0.78	0.87	0.82

Table 4: Results of case 2

Table 4 shows the results of the model with balanced data. In this case for non-fatal class (0), the precision, recall, and f1 score are 0.85, 0.75, and 0.80, respectively. For fatal (1), the precision, recall, and f1 score are 0.78, 0.87, and 0.82, respectively. The accuracy is 0.81, which is acceptable. Overall, the model performs better when it comes to classification. When the data is imbalanced, the accuracy is 0.99, but the classification is poor, but when the data is balanced, the categorization is excellent. However, It is observed that the f1 scores of the fatal and non-fatal classes are quite close, indicating that the model with balanced data is considerably superior[21] to the model with unbalanced data.

VII. UNDERSTANDING THE MODEL PREDICTION USING LIME

The lime, which stands for "local interpretable model agnostic explanations, accepts any machine learning model as input and creates explanations regarding feature contributions in creating a prediction. It assumes that it is a black-box model, which means that it does not understand how models function and creates explanations based on this assumption"[9-10].

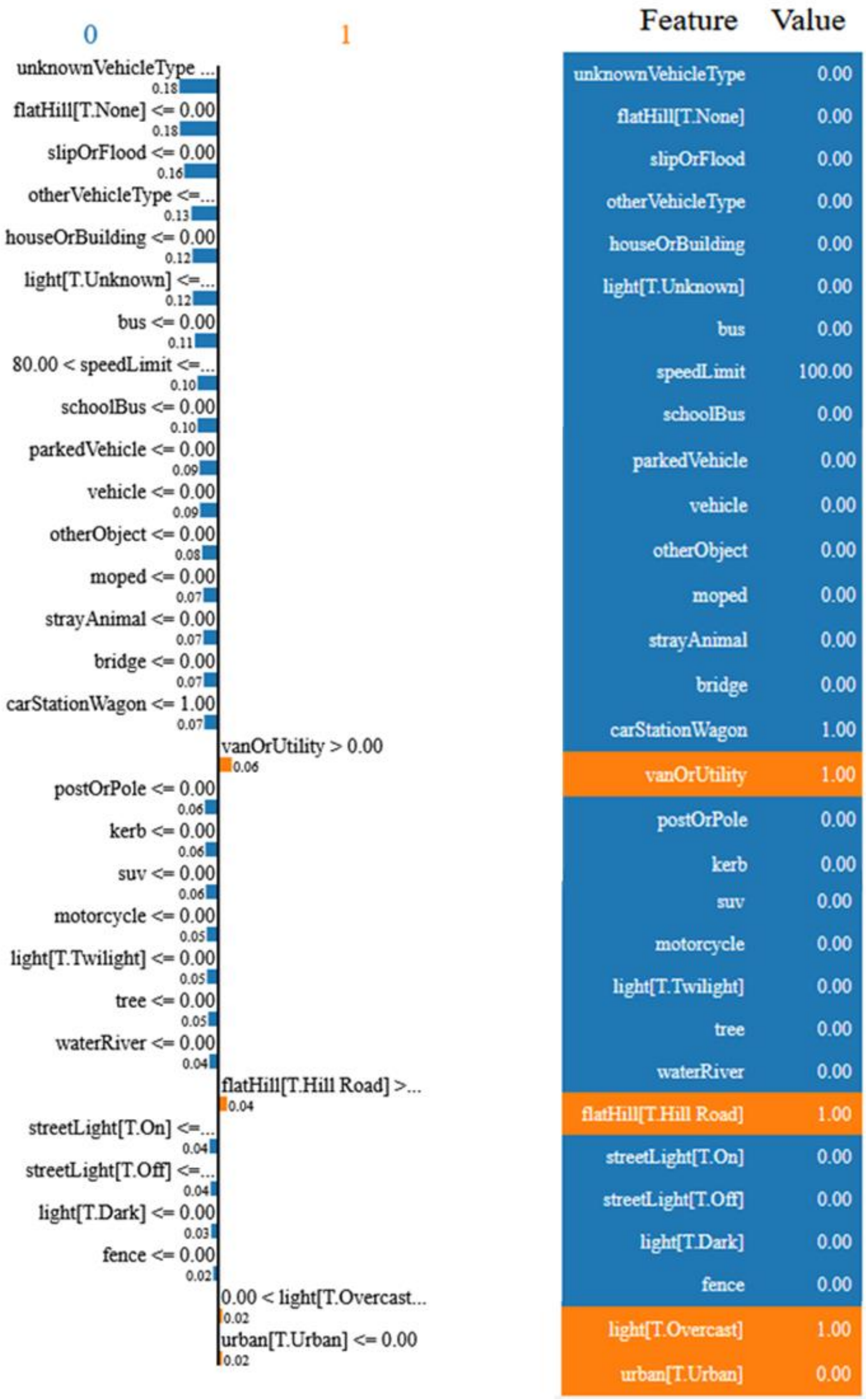


Figure-8: Lime output for Fatal classification

The most effective factors for a given data point is depicted in figure-8. The features on the 0 (non-fatal injury) side, such as unknown vehicle type= 0, other vehicle type= 0, slip or food= 0, and so on, are intended to improve the chance of the non-fatal injury class, that is, if an accident occurs, there will be no deaths. On the 1 (fatal injury) side, features like van or utility > 0, flat hill > 0, light > 0, and urban = 0 (open) are striving to enhance the chance of fatal class, that is, if a crash occurs, there will be a death. The interpretation is very reasonable that if it is an open road the chances of death is more because speed on open roads is high compared to urban roads. Similarly, we can interpret all the variables.

VIII. Conclusion

The classification model's performance on unbalanced and balanced datasets is compared, and it is seen that the model with balanced data performs well and achieves a good degree of f1 score. The classification report for the fatal injury classification model with balanced data using SMOTE is shown in table 5.

Class	Precision	Recall	F1-score
Non-fatal (0)	0.85	0.75	0.80
Fatal (1)	0.78	0.87	0.82
Weighted average	0.81	0.81	0.81

Table 5: Results of the model with balanced data

The classification model with balanced data using SMOTE is giving accurate predictions compared to unbalanced data. So, we can conclude that balancing automobile data with SMOTE and then applying the neural network classification model is helping us to classify the data better. The lime interpretation is reasonable to state our model with balanced data is correct. The lime also provided insights into the variables causing the fatal crashes.

IX. Future scope

Instead of directly estimating fatal counts, one may take the fatal classes using the ANN classification model, and model fatal counts using GLM to get the exact fatal count which will work better because non-fatal have been separated from data. The same methods which are implemented in Python can be implemented in PySpark to handle big data.

X. Acknowledgment

We dedicate this work to Bhagawan Sri Sathya Sai Baba, the Revered Founder Chancellor of Sri Sathya Sai Institute of Higher Learning. We acknowledge expert advice received from Mr. T.V.S Aditya, Actuarial Data Scientist.

References

[1] West's Encyclopedia of American Law, edition 2. S.v. "fatality." Retrieved June 9 2022 from https://legal-dictionary.thefreedictionary.com/fatality

[2] WHO, W. H. O. (2021, June 21). Road traffic injuries. *World Health Organization: WHO*. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

[3] Centers for Disease Control and Prevention (CDC), National Center for Injury Prevention and Control (NCIPC). Web-based Injury Statistics Query and Reporting System (WISQARS). [cited 2020 October 28]. Available from URL: http://www.cdc.gov/injury/wisqars

[4] Chen S, Kuhn M, Prettner K, Bloom DE.*The Lancet Planetary Health*. 2019 Sep 1;3(9):e390–398. The global macroeconomic burden of road injuries: estimates and projections for 166 countries external icon

[5] Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2019 (GBD 2019) Results. Seattle, United States: Institute for Health Metrics and Evaluation (IHME),2021 https://ourworldindata.org/grapher/death-rates-road-incidents?time=2019

[6] *SMOTE - Azure machine learning*. Microsoft Docs. Retrieved June 9, 2022, https://docs.microsoft.com/en-us/azure/machine-learning/component-reference/smote

[7] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

[8] Gupta, Rohan & Mudigonda, Satya & Baruah, Pallav Kumar. (2021). TGANs with Machine Learning Models in Automobile Insurance Fraud Detection and Comparative Study with Other Data Imbalance Techniques. International Journal of Recent Technology and Engineering. 9. 236-244. https://www.researchgate.net/publication/349095432_TGANs_with_Machine_Learning_Models_in_Automobile_Insurance_Fraud_Detection_and_Comparative_Study_with_Other_Data_Imbalance_Techniques

[9] H2O.ai. (2017). Interpretable machine learning using a LIME framework - Kasia Kulma (Ph.D.), data scientist, Aviva [Video]. On YouTube. https://www.youtube.com/watch?v=CY3t11vuuOM

[10] Linardatos, Pantelis & Papastefanopoulos, Vasilis & Kotsiantis, Sotiris. (2020). Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy. 23. 18. 10.3390/e23010018.

[11] Charpentier, A. (2015). Computational actuarial science with R. CRC Press.

[12] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2015). Introduction to linear regression analysis. John Wiley & Sons.

[13] Géron, A. (2019b). Hands-On machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. "O'Reilly Media, Inc."

[14] Momeni Kho, S., Pahlavani, P., & Bigdeli, B. (2021). Classification and association rule mining of road collisions for analyzing the fatal severity, a case study. *Journal of Transport & Health*, *23*, 101278. https://doi.org/10.1016/j.jth.2021.101278

[15] Yunos, Z. M., Ali, A., Shamsuddin, S. M., & Sallehuddin, R. (2016, December 1). Predictive modeling for motor insurance claims using artificial neural networks. https://www.researchgate.net/publication/311921800_Predictive_Modelling_for_Motor_Insurance_Claims_Using_Artificial_Neural_Networks

[16] Kafková, S., & Křivánková, L. (2014). Generalized linear models in vehicle insurance. Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis, 62(2), 383–388. https://doi.org/10.11118/actaun201462020383

[17] Crash Analysis System (CAS) data. (n.d.). Retrieved May 13, 2022, from https://opendata-nzta.opendata.arcgis.com/datasets/NZTA::crash-analysis-system-cas-data-1/explore

[18] *CAS data field descriptions*. (n.d.). Retrieved June 9, 2022, from https://opendata-nzta.opendata.arcgis.com/pages/cas-data-field-descriptions

[19] Mcleod, S. (n.d.). *Correlation definitions, examples & interpretation*. Simply Psychology. Retrieved June 9, 2022, from https://www.simplypsychology.org/correlation.html

[20] Artificial neural network tutorial - Javatpoint. (n.d.). Www.Javatpoint.Com. Retrieved May 13, 2022, from https://www.javatpoint.com/artificial-neural-network

[21] Goutte, Cyril & Gaussier, Eric. (2005). A Probabilistic Interpretation of Precision, Recall, and F-Score, with Implication for Evaluation. Lecture Notes in Computer Science. 3408. 345-359. 10.1007/978-3-540-31865-1_25.