Cilt 11, Sayı 3, Sayfa 79-83, 2022 Volume 11, Issue 3, Page 79-83, 2022

Araştırma Makalesi https://doi.org/10.46810/tdfd.1136432 Research Article

Türk Doğa ve Fen Dergisi
Turkish Journal of Nature and Science
www.dergipark.gov.tr/tdfd

Predicting Traffic Accident Severity Using Machine Learning Techniques

Ali ÇELİK[1*], Onur SEVLİ[2]

[1] Burdur Mehmet Akif Ersoy Üniversitesi, Fen-Edebiyat Fakültesi, Fizik Bölümü, Burdur, Türkiye
[2] Burdur Mehmet Akif Ersoy Üniversitesi, Eğitim Fakültesi, Bilgisayar ve Öğretim Teknolojileri Eğitimi Bölümü, Burdur, Türkiye
Ali ÇELİK ORCID No: 0000-0001-8218-6512
Onur SEVLİ ORCID No: 0000-0002-8933-8395

**Corresponding author: ali.celik@cern.ch*

(Received: 27.06.2022, Accepted: 18.08.2022, Online Publication: 29.09.2022)

Keywords
Machine Learning, Deep Learning, Traffic Accident, Data Mining, Crash Severity

Abstract: Road accidents, harming countries' economies, national assets as well as people's lives, are one of the major problems for countries. Thus, investigating contributing factors to the accidents and developing an accurate accident severity prediction model is critical. Using the traffic accident data collected in Austin, Dallas, and San Antonio city of Texas between 2011 and 2021, the primary contributing factors in crashes are probed and the performance of a deep learning model and five different machine learning techniques, such as Logistic Regression, XGBoost, Random Forest, KNN, and SVM, are investigated. The finding shows that the Logistic Regression algorithm shows the best performance among the others with an accuracy of 88% in classifying accident severity.

Makine Öğrenmesi Tekniklerini Kullanarak Trafik Kazalarının Sonucunu Tahmin Etme

Anahtar Kelimeler
Makine Öğrenmesi, Derin Öğrenme, Trafik Kazaları, Veri Madenciliği, Kaza Şiddeti

Öz: Ülkelerin ekonomilerine, milli varlıklarına zarar verip insanların yaşamlarına sebep olan trafik kazaları, ülkelerin en büyük sorunlarından biridir. Dolayısıyla, kazaların meydana gelmesine katkıda bulunan faktörlerin araştırılması ve doğru bir kaza şiddeti tahmin modelinin geliştirilmesi kritik öneme sahiptir. Bu çalışmada, 2011-2021 yılları arasında Teksas'ın Austin, Dallas ve San Antonio şehirlerinden toplanan trafik kazası verileri kullanılarak, kazalara sebep olan faktörler incelenip, Derin Öğrenme, Lojistik Regresyon, XGBoost, Random Forest, KNN ve SVM gibi 6 farklı makine öğrenme tekniğinin kaza şiddet tahmin performans sonuçları karşılaştırılırdı. Elde edilen bulgular, Lojistik Regresyon algoritmasının kaza şiddetini sınıflandırmada %88 doğrulukla diğerleri arasında en iyi performansı gösterdiğini göstermektedir.

1. INTRODUCTION

Traffic accidents are happening every second worldwide and are causing both people's lives and negative impacts on countries' economies. Although it might be difficult to avoid traffic accidents altogether, reducing the occurrence rate and death rate by taking some pre-measures is possible. As traffic accidents result from road conditions, weather conditions, driver's behavior, or any combinations, machine learning techniques could help model the accidents and classify the severity of the accidents.

Several studies have been conducted on traffic accident classification. Two of them are traffic accident analyses carried out by the same authors with different methods in the studies [1] and [2]. Another study on the classification of traffic accidents is presented in work [3] in Korea. The research [4] probes driver injury severity, which is divided into three classes: no injury, possible injury, and disabling injury, at the signalized intersections in central Florida, while another study reveals the relation between speed limit increase and fatal crash rate in Washington State [5]. A similar study to detect the severity of accidents is conducted with data containing close to 35000 records in Hong Kong using the WEKA tool [6]. Authors of work [7] use the support vector machines to find a pattern in crash injury severity by using collected data from rollover accidents within a

period of two years in New Mexico. Using more than 270000 traffic accident records collected in Michigan, USA, from 2010 to 2016, the study predicts accident severity utilizing machine learning algorithms: Logistic Regression, Random Forest Model, Naïve Bayesian Classifier, AdaBoost Classification Tree [8]. Work [9] presents a case study of traffic accident classification and severity prediction in Spain using data collected over a six-year period (2011–2015) by the Spanish traffic agency. Reference [10] investigated the key factors associated with fatal severity by analyzing 971 accidents in Abu Dhabi in 2014. Authors of [11] investigate prominent factors in traffic accidents in Adana province, Turkey, and classify them according to their injury severity (i.e., fatal, non-fatal). Work [12] conducts a case study in the example of rural roads in Texas. Authors probe crash factor identification and severity prediction in accidents involving teen drivers. The outcomes are evaluated in terms of prediction performance and speed, and XGBoost is concluded to be the best-performing one in both categories.

This paper presents the utilization of deep learning and machine learning algorithms to predict traffic accident severity and identify underlying reasons causing both casualties and damages to national assets. We analyzed data recorded by the Texas Department of Transportation (TxDOT) in Austin, Dallas, and San Antonio cities from 2011 through 2021. Data is accessed through the Crash Records Information System (CRIS) [13]. Deep learning and five different machine learning algorithms, Logistic regression, K-Nearest Neighbors, Support Vector Machine, Random Forest, and XGBoost classifiers, are considered, and their results are compared.

2. MATERIAL AND METHOD

2.1. Dataset and Datamining Process

CRIS database accommodates a variety of traffic accident data from 2011 through 2021 for any city or county in the state of Texas. Dataset can be created with multiple features, such as weather conditions, road surface conditions, light conditions, crash severity, crash time, crash date, location, airbag deployment status, human-related factors contributing to the crash, vehicle-related attributes, and many others. The dataset for this study contains close to 1.1 million accident records. We constructed our data set with attributes; weather conditions, road surface conditions, crash severity, airbag deployment status, human-related factors, and person injury severity. Figure 1 shows the comparison of the number of Fatal/Serious and Other injuries by year from the constructed dataset.

The classification of traffic accident severity is performed in different stages, which involves cleaning, feature selection, and transformation before training each model. We make use of python, deep learning framework Keras [14], machine learning library scikit-learn [15], and pandas [16] for cleaning and training a model. After cleaning the raw data as some of the attributes have many missing or "unknown" data, all the categorical values are encoded into numerical ones (0-1) using the pandas "get_dummies" function for further processing. The cleaned data set is then split into two as train and test set at a rate of 75% and 25%, respectively. The attributes "Speed limit," "Car_Age," and "Vehicle_Damage_Rating" have values varying in different ranges, and it is an issue for machine learning algorithms as they do not contribute equally to the model. Thus, we used Scikit-learn's "StandardScaler" class to scale all the features with a mean of zero and a standard deviation equal to one.

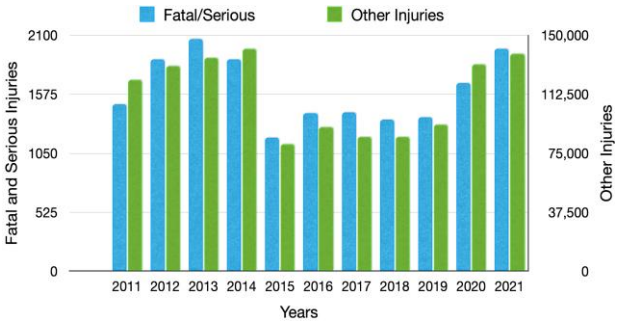

Figure 1. The number of Fatal/Serious and Other injuries by year. Data from San Antonio city is included partially

Classification algorithms usually perform poorly with imbalanced data sets; hence obtained accuracy results are likely to be misleading. As the classes in our data are distributed unevenly, as seen in Figure 2, the down-sampling method is applied to the majority class using python's resampling library [17] to overcome the performance issue. After getting the data ready for building predictive models for person injury severity, the deep learning classification technique and the scikit-learn library are utilized to implement Random Forest, XGBoost, Logistic Regression, K-Nearest Neighbors, and SVM classifiers algorithms.

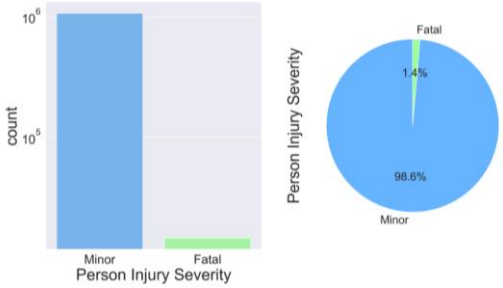

Figure 2. The number of instances corresponding to each class

2.2. Classification Algorithms

2.2.1. Logistic regression (LR)

As opposed to regression in its name, Logistic regression formulated in 1958 by David Cox is a classification model [18]. It is widely used for both binary and multi-class classification problems and achieves excellent performance for linearly separable classes. LR uses the sigmoid function shown below in equation 1, and the y-axis corresponds to the classification's probability.

$$Logistic\ function = \frac{1}{1+e^{-x}} \quad (1)$$

2.2.2. Random forests (RF)

Random forests [19], a supervised learning algorithm, are an ensemble method of decision trees trained on a subset of the training set. Random forest is widely used for classification and regression tasks. Since RF is a combination of learning models, it performs better than any single predictor itself.

2.2.3. XGBoost

XGBoost [20] algorithm stands for "Extreme Gradient Boosting" and is an implementation of a gradient boosting library. XGBoost is known for a better execution speed on a large number of data set as well as utilizing memory resources efficiently.

2.2.4. K-Nearest neighbors (KNN)

K-Nearest Neighbors [21] algorithm is a supervised machine learning algorithm that can be used for both classification and regression problems. KNN is a distance-based algorithm that predicts to which class an unknown data point might belong.

2.2.5. Support vector machine (SVM)

Support Vector Machine [22] is also a supervised learning model used for both classification and regression tasks. The idea behind how it works is to construct an optimum hyperplane in multi-dimensional space to separate classes and predict which classes a new example belongs to. The optimum hyperplane is obtained when the distance from the hyperplane to the closest data points of any class is maximized. This optimum hyperplane is also called a maximum-margin hyperplane.

2.2.6. Deep learning

Deep learning [23] is a subfield of machine learning that uses neural networks to generate data-learning and prediction-capable models. Neural networks are systems of interconnected nodes, known as neurons, that imitate the functioning of the human brain. A neural network is constructed of layers of nodes, with each layer transforming the incoming data in a unique manner. The input layer accepts unprocessed data as input and transmits it to subsequent processing layers. Each layer computes a function on the previous layer's output and then passes its output to the next layer for further processing. Finally, the outputs from all layers are integrated into a single set that represents the final prediction or classification. In order to minimize prediction errors and achieve higher accuracy in a deep learning model, the back-propagation algorithm, altering the weights of connections between nodes, is made use of.

3. RESULTS

In this section, we present and discuss results from different classification techniques as well as analyze the major contributing factors to the accidents. Figure 3 shows contributing factors to traffic crashes on the roads of Austin, Dallas, and San Antonio. As seen from the histogram, driver's inattention (19%), failing speed control (15%), and following too closely (10%) are some of the major contributing factors. Other minor factors are cell/mobile device use, speeding (over limit), impaired visibility, being under drug influence, and failing to yield right of way to pedestrians. While disregarding the stop signs and traffic lights contributes to crashes 6.5%, speed-related crashes are about 18% overall.

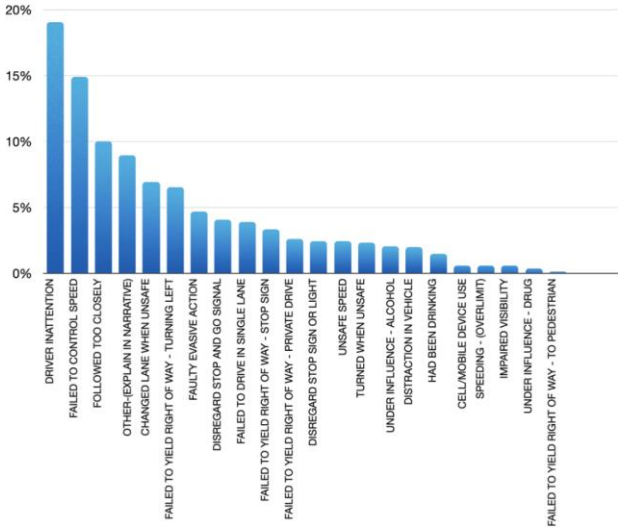

Figure 3. Major contributing factors to crashes in Austin, Dallas, and San Antonio

As the performance measures of the models, we utilized the weighted average for recall, the weighted average for f1-score, the Receiver Operator Characteristic (ROC), and the Area Under the Receiver Operating Characteristic Curve (AUC). Table 1 shows the performance of all classification techniques used in this work. While Figure 4 shows the confusion matrix for the best classifier in this work, Figure 5 and Figure 6 depict ROC comparison and AUC values for each model, respectively. AUC value varies between 0-1, and the bigger AUC indicates how better the model's classification performance is. An AUC value of 1 indicates that the model is excellent, whereas 0.5 or less means the model is poor. An AUC value greater than 0.7 generally indicates that a model has good prediction ability for classification.

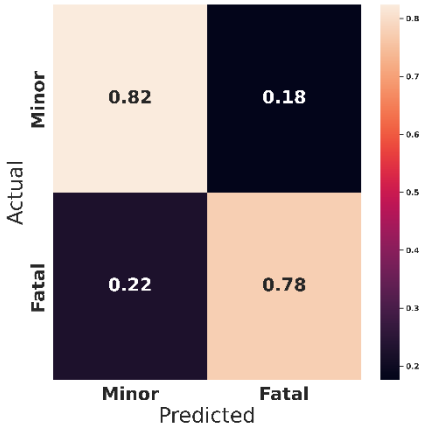

Figure 4. Confusion Matrix for Logistic Regression Model

Although all of the AUC values from different models are fairly close, the LR algorithm outperforms the others with an 88.1% accuracy. XGBoost has a better performance with 87.9% accuracy than SVM with 87.4%. An accuracy of 86.0%, 85.8%, and 80.6% is obtained with deep learning, RF, and KNN, respectively. A comparison of the AUC value for all the classifiers is shown in Figure 6. The result demonstrates that Logistic regression is the best classifier, although the second-best model, XGBoost, performs nearly well.

Table 1. Performance measures with different machine learning techniques

Models	Recall	F1-Score
XGBOOST	81%	88%
KNN	79%	87%
LR	82%	89%
RF	79%	87%
SVM	77%	86%
Deep Learning	81%	88%

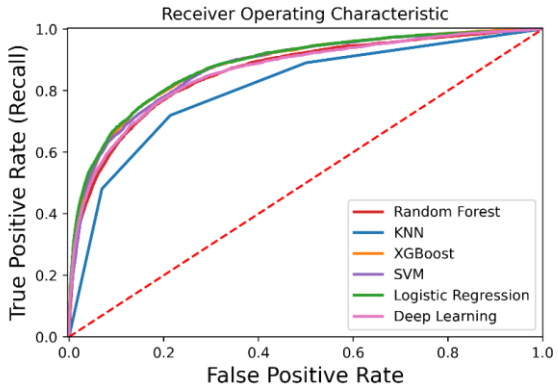

Figure 5. ROC distribution for all the trained models

4. DISCUSSION AND CONCLUSION

The second most populated state in the United States [24], Texas also has more than 22 million registered vehicles [25]. These figures indicate that there are approximately 75 vehicles per 100 people, implying that there might be a significant number of vehicles on the roads on a daily basis. In this study, we investigate the primary contributing factors in crashes and forecast crash severity in three major cities: Austin, Dallas, and San Antonio. According to the findings, driver inattention, failure to manage speed, and following too closely are the top three most common contributing causes in collisions. As one might expect, these three contributing factors might be reasonable findings as there are too many distracting elements around us today, which may cause drivers to lose focus and miss the instructions and traffic signs/warnings on the roads that ultimately trigger accidents. The inability to manage speed and following too close may be the outcome of over-reliance on automobiles and disregarding other considerations since modern vehicles include new features and safety precautions that might mislead the drivers to over-rely on them. Consequently, identifying important factors causing crashes can assist policymakers in developing new road safety policies and engineers in building safer roads. Additionally, predicting accident severity in real-time with our high-accuracy model might assist in taking the necessary prompt action in arriving at the crash scene to reduce the accident's severity. We also probed the performance of different classification techniques in classifying traffic accident severity, grouped into two categories: Fatal/Serious and other injuries. Based on the performance metrics considered in this study, Logistic Regression shows the best performance, with 88.1% accuracy in classifying accident severity (see Figure 6).

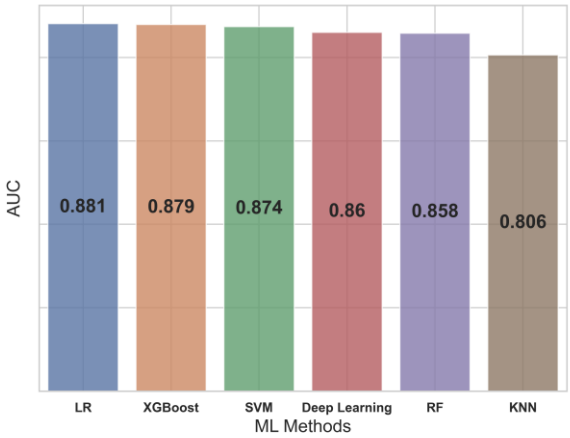

Figure 6. Comparison of AUC values for different trained models

Various research groups conduct similar studies for predicting the severity of accidents. While other works are evaluated with different performance metrics, work [8] adopts AUC as a performance metric to evaluate their findings. For predicting the severity of accidents in the Michigan example, 75.5% accuracy is obtained. However, a better result is presented in this study by analyzing the dataset collected in Texas's three major cities between 2011 and 2021.

REFERENCES

[1] Chong M, Abraham A, Paprzycki M. Traffic accident data mining using machine learning paradigms. In: *Fourth International Conference on Intelligent Systems Design and Applications (ISDA'04), Hungary*. 2004, pp. 415–420.

[2] Chong MM, Abraham A, Paprzycki M. Traffic accident analysis using decision trees and neural networks. *ArXiv Prepr Cs0405050*.
[3] Sohn SY, Lee SH. Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea. *Saf Sci* 2003; 41: 1–14.
[4] Abdelwahab HT, Abdel-Aty MA. Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. *Transp Res Rec* 2001; 1746: 6–13.
[5] Ossiander EM, Cummings P. Freeway speed limits and traffic fatalities in Washington State. *Accid Anal Prev* 2002; 34: 13–18.
[6] Krishnaveni S, Hemalatha M. A perspective analysis of traffic accident using data mining techniques. *Int J Comput Appl* 2011; 23: 40–48.
[7] Chen C, Zhang G, Qian Z, et al. Investigating driver injury severity patterns in rollover crashes using support vector machine models. *Accid Anal Prev* 2016; 90: 128–139.
[8] Comparison of Machine Learning Algorithms for Predicting Traffic Accident Severity | IEEE Conference Publication | IEEE Xplore, https://ieeexplore.ieee.org/abstract/document/8717393 (accessed 4 December 2021).
[9] Traffic Accidents Classification and Injury Severity Prediction | IEEE Conference Publication | IEEE Xplore, https://ieeexplore.ieee.org/document/8492545 (accessed 3 January 2022).
[10] Taamneh M, Alkheder S, Taamneh S. Data-mining techniques for traffic accident modeling and prediction in the United Arab Emirates. *J Transp Saf Secur* 2017; 9: 146–166.
[11] Aci C, Ozden C. Predicting the Severity of Motor Vehicle Accident Injuries in Adana-Turkey Using Machine Learning Methods and Detailed Meteorological Data. *Int J Intell Syst Appl Eng* 2018; 6: 72–79.
[12] Lin C, Wu D, Liu H, et al. Factor Identification and Prediction for Teen Driver Crash Severity Using Machine Learning: A Case Study. *Appl Sci* 2020; 10: 1675.
[13] CRIS Query, https://cris.dot.state.tx.us/public/Query/app/home (accessed 4 December 2021).
[14] Chollet F, others. Keras, https://github.com/fchollet/keras (2015).
[15] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res* 2011; 12: 2825–2830.
[16] McKinney W, others. Data structures for statistical computing in python. In: *Proceedings of the 9th Python in Science Conference*. Austin, TX, 2010, pp. 51–56.
[17] Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *J Mach Learn Res* 2017; 18: 559–563.
[18] Cox DR. The regression analysis of binary sequences. *J R Stat Soc Ser B Methodol* 1958; 20: 215–232.
[19] Ho TK. Random decision forests. In: *Proceedings of 3rd international conference on document analysis and recognition*. IEEE, 1995, pp. 278–282.
[20] XGBoost | Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, https://dl.acm.org/doi/10.1145/2939672.2939785 (accessed 26 June 2022).
[21] Peterson LE. K-nearest neighbor. *Scholarpedia* 2009; 4: 1883.
[22] Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995; 20: 273–297.
[23] LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; 521: 436–444.
[24] US States - Ranked by Population 2022, https://worldpopulationreview.com/states (accessed 18 May 2022).
[25] Motor vehicles in the U.S. - registrations by state. *Statista*, https://www.statista.com/statistics/196505/total-number-of-registered-motor-vehicles-in-the-us-by-state/ (accessed 18 May 2022).