IJTE, Vol. 13, No. 1, Summer 2025, pp. 2039-2054
Research Paper

How Threshold-Moving Technique May Change the Performance of Different Machine Learning Models in Crash Severity Prediction Problems

Alireza Mahpour*[,1], Mostafa Shafaati[2,], Mahmoud Saffarzadeh[3]

Received: 2023/09/24 *Accepted: 2024/12/21*

Abstract

To predict crash severity using Machine Learning (ML) models, dealing with imbalanced classification problems could be inevitable. Threshold-moving can address such problems. Based on a review of the literature, this technique seems to be underutilized. Also, the issue of comparing the performance of different machine learning models in the prediction of crash severity seems to be an open one. Thus, this research focuses on comparing the performance of Random Forest (RF), Logistic Regression (LR) and Naïve Bayes (NB) models by analyzing the trade-off between accuracy and recall for the minority class (both measures change as a result of thresholding). The minority class in our problem is fatal and serious injuries crashes. We use a state-wide crash database from California which contains 143310 records in order to address this issue. Various thresholds are used in the comparison, which are determined by Receiver Operating Characteristic Curves (ROC) and Precision-Recall Curves. There are three thresholds chosen for this study: 0.05, 0.10, and 0.15. Based on the results, the LR with a threshold of 0.1, the RF with 250 trees and the Bernoulli Naive Bayes (BNB) with a threshold of 0.05 are the best models. In addition, LR outperforms the rest of these three models. After threshold moving is employed, even simple models such as the LR can outperform more complicated ones like the RF in this paper, contradicting several previous studies in which the RF is found to be the best model.

Keywords: Crash Severity Prediction, Threshold-moving Technique, Random Forest, Logit, Naïve Bayes Models

* Corresponding author. E-mail: a_mahpour@sbu.ac.ir

[1] Faculty of Civil, Water, and Environmental Engineering, Shahid Beheshti University, Tehran, Iran

[2] PhD, Faculty of Civil, and Environmental Engineering, Tarbiat Modares University, Tehran, Iran

[3] Professor, Faculty of Civil, and Environmental Engineering, Tarbiat Modares University, Tehran, Iran

1. Introduction

Road crashes result in the deaths of about 1.35 million people and severe injuries for 50 million people each year (WHO, 2018). Therefore, reducing crash severity is a global goal (Islam et al., 2023) Data analysis using statistical and Machine Learning (ML) models has been used extensively to attain this goal. It has been commonly considered valuable to analyze crash data with statistic models in the past decades (Eluru et al., 2008; Mannering et al., 2016; Vajari et al., 2020; Liu, 2022; Ahmed et al., 2023; Kabli et al., 2023; Feknssa et al., 2023). However, such techniques are limited by the reliance on significant statistical assumptions. In contrast, for building ML models, there is no need to make statistical assumptions before modeling (Santos et al., 2022). Hence, the use of ML models for crash prediction has gained a great deal of attention over the past few years (Tayarani Yousefabadi et al., 2020).

In analyzing crash severity, dealing with highly imbalanced data could be unavoidable (Jeong et al., 2018; Fiorentini and Losa, 2020). The simple, yet often overlooked method for addressing this issue is threshold-moving (Zhou et al., 2005; He et al., 2013; Fernandez et al., 2016). In addition, multiclass classification problems can be converted into multiple binary-class classifications, and threshold-moving can be used to solve such classification problems. The number of binary classifications will be n-1 if the problem has n classes. According to (Satnos et al., 2022) only a few studies have compared different ML models for predicting crash severity. Literature review will address most of these studies. Therefore, one of the objectives of this research is to provide insights into the potential differences between various classification models using threshold-moving. We emphasize that unlike (Amiri et al., 2020) in which ML models are compared to predict crash frequency, our paper will focus on predicting crash severity using ML models (e,g, Mahpour et al., 2022; Mahpour et al., 2023).

This study utilizes crash data from California for 2012 (Nujjetty et al., 2014). In total, there are 143310 observations in the dataset. The models included in this paper are Logistic Regression, Random Forest (RF) Model, and Nave Bayes (NB) Model.

The rest of the paper is as follows: Literature Review, Methodology, Results and Discussions, Study Limitations, Future Studies, and References.

2. Literature Review

There is a wide use of machine learning models in the prediction and analysis of crash severity (e.g. Haery, et a., 2024; Mahpour and Shafaati, 2024; Tselentis et al., 2023). A comparison between Decision Tree (DT), Nearest Neighbor (NB), and K-Nearest Neighbor (K-NN) classifiers in Ethiopia demonstrated that the KNN outperforms the others (Beshah and Hill, 2020). Among RF, Naive Bayes, AdaBoost, PART Rule, and DT, RF is the best (Krishnaveni and Hemalatha, 2011). It is shown by (Singh et al., 2018) that RF outperforms DT and multinomial logit.

(Wang and Kim, 2019) compares RF and Multinomial Logit models. The researchers concluded that the RF outperforms Multinomial Logit. The data used in this study consists of imbalanced data with three classes in which fatal crashes represent only 0.47% of all observations. According to (Wahab et al., 2019), RF, J48 DeT, Instance-based learning, and multinomial logit are all useful for predicting the severity of motorcycle crashes in Ghana. Results show that RF outperformed the others in accuracy. Gan et al. (2020) uses Multinomial Logit and RF to predict crash severity in highways with and without traffic hazards. The RF appears to be the most accurate model. According to Umer et al. (2020), the RF proved to be the most effective model after trying several different models. According to Chen et al., (2016) the RF outperforms the LR and Regression Tree in performance. A study by Al-Moqri et al., (2020) investigates the

performance of six machine learning models, including the RF. In addition to these models, Multinomial LR, BNB, DT, RF, SVM, and multilayer perceptron are also used. In this case, the RF is the most accurate. Lee et al., (2019) finds that RF is more effective than artificial neural networks and deep neural networks. Bokaba et al., (2022) compares K-NNs with LRs, NBs, AdaBoosts, SVMs and RFs. Five evaluation metrics were used to determine the RF's performance. Yang et al., (2023) finds that RF outperforms SVM and two types of neural networks in prediction accuracy. After comparing DT, K-NN, and NB, Beshah et al., (2013) finds that K-NN outperformes the others. To deal with imbalanced dataset, Fiorentini et al., (2020) uses Random Under Sampling the Majority Class (RUMC) and then the RF, K-NN, and LR. It is found that the true positive rate for the minority class (which consisted of fatal crashes and injury crashes) is the best in K-NN. For predicting the crash severity involving Wheeled Motorized Rickshaw in Pakistan, (Ijaz et al., 2021) finds that among Decision Jungle (DJ), RF, and DT, DJ is the best with its overall accuracy being 83.7 %. (Azhar et al., 2022) compares DT and RF for crashes in which heavy vehicles are involved. It is found that "the RF classifier achieved slightly better accuracy". Chakraborty et al. (2021) finds that for the rarest class of their data, the Deep Neural Net Classifier outperforms the RF. The RF model has consistently outperformed other models in various studies and according to Santos et al. (2020); "…the latter (RF) being, at this point, the most promising machine learning algorithm to develop road traffic crash injury severity prediction models". Therefore, this paper will primarily focus on the RF model. According to the studies reviewed in this section, it seems that threshold-moving has been overlooked when dealing with severity prediction problems. Another issue is that according to (Santos et al., 2020) more studies are needed concerning comparative studies. Therefore, it seems that launching a research in which both threshold-moving and comparing are the main focuses, is necessary. Hence, this paper will focus on not only comparing the RF to other ML models but also investigating the effects of moving thresholds and changing hyper-parameters such as the number of trees in the RF model. For comparison purposes, we have chosen Logit and Naïve Bayes because both of them are simple, easy and fast algorithms which make them preferable models for researchers and practitioners. Also, these models have been used in several studies (AlMamlook et al., 2019; Jeong et al., 2018) and some of above-mentioned studies. These two models will be executed at different thresholds too. Overall, considering the reviewed studies and the introduction section, this paper will seek answers to two important questions:

- Does RF remain the most powerful model after using the threshold-moving technique?
- If the first question is answered in the affirmative, then why? The second question should still be asked if the first question is answered negatively.

3. Data Description

A brief description of the machine learning classification algorithms used in this research is provided in this section. Machine learning classifiers are supervised training algorithms used in classifying datasets that can produce promising results due to their multi-dimensional data processing capability, flexibility in implementation, versatility, and superior predictive capabilities. Three algorithms, including Logistic Regression (LR), Naïve Bayes (NB), and Random Forest (RF) are implemented in this research, using Python analytic platform. The detailed methodology is further discussed in the following paragraphs.

3.1. Logistic Regression Classification

The LR function is commonly used in logistic classification for classifying data. The LR model output is a probability, and it can be used

as a classifier by defining a cut-off point (Mamdoohi, et al., 2016).

3.2. Ensemble Learning

Ensemble learning is an ML technique that combines multiple individual models to improve the overall performance of a prediction task. The individual models in an ensemble can be different types of models, such as DT, neural networks, or SVM. The most common approach in ensemble learning is to use a voting mechanism to combine the predictions of the individual models (Ryu et al., 2010).

3.3. Decision Tree Learning

It is a supervised learning approach which is mostly used as a predictive model. This model is literally a tree with leaves and branches representing class labels and conjunctions of features that make the class labels appear (Studer et al., 2018).

3.4. Random Forest

RF model consists of a lot of trees. Therefore, it is an ensemble learning method. This model usually works based on voting concept. It means that every single tree has a prediction, the model counts all predictions and choses the prediction with the biggest number of trees supporting it, as the final outcome (Mahpour and Kazemi Naeini, 2021).

3.5. Naïve Bayes

This model is based on Bayes theorem in which, the probability of an event can be obtained based on conditions that may be increasing the likelihood of occurring that event (Zhang, 2004). NB algorithms, naively assume that there is a conditional independence between each two features, given the class of a particular observation (Zhang, 2004). If the class variable is y and features are x_i then (Murty et al., 2011);

$$\begin{aligned} & P(y|x_1, \dots, x_n) \\ & \propto P(y) \prod_{i=1}^{n} P(x_i|y) \xrightarrow{yields} \hat{y} \\ & = argmax\, P(y) \prod_{i=1}^{n} P(x_i|y) \end{aligned} \quad (1)$$

If features' likelihood is assumed to be Gaussian, then the type of NB will be Gaussian. Another NB method used in this paper is Bernoulli Naïve Bayes (BNB). This model works well especially when the features are binaries (Schütze et al., 2008; Metsis et al., 2006).

3.6. Evaluation Metrics

In this paper, because of the goals of the research, Precision, Recall or True Positive Rate (TPR), False Positive Rate (FPR), True Positive, False Positive, and Accuracy are used as evaluation metrics. Also, we introduce a new evaluation measure as well. In the following paragraphs, those evaluation metrics are explained. As shown in table 1, For classification problems, and in the case of this research, there is a matrix called the confusion matrix which is made up of four possible scenarios: True Positive (TP), True Negative (TN), False Positive (FP), and False Negatives (FN)

Table 1. Four possible scenarios in the models

Actual \ Predicted	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

Considering table 1, the evaluation metrics needed for this research can be calculated using equation 2 to 5:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$Recall\ (TPR) = \frac{TP}{TP + FN} \quad (3)$$

$$Precision = \frac{TP}{TP + FN} \quad (4)$$

$$FPR = \frac{FP}{FP + TN} \quad (5)$$

Also for the purpose of this research, we introduce a measure which is called δ in this paper and that is:

$$\delta = \begin{cases} \mathbf{1} & if\ (R1 > R0\ and\ (R1 - R0) < 0.2 \\ \mathbf{0} & otherwise \end{cases} \quad (6)$$

where:

R1 = Recall for class 1 (Positive Class), R0 = Recall for class 0 (Negative Class)

3.7. Data Preprocessing

The data used in this research does not suffer from missing values, except for some variables such as "Gender". To address this problem, the missing observations are filled by the most frequent cases. Filling missing values by mode is a common solution (Kuhn and Johnson, 2013). There are 9320 observations with missing "Gender". It is true that removing the observations with missing values is a solution too, but doing so, will result in a relatively high loss of information (in this case, 6 percent of all observations have missing values in the variable "Gender"). Here, the most frequent one is 'male', thus the empty cells are filled by 'male'. However, for some variables, the number of missing values are too high that we have to remove the variable completely. Only two variables in the dataset has this problem: Weather 2 and Contrib2. For Contrib2, the number of missing observations were 139422 and for Weather2, the number of missing observations are 137294.

The dataset suffers heavily from outliers. in order to address the problem of outliers, the quantile method is employed. Subsequently, the skewness of the continuous variables was assessed to ensure that they fell within the range of -1 to +1. It is worth noting that deleting the observations with outliers could be an option. However, deleting them would have caused a great loss of information. For example, for the variable "MEDWID" which represents median width, even if the upper quantile is assumed to be 0.85 and the lower, 0.15, the number of outliers would be 21902 which accounts for 15 percent of all observations. It couldn't be reasonable to remove at least 15 percent of our observations because of the outliers of one variable. Hence, replacing the outliers would be a better choice. After replacing the outliers by upper and lower quantiles, the continuous variables are normalized, and the categorical variables are converted into dummies. To choose the most important variables, a feature selection analysis is launched using an RF model. Considering the goal of this research, a binary imbalanced classification problem should be solved. These two classes are 1 (killed, severe injury, and other visible injury), and 0 (complaint of pain, and non-injury crash). The total number of observations is 143310. It should be noted that merging two or more classes to achieve a better analysis and results has been done in other studies too (Chen et al., 2016; Iranitalab and Khattak, 2017; Li et al., 2012; Tang et al., 2019; Fiorentini et al., 2020). The data is split into train (60 %) and test (40 %) sets randomly. Table 2 demonstrates the size of train and test sets for each class.

Table 2. Train and Test Set

Set	Class	N	Percent
Train	1	10954	12.739
	0	75031	87.261
Test	1	7303	12.739
	0	50021	87.261

3.8. Models' Features

For selecting the most influential variables, a feature selection analysis is carried out, using an RF. The feature selection analysis resulted in the variables defined in table 3. It should be noted that to summarize the article, we have avoided introducing all variables but only the feature-selected ones.

Table 3. Variables and Their Definitions

Variable	Definition	Mean	Min	Max
MILEPOST	Reference point where the crash occurred	20.04	0	186
LANEWID	Calculated average lane width of the roadway segment	40.97	3	89
SEG_LNG	Section length in miles.	0.18	0	6.43
MEDWID	Median width (in feet).	32.67	0	99
AADT	Calculated average AADT	111947	0	354772
LSHLDWID	Width of left shoulder of the roadway segment	4.82	0	26
PAV_WDL	Width of left paved shoulder of the roadway segment	4.53	0	26
SURF_WID	Width of traveled way of the roadway segment	37.12	0	83
RSHLDWID	Right shoulder width.	7.05	0	20
PAV_WIDR	Width of right paved shoulder of the roadway segment	6.80	0	20
DRV_AGE	The age of the driver of the vehicle involved in the crash.	117.41	0	998
CAUSE1_1	Primary collision factor of the crash (Under Influence of Alcohol)	0.07	0	1
CAUSE1_4	Primary collision factor of the crash (Improper Turn)	0.17	0	1
CAUSE1_5	Primary collision factor of the crash (Speeding)	0.47	0	1
CAUSE1_6	Primary collision factor of the crash (Other Violations (Hazardous))	0.20	0	1
ACCTYPE_B	Type of accident that occurred (Sideswipe)	0.18	0	1
ACCTYPE_C	Type of accident that occurred (Rear End)	0.48	0	1
ACCTYPE_E	Type of accident that occurred (Hit Object)	0.20	0	1
ACCTYPE_F	Type of accident that occurred (Overturned)	0.04	0	1
ACCTYPE_G	Type of accident that occurred (Auto-Pedestrian)	0.01	0	1
POP_GRP_5	Population group. Incorporated (50000 To 100000)	0.16	0	1
POP_GRP_6	Population group. Incorporated (100000 To 250000)	0.17	0	1
POP_GRP_7	Population group. Incorporated (Greater Than 250000)	0.27	0	1
POP_GRP_9	Population group. Unincorporated (Rural)	0.27	0	1
WEATHER1_A	Weather conditions when the crash occurred. Clear	0.80	0	1
WEATHER1_B	Weather conditions when the crash occurred. Cloudy	0.16	0	1
LIGHT_A	The type/level of light that existed at the time of the crash. Daylight	0.69	0	1
LIGHT_C	The type/level of light that existed at the time of the crash. Dark - Street Lights	0.15	0	1
LIGHT_D	The type/level of light that existed at the time of the crash. Dark - No Street Lights	0.12	0	1
NUMVEHS_1	Total number of vehicles involved in the crash (1 vehicle)	0.23	0	1
NUMVEHS_2	Total number of vehicles involved in the crash (2 vehicles)	0.60	0	1
DRV_SEX_F	Driver gender (Female)	0.34	0	1
CONTRIB1_A	Violation or factor contributing to the crash. Vehicle Code Violation	0.10	0	1
CONTRIB1_N	Violation or factor contributing to the crash. None Apparent	0.74	0	1

3.9. Selecting Thresholds

Two techniques are employed in this research in order to estimate primary thresholds. The first method uses a ROC curve and the other uses a precision-recall curve. There is a diagnostic plot known as an ROC curve that evaluates a set of probability predictions made by a model on a test dataset (Fawcett, 2004). Various thresholds are used to interpret the TPR and FPR of predictions on the positive (minority) class, and the scores are plotted in a line of increasing thresholds. This graph shows the FPR on the x-

axis and the TPR on the y-axis. It is this plot that is known as a Receiver Operating Characteristic (ROC) curve. From bottom-left to top-right, there is a diagonal line that indicates a model that has no skill (predicts the majority class in every case), and at the top left there is a point that indicates a model that has perfect skills. As a diagnostic tool, the ROC curve can be used to assess the trade-off between different thresholds. According to the ROC curve, 0.11 is the optimal threshold. Fig. 1 illustrates our ROC curve and the location of the best point in our data.

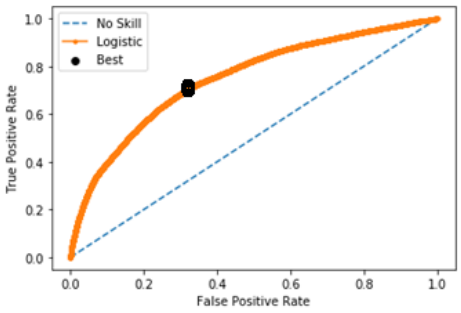

Figure 1. ROC curve for calculating the primary threshold

Precision-recall curves, on the other hand, focus on the performance of a classifier only on the positive (minority class) data (Fernandez et al., 2018). For probability predictions, precision-recall curves are calculated by creating crisp class labels and calculating precision and recall for each threshold. Using a line plot with recall on the x-axis and precision on the y-axis, the thresholds are arranged in ascending order. There is a horizontal line representing a no-skill model, whose precision is the ratio of positive examples in the dataset (for example, TP / (TP + TN)). There is a dot in the upper right corner of the perfect skill classifier indicating its full precision and recall. In Figure 2, we show the precision-recall curve for our data. This curve determined that 0.17 was the best threshold.

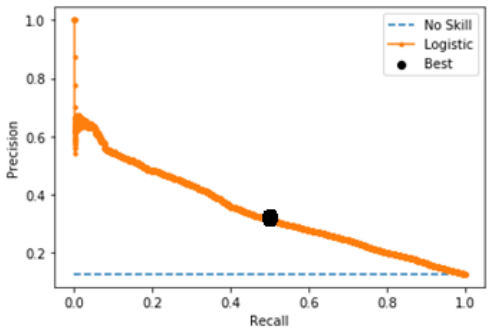

Figure 2. Precision-Recall curve for calculating the primary threshold

ROC curves and precision-recall curves suggest that the best threshold should fall between 0.11 and 0.17 (e.g. Mamdoohi et al., 2013; Mahpour et al., 2020). It is possible that 0.15 would be a suitable threshold, but the objective in this research is to construct a model in which the measure of δ is 1. This objective may not be achieved by choosing 0.15 as the threshold. Consequently, we have also experimented with smaller thresholds. Due to this, three thresholds are used in this research, including 0.15, 0.10, and 0.05.

4. Results and Discussions

4.1. Logit Model Results

A threshold can be determined in several different ways, each of which optimizes a different aspect of the ML models. To begin with, we need to determine which of two classes is most important to us. The prediction of the harshest consequences (fatalities and severe injuries) appears more important when it comes to crash severity prediction (e.g. Ahadi et al., 2018). The severity of these injuries is classified as class 1. Our goal in this paper is to maximize "Recall" for class 1 in this study. Following the construction of each model, its performance is evaluated using three different thresholds: 0.05, 0.10, and 0.15. The evaluation metrics and confusion matrices for all models are presented in table 5 and table 6. According to table 6, as the threshold decreases, the number of TPs increases dramatically, resulting in lower "Recall" values. The cost of this success, however, is a reduction in accuracy and an

increase in FP. Using a threshold of 0.05, the FP is almost three times larger than the TN, which is not desirable. Even though making accurate predictions for Class 1 observations is imperative, what is the point of modeling if we incorrectly predict a large number of Class 0 observations to belong to Class 1? When all observations are regarded as class 1, a perfect Recall can be achieved. As a result, it is not desirable to set the threshold at 0.05. It is better to have FP not exceed TN, therefore, as shown in table 5, the threshold for logit should be greater than 0.05 and less than 0.1. Figure 3 illustrates the results of the confusion matrix. Figure 3, table 5, and table 6 indicate that a good scenario occurs when "Recalls" of both models are close to each other, with the "Recall" of class 1 (positive class) being slightly greater than that of class 0. This is why we have introduced the concept of "δ" as a new measure. Based on the goals of this research, a model with δ of 1 will be considered appropriate. Thus, the model appears to provide a better balance between the costs of having too many false positives and the necessity of increasing true positives. In addition, from another point of view, the TP should be larger than the FP in a good model. In general, the greater the difference between TP and FP, the better the model. Table 6 allows comparison of these two measures. There is also a visual comparison between these two measures in Figure 3.

4.2. GNB and BNB Model Results

Tables 5 and 6 present the results of both GNB and BNB. Figure 3 also illustrates a visual comparison of TP and FP for both models with different thresholds. By comparing GNB and BNB results with those of logit, it becomes evident that GNB's best scenario for class 1 is less desirable than that of LR's. This implies that, in the case of GNB with different thresholds compared with logit, the LR is preferred. Moreover, according to table 5, none of the GNB models presented in this paper have a "δ" equal to 1, indicating that this model did not meet our expectations. As shown in table 7, in a model with a threshold of 0.05, R0 and R1 are close in value. It means that a GNB model with a δ of 1 can be built by modifying the threshold value. The biggest TP for GNB is smaller than that of Logit, however their FPs are almost the same. This is shown in table 6 and figure 3. In the case of BNB, the situation is somewhat different. When the threshold is set at 0.05, class 1 has a higher "recall" than class 0. Additionally, δ is 1 for this model. As can be seen in Figure 1, BNB with a threshold of 0.05 is somewhat similar to the LR model with a threshold of 0.1. Based on these results, BNB and LR perform better than the other two models so far.

4.3. RF Model Results

Several RF models with different number of trees and different threshold values are built, but only the best model is presented in tables 4 and 5. The remaining RFs are presented in table 6. Similarly, to other models, both tables indicate that the variation of threshold is an important factor in determining and increasing recall for RFs. It is expected that when the number of trees is similar, models with lower thresholds will perform better as far as maximizing TP is concerned. There is, however, a trade-off to consider, since increasing TP costs increasing FP.

The relationship between the number of trees and the TP is neither direct nor indirect, as shown in table 6. There can be no guarantee of improved performance when the number of trees exceeds or falls below a certain number. Also, a large number of trees will not necessarily improve performance. Therefore, it may not be beneficial to use a large number of trees in the forest since it will consume a great deal of memory. In other words, an increase in the number of trees is not an efficient solution. As an additional note, all models in table 6 have a "δ" of 0. As a result, we have removed the column representing "δ" from the table. Tables 5 and 6 indicate the best outcome that is achieved when the threshold is equal to 0.1, and the number of trees is equal to 250. Taking into

account table 6 and figure 3, the FP is relatively small (not the smallest), while at the same time, the TP is one of the highest.

Table 4. Evaluating the performance of the LR model with three different thresholds

Model	Threshold	R0	R1	A	P0	P1	δ
Logit	0.05	0.213	0.94	0.306	0.961	0.149	0
	0.10	**0.645**	**0.731**	**0.656**	**0.943**	**0.231**	**1**
	0.15	0.778	0.592	0.755	0.929	0.281	0
GNB	0.05	0.701	0.647	0.694	0.931	0.240	0
	0.10	0.736	0.604	0.719	0.927	0.250	0
	0.15	0.753	0.579	0.731	0.925	0.255	0
BNB	**0.05**	**0.628**	**0.711**	**0.639**	**0.937**	**0.218**	**1**
	0.10	0.714	0.619	0.702	0.928	0.240	0
	0.15	0.749	0.574	0.726	0.923	0.250	0
RF_{250}	0.05	0.441	0.835	0.491	0.948	0.179	0
	0.10	**0.607**	**0.717**	**0.621**	**0.939**	**0.210**	**1**
	0.15	0.723	0.603	0.709	0.926	0.242	0

R0: Recall for class 0, R1: Recall for class 1, A: Accuracy, P0: Precision for class 0, P1: Precision for class 1, δ: It takes 1 if R1>R0 and R1-R0<0.2

Table 5. Confusion matrices for the LR model

Model	Threshold / Class	0.15		0.10		0.05	
		0	**1**	**0**	**1**	**0**	**1**
Logit	**0**	38695	11326	32286	17735	39353	10668
	1	2951	4352	1964	5339	435	6868
GNB	**0**	37680	12341	36808	13213	35050	14971
	1	3075	4228	2894	4409	2578	4725
BNB	**0**	37444	12577	35739	14282	31407	18614
	1	3103	4200	2784	4519	2108	5195
RF_{250}	**0**	36242	13779	30369	19652	22059	27962
	1	2897	4406	2066	5237	1208	6095

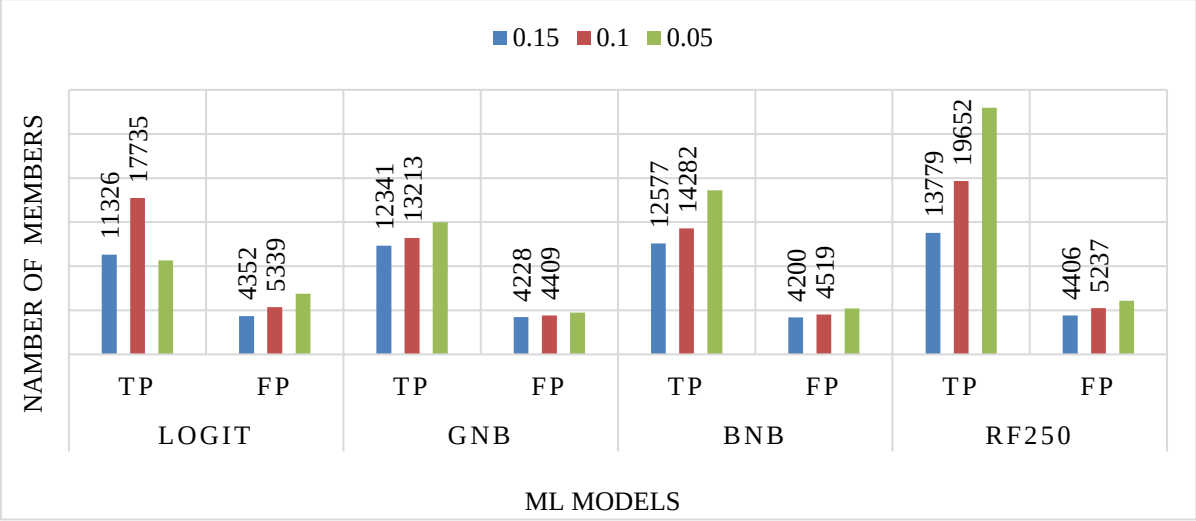

Figure 3. A visual comparison between TP and FP for each model

Table 6. Evaluating the performance of the RF models with different number of trees and different thresholds

No of Trees	Threshold	R0	R1	A	P0	P1
10	0.03	0.501	0.759	0.534	0.934	0.182
20	0.03	0.374	0.854	0.435	0.946	0.166
30	0.03	0.308	0.895	0.383	0.953	0.159
32	0.03	0.296	0.901	0.373	0.953	0.157
33	0.03	0.292	0.903	0.37	0.954	0.157
34	0.03	0.432	0.829	0.482	0.945	0.176
40	0.03	0.399	0.848	0.456	0.947	0.171
250	0.03	0.341	0.888	0.411	0.954	0.164
450	0.03	0.337	0.892	0.408	0.955	0.164
500	0.03	0.333	0.894	0.404	0.956	0.164
600	0.03	0.333	0.893	0.405	0.955	0.164
250	0.05	0.441	0.835	0.491	0.948	0.179
250	0.15	0.723	0.603	0.709	0.926	0.242

In accordance with this research, models where R1 exceeds R0 are the best, but the difference between them should not be too large, because if we increase the number of false positives recklessly, our work will be completely pointless. Thus, it is important to determine an approximate threshold for the difference between R0 and R1. Therefore, we define "δ" as a measure that takes into account this important point. This measure takes a 0 or a 1. When the difference between R1
and R0 is positive and less than 0.2, it takes 1, otherwise, it takes 0. Therefore, the best models are those whose δ is 1. The value 0.2 for "δ" should be considered as an estimate only. It could be the subject of a new paper to determine the best value for this proposed measure. In Table 7, the best ML models are presented.

Table 7. The best models based on the logic of this research

Model	Threshold	R0	R1	A	P0	P1	δ
LR	0.10	0.645	0.731	0.656	0.943	0.231	1
BNB	0.05	0.628	0.711	0.639	0.937	0.218	1
RF250	0.10	0.607	0.717	0.621	0.936	0.210	1

A comparison of the other metrics reveals that Logistic Regression (LR) outperforms the other two good models. While most of the studies discussed in section 2 found the RF to be the best model, in this paper the LR with a threshold of 0.10 is found to be the most effective. However, it is important not to jump to conclusions regarding the superiority of LR over RF. Due to the fact that only three thresholds were tested in this study, we may have obtained different results if other thresholds had been tested, such as 0.02 or 0.7 or etc. Using the simple threshold-moving technique for binary classification problems, it is highly likely that different models will have similar performances when thresholds are set at different values. As a consequence, more work is required concerning the different aspects of threshold-moving techniques.

5. Conclusion

The purpose of this study was to compare the performance of three Machine Learning (ML) models, namely Logistic Regression (LR), Random Forest (RF), and Naive Bayes (NB), in predicting crash severity levels using threshold-moving. RF method was used for feature selection after preprocessing. As a result, the severity levels were reduced to two, simplifying the classification task to a binary one. In order to determine the first thresholds, precision-recall and true positive/false positive rates were

considered. Using threshold-moving, the highly imbalanced nature of the classification problem was addressed. The majority of observations in our dataset belong to class 2, with only 12 percent belonging to class 1. This study demonstrates that class 1 is both a minority class and is of greater importance than class 0. As a result, even at the expense of an increase in the number of FPs in the confusion matrix, an increase in TPs and a decrease in FNs are desirable. As a result of this preference, the accuracy of the model could be decreased. Therefore, having a high accuracy value might not be beneficial in this situation. A high degree of accuracy would be achieved by simply placing all observations into class 0, which would result in an accuracy of up to 88 percent. It is therefore inappropriate to rely on accuracy in such a situation, since the measure of accuracy is totally problematic in this case.

In the case of problems such as this one, there is, however, a question that may be asked: "What is the true strategy?"? There is no doubt that focusing solely on increasing TPs without regard to diminishing FPs is not acceptable, as this will render the process of building models entirely pointless. Therefore, TP must be increased as much as possible, while FP must be decreased as much as possible. Therefore, the efforts may be directed toward obtaining an optimum threshold value in order to achieve the purpose of this paper. As a result, this study developed a strategy in which the first focus is on increasing the value of R1 (Recall for class 1) in a way that R1 is greater than R0 at the end and that the difference between R1 and R0 does not exceed a certain point. In this case, the measure "δ" is applicable. For models that meet these two conditions, "δ" will be 1, while for those that do not meet these conditions, it will be 0. Among the different models, the ones with a "δ" value of 1 are preferred.

In this study, two types of NB models were used: Gaussian and Bernoulli. Additionally, several models of random forest were developed, varying in the number of trees from 10 to 600. For the LR models, recall for minority classes, which are more important to us than majority classes, ranges from 0.592 to 0.940. In the GNB, R1 ranges between 0.503 and 0.647. The BNB shows a range of R1 values between 0.574 and 0.711. In the RF models, the R1 varies from 0.759 to 0.903. According to this paper's logic, three models perform better than the others: Logistic Regression, BNB, and RF with 250 trees, each having a threshold of 0.1, 0.05, and 0.1. Therefore, the "δ" in these models was 1.

Traditionally, comparison of different models using evaluation metrics has not been a difficult task, but once threshold-moving is employed, it appears that any model can achieve a good performance. The question remains unanswered, however, and that is “How should the certain value for the difference between R1 and R0 be calculated? In this study, an approximate value of 0.2 was chosen, however, the value may be changed depending on the nature of the research questions. As a result, new avenues could be opened for future research.

6. Study Limitations

The main limitation of this paper is the number of models used in it. Utilizing different models, and reporting different models, and their results in one paper is almost impossible. Therefore, we had to limit our options, and choose only three main models. Overcoming this limitation by using other models can be considered in future studies.

7. Future Studies

For future research, it may be of interest to determine this value based on minimizing the cost of fatalities and possible countermeasures to reduce the risk of crashes. It is likely that future research will need to focus on determining the costs as another important issue. The development of mathematical and statistical methods that can be used to convert those costs to the best threshold could also be a

concern. Moreover, future studies can explore models other than Logit, Nave Bayes, and Random Forest. Additionally, other threshold values could be tested. Future research could also focus on finding the best threshold for δ.

8. References

-Laskaris, R. (2015). Artificial Intelligence: a modern approach.

-Mokhtarimousavi, S., Anderson, J. C., Azizinamini, A., & Hadi, M. (2020). Factors affecting injury severity in vehicle-pedestrian crashes: A day-of-week analysis using random parameter ordered response models and Artificial Neural Networks. International journal of transportation science and technology, 9(2), 100-115.

-Zhang, H. (2004). The optimality of naive Bayes. Aa, 1(2), 3.

-Ahadi, M. R., Mahpour, A. R., & Taraghi, V. (2018). A Combined Fuzzy Logic and Analytical Hierarchy Process Method for Optimal Selection and Locating of Pedestrian Crosswalks. Journal of Optimization in Industrial Engineering, 11(2), 79-89.

-Ahmed, S. S., Corman, F., & Anastasopoulos, P. C. (2023). Accounting for unobserved heterogeneity and spatial instability in the analysis of crash injury-severity at highway-rail grade crossings: A random parameters with heterogeneity in the means and variances approach. Analytic methods in accident research, 37, 100250.

-AlMamlook, R. E., Kwayu, K. M., Alkasisbeh, M. R., & Frefer, A. A. (2019, April). Comparison of machine learning algorithms for predicting traffic accident severity. In 2019 IEEE Jordan international joint conference on electrical engineering and information technology (JEEIT) (pp. 272-276). IEEE.

-Al-Moqri, T., Haijun, X., Namahoro, J. P., Alfalahi, E. N., & Alwesabi, I. (2020). Exploiting Machine Learning Algorithms for Predicting Crash Injury Severity in Yemen: Hospital Case Study. Appl. Comput. Math, 9(5), 155-164.

-Amiri, A. M., Nadimi, N., & Yousefian, A. (2020). Comparing the efficiency of different computation intelligence techniques in predicting accident frequency. IATSS research, 44(4), 285-292.

-Azhar, A., Ariff, N. M., Bakar, M. A. A., & Roslan, A. (2022). Classification of driver injury severity for accidents involving heavy vehicles with decision tree and random forest. Sustainability, 14(7), 4101.

-Beshah, T., & Hill, S. (2010, March). Mining road traffic accident data to improve safety: role of road-related factors on accident severity in Ethiopia. In 2010 AAAI Spring symposium series.

-Beshah, T., Ejigu, D., Abraham, A., Snasel, V., & Kromer, P. (2013). Mining pattern from road accident data: role of road user's behaviour and implications for improving road safety. International journal of tomography and simulation, 22(1), 73-86.

-Bokaba, T., Doorsamy, W., & Paul, B. S. (2022). Comparative study of machine learning classifiers for modelling road traffic accidents. Applied Sciences, 12(2), 828.

-Chakraborty, M., Gates, T., & Sinha, S. (2021). Causal Analysis and Classification of Traffic Crash Injury Severity Using Machine Learning Algorithms. arXiv preprint arXiv:2112.03407.

-Chen, C., Zhang, G., Qian, Z., Tarefder, R. A., & Tian, Z. (2016). Investigating driver injury severity patterns in rollover crashes using support vector machine models. Accident Analysis & Prevention, 90, 128-139.

-Chen, M. M., & Chen, M. C. (2020). Modeling road accident severity with comparisons of logistic regression, decision tree and random forest. Information, 11(5), 270.

-Eluru, N., Bhat, C. R., & Hensher, D. A. (2008). A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. Accident Analysis & Prevention, 40(3), 1033-1054.

-Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers. Machine learning, 31(1), 1-38.

-Feknssa, N., Venkataraman, N., Shankar, V., & Ghebrab, T. (2023). Unobserved heterogeneity in ramp crashes due to alignment, interchange geometry and truck volume: Insights from a random parameter model. Analytic Methods in Accident Research, 37, 100254.

-Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Learning from imbalanced data sets (Vol. 10, pp. 978-3). Cham: Springer.

-Fiorentini, N., & Losa, M. (2020). Handling imbalanced data in road crash severity prediction by machine learning algorithms. Infrastructures, 5(7), 61.

-Gan, X., & Weng, J. (2020). Predicting Crash Injury Severity for the Highways Involving Traffic Hazards and Those Involving No Traffic Hazards. In CICTP 2020 (pp. 4195-4206).

-Haery, S., Mahpour, A. and Vafaeinejad, A., 2024. Forecasting urban travel demand with geo-AI: a combination of GIS and machine learning techniques utilizing uber data in New York City. Environmental Earth Sciences, 83(20), p.594.

-He, H., & Ma, Y. (Eds.). (2013). Imbalanced learning: foundations, algorithms, and applications.

-Ho, T. K. (1995, August). Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278-282). IEEE.

-Ijaz, M., Zahid, M., & Jamal, A. (2021). A comparative study of machine learning classifiers for injury severity prediction of crashes involving three-wheeled motorized rickshaw. Accident Analysis & Prevention, 154, 106094.

-Iranitalab, A., & Khattak, A. (2017). Comparison of four statistical and machine learning methods for crash severity prediction. Accident Analysis & Prevention, 108, 27-36.

-Islam, A. M., Shirazi, M., & Lord, D. (2023). Grouped Random Parameters Negative Binomial-Lindley for accounting unobserved heterogeneity in crash data with preponderant zero observations. Analytic Methods in Accident Research, 37, 100255.

-Jeong, H., Jang, Y., Bowman, P. J., & Masoud, N. (2018). Classification of motor vehicle crash injury severity: A hybrid approach for imbalanced data. Accident Analysis & Prevention, 120, 250-261.

-Kabli, A., Bhowmik, T., & Eluru, N. (2023). Exploring the temporal variability of the factors

affecting driver injury severity by body region employing a hybrid econometric approach. Analytic Methods in Accident Research, 37, 100246.

-Krishnaveni, S., & Hemalatha, M. (2011). A perspective analysis of traffic accident using data mining techniques. International Journal of Computer Applications, 23(7), 40-48.

-Kuhn, M., & Johnson, K. (2013). Applied predictive modeling (Vol. 26, p. 13). New York: Springer.

-Lee, J., Yoon, T., Kwon, S., & Lee, J. (2019). Model evaluation for forecasting traffic accident severity in rainy seasons using machine learning algorithms: Seoul city study. Applied Sciences, 10(1), 129.

-Li, Z., Liu, P., Wang, W., & Xu, C. (2012). Using support vector machine models for crash injury severity analysis. Accident Analysis & Prevention, 45, 478-486.

-Liu, D. X. (2022). A spatial data statistical model of urban road traffic accidents. Advances in transportation studies, 1.

-Mahpour, A. and Kazemi Naeini, K., 2021. Investigating the social effects of Covid-19 pandemic in the passenger sector of railroad transportation (Case study: Railways of the Islamic Republic of Iran). International Journal of Railway Research, 8(1), pp.43-52.

-Mahpour, A. and Shafaati, M., 2024. Developing a Framework for Selecting an Appropriate Model based on the Ensemble Learning. International Journal of Transportation Engineering, 12(2), pp.1719-1745.

-Mahpour, A., Forsi, H., Vafaeenejad, A. and Saffarzadeh, A., 2022. An improvement on the topological map matching algorithm at junctions: a heuristic approach. International journal of transportation engineering, 9(4), pp.749-761.

-Mahpour, A., Hashemi, M., Asadi, I., Yan, K., You, L., Maghfouri, M. and Haerinia, B., 2023. Evaluation of the optimum value of lightweight expanded clay aggregate incorporation into the roller-compacted concrete pavement through experimental measurement of mechanical and thermal properties. International Journal of Pavement Engineering, 24(2), p.2065489.

-Mahpour, A., Mamdoohi, A. and Hakimelahi, A., 2020. A heuristic technique for traffic assignment with variable step size and number of iterations. Transportation Research Procedia, 48, pp.2569-2579.

-Mahpour, A.R., Amiri, A. and Ebrahimi, E.S., (2019). Do drivers have a good understanding of distraction by wrap advertisements? Investigating the impact of wrap advertisement on distraction-related driver's accidents. Advances in transportation studies, 48, 19-30.

-Mamdoohi, A., Axhausen, K.W., Mahpour, A., Rashidi, T.H. and Saffarzadeh, M., 2016. Are there latent effects in shopping destination choice?: survey methods and response behavior. In 16th Swiss Transport Research Conference (STRC 2016). Swiss Transport Research Conference (STRC).

-Mamdoohi, A.R., Yousefikia, M. and Mahpour, A.R., 2013. Increasing Minimum Spanning Tree estimation precision; implemented for Tehran province. Advances in Civil Engineering & Building Materials,

Routledge Taylor & Francis Group, pp.879-882.

-Tayarani Yousefabadi, A., Mahpour, A. and Javanshir, H., 2020. Modeling share change of non-public vehicles and the rate of emissions due to the implementation of demand management policies. Journal of Transportation Research, 17(3), pp.203-216.

-Mannering, F. L., Shankar, V., & Bhat, C. R. (2016). Unobserved heterogeneity and the statistical analysis of highway accident data. Analytic methods in accident research, 11, 1-16.

-Metsis, V., Androutsopoulos, I., & Paliouras, G. (2006, July). Spam filtering with naive bayes-which naive bayes?.bayes? In CEAS (Vol. 17, pp. 28-69).

-Murty, M. N., & Devi, V. S. (2011). Pattern recognition: An algorithmic approach. Springer Science & Business Media.

-Nujjetty, A. P., Mohamedshah, Y. M., & Council, F. M. (2014). Highway safety information system: Guidebook for data files California. Washington, DC: Federal Highway Administration.

-Provost, F. (2000, July). Machine learning from imbalanced data sets 101. In Proceedings of the AAAI'2000 workshop on imbalanced data sets (Vol. 68, No. 2000, pp. 1-3). AAAI Press.

-Ryu, J. W., Kantardzic, M., & Walgampaya, C. (2010). Ensemble classifier based on misclassified streaming data. In Proc. of the 10th IASTED int. Conf. on artificial intelligence and applications, austria (pp. 347-354).

-Sahebi, S., Mirbaha, B., Mahpour, A. and Norouz Oliaee, M., (2015). Predicting pedestrian accidents in rural roads using ordered logit model. Quarterly Journal of Transportation Engineering, 6(4), pp.581-592.

-Santos, K., Dias, J. P., & Amado, C. (2022). A literature review of machine learning algorithms for crash injury severity prediction. Journal of safety research, 80, 254-269.

-Schutze, H., Manning, C. D., & Raghavan, P. (2008). Introduction to information retrieval. Cambridge University Press.

-Singh, G., Sachdeva, S. N., & Pal, M. (2018). Comparison of three parametric and machine learning approaches for modeling accident severity on non-urban sections of Indian highways. Advances in transportation studies, 45.

-Studer, M., Struffolino, E., & Fasang, A. E. (2018). Estimating the relationship between time-varying covariates and trajectories: The sequence analysis multistate model procedure. Sociological Methodology, 48(1), 103-135.

-Tang, J., Liang, J., Han, C., Li, Z., & Huang, H. (2019). Crash injury severity analysis using a two-layer Stacking framework. Accident Analysis & Prevention, 122, 226-238.

-Tselentis, D. I., Papadimitriou, E., & van Gelder, P. (2023). The usefulness of artificial intelligence for safety assessment of different transport modes. Accident Analysis & Prevention, 186, 107034.

-Umer, M., Sadiq, S., Ishaq, A., Ullah, S., Saher, N., & Madni, H. A. (2020). Comparison analysis of tree based and ensembled regression

algorithms for traffic accident severity prediction. arXiv preprint arXiv:2010.14921.

-Vajari, M. A., Aghabayk, K., Sadeghian, M., & Shiwakoti, N. (2020). A multinomial logit model of motorcycle crash severity at Australian intersections. Journal of safety research, 73, 17-24.

-Wahab, L., & Jiang, H. (2019). A comparative study on machine learning based algorithms for prediction of motorcycle crash severity. PLoS one, 14(4), e0214966.

-Wang, X., & Kim, S. H. (2019). Prediction and factor identification for crash severity: Comparison of discrete choice and tree-based models. Transportation research record, 2673(9), 640-653.

-WHO. (2018). Global status report on road safety 2018. Geneva: World Health Organization; 2018. Licence: CC BYNC-SA 3.0 IGO.

-Yan, X. T., & Shang, Z. L. (2023). Vehicle lane change behavior detection method based on machine learning. Advances in Transportation Studies.

-Yang, J., Han, S., & Chen, Y. (2023). Prediction of Traffic Accident Severity Based on Random Forest. Journal of Advanced Transportation, 2023.

-Zhou, Z. H., & Liu, X. Y. (2005). Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Transactions on knowledge and data engineering, 18(1), 63-77.