

Article

# Identify Risk Pattern of E-Bike Riders in China Based on Machine Learning Framework

Chen Wang \*, Siyuan Kou and Yanchao Song

Intelligent Transportation System Research Center, Southeast University, Nanjing 211189, China; 220183009@seu.edu.cn (S.K.); 220173180@seu.edu.cn (Y.S.)

\* Correspondence: chen\_david\_wang@seu.edu.cn; Tel.: +86-150-0517-2397

Received: 28 September 2019; Accepted: 31 October 2019; Published: 6 November 2019



**Abstract:** In this paper, the risk pattern of e-bike riders in China was examined, based on tree-structured machine learning techniques. Three-year crash/violation data were acquired from the Kunshan traffic police department, China. Firstly, high-risk (HR) electric bicycle (e-bike) riders were defined as those with at-fault crash involvement, while others (i.e., non-at-fault or without crash involvement) were considered as non-high-risk (NHR) riders, based on quasi-induced exposure theory. Then, for e-bike riders, their demographics and previous violation-related features were developed based on the crash/violation records. After that, a systematic machine learning (ML) framework was proposed so as to capture the complex risk patterns of those e-bike riders. An ensemble sampling method was selected to deal with the imbalanced datasets. Four tree-structured machine learning methods were compared, and a gradient boost decision tree (GBDT) appeared to be the best. The feature importance and partial dependence were further examined. Interesting findings include the following: (1) tree-structured ML models are able to capture complex risk patterns and interpret them properly; (2) spatial-temporal violation features were found as important indicators of high-risk e-bike riders; and (3) violation behavior features appeared to be more effective than violation punishment-related features, in terms of identifying high-risk e-bike riders. In general, the proposed ML framework is able to identify the complex crash risk pattern of e-bike riders. This paper provides useful insights for policy-makers and traffic practitioners regarding e-bike safety improvement in China.

**Keywords:** e-bike rider; crash risk; machine learning; traffic violation

## 1. Introduction

As a convenient, economical, energy-saving, and environmentally-friendly travel tool, electric bicycles (e-bikes) can not only meet the needs of people traveling short or medium distances, but they are also able to propel the development of the green sustainability concept. At present, China has become the largest e-bike producer and consumer country around the world. According to statistics, e-bike ownership in China has sharply increased from 58 thousand to 29.96 million, from 1998 to 2014 [1,2]. Experts also predict that the ownership of e-bikes may be in a continuous growth condition with the technological innovation and policy improvement of e-bikes in the future [3].

However, increased e-bike ownership also brings a lot of challenges for traffic safety [4]. Yao claimed that the number of crashes reached nearly 56.2 thousand, with an approximately 8.6% annual increase until the year of 2017 in China [5]. Feng et al. revealed the fact that although the total crash rate in China was in a declining state, the crash rate of e-bikes was still a rising trend [6]. Many studies have also shown that the driving behavior of e-bike riders is non-identical to the regular bike riders, that they tend to be more aggressive and careless. Moreover, compared with other drivers, e-bike riders are more prone to violate traffic rules [7]. Those findings suggest that identifying the risk pattern of e-bike riders is important and necessary. Previous literature has revealed many findings

regarding the contributory factors of e-bike crash risks, including e-bike violation behaviors, e-bike demographics, and the psychological factors of e-bike riders. However, there is limited literature, to our best knowledge, for identifying the complex risk pattern of e-bike riders.

Thus, in our study, the risk pattern of e-bike riders in Kunshan, China, was deeply examined. At-fault e-bike riders were considered as a high-risk group, while other e-bike riders were used as a non-high-risk group. A number of violation-related features were developed, and four tree-structured machine learning models were utilized to capture the complex risk patterns of e-bike riders. This study is expected to provide policy-makers and traffic practitioners with useful insights into e-bike safety.

## 2. Literature Review

Previous studies have concluded that the demographics, aberrant riding behaviors, and environmental conditions can be selected as predictors of the driving risk of e-bike riders [8]. For example, Haustein et al. found that age has a negative correlation with e-bike riding safety. They also found that male riders were more likely to encounter riding risk [9]. A study implemented in China indicated that male and aggressive riders were more likely to increase riding risk [10]. Petzoldt et al. claimed that riding e-bikes at a high speed can induce a higher crash risk [11]. Du proved that unregulated riding behavior and riding without a helmet are prone to increase riding risk [12]. Hertach et al. also found that poor roadway conditions would increase riding risk [13]. In addition, some literature has also found that the crash risk of e-bike riders was related to whether they were registered or not [14].

Regarding the crash modeling technique, most previous literature adopted traditional statistical methods to identify crash-related contributing factors. For example, Cherry et al. used simple mathematical statistical indicators to analyze the riding behavior of e-bike riders [15]. Although this method can analyze the driving behavior of e-bike riders, it requires a large amount of historical data and failed to predict the driver's future driving behavior quantitatively [16]. In addition, some mathematical analysis models have also been introduced into the study of the riding behavior of e-bike riders. For example, the logit model was selected to analyze the risk behavior of e-bike riders [17]. Guo et al. adopted a probit model to analyze the factors related to e-bike riders' riding risk [18]. Wang et al. applied the regression model to identify the association between the riding risk and the related factors [19]. Although these studies reasonably identify some risk aspects of e-bike riders, a comprehensive risk pattern recognition of e-bike riders is still lacking. In particular, a complex risk pattern (e.g., non-linear relationships) was not found.

In general, considerable accomplishments have been achieved in e-bike-related safety research. However, few studies have been found to focus on exploring the complex risk pattern of e-bike riders, based on a machine learning framework.

## 3. Study Design

This study was conducted based on e-bike violation records and crash records during the years between 2015 and 2018, throughout the Kunshan City of Jiangsu Province in China, which were collected by the local traffic police department. There were a total of 242,030 e-bike riders identified in both of the two records. The combined dataset contains the demographic, violation-related information, and crash-related information for each e-bike rider, by matching their previous violation and crash records.

Previous studies have shown that the crash rate and fatality rate of e-bike riders were significantly higher than for other vehicle drivers [20]. Once a crash occurs, it is very likely to induce road congestion [21,22]. These findings indicate that when an e-bike accident occurs, it is a significant hazard to both the rider and the traffic condition around him/her. Therefore, it is essential to identify e-bike riders who are prone to crashes. According to previous literature, at fault e-bike riders in crashes were largely considered as a high-risk group [23], while those not at-fault were often used as quasi-induced exposure [24]. There were some arguments on the quasi-induced exposure method about whether not

at-fault drivers are representatives of the total population. However, it has been agreed that at-fault drivers are a specific group with a higher crash risk.

Thus, in this study, e-bike riders were divided into two groups, namely: high-risk (HR) riders and non-high-risk (NHR) riders. Those with at-fault crash involvements were determined as HR, while all of the other riders (non-at-fault and no crash involvement) were considered NHR. After cleaning the origin data, 2605 e-bike riders were labeled as HR and 239,425 riders were labeled as NHR.

In order to identify the HR e-bike riders based on the machine learning methods, a number of features were developed. As for the demographics, the age groups were divided into four categories based on exclusive class intervals, namely, teenagers (<18 years old), young-aged riders (18~35 years old), middle-aged riders (35~65 years old), and old-aged riders (>65 years old), according to previous literature [25].

Temporal violation features were developed. For example, if one has two prior violation records during the morning peak-hour traffic in Kunshan (i.e., 07:00–09:00), the feature value for morning peak-hour violation will be coded as two. Spatial violation features were also considered. All of the violation locations were classified into the three groups (high, medium, and low) based on their prior cumulative violation frequency. For determining the high and low locations, 15 and 85 percentiles were used. For example, if one has three violation records at low-violation locations, the low-violation location feature value will be coded as three.

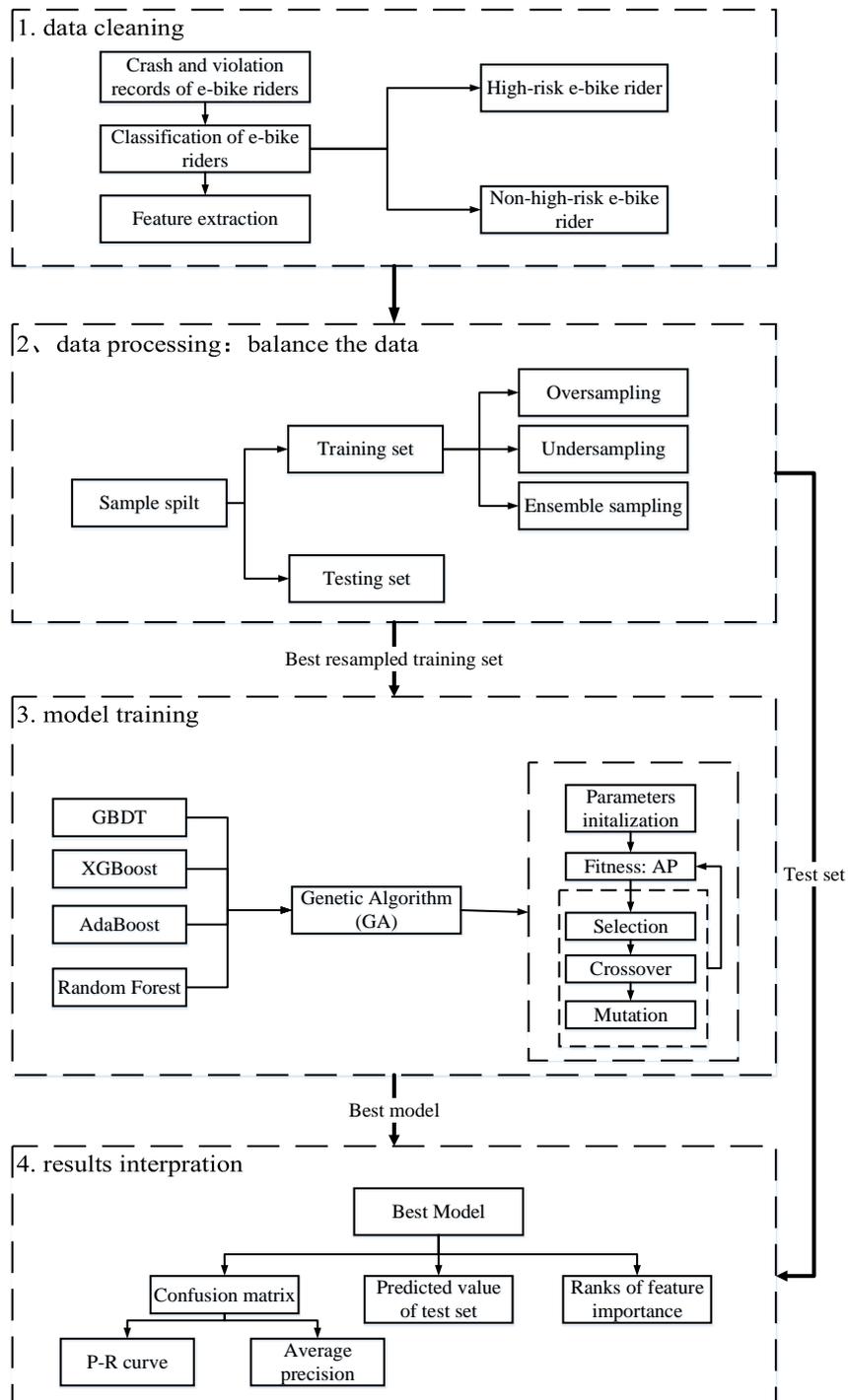
From the original violation dataset, the violation behavior features in the three years were directly extracted for each e-bike rider. For example, if one had two red-light running violation records, the corresponding feature value (i.e., type three: violating traffic signal) would be coded as two. Violation punishment-related features were also developed, including the frequency of violations, frequency of violation type, cumulative penalty, amount of violation penalty, and so on. Table 1 depicts a detailed explanation of the features extracted from the original dataset.

**Table 1.** The features extracted from the original dataset.

Type	Feature	Variables
Demographic Information	Gender	Male Female
	Age	Teenager (<18) Young age (18~35) Middle age (35~65) Old age (>65)
Environmental Information	Temporal violation features	Late night violation (LNV; 0–6) Morning peak hour violation (MPHV; 7–9) Evening peak hour violation (EPHV; 17–19) Night violation (NV; 20–24) Other time (OT; 10–16)
	Spatial violation features	High-frequency location (H_LOC) Middle-frequency location (M_LOC) Low-frequency location (L_LOC)
Violation Information	Violation behavior features	Not riding on the non-vehicle lane (Type 1) Retrograding (Type 2) Red-light running (Type 3) Aggressive riding (Type 4) Speeding (Type 5) Drunk driving (Type 6) Overloading (Type 7) Other violations (Type 8)
	Punishment-related features	Cumulative violation frequency (CVF) Cumulative violation type frequency (CVTF) Cumulative prior penalty (CPP) Average prior penalty (APP) Maximum prior penalty (MPP)

### 4. Methodology

In order to reasonably identify the risk patterns of e-bike riders, a systematic machine learning-based approach was proposed to deal with a number of data mining issues, including imbalanced datasets, machine learning model selection, hyper-parameter tuning, and model validation. These issues will be discussed in the following sub-sections in Section 4. Figure 1 briefly describes the technical flow of the approach.



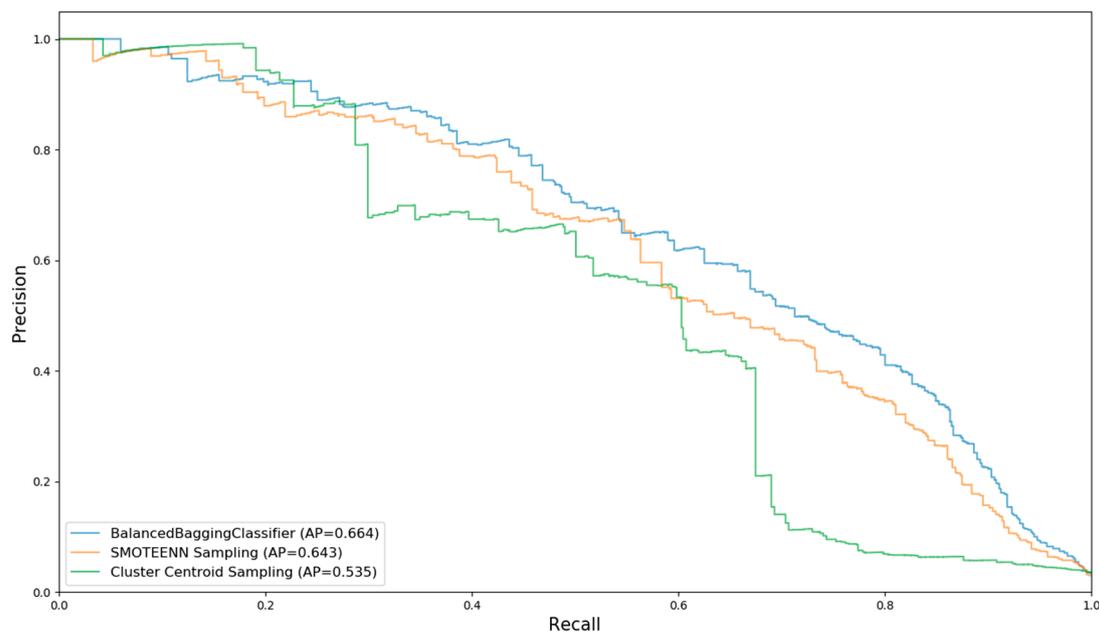
**Figure 1.** Technique flow chart. P-R—precision–recall; GBDT—gradient boosting decision tree; AP—average precision value.

#### 4.1. Sampling Techniques

In the dataset, the number of NHR riders greatly exceeds the number of HR riders, which could cause over-fitting issues. Therefore, sampling techniques should be adopted in order to balance the training dataset. In this study, an over-sampling technique (i.e., synthetic minority oversampling technique (SMOTE)), an under-sampling technique (i.e., cluster-centroid method (CC)), and an ensemble method (i.e., balanced bagging classifier (BBC)) were compared. The average precision values (APs) were used as performance measures.

The under-sampling method adopted in this study is the cluster-centroid (CC) method, which makes use of a k-means algorithm to cut down the number of samples in a majority class. The SMOTE method was considered to be an easy and practical over-sampling method [26], the main idea of which is to synthesize new samples based on the K-nearest neighbors of the original samples in the minority class [27]. The ensemble method used in this study is the balanced bagging method, which combines both the sampling and classification techniques. This method first divides the imbalanced dataset into several subsets first. Then, the under-sampling technique and the classifying estimators will be applied to each subset. At last, the final result can be calculated by synthesizing the result of each subset.

After balancing the training set with the above three sampling techniques, a random forest (RF) classifier was provided for the model training. Figure 2 shows the precision–recall (P–R) curves and AP values of the three sampling methods using RF.



**Figure 2.** P–R curve of each sampling technique.

From the figure, it can be concluded that the balanced bagging method has the best performance. Thus, it was chosen as the standard sampling method, based on which four different tree-structured machine learning models (RF, Adaboost, XGboost, and GBDT) were further trained and compared.

#### 4.2. Model Training

In this study, four tree-structured machine learning methods were utilized for the model training, including the RF algorithm, AdaBoost algorithm with a decision tree, XGBoost algorithm, and the gradient boosting decision tree (GBDT). They are all ensemble machine learning algorithms based on decision trees, so that they are more interpretable than other machine learning models.

#### 4.2.1. Random Forest

The random forest model is an optimization of the decision tree-based algorithm. The random forest introduces two kinds of randomness, namely: (1) random selected samples and (2) random selected feature variables, which make the algorithm insensitive to noise and difficult to over fit. Therefore, the prediction precision of the random forest algorithm can be improved effectively.

The principle of this algorithm is to randomly extract  $n$  subsets from the dataset, and to randomly extract  $m$  features for each subset. Then,  $n$  incomplete decision trees can be established by the randomly selected samples and features. At last, the output result of each decision tree can be integrated, and then the final classification result of each sample is determined through means of voting.

#### 4.2.2. AdaBoost

The AdaBoost classifying technique was first discussed in 1995 by Schapire and Freund [28]. The main idea of AdaBoost is to combine the weak classifiers with suitable weights so as to form a strong classifier [29]. The weighted value can be calculated over multiple iterations. For example, if the sample is classified into the wrong class, its weights will be increased, and vice versa. If the weak classifier has a worse performance than the others, its weight will be reduced, and vice versa. In this study, the selected weak classifier in AdaBoost will also choose a decision tree. The steps of this algorithm can be concluded as follows:

1. Initialize weights  $D_1 = \{w_{11}, w_{12}, w_{13}, \dots, w_{1n}\}$ ,  
where,  $w_{1i} = \frac{1}{n}$ ,  $i = 1, 2, 3 \dots n$
2. For  $m = 1$  to  $M$ :
  - (1) Update training set  $\{(x_i, y_i)\}_{i=1}^n$  by weighted values;
  - (2) Fit a decision tree  $T_m$ , namely:

$$T_m(x) : x \rightarrow \{-1, +1\}$$

- (3) Compute classification error  $e_m$ , as follows:

$$e_m = \sum_{i=1}^n w_i I[T(x_i)]$$

If  $T(x_i) \neq y_i$ , then,  $I[T(x_i)] = 1$ ; if  $T(x_i) = y_i$ , then  $I[T(x_i)] = 0$ ;

- (4) Compute weights  $\beta_m$  of decision tree  $T_m(x)$ , as follows:

$$\beta_m = \frac{1}{2} \log \frac{1 - e_m}{e_m}$$

- (5) Update the model, as follows:

$$F_m(x) = \text{sign} \left( \sum_m \beta_m T_m(x; w_{m,i}) \right)$$

- (6) For  $i = 1$  to  $n$ , update weights  $D_{m+1}$  of the training set, as follows:

$$w_{m+1,i} = \frac{w_{m,i}}{z_m} \exp\{-\beta_m y_i I[T(x_i)]\}$$

Let  $z_m = \sum_{i=1}^n w_{m,i} \exp\{-\beta_m y_i I[T(x_i)]\}$ .

3. Output  $F_M(x)$ .

### 4.2.3. Gradient Boosting Decision Tree

As for GBDT, it generates a strong classifier by weighting the weak classifier. The purpose of this algorithm is to minimize the loss function. The loss function can be defined as  $L(y, F(x)) = \log(1 + e^{-yF(x)})$ , in which  $F(x)$  is the predicted value. The function of  $F(x)$  is  $F(x) = \sum_{m=1}^M \beta_m h_m(x; s_{m,i})$ . To minimize the loss function, a residual term can be added to the prediction function  $F(x)$  in each iteration [30,31]. The steps of this algorithm can be concluded as follows:

1. Initialize model as constant  $c$ :

$$F_0(x) = \operatorname{argmin} \sum_{i=1}^N L(y, c);$$

2. For  $m = 1$  to  $M$ , as follows:

- (1) For  $i = 1$  to  $N$ , compute the so-called pseudo-residuals, as follows:

$$r_{im} = - \left[ \frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right]_{F(x)=F_{m-1}(x)};$$

- (2) Fit decision tree  $h_{mi}$  to  $r_{im}$  by training set  $\{(x_i, y_i)\}_{i=1}^n$ ;

$$h_{mi} = \frac{y_i}{1 + \exp(y_i F_{m-1}(x_i))};$$

- (3) Calculate weights  $\beta_m$  by minimizing the loss function  $L$ :

$$\beta_m = \operatorname{argmin}_{\beta} L(y, F_{m-1}(x) + \beta h_{mi});$$

- (4) Update  $F_m = F_{m-1}(x) + \beta_m h_{mi}$ ;

3. Output  $F_M(x)$

### 4.2.4. XGBoost

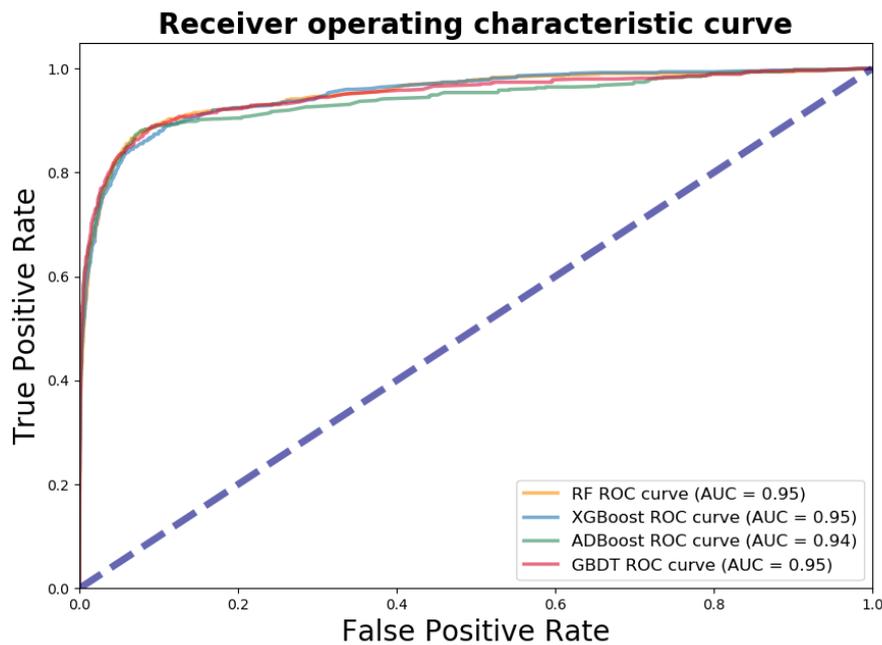
The last adopted algorithm in this study is XGBoost. The main idea of this algorithm is similar to GBDT. Both of them are trying to form a strong classifier by iterating the previous classifier. The differences between XGBoost and GBDT reflect in the following two aspects. First, the objective function of XGBoost is not just a loss function as defined in GBDT. It will add a new term  $\Omega(h(x; s))$  after the loss function, which is used to reduce the complexity of the decision tree. Secondly, in GBDT, the tree just fits the first order gradient of the loss function. But XGBoost will replace the first order gradient with both first order and second order gradients when generating a new decision tree in each iteration [32].

## 5. Model Results

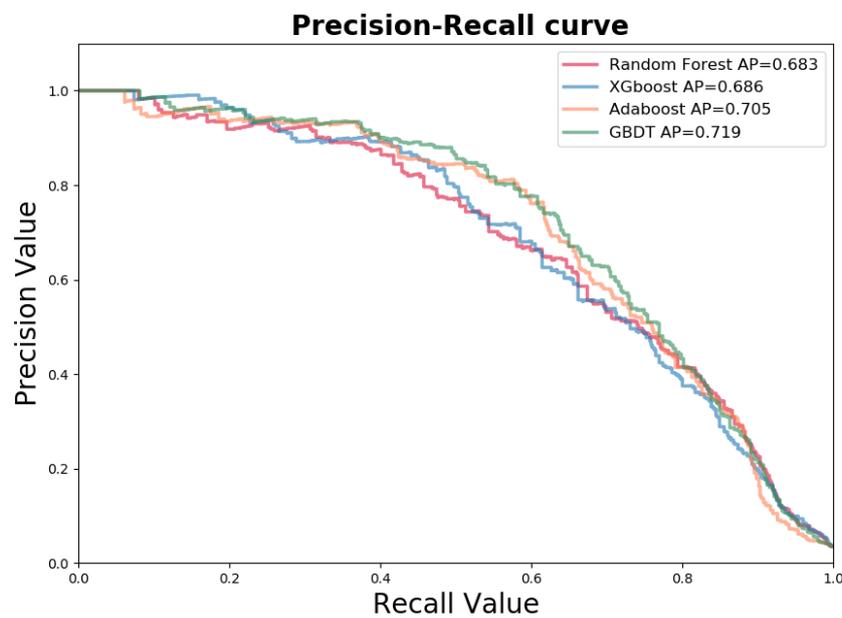
### 5.1. Model Performance

In this study, 75% of the data were used for the model training, while the remaining 25% were used for the model testing. Figure 3 shows the receiver operating characteristic (ROC) curves and precision–recall (P–R) curves of the models. The ROC curve reports the false positive rate and true positive rate. The false positive rate was calculated as the ratio between the number of NHRs wrongly categorized as HR and the total number of actual NHRs. The true positive rate refers to the proportion of HR correctly categorized by the model. A P–R curve reports the precision and recall, which is often used to evaluate the model performance for an imbalanced dataset. The precision is the percentage of

correctly predicted HR over all of the predicted HR. The recall is calculated in the same way as the true positive rate.



(a) P-R Curve.



(b) ROC Curve

**Figure 3.** Estimation of the model performance. ROC—receiver operating characteristic.

The four models all have a large area-under-curve (AUC), indicating that they are capable of identifying both the HR and NHR in general. According to the P-R curves, the average precision of the four machine learning models on the testing dataset were 0.683, 0.686, 0.705, and 0.719 (RF, XGBoost, AdaBoost, and GBDT, respectively). The results show that GBDT appears to be the best model, as it is able to better identify the HR of the e-bike riders.

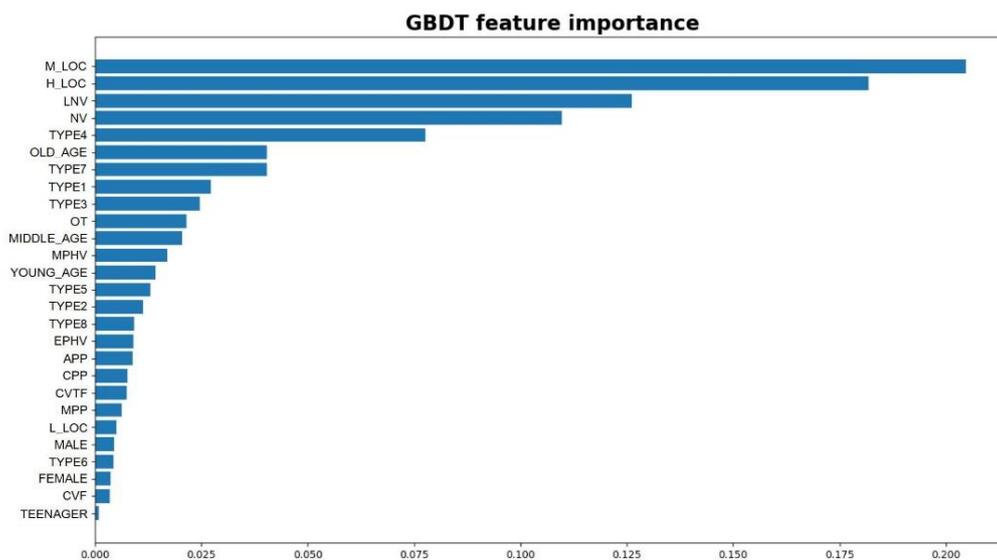
Moreover, the thresholds can be further adjusted to obtain an even higher precision for the HR of e-bike riders, while lowering the recall. The testing dataset contains a total of 18,609 samples, in which the number of HR riders is 651, and the number of NHR riders is 17,958. The adjusted GBDT model prediction results are shown in Table 2, below.

**Table 2.** Adjusted confusion table of the gradient boost decision tree (GBDT) model. HR—high risk; NHR—non-high risk.

	NHR Rider	HR Rider	Threshold	Evaluation Index
<b>NHR Rider</b>	17,829	129	0.959	Recall = 0.596
<b>HR Rider</b>	263	388		Precision = 0.750
<b>NHR Rider</b>	17,682	276	0.909	Recall = 0.70
<b>HR Rider</b>	196	455		Precision = 0.622

### 5.2. Feature Importance

As for the tree-structured machine learning models, the Gini index can be used to evaluate the feature importance. The higher the Gini index, the more importance a feature has. Figure 4, below, presents the importance ranking of all of the features, based on the GBDT model.



**Figure 4.** Feature importance extracted from the GBDT model.

From the results of the feature importance, it can be seen that the top three most influential features are the medium-violation location (M\_LOC), high-violation location (H\_LOC), and late-night violation period (LNV). It is surprising to find that the spatial and temporal characteristics of the previous violation experience have the largest effects on the crash risk of e-bike riders. A possible reason could be that those features are related to risk exposure. Violation behaviors were also found to be important, such as aggressive riding (type 4), overloading (TYPE7), not riding on the non-vehicle lane (type 1), and red-light running (type 3). On the other hand, however, violation punishment-related features were not found to be as important as other violation features. As for demographics, older drivers were also found to be an important feature.

### 5.3. Partial Dependence

To deeply examine the complex risk patterns of e-bike riders, the partial dependence was calculated for each feature [31].

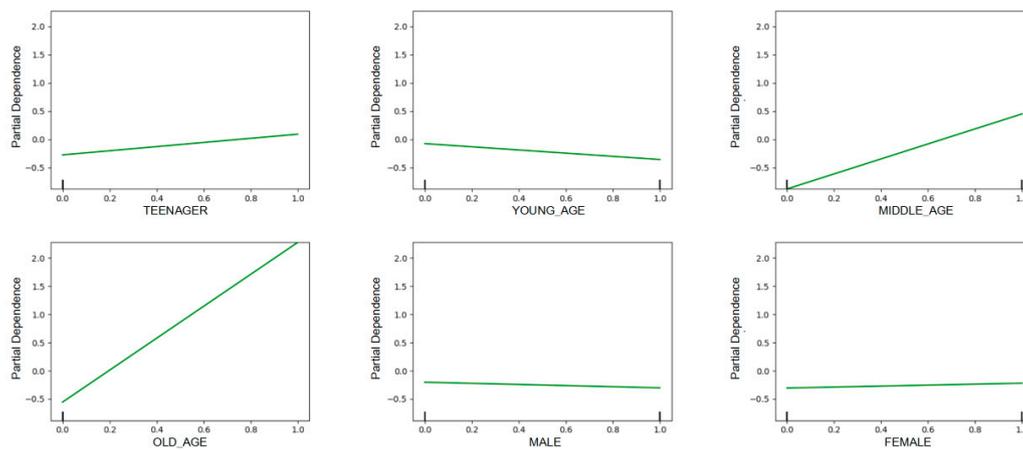
$$f_{x_i}(x_i) = E_{x_{-i}}[f(x_i, x_C)] = \int f(x_i, x_C)d\mathbb{P}(x_C) \quad (3)$$

where  $x_i$  is the feature  $i$ , for which the partial dependence function needs to be calculated, and  $x_C$  is the other features used in model  $f$ . The partial function is estimated by calculating the averages in the training data, based on the Monte Carol method, as follows:

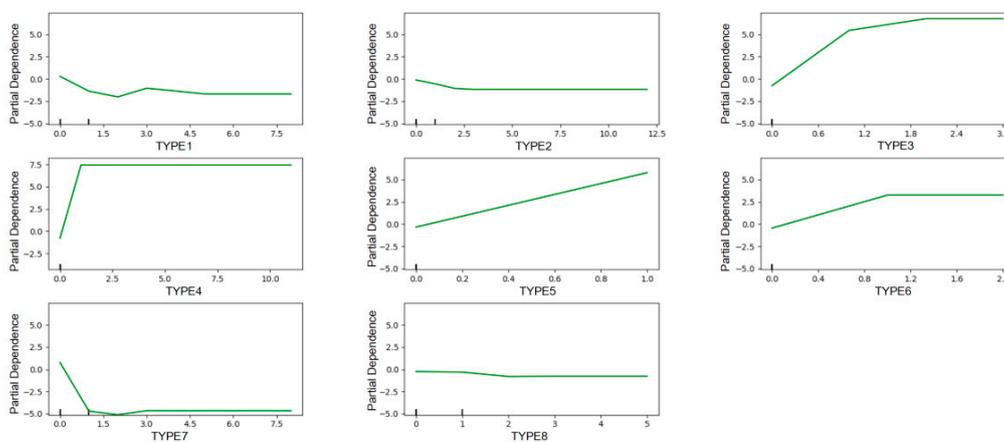
$$f_{x_i}(x_i) = \frac{1}{n} \sum_{i=1}^n f(x_i, x_{C_i}) \quad (4),$$

where  $x_{C_i}$  is the actual values of the feature set  $x_C$ , according to the dataset, and  $n$  is the number of instances in the dataset.

The major advantage of this method is to reveal complex relationships between each factor and outcome. As shown in Figure 5, for each graph, the  $x$ -axis indicates a feature, while the  $y$ -axis presents its partial dependence. It can be found that the effects of each feature tend to be non-constant across different levels. In doing so, a complex relationship between the features and outcomes can be properly identified.

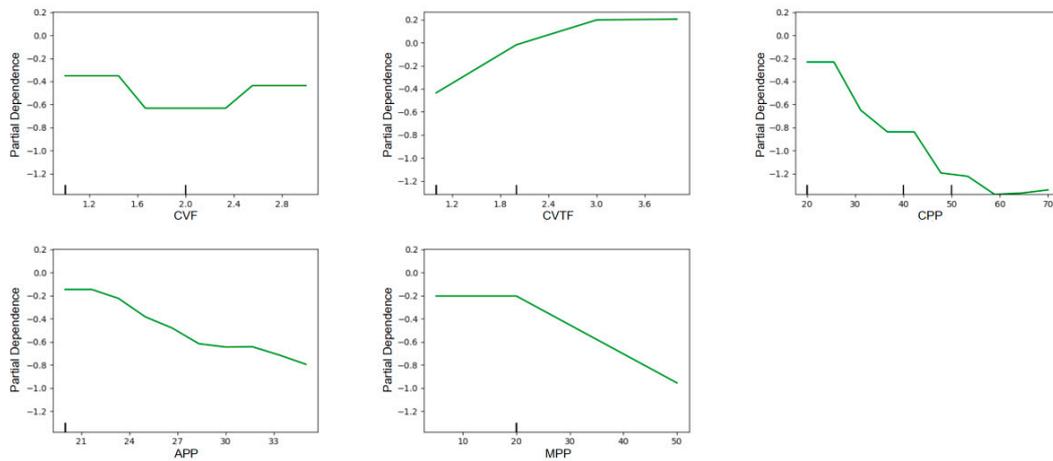


(a) Demographics

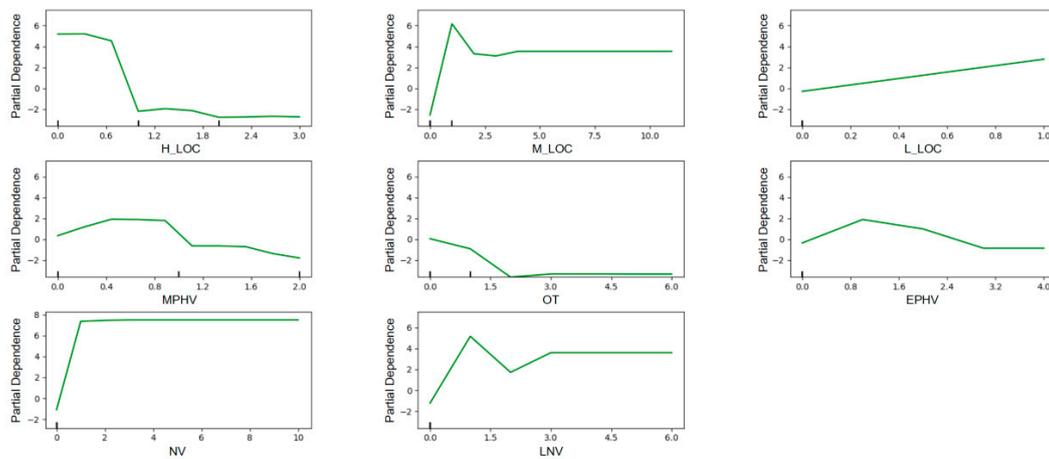


(b) Violation Behaviors

Figure 5. Cont.



(c) Violation Punishment-related Features



(d) Spatial-temporal Violation Features

**Figure 5.** Partial dependence plots of the different features.

It appears that gender has little effect on the risk pattern of e-bike riders (Figure 5a). There is a slight trend that females are more likely to be high-risk riders. Regarding age, it has been shown that older and middle-aged e-bike riders tend to have a higher crash risk. According to previous literature, middle-aged riders are more likely to be involved in a crash [18], because their cognitive ability regarding the surrounding roadway risk is weaker than young-aged riders. Older e-bike drivers were also found to be at high risk, as a result of their diminished physical capabilities [17].

From Figure 5b, the e-bike riders with more previous violation experiences of signal violation (type 3), aggressive riding (type 4), over-speed riding (type 5), and drunk driving (type 6), are more likely to be high-risk riders. Those with more experience in lane violation (Type 1) and overloading (type 7) tend to be in the low-risk group. These findings are interesting. Sometimes e-bike riders commit lane violations (i.e., not riding on the non-vehicle lane) may not be intentional, when the non-motorized lane is occupied by other vehicles or e-bikes. E-bike riders with more overloading violation could be more skilled riders. However, these violation behaviors still deserve further investigation.

From Figure 5c, e-bike riders with many previous violation penalty fees and points tend to be NHR riders. This indicates that riders with more penalty points and fees tend to be less aggressive and more cautious, based on the current violation penalty system. Thus, the current violation penalty system can be considered to effectively alter the dangerous driving habits of HR riders.

According to Figure 5d, e-bike riders with more previous violation experience at high-frequency violation locations tend to be NHR. These violate traffic rules at middle- and low-frequency violation locations tend to be HR. At low-frequency violation locations, most e-bike riders obey traffic rules because of their awareness of traffic enforcement, dangerous roadway conditions, and so on. Thus, those who still violated the rules could be either much more aggressive or have a lack of safety awareness, exposing themselves to dangerous situations. Moreover, the violation behaviors of e-bike riders at low-frequency violation locations could be unexpected for other road users, and more likely to cause crashes. E-bike riders with more violation experiences at night (i.e., night violation (20–24) and late night violation (0–6)) are more likely to be high-risk riders. During the night, visibility could be largely decreased for drivers, because of poor light conditions. Thus, e-bike riders are more likely to be involved in a crash with motor-vehicles when they violate traffic rules.

## 6. Conclusions

In this study, the complex risk pattern of e-bike riders was examined based on a systematic machine learning framework. Three-year crash/violation records were acquired from the Kunshan traffic police department. Based on the quasi-induced exposure theory, at-fault e-bike riders were considered as a high-risk (HR) group, while other e-bike riders (including non-at-fault and non-crash-involved) were determined as a non-high-risk (NHR) group. Demographics and violation-related features were created, and four tree-structured ML models were developed to differentiate the HR group from the NHR group. The following conclusions can be drawn:

- (1) High-risk e-bike riders can be reasonably identified based on a machine learning approach. The major advantage of ML models is that they can better identify complex relationships between the features and e-bike riders' riding risk (i.e., the complex risk pattern), which are not easily captured in traditional statistical methods. Explicable methods, such as tree-based ensemble methods, are highly recommended for such tasks.
- (2) Spatial-temporal violation features were found to be important. Riders with more violation records in low- and medium-violation locations appeared to be high-risk riders. Meanwhile, those with more violation records during night were likely to be high-risk riders. Those features could be useful for identifying and characterizing high-risk e-bike riders.
- (3) Riders' violation behavior was found to be highly correlated with crash risk, as it is able to reflect riders' risky riding habits and attitudes. Violation behavior features appear to be more effective than violation punishment-related features, in terms of capturing the risk pattern of e-bike riders.

In general, the proposed method appears to be a promising tool to identify the risk pattern of e-bike riders in China. Moreover, as such, potential high-risk e-bike riders can be early identified, and certain safety interventions can be applied on this group. For example, they can be mandatorily required to attend traffic school, or can receive safety warning messages routinely. In doing so, e-bike safety in China is expected to be improved in the future.

Admittedly, this research also has some limitations. First, it would be interesting to understand if/how the selected features could support the prediction of new e-bike riders that might have some features missing. Second, it would be more convincing to compare the proposed model to other statistical/machine learning models. However, to our best knowledge, this is the first paper to identify the complex risk pattern of e-bike riders based on machine learning models. Third, the full database containing all of the registered e-bike riders can be used to validate the model. The current database only contains those with either violation or crash records. However, according to quasi-induced exposure theory [24], the current model can still be considered as a reliable tool. We recommend that future research can focus on these directions.

**Author Contributions:** Conceptualization, C.W.; Data curation, S.K and Y.S.; Methodology, C.W. and S.K.; Resources, C.W.; Supervision, C.W.; Writing—review & editing, C.W.

**Funding:** This research was supported by the National Key R&D Program of China (2018YFE0102700).

**Conflicts of Interest:** The authors declare there is no conflict of interest.

## References

1. National Bureau of Statistics of People's Republic of China. *China Statistical Yearbook*; National Bureau of Statistics of People's Republic of China: Beijing, China, 2015.
2. Guo, Y.; Li, Z.; Wu, Y.; Xu, C. Exploring unobserved heterogeneity in bicyclists' red-light running behaviors at different crossing facilities. *Accid. Anal. Prev.* **2018**, *115*, 118–127. [[CrossRef](#)] [[PubMed](#)]
3. Chen, J. Development Trend of Electric Bicycles. *For. Mach. Woodwork. Equip.* **2017**, *45*, 7–9. [[CrossRef](#)]
4. Wang, C.; Xu, C.; Xia, J.; Qian, Z.; Lu, L. A combined use of microscopic traffic simulation and extreme value methods for traffic safety evaluation. *Transp. Res. Part C Emerg. Technol.* **2018**, *90*, 281–291. [[CrossRef](#)]
5. Yao, X. The New National Standard for the Rapid Development of Electric Bicycles. *Electr. Bicycl.* **2018**, *2*, 1–5. (In Chinese)
6. Feng, Z.; Xu, Z.; Huang, D.; Jin, T.; Raghuwanshi, R.P.; Zhang, C. Electric-bicycle-related injury: A rising traffic injury burden in China. *Inj. Prev.* **2010**, *16*, 417–419. [[CrossRef](#)]
7. Brustman, R. *An Analysis of Available Bicycle and Pedestrian Accident Data: A Report to the New York Bicycling Coalition*; New York Bicycling Coalition: Albany, NY, USA, 1999.
8. Hu, F.; Lv, D.; Zhu, J.; Fang, J. Related Risk Factors for Injury Severity of E-Bike and Bicycle Crashes in Hefei. *Traffic Inj. Prev.* **2014**, *15*, 319–323. [[CrossRef](#)]
9. Hausteiner, S.; Möller, M. E-bike safety: Individual-level factors and incident characteristics. *J. Transp. Health* **2016**, *3*, 386–394. [[CrossRef](#)]
10. Yao, L.; Wu, C. Traffic Safety for Electric Bike Riders in China Attitudes, Risk Perception, and Aberrant Riding Behaviors. *Transp. Res. Rec.* **2012**, *24*, 49–56. [[CrossRef](#)]
11. Petzoldt, T.; Schleinitz, K.; Heilmann, S.; Gehlert, T. Traffic Conflicts and Their Contextual Factors When Riding Conventional vs. Electric Bicycles. *Transp. Res. Part F Traffic Psychol. Behav.* **2017**, *46*, 477–490. [[CrossRef](#)]
12. Du, W.; Yang, J.; Powis, B.; Zheng, X.; Ozanne-Smith, J.; Bilston, L.; Wu, M. Understanding on-road practices of electric bike riders: An observational study in a developed city of China. *Accid. Anal. Prev.* **2013**, *59*, 319–326. [[CrossRef](#)]
13. Hertach, P.; Uhr, A.; Niemann, S.; Cavegn, M. Characteristics of single-vehicle crashes with e-bikes in Switzerland. *Accid. Anal. Prev.* **2018**, *117*, 232–238. [[CrossRef](#)] [[PubMed](#)]
14. Zhou, J.; Guo, Y.; Wu, Y.; Dong, S. Assessing Factors Related to E-Bike Crash and E-Bike License Plate Use. *J. Transp. Syst. Eng. Inf. Technol.* **2017**, *17*, 229–234. [[CrossRef](#)]
15. Cherry, C.; Cervero, R. Use characteristics and mode choice behavior of electric bike users in China. *Transp. Policy* **2007**, *14*, 247–257. [[CrossRef](#)]
16. Wang, C.; Xu, C.; Dai, Y. A crash prediction method based on bivariate extreme value theory and video-based vehicle trajectory data. *Accid. Anal. Prev.* **2019**, *123*, 365–373. [[CrossRef](#)]
17. Wang, C.; Xu, C.; Xia, J.; Qian, Z. Modeling Faults among E-Bike-Related Fatal Crashes in China. *Traffic Inj. Prev.* **2017**, *18*, 175–181. [[CrossRef](#)]
18. Guo, Y.; Zhou, J.; Wu, Y.; Chen, J. Evaluation of Factors Affecting E-Bike Involved Crash and E-Bike License Plate Use in China Using a Bivariate Probit Model. *J. Adv. Transp.* **2017**, *2017*, 1–12. [[CrossRef](#)]
19. Wang, C.; Xu, C.; Xia, J.; Qian, Z. The effects of safety knowledge and psychological factors on self-reported risky driving behaviors including group violations for e-bike riders in China. *Transp. Res. Part F Traffic Psychol. Behav.* **2018**, *56*, 344–353. [[CrossRef](#)]
20. Jianan, Z.; Shuai, D.; Xinyu, Z. Characteristics of Electric Bike Accidents and Safety Enhancement Strategies. *Urban Transp. China* **2018**, *16*, 15–20. [[CrossRef](#)]
21. Shi, X.; Wang, W. Study on Accident Effects in Urban Transportation Network. *J. Highw. Transp. Res. Dev.* **2000**, *17*, 38–41.
22. Chen, J.; Li, Z.; Wang, W.; Jiang, H. Evaluating Bicycle–Vehicle Conflicts and Delays on Urban Streets with Bike Lane and on-Street Parking. *Transp. Lett.* **2018**, *10*, 1–11. [[CrossRef](#)]
23. Rose, G. E-Bikes and Urban Transportation: Emerging Issues and Unresolved Questions. *Transportation* **2012**, *39*, 81–96. [[CrossRef](#)]

24. Chandraratna, S.; Stamatiadis, N. Quasi-induced exposure method: Evaluation of not-at-fault assumption. *Accid. Anal. Prev.* **2009**, *41*, 308–313. [[CrossRef](#)] [[PubMed](#)]
25. Wang, C.; Liu, L.; Xu, C.; Lv, W. Predicting Future Driving Risk of Crash-Involved Drivers Based on a Systematic Machine Learning Framework. *Int. J. Environ. Res. Public Health* **2019**, *16*, 334. [[CrossRef](#)] [[PubMed](#)]
26. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [[CrossRef](#)]
27. Elreedy, D.; Atiya, A.F. A Comprehensive Analysis of Synthetic Minority Oversampling TEchnique (SMOTE) for Handling Class Imbalance. *Inf. Sci.* **2019**, *505*, 32–64. [[CrossRef](#)]
28. Freund, Y.; Schapire, R.E. A Decisiontheoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.* **1995**, *55*, 119–139. [[CrossRef](#)]
29. Hu, W.; Member, S.; Hu, W.; Maybank, S. AdaBoost-Based Algorithm for Network. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* **2008**, *38*, 577–583.
30. Ding, C.; Cao, X.J.; Næss, P. Applying gradient boosting decision trees to examine non-linear effects of the built environment on driving distance in Oslo. *Transp. Res. Part A Policy Pract.* **2018**, *110*, 107–117. [[CrossRef](#)]
31. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [[CrossRef](#)]
32. Gómez-Ríos, A.; Luengo, J.; Herrera, F. A Study on the Noise Label Influence in Boosting Algorithms: AdaBoost, GBM and XGBoost. *Lect. Notes Comput. Sci.* **2017**, *10334*, 268–280.



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).