Assessing Accident Risk using Ordinal Regression and Multinomial Logistic Regression Data Generation

Gulsum Alicioglu
Electrical and Computer Engineering
Rowan University
Glassboro, US
alicio87@students.rowan.edu

Bo Sun
Computer Science
Rowan University
Glassboro, US
sunb@rowan.edu

Shen Shyang Ho
Computer Science
Rowan University
Glassboro, US
hos@rowan.edu

***Abstract*— Robust and accurate modeling of motor vehicle accident and injury severities have significant impact on transportation safety and economy. The capability to assess accident risk based on external driving conditions (e.g., weather, road condition, etc.) and driver behavior and characteristics can reduce accident occurrences by alerting drivers to alleviated risk. In this paper, we propose a novel accident risk assessment framework driven by ordinal regression. One challenge of the risk assessment problem is that non-accident data are not collected by any agency in their study of transportation safety. Hence, we also propose a realistic negative data generation scheme based on feature weighs derived from multinomial logistic regression to overcome this challenge. Experimental results on two different real-world datasets from the US National Highway Traffic Safety Administration and UK Transport for Greater Manchester are used to demonstrate the feasibility and robustness of our proposed ordinal regression framework. Performance on four ordinal regression algorithms, namely: logistic all-threshold, logistic immediate-threshold, ordinal ridge, and least absolute deviations are compared. In addition, for US dataset, we investigate the effect of random oversampling and undersampling on the proposed risk assessment framework. We empirically show that bagging with random oversampling using logistic all-threshold ordinal regression method has the best prediction performance among ordinal regression models.**

Keywords*— ordinal regression, injury severity classification, machine learning, data generation, accident risks.

I. Introduction

With the increase in population and the rise in the number of vehicles, the visible and hidden costs happened due to traffic accidents and injuries have been increased over the years [1]. The capability to model and assess motor vehicle accident and injury severities have significant impact on transportation safety and economy. Therefore, we propose an accident prevention and alerting system to reduce these risks and costs by predicting the risks. Moreover, we determine the factors that cause accidents and injuries before training the prediction model. Using the prevention and the alerting system, accident-prone situations and dangerous human driving behaviors will be detected. Thus, the risks of accidents and the severity of accidents can be reduced by providing real-time warnings to the drivers about accident risk. To train a prediction model, a system requires both the non-accident training data, as well as accident training data. However, in reality, government agencies do not collect any non-accident data. Hence, generating non-accident data to be used in the training of the prediction model is necessary or the system needs to adopt a prediction model that can be trained without using the non-accident data.

Accident injury severities typically vary from non-fatal injury to fatal injury level. An ideal prevent system would change the course of the precautions to be taken according to a predicted injury severity for a potential fatal injury accident. For example, estimating fatal injury as a serious injury instead of possible injury provides consistent results; however, precautions issued by the system due to incorrect predictions would mislead the drivers. To ensure road and driver safety, accurate prediction results need to be obtained to alert drivers to take precautions. Many studies were done to determine accident risks using machine learning algorithms [2-5]. However, none of them used ordinal regression algorithms which demonstrate better results than conventional machine learning algorithms in a classification problem where the class order is important. In ordinal regression models, estimated output is ordinal, i.e., the classes have an inherent order [6]. In the literature, ordinal regression models have been used in applications where human assessment plays a significant role [7-9]. Some examples of the applications include evaluating disease severity in plant [8] and assessing credit-rating agencies [9].

The contributions of this paper are as follows:

- The first use of ordinal regression for accident (or injury) risk prediction.
- A novel ordinal regression-based framework for accident (or injury) risk prediction that utilizes negative (non-accident) data generated based on feature weighs derived from multinomial logistic regression.

The rest of the paper is organized as follows. Section II provides a brief review on accident (or injury) predictions and ordinal regression methods. Section III describes the proposed method in detail. Section IV shows the experimental results and discussions. Section V concludes the paper.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

II. Literature Review

Assessment of injury severity and determination of critical factors for motor vehicle accidents using machine learning algorithms have been extensively investigated [2-5, 10-12]. Jeong et al. [4] addressed certain issues in imbalanced datasets with a hybrid method to improve the accuracy performance in injury severity classifications. They showed that oversampling treatment with bagging has the best accuracy performance in decision tree (DT) method. Similarly, Yuan et al. [3] used informative negative sampling approach to handle imbalanced motor vehicle dataset. They evaluated support vector machine (SVM), DT, neural network (NN), and random forest (RF) methods for their binary classification problem. Zhu et al. [12] proposed a machine learning based framework to detect driver injury patterns using NN and RF. They concluded that female drivers, truck usage, driver distraction, vehicle rollover and dawn/dusk are some contributing factors for severe injury levels and fatalities. Aci and Ozden [5] investigated the effect of weather on injury severity comparing the results of numerous machine learning methods (i.e. k-Nearest Neighbor, Naive Bayes, NN, SVM, and DT), and logistic regression. The most common machine learning algorithms (DT, SVM, k-NN, NN, RF, etc.) are frequently applied in assessing injury severity in accidents [3-5, 12]. These researches tried to improve the performance of the algorithms classifying accident risk problems. Table I shows the natural order from low level severity to high level severity for the classes in seven related work and the numerous methods used and compared.

Ordinal regression models are used when the classes represent levels of an inherent order [6-9, 13,14]. Xia et al. [13] using synthetic data that ranked classes from 1 to 4 and benchmark dataset contains 10 ordered classes, compared the performance of support vector ordinal regression with perceptron and gaussian kernel. Pérez-Ortiz el al. [6] conducted a study of a healthcare application using ordinal regression. They used kernel discriminant learning ordinal regression to investigate whether depression has a spatial dependence of prevalence. They had three ordered classes: spatial unit with depression, depression could be present and no depression. Landschoot et al. [8] converted their continuous DON (deoxynivalenol) classes, which is one of the most prevalent toxins in wheat samples, into natural ordered and partition classes using thresholds. They highlighted that predicting and assessing DON values with ordinal regression achieves better accuracy.

To the best of our knowledge, there was no previous study on using ordinal regression to classify injury severities and risks. As accident classifications are ranked based on categories ranging from no injury, possible injury, non-incapacitating injury, incapacitating injury to fatal injury, ordinal regression is a natural approach to deliver better risk assessment performance. Table I lists out all the classes used for accident classification along with proposed machine learning algorithms for some related work. In our paper, we compare four different ordinal regression models used in our prediction framework (see Section III. *B*).

TABLE I. Summary of Research Related to Accident Risk Assessment

Studies	Class Descriptions	Algorithms
[2]	Slight Injured Killed or Seriously Injured	Bayesian Networks
[3]	Accident No Accident	SVM DT RF Deep NN
[4]	Fatal Injury Incapacitating Injury Non-Incapacitating Injury Possible Injury No Injury	Logistic Regression (LR) DT NN Gradient Boosting Model Naïve Bayes
[5]	Non-Fatal Injury Fatal Injury	k-NN Naïve Bayes NN DT SVM LR
[10]	No Injury Possible Injury Non-Incapacitating Injury Incapacitating Injury Fatal Injury	DT SVM Hybrid DT-Artificial NN (DTANN)
[11]	Property Damage Only Possible Injury Visible Injury Fatal Injury	Multinomial Logit k-NN SVM RF k-Means Latent Class Clustering
[12]	No Injury Possible Injury Evident Injury Fatal Injury	RF NN

III. Methodology

A. Overview of Accident Prevention and Alert System

This section provides an overview of the prevention and alerting system framework and details of the prediction model. The framework of the prevention and alerting system is shown in Fig. 1. The system contains four steps: inputs to the system, training models, estimating/predicting accident risks and providing outputs (say, precaution messages) to drivers.

The prevention and alert system take 5 categories of inputs to estimate possible injury related accident. The 5 input categories are driver information (such as driver's gender, age), GPS data (such as vehicle location), weather information, and road information (such as surface type, surface condition, traffic lines, light condition) as well as former recorded accident datasets as training dataset. The features from these categories are used to obtain estimations through our proposed prediction model. The system obtains the accident risks according to prediction results. Then the system provides alert messages to warn drivers to take precaution. The outputs vary from safe zone to high crash zone or to signal alleviated accident risk. Drivers can take some precautions according to these outputs like reducing the speed, keeping a safe following distance and etc. This system aims to reduce accident risks and severity of the injuries. Therefore, the prediction model plays a significant role in the whole system.

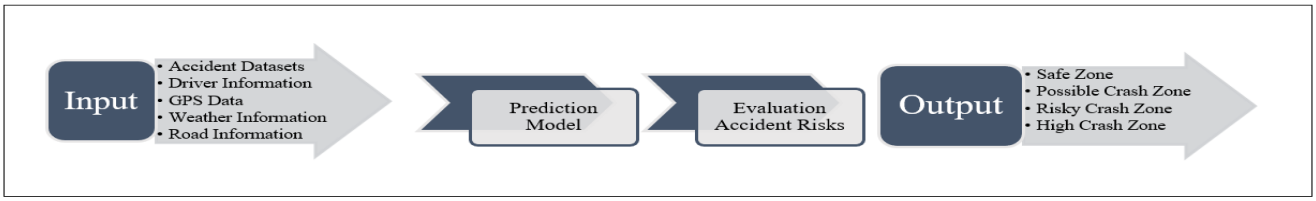

Fig. 1. The framework of the prevention and alerting system

Our paper focuses on the prediction model. In Fig. 2, the framework of the prediction model is given in detail. Negative data generator creates negative (non-accident) samples. Then random oversampling or undersampling techniques are applied to US accident dataset before the training stage. Since the UK accident dataset has balanced classes, random oversampling and undersampling techniques were not applied. Prediction results for binary classification as accident/non accident and multiclass classification as from non-fatal injury severity to fatal injury severity are obtained using ordinal regression models as prediction model. The prediction framework is described in detailed in Section III.*B*.

B. Proposed Prediction Framework

The proposed prediction method consists of two steps: 1) Creating the missing class and 2) applying ordinal regression models. Sections below will provide information about the negative data generation process (Step 1) and ordinal regression models along with preprocessing techniques (Step 2).

1) Negative Data Generator: The data generator (only for US dataset) creates negative examples (non-accident data) used in training. The method used the combined value ranges of significant features in dataset to create sample data on non-accidental data. The weights of the features reflect the importance degree of the related feature. Features with a higher weight is important for classification of the dataset. Fig. 3 describes steps to create non-accident data. The weights of the features were obtained using Multinomial Logistic Regression (MLR). In the present paper, the negative data generation process is mainly used to generate non-accident instances. Equation (1) shows the MLR model [15].

$$P(Y_i = m) = \frac{e^{\alpha_m + \sum_{k=1}^{K} \beta_{mk} X_{ik}}}{1 + \sum_{h=2}^{M} e^{\alpha_h + \sum_{k=1}^{K} \beta_{hk} X_{ik}}} \quad (1)$$

where α is the constant, β is a vector of regression coefficients, x is data and M is class label.

Random oversampling and undersampling methods are used. Random sampling is a technique used to create balanced datasets by generating or removing samples randomly with current samples. Random oversampling is used by generating new samples randomly when a class is underrepresented in dataset. On the other hand, in cases where a class is overrepresented, random undersampling is used by removing samples randomly from current samples.

2) Ordinal Regression Models: Ordinal regression models developed by McCullagh, uses ordinal nature of data by defining various stochastic sorting paradigms [16]. These methods resolve the requirement of assigning scores to classes instead of ordinality [16]. Ordinal regression is a supervised learning problem where the label of the classes has an inherent order [13]. Ordinal regression algorithms benefit this order information to improve classification performance [7]. Ordinal regression implementations occur in areas where human-source data is significant and where the output variable cannot be measured with high sensitivity [7,8]. The accident dataset used in the paper presents such characteristics. In this paper, ordinal regression methods are divided into two main groups as: Threshold-based and Regression based methods. Equation (2) shows general ordinal regression model [16].

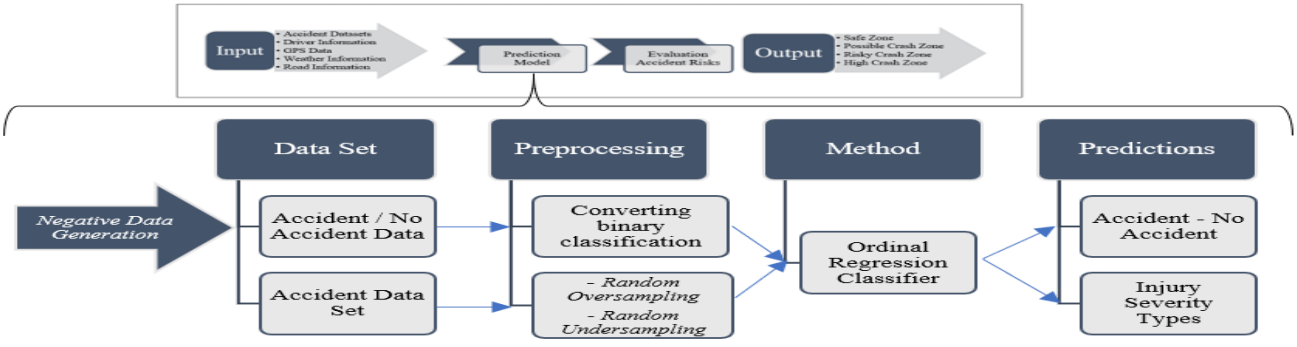

Fig. 2. The framework of the proposed prediction model

$$\log\left[\frac{\gamma_j(x)}{\{1-\gamma_j(x)\}}\right] = \theta_j - \beta' x \qquad (1 \le j < k) \tag{2}$$

where $\gamma j = p_1(x) + \ldots + p_j(x)$, β is a vector of regression coefficients and $\theta_j = log_{kj}$, $k_j(x)$ be the odds.

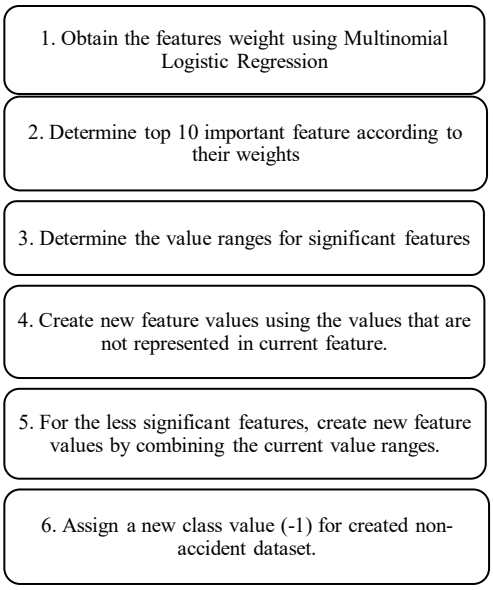

Fig. 3. Negative (Non-accident) data generation process

a) Threshold-based Methods: In threshold-based ordinal regression models, threshold values are determined to create consecutive intervals between ordered classes [13]. In the current study two type threshold-based ordinal regression models are used: Logistic All-threshold (AT) and Logistic Immediate threshold (IT). Shashua and Levin [17] introduced the immediate-threshold model, also called fixed margins, by proposing the application of large-margin classifiers for ordinal regression models. If the threshold values defined for each class are violated, the penalty is imposed. However, the immediate threshold method does not guarantee that the threshold values will be consecutive. Thus, all-threshold based method was introduced to guarantee that the thresholds will be ordered by imposing more penalties [18]. The all-threshold loss corresponds to a total value of all threshold violation penalties. Therefore, solutions in the all-threshold method are desired to have the minimum number of crossed thresholds [19]. In this paper, α regularization parameter is taken as 50 for threshold-based methods.

b) Regression-based Methods: Two different regression-based ordinal regression models are applied to dataset: Ordinal Ridge and Least Absolute Deviation (LAD). To estimate the regression coefficients without removing the variables in the model, Ridge regression and LAD methods are used. Ridge regression provides biased estimates and is the best-known penalization approach [20]. Ridge regression approximates parameter estimates to zero value without making them completely zero [8]. The Least Absolute Deviation model, also known as a L1 regression, is a statistical optimization technique that minimizes the sum of the absolute values of the residuals. LAD can be classified as a nonlinear optimization problem [21]. This provides a robust estimator. However, LAD regression is not robust when the data has outliers in the illustrative variables [9]. Ordinal regression methods that are based on Support Vector Machines also have high computational complexity due to optimization. In LAD method, ϵ parameter is taken as 0.001, tolerance value is taken 0.0001 and regularization parameter is taken as 10 in this study. For Ordinal Ridge method, regularization parameter and tolerance values are equal to 10 and 0.0001, respectively. The performance of the ordinal regression models is compared in terms of training and testing time.

IV. Experimental Results

A. Data Description

Experiments are performed using two different real-world accident datasets. Motor vehicle accident data used for accident risk analysis are retrieved from the US National Highway Traffic Safety Administration website, particularly the Fatality Analysis Reporting System [22] and UK Transport for Greater Manchester website [23]. The first dataset contains accident records from the year of 2015 to 2016 for top five states where highest number of accident records were found in US. The states include California, Florida, Georgia, North Carolina and Texas. The original dataset went through data cleaning process for missing and incorrect values; the data has 22380 entries and 17 features related to driving conditions including atmospheric condition, accident day, accident hour, accident month, day of week, holiday related, light condition, intersection type, age, person type, sex, travel speed, vehicle make, driver alcohol involvement, surface condition, surface type and traffic lanes. The second dataset contains accident records from the year of 2018 in UK. Data cleaning process for missing and incorrect values was applied; the data has 14593 entries and 10 features related to driving conditions including atmospheric condition, day of week, road type, speed limit, junction detail, junction control, pedestrian crossing-human control, pedestrian crossing-physical facilities, light condition, weather condition and road surface condition.

The US dataset provides five different injury severities and the UK dataset provides three different injury severities as classification classes. Table II summarizes the information about injury severity levels of accidents. The datasets ranged from no apparent/slight injury to fatal injury. Thereby, ordinal regression methods are applied to predict the classification results for prevention and alerting system. All codes for the experiments are written in Python.

In the US accident dataset, fatal injury and no apparent injury classes have the two largest percentage in all injury severities, which are 38% and 28.6%, respectively. Slight injury severity has the largest percentage (57.4%) in the UK dataset. Besides conducting five-class and three-class classification experiments with ordinal regression models, for

US dataset binary classification experiments also carried out by generating missing class which contains non-accident data/records. The ordered injury severity levels are coded in the datasets as 0, 1, 2, 3, and 4, respectively.

TABLE II. DESCRIPTION OF INJURY SEVERITY LEVELS IN ACCIDENT DATASETS

US Accident Dataset (2015-2016)			**UK Accident Dataset (2018)**	
Injury severity		***# of accidents***	***Injury severity***	***# of accidents***
Class 0	No apparent injury	6405 (28.6 %)	Slight	8381 (57.4 %)
Class 1	Possible injury	2697 (12.1 %)	Serious	4541 (31.1 %)
Class 2	Suspected minor	2967 (13.3 %)	Fatal	1671 (11.5 %)
Class 3	Suspected serious	1812 (8.1 %)		
Class 4	Fatal	8499 (38.0 %)		

B. Feature Extraction and Negative Data Generation

For the US dataset, among all 17 driving related features, we only picked high impact features for negative sampling to create non-accident data. The weights of the features are obtained using multinomial logistic regression. Features with a higher weight are important on the injury severity of the accident. The top five features and their corresponding weights are provided in the order in Table III.

TABLE III. THE TOP FIVE FEATURES AND CORRESPONDING WEIGHTS

Non-fatal injury		**Possible injury**		**Minor injury**		**Major injury**		**Fatal injury**	
Light cond.	0.166	P. type	0.264	Alc.	0.262	Alc.	0.490	Alc.	0.918
Lane	0.161	Intsec type	0.213	P. type	0.259	P. type	0.442	Surf type	0.099
Intsec type	0.064	Sex	0.189	Surf cond	0.122	Surf type	0.127	Age	0.013
Hold	0.016	Lane	0.081	Surf type	0.099	Sex	0.106	V. make	0.005
Acc. hour	0.012	Surf cond	0.032	Acc. hour	0.004	Surf cond.	0.022	Surf cond	0.002

The most important feature from the minor injury severity level to fatal injury is alcohol. Among these levels, surface type, surface condition, person type, age, and sex are also common. For the accidents that have low injury severities, such as non-fatal and possible injury, light condition, intersection type, number of traffic lanes are among the important features.

In Table 3, the important factors are identified for the five accident classes. For the non-accident data generation process, the top ten features' combined range of values is examined. For instance, surface condition, which is one of the important factors, ranges from 1 to 2 values for all classes, in accident dataset. Thus, other surface condition values should range randomly from 3 to 5 for non-accident class. With this information, by using negative sampling surface condition values ranged from 3 to 5 in non-accident class. Similar approaches are applied to other ten important features to generate random values for non-accident class. For other less significant features, the values are randomly chosen from the combined range value of the five classes.

After the data generation process, the dataset has 44380 records with accident and non-accident classes. Non-accident class has 22000 entries labeled "-1" and accident class has 22380 entries labeled "+1", respectively; specifically, accident class was represented as one class that contains all injury severity classes.

C. Performance Metrics

To examine the performances of the ordinal regression models MSE, measuring the average of the squares of the errors, is used as performance evaluation criteria. Equation (3) shows how to assess MSE.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{} \quad (3)$$

where n is the number of sample size in the dataset, y_i is the actual values and $\tilde{y}_i$ is the predicted values of the target. MSE defined as the average squared difference between the actual values and predicted values, where lower difference is preferred.

D. Experimental Results, Comparisons and Discussion

This section presents the experimental results on the accident datasets. Before training-testing and cross validation process, data cleaning process and random oversampling (ROS), random under sampling (RUS) methods are implemented and applied to the datasets. In all experiments, 10-fold cross-validation is applied to avoid the effect of randomness by dividing dataset randomly into 80% for training and 20% for testing. Bagging method, also called Bootstrap aggregating is an ensemble meta-algorithm, which is proposed by [24] to improve the performance of weak classifier. In the current study, bagging method is also implemented in both datasets. Bagging method not only reduces the variance but also avoid overfitting [24].

Table IV indicates the results of ordinal regression models in accident and non-accident data (i.e. binary classification). The values represent the mean MSE scores and standard deviation. The best score is marked as bold.

TABLE IV. BINARY CLASSIFICATION RESULTS IN ORDINAL REGRESSION MODELS

Dataset	**Methods**			
	Regression-based		***Threshold-based***	
	Ordinal Ridge	***LAD***	***Logistic IT***	***Logistic AT***
Non-Accident/ Accident Dataset	0.0009 ± 0.0011	0.0217 ± 0.0458	0.0011 ± 0.0004	**0.0005 ± 0.0004**[a]

[a] Mean Squared Error ± Standard Deviation

Logistic AT method has the best MSE score for binary classification. Since the Logistic AT is introduced to annihilate

the disadvantages of Logistic IT, getting a better result was expected. The worst result belongs to LAD, which also uses more time to solve the problem because of the optimization.

After the binary classification, the experiments are conducted using ordinal regression models to determine injury severity levels and accident risks. The experiments carried out for original datasets, and improved datasets (i.e., after applying random oversampling and under sampling). The corresponding results are shown in Table V.

TABLE V. MULTICLASS CLASSIFICATION RESULTS IN ORDINAL REGRESSION MODELS

US Dataset		Methods			
		Regression-based		*Threshold-based*	
		Ordinal Ridge	***LAD***	***Logistic IT***	***Logistic AT***
No Bagging	Original	**1.265 ± 0.038**	1.269 ± 0.034	2.570 ± 0.076	1.338 ± 0.045
	ROS[b]	0.643 ± 0.015	0.650 ± 0.022	0.845 ± 0.019	**0.625 ± 0.016**
	RUS[c]	1.064 ± 0.037	**1.062 ± 0.047**	1.932 ± 0.101	1.082 ± 0.056
Bagging	Original	**1.251 ± 0.033**	1.253 ± 0.034	2.507 ± 0.072	1.325 ± 0.043
	ROS	*0.635 ± 0.015*	*0.642 ± 0.021*	*0.835 ± 0.018*	***0.616 ± 0.017***
	RUS	1.035 ± 0.035	**1.042 ± 0.049**	1.886 ± 0.102	1.056 ± 0.056
Training-Testing Time (Secs)		0.097	671.58	3.632	4.120

UK Dataset		***Ordinal Ridge***	***LAD***	***Logistic IT***	***Logistic AT***
No Bagging	Original	**0.372 ± 0.025**	0.516 ± 0.170	0.438 ± 0.062	0.396 ± 0.022
Bagging	Original	**0.363 ± 0.022**	0.501 ± 0.092	0.426 ± 0.059	0.387 ± 0.023
Training-Testing Time (Secs)		0.054	1352.7	11.767	11.048

[b] ROS: Random Oversampling, [c] RUS: Random Undersampling

For all methods, the best scores are obtained in random oversampling US dataset with bagging, shown in italic. Specifically, Logistic AT method has the best MSE score for ROS with bagging. In each variation of US dataset, performance of ordinal regression model varies: the best score (shown in bold): for original dataset ordinal ridge; the best score for ROS dataset is Logistic AT; and LAD has the best MSE score for RUS dataset. For UK dataset, ordinal ridge has the best MSE score with bagging and no bagging. When compare two datasets, UK accident data has the best MSE score among all ordinal regression models for all datasets. This is mainly because the UK dataset has three-class and feature values are not overlapping among classes. In other words, the feature values can better distinguish the classes. These data characteristics enabled prediction models to get better results. Confusion matrices for ordinal ridge method in both datasets are shown in Fig. 4 and Fig. 5. Each row represented the actual class and each column represented the predicted class and values in the matrices demonstrated the number of instances. As a result of ordered classes, they often confused with classes adjacent to them. In Fig. 4, class 0 and class 1 are confused with 655 number of instances, indicating Ordinal Ridge algorithm misclassified these classes. There is a diagonal trend in confusion matrices which is desired and expected from upper left to lower right. This trend means the classes are confused with the adjacent class next to them. Fig. 5 showed ordinal ridge classified class 0 and class 2 correctly with high accuracy scores. Specifically, class 1 was mostly confused with class 2. However, the algorithm classified class 0 well with relative low number of instances presented in the confusion matrix. Color-coded trend lies down from light orange to lilac, as seen in Fig. 4 and Fig. 5.

Training and testing time are provided in Table V. Ordinal ridge is faster comparing with other algorithms. This method also provides the best MSE score for both original datasets. LAD method is based on optimization technique, so the training and testing time are longer than other ordinal regression algorithms. Logistic AT and Logistic IT have reasonable running time, besides better MSE scores relatively. Solution times vary according to different datasets, their size and distributions of features, and experiment environment.

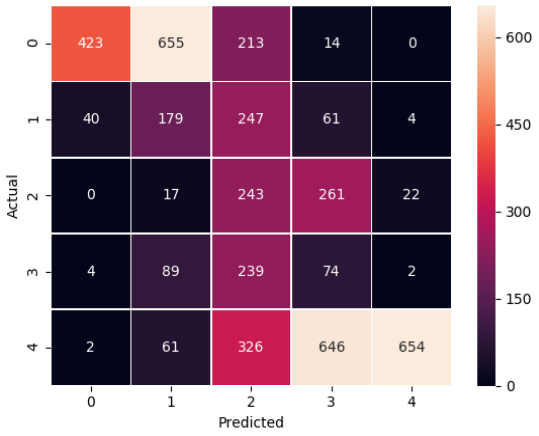

Fig. 4. Confusion matrix for Ordinal Ridge (US dataset)

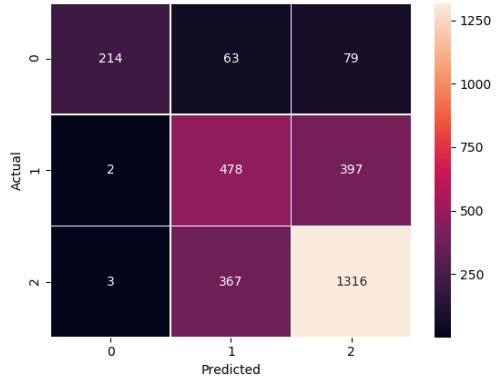

Fig. 5. Confusion matrix for Ordinal Ridge (UK dataset)

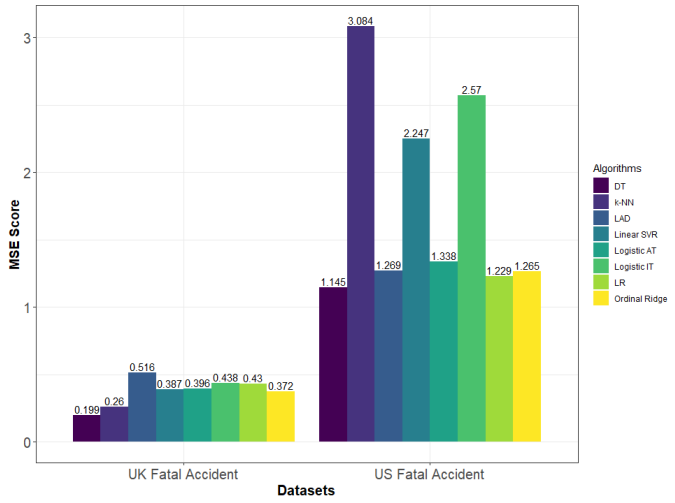

Fig. 6. The performance comparison of machine learning algorithms

Fig. 6. shows the comprehensive comparison between ordinal regression algorithms and other methods from literatures for both UK fatal accident and US fatal accident datasets. We conducted the comparision using k-NN, Decision Tree, Logistic Regression, and Linear Support Vector Regression algorithms. Most of the studies [4,5,11] in injury severity prediction used these algorithms for classification. For comparison, only original datasets with no bagging were considered. Fig 6 shows that UK dataset achieved better results than US dataset with lower mean square errors. This is mainly because the UK dataset (3 classes) has less classes than US dataset and the data features have more distinguished boundary in values among classes. In contrast, the US dataset was extracted from the National Highway Traffic Safety Administration website and carried out preprocessing and cleaning steps. It has 5 injury severity classes and overlapped feature values between classes. Decision Tree has the better performance with 0.199 and 1.145 mean squared error scores for both the UK and the US datasets, respectively. Specifically in the UK data, Logistic AT (0.396) and Ordinal Ridge (0.372) algorithms performed well than Logistic Regression (0.43) and Linear SVR (0.387), and LAD achieved the worst performance with 0.516 mean squared error score. In the US data, LAD (1.269), Logistic AT (1.338), and Ordinal Ridge (1.265) methods outperformed k-NN (3.084) and Linear SVR (2.247) methods. K-NN has the worst performance with 3.084 mean squared error. Logistic AT and Ordinal Ridge algorithms achieved better results among other ordinal regression algorithms for both datasets.

V. Conclusions

In this paper, we propose a novel accident risk assessment framework driven by ordinal regression. One challenge of the risk assessment problem is that non-accident data are typically not collected by agency in transportation safety. As the system needs both accident and non-accident data to train the classification model in order to prevent accidents and alert driver, we also propose a realistic negative data generation scheme based on feature weights derived from multinomial logistic regression to overcome this challenge. Thus, the model learned the pattern in traffic data properly. Experimental results on two real-world datasets from the US National Highway Traffic Safety Administration and UK Transport for Greater Manchester are used to demonstrate the feasibility and robustness of our proposed ordinal regression framework. Ordinal regression models played a significant role where class category exists in ordinal order. Predicting and assessing accident risks with ordinal regression is the main contributions of the paper.

Performance on four ordinal regression algorithms, namely: logistic all-threshold, logistic immediate-threshold, ordinal ridge, and least absolute deviations are compared. Since the prediction results will be used in the prevention and alerting system, original datasets should be considered in the comparison of ordinal regression performance. In this context, the Ordinal Ridge method provided the best MSE score and fastest prediction time. Bagging method, also called Bootstrap aggregating, which improved the ordinal regression models' performance is also implemented both accident datasets. In addition, we investigated the effect of random oversampling and undersampling on the proposed risk assessment framework.

We also conducted a comprehensive comparison between Ordinal Regression algorithms and other machine learning algorithms that are often used in injury severity classification. It has demonstrated that ordinal regression algorithms are usable when class of an accident dataset is ranked in ordinal order. Among ordinal regression algorithms, Logistic AT and Ordinal Ridge algorithms performed well.

The proposed prediction framework can be integrated into an accident prevention and alert system to be used by drivers. Furthermore, we also identified factors and situations that caused accidents and injuries which can contribute to the design of autonomous vehicles. For future work, spatiotemporal characteristics and driver behavior patterns can be examined with more comprehensive data. Moreover, other machine learning approaches like neural network will be integrated into our current study to ensure a more comprehensive comparison before the system integration.

References

[1] National Center for Statistics and Analysis. Motor Vehicle Crashes: Overview. Traffic Safety Facts Research Note. Report No. DOT HS 812 318, 2015.

[2] R. O Mujalli and J. D Oña, "A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks," Journal of Safety Research, vol. 42(5), pp. 317-26, 2011.

[3] Z Yuan, X Zhou, T Yang and J Tamerius, "Predicting traffic accidents through heterogeneous urban data: a case study," International Workshop on Urban Computing (KDD), 2017.

[4] H Jeong, Y Jang, P. J Bowman and N Masoud, "Classification of motor vehicle crash injury severity: A hybrid approach for imbalanced data," Accident; Analysis and Prevention, vol. 120, pp. 250-261, 2018.

[5] C Aci and C Ozden, "Predicting the severity of motor vehicle accident injuries in Adana-Turkey using machine learning methods and detailed meteorological data," International Journal of Intelligent Systems and Applications in Engineering, vol. 6(1), pp. 72-79. 2018.

[6] M Pérez-Ortiz, P.A Gutiérrez, C.R García-Alonso, L Salvador-Carulla, J. A Salinas-Perez, and C Hervás-Martínez, "Ordinal classification of

depression spatial hot-spots of prevalence," 11th International Conference on Intelligent Systems Design and Applications, pp. 1170-1175, 2011.

[7] F Fernández-Navarro, P Campoy-Muñoz, M.L Paz-Marin, C Hervás-Martínez and X Yao, "Addressing the EU sovereign ratings using an ordinal regression approach," IEEE Transactions on Cybernetics, vol. 43, pp. 2228-2240, 2013.

[8] S Landschoot, W Waegeman, K Audenaert, G Haesaert, and B.D Baets, "Ordinal regression models for predicting deoxynivalenol in winter wheat," Plant Pathology, vol. 62, pp. 1319-1329, 2013.

[9] X Gao and Y Feng, "Penalized weighted least absolute deviation regression," Statistics and Its Interface, vol. 11, pp. 79-89, 2018.

[10] M Chong, A. Abraham and M Paprzycki, "Traffic accident analysis using machine learning paradigms," Informatica, vol. 29, pp. 89-98, 2005.

[11] A Iranitalab and A.J Khattak, "Comparison of four statistical and machine learning methods for crash severity prediction," Accident; Analysis and Prevention, vol. 108, pp. 27-36, 2017.

[12] M Zhu, Y Li and Y Wang, "Design and experiment verification of a novel analysis framework for recognition of driver injury patterns: From a multi-class classification perspective," Accident analysis and prevention, vol. 120, pp. 152-164, 2018.

[13] F Xia, L Zhou, Y Yang, and W Zhang, "Ordinal regression as multiclass classification," International Journal of Intelligent Control and Systems, vol. 12(3), pp. 230-236, 2007.

[14] F. M Zahid and S Ramzan, "Ordinal ridge regression with categorical predictors," Journal of Applied Statistics, vol. 39(1), pp. 161-171, 2012.

[15] P McCullagh, "Regression models for ordinal data," Journal of the Royal Statistical Society. Series B (Methodological), vol. 42(2), pp. 109–142, 1980.

[16] A Shashua and A Levin, "Ranking with large margin principle: Two approaches," In Advances in Neural Information Processing Systems, vol. 15, 2003.

[17] J.D.M Renni Rennie, "Ordinal logistic regression," http://people.csail.mit.edu/jrennie/writing/olr.pdf , 2005.

[18] J.D.M Rennie and N Srebro, "Loss functions for preference levels: regression with discrete ordered labels," Proceedings of the IJCAI Multidisciplinary Workshop on Advances in Preference Handling, 2005.

[19] M Topal, E Eyduran, A.M Yaganoglu, A.Y Sonmez and S Keskin, "Use of ridge and principal component regression analysis methods in multicollinearity," Journal of Agricultural Faculty of Atatürk University, vol. 41(1), pp. 53-57, 2010.

[20] J. P Brooks and J.H Dulá, "The L1-norm best-fit hyperplane problem," Applied Mathematics Letters, vol. 26(1), pp. 51-56, 2012.

[21] National Highway Traffic Safety Administration, https://www-fars.nhtsa.dot.gov, (Accessed 18 February 2019).

[22] UK Transport for Greater Manchester, https://data.gov.uk/, (Accessed 20 December 2019).

[23] L Breiman, "Bagging predictors," Machine Learning, vol. 24(2), pp. 123-140, 1996.

[24] C Chen, G Zhang, R. A Tarefder, J Ma, H Wei and H Guan, "A multinomial logit model-Bayesian network hybrid approach for driver injury severity analyses in rear-end crashes," Accident; analysis and prevention, 80, 76-88, 2015.