applied sciences

MDPI

Article

Factor Identification and Prediction for Teen Driver Crash Severity Using Machine Learning: A Case Study

Ciyun Lin [1], Dayong Wu [2,*], Hongchao Liu [3], Xueting Xia [3] and Nischal Bhattarai [3]

[1] Department of Traffic Information and Control Engineering, Jilin University, Changchun 130022, China; linciyun@jlu.edu.cn
[2] Texas A & M Transportation Institute, Texas A & M University, College Station, TX 77843, USA
[3] Department of Civil, Environmental and Construction Engineering, Texas Tech University, Lubbock, TX 79409, USA; hongchao.liu@ttu.edu (H.L.); xueting.xia@ttu.edu (X.X.); nischal.bhattarai@ttu.edu (N.B.)
* Correspondence: J-Wu@tti.tamu.edu

Received: 8 January 2020; Accepted: 24 February 2020; Published: 2 March 2020



Abstract: Crashes among young and inexperienced drives are a major safety problem in the United States, especially in an area with large rural road networks, such as West Texas. Rural roads present many unique safety concerns that are not fully explored. This study presents a complete machine leaning pipeline to find the patterns of crashes involved with teen drivers no older than 20 on rural roads in West Texas, identify factors that affect injury levels, and build four machine learning predictive models on crash severity. The analysis indicates that the major causes of teen driver crashes in West Texas are teen drivers who failed to control speed or travel at an unsafe speed when they merged from rural roads to highways or approached intersections. They also failed to yield on the undivided roads with four or more lanes, leading to serious injuries. Road class, speed limit, and the first harmful event are the top three factors affecting crash severity. The predictive machine learning model, based on Label Encoder and XGBoost, seems the best option when considering both accuracy and computational cost. The results of this work should be useful to improve rural teen driver traffic safety in West Texas and other rural areas with similar issues.

Keywords: teen driver; rural roads; crash severity; machine learning

1. Introduction

Teen drivers are at greater risk than any other drivers on the highway system in the United States, especially in the areas like West Texas with large rural road networks. According to the Texas Department of Transportation (TxDOT), the number of crashes in the West Texas rural regions has been significantly increasing in the past years (11,652 in 2016, 12,801 in 2017, and 14,609 in 2018). Among them, around 23% on average are involved with teen drivers, which is disproportionately higher than other age driver groups [1].

Rural roads commonly have unique hazards that include different geometric and access properties than urban roads, such as increased speed limits, less cover against adverse weather, an increased association with driver alcohol use, longer trip lengths, less enforcement and use of safety devices, underage driving in farm communities, and more [2]. The problem is often made severe for teen drivers, since they are particularly vulnerable to hazards on rural roads. This vulnerability is because teen drivers are typically immature and lack safe driving habits. They are less likely to wear safety restraints and more likely to speed, drive late at night, drive impaired, and transport teenage passengers [3].

In order to address rural teen driver safety issues, which has been raising public concerns recently, factors contributing to these problems need to be identified. It makes significant sense that an understanding of how the number of teen driver crashes might be reduced begins with an understanding of how they happened. In this article, machine learning approaches are used to analyze factors that affect the occurrence of teen driver crashes and resulting severity. A complete machine learning pipeline has been conducted from collecting/cleaning data, performing exploratory data analysis, and formulating a real-world problem into machine-learning predictive models. The study is performed in two steps:

- Step 1: Firstly, perform an Exploratory Data Analysis (EDA) on the collected dataset. The purpose of performing this EDA is to first to explore the collected raw dataset by using methods from descriptive statistics and data visualization. This EDA is done without any pre-conceived notions or hypotheses, and the results of this exploration is used to guide and to identify the factors in the subsequent machine learning (ML) models. Remove any features that are not related to crash severity by nature and any features with a large percentile of "no-data" or "missing data". Also remove features with notable positive correlation (e.g., object stuck and first harmful event, etc.). Lastly, select features based on the existing literature on crash severity and teen crashes and your best engineering judgement.
- Step 2: By performing data dimensionality reduction in the first step, build four machine learning models to identify the most important factors associated with the severity of crashes on rural roads and predict crash severity using these identified factors. Approach the problem as a multivariate regression problem and predict the severity of the crash based on the collected crash dataset.

The rest of the paper is organized as follows: Section 2 gives a thorough view of various literatures about teen driver crash severity and machine learning approaches applied in this area. Section 3 highlights the data collection/filtering process for this research. Section 4 performs an exploratory data analysis and the results from easily understandable visual presentations. Section 5 ranks the importance of factors on crash severity and building four machine learning models on crash severity prediction based on the teen driver crash data in rural West Texas alone. In the last section, a conclusion is discussed.

2. Literature Review

Crash severity analysis is an important and widely studied area in safety research. Past studies are usually focused on two important topics: (1) identification of risk factors associated with crash injury severity and (2) prediction of crash injury levels from the identified contributing variables.

In the first area, researchers in the past have investigated various factors that are contributing to crashes involved with teen drivers, such as speeding [4,5], distracted driving [6,7], driver age and gender [8,9], risky behavior [10], impaired driving [11], night-time driving [12], and peer passengers [13,14]. In the second field, many researchers have developed prediction models on crash severity, which can be generally divided into two groups: parametric models and nonparametric (or machine learning) methods. In crash severity analysis, the dependent variable is discrete, with multiple categories corresponding to different crash severity levels (for example, in this study, or in the TxDOT Crash Records Information System (CRIS), crash severity is defined with the KBACO scale: K: Killed, A: Incapacitating injury, B: Non-incapacitating injury, C: Possible injury, O: Not injured). Since crash severity levels can be considered as either nominal or ordinal variables, different types of statistical modeling techniques for nominal variables can be applied, such as multinomial logit (MNL) models (for example, investigating the impact of road features on crash injury severity [15], examining the impact of alcohol and drug use cyclist injury severities [16], and analyzing the severity of large-truck-involved crashes [17]) and ordered logit (OL) models (for example, investigating the impact of rural highway geometric features on crash injury severity [18], comparing the severity of

freight-involved crashes and non-freight crashes [19], and finding the impact of contributing variables on the severity for highway-rail crossing crashes [20]).

It should be noted that parametric models need some model assumptions and pre-defined relationships between dependent and explanatory variables, which restrict their application when the assumptions are violated [21]. Meanwhile, nonparametric or machine learning methods usually do not require such assumptions and can identify and present underlying knowledge in large datasets with diverse variable categories without priori model assumptions or predefined relationships. This makes machine learning approaches particularly useful for traffic safety data analysis. The machine learning technologies applied in the past research include: Random forest [22,23], SVM (Support Vector Machine) [24,25], neural networks [26,27], and XGBoost [28].

Some scholars also performed comparison studies on these two kinds of approaches. Their studies consistently show the advantages of machine learning approaches over traditional statistical modelling methods. For example, Wang and Kim performed a comprehensive comparison between discrete models and tree-based models [21,29].

As discussed previously, rural roads are particularly dangerous to teen drivers due to geographically isolated rural areas and longer travel distances within rural communities. Combined with inexperience and risky driving behaviors of teen drivers, the situation becomes even more serious, and a comprehensive analysis of teen driver crashes on rural roads is necessary. Texas has one of the largest rural road networks in the nation (nearly 57,000 centerline miles and over 128,000 lane miles of state highway in rural areas, including 50,000 miles of farm-to-market (FM) roads). Therefore, we perform a case study in West Texas rural areas in this study.

The major contribution of this study is to conduct a systematic and panoramic machine learning methodology from collecting, cleaning, and coding crash data, performing a comprehensive exploratory data analysis, identifying the effect of factors contributing to the injury severity, and formulating four machine learning models on crash severity prediction. We use West Texas rural teen driver crash data as an example case. The research findings are helpful to better understand teen driver crashes and reduce teen driver injury severities by adopting effective solutions in West Texas or any other rural areas facing similar issues.

It is worthy of note that the proposed methodology has no any specific geospatial coverage limit and could be applied to many other areas facing similar safety issues. The proposed method can also be applied to other driver groups, such as adults or elderly drivers. The authors hope that other researchers could benefit from this study by applying this systemic approach to their own crash data set and hence better understand their own safety issues. In addition, we compared the performance of four ML models in terms of data coding methods, accuracy, and efficiency. We believe these comparisons could provide some insightful ideas for other researchers when they choose an appropriate ML method to process their own crash dataset, particularly to select data coding methods.

3. Data Collection

3.1. Study Area

West Texas is a loosely defined part of the US state of Texas, generally encompassing the arid and semiarid lands west of a line drawn between the cities of Wichita Falls, Abilene, and Del Rio [30]. Although there is no consensus on the boundary between East Texas and West Texas, we included five geographical TxDOT districts in this definition based on four principal metropolitan areas: Lubbock, Abilene, Midland/Odessa, and San Angelo, as highlighted in the map of Figure 1. Considering the huge binational border crossing transportation activities between the US and Mexico, we don't include the TxDOT El Paso district in this West Texas focused research. The Lubbock district sponsored and managed this teen driver safety research as the home base for the researchers.

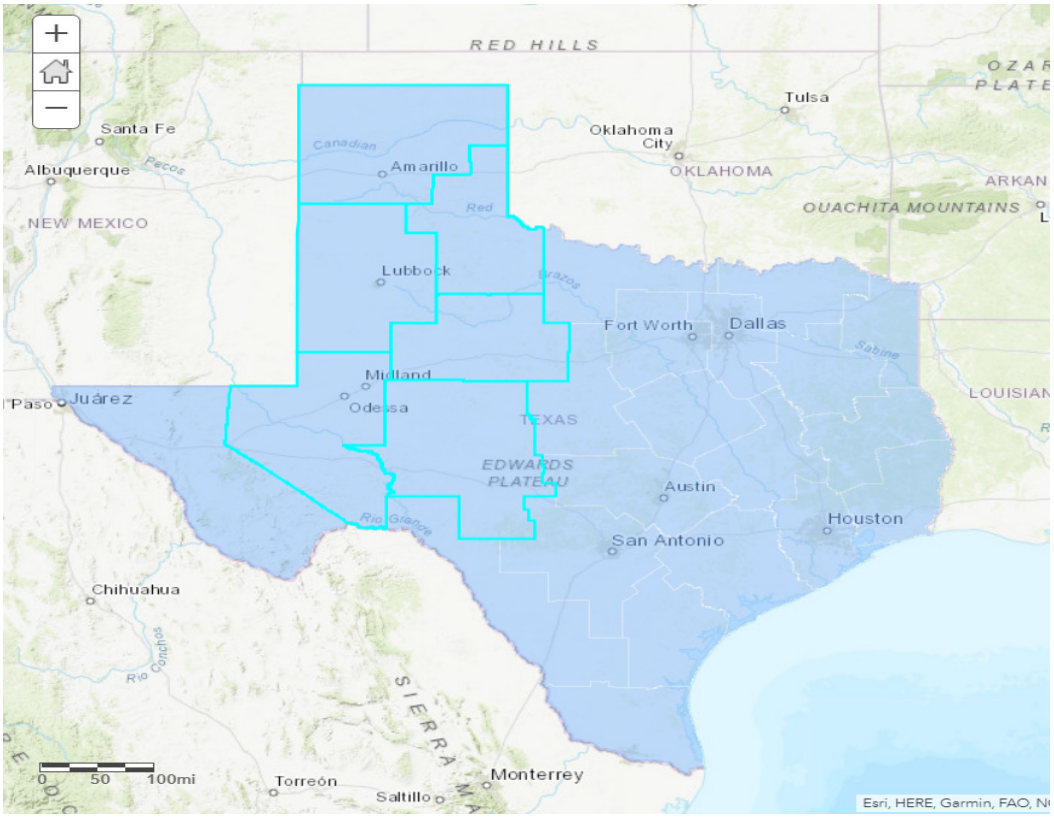


Figure 1. Study area: Five Texas Department of Transportation (TxDOT) districts in the West Texas rural area.

3.2. Crash Data

TxDOT maintains a statewide automated database for all reported motor vehicle traffic crashes received from the department of public safety, which is called the Crash Records Information System (CRIS). The CRIS database contain many attributes of crashes, such as crash identifier, crash date/time, location, facility, crash characteristics, human factors, environmental variables, and vehicle characteristics. Users may obtain publicly available crash data through the TxDOT's CRIS Query tool. This application allows public users to query, extract, and/or analyze publicly available crash data. In this study, this Query tool was used to firstly collect the crashes with the rural area flag in the selected five TxDOT districts during the period of 2010 to 2018, and then filtered them to get the rural crashes involved with teen drivers only (aged from 15 to 20). Note that we included the drivers of age 20 to represent the teen driver group with longer trip distances on rural roads. Finally, we have collected 8859 teen driver crashes in rural West Texas. The detailed descriptions of the collected data can be found in Table A1.

4. Exploratory Data Analysis (EDA)

A comprehensive EDA was performed for the collected rural teen driver crash data in three stages. The first stage was all about the detailed analysis of the causes of teen driver crashes, nature, and class. The second stage was about the trend analysis related to teen driver crashes, and the third stage focused on the demographics of teen drivers involved with the crashes. Finally, we applied the EDA for dimensionality reduction to the raw data so that further processing could be accomplished by great reducing the computational cost of the developed machine learning models.

4.1. Analysis of Major Causes

As observed in Figure 2, the major causes of teen driver crashes in rural West Texas are: unsafe speed (23%), failure to control speed (19%), failure to yield right of way (13%), driver inattention (6%), fatigue or sleep (5%), and animals on roads (6%). Other minor causes are mainly about risky driving behaviors, such as turning or changing lane when unsafe, following too closely, and backibg without safety, etc.

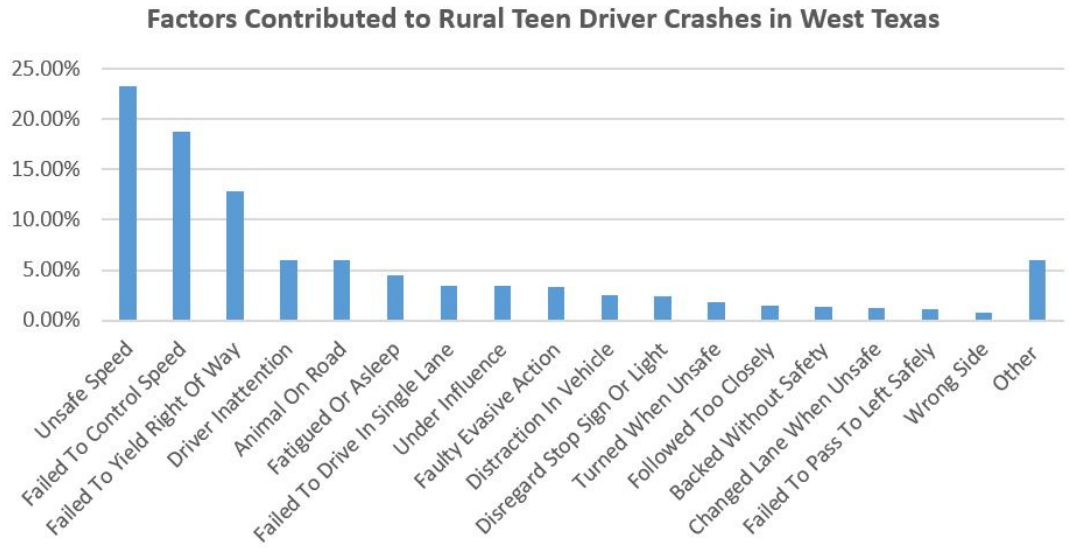


Figure 2. Major contributing factors to teen driver crashes in West Texas rural areas.

It is evident that speeding appeared to be the most important cause of teen driver crashes in West rural Texas regions (nearly 42% of crashes occurred due to high speed of the vehicle (unsafe speed and failure to control speed)). The yearly number and causes of teen driver crashes in West Texas rural areas can be found in Table A2.

Speeding increases the stopping distance of the vehicles while reducing reaction time to avoid a potential collision. It also increases the likelihood of injury in a crash. There is also evidence from naturalistic driving studies that teens' speeding behavior increases over time, possibly as they gain confidence. In general, they take more risks than adults and are more likely to involve speeding, which becomes more prevalent when driving on rural roads in West Texas at night or with other teen passengers on board. The problem is even worse since Texas has the highest speed limit on its rural highways in the nation (75 mph is a common speed limit setting and 80 or 85 mph will be allowed if the highway is designed to accommodate that speed.) It should also be noted that not all speed-related teen crashes are due to intentional risk-taking. Instead, most are caused by inexperience of teen drivers and a lack of driving skills. Teens or new drivers need to be educated on speed management based on traffic and road conditions and keep a safe distance from other vehicles.

The second important cause is mainly related to the improper driving behaviors of teen drivers in West Texas: failure to yield right of way and being distracted while driving (roughly 19% of the total crashes). In some cases, teen drivers of rural West Texas failed to yield to other vehicles while they were merging into the interstate highway, which resulted in crashes. A merging vehicle is usually required to execute a lane-changing maneuver along a limited length of a merge lane, which requires a good control while timing the merging and possesses a high chance of collision [31]. Therefore, teen drivers might not be experienced enough to perform the maneuver properly and get involved in crashes. Additionally, teens' inexperience behind the wheel makes them more susceptible to distraction of electronic devices or passengers in the vehicle.

The third important cause is from the unique hazards on rural roads: fatigue or asleep and animals on the road, which accounts for 11% of the overall crashes. One major concern for teen drivers on rural roads is highway hypnosis. Many rural highways in Texas consist of long, straight roadway sections. This fact, coupled with the propensity for teens to drive late at night, is a likely explanation for high fatal crash counts during night hours in Texas. Highway hypnosis can occur when drivers operate vehicles for extended periods of time in monotonous surroundings, entering a dulled and drowsy state. Low light and lack of sleep can contribute to the issue [2]. Additionally, in West Texas, animal crossings are common in rural terrains. Teen drivers often carelessly endanger themselves by overlooking the unforeseen hazards from animals. Wild animals, such as deer and domestic animals, escape from farms or ranches and dash into oncoming traffic. Inexperienced teen drivers often overcompensate when trying not to collide with the animal and, in fact, cause a collision. Other environmental issues coinciding with rural road animal accidents include lack of lighting and obstructions to visibility. It is important to teach teen drivers to be more cautious in areas with low visibility, more animals, and higher volume of slow-moving vehicles, such as farm machinery, in West Texas regions in the safety education program [2].

It is noted that driving under the influence is a minor cause for teen driver crashes in West Texas (only about 3% of the overall crashes). According to the National Highway Traffic Safety Administration (NHTSA), in 2016, alcohol-impaired driving crashes accounted for 18% of drivers involved in fatal crashes in the age group from 16–20 [32]. This indicates that the current Texas alcohol regulation that bans anyone under the age of 21 to purchase or consume alcohol successfully reduced the occurrences of teen driver alcohol-related crashes in West Texas.

4.2. Analysis of Trends

As overserved in Figure 3, the number of crashes involved with teen drivers in West Texas has experienced an overall increase in trend in the past nine years. The number almost doubled in 2018 compared with the number in 2010. However, according to a national study that examined the teen driver crash data in the entire 20-year period from 1994 through 2013, a substantial downward trend was presented in crash involvements of teen drivers [33]. In this regard, great efforts are needed to be devoted exclusively to teen driving safety in West rural Texas.

Additionally, as observed in Figure 4, there are two daily peaks for the occurrence of teen driver crashes on rural roads in West Texas: 7:00 to 8:00 in the morning and 16:00 to 17:00 in the afternoon. The researchers originally linked these two peaks with the daily peak hour traffic. However, these trends have not been found for other age driver groups. The researcher assumed that these two peaks are associated with teen driver traffic at school starting and ending time.

4.3. Other Exploratory Analysis

Other analyses were also performed to find the demographics of teen drivers and crash natures. As shown in Figure 5, male teen drivers were more likely to be involved in crashes than their female peers in rural West Texas (60% vs. 40%). In different ages, the occurrence trend of crashes was almost equally distributed (except the age of 15, only accounting for 2%). From the aspect of ethnicities, the majority of teen drivers were White and Hispanic, which was consistent with the demographics of residents in West Texas (almost 90% in total). As for the levels of crash severity, more than two thirds of the crashes did not end in injury. About 4% of teen drivers were involved with crashes with serious injuries and 1% were involved with fatal crashes, which was much lower than the average level of teen fatal crashes in the nation. In 2017, teen drivers involved in fatal crashes were 3225 in the nation, which was about 8% of total crashes [34].

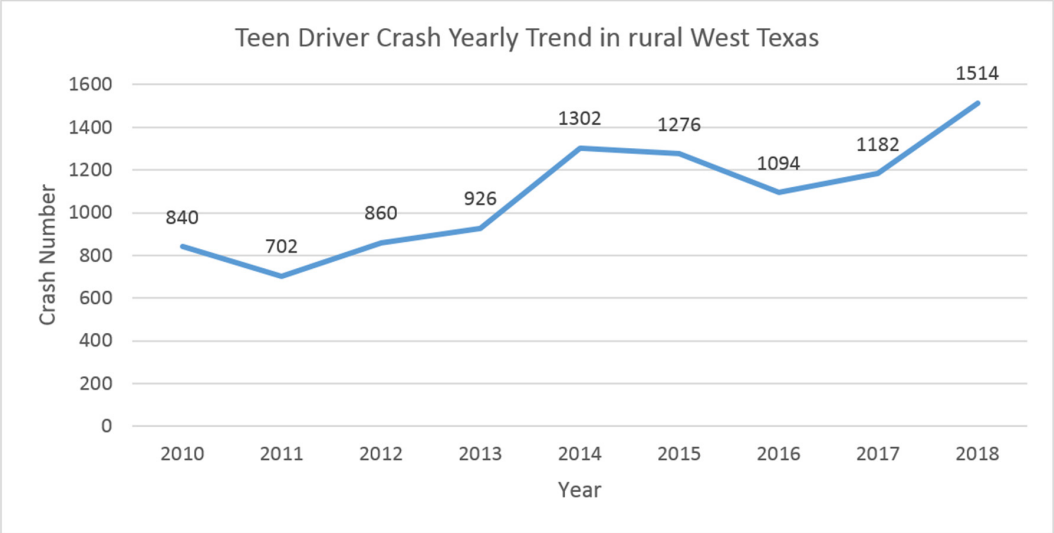


Figure 3. Yearly trend of teen driver crashes in the West Texas rural area.

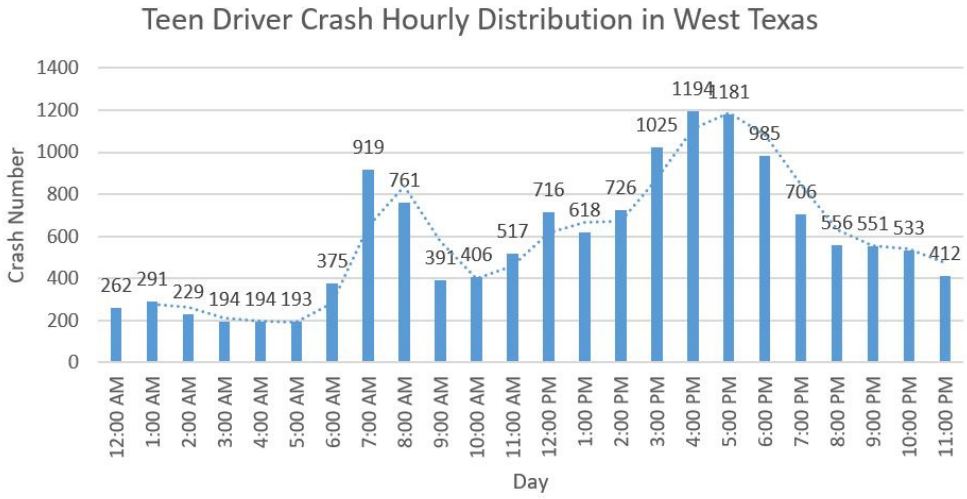


Figure 4. Hourly distribution of teen driver crashes in the West Texas rural area.

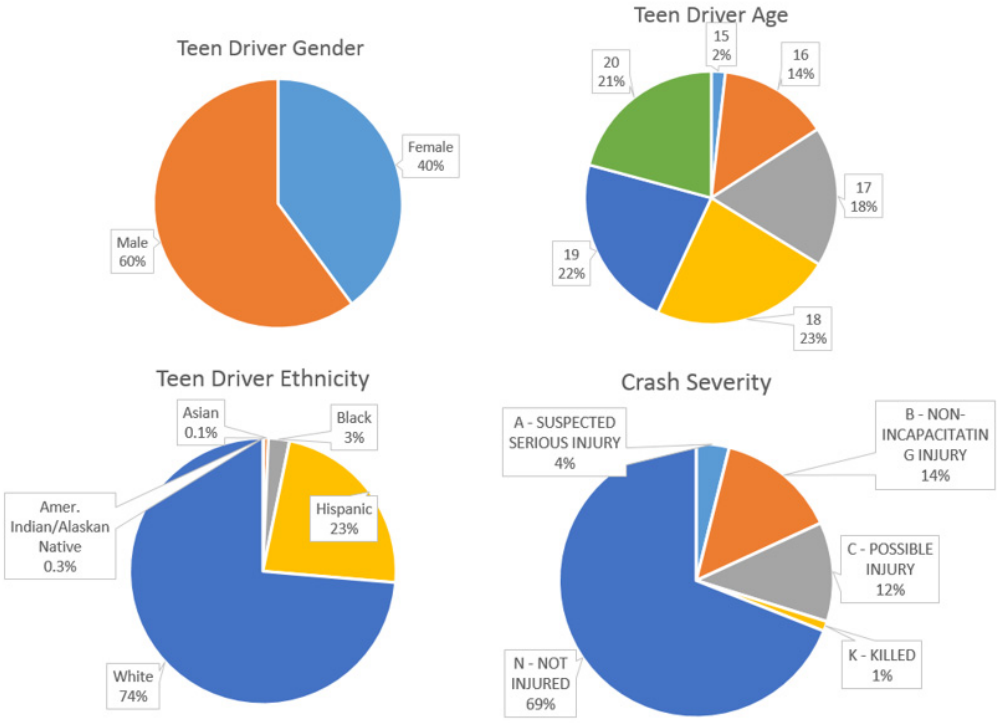


Figure 5. Other exploratory analysis on teen drivers and crash severity.

5. Factor Identification and Predictive Models

Predictive analytics is defined as finding patterns in data and predicting values in future trends and outcomes by using those patterns. Moreover, machine learning techniques have become more popular in accomplishing predictive analytics due to their exceptional usage in managing large scale data in recent years. In addition to the EDA, in this study, we built several machine learning models to rank the important factors associated with the severity of crashes and performed crash severity prediction using these identified factors from the EDA. The machine learning process was implemented using Python and the related Python data learning libraries, such as Pandas and SKlearn. The detailed steps are described as follows:

Step 1: W Cleaned and filter raw crash data from the CRIS. "No data" and "Unknown data" of crash severity are filtered out of the dataset. Choose the features based on the previous EDA and remove the features with the large percentile of missing data. In the end, we obtained 8859 valid crash records during the period of year 2010 to 2018. Each crash record had 46 features about the characteristics of crash, teen driver, vehicle, road, and driving environment.

Step 2: Since most of the features in the crash dataset are categorical, encode categorical data to numerical data for further construction of predictive models. Use two encoders that are available in the Python Sklearn library: label encoder and one hot encoder.

Label encoder assigns a non-repeat numerical value to each category in the data type of object. However, one weakness of the label encoder is that, by assigning different values to different categories, it gives an order to all categories, even though all categories are weighted equally and have no relationships between each category. One hot encoder splits one categorical feature into a binary sparse matrix. For example, in the categorical feature "Day of the week", there are seven different values representing each day of the week. One hot encoder splits the feature into a binary matrix with seven columns, where 1 represents true and 0 represents false for the value in each column. The differences between these two coding methods are illustrated in Figure 6.

Label Encoding

Days of week	Cateory	Crash Severity
Monday	1	0
Tuesday	2	1
Wednesday	3	2
Thursday	4	2
Friday	5	1

One Hot Encoding

Monday	Tuesday	Wednesday	Thursday	Friday	Crash Severity
1	0	0	0	0	0
0	1	0	0	0	1
0	0	1	0	0	2
0	0	0	1	0	2
0	0	0	0	1	1

Standard Scalar examples

AADT	AADT_Trucks	at_intersection_flag	crash_time	speed_limit	person_age
0.700426	1.64815	-0.624392	-0.765932	-1.116884	-1.68639
1.066361	1.338284	-0.624392	1.151681	0.273317	0.526138
0.464807	1.923587	-0.624392	-0.085319	0.620867	-0.211371
1.066361	1.338284	-0.624392	-0.904154	0.620867	1.263648
0.700426	1.64815	-0.624392	0.852492	0.273317	-0.211371

Figure 6. The coding and the standard scalar examples in this machine learning process.

After encoding the data, all data should be numerical in all cases, but the values of some continuous features are in large numbers by nature (such as the average daily traffic), which means it may have a higher impact on prediction than other features. Standardize all continuous features so that all variability is measured at the same scale. We used the Standardscaler function in the SKlearn library to normalize all the features in the dataset. The standard scalar examples are also shown in Figure 6.

Step 3: With the encoded and scaled data, start building predictive models for crash severity prediction. Here, approach the prediction as a multivariate regression problem and choose Random Forest (RF) and XGBoost as predictive methods. We also used the Python SKlearn library to implement the selected algorithms.

Random Forest is an ensemble machine learning algorithm that is used for classification and regression problems. It employs the decision tree algorithm for parametrization, but it mixes a sampling procedure, a subspace technique, and an ensemble tactic to optimize the model building. Bootstrap is the name of the sampling method, which uses a random sampling tactic with replacement. The subspace technique also takes a random sampling method, but it assists in removing smaller subsets (i.e., subspaces) of variables. The overfitting problem in decision trees is corrected in random decision forests by providing manifold. Use trained decision tree algorithms for the testing stage [21]. This property makes the RF preferable over the regular decision trees algorithm, since it was the most popular machine learning algorithm amongst data scientist until XGBoost took over.

XGBoost is an optimized distributed gradient boosting library, designed to be highly efficient, flexible, and portable. It implements machine learning algorithms under the gradient boosting framework. XGBoost provides a parallel tree boosting, also known as GBDT (Gradient Boosting Decision Tree), GBM (Gradient Boosting Machine), which solves many data science problems in a fast and accurate way. It has recently been dominating applied machine learning for structured or tabular data [28].

Since we had two predictive algorithms, combined with the two coding methods, we built four models for the crash severity perdition: Model 1 (one hot encoder/Random Forest); Model 2 (one hot encoder/XGBoost); Model 3 (label encoder/Random Forest); and Model 4 (label encoder/XGBoost).

Step 4: In this step, apply the four models to the West Texas rural teen driver crash dataset. The dataset is randomly split into training and testing datasets (75%/25%) to avoid any bias from the data itself. Finally, obtain the feature importance graph (as shown in Figure 7) and compare the performance measures in both accuracy and speed for the four models (as illustrated in Tables 1 and 2).

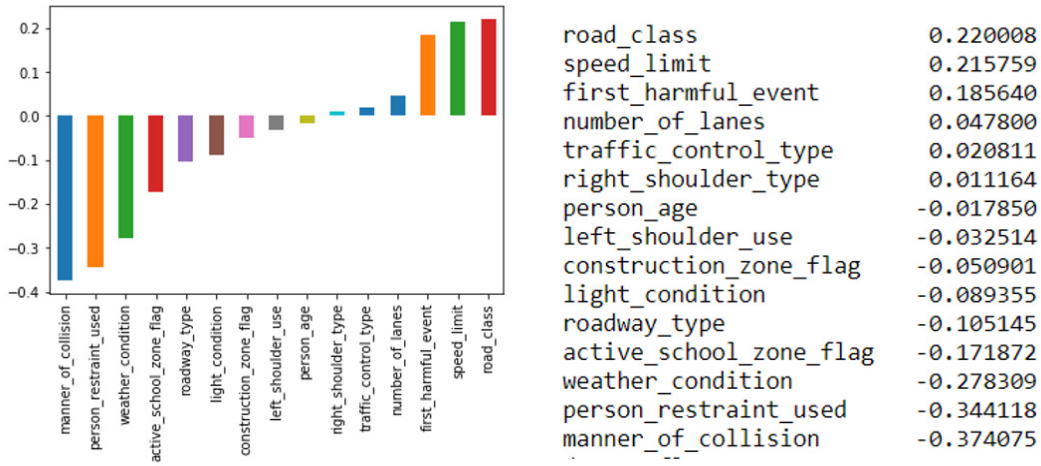


Figure 7. Ranking of feature importance on severity prediction of teen driver crashes in West Texas rural areas.

Table 1. The prediction accuracy of the developed four models.

Encoding Method	Predictive Algorithm	Mean Absolute Error (MAE)
One hot encoder	Random Forest	0.7271
One hot encoder	XGBoost	0.7140
Label encoder	Random Forest	1.0634
Label encoder	XGBoost	0.7804

Table 2. The prediction speeds of the developed four models.

Encoding Method	Predictive Algorithm	Time (second)
One hot encoder	Random Forest	3.51
One hot encoder	XGBoost	2.49
Label encoder	Random Forest	1.12
Label encoder	XGBoost	0.76

Note that, in Figure 7, the features with positive values mean that these features are the values that may lead to high-level crash severity (injury or even fatal crashes), while the features with negative values indicate that these are the values that may lead to low-level crash severity or even prevent the occurrence of crash injury. The features that rank highest in importance to crash severity prediction for West rural Texas teen driver crashes are: manner of collision, person restraint used, weather condition, school zone, roadway type, light condition, construction zone, left shoulder use, person age, right should type, traffic control type, number of lanes, speed limit, and road class.

In Tables 1 and 2, we can find that, for both coding methods, XGBoost has the highest accuracy (Model 2). As for the speed evaluation, Model 4 takes the shortest time to process the same dataset, which is critical for real-time crash prediction applications.

6. Conclusions

In this study, we presented a complete machine learning pipeline on teen driver crashes in rural West Texas, including collecting, cleaning, and coding crash data, a comprehensive exploratory data analysis, identification of contributing factors to crash severity, and four machine learning predictive models. We have the following major findings:

(1) The major causes of teen driver crashes in rural West Texas are unsafe speed (23%), failure to control speed (19%), failure to yield right of way (13%), driver inattention (6%), fatigue or sleep

(5%), and animals on the road (6%). Other minor factors are mainly about risky driving behaviors, such as turning = or changing lane when unsafe, following too closely, and backing without safety.

(2) The features that rank highest on importance to crash severity prediction for West rural Texas teen driver crashes are road class, speed limit, first harmful event, number of lanes, traffic control type, right shoulder type, person age, left shoulder use, construction zone, light condition, roadway type, school zone, weather condition, person restraint used, and manner of collision.

(3) Machine learning approaches are particularly useful to uncover new statistical patterns in the large heterogeneous crash datasets. XGBoost seems to be an effective option when considering both predictive performance and computational cost.

(4) It became apparent in this study that many teens do not have adequate knowledge of safe driving behavior on rural roads in West Texas. Policies and programs that can be built to curb this upward trend are in great need.

However, the dataset created contains information that has not been exploited yet, for example tempo-spatial patterns of teen driver crashes in rural West Texas. These will be included in further work. Additionally further time series analyses, for example, a comparison of the monthly time series through the years, as well as additional clustering experiments, for example, clustering of the day of week profiles by districts, might provide additional useful information. In cooperation with the TxDOT districts, we also hope to find explanations for some interesting features of the data discovered and possible ways to use the results of this work in improving rural teen driver traffic safety in West Texas.

Author Contributions: Conceived and designed the experiments: C.L., D.W., and H.L. Performed the experiments: X.X. and N.B. Analyzed the data: C.L., D.W., and H.L. Contributed reagents/materials/analysis tools: C.L., D.W., X.X., and N.B. Wrote the paper: C.L., D.W., and H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research is partly support by the National Natural Science Foundation of China with Grant No. 51408257 and the Youth Scientific Research Fund of Jilin with Grant No. 20180520075JH.

Acknowledgments: We are very thankful to the reviewers for their time and efforts; their comments and suggestions have greatly improved the quality of this paper.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. The descriptive of the collected teen driver crashes.

Variable Category		Frequency	Percentage
Commercial motor vehicle flag	No	8220	92.79
	Yes	639	7.21
First harmful event	Motor Vehicle in Transport	5355	60.45
	Fixed Object	2233	25.21
	Overturned	883	9.97
	Animal	240	2.71
	Other Object	61	0.69
	Parked Car	39	0.44
	Other Non-Collision	33	0.37
	Pedestrian	10	0.11
	Pedal cyclist	4	0.05
	RR Train	1	0.01
Highway lane design	Two-Way	5069	57.22
	Freeway	2078	23.46
	Boulevard	1287	14.53
	Expressway	425	4.80

Table A1. *Cont.*

Variable Category		Frequency	Percentage
Intersection related	Non-Intersection	5188	58.56
	Intersection	1774	20.02
	Intersection Related	1286	14.52
	Driveway Access	611	6.90
Median type	No Median	5069	57.22
	Unprotected	2791	31.50
	Positive Barrier	815	9.20
	Curbed	184	2.08
Road class	US and State Highways	3799	42.88
	Farm to Market	3076	34.72
	Interstate	1975	22.29
	City Street	9	0.10
Roadway function	Rural Prin Arterial	1655	18.68
	Rural Major Coll	1621	18.30
	Urban Prin Arterial (Other)	1336	15.08
	Rural Interstate	1280	14.45
	Rural Minor Arterial	1039	11.73
	Urban Prin Arterial (IH)	674	7.61
	Urban Minor Arterial	593	6.69
	Urban Collector	442	4.99
	Urban Prin Arterial (Other Freeway)	117	1.32
	Rural Minor Coll	95	1.07
	Urban Local	4	0.05
	Rural Local	3	0.03
Roadway type	2 Lane, 2 Way	4321	48.78
	Four or More Lanes, Divided	3781	42.68
	Four or More Lanes, Undivided	748	8.44
	Other Road Type	9	0.10
Vehicle body style	Pickup	3162	35.69
	Passenger Car, 4-Door	2967	33.49
	Sport Utility Vehicle	1475	16.65
	Passenger Car, 2-Door	1046	11.81
	Van	78	0.88
	Truck	56	0.63
	Unknown	27	0.30
	Truck Tractor	25	0.28
	Farm Equipment	8	0.09
	No Data	6	0.07
	Other (Explain in Narrative)	6	0.07
	Bus	1	0.01
	Ambulance	1	0.01
	Fire Truck	1	0.01
Light condition	Daylight	5946	67.12
	Dark, Not Lighted	2099	23.69
	Dark, Lighted	511	5.77
	Dusk	138	1.56
	Dawn	125	1.41
	Dark, Unknown Lighting	34	0.38
	Other (Explain in Narrative)	5	0.06
	Unknown	1	0.01
Crash severity	N-NOT INJURED	5936	67.01
	B-NON-INCAPACITATING INJURY	1307	14.75
	C-POSSIBLE INJURY	1077	12.16
	A-SUSPECTED SERIOUS INJURY	402	4.54
	K-KILLED	137	1.55

Table A1. *Cont.*

Variable Category		Frequency	Percentage
Surface type	Dry	6746	76.15
	Wet	1140	12.87
	No Data	520	5.87
	Ice	257	2.90
	Standing Water	79	0.89
	Snow	64	0.72
	Slush	32	0.36
	Sand, Mud, Dirt	10	0.11
	Other (Explain in Narrative)	8	0.09
	Unknown	3	0.03
Traffic control type	Center Stripe/Divider	3075	34.71
	Marked Lanes	2283	25.77
	Stop Sign	1187	13.40
	No Passing Zone	692	7.81
	None	623	7.03
	Signal Light	555	6.26
	Yield Sign	172	1.94
	Warning Sign	96	1.08
	Other (Explain in Narrative)	42	0.47
	Flashing Red Light	29	0.33
	Officer	27	0.30
	Signal Light with Red Light Running Camera	26	0.29
	Flashing Yellow Light	22	0.25
	Flagman	17	0.19
	RR Gate/Signal	6	0.07
	Inoperative (Explain in Narrative)	5	0.06
	Crosswalk	1	0.01
	Bike Lane	1	0.01
Weather condition	Clear	6528	73.69
	Cloudy	992	11.20
	Rain	979	11.05
	Snow	140	1.58
	Sleet/Hail	96	1.08
	Fog	76	0.86
	Blowing Sand/Snow	21	0.24
	Severe Crosswinds	16	0.18
	Other (Explain in Narrative)	8	0.09
	Unknown	3	0.03
Person gender	Male	5228	59.01
	Female	3593	40.56
	Unknown	38	0.43
Person ethnicity	White	6392	72.15
	Hispanic	1987	22.43
	Black	250	2.82
	Other	80	0.90
	Unknown	80	0.90
	Asian	55	0.62
	Amer. Indian/Alaskan Native	9	0.10
	No Data	6	0.07
Person restraint used	Shoulder and Lap Belt	8468	95.59
	None	213	2.40
	Unknown	141	1.59
	Shoulder Belt Only	19	0.21
	Lap Belt Only	8	0.09
	Not Applicable	6	0.07
	Other (Explain in Narrative)	2	0.02
	No Data	2	0.02

Table A2. Yearly number and causes of teen driver crashes in West Texas rural areas.

Contributing Factor	2010	2011	2012	2013	2014	2015	2016	2017	2018	Percentage
Unsafe Speed	251	239	221	298	290	306	250	272	253	23.09
Failed to Control Speed	173	162	183	203	195	232	220	266	287	19.79
Failed to Yield Right of Way	123	116	120	134	171	149	123	140	171	12.95
Driver Inattention	73	66	85	94	78	66	54	52	45	6.40
Animal on Road	84	70	65	73	61	64	63	64	63	6.36
Fatigued or Asleep	54	48	55	54	57	46	52	45	49	5.30
Under Influence	54	66	46	66	49	50	64	47	44	3.49
Failed to Drive in Single Lane	33	25	28	22	41	31	36	66	71	2.53
Faulty Evasive Action	29	43	49	45	51	39	33	18	30	2.42
Distraction in Vehicle	25	27	44	34	28	31	28	29	15	1.87
Disregard Stop Sign or Light	28	31	25	31	27	23	16	27	32	1.72
Turned When Unsafe	15	19	18	24	22	19	25	25	16	1.31
Taking Medication (Explain in Narrative)	19	19	19	19	19	19	19	19	19	1.23
Passed on Right Shoulder	18	18	18	18	18	18	18	18	18	1.16
Road Rage	18	18	18	18	18	18	18	18	18	1.16
Overtake and Pass Insufficient Clearance	17	17	17	17	17	17	17	17	17	1.10
Parked in Traffic Lane	17	17	17	17	17	17	17	17	17	1.10
Load Not Secured	16	16	16	16	16	16	16	16	16	1.03
Opened Door into Traffic Lane	16	16	16	16	16	16	16	16	16	1.03
Backed Without Safety	19	18	21	20	7	19	14	13	8	1.00
Ill (Explain in Narrative)	15	15	15	15	15	15	15	15	15	0.97
Impaired Visibility (Explain in Narrative)	15	15	15	15	15	15	15	15	15	0.97
Followed Too Closely	14	14	14	14	14	14	14	14	14	0.90
Changed Lane When Unsafe	9	9	17	19	17	7	17	11	18	0.89
Fire in Vehicle	13	13	13	13	13	13	13	13	13	0.84
Failed to Pass to Left Safely	13	17	16	11	15	10	7	12	8	0.78
Had Been Drinking	11	8	6	6	11	10	7	3	6	0.49
Speeding (Over limit)	10	5	13	9	6	8	11	4	2	0.49
Failed to Give Half of Roadway	6	4	7	1	6	7	8	8	10	0.41
Fleeing or Evading Police	7	7	4	3	5	5	2	8	9	0.36
Turned Improperly	3	3	8	7	3	1	4	6	10	0.32
Disregard Stop and Go Signal	2	3	2	2	5	5	5	4	11	0.28
Cell/Mobile Device Use	0	0	0	0	0	8	14	8	2	0.23
Passed in No Passing Lane	3	3	2	0	3	3	4	5	3	0.19
Failed to Pass to Right Safely	2	2	3	3	0	1	0	4	2	0.12
Improper Start from Parked Position	2	1	1	3	1	1	1	0	5	0.11
Failed to Stop at Proper Place	1	4	0	1	2	1	2	2	0	0.09
Failed to Signal or Gave Wrong Signal	1	2	2	0	2	3	1	0	1	0.09
Failed to Heed Warning Sign	1	3	2	0	1	1	0	1	0	0.06
Failed to Stop for Train	0	2	0	0	1	1	0	2	2	0.06
Disabled in Traffic Lane	0	0	2	0	0	0	0	2	1	0.04
Drove Without Headlights	1	1	0	0	0	0	0	1	1	0.03
Disregard Warning Sign at Construction	1	0	0	0	1	0	1	0	0	0.02
Wrong Way/One Way Road	0	1	1	0	0	1	0	0	0	0.02
Disregard Turn Marks at Intersection	0	0	1	0	0	0	1	0	0	0.01
Oversized Vehicle or Load	0	0	1	1	0	0	0	0	0	0.01
Failed to Stop for School Bus	0	0	0	0	0	1	0	0	0	0.01

References

1. TxDOT. TxDOT Crash Report Information System (CRIS). Available online: https://cris.dot.state.tx.us/public/Query/app/welcome (accessed on 30 July 2019).
2. Kumfer, W.; Liu, H.; Wu, D.; Wei, D.; Sama, S. Development of a supplementary driver education tool for teenage drivers on rural roads. *Saf. Sci.* **2017**, *98*, 136–144. [CrossRef]
3. Compton, R.P.; Ellison-Potter, P. *Teen Driver Crashes: A Report to Congress*; National Highway Traffic Safety Administration: Washington, DC, USA, 2008.
4. Ankem, G.; Gorman, T.; Klauer, C.; Ehsani, J.P.; Simons-Morton, B.; Gershon, P.; Dingus, T. An Objective Evaluation of Novice Teen Driver Speeding Behavior. In Proceedings of the Transportation Research Board 97th Annual Meeting, Washington, DC, USA, 7–11 January 2018.
5. Ferguson, S.A. *Speeding-Related Fatal Crashes among Teen Drivers and Opportunities for Reducing the Risks*; Governors Highway Safety Association: Washington, DC, USA, 2013.

6. Gershon, P.; Sita, K.R.; Zhu, C.; Ehsani, J.P.; Klauer, S.G.; Dingus, T.A.; Simons-Morton, B.G. Distracted driving, visual inattention, and crash risk among teenage drivers. *Am. J. Prev. Med.* **2019**, *56*, 494–500. [CrossRef] [PubMed]
7. Gershon, P.; Zhu, C.; Klauer, S.G.; Dingus, T.; Simons-Morton, B. Teens' distracted driving behavior: Prevalence and predictors. *J. Saf. Res.* **2017**, *63*, 157–161. [CrossRef] [PubMed]
8. Keating, D.P.; Halpern-Felsher, B.L. Adolescent drivers: A developmental perspective on risk, proficiency, and safety. *Am. J. Prev. Med.* **2008**, *35*, S272–S277. [CrossRef] [PubMed]
9. Rhodes, N.; Pivik, K. Age and gender differences in risky driving: The roles of positive affect and risk perception. *Accid. Anal. Prev.* **2011**, *43*, 923–931. [CrossRef] [PubMed]
10. Simons-Morton, B.; Lerner, N.; Singer, J. The observed effects of teenage passengers on the risky driving behavior of teenage drivers. *Accid. Anal. Prev.* **2005**, *37*, 973–982. [CrossRef]
11. Simons-Morton, B.; Li, K.; Ehsani, J.; Vaca, F.E. Covariability in three dimensions of teenage driving risk behavior: Impaired driving, risky and unsafe driving behavior, and secondary task engagement. *Traffic Inj. Prev.* **2016**, *17*, 441–446. [CrossRef]
12. Doherty, S.T.; Andrey, J.C.; MacGregor, C. The situational risks of young drivers: The influence of passengers, time of day and day of week on accident rates. *Accid. Anal. Prev.* **1998**, *30*, 45–52. [CrossRef]
13. Bingham, C.R.; Simons-Morton, B.G.; Pradhan, A.K.; Li, K.; Almani, F.; Falk, E.B.; Shope, J.T.; Buckley, L.; Ouimet, M.C.; Albert, P.S. Peer passenger norms and pressure: Experimental effects on simulated driving among teenage males. *Transp. Res. Part F Traffic Psychol. Behav.* **2016**, *41*, 124–137. [CrossRef]
14. Micucci, A.; Mantecchini, L.; Sangermano, M. Analysis of the Relationship between Turning Signal Detection and Motorcycle Driver's Characteristics on Urban Roads; A Case Study. *Sensors* **2019**, *19*, 1802. [CrossRef]
15. Penmetsa, P.; Pulugurtha, S.S. Modeling crash injury severity by road feature to improve safety. *Traffic Inj. Prev.* **2018**, *19*, 102–109. [CrossRef] [PubMed]
16. Kwigizile, V.; Sando, T.; Chimba, D. Understanding the role of alcohol and drugs on bicyclist injury severity. *Adv. Transp. Stud.* **2014**, *34*, 43–54.
17. Dong, C.; Dong, Q.; Huang, B.; Hu, W.; Nambisan, S.S. Estimating factors contributing to frequency and severity of large truck–Involved crashes. *J. Transp. Eng. Part A Syst.* **2017**, *143*, 04017032. [CrossRef]
18. Wu, Q.; Chen, F.; Zhang, G.; Liu, X.C.; Wang, H.; Bogus, S.M. Mixed logit model-Based driver injury severity investigations in single-And multi-Vehicle crashes on rural two-Lane highways. *Accid. Anal. Prev.* **2014**, *72*, 105–115. [CrossRef] [PubMed]
19. Taylor, S.G.; Russo, B.J.; James, E. A comparative analysis of factors affecting the frequency and severity of freight-Involved and non-Freight crashes on a major freight corridor freeway. *Transp. Res. Rec.* **2018**, *2672*, 49–62. [CrossRef]
20. Fan, W.D.; Gong, L.; Washing, E.M.; Yu, M.; Haile, E. Identifying and Quantifying Factors Affecting Vehicle Crash Severity at Highway-Rail Grade Crossings: Models and Their Comparison. In Proceedings of the Transportation Research Board 97th Annual Meeting, Washington, DC, USA, 10–14 January 2016.
21. Wang, X.; Kim, S.H. Prediction and Factor Identification for Crash Severity: Comparison of Discrete Choice and Tree-Based Models. *Transp. Res. Rec.* **2019**, *2673*, 640–653. [CrossRef]
22. Li, D.; Ranjitkar, P.; Zhao, Y.; Yi, H.; Rashidi, S. Analyzing pedestrian crash injury severity under different weather conditions. *Traffic Inj. Prev.* **2017**, *18*, 427–430. [CrossRef]
23. Mafi, S.; Abdelrazig, Y.; Doczy, R. Machine learning methods to analyze injury severity of drivers from different age and gender groups. *Transp. Res. Rec.* **2018**, *2672*, 171–183. [CrossRef]
24. Li, Z.; Liu, P.; Wang, W.; Xu, C. Using support vector machine models for crash injury severity analysis. *Accid. Anal. Prev.* **2012**, *45*, 478–486. [CrossRef]
25. Aghayan, I.; Hosseinlou, M.H.; Kunt, M.M. Application of support vector machine for crash injury severity prediction: A model comparison approach. *J. Civ. Eng. Urban.* **2015**, *5*, 193–199.
26. Zeng, Q.; Huang, H.; Pei, X.; Wong, S. Modeling nonlinear relationship between crash frequency by severity and contributing factors by neural networks. *Anal. Methods Accid. Res.* **2016**, *10*, 12–25. [CrossRef]
27. Sameen, M.; Pradhan, B. Severity prediction of traffic accidents with recurrent neural networks. *Appl. Sci.* **2017**, *7*, 476. [CrossRef]
28. Lee, D.; Warner, J.; Morgan, C. Discovering Crash Severity Factors of Grade Crossing With a Machine Learning Approach. In Proceedings of the 2019 Joint Rail Conference, Snowbird, UT, USA, 10–12 April 2019. [CrossRef]

29. Iranitalab, A.; Khattak, A. Comparison of four statistical and machine learning methods for crash severity prediction. *Accid. Anal. Prev.* **2017**, *108*, 27–36. [CrossRef] [PubMed]
30. Wikipedia. Available online: https://en.wikipedia.org/wiki/West_Texas (accessed on 30 July 2019).
31. Yang, H.; Ozbay, K. Estimating Traffic Control Risk Associated with Merging Vehicles on a Highway Merge Section. *Transp. Res. Rec.* **2011**, *2236*, 58–65. [CrossRef]
32. National Highway Traffic Safety Administration. Traffic Safety Facts 2016 data: Alcohol-Impaired Driving. Available online: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812630 (accessed on 30 July 2019).
33. Tefft, B.C. *Teen Driver Crashes: 1994–2013*; AAA Foundation for Traffic Safety: Washington, DC, USA, 2015.
34. NHTSA. Teen Distracted Driver Data. Available online: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812667 (accessed on 30 July 2019).



© 2020 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/).