International Journal of Technology 14(8) 1717-1727 (2023)
Received September 2023 / Revised December 2023 / Accepted December 2023

International Journal of Technology

http://ijtech.eng.ui.ac.id

Prediction of the Road Accidents Severity Level: Case of Saint-Petersburg and Leningrad Oblast

Angi Skhvediani[1*], Maria Rodionova[1], Natalia Savchenko[1], Tatiana Kudryavtseva[1]

[1]*Graduate school of industrial economics, Peter the Great St. Petersburg Polytechnic University, Saint – Petersburg, Russia, 195251*

Abstract. This article examines the factors influencing the severity of road accidents in St. Petersburg and Leningrad oblast for 2015–2023. The study is carried out on the analysis of 69190 road accidents and 6 groups of factors using the logit model and testing the oversampling technique to predict the probability of severe injuries and fatal cases after road accidents. The main factors in the study were lighting, deficiencies in road maintenance, and mean of transport. In particular, the logit model made for a joint sample on Saint – Petersburg and Leningrad oblast showed that the absence of lighting increases the probability of a serious accident by 19.6%, the presence of a vehicle such as a truck or motorcycle in a traffic accident increases the probability by 10.9%, and the presence of fog raises the probability by 17.6%. The usage of Synthetic Minority Over-sampling Technique (SMOTE) did not lead to a significant increase in the prediction accuracy of the models. The results of the study can be useful for organizing safe traffic in the city and providing recommendations for road users and public officials involved in improving the city's infrastructure.

Keywords: Logit model; Machine learning; Road safety; SMOTE; Traffic accident

1. Introduction

The analysis of the causes of road accidents is highly relevant, as the number of road accidents worldwide continues to increase, resulting in a significant number of injuries and deaths (Chang *et al.,* 2020). Hence, understanding the main causes and factors influencing the occurrence of road accidents is extremely important for developing effective measures to prevent them and reduce the number of victims on the roads. In addition, road accidents cause significant economic damage, which also makes this topic relevant for various countries and organizations (Zuraida and Abbas, 2020; Savolainen *et al.,* 2011). Therefore, this research topic is dedicated to numerous studies focused on developing effective measures to prevent road accidents, aiming to preserve the lives and health of individuals while also mitigating economic losses.

Many authors investigate the problem of road accident occurrence. Several works are based on statistical data collected by surveying respondents (Karim and Ali, 2020), here, authors assess the most influential factors influencing road accidents in Lebanon from data collected from a questionnaire designed using a Likert scale. In a work devoted to fatal accidents (Khurshid *et al.,* 2021), an analysis is carried out based on medical records of victims of road accidents. The authors of these works concluded that the most influential

*Corresponding author's email: shvediani_ae@spbstu.ru, Tel.: +7 (812) 775-05-30
doi: 10.14716/ijtech.v14i8.6859

factor among human factors is "Non-compliance with driving rules," followed by "Inexperience in driving," followed by "Drowsiness and fatigue."

Over the past five years, a significant amount of research has been carried out on the causes and consequences of road traffic accidents in various countries around the world. J. Brown's study looked at recent studies of traffic accidents in the United States. The authors found that factors such as distracted driving, speeding, and alcohol consumption are the leading causes of accidents on American roads (Brown *et al.,* 2017).

With the use of mathematical statistics in the analysis of road accidents, many scientists have tried to determine the causes of road accidents from different points of view, so let us consider the methods of data analysis used in various studies.

Various machine learning methods are used in many works, for example, in the articles (Santos *et al.,* 2021; Lin, Wang, and Sadek, 2014; Bohn *et al.,* 2013). Also, the logit model is used in many papers (Gilani *et al.,* 2021). It uses multiple logistic regression to determine the effect of each independent variable on the accident severity. In addition, this method is used in other papers (Milton, Shankar, and Mannering, 2008; Al-Ghamdi *et al.,* 2002). In addition, machine learning methods are used in work, where the influence of the condition of the road surface and the speed characteristics inherent in certain vehicles are analyzed (Siregar and Yusuf, 2022). However, the authors who investigate accident severity highlight unbalanced data for the output variable. Severe and fatal cases much less, than slight one (Wei, Zhang, and Das, 2023; Morris and Yang, 2021; Chen, Chen, and Ma, 2018). To address this issue, they employ various methods before modeling, such as the SMOTE method, clustering analysis, and data undersampling techniques, among others. An example of using the SMOTE method can be the works (Mostafa, Salem, and Habashyis, 2022) and (Mehrannia *et al.,* 2023), where using this method the sample was balanced, and further model construction was carried out. The method of synthetic oversampling of the minority was also used in the works (Shirwaikar *et al.,* 2022) and (Sobhana *et al.,* 2022) devoted to the analysis of the road accident severity levels.

Therefore, the aim of the research is to estimate the effect of different factors on the accident severity level in Saint – Petersburg and Leningrad oblast for 2015 – 2023 considering the problem of unbalanced data.

The paper is organized in the following way:

1. Description of the data and research methods (Chapter 2, "Data and methods").
2. Obtained results and their discussion with the other authors' results (Chapter 3 "Results and discussion").
3. Conclusions of the research (Chapter 4 "Conclusions").

2. Data and Methods

Healthcare To conduct the study, data on road traffic accidents that occurred in St. Petersburg and the Leningrad region from 2015 to May 2023 was obtained from Karta DTP as well as from the earlier study (Rodionova, Skhvediani, and Kudryavtseva, 2021). The research sample consists of 69,190 observations. For the analysis, we divide it into training and test sets in the proportion of 33% for the test sample (15,502 observations for Saint – Petersburg and 7332 for Leningrad Oblast) and 67% (31,472 observations for Saint – Petersburg and 14884 Leningrad Oblast) (Figure 1).

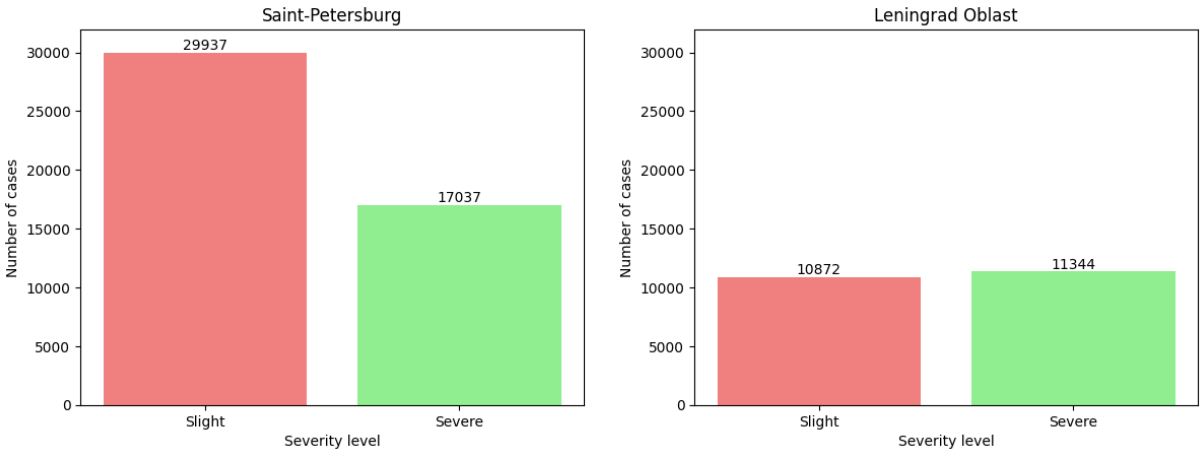
The study examines the dependent variable – accident severity level, that is binary variable (severe and slight accidents), and the influence of independent variables (Table 1) on the severity level. The independent variables were selected based on previous studies.

Table 1 presents the categories of factors influencing the severity level of road accidents, with the authors examining similar factors and employing variables used in their research.

Table 1 Independent variables

Number	Factor	Authors	Values
1	Illumination	(Mostafa, Salem, and Habashyis, 2022; Azhar *et al.,* 2022)	Daylight_hours, Dark_light_on, Twilight, light_Dark_light_absent)
2	Weather	(Elassad *et al.,* 2023; Azhar *et al.,* 2022)	Clearly, Cloudy, Rain, Snowfall, Fog, Other
3	Vehicle color	(Eustace, Alanazi, and Hovey, 2019)	Black, Grey, Blue, Red, Brown, Many, Green, Yellow, Orange, Purple, Other
4	Type of accident	(Boo and Choi, 2022; Azhar *et al.,* 2022)	Collision, Hitting_pedestrian, Hitting_cyclist, Hitting_standing_vehicle, Hitting_obstacle, Hitting_animal, Passenger_fall, Rollover, Ran_of_road, Other
5	Road conditions	(Sobhana *et al.,* 2022; Azhar *et al.,* 2022)	Dry, Wet, Traffic_Management_Facilities (technical means of traffic management), RC_Road_signs (Disadvantages of road signs), RC_Winter_maintenance (Disadvantages of winter maintenance), Other
6	Type of vehicle	(Boo and Choi, 2022; Azhar *et al.,* 2022)	Individual_mobility(Individual mobility equipment), Other, Special_equipment, Public_Transport, TRUCKS, Motorcycle_Transport, Passenger_Cars

Saint – Petersburg subsample contains higher amount of cases with slight injuries comparing to the severe, while in Leningrad oblast this proportion is approximately equal. Therefore, in total sample we have much more accidents with slight injuries, than the accidents with severe accidents (including fatal ones). It means that we meet with the imbalance in the examined dataset.



(a) (b)

Figure 1 Severity level of road accidents

The logit model is used for the analysis since the output variable is binary. In addition, this method has been used by many authors of similar research (Gilani *et al.,* 2021; Shiran, Imaninasab, and Khayamim, 2021; Ahmadi *et al.,* 2020). The python language is used for the model implementation and analysis.

In logistic regression, the dependent variable is a logit, which is the natural log of the odds. This is presented in equation 1.

$$\log(odds) = logit(P) = \ln\left(\frac{P}{1-P}\right), \quad (1)$$

where P – probability.

Hence, a logit is a log of odds, and odds are a function of the probability. In logistic regression, we find the log odds (logit) is assumed to be linearly related to X (2).

$$logit(P) = a + bX \quad (2)$$

To interpret the logit model, logits is needed to be converted to probability. For this aim, marginal effects are estimated after logit model calculation. Marginal effects show the change in probability when the predictor or independent variable increases by one unit. For continuous variables, this represents the instantaneous change given that the 'unit' may be very small. For binary variables, the change is from 0 to 1.

For the estimation of the obtained prediction quality is used confusion matrix with the following metrics. Formula for accuracy metric presented by equation 3.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}}, \quad (3)$$

where TP – true positive prediction in confusion matrix; TN – true negative prediction; FP – false positive prediction; FN – false negative prediction.

But in our case to assess the quality of problems with multiple classes, we consider macro F1-score (short for macro-averaged F1 score). Formula for F1 score metric presented by equation 4.

$$\text{F1 score} = \frac{2*(\text{precision}*\text{recall})}{(\text{precision+recall})}, \quad (4)$$

where precision – positive predictive value or the fraction of relevant instances among the retrieved instances;

recall – sensitivity or the fraction of relevant instances that were retrieved.

Furthermore, the area under the receiver operating characteristic curve (ROC AUC) was computed from prediction scores.

Given the imbalance in our dataset, we implement an oversampling method to address the scarcity of instances related to severe accidents. The chosen approach is Synthetic Minority Over-sampling Technique (SMOTE), which involves generating synthetic elements in close proximity to the existing ones within the minority class. In order to see how prediction accuracy changes depending on the usage of the SMOTE algorithm and sample, we estimate logit models using subsamples for Saint – Petersburg and Leningrad oblast and combined sample. In addition, for each case, we conduct modeling using both initial data and oversampled data (Mehrannia *et al.,* 2023; Mostafa, Salem, and Habashyis, 2022; Shirwaikar *et al.,* 2022; Sobhana *et al.,* 2022).

For the comparison of the obtained models, the ROC curve is used, which is a graphical representation of the performance of a binary classifier at different classification thresholds. The curve plots the possible True Positive rates (TPR) against the False Positive rates (FPR). The area under the ROC curve is measured by the ROC-AUC score, which is a single number that summarizes the classifier's performance across all possible classification thresholds. ROC-AUC score shows how well the classifier distinguishes positive and negative classes. It can take values from 0 to 1. A higher ROC-AUC indicates better performance.

The SPB, SPB_SMOTE, LO, LO_SMOTE, SPBLO, and SPBLO_SMOKE models were considered, information on which is presented in Table 2. This table provides information on the analyzed data collected for St. Petersburg and the Leningrad Region, as well as combined data for these regions.

Table 2 The models in question

Model Number	Model Name	Sample	Number of observations	Number of synthetic observations	Total number of observations
Model0	SPB	Sample for St. Petersburg	46974		46974
Model1	LO	Sample for Leningrad Region	22216		22216
Model2	SPB_SMOTE	Sample for St.Petersburg using the SMOTE method	46974	12900	59874
Model3	LO_SMOTE	Sample for St.Petersburg using the SMOTE method	22216	472	22688
Model4	SPBLO	Combined sample for St. Petersburg and the Leningrad region	69190		69190
Model5	SPBLO_SMOTE	Combined sample for St. Petersburg and the Leningrad region using the SMOTE method	69190	12428	81618

3. Results and Discussion

3.1. Regression analyses

As mentioned earlier, the work is carried out using machine learning on the training sample, and then a prediction is made on the test sample. Thus, we look at how the working algorithm was trained and what results were obtained on test data. The results obtained are presented in Table 3.

Table 3 Estimation results of logit model for severity level prediction

Factors \ Models	SPb	SPb_Smote	LO	LO_Smote	SPb&LO	SPb&LO_ Smote
	Marginal eff.	Marginal eff.	Marginal eff.	Marginal eff.	Marginal eff.	Marginal eff.
Weather conditions (reference: clear)						
Cloudy	0.0124 (0.0293)	0.0082 (0.0251)	0.0071 (0.0415)	0.0169* (0.0576)	0.0056 (0.0322)	0.0067 (0.0215)
Rain	0.0634*** (0.0554)	0.0594*** (0.0484)	0.0531** (0.0780)	0.0609*** (0.0410)	0.0594*** (0.0237)	0.0617*** (0.0415)
Snowfall	0.0793*** (0.0829)	0.0709*** (0.0734)	0.0299 (0.0911)	0.0328 (0.0772)	0.0520*** (0.0454)	0.0424*** (0.0567)
Fog	0.2504 (0.4521)	0.1409 (0.4507)	0.1173 (0.2977)	0.1432** (0.0911)	0.2453*** (0.0613)	0.1775*** (0.2566)
Type of accident (reference: Collision)						
Hitting_ animal	0.1270 (0.9408)	0.0362 (0.9406)	0.0380 (0.1757)	0.0443 (0.2989)	0.0079 (0.2593)	0.0035 (0.1597)
Hitting_ pedestrian	0.1254*** (0.0313)	0.1083*** (0.0262)	0.0936*** (0.0493)	0.0995*** (0.1743)	0.1044*** (0.1781)	0.0847*** (0.0232)
Hitting_ cyclist	0.1490*** (0.1483)	0.2106*** (0.1359)	0.1535 (0.4346)	0.0593 (0.0490)	0.1380*** (0.0259)	0.1596*** (0.1278)
Hitting_ standing_ vehicle	0.1034*** (0.0627)	0.0756*** (0.0551)	-0.0017 (0.0916)	-0.0035 (0.4425)	0.0580*** (0.1370)	0.0438*** (0.0472)
Hitting_ obstacle	0.1636*** (0.0541)	0.1451*** (0.0472)	0.1353*** (0.0680)	0.1397*** (0.0930)	0.1637*** (0.0517)	0.1506*** (0.0381)
Passenger_ fall	-0.0789 ***(0.0617)	-0.1159*** (0.0527)	-0.2790*** (0.2201)	-0.2853*** (0.0674)	-0.1411*** (0.0417)	-0.1578*** (0.0501)
Rollover	0.0608 (0.1278)	0.0416 (0.1136)	0.1099*** (0.0737)	0.1184*** (0.2210)	0.1456*** (0.0564)	0.1083*** (0.0573)
Ran_ of_road	0.1428*** (0.1362)	0.0832*** (0.1235)	0.0678*** (0.0615)	0.0793*** (0.0726)	0.1463*** (0.0610)	0.1102*** (0.0493)
Other	-0.1013** (0.1818)	-0.2579*** (0.1813)	-0.1097** (0.2104)	-0.0744 (0.0609)	-0.1197*** (0.0528)	-0.2124*** (0.1379)

Table 3 Estimation results of logit model for severity level prediction (Cont.)

Models / Factors	SPb	SPb_Smote	LO	LO_Smote	SPb&LO	SPb&LO_ Smote
	Marginal eff.	Marginal eff.	Marginal eff.	Marginal eff.	Marginal eff.	Marginal eff.
Road conditions (reference: dry)						
Wet	0.0018 (0.0332)	-0.0035 (0.0285)	-0.0293** (0.0489)	-0.0385*** (0.2068)	-0.0075 (0.1418)	-0.0055 (0.0248)
Traffic_ Management_ Facilities	0.0700*** (0.0293)	0.0600*** (0.0254)	0.0634*** (0.0405)	0.0665*** (0.0487)	0.0664*** (0.0273)	0.0688*** (0.0215)
Road_signs	0.1054*** (0.1651)	-0.0419 (0.1675)	0.0149 (0.0883)	0.0360** (0.0402)	0.0646*** (0.0236)	0.0141 (0.0757)
Winter_ maintenance	-0.0197** (0.0450)	-0.0351*** (0.0387)	-0.0425*** (0.0557)	-0.0466*** (0.0877)	-0.0232*** (0.0786)	-0.0258*** (0.0316)
Vehicle color (reference: white)						
Black	0.0161** (0.0324)	-0.0159** (0.0276)	0.0119 (0.0464)	0.0142 (0.0555)	0.0182*** (0.0348)	-0.0063 (0.0238)
Grey	0.0008 (0.0344)	-0.0247*** (0.0292)	0.0014 (0.0485)	0.0065 (0.0463)	0.0078 (0.0264)	-0.0159*** (0.0250)
Blue	0.0230*** (0.0382)	-0.0130* (0.0328)	0.0128 (0.0509)	0.0133 (0.0482)	0.0252*** (0.0277)	0.0100 (0.0272)
Red	0.0006 (0.0437)	-0.0231** (0.0373)	0.0219 (0.0567)	0.0159 (0.0506)	0.0106 (0.0302)	-0.0103 (0.0310)
Brown	-0.0064 (0.0621)	-0.0837*** (0.0546)	0.0230 (0.0845)	0.0042 (0.0561)	-0.0021 (0.0342)	-0.0383*** (0.0453)
Many	0.0593*** (0.0755)	0.0226 (0.0681)	-0.0213 (0.1347)	0.0052 (0.0850)	0.0658*** (0.0496)	0.0171 (0.0611)
Green	0.0284** (0.0617)	-0.0221** (0.0539)	0.0362** (0.0681)	0.0381** (0.1338)	0.0551*** (0.0649)	0.0296*** (0.0414)
Yellow	0.0233 (0.0789)	-0.0520** (0.0721)	0.0228 (0.1108)	0.0355 (0.0672)	0.0358** (0.0452)	-0.0038 (0.0586)
Orange	0.0553*** (0.0960)	0.0158 (0.0864)	0.0565** (0.1112)	0.0350 (0.1088)	0.0531*** (0.0625)	0.0110 (0.0677)
Purple	0.0107 (0.1422)	-0.0994*** (0.1314)	0.0769** (0.1618)	0.0756** (0.1083)	0.0299 (0.0717)	-0.0115 (0.1003)
Other	-0.0644*** (0.0395)	-0.1006*** (0.0333)	-0.0225** (0.0577)	-0.0234** (0.1636)	-0.0438*** (0.1082)	-0.0708*** (0.0290)
Type of vehicle (reference: Passenger cars)						
Individual_ mobility	-0.1103*** (0.1385)	-0.1993*** (0.1288)	-0.1156 (0.4278)	-0.0149 (0.0667)	-0.1264*** (0.0409)	-0.1614*** (0.1220)
Special_ equipment	0.0877*** (0.0891)	0.0271 (0.0811)	0.0705*** (0.1105)	0.0713*** (0.0552)	0.0845*** (0.0340)	0.0541*** (0.0645)
Public_ Transport	0.0718*** (0.0596)	0.0389*** (0.0528)	0.0803*** (0.1076)	0.0885*** (0.1096)	0.0802*** (0.0687)	0.0563*** (0.0465)
TRUCKS	0.1186*** (0.0434)	0.1033*** (0.0381)	0.1066*** (0.0553)	0.1164*** (0.1041)	0.1197*** (0.0507)	0.1093*** (0.0308)
Motorcycle_ Transport	0.1411*** (0.0536)	0.1218*** (0.0475)	0.1276*** (0.0674)	0.1308*** (0.0548)	0.1217*** (0.0335)	0.1091*** (0.0380)
Other	-0.0734*** (0.0443)	-0.1091*** (0.0380)	-0.0173 (0.0560)	-0.0114 (0.0573)	-0.0531*** (0.0322)	-0.0677*** (0.0308)
Illumination (reference: daylight)						
Dark_ light_on	0.0356*** (0.0273)	0.0260*** (0.0234)	0.0178 (0.0481)	0.0210** (0.4350)	0.0194*** (0.1301)	0.0075 (0.0210)
Twilight	-0.0244 (0.0935)	-0.0940*** (0.0849)	0.0077 (0.1001)	0.0089 (0.0479)	0.0047 (0.0232)	-0.0399*** (0.0638)
Dark_ light_absent	0.2413*** (0.1155)	0.1974*** (0.1101)	0.1480*** (0.0477)	0.1532*** (0.1012)	0.1975*** (0.0678)	0.1959*** (0.0396)

significance level: *** p < 0.01, ** p < 0.05, * p < 0.1
Standard error in parentheses

Most of the coefficient estimates are significant and stable across all combinations of subsamples and generated data. For further analysis, we focus on model 5, which was built

using both Saint–Petersburg and Leningrad oblast observations and generated observations. Model 5 demonstrates that such weather conditions as rain, snowfall, and fog appeared to be significant at 0.01 level and increased the probability of severe outcomes by 6.17, 4.24, and 17.75%, respectively compared to the clear weather. In addition, compared to collision type of accident, the probability of having severe injuries increases by 8.47, 15.96, 4.38, 15.06, 10.83 and 11.02% in hitting pedestrian, cyclist, standing vehicle, obstacle, rollover, and exit from the road types of accidents respectively at 0.01 significance level, while this probability decreases by 15.78% in passenger fall type of accident. Next, the absence of specific traffic management facilities increases the probability of severe outcomes by 6.88%. Also, if accident participants used special equipment, public transport, trucks or motorcycling transport, then the probability of severe outcome was higher by 5.41, 5.63, 10.93 and 10,91% compared to the vehicle–vehicle type of collisions, while in vehicle personal mobility devise type of collisions probability of severe outcomes lower 16.14%. Finally, the absence of lightning at nighttime increases the probability of severe outcomes by 19.59%.

The prediction quality was estimated using a confusion matrix and classification metrics. The classification report is presented in Table 4. As is seen, the prediction accuracy for the joint sample is 62% and if the 0 class (slight severity) is predicted by 73%, the 1st class (severe and fatal accidents) is predicted by only 35% of the f1-score. If we implement the SMOTE method, our results are better for the 1st class but lower for the slight accidents. After adding synthetic data to the 1st class observations, we have increased the f1-score for the 1st class from 35% to 58% but decreased the f1-score metric for the 0 class (from 73% to 58%). Therefore, the average model accuracy is less than the previous (58%).

Figure 2 presents results of ROC-AUC scores, which provide an opportunity to compare different models. As is seen, the models with the SMOTE algorithm present better performance of ROC-AUC score in all three cases (SPB_SMOTE, LO_SMOTE, SPBLO_SMOTE), but not significantly.

Table 4 Classification report

Model	Severity Level	Precision	Recall	F1-Score	Accuracy
SPb	0	0.65	0.94	0.77	0.64
	1	0.53	0.12	0.20	
LO	0	0.57	0.56	0.56	0.57
	1	0.58	0.59	0.59	
SPB_SMOTE	0	0.58	0.59	0.58	0.58
	1	0.58	0.57	0.58	
LO_SMOTE	0	0.57	0.60	0.58	0.57
	1	0.57	0.54	0.56	
SPBLO	0	0.63	0.88	0.73	0.62
	1	0.59	0.25	0.35	
SPBLO_SMOTE	0	0.58	0.59	0.59	0.58
	1	0.59	0.58	0.58	

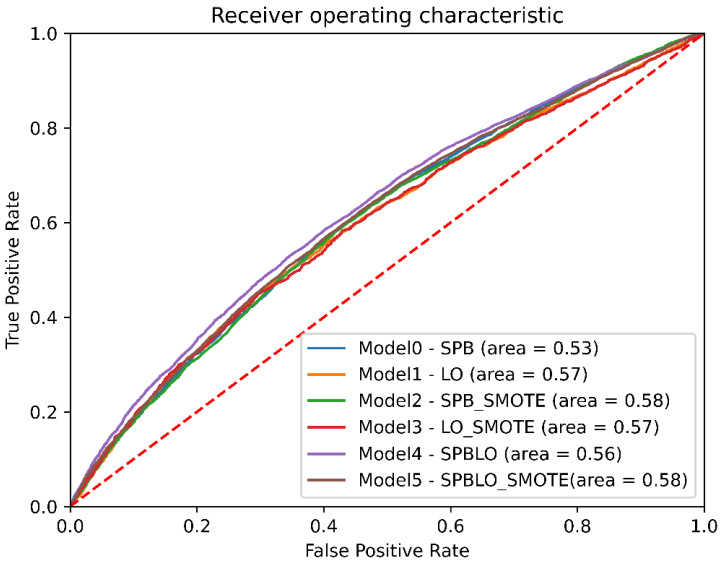


Figure 2 ROC AUC estimation results

3.2. Discussion

Results of the study are consistent with previous works. Deterioration of weather conditions leads to higher probabilities of severe injuries. This study finds that nighttime combined with the absence of lightning has a significant effect on crash severity. Behavior of traffic participants and visibility at nighttime are key factors, which contribute to higher probabilities of severe outcomes. The absence of lightning during dark hours significantly diminishes visibility and, consequently, the reaction time to avoid crashes (Riccardi *et al.,* 2023; Azhar *et al.,* 2022; Zhu, Li, and Wang 2018). Therefore, improved lightning conditions on the road at night can decrease the probability of severe outcomes.

This study finds foggy weather among the most influential factors contributing to severe outcomes in auto crashes. Foggy weather reduces visibility, limits contrast, and distorts perception. In heavy fog, drivers tend to perform more cautiously and reduce speed. However, it is usually not sufficient for the prevention of auto crashes with severe outcomes (Li, Yan, and Wong, 2015). Recent studies found that the usage of in-vehicle information systems can help drivers to adjust speed better at different road sections and, as a consequence, improve road safety (Calsavara, Kabbach, and Larocca, 2021).

The involvement of specific types of vehicles also influences the likelihood of severe outcomes in crashes. For instance, road accidents involving trucks tend to have a higher probability of severe outcomes compared to car–car accidents, primarily due to the larger mass and impact area (Chang and Chien, 2013). Participation in motorcycles also increases the probability of severe outcomes due to the higher tendency of motorcycles to speeding and reckless riding, lower safety of motorcycles, and pillion riders (Salum *et al.,* 2019).

Such types of collisions as hitting obstacles, rollovers, and running off-road also positively contribute to the crash severity levels and tend to have a higher probability of severe outcomes, which is consistent with (Roque, Moura, and Cardoso, 2015).

During the work, the problem of sample imbalance was identified since severe road accidents (including fatal ones) account for 40% of all observations. That is why the obtained results are tested using the SMOTE method, and the sample includes data from both metropolitan agglomeration (Leningrad oblast). According to the literature, usage of SMOTE may increase accuracy for severe or fatal outcomes by 15 – 25%, depending on the estimation method (Mohammadpour, Khedmati, and Zada, 2023). However, in our case,

there was not significant increase in the accuracy of the obtained results. A possible explanation may be the low number and level of detail of factors included in the model.

4. Conclusions

The main contribution of the article is the provision of a trained logit model for analyzing the influence of factors on the level of severity of road accidents and testing the oversampling technique. A model with a forecast accuracy of 63% and marginal effects for it is obtained. To increase the accuracy of the forecast, it is necessary to provide a more appropriate set of variables and test other options for constructing models. The results obtained can be useful to the state when building or implementing Traffic Management Facilities, building roads, and organizing traffic. In particular, the state can identify places of road accident concentration and to elaborate measures, which will decrease both probability of occurrence and severity level of road accident outcome. This analysis examines the influence of factors on road accidents in Saint – Petersburg and Leningrad oblast, but in the future, it is planned to continue the study by financial analysis of the risks of the budget from the municipality from the occurrence of an accident, thereby forming recommendations to the municipality on the effectiveness of financing infrastructure projects in the city. It is also planned to continue this research towards the development of a methodology for calculating the cost of human life since, at the moment, there is no single accepted methodology, and this issue directly affects the justification of investments in road transport infrastructure and other socially significant projects.

Acknowledgments

This research was funded by the Russian Science Foundation (project No. 23-78-10176, https://rscf.ru/en/project/23-78-10176/).

References

Ahmadi, A., Jahangiri, A., Berardi, V., Machiani, S. G., 2020. Crash Severity Analysis of Rear-End Crashes in California Using Statistical and Machine Learning Classification Methods. *Journal of Transportation Safety & Security*, Volume 12(4), pp. 522–546

Al-Ghamdi, A. S., 2002. Using Logistic Regression to Estimate the Influence of Accident Factors on Accident Severity. *Accident Analysis & Prevention*, Volume 34(6), pp. 729–741

Azhar, A., Ariff, N.M., Bakar, M.A.A., Roslan, A., 2022. Classification of Driver Injury Severity for Accidents Involving Heavy Vehicles with Decision Tree and Random Forest. *Sustainability*, Volume 14(7), p. 4101

Bohn, B., Garcke, J., Iza-Teran, R., Paprotny, A., Peherstorfer, B., Schepsmeier, U., Thole, C. A., 2013. Analysis of Car Crash Simulation Data with Nonlinear Machine Learning Methods. *Procedia Computer Science*, Volume 18, pp. 621–630

Boo, Y., Choi, Y., 2022. Comparison of Mortality Prediction Models for Road Traffic Accidents: An Ensemble Technique for Imbalanced Data. *BMC Public Health*, Volume 22(1), p. 1476

Brown, J.B., Rosengart, M.R., Billiar, T.R., Peitzman, A.B., Sperry, J.L., 2017. Distance Matters: Effect of Geographic Trauma System Resource Organization on Fatal Motor Vehicle Collisions. *The Journal of Trauma and Acute Care Surgery*, Volume 83(1), pp. 111–118

Calsavara, F., Kabbach Jr, F.I., Larocca, A.P.C. 2021. Effects of Fog in a Brazilian Road Segment Analyzed by a Driving Simulator for Sustainable Transport: Drivers' Speed Profile Under In-Vehicle Warning Systems. *Sustainability*, Volume 13(19), p. 10501

Chang, F.R., Huang, H.L., Schwebel, D.C., Chan, A.H., Hu, G. Q., 2020. Global Road Traffic Injury Statistics: Challenges, Mechanisms and Solutions. *Chinese journal of traumatology,* Volume 23(4), pp. 216–218

Chang, L.Y., Chien, J.T. 2013. Analysis of Driver Injury Severity in Truck-Involved Accidents Using a Non-Parametric Classification Tree Model. *Safety science*, Volume 51(1), pp. 17–22

Chen, F., Chen, S., Ma, X., 2018. Analysis Of Hourly Crash Likelihood Using Unbalanced Panel Data Mixed Logit Model and Real-Time Driving Environmental Big Data. *Journal of Safety Research*, Volume *65*, pp. 153–159

Elassad, Z.E.A., Ameksa, M., Elamrani Abou Elassad, D., Mousannif, H., 2023. Efficient Fusion Decision System for Predicting Road Crash Events: A Comparative Simulator Study for Imbalance Class Handling. *Transportation Research Record*, pp. 1–23

Eustace, D., Alanazi, F.K., Hovey, P.W., 2019. Investigation of the Effect of Vehicle Color on Safety. *Advances in Transportation Studies*, Volume 47, p. 69

Karim, F., Ali, S.IA., 2020. Evaluation of Most Influential Factors Affecting Road Traffic Accidents in Sidon, Lebanon*. Jurnal Kejuruteraan*, Volume 32(3), pp. 467–473

Khurshid, A., Sohail, A., Khurshid, M., Shah, M. U., Jaffry, A.A., 2021. Analysis of Road Traffic Accident Fatalities in Karachi, Pakistan: An Autopsy-Based Study. *Cureus,* Volume 13(4), p. e14459

Li, X., Yan, X., Wong, S. C. 2015. Effects of Fog, Driver Experience and Gender On Driving Behavior On S-Curved Road Segments. *Accident Analysis & Prevention*, Volume 77, pp. 91–104

Lin, L., Wang, Q., Sadek, A.W., 2014. Data Mining And Complex Network Algorithms For Traffic Accident Analysis. *Transportation Research Record*, Volume 2460(1), pp. 128–136

Mehrannia, P., Bagi, S.S.G., Moshiri, B., Al‐Basir, O.A., 2023. Deep Representation of Imbalanced Spatio‐Temporal Traffic Flow Data for Traffic Accident Detection. IET Intelligent Transport Systems, Volume 17(3), pp. 606–619

Milton, J.C., Shankar, V.N., Mannering, F.L., 2008. Highway Accident Severities and the Mixed Logit Model: An Exploratory Empirical Analysis. *Accident Analysis & Prevention*, Volume 40(1), pp. 260–266

Mohammadpour, S.I., Khedmati, M., Zada, M.J.H. 2023. Classification of Truck-Involved Crash Severity: Dealing with Missing, Imbalanced, and High Dimensional Safety Data. *PLoS one*, Volume 18(3), p. e0281901

Morris, C., Yang, J. J., 2021. Effectiveness of Resampling Methods in Coping With Imbalanced Crash Data: Crash Type Analysis and Predictive Modeling. *Accident Analysis & Prevention*, Volume 159, p. 106240

Mostafa, S.M., Salem, S.A., Habashyis, S.M., 2022. Predictive Model for Accident Severity. *International Journal of Computer Science*, Volume 49, pp. 110–124

Gilani, V.N.M., Hosseinian, S.M., Ghasedi, M., Nikookar, M. 2021. Data-Driven Urban Traffic Accident Analysis and Prediction Using Logit and Machine Learning-Based Pattern Recognition Models. *Mathematical problems in engineering*, Volume 2021, pp. 1–11

Riccardi, M.R., Mauriello, F., Scarano, A., Montella, A. 2023. Analysis Of Contributory Factors of Fatal Pedestrian Crashes by Mixed Logit Model and Association Rules. *International Journal of Injury Control And Safety Promotion*, Volume 30(2), pp. 195–209

Rodionova, M., Skhvediani, A., Kudryavtseva, T., 2021. Determinants of Pedestrian–Vehicle Crash Severity: Case of Saint Petersburg, Russia. *International Journal of Technology*, Volume 12(7), pp. 1427–1436

Roque, C., Moura, F., Cardoso, J.L. 2015. Detecting Unforgiving Roadside Contributors Through the Severity Analysis of Ran-Off-Road Crashes. *Accident Analysis & Prevention*, Volume 80, pp. 262–273

Salum, J. H., Kitali, A.E., Bwire, H., Sando, T., Alluri, P. 2019. Severity of Motorcycle Crashes in Dar Es Salaam, Tanzania. *Traffic injury prevention*, 20(2), pp. 189–195

Santos, D., Saias, J., Quaresma, P., Nogueira, V. B., 2021. Machine Learning Approaches to Traffic Accident Analysis and Hotspot Prediction. *Computers*, Volume 10(12), pp.157

Savolainen, P.T., Mannering, F.L., Lord, D., Quddus, M.A., 2011. The Statistical Analysis of Highway Crash-Injury Severities: A Review and Assessment of Methodological Alternatives. *Accident Analysis & Prevention*, Volume 43(5), pp. 1666–1676

Shiran, G., Imaninasab, R., Khayamim, R., 2021. Crash Severity Analysis of Highways Based on Multinomial Logistic Regression Model, Decision Tree Techniques, and Artificial Neural Network: A Modeling Comparison. *Sustainability*, Volume 13(10), p. 5670

Shirwaikar, R., KP, P., H Simha, 2022 A. Machine Learning Approach for Predicting Accident Severity. Machine Learning Approach for Predicting Accident Severity. Available online at SSRN: https://ssrn.com/abstract=4183574

Siregar, M.L., Tjahjono, T., Yusuf, N., 2022. Predicting the Segment-Based Effects of Heterogeneous Traffic and Road Geometric Features on Fatal Accidents. *International Journal of Technology*, Volume 13(1), pp. 92–102

Sobhana, M., Rohith, V. K., Avinash, T., Malathi, N., 2022. A Hybrid Machine Learning Approach for Performing Predictive Analytics on Road Accidents. *In*: 6th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)

Wei, Z., Zhang, Y., Das, S., 2023. Applying Explainable Machine Learning Techniques in Daily Crash Occurrence and Severity Modeling for Rural Interstates. *Transportation research record*, Volume 2677(5), pp. 611–628

Zhu, M., Li, Y., Wang, Y. 2018. Design and Experiment Verification of a Novel Analysis Framework for Recognition of Driver Injury Patterns: From A Multi-Class Classification Perspective. *Accident Analysis & Prevention*, Volume 120, pp. 152–164

Zuraida, R., Abbas, B.S., 2020. The Factors Influencing Fatigue Related to the Accident of Intercity Bus Drivers in Indonesia. *International Journal of Technology*, Volume 11(2), p. 342–352