Transportation Research Interdisciplinary Perspectives 32 (2025) 101495

Contents lists available at ScienceDirect

Transportation Research Interdisciplinary Perspectives

journal homepage: www.sciencedirect.com/journal/transportation-research-interdisciplinary-perspectives

Assessing the risk of traffic accidents in lisbon using a gradient boosting algorithm with a hybrid classification/regression approach

Nuno Alpalhão [*], Pedro Sarmento, Bruno Jardim, Miguel de Castro Neto

NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal

ARTICLE INFO

Keywords:
Traffic accidents
Emergency Response Planning
Urban Planning
Machine Learning
Risk Simulation

ABSTRACT

Traffic accidents significantly impact public health and economy through injuries, fatalities, and property damage. Effective emergency response planning requires sophisticated risk prediction tools with precise spatial and temporal resolution. While previous studies have assessed accident risk, they typically employed coarse spatial grids that lack the street-level detail crucial for emergency operations. This research presents a novel two-stage gradient-boosting predictive model, using tree-based learning algorithms to analyze traffic accidents requiring firefighter intervention in Lisbon, Portugal. To address the inherently unbalanced nature of accident data, we developed a sequential approach: first, a classification model identifies locations with non-zero accident probability; second, a regression model quantifies accident probabilities at street level across different time periods. The resulting risk simulator enables emergency planners to recalculate accident probabilities when street characteristics or weather conditions change, providing actionable insights for resource allocation and response planning. This research contributes both methodologically, through its innovative handling of spatially imbalanced data, and practically, by delivering an operational tool that supports evidence-based emergency service management. Validation results demonstrate the model's effectiveness in predicting high-risk locations and times, allowing for proactive deployment of emergency resources.

1. Introduction

Road traffic accidents constitute a major global health and economic challenge, causing significant social and economic losses through fatalities, injuries, and property damage. According to the World Health Organization (WHO, 2018), approximately 1.2 million people die annually and 50 million suffer non-fatal injuries due to road traffic accidents, which are projected to become the 7th leading cause of death worldwide by 2030.

To mitigate these devastating effects, emergency services must develop effective strategies to reduce response times and provide timely aid to injured individuals (Racioppi et al., 2004). Traffic accident prediction plays a crucial role in developing these strategies, enabling authorities to understand the factors influencing accident occurrence and anticipate where and when accidents are most likely to happen (Abdulhafedh, 2017; Verma et al., 2022).

Historically, modeling approaches for traffic accident prediction have focused primarily on identifying relationships between accidents and their influencing factors (Yassin & Pooja, 2020). Recent technological advancements in computing power and data availability have enabled machine learning models to gain prominence in this field. Numerous studies have employed these algorithms to both understand traffic accident factors and predict their occurrence (Campos et al., 2021; Fancello et al., 2020). These include decision tree classifiers and multilayer perceptrons (Taamneh et al., 2017), association rules mining frameworks (Ait-Mlouk and Agouti, 2019), neural networks with image processing techniques (Tanprasert et al., 2020), deep forests (Gan et al., 2020), and deep learning approaches (Shen et al., 2020).

Despite these advances, several critical gaps remain in current traffic accident prediction research:

1) Many studies utilize coarse spatial units, limiting their practical implementation by city planners and emergency services in operational and planning contexts (Ebrahimi et al., 2022; T. Huang et al., 2020);
2) Traffic accident datasets are inherently unbalanced, with significantly more spatial–temporal points lacking accidents than those

---

[*] Corresponding author.

E-mail addresses: nalpalhao@novaims.unl.pt (N. Alpalhão), psarmento@novaims.unl.pt (P. Sarmento), bjardim@novaims.unl.pt (B. Jardim), mneto@novaims.unl.pt (M. de Castro Neto).

https://doi.org/10.1016/j.trip.2025.101495
Received 5 December 2022; Received in revised form 17 May 2025; Accepted 16 June 2025

with accidents, yet studies often lack methodologies to address this imbalance, resulting in high false positive rates (Basso et al., 2018);

3) Computing limitations in processing high-volume data constrain the spatial and temporal granularity of proposed solutions (Bao et al., 2019);
4) As Yu et al. (2021) note, most traffic accident prediction studies focus on non-urban contexts, with minimal attention given to urban environments where accident patterns and influencing factors differ significantly.

To address these limitations, we propose a model that predicts traffic accident occurrence risk using a Light Gradient Boosting Machine (LightGBM) algorithm (Ke et al., 2017) for the city of Lisbon, Portugal. LightGBM is a decision tree framework based on the eXtreme Gradient Boosting (XGBoost) algorithm (T. Chen & Guestrin, 2016), offering faster training speeds while maintaining accuracy, particularly with sparse feature spaces (Li et al., 2022). LightGBM models have demonstrated effectiveness in urban contexts with strong result (Jardim et al., 2022).

Our approach tackles the unbalanced data challenge through a hybrid classification/regression methodology that provides traffic accident risk predictions at road segment level across multiple temporal partitions. This fine-grained spatial detail enables emergency services to optimize vehicle allocation and supports city planners in developing safer road infrastructure policies. Additionally, we developed a risk simulator prototype with descriptive and predictive capabilities to support emergency operations management and urban infrastructure planning.

The remainder of this paper is organized as follows: Section 2 presents a comprehensive literature review covering traffic accident influence factors and prediction methods; Section 3 describes the data sources and methodology, including our hybrid classification/regression modeling approach; Section 4 details the results and discussion, analyzing traffic accident risk factors and evaluating model performance; Section 5 introduces our risk simulator prototype and demonstrates its practical applications; and Section 6 provides conclusions, limitations, and suggestions for future research.

2. Literature review

2.1. Traffic accidents influence factors

Numerous studies have explored the factors influencing traffic accidents. Marcillo et al. (2022) identified several common data categories used in accident prediction research, including vehicle characteristics (Basso et al., 2018; Kiviluoto et al., 2022), driver attributes (Effati et al., 2014; Effati & Sadeghi-Niaraki, 2015), weather conditions (Wenqi et al., 2017; Xu et al., 2013), and light conditions (Effati et al., 2014; Xiong et al., 2018). Additional factors include traffic flow (Lin et al., 2015; Wenqi et al., 2017), traffic events (Y. Chen et al., 2020; Liu et al., 2021), road infrastructure characteristics (Bao et al., 2019; Effati & Sadeghi-Niaraki, 2015; Umair et al., 2022), taxi trip patterns (Bao et al., 2019; Zhu et al., 2019), points of interest (C. Huang et al., 2019; Paikari et al., 2014), and population density (Dong et al., 2015).

Other researchers have incorporated diverse contextual variables such as topographic maps, digital elevation models, land use patterns, satellite imagery, census block characteristics, special calendar dates, geographical features, trip surveys, and bicycle trip data. Given this extensive range of variables with varying contextual relevance, a thorough a priori evaluation of potential predictors is essential to determine their appropriateness for inclusion in predictive models.

For our study, we selected variables based on their demonstrated predictive power in urban environments and data availability for Lisbon. Road infrastructure characteristics (number of lanes, maximum speed, presence of radars and traffic lights), temporal factors (time of day, weekday/weekend distinction), and weather conditions (temperature, precipitation) were prioritized as they have shown consistent relationships with accident occurrence across multiple urban studies while being accessible at the required spatial and temporal resolution.

2.2. Traffic accidents prediction

The prediction of traffic accidents has evolved significantly through various methodological approaches (Vlahogianni et al., 2014). Traditional statistical methods include generalized linear models using Poisson and negative binomial distributions (Fancello et al., 2018). While these models offer interpretability and statistical rigor, they struggle with the non-linear relationships and complex interactions common in traffic accident data, particularly at fine spatial resolutions.

Time series approaches have also been widely applied, with autoregressive integrated moving average (ARIMA) and autoregressive integrated moving average with explanatory variables (ARIMAX) models tested by Ihueze & Onwurah (2018) and Smith et al. (2002), who found that incorporating human, vehicle, and environmental factors improved prediction accuracy. Seasonal autoregressive integrated moving average (SARIMA) models have demonstrated acceptable accuracy in predicting traffic accidents in Belgrade Deretić et al. (2022). However, time series models are primarily designed for temporal forecasting and lack the spatial component essential for street-level predictions in urban environments.

Machine learning techniques have gained prominence in traffic accident prediction research. Bayesian networks have been employed for real-time crash prediction, achieving 66 % accuracy for 250-meter road sections (Hossain & Muromachi, 2012) and 70 % accuracy for 22.2-kilometer road segments (Wu et al., 2019). While Bayesian networks effectively capture probabilistic relationships, they can become computationally prohibitive when scaled to city-wide street-level analysis.

Decision trees have proven valuable for establishing relationships between traffic accidents and various factors including road geometry, traffic characteristics, and environmental conditions (Chang & Chen, 2005), while also serving as feature selection tools to manage input data complexity (Lin et al., 2015). Their hierarchical structure makes them particularly suited for capturing non-linear interactions, but individual trees often suffer from high variance and potential overfitting.

Neural networks have been applied to forecast traffic accidents using traffic flow and weather data (Ozbayoglu et al., 2016; Vlahogianni et al., 2014; Wenqi et al., 2017). More recently, deep learning approaches have shown promise in predicting traffic accident risk (Bao et al., 2019; Q. Chen et al., 2016; T. Huang et al., 2020; Ren et al., 2018). While deep learning offers powerful representation capabilities, it typically requires substantial training data and computational resources, making it challenging to implement for fine-grained spatial analysis across an entire city network.

After evaluating these methodological alternatives, we selected the Light Gradient Boosting Machine (LightGBM) algorithm for our study due to several key advantages: (1) it efficiently handles sparse feature spaces common in street-level accident data; (2) its leaf-wise growth strategy allows for more nuanced capture of rare events, critical for addressing the class imbalance inherent in accident data; (3) it offers faster training speeds compared to other ensemble methods while maintaining prediction accuracy; and (4) it provides a framework amenable to our proposed hybrid classification/regression approach.

Our hybrid modeling approach specifically addresses the severe class imbalance challenge in accident prediction. Rather than forcing a single model to handle both classification (accident/no accident) and probability estimation simultaneously, we decompose the problem into two specialized models: first determining locations with non-zero accident probability, then quantifying those probabilities. This approach significantly reduces false positives—a common limitation in previous studies—while providing the fine spatial and temporal resolution required for operational emergency response planning.

The main limitation of our approach is the dependency on sufficient historical accident data for model training, particularly at street segments with varying characteristics. We address this through careful feature engineering and our two-stage modeling process, which maximizes the utility of available data while maintaining prediction reliability.

3. Data and methods

3.1. Traffic accidents data

We applied the proposed methodology to Lisbon, Portugal's capital city. Lisbon covers an area of 100 km2 with a resident population of 544,581 inhabitants (INE, 2021), which can increase by approximately 70 % due to daily commuters (INE, 2011).

The Lisbon City Council (CML) provided data on traffic accidents requiring intervention by Lisbon professional firefighters. The dataset spans from January 1, 2013, to December 15, 2020, comprising 6,076 traffic accident records. Each record contains information on: 1) the timestamp of occurrence; 2) accident type code and description (e.g., 2102 – accidents involving vehicles, 2103 – accidents with trapped individuals); and 3) geographical coordinates (latitude and longitude).

To identify spatial concentrations of traffic accidents across the entire temporal range, we employed Kernel Density Estimation (KDE) (Silverman, 1998) to estimate accident counts within a grid of 50 × 50 m cells covering Lisbon (Fig. 1). This visualization highlights accident hotspots throughout the city, particularly along major thoroughfares and intersections.

Fig. 1 illustrates the spatial distribution of traffic accident densities, revealing clear patterns of concentration along major transportation corridors, particularly at intersections of primary roads. The highest densities appear along the city's main east–west and north–south arteries.

The temporal distribution of accidents exhibits distinct patterns across different scales. Fig. 2 displays the monthly distribution of traffic accidents, showing higher frequencies between September and January. This seasonal pattern may relate to changing weather conditions, reduced daylight hours, and increased traffic during these months.

Fig. 3 presents the hourly distribution of accidents, differentiated between weekdays and weekends. The distribution reveals pronounced peaks during morning (7–10 h) and afternoon (17–20 h) rush hours. A notable difference between weekdays and weekends appears in the midday period, where weekdays maintain relatively high accident rates while weekend accidents decline significantly between morning and afternoon peaks.

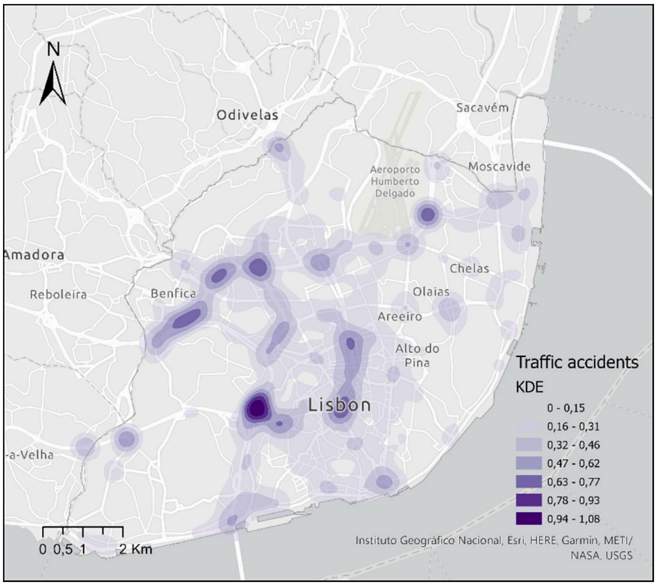


Fig. 1. KDE of traffic accidents estimated counts in a square grid of 50x50 m cells using traffic accidents data from 2013 to 2020.

3.2. Spatial and temporal data preprocessing

For model development, we collected relevant datasets based on contextual variables identified in previous traffic accident prediction studies. Table 1 summarizes these datasets, which were provided by CML and the Portuguese Institute for Sea and Atmosphere (IPMA). Some datasets are publicly available through Lisbon's Open Data Portal (Lisboa Aberta) (CML, 2018).

We aggregated traffic accidents, traffic jams (from Waze data), radar locations, and traffic light areas to their nearest road segments. Temperature and precipitation data from three weather stations were spatially interpolated to each road segment using linear interpolation weighted by distance—values from stations nearer to a road segment received greater weight than those from more distant stations.

Temporal variables (traffic accidents, temperature, and precipitation) were aggregated into seven time bins that provide semi-homogeneous distribution while capturing traffic peak hours: 1 = [0 h-4 h[; 2 = [4 h-7 h[; 3 = [7 h–10 h[; 4 = [10 h–14 h[; 5 = [14 h–17 h[; 6 = [17 h–20 h[; and 7 = [20 h–24 h[.

Fig. 4 displays the spatial distribution of traffic accidents after aggregation to their nearest road segments for the entire study period. This visualization presents a more detailed view than the KDE analysis, showing the specific road segments with the highest accident frequencies. Particularly high concentrations appear at major road intersections and along primary arterial routes.

3.3. Hybrid classification/regression modeling

Traffic accidents are rare events in both spatial and temporal dimensions, resulting in highly unbalanced datasets when aggregated for analysis. This imbalance presents significant challenges for modeling, typically leading to excessive false positive predictions. To address these challenges, we implemented a Light Gradient Boosting Machine (LGBM) model with a two-phase approach.

LGBM is a gradient boosting framework that ensembles small prediction models using tree-based learning algorithms. It grows trees leaf-wise based on the highest expected decrease in loss function. To optimize tree construction, LGBM implements two key techniques: Gradient-Based One-Side Sampling (GOSS), which selectively samples data instances with higher information gain potential, and Exclusive Feature Bundling (EFB), which efficiently reduces feature dimensionality in sparse feature spaces without sacrificing accuracy.

Our modeling strategy comprised two distinct phases. In the first phase, we calculated the probability of traffic accident occurrence by road segment for specific day periods, considering weather conditions (temperature and precipitation). Table 2 describes the variables used in this calculation.

For computing conditional probabilities, we employed Bayes' Theorem (Equation (1), which calculates the probability of an event based on prior knowledge related to the event (Joyce, 2021).

$$P(I_i|C) = P(C|I_i)P(I_i)/P(C) \tag{1}$$

In the second modeling phase, we addressed the statistical significance issues in our dataset. All variable permutations with occurrence counts less than one hundred were excluded from direct probability estimation due to their insufficient statistical significance. To estimate probabilities for these low-frequency permutations, we implemented the LGBM algorithm in two sequential steps:

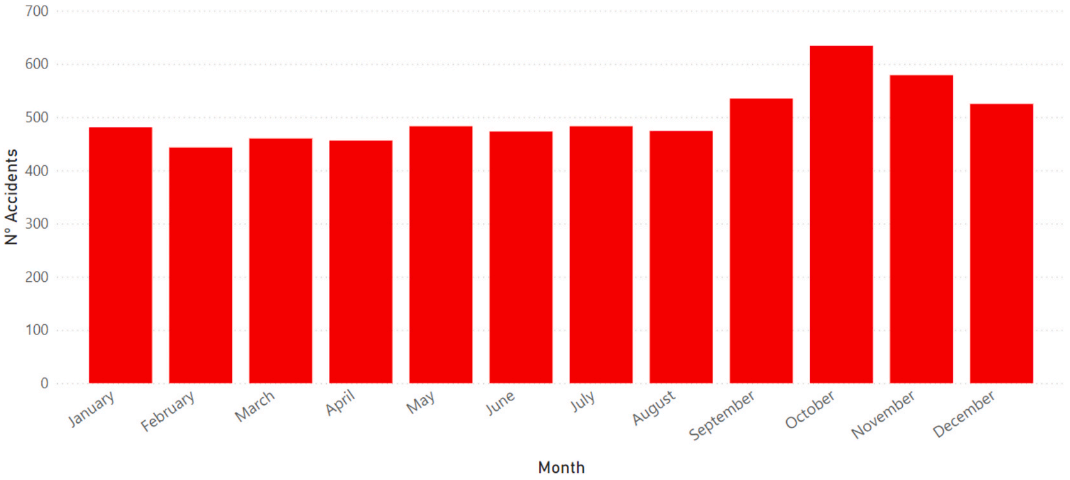


Fig. 2. Traffic Accidents number per month of the year.

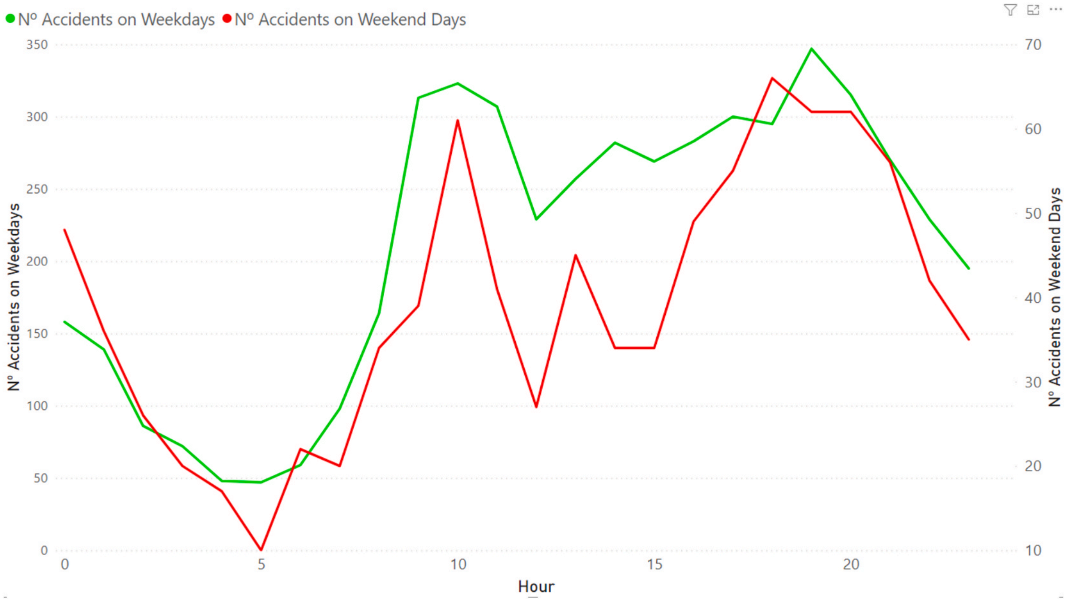


Fig. 3. Traffic Accidents number per hour of the day, segregated by week and weekend days.

Table 1

Relevant datasets used for the modeling phase.

Dataset	Description	Source	Open data
Traffic accidents occurrences	Location and time recorded of traffic accidents	CML	No
Weather data	Data on temperature and precipitation were collected hourly in three weather stations in Lisbon	IPMA	No
Roads	Road segments of the streets of Lisbon with the number of lanes, maximum velocity allowed, and slope	CML	No
Waze jams	The intensity of traffic jams in Lisbon collected by Waze	CML	No
Radars	Location of radars for velocity control in Lisbon	CML	Yes
Traffic light areas	Intersection areas with the presence of traffic lights	CML	Yes

1. Classification step: We trained a classification model to identify which permutations (among those with insufficient direct observations) had non-zero accident probabilities.
2. Regression step: For permutations identified as having non-zero probabilities in step 1, we trained a regression model to quantify the specific probability of traffic accident occurrence.

We divided the dataset into training (70 %) and testing (30 %) subsets for model development and evaluation. Table 3 presents the complete set of input variables used for predicting traffic accident probabilities in cases where direct statistical estimation was not possible due to insufficient observations.

To evaluate model performance, we calculated the Area Under Curve (AUC) (J. Huang & Ling, 2005) for the classification model and the Mean Absolute Percentage Error (MAPE) (de Myttenaere et al., 2016) for the regression model. These metrics provide comprehensive assessments of model accuracy for binary classification and continuous prediction tasks, respectively.

Fig. 5 illustrates the complete pipeline of our road accident risk calculation methodology, from data preparation through the two-phase modeling approach to final risk estimation. This workflow addresses the challenges of data imbalance while maintaining prediction accuracy at fine spatial and temporal resolutions. The modeling process begins with data preprocessing and aggregation to road segments and time periods.

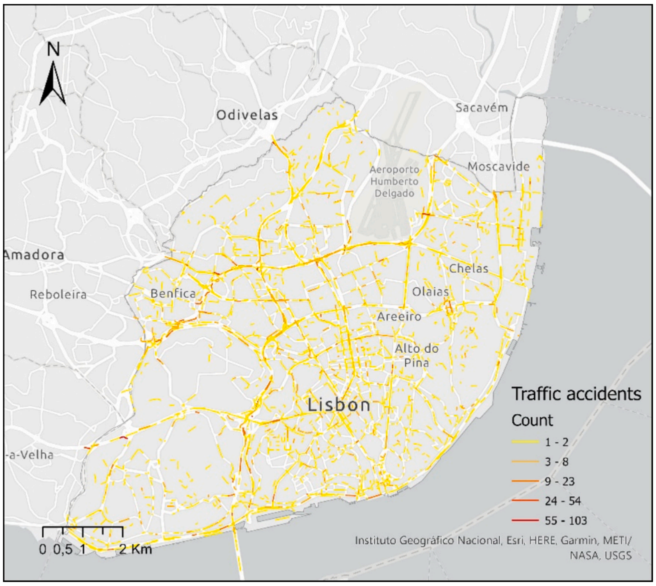


Fig. 4. The number of traffic accidents associated with the nearest road segment from 01/01/2013 and 15/12/2020.

Table 2
Features used for the computation of traffic accident probability.

Feature	Description	Type
road_id	Unique identifier of the road segment	Integer
temperature	Code for temperature recorded during a specific day period: 10 =]∞, 10 °C]; 20 =]10 °C – 20 °C]; 30 =]20 °C – 30 °C]; 40 =]30 °C – 40 °C]	Integer
precipitation	Code for precipitation recorded during a specific day period: 0.01 = [0 mm – 0,01 mm]; 2.5 =]0,01 mm – 2,5 mm]; 5 =]2,5 mm – 5 mm]; 10 =]5 mm – ∞[	Integer
period	Code to identify the period of day: 1 = [0 h – 4 h[; 2 = [4 h – 7 h[; 3 = [7 h – 10 h[; 4 = [10 h – 14 h[; 5 = [14 h – 17 h[; 6 = [17 h – 20 h[; 7 = [20 h – 24 h[	Integer
off_day	Flag identifying weekends and holidays: 0 = business day; 1 = weekend or holiday	Integer
count	group by the count of the combination of [road_id], [temperature], [precipitation], [period], and [off_day]	Integer

The first modeling phase computes direct probabilities for variable combinations with sufficient observations. The second phase employs the two-step LGBM approach (classification followed by regression) for combinations lacking sufficient observations. The outputs from both phases are combined to produce comprehensive risk estimations for all road segments across all time periods, which then feed into the risk simulator tool for practical applications.

4. Results and discussion

4.1. Risk of traffic accidents analysis

The risk of traffic accidents is associated with several factors. We analyzed the risk of traffic accidents for each road segment considering: 1) time of day; 2) temperature and precipitation; 3) maximum permitted velocity; 4) number of lanes; and 5) presence of speed enforcement radars and traffic lights. To facilitate comprehension of the computed probabilities, the values of traffic accident risk and standard deviations were multiplied by 100,000.

During weekdays, the risk of traffic accidents is higher between 10 h and 24 h, with the peak occurring between 17 h and 20 h. This peak period also exhibits higher dispersion of risk values, as evidenced by the elevated standard deviation. A similar pattern is observed on weekends, albeit with generally lower risk values. The only exception occurs between 0 h and 7 h, when the risk of traffic accidents is higher on weekends compared to weekdays (Table 4).

Table 3
Input features for the prediction of the probability of traffic accident occurrences in the observations were the combination of the features [road_id], [temperature], [precipitation], [period], and [off_day] is < 100.

Feature	Description	Type
road_id	Road segment unique identifier	Integer
road_name	Road Name	String
is_off_day	Flag identifying weekends and holidays: 0 = business day; 1 = weekend or holiday	Integer
temperature	Bin for temperature recorded: 10 =]∞, 10 °C]; 20 =]10 °C – 20 °C]; 30 =]20 °C – 30 °C]; 40 =]30 °C – 40 °C]	Integer
precipitation	Bin for precipitation: 0.01 = [0 mm – 0,01 mm]; 2.5 =]0,01 mm – 2,5 mm]; 5 =]2,5 mm – 5 mm]; 10 =]5 mm – ∞[	Integer
day_period_1	Boolean value identifying hour period [0 h – 4 h[	Integer
day_period_2	Boolean value hour period [4 h – 7 h[	Integer
day_period_3	Boolean value hour period [7 h – 10 h[	Integer
day_period_4	Boolean value hour period [10 h – 14 h[	Integer
day_period_5	Boolean value hour period [14 h – 17 h[	Integer
day_period_6	Boolean value hour period [17 h – 20 h[	Integer
day_period_7	Boolean value hour period [20 h – 24 h[	Integer
waze_proxy	Number of congestion events	Integer
lane_number	Number of lanes	Integer
vel_max	Maximum velocity accepted	Integer
comp	Road Length	Float
radar	Boolean value to identify radars	Integer
traffic light	Boolean value to identify if the road segment is in a traffic light area: 1 = road segment in a traffic light area	Integer
probability	Probability of the occurrence of a traffic accident	Float

The higher nighttime risk during weekends (0 h-7 h) likely reflects increased late-night social activities and potential alcohol consumption during weekend evenings, which are well-documented risk factors for traffic accidents. The afternoon peak (17 h-20 h) on both weekdays and weekends corresponds to rush hour traffic combined with reduced visibility during sunset, especially in winter months, and driver fatigue at the end of the workday.

The risk of traffic accidents is higher at temperature ranges of 10 °C to 20 °C and 30 °C to 40 °C, and when precipitation is between 0.01 mm and 2.5 mm (Table 5).

Given the significantly low number of conditions with precipitation levels above 2.5 mm, these classes were omitted from the analysis, as they were not deemed statistically significant.

The elevated risk during light precipitation (0.01–2.5 mm) compared to dry conditions aligns with previous research indicating that light rain can create particularly hazardous road conditions due to oil and dust mixing with water to create slippery surfaces. The fact that moderate temperatures (10 °C-20 °C) and high temperatures (30 °C-40 °C) both show increased risk suggests different mechanisms may be at play—reduced tire grip in colder temperatures and potential driver fatigue or discomfort in higher temperatures.

Table 6 presents the mean risk of traffic accidents conditioned on the maximum velocity permitted on each road segment. The mean risk is highest for road segments with maximum velocities of 70 km/h and 80 km/h, respectively. Road segments with a maximum permitted velocity of 40 km/h presented the lowest risk values.

These findings confirm the well-established relationship between higher speeds and increased accident risk, as documented extensively in road safety literature. The notable exception is the relatively low risk at 40 km/h compared to 30 km/h, which may be explained by the specific urban context of these speed zones; 40 km/h zones in Lisbon are often implemented in areas with better road design and clearer sight lines than the more confined 30 km/h zones, which are typically found in older neighborhoods with narrower streets.

Road segments with three lanes presented the highest mean risk value, followed by similar risk levels for road segments with four, five, and six lanes. Road segments with one lane presented the lowest mean

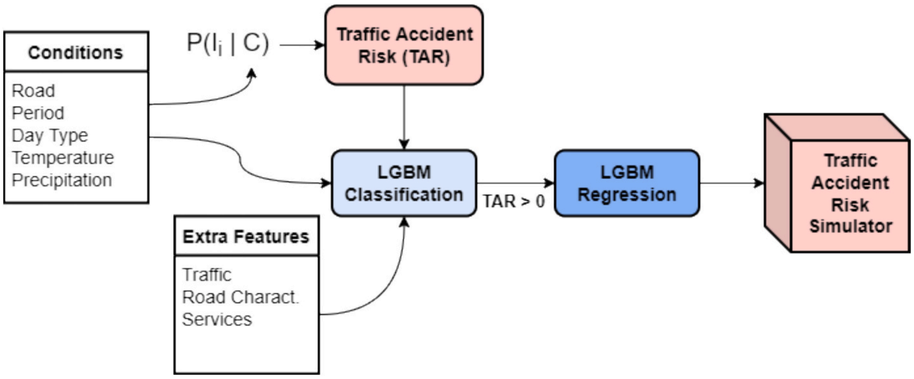


Fig. 5. The pipeline of the Road Accident Risk Calculation.

Table 4
Mean risk (μ) of traffic accidents in each daytime slot along with the respective standard deviation (σ). Values of risk and standard deviation were multiplied by 100,000.

		[0 h – 4 h[	[4 h – 7 h[	[7 h – 10 h [	[10 h – 14 h[	[14 h – 17 h[	[17 h – 20 h[	[20 h – 24 h[
weekdays	μ	1,5	0,7	2,7	3,9	4,4	5,0	3,9
	σ	13,2	18,2	26,0	29,0	28,0	36,7	23,8
weekends	μ	2,1	1,1	2,3	3,2	3,7	4,6	3,6
	σ	24,5	20,1	25,8	38,1	36,5	47,9	34,5

Table 5
Mean risk (μ) of traffic accidents in each temperature and precipitation bin with the respective standard deviation (σ). Values of risk and standard deviation were multiplied by 100,000.

	Temperature				Precipitation	
	]∞ – 10°]	]10 °C – 20 °C]	]20 °C – 30 °C]	]30 °C – 40 °C]	[0 mm – 0,01 mm]	]0,01 mm – 2,5mm]
μ	2,1	3,5	2,6	3,9	2,8	3,9
σ	30,7	30,1	17,8	46,4	25,0	40,7

Table 6
Mean risk (μ) of traffic accidents for maximum velocity permitted on each road segment along with standard deviation (σ). Values of risk and standard deviation were multiplied by 100,000.

	20 km/h	30 km/h	40 km/h	50 km/h	70 km/h	80 km/h
μ	2,0	2,7	1,8	3,0	4,7	3,8
σ	24,5	26,1	18,3	29,9	39,3	33,3

risk. In road segments with the presence of speed enforcement radars, the risk of traffic accidents is double compared to road segments without radars (6.0 and 3.0, respectively). The presence of traffic lights is also associated with slightly higher risk values compared to road segments without traffic lights (3.2 and 2.8, respectively) (Table 7).

The higher risk observed on roads with speed enforcement radars and traffic lights does not indicate that these safety measures increase accident risk. Rather, this correlation reflects the strategic placement of these safety measures in locations that were already identified as high-risk areas. This represents a classic case of reverse causality in observational studies—radars and traffic lights are installed as a response to high accident rates, not as their cause.

Our data does not include information on when these safety features were installed, making it impossible to conduct a proper before-and-after analysis. Previous studies employing such methodologies have consistently shown that both speed cameras and traffic signals reduce accident severity, if not always frequency. The persistence of elevated risk despite these interventions suggests that additional complementary measures may be necessary to effectively reduce accident risk in these particularly challenging locations.

4.2. Predicting the risk of traffic accidents

The LightGBM model provided good performance levels, with an AUC of 0.8 for the classification model and a MAPE of 20 % for the regression model. Fig. 6 presents a scatter plot showing the relationship between observations ordered by magnitude and the predicted probability scores of traffic accidents.

With the prediction of traffic accident risk for each road segment, considering the features presented in Table 3, we computed the mean predicted risk value for each road segment (Fig. 7).

Fig. 7c and 7d display the observed risk and predicted risk, respectively, at an intersection between two major roads that cross Lisbon: Eixo Norte-Sul (Eixo N-S) (north–south direction) and 2ª Circular (west-east direction). The highest risk in this example is observed at both entrances to 2ª Circular from Eixo Norte-Sul.

We compared the observed and predicted risks using several accuracy measures: mean absolute error (MAE), root mean square error (RMSE), and symmetric mean absolute percentage error (sMAPE). The results demonstrated good prediction quality with MAE, RMSE, and sMAPE values of 0.003, 0.009, and 13 %, respectively. Fig. 8 shows the MAE and sMAPE for each road segment regarding the mean risk of traffic accidents, along with the same measures for the intersection between Eixo Norte-Sul and the 2ª Circular entrance.

MAE is higher at the entrance and exit of 2ª Circular as well as in the southern segments of Eixo Norte-Sul (Fig. 8a). Regarding sMAPE, all road segments present highly accurate forecasts except for the entrance

Table 7
Mean risk (μ) of traffic accidents considering the number of lanes of each road segment along with standard deviation (σ). Values of risk and standard deviation were multiplied by 100,000.

	1	2	3	4	5	6	With radar	Without radar	With traffic lights	Without traffic lights
μ	2,5	3,0	4,1	3,6	3,7	3,5	6,0	3,0	3,2	2,8
σ	25,7	28,6	39,5	32,4	31,0	16,6	41,4	30,0	31,3	28,8

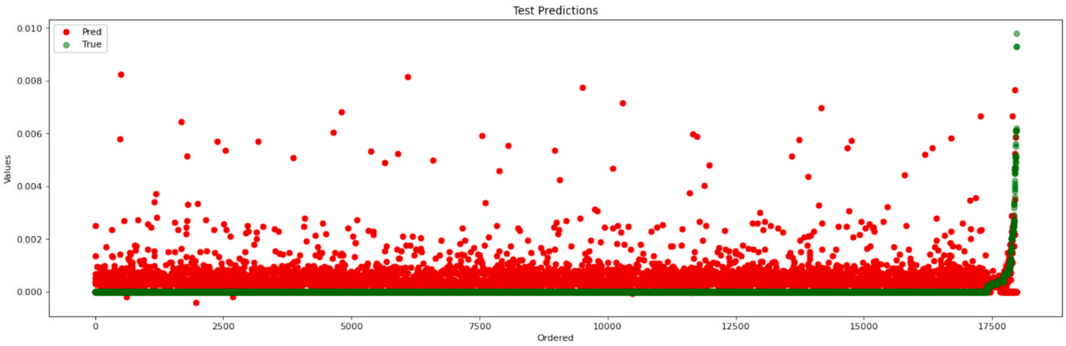
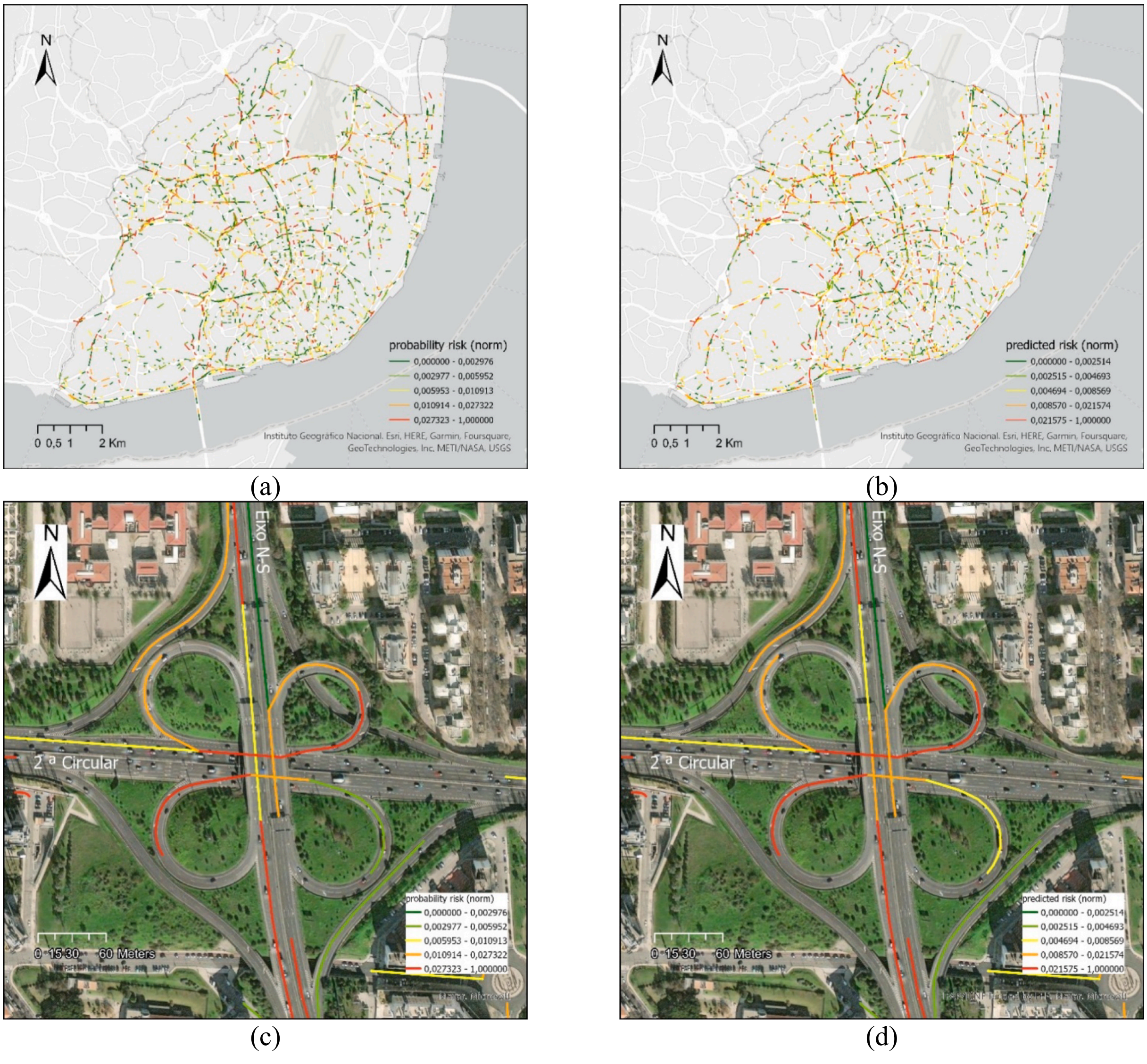


Fig. 6. Scatter plot between ordered observations and predicted traffic accident values.



(a) (b)

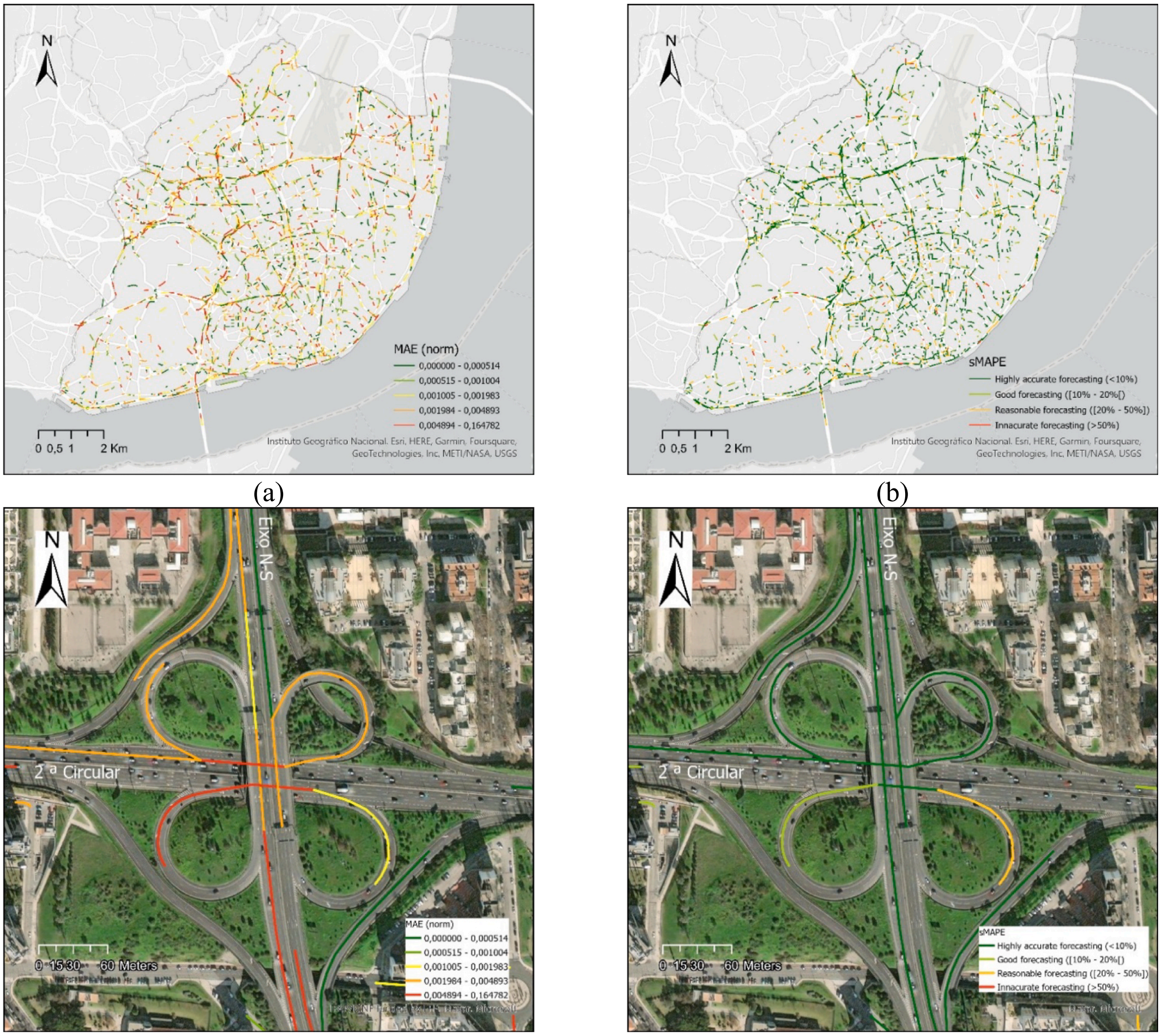
(c) (d)

Fig. 7. Mean observed risk (a) and predicted risk (b) for each road segment. The intersection between Eixo Norte-Sul and 2ª Circular for observed probability risk (c) and predicted risk (d). Presented risk values were normalized using the min–max normalization method.

and exit of 2ª Circular in the west-east direction (Fig. 8d).

These results suggest design deficiencies in the entrances and exits, especially on 2ª Circular, notably the insufficient length of acceleration lanes. This finding is consistent with international road design standards that recommend longer acceleration lanes for high-speed roadways to allow safer merging. Such detailed spatial analysis enables urban mobility planners to identify and prioritize specific road segments for redesign or modification with the goal of decreasing traffic accident risk.

Potential interventions could include increasing the length of acceleration and deceleration lanes, implementing variable speed limits during peak hours, improving signage, or redesigning the geometric characteristics of these high-risk intersections. These targeted



(a) (b)

(c) (d)

Fig. 8. Normalized MAE (a) and sMAPE (b) of the risk of traffic accidents for each road segment and MAE (c) and sMAPE (d) for the intersection between Eixo Norte-Sul and 2ª Circular. The categories defined for sMAPE were based on (Lewis, 1982).

modifications would likely yield greater safety improvements than blanket interventions applied across all road segments.

In addition, emergency services can utilize this fine-grained spatial risk information to optimize the allocation of emergency response vehicles, potentially reducing response times to accidents on roads with higher traffic accident risk. This application demonstrates how the model outputs can benefit multiple stakeholders beyond urban planners, extending to emergency management agencies.

4.3. Risk of traffic accidents simulator

We've developed a prototype simulator for traffic accident risk that allows users to modify a set of road characteristics to generate new risk projections for specific road segments. Users can select a specific road segment and modify its initial characteristics, including the number of lanes, maximum permitted speed, and the presence or absence of speed enforcement radars and traffic lights. The simulator then computes and displays both the mean probability risk and the predicted risk given the selected conditions.

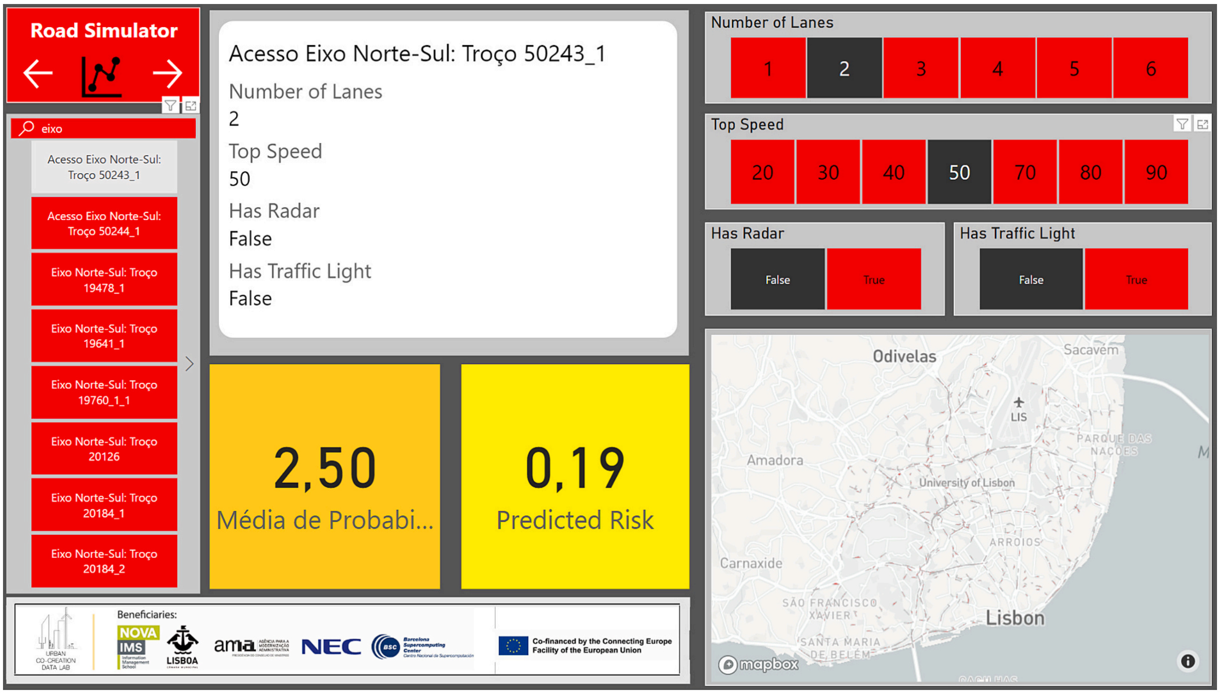
Fig. 9 presents the information for a specific road segment with two lanes, a maximum permitted speed of 50 km/h, and no radars or traffic lights. Under these conditions, the predicted risk of traffic accidents is 0.19, compared to the initial risk of 2.50.

Increasing the maximum permitted speed to 70 km/h causes the predicted risk to increase to 2.57 (Fig. 10). This dramatic increase of over 1,300 % illustrates the powerful influence of speed limits on accident risk, consistent with the exponential relationship between speed and kinetic energy involved in collisions.

The simulator serves as a practical tool for mobility planners to evaluate the potential impacts of changing road characteristics on traffic accident risk. This provides decision support for planning and managing Lisbon's road infrastructure. By allowing planners to test multiple scenarios before implementing physical changes, the tool facilitates evidence-based decision-making and potentially more efficient allocation of limited road safety resources.

5. Conclusion

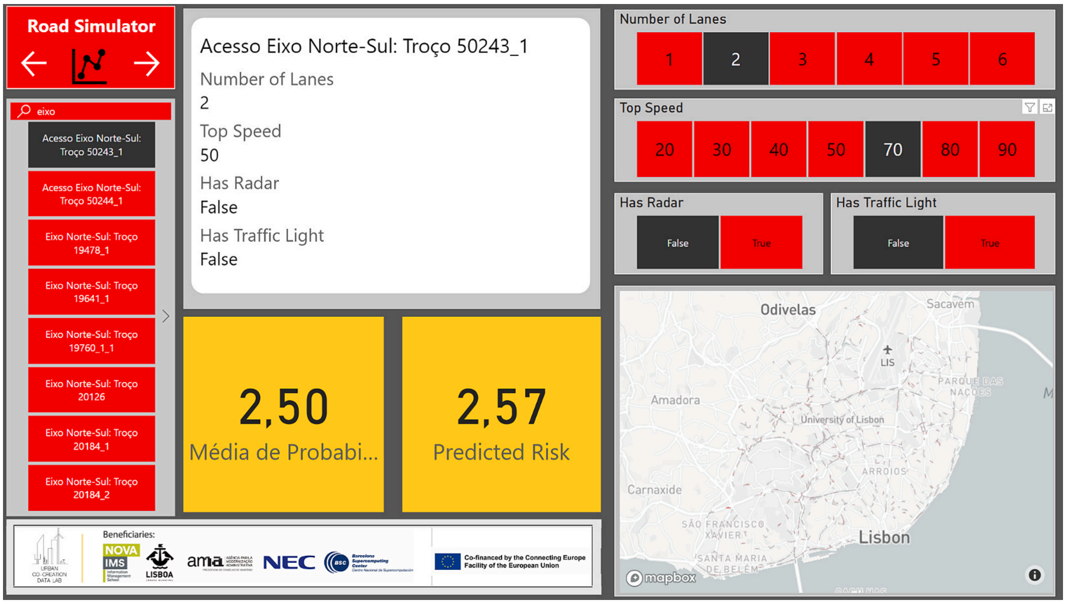
This study proposed a hybrid classification/regression approach using the LightGBM model to predict traffic accident risk in Lisbon at a high spatial resolution. The model was designed to address the challenges of unbalanced datasets and the need for fine-grained urban traffic predictions, which are crucial for emergency services and urban planning. The two-stage modeling process first identified road segments with



Acesso Eixo Norte-Sul: Troço 50243_1
Number of Lanes
2
Top Speed
50
Has Radar
False
Has Traffic Light
False

2,50 Média de Probabi...
0,19 Predicted Risk

Fig. 9. The user interface of the risk of traffic accidents simulator. The selected road segment has 2 lanes, with a maximum speed allowed of 50 km/h, and doesn't have radars or traffic lights.



Acesso Eixo Norte-Sul: Troço 50243_1
Number of Lanes
2
Top Speed
50
Has Radar
False
Has Traffic Light
False

2,50 Média de Probabi...
2,57 Predicted Risk

Fig. 10. Simulation for a specific road segment of the risk of traffic accidents, increasing the maximum velocity allowed from 50 km/h to 70 km/h.

a non-null risk of traffic accidents and then estimated the probability of accidents for different times of the day, considering contextual features such as weather conditions, road characteristics, and traffic signals. The results demonstrated robust predictive performance, achieving an AUC of 0.8 for classification and a MAPE of 20 % for regression.

The findings of this research reveal significant spatial and temporal patterns in traffic accident risk. Notably, accident risk was observed to be higher during the afternoon and evening on weekdays, and during early morning hours on weekends. Furthermore, road segments with speed radars and traffic lights surprisingly exhibited elevated accident risks. This counterintuitive outcome suggests that the mere presence of these traffic control measures may not suffice to reduce risk and that complementary strategies might be necessary, such as optimizing radar placement or improving driver awareness programs.

To bridge the gap between predictive modeling and urban safety planning, a simulation tool was developed, enabling urban planners and emergency services to evaluate how modifications to road characteristics—such as speed limits, lane configurations, and the presence of traffic lights—could impact accident risk. This simulator provides a practical application for policymakers, offering a data-driven mechanism to prioritize safety interventions in high-risk areas and optimize emergency response allocation.

While the proposed model demonstrated strong predictive capabilities, some limitations should be acknowledged. First, the dataset used only includes traffic accidents that required intervention from Lisbon's professional firefighters, which may omit minor accidents or those managed by other emergency services. Additionally, the model does not differentiate accident severity, which could be valuable for prioritizing high-risk segments. The lack of real-time traffic data, pedestrian flow information, and detailed temporal road usage patterns also constrains the model's ability to capture dynamic traffic behaviors. Moreover, the model assumes static traffic configurations, disregarding temporary changes such as road works, events, or seasonal variations.

Future research could address these limitations by integrating richer data sources, including real-time traffic monitoring, pedestrian movement data, and multi-agency accident reports. Extending the model to predict not only the occurrence but also the severity of accidents would provide even greater value to emergency services. Further development of the simulator could enable real-time risk assessment, enhancing its utility for live traffic management and emergency response optimization. Finally, the approach could be replicated in other urban environments, allowing for comparative analysis and broader validation of the proposed methodology.

CRediT authorship contribution statement

Nuno Alpalhão: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Visualization, Writing – original draft. **Pedro Sarmento:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Validation, Writing – original draft. **Bruno Jardim:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Visualization, Writing – original draft, Writing – review & editing. **Miguel de Castro Neto:** Conceptualization, Funding acquisition, Investigation, Project administration, Supervision, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Funding: This work was supported by the Connecting Europe Facility (CEF) – Telecommunications sector in the framework of the project Urban Co-Creation Data Lab [INEA/CEF/ICT/A2018/1837945]; This work was also supported by Portuguese national funds through FCT (Fundação para a Ciência e a Tecnologia) under research grant FCT UIDB/04152/2020 - Centro de Investigação em Gestão de Informação (MagIC); As well as the project C-TECH—Climate Driven Technologies for Low Carbon Cities (POCI-01-0247 FEDER-045919 | LISBOA-01-0247-FEDER-045919) co-financed by the ERDF European Regional Development Fund through the Operational Program for Competitiveness and Internationalization COMPETE 2020, the Lisbon Portugal Regional Operational Program LISBOA 2020 and by the Portuguese Foundation for Science and Technology FCT under MIT Portugal Program.

Data availability

The authors do not have permission to share data.

References

Abdulhafedh, A., 2017. Road Crash Prediction Models: Different Statistical Modeling Approaches. Journal of Transportation Technologies 07 (02), 190–205. https://doi.org/10.4236/jtts.2017.72014.

Ait-Mlouk, A., Agouti, T., 2019. DM-MCDA: a web-based platform for data mining and multiple criteria decision analysis: a case study on road accident. SoftwareX 10, 100323. https://doi.org/10.1016/j.softx.2019.100323.

Bao, J., Liu, P., Ukkusuri, S.V., 2019. A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data. Accid. Anal. Prev. 122, 239. https://doi.org/10.1016/j.aap.2018.10.015.

Basso, F., Basso, L. J., Bravo, F., & Pezoa, R. (2018). Real-time crash prediction in an urban expressway using disaggregated data. *Transportation Research Part C: Emerging Technologies*, *86*(November 2017), 202–219. doi: 10.1016/j.trc.2017.11.014.

de Campos, C.I., Pitombo, C.S., Delhomme, P., Ferreira, S., 2021. What are the risk behaviors of Brazilian and Portuguese drivers? an exploratory approach using self-reported data. Case Studies on Transport Policy 9 (4), 1746–1756. https://doi.org/10.1016/J.CSTP.2021.07.021.

Chang, L.Y., Chen, W.C., 2005. Data mining of tree-based models to analyze freeway accident frequency. J. Saf. Res. 36 (4), 365–375. https://doi.org/10.1016/J.JSR.2005.06.013.

Chen, Q., Song, X., Yamada, H., Shibasaki, R., 2016. Learning deep representation from big and heterogeneous data for traffic accident inference. *30th AAAI Conference on Artificial Intelligence*.

Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* https://doi.org/10.1145/2939672.2939785.

Chen, Y., Shafi, S.Y., Chen, Y., fan., 2020. Simulation pipeline for traffic evacuation in urban areas and emergency traffic management policy improvements through case studies. Transp. Res. Interdiscip. Perspect. 7, 100210. https://doi.org/10.1016/J.TRIP.2020.100210.

CML. (2018). *Lisbon Open Data Portal*. http://lisboaaberta.cm-lisboa.pt/index.php/pt/.

de Myttenaere, A., Golden, B., Le Grand, B., Rossi, F., 2016. Mean absolute percentage Error for regression models. Neurocomputing 192, 38–48. https://doi.org/10.1016/j.neucom.2015.12.114.

Deretić, N., Stanimirović, D., Al Awadh, M., Vujanović, N., Djukić, A., 2022. SARIMA Modelling Approach for forecasting of Traffic Accidents. Sustainability (switzerland) 14 (8). https://doi.org/10.3390/su14084403.

Dong, N., Huang, H., Zheng, L., 2015. Support vector machine in crash prediction at the level of traffic analysis zones: Assessing the spatial proximity effects. Accid. Anal. Prev. 82, 192–198. https://doi.org/10.1016/j.aap.2015.05.018.

Ebrahimi, S., Vachal, K., Szmerekovsky, J., 2022. A Delphi-FMEA model to assess county-level speeding crash risk in North Dakota. Transp. Res. Interdiscip. Perspect. 16, 100688. https://doi.org/10.1016/J.TRIP.2022.100688.

Effati, M., Rajabi, M., Hakimpour, F., Shabani, S., 2014. Prediction of Crash Severity on Two-Lane, Two-Way roads based on Fuzzy Classification and Regression tree using Geospatial Analysis. J. Comput. Civ. Eng. 28. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000432.

Effati, M., Sadeghi-Niaraki, A., 2015. A semantic-based classification and regression tree approach for modelling complex spatial rules in motor vehicle crashes domain. Wires Data Mining Knowl Discov 5, 181–194. https://doi.org/10.1002/widm.1152.

Fancello, G., Carta, M., Serra, P., 2020. Data Envelopment Analysis for the assessment of road safety in urban road networks: a comparative study using CCR and BCC models. Case Studies on Transport Policy 8 (3), 736–744. https://doi.org/10.1016/J.CSTP.2020.07.007.

Fancello, G., Soddu, S., Fadda, P., 2018. An accident prediction model for urban road networks. Journal of Transportation Safety and Security 10 (4), 387–405. https://doi.org/10.1080/19439962.2016.1268659.

Gan, J., Li, L., Zhang, D., Yi, Z., Xiang, Q., 2020. An Alternative Method for Traffic Accident Severity Prediction: using Deep Forests Algorithm. J. Adv. Transp. 2020, 1257627. https://doi.org/10.1155/2020/1257627.

Hossain, M., Muromachi, Y., 2012. A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways. Accid. Anal. Prev. 45, 373–381. https://doi.org/10.1016/J.AAP.2011.08.004.

Huang, C., Zhang, C., Dai, P., Bo, L., 2019. Deep Dynamic Fusion Network for Traffic Accident forecasting. In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 2673–2681. https://doi.org/10.1145/3357384.3357829.

Huang, J., Ling, C.X., 2005. Using AUC and Accuracy in evaluating Learning Algorithms. IEEE Trans. Knowl. Data Eng. 17 (3), 299–310.

Huang, T., Wang, S., & Sharma, A. (2020). Highway crash detection and risk estimation using deep learning. *Accident Analysis and Prevention*, *135*(December 2019), 105392. doi: 10.1016/j.aap.2019.105392.

Ihueze, C.C., Onwurah, U.O., 2018. Road traffic accidents prediction modelling: an analysis of Anambra State, Nigeria. Accid. Anal. Prev. 112, 21–29. https://doi.org/10.1016/j.aap.2017.12.016.

INE. (2011). *Censos 2011*. https://censos.ine.pt/xportal/xmain?xpid=CENSOS&xpgid=censos2011_apresentacao.

INE. (2021). *Censos 2021 - Preliminary results*. https://www.ine.pt/scripts/db_censos_2021.html.

Jardim, B., Alpalhão, N., Sarmento, P., de Castro Neto, M., 2022. The Illegal Parking score – Understanding and predicting the risk of parking illegalities in Lisbon based on spatiotemporal features. Case Studies on Transport Policy. https://doi.org/10.1016/j.cstp.2022.07.011.

Joyce, J., 2021. Bayes' Theorem. In: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy ({f}all 202). Stanford University, Metaphysics Research Lab.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *2017-Decem*.

Kiviluoto, K., Tapio, P., Tuominen, A., Lyytimäki, J., Ahokas, I., Silonsaari, J., Schwanen, T., 2022. Towards sustainable mobility – Transformative scenarios for

2034. Transp. Res. Interdiscip. Perspect. 16, 100690. https://doi.org/10.1016/J.TRIP.2022.100690.

Lewis, C.D., 1982. *Industrial and business forecasting methods*. Butterworths.

Li, K., Xu, H., Liu, X., 2022. Analysis and visualization of accidents severity based on LightGBM-TPE. Chaos Solitons Fractals 157, 111987. https://doi.org/10.1016/j.chaos.2022.111987.

Lin, L., Wang, Q., Sadek, A.W., 2015. A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction. Transp. Res. Part C Emerging Technol. 55, 444–459. https://doi.org/10.1016/J.TRC.2015.03.015.

Liu, X., Lan, Y., Zhou, Y., Shen, C., Guan, X., 2021. A real-time explainable traffic collision inference framework based on probabilistic graph theory. Knowl.-Based Syst. 212, 106442. https://doi.org/10.1016/j.knosys.2020.106442.

Marcillo, P., Leonardo, Á., Caraguay, V., Hernández-Álvarez, M., 2022. A Systematic Literature Review of Learning-Based Traffic Accident Prediction Models Based on Heterogeneous Sources. https://doi.org/10.3390/app12094529.

Ozbayoglu, M., Kucukayan, G., Dogdu, E., 2016. A real-time autonomous highway accident detection model based on big data processing and computational intelligence. IEEE International Conference on Big Data (big Data) 2016, 1807–1813. https://doi.org/10.1109/BigData.2016.7840798.

Paikari, E., Moshirpour, M., Alhajj, R., & Far, B. H. (2014). Data integration and clustering for real time crash prediction. *Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014)*, 537–544. doi: 10.1109/IRI.2014.7051936.

Racioppi, F., Eriksson, L., Tingvall, C., Villaveces, A., 2004. Preventing road traffic injury: a public health perspective for europe. World Health Organization 9–79.

Ren, H., Song, Y., Wang, J., Hu, Y., Lei, J., 2018. A Deep Learning Approach to the Citywide Traffic Accident Risk Prediction. In: *IEEE Conference on Intelligent Transportation Systems*. https://doi.org/10.1109/ITSC.2018.8569437.

Shen, G., Guan, L., Tan, J., Kong, X., 2020. DeepTSW: an Urban Traffic Safety Warning Framework based on Bayesian Deep Learning. In: Ning, H., Shi, F. (Eds.), *Cyberspace Data and Intelligence, and Cyber-Living, Syndrome, and Health*. Springer Singapore, pp. 50–63.

B.W. Silverman Density Estimation for Statistics and Data Analysis (1st Editio). 1998 10.1201/9781315140919.

Smith, B.L., Williams, B.M., Keith Oswald, R., 2002. Comparison of parametric and nonparametric models for traffic flow forecasting. Transp. Res. Part C Emerging Technol. 10 (4), 303–321. https://doi.org/10.1016/S0968-090X(02)00009-8.

Taamneh, M., Alkheder, S., Taamneh, S., 2017. Data-mining techniques for traffic accident modeling and prediction in the United Arab Emirates. Journal of Transportation Safety and Security 9 (2), 146–166. https://doi.org/10.1080/19439962.2016.1152338.

Tanprasert, T., Siripanpornchana, C., Surasvadi, N., Thajchayapong, S., 2020. Recognizing Traffic Black spots from Street View Images using Environment-Aware image Processing and Neural Network. IEEE Access 8, 121469–121478. https://doi.org/10.1109/ACCESS.2020.3006493.

Umair, M., Rana, I.A., Lodhi, R.H., 2022. The impact of urban design and the built environment on road traffic crashes: a case study of Rawalpindi. Pakistan. Case Studies on Transport Policy 10 (1), 417–426. https://doi.org/10.1016/J.CSTP.2022.01.002.

Verma, A., Sasidharan, S., Bhalla, K., Allirani, H., 2022. Fatality risk analysis of vulnerable road users from an Indian city. Case Studies on Transport Policy 10 (1), 269–277. https://doi.org/10.1016/J.CSTP.2021.12.008.

Vlahogianni, E.I., Karlaftis, M.G., Golias, J.C., 2014. Short-term traffic forecasting: where we are and where we're going. Transp. Res. Part C Emerging Technol. 43, 3–19. https://doi.org/10.1016/J.TRC.2014.01.005.

Wenqi, L., Dongyu, L., Menghua, Y., 2017. A model of traffic accident prediction based on convolutional neural network. In: *2017 2nd IEEE International Conference on Intelligent Transportation Engineering (ICITE)*, pp. 198–202. https://doi.org/10.1109/ICITE.2017.8056908.

WHO. (2018). *Global Status Report on Road Safety 2018*.

Wu, M., Shan, D., Wang, Z., Sun, X., Liu, J., Sun, M., 2019. A Bayesian Network Model for Real-time Crash Prediction based on selected Variables by Random Forest. In: *2019 5th International Conference on Transportation Information and Safety (ICTIS)*, pp. 670–677. https://doi.org/10.1109/ICTIS.2019.8883694.

Xiong, X., Chen, L., Liang, J., 2018. A New Framework of Vehicle Collision Prediction by Combining SVM and HMM. IEEE Trans. Intell. Transp. Syst. 19 (3), 699–710. https://doi.org/10.1109/TITS.2017.2699191.

Xu, C., Wang, W., Liu, P., 2013. A Genetic programming Model for Real-Time Crash Prediction on Freeways. IEEE Trans. Intell. Transp. Syst. 14 (2), 574–586. https://doi.org/10.1109/TITS.2012.2226240.

Yassin, S.S., Pooja., 2020. Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach. SN Appl. Sci. 2 (9), 1576. https://doi.org/10.1007/s42452-020-3125-1.

Yu, L., Du, B., Hu, X., Sun, L., Han, L., Lv, W., 2021. Deep spatio-temporal graph convolutional network for traffic accident prediction. Neurocomputing 423, 135–147. https://doi.org/10.1016/J.NEUCOM.2020.09.043.

Zhu, L., Li, T., Du, S., 2019. TA-STAN: a Deep Spatial-Temporal attention Learning Framework for Regional Traffic Accident Risk Prediction. International Joint Conference on Neural Networks (IJCNN) 2019, 1–8. https://doi.org/10.1109/IJCNN.2019.8852212.