Group 11
Yacouba Coulibaly
May 5th, 2024
DNSC 6315 – Final Paper

SEVERE ROAD ACCIDENTS RISK PREDICTION

PUBLIC SAFETY

Columbia City, South Carolina

Compared to other modes of transportation, automobiles rank second only to motorcycles in terms of lethality. In response to the safety concerns raised by the Amtrak train derailment in Philadelphia, Ingraham highlighted in his Washington Post article that automobiles posed a much higher risk, with 7.28 deaths per billion passenger miles.

The adage "time is money" underscores the importance of time management in our civilization's success. While automobiles have undoubtedly facilitated time-saving travel, they have also introduced heightened risks to life and safety. According to National Highway Traffic Safety Administration (NHTSA), an overwhelming majority, an estimated 94%, of accidents can be attributed to human error.

Of particular concern are distracted driving, which has been exacerbated by the pervasive use of electronic devices; Impaired driving, which involves drivers under the influence of alcohol and therefore unable to make sound judgments, magnifies not only the likelihood but also the severity of accidents; And finally, amongst the major causes, speeding as well as adverse weather conditions exert considerable influence on accident rates. Beyond the human factors, deficiencies in road infrastructure such as poor lighting, design flaws, or inadequate signage contribute significantly to accidents occurrences.

To mitigate the incidence of car accidents, most countries have adopted a multifaceted approach which involves on one hand preventing measures such as campaigns and sensibilization, and on the other hand punitive measures to sanction dangerous drivers and de-incentivize dangerous

driving. The use of technology has been paramount in those efforts and new technologies are expanding our expectations to promising new frontiers in accident prevention.

Addressing the challenge of reducing car accident fatalities, I have chosen to focus my final project on applying machine learning techniques to this pressing issue. Thus, the ideas in this paper aim to re-imagine accident prevention through the lenses of machine learning to develop a comprehensive and an applicable approach for road safety in the United States.

Studies have indicated that increased police presence can substantially reduce the occurrence of severe crashes. My objective is to predict severe car accidents hotspots based on a combination of weather and road infrastructure features. By utilizing these predictions, strategic allocation of police vehicles can be implemented daily, effectively reducing the frequency of severe crashes, and preventing fatalities.

In the subsequent sections, I will undertake the following tasks in a structured manner: (1) conduct a concise review of relevant literature, (2) articulate my approach to addressing the issue, (3) describe the dataset that I will use for this problem, (4) present and discuss the statistical performance of the models in-sample as well as out-of-sample, (5) assess models performance under the overarching goal of reducing car accident fatalities through strategic police cars allocation, and (6) shed some light on the limitations inherent in this analysis.

(1) Literature review

Accident risk prediction has been a subject of extensive research in recent decades, with studies typically falling into three main categories, each offering unique insights and methodologies.

One prominent category of research in accident risk prediction involves analyzing the influence of environmental stimuli on the likelihood or severity of traffic accidents. Environmental factors such as weather conditions, traffic flow patterns, and properties of the road network play significant roles in shaping accident occurrences. Studies within this category delve into various aspects of environmental stimuli and their impact on accidents.

For instance, researchers have investigated the correlation between weather factors, such as precipitation, and road accidents. Through data mining techniques and statistical analysis, associations between weather conditions and accident rates have been explored, providing valuable insights into causality and risk factors. Additionally, efforts have been made

to understand the impact of unobserved variables, such as missing data, on the severity of traffic accidents, highlighting the complexities involved in predicting accident outcomes solely based on environmental stimuli.

While studies in this category offer significant insights into the relationship between environmental factors and accident risk, their applicability to real-time prediction and planning is limited. Nonetheless, the findings contribute to a deeper understanding of the multifaceted nature of accident causation.

Another area of research focuses on predicting the frequency of traffic accidents within specific road segments or geographical regions. These studies aim to forecast the expected number of accidents based on a range of factors, including road geometry, traffic volume, and historical accident data. Various modeling techniques, such as neural networks, convolutional neural networks, and Long Short-Term Memory (LSTM) models, have been employed to predict accident frequency accurately.

For example, early studies utilized road-related attributes and weather data to develop predictive models for accident frequency. More recent research has explored the use of satellite imagery and advanced machine learning algorithms to enhance prediction accuracy. These approaches leverage large-scale datasets and sophisticated modeling techniques to provide insights into accident patterns and trends.

However, despite the advancements in accident frequency prediction, challenges remain in integrating these models into real-time applications due to the complexity and volume of data involved. Nevertheless, the predictive capabilities offered by these studies contribute to proactive planning and resource allocation for accident prevention and mitigation.

The third category of research focuses on predicting the risk of accidents as a binary classification task, making it more suitable for real-time applications. These studies aim to classify road segments or geographical areas based on the likelihood of accident occurrence, enabling timely interventions and proactive measures.

Researchers have employed various approaches, including decision tree models, autoencoder models, and logistic regression models, to predict accident risk. Factors such as weather conditions, traffic volume, and road characteristics are commonly used as predictors in these models. By leveraging machine learning techniques and heterogeneous urban data sources, researchers have been able to capture spatial heterogeneity and temporal trends, enhancing the accuracy of accident risk prediction.

For example, studies have utilized human mobility data, satellite imagery, and radar-based rainfall data to predict the probability of accident occurrence in specific areas. These approaches highlight the importance of integrating diverse data sources and analytical methods to achieve robust and reliable predictions.

In the last category, the study conducted by Sobhan Moosavi, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath stands out. Their work, titled "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights," was presented at the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems in 2019. This insightful research has served as inspiration for the methodology of my own project.

Their approach, named DAP (Deep Accident Prediction), leverages a deep neural network model that integrates various data attributes including traffic events, weather data, points-of-interest, and time. DAP incorporates multiple components, including recurrent and fully connected layers, as well as a trainable embedding component to capture spatial heterogeneity. To address data gaps, they have curated a comprehensive database of accident information, named US-Accidents spanning from 2016 for 2023 *(the dataset they used at the time was up to 2019)*. The dataset has been made publicly available, and I will utilize a subset containing data from the city of Columbia in South Carolina to predict severe accidents hotspots, identifying areas where severe accidents are more likely to occur. In the following subsection, I will detail the approach I have formulated to predict accident hotspots and subsequently explore its integration with law enforcement agencies to bolster efforts in reducing road accidents.

(2) Approach

The approach to addressing the complex issue of accident risk prediction relies partly on a comprehensive literature review conducted prior to the analysis. As demonstrated by the referenced works in the preceding section, conceptualizing accident risk prediction as a classification problem can aid in simplifying the issue, rendering results more conducive to real-time application. For example, employing logistic regression enables us to rank predictions by their probabilities, offering insights into the prioritization of law enforcement efforts.

Regarding prediction, our methodology involves utilizing weather and road infrastructure features known to influence accidents. These features will be elaborated upon further in the subsequent section of this report.

To identify accident hotspots, it is crucial to divide the relevant city into uniformly sized grids. Following prediction generation, these grids will be ranked based on the frequency of highly probable severe accidents. Subsequently, law enforcement resources can be allocated to these areas accordingly. Upon completion of preprocessing, we will employ cross-validation to compare the performance of **four baseline models**: (1) logistic regression, (2) random forest, (3) decision tree, and (4) support vector machine (SVM). Due to the length constraint of the paper, we will conduct an in-depth analysis of the top two performing models and assess their efficacy in meeting the predefined public safety objectives.

This transition segues into the next part of this subsection, which outlines the methodology for evaluating prediction performance. To facilitate this assessment, certain assumptions are necessary:

1. The city of Columbia in South Carolina possesses only 30 police vehicles available for daily patrols.

2. Each accident recorded in our dataset is presumed to involve exactly two fatalities, reflecting the severity of crashes typically considered.

3. The presence of a police vehicle in a zone reduces the total number of fatalities by 2. However, once a zone has 5 police vehicles, additional units yield no further impact. (Not sending out a car cost 0)

4. It is assumed that the cost of dispatching a police vehicle to a location is 2 units, while each life saved yields 4 units in return.

5. Police cars will be patrolling the squares where they are sent instead of remaining idle at specific spots.

6. All predictions made are valid for an entire day.

Below is a summary table illustrating how key performance indicators (KPIs) will be assessed *(using fake data in the table).*

	A	B	C	D	E	F	G
	Total deaths	**Allocation**	**Total lives saved**	**Value of saved lives**	**Cost of allocation**	**KPI_IS**	**KPI_OS**
	(2* sum predicted (1) in ZONE A) y_pred	Nb of cars <= 5		C*4	*-2*Allocation	D+E	Use y_test to evaluate business decision
ZONE A	6	5	6	24	-10	14	
ZONE B	10	1	2	8	-2	6	
ZONE C	2	5	2	8	-10	-2	
ZONE D	3	5	3	12	-10	2	
ZONE E	30	5	10	40	-10	30	
ZONE F	10	2	4	16	-4	12	
						62	

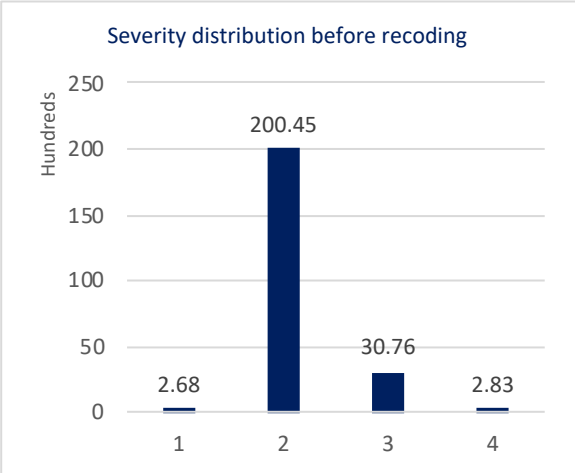
The cells emphasized at the table's bottom represent the combined totals of each zone's performance. A positive figure indicates that the value of saved lives exceeds the expense of deploying patrol units. Therefore, the mode yielding a higher total in out-of-sample scenarios is preferable. The out of sample KPI will also be compared with **the optimal allocation** of police cars, that is the allocation yielding the highest KPI once we have observed y_test. Additionally, we can examine column D to assess the value of lives saved independently of cost considerations. A more detailed discussion on this aspect will follow in the relevant section. The subsequent subsection explores the project's dataset and provides an initial data analysis.

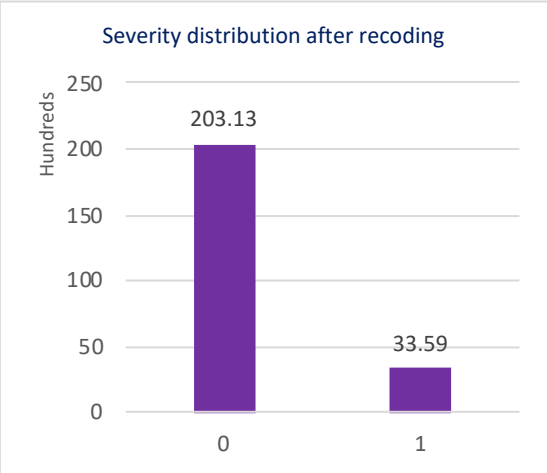
(3) dataset and descriptive analytics

We will be using publicly available data pulled from Kaggle (US Accidents (2016 - 2023)). This dataset includes car accidents data spanning 49 states across the USA. It includes incidents recorded from February 2016 to March 2023, gathered through several APIs that stream traffic incident data. These APIs collate information from diverse sources, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and road network sensors.

The original dataset comprises approximately 7.7 million accident records. The initial step involved creating a subset of records specific to the city of Columbia in South Carolina. This task was executed using EC2 for enhanced efficiency. Subsequently, the resulting subset contained 35,018 records. After removing NA (not available) values, the final dataset utilized for the project consisted of 23,672 records.

As indicated in the project proposal, the target variable (severity) encompasses values ranging from 1 to 4. Given the project's focus on severe crashes, values 1 and 2 were recoded as 0, while values 3 and 4 were recoded as 1. The distributions before and after recoding are illustrated in the graphs below.





As depicted in the left graph, severe crashes (severity levels 3 and 4) occur less frequently compared to incidents with a severity level of two. Incidents with a severity level of two constitute the majority of the sample, accounting for 84.68%. Following this, incidents with a severity level of three rank second, comprising 12.99% of the sample, while severity levels 4 and 1 are the least frequent, with respective proportions of 1.2% and 1.13%. Upon recoding the data

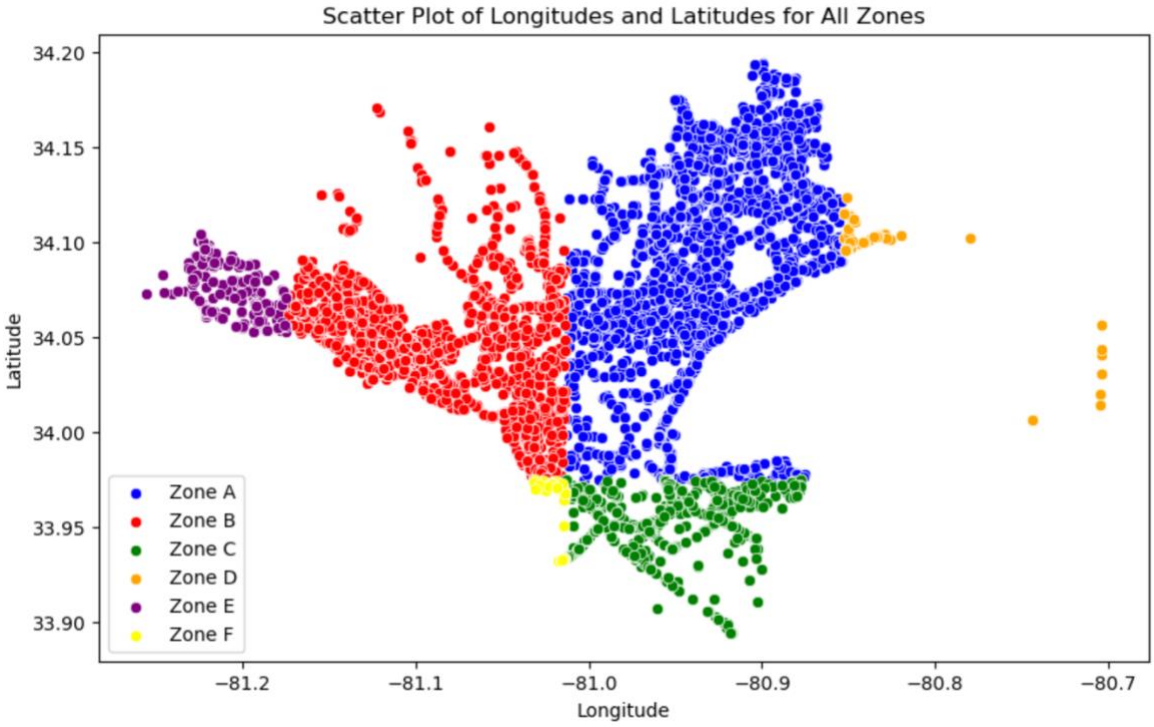
to merge severity levels 1 and 2 as well as 3 and 4 into binary categories of 0 and 1 respectively, the proportion of severe incidents becomes 14.19%.

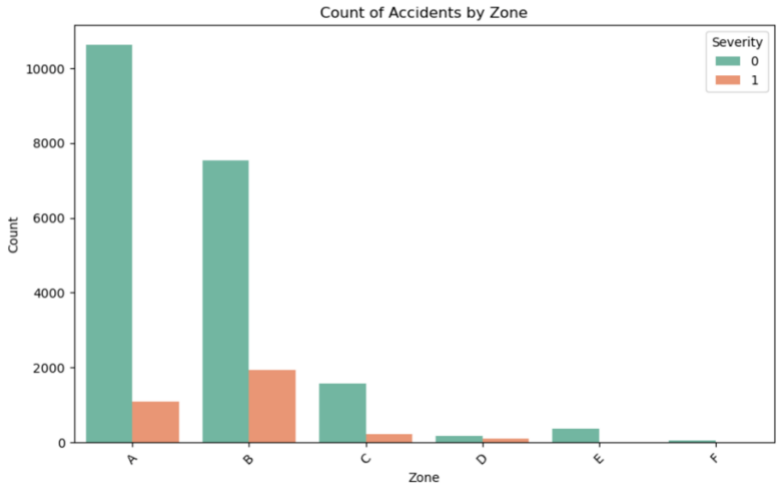
We will use a total of 12 predictor variables, consisting of seven weather-related features and five road infrastructure features, as depicted in the table provided below.

Among the 12 features, five are categorical and will necessitate recoding as dummy variables prior to training. This process is elaborated upon in greater detail in the section focusing on the training and performance of the models.

Feature type		Variable name	Description	Data type
Weather features	1	Temperature	Shows the temperature (in Fahrenheit)	*numeric*
	2	Wind_Chill	Shows the wind chill (in Fahrenheit)	*numeric*
	3	Humidity	Shows the humidity (in percentage)	*numeric*
	4	Pressure	Shows the air pressure (in inches)	*numeric*
	5	Visibility	Shows visibility (in miles)	*numeric*
	6	Wind_Speed	Shows wind speed (in miles per hour)	*numeric*
	7	Precipitation	Shows precipitation amount in inches, if there is any	*numeric*
Road infrastructure features	1	Bump	Indicates presence of speed bump or hump in a nearby location	*categorical*
	2	Crossing	Indicates presence of crossing in a nearby location	*categorical*
	3	Give_Way	Indicates presence of give_way in a nearby location	*categorical*
	4	Junction	Indicates presence of junction in a nearby location	*categorical*
	5	Stop	Indicates presence of stop in a nearby location	*categorical*

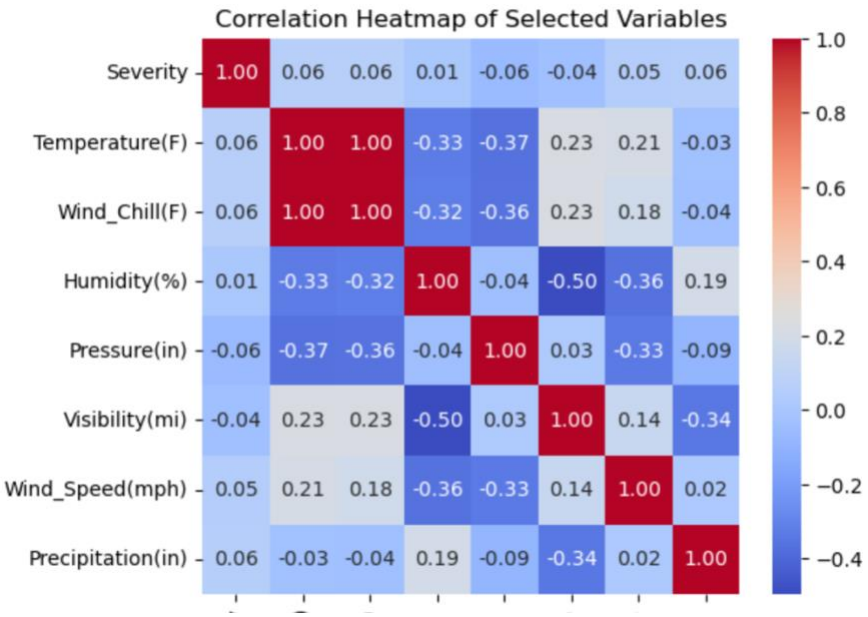
Since one of our objectives is to forecast hotspots, it's essential to establish a grid system for the city. I opted for a 3km-by-3km grid, resulting in six zones. The scatter plot below illustrates the distribution of accidents in Columbia city across these zones. It's notable that zones A and B exhibit the highest incident rates, followed by zones C and E, with zones D and F showing the lowest incidents. These findings are significant as we proceed with the training process and evaluate the predictions generated by the trained models. A logical subsequent inquiry arose regarding the distribution of accident severity within each zone and whether these outcomes aligned with the distributions observed before the creation of the grid. The bar chart juxtaposed to the grid illustrates the distribution of accidents across zones





The results maintain consistency with the initial distribution, where Severity 0 was more prevalent than Severity 1. Each zone exhibits a similar distribution pattern.

Next, I looked at the correlation heatmap of the numerical features. In general, this step revealed that the target variable exhibits minimal correlation with the other weather variables. The highest correlation observed with a weather feature is 0.06. This detail is noteworthy as we proceed to train the model in the subsequent subsection of the report.



(4) Model training and testing

As demonstrated in the literature review, the task of predicting road accident risk through classification is common. frequently utilized models include decision trees, logistic regressions, and others. In this paper's context, I've structured my approach to modeling into two primary steps.

Initially, I conducted cross-validation to assess the performance of four models: logistic regression, decision tree, random forest, and SVM. Evaluating their respective accuracy means and the standard deviation of those means, I ranked them and focused on analyzing the top two performers in greater detail.

With a fold number of 10 and a test size of 20%, each model was trained using the same predictor variables, and no hyperparameters were tuned at this stage. The table below summarizes the outcomes of this initial step.

Note: a smaller subset of 5000 observations is used for efficiency in computations

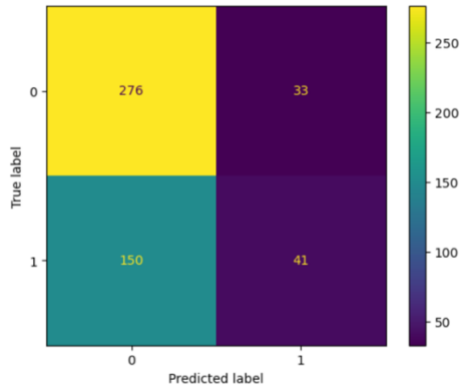
	Algorithm	AUC Mean	AUC STD	Accuracy Mean	Accuracy STD
3	Random Forest	67.36	2.48	63.37	2.21
0	Logistic Regression	65.37	2.85	62.38	1.95
1	Kernel SVM	63.79	2.17	59.75	2.35
2	Decision Tree Classifier	57.28	3.24	59.10	2.26

The Random Forest and Logistic Regression emerged as the top performers in the ranking. In terms of the AUC mean, the Random Forest achieved a mean of 67.36, while Logistic Regression attained a mean of 65.37. With standard deviations of 2.48 and 2.85 respectively for Random Forest and Logistic Regression, the former appears to exhibit greater stability.

In terms of accuracy mean, Random Forest yielded a mean of 63.37 with a standard deviation of 2.21, while Logistic Regression achieved a mean of 62.38 with a standard deviation of 1.95. Despite their relatively equal means, Logistic Regression presents a lower standard deviation, suggesting it may be a more suited choice.

Following closely behind are SVM and Decision Tree, both of which performed similarly to the top models in terms of accuracy mean. In the subsequent pages, we will delve into a detailed analysis of Logistic Regression, examining its performance results, confusion matrix, and specificity and sensitivity. An examination of the Random Forest hyperparameter tuning will ensue, with results presented thereafter.

The logistic regression model was trained utilizing the 12 predictor variables outlined in the data description. Categorical variables were transformed into dummy variables for training purposes. Subsequently, the dataset was partitioned into training and testing sets, with a test size of 10%. Accuracy was evaluated both within the sample and out of sample. The in-sample accuracy score was determined to be 63.17%, while the out-of-sample accuracy score was 63.4%. Following this, I utilized sklearn to construct the confusion matrix as depicted below. Next to the confusion matrix, I have also calculated the specificity, the sensitivity, the true positive and the true negative rates.



Specificity	Sensitivity
55%	65%
True positive rate	**True negative rate**
21%	89%

In the context of predicting severe car crashes, these performance metrics provide valuable insights into the model's effectiveness.

1. **Sensitivity of 65%:** This indicates that the model correctly identifies 65% of actual severe car crashes among all instances of severe crashes. A higher sensitivity suggests that the model is adept at capturing most severe crash incidents, which is crucial for proactive measures and intervention.

2. **Specificity of 55%:** This indicates that the model correctly identifies 55% of non-severe car crashes among all instances of non-severe crashes. A lower specificity suggests that the model may misclassify a notable portion of non-severe incidents as severe, potentially leading to unnecessary interventions or resource allocation.

3. **TPR (True Positive Rate) of 21%:** This indicates that only 21% of actual severe car crashes are correctly identified by the model. This lower value suggests that the model may miss a significant portion of severe crash incidents, potentially impacting its effectiveness in preventing severe accidents.

4. **TNR (True Negative Rate) of 89%:** This indicates that the model correctly identifies 89% of non-severe car crashes among all instances of non-severe crashes. A higher TNR suggests that the model is proficient at accurately classifying non-severe incidents, reducing the likelihood of misclassification and false alarms.

Overall, while the model demonstrates relatively high TNR, its sensitivity and TPR are notably lower, this indicates that there are challenges in accurately identifying and predicting severe car crashes.

From this model, I generated class and probability predictions for the test instances, which were subsequently ranked. Below is a segment of the output, which will be utilized to assess the model's performance in achieving public safety objectives.

	0	1	predicted_class	actual_class	z_prime
448	0.261126	0.738874	1	0	B
416	0.290819	0.709181	1	0	B
413	0.304445	0.695555	1	0	B

After completing the training of the logistic regression model, I proceeded to train a random forest model using the same set of variables.

Initially, the accuracy score of the model, prior to hyperparameter tuning, stood at 63.8%. Subsequently, I employed a randomized grid search technique to identify the optimal hyperparameters through 3-fold cross-validation. The resulting hyperparameters from the grid search were as follows:

'n_estimators': 522, 'min_samples_split': 2, 'min_samples_leaf': 2, 'max_features': 'log2', 'max_depth': 10, and 'bootstrap': False.

Following the application of these tuned hyperparameters, the accuracy achieved on the test set (y_test) increased to **74.4%.** Like the logistic regression model, the evaluation of class probabilities and predictions will be instrumental in assessing the model's performance in meeting public safety objectives. Additionally, I generated a series containing the features importance from the model. Pressure, humidity, wind_chill emerged as the top three whereas stop, give_way and bump were ranked last. Removing those features from the model will have minor impact on the model's performance.

(5) Performance under public safety objective

Upon examining the models' performance from a statistical perspective, it is crucial to further evaluate their effectiveness in achieving the primary objective of public safety. As previously mentioned, this objective has been the primary motivation behind training the models. To accomplish this, we will utilize both in-sample and out-of-sample key performance indicators

(KPIs). The in-sample KPI will inform our decisions of police cars allocation based on the predictions generated by our models, while the out-of-sample KPI will assess how these decisions fare when applied to the actual y_test dataset.

Taking into consideration that we are trying to attain a public safety objective; we also need to determine the value that an optimal allocation of police cars available would have yielded and compare that number to the out of sample value obtained. We will commence by examining the logistic regression results followed by an analysis of the random forest outcomes. Each result will be supplemented with relevant comments to provide context and insights.

Logistic:

	A	B	C	D	E	F
	Total deaths	Allocation	Total lives saved	Value of saved lives	Cost of allocation	KPI_IS
	(2* sum predicted (1) in ZONE A) y_pred	Nb of cars <= 5		C*4	*-2*Allocation	D+E
ZONE A	12	5	10	40	-10	30
ZONE B	59	5	10	40	-10	30
ZONE C	2	1	2	8	-2	6
ZONE D	3	2	3	12	-4	8
ZONE E	0	0	0	0	0	0
ZONE F	1	1	1	4	-2	2
						76

This first table has in the first column the predicted values of severe crashes in each of our zone. The second column represent the allocation of police cars decision (14 in total) and the kpi (value of lives saved minus cost of allocation) is **76**. In the next table, we compute the kpi based on our allocation using in the first column the actual y_test for each zone. As we can see, the KPI remains 76.

	A	B	C	D	E	F
	Total deaths	Allocation	Total lives saved	Value of saved lives	Cost of allocation	KPI_OS
	(2* sum actual (1) in ZONE A) y_test	Nb of cars <= 5		C*4	*-2*Allocation	D+E
ZONE A	58	5	10	40	-10	30
ZONE B	115	5	10	40	-10	30
ZONE C	14	1	2	8	-2	6
ZONE D	4	2	4	16	-4	12
ZONE E	0	0	0	0	0	0
ZONE F	0	1	0	0	-2	-2
						76

In the last table below, we compute the kpi based on the optimal allocation of police cars using y_test.

	A	B	C	D	E	F
	Total deaths	Allocation	Total lives saved	Value of saved lives	Cost of allocation	KPI_OS
	(2* sum actual (1) in ZONE A) y_test	Nb of cars <= 5		C*4	*-2*Allocation	D+E
ZONE A	58	5	10	40	-10	30
ZONE B	115	5	10	40	-10	30
ZONE C	14	5	10	40	-10	30
ZONE D	4	2	4	16	-4	12
ZONE E	0	0	0	0	0	0
ZONE F	0	0	0	0	0	0
						102

We see that the optimal allocation (17 cars) would have yielded a KPI of 102. That corresponds to 26 value points of difference. Let's now look at how that random forest performs.

Random Forest:

The first table below has in A our predictions, and in B the allocation decision. We can see that the KPI value is 82.

	A	B	C	D	E	F
	Total deaths	Allocation	Total lives saved	Value of saved lives	Cost of allocation	KPI_IS
	(2* sum predicted (1) in ZONE A) y_pred	Nb of cars <= 5		C*4	*-2*Allocation	D+E
ZONE A	28	5	10	40	-10	30
ZONE B	56	5	10	40	-10	30
ZONE C	3	2	3	12	-4	8
ZONE D	5	3	5	20	-6	14
ZONE E	0	0	0	0	0	0
ZONE F	0	0	0	0	0	0
						82

The next table computes the KPI based on our allocation and the y_test values in column A.

	A	B	C	D	E	F
	Total deaths	Allocation	Total lives saved	Value of saved lives	Cost of allocation	KPI_OS
	(2* sum actual (1) in ZONE A) y_test	Nb of cars <= 5		C*4	*-2*Allocation	D+E
ZONE A	69	5	10	40	-10	30
ZONE B	101	5	10	40	-10	30
ZONE C	14	2	4	16	-4	12
ZONE D	7	3	6	24	-6	18
ZONE E	1	0	0	0	0	0
ZONE F	0	0	0	0	0	0
						90

As we can see, the out of sample performance has improved to a score of 90. Let's look at the last table, which computes the KPI based on the optimal allocation.

	A	B	C	D	E	F
	Total deaths	Allocation	Total lives saved	Value of saved lives	Cost of allocation	KPI_OS
	(2* sum actual (1) in ZONE A) y_test	Nb of cars <= 5		C*4	*-2*Allocation	D+E
ZONE A	69	5	10	40	-10	30
ZONE B	101	5	10	40	-10	30
ZONE C	14	5	10	40	-10	30
ZONE D	7	4	7	28	-8	20
ZONE E	1	1	1	4	-2	2
ZONE F	0	0	0	0	0	0
						112

The optimal value stands at 112, that a difference of 22 value points when compared to the allocation based on the predictions.

We note that the random forest performs better than the logistic model under the safety objective evaluation.

(6) Limitations

Maintaining intellectual integrity necessitates a critical examination of the preceding research, highlighting inherent limitations in the analysis.

One critique pertains to the data utilized for model training. Absence of specific data regarding the number of deaths per accident necessitated an assumption of a minimum of two deaths to facilitate model evaluation.

Furthermore, the predictive power of our weather-related features is notably weak with upper ranges of .06, compounded by a disproportionate representation of non-accident instances in the dataset. Consequently, the models tend to predict zero occurrences more frequently. This observation is evident upon scrutiny of the logistic regression analysis.

Although the underlying concept holds promise, real-time application of this model for police car allocation presents significant challenges. Forecasted weather data, while feasible, is subject to rapid fluctuations, rendering the model outdated long before implementation.

Additionally, deploying police cars to grid zones without explicit specifications poses a dilemma. The assumption of patrol duty for each car overlooks the potential for a more refined model capable of optimizing patrol locations to maximize efficiency.

Moreover, the practical implementation of such a solution entails substantial costs. A robust model requires extensive data and computational power to process information rapidly.

Although we can pinpoint a few areas for improvement in the research outlined in these pages, it's crucial to acknowledge its merits as well. This study aimed to demonstrate the application of machine learning techniques in enhancing public safety, and in that aspect, it has proven successful. Scaling up a similar framework with adequate funding could significantly reduce car accident fatalities. However, it's important to recognize that these methods must complement other strategies and initiatives implemented by authorities.

SOURCES:

Washington post: Igraham article (https://www.washingtonpost.com/news/wonk/wp/2015/05/14/the-safest-and-deadliest-ways-to-travel/)

Sobhan Moosavi. (2023). US Accidents (2016 - 2023) [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DS/199387

Sobhan Moosavi, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. 2019. Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights. In 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '19), November 5–8, 2019, Chicago, IL, USA. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3347146.3359078