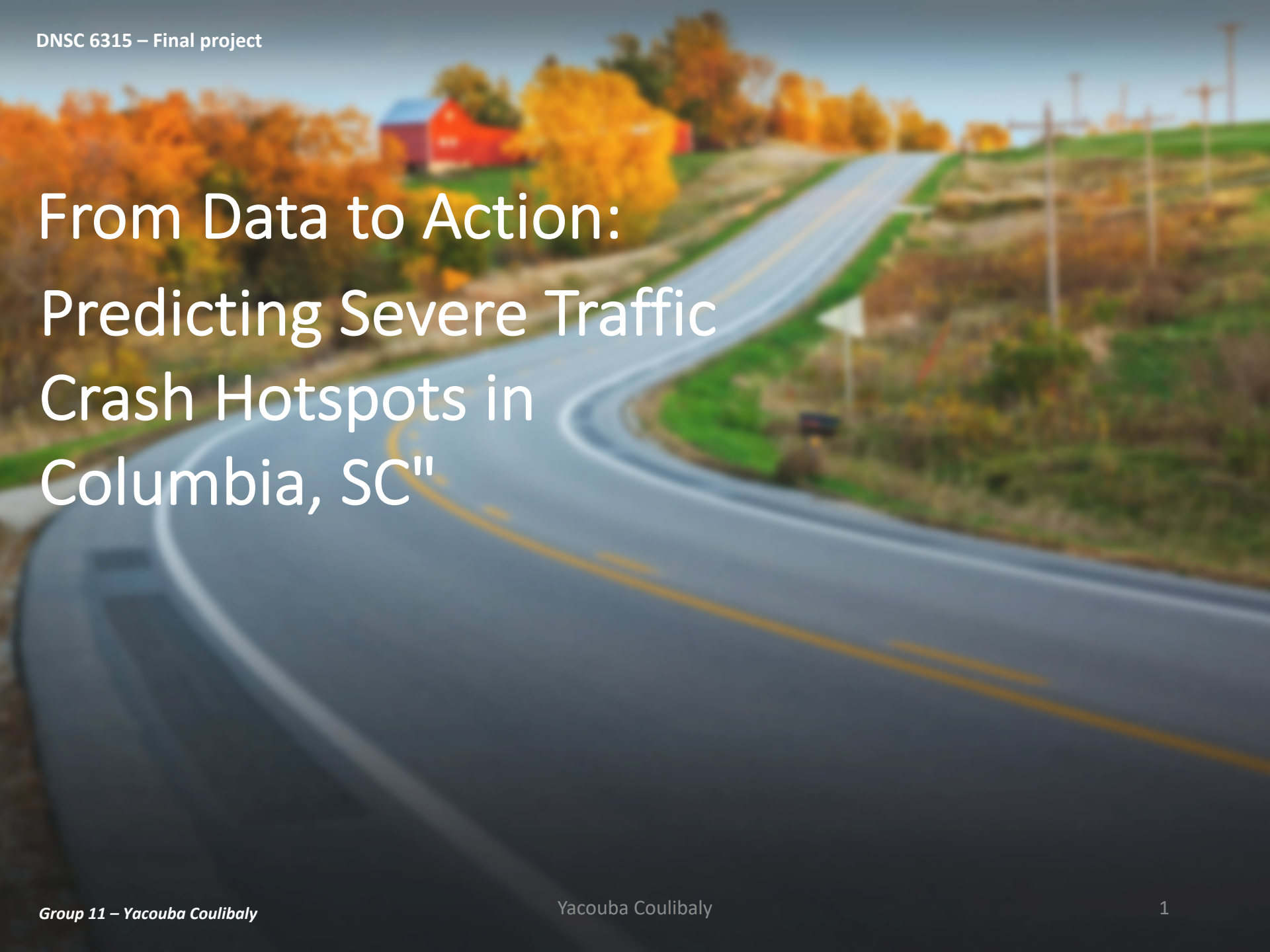


From Data to Action: Predicting Severe Traffic Crash Hotspots in Columbia, SC"

Agenda

01- Introduction

02- Literature review

03- Project proposal

04- Data description

05- Model training

06- Performance

07- Limitations & Conclusion

01. Introduction

Compared to other modes of transportation, automobiles rank second only to motorcycles in terms of lethality with **7.28 deaths per billion passenger miles**. According to National Highway Traffic Safety Administration (NHTSA), an overwhelming majority, an estimated **94%, of accidents can be attributed to human error.** 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.

Impaired driving

Refers to operating a vehicle while under the influence of alcohol, drugs (both illegal substances and prescription medications that affect one's ability to drive safely),

Distracted Driving

Refers to any activity that diverts a driver's attention away from the primary task of operating a vehicle.

Speeding

Refers to driving a vehicle at a speed that exceeds the posted speed limit or is too fast for the current road conditions

Adverse weather

Refers to weather conditions that pose hazards or challenges to safe driving. These conditions may include heavy rains or storms, snow and ice, fog, strong winds, etc.

Deficiencies in road infrastructure such as poor lighting, design flaws, or inadequate signage contribute significantly to accidents occurrences

02. 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.

Environmental stimuli	Frequency of traffic accidents	Predicting risk of accidents
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.	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.	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.
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.	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.	This category highlights the importance of integrating diverse data sources and analytical methods to achieve robust and reliable predictions. One major drawback is the cost involved in these studies, a goof example is the study by Sobhan Moosavi, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath

03. Project Overview

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.

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.

The prediction methodology involves utilizing weather and road infrastructure features known to influence accidents. Hotspots will be obtained by dividing the 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:

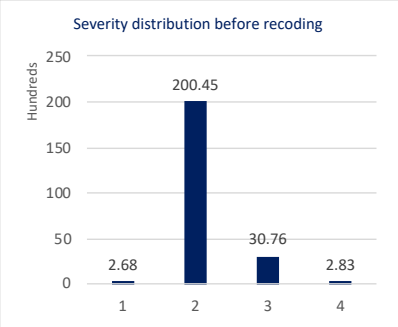
I. **Logistic regression,**
II. **Random forest,**
III. **Decision tree,**
IV. **Support vector machine (SVM).**

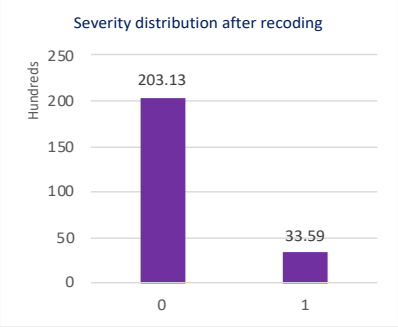
Due to the length constraint of the paper, we will conduct an in-depth analysis of the top two performing models and assess their efficacy in meeting the predefined public safety objectives. To facilitate this assessment, certain assumptions are necessary:

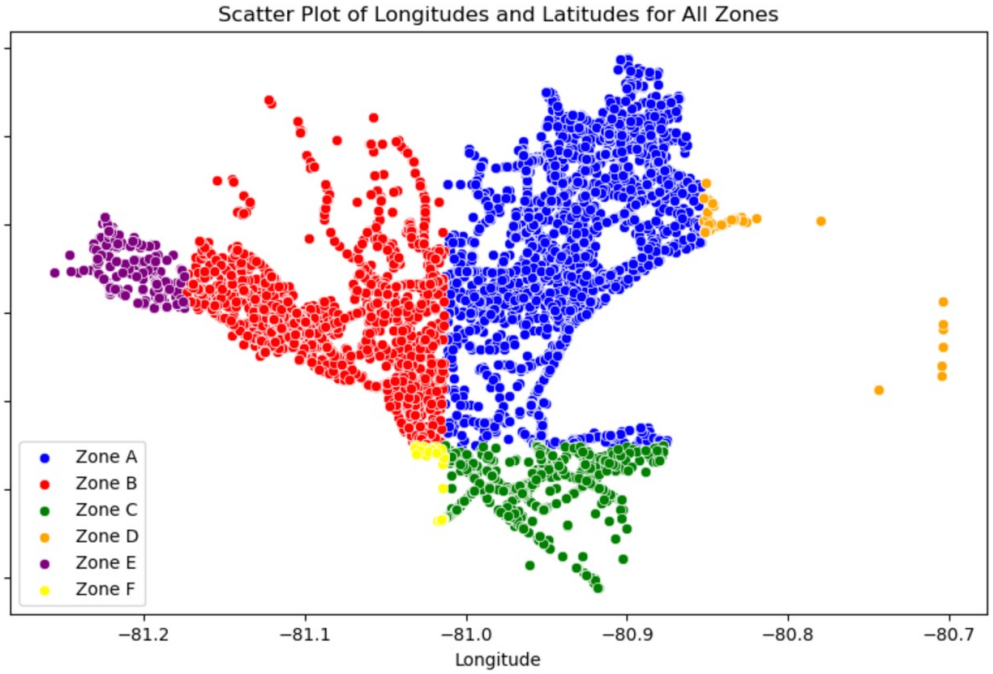
1. The city of Columbia in South Carolina possesses only 30 police vehicles available for daily patrols.
2. Each accident recorded in our dataset is presumed to involve exactly two fatalities, reflecting the severity of crashes typically considered.
3. The presence of a police vehicle in a zone reduces 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.

04. Data Description

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.



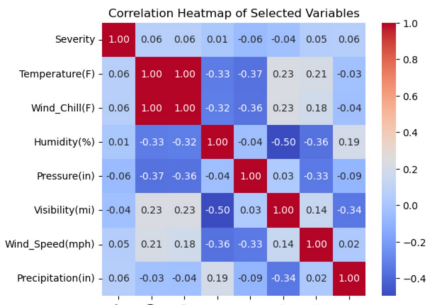




04. Data Description

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.

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*



05. Model Training

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.

	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

Comments :

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.

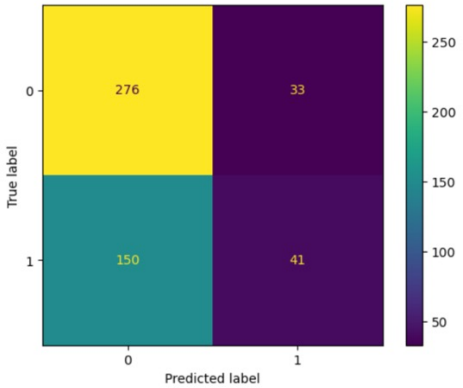
In terms of accuracy mean, Random Forest yielded a mean of 63.37 with a standard deviation of 2.21, while Logistic Regression achieved a mean of 62.38 with a standard deviation of 1.95. Despite their relatively equal means, Logistic Regression presents a lower standard deviation, suggesting it may be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

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

05. Model Training

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.

Logistic regression:



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

Comments :

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.

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

05. Model Training

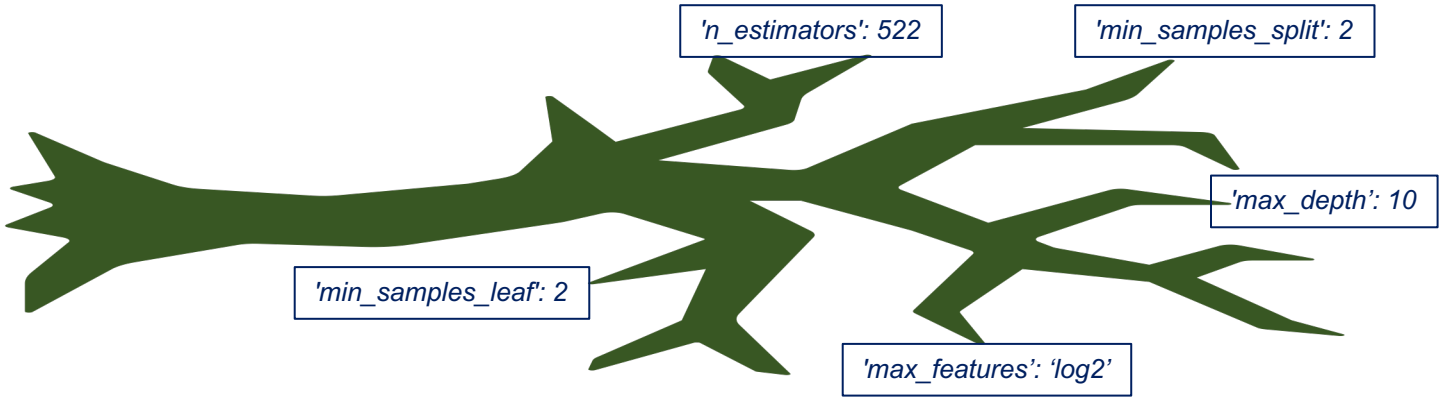
As demonstrated in the literature review, the task of predicting road accident risk through classification is common. frequently utilized models include decision trees, logistic regressions, and others. In this paper's context, I've structured my approach to modeling into two primary steps. Initially,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

Random Forest:

The accuracy score of the random forest 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%.



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

06. Model Performance

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

06. Model Performance

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_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

06. Model Performance

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_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

06. Model Performance

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.

Random Forest:

	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

06. Model Performance

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.

Random Forest:

	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

06. Model Performance

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.

Random Forest:

	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

07. Limitations & Conclusion

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

1	**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.**
2	**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.**
3	**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.**
4	**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.**
5	**The practical implementation of such a solution entails substantial costs. A robust model requires extensive data and computational power to process information rapidly.**