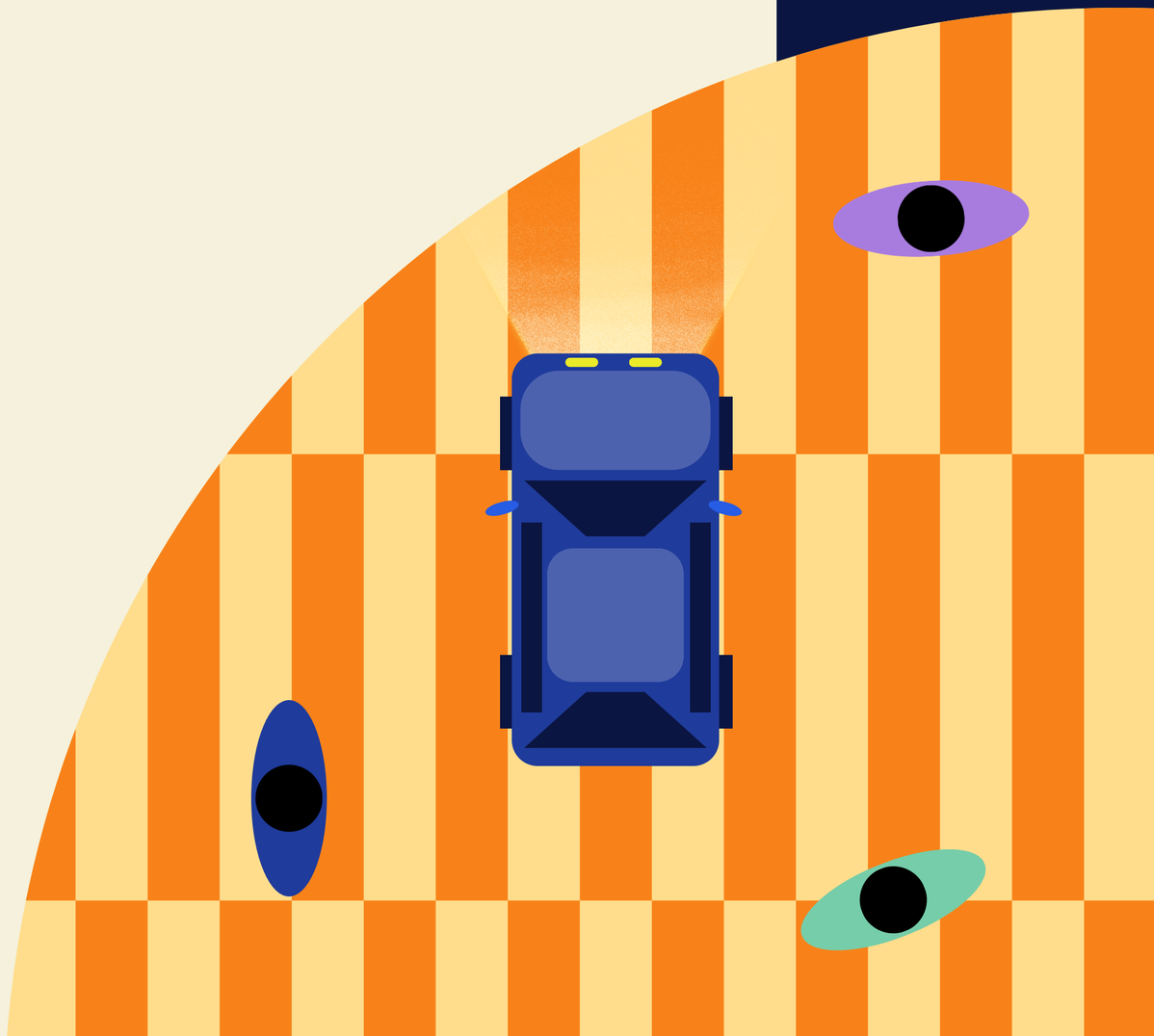
PREDICTING ACCIDENT SEVERITY IN MARYLAND

INSY 662 | DATA MINING & VISUALIZATION

Final Report | Group Project

Meriem | Vincent | Abdul | Xingchen | Chien

AGENDA



EXECUTIVE SUMMARY

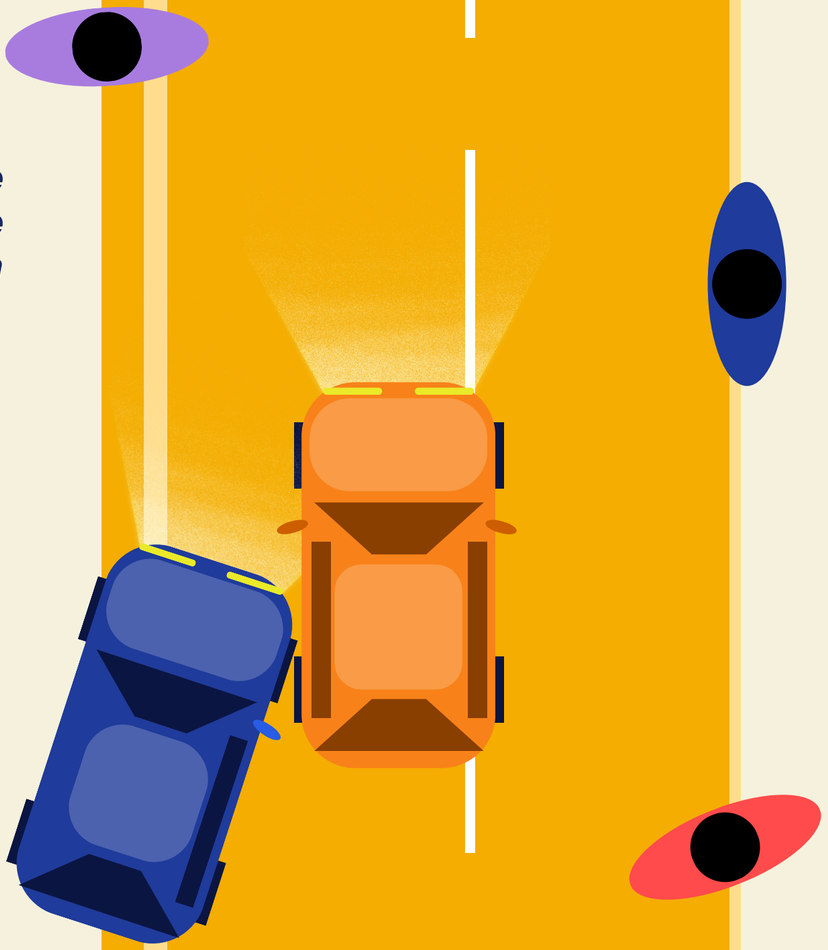
DRIVING TOWARD SAFER ROADS: OUR PROJECT GOAL

Our main goal is to develop a predictive model assessing car crash severity - damage extent, and passenger injury - in Maryland, USA. This endeavor serves as the cornerstone for formulating policy recommendations to bolster road safety through the effective implementation of traffic regulations.

- To achieve this, our approach involves consolidating information from a comprehensive dataset, ensuring relevance and precision. Key factors such as **traffic patterns**, **road infrastructure**, **demographics**, and **weather conditions** will be thoroughly examined both in our analysis and predictive modeling.

- The **core focus of our project is to expedite accident responses** and minimize their impact. This will be achieved through resource optimization - *e.g., roads redesign, real-time warnings, traffic lights adjustments, etc.* - and the formulation of **policy recommendations grounded in data-driven insights**.

- Recognizing the need for clear and understandable policies, we commit to employing a model that is **straightforward**, **instrumental** and **transparent**, facilitating informed decision-making for authorities. Looking ahead, we also plan to explore the possibility of grouping similar cities like Toronto or Chicago to tailor policy recommendations for even greater **effectiveness** and **scalability**.

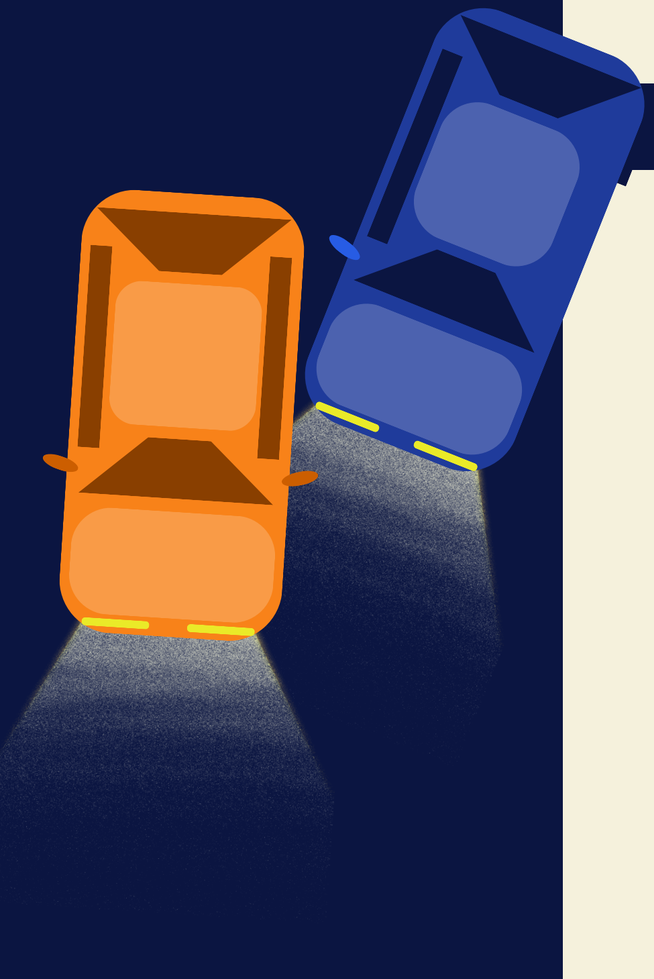


Our initiative aims to usher in a new era of road safety through innovation and strategic policymaking

PROBLEM STATEMENT

ENHANCING TORONTO ROAD SAFETY VIA MACHINE LEARNING AND MARYLAND INSIGHTS

*Our objective is to leverage machine learning to predict **car incident severity in Maryland**, drawing valuable insights from this specific location data. Our models aim to contribute to road safety by offering accurate incident predictions, empowering city authorities and drivers to adopt proactive measures. The integration of data of various types including - weather, driving styles, etc. - will enhance the expected predictive capabilities.*



ANTICIPATED MODEL BENEFITS & OUTCOMES

- **IMPROVED ROAD SAFETY**

The **final** model's accurate predictions will enable city authorities and law enforcement agencies to take **proactive steps**, preventing accidents and **reducing road fatalities**. It will also contribute to minimizing the likelihood of **incidents in high-risk areas**.

- **OPTIMIZED RESOURCE ALLOCATION**

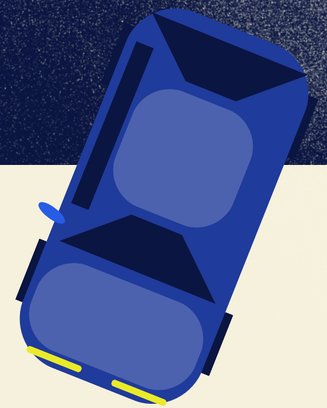
The model's insights can facilitate **the strategic allocation of resources**, such as **police patrols** (1) and **emergency response teams** (2) , to areas with a higher predicted risk of incidents. This ensures a more efficient and targeted use of resources.

- **INFORMED POLICY RECOMMENDATIONS**

The data-driven insights generated by the models will serve as a foundation for **developing policies** and **regulations** aimed at **enhancing road safety in Maryland**. For instance, the city can consider implementing stricter regulations during adverse weather conditions or in areas with a historically high incident risk. The policies may have **the potential to be scaled up** and reproduced in other cities with similar characteristics.

DATA SOURCES

RATIONALE BEHIND SELECTING SPECIFIC DATA SOURCES



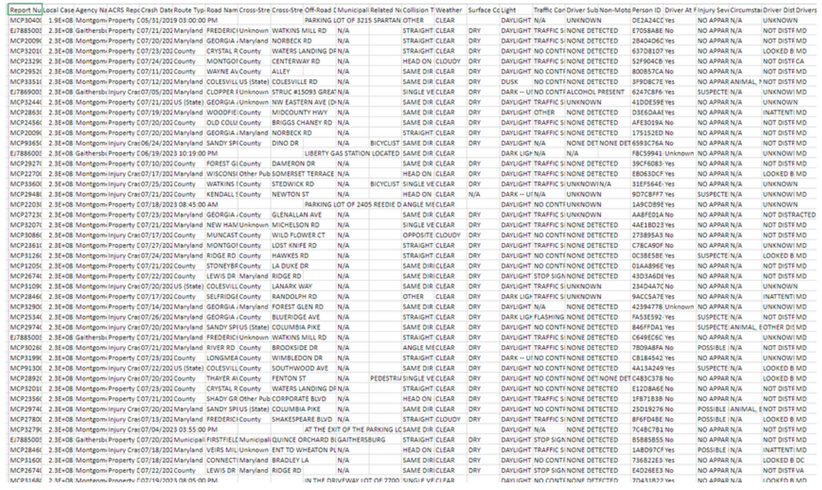
After comparing multiple data sources, we identified that the **metadata** and **accessibility** of data specific to **Maryland** played a crucial role in our selection process. Consequently, we opted for the "***Crash Reporting - Drivers Data***" dataset over others. Our **data acquisition strategy** is centered on utilizing information from certified and varied sources to enhance the robustness of our predictive model. The emphasis on metadata is pivotal, outlining the model's scope in predicting car crash severity, damage extent, and passenger injury levels within the city of Maryland. Our commitment to **credible** and **diverse** sources ensures a comprehensive **understanding of the variables that influence our predictive model**, leading to more accurate incident predictions and impactful policy recommendations.

Dataset	***Motor Vehicle Collisions Crashes, Chicago***	***Crash Reporting - Drivers Data***	***Transport Canada's "National Collision Database Online 1.0"***
Brief Description	*Offers detailed crash information within Chicago city limits under the jurisdiction of the Chicago Police Department.*	*Provides valuable insights into motor vehicle operators involved in traffic collisions on county and local roadways in Montgomery County, Maryland.*	*Contributes a comprehensive subset, encompassing police-reported motor vehicle collisions on public roads in Canada.*
	Link	*Link*	*Link*

DATA SOURCES

RATIONALE BEHIND SELECTING SPECIFIC DATA SOURCES

Outlook of our dataset



Report Nu	Local Case	Agency Na	ACRS Repo	Crash Date	Route Typ	Road Nam	Cross-Stre	Cross-Stre	Off-Road D	Municipali	Related Ni	Collision T	Weather	Surface Co	Light	Traffic Con	Driver Sub	Non-Moto	Person ID	Driver At F	Injury Sev	Circumsta	Driver Dist	Drivers
MCP3040(	1.9E+08	Montgom	Property	C05/31/2019	03:00:00	PM			PARKING	LOT OF 3215	SPARTAN	OTHER	CLEAR		DAYLIGHT	N/A	UNKNOWN		DE2A24CE	Yes	NO APPAR	N/A		UNKNOWN
EJ7885001	2.3E+08	Gaithersbı	Property	C07/21/202	Maryland	FREDERICI	Unknown	WATKINS	MILL RD	N/A		STRAIGHT	CLEAR	DRY	DAYLIGHT	TRAFFIC SI	NONE DETECTED		E7058ABE	No	NO APPAR	N/A		NOT DISTR	MD
MCP2009(	2.3E+08	Montgom	Property	C07/20/202	Maryland	GEORGIA /	Maryland	NORBECK	RD	N/A		STRAIGHT	CLEAR	DRY	DAYLIGHT	TRAFFIC SI	NONE DETECTED		28404D6D	Yes	NO APPAR	N/A		NOT DISTR	MD
MCP3201(	2.3E+08	Montgom	Property	C07/23/202	County	CRYSTAL R	County	WATERS	LANDING DF	N/A		STRAIGHT	CLEAR	DRY	DAYLIGHT	NO CONTI	NONE DETECTED		6370B107	Yes	NO APPAR	N/A		LOOKED B	MD
MCP2329(	2.3E+08	Montgom	Property	C07/24/202	County	MONTGOI	County	CENTERWAY	RD	N/A		HEAD ON	CLOUDY	DRY	DAYLIGHT	TRAFFIC SI	NONE DETECTED		52F904CB	Yes	NO APPAR	N/A		NOT DISTR	CA
MCP2952(	2.3E+08	Montgom	Property	C07/11/202	County	WAYNE AV	County	ALLEY		N/A		SAME DIR	CLEAR	DRY	DAYLIGHT	NO CONTI	NONE DETECTED		B00857CA	No	NO APPAR	N/A		NOT DISTR	MD
MCP3155(	2.3E+08	Montgom	Property	C07/12/202	Maryland	COLESVILL	US (State)	COLESVILL	RD	N/A		SAME DIR	CLEAR	DRY	DUSK	NO CONTI	NONE DETECTED		3F9D8C7E	Yes	NO APPAR	ANIMAL, I	NOT DISTR	MD
EJ7869001	2.3E+08	Gaithersbı	Injury Crac	07/05/202	Maryland	CLOPPER I	Unknown	STRUC #15093	GREAT	N/A		SINGLE VE	CLEAR	DRY	DARK -- UI	NO CONTI	ALCOHOL PRESENT		6247CBF6	Yes	SUSPECTE	N/A		UNKNOWI	MD
MCP3244(	2.3E+08	Montgom	Property	C07/21/202	US (State)	GEORGIA /	Unknown	NW EASTERN	AVE (DI	N/A		SAME DIR	CLEAR	DRY	DAYLIGHT	TRAFFIC SI	UNKNOWN		41DDE59E	Yes	NO APPAR	N/A		UNKNOWN	
MCP2863(	2.3E+08	Montgom	Property	C07/19/202	Maryland	WOODFIE	County	MIDCOUNTY	HWY	N/A		SAME DIR	CLEAR	DRY	DAYLIGHT	OTHER	NONE DETECTED		D3E6DAAE	Yes	NO APPAR	N/A		INATTENTI	MD
MCP2456(	2.3E+08	Montgom	Property	C07/20/202	County	OLD COLU	County	BRIGGS CHANEY	RD	N/A		SAME DIR	CLEAR	DRY	DAYLIGHT	TRAFFIC SI	NONE DETECTED		AFE30194	No	NO APPAR	N/A		NOT DISTR	MD
MCP2009(	2.3E+08	Montgom	Property	C07/20/202	Maryland	GEORGIA /	Maryland	NORBECK	RD	N/A		STRAIGHT	CLEAR	DRY	DAYLIGHT	TRAFFIC SI	NONE DETECTED		175152ED	No	NO APPAR	N/A		NOT DISTR	MD
MCP9365(	2.3E+08	Montgom	Injury Crac	06/24/202	Maryland	SANDY SPI	County	DINO DR		N/A	BICYCLIST	SAME DIR	CLEAR	DRY	DAYLIGHT	N/A	NONE DET	NONE DET	6593C76A	No	NO APPAR	N/A		NOT DISTR	MD
EJ7886001	2.3E+08	Gaithersbı	Property	C06/19/2023	10:19:00	PM			LIBERTY GAS	STATION	LOCATED	SAME DIR	CLEAR		DARK - LIGI	N/A	N/A		F8C59941	Unknown	NO APPAR	N/A		UNKNOWI	MD
MCP2927(	2.3E+08	Montgom	Property	C07/10/202	County	FOREST GI	County	DAMERON	DR	N/A		SAME DIR	CLEAR	DRY	DAYLIGHT	TRAFFIC SI	NONE DETECTED		39CF60B3	Yes	NO APPAR	N/A		NOT DISTR	MD
MCP2270(	2.3E+08	Montgom	Injury Crac	07/17/202	Maryland	WISCONSI	Other Pub	SOMERSET	TERRACE	N/A		HEAD ON	CLEAR	DRY	DAYLIGHT	TRAFFIC SI	NONE DETECTED		EB063DCF	Yes	NO APPAR	N/A		LOOKED B	MD
MCP3360(	2.3E+08	Montgom	Injury Crac	07/25/202	County	WATKINS I	County	STEDWICK	RD	N/A	BICYCLIST	SINGLE VE	CLEAR	DRY	DAYLIGHT	TRAFFIC SI	UNKNOWN	N/A	31EF564E	Yes	NO APPAR	N/A		UNKNOWI	
MCP3948(	2.3E+08	Montgom	Injury Crac	07/21/202	County	KENDALL (	County	NEWTON	ST			HEAD ON	CLEAR	N/A	DARK -- UI	N/A	UNKNOWN		807CBFF7	Yes	SUSPECTE	N/A		UNKNOWI	MD
MCP2203(	2.3E+08	Montgom	Property	C07/18/2023	08:45:00	AM			PARKING	LOT OF 2405	REEDIE D	ANGLE ME	CLEAR		DAYLIGHT	NO CONTI	UNKNOWN		1A9CDB9E	Yes	NO APPAR	N/A		UNKNOWN	
MCP2723(	2.3E+08	Montgom	Property	C07/23/202	Maryland	GEORGIA /	County	GLENALLAN	AVE	N/A		SAME DIR	CLEAR	DRY	DAYLIGHT	TRAFFIC SI	UNKNOWN		AABFE01A	No	NO APPAR	N/A		NOT DISTR	ACTED
MCP3207(	2.3E+08	Montgom	Property	C07/21/202	Maryland	NEW HAM	Unknown	MICHELSON	RD	N/A		SINGLE VE	CLEAR	DRY	DAYLIGHT	TRAFFIC SI	NONE DETECTED		4AE1BD23	Yes	NO APPAR	N/A		NOT DISTR	MD
MCP3086(	2.3E+08	Montgom	Injury Crac	07/17/202	County	MUNCAST	County	WILD FLOWER	CT	N/A		OPPOSITE	CLOUDY	DRY	DAYLIGHT	NO CONTI	NONE DETECTED		273895A3	No	NO APPAR	N/A		NOT DISTR	MD
MCP2361(	2.3E+08	Montgom	Property	C07/27/202	Maryland	MONTGOI	County	LOST KNIFE	RD	N/A		STRAIGHT	CLEAR	DRY	DAYLIGHT	TRAFFIC SI	NONE DETECTED		C7BCA90F	No	NO APPAR	N/A		UNKNOWI	MD
MCP3126(	2.3E+08	Montgom	Injury Crac	07/24/202	Maryland	RIDGE RD	County	HAWKES	RD	N/A		STRAIGHT	CLEAR	DRY	DAYLIGHT	NO CONTI	NONE DETECTED		0C385E8E	Yes	SUSPECTE	N/A		LOOKED B	MD
MCP3205(	2.3E+08	Montgom	Property	C07/21/202	County	STONEYBR	County	LA DUKE	DR	N/A		SAME DIR	CLEAR	DRY	DAYLIGHT	NO CONTI	NONE DETECTED		01AA8969	Yes	NO APPAR	N/A		NOT DISTR	MD
MCP2674(	2.3E+08	Montgom	Property	C07/22/202	County	LEWIS DR	Maryland	RIDGE	RD	N/A		SAME DIR	CLEAR	DRY	DAYLIGHT	STOP SIGN	NONE DETECTED		43D3A6D8	Yes	NO APPAR	N/A		NOT DISTR	MD
MCP3109(	2.3E+08	Montgom	Injury Crac	07/20/202	US (State)	COLESVILL	County	LANARK	WAY	N/A		SAME DIR	CLEAR	DRY	DAYLIGHT	TRAFFIC SI	UNKNOWN		234D4A7C	No	NO APPAR	N/A		UNKNOWN	
MCP2846(	2.3E+08	Montgom	Property	C07/17/202	County	SELFRIDG	County	RANDOLPH	RD	N/A		OTHER	CLEAR	DRY	DARK - LIGI	TRAFFIC SI	UNKNOWN		9ACC5A7E	Yes	NO APPAR	N/A		INATTENTI	MD
MCP3290(	2.3E+08	Montgom	Property	C07/14/202	Maryland	GEORGIA /	Maryland	FOREST GLEN	RD	N/A		SAME DIR	CLEAR	DRY	DAYLIGHT	N/A	NONE DETECTED		42394778	Unknown	NO APPAR	N/A		UNKNOWI	MD
MCP2534(	2.3E+08	Montgom	Property	C07/26/202	Maryland	GEORGIA /	County	BLUERIDGE	AVE	N/A		STRAIGHT	CLEAR	DRY	DARK - LIGI	FLASHING	NONE DETECTED		FA53E592	Yes	SUSPECTE	N/A		NOT DISTR	MD
MCP2974(	2.3E+08	Montgom	Property	C07/20/202	Maryland	SANDY SPI	US (State)	COLUMBIA	PIKE	N/A		SAME DIR	CLEAR	DRY	DAYLIGHT	NO CONTI	NONE DETECTED		846FFDA1	Yes	SUSPECTE	ANIMAL, E	OTHER DIS	MD
EJ7885001	2.3E+08	Gaithersbı	Property	C07/21/202	Maryland	FREDERICI	Unknown	WATKINS	MILL RD	N/A		STRAIGHT	CLEAR	DRY	DAYLIGHT	TRAFFIC SI	NONE DETECTED		C649EC6C	Yes	NO APPAR	N/A		UNKNOWI	MD
MCP3025(	2.3E+08	Montgom	Injury Crac	07/21/202	Maryland	RIVER RD	County	BROOKSIDE	DR	N/A		ANGLE ME	CLEAR	DRY	DAYLIGHT	TRAFFIC SI	NONE DETECTED		7809A8FA	No	POSSIBLE	N/A		NOT DISTR	MD
MCP3199(	2.3E+08	Montgom	Property	C07/23/202	County	LONGMEA	County	WIMBLEDON	DR	N/A		STRAIGHT	CLEAR	DRY	DARK -- UI	NO CONTI	NONE DETECTED		CB1B4542	Yes	NO APPAR	N/A		UNKNOWI	MD
MCP9130(	2.3E+08	Montgom	Injury Crac	07/22/202	US (State)	COLESVILL	County	SOUTHWOOD	AVE	N/A		SAME DIR	CLEAR	DRY	DAYLIGHT	NO CONTI	NONE DETECTED		4A13A249	Yes	SUSPECTE	N/A		LOOKED B	MD
MCP2892(	2.3E+08	Montgom	Property	C07/20/202	County	THAYER AV	County	FENTON	ST	N/A	PEDESTRI	SINGLE VE	CLEAR	DRY	DAYLIGHT	NO CONTI	NONE DET	NONE DET	C4B1C37B	No	NO APPAR	N/A		LOOKED B	MD
MCP3201(	2.3E+08	Montgom	Property	C07/23/202	County	CRYSTAL R	County	WATERS	LANDING DF	N/A		STRAIGHT	CLEAR	DRY	DAYLIGHT	NO CONTI	NONE DETECTED		E120BA6E	No	NO APPAR	N/A		NOT DISTR	MD
MCP2356(	2.3E+08	Montgom	Property	C07/21/202	County	SHADY GR	Other Pub	CORPORATE	BLVD	N/A		HEAD ON	CLEAR	DRY	DAYLIGHT	TRAFFIC SI	NONE DETECTED		1FB71B3B	No	NO APPAR	N/A		NOT DISTR	MD
MCP2974(	2.3E+08	Montgom	Property	C07/20/202	Maryland	SANDY SPI	US (State)	COLUMBIA	PIKE	N/A		SAME DIR	CLEAR	DRY	DAYLIGHT	NO CONTI	NONE DETECTED		25D19276	No	POSSIBLE	ANIMAL, E	NOT DISTR	MD
MCP2780(	2.3E+08	Montgom	Injury Crac	07/13/202	Maryland	FREDERICI	County	SHAKESPEARE	BLVD	N/A		STRAIGHT	CLOUDY	DRY	DAYLIGHT	TRAFFIC SI	NONE DETECTED		8F6FD4BE	No	POSSIBLE	N/A		LOOKED B	MD
MCP3279(	2.3E+08	Montgom	Injury Crac	07/04/2023	03:55:00	PM			AT THE EXIT	OF THE PARKING LC		SAME DIR	CLEAR		DAYLIGHT	N/A	NONE DETECTED		7C48C7B1	No	NO APPAR	N/A		NOT DISTR	MD
EJ7885001	2.3E+08	Gaithersbı	Property	C07/20/202	Municipali	FIRSTFIELI	Municipali	QUINCE ORCHARD BI		GAITHERSBURG		STRAIGHT	CLEAR	DRY	DAYLIGHT	STOP SIGN	NONE DETECTED		85B85B55	No	NO APPAR	N/A		NOT DISTR	MD
MCP2846(	2.3E+08	Montgom	Injury Crac	07/18/202	Maryland	VEIRS MIL	Unknown	ENT TO WHEATON	PLN	N/A		HEAD ON	CLEAR	DRY	DAYLIGHT	TRAFFIC SI	NONE DETECTED		1ABD97CF	Yes	POSSIBLE	N/A		INATTENTI	MD
MCP3160(	2.3E+08	Montgom	Property	C07/18/202	Maryland	CONNECTI	Maryland	BRADLEY	LA	N/A		SAME DIR	CLEAR	DRY	DAYLIGHT	NO CONTI	NONE DETECTED		73682261	Yes	NO APPAR	N/A		LOOKED B	DC
MCP2674(	2.3E+08	Montgom	Property	C07/22/202	County	LEWIS DR	Maryland	RIDGE	RD	N/A		SAME DIR	CLEAR	DRY	DAYLIGHT	STOP SIGN	NONE DETECTED		E4D26EE8	No	NO APPAR	N/A		NOT DISTR	VA
MCP3168(	2.3E+08	Montgom	Property	C07/19/2023	08:05:00	PM			IN THE DRIVEWAY	LOT OF 7200		SINGLE VE	CLEAR		DAYLIGHT	NO CONTI	NONE DETECTED		3D451B23	Yes	NO APPAR	N/A		LOOKED B	MD

Selected Dataset

Crash Reporting - Drivers Data

Provides valuable insights into motor vehicle operators involved in traffic collisions on county and local roadways in Montgomery County, Maryland.

Link

Transport Canada's "National Collision Database Online 1.0"

Contributes a comprehensive subset, encompassing police-reported motor vehicle collisions on public roads in Canada.

Link

Brief Description

city limits under the jurisdiction of the Chicago Police Department.

Link

DATASET DESCRIPTION

DATA LANDSCAPE: KEY COMPONENTS SHAPING OUR PREDICTIVE MODEL

Variable	Description	Metadata	Example
Report Number	***Unique identifier for the crash report***	*Alphanumeric*	*'MCP3040003N'*
Local Case Number	***Local case identifier for the crash report***	*Numerical*	*'190026050'*
Agency Name	***Name of the agency reporting the crash***	*String*	*'Montgomery County Police*
ACRS Report Type	***Type of report generated by the ACRS***	*String*	*'Property Damage Crash', 'Injury Crash'*
Crash Date/Time	***Date and time when the crash occurred***	*Date/Timestamp*	*'05/31/2019 03:00:00 PM'*
Route Type	***Type of route where the crash occurred***	*String*	*'Maryland (State)'*
Road Name	***Name of the road where the crash occurred***	*String*	*'FREDERICK RD'*

(A)

DATASET DESCRIPTION

DATA LANDSCAPE: KEY COMPONENTS SHAPING OUR PREDICTIVE MODEL

(B)

Variable	Description	Metadata	Example
Cross-Street Type	***Type of the intersecting street***	*String*	*'Maryland (State)'*
Cross-Street Name	***Name of the intersecting street***	*String*	*'WATKINS MILL RD', 'NORBECK RD'*
Off-Road Description	***Description of the off-road crash location***	*Text*	*'PARKING LOT OF 3215 SPARTAN RD'*
Municipality	***Municipality where the crash occurred***	*String*	*'GAITHERSBURG', 'CHEVY CHASE #4'*
Non-Motorist	***Info about non-motorists involved***	*String*	*'BICYCLIST', 'PEDESTRIAN'*
Collision Type	***Type of collision (e.g., rear-end, head-on)***	*Text*	*'STRAIGHT MOVEMENT ANGLE'*
Weather	***Weather conditions - crash time***	*String*	*'CLEAR', 'CLOUDY'*

DATASET DESCRIPTION

DATA LANDSCAPE: KEY COMPONENTS SHAPING OUR PREDICTIVE MODEL

(C)

Variable	Description	Metadata	Example
Surface Condition	***Road surface conditions - crash time***	*String*	*'DRY', 'WET'*
Light	***Lighting conditions - crash time***	*String*	*'DAYLIGHT', 'DUSK'*
Traffic Control	***Type of traffic control - crash location***	*String*	*'TRAFFIC SIGNAL', 'NO CONTROLS'*
Driver Subs. Abuse	***Info - driver substance abuse***	*String*	*'NONE DETECTED', 'ALCOHOL PRESENT'*
N-Motorist Subs Ab.	***Info - non-motorist substance abuse***	*String*	*'NONE DETECTED'*
Person ID	***Unique identifier for individuals involved***	*Alphanumeric*	*'DE2A24CD-7919-4F8D-BABF-5B75CE12D21E'*
Driver At Fault	***Indicates whether the driver was at fault***	*Binary*	*'Yes', 'No',*

DATASET DESCRIPTION

DATA LANDSCAPE: KEY COMPONENTS SHAPING OUR PREDICTIVE MODEL

Variable	Description	Metadata	Example
Injury Severity	***Severity of injuries sustained in the crash***	*String*	*'NO APPARENT INJURY', 'SUSPECTED'*
Circumstance	***Circumstances surrounding the crash***	*String*	*'ANIMAL, N/A'*
Driver Distracted By	***Factors of driver distraction - crash time***	*String*	*'NOT DISTRACTED'*
Vehicle ID	***Unique identifier for vehicles involved in the crash***	*Alphanumeric*	*'165AD539-A8C8-4004-AF73-B7DCAAA8B3CC'*
Vehicle Damage Extent	***Extent of damage to the vehicle***	*String*	*'SUPERFICIAL', 'DISABLING'*
Vehicle 1st Impact Location	***Location on the vehicle where the first impact occurred***	*String*	*'ONE OCLOCK'*

(D)

DATASET DESCRIPTION

DATA LANDSCAPE: KEY COMPONENTS SHAPING OUR PREDICTIVE MODEL

Variable	Description	Metadata	Example
Vehicle 2nd Impact Location	***Location on the vehicle where the 2nd impact occurred***	*String*	*'ONE OCLOCK'*
Vehicle Body Type	***Type of vehicle body***	*String*	'PASSENGER CAR', 'PICKUP TRUCK'
Vehicle Movement	***Movement of the vehicle - crash time***	*String*	'PARKING', 'MAKING LEFT TURN'
Vehicle Cont. Dir	***Direction in of vehicle continuing***	*String*	'North', 'East'
Vehicle Going Dir	***Direction in of vehicle going***	*String*	'North', 'East'
Speed Limit	***Posted speed limit at the crash location***	*Numerical*	15, 40, 35
Driverless Vehicle	***Indicates if the vehicle was driverless***	*String*	'No', 'Unknown'

(E)

DATASET DESCRIPTION

DATA LANDSCAPE: KEY COMPONENTS SHAPING OUR PREDICTIVE MODEL

Variable	Description	Metadata	Example
Parked Vehicle	***Indicates if the vehicle was parked***	*Binary*	*'No', 'Yes'*
Vehicle Year	***Year of manufacture of the vehicle***	*Numerical*	*2004, 2011, 2019*
Vehicle Make	***Make or manufacturer of the vehicle***	*String*	*'HONDA', 'GMC', 'FORD'*
Vehicle Model	***Model of the vehicle***	*Alphanumeric*	*'TK', 'F150', 'SW'*
Equipment Problems	***Any equipment problems reported***	*String*	*'NO MISUSE', nan*
Latitude/Longitude	***Latitude/Longitude coordinates of the crash location***	*Numerical*	*'(39.15004368, -77.06308884)'*
Location	***Crash location***	*Numerical*	*'NO MISUSE', nan*

(F)

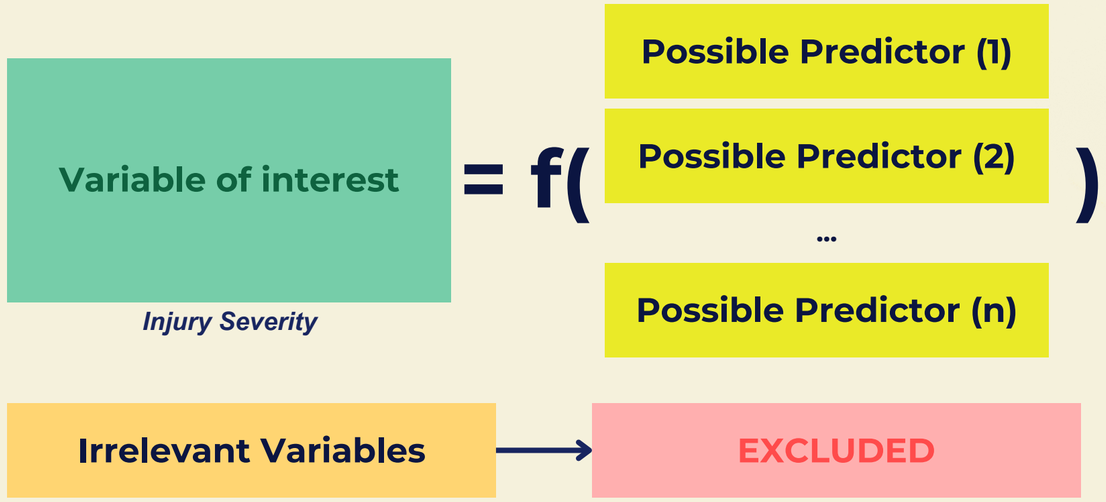
DATASET DESCRIPTION

EXPLORING OUR DATASET: UNRAVELING DIMENSIONS AND RELATIONSHIPS

Target Variable: Injury Severity serves as the pivotal target variable. This **categorical variable** encompasses various possible values, playing a crucial role in categorizing and understanding the diverse outcomes associated with car incidents. Utilizing a range of **predetermined predictors within our dataset**, our model seeks to unveil the diverse variables influencing injury severity, offering **nuanced insights** into potential outcomes. This approach contributes to a more comprehensive understanding of the multifaceted impact of accidents.

Variable Insights and Assumptions

- **Variable Characteristics:**
  - *Predominance of categorical variables.*
- **Weather and Surface Conditions:**
  - *Anticipated correlation between poor weather & surface conditions leading to more severe injuries/accidents.*
- **Driver Factors:**
  - *Expect drivers under the influence of substance abuse and those distracted to be associated with more severe injuries and accidents.*
- **Speed Limit Relationship:**
  - *Potential relationship between the speed limit and the severity of accidents, recognizing it as a factor influencing outcomes.*



The relationship between the target variable (**Y**) and predictor variables (**Xi**) is fundamental to our analysis. Y represents the target variable, which is the focal point of our investigation, while Xi encompasses the predictor variables that contribute to shaping the relationship with Y. We will **exclude irrelevant variables** from our analysis.

DATA PRE-PROCESSING

STEP 1

Data Cleaning and Handling Missing Values: Initiating the Pre-processing
The first step in preparing our dataset involved thorough data cleaning and addressing missing values. This initial phase set the foundation for a robust pre-processing strategy, ensuring the integrity and completeness of our data before further analysis and modeling.

Original Dataset

43 variables (columns)	160,000 records (rows)

1. **Drop the significant missing values variables:**
   'Off-Road Description', 'Municipality', 'Related Non-Motorist', 'Non-Motorist Substance Abuse', 'Circumstance', 'Equipment Problems'
2. **Dropped rows with '*OTHER*', and '*UNKNOWN*' values.**
3. **Removed columns with extensive *NA* values.**
4. **Dropped irrelevant columns/variables**
   Eliminated irrelevant columns like 'Report Number', 'Road Name', etc.
   - *N.B.: Irrelevant variables have been identified beforehand; Cf. Data Description*

27 variables (columns)	66, 640 records (rows)

Final Dataset*

****Before Encoding & Feature Engineering***

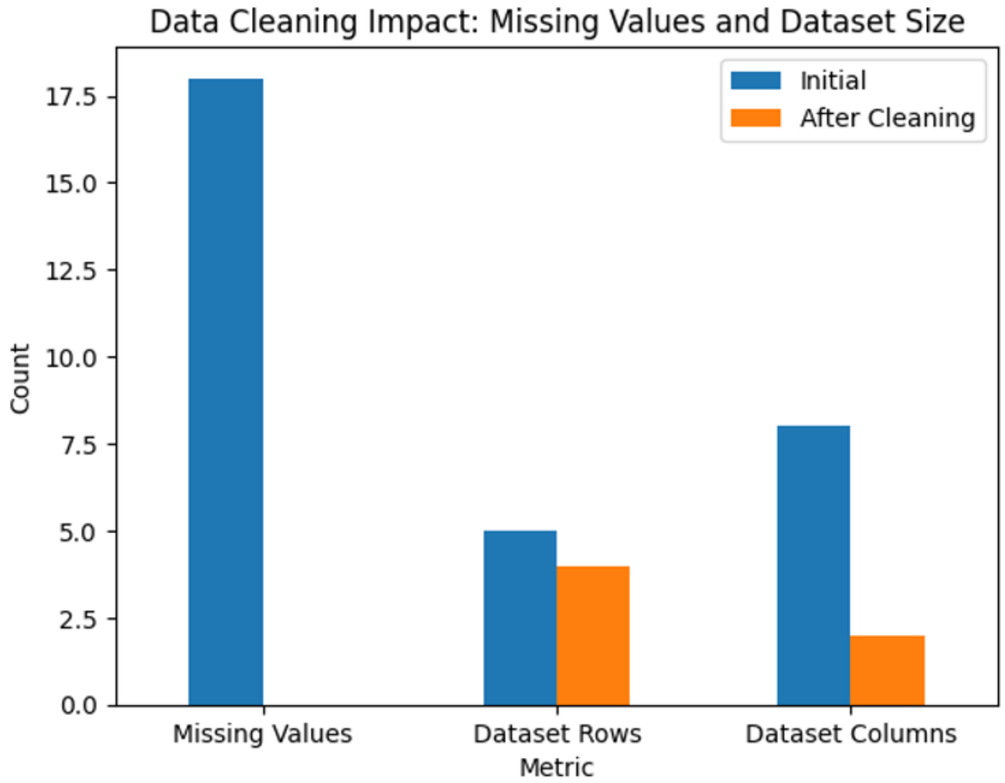


Figure 1. Influence of Data Cleaning on Missing Values and Dataset Size

DATA PRE-PROCESSING

STEP 2

Categorical Encoding and Temporal Insight Enhancement: Elevating the Data Structure

In the second phase of pre-processing, we elevated our data's structure by strategically encoding categorical variables, a decision guided by a meticulous assessment of unique category counts per column. Simultaneously, we enhanced temporal insights by transforming the 'Crash Date/Time' variable into a datetime format and extracting key temporal components, fostering a higher-level understanding of the dataset's temporal dynamics. These comprehensive adjustments set the stage for more refined and insightful model development.

Categorical Encoding

- **Checked with the unique category count per column:**
  Define which variable needs to become categorical.
- **Emphasized the need for encoding by showcasing the count of unique categories per column.**

Date/Time Feature Engineering

- **Transformed 'Crash Date/Time' into datetime format.**
- **Extracted and created 'Hour', 'Day', 'Month', 'Year', 'Season', 'Time of Day', and 'Month Segment'.**

	Crash Date/Time	Hour	Day	Month	Year	Some Other Column	Date Day	Date Year	Season
0	2023-01-01 15:30:00	15	1	1	2023	1	1	2023	Winter
1	2023-02-15 08:45:00	8	15	2	2023	2	15	2023	Winter
2	2023-03-20 22:10:00	22	20	3	2023	3	20	2023	Spring

The following table provides precise details on the date and time of each crash, along with corresponding seasonal information.

For instance, the second entry indicates that the crash took place on February 15, 2023, during winter, at 8:45 am.

DATA PRE-PROCESSING

STEP 3

Geospatial Insight Unveiling and Dataset Refinement: Preparing for Analysis

In the third pre-processing step, our focus is on unlocking geospatial patterns within crash data. By centering on 'Longitude' and 'Latitude' and utilizing the K-Means clustering technique tailored for geographical insights, we lay the groundwork for insightful clustering. The final dataset is thoughtfully refined, excluding 'Latitude,' 'Longitude,' and extraneous columns, ensuring a streamlined and optimized structure for robust analysis and exploration of geospatial patterns in crash occurrences.

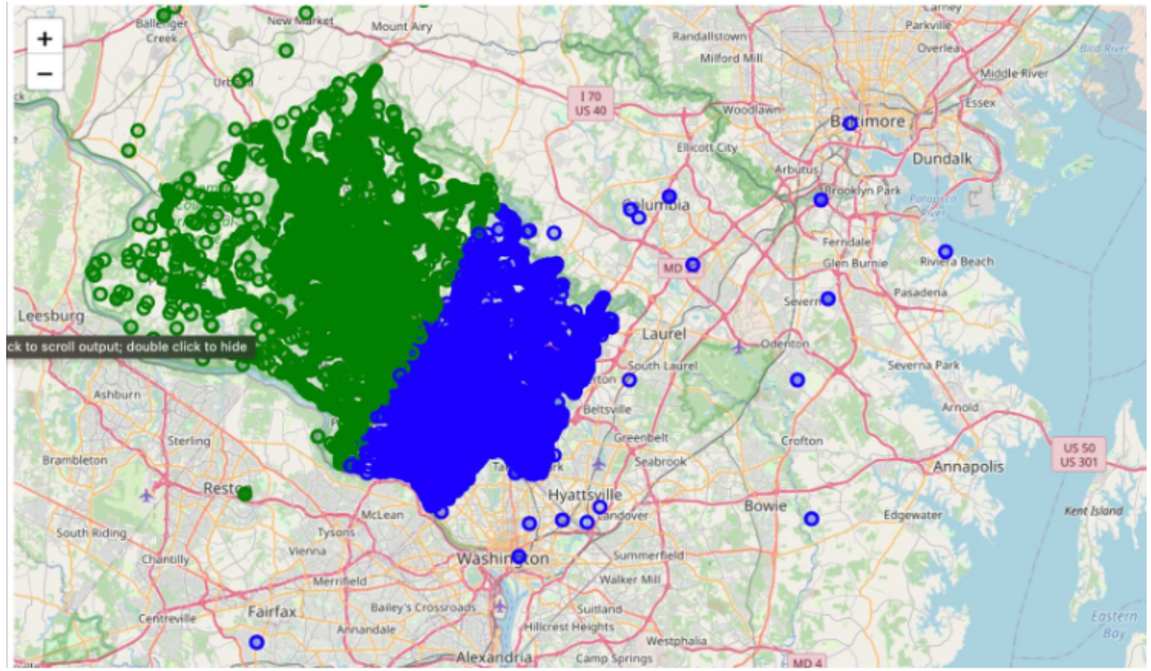


Figure 2. Map-Based Geospatial Pattern Analysis: K-Means Clustering in Crash Data Pre-processing

Clustering Rationale:

1. **Exploring Geographical Patterns in Crash Data:**
   - *Unveiling insights into the spatial distribution and clustering tendencies of crash occurrences.*

2. **Spatial Focus**
   - *Concentrating on 'Longitude' and 'Latitude' as key variables for meaningful spatial clustering analysis.*

Scaling for Clustering: Tailoring Techniques for Geospatial Insights

In the process of scaling our data for clustering, we strategically employ the K-Means clustering technique with a specific focus on its applicability for geospatial analysis. Recognizing the unique requirements of our geographical purpose, K-Means provides an effective framework for uncovering spatial patterns and clusters within the crash data. This tailored approach enhances the precision and relevance of our clustering analysis for geographic insights.Clustering techniques: K-Means for geo purpose

Final DataFrame for Analysis
Excluded 'Latitude', 'Longitude', and additional non-relevant columns.

The dataset is now streamlined for effective modelling and cluster analysis

DATA PRE-PROCESSING

STEP 4

Streamlining Injury Severity for Enhanced Model Interpretation

In the fourth pre-processing step, we code the 'Injury Severity' variable, thus reducing complexity, focusing on injury presence, mitigating potential class imbalance, and enhancing model interpretability by emphasizing the impact of 'injury vs. no injury' in our analyses.

Original Categorical Levels

- 'POSSIBLE INJURY', 'SUSPECTED MINOR INJURY', 'SUSPECTED SERIOUS INJURY', 'FATAL INJURY', 'NO APPARENT INJURY'.

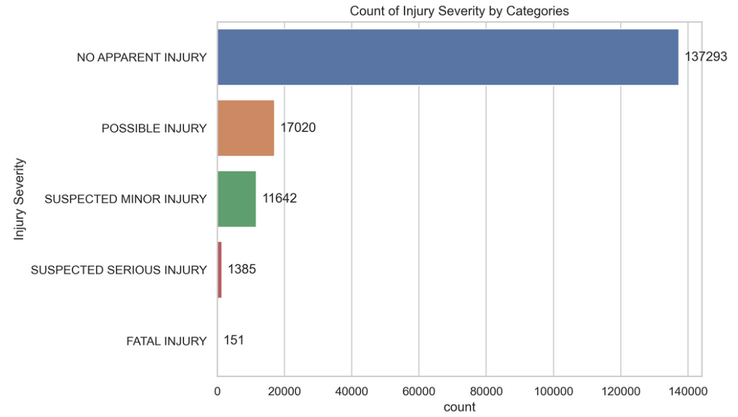
Transformation Approach

- Any injury ('POSSIBLE' to 'FATAL') is coded as 1, while 'NO APPARENT INJURY' is coded as 0.

Reasons for Transformation

- Reduces complexity by focusing on the presence of injury irrespective of severity.
- Addresses potential class imbalance by consolidating underrepresented injury levels.
- Improves interpretability: coefficients in models explain the effect of 'injury vs. no injury'.

Pre-transformation



Post-transformation

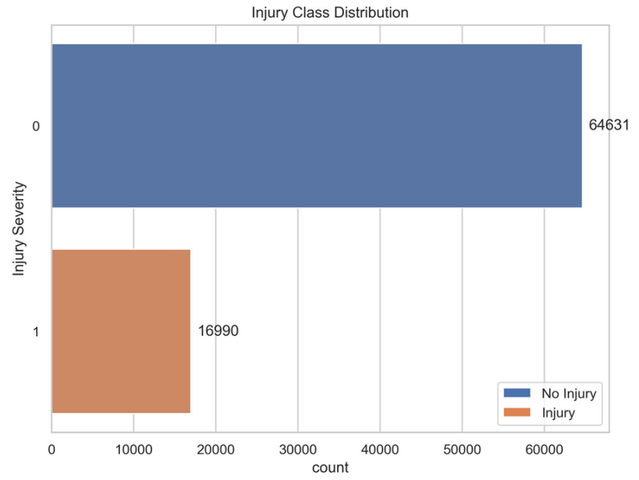


Figure 3. Mitigating Class Imbalance in Injury Severity through Encoding

Post Processing

In the post-processing phase, we examine the distribution after encoding, revealing that around 79% of instances are categorized as 'no injury' (coded as 0), while approximately 21% indicate 'injury' (coded as 1). This distribution serves as a crucial foundation for predictive modeling focused on injury occurrence, providing insight into the prevalence of injuries within the dataset.

Figure 1. XXXXXXXXXXXXX

EXPLORATORY DATA ANALYSIS

UNCOVERING MEANINGFUL INSIGHTS

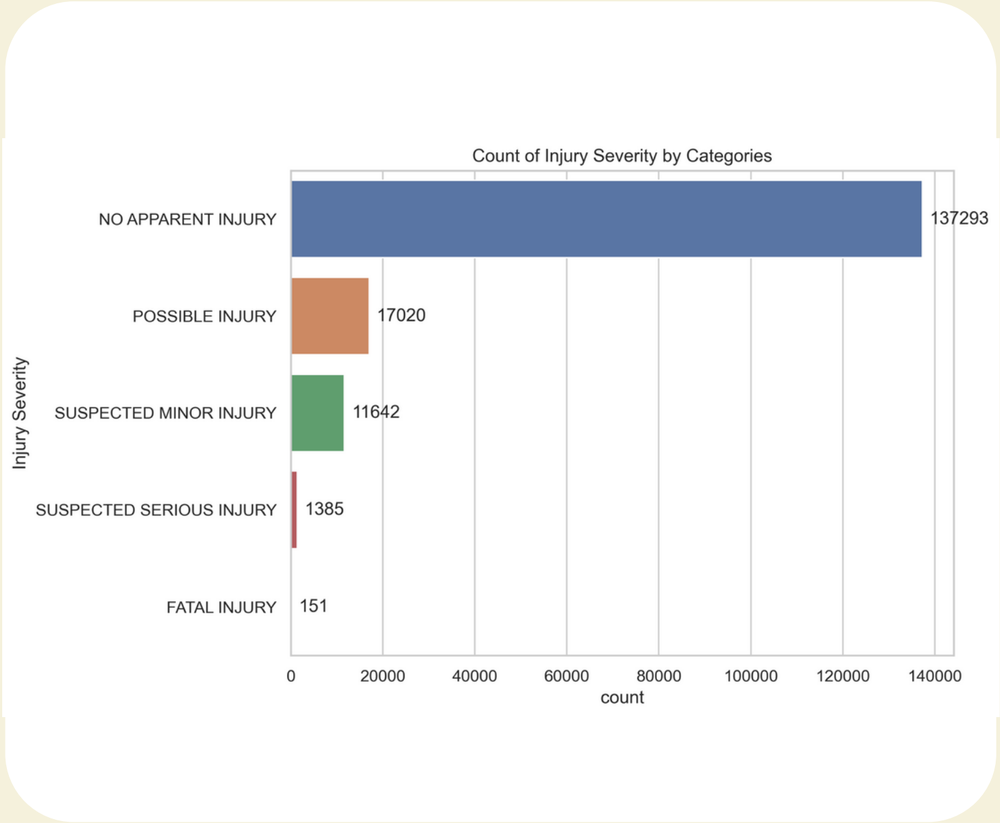


Figure 4. Injury Severity Distribution Across Categories

Exploratory Data Analysis: Navigating Insights Through Data Exploration

Exploratory Data Analysis (EDA) acts as our guide through the intricate dataset landscape, unveiling hidden patterns and illuminating crucial insights. From deciphering injury severity distribution to understanding temporal and geographical influences on accidents, each facet undergoes meticulous examination. EDA serves as the lens through which we decode the data's narrative, laying the groundwork for robust modeling and informed decision-making, transforming raw data into actionable knowledge.

1. The majority of accidents resulted in **no apparent injuries**, underscoring that a significant proportion of incidents do not cause physical harm.

2. Crashes with **possible injuries** or **suspected minor injuries** are comparatively less frequent, indicating that severe outcomes are less prevalent.

3. Instances of **suspected serious injuries** and **fatal injuries** are minimal in comparison to other categories, emphasizing the rarity of these tragic outcomes.

EXPLORATORY DATA ANALYSIS

UNCOVERING MEANINGFUL INSIGHTS

NUMBER OF CRASHERS PER YEAR

Crash Trends Over the Years

- The trend line reveals a peak in crash numbers in the initial years, particularly in 2016, suggesting a period of heightened accident rates.
- Subsequently, there is a slight decline until the beginning of 2019, which may be attributed to effective safety measures or changes in reporting practices.
- The following years exhibit fluctuations, with a noticeable dip in 2020, likely influenced by reduced travel during the pandemic.
- A recovery in crash numbers is evident in 2021, followed by another in 2022, signifying a gradual return to office activities observed after the initial COVID-19 lockdown, resulting in increased traffic on the roads.

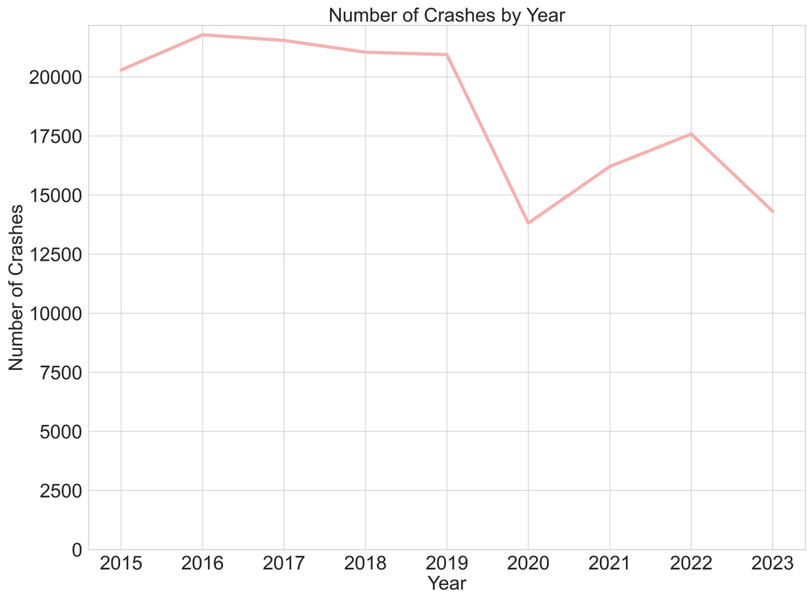


Figure 5. Yearly Crash Incidents: Count of Collisions

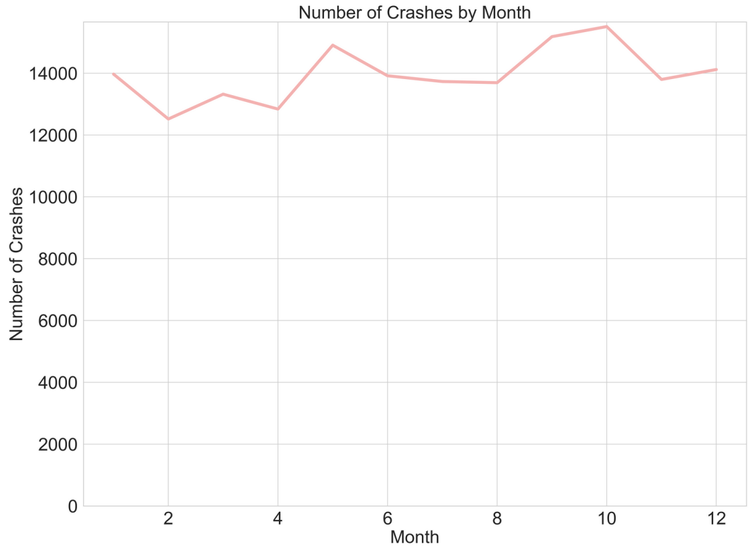


Figure 6. Monthly Crash Incidents: Count of Collisions

NUMBER OF CRASHERS PER MONTH

Crash Frequency By Month

- The data reflects a generally stable trend with subtle fluctuations in the number of crashes across the months.
- Observable peaks in certain months may align with seasonal factors influencing driving conditions or travel patterns.
- The graph exhibits a mild seasonal pattern, suggesting a correlation between crashes and seasonal variations, such as a decrease in outdoor activities during winter.
- A notable spike in October could potentially be linked to the period of tire change between summer and winter, where the slipperiness of summer tires may contribute to increased accidents if not replaced.

EXPLORATORY DATA ANALYSIS

UNCOVERING MEANINGFUL INSIGHTS

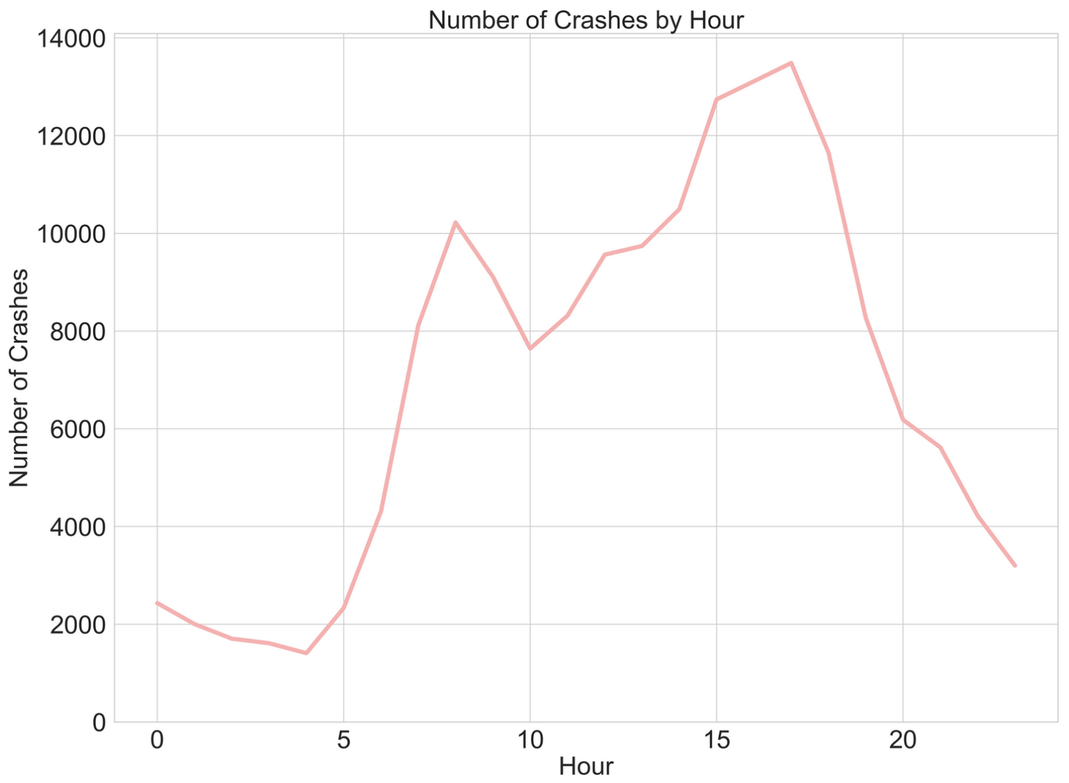


Figure 7. Hourly Crash Incidents: Count of Collisions

NUMBER OF CRASHERS PER HOUR

Crash Trends By Hour

- A notable surge in crashes occurs during the early hours, peaking around the morning commute, indicating heightened risk during this period.
- A subsequent decline in crash frequency is observed during midday hours, potentially correlating with reduced traffic volume.
- The late afternoon to early evening witnesses another significant increase, aligning with typical evening rush hours.
- The heightened crash rate in the afternoon may be attributed to factors like fatigue and hunger, as individuals drive back home after a long workday.
- A sharp decline in crashes occurs after peak evening hours, reflecting reduced traffic as the night progresses.
- The lowest crash occurrences are noted during late-night to early-morning hours, likely due to fewer vehicles on the road during this time.

EXPLORATORY DATA ANALYSIS

UNCOVERING MEANINGFUL INSIGHTS

Geospatial Analysis for Traffic Safety: Plotting Crash Data

The map offers a vivid representation of a concentrated cluster of crashes within a specific geographic area, notably reported by the department. Fatal crashes, depicted in red, are widespread within this central cluster, underscoring the necessity for targeted safety measures in this region. Conversely, the outer regions exhibit fewer crashes, suggesting potential out-of-area reporting. Notably, the concentration of crashes near major roads and intersections may indicate high traffic volumes or potentially hazardous driving conditions. This geospatial distribution pattern becomes a valuable tool for prioritizing areas for traffic safety improvements and optimizing resource deployment for emergency response.

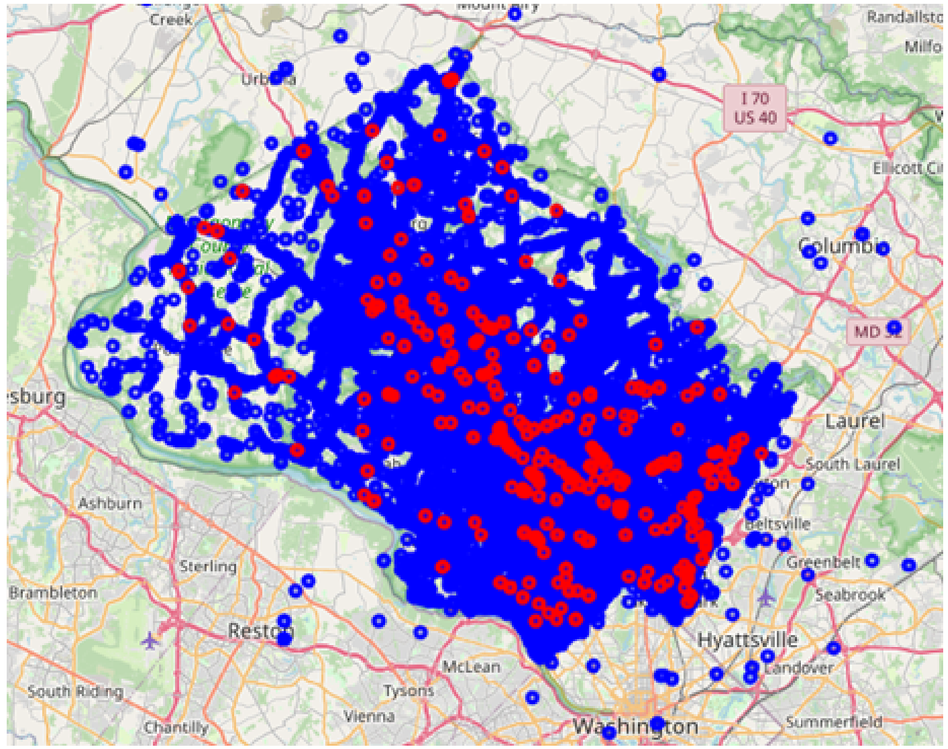
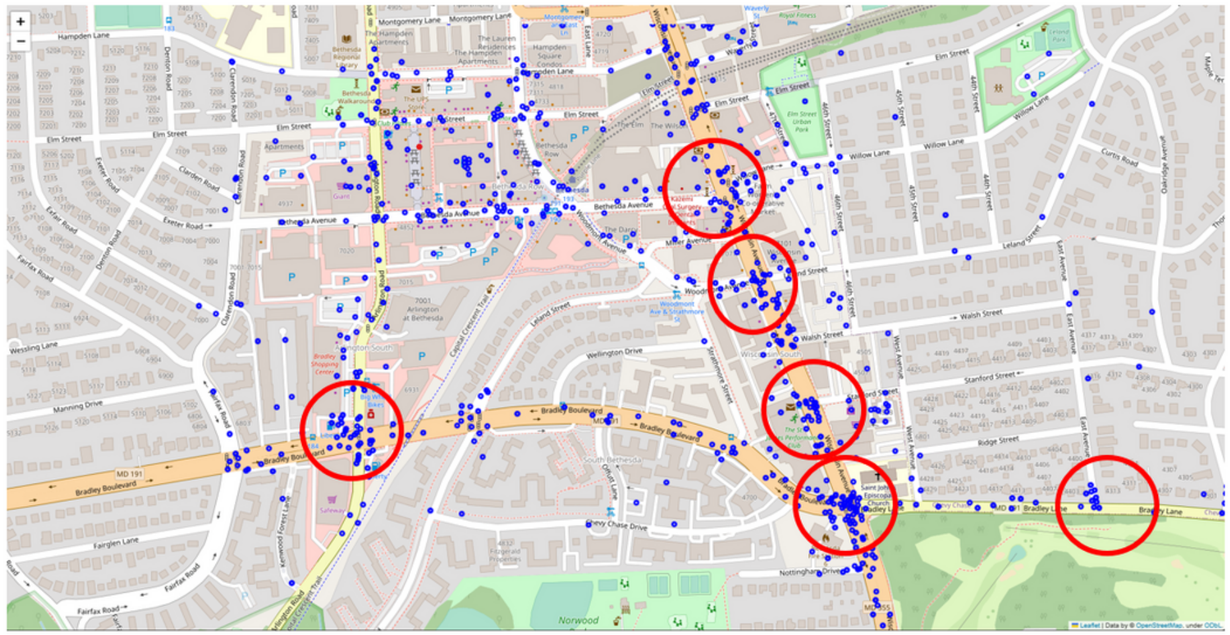


Figure 8. Map of Concentrated Crashes: Identifying Hotspots and Prioritizing Safety Measures

EXPLORATORY DATA ANALYSIS

UNCOVERING MEANINGFUL INSIGHTS

ZOOMED-IN VIEW



Observing the Map at a Closer Scale

- Upon closer examination of the map, it becomes evident that the majority of crashes are concentrated on major roads, particularly at intersections.

MODELING

MODEL SELECTION CRITERIA

In selecting algorithms for our use case, three primary criteria guided our choices: model interpretability, performance considering the dataset's size, and overall predictive capability.

Considered Algorithms:

- Two tree-based ensemble methods:
  - Random Forest
  - Extreme Gradient Boosting (XGBoost)
- Logistic Regression

Logistic Regression	XGBoost & Random Forest
Better Interpretability	Balance of Interpretability + Performance

Considerations

- Tree-based algorithms are generally effective with high-dimensional data. The dataset, comprising 166,537 observations and 43 features, benefits from this approach.
- XGBoost was preferred over Gradient Boosting due to its incorporation of regularization (L1 and L2) in the loss function, crucial for preventing overfitting in high-dimensional datasets.
- Additionally, XGBoost offers faster performance.

MODELING

MODEL BUILDING PROCESSS

The model-building process was iterative, involving hyperparameter tuning and data manipulation.

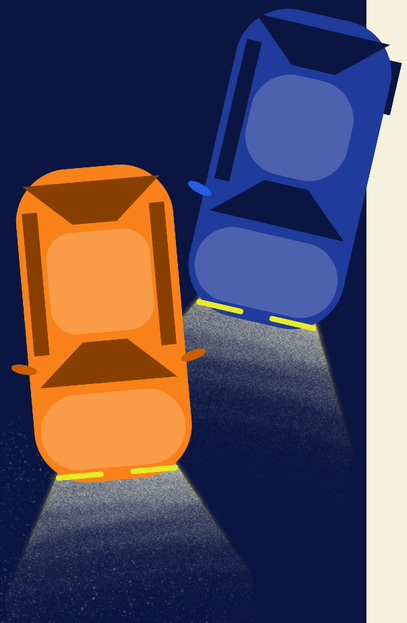
- A significant class imbalance was noted, with the 'Injury' class making up only 20% of the dataset. This impacts the model's ability to classify the minority class effectively.
- Initial models were built with the *class_weight* hyperparameter set to 'balanced' to internally address this imbalance.

- In the absence of a ***class_weight*** parameter in **XGBoost**, we utilized the ***scale_pos_weight*** parameter to address the challenge of class imbalance.

- Despite initial efforts to tackle class imbalance, further refinement was necessary. We explored **Randomized Undersampling, Randomized Oversampling**, and **Synthetic Minority Oversampling Technique (SMOTE)** as potential solutions to rebalance the classes.

- The **XGBoost** model exhibited the best performance when trained on the undersampled dataset. This success can be attributed to the extensive number of observations, enabling effective downsampling without significant loss of information.

- Contrarily, oversampling techniques resulted in **overfitting** and **unsatisfactory model performance**, highlighting the importance of selecting the appropriate method to address class imbalance.

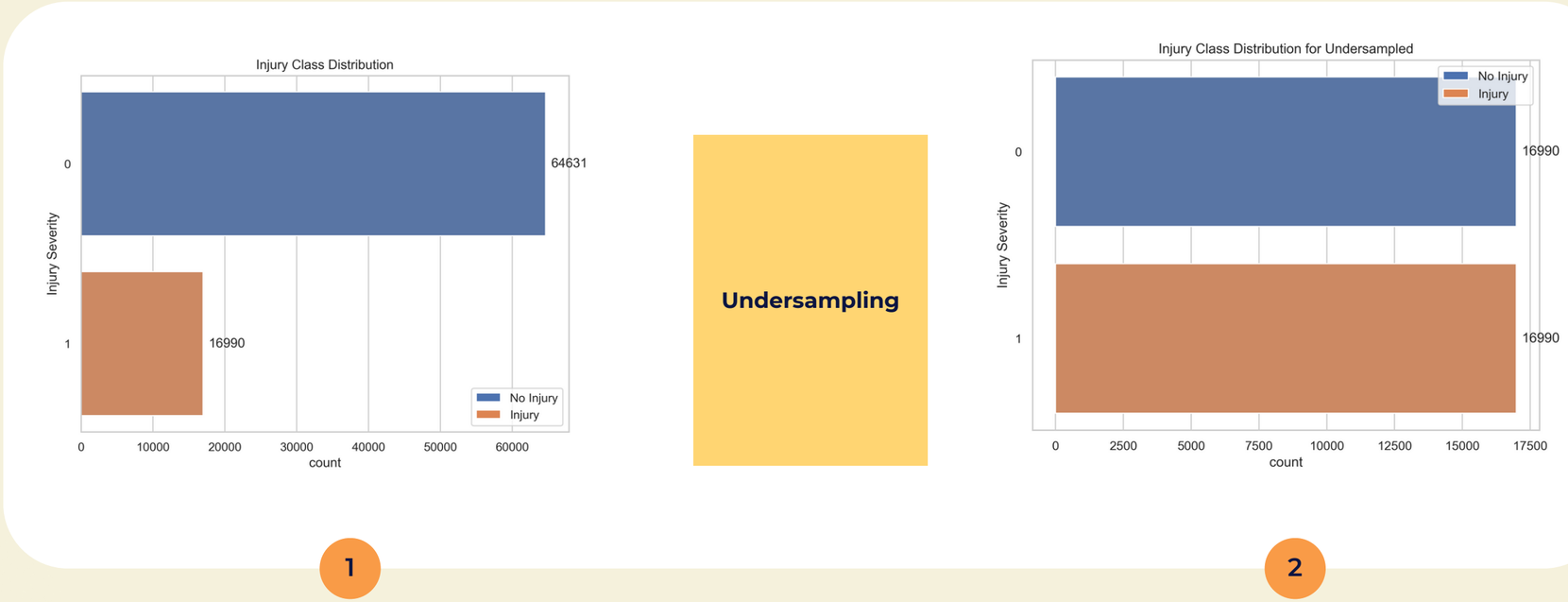


MODELING

ADDRESSING CLASS IMBALANCE

In this critical step of our modeling process, we grapple with the **challenge of class imbalance.** After exploring various class resampling techniques, **undersampling** is found to be the **most effective** on the dataset. Specifically, observations from the majority class are randomly removed with replacement until a desired level of balance is achieved. This approach ensures that our model is not disproportionately influenced by the overrepresented class, enhancing its ability to discern patterns and make accurate predictions across all classes. After addressing the class imbalance, **hyperparameter tuning** is undertaken, **optimizing the configuration of our model** to achieve the best performance. This meticulous process contributes to the robustness and efficacy of our predictive model in handling the intricacies of the dataset.

Figure 9. Under-sampling for Class Imbalance Mitigation in Model Optimization



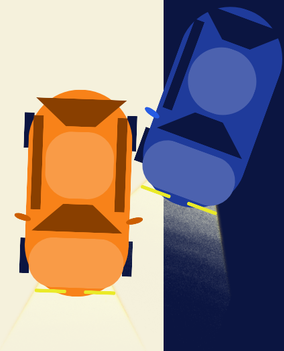
1

2

MODELING

HYPERPARAMETER TUNING AND THRESHOLD MOVING

To enhance the XGBoost model performance on the **downsampled** data, Randomized Search CV was employed for tuning parameters such as *alpha*, *max_depth*, *n_estimators*, *learning_rate*, and *subsample*. Randomized Search CV was more efficient than Grid Search due to the dataset's size and the number of parameters.



Baseline Model

	precision	recall	f1-score	support
No Injury	0.90	0.63	0.74	12912
Injury	0.33	0.72	0.46	3313
accuracy			0.65	16225
macro avg	0.62	0.67	0.60	16225
weighted avg	0.78	0.65	0.68	16225

Performance after Undersampling & Tuning

	precision	recall	f1-score	support
0	0.70	0.58	0.64	3384
1	0.64	0.75	0.69	3374
accuracy			0.67	6758
macro avg	0.67	0.67	0.66	6758
weighted avg	0.67	0.67	0.66	6758

Final Model Performance with 0.38 threshold

```
Average Best Threshold: 0.382
Average Best F1 Score: 0.7201115921857997
----------------------------------------------------------
Final Model Performance with Threshold of 0.382
              precision    recall  f1-score   support

   No Injury       0.81      0.40      0.53     16893
      Injury       0.60      0.91      0.72     16893

    accuracy                           0.65     33786
   macro avg       0.70      0.65      0.63     33786
weighted avg       0.70      0.65      0.63     33786
```

Further refinement of the model was achieved by adjusting the classification threshold, finding an optimal threshold of 0.38 through threshold moving. A threshold of 0.38 implies that the model will make a positive prediction when the predicted probability of an injury crash is at least 0.38.

MODELING

MODEL EVALUATION

In a thorough evaluation, we considered the critical cost implications of false positives and negatives.
Given the paramount goal of optimizing resources and accurately identifying emergencies, both types of errors were deemed equally costly. Consequently, the F1 score emerged as the primary evaluation metric, supplemented by the examination of micro precision and recall for both classes. The optimization of the classification threshold was strategically centered around maximizing the F1 score.

Final Model Performance with 0.38 threshold

```
Average Best Threshold: 0.382
Average Best F1 Score: 0.7201115921857997
------------------------------------------------------------
Final Model Performance with Threshold of 0.382
              precision    recall  f1-score   support

   No Injury       0.81      0.40      0.53     16893
      Injury       0.60      0.91      0.72     16893

    accuracy                           0.65     33786
   macro avg       0.70      0.65      0.63     33786
weighted avg       0.70      0.65      0.63     33786
```

• The final model demonstrated notable performance metrics, boasting an average F1 score of 72%, a **recall** of 91%, and a **precision** of 63%. Particularly noteworthy is the micro precision for non-emergency predictions, which stood at 81%. This signifies that the model adeptly identifies 91% of high-emergency crashes, and of all crashes predicted as high-emergency, 63% are indeed critical. Although the model didn't capture a large percentage of non-emergent cases (39% recall), its accuracy reaches 81% when predicting a crash as not high emergency.

• Striving for optimal performance with sufficient model complexity, we leveraged XGBoost's feature importance. Through a systematic reduction of features, it was determined that the top 20 features proved as effective as utilizing the entire feature set. Further reduction led to a decline in performance, solidifying the choice of 20 features for the most parsimonious model.

FURTHER EXTENSIONS & CONSIDERATIONS

MODELING METHOLOGY

Neural Networks

Neural Networks algorithm was not used due to its complexity and interpretability challenges. Scaling was not utilized as tree-based models are not sensitive to the scale of features.

Principal Component Analysis

Principal Component Analysis (PCA) was not applied to preserve interpretability and because tree-based models can handle high dimensionality effectively.

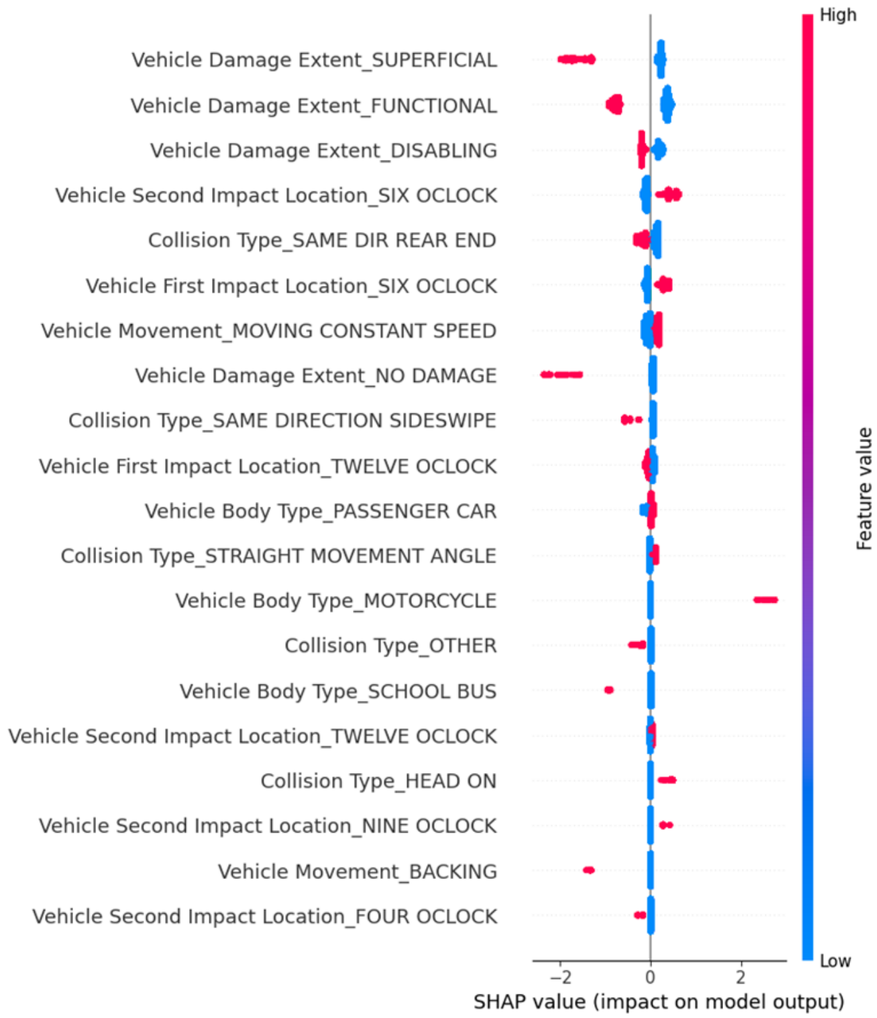
Model Interpretability

While ensemble algorithms are often viewed as 'black box' due to their complexity, extensive exploratory data analysis (EDA) provided valuable insights for interventions for the use case. Feature importance gives insight into influential factors for crash injuries.

Figure 1: xxxxxxxxxxxxxx

SHAP BEESWAR

PLOT & INTEPRETATION

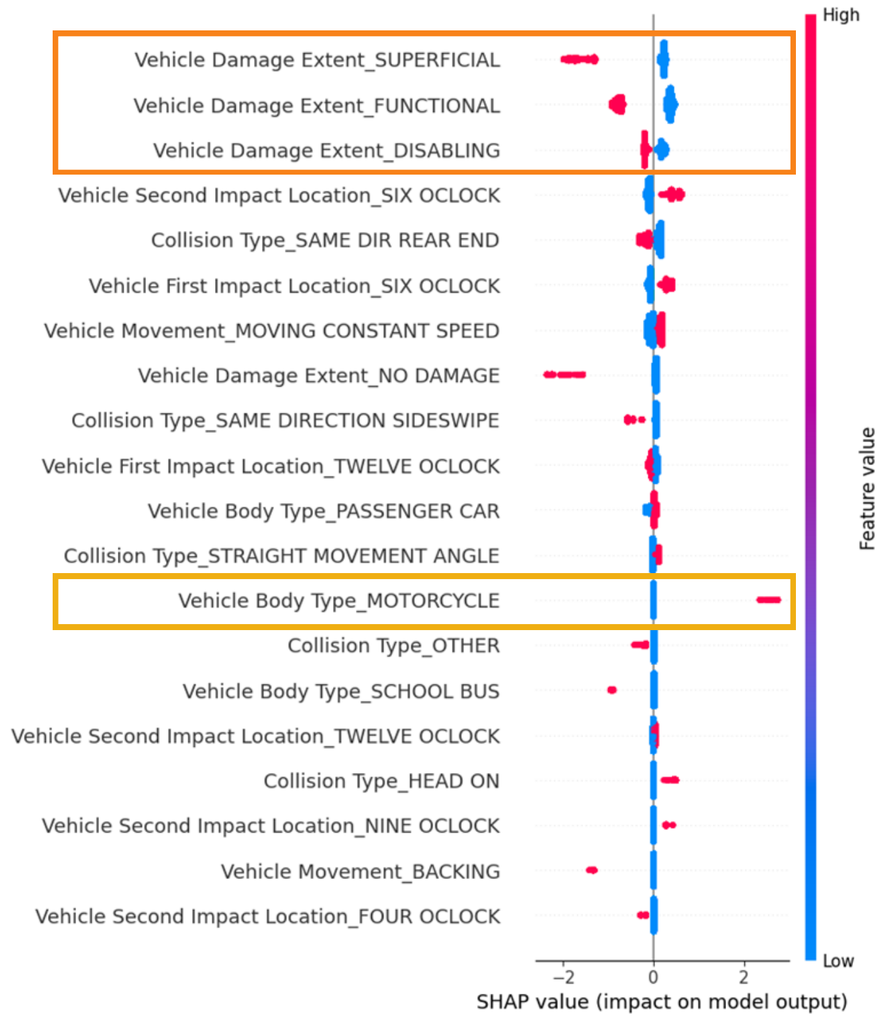


SHAP (SHapley Additive exPlanations) was employed to comprehend how each feature influences the probability of crash injuries and the specifics of individual prediction outcomes.

- Input variables ranked from **top** to **bottom** by their mean absolute SHAP values for the entire dataset.
- **Most important variables** are listed from top to bottom. The most important variable in predicting injury when a car accident occurs is ‘Vehicle Damage Extent_SUPERFICIAL’ followed by ‘Vehicle Damage Extent_FUNCTIONAL’ and so on.
- Feature selection from **XGB** coupled with **SHAP** allows emergency operators to ask better questions to determine injuries **after** a car accident occurs.

SHAP BEESWAR

PLOT & INTEPRETATION



- For each variable, every instance of the dataset appears as a point. Points distributed horizontally **along x-axis according to SHAP value**. Places with high density of SHAP values are stacked vertically.

- More **negative SHAP** values indicate a lower probability of an accident classified as “Injury”. More **positive SHAP** value indicate higher probability of an accident classified as “Injury”.

- Color bar corresponds to the raw values **(different from SHAP values)** of the variables for each instance.

- Most of the variables here are **dummy variables**. Red points indicate instances where the dummy variable is positive **(value of 1)** and blue points indicate instances where dummy variable is negative **(value of 0)**.

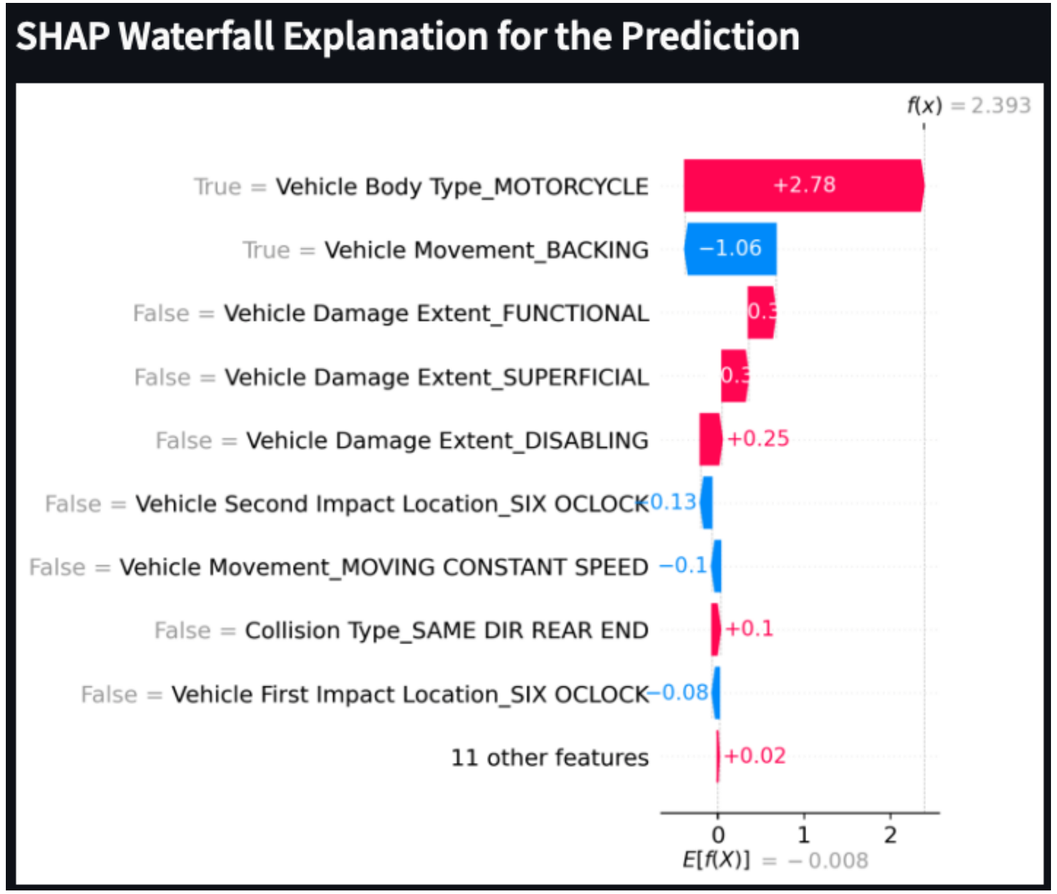
- Distribution of variables **across the x-axis could suggest some policy intervention** upstream such as stricter training for motorcyclists.

SHAP BEESWAR

PLOT & INTEPRETATION

SCAN ME

Link for Streamlit Demo: https://predcrashseverity.streamlit.app



- f(x) is the result from the summation of the factors from E[f(X)]
- f(x) measures the log-odds.
- More positive number indicates higher probability of the class ‘injury’.

POLICY RECOMMENDATIONS

STRATEGIC & DATA-DRIVEN PROPOSALS FOR ENHANCING ROAD SAFETY

1

Refining Standard Operating Procedures for Emergency Operators

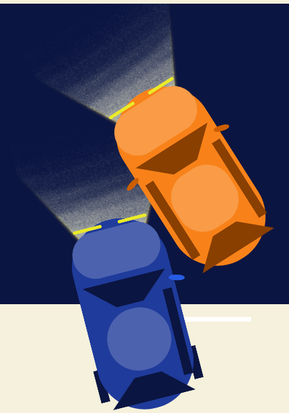
Feature importance could inform Operators on the questions to ask. SHAP Waterfall allows Operators to know the specific variable contributing to the prediction. Allowing for man-in-the-loop interventions.

Maryland can take a further step in identifying if the motorcyclists involved in the accidents are from within the state to determine relevant policies to intervene such as increasing/ introducing more defensive riding training for motorcyclists or passing laws that protect motorcyclists (e.g. refusing to legalise lane splitting for motorcyclists in Maryland)

Mitigating High Injury Risks in Motorcyclist-involved Incidents

2

In the long run, Maryland can build onto the clustering approach to identify if the current A&E wards are optimally located to attend car accidents. Our model was unable to show any significance of the clusters but more efforts can be invested to improve the clustering model should Maryland state aims to reduce fatalities due to car accidents in the long run.



REFERENCES

- Bhuiyan, H., Ara, J., Hasib, K.M., et al. (2022). **Crash severity analysis and risk factors identification based on an alternate data source: a case study of developing country**. *Scientific Reports*, 12, 21243. https://doi.org/10.1038/s41598-022-25361-5

- Wang, C., Chen, F., Yu, B., & Cheng, J. (2022). **Injury severity assessment of rear-end crashes via approaches based on generalized estimating equations**. *Canadian Journal of Civil Engineering*. Advance online publication. https://doi.org/10.1139/cjce-2022-0197

- Safari, M., Alizadeh, S.S., Sadeghi-Bazargani, H., Aliashrafi, A., Maleki Ghahfarokhi, A., Moshashaei, P., & Shakerkhatibi, M. (2019). **A Comprehensive Review on Risk Factors Affecting the Crash Severity**. *Journal Article*. Pages 1366-1376.

- Al-Mistarehi, B. W., Alomari, A. H., Imam, R., & Mashaqba, M. (2022). **Using Machine Learning Models to Forecast Severity Level of Traffic Crashes by R Studio and ArcGIS**. *Frontiers in Built Environment*, 8, Article 860805. https://doi.org/10.3389/fbuil.2022.860805

Link to the dataset used for this project:

Crash Reporting - Drivers Data

APPENDIX

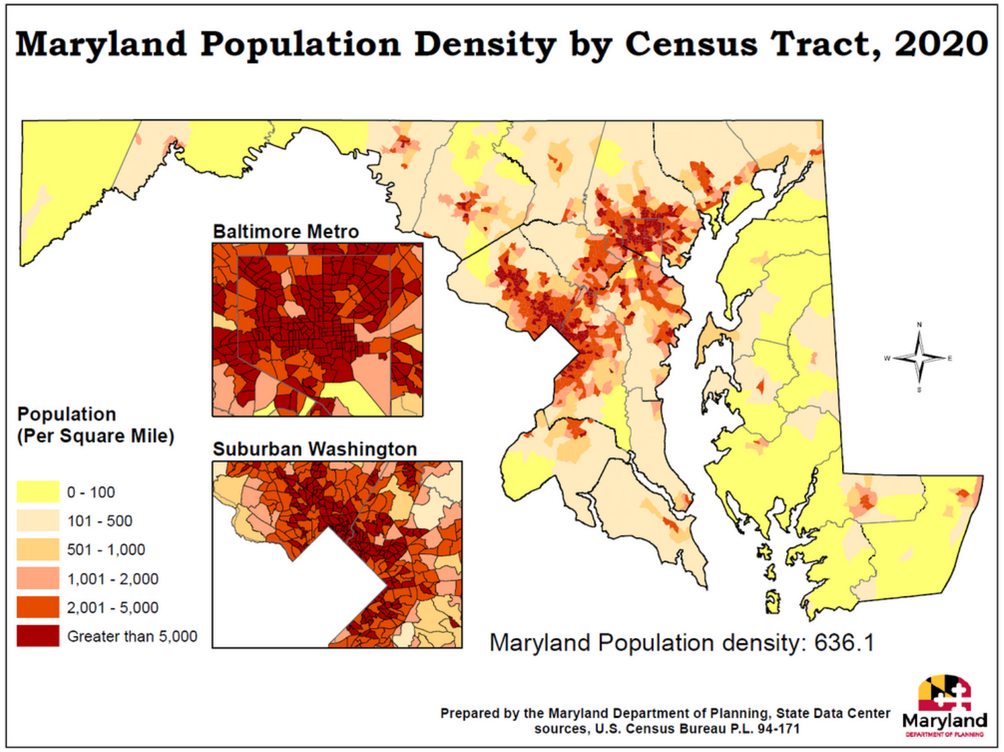


Figure A. **Maryland Census Tract Population Density Chart**

APPENDIX

Considered Models and Their Application for Car Crash Prediction

In our analysis, we will leverage various methods introduced in the course to assess their performance in achieving a balanced outcome between accuracy and interpretability. Below are details on three considered models:

- **Logistic Regression:**
  - Applicability: Ideal for binary classification tasks, making it suitable for predicting car crashes.
  - Strengths: Provides clear insights into the likelihood of a crash, facilitating a comprehensive understanding.
  - Use Case: Particularly useful when seeking a model for explanatory purposes, helping to identify influential factors contributing to accidents.

- **KNN (K-Nearest Neighbors):**
  - Applicability: Effective for identifying similar patterns or clusters of accidents within the dataset.
  - Strengths: Valuable for assessing the risk of a location or road segment based on historical accidents in nearby areas.
  - Use Case: Particularly beneficial when aiming to understand the localized patterns and risks associated with specific areas.

- **Gradient Boosting:**
  - Applicability: Capable of handling complex, nonlinear relationships in the data.
  - Strengths: Particularly beneficial when precision is crucial, as it can capture intricate interactions between various features.
  - Use Case: Useful for predicting car crashes with high accuracy, especially when considering nuanced relationships between factors such as road conditions, driver behavior, and weather.