PREDICTING ACCIDENT SEVERITY IN MARYLAND

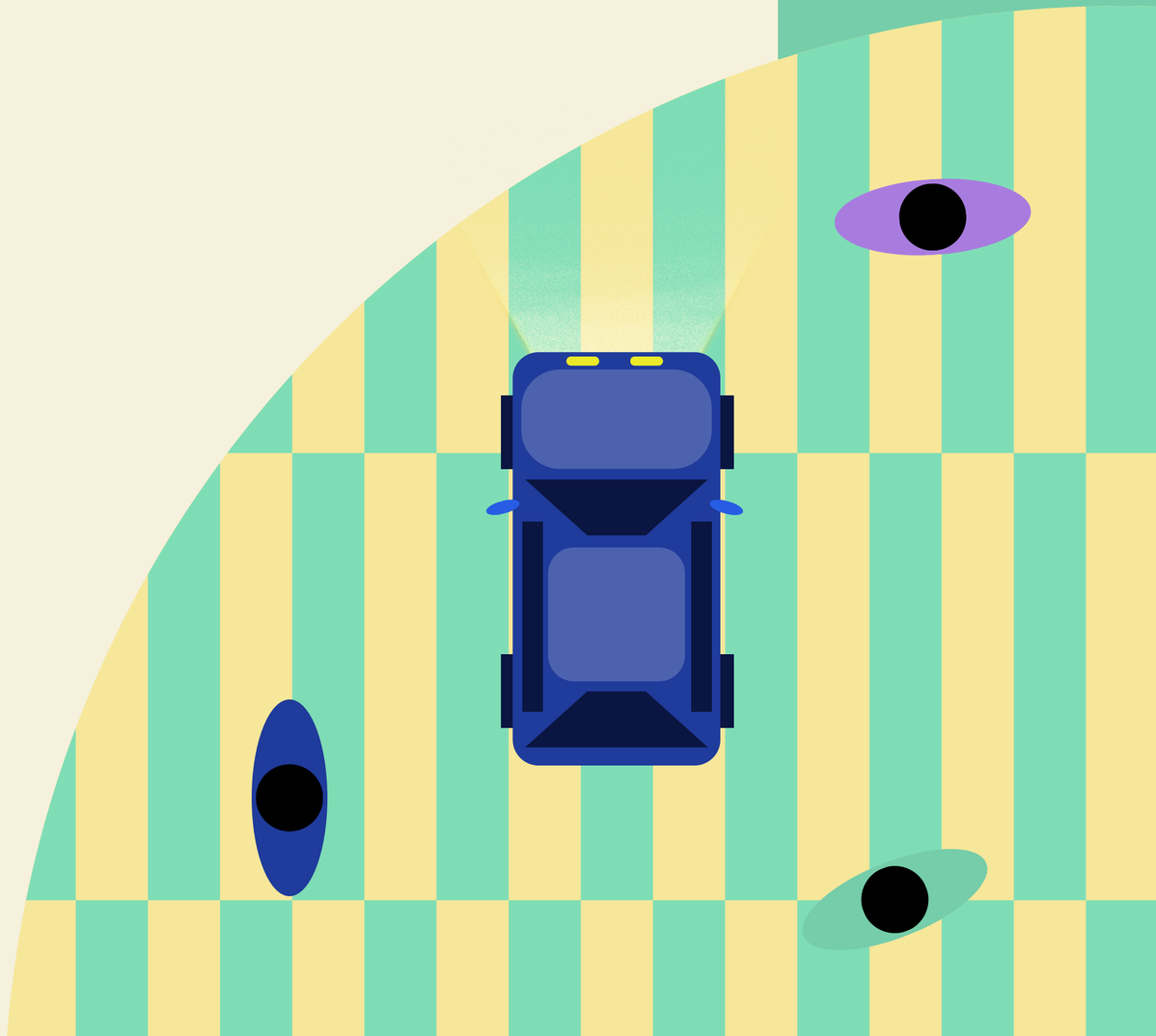
INSY 662 | DATA MINING & VISUALIZATION

Final Presentation | Group Project

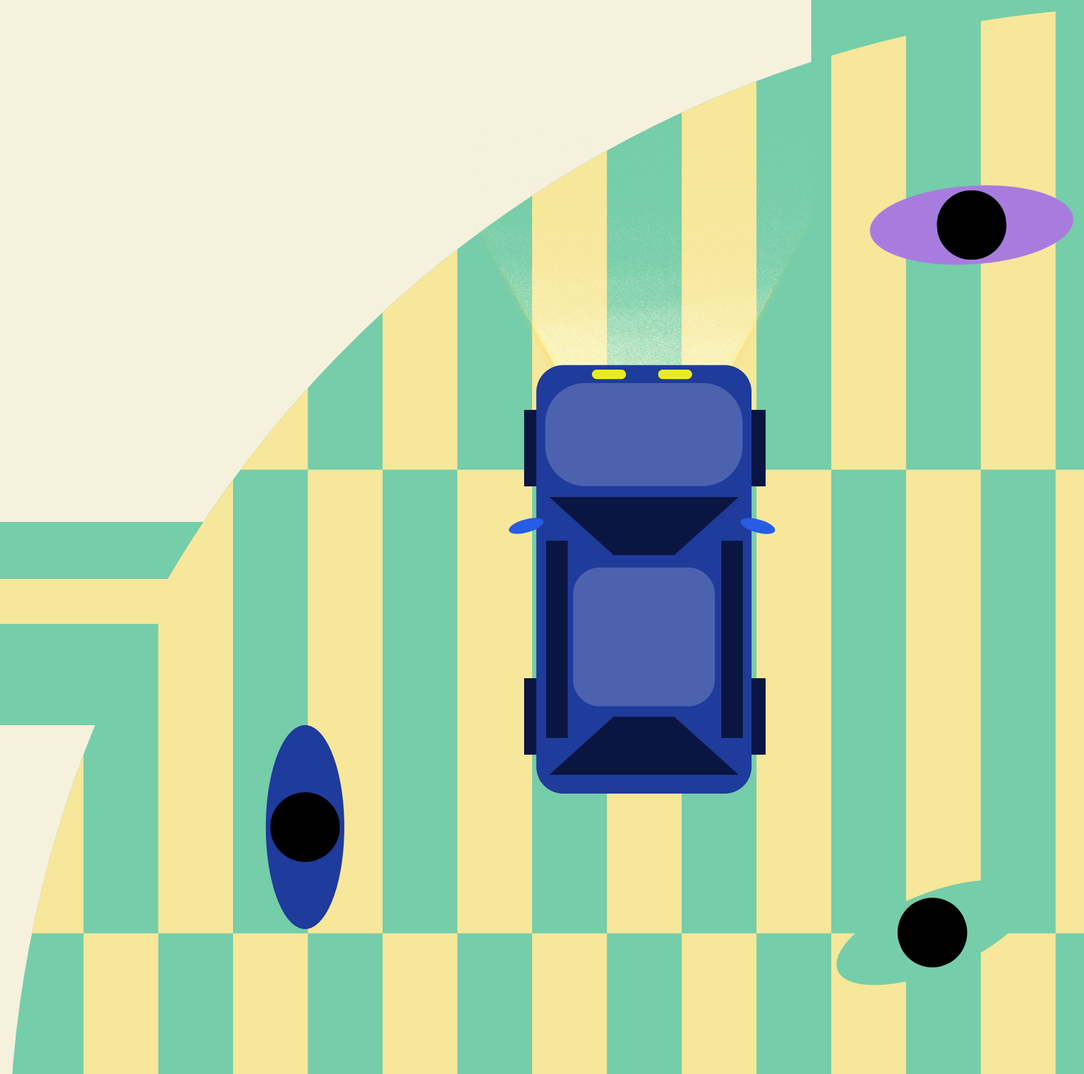
Meriem | Vincent | Abdul | Xingchen | Chien

AGENDA

1. EXECUTIVE SUMMARY
2. PROBLEM STATEMENT
3. DATA SOURCES & PRE-PROCESSING
4. EXPLORATORY DATA ANALYSIS
5. PREDICTIVE MODELING & RESULTS
6. FURTHER MODELING EXTENSIONS
7. POLICY RECOMMENDATIONS



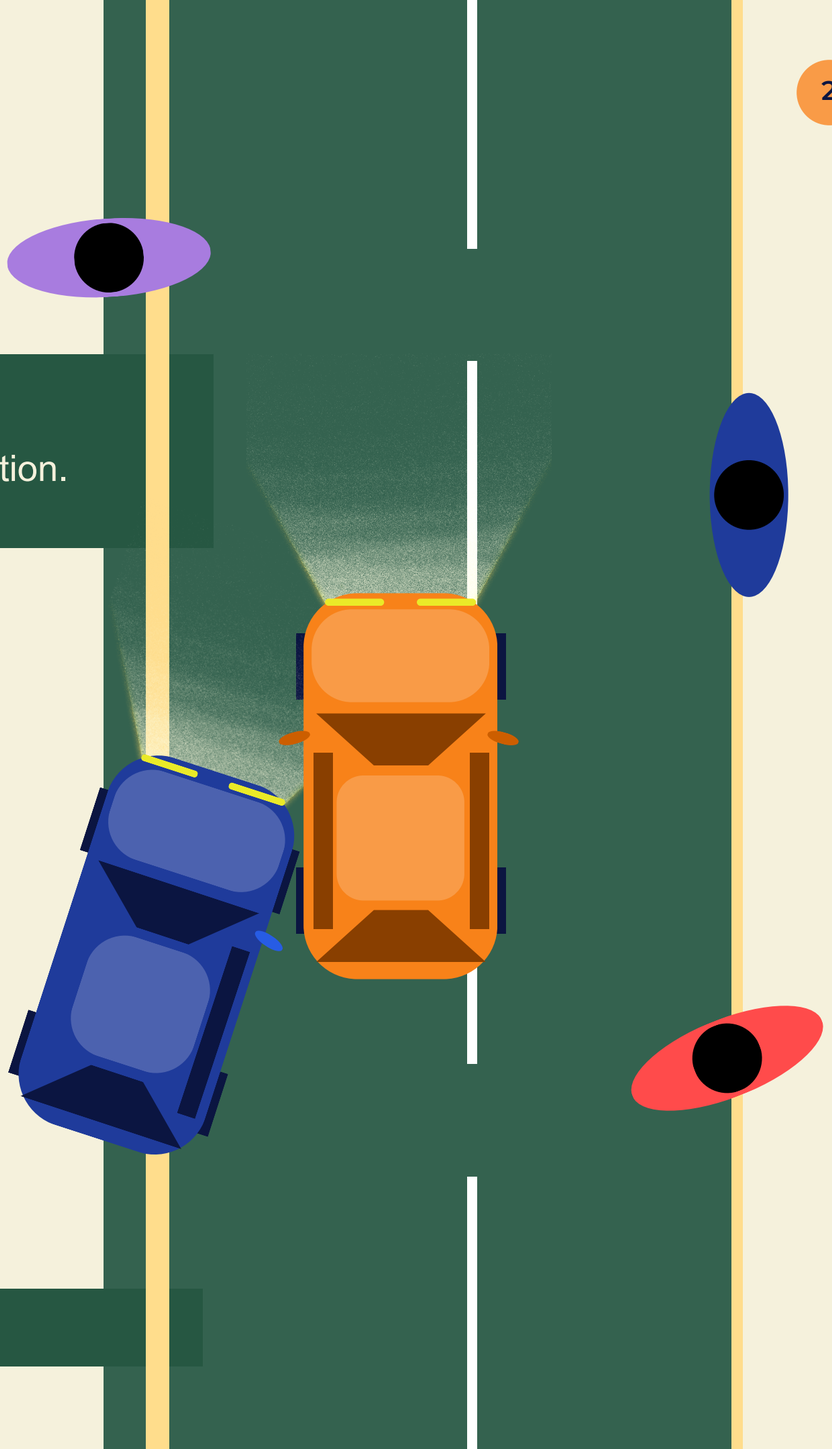
EXECUTIVE SUMMARY



EXECUTIVE SUMMARY

DRIVING TOWARD SAFER ROADS: OUR PROJECT GOAL

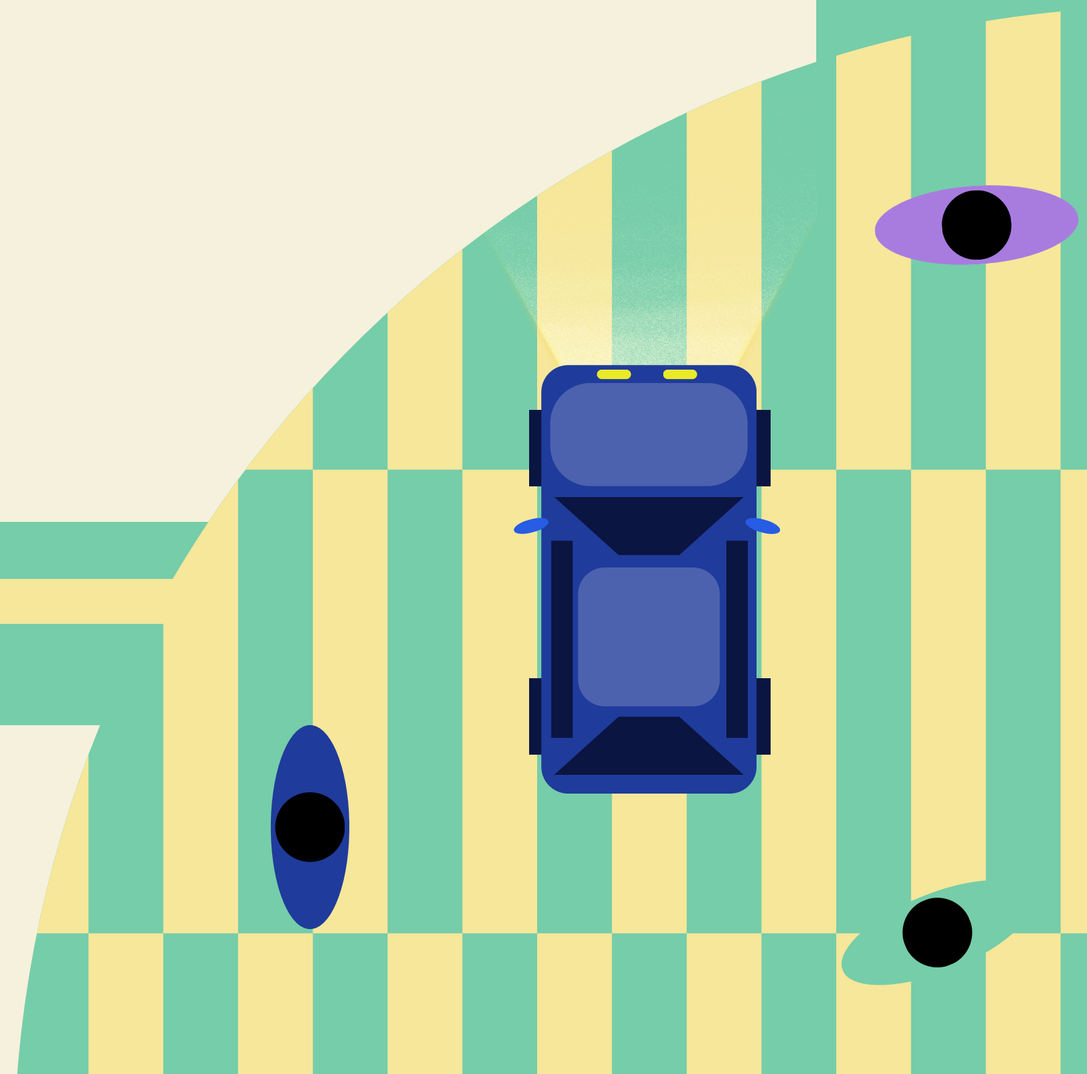
- Develop a predictive model for car crash severity (injury) in Maryland.
- Foundation for policy recommendations aimed at enhancing road safety & effective traffic regulation.



- Consolidating diverse datasets for relevance through city-specific road features, examining factors like traffic, demographics, and weather.
- **Core focus:** Expedite accident responses and minimize impact through resource optimization and data-driven policy recommendations.
- **Commitment to clear policies:** Employing a transparent model for informed decision-making, with future exploration of city grouping for tailored recommendations.

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

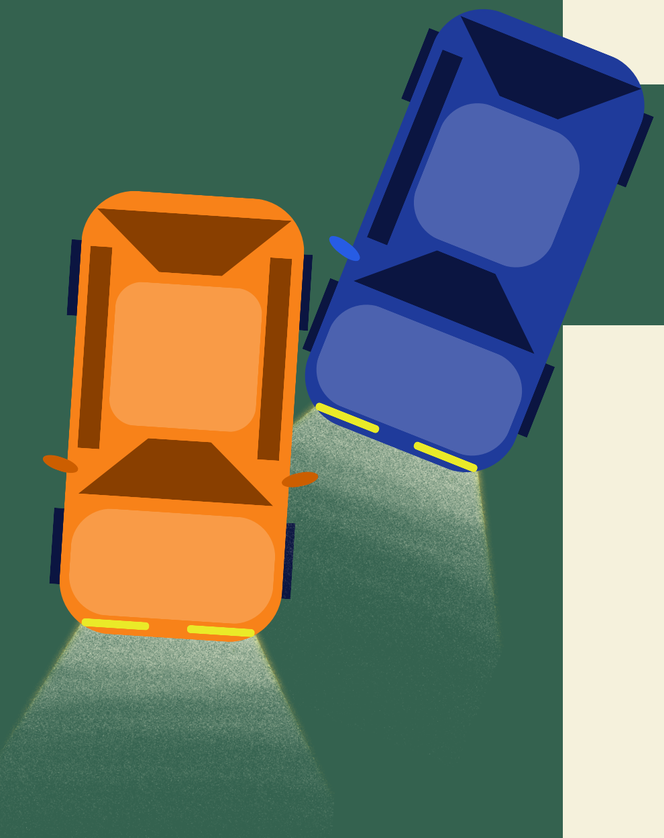
PROBLEM STATEMENT



PROBLEM STATEMENT

ENHANCING MARYLAND ROAD SAFETY VIA MACHINE LEARNING

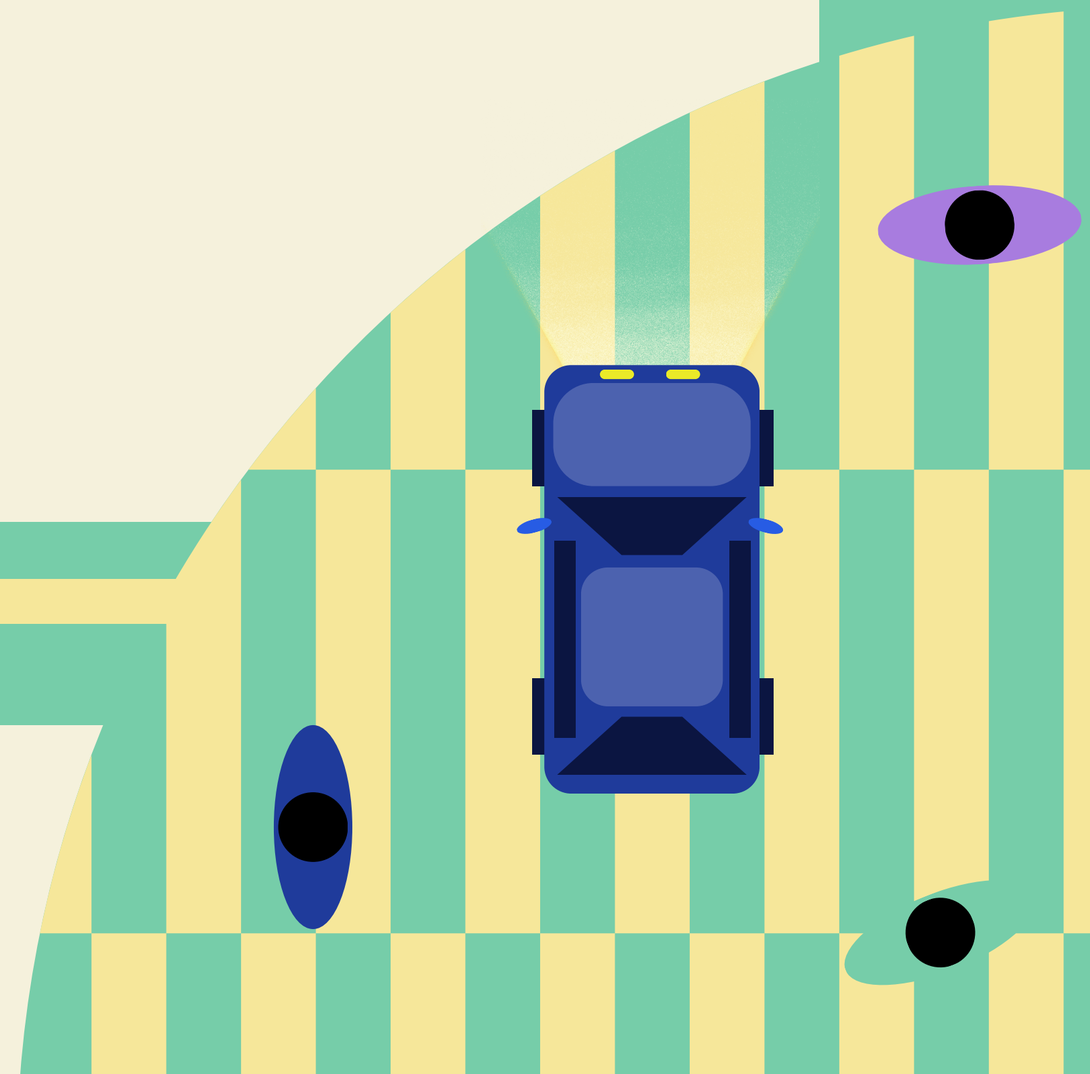
- **Leverage machine learning for accurate car incident predictions in Maryland.**
- **Contribute to road safety by empowering authorities with proactive measures.**
- **Integrate diverse data sources for enhanced predictive capabilities.**
  - *roads features, weather, driving styles, etc.*



ANTICIPATED MODEL BENEFITS & OUTCOMES

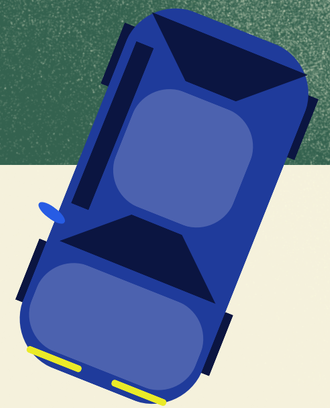
- **IMPROVED ROAD SAFETY**
  - Accurate predictions empower authorities, preventing accidents and reducing road fatalities.
  - Contributes to minimizing incidents in high-risk areas.
- **OPTIMIZED RESOURCE ALLOCATION**
  - Model insights guide strategic resource allocation for efficient and targeted deployment to high-risk areas.
- **INFORMED POLICY RECOMMENDATIONS**
  - Model insights inform policy development for enhanced road safety in Maryland.
  - Recommendations may include stricter regulations in adverse weather or high-risk areas.

DATA SOURCES & PRE-PROCESSING



DATASET SOURCES

RATIONALE BEHIND SELECTING SPECIFIC DATA SOURCES



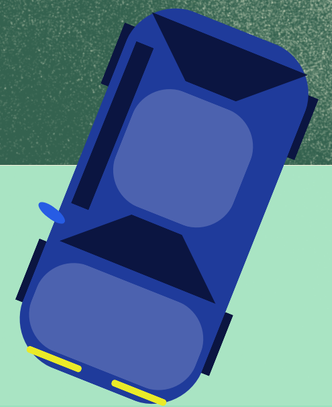
Data Acquisition Strategy:

- *Acquire diverse and certified data for robust predictive models.*
- *Commitment to credible sources ensures accurate 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*

DATASET SOURCES

RATIONALE BEHIND SELECTING SPECIFIC DATA SOURCES



Data Acquisition Strategy:

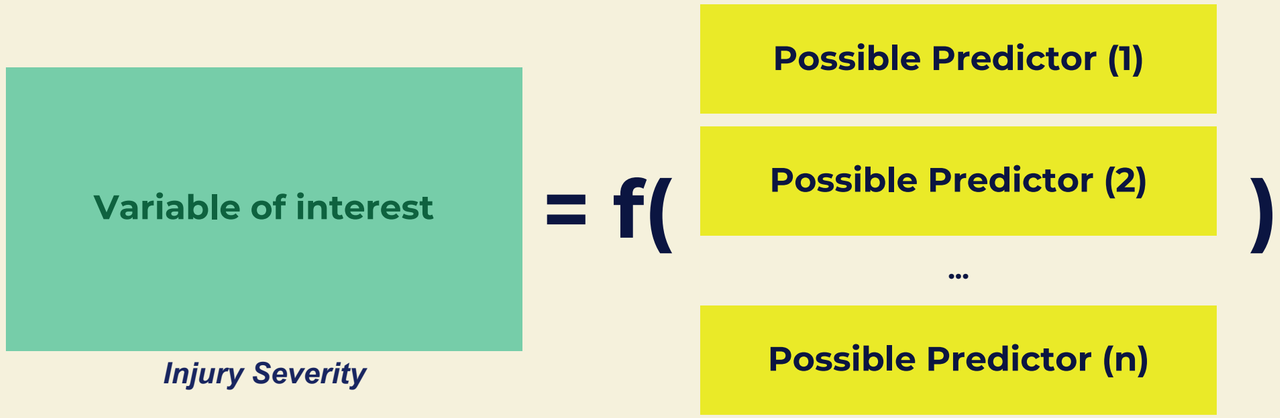
- *Acquire diverse and certified data for robust predictive models.*
- *Commitment to credible sources ensures accurate predictions and impactful policy recommendations.*

		Selected Dataset	
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*

DATASET DESCRIPTION

EXPLORING OUR DATASET: UNRAVELING DIMENSIONS AND RELATIONSHIPS

- **Target Variable: Injury Severity**, a categorical variables.
- Model uses predetermined predictors to unveil variables influencing severity, providing nuanced insights.
- Contributes to a comprehensive understanding of the multifaceted impact of accidents in the city of Maryland.



Irrelevant Variables → **EXCLUDED**

The analysis hinges on the fundamental relationship between the target variable (Y) and predictor variables (Xi), with Y being the focal point of investigation and Xi contributing to shaping this relationship. Irrelevant variables will be excluded from our analysis.

Variable Insights and Assumptions

- Predominance of categorical variables.
  - Correlation between Weather & Surface Conditions
  - Driver Factors: *Substance abuse, etc.*
  - Speed Limit Relationship with Severity

DATA PRE-PROCESSING

STEP 1

Data Cleaning and Handling Missing Values: Initiating the Pre-processing

Original Dataset

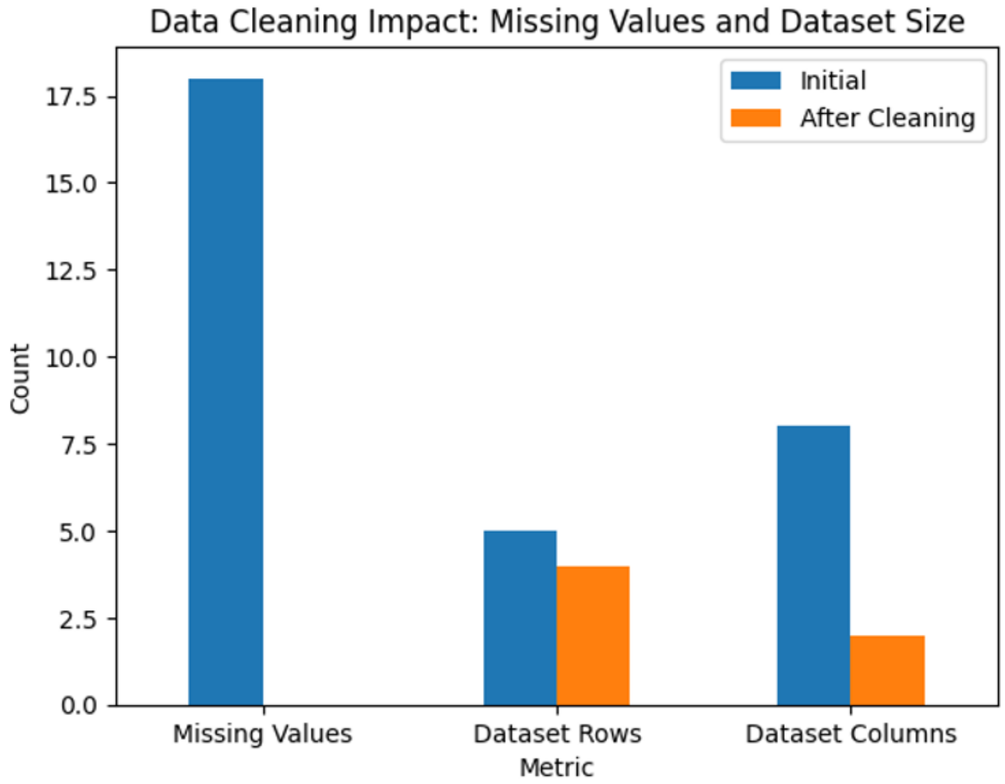
43 variables (columns)	160,000 records (rows)

1. **Drop the significant missing values variables:**
   'Off-Road Description', 'Municipality', 'Related Non-Motorist', 'Non-Motorist Substance Abuse', , '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***



DATA PRE-PROCESSING

STEP 2

Categorical Encoding and Temporal Insight Enhancement: Elevating the Data Structure

- Strategically encoded categorical variables based on unique category counts per column.
- Elevated data structure by transforming 'Crash Date/Time' into datetime format and extracting key temporal components.
- Enhancements in data preparation 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'.**

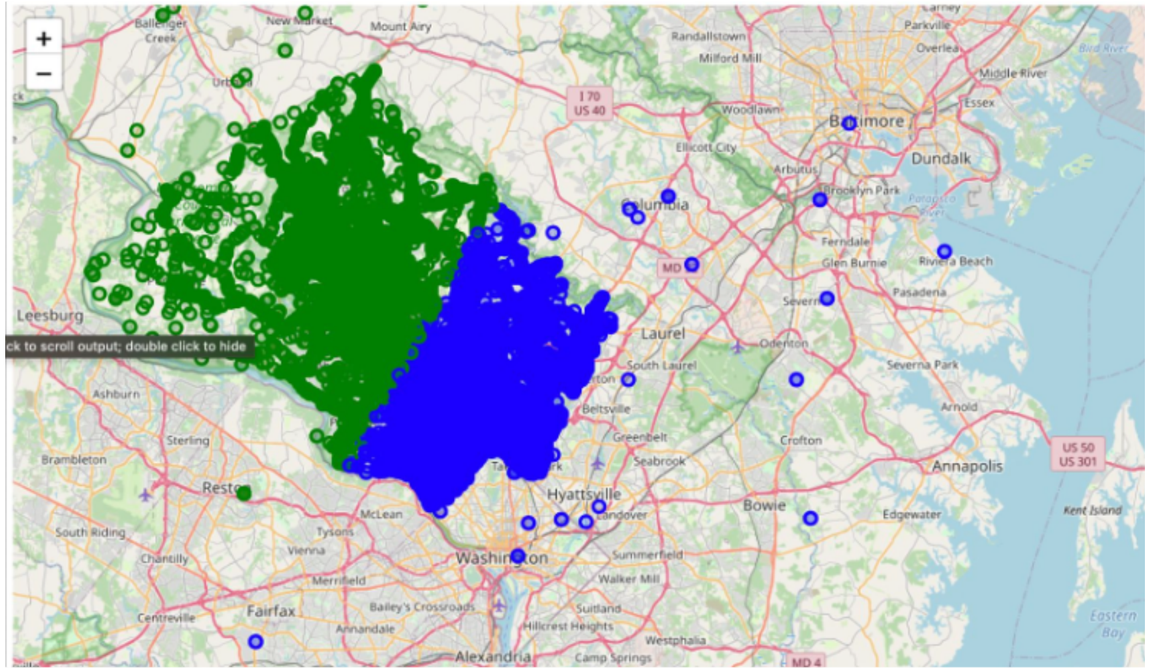
	Crash Date/Time	Hour	Day	Month	Year	Some Other Column	Date Day	Date Year	Season	Time of Day
0	2023-01-01 15:30:00	15	1	1	2023	1	1	2023	Winter	12pm-18pm
1	2023-02-15 08:45:00	8	15	2	2023	2	15	2023	Winter	6am-12pm
2	2023-03-20 22:10:00	22	20	3	2023	3	20	2023	Spring	18pm-24am

DATA PRE-PROCESSING

STEP 3

Geospatial Insight Unveil: Preparing for Analysis

- Unlocked geospatial patterns using 'Longitude' and 'Latitude' with K-Means clustering.
- Refined the final dataset for streamlined analysis, excluding extraneous columns.
- Prepared for robust geospatial pattern exploration in crash occurrences.



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

- Strategically applied K-Means clustering for geospatial analysis during data scaling.
- K-Means tailored for effective uncovering of spatial patterns and clusters.

Final DataFrame for Analysis

Excluded 'Latitude', 'Longitude', and additional non-relevant columns.

DATA PRE-PROCESSING

STEP 4

Streamlining Injury Severity for Enhanced Model Interpretation

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.

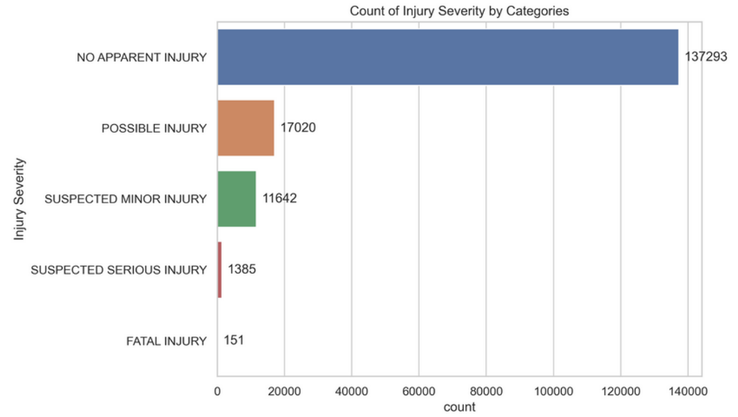
Reasons for Transformation

- 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'.

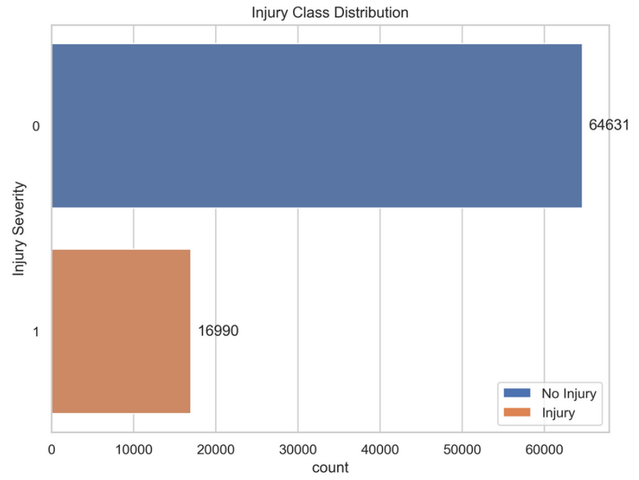
Post Processing

- Examined post-encoding distribution: 79% 'no injury' (0) and 21% 'injury' instances (1).
- Foundation for injury-focused predictive modeling, revealing prevalence insights in the dataset.

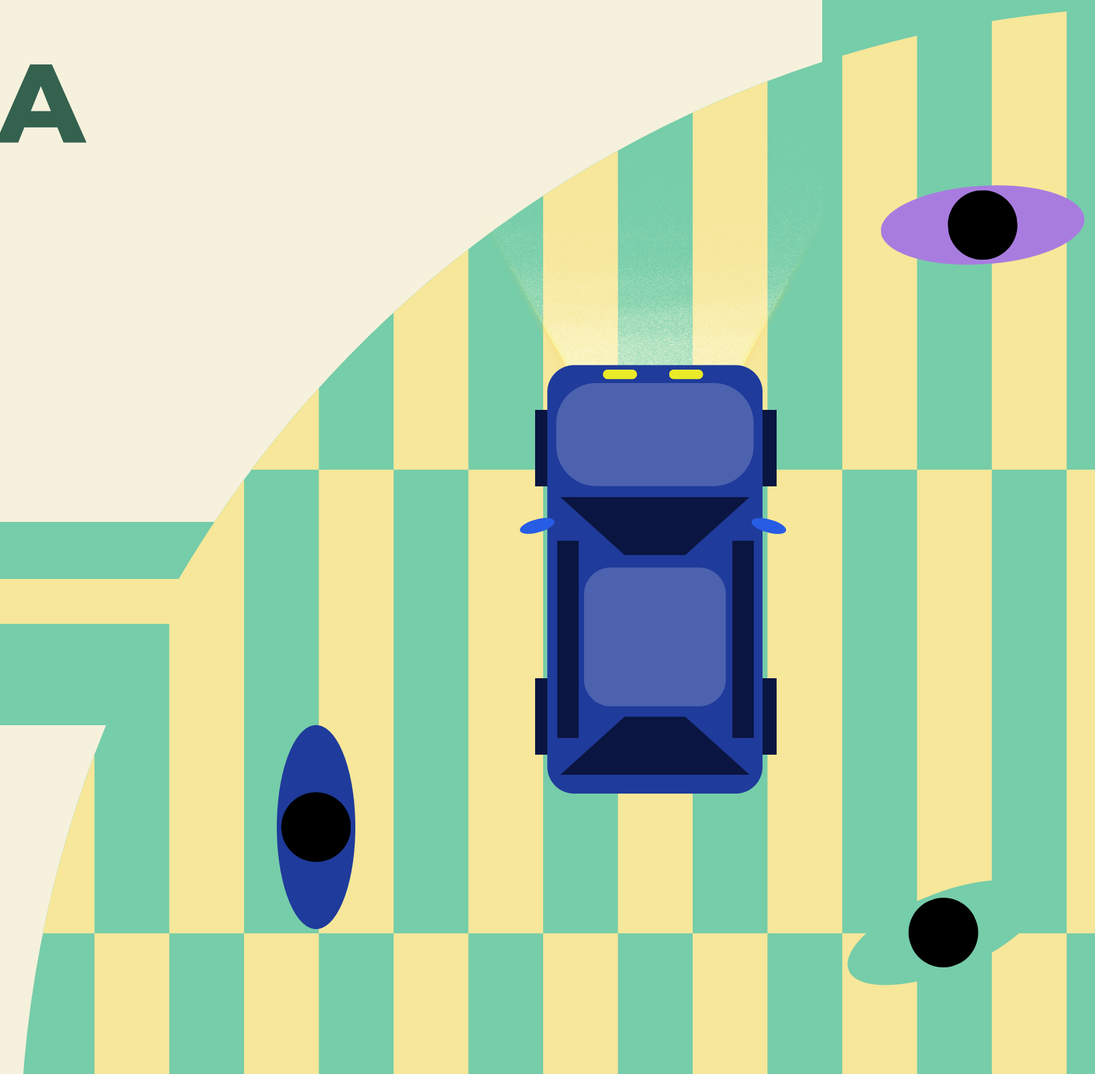
Pre-transformation



Post-transformation



EXPLORATORY DATA ANALYSIS (EDA)

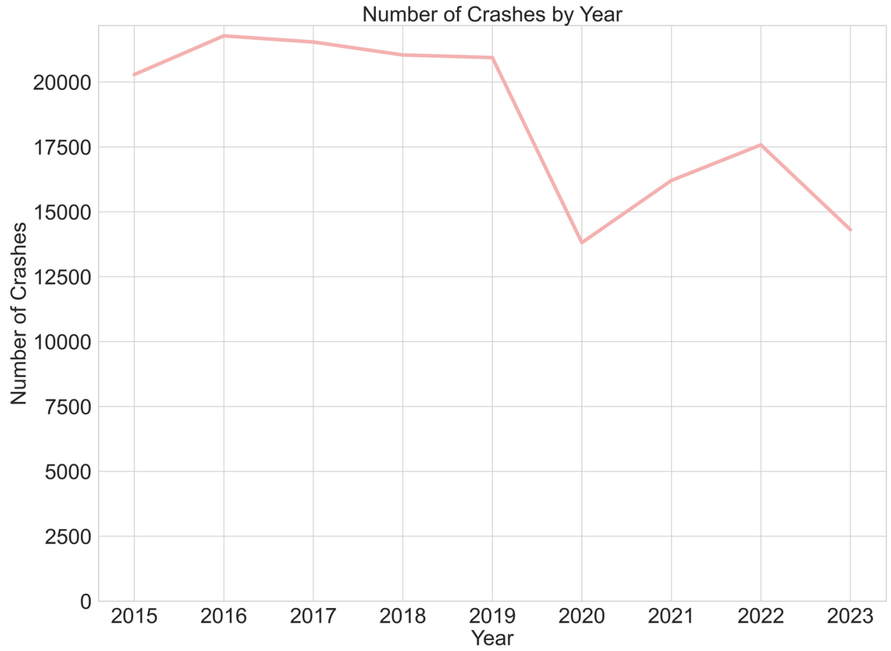


EXPLORATORY DATA ANALYSIS

UNCOVERING MEANINGFUL INSIGHTS

NUMBER OF CRASHERS PER YEAR

Crash Trends Over the Years



- Trend line shows **peak crash numbers in 2016**, indicating heightened accident rates.
- Slight **decline** until early 2019, possibly due to safety measures.
- Fluctuations in subsequent years, with a dip in 2020 linked - ***pandemic***.
- **Recovery in 2021 and 2022** suggests a gradual return to office + increased road traffic.

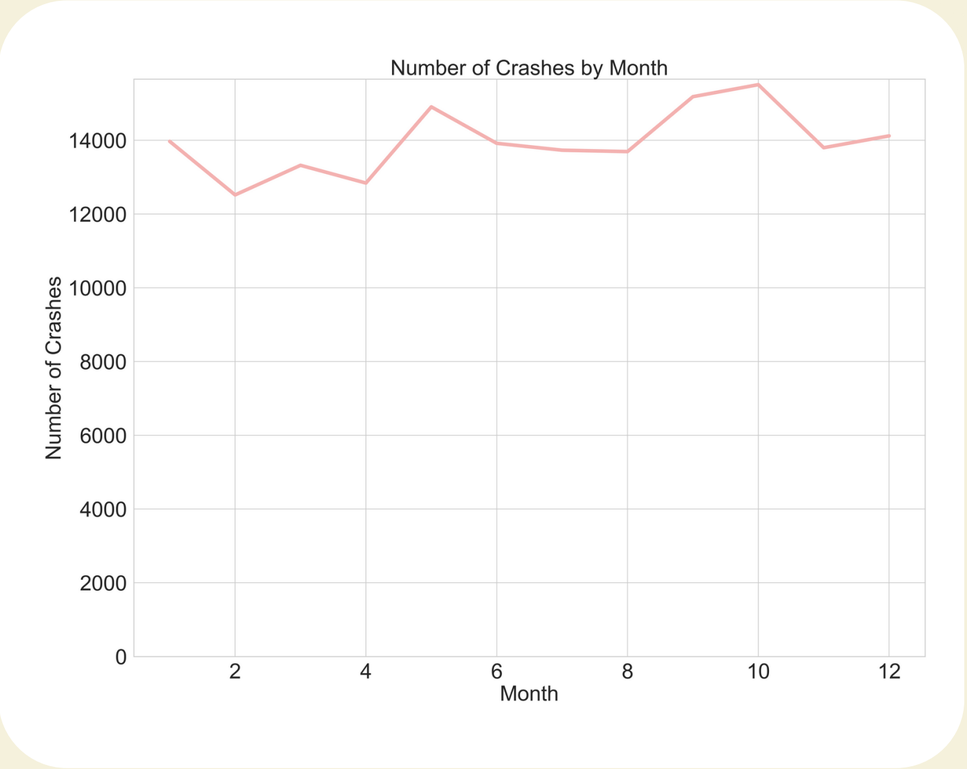
EXPLORATORY DATA ANALYSIS

UNCOVERING MEANINGFUL INSIGHTS

NUMBER OF CRASHERS PER MONTH

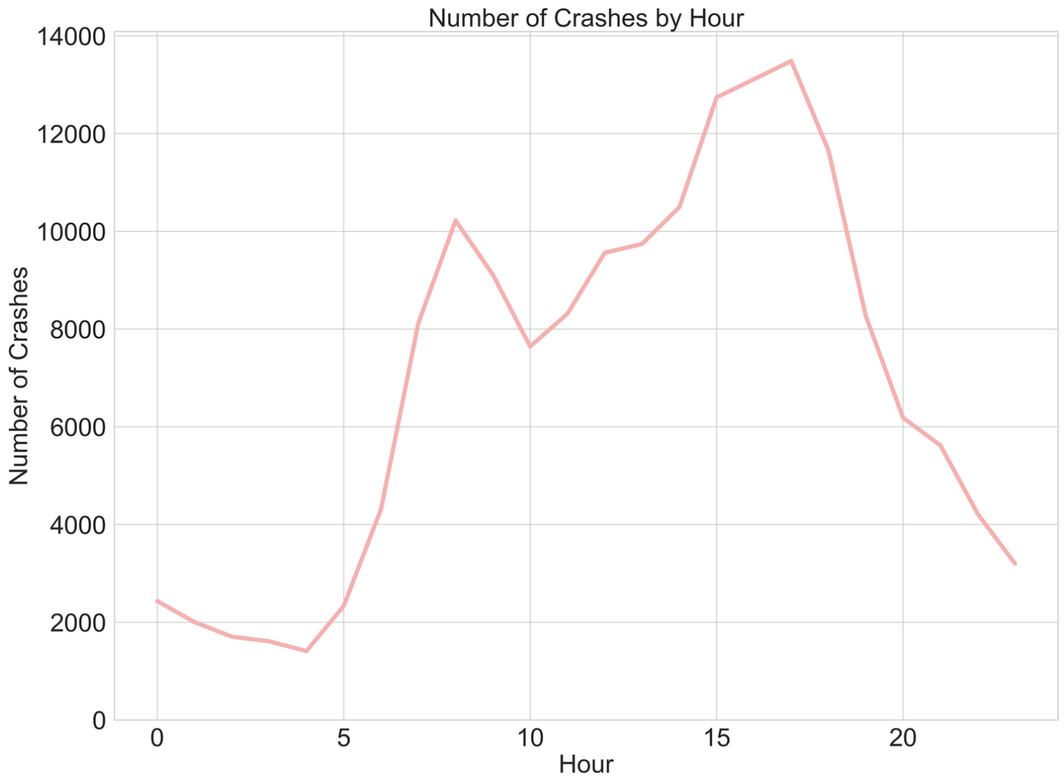
Crash Trends By Month

- **Generally stable trend** with subtle monthly fluctuations in crash numbers.
- Peaks may align with **seasonal factors** affecting driving conditions or travel patterns.
- **Mild seasonal pattern** suggests a correlation between crashes and seasonal variations.
- **October spike** possibly linked to tire change, where slippery summer tires may contribute to accidents if not replaced.



EXPLORATORY DATA ANALYSIS

UNCOVERING MEANINGFUL INSIGHTS



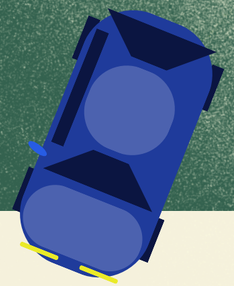
NUMBER OF CRASHERS PER HOUR

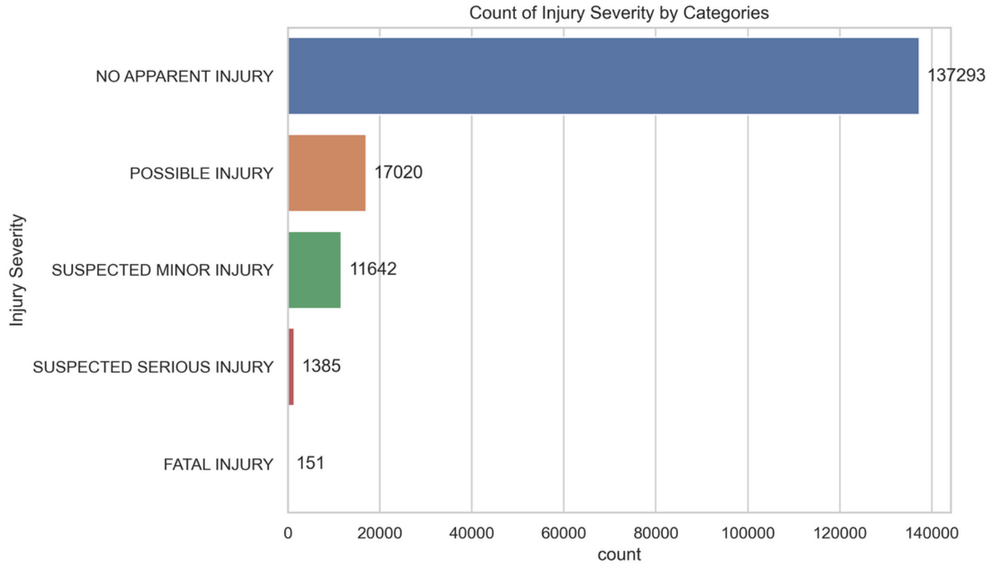
Crash Trends By Hour

- Notable surge in crashes during **early hours**, peaking in morning commute, indicating heightened risk.
- **Midday sees a decline**, potentially correlating with reduced traffic volume.
- **Late afternoon** to early evening witnesses another increase, aligning with **evening rush hours.**
- Afternoon heightened crash rate attributed to factors like **fatigue** and **hunger** during the commute.
- **Sharp decline after peak evening hours** reflects reduced traffic, with lowest occurrences in late-night to early-morning hours.

EXPLORATORY DATA ANALYSIS

UNCOVERING MEANINGFUL INSIGHTS





Exploratory Data Analysis: Navigating Insights Through Data Exploration

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.

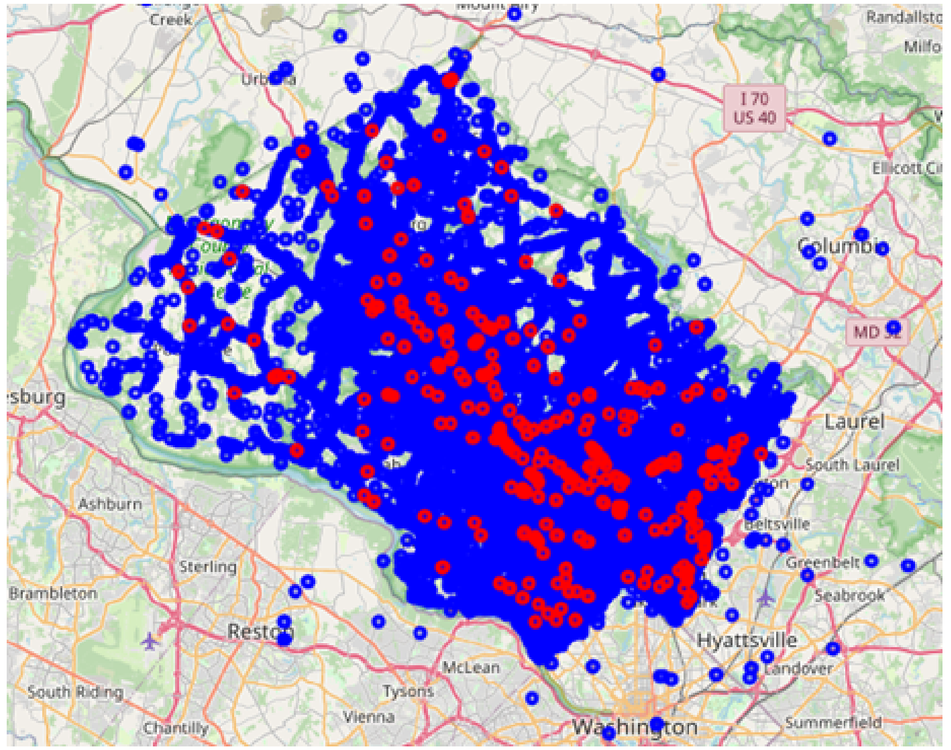
3. Instances of **suspected serious injuries** and **fatal injuries** are minimal in comparison to other categories, emphasizing the rarity of these tragic outcomes.

EXPLORATORY DATA ANALYSIS

UNCOVERING MEANINGFUL INSIGHTS

Geospatial Analysis for Traffic Safety: Plotting Crash Data

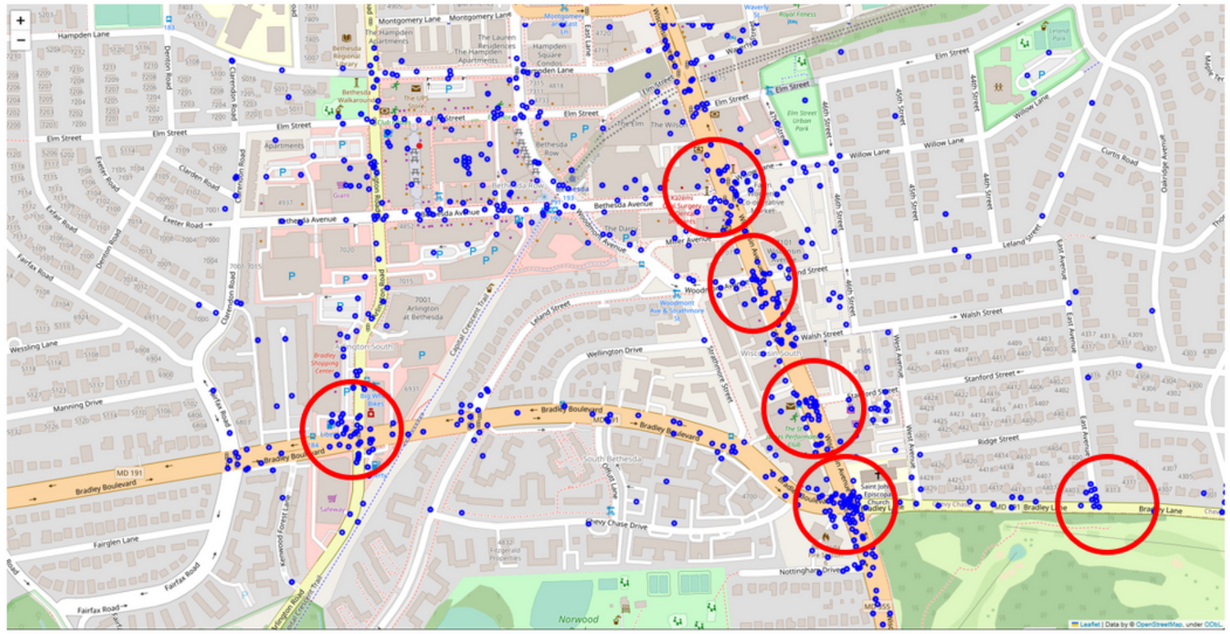
- Map shows concentrated **crash cluster** in a specific geographic area.
- Widespread red markings denote **fatal crashes**, emphasizing the need for targeted safety measures.
- Concentration near **major roads** and **intersections** indicates high traffic or hazardous conditions.
- Geospatial distribution **aids in prioritizing areas for safety improvements.**



EXPLORATORY DATA ANALYSIS

UNCOVERING MEANINGFUL INSIGHTS

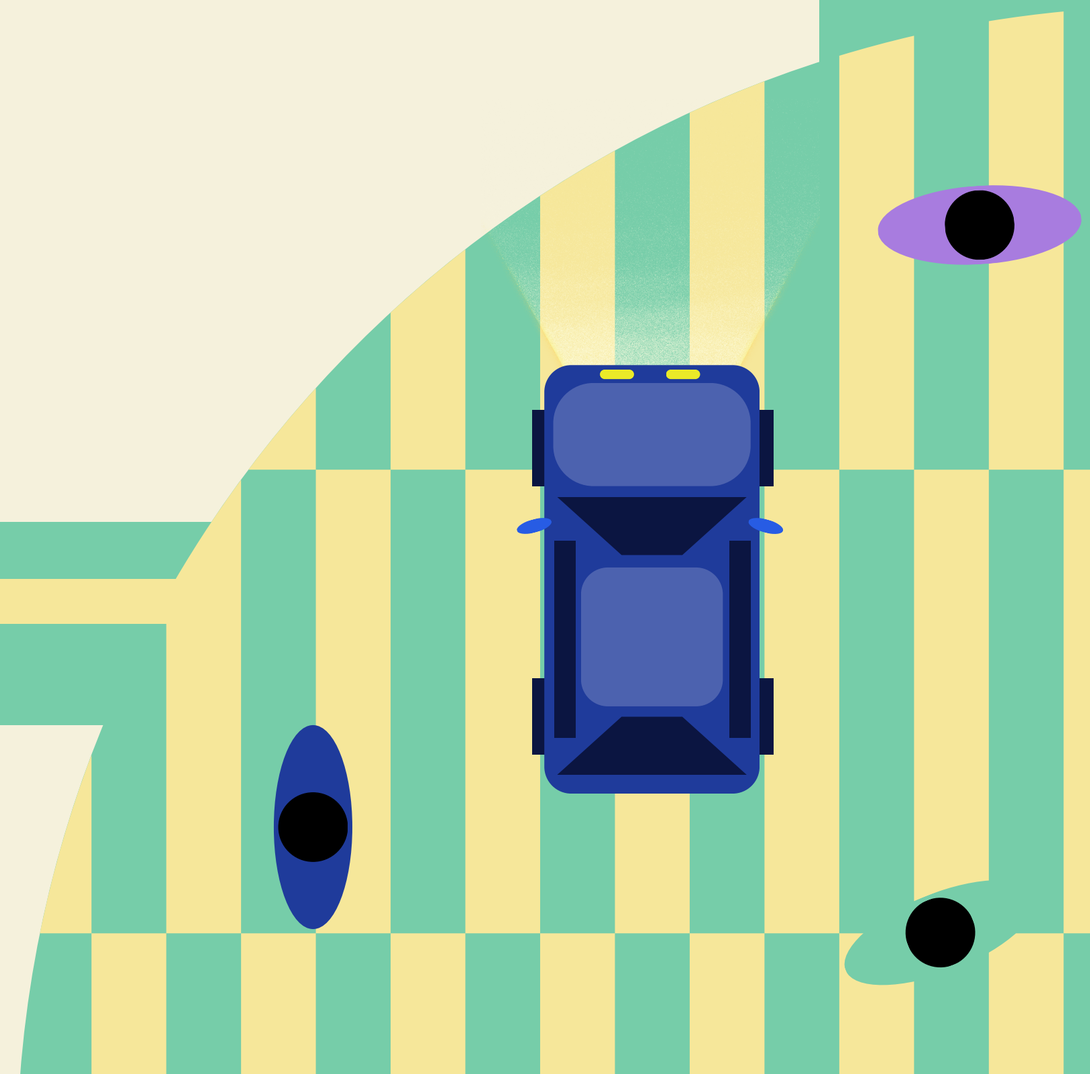
ZOOMED-IN VIEW



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.

PREDICTIVE MODELING



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 effective with high-dimensional data, benefiting the dataset with 166,537 observations and 43 features.
- XGBoost preferred over Gradient Boosting for regularization (L1 and L2) in the loss function, preventing overfitting in high-dimensional datasets.
- XGBoost chosen for faster performance.

MODELING

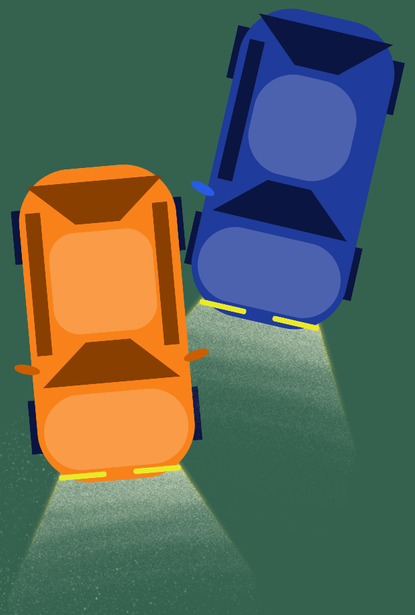
MODEL BUILDING PROCESSS

ITERATIVE PROCESS

- A significant class imbalance was noted, with the **'Injury'** class making up only 20% of the dataset - impacting 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.

Key Takeaways

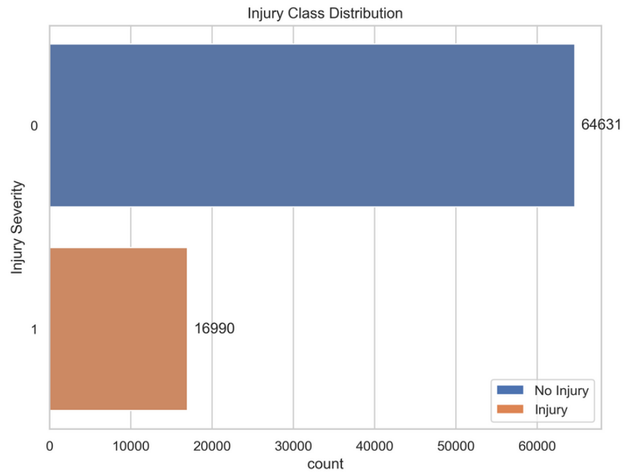
- Utilized ***scale_pos_weight*** in **XGBoost** to address class imbalance.
- Explored Randomized **Undersampling**, Randomized **Oversampling**, and **SMOTE** for class rebalancing.
- **XGBoost** performed best with undersampled data, avoiding overfitting seen with oversampling techniques.



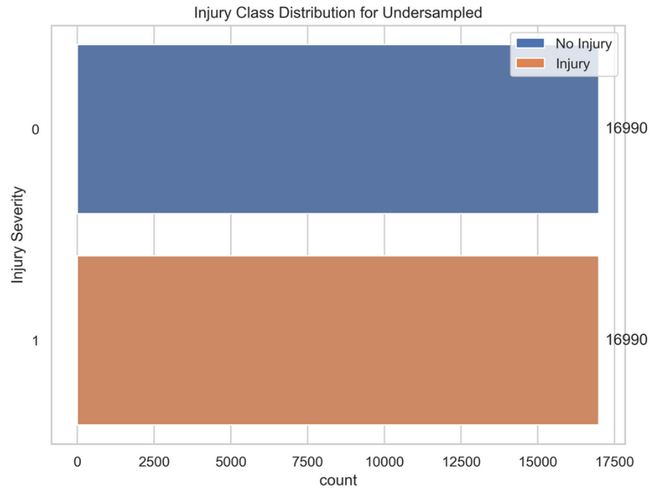
MODELING

ADDRESSING CLASS IMBALANCE

- Undersampling ensures balance, preventing disproportionate influence, and improving model discernment.
- Simultaneously, hyperparameter tuning optimizes model configuration for the best performance.



Undersampling



1

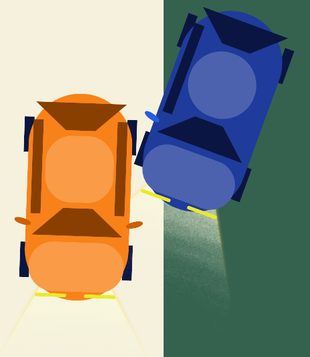
2

- Critical modeling step addresses class imbalance and fine-tunes hyperparameters.
- Class imbalance tackled through undersampling, strategically removing observations from the majority class.

MODELING

HYPERPARAMETER TUNING

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
```

- *Model refinement achieved by adjusting classification threshold through threshold moving.*
- *Optimal threshold set at 0.38, indicating a positive prediction when predicted probability of injury crash is at least 0.38.*

MODELING

MODEL EVALUATION

Final Model Performance with 0.38 threshold

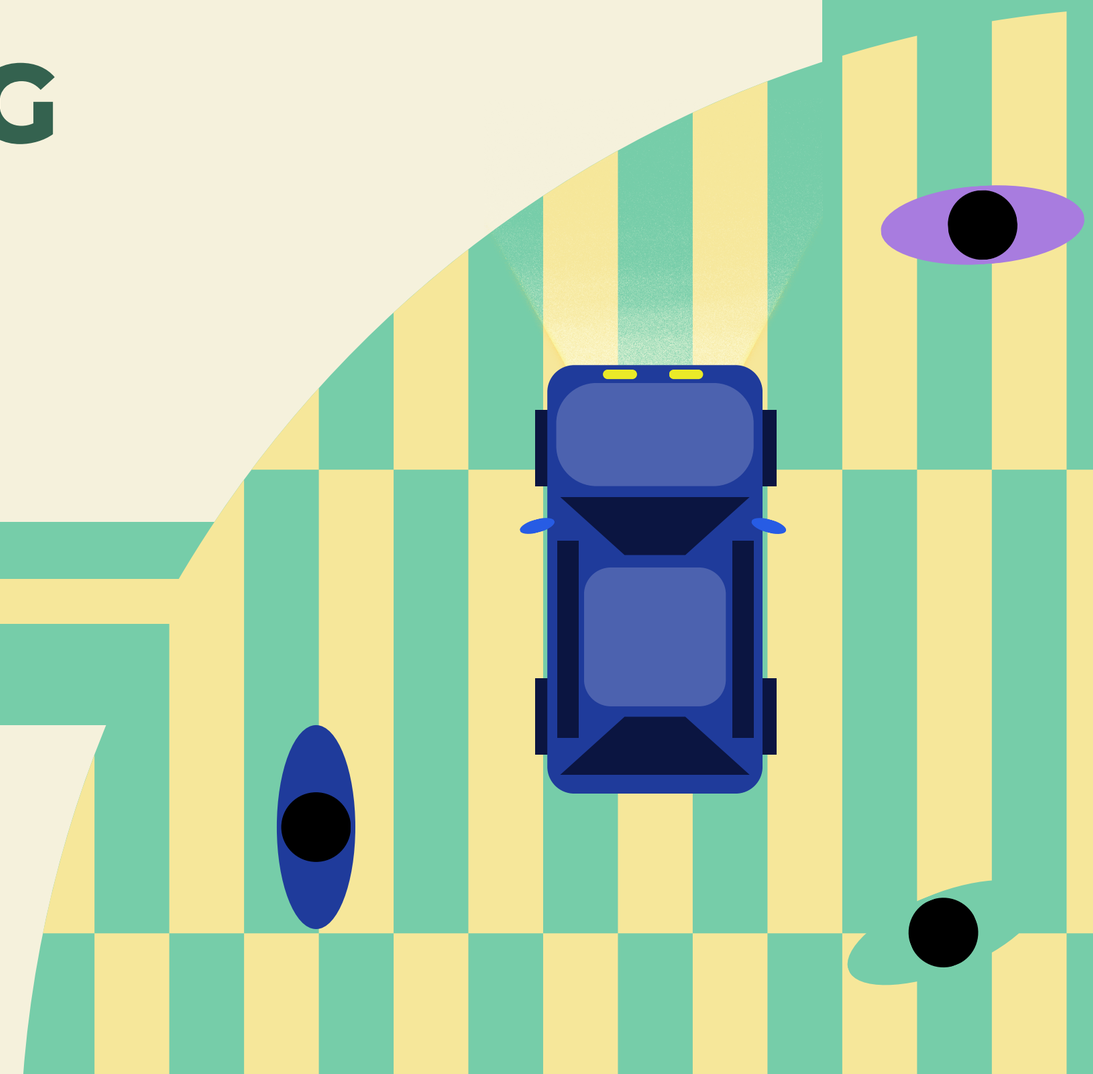
```
Average Best Threshold: 0.382
Average Best F1 Score: 0.7201115921857997
----------------------------------------------------------
Final Model Performance with Threshold of 0.382
              precision    recall  f1-score   support

   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
```

- Critical evaluation considering cost implications of false positives and negatives.
- F1 score as primary metric, complemented by precision and recall examination for both classes.
- Optimization of classification threshold focused on maximizing the F1 score.
- Final model performance metrics: average F1 score of 72%, recall of 91%, precision of 63%.
- Micro precision for non-emergency predictions impressive at 81%.
- Top 20 features found as effective as the entire feature set, ensuring a parsimonious model.

FURTHER MODELING EXTENSIONS



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.

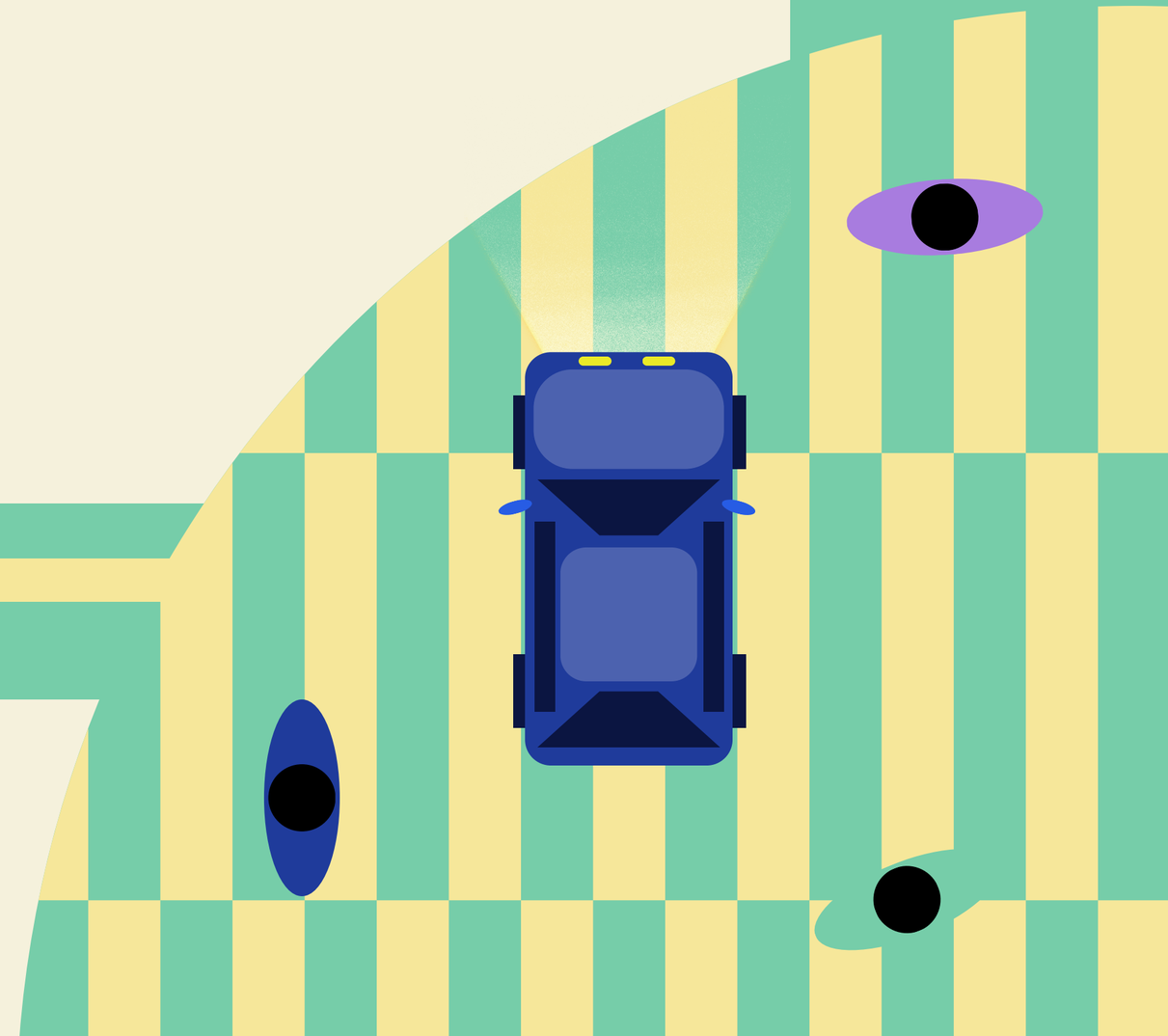
Principal Component Analysis

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

Model Interpretability

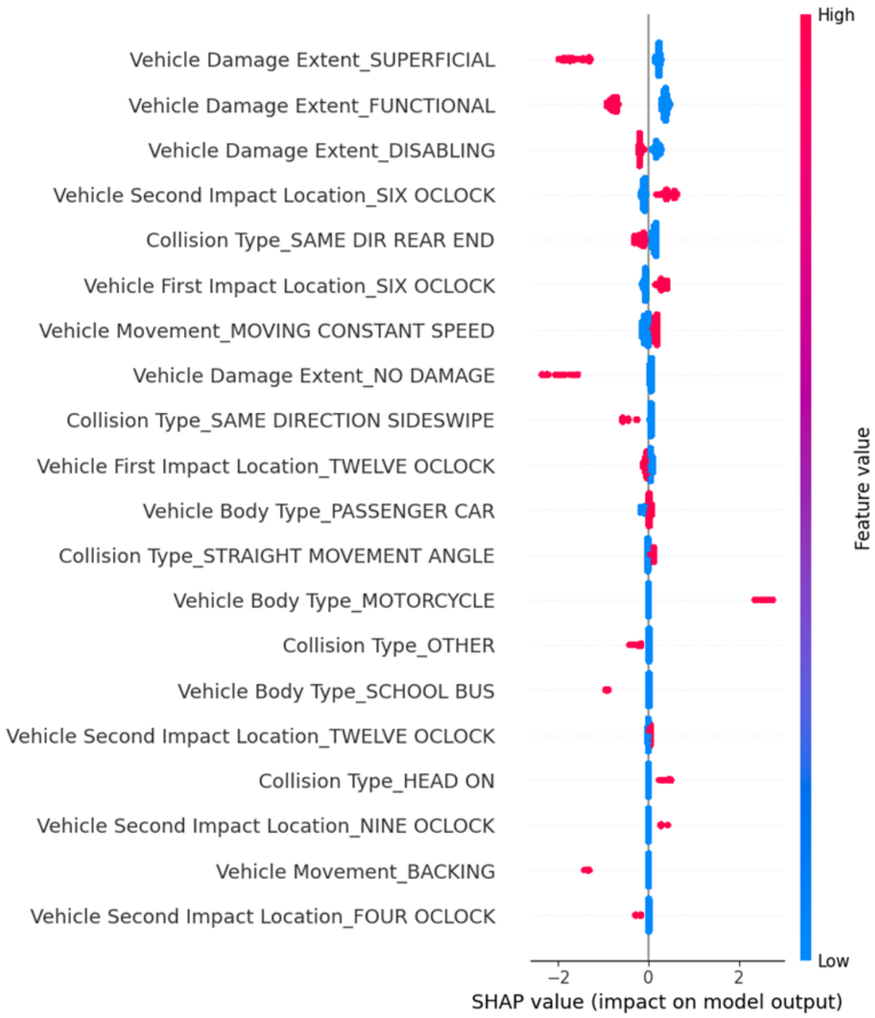
While ensemble algorithms are often viewed as 'black box' due to their complexity, extensive exploratory data analysis (EDA) provided valuable insights for interventions for the use case. In addition, Feature importance gives insight into influential factors for crash injuries.

SHAP BEESWAR



SHAP BEESWAR

PLOT & INTEPRETATION



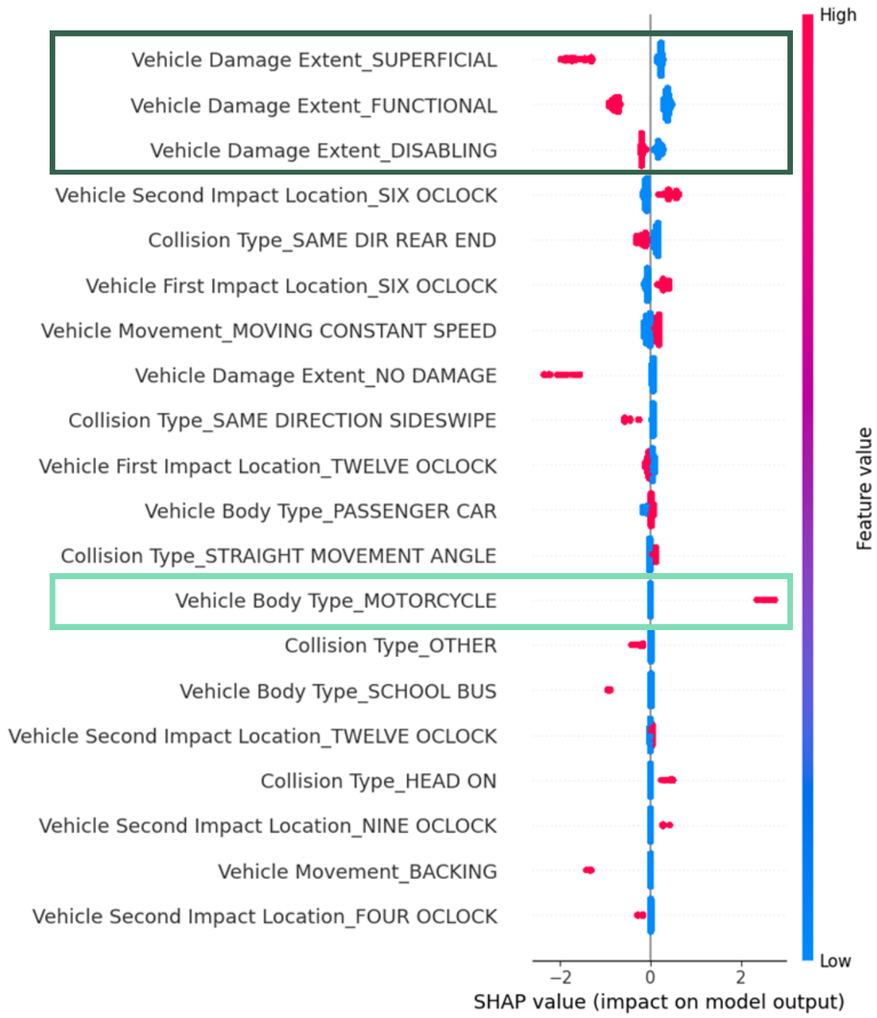
- Input variables ranked from **top** to **bottom** by their mean absolute SHAP values for the entire dataset.

- **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



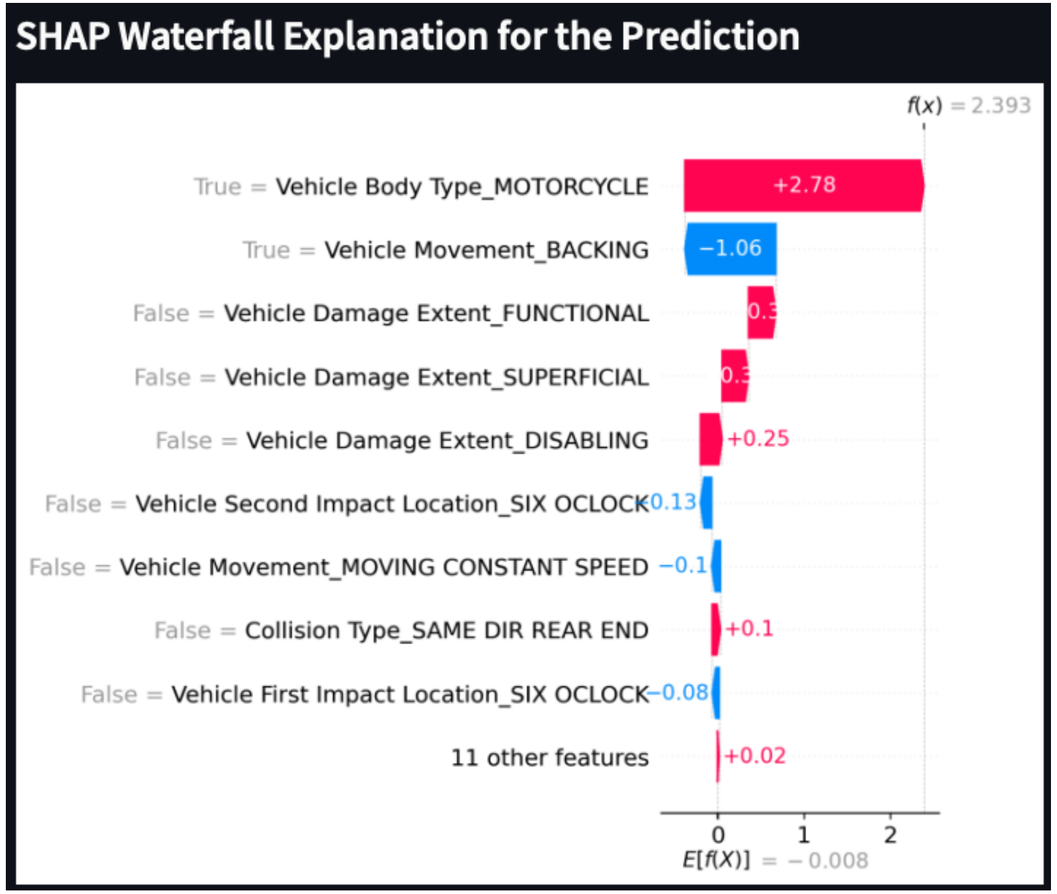
- SHAP values represent **probabilities**, with more **negative** values indicating **lower** injury probability and more **positive** values indicating **higher** injury probability.
- The color bar corresponds to raw variable values, revealing instances of positive (**red**) and negative (**blue**) dummy variables.
- **Horizontal distribution of variables** along the x-axis suggests potential policy interventions, like enhanced training for motorcyclists.

SHAP BEESWAR

PLOT & INTEPRETATION

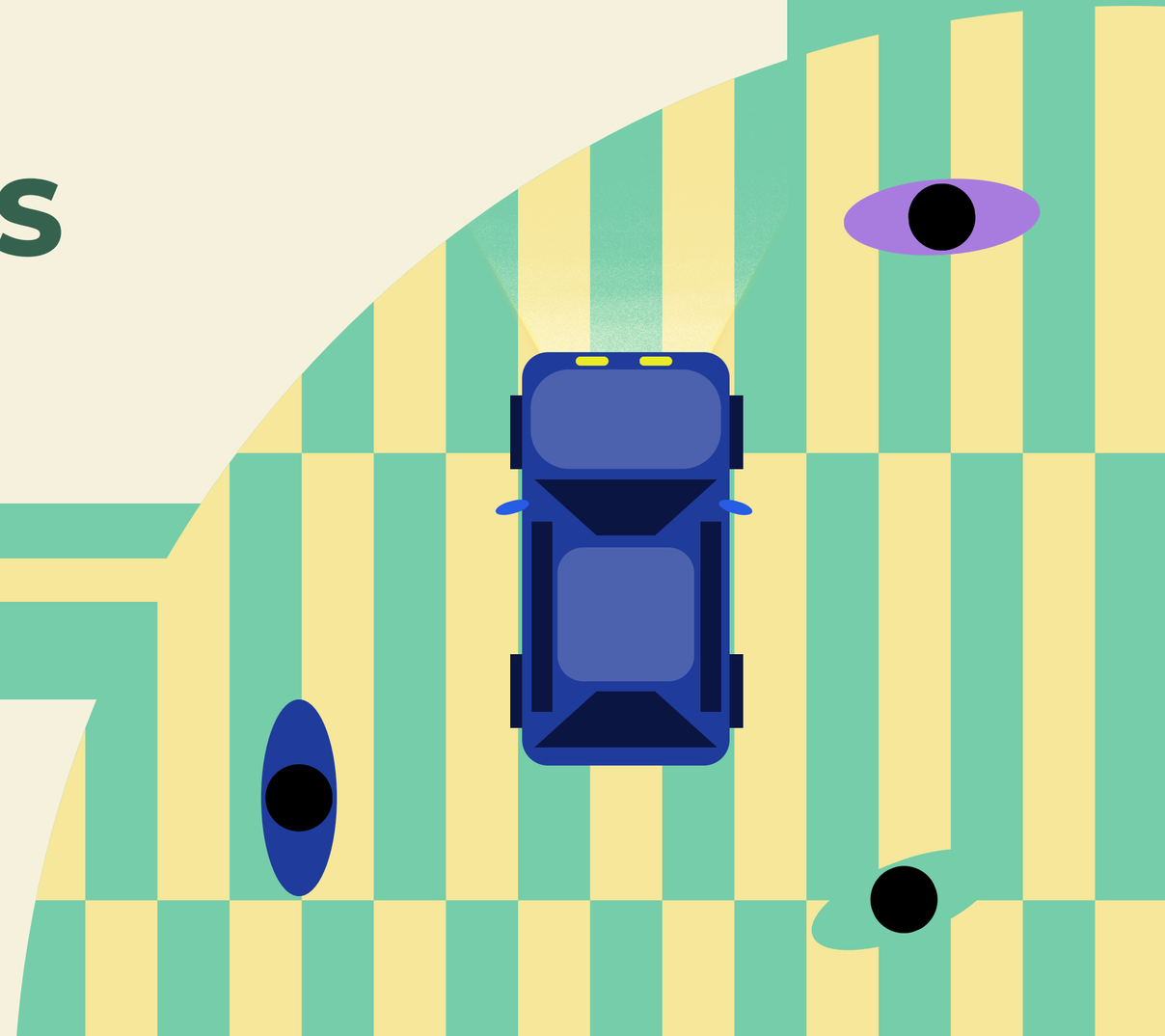
SCAN ME

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



POLICY RECOMMENDATIONS

STRATEGIC & DATA-DRIVEN PROPOSALS FOR ENHANCING ROAD SAFETY

1

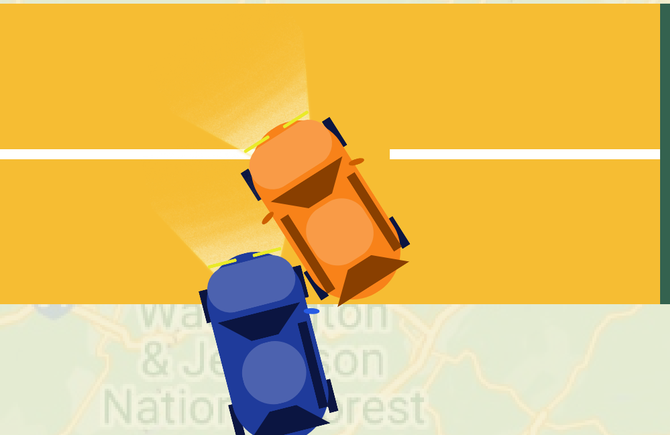
Refining Standard Operating Procedures for Emergency Operators

- Feature importance guides Operators in formulating relevant questions.
- SHAP Waterfall provides insight into the specific variables influencing predictions.
- Enables man-in-the-loop interventions based on interpretable model outputs.

- Evaluate the residency status of motorcyclists involved in accidents within Maryland.
- Consider policies tailored to resident motorcyclists, such as enhanced defensive riding training.
- Explore legislative measures, such as restricting or refusing to legalize lane splitting for motorcyclists in Maryland.

Mitigating High Injury Risks in Motorcyclist-involved Incidents

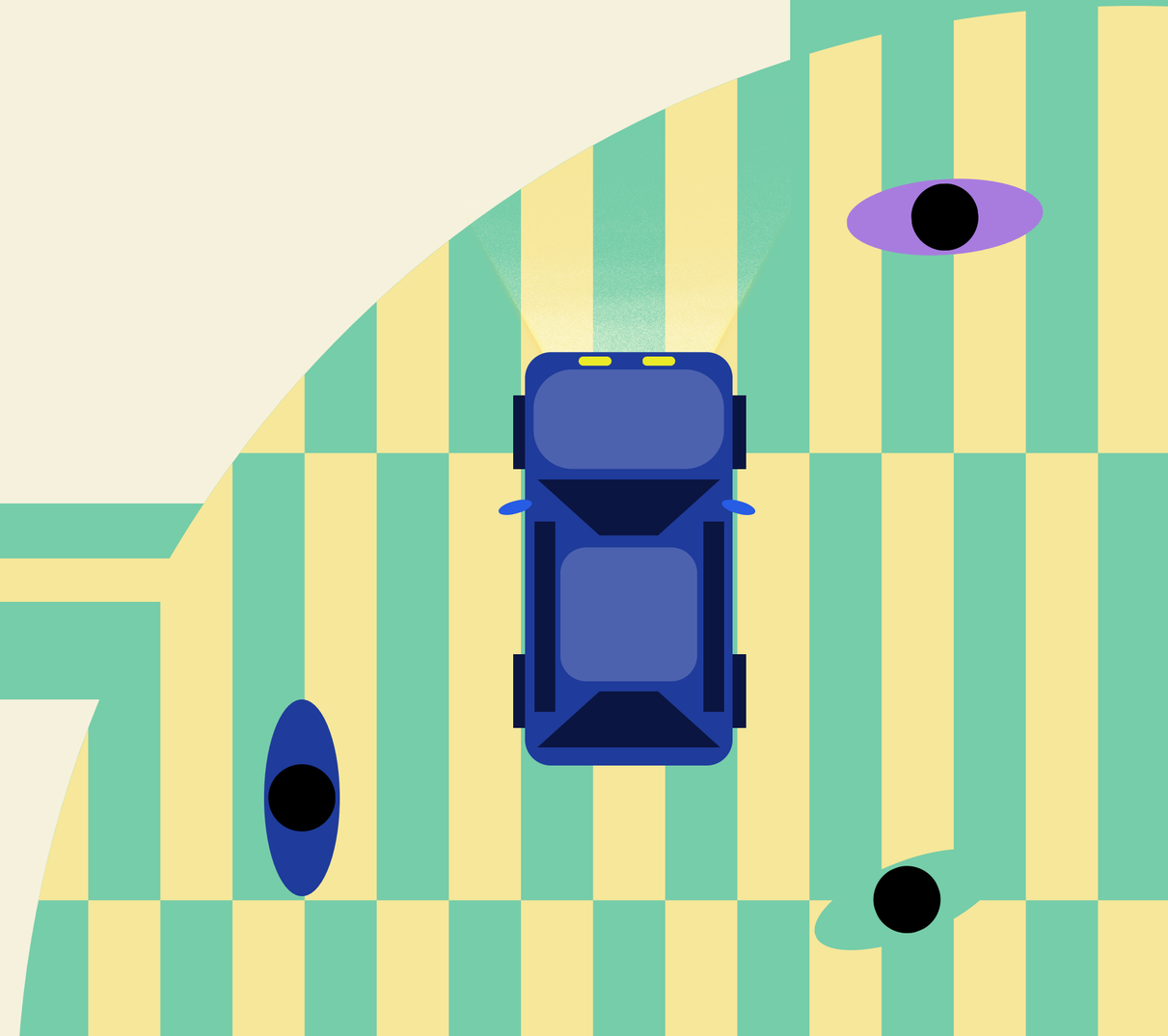
2



Moving Forward...

- Maryland can further develop the clustering approach to assess optimal locations for A&E wards.
- Despite current limitations, additional efforts can enhance the clustering model for improved insights.
- This initiative aims to reduce fatalities from car accidents in the long run.

QUESTIONS?



THANK YOU!