From Crash Records to Machine Learning:

Data Quality Challenges and Decision-Support Opportunities

Yusuf W. Zakaria, Transportation Specialist | WWW Interstate Planning Commission

2026 WV Transportation Planning Conference | June 3, 2026

Presentation Outline

PART 01

Study Background & Data Access

MPO context, GCAT, and the crash record

INTERSTATE PLANNING COMMISSION

Study Area: WWW MPO: Ohio Side

Bi-state MPO spanning West Virginia & Ohio

~4,534
Crash incidents in the record (2020-2023)

GCAT
Primary data access tool (ODOT GIS Crash Analysis Tool)

Multi-year
Period of review: Ohio-side corridor focus

82
Engineered features used in ML model

100%
Neural network classification accuracy (overfitted)

PART 02

Descriptive Overview of Crash Data

Crash types, spatial hotspots, and executive summary

INTERSTATE PLANNING COMMISSION

Summary of Crash Incidents

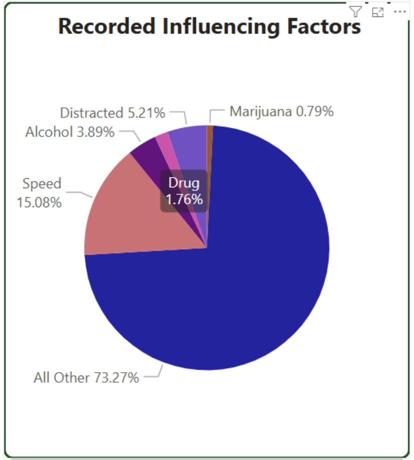
Executive summary: key metrics from the complete crash record

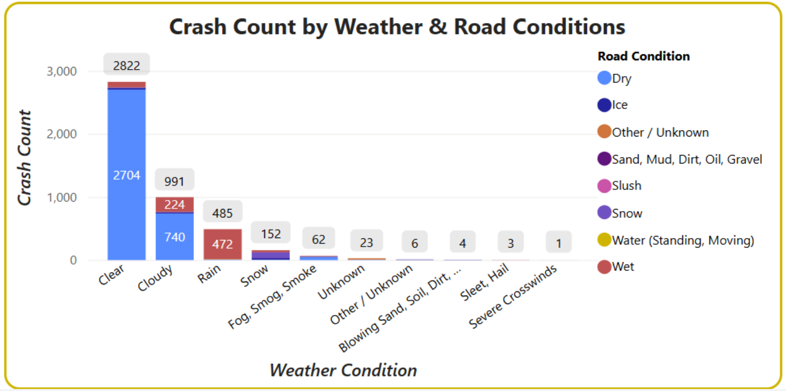
Year

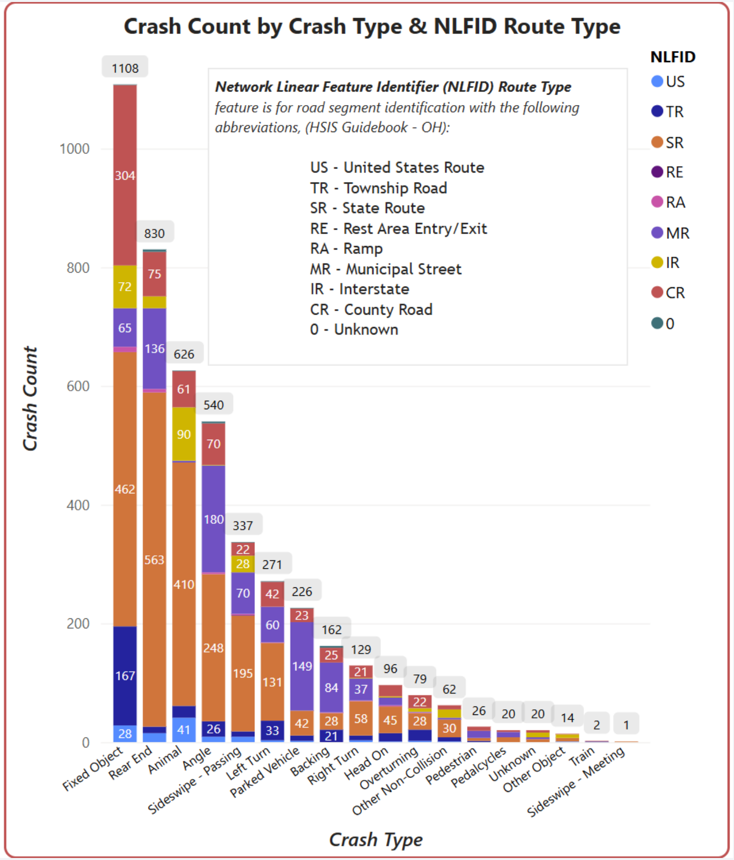
2020 | 2021 | 2022 | 2023

Crash Incidents by NLFID Route ID

NLFID	Crashes	Injuries	Fatalities
SR	2265	539	14
CR	707	161	9
TR	347	90	1
MR	825	75	2
IR	248	27	1
US	115	27	0
RA	26	2	0
0	15	1	0
RE	1	0	0
Total	**4549**	**922**	**27**

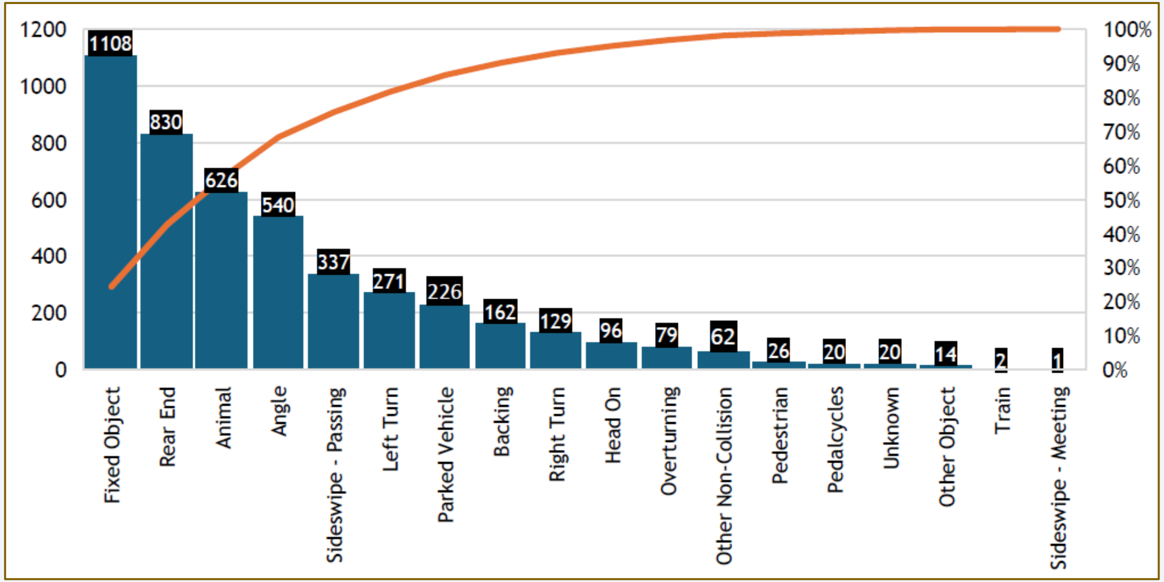






Overview of Crash Types — Raw Crash Data

Pareto analysis: the 80/20 rule holds strongly across this record



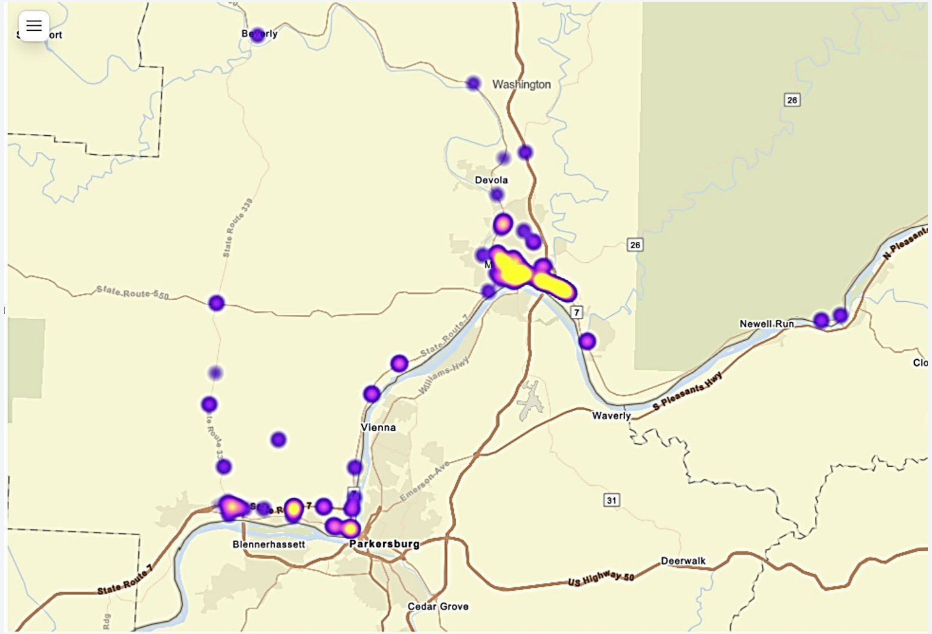
≈80%

of crashes fall under just 5 of 18 crash types

- Rear-end, fixed object, animal & angle crashes dominate
- Intersection types (left/right turn) rank lower than expected
- Non-intersection crashes account for >75% of all incidents (*pp. 23*)
- The Pareto pattern holds consistently across speed-limit subgroups

Crash Hotspot Map & Route Incidence Matrix

Crash hotspots for intersections on the long list



Incidence matrix with row-level heatmap

ROUTE TYPE	2020			2021			2022			2023			TOTALS		
	Crashes	Injuries	Fatalities	Crashes	Injuries	Fatalities	Crashes	Injuries	Fatalities	Crashes	Injuries	Fatalities	Crashes	Injuries	Fatalities
N/A	2	0	0	13	1	0	-	-	-	-	-	-	15	1	0
County Route (CR)	169	26	0	178	47	4	156	40	2	204	48	3	707	161	9
Interstate (IR)	53	4	0	75	2	0	66	4	0	54	17	1	248	27	1
Municipal Route (MR)	193	12	1	227	24	1	183	11	0	222	28	0	825	75	2
Interstate On/Off Ramp (RA)	4	0	0	7	1	0	7	0	0	8	1	0	26	2	0
Rest Area On/Off Ramp (RE)	-	-	-	-	-	-	-	-	-	1	0	0	1	0	0
State Route (SR)	582	125	2	554	140	2	592	132	5	537	142	5	2265	539	14
Township Route (TR)	85	21	0	93	25	0	86	20	1	83	24	0	347	90	1
US Route (US)	23	6	0	32	11	0	26	3	0	34	7	0	115	27	0
TOTALS	1111	194	3	1179	251	7	1116	210	8	1143	267	9	4549	922	27

PART 03

Beyond Descriptive Analytics

Why machine learning? — opportunity and data readiness

INTERSTATE PLANNING COMMISSION

From Visualization to Machine Learning

~4,500 data points: significant, but constrained

THE OPPORTUNITY

- ML reveals contribution of factors within each record beyond counts and maps
- Identifies patterns invisible to cross-tabulation or summary statistics
- Supports targeted safety campaigns and corridor-specific countermeasures
- Transferable framework to similar roadway contexts in the MPO region

DATA CHALLENGES

1 Erroneous or missing values

One or more attributes incomplete in a significant share of records

2 Roadway upgrades within study period

Network changes between crash years affect attribute validity

3 Attribute Homogeneity

Some categorical variables lack sufficient spread for ML

4 Nuanced variables

Granular attributes need careful encoding or consolidation

PART 04

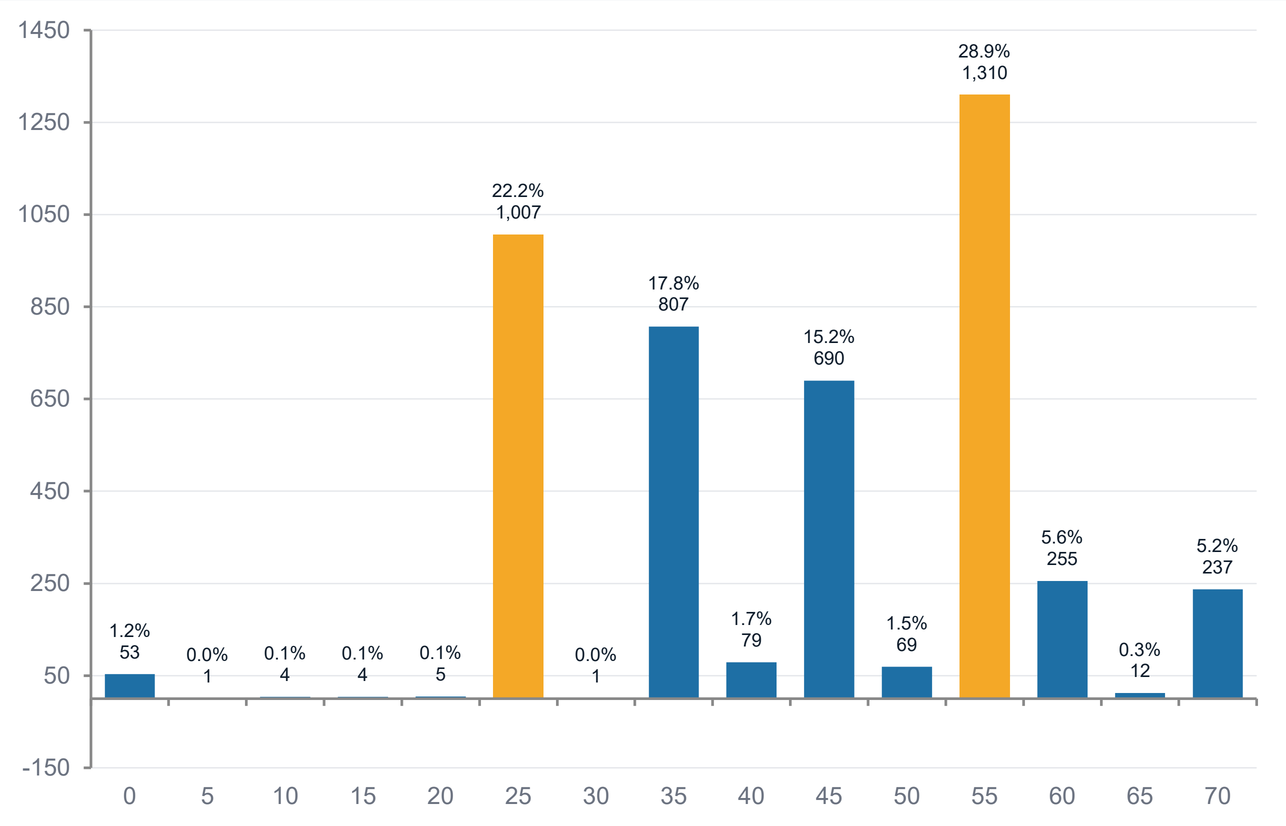
Feature Analysis & Engineering

Preparing crash variables for machine learning

INTERSTATE PLANNING COMMISSION

Feature: Posted Speed Limit Distribution

Unit 1 as principal reference: proximity assumption for other units



28.9%

of crashes at 55 mph (1,310 incidents)

22.2%

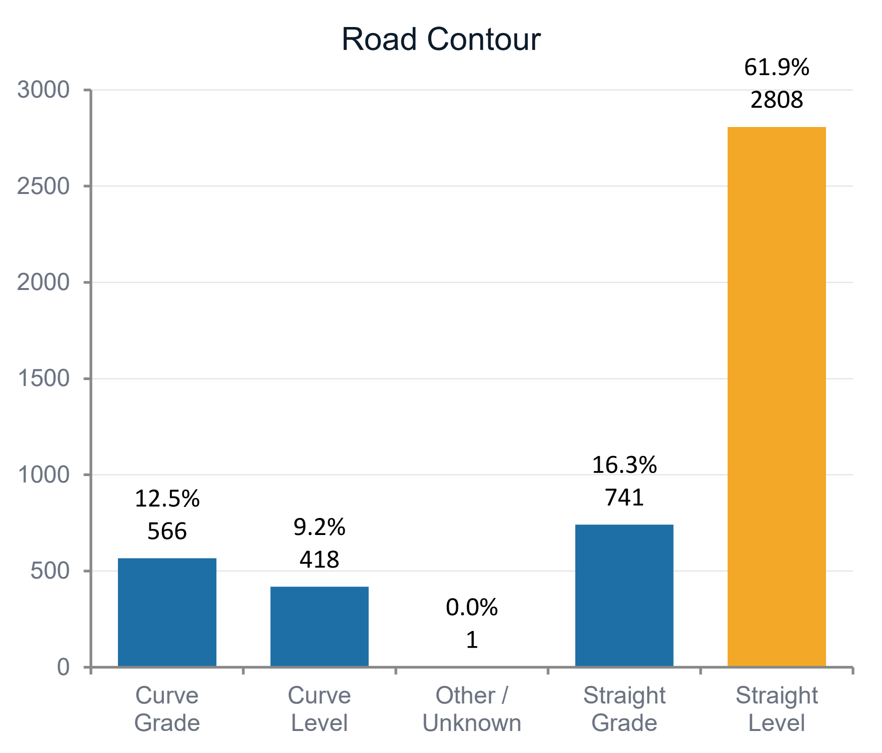
at 25 mph, urban corridor (1,007)

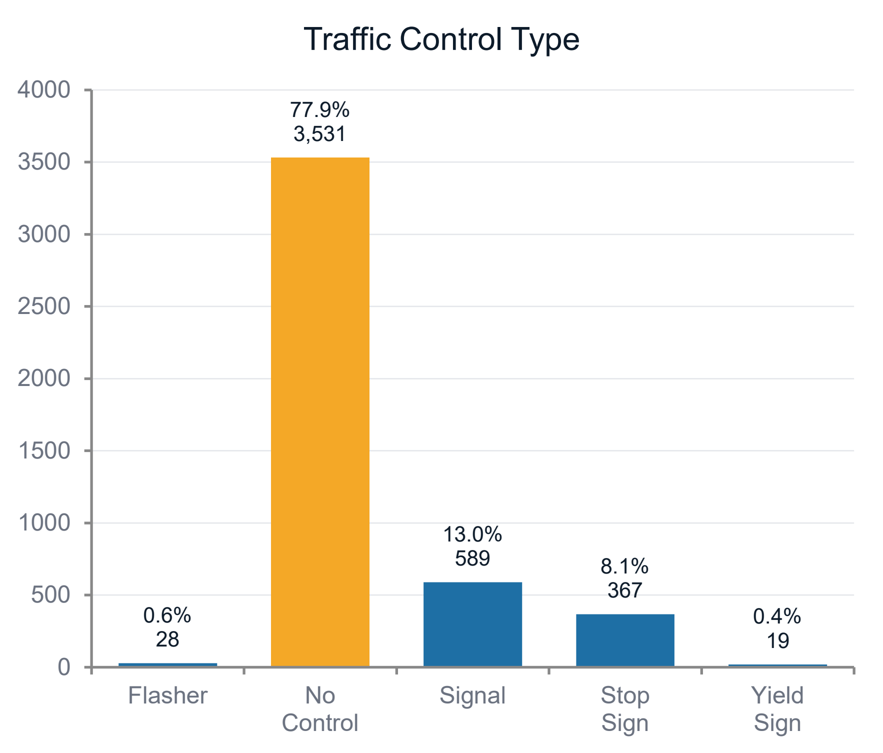
<2%

at speeds 0–20 mph combined

Road Contour & Traffic Control Type

Roadway geometry and control characteristics at crash location



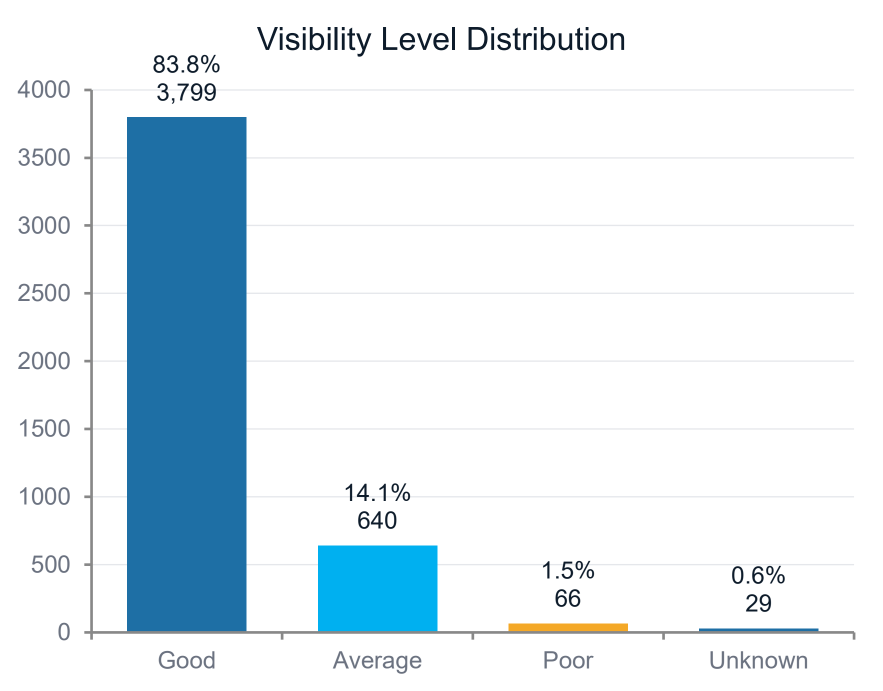


78% straight roads • 62% straight & level • 78% uncontrolled traffic: crash profile skews heavily toward open-road, non-intersection events

Feature Engineering: Visibility from Weather Condition

Simplified 4-category encoding replaces 10+ original weather variables with perceived visibility

Category	Original Variable	% of Record
Good	Clear	**62.0%**
	Cloudy	**21.7%**
	Severe Crosswinds	**0.0%**
Average	Rain	**10.7%**
	Sleet / Hail	**0.1%**
	Snow	**3.4%**
Poor	Blowing Sand, Soil, Dirt, Snow	**0.1%**
	Fog, Smog, Smoke	**1.4%**
Unknown	Other / Unknown	**0.1%**
	Unknown	**0.5%**

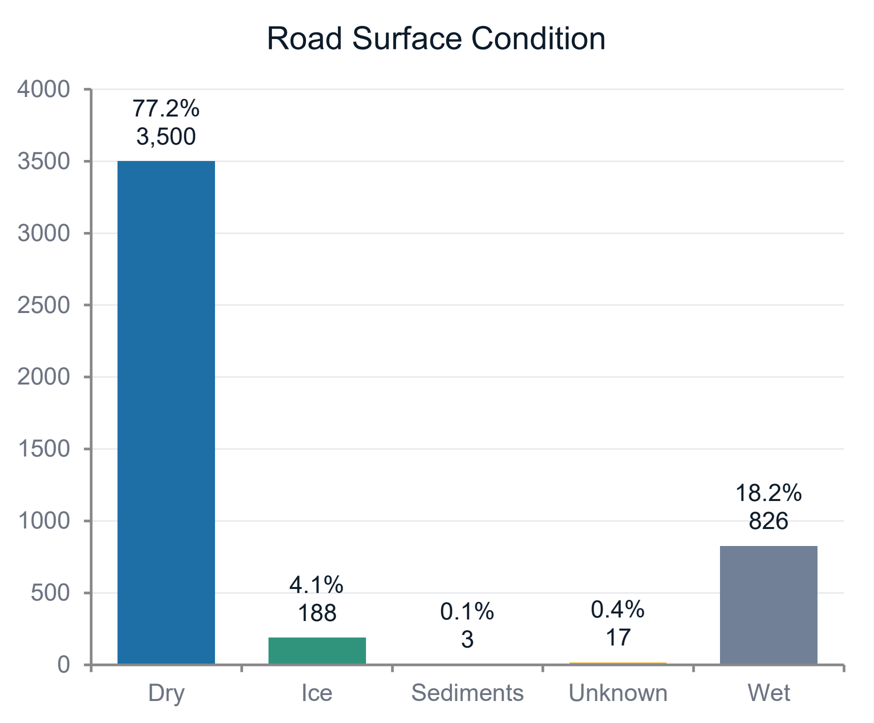


83.8% of crashes occurred in GOOD visibility; weather is not the primary risk driver

Feature Engineering: Road Surface Condition

Simplified 5-category encoding replaces 8+ original road condition

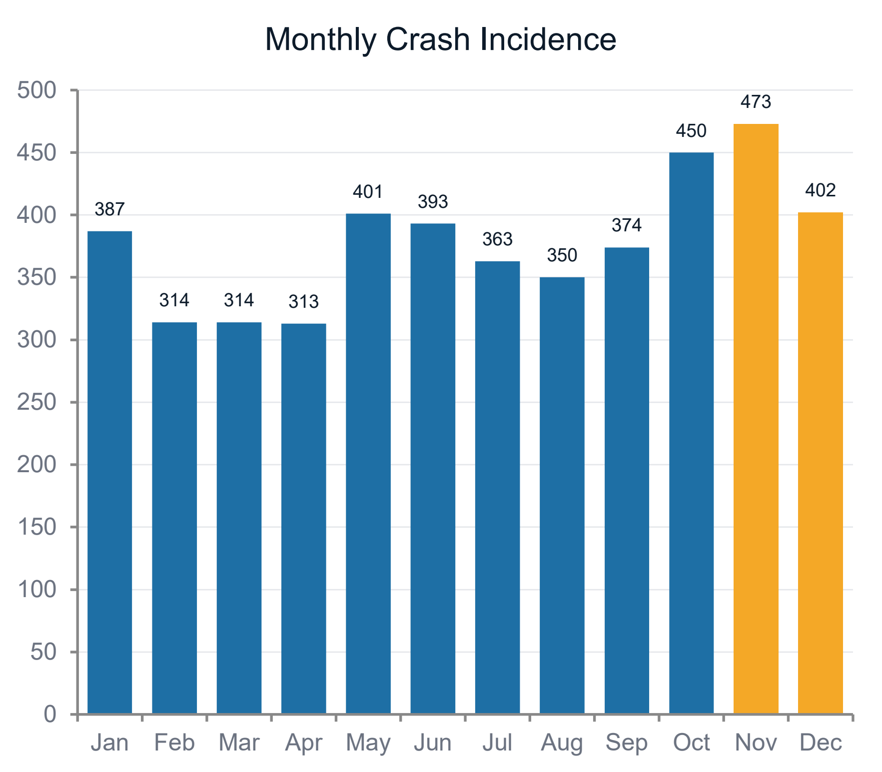
Category	Original Variable	% of Record
Dry	Dry	**77.2%**
Ice	Ice	**1.5%**
	Slush	**0.2%**
	Snow	**2.4%**
Sediments	Sand, Mud, Dirt, Oil, Gravel	**0.1%**
Unknown	Other / Unknown	**0.4%**
Wet	Water (Standing, Moving)	**0.2%**
	Wet	**18.1%**

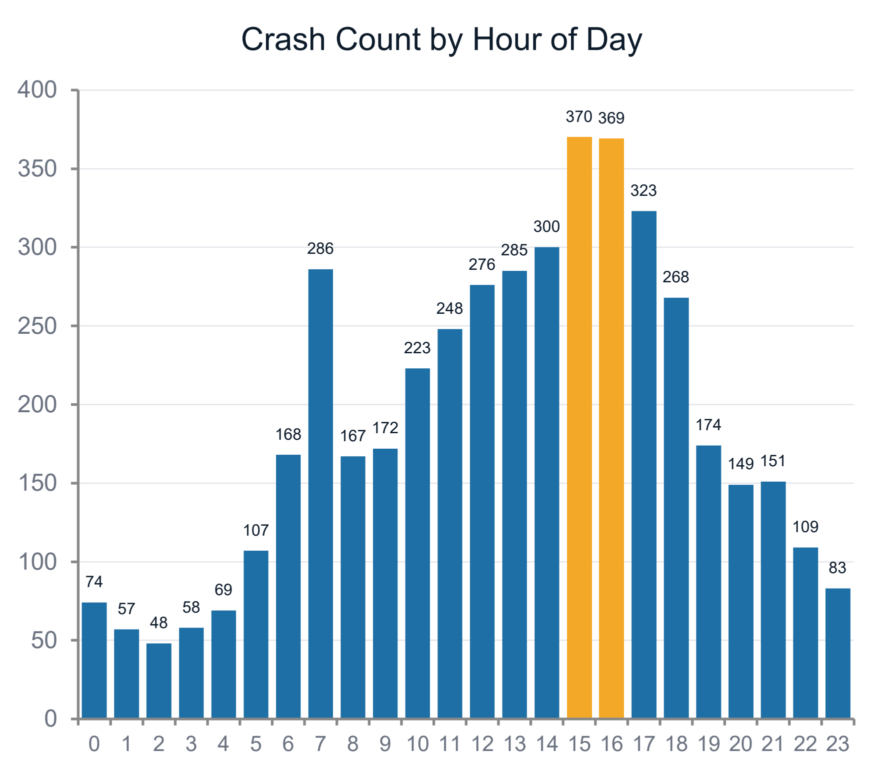


77.2% dry road crashes

Temporal Distribution: Monthly & Hourly Patterns

When crashes happen — seasonal trends and daily peaks

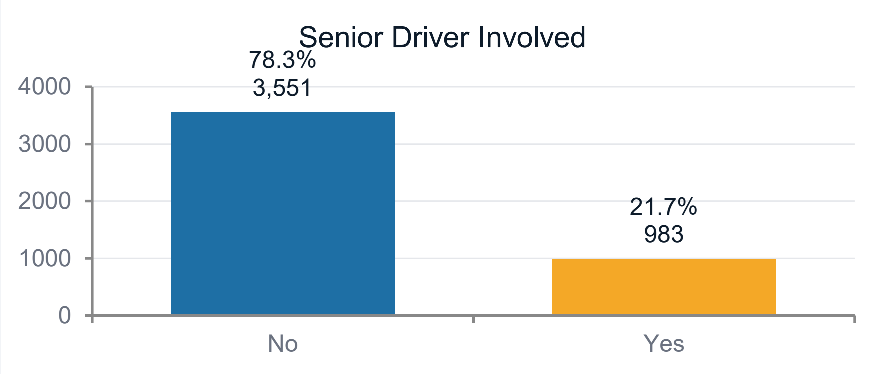


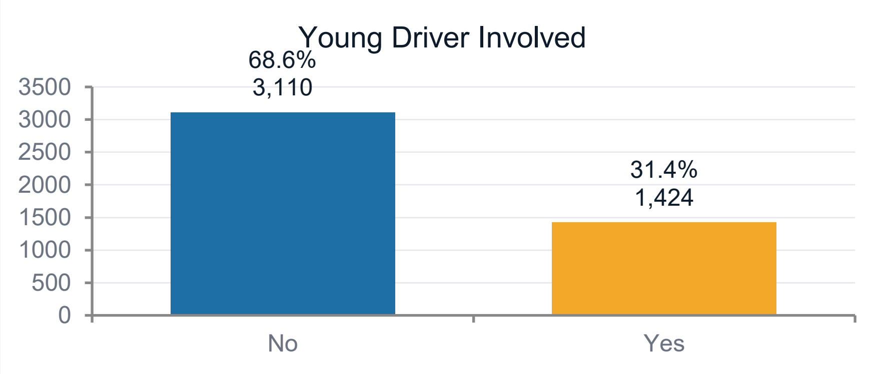


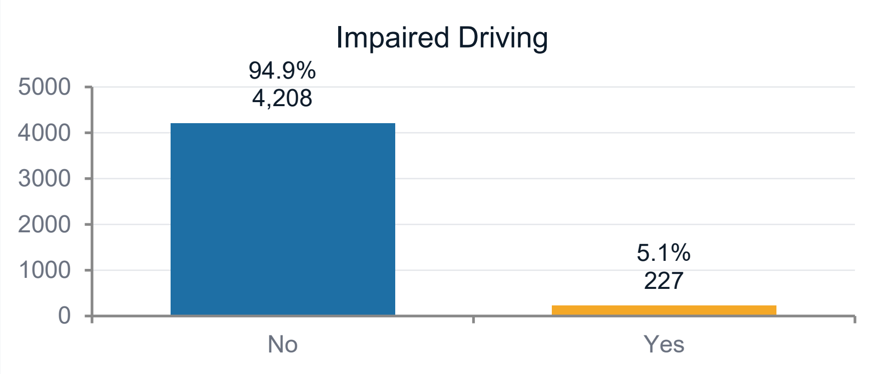
Oct–Nov peak seasonally • Afternoon peak 3–4 PM • Low overnight hours; behavioral and exposure patterns drive timing

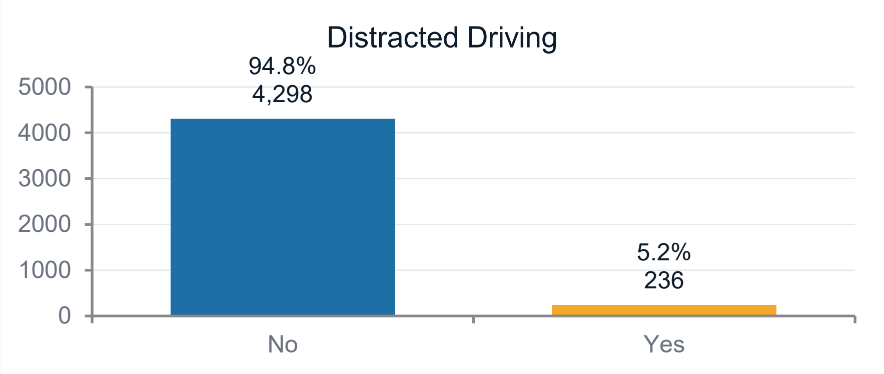
Driver Age & Behavioral Influencing Factors

Binary features: senior/young driver involvement and impairment indicators





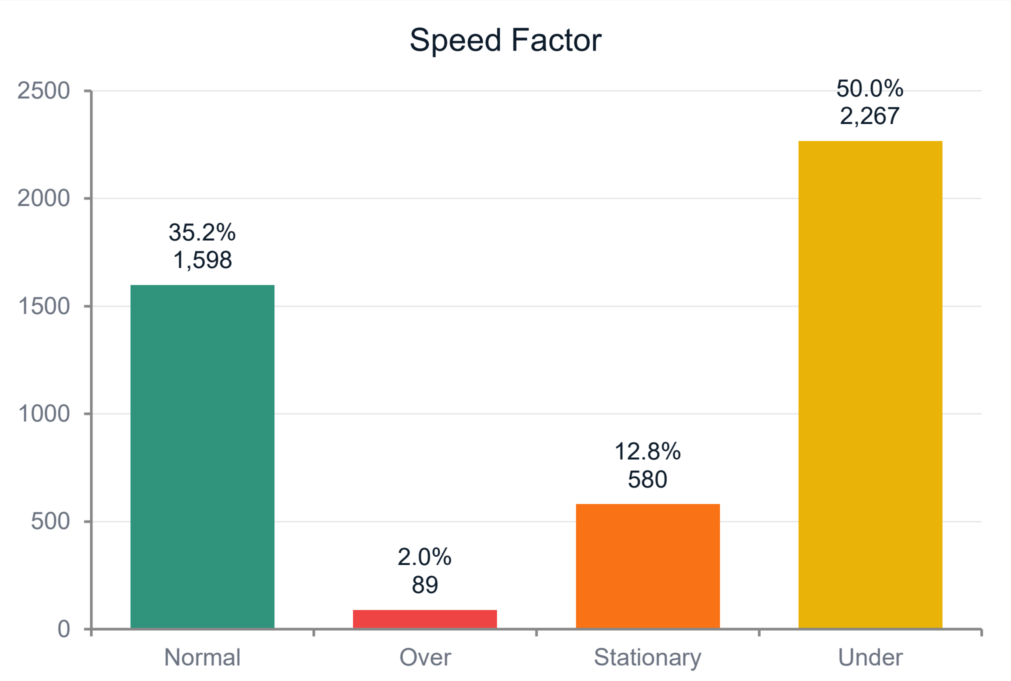




Senior 21.7% • Young 31.4% • Impaired 5.1% • Distracted 5.2%; yet these 4 factors drive 81% of fatalities in the record

Feature Engineering: Speed Limit

Discrete numerical values of principal driver speed compared to their posted/legal speed limit *+/-5MPH vs 0* was used for encoding.
Numerical representation for speed is likely not accurate.



50% of incidents: under-speed or stationary vehicles • Only 2% classified as over-speeding • Speed encoding derived from Unit 1 posted vs. recorded speed (±5 mph threshold)

Influencing Factors vs. Crash Severity

4 factors responsible for 81% of all fatalities in the record despite covering only 21% of crashes

Filter	Crashes	Injuries	Serious Injuries	Fatalities
Complete Record	***4,534***	***921***	***161***	***27***
Alcohol Related	177	71	15	**6**
Drug Related	80	48	16	**9**
Marijuana Related	36	23	9	**3**
Speed Related	685	221	44	**4**
% of Complete Record	**21.57%**	**39.41%**	**52.17%**	**81.48%**

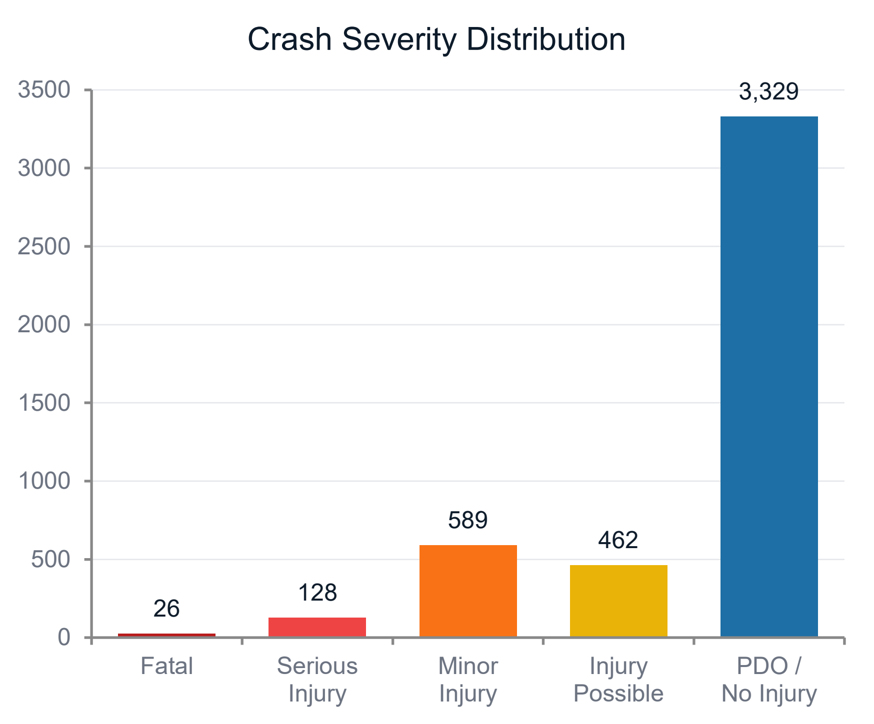
These factors are not mutually exclusive; together they account for 21 of 27 fatalities in the record

Classification Encoding: Crash Severity

5-level severity hierarchy re-encoded to 3 severity levels (for possible use as target for classification)

Resulting "subjective encoding" is inherently **imbalanced** and rational.

Original	% of original	New encoding	% of new encoding
Fatal Injuries	**0.57%**	**Fatal/Serious**	3.396%
Serious Injuries	**2.82%**		
Minor Injury	**13%**	**Minor Injury/Injury Possible**	23.18%
Injury Possible	**10.1%**		
No Injury/ PDO (Property Damage Only)	**73.4%**	**PDO/No Injury**	73.42%



77.2% dry road condition crashes • 73.4% property-damage-only • "Only" 0.57% fatal crashes

PART 05

Machine Learning (ML) Model Results & Potential Policy Direction

Neural network classification and SHAP feature importance by speed-limit segment

INTERSTATE PLANNING COMMISSION

Final ML Model: Approach & SHAP Interpretation Guide

Neural network binary classifier

Model:	Neural network binary classifier — Fatal/Serious vs. Minor+
Features:	82 engineered features from all crash record and feature engineering
Accuracy:	100% correctly classified all speed-limit segment observations
Explainer:	SHAP (SHapley Additive exPlanations)
Segments:	NLFID 0 • 25 mph • 35 mph • 40 mph • 45 mph • 50 mph • 55 mph • 60 mph • 65–70 mph • Hotspot Intersections

How to Read the SHAP Plots

- Red points = attribute present/high value
- Blue points = attribute absent/low value
- Right of axis = FOR Fatal/Serious outcome
- Left of axis = AGAINST Fatal/Serious
- Higher score = greater feature importance
- Ambiguous = unclear directional impact

⚠️ CAUTION:

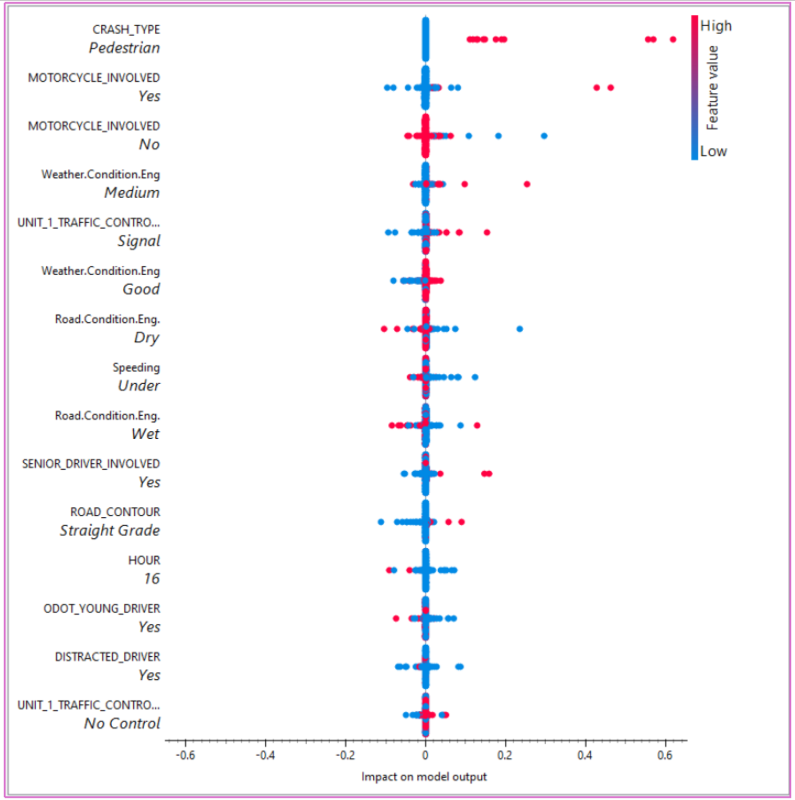
- ❑ The presence of some features should be interpreted carefully – especially features that induce certain driving habits.
- ❑ Opposing features as well as “uncommon/unusual” states CAN show up.

SHAP Feature Importance By Speed-Limit Segment

Top contributing features predicting Fatal/Serious crash outcome at 25 mph

Rank	Feature	Score	Impact
1	Pedestrian	0.003087	For
2	Motorcycle involved=Yes	0.002075	For
3	Motorcycle involved=No	0.001637	For
4	Weather condition=Medium	0.001586	Against
5	Traffic control=Signal	0.001529	For
6	Weather condition=Good	0.001491	For
7	Road condition=Dry	0.001485	Against
8	Speeding=Under	0.001441	Against
9	Road condition=Wet	0.001391	Against
10	Senior driver involved=Yes	0.001265	For
11	Road contour=Straight Grade	0.001229	For
12	Hour=16	0.001164	For
13	Young driver involved =Yes	0.001114	Against
14	Distracted driver=Yes	0.001112	Against
15	Traffic control=No Control	0.001103	For

999 observations (sampled), 82 features | 11 positive (100% correct)

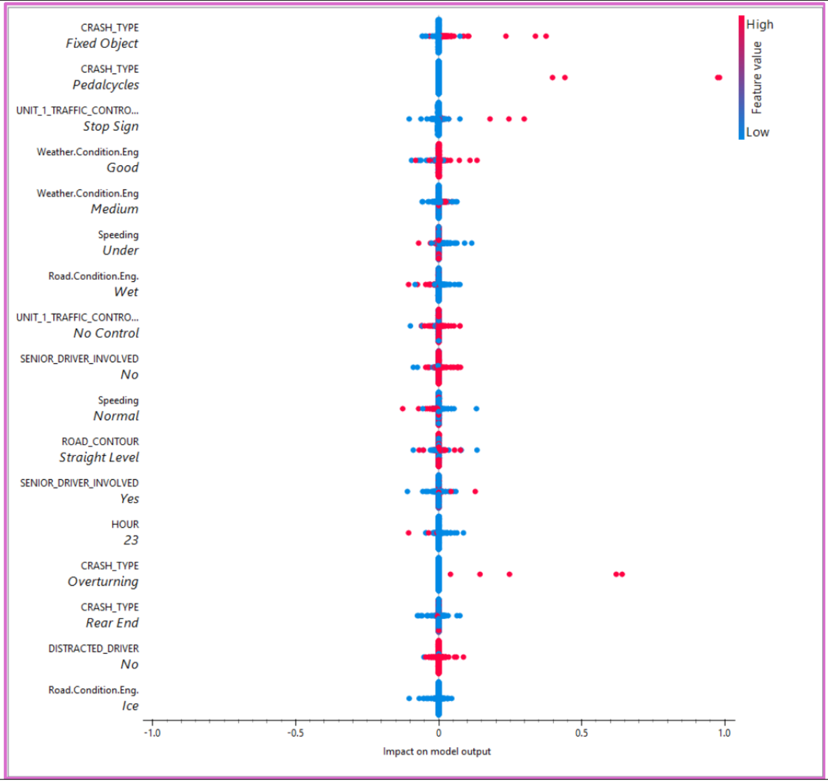


SHAP Feature Importance By Speed-Limit Segment

Top contributing features predicting Fatal/Serious crash outcome from 35 mph to 40 mph

Rank	Feature	Score	Impact
1	Crash type=Fixed Object	0.004666	For
2	Crash type =Cyclist	0.003231	For
3	Traffic control=Stop Sign	0.002946	For
4	Weather condition=Good	0.002649	For
5	Weather condition=Medium	0.002601	For
6	Speeding=Under	0.002576	Against
7	Road condition=Wet	0.00249	Against
8	Traffic control=No Control	0.002405	For
9	Senior driver =No	0.002387	For
10	Speeding=Normal	0.002361	Against
11	Road contour=Straight Level	0.002231	For
12	Senior driver involved=Yes	0.002102	For
13	Hour=23	0.001971	Against
14	Crash type=Overturning	0.001957	For
15	Crash type=Rear End	0.001919	?
16	Distracted driver=No	0.001917	For
17	Road condition=Ice	0.001887	?

866 Observations, 82 features | 16 positive (100% correct)

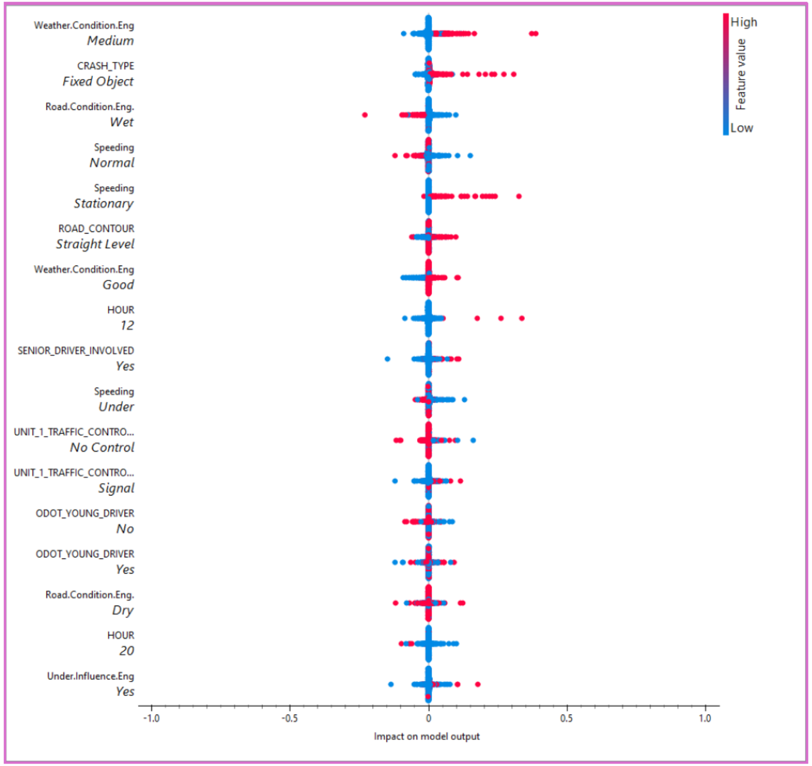


SHAP Feature Importance By Speed-Limit Segment

Top contributing features predicting Fatal/Serious crash outcome from 45 mph to 50 mph

Rank	Feature	Score	Impact
1	Weather condition=Medium	0.0128001	For
2	Crash type=Fixed Object	0.00829962	For
3	Road condition=Wet	0.00610447	Against
4	Speeding=Normal	0.00520287	Against
5	Speeding=Stationary	0.00517435	For
6	Road contour=Straight Level	0.00510881	For
7	Weather condition=Good	0.00491014	For
8	Hour=12	0.00486565	For
9	Senior driver involved=Yes	0.00483998	For
10	Speeding=Under	0.00460205	Against
11	Traffic control=No Control	0.00454589	Against
12	Traffic control=Signal	0.00429794	For
13	Young driver=No	0.00428204	Against
14	Young driver=Yes	0.00427187	?
15	Road condition=Dry	0.00422343	Against
16	Hour=20	0.00408634	Against
17	Under influence=Yes	0.00406031	For

743 observations, 82 features | 25 positives

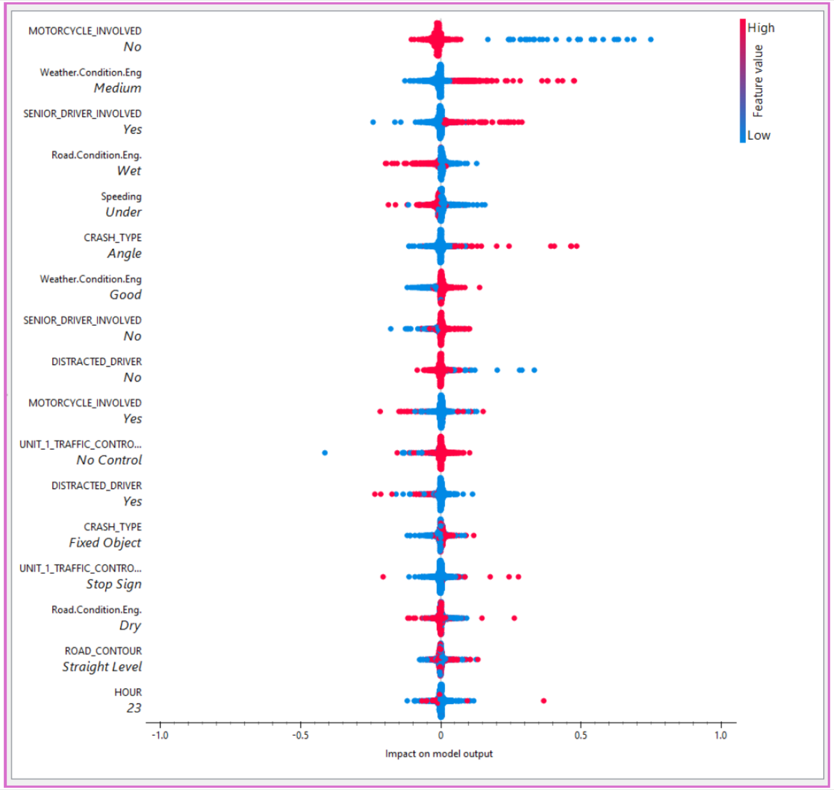


SHAP Feature Importance By Speed-Limit Segment

Top contributing features predicting Fatal/Serious crash outcome at 55 mph

Rank	Feature	Score	Impact
1	Motorcycle involved=No	0.0278751	Against
2	Weather condition=Average	0.0275047	For
3	Senior driver involved=Yes	0.0169162	For
4	Road condition=Wet	0.0154724	Against
5	Speeding=Under	0.0130703	Against
6	Crash type=Angle	0.0128399	For
7	Weather condition =Good	0.0117153	For
8	Senior driver involved=No	0.0104429	For
9	Distracted driver=No	0.010173	Against
10	Motorcycle involved=Yes	0.0100304	?
11	Traffic control=No Control	0.0100276	For
12	Distracted driver=Yes	0.00990674	Against
13	Crash type=Fixed Object	0.00967049	For
14	Traffic control=Stop Sign	0.00936721	For
15	Road condition=Dry	0.00936313	Against
16	Road contour=Straight Level	0.00930023	For
17	Hour=23	0.00908578	?

1276 observations (sampled), 82 features | 81 positive

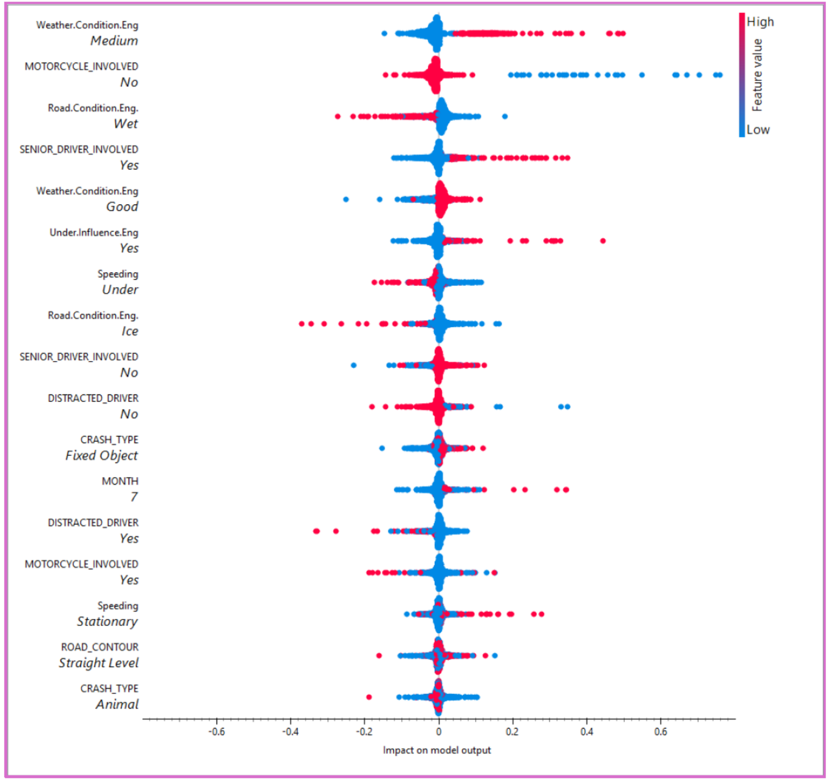


SHAP Feature Importance By Speed-Limit Segment

Top contributing features predicting Fatal/Serious crash outcome from 55 mph to 60 mph

Rank	Contributing Feature	Score	Impact
1	Weather condition=Average	0.0382039	For
2	Motorcycle involved=No	0.0278771	Against
3	Road condition=Wet	0.0214904	Against
4	Senior driver involved=Yes	0.0212895	For
5	Weather condition=Good	0.0162219	For
6	Under influence=Yes	0.0149581	For
7	Speeding=Under	0.014073	Against
8	Road condition=Ice	0.013592	Against
9	Senior driver involved =No	0.0126537	For
10	Distracted driver=No	0.0124853	Against
11	Crash type=Fixed Object	0.0124704	For
12	Month=7	0.0121388	For
13	Distracted driver=Yes	0.0119581	Against
14	Motorcycle involved=Yes	0.0113056	?
15	Speeding=Stationary	0.0111152	For
16	Road contour=Straight Level	0.0105726	For
17	Crash type=Animal	0.0105344	?

1522 observations (sampled), 82 features | 91+ve (100% correct)

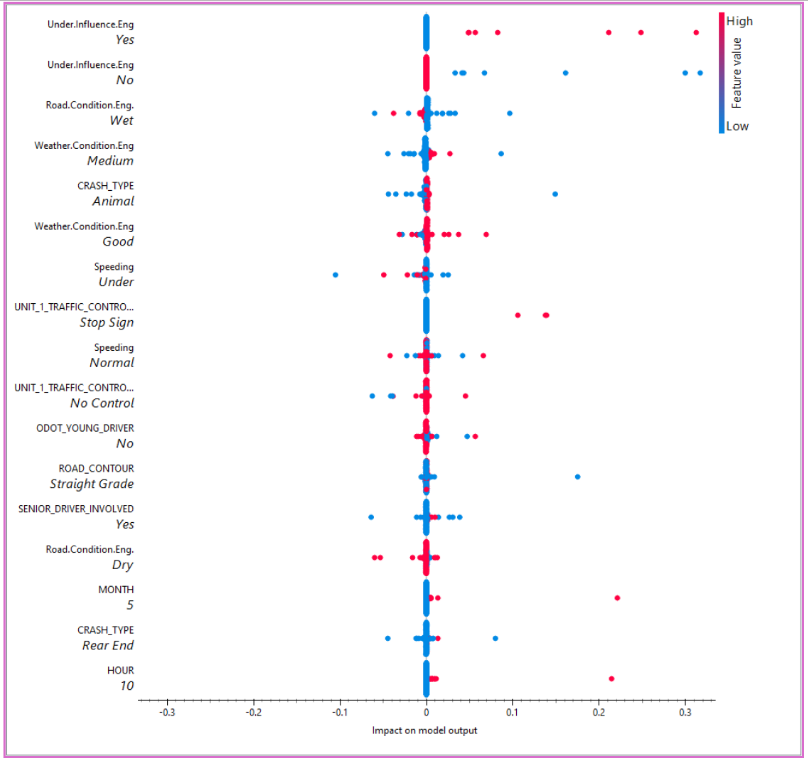


SHAP Feature Importance By Speed-Limit Segment

Top contributing features predicting Fatal/Serious crash outcome from 65 mph to 70 mph

Rank	Contributing Feature	Score	Impact
1	Under influence=Yes	0.00477848	For
2	Under influence=No	0.00456556	Against
3	Road condition=Wet	0.00299182	Against
4	Weather condition=Medium	0.00297561	For
5	Crash type=Animal	0.00230531	For
6	Weather condition=Good	0.00221277	For
7	Speeding=Under	0.00190368	Against
8	Traffic control=Stop Sign	0.00181295	For
9	Speeding=Normal	0.00157738	?
10	Traffic control=No Control	0.0015289	For
11	Young driver involved=No	0.00149408	Against
12	Road contour=Straight Grade	0.00144831	?
13	Senior driver involved=Yes	0.0014402	For
14	Road condition=Dry	0.00137278	Against
15	Month=5	0.00134704	For
16	Crash type=Rear End	0.00134298	For
17	Hour=10	0.00132947	For

211 observations, 82 features | 3 positive (100% correct)

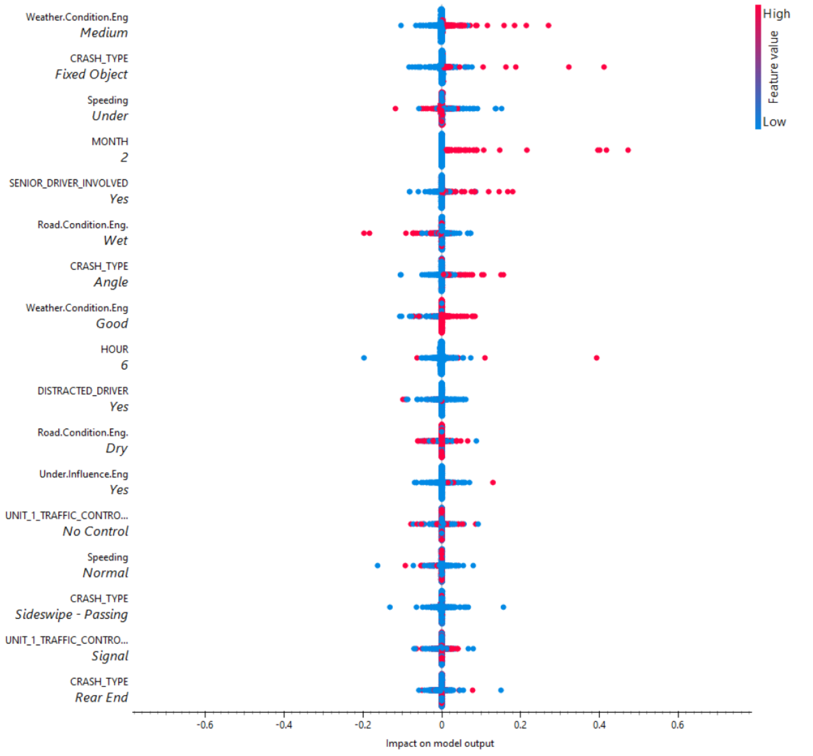


SHAP Feature Importance By Speed-Limit Segment

Top contributing features predicting Fatal/Serious crash outcome at "Hotspot" Intersections

Rank	Feature	Score	Impact
1	Weather condition = Medium	0.006172	For
2	Crash type=Fixed Object	0.004516	For
3	Speeding=Under	0.004063	Against
4	Month =2	0.004062	For
5	Senior driver involved=Yes	0.003877	For
6	Road Condition=Wet	0.003812	Against
7	Crash type=Angle	0.003766	For
8	Weather condition=Good	0.00376	For
9	Hour=6	0.003462	?
10	Distracted driver=Yes	0.003288	?
11	Road condition=Dry	0.00309	?
12	Under influence=Yes	0.003086	For
13	Traffic control=No Control	0.003081	For
14	Speeding=Normal	0.003012	Against
15	Crash type=Sideswipe-Passing	0.002931	?
16	Traffic control=Signal	0.002906	For
17	Crash type=Rear End	0.002891	?

1051 observations (sampled) 82 features | 27+ves (100% correct)



PART 06

From Findings to Actions

The Road Forward: Data Strategy for ML-Ready Crash Records

INTERSTATE PLANNING COMMISSION

From Reporting to Intelligence: A Forward-Looking Data Strategy

Rethinking the crash data pipeline: from the scene report to the machine learning model

Scene Report — Officer capture ▶ **Structured Entry** — Standardized fields ▶ **Validation & QC** — Flags & review ▶ **GIS Integration** — Spatial enrichment ▶ **ML-Ready Pipeline** — Analysis & insight

Descriptive fields have value: inputs should be consolidated, not elimination entirely

- Narrative and detailed descriptive fields remain important for **reporting, public communication, and contextual understanding**. The problem for use in ML pipelines is not richness but **redundancy**.
- Fields with duplicative intent dilute signal for ML without adding informational value.

❑ **THE BALANCING ACT:** Thoughtful field consolidation preserves meaning while improving model fitness.

Every form design decision has downstream consequences

- A crash record that was not captured cannot be reconstructed.
- Poorly structured fields cannot be reliably cleaned and introduced subjectivity in preparations for ML works.
- Decisions made at the reporting form level either open or permanently foreclose analytical possibilities.
- A forward-looking database must be designed with its end use, including machine learning, explicitly in mind from the outset.

Smarter Collection: Standardization, Automation & GIS Integration

Practical design principles that improve data quality without increasing officer burden

01 Standardize Core Fields

Adopt **uniform drop-down selection** across all jurisdictions for crash type, contributing factors, road condition, and injury severity. Free-text in high-signal fields compounds into thousands of inconsistencies. Standardization is the highest-leverage single intervention available.

02 Capture ML-Relevant Metrics

Future reporting forms should deliberately **include variables with predictive value**: precise conflict point, pre-crash maneuver, infrastructure condition rating at site, and richer temporal context.

03 Enforce Standards Where It Matters Most

Not all fields carry equal weight. Injury severity, contributing factor, and precise location are critical. Tiered enforcement, **including mandatory review for high-severity incidents**, protects the fields that drive the most analytical value.

AUTOMATE & SIMPLIFY: Example Implementations

📍 Location Auto-Population

GPS coordinates cross-referenced with road centerline GIS: address, route, and milepost, speed limit, signal types, etc. populate automatically. Officer verifies rather than types. Reduces location errors at source.

🛣️ GIS Pre-Fill & Field Graying

Speed limit, lane count, functional class, median type, and roadway conditions pre-load from existing GIS layers. **Fields gray out** unless officer actively disputes; only deviations require input.

🔀 Conditional Field Logic

Pedestrian and cyclist **fields appear only when crash type requires them.** School zone fields activate only when location falls within a designated zone. Cognitive load drops significantly.

🏴 Pre-Submission Cross-Validation

Form flags internally inconsistent entries before submission & especially for late submitted records and manual entry fields: for example, recorded travel speed inconsistent with road type, or injury level mismatched with crash severity code. Catches errors at source.

Data Quality as a Long-Term Public Safety Investment

A crash record cannot be retroactively corrected: quality at the point of collection is effectively permanent. This makes upstream data investment uniquely high-return.

Targeted Enforcement

Consistently collected data enables deployment of resources to specific road segment types, conditions, and time windows where serious injury risk is demonstrably highest, shifting from reactive response to proactive positioning.

Better data → sharper risk mapping → more effective resource allocation

Factor-Specific Countermeasures

Isolating the contribution of individual variables (speed, impairment, geometry, time) on comparable roadway segments enables countermeasures designed for the actual cause. Similar segments become a significant study population, not just context.

- ❑ Can be effective in studying combination of factors that result in specific outcomes using ML techniques

Comparable corridors → transferable lessons → evidence-based design

Collective National Intelligence

Uniform standards adopted across jurisdictions transform isolated MPO datasets into nodes of a shared analytical infrastructure. Patterns invisible at the local level emerge at scale. More quality data improves model accuracy non-linearly; the compounding returns can be substantial.

Consistent standards → pooled national data → population-level insight

A strict, sustained commitment to data quality is not an administrative burden; it is a long-term investment in the precision of every safety decision that follows.

Thank You

Questions & Discussion

Yusuf W. Zakaria, Transportation Specialist | WWW Interstate Planning Commission
2026 WV Transportation Planning Conference | June 3, 2026