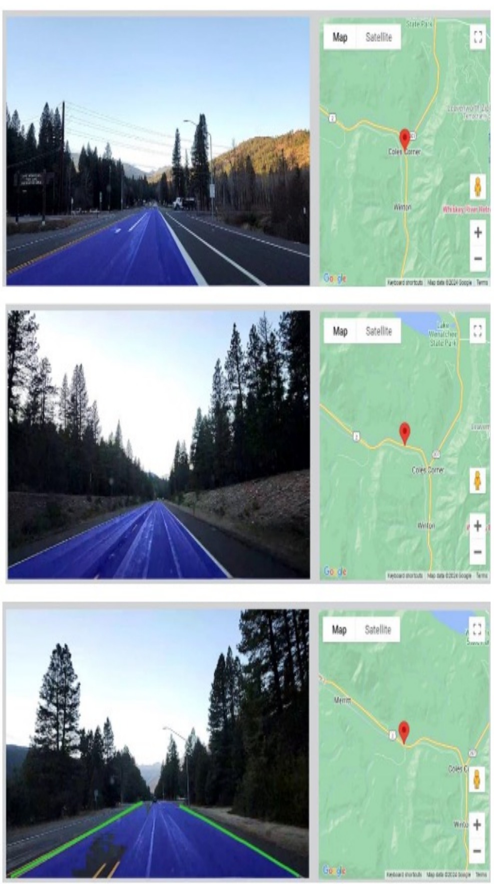
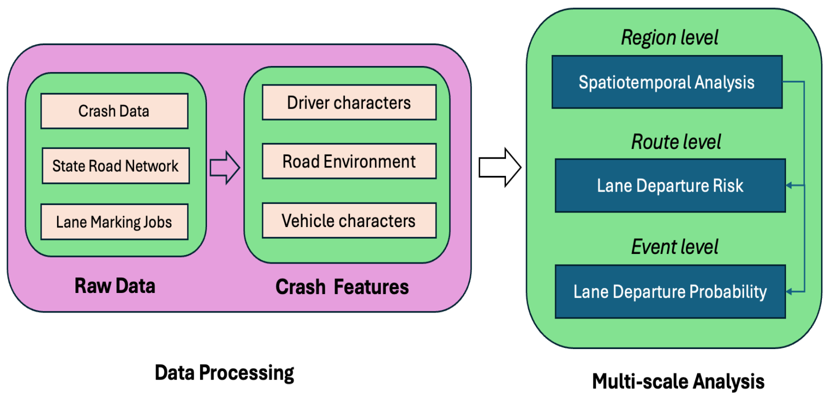
Analysis of Lane Departure Crashes with Emerging Data, Machine Learning and Field Evaluation

WA-RD 947.1

Jia Li

June 2025





WSDOT - Locator Map

Washington State Department of Transportation
Office of Research & Library Services

WSDOT Research Report

Research Report
Agreement T1462-39
WA-RD 947.1

ANALYSIS OF LANE DEPARTURE CRASHES WITH EMERGING DATA, MACHINE LEARNING AND FIELD EVALUATION

By

Jia Li, Ph.D.
Assistant Professor

Department of Civil and Environmental Engineering
Washington State University
Pullman, Washington 99164

Washington State Department of Transportation
Technical Monitors
Scott Davis & Dina Swires
Traffic Operations

Prepared for

The State of Washington
Department of Transportation
Julie Meredith, Secretary

June 2025

1. Report No. WA-RD 947.1	2.Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Analysis of Lane Departure Crashes with Emerging Data, Machine Learning and Field Evaluation		5. Report Date June 2025
		6. Performing Organization Code
7. Author(s) Jia Li		8. Performing Organization Report No.
9. Performing Organization Name and Address Washington State Transportation Center (TRAC) Washington State University Sloan Hall 101 Pullman, Washington 99164-2910		10. Work Unit No.
		11. Contract or Grant No. Agreement T1462 Task 39
12. Sponsoring Agency Name and Address Research Office Washington State Department of Transportation Transportation Building, MS 47372 Olympia, Washington 98504-7372 Project Manager: Doug Brodin		13. Type of Report and Period Covered Final Research Report
		14. Sponsoring Agency Code
15. Supplementary Notes		

16. Abstract

This study develops a technical framework to characterize and quantify the contributing factors of lane departure crashes in the State of Washington. The technical framework consists of machine learning models for route- and event-level lane departure risk prediction leveraging different sources of emerging data. Field study is also conducted to assess lane marking conditions and how well lane detection function works using ADAS (Advanced Driver Assistance System) technology. Using the proposed framework and data from 2020 to 2022, we identify that driver age, driving speed, lighting condition, and surface condition are consistently the top factors contributing to lane departure crashes on state routes in Washington. Field study shows that though weak spots exist, overall lane marking conditions on the surveyed segments are compatible with the lane detection function of ADAS.

17. Key Words Lane departure crash, crash risk, lane marking, ADAS, machine vision data, Washington state		18. Distribution Statement No restrictions.	
19. Security Classif. (of this report) None	20. Security Classif. (of this page) None	21. No of Pages 84	22. Price

Form DOT F 1700.7 (8-69)

Disclaimer

CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

Objectives

The objective of this study is to develop a technical framework to characterize and quantify risk and contributing factors of lane departure crashes. This framework leverages crash, lane marking, and road network data sets in state of Washington and deploys interpretable machine learning techniques to mine information from the datasets and make route- and event-level predictions.

Background

Lane departure is one dominant crash type that leads to serious driver injuries and fatalities in Washington, which accounts for 48.2% fatalities and 37.6% serious injuries in Washington between 2015-2017, representing 9.3% and 5.8% increases compared to 2012-2014. The Washington State Strategic Highway Safety Plan 2019 lists lane departure crashes as one of the two major types of crashes that lead to fatalities and serious injuries in Washington.

To reduce lane departure crashes, a more systematic and detailed analysis of their causes is needed. A detailed assessment of contributing factors to the lane departure crashes will help prioritize the state's investment in safety improvement to ultimately reduce the fatalities and serious injuries on Washington roads. Among others, lane marking is one of the key measures to increase roadway visibility and effectively reduce lane departure crashes. The lack of detailed lane markings and road condition data hinders the systematic understanding of lane departure crashes, and no model has been developed to evaluate the benefits of lane marking and other factors to reduce lane departure crashes.

With Advanced Driver Assistive Systems (ADAS) becoming mature, it is now feasible and cost-effective to collect extensive machine vision (video and Lidar) data under different conditions. This data provides detailed information on driving conditions such as lane markings, road signs, curves, and other roadside objects. In this project, we propose to develop a new technical framework to characterize contributing factors of lane departure crashes using emerging ADAS-based data. The data will be integrated with existing pavement lane marking asset management data and roadway crash data in Washington, and regression models will be developed to identify and quantify the key contributing factors in lane departure crashes.

Research Activities

A technical framework was developed to characterize and quantify the contributing factors of lane departure crashes in state of Washington. The technical framework leverages multiple sources of data and is developed using interpretable machine learning techniques. Machine vision data is collected from the field to assess lane marking conditions using the lane detection function in ADAS systems.

Conclusion

Using the proposed framework and data from 2020 to 2022, we identify that driver age, driving speed, lighting condition, and surface condition are consistently the top factors underlying lane departure crashes on state routes in Washington. Field study shows that overall lane marking conditions on the surveyed segments are compatible with lane detection function using ADAS, though certain callouts are noticed.

INTRODUCTION AND BACKGROUND

Road crashes present multiple threats to public health, economy, and overall quality of life at both national and regional levels. In 2022, the United States recorded 42,514 motor vehicle crash fatalities, equating to 12.8 deaths per 100,000 people. In the U.S., the National Highway Traffic Safety Administration (NHTSA) estimates that road crashes cost the economy $340 billion annually, including medical expenses, lost productivity, and property damage (IIHS, 2022). In 2023, Washington experienced 810 traffic deaths, marking a 10% increase from 2022 and the highest number since 1990 (Washington Traffic Safety Commission, 2024). Besides serious injuries and fatalities, traffic crashes also lead to travel delay and undermine the travel reliability for multimodal transportation, for both travelers and goods. Compared to metropolitans, in the rural areas of Pacific Northwest, the issue is even more significant, due the limited lane number, lack of alternative highways and re-routing options, scarce emergency response resources, severe weather and dark road conditions in winters.

Lane departure is one dominant crash type that leads to serious driver injuries and fatalities in Washington, which accounts for 48.2% fatalities and 37.6% serious injuries in Washington between 2015-2017, representing 9.3% and 5.8% increases compared to 2012-2014. Washington Traffic Safety Commission (2019) lists lane departure crashes as one of the two major types of crashes that lead to fatalities and serious injuries in Washington.

To reduce lane departure crashes, a more systematic and detailed analysis of their causes is needed. A detailed assessment of contributing factors to the lane departure crashes will help prioritize the state's investment in safety improvement to ultimately reduce the fatalities and serious injuries on Washington roads. Among others, lane marking is one of the key measures to increase roadway visibility and effectively reduce lane departure crashes. The lack of detailed lane markings and road condition data hinders the systematic understanding of lane departure crashes, and no model has been developed to evaluate the benefits of lane marking and other factors to reduce lane departure crashes.

Meanwhile, Advanced Driver Assistive Systems (ADAS) is becoming a mature technology that can potentially mitigate lane departure crashes through functions such as lane detection, lane keeping, and lane departure alert. It is now feasible and desirable to collect machine vision data based on ADAS and evaluate the lane detection function against different lane marking conditions. Such data can help to unveil the compatibility of ADAS with existing lane markings and spots that need additional attentions.

LITERATURE REVIEW

Contributing Factors of Lane Departure Crashes

Lane departure crashes, which occur when a vehicle unintentionally veers out of its travel lane, are influenced by a complex interplay of contributing factors in three categories: driver-related factors, road and environment conditions, and vehicle condition and technologies.

1. Driver-Related Factors

Human factors dominate lane departure causation, which mainly include fatigue and drowsiness, distraction, impaired driving, as well as speeding.

- Fatigue and Drowsiness are especially critical risks, particularly on rural highways during nighttime hours. According to the National Highway Traffic Safety Administration (NHTSA, 2017), drowsy driving is responsible for thousands of crashes annually, many of which involve lane departures. AAA Foundation for Traffic Safety (2018) reported that 17.6% of fatal crashes involve drowsy drivers. Drowsy driving contributes to 17.6% of fatal departures, most frequently between 3–5 AM (AAA Foundation, 2018). Commercial drivers face unique risks; Hickman et al. (2020) identified microsleeps in 15% of truck lane departures.
- Distraction is another major concern—whether from mobile phone use, in-vehicle infotainment systems, or external events—which can delay reaction times and compromise lane control (Dingus et al., 2016; NHTSA, 2021).

Distraction is a leading cause of lane departure crashes, with texting increasing lane departure risk by 300% (Klauer et al., 2014).

- Impaired Driving, due to alcohol or drug consumption, also dramatically increases the probability of unintentional lane deviations (Fell & Voas, 2014; Fell et al., 2016). Alcohol impairment remains prevalent in United States. In 2021, 13,384 people were killed in crashes involving alcohol-impaired drivers (Kirley et al., 2023).
- Speeding and Aggressive Driving reduce the driver's margin of error, while inexperience, particularly among younger drivers, is linked to increased crash rates due to poor lane discipline (IIHS, 2023). Speeding increases lane departure likelihood by 2.6 times (IIHS, 2020). Aggressive driving a factor in 54% of fatal crashes. Overcorrection is particularly dangerous; Hallmark et al. (2018) found it triggers 4% of rollovers.

2. Road & Environmental Conditions

- Roadway Design and Conditions play a crucial role. Faded lane markings, narrow lanes, and insufficient shoulder width limit drivers' ability to maintain proper lane positioning. Curves, especially when combined with poor sight distance or high speeds, increase the risk of drifting off the road. Rural roads, which often lack lighting, guardrails, or modern design standards, are disproportionately represented in lane departure fatalities (FHWA, 2020). Work zones, with their temporary and sometimes confusing lane configurations, can further heighten crash risk. A 25% increase in night-time lane departures occurs with faded markings (Craig et al., 2017). Donnell et al. (2016) demonstrated that horizontal curves account for 30% of fatal lane departures. Faded lane markings raise crash likelihood by 25% at night (Craig et al., 2017).

- Environmental Conditions can exacerbate driving risk. Low visibility from fog, heavy rain, or nighttime driving can obscure lane markings and reduce a driver's ability to judge distances and lateral position. FHWA (2016) estimated about half of traffic fatalities occur at night. Slippery road surfaces—due to snow, ice, or standing water—reduce tire traction and vehicle stability. Additionally, glare from the sun or other vehicles' headlights can impair visual perception, contributing to unintended departures. Weather impacts are substantial; Andrey (2010) found wet pavement increases crash rates by 34%, while snow elevates lane departure risk by 52%.

3. Vehicle Condition and Technologies

Vehicle technologies and mechanical failures also contribute substantially to lane departure crashes. Mechanical failures, especially involving steering or braking systems, can cause sudden loss of control. Tire issues, such as blowouts or uneven wear, may induce unexpected veering.

- ADAS system: Advanced driver-assistance systems (ADAS) functions, such as lane departure warnings (LDW) and lane keeping assist (LKA) can increase opportunities for early correction (NHTSA, 2019). Research shows that vehicles equipped with these systems have significantly lower rates of single-vehicle, sideswipe, and head-on crashes (Cicchino, 2017). Jermakian (2011) showed Lane

Keeping Assist reduces departures by 11%. Recent IIHS (2022) data confirms 21% fewer lane departures with warning systems. Lane Departure Warning (LDW) reduces crashes by 21% (Cicchino, 2017).

Preventive Measures

Preventive measures such as rumble strips, reflective pavement markings, and ADAS technologies have been shown to reduce the incidence of such crashes. In below (see Table 1), we list a few preventive measures can be adopted to reduce lane departure risks.

Causal Factor	**Example of Preventive Measures**
Distracted Driving	"Smartphone Ban" by Washington State Legislation
Speeding	Driver education programs
Surface and Lighting Condition	3rd Amendment to 2009 MUTCD that added the requirements for retroreflective paint and maintenance planning.
Driver Fatigue & Drowsiness, Impaired Driving	ADAS Technologies, including Lane Departure Warning (LDW) and Lane Keeping Assist (LKA)

Table 1. Example Preventive Measures to Mitigate Lane Departure Crashes

Many of the above-mentioned factors causing lane departure crashes interact in complicated ways that may amplify risk. For instance, a fatigued or impaired driver navigating a poorly marked, curved rural road at night during inclement weather is far more likely to depart from their lane. Understanding and mitigating the confounding effect of different factors is still an open research problem.

LANE DEPARTURE RISK MODELING

Overview

In this project, we aim to analyze the causal factors and their levels of significance in lane departure crashes leveraging multiple sources of data. To achieve this goal, the technical framework is described as follows (refer to Figure 1). First, the data from multiple sources are collected. These data are then processed to retrieve the features of individual crashes as well as the associated driver and road environment characters. Then the analysis is conducted at three scales, namely region, route, and event level, to reveal the spatial and temporal patterns of lane departure crashes, lane departure tendency, and lane departure risks. We introduce the SHAP analysis to identify the significant factors leading to lane departure crashes in state of Washington.

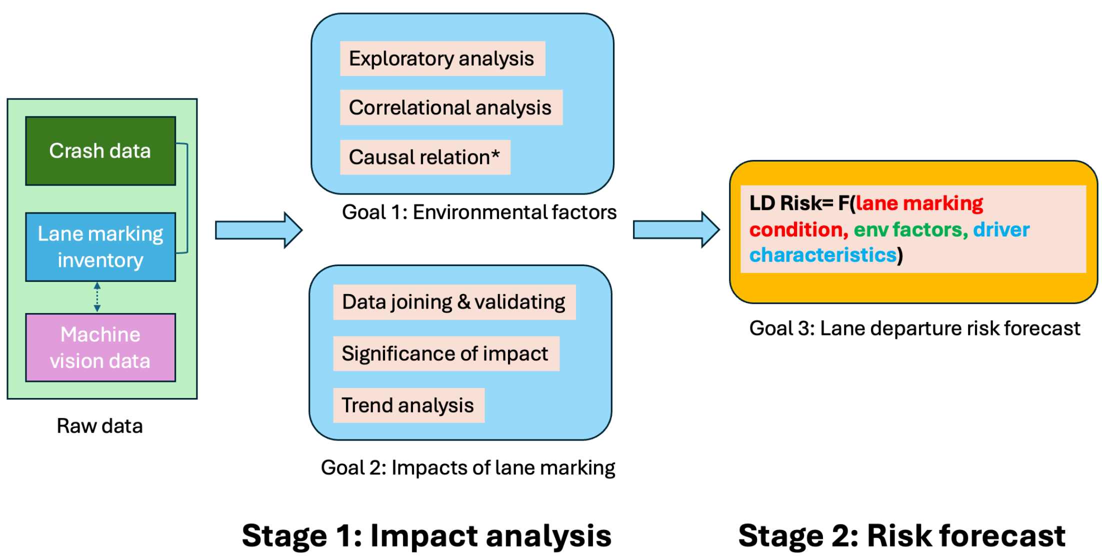


Figure 1. Project Overview

Data Sources and Processing

This project used the following raw data.

1. Crash Data: Crash data was acquired from WSDOT through the Crash data portal. The data consists of descriptions of crash events on road systems in state of Washington based on police reports. At the time of this project, the requested data covers crashes in calendar years of 2014 until 2022, and partial data in 2023. The data consists of 253 columns (variables), covering characteristics of drivers, vehicles, and driving condition (e.g. lighting and surface condition) when crashes occurred. The data contains around 100,000 crash records for each year.
2. State Routes: Route data was acquired from Washington State Geospatial Open Data Portal (https://geo.wa.gov). State route characteristics, such as milepost and route length, are retrieved from this data set. The data set was last updated in May 2025. The route data is used for locating lane marking job locations and crash locations and calculating crash rates (per mile) and lane marking materials used per unit distance.
3. Lane Marking Data: Lane marking data was acquired from Spec-Rite pavement marking management system (https://spec-rite.io). Lane marking data covers statewide lane marking maintenance job time and locations, as well as corresponding lane marking materials usage. This dataset consists of a total of 23 variables and around 4300 rows, covering all WSDOT lane marking job activities on state routes dated back to 2019.
4. ADAS Computer Vision Data: Computer vision data was collected using dedicated cameras by VSI Labs on selected routes in Washington. The survey

was conducted in November 2023 on 7 segments on state routes. VSI provided lane detection analysis using proprietary algorithms. Detailed description of the data collection steps and findings are presented in forthcoming chapter.

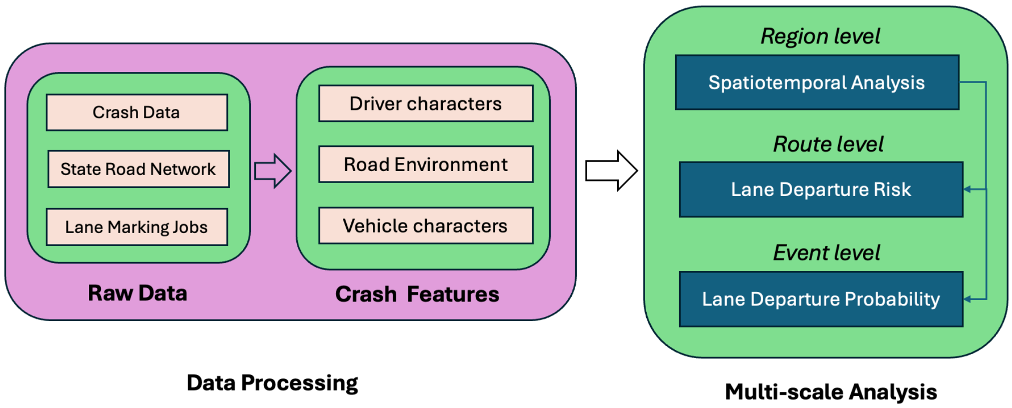


Figure 2. Data Processing Workflow

Figure 2 depicts the workflow of processing the data. The crash, route, and lane marking data were jointly analyzed to uncover characteristics of lane departures at regional, route, and event levels. This process involves data cleaning, matching lane marking data to state routes, and alignment and grouping of the data. The detailed procedure can be referred to code in the appendix. In the following analysis, crash data from 2020 to 2022 was used due to availability.

After examining the above datasets, we find the lane marking and crash data overlap in 2020 through 2022. Therefore, our model selects these three years' data for training and testing. Other years' data are used in exploratory analysis to characterize trend and spatial patterns.

Exploratory Analysis

We first conduct exploratory analysis of the data to reveal lane departure crash patterns. Overall, we found lane departure crashes have distinct spatial and temporal patterns compared to non-lane departure crashes. The initial analysis also indicates that lane departure crashes are related to lane markings.

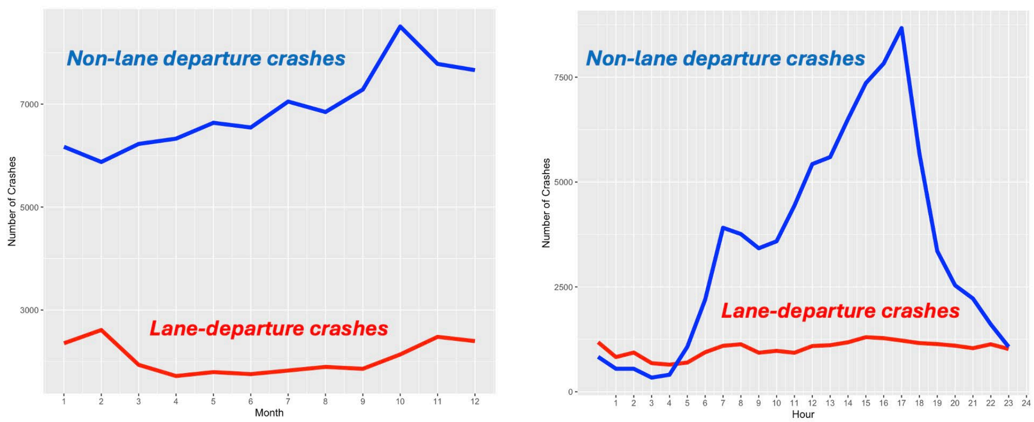


Figure 3. Temporal Pattern of Non-Lane Departure and Lane Departure Crashes

In Figure 3, we compare the temporal distribution of non-lane departure and lane departure crashes, namely variations over different months and time of day, using the data of Year 2014 as one example. The two types of crashes show distinct patterns. In particular, we find the lane departure crashes occurred more in winter months (November to February). In addition, frequency of lane departure crashes is relatively independent of the time of day compared with non-lane departure crashes.

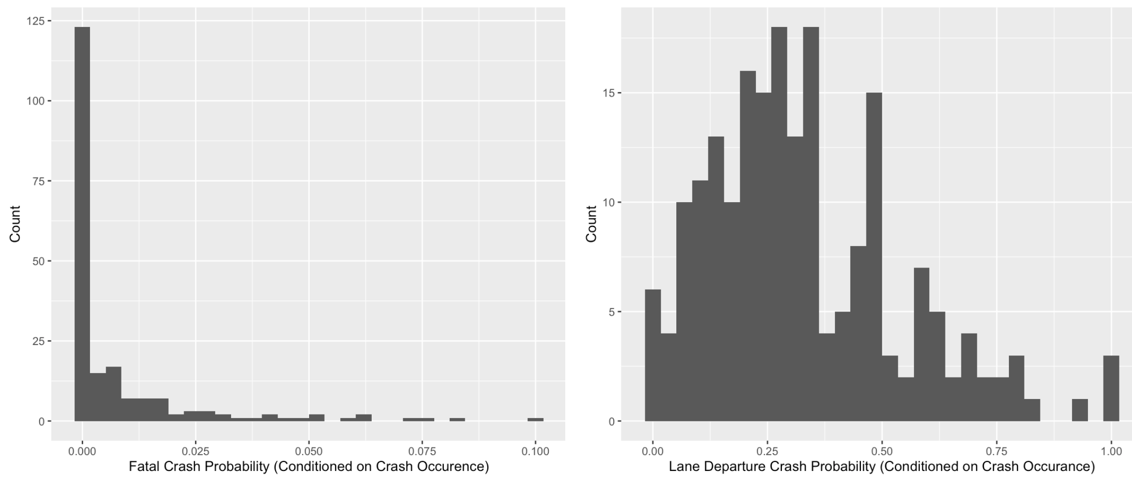


Figure 4. Distribution of Fatal and Lane Departure Crashes on State routes

In Figure 4, we compare the distribution of fatal and lane departure crash frequencies on state routes. The distributions suggest that fatal crashes are less likely to occur on most routes, and lane departure crashes are distributed on evenly on all state routes. In another word, fatal crashes only likely to occur on a small number of routes. Most routes have zero or few fatal crashes. In contrast, lane departure crashes are more evenly distributed over different routes.

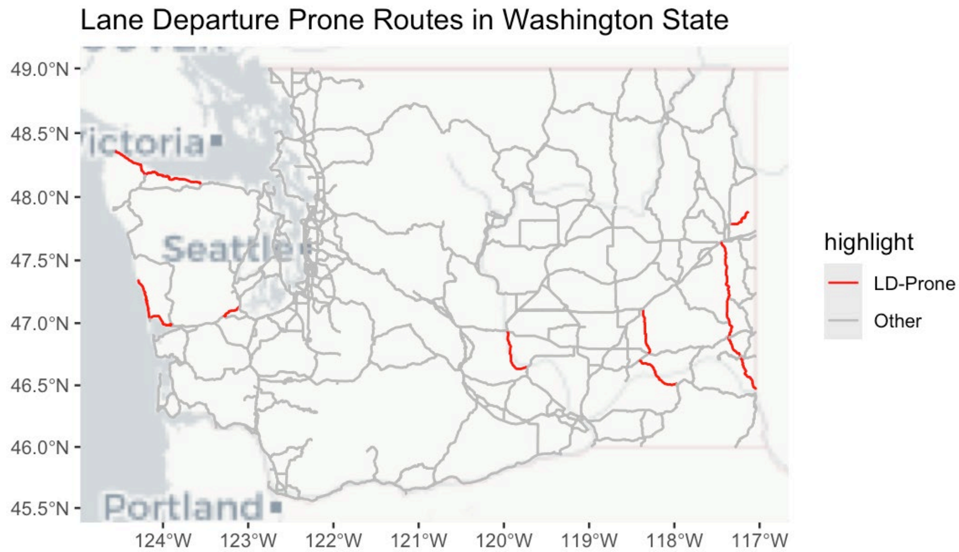


Figure 5. Lane Departure Prone Routes in Washington

In Figure 5, we show the routes with the highest probability of lane departure crashes, conditioned that a crash occurs. Roughly, those are the routes with the highest percentage of lane departure crashes among all crashes. To capture this performance, we adopt the following criteria: 1) the probability of lane departure crashes conditioned on all crashes is larger than 0.5; 2) the probability of fatal crashes is larger than 0; and 3) the number of lane departure crashes is larger than 5. From this Figure, it is found most of the lane departure prone routes are in rural areas.

Next, we examine the spatial and temporal pattern of lane marking jobs. Figure 6 shows the job starting locations (red points) of lane marking maintenance activities in State of Washington since 2019, where the lines show directions to the job completion locations. Figure 7 shows the number of lane marking job activities in different months. It is found that most lane marking jobs were conducted between May and October.

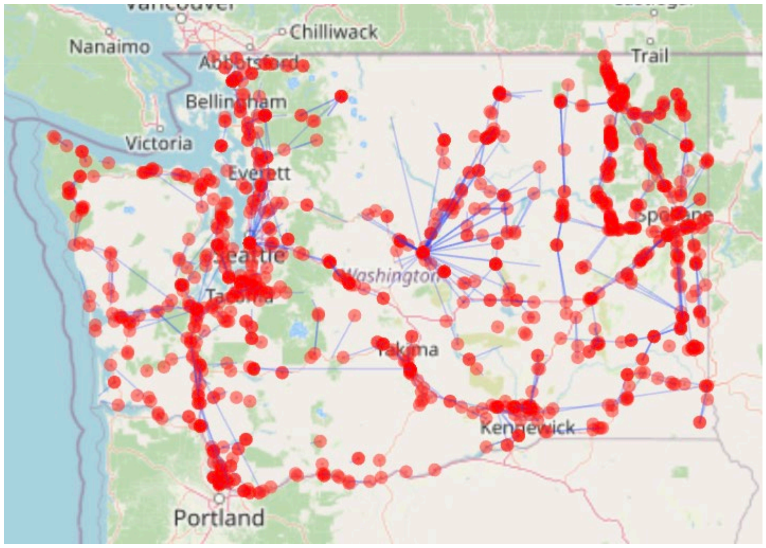


Figure 6. Lane Marking Job Locations from 2019 to 2024

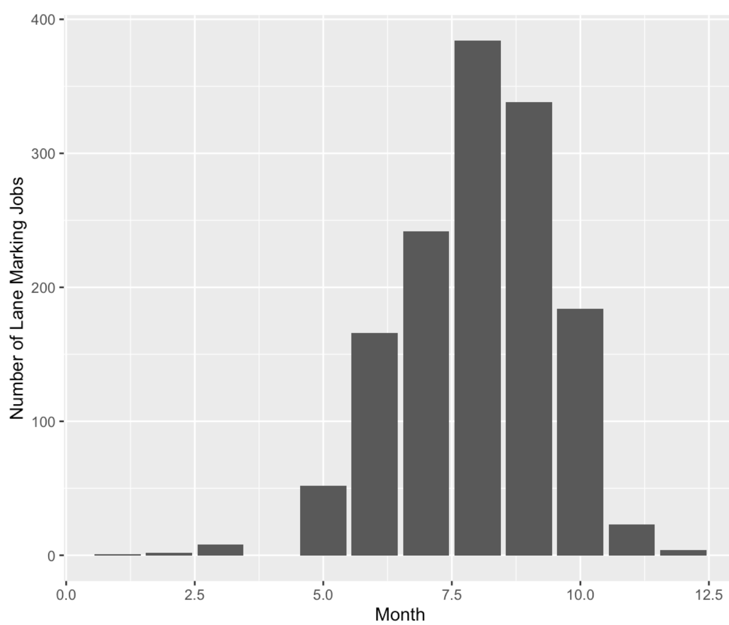


Figure 7. Lane Marking Job Activities in Different Months

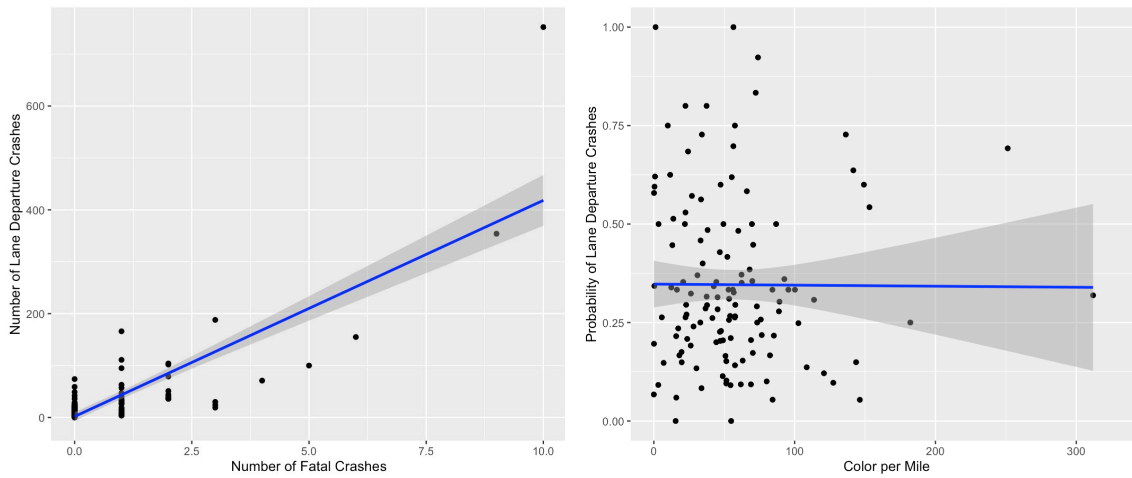


Figure 8. Relation between Lane Departure Crash and Number of Fatality and Lane Marking Materials Used

Finally, we explore the relation between lane departure crashes and fatality and lane marking materials used per mile on a route. The scatter plots in Figure 8 indicate a positive relationship between the number of lane departure crashes, while the probability of lane departure crashes are independent of lane marking activities (measured by paint used per mile). The exact impact of lane marking materials used on the probability of lane departure crashes cannot be told from the plot. Therefore, a more in-depth modeling that takes other factors into account is necessary.

Route-level Lane Departure Risk Modeling

In this section, we aim to give a more accurate characterization of the lane departure risk and its relationship with observable factors. For this purpose, we propose a lane departure tendency indicator for a route as follows,

$$y_i = \frac{\#\ lane\ departure\ crashes\ on\ route\ i}{\#\ total\ crashes\ on\ route\ i}$$

Because y_i is always in the interval $[0,1]$, we employ the beta regression to model the lane departure tendency as a function of explanatory variables. Beta regression is a specialized statistical model designed for analyzing continuous dependent variables that are bounded between 0 and 1, such as Proportions (e.g., 0.2, 0.95), Percentages (scaled to [0, 1]), and Rates (e.g., response rates, market shares). Unlike linear regression (which assumes unbounded Y), beta regression accounts for the unique properties of bounded data using the beta distribution, which is flexible in modeling skewness and heteroscedasticity. Specifically, the proposed model reads,

$$Y_i \sim Beta(\mu_i, \phi)$$

where $Beta(\cdot,\cdot)$ is a Beta function, and μ_i and ϕ are respectively mean and precision parameters. The route-specific parameter μ_i is a function of observable variables from the data through a link function $g(\cdot)$,

$$\mu_i = g(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)$$

The unknown coefficients $(\beta_1, \dots, \beta_n)$ are estimated from sample observations of variables $(x_1, \dots, x_n)$. The variables can either be numeric or categorical. When categorical variables are involved, they are encoded as integers.

Based on the exploratory analysis and initial evidence from data, we propose a Beta regression model of lane departure tendency using the following predictors for each route:

- Weather
- Surface Condition
- Lighting Condition

- Lane Marking Usage (color per mile)
- Probability of Fatal Crashes

Among these variables, those categorical variables (such as weather, surface condition, and lighting condition) are encoded at integers. Other data preconditioning and details of the model calibration and validation can by referred to the R code in the appendix.

Event-level Lane Departure Risk Modeling

Due to its aggregate nature, route-level lane departure risk model cannot consider driver and vehicle characters. Due to this reason, we propose an event-level model that predict if a crash event is lane departure or not, based on characteristics of driver, vehicle, driving environment, and route characteristics.

To achieve the goal, we employ a machine learning model called XGBoost (Extreme Gradient Boosting). XGBoost is a scalable machine learning algorithm based on gradient-boosted decision trees. To predict if a crash is a lane departure one, the problem is reduced to a classification problem from observations. We experimented different combinations of features and considered domain knowledge. The following eight predictors were chosen to make the predictions,

- Vehicle and drivers: driver age, posted speed of vehicles
- Road environment: lighting condition, road surface condition, weather
- Lane marking condition: color per mile
- Time: month of year
- Route characteristics: probability of fatal crashes

SHAP Analysis

As machine learning (ML) models become increasingly complex (e.g., deep neural networks, ensemble methods), their "black-box" nature poses challenges for interpretability. SHAP (SHapley Additive exPlanations), introduced by Lundberg and Lee (2017), is a unified framework for interpreting the output of machine learning models. It is based on the concept of Shapley values from cooperative game theory, which were originally designed to fairly distribute payouts among players depending on their contributions to the total outcome. In the context of machine learning, the “game” is the model's prediction, and the “players” are the input features. SHAP estimates the contribution of each feature by considering all possible combinations of features and calculating how the prediction changes when a feature is included or excluded. The result is a SHAP value for each feature, quantifying how much that feature pushed the model’s prediction up or down relative to a baseline (typically the average prediction).

One of SHAP’s key strengths is that it provides consistent and locally accurate explanations, meaning that if a model changes such that a feature has more impact on the output, its SHAP value will also increase. SHAP can be used with a wide variety of models and offers both local explanations (why a model made a specific prediction) and global insights (overall feature importance across all predictions). This makes it a powerful and interpretable tool for understanding complex machine learning models, particularly in high-stakes domains like healthcare, finance, and transportation.

FINDINGS AND DISCUSSION

Prediction Performance

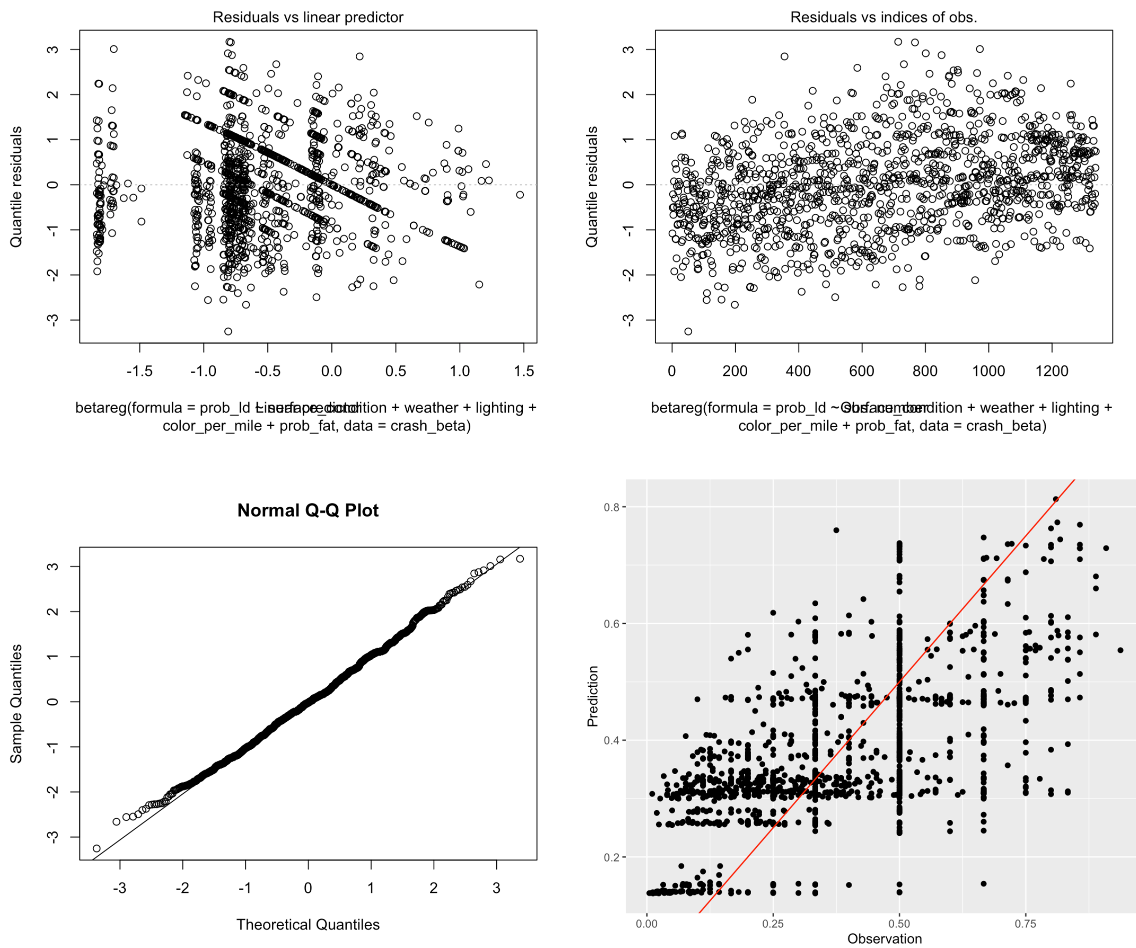


Figure 9. Performance of Event-level Lane Departure Risk Forecast

First, we show verification of the model using four visualization techniques, respectively quantile residuals versus a linear prediction, magnitude of lane departure crashes and their relation to indices of observations, and comparison of sample and theoretical quantiles. The results are shown in Figure 9. Overall, the results indicate the goodness-of-fit of the model is acceptable.

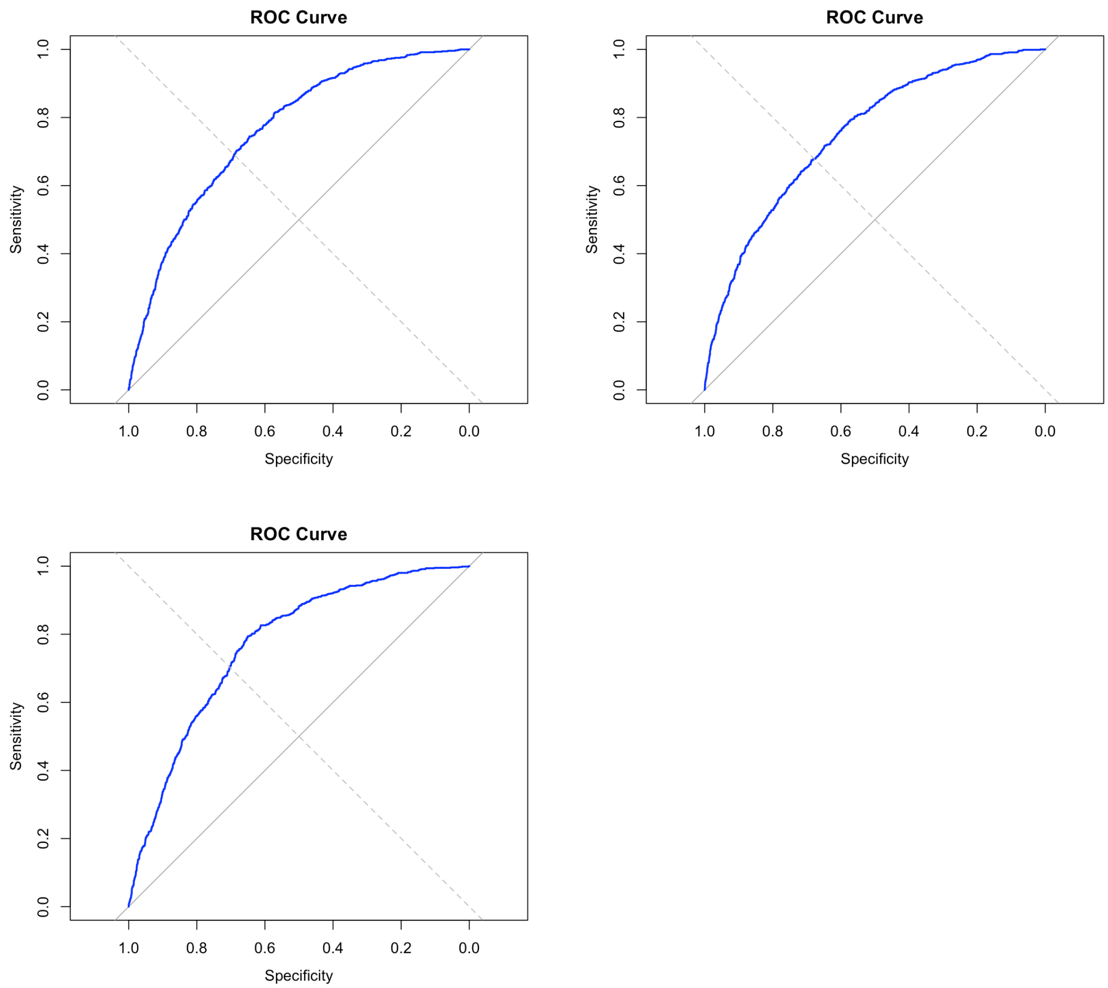


Figure 10. ROC Curve of the Proposed Model

Figure 10 shows the ROC Curves (Receiver Operating Characteristic Curve) of the proposed model for data in Year 2020, 2021, and 2022. The ROC Curve is a graphical tool used to evaluate the performance of a binary classifier across all possible classification thresholds. It shows how well the model separates the two classes across all thresholds, rather than just at a single cutoff. AUC-ROC (Area Under the ROC Curve) in those plots quantifies overall performance of the model. From 2020 to 2022, AUC-ROC values are respectively 0.7626, 0.7518, 0.7694, which are close to 1 that represents perfect classification. The results in Figure 11 indicates the proposed model performs reasonably well to differentiate lane departure crashes from non-lane departure crashes, and the performance is consistent over different years' data.

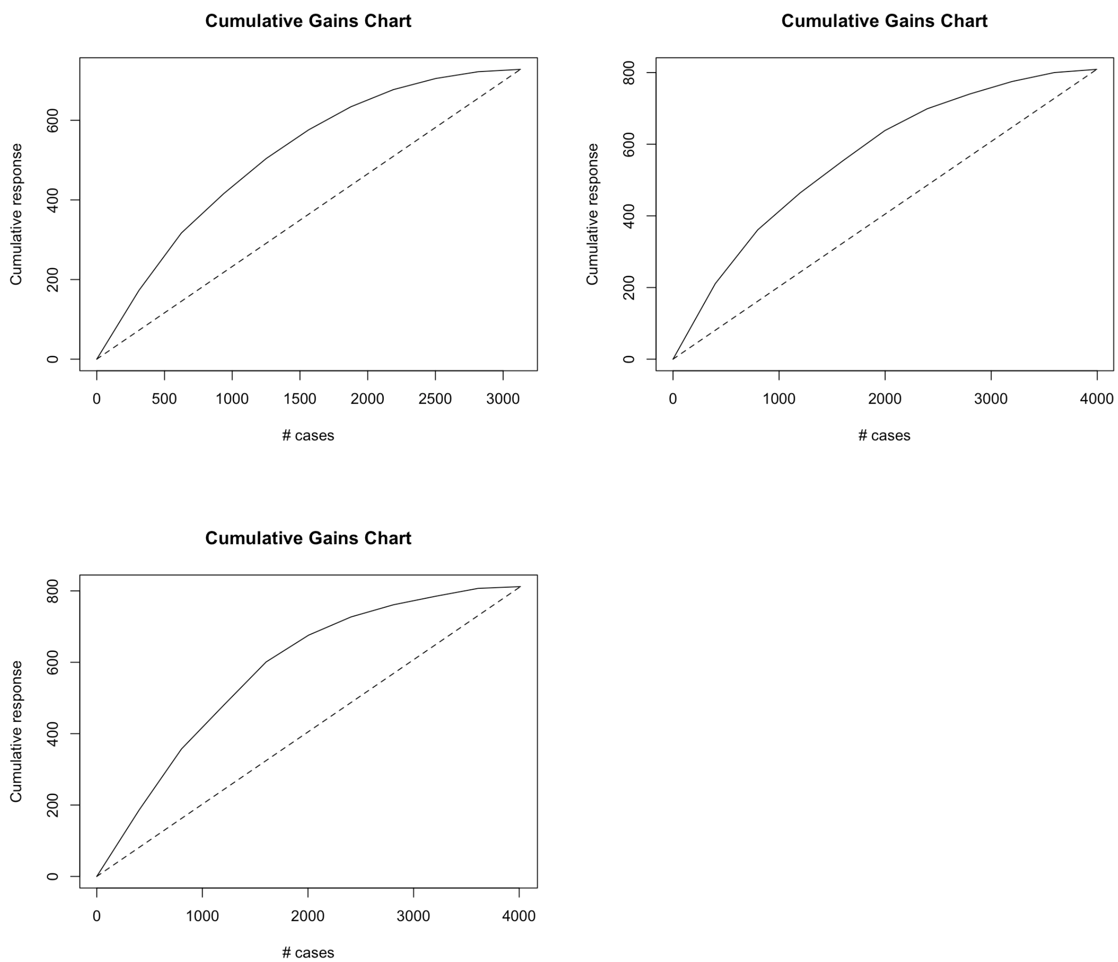


Figure 11. Cumulative Gains Chart

Figure 11 shows the cumulative gains chart for Year 2020 to 2022. The Cumulative Gains Chart (or Gains Curve) is a visualization tool used to assess the effectiveness of a binary classification model by showing how much better it performs at identifying positive cases compared to random selection. In this study, the cumulative gains chart tells how well the event-based model identifies lane departure crashes from the samples. The diagonal line represents a baseline where a model performs random selection. The separation of the curve from the diagonal line in plots indicate the proposed model consistently perform well in identifying lane departure risks based on the selected features.

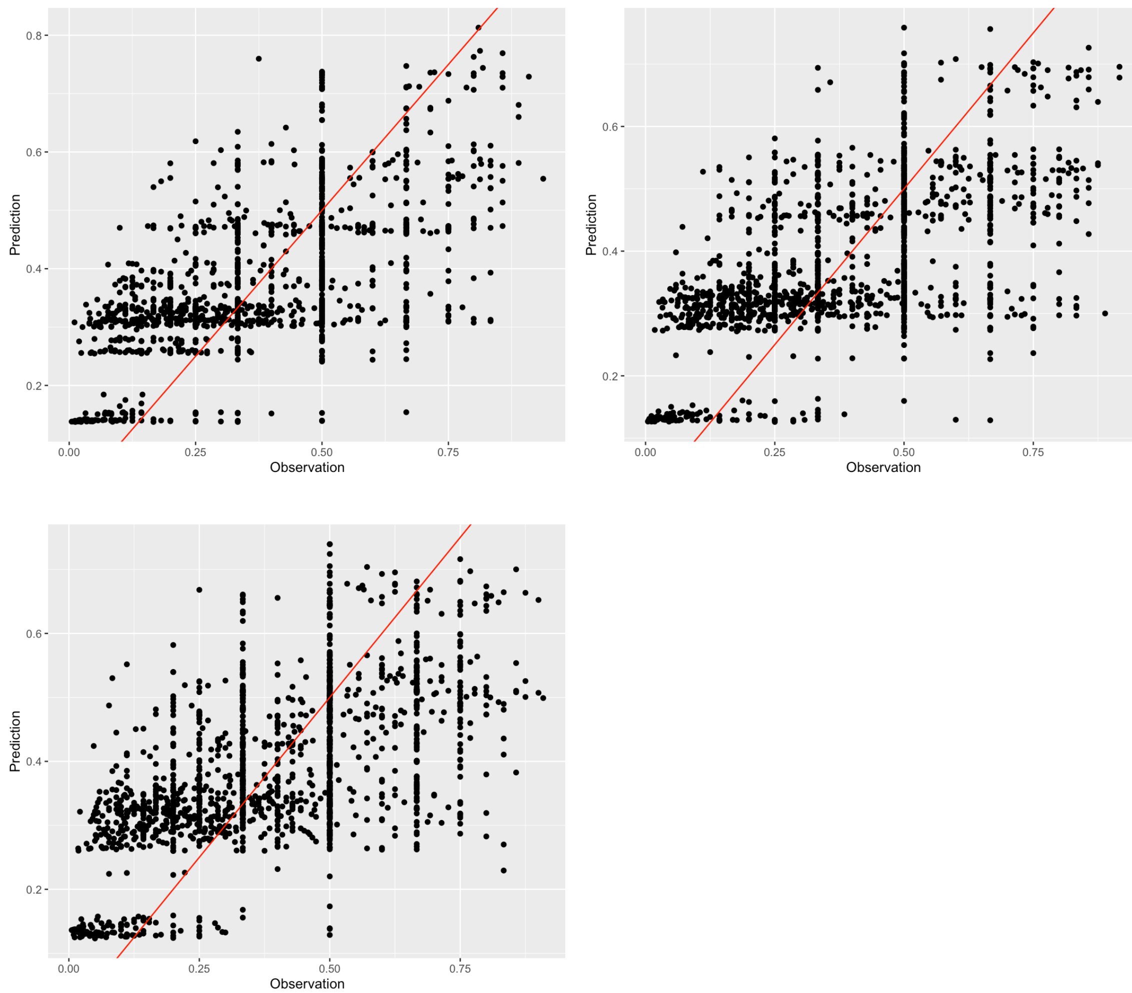
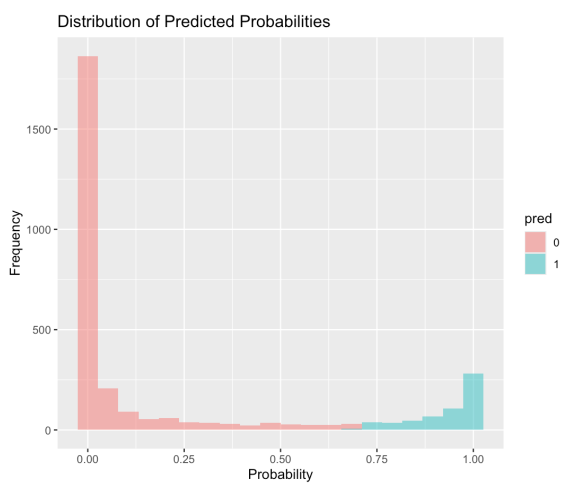


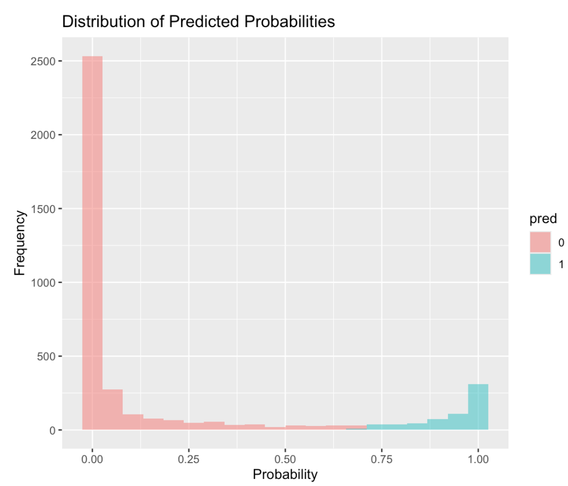
Figure 12. Prediction vs. Observation in Route-based Model

Figure 12 shows the performance of the route-based model to predict lane departure probability, where each point represents an observation-prediction pair. The red line represents the perfect prediction, i.e. prediction is equal to observation. The model predictions are distributed around this line, indicating the unbiasedness of the predictions. Table 2 further compares the model performance using MSE (Mean Square Error) and MAE (Mean Absolute Error) as metrics. The model has similar performance over three years' data.

	Mean Square Error (MSE)	Mean Absolute Error (MAE)
2020	0.026	0.131
2021	0.028	0.135
2022	0.027	0.133

Table 2. Performance of Route-based Prediction





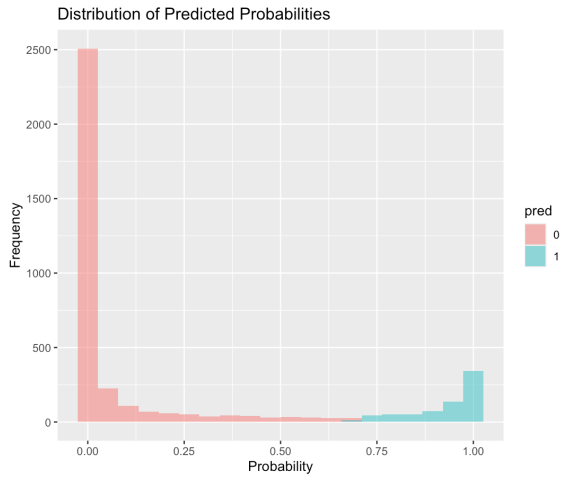


Figure 13. Predicted Probabilities and Corresponding Classification of Crash Types

Figure 13 depicts the outcome of the model on the test data for each year. It is observed that the predicted probabilities are distributed bimodally for all the years, and the lane departure crashes (green color) and non-lane departure crashes (red color) are clearly delineated from each other, with negligible overlap with each other. This is another evidence that the event-based model can effectively separate lane departure crashes from the other samples.

Finally, Figure 14 shows the confusion matrices for the three years' data. The performance is similar to each other. Overall prediction accuracy of the event-based model is around 80%.

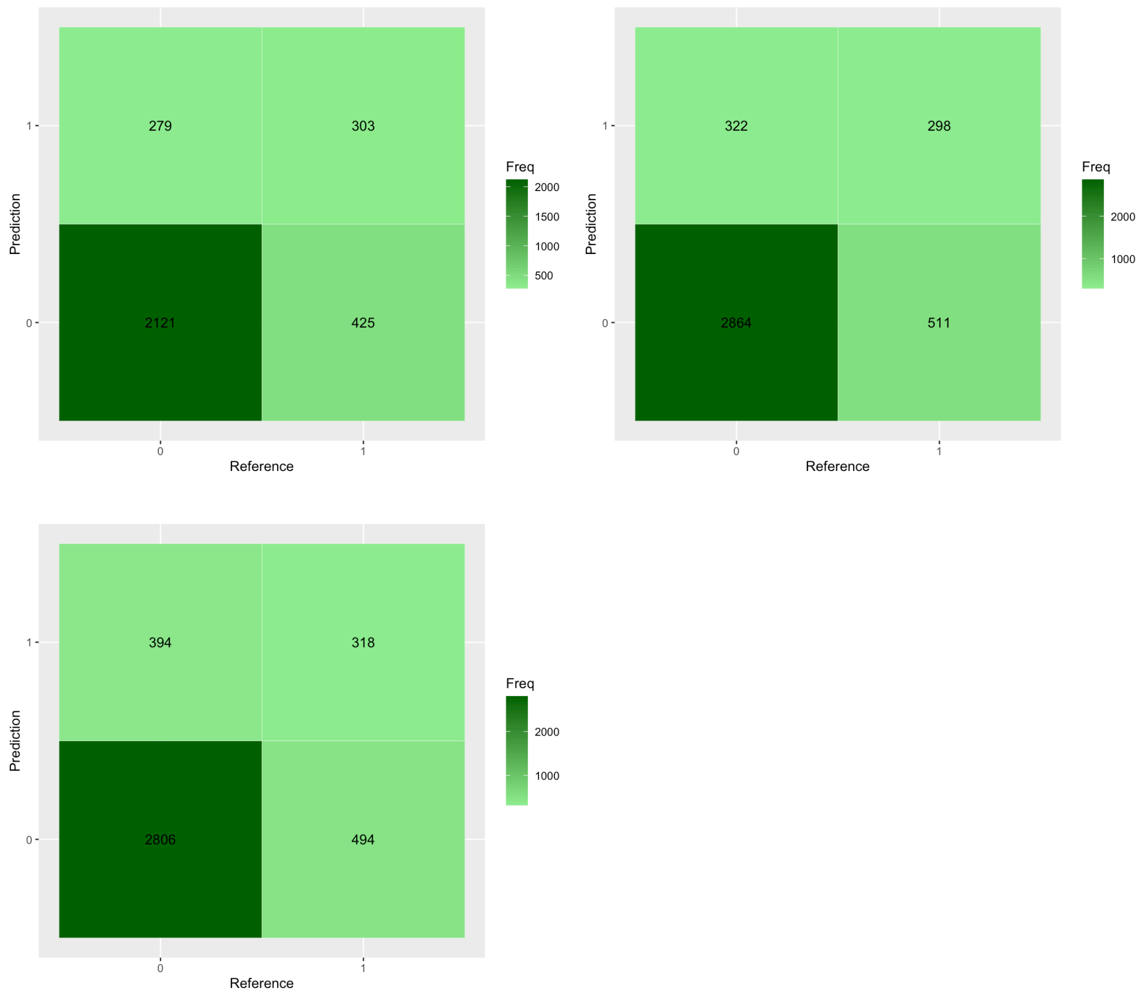


Figure 14. Confusion Matrices for the Predictions

Feature Significance

We apply the SHAP analysis to identify significant factors contributing to lane departure crashes. SHAP is a unified framework for interpreting machine learning models by quantifying the contribution of each feature to individual predictions. Rooted in game theory (Shapley values), it provides both local (per-instance) and global (model-wide) interpretability.

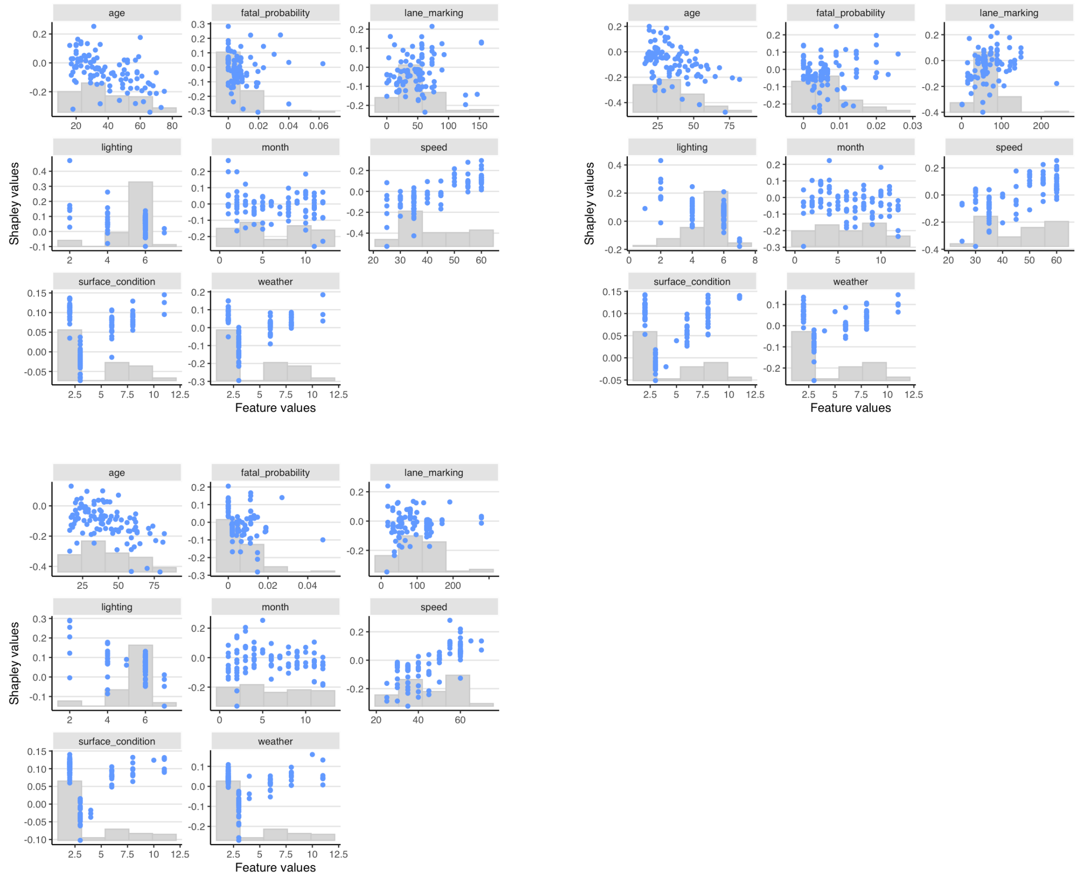


Figure 15. SHAP Values for All Features

Figure 15 shows SHAP values for all features across different feature values. Each subplot shows the SHAP value against the values of eight selected features. Consistent patterns of SHAP values can be observed over the multiple years of data. This indicates the stability of results when the time changes. This is a desired property when we want to infer consistent causal relationships from the data.

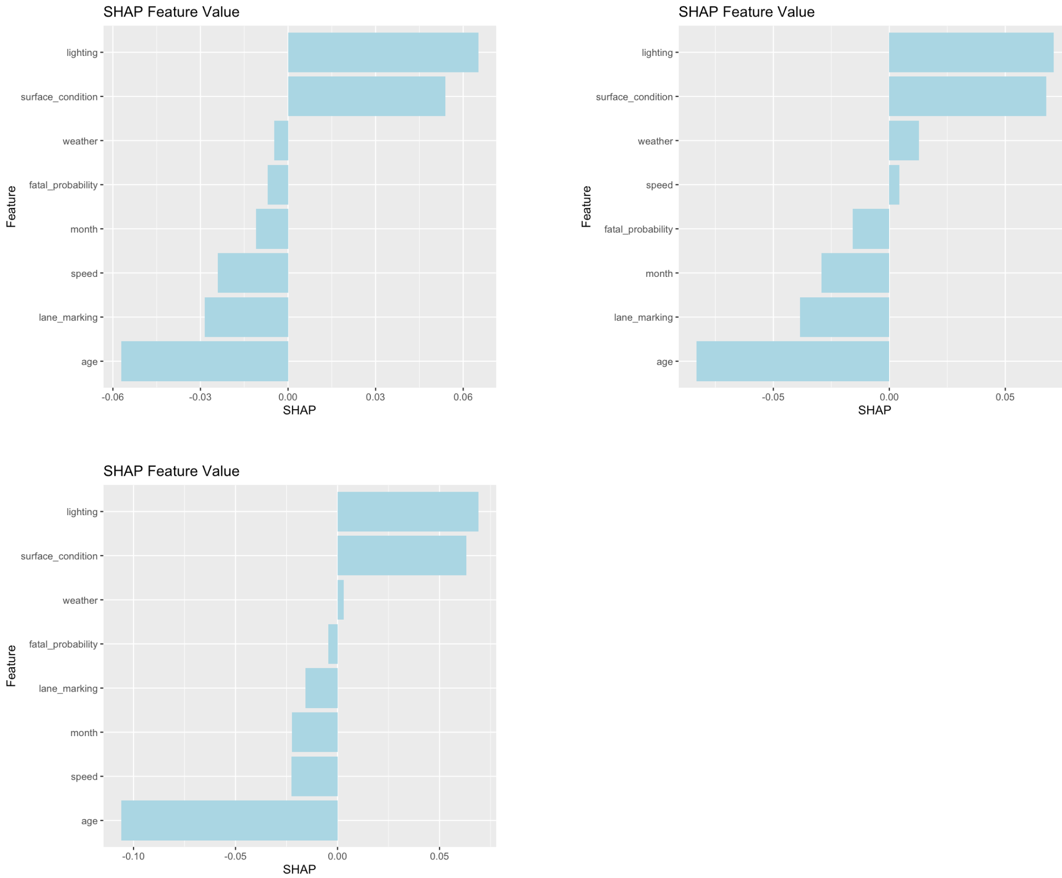


Figure 16. SHAP Values for All Features

Figure 16 shows average SHAP values for all features (predictors) of the model. It is found that lighting condition, surface condition, and weather constantly have positive average SHAP values. Driver age and color per mile constantly have negative average SHAP values. The consistency over different data confirms robustness of the result. The mean absolute SHAP values are further calculated and ranked, and the result is presented in Figure 17. It is found driver age, driving speed, and lighting condition constantly rank high. Color per mile and surface condition are also relatively significant factors contributing to explaining the lane departure crashes.

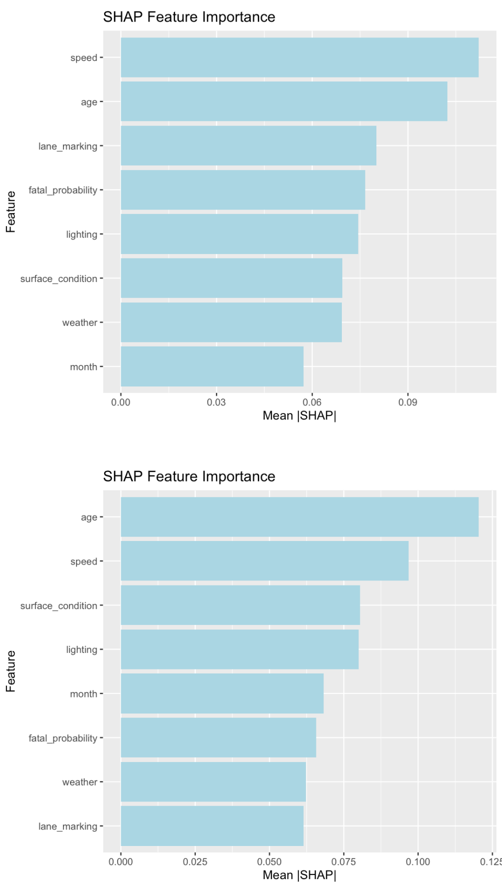
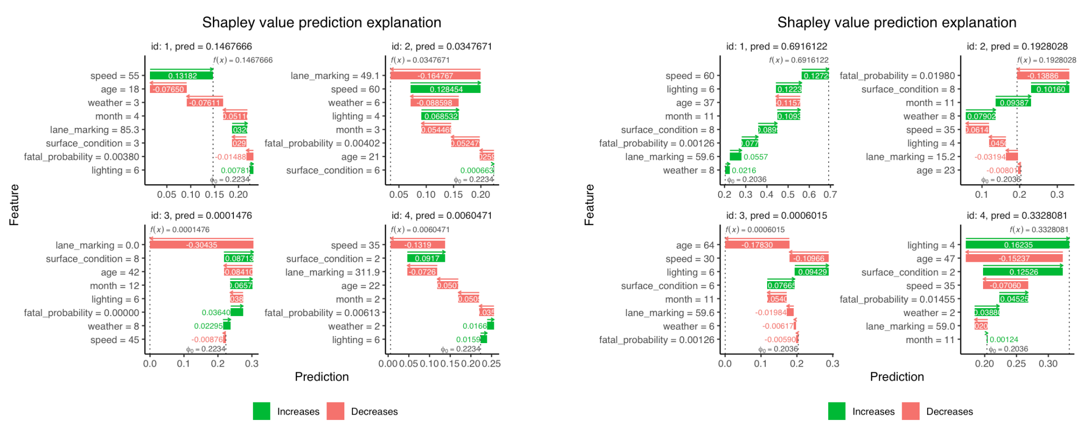


Figure 17. Feature Importance for all Features



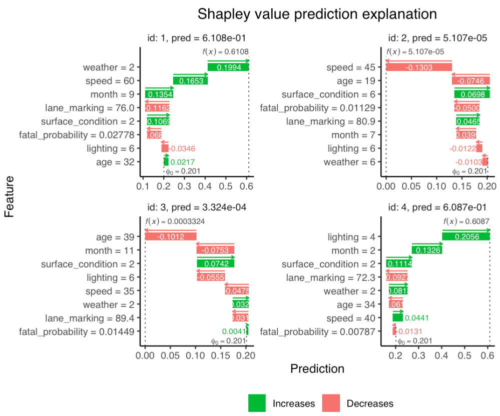
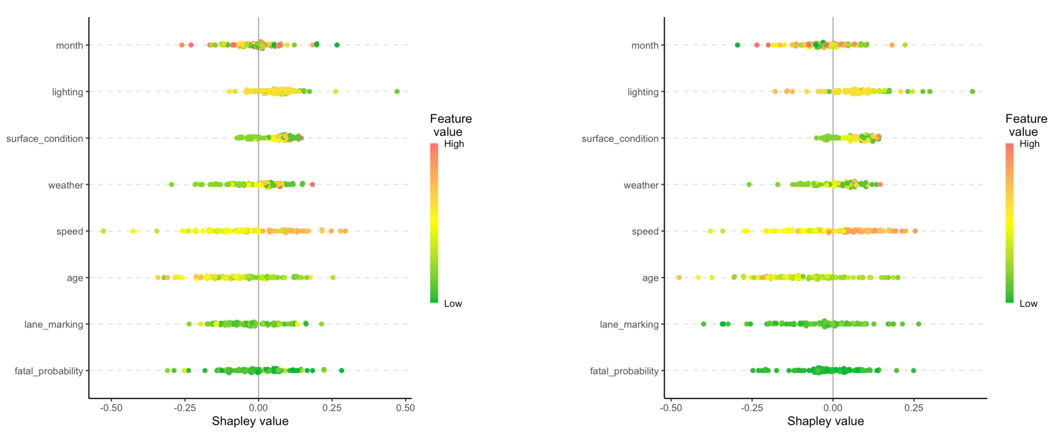


Figure 18. SHAP Values for Sample Cases

In Figure 18, SHAP values for four random sample cases are shown. For these specific cases, the contributions of different factors can be different. In each case, the model predicts a different lane departure probability due to the different values of driver, road environment, and lane marking condition values. To how the SHAP values distribute over a large sample, we randomly select 100 samples for each year and examine the SHAP values versus feature values. Figure 19 presents the result. Similar overall patterns are observed, which further confirms the stability of SHAP analysis over different years' data.



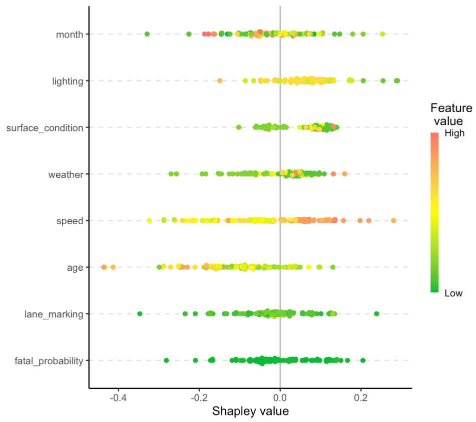
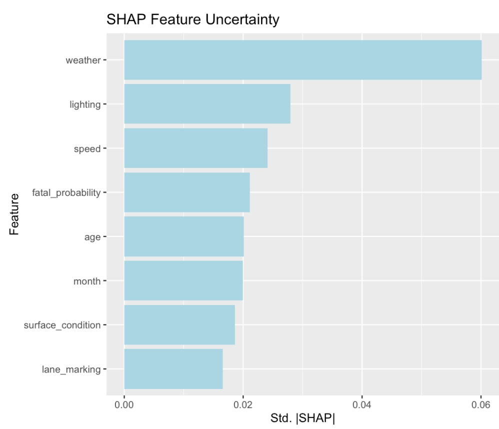
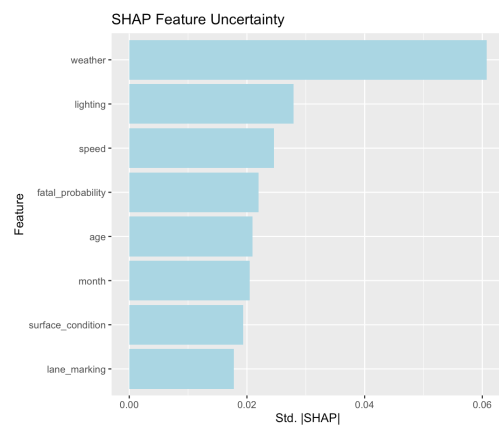


Figure 19. SHAP Values for 100 Random Samples

Finally, we examine uncertainties underlying each factor as shown in Figure 20. The three years' data reveal similar results. Weather, lighting condition, and vehicle speed are most uncertain factors impacting the occurrence of lane departure crashes. On the other hand, color per mile, surface condition, and month of year are the most certain factors impacting the lane departure crashes.





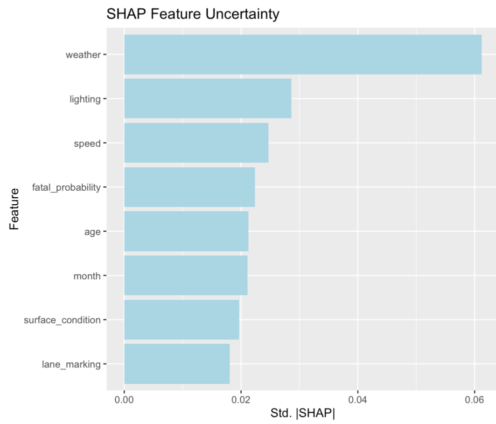


Figure 20. SHAP Feature Uncertainties

Model Limitations

The current analysis is limited by several factors related to data, and some assumptions are made. The current model has the following limitations,

- Lane marking data lacks accurate route information, and current analysis is based on an estimate using map matching.
- Lane marking data is only available since 2019. This limits the analysis of the impacts of lane marking for a longer period, e.g. five years or ten years.
- In this analysis, we do not have data to characterize current lane marking condition. As a surrogate measure, we assume the lane marking activities have impacts on current and next-year's crashes. This assumption may be refined in future work.
- In current analysis, the lane marking activities are still aggregated at route level and this may introduce inaccuracies when the route length is long, because the lane marking activities can be non-uniform. It is desirable to refine the spatial resolution from route to segment in future studies.

FIELD EVALUATION OF ADAS-BASED DETECTION

Machine vision data was collected by VSI labs to understand conditions of lane markings on selected segments and whether these segments satisfy the need of operating ADAS (Advanced Driver Assistance System) equipped vehicles. Following is the details on collection of machine vision data and key outcomes.

Data Collection Process

The machine vision data survey was conducted by the VSI labs in the week of November 6th. The survey machine is a mobile unit installed on rental vehicle, which consists of:

- Two HDR GigE visible cameras, one mounted inside and one mounted outside.
- One GPS
- Computer system and data logging tool

The survey was conducted one-way that best fitted the logistics. The routes covered in this survey are as follows.

- Segment 1: Discovery Bay to Olympia (US 101) – 86 miles
- Segment 2: Olympia to Elma (SR 8) – 29 Miles
- Segment 3: New London to McGowan (US 101) – 83 Miles
- Segment 4: Everett to Wenatchee (US 2) – 170 Miles
- Segment 5: Wenatchee to Stratford (SR 28) – 64 Miles
- Segment 6: Mead to Newport (US 2) – 47 Miles

The segments were shown in the following map (Figure 21).

Figure 21. Routes for Machine Vision Data Collection with ADAS

The survey is intended to detect lane marking conditions in various scenarios for ADAS systems. The scenarios under consideration and potential lane marking conditions are listed in Table 3. We note that the survey is intended to identify potential issues of lane detection for ADAS systems. However, due to the limited temporal and spatial coverage and main focus on normal driving conditions, it does not provide direct evidence of factors leading to historical lane departure crashes.

Scenarios

- Merge/Blend Area Gap
- Tunnel/Overpass
- Curvature Situation
- Entering Urban
- Traffic Jam Scenario
- Workzone – Temporary Markings
- Exit Gap (non-ramp)
- Ghost Markings/Scars

Markings:

- Positive Call-Out
- High Contrast Markings
- Marker Misdetection
- Missing Markers
- Off-Ramp Item (blend area)
- On-Ramp Item (blend area)
- Poor Centerline
- Poor Contrast (light surface)
- Poor Contrast (polished surface)
- Poor Contrast (wet surface)
- Faded Markings

Table 3. Lane Marking Detection Scenarios in Field Evaluation

Outcome and Finding

The raw data, including the collected videos, images, and analysis results, were provided by VSI labs through https://analytics-washingtonstate.vsi-labs.com/ (accessed in February, 2024). A screenshot of the analysis interface is shown below (Figure 22).

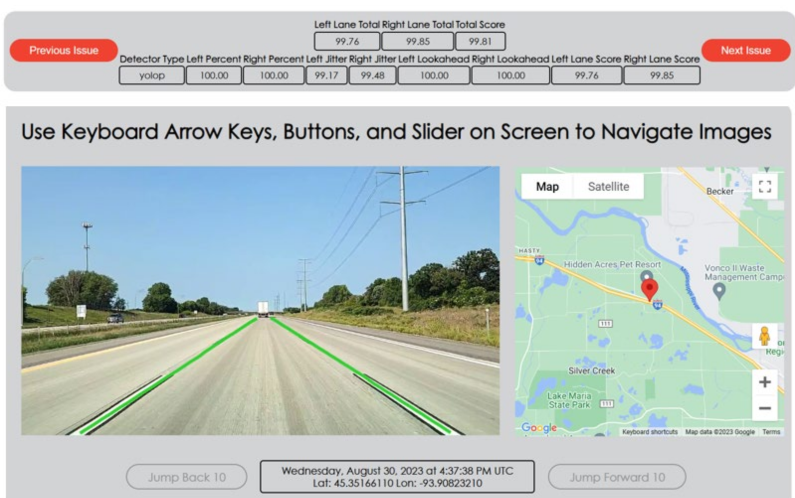


Figure 22. Snapshot of Lane Marking Analysis Interface

Route	Start and Ending Location	Miles	# Call-out	#/Mile	LD Prob (2022)
US 101	Olympia to Discovery Bay	86	74	0.860	0.354
SR 8	Olympia to Elma	29	26	0.897	0.667
US 101	McGowan to New London	83	39	0.470	0.354
US 2	Wenatchee to Everett	170	54	0.318	0.147
SR 28	Stratford to Wenatchee	64	40	0.625	0.049
US 2	Newport to Mead	47	37	0.787	0.147

Table 4. Events Logged in the Survey

Table 4 summarizes the number of call-out items and notable issues for all the segments. We also add the lane departure probability to the last column for comparison. A call-out item generally means a lane marking condition that causes lane detection failure and needs attention, such as faded centerline. From the result, it is found the average number of call-outs for each segment varies, ranging from 0.318 (US 2, Wenatchee to Everett) to 0.897 (SR 8, Elma to Olympia).

We further compare the callout rate (defined as number of callouts per mile) with lane departure probability on all the segments (the probability is calculated using 2022 data). Figure 23 shows the relation between the two, where the blue line is a fit to the relation. It is observed that the call-out rates are positively related to lane departure probability of a segment, which indicates lane marking condition contributes to lane departure crashes. The correlation between the two is 0.501 in current study.

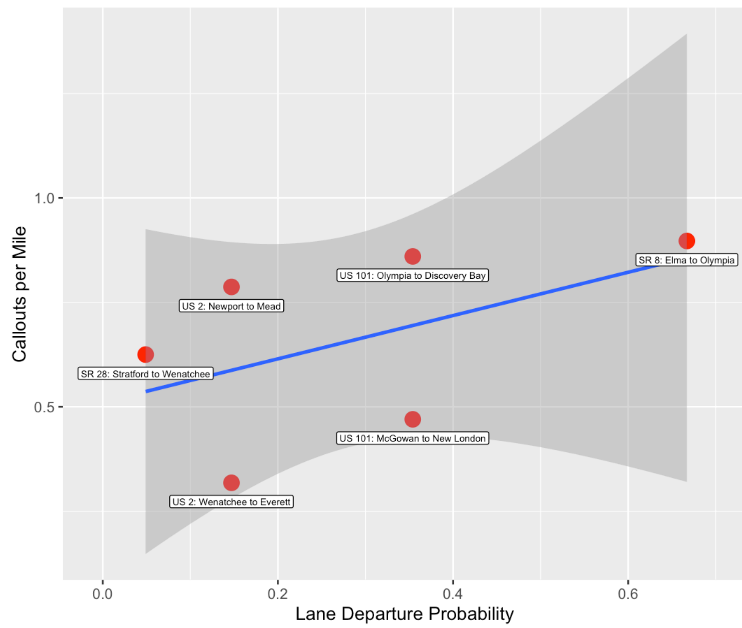


Figure 23. Relation between Call-out Rate and Lane Departure Probability

We then check the call-out record and identify the top factors that lead to compromised ADAS lane detection for each segment. The result is summarized in Table 5. Unmarked off-ramps and faded and missing lane marking are among the top factors leading to compromised lane detection in this study.

Segment	**Top call-outs**	**Call-out ranking**
US 101, Discovery Bay to Olympia	Unmarked off-ramp, Faded centerline	2
SR 8, Olympia to Elma	Unmarked off-ramp, Weather	1
US 101, McGowan to New London	Faded centerline, Curvature, Glare	5

US 2, Wenatchee to Everett	Left-turn exit, entering and exiting town	6
SR 28, Stratford to Wenatchee	Intersection gap, Left turn exit	4
US 2, Newport to Mead	Faded and missing marking	3

Table 5. Top Call-outs Compromising Lane Detection with ADAS

Conclusion

Based on the above results, we can draw the following conclusions:

- The lane detection function of ADAS is often compromised when dashed lane marking are missing at on-ramps or off-ramps, or right lane and left lane exits on 2-lane roads. ADAS lane detection is also compromised when the centerline or edge lines are faded out. Refer to Figure 24 for examples.
- All the surveyed segments are considered to be safe for the operation of assisted and automated driving features. However, continuous autonomous operation would require human attention and interventions.

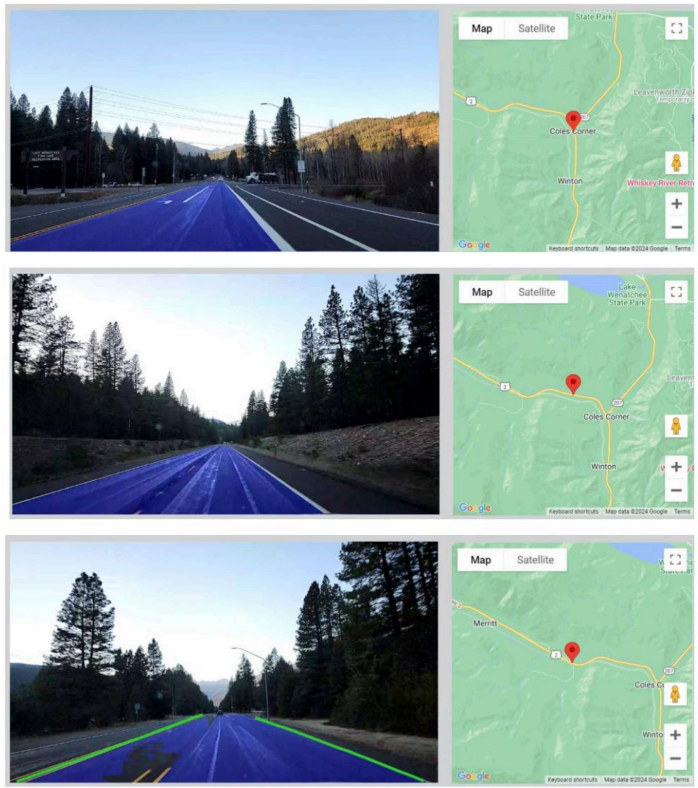


Figure 24. Examples of Compromised ADAS Lane Detections

Potential Applications of Computer Vision

Lighting and lane marking conditions are important contributing factors of lane departure crashes (see Figure 17). This field study confirms the overall good lane marking condition on the surveyed segments for lane detection function in ADAS. This indicates the potential of ADAS to enhance the perception of human drivers and reduce lane departure risks. Along this line, it would be useful to combine machine vision data with historical crash data to predict lane departure risks in real time, which can enable a safety alert system leveraging vehicle-to-everything (V2X) technologies. Furthermore, machine vision data from a single ADAS vehicle can be integrated with data from infrastructure-based sensors. Figure 25 shows a sample snapshot from the current traffic monitoring system in Washington. Video streams from multiple cameras can be jointly used to establish a collaborative perception system for road safety.

Figure 25. Road Camera Detection in Current Traffic Monitoring System

CONCLUSION AND FUTURE WORK

Summary of Findings

A technical framework was developed to characterize and quantify the contributing factors of lane departure crashes in state of Washington. The technical framework consists of two key models and leverages multiple sources of data and interpretable machine learning techniques. Machine vision data is collected from the field to demonstrate feasibility of collecting leveraging ADAS technology. The potential of incorporating machine vision data to enhance lane departure risk forecast was discussed.

Driver age, driving speed, lighting condition, and surface condition are identified to be the top factors leading to lane departure crashes. This finding is consistent across the three years' analysis period.

Policy Implications

While this study primarily focuses on the characterization and modeling of lane departure crashes, the findings provide insights into policymaking to enhance road safety. It may be anticipated that those significant contributing factors call for corresponding safety countermeasures, such as young driver education, speed limit enforcement, improved lighting, and adoption of new vehicle-infrastructure detection and communication technologies. Detailed discussions of the policy implications based on the current findings are desired in future works.

Future Work

This study paves the way for identifying significant factors leading to lane departure crashes using various sources of emerging data and machine learning techniques. In future work, we envision the following directions are worth exploring.

1. Improved model granularity: while this study focuses on the difference between lane departure crashes compared to non-lane departure crashes, it is worth investigating the lane departure crashes that involve fatal/serious injuries and differentiating them with lane departure crashes with only property damages. This is due to the potential differential causal factors in these crashes. Such refined model granularity is desired for both route-level and event-level models.

2. Model extension with other data: This study leverages crash, route, and lane marking data to predict lane departure risks. We envision that the technical framework can be extended to handle other data and improve prediction accuracy of lane departure crashes at both route and event level. The data include but are not limited to traffic volume and composition, road geometry (e.g. curvature and number of lanes), pavement performance, roadside signs and safety countermeasures, etc.

3. Leveraging machine vision data for real-time prediction: this study evaluates the lane detection function with current ADAS system in the field. Due to problem of data alignment, the machine vision data was not used in the prediction of lane departure risks. In future work, it is worth to investigate if machine vision data can be combined with historical data and used to predict lane departure risks in real time. Such a model would help to develop a safety alert system to reduce human errors, and injuries and fatalities leveraging vehicle-to-everything (V2X) technologies.

ACKNOWLEDGMENTS

The author would like to thank WSDOT for providing financial support and access to data for this study. The author would like to thank WSDOT subject matter experts Scott Davis and Dina Swires for the discussion and detailed technical guidance. The author also wants to express sincere gratitude to the Project Manager Doug Brodin for his strong and effective support throughout the project.

REFERENCES

AAA Foundation for Traffic Safety. (2018). *Prevalence of drowsy driving crashes: Estimates from a large-scale naturalistic driving study* (Report No. 18-024). https://aaafoundation.org/wp-content/uploads/2018/10/Prevalence-of-Drowsy-Driving-Crashes-Full-Report.pdf

Andrey, J. (2010). Long-term trends in weather-related crash risks. *Journal of Transport Geography, 18*(5), 619-631. https://doi.org/10.1016/j.jtrangeo.2009.10.005

Cicchino, J. B. (2017). Effectiveness of forward collision warning and autonomous emergency braking systems in reducing front-to-rear crash rates. *Accident Analysis & Prevention*, 99, 142–152. https://doi.org/10.1016/j.aap.2016.11.009

Cicchino, J. B. (2017). Effectiveness of lane departure warning systems. *Accident Analysis & Prevention, 102*, 121-128. https://doi.org/10.1016/j.aap.2017.02.016

Craig, C. M., Huffman, S., & Knox, D. (2017). Nighttime visibility of pavement markings: Effects of retroreflectivity on driver behavior. *Transportation Research Record, 2637*(1), 42-50. https://doi.org/10.3141/2637-06

Dingus, T. A., Guo, F., Lee, S., et al. (2016). *Driver crash risk factors and prevalence evaluation using naturalistic driving data*. Proceedings of the National Academy of Sciences, 113(10), 2636–2641. https://doi.org/10.1073/pnas.1513271113

Donnell, E. T., Porter, R. J., & Shankar, V. (2016). Crash frequency analysis of horizontal curves on rural two-lane highways. *Journal of Safety Research, 57*, 105-113. https://doi.org/10.1016/j.jsr.2016.05.001

Federal Highway Administration (FHWA). (2020). *Rural Road Safety Overview*. https://highways.dot.gov/safety/rural/overview

Fell, J. C., & Voas, R. B. (2014). *The effectiveness of a 0.05 blood alcohol concentration (BAC) limit for driving in the United States*. Addiction, 109(6), 869–874. https://doi.org/10.1111/add.12458

Fell, J. C., Scherer, M., & Thomas, S. (2016). Effectiveness of enforcement programs for impaired driving. *Traffic Injury Prevention, 17*(3), 306-311. https://doi.org/10.1080/15389588.2015.1087643

Hallmark, S. L., Tyner, S., & Oneyear, N. (2018). Driver overcorrection in run-off-road crashes. *Journal of Transportation Safety & Security, 10*(1), 1-15. https://doi.org/10.1080/19439962.2017.1315612

Hickman, J. S., Hanowski, R. J., & Bocanegra, J. (2020). Distraction and drowsiness in commercial motor vehicle drivers. *Transportation Research Part F: Traffic Psychology and Behaviour, 68*, 403-418. https://doi.org/10.1016/j.trf.2019.12.001

Insurance Institute for Highway Safety (IIHS). (2020). *Speeding: Speeding facts* [Fact sheet]. https://www.iihs.org/topics/speeding

for Highway Safety (IIHS). (2022). Fatality Fact. Retrieved from https://www.iihs.org/topics/fatality-statistics/detail/state-by-state

Insurance Institute for Highway Safety (IIHS). (2022). *Lane departure warning lowers crash rates* [Press release]. https://www.iihs.org/news/detail/lane-departure-warning-lowers-crash-rates

Insurance Institute for Highway Safety (IIHS). (2023). *Fatality Facts 2021: Teenagers*. https://www.iihs.org/topics/fatality-statistics/detail/teenagers

Jermakian, J. S. (2011). Crash avoidance potential of four passenger vehicle technologies. *Accident Analysis & Prevention, 43*(1), 326-336. https://doi.org/10.1016/j.aap.2010.09.003

Kirley, B., Robison, K., Goodwin, A., Harmon, K. J., O’Brien, N. P., West, A., ... & Brookshire, K. (2023). *Countermeasures that work: A highway safety countermeasure guide for state highway safety offices, 2023* (No. DOT HS 813 490). United States.

Department of Transportation. National Highway Traffic Safety Administration. Office of Behavioral Safety Research.

Klauer, S. G., Guo, F., Simons-Morton, B. G., Ouimet, M. C., Lee, S. E., & Dingus, T. A. (2014). Distracted driving and risk of road crashes among novice and experienced drivers. *New England Journal of Medicine, 370*(1), 54-59. https://doi.org/10.1056/NEJMsa1204142

National Highway Traffic Safety Administration (NHTSA). (2017). *Drowsy Driving Research and Program Plan*. https://www.nhtsa.gov/sites/nhtsa.gov/files/documents/812446-drowsy-driving-research-program-plan.pdf

National Highway Traffic Safety Administration (NHTSA). (2019). *Lane departure warning system effectiveness* (Report No. DOT HS 812 682). https://www.nhtsa.gov/sites/nhtsa.gov/files/documents/ldw_final_report_092018_v5_tag.pdf

National Highway Traffic Safety Administration (NHTSA). (2021). *Distracted driving 2019* (Report No. DOT HS 813 111). https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813111

U.S. Department of Transportation, Federal Highway Administration (FHWA). (2016). *Nighttime visibility of pavement markings* (Report No. FHWA-SA-16-071). https://safety.fhwa.dot.gov/roadway_dept/night_visib/

Washington Traffic Safety Commission. (2019). *Washington State Strategic Highway Safety Plan.*

Washington Traffic Safety Commission. (2024). *Washington Traffic Deaths Reach 33-Year High.* Retrieved from https://wtsc.wa.gov/wtsc-2023-traffic-fatalities

APPENDIX

A1. Acronyms and Definitions

- **ADAS**: Advanced Driver Assistance System
- **FHWA**: Federal Highway Administration
- **IIHS**: Insurance Institute for Highway Safety
- **Lane Departure Warning (LDW)**: Lane Departure Warning (LDW) is a passive safety feature designed to alert the driver when the vehicle unintentionally begins to drift out of its lane without the use of a turn signal. It typically uses cameras to monitor lane markings and provides visual, audible, or tactile warnings—such as a beep, flashing icon, or steering wheel vibration—when lane departure is detected. However, LDW does not take any corrective action to keep the vehicle in its lane; it simply notifies the driver so they can steer back manually.
- **Lane Keeping Assist (LKA)**: Lane Keeping Assist (LKA) goes a step further by actively intervening to help keep the vehicle within its lane. Like LDW, it uses cameras to detect lane markings, but in addition to issuing warnings, LKA applies gentle steering inputs or braking on specific wheels to guide the vehicle back toward the center of the lane. LKA is considered an active safety system and is more effective in preventing lane departure-related crashes than LDW alone.
- **NHTSA**: National Highway Traffic Safety Administration
- **SHAP value**: SHapley Additive exPlanation value

A2. R Code for Lane Departure Crash Modeling

load packages

```
library(readxl)
library(tidyverse)
library(lubridate)
library(leaflet)
library(osmdata)
library(sf)
library(ggspatial)
library(geosphere)
library(betareg)
library(tidymodels)
library(caret)
library(pROC)
library(gains)
library(shapr)
library(leaflet.extras)
library(prettymapr)
library(performance)
library(xgboost)

### initialization

graphics.off()
rm(list = ls())

year_analysis <- 2021 # 2020--2022

set.seed(123)
```

```
### load data

## crash data

crash_file <- paste("Crash_",as.character(year_analysis),".csv",sep="")

crash_raw <- read_csv(crash_file,skip=1,na=c("NA","-","")) # skip the first
line

## route data
route_file <- "WSDOT_-_Functional_Class_Data_for_State_Routes/WSDOT_-
_Functional_Class_Data_for_State_Routes.shp"

state_routes <- st_read(route_file)

## lane marking data

lanemarking <- read_csv("LaneMarking_CSV.csv",skip=1)

### part 0. data preprocessing

# crash data - rename variables
name_set <- as_tibble(colnames(crash_raw)) # all column names

which(name_set$value=="WA STATE PLANE SOUTH - X") # EPSG:2927 for WA South
which(name_set$value=="WA STATE PLANE SOUTH - Y")

names(crash_raw)[c(2,3,15,19,22,29,33,37,40,41,42,47,48,53,104,105,107,207,208
,234)] <-
c("county","city","date","time","severity","total_fatality","total_veh","first
_collision","weather","surface_condition","lighting","first_location","second_
```

```
location","veh1_type","veh1_age","veh1_gender","veh1_cause","plane_x","plane_y
","ld_ind")


names(crash_raw)[which(names(crash_raw)=="MONTH")]<-"month"

names(crash_raw)[which(names(crash_raw)=="QTR #")]<-"qtr"

names(crash_raw)[which(names(crash_raw)=="FULL TIME")]<-"full_time"

names(crash_raw)[which(names(crash_raw)=="VEH 2 MV DRIVER AGE")]<-"veh2_age"

names(crash_raw)[which(names(crash_raw)=="VEH 2 MV DRIVER GENDER")]<-
"veh2_gender"

names(crash_raw)[which(names(crash_raw)=="VEH 1 POSTED SPEED")]<-
"veh1_posted_speed"

names(crash_raw)[which(names(crash_raw)=="VEH 2 POSTED SPEED")]<-
"veh2_posted_speed"


# crash data - add longitude and latitude


state_plane_data <- crash_raw[c("plane_x","plane_y")]


state_latlon <- state_plane_data %>% # add longitude and latitude to crash
data

  st_as_sf(coords = c("plane_x", "plane_y"), crs = 2927) %>%

  st_transform(state_plane_sf, crs = 4326) %>%

  st_coordinates() %>%

  as_tibble() %>%

  rename("longitude"="X","latitude"="Y")


crash_raw <- crash_raw %>%

  cbind(state_latlon) %>%

  select(date,time,longitude,latitude,county,city,`PRIMARY
TRAFFICWAY`,`MILEPOST`,everything()) %>%

  rename("route_name"="PRIMARY TRAFFICWAY")
```

```
# crash data - subset based on jurisdiction

unique(crash_raw$JURISDICTION)

crash_state <- crash_raw %>% filter(JURISDICTION=="State Route")
crash_city <- crash_raw %>% filter(JURISDICTION=="City Street")
crash_county <- crash_raw %>% filter(JURISDICTION=="County Road")
crash_misc <- crash_raw %>% filter(JURISDICTION=="Miscellaneous Trafficway")

dim(crash_state)
dim(crash_city)
dim(crash_county)
dim(crash_misc)

# crash data - crash probability

route_set <- unique(str_sub(crash_state$route_name,1,3)) # the set of all
routes

crash_state <- crash_state %>%
  mutate(route=str_sub(route_name,1,3)) %>%

select(date,time,route,longitude,latitude,,severity,total_veh,weather,surface_
condition,lighting,veh1_type,veh1_age,veh1_gender,veh1_cause,ld_ind,everything
())

crash_prob <- crash_state %>%
  group_by(route) %>%
  summarize(n_rec=n(),n_ld=sum(ld_ind),n_fat=sum(total_fatality)) %>%
  arrange(desc(n_rec)) %>%
  mutate(prob_ld=n_ld/n_rec,prob_fat=n_fat/n_rec)
```

```
# lane marking data - rename variables

col_names <- c("job_name", "route_number", "start_time", "end_time",
"yellow_useage", "yellow_thickness", "yellow_skip_dist", "yellow_solid_dist",
"white_useage", "white_thickness", "white_skip_dist", "white_solid_dist",
"total_bead", "total_bead_rate", "bead_1_rate", "start_lat", "start_long",
"end_lat", "end_long", "air_temp", "road_temp", "humidity", "dew")


lanemarking <- lanemarking %>% set_names(col_names)




# lane marking - add region


lanemarking %>% filter(str_detect(job_name,"Region")) %>% print() # find all
Regions


region_pos <- which(str_detect(lanemarking$job_name,"Region")==1) # positions
of all rows containing regions


region_pos_aug <- c(region_pos,99999)


lanemarking_new <- tibble()


for (i in seq(region_pos)){ # include region as one row


  lanemarking_temp <- lanemarking %>% filter(row_number()>region_pos[i] &
row_number() < region_pos_aug[i+1])

  lanemarking_temp <- lanemarking_temp %>%
mutate(region=lanemarking$job_name[region_pos[i]])


  lanemarking_new <- rbind(lanemarking_new,lanemarking_temp)

}
```

```
lanemarking_new <- lanemarking_new %>% # clean the data

  select(region,everything()) %>%

  mutate(across(seq(6,23),parse_number)) %>%

  mutate(across(seq(4,5),mdy_hm)) %>%

  mutate(year=year(start_time),month=month(start_time)) %>%

  filter(!is.na(start_long) & !is.na(start_lat) & !is.na(end_long)
& !is.na(end_lat)) %>%

  select(year,month,everything()) %>%

  filter(year<=year_analysis & year>=year_analysis-1) # year and jurisdiction



# lane marking - add route information

start_df <- data.frame(

  id=lanemarking_new$job_name,

  longitude=lanemarking_new$start_long,

  latitude=lanemarking_new$start_lat

)


end_df <- data.frame(

  id=lanemarking_new$job_name,

  longitude=lanemarking_new$end_long,

  latitude=lanemarking_new$end_lat

)



start_sf <- st_as_sf(start_df, coords = c("longitude", "latitude"), crs =
4326) %>%

  st_transform(st_crs(state_routes))


end_sf <- st_as_sf(end_df, coords = c("longitude", "latitude"), crs =
4326) %>%
```

```
  st_transform(st_crs(state_routes))

nearest_routes_start <- st_nearest_feature(start_sf, state_routes)
nearest_routes_end <- st_nearest_feature(end_sf, state_routes)

matched_routes_start <- state_routes[nearest_routes_start, ] %>%
select(StateRoute)
matched_routes_end <- state_routes[nearest_routes_end, ] %>%
select(StateRoute)


# lane marking - job starting and ending state route
lanemarking_new <- lanemarking_new %>%
  cbind(route_start=matched_routes_start$StateRoute) %>%
  cbind(route_end=matched_routes_end$StateRoute) %>%
  mutate(route_number=route_end) %>%
  select(route_number,everything())


# lane marking - aggregate color usage by route number, year, and month
lanemarking_summary <- lanemarking_new %>%
  mutate(yellow_useage = ifelse(is.na(yellow_useage), 0, yellow_useage)) %>%
  mutate(white_useage = ifelse(is.na(white_useage), 0, white_useage)) %>%
  group_by(route_number,year,month) %>%
  summarize(total_yellow=sum(yellow_useage),total_white=sum(white_useage)) %>%
  mutate(total_color=total_yellow+total_white)

lanemarking_summary2 <- lanemarking_summary %>% # color on each route
  group_by(route_number) %>%
  summarize(total_yellow=sum(total_yellow),total_white=sum(total_white)) %>%
  mutate(total=total_yellow+total_white)
```

```
# route data - get route length

state_routes_len_stat <- state_routes %>% as.tibble() %>%
  group_by(StateRoute) %>%
  summarise(route_length=max(EndAccumul))

# merge data - route level

crash_state_byroute <- crash_state %>% # number of crashes on state route
  filter(ld_ind==1) %>%
  group_by(route) %>%
  summarize(total_crash=n()) %>%
  select(total_crash,everything())

crash_lm_combine <-
full_join(crash_state_byroute,lanemarking_summary2,by=join_by(route==route_num
ber)) # crash and lane marking

crash_lm_full <-
full_join(crash_lm_combine,state_routes_len_stat,by=join_by(route==StateRoute)
) %>%
  full_join(crash_prob,by="route") %>%
  select(route,everything()) %>%
  mutate(color_per_mile = total/route_length) %>%
  filter(color_per_mile<1000)

# merge data - event level

crash_event_full <-
left_join(crash_state,crash_lm_full,by=join_by(route==route))
```

```
### part 1. exploratory analysis

## crash location

# Create leaflet heatmap
leaflet(crash_state) %>%
  addProviderTiles("CartoDB.Positron") %>%
  addTiles() %>%
  addHeatmap(lng = ~longitude, lat = ~latitude, blur = 5, max = 0.05, radius = 
5)

leaflet(crash_city) %>%
  addProviderTiles("CartoDB.Positron") %>%
  addTiles() %>%
  addHeatmap(lng = ~longitude, lat = ~latitude, blur = 5, max = 0.05, radius = 
5)

leaflet(crash_county) %>%
  addProviderTiles("CartoDB.Positron") %>%
  addTiles() %>%
  addHeatmap(lng = ~longitude, lat = ~latitude, blur = 5, max = 0.05, radius = 
5)

leaflet(crash_misc) %>%
  addProviderTiles("CartoDB.Positron") %>%
  addTiles() %>%
  addHeatmap(lng = ~longitude, lat = ~latitude, blur = 5, max = 0.05, radius = 
5)

## lane marking impact
```

```
### map job locations

lanemarking_joblocations <- lanemarking_new %>%
  select(job_name,start_lat,start_long,end_lat,end_long,year,month) %>%
  filter(start_lat>min(start_lat,na.rm = T)+0.1)

unique(lanemarking_joblocations$year)

leaflet(lanemarking_joblocations %>% filter(year<=year_analysis &
year>=year_analysis-1)) %>% # job location plot
  addProviderTiles("CartoDB.Positron") %>%
  addTiles() %>%
  addCircleMarkers(
    lng = ~start_long, lat = ~start_lat,
    label = ~job_name,
    radius = 1,
    color = "blue",
    fillOpacity = 0.1
  )

# map showing the starting locations of jobs
map <- leaflet() %>%
  addProviderTiles("CartoDB.Positron") %>%
  addTiles()  # Add default OpenStreetMap tiles

line_data <- lanemarking_joblocations

# Loop through each row of the data and add a line for each segment
for (i in 1:nrow(line_data)) {
  map <- map %>%
```

```
    addPolylines(
      lng = c(line_data$start_long[i], line_data$end_long[i]),  # Longitudes
      lat = c(line_data$start_lat[i], line_data$end_lat[i]),  # Latitudes
      color = "#03F",  # Line color
      stroke = TRUE,
      weight = 1,  # Line thickness
      opacity = 0.2  # Line transparency
    ) %>%
    addCircleMarkers(
      lng = line_data$start_long[i],  # Longitude of the arrow
      lat = line_data$start_lat[i],  # Latitude of the arrow
      radius = 1,
      color = "red",
      fillOpacity = 0.2
    )
}

map # Print the map


ggplot() + # start job locations
  borders("state", region="Washington", colour = "gray85", fill = "gray80") +
# World map
  geom_point(data = lanemarking_joblocations, aes(x = start_long, y =
start_lat), color="blue", size = 0.5) +
  theme_minimal() +
  labs(title = "Job start locations", x = "Longitude", y = "Latitude")+
  coord_fixed(1.33)

ggplot() + # end job locations
```

```
  borders("state", region="Washington", colour = "gray85", fill = "gray80") +
# World map

  geom_point(data = lanemarking_joblocations, aes(x = end_long, y = end_lat),
color="red", size = 0.5) +

  theme_minimal() +

  labs(title = "Job ending locations", x = "Longitude", y = "Latitude")+

  coord_fixed(1.33)


# lane marking job frequency

lanemarking_jobmonth <- table(lanemarking_new$month)


ggplot(data=data.frame(month=lanemarking_new$month))+

  geom_bar(aes(x=month))+

  xlab("Month")+

  ylab("Number of Lane Marking Jobs")


ggplot(data=crash_lm_full %>% filter(color_per_mile<500000))+

  geom_point(aes(x=color_per_mile,y=prob_ld))+

  geom_smooth(aes(x=color_per_mile,y=prob_ld),method = "lm", se = TRUE, color
= "blue")+

  xlab("Color per Mile")+

  ylab("Probability of Lane Departure Crashes")


ggplot(data=crash_lm_full %>% filter(color_per_mile<500 & prob_fat>0))+

  geom_point(aes(x=color_per_mile,y=prob_fat))


ggplot(data=crash_lm_full)+

  geom_point(aes(x=route_length,y=n_rec))


ggplot(data=crash_lm_full)+

  geom_point(aes(x=route_length,y=n_fat))
```

```
ggplot(data=crash_lm_full)+
  geom_point(aes(x=n_fat,y=n_ld))+
  geom_smooth(aes(x=n_fat,y=n_ld),method = "lm", se = TRUE, color = "blue")+
  xlab("Number of Fatal Crashes")+
  ylab("Number of Lane Departure Crashes")

ggplot(data=crash_lm_full)+
  geom_point(aes(x=total_yellow,y=total_white))

# lane marking spatial visualization

lm_route <- lanemarking_summary2 %>%
  arrange(desc(total)) %>%
  filter(total>70000)

highlight_routes_lm <- lm_route$route_number  # Route IDs to highlight

# Create a new column for color-coding
state_routes <- state_routes %>%
  mutate(highlight = ifelse(StateRoute %in% highlight_routes_lm,
"Marking_high", "Other"))

ggplot() +
  geom_sf(data = state_routes, aes(color = highlight), size = 0.8) +
  scale_color_manual(values = c("Marking_high" = "blue", "Other" = "gray")) +
  scale_size_manual(values = c("Marking_high" = 3, "Other" = 0.1))+
  labs(title = "State Routes with High Lane Marking Activities",
       color = "Route Type") +
  theme_minimal()
```

```
### part 2. lane departure risk


## lane departure risk - computation

crash_state %>%

  group_by(route_name) %>%

  summarise(total_crash=n()) %>%

  arrange(desc(total_crash))


crash_005 <- crash_state %>% filter(str_sub(route_name,1,3)=="005")


fatality_prob <- sum(crash_005$total_fatality)/nrow(crash_005) # fatal crash probability

ld_prob <- sum(crash_005$ld_ind)/nrow(crash_005) # lane departure probability


glimpse(crash_005)


route_set <- unique(str_sub(crash_state$route_name,1,3)) # the set of all routes


crash_state <- crash_state %>%

  mutate(route=str_sub(route_name,1,3)) %>%

select(date,time,route,longitude,latitude,,severity,total_veh,weather,surface_condition,lighting,veh1_type,veh1_age,veh1_gender,veh1_cause,ld_ind,everything())


crash_prob <- crash_state %>%

  group_by(route) %>%

  summarize(n_rec=n(),n_ld=sum(ld_ind),n_fat=sum(total_fatality)) %>%

  arrange(desc(n_rec)) %>%

  mutate(prob_ld=n_ld/n_rec,prob_fat=n_fat/n_rec)
```

```
ggplot(data=crash_prob)+
  geom_point(aes(x=prob_ld,prob_fat))

ggplot(data=crash_prob %>% filter(prob_ld>=0 & prob_ld<=1))+
  geom_histogram(aes(x=prob_ld))+
  xlab("Lane Departure Crash Probability (Conditioned on Crash Occurance)")+
  ylab("Count") # distribution of lane departure percentage on state routes

ggplot(data=crash_prob %>% filter(prob_fat>=0 & prob_fat<=1))+
  geom_histogram(aes(x=prob_fat))+
  xlab("Fatal Crash Probability (Conditioned on Crash Occurence)")+
  ylab("Count")
# distribution of lane departure percentage on state routes

## lane departure risk - high risk route mapping

ld_prone_route <- crash_prob %>% # find the most lane departure prone route
  filter(prob_ld>0.5 & n_ld>5 & prob_fat>0) %>%
  arrange(desc(prob_ld))

highlight_routes <- ld_prone_route$route  # Route IDs to highlight

state_routes <- state_routes %>%
  mutate(highlight = ifelse(StateRoute %in% highlight_routes, "LD-Prone",
"Other"))

osm_basemap <- ggplot() + # Get OSM tiles for Washington State
  annotation_map_tile(type = "cartolight") +  # Uses Carto Light basemap
  coord_sf(xlim = c(-125, -116), ylim = c(45, 50), expand = FALSE)  # Adjust
for WA
```

```
osm_basemap + # Overlay with highway network
  geom_sf(data = state_routes, aes(color = highlight, size = highlight),
inherit.aes = FALSE) +
  scale_color_manual(values = c("LD-Prone" = "red", "Other" = "gray")) +
  scale_size_manual(values = c("LD-Prone" = 2, "Other" = 0.5)) +
  labs(title = "Lane Departure Prone Routes in Washington State")


## lane departure risk - exploratory bivariate relations


ggplot(data=crash_prob)+
  geom_point(aes(x=prob_ld,y=n_fat))


ggplot(data=crash_lm_combine)+
  geom_point(aes(x=total,y=total_crash))


ggplot(data=crash_lm_combine %>% filter(total>0 &
total<500000),aes(x=total,y=total_crash))+
  geom_point()+
  geom_smooth(method="lm",color="red")


## lane departure risk - beta regression model


# preparing data


lanemarking_summary3 <- lanemarking_summary2
lanemarking_summary3$route_number <-
as.numeric(lanemarking_summary3$route_number)


route_len <- state_routes_len_stat
route_len$StateRoute <- as.numeric(route_len$StateRoute)
```

```
crash_beta <- crash_state %>%
  group_by(route,weather,surface_condition,lighting) %>%
  summarize(n_rec=n(),n_ld=sum(ld_ind),n_fat=sum(total_fatality)) %>%
  arrange(desc(n_rec)) %>%
  mutate(prob_ld=n_ld/n_rec,prob_fat=n_fat/n_rec) %>%
  na.omit() %>%
  filter(prob_ld>0 & prob_ld <1)

crash_beta$weather <- factor(crash_beta$weather)
crash_beta$surface_condition <- factor(crash_beta$surface_condition)
crash_beta$lighting <- factor(crash_beta$lighting)
crash_beta$route <- as.numeric(crash_beta$route)

crash_beta <- crash_beta %>%
  left_join(lanemarking_summary3,by=join_by(route==route_number)) %>%
  left_join(route_len, by=join_by(route==StateRoute)) %>%
  mutate(color_per_mile=min(total/route_length,1000)) %>%
  na.omit()

# model fit
beta_model <- betareg(formula = prob_ld ~ surface_condition + weather +
lighting + color_per_mile+ prob_fat, data = crash_beta)

summary(beta_model)

# model assessment

data_validation <- crash_beta %>% slice_sample(prop=0.2)
```

```
predictions <- predict(beta_model, newdata=data_validation)



r2_mcfadden <- r2(beta_model, which = "McFadden")

r2_nagelkerke <- r2(beta_model, which = "Nagelkerke")

r2_coxsnell <- r2(beta_model, which = "CoxSnell")


logLik(beta_model)

AIC(beta_model)

BIC(beta_model)


beta_mse <- mean((crash_beta$prob_ld - fitted(beta_model))^2)

beta_mae <- mean(abs(crash_beta$prob_ld - fitted(beta_model)))


ggplot(data=tibble(Observation=crash_beta$prob_ld,Prediction=fitted(beta_model
)))+

  geom_point(aes(x=Observation,y=Prediction))+

  geom_abline(slope=1,intercept = 0,color="red")


# residuals <- residuals(beta_model, type = "sweighted2") # Compute residuals


residuals <- residuals(beta_model)


ggplot(data=tibble(Fitted=fitted(beta_model),Residuals=residuals) %>%
filter(!is.nan(Residuals)))+

  geom_point(aes(x=Fitted,y=Residuals))+

  geom_abline(slope=0,intercept = 0,color="red")


qqnorm(residuals)

qqline(residuals)
```

```
plot(beta_model,which = 1)

plot(beta_model,which = 2)

plot(beta_model,which = 3)

plot(beta_model,which = 4)


### part 3. lane departure event - predictive model


## split training and test sets


crash_event_full$ld_ind <- factor(crash_event_full$ld_ind)

crash_event_full$weather <- factor(crash_event_full$weather)

crash_event_full$surface_condition <-
factor(crash_event_full$surface_condition)

crash_event_full$lighting <- factor(crash_event_full$lighting)

crash_event_full$month <- factor(crash_event_full$month)


crash_event_select <- crash_event_full %>%

  filter() %>% # jurisdiction


select(ld_ind,prob_ld,prob_fat,color_per_mile,veh1_age,veh1_posted_speed,weath
er,surface_condition,lighting,month) %>%

  na.omit()


names(crash_event_select)[which(names(crash_event_select)=="veh1_age")]<-"age"

names(crash_event_select)[which(names(crash_event_select)=="veh1_posted_speed"
)]<-"speed"

names(crash_event_select)[which(names(crash_event_select)=="color_per_mile")]<
-"lane_marking"

names(crash_event_select)[which(names(crash_event_select)=="prob_fat")]<-
"fatal_probability"

names(crash_event_select)[which(names(crash_event_select)=="surface_condition"
)]<-"surface_condition"
```

```
crash_split <- initial_split(crash_event_select)


## xgboost model for event classification

crash_train_subset <- training(crash_split)

crash_test_subset <- testing(crash_split)


# encode categorical variables

crash_train_subset$weather <- as.numeric(factor(crash_train_subset$weather))

crash_train_subset$surface_condition <-
as.numeric(factor(crash_train_subset$weather))

crash_train_subset$lighting <- as.numeric(factor(crash_train_subset$lighting))

crash_train_subset$month <- as.numeric(factor(crash_train_subset$month))


crash_train <- crash_train_subset %>% select(-ld_ind,-prob_ld) # features for
training data

crash_label <- crash_train_subset %>% select(ld_ind) # response


p0 <- mean(as.numeric(crash_label$ld_ind)-1)


crash_resampled <-
upSample(crash_train,factor(crash_label$ld_ind),list=FALSE,yname="ld_ind") #
upsampling


crash_train <- crash_resampled %>% select(-ld_ind)

crash_label <- crash_resampled %>% select(ld_ind)


crash_test_subset$weather <- as.numeric(factor(crash_test_subset$weather))

crash_test_subset$surface_condition <-
as.numeric(factor(crash_test_subset$weather))

crash_test_subset$lighting <- as.numeric(factor(crash_test_subset$lighting))
```

```
crash_test_subset$month <- as.numeric(factor(crash_test_subset$month))

crash_test <- crash_test_subset %>% select(-ld_ind,-prob_ld) # feature for
test data

crash_test_xgb <- xgb.DMatrix(data = as.matrix(crash_test)) # needed format
for xgboost

# train xgboost model

xgboost_params <- list(
  objective = "binary:logistic",
  eval_metric = "logloss",
  scale_pos_weight = 1.2,  # weight for the minority class
  max_depth = 20,
  eta = 0.1
)

crash_xgboost_model <- xgboost(
  data = as.matrix(crash_train),
  label = as.matrix(crash_label),
  params = xgboost_params,
  nround = 2500,
  verbose = TRUE,
  print_every_n = 100
)

# prediction

pred_xgb <- predict(crash_xgboost_model,crash_test_xgb) # probability of lane
departure
```

```
ld_threshold <- 0.7 # lane departure event threshold

pred_xgb_binary <- ifelse(pred_xgb>ld_threshold,1,0) # lane departure outcome

# model evaluation - multiple metrics

combined_result <-
bind_cols(crash_test_subset,pred=factor(pred_xgb_binary),prob=pred_xgb)

combined_result %>% metrics(truth=ld_ind,estimate=pred,prob)

# model evaluation - confusion matrix

cm <- confusionMatrix(factor(combined_result$pred),
factor(combined_result$ld_ind))

cm # show metrics

print(cm$byClass["F1"])

ggplot(data = as.data.frame(cm$table), aes(x = Reference, y = Prediction)) +
  geom_tile(aes(fill = Freq), color = "white") +
  geom_text(aes(label = Freq)) +
  scale_fill_gradient(low = "lightgreen", high = "darkgreen")

# model evaluation - histogram of predicted probability

ggplot(combined_result, aes(x = prob, fill = pred)) +
  geom_histogram(alpha = 0.5, position = "identity", bins = 20) +
  labs(title = "Distribution of Predicted Probabilities",
```

```
       x = "Probability", y = "Frequency")

# model evaluation - roc curve

ld_actual <- combined_result$ld_ind
ld_predicted_prob <- combined_result$prob

roc_obj <- roc(ld_actual,ld_predicted_prob)
plot(roc_obj, main = "ROC Curve", col = "blue")
abline(0, 1, lty = 2, col = "gray")
auc(roc_obj)  # Area under the curve

# model evaluation - cumulative gain (lift) curve

y_true <- as.numeric(ld_actual)-1
y_score <- ld_predicted_prob

gain <- gains(y_true, y_score)

plot(c(0,gain$cume.pct.of.total*sum(y_true))~c(0,gain$cume.obs),
     type="l", xlab="# cases", ylab="Cumulative response", main="Cumulative
Gains Chart")

lines(c(0,sum(y_true))~c(0,length(y_true)), lty=2)

### part 4. SHAP analysis

## SHAP model

xx_explain_4 <- crash_test %>% slice_sample(n=4) # samples to be explained
```

```
xx_explain_100 <- crash_test %>% slice_sample(n=100)

crash_test

xx_train <- crash_train # samples to estimate Shapley value

explanation_crash_4 <- explain(
  model=crash_xgboost_model,
  x_explain=xx_explain_4,
  x_train=xx_train,
  approach= "empirical",
  phi0=p0,
  seed = 1,
  n_MC_samples = 1e2,
  MSEv_uniform_comb_weights = TRUE,
  max_n_coalitions=128
)

explanation_crash_100 <- explain(
  model=crash_xgboost_model,
  x_explain=xx_explain_100,
  x_train=xx_train,
  approach= "empirical",
  phi0=p0,
  seed = 1,
  n_MC_samples = 1e2,
  MSEv_uniform_comb_weights = TRUE,
  max_n_coalitions=128
)
```

```
## visualization

plot(explanation_crash_4)

plot(explanation_crash_4,plot_type="waterfall")

plot(explanation_crash_100,plot_type = "beeswarm")

plot(explanation_crash_100,plot_type = "scatter")

## identify significance factors

shap_df <- as.data.frame(explanation_crash_100$shapley_values_est) %>%
select(-explain_id,-none)  # Convert to data frame

shap_value <- colMeans(shap_df)  # Mean absolute SHAP

shap_importance <- colMeans(abs(shap_df))  # Mean absolute SHAP

ggplot(data.frame(Feature = names(shap_value), Value = shap_value),
       aes(x = reorder(Feature, Value), y = Value)) +
  geom_col(fill = "lightblue") +
  coord_flip() +
  labs(title = "SHAP Feature Value", x = "Feature", y = "SHAP")

ggplot(data.frame(Feature = names(shap_importance), Importance =
shap_importance),
       aes(x = reorder(Feature, Importance), y = Importance)) +
  geom_col(fill = "lightblue") +
  coord_flip() +
  labs(title = "SHAP Feature Importance", x = "Feature", y = "Mean |SHAP|")

shap_sd <- as.data.frame(explanation_crash_100$shapley_values_sd) %>% select(-
explain_id,-none)  # Convert to data frame

shap_importance_sd <- colMeans(abs(shap_sd))  # Mean absolute SHAP
```

```
ggplot(data.frame(Feature = names(shap_importance), Uncertainty =
shap_importance_sd),

       aes(x = reorder(Feature, Uncertainty), y = Uncertainty)) +

  geom_col(fill = "lightblue") +

  coord_flip() +

  labs(title = "SHAP Feature Uncertainty", x = "Feature", y = "Std. |SHAP|")


# Permutation test for a feature (e.g., "age")

original_shap <- shap_df$veh1_age

permuted_shap <- sample(shap_df$veh1_age)  # Shuffled version


paste(as.character(year_analysis), "analysis completed")
```

Title VI Notice to Public

It is the Washington State Department of Transportation's (WSDOT's) policy to assure that no person shall, on the grounds of race, color, or national origin, as provided by Title VI of the Civil Rights Act of 1964, be excluded from participation in, be denied the benefits of, or be otherwise discriminated against under any of its programs and activities. Any person who believes his/her Title VI protection has been violated, may file a complaint with WSDOT's Office of Equity and Civil Rights (OECR). For additional information regarding Title VI complaint procedures and/or information regarding our non-discrimination obligations, please contact OECR's Title VI Coordinator at (360) 705-7090.

Americans with Disabilities Act (ADA) Information

This material can be made available in an alternate format by emailing the Office of Equity and Civil Rights at wsdotada@wsdot.wa.gov or by calling toll free, 855-362-4ADA(4232). Persons who are deaf or hard of hearing may make a request by calling the Washington State Relay at 711.