Report No. UT-24.08

CONSTRUCTION WORK ZONE SAFETY: SPATIO-TEMPORAL ANALYSIS OF CONSTRUCTION WORK ZONE CRASHES

Prepared For:

Utah Department of Transportation
Research & Innovation Division

Final Report
April 2024

RESEARCH
UDOT Keeping Utah Moving
Utah Department of Transportation - Research Division
4501 South 2700 West - P.O. Box 148410 - SLC,UT 84114-8410

DISCLAIMER

The authors alone are responsible for the preparation and accuracy of the information, data, analysis, discussions, recommendations, and conclusions presented herein. The contents do not necessarily reflect the views, opinions, endorsements, or policies of the Utah Department of Transportation or the U.S. Department of Transportation. The Utah Department of Transportation makes no representation or warranty of any kind and, therefore, assumes no liability.

ACKNOWLEDGMENTS

The authors acknowledge the Utah Department of Transportation (UDOT) for funding this research, and the following individuals from UDOT on the Technical Advisory Committee for helping to guide the research:

- Robert Chamberlin (Consultant)
- Matt Lazanich
- Michael Adams
- Kelly Njord
- Jared Beard
- Suyanka Neupaney
- Rodney Ruby
- Barbra Freeman

TECHNICAL REPORT ABSTRACT

1. Report No.
UT- 24.08

2. Government Accession No.
N/A

3. Recipient's Catalog No.
N/A

4. Title and Subtitle
Construction Work Zone Safety: Spatio-Temporal Analysis of Construction Work Zone Crashes

5. Report Date
April 2024

6. Performing Organization Code

7. Author(s)
Ali Hassandokht Mashhadi, Abbas Rashidi, Nikola Markovic

8. Performing Organization Report No.

9. Performing Organization Name and Address
The University of Utah
Department of Civil and Environmental Engineering
201 Presidents Circle
Salt Lake City, Utah 84112

10. Work Unit No.
5H088 45H

11. Contract or Grant No.
238192

12. Sponsoring Agency Name and Address
Utah Department of Transportation
4501 South 2700 West
P.O. Box 148410
Salt Lake City, UT 84114-8410

13. Type of Report & Period Covered
Final. October 2022 to Feb 2024

14. Sponsoring Agency Code
PIC No. UT22.31

15. Supplementary Notes
Prepared in cooperation with the Utah Department of Transportation and the U.S. Department of Transportation, Federal Highway Administration

16. Abstract

This report presents a comprehensive analysis of work zone safety considering multiple factors, including crash severity, speed analysis, countermeasure analysis, and state-of-the-practice in DOTs. It encompasses a broad literature review, drawing from DOT reports, NCHRP publications, MUTCD guidelines, and academic research, to explore a range of traffic control approaches within work zones. The study also includes findings from a survey distributed to all DOTs, with 24 recorded responses across the nation, providing valuable perspectives on safety and satisfaction in work zones. In the context of specific safety measures, the report delves into the role of longitudinal rumble strips. While generally effective in reducing roadway departure crashes, their impact within work zones appears less significant. This observation prompts a call for further investigation and possible adjustments to their use in these areas. The report culminates in a comprehensive evaluation of work zone safety countermeasures and their implications on driver behavior, particularly speed. By integrating survey results from state DOTs and reviewing the efficacy of various safety interventions, the report offers substantive recommendations aimed at enhancing safety in construction work zones. In conclusion, this report provides a comprehensive analysis of work zone safety countermeasures, survey findings from state DOTs, and insights into the impact of work zones on drivers' speed, offering valuable recommendations for improving safety within construction work zones.

17. Key Words
Work zone safety, Work zone crashes

18. Distribution Statement
Not restricted. Available through:
UDOT Research Division
4501 South 2700 West
P.O. Box 148410
Salt Lake City, UT 84114-8410

23. Registrant's Seal
N/A

19. Security Classification
(of this report)
Unclassified

20. Security Classification
(of this page)
Unclassified

21. No. of Pages

22. Price
N/A

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

UNIT CONVERSION FACTORS

NO UNIT CONVERSIONS

EXECUTIVE SUMMARY

This report presents a comprehensive analysis of work zone safety considering multiple factors, including crash severity, speed analysis, countermeasure analysis, and state-of-the-practice in DOTs.

Machine learning models were utilized to interpret the influence of various factors on work zone crash severity. The findings underscore the capability of these models to provide insights into the complex interplay of elements affecting crashes, laying a groundwork for future explorations in this domain. The study analyzed the effect of different contract types on crash occurrence. CMGC contracts exhibited a notable increase in the number of crashes as vehicles approached work zones, indicating the importance of considering contract specifications in relation to safety measures. Moreover, CMGC has a much higher crash rate per 100 million VMT compared to design-build or design-bid-build contract types. This report thoroughly examines work zone safety countermeasures, drawing from an extensive array of sources, including DOT reports, NCHRP publications, MUTCD guidelines, and academic research, and categorized them into 5 groups, including speed control, intrusion prevention, human-machine interaction, smart work zone, and traditional approaches.

Furthermore, insights from a survey distributed to all DOTs, with 24 responses from 22 states, are also incorporated. The feedback from these states, which span a wide geographic area, offers valuable perspectives on factors that influence safety and satisfaction in work zones, thus enriching our understanding of implementing effective countermeasures. Overall, this report provides valuable insights into work zone crash severity and offers recommendations for enhancing safety. Future research opportunities include exploring the effectiveness of various countermeasures, incorporating real-time data for improved prediction accuracy, and investigating the impact of additional variables on work zone crash severity. By addressing these areas and implementing evidence-based safety measures, we can work towards creating safer work zones, reducing the occurrence and severity of crashes, and improving overall road safety.

1.0 INTRODUCTION

1.1 Introduction

Work zone crashes in transportation systems pose a significant threat to road users and transportation agencies. The Federal Highway Administration (FHWA) reports an average of 794 fatalities annually in the United States between 2015 and 2020, resulting in an estimated cost of $17.5 billion annually (Work Zone Crashes, n.d.).

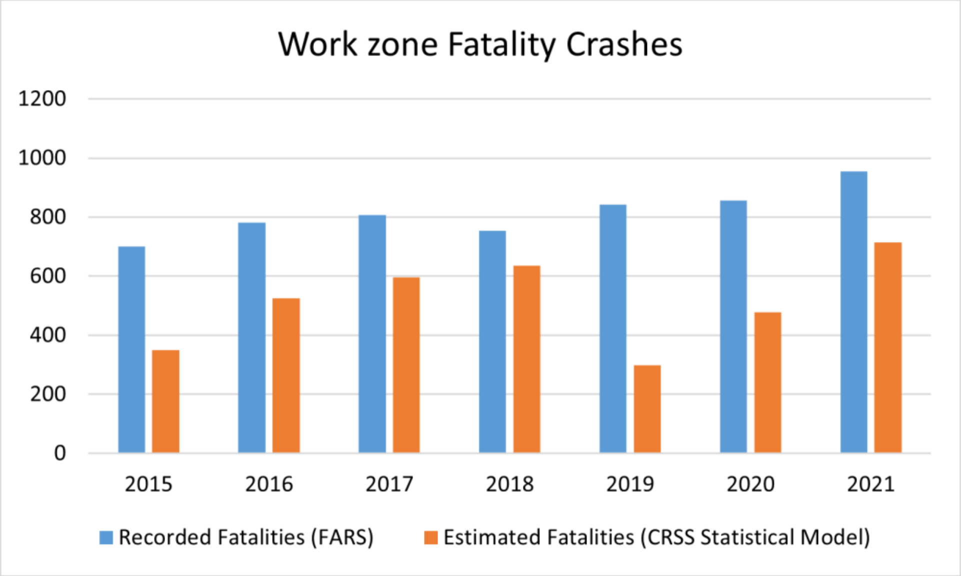


Figure 1. Number of Work Zone Fatality Crashes Between 2015-2021

Even with reduced traffic volumes during the COVID-19 pandemic, work zone crashes in 2020 alone accounted for over 102,000 incidents, causing more than 45,000 injuries and over 850 fatalities, surpassing the previous year's records (Work Zone Crashes, n.d.). These alarming statistics highlight the urgent need to understand and mitigate the impact of work zones on traffic safety. To design effective mitigation and improvement strategies, it is crucial to accurately comprehend the factors influencing work zone crash severity. Local data capturing unique conditions such as driving behavior, regulations, geography, weather, and road conditions are essential for maximizing the effectiveness of the strategies.

Despite the critical need for comprehensive analysis, there is currently no study investigating the state of practice in DOTs regarding work zones, speed analysis in work zones in

Utah, and the effectiveness of safety countermeasures implemented by the Utah Department of Transportation (UDOT). Previous studies have primarily relied on analytical methods to establish relationships between work zone attributes and crash occurrence or severity. However, the dynamic and complex nature of work zones makes mathematical models challenging to apply. As a result, machine learning techniques have emerged as powerful tools for modeling such intricate systems. These algorithms can learn patterns and relationships from data, making them suitable for capturing the complexities of work zone crashes.

This study aims to address the pressing need for a more accurate and comprehensive understanding of work zone crash severity factors to inform the development of effective safety management strategies. By employing advanced machine learning techniques, this research endeavors to overcome the limitations of traditional analytical models and provide insights into the intricate relationships between work zone features and crash severity. Additionally, this study conducted a thorough literature review on countermeasures implemented to enhance work zone safety, exploring the state of practice in various DOTs. Furthermore, the research investigated the speed effect of work zones, analyzing traffic data to understand the impact of work zones on drivers' speed behavior. Moreover, the study examined the effect of different contract types on work zone crashes, aiming to identify potential correlations between contract specifications and safety outcomes. Through these comprehensive analyses, this research seeks to provide valuable insights that can guide the development of targeted and effective safety management strategies for mitigating work zone crashes. The findings from this study can contribute to the development of targeted and tailored interventions to mitigate work zone crashes, ultimately improving traffic safety for all road users.

1.2 Background

Work zones play a crucial role in infrastructure development and maintenance but pose significant safety risks for both workers and motorists. In recent years, there has been a growing interest in utilizing statistical and machine learning models to enhance our understanding and prediction of transportation safety outcomes in work zones. This literature review discusses the research on work zone safety, dividing it into separate sections to discuss the findings of studies that utilize statistical methods and machine learning approaches.

1.2.1 Statistical Approaches

Statistical approaches have been widely used to estimate crash severity/frequency and identify factors contributing to more severe crashes. Regression analysis is a commonly employed statistical tool to examine the relationship between speed, traffic volume, road geometry, and crash severity/frequency variables. Logistic and probit regression models are frequently used for analyzing discrete outcomes, allowing the estimation of the probability of a specific outcome based on explanatory variables. Various studies have utilized statistical methods to investigate work zone crash severity. For example, Coburn et al. (2013) aimed to quantify injury outcomes and develop comprehensive injury costs for work zone crashes based on the crash type and severity using a three-step methodology and crashes in Wisconsin between 2001 and 2010. The study found that the KABCO scale, which classifies injuries as killed, incapacitating injury, non-incapacitating injury, possible injury, or property damage only, may need reconsideration due to discrepancies between injury types and severities. The calculated comprehensive costs for different crash types were significantly higher than the default values provided by FHWA. This highlights the importance of developing crash-specific costs for more accurate benefit-cost analysis and implementing countermeasures in work zones. In another study, Chen & Tarko (2014) examined traffic safety in highway work zones using detailed data from a survey of project engineers and existing datasets. Monthly clusters of observations corresponding to individual work zones are analyzed using a two-level random parameter negative binomial model. The safety effects of various work zone design and traffic management features, including lane shift, lane split, and detour, are identified. The study also explores the viability of a fixed parameters negative binomial model with random effects as an alternative. The results show that both models yield similar marginal effects on crash frequency, suggesting the potential practicality of using fixed parameters models in certain cases. The obtained model with random effects is found to be useful for programming police enforcement in highway work zones in Indiana.

Osman et al. (2016) focused on investigating the factors contributing to the injury severity of large truck crashes in work zones. Various econometric models, including multinomial logit, nested logit, ordered logit, and generalized ordered logit, were compared to analyze the injury severity data. The database consisted of work zone crashes involving large trucks in Minnesota over 10 years. The empirical findings indicate that the generalized ordered logit model provided the best fit for the data. Elasticity analysis revealed that factors such as daytime crashes, lack of access control, higher speed limits, and crashes on rural principal arterials increased the risk of severe crashes in work zones. Liu et al. (2016) investigated the correlation between precrash actions and driver injury severity in work and non-work zone crashes. Using a large-scale statewide crash database, hierarchical models were employed to account for the injury severity of each driver involved. The analysis reveals that intentional improper actions or violations increase the chances of driver injury by 9.9% to 10.3% in work zone crashes, compared to 1.7% to 5.7% in non-work zone crashes**. Speeding, following too closely, and disregarding traffic regulations were identified as significant contributing factors**. These findings highlight the **importance of effective speed enforcement and traffic regulations** to improve work zone safety and reduce the risk of injuries.

Anderson & Hernandez (2017) addressed the gap in previous research by examining injury severity factors for heavy-vehicle crashes based on roadway classification. A mixed logit modeling framework is used, and the results indicate that roadway classifications should be considered separately due to statistically significant differences in estimated parameters. The findings emphasize the **importance of considering roadway classification** in safety analyses and suggest the need for further research on injury severity and other safety measures within different subpopulations of crash datasets. Osman et al. (2018) examined factors influencing injury severity in passenger-car crashes within various work zone configurations. A Mixed Generalized Ordered Response Probit (MGORP) model is developed using a 10-year crash database. Results indicate that **factors such as partial access control, rural roads, evening and weekend crashes, and curved roadways contribute to higher severity outcomes**. Covariate effects vary across different work zone configurations, highlighting the importance of tailored safety measures for specific layouts.

Ravani & Wang (2018) examined the impact of police presence on work zone safety and speeding in highway work zones. Speed data were collected from six work zone locations in California, and data analysis was conducted using statistical methods. Four measures of effectiveness (MOEs) were evaluated, including average speed reduction, speed variance, 85th percentile speed, and proportion of high-speed vehicles. The results indicate that **all levels of police presence led to statistically significant improvements** in one or more of the MOEs, highlighting the positive impact of police presence in mitigating work zone safety risks and reducing speeding incidents. K. Zhang & Hassan (2019a) developed a random parameter-ordered probit model to analyze factors affecting work zone crash severity. Their study found that speeding and foggy weather are important factors that can influence the parameters of a random parameter model and identified **weekdays and nighttime as having a higher risk of rear-end crashes in work zones**. Santos et al. (2021) employed statistical models to identify primary risk factors causing work zone crashes. Their analysis revealed that the major **contributing factors were speeding, disregard for vertical signs, lighting, locations that include intersections, and involvement of motorcycles and heavy vehicles.**

While statistical approaches have shown promise in estimating crash severity, it is important to consider their potential limitations, such as the oversimplification of complex relationships and dependence on assumptions and model specifications. These factors can affect the accuracy and reliability of the predictions. Nonetheless, these studies contribute valuable insights into understanding work zone safety and identifying factors that can mitigate crash severity.

Table 1. Work Zone Crash Literature and Findings

Authors	Findings
(Akepati & Dissanayake, 2011)	The lane-closure work zone type had the highest percentage of crashes, followed by work on the shoulder or median type of work zone.
(Al-Bdairi, 2020)	Contributing factors such as lighting, driver behavior, and age are uniquely significant for a specific time of day period. Whereas undeployed airbags, single-vehicle crashes and rear-end collisions tend to have higher injury severity regardless of the time of day.
(Z. Zhang et al., 2022)	It appears that conducting work zones during the nighttime with the current deployment strategies on Pennsylvania state roads does not necessarily increase crash risks, but a work zone significantly increases crash risks during daytime

(Mokhtarimousavi et al., 2019, 2020)	Work on the shoulder or median, the presence of advance warning areas, daytime non-peak construction, and vehicles that are not carrying multiple passengers are more likely to decrease injury severity.
(Mokhtarimousavi et al., 2021)	The termination area of the work zone is most critical for both daytime and nighttime crashes, as this location has the highest increase in severe injury likelihood.
(Santos et al., 2021)	Excessive speed, disregard for vertical signs, poor lighting, locations with intersections, and motorcycle and heavy vehicle involvement as the most significant risk factors.
(K. Zhang & Hassan, 2019b)	Weather conditions (rain) and driver characteristics, such as gender and age group, work zones with multiple lane closures and the presence of heavy vehicles increase the crash fatality risk.
(Islam, 2022)	Poor lighting and areas with older motorcyclists (50-65) are more likely to experience higher crash severities.

1.2.2 Machine Learning Approaches

Machine learning approaches provide an alternative means to estimate crash severity and frequency, addressing some of the limitations of statistical methods. These algorithms do not rely on specific assumptions about variable relationships, allowing greater flexibility in handling complex data and capturing nonlinear relationships. Several studies have utilized machine learning techniques to analyze work zones (Mashhadi et al., n.d., 2021a, 2021b; Mashhadi & Rashidi, 2021). Effati et al. (2015) introduced a geospatial approach, using fuzzy classification and regression tree (FCART), to predict motor vehicle crashes and their severity on two-lane, two-way roads. The FCART model combines fuzzy logic and decision tree techniques to handle uncertain input data and improve interpretability. The model is compared with other methods, such as CART and SVM, and the results demonstrate that the bagged-FCART model outperforms the others in predicting crash severity. Factors such as vehicle failure, seat belt usage, weather conditions, and geographic features like curves and adjacent facilities were identified as significant contributors to crash severity. This approach highlights the importance of targeted and behaviorally informed safety measures on regional roads.

Iranitalab & Khattak (2017) compared the performance of four methods (MNL, NNC, SVM, RF) in predicting traffic crash severity and developed a crash costs-based approach for evaluation. Two vehicle crashes were analyzed and split into training and validation subsets using reported crash data from Nebraska. NNC showed the best overall prediction performance, followed by RF and SVM, while MNL performed the weakest. Data clustering improved MNL, NNC, and RF prediction performance but had mixed effects on NNC. The proposed crash costs-based accuracy measure highlighted the importance of considering crash costs for accurate prediction. Alkheder et al. (2017) developed an Artificial Neural Network (ANN) classifier to predict crash severity in normal conditions, using a k-means algorithm for data clustering and an ordered probit model for benchmarking. Their ANN model achieved 74.6% accuracy in predicting crash severity. Park et al. (2017) addressed the limitations of existing proximity sensing and alert systems in roadway work zones by developing a Bluetooth Low Energy (BLE)-based system. The study focuses on parameter adjustment and adaptive signal processing (ASP) methods to account for variations in equipment types, approach speeds, and dynamic conditions. Field trials demonstrate that the system's parameter adjustment reduces inconsistency in alert distances, while the ASP method minimizes time delays caused by high approaching speeds. Overall, the developed system enhances construction work zone safety by better understanding spatial relationships among equipment, operators, and workers in real time.

In addition to the studies mentioned earlier, (Jeong et al., 2018) utilized a dataset of 297,113 vehicle crashes from the Michigan Traffic Crash Facts (MTCF) to classify injury severity. Techniques like under-sampling and over-sampling are employed to address imbalanced classes. Five classification models are used, and bagging with decision trees and over-sampling yields the highest performance. Mokhtarimousavi et al. (2019) employed a mixed logit model and Support Vector Machine (SVM) to predict work zone crash severity. They also utilized metaheuristic algorithms such as particle swarm optimization, harmony search, and the whale optimization algorithm to enhance SVM performance. SVM outperformed the mixed logit model by 16 percentage points, highlighting its effectiveness. In a subsequent study, Mokhtarimousavi et al. (2020) utilized mixed logit and random forest algorithms to evaluate the importance of variables on work zone crash severity. Their findings revealed four influential factors: work on the shoulder or median, advance warning area, daytime nonpeak, and multi-occupant, directly affecting crash severity.

Machine learning approaches offer flexibility in handling complex data, capturing nonlinear relationships, and identifying patterns that traditional statistical models may overlook. However, it is important to note that these methods may require substantial amounts of data, are prone to overfitting, and demand significant computing power and time for processing extensive datasets. Nonetheless, they provide valuable insights into understanding and predicting work zone crash severity.

1.3 Objectives

The primary objective of this study is to enhance the prediction of work zone crash severity by employing different machine learning techniques and analyzing their effectiveness when applied to a dataset containing a wide range of work zone crash and roadway attributes. Specifically, the objectives of this study are as follows:

1. Perform comprehensive data analysis of work zone crashes: Conduct a detailed analysis of the work zone crash dataset to identify patterns, trends, and influencing factors associated with crash severity. Explore the relationships between various factors such as driver behavior, work zone characteristics, traffic flow, and environmental conditions to gain insights into their impact on crash severity outcomes. This analysis will provide a deeper understanding of the dynamics and interactions among these factors and their contribution to work zone crash severity.
2. Develop and implement a comprehensive machine learning framework: Establish a framework incorporating various machine learning algorithms to predict work zone crash severity, including probabilistic and non-probabilistic models. This framework will enable the comparison of different algorithms and their performance in predicting the severity of work zone crashes.
3. Conduct a feature importance analysis: Identify and analyze the key factors influencing work zone crash severity through a feature importance analysis. Determine the relative importance of various work zone attributes, such as weather conditions, road geometries, traffic characteristics, and work zone configurations, in predicting the severity of crashes.

4. Analyzing the effects of different factors on work zone safety: Investigate the impact of various factors on work zone safety, including contract types, traffic countermeasures, and rumble strips.
5. State of the practice in Work Zone Countermeasures: Evaluate the current state of practice in work zone safety countermeasures among DOTs, including both traditional approaches and emerging technologies. Conduct a comprehensive review of existing literature, guidelines, and best practices related to work zone countermeasures.

By achieving these objectives, this study aims to contribute to advancing work zone safety management by providing a more accurate and comprehensive understanding of the factors influencing crash severity. The findings will assist transportation agencies in designing evidence-based interventions and strategies to mitigate work zone crashes, improve traffic safety, and reduce the economic burden of these incidents.

1.4 Outline of Report

1. Introduction
   - Overview of work zone safety and the importance of studying crash severity
   - Research objectives and significance
   - Review of existing studies on work zone crash severity and influencing factors
   - Discussion of previous research methods and findings
   - Identification of research gaps and the need for the current study
   - Brief description of the report structure

2. Research Methods
   - Explanation of any preprocessing steps performed on the data, such as data cleaning or feature engineering
   - Explanation of the machine learning techniques employed for crash severity prediction.

3. Data Collection
   - Description of the dataset used and its characteristics

- Overview of the data collection process, including the sources and methods used
- Description of the work zone crash data and associated attributes

4. Results and Findings
   - Presentation and interpretation of the findings
   - Discussion of the feature importance analysis and the relative significance of different variables

5. Conclusion
   - Summary of the main findings and their implications
   - Reflection on the research limitations and suggestions for future studies

The report will follow this structure to provide a comprehensive understanding of the research methods, data collection process, model evaluation, and the resulting findings and conclusions related to work zone crash severity prediction and influencing factors.

2.0 RESEARCH METHODS

2.1 Overview

This section encompasses several key components, including data cleaning, statistical modeling, deterministic machine learning modeling, and probabilistic machine learning modeling. These methods were employed to analyze work zone crash data and predict crash severity based on various influencing factors.

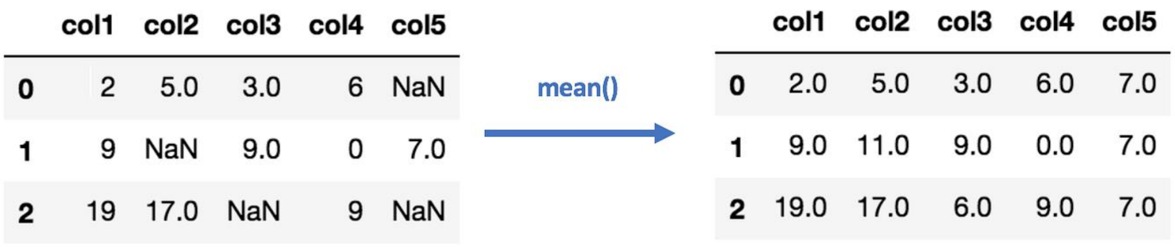
2.2 Data Cleaning and Preprocessing

Data cleaning and preprocessing are crucial in ensuring the quality and reliability of tabular data used for analysis. This study conducted a comprehensive data cleaning process to prepare the dataset for subsequent modeling. The first step involved identifying and handling missing values in the dataset. Missing data can introduce biases and affect analysis accuracy, so various techniques, such as imputation, were applied to fill in missing values based on statistical methods or pattern recognition. Here are some commonly used data cleaning approaches:

2.2.1 Missing Data

Missing data is a common challenge in datasets. There are several strategies to handle missing data, including:

- Deletion: Removing rows or columns with missing values. This approach should be used cautiously as it may result in data loss and biased analysis.
- Imputation: Filling in missing values using statistical methods such as mean, median, mode, or regression imputation.



	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0

Figure 2. Imputation Example with Column Mean Values

2.2.2 Outlier Detection and Treatment

Outliers are extreme or unusual observations that can significantly affect the analysis. Various methods can be used to detect outliers, such as:

- Statistical methods: Identifying outliers based on z-scores, standard deviations, or boxplot measures.
  - A z-score is just the number of standard deviations away from the mean that a certain data point is.
  - A boxplot is a simple way of detecting outliers by drawing a box representing the central 50% of the data. The line drawn in the middle shows the median value. The lines extending from the box (whiskers) capture the range of the remaining data outside of the middle 50% (for example, the upper 25% and the lower 25%). Any point that falls outside the lines indicates an outlier.

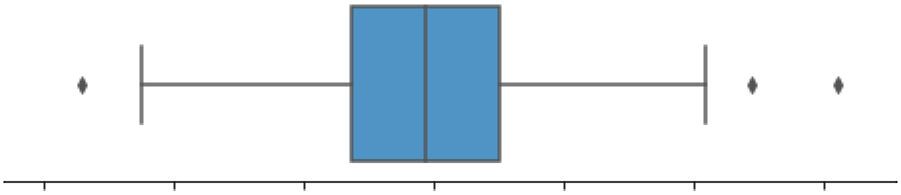


Figure 3. Statistical Methods for Outlier Detection

- Visualization techniques: Plotting the data to visually identify data points that deviate significantly from the overall pattern.

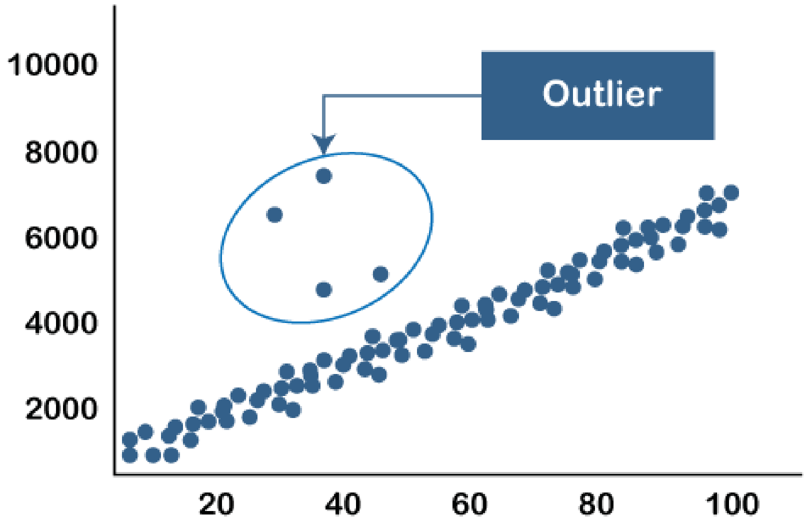


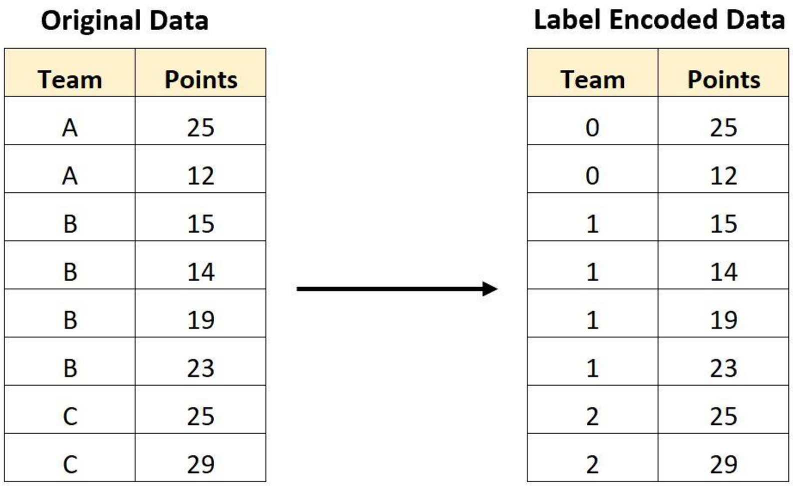
Figure 4. Outlier Detection Using Visualization

- Winsorization or trimming: Winsorization replaces extreme values with the nearest non-outlier value to reduce their impact, while trimming removes outliers from the data set entirely.

2.2.3 Transformation and Encoding

Data may need to be transformed or encoded depending on the analysis requirements. Examples include:

- Feature scaling: Scaling numerical features to a standard range (e.g., normalization or standardization).
- Label Encoding: Assigning numeric labels to categorical variables with an inherent order.



Original Data

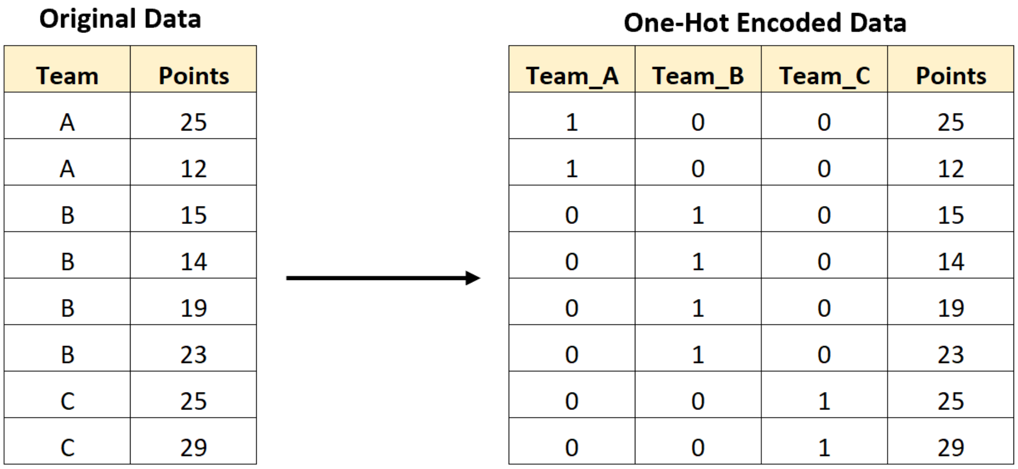
Team	Points
A	25
A	12
B	15
B	14
B	19
B	23
C	25
C	29

Label Encoded Data

Team	Points
0	25
0	12
1	15
1	14
1	19
1	23
2	25
2	29

Figure 5. Label Encoding Technique Example

- One-Hot Encoding: machine learning algorithms require numeric input and output variables. One-hot encoding transforms categorical data into numeric variables.
    - For example, imagine a data set with a column of different basketball teams, each with a number of points scored. One-hot encoding will create new columns to reflect each of the unique team names in the “team name” column, and the new columns will be filled with 0s and 1s.



Original Data

Team	Points
A	25
A	12
B	15
B	14
B	19
B	23
C	25
C	29

One-Hot Encoded Data

Team_A	Team_B	Team_C	Points
1	0	0	25
1	0	0	12
0	1	0	15
0	1	0	14
0	1	0	19
0	1	0	23
0	0	1	25
0	0	1	29

Figure 6. One-Hot Encoding Technique Example

2.2.4 Feature Selection

Feature selection is an essential step in machine learning because it helps identify the most important variables that influence the outcome of the target variables. The remaining features may be irrelevant to the target variable. Narrowing down the feature selection reduces the model's complexity, decreases the time it takes for the model to be trained, and prevents a dumb model, filled with inaccurate or less reliable predictions, from being created. Common approaches include:

- Filter methods: Select features based on statistical measures like correlation or mutual information and “filter” the remaining features out.
    - Mutual information measures how much one random variable tells us about another. In other words, it quantifies how similar or how different two variables are.
- Wrapper methods: Selects features based on a specific machine learning algorithm that we are trying to fit into a given data set. All of the possible combinations of the features are considered. The combination of features that gives the optimal results for the specific machine learning algorithm is selected.
- Embedded methods: Select features by embedding features (creating a lot of subsets from the particular dataset) during the model building process and observing each iteration of model training. Every subset that results in the maximum accuracy will be selected as a subset of features, which will later be given to the dataset for training.

2.2.5 Overfitting

One of the most common challenges in machine learning is overfitting, where the model can perform well on trained data but cannot accurately predict values on test data. Regularization is a technique used to prevent overfitting by applying a penalty term to the loss function during training. The penalty prevents the modeling from becoming too complex and helps control the model’s ability to fit noise within the trained data.

2.3 Machine Learning Modeling

Machine learning modeling is a process used to train computer algorithms to make predictions or decisions based on data. These techniques have been used and applied to different areas of science, including safety assessments (Hassandokht Mashhadi et al., 2024; Mashhadi et al., 2023; Mashhadi & Rashidi, 2021), condition assessments (Mohammadi, Rashidi, et al., 2023;

Mohammadi, Sherafat, et al., 2023), and contractual issues (Erfani, Tavakolan, et al., 2021; Erfani, Zhang, et al., 2021; Erfani & Tavakolan, 2020). It involves several key steps, starting with the definition of a train and test set.

Train and Test Set: The first step in building a machine learning model is splitting the available data into two subsets: the training set and the test set. Typically, this division is done with a ratio of 70/30 or 80/20, where 70% or 80% of the data is used for training, and the remaining 30% or 20% is used for testing. The training set is used to train the model, while the test set is used to evaluate its performance. This division helps ensure that the model's effectiveness is assessed on unseen data, simulating how it might perform in the real world.

Model Development: The model development process begins once the data is divided. This involves selecting an appropriate algorithm or set of algorithms based on the nature of the problem and the type of data available. Different algorithms are suited for classification, regression, or clustering tasks.

Training the Model: With the algorithm chosen, the model is trained using the data in the training set. During training, the model learns the underlying patterns and relationships in the data. This typically involves adjusting the model's parameters iteratively to minimize the difference between its predictions and the actual outcomes in the training data.

Evaluation of Test Set: The model's performance is evaluated using the test set after training. This involves making predictions on the test data and comparing them to the actual outcomes. Common evaluation metrics include accuracy, precision, recall, and F1 score for classification tasks and mean squared error or R-squared for regression tasks.

Fine-Tuning and Validation: Further adjustments may be made Depending on the model's performance on the test set. This could involve fine-tuning hyperparameters, such as learning rate or regularization strength, or selecting different features or algorithms. It's important to validate the model on separate validation data to avoid overfitting, where the model performs well on the training data but poorly on unseen data.

Deployment and Monitoring: Once a satisfactory model is developed and validated, it can be deployed for use in real-world applications. However, the process doesn't end there; models should be continually monitored and updated as new data becomes available or as the underlying patterns in the data change over time.

2.4 Evaluation Metrics

Accuracy, precision, recall, and ROC-AUC (i.e., Receiver Operating Characteristic – Area Under the Curve) are used to evaluate the performance and effectiveness of different models. The ROC-AUC metric is particularly valuable when dealing with imbalanced datasets, as it measures a model's ability to differentiate between positive and negative samples. Accuracy measures the percentage of correct predictions (Eq. 1), while precision measures the percentage of true positives among the total predicted positives (Eq. 2), and recall measures the percentage of true positives among the actual positives (Eq. 3). Overall, a combination of these metrics can provide a comprehensive evaluation of a model's performance in different classification tasks.

$$Accuracy \ = \frac{TN + TP}{TN + FN + FP + TP} \tag*{1)}$$

$$Precision \ = \frac{TP}{FP + TP} \tag*{2)}$$

$$Sensitivity = Recall \ = \ \frac{TP}{TP + FN} \tag*{3)}$$

where TP, TN, FP, and FN are the true positive, true negative, false positive, and false negative values, respectively, where:

True Positive (TP):

- Definition: In a binary classification task, a true positive (TP) occurs when the model correctly predicts a positive outcome (e.g., severe crash) for an instance that actually belongs to the positive class.

- Example: If the model correctly predicts that a work zone crash resulted in severe injuries, it is considered a true positive.

True Negative (TN):

- Definition: A true negative (TN) occurs when the model correctly predicts a negative outcome (e.g., non-severe crash) for an instance that actually belongs to the negative class.
- Example: If the model correctly predicts that a work zone crash did not result in severe injuries, it is considered a true negative.

False Positive (FP):

- Definition: A false positive (FP) occurs when the model incorrectly predicts a positive outcome (e.g., severe crash) for an instance that actually belongs to the negative class.
- Example: If the model incorrectly predicts that a work zone crash resulted in severe injuries when it did not, it is considered a false positive.

False Negative (FN):

- Definition: A false negative (FN) occurs when the model incorrectly predicts a negative outcome (e.g., non-severe crash) for an instance that actually belongs to the positive class.
- Example: If the model incorrectly predicts that a work zone crash did not result in severe injuries when it did, it is considered a false negative.

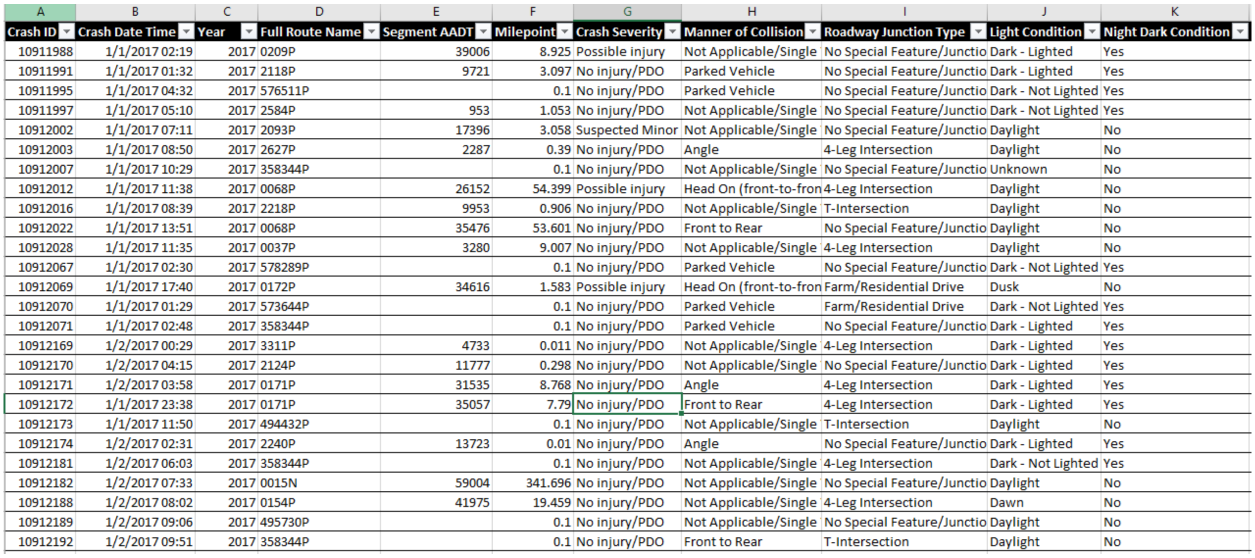
3.0 DATA COLLECTION

3.1 Overview

In this project, two distinct datasets were utilized for comprehensive data analysis. The first dataset consisted of crash data obtained from Numetric, a reliable source of transportation data. The second dataset encompassed work zone data collected from Masterworks, a comprehensive platform that manages and tracks information related to construction projects. By combining these two datasets, a holistic view of the interactions between work zones and crashes could be achieved, facilitating a comprehensive analysis of the factors influencing crash severity and frequency within work zones.

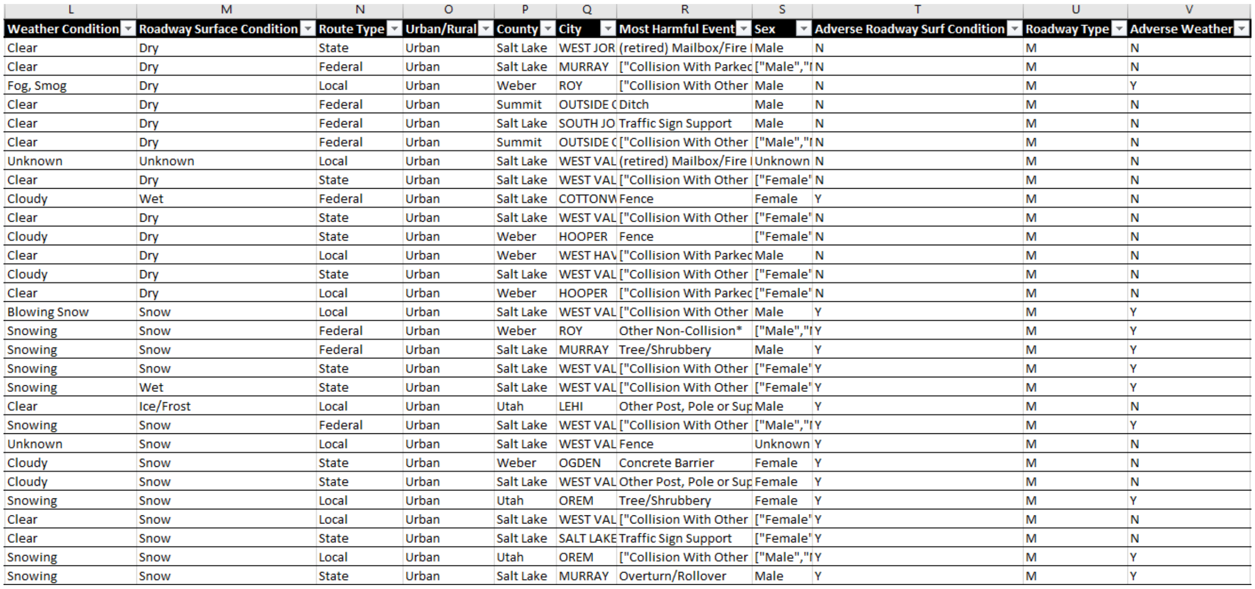
3.2 Crash Data

The crash dataset used in this study comprised over 300,000 crashes from the state of Utah, spanning from 2017 to 2021. It included an extensive set of features, more than 80 variables, capturing various aspects of the crashes. These features encompassed a wide range of information, including demographic details of the involved parties, road and weather conditions, crash types, contributing factors, vehicle attributes, and injury severity levels. The dataset provided a comprehensive and detailed representation of the crashes, enabling a comprehensive analysis of the factors influencing crash outcomes. The extensive feature set allowed for a comprehensive exploration of the relationships and interactions between different variables and their impact on crash severity and frequency. Considering such a diverse range of features, this study aimed to provide a thorough understanding of the complex dynamics associated with crashes in Utah. Sample examples of the dataset are shown below.



A	B	C	D	E	F	G	H	I	J	K
Crash ID	Crash Date Time	Year	Full Route Name	Segment AADT	Milepoint	Crash Severity	Manner of Collision	Roadway Junction Type	Light Condition	Night Dark Condition
10911988	1/1/2017 02:19	2017	0209P	39006	8.925	Possible injury	Not Applicable/Single	No Special Feature/Junctio	Dark - Lighted	Yes
10911991	1/1/2017 01:32	2017	2118P	9721	3.097	No injury/PDO	Parked Vehicle	No Special Feature/Junctio	Dark - Lighted	Yes
10911995	1/1/2017 04:32	2017	576511P		0.1	No injury/PDO	Parked Vehicle	No Special Feature/Junctio	Dark - Not Lighted	Yes
10911997	1/1/2017 05:10	2017	2584P	953	1.053	No injury/PDO	Not Applicable/Single	No Special Feature/Junctio	Dark - Not Lighted	Yes
10912002	1/1/2017 07:11	2017	2093P	17396	3.058	Suspected Minor	Not Applicable/Single	No Special Feature/Junctio	Daylight	No
10912003	1/1/2017 08:50	2017	2627P	2287	0.39	No injury/PDO	Angle	4-Leg Intersection	Daylight	No
10912007	1/1/2017 10:29	2017	358344P		0.1	No injury/PDO	Not Applicable/Single	No Special Feature/Junctio	Unknown	No
10912012	1/1/2017 11:38	2017	0068P	26152	54.399	Possible injury	Head On (front-to-fron	4-Leg Intersection	Daylight	No
10912016	1/1/2017 08:39	2017	2218P	9953	0.906	No injury/PDO	Not Applicable/Single	T-Intersection	Daylight	No
10912022	1/1/2017 13:51	2017	0068P	35476	53.601	No injury/PDO	Front to Rear	No Special Feature/Junctio	Daylight	No
10912028	1/1/2017 11:35	2017	0037P	3280	9.007	No injury/PDO	Not Applicable/Single	4-Leg Intersection	Daylight	No
10912067	1/1/2017 02:30	2017	578289P		0.1	No injury/PDO	Parked Vehicle	No Special Feature/Junctio	Dark - Not Lighted	Yes
10912069	1/1/2017 17:40	2017	0172P	34616	1.583	Possible injury	Head On (front-to-fron	Farm/Residential Drive	Dusk	No
10912070	1/1/2017 01:29	2017	573644P		0.1	No injury/PDO	Parked Vehicle	Farm/Residential Drive	Dark - Not Lighted	Yes
10912071	1/1/2017 02:48	2017	358344P		0.1	No injury/PDO	Parked Vehicle	No Special Feature/Junctio	Dark - Lighted	Yes
10912169	1/2/2017 00:29	2017	3311P	4733	0.011	No injury/PDO	Not Applicable/Single	4-Leg Intersection	Dark - Lighted	Yes
10912170	1/2/2017 04:15	2017	2124P	11777	0.298	No injury/PDO	Not Applicable/Single	No Special Feature/Junctio	Dark - Lighted	Yes
10912171	1/2/2017 03:58	2017	0171P	31535	8.768	No injury/PDO	Angle	4-Leg Intersection	Dark - Lighted	Yes
10912172	1/1/2017 23:38	2017	0171P	35057	7.79	No injury/PDO	Front to Rear	4-Leg Intersection	Dark - Lighted	Yes
10912173	1/1/2017 11:50	2017	494432P		0.1	No injury/PDO	Not Applicable/Single	T-Intersection	Daylight	No
10912174	1/2/2017 02:31	2017	2240P	13723	0.01	No injury/PDO	Angle	No Special Feature/Junctio	Dark - Lighted	Yes
10912181	1/2/2017 06:03	2017	358344P		0.1	No injury/PDO	Not Applicable/Single	4-Leg Intersection	Dark - Not Lighted	Yes
10912182	1/2/2017 07:33	2017	0015N	59004	341.696	No injury/PDO	Not Applicable/Single	No Special Feature/Junctio	Daylight	No
10912188	1/2/2017 08:02	2017	0154P	41975	19.459	No injury/PDO	Not Applicable/Single	4-Leg Intersection	Dawn	No
10912189	1/2/2017 09:06	2017	495730P		0.1	No injury/PDO	Not Applicable/Single	No Special Feature/Junctio	Daylight	No
10912192	1/2/2017 09:51	2017	358344P		0.1	No injury/PDO	Front to Rear	T-Intersection	Daylight	No

Figure 7. Features of the Crash Dataset (Part I)

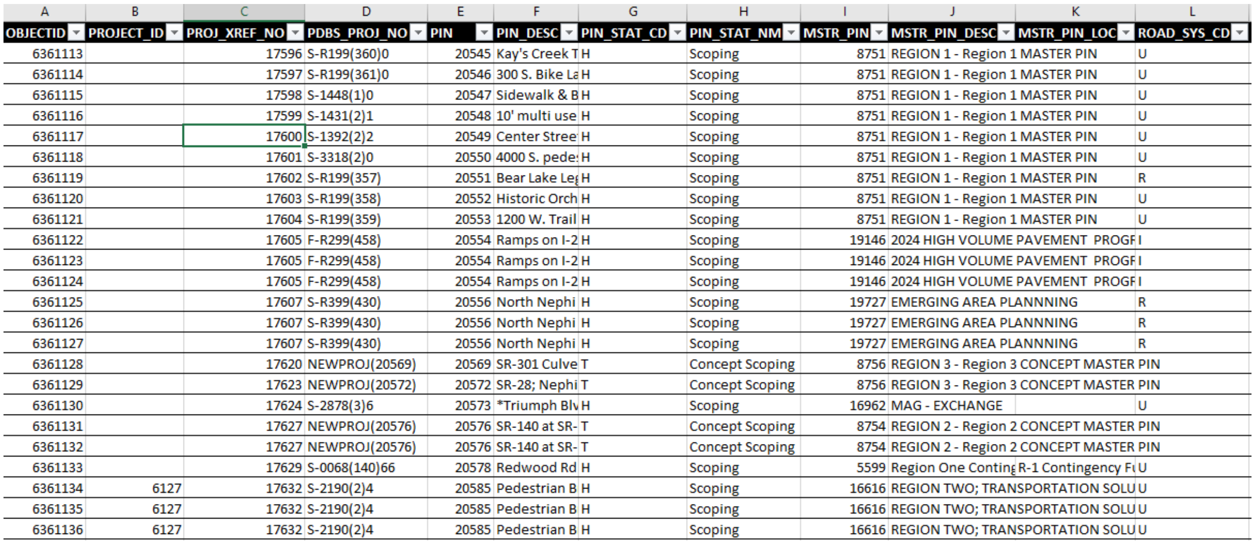


L	M	N	O	P	Q	R	S	T	U	V
Weather Condition	Roadway Surface Condition	Route Type	Urban/Rural	County	City	Most Harmful Event	Sex	Adverse Roadway Surf Condition	Roadway Type	Adverse Weather
Clear	Dry	State	Urban	Salt Lake	WEST JOR	(retired) Mailbox/Fire	Male	N	M	N
Clear	Dry	Federal	Urban	Salt Lake	MURRAY	["Collision With Parkec	["Male","	N	M	N
Fog, Smog	Dry	Local	Urban	Weber	ROY	["Collision With Other	Male	N	M	Y
Clear	Dry	Federal	Urban	Summit	OUTSIDE (	Ditch	Male	N	M	N
Clear	Dry	Federal	Urban	Salt Lake	SOUTH JO	Traffic Sign Support	Male	N	M	N
Clear	Dry	Federal	Urban	Summit	OUTSIDE (	["Collision With Other	["Male","	N	M	N
Unknown	Unknown	Local	Urban	Salt Lake	WEST VAL	(retired) Mailbox/Fire	Unknown	N	M	N
Clear	Dry	State	Urban	Salt Lake	WEST VAL	["Collision With Other	["Female"	N	M	N
Cloudy	Wet	Federal	Urban	Salt Lake	COTTONW	Fence	Female	Y	M	N
Clear	Dry	State	Urban	Salt Lake	WEST VAL	["Collision With Other	["Female"	N	M	N
Cloudy	Dry	State	Urban	Weber	HOOPER	Fence	["Female"	N	M	N
Clear	Dry	Local	Urban	Weber	WEST HAV	["Collision With Parkec	Male	N	M	N
Cloudy	Dry	State	Urban	Salt Lake	WEST VAL	["Collision With Other	["Female"	N	M	N
Clear	Dry	Local	Urban	Weber	HOOPER	["Collision With Parkec	["Female"	N	M	N
Blowing Snow	Snow	Local	Urban	Salt Lake	WEST VAL	["Collision With Other	Male	Y	M	Y
Snowing	Snow	Federal	Urban	Weber	ROY	Other Non-Collision*	["Male","	Y	M	Y
Snowing	Snow	Federal	Urban	Salt Lake	MURRAY	Tree/Shrubbery	Male	Y	M	Y
Snowing	Snow	State	Urban	Salt Lake	WEST VAL	["Collision With Other	["Female"	Y	M	Y
Snowing	Wet	State	Urban	Salt Lake	WEST VAL	["Collision With Other	["Female"	Y	M	Y
Clear	Ice/Frost	Local	Urban	Utah	LEHI	Other Post, Pole or Sup	Male	Y	M	N
Snowing	Snow	Federal	Urban	Salt Lake	WEST VAL	["Collision With Other	["Male","	Y	M	Y
Unknown	Snow	Local	Urban	Salt Lake	WEST VAL	Fence	Unknown	Y	M	N
Cloudy	Snow	State	Urban	Weber	OGDEN	Concrete Barrier	Female	Y	M	N
Cloudy	Snow	State	Urban	Salt Lake	WEST VAL	Other Post, Pole or Sup	Female	Y	M	N
Snowing	Snow	Local	Urban	Utah	OREM	Tree/Shrubbery	Female	Y	M	Y
Clear	Snow	Local	Urban	Salt Lake	WEST VAL	["Collision With Other	["Female"	Y	M	N
Clear	Snow	State	Urban	Salt Lake	SALT LAKE	Traffic Sign Support	["Female"	Y	M	N
Snowing	Snow	Local	Urban	Utah	OREM	["Collision With Other	["Male","	Y	M	Y
Snowing	Snow	State	Urban	Salt Lake	MURRAY	Overturn/Rollover	Male	Y	M	Y

Figure 8. Features of the Crash Dataset (Part II)

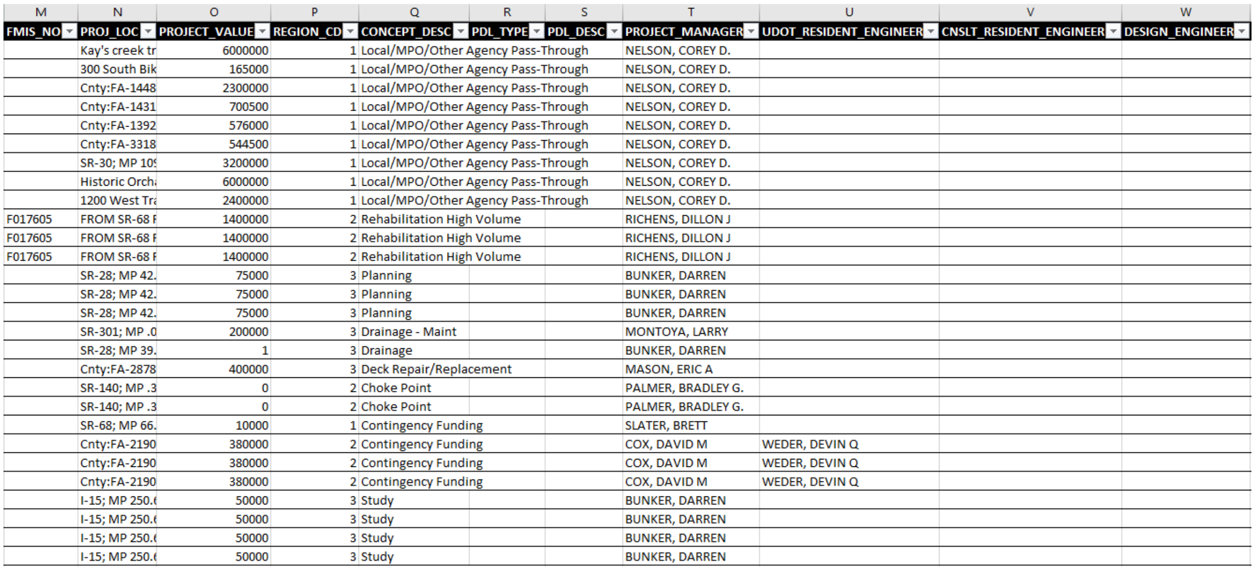
3.3 Work Zone Data

This study utilized work zone data from the state of Utah spanning from 2017 to 2021. The dataset was obtained from Masterworks, a database maintained by the Utah Department of Transportation (UDOT) that stores work zone data along with other traffic-related information. The UDOT databases are regularly updated to reflect the latest work zone configurations and conditions. Crashes associated with specific work zones were identified by cross-referencing the work zone dataset with the Numetric dataset. This cross-referencing was achieved by matching the location and date of each crash with the corresponding work zone information in the dataset. It allowed for a comprehensive analysis of the relationship between work zones and safety conditions, providing valuable insights into the impact of work zones on crash occurrences and severity. It is worth noting that certain attributes deemed irrelevant to the analysis of road safety conditions, such as the contractor, project cost, and engineering company, were excluded from further consideration to focus on factors directly related to crash outcomes. Sample examples of the dataset are shown below.



A	B	C	D	E	F	G	H	I	J	K	L
OBJECTID	PROJECT_ID	PROJ_XREF_NO	PDBS_PROJ_NO	PIN	PIN_DESC	PIN_STAT_CD	PIN_STAT_NM	MSTR_PIN	MSTR_PIN_DESC	MSTR_PIN_LOC	ROAD_SYS_CD
6361113		17596	S-R199(360)0	20545	Kay's Creek T	H	Scoping	8751	REGION 1 - Region 1 MASTER PIN		U
6361114		17597	S-R199(361)0	20546	300 S. Bike La	H	Scoping	8751	REGION 1 - Region 1 MASTER PIN		U
6361115		17598	S-1448(1)0	20547	Sidewalk & B	H	Scoping	8751	REGION 1 - Region 1 MASTER PIN		U
6361116		17599	S-1431(2)1	20548	10' multi use	H	Scoping	8751	REGION 1 - Region 1 MASTER PIN		U
6361117		17600	S-1392(2)2	20549	Center Stree	H	Scoping	8751	REGION 1 - Region 1 MASTER PIN		U
6361118		17601	S-3318(2)0	20550	4000 S. pede	H	Scoping	8751	REGION 1 - Region 1 MASTER PIN		U
6361119		17602	S-R199(357)	20551	Bear Lake Leg	H	Scoping	8751	REGION 1 - Region 1 MASTER PIN		R
6361120		17603	S-R199(358)	20552	Historic Orch	H	Scoping	8751	REGION 1 - Region 1 MASTER PIN		U
6361121		17604	S-R199(359)	20553	1200 W. Trail	H	Scoping	8751	REGION 1 - Region 1 MASTER PIN		U
6361122		17605	F-R299(458)	20554	Ramps on I-2	H	Scoping	19146	2024 HIGH VOLUME PAVEMENT PROGF		I
6361123		17605	F-R299(458)	20554	Ramps on I-2	H	Scoping	19146	2024 HIGH VOLUME PAVEMENT PROGF		I
6361124		17605	F-R299(458)	20554	Ramps on I-2	H	Scoping	19146	2024 HIGH VOLUME PAVEMENT PROGF		I
6361125		17607	S-R399(430)	20556	North Nephi	H	Scoping	19727	EMERGING AREA PLANNNING		R
6361126		17607	S-R399(430)	20556	North Nephi	H	Scoping	19727	EMERGING AREA PLANNNING		R
6361127		17607	S-R399(430)	20556	North Nephi	H	Scoping	19727	EMERGING AREA PLANNNING		R
6361128		17620	NEWPROJ(20569)	20569	SR-301 Culve	T	Concept Scoping	8756	REGION 3 - Region 3 CONCEPT MASTER PIN		
6361129		17623	NEWPROJ(20572)	20572	SR-28; Nephi	T	Concept Scoping	8756	REGION 3 - Region 3 CONCEPT MASTER PIN		
6361130		17624	S-2878(3)6	20573	*Triumph Blv	H	Scoping	16962	MAG - EXCHANGE		U
6361131		17627	NEWPROJ(20576)	20576	SR-140 at SR-	T	Concept Scoping	8754	REGION 2 - Region 2 CONCEPT MASTER PIN		
6361132		17627	NEWPROJ(20576)	20576	SR-140 at SR-	T	Concept Scoping	8754	REGION 2 - Region 2 CONCEPT MASTER PIN		
6361133		17629	S-0068(140)66	20578	Redwood Rd	H	Scoping	5599	Region One Conting	R-1 Contingency Fu	U
6361134	6127	17632	S-2190(2)4	20585	Pedestrian B	H	Scoping	16616	REGION TWO; TRANSPORTATION SOLU		U
6361135	6127	17632	S-2190(2)4	20585	Pedestrian B	H	Scoping	16616	REGION TWO; TRANSPORTATION SOLU		U
6361136	6127	17632	S-2190(2)4	20585	Pedestrian B	H	Scoping	16616	REGION TWO; TRANSPORTATION SOLU		U

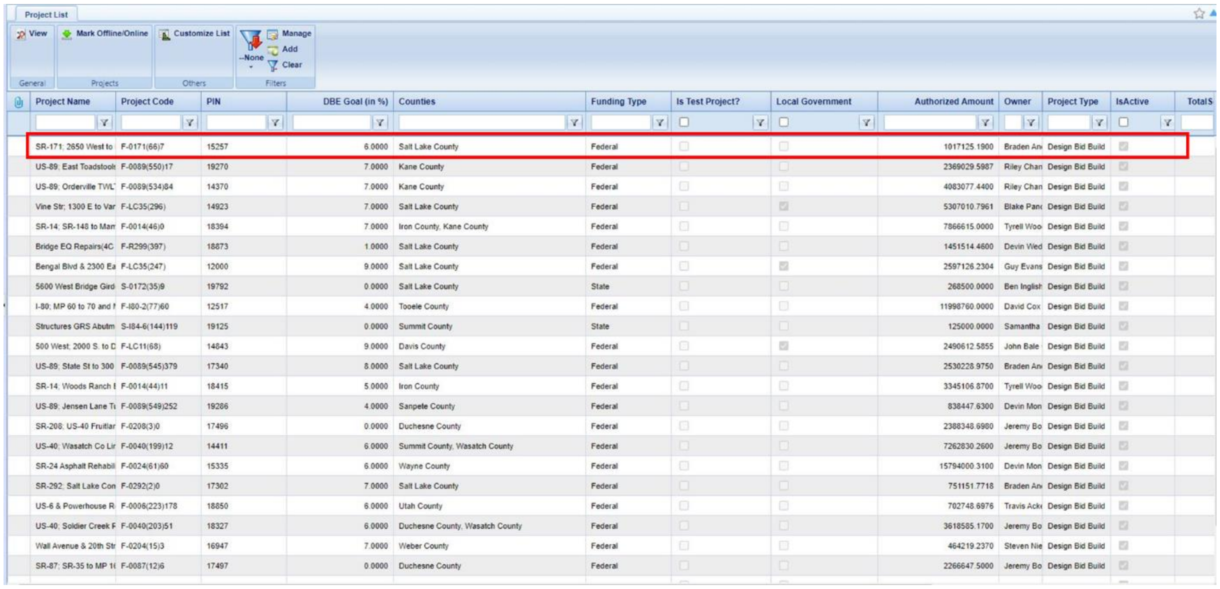
Figure 9. Features of the Work Zone Dataset (Part I)



M	N	O	P	Q	R	S	T	U	V	W
FMIS_NO	PROJ_LOC	PROJECT_VALUE	REGION_CD	CONCEPT_DESC	PDL_TYPE	PDL_DESC	PROJECT_MANAGER	UDOT_RESIDENT_ENGINEER	CNSLT_RESIDENT_ENGINEER	DESIGN_ENGINEER
	Kay's creek tr	6000000	1	Local/MPO/Other Agency Pass-Through			NELSON, COREY D.			
	300 South Bik	165000	1	Local/MPO/Other Agency Pass-Through			NELSON, COREY D.			
	Cnty:FA-1448	2300000	1	Local/MPO/Other Agency Pass-Through			NELSON, COREY D.			
	Cnty:FA-1431	700500	1	Local/MPO/Other Agency Pass-Through			NELSON, COREY D.			
	Cnty:FA-1392	576000	1	Local/MPO/Other Agency Pass-Through			NELSON, COREY D.			
	Cnty:FA-3318	544500	1	Local/MPO/Other Agency Pass-Through			NELSON, COREY D.			
	SR-30; MP 10	3200000	1	Local/MPO/Other Agency Pass-Through			NELSON, COREY D.			
	Historic Orch	6000000	1	Local/MPO/Other Agency Pass-Through			NELSON, COREY D.			
	1200 West Tr	2400000	1	Local/MPO/Other Agency Pass-Through			NELSON, COREY D.			
F017605	FROM SR-68 F	1400000	2	Rehabilitation High Volume			RICHENS, DILLON J			
F017605	FROM SR-68 F	1400000	2	Rehabilitation High Volume			RICHENS, DILLON J			
F017605	FROM SR-68 F	1400000	2	Rehabilitation High Volume			RICHENS, DILLON J			
	SR-28; MP 42.	75000	3	Planning			BUNKER, DARREN			
	SR-28; MP 42.	75000	3	Planning			BUNKER, DARREN			
	SR-28; MP 42.	75000	3	Planning			BUNKER, DARREN			
	SR-301; MP .0	200000	3	Drainage - Maint			MONTOYA, LARRY			
	SR-28; MP 39.	1	3	Drainage			BUNKER, DARREN			
	Cnty:FA-2878	400000	3	Deck Repair/Replacement			MASON, ERIC A			
	SR-140; MP .3	0	2	Choke Point			PALMER, BRADLEY G.			
	SR-140; MP .3	0	2	Choke Point			PALMER, BRADLEY G.			
	SR-68; MP 66.	10000	1	Contingency Funding			SLATER, BRETT			
	Cnty:FA-2190	380000	2	Contingency Funding			COX, DAVID M	WEDER, DEVIN Q		
	Cnty:FA-2190	380000	2	Contingency Funding			COX, DAVID M	WEDER, DEVIN Q		
	Cnty:FA-2190	380000	2	Contingency Funding			COX, DAVID M	WEDER, DEVIN Q		
	I-15; MP 250.	50000	3	Study			BUNKER, DARREN			
	I-15; MP 250.	50000	3	Study			BUNKER, DARREN			
	I-15; MP 250.	50000	3	Study			BUNKER, DARREN			
	I-15; MP 250.	50000	3	Study			BUNKER, DARREN			

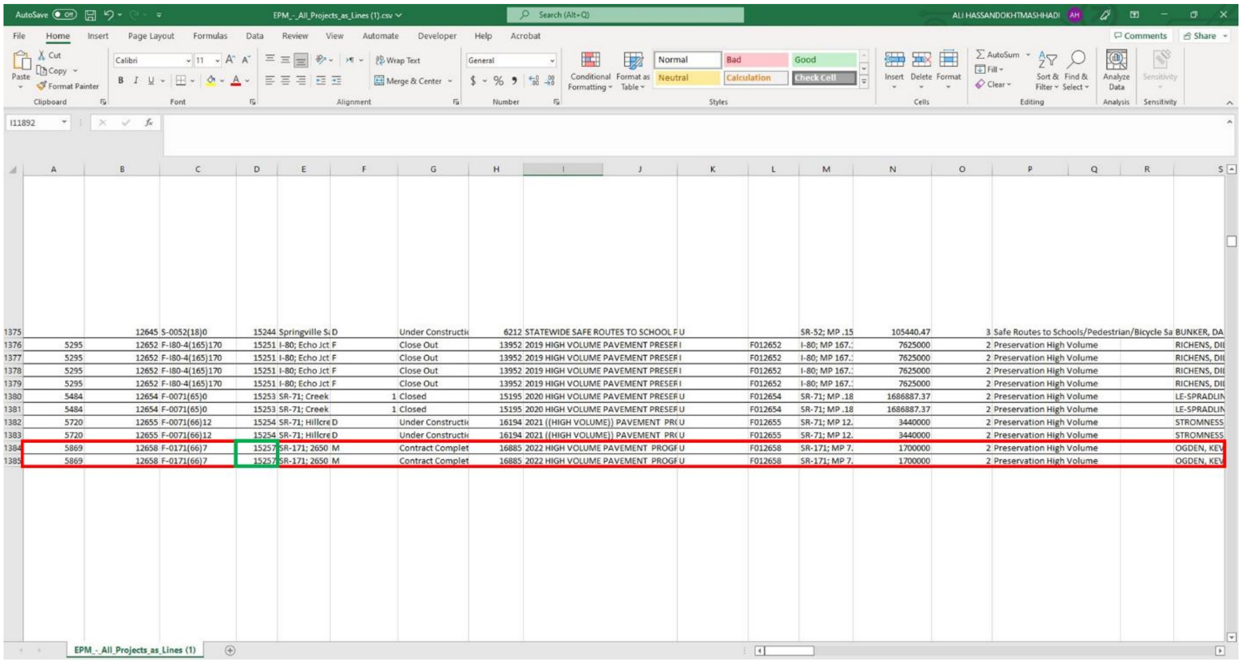
Figure 10. Features of the Work Zone Dataset (Part II)

Among the three available resources, Incident Data, ProjectWise, and Masterworks, the latter is the most useful one in extracting lane closure activities. Also, the results of cross-referencing information from ePM (Electronic Program Management) and Masterworks show the consistency of the two resources.



Project Name	Project Code	PIN	DBE Goal (in %)	Counties	Funding Type	Is Test Project?	Local Government	Authorized Amount	Owner	Project Type	IsActive	TotalS
SR-171; 2650 West to	F-0171(66)7	15257	6.0000	Salt Lake County	Federal	☐	☐	1017125.1900	Braden An	Design Bid Build	☑	
US-89; East Toadstool	F-0089(550)17	19270	7.0000	Kane County	Federal	☐	☐	2369029.5987	Riley Chan	Design Bid Build	☑	
US-89; Orderville TWL	F-0089(534)84	14370	7.0000	Kane County	Federal	☐	☐	4083077.4400	Riley Chan	Design Bid Build	☑	
Vine Str; 1300 E to Var	F-LC35(296)	14923	7.0000	Salt Lake County	Federal	☐	☑	5307010.7961	Blake Pan	Design Bid Build	☑	
SR-14; SR-148 to Mar	F-0014(46)0	18394	7.0000	Iron County, Kane County	Federal	☐	☐	7866615.0000	Tyrell Woo	Design Bid Build	☑	
Bridge EQ Repairs(4C	F-R299(397)	18873	1.0000	Salt Lake County	Federal	☐	☐	1451514.4600	Devin Wed	Design Bid Build	☑	
Bengal Blvd & 2300 Ea	F-LC35(247)	12000	9.0000	Salt Lake County	Federal	☐	☑	2597126.2304	Guy Evans	Design Bid Build	☑	
5600 West Bridge Gird	S-0172(35)9	19792	0.0000	Salt Lake County	State	☐	☐	268500.0000	Ben Inglish	Design Bid Build	☑	
I-80; MP 60 to 70 and	F-I80-2(77)60	12517	4.0000	Tooele County	Federal	☐	☐	11998760.0000	David Cox	Design Bid Build	☑	
Structures GRS Abutm	S-I84-6(144)119	19125	0.0000	Summit County	State	☐	☐	125000.0000	Samantha	Design Bid Build	☑	
500 West; 2000 S. to D	F-LC11(68)	14843	9.0000	Davis County	Federal	☐	☑	2490612.5855	John Bale	Design Bid Build	☑	
US-89; State St to 300	F-0089(545)379	17340	8.0000	Salt Lake County	Federal	☐	☐	2530228.9750	Braden An	Design Bid Build	☑	
SR-14; Woods Ranch	F-0014(44)11	18415	5.0000	Iron County	Federal	☐	☐	3345106.8700	Tyrell Woo	Design Bid Build	☑	
US-89; Jensen Lane T	F-0089(549)252	19286	4.0000	Sanpete County	Federal	☐	☐	838447.6300	Devin Mon	Design Bid Build	☑	
SR-208; US-40 Fruitlar	F-0208(3)0	17496	0.0000	Duchesne County	Federal	☐	☐	2388348.6980	Jeremy Bo	Design Bid Build	☑	
US-40; Wasatch Co Lir	F-0040(199)12	14411	6.0000	Summit County, Wasatch County	Federal	☐	☐	7262830.2600	Jeremy Bo	Design Bid Build	☑	
SR-24 Asphalt Rehabil	F-0024(61)60	15335	6.0000	Wayne County	Federal	☐	☐	15794000.3100	Devin Mon	Design Bid Build	☑	
SR-292; Salt Lake Con	F-0292(2)0	17302	7.0000	Salt Lake County	Federal	☐	☐	751151.7718	Braden An	Design Bid Build	☑	
US-6 & Powerhouse R	F-0006(223)178	18850	6.0000	Utah County	Federal	☐	☐	702748.6976	Travis Ack	Design Bid Build	☑	
US-40; Soldier Creek F	F-0040(203)51	18327	6.0000	Duchesne County, Wasatch County	Federal	☐	☐	3618585.1700	Jeremy Bo	Design Bid Build	☑	
Wall Avenue & 20th Str	F-0204(15)3	16947	7.0000	Weber County	Federal	☐	☐	464219.2370	Steven Nie	Design Bid Build	☑	
SR-87; SR-35 to MP 1	F-0087(12)6	17497	0.0000	Duchesne County	Federal	☐	☐	2266647.5000	Jeremy Bo	Design Bid Build	☑	

Figure 11. Masterworks Interface, Including Project Information



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1375		12645	S-0052(18)0	15244	Springville S	D	Under Constructi	6212	STATEWIDE SAFE ROUTES TO SCHOOL F		U		SR-52; MP .15	105440.47	3	Safe Routes to Schools/Pedestrian/Bicycle Sa		BUNKER, DA	
1376	5295	12652	F-I80-4(165)170	15251	I-80; Echo Jct	F	Close Out	13952	2019 HIGH VOLUME PAVEMENT PRESER		I	F012652	I-80; MP 167.	7625000	2	Preservation High Volume		RICHENS, DI	
1377	5295	12652	F-I80-4(165)170	15251	I-80; Echo Jct	F	Close Out	13952	2019 HIGH VOLUME PAVEMENT PRESER		I	F012652	I-80; MP 167.	7625000	2	Preservation High Volume		RICHENS, DI	
1378	5295	12652	F-I80-4(165)170	15251	I-80; Echo Jct	F	Close Out	13952	2019 HIGH VOLUME PAVEMENT PRESER		I	F012652	I-80; MP 167.	7625000	2	Preservation High Volume		RICHENS, DI	
1379	5295	12652	F-I80-4(165)170	15251	I-80; Echo Jct	F	Close Out	13952	2019 HIGH VOLUME PAVEMENT PRESER		I	F012652	I-80; MP 167.	7625000	2	Preservation High Volume		RICHENS, DI	
1380	5484	12654	F-0071(65)0	15253	SR-71; Creek	1	Closed	15195	2020 HIGH VOLUME PAVEMENT PRESER		U	F012654	SR-71; MP .18	1686887.37	2	Preservation High Volume		LE-SPRADLIN	
1381	5484	12654	F-0071(65)0	15253	SR-71; Creek	1	Closed	15195	2020 HIGH VOLUME PAVEMENT PRESER		U	F012654	SR-71; MP .18	1686887.37	2	Preservation High Volume		LE-SPRADLIN	
1382	5720	12655	F-0071(66)12	15254	SR-71; Hillcre	D	Under Constructi	16194	2021 ((HIGH VOLUME)) PAVEMENT PR		U	F012655	SR-71; MP 12.	3440000	2	Preservation High Volume		STROMNESS	
1383	5720	12655	F-0071(66)12	15254	SR-71; Hillcre	D	Under Constructi	16194	2021 ((HIGH VOLUME)) PAVEMENT PR		U	F012655	SR-71; MP 12.	3440000	2	Preservation High Volume		STROMNESS	
1384	5869	12658	F-0171(66)7	15257	SR-171; 2650	M	Contract Complet	16885	2022 HIGH VOLUME PAVEMENT PROGF		U	F012658	SR-171; MP 7.	1700000	2	Preservation High Volume		OGDEN, KEV	
1385	5869	12658	F-0171(66)7	15257	SR-171; 2650	M	Contract Complet	16885	2022 HIGH VOLUME PAVEMENT PROGF		U	F012658	SR-171; MP 7.	1700000	2	Preservation High Volume		OGDEN, KEV	

Figure 12. ePM Database, Including Projects Information

3.4 Summary

This study comprehensively analyzed road safety conditions in work zones using datasets from the state of Utah. The crash dataset, comprising over 300,000 crashes from 2017 to 2021, was cross-referenced with the work zone dataset obtained from Masterworks. By linking crashes to specific work zones based on location and date, the study examined the impact of work zones on crash occurrences and severity. Detailed information from the work zone dataset allowed for identifying influential factors. The study aimed to uncover patterns, identify risk factors, and inform the development of effective safety strategies for work zones through rigorous data collection, cleaning, and analysis using statistical and machine learning models. The findings contribute to enhancing work zone safety management and have the potential to improve road safety outcomes.

4.0 RESULTS AND FINDINGS

4.1 Overview

UDOT provided the research team access to incident data, ProjectWise, and Masterworks. Before processing data, the research team conducted a comprehensive literature review to extract the most influential factors affecting work zone safety. Based on the literature, the following features are among the most influential factors in work zone safety:

1. Daytime/Nighttime
2. Traffic Volume
3. Closed Lane Counts
4. Speeding
5. Road Class
6. Number of Intersections
7. Portable Rumble Strips (PRS) or Rumble Strips
8. Speed Feedback Display
9. Automated Speed-Camera Enforcement
10. Live Police Presence
11. Advanced Information Availability
12. Construction Type
13. Weather (Foggy, Clear)
14. Light Condition
15. Dry/Wet Surface
16. ITS Technologies, such as variable speed limit (VSL) and dynamic message signs (DMS) at an appropriate distance
17. Shoulder Width
18. Work-Zone Types: lane closure, work on shoulder-median

These factors are extracted from more than 20 papers published in recent years.

4.2 Data Analysis

Some initial data analysis has been undertaken on crashes within work zone areas and those without work zones. Figure 13 shows the distribution of work zones and regular crashes in different months. The diagrams reveal fewer work zone crashes by the end of the year, probably due to the limited number of projects happening around the state.

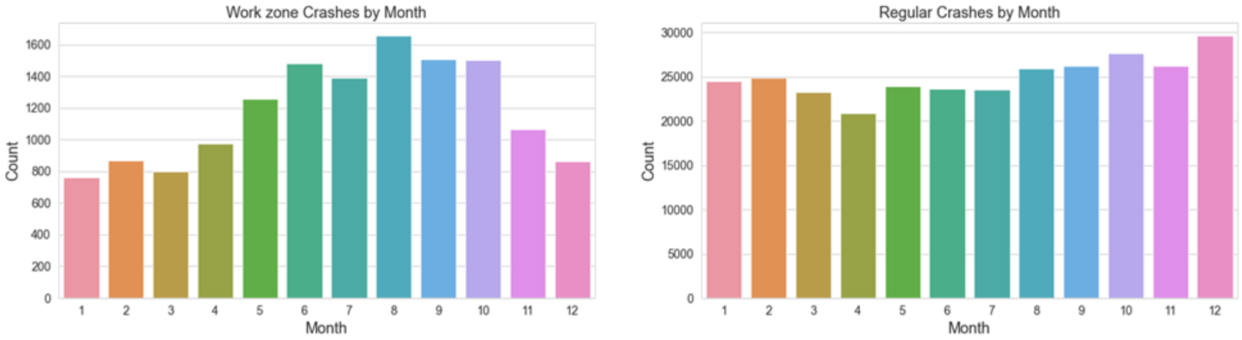


Figure 13. Work Zone and Non-Work Zone Crashes by Month

Figure 14 compares work zone incidents and regular crashes within rural and urban settings. The findings indicate a slight discrepancy in the proportion of rural locations when comparing regular crashes to those occurring in work zones.

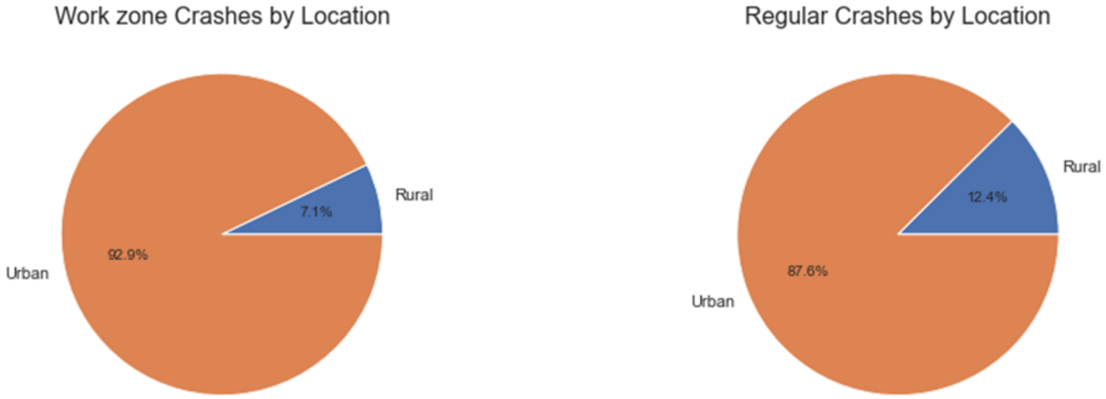


Figure 14. Work Zone and Non-Work Zone Crashes by Location

Additionally, when examining the DUI rates in work zone crashes versus regular crashes, the proportions were found to be nearly identical (Figure 15).

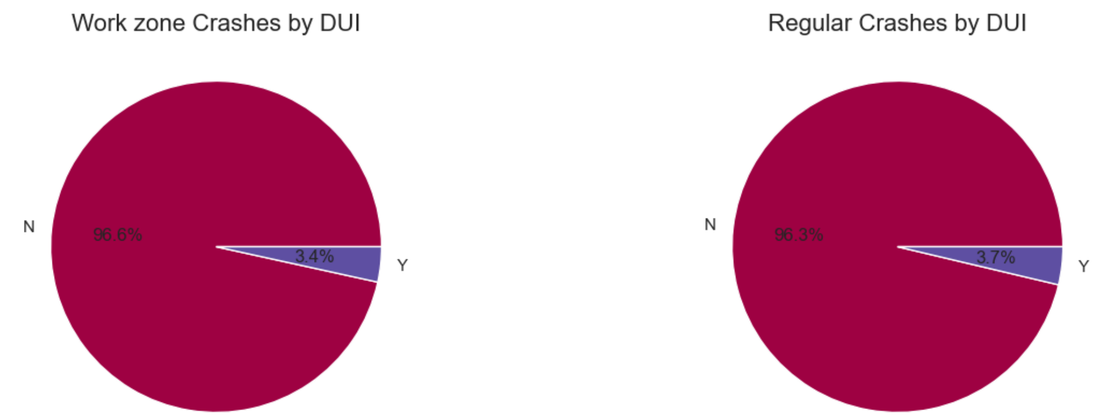


Figure 15. Work Zone and Non-Work Zone Crashes by DUI

When comparing the rate of collisions with fixed objects, work zone crashes, and regular crashes exhibit almost the same frequency.

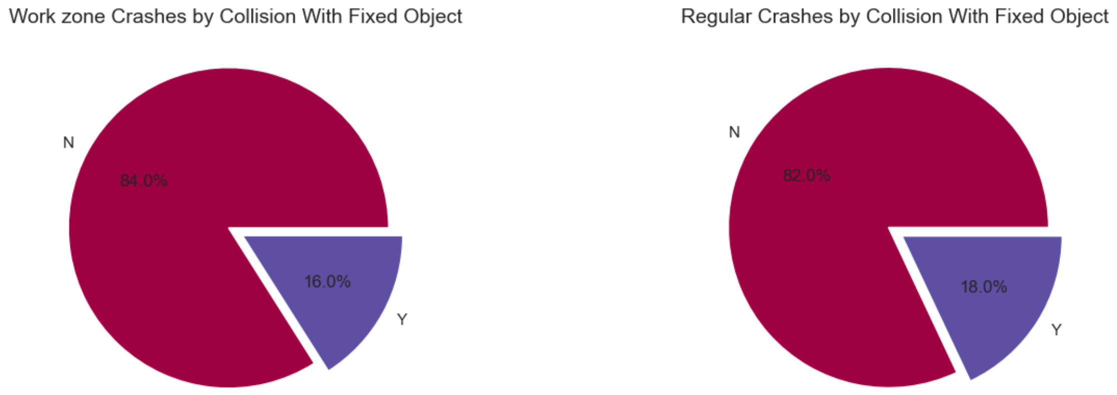


Figure 16. Work Zone and Non-Work Zone Crashes by Collision with Fixed Object

Figure 17 displays the distribution of severity levels for work zones and regular crashes.

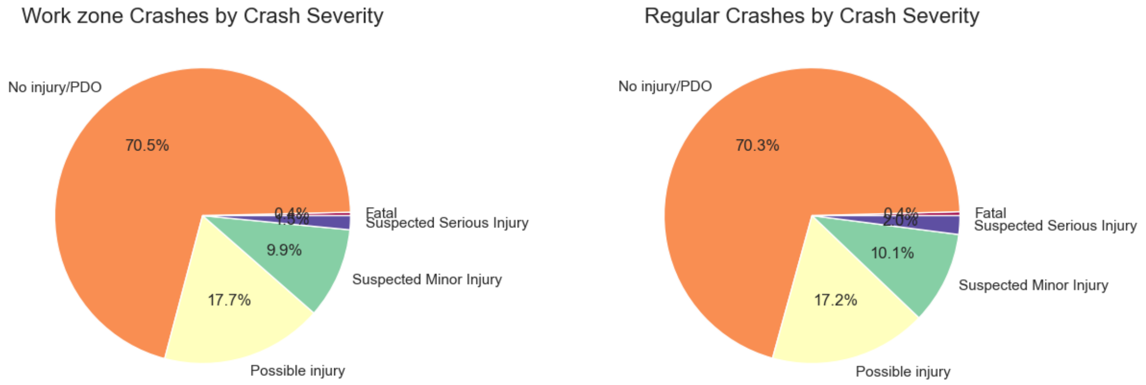


Figure 17. Work Zone and Non-Work Zone Crashes by Crash Severity

Figure 18 compares work zone and regular crashes by weather condition, showing similar rates across different weather conditions.

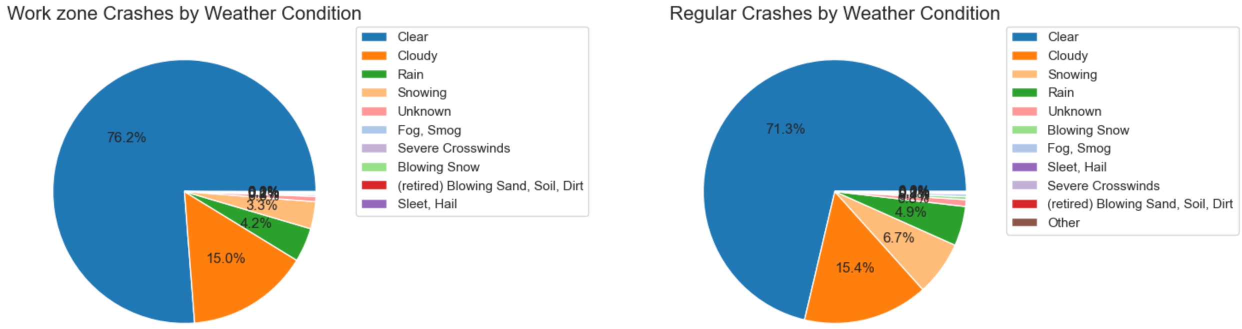


Figure 18. Work Zone and Non-Work Zone Crashes by Weather Condition

Figure 19 illustrates the impact of lighting conditions on work zones and regular crashes.

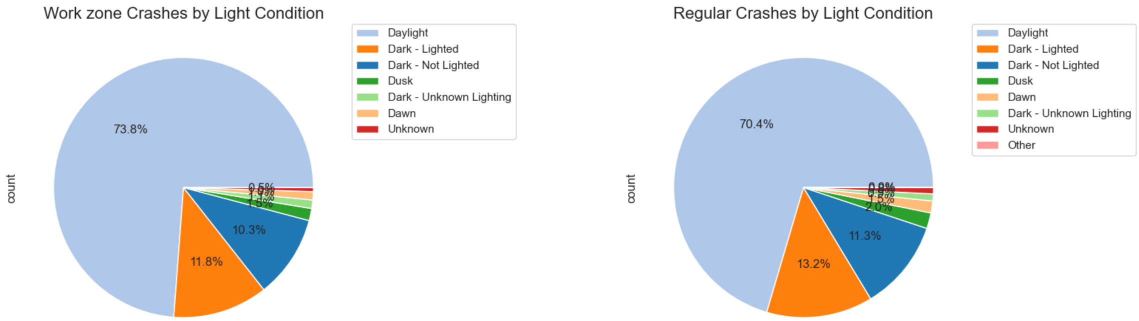


Figure 19. Work Zone and Non-Work Zone Crashes by Light Condition

Figure 20 depicts the influence of surface conditions on work zones and regular crashes.

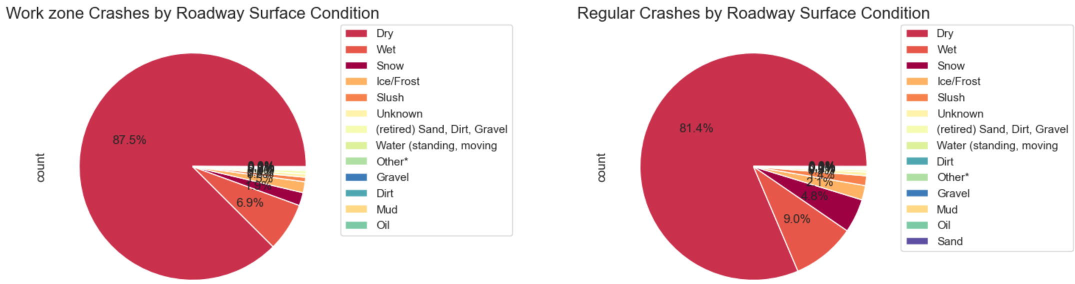


Figure 20. Work Zone and Non-Work Zone Crashes by Surface Condition

Figure 21 showcases the effectiveness of different traffic control approaches in work zones and regular crash scenarios.

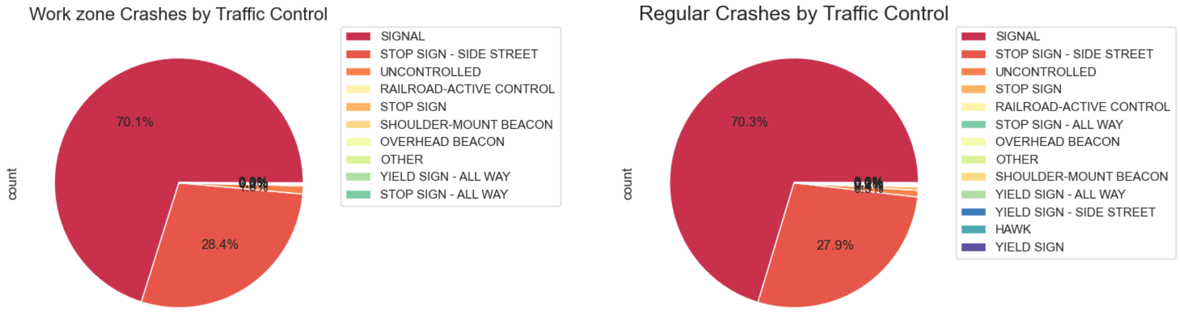


Figure 21. Work Zone and Non-Work Zone Crashes by Traffic Control

Figure 22 illustrates the manner of collision comparison, indicating that work zone crashes have a 10 percent higher rate of front-to-rear collisions attributable to sudden changes in speed.

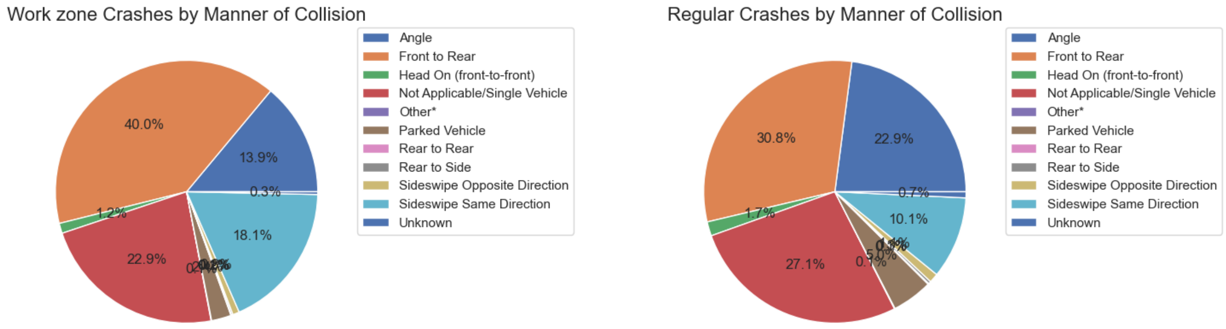


Figure 22. Work Zone and Non-Work Zone Crashes by Manner of Collision

Figure 23 compares crash types in queue zones and regular crashes, revealing a similar pattern as Figure 22.

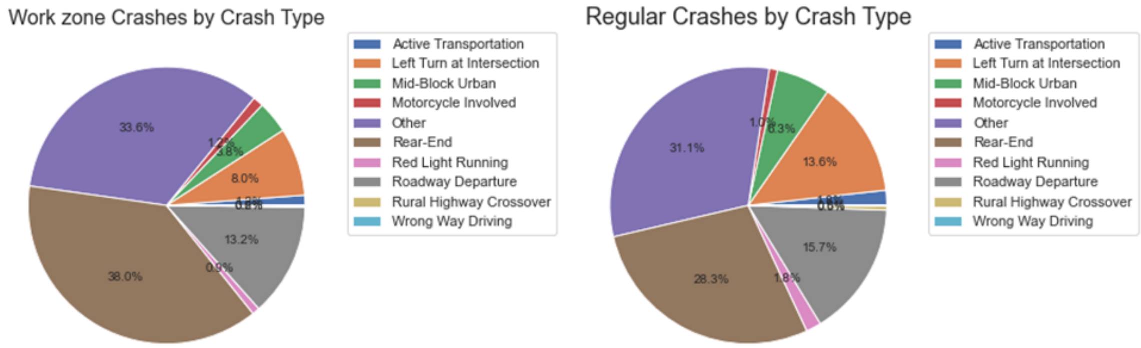


Figure 23. Work Zone and Non-Work Zone Crashes by Crash Type

Figure 24 lists the roads with the highest number of work zones and regular crashes.

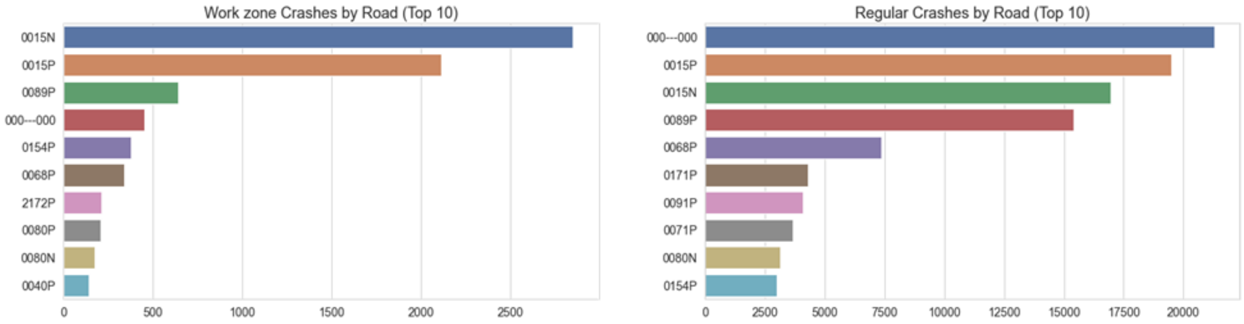


Figure 24. Work Zone and Non-Work Zone Crashes by Road (000—000 refers to crashes where the road name was not recorded)

Figure 25 displays the distribution of work zones and regular crashes along I-15 in the positive (northbound) direction.

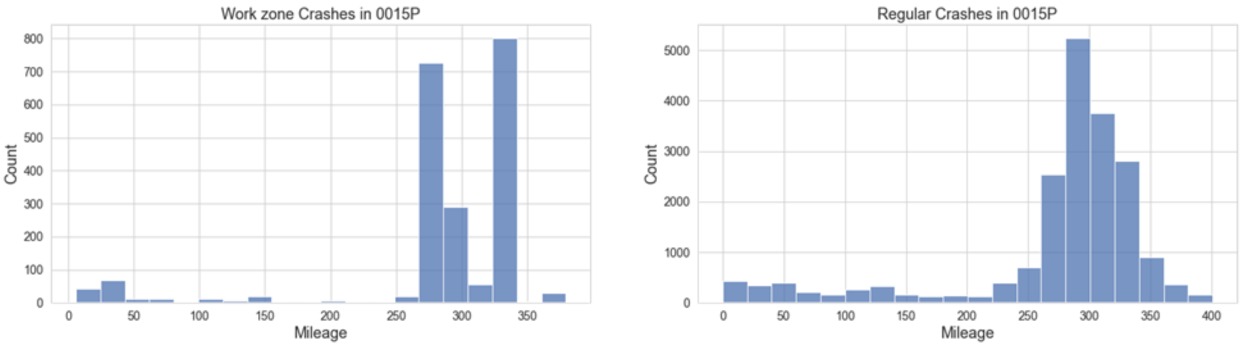


Figure 25. Work Zone and Non-Work Zone Crashes in I-15P

Figure 26 displays the distribution of work zones and regular crash types along I-15 in the positive (northbound) direction.

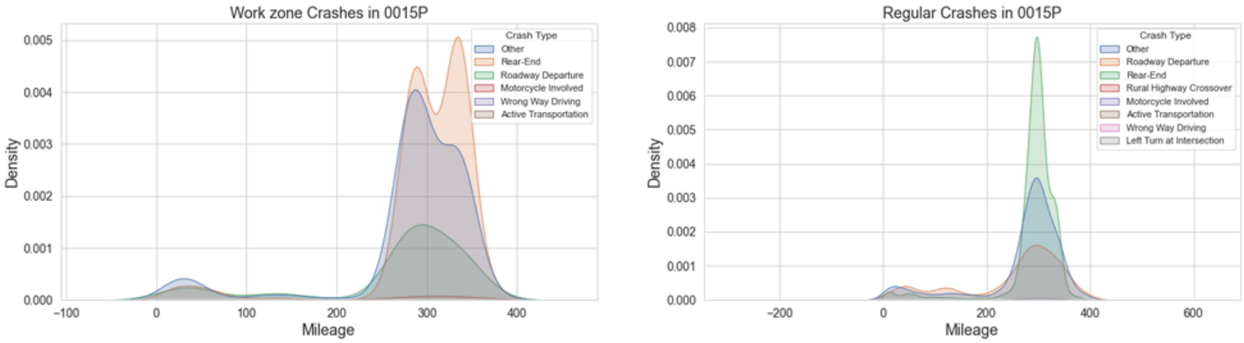


Figure 26. Work Zone and Non-Work Zone Crashes in I-15P

Figure 27 displays the distribution of work zones and regular crashes along I-15 in the negative (southbound) direction.

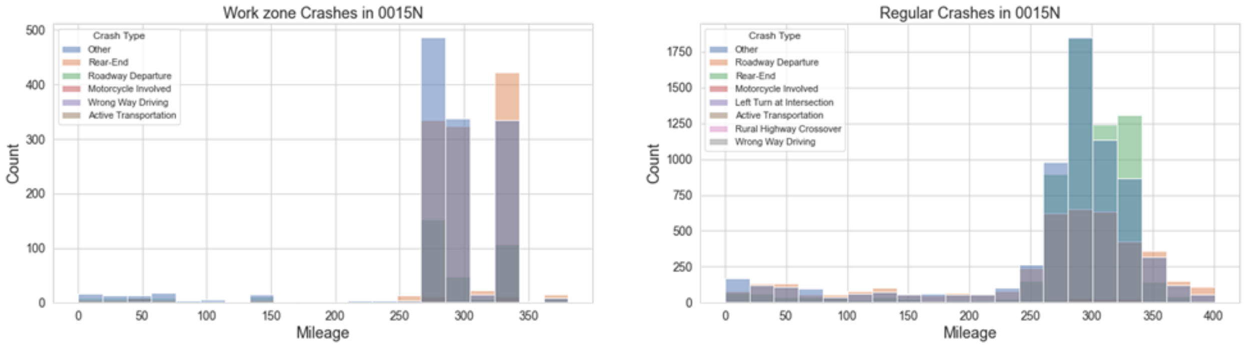


Figure 27. Work Zone and Non-Work Zone Crashes in I-15N

Figure 28 displays the distribution of work zones and regular crash types along I-15 in the negative (southbound) direction.

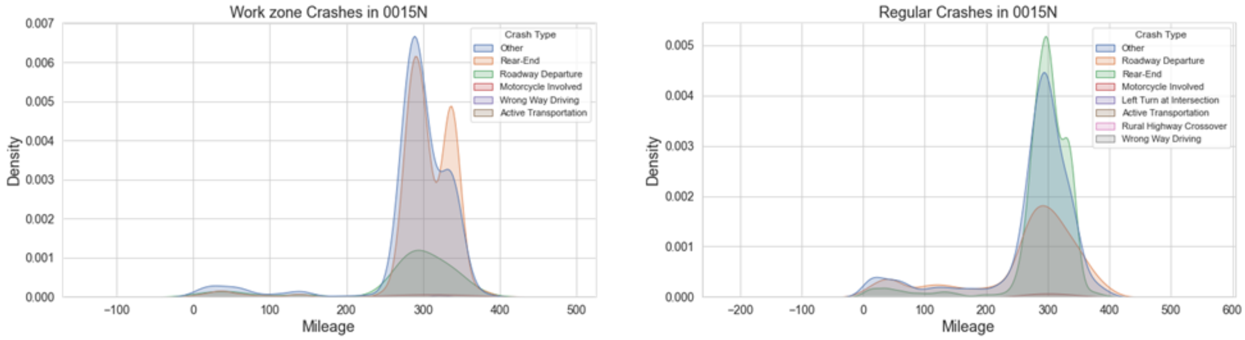


Figure 28. Work Zone and Non-Work Zone Crashes in I-15N

4.3 Rumble Strips Analysis

The location of existing rumble strips around the state was extracted from https://digitaldelivery.udot.utah.gov/datasets/uplan::rumble-strips/about and integrated with the extracted crashes dataset and Masterworks dataset. The following table summarizes the crashes at 3 miles before and after work zones. This 3-mile distance was chosen based on a comprehensive review of the literature, where various research papers proposed different distances for analysis. After evaluating these studies, the research team concluded that a 5-kilometer (approximately 3 miles) range serves as an optimal distance to assess the impact of work zones on crash rates, balancing the need for comprehensive data analysis with the practical considerations of crash data availability and relevance to work zone safety evaluations.

Table 2. Frequency of Work Zone Crashes in the Presence of Rumble Strips

Rumble Strips	**3 Miles Before WZ**	**2 Miles Before WZ**	**1 Mile Before WZ**	**WZ**	**1 Mile After WZ**	**2 Miles After WZ**	**3 Miles After WZ**
Total # Crashes (Crashes & Masterworks)	100	140	212	1710	202	115	95
Road Segments in Rumble Dataset	92	125	169	1614	174	111	90
Total # Roadway Departure Crashes	20	21	25	241	21	21	13
Rumble Presence	14 (70%)	4 (19%)	9 (36%)	111 (46%)	10 (48%)	4 (19%)	5 (38%)
No Rumble	6 (30%)	17 (80%)	16 (64%)	130 (54%)	11 (52%)	17 (81%)	8 (62%)

These figures show that the presence of rumble strips was generally associated with a lower percentage of roadway departure crashes compared to the absence of rumble strips. Interestingly, the table also suggests that rumble strips have less impact in work zone areas compared to areas before and after the work zone. **While most roadway departure crashes in areas before and after a work zone occurred in areas with no rumble strips, there was almost the same number of roadway departure crashes in areas with and without rumble strips within the work zone itself**.

4.4 Traffic Countermeasure Analysis

The traffic countermeasure strategies most commonly used by UDOT are as follows:

1. Pave or Widen Shoulder
2. Left-Turn Lane
3. Shoulder Rumble Strips

4. Roundabout or Signal
5. Horizontal Curve Improvements
6. Left-Turn Phase Change
7. Clear Zone Improvements
8. Right-Turn Lane
9. Active Transportation Improvement
10. Shoulder Barrier
11. Intersection Lighting
12. Raised Median
13. Centerline Rumble Strips
14. Median Barrier

In order to better understand the effect of each countermeasure, the number of crashes that occurred within a 3-mile distance from and within the work zones are summarized in Table 2. The table presents the following information:

- The table presents the cross-referenced data from the Numetric and Masterworks datasets.
- The first line indicates the number of crashes for which information was available in the rumble Masterworks and Numetric Crashes dataset.
- The next 14 lines show the number of crashes that happened in the presence of each safety countermeasure.

The table provides a comprehensive overview of the number of crashes within the 3 miles from and within the work zones for each countermeasure strategy. This analytical approach of examining crashes within specific distances from work zones, especially extending to 3 miles, is instrumental for traffic engineers seeking to comprehend the effectiveness of various traffic control and safety measures at different proximities to work zones. This tiered distance analysis (1, 2, and 3 miles) before and after work zones is critical for several reasons:

1. **Early Warning and Driver Behavior**: It helps understand how early warning signs and other preemptive measures influence driver behavior well before the work zone. Drivers' responses to such measures can vary significantly, and the extended analysis helps identify the optimal placement for these warnings to enhance safety.
2. **Traffic Flow and Congestion Analysis**: By analyzing crash rates at varying distances, engineers can gauge the impact of work zones on traffic flow and congestion, which often

begins to manifest several miles before a work zone. This can inform strategies to mitigate congestion and reduce crash risks.

3. **Evaluating the Impact of Countermeasures Over Distance**: Different countermeasures may have varying degrees of effectiveness based on distance from the work zone. For instance, some measures might be more effective in immediate proximity, while others have a broader impact, reducing the likelihood of crashes due to traffic buildup or changes in traffic patterns several miles away.
4. **Comprehensive Safety Planning**: This approach allows for a more nuanced safety analysis, facilitating the development of tailored strategies that address both immediate and distant risks associated with work zones. It acknowledges that the influence of a work zone on driver behavior and safety extends beyond its physical boundaries.

The analysis demonstrates the impact of these countermeasures in reducing the number of crashes. They are sorted based on their popularity (i.e., how frequently they are implemented). The results reveal that the presence of countermeasures is generally associated with a lower percentage of work zone crashes compared to their absence. However, the effect of countermeasures in reducing the number of crashes is almost the same for areas before, after, and within the work zone. Moreover, the analysis shows that nearly 60% of work zone crashes happened in areas without traffic countermeasures.

Table 3. Frequency of Work Zone Crashes Considering the Traffic Safety Countermeasures

Traffic Countermeasures	3 Miles Before WZ	2 Miles Before WZ	1 Mile Before WZ	WZ	1 Mile After WZ	2 Miles After WZ	3 Miles After WZ
Total # crashes (cross-referencing Numetric crashes & Masterworks)	100	140	212	1710	202	115	95
Paved or widened shoulder	11	9	14	153	15	9	7
Left turn lane	4	4	12	69	6	10	6
Shoulder rumble strips	11	7	14	95	9	6	5

Roundabout or signal	0	3	5	41	3	5	0
Horizontal curve improvements	6	5	8	87	7	4	5
Left-turn phase change	3	4	8	47	3	6	3
Clear zone improvements	6	3	9	91	10	3	10
Right-turn lane	1	5	2	18	9	3	2
Active transportation improvement	2	0	3	11	0	0	2
Shoulder barrier	0	1	3	37	3	0	4
Intersection lighting	2	2	1	17	3	3	2
Raised median	0	0	3	19	3	2	0
Centerline rumble strips	1	1	0	12	1	1	2
Median barrier	0	1	0	8	0	0	1
No countermeasure	53	95	130	1005	130	63	46
Percentage of No Countermeasures	53%	68%	61%	59%	64%	55%	48%

4.5 Contract Type Analysis

This analysis aims to understand how different contract types may influence the occurrence of crashes. The findings of this analysis have been summarized in Table 4. Our analysis reveals that CMGC contracts exhibit a more significant increase in the number of crashes as vehicles approach work zones compared to other contract types, which could be related to both the sample size and poor safety management. Also, based on normalization results (Table 5), Desing-Bid-Build contracts are the safest ones, and CMGCs are the most dangerous ones. Moreover, based on the results, work zones have a total crash (all 5 classes) rate of 0.63 per 100 million VMT. At the same time, they have a fatality rate of 0.004 per 100 million VMT.

Table 4. Effect of Contract Types on the Frequency of Work Zone Crashes

Contract Type	3 Miles Before WZ	2 Miles Before WZ	1 Mile Before WZ	WZ	1 Mile After WZ	2 Miles After WZ	3 Miles After WZ

Total # Crashes (Cross Referencing Numetric Crashes & ProjectWise)	100	140	212	1710	202	115	95
CMGC	2%	2%	7%	4%	6%	0	1%
Design-Build	28%	31%	10%	14%	11%	41%	34%
Design-Bid-Build	70%	67%	83%	82%	83%	59%	65%

Table 5. Crash Rates Based on Contract Types

Contracts	Count	Average Duration (Days)	Average Length (Miles)	Total Crash Per 100M VMT
CMGC	71	469	1.7	5.45
Design - Build	238	830	4.92	1.02
Design, Bid, Build	1401	223	10.1	0.57

Additionally, the following table lists the number of non-work zone crashes in Utah.

Table 6. Non-Work Zone Crashes in the State of Utah

Year	VMT	Fatal	Suspected Serious Injury	Suspected Minor Injury	Possible Injury	No Injury/ PDO	Total
2017	31,510,020,465	236	1,167	5,678	10,404	42,608	60,093
2018	32,258,369,802	226	1,094	5,588	10,314	41,490	58,712
2019	32,933,228,764	205	1,055	5,711	10,660	43,254	60,885
2020	30,189,193,125	245	1,240	5,412	8,256	33,132	48,285
2021	33,755,013,902	289	1,378	6,615	9,532	41,215	59,029
Total	**160,645,826,058**	**1,201**	**5,934**	**29,004**	**49,166**	**201,699**	**287,004**

Based on data in Table 6, Table 7 summarizes the non-work zone crashes per 100 million VMT.

Year	Fatal	Suspected Serious Injury	Suspected Minor Injury	Possible Injury	No Injury/PDO	Total
2017	0.75	3.70	18.02	33.02	135.22	190.71
2018	0.70	3.39	17.32	31.97	128.62	182.01
2019	0.62	3.20	17.34	32.37	131.34	184.87
2020	0.81	4.11	17.93	27.35	109.75	159.94

2021	0.86	4.08	19.60	28.24	122.10	174.87
Average	**0.75**	**3.69**	**18.05**	**30.61**	**125.56**	**178.66**

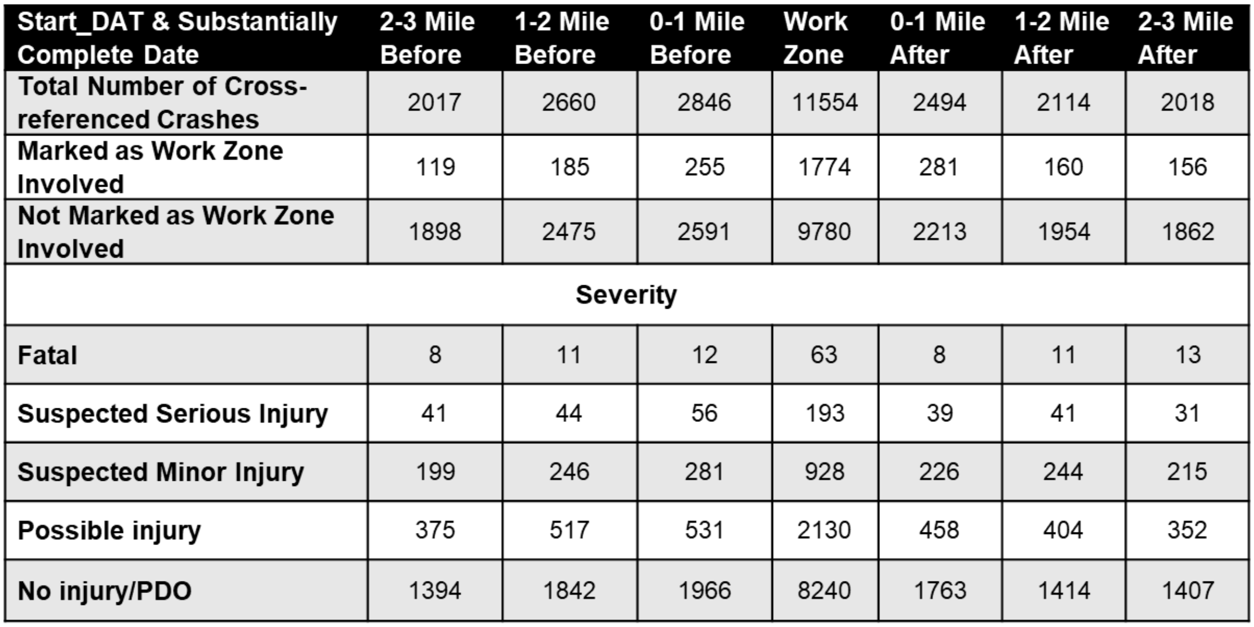
When contrasting the crash rates between work zones and non-work zones in Utah, it's evident that work zones exhibit significantly higher safety levels, evidenced by lower crash rates.

4.6 Potential Work Zone Crashes

Effective road safety management requires a comprehensive understanding and analysis of crash data, particularly those occurring in work zones. In this section, we examine work zone and non-work zone crashes, focusing on the meticulous process of identifying unmarked work zone incidents through cross-referencing location and date data. Additionally, we address discrepancies observed in crash data and propose further investigation methods to enhance data accuracy and alignment. Through this analysis, we aim to shed light on the intricacies of work zone crash data and underscore the importance of robust data management practices in ensuring road safety.

As indicated in Table 8, a significant portion of unmarked work zone crashes were identified by cross-referencing the location and date of the incidents with known work zones. This meticulous process allowed for the identification of crashes that occurred in close proximity to work zones but were not explicitly labeled as 'work zone related.' For these instances, further examination using ClearGuide data is proposed. ClearGuide data analysis could unveil additional insights, particularly regarding incidents that occurred near work zones during periods of reduced speed, which are commonly associated with such construction areas.

Table 7. Potential Work Zone Crashes



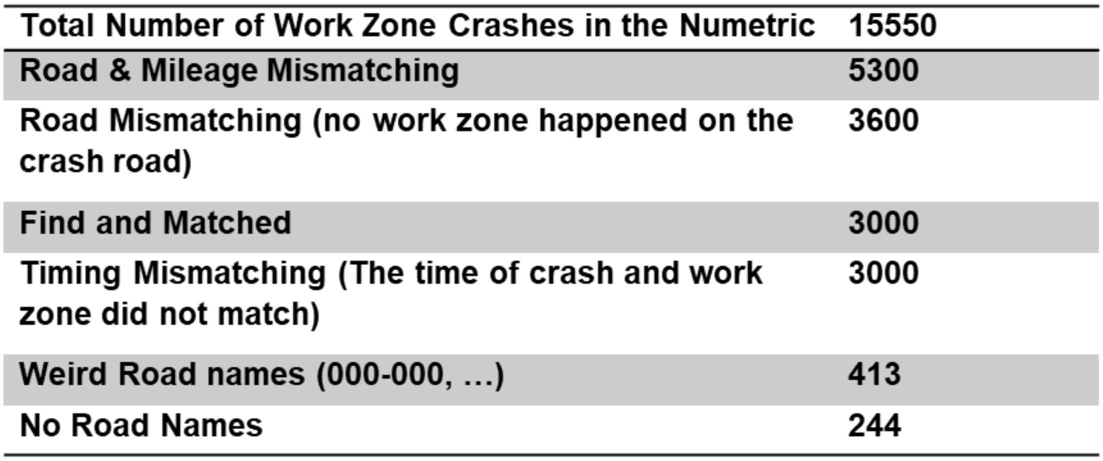
Start_DAT & Substantially Complete Date	2-3 Mile Before	1-2 Mile Before	0-1 Mile Before	Work Zone	0-1 Mile After	1-2 Mile After	2-3 Mile After
Total Number of Cross-referenced Crashes	2017	2660	2846	11554	2494	2114	2018
Marked as Work Zone Involved	119	185	255	1774	281	160	156
Not Marked as Work Zone Involved	1898	2475	2591	9780	2213	1954	1862
Severity							
Fatal	8	11	12	63	8	11	13
Suspected Serious Injury	41	44	56	193	39	41	31
Suspected Minor Injury	199	246	281	928	226	244	215
Possible injury	375	517	531	2130	458	404	352
No injury/PDO	1394	1842	1966	8240	1763	1414	1407

As shown in Table 9, out of the 15,550 work-zone-involved crashes in Numetric:

- Around 5300 did not occur within the work zone activities' reported start and end mileage.
- Approximately 3000 of them occurred on roads where there were no reported work zones in Masterworks.
- Approximately 3000 occurred in the reported location of work zones but not within the reported start and end times of the work zones.
- Finally, 700 were either recorded with peculiar road names (e.g., 5700000, 000-000, ...) or had no road names provided.

The substantial number of unreported work zone crashes highlights the serious issue of underreported incidents that may occur within work zones.

Table 8. Reasons for Differences in Detected Work Zone Crashes



Total Number of Work Zone Crashes in the Numetric	15550
Road & Mileage Mismatching	5300
Road Mismatching (no work zone happened on the crash road)	3600
Find and Matched	3000
Timing Mismatching (The time of crash and work zone did not match)	3000
Weird Road names (000-000, …)	413
No Road Names	244

4.7 Safety Countermeasures

In this section, an extensive review and analysis of work zone safety countermeasures drawn from a comprehensive selection of sources, including DOT reports, National Cooperative Highway Research Program (NCHRP) publications, the Manual on Uniform Traffic Control Devices (MUTCD), and various research papers will be presented. The objective is to assess and compare the effectiveness of these countermeasures in mitigating the risk of crashes within construction work zones. The primary metric used for this comparison is the crash modification factor (CMF), a parameter that quantifies the impact of safety measures on crash reduction. A CMF is a statistical parameter used to evaluate the effectiveness of a safety intervention or countermeasure. It quantifies the change in the expected number of crashes after implementing a specific safety measure when compared to a baseline or control condition. CMFs are typically calculated by analyzing historical crash data for sites with and without safety measures.

$$CMF = \frac{Crash\ Frequency\ After\ Countermeasure}{Crash\ Frequency\ Before\ Countermeasure} \tag{12}$$

For example, if the baseline crash frequency before implementing a new work zone safety measure is 100 accidents per year, and after implementation, the crash frequency decreases to 80 accidents per year, the CMF would be:

$$CMF = \frac{80}{100} = 0.8$$

This CMF value of 0.8 indicates that the safety measure resulted in a 20% reduction in crashes compared to the baseline condition. A CMF less than 1 suggests that the intervention effectively reduces crashes, while a CMF greater than 1 indicates that it may increase crash risk. Hence, a lower CMF indicates a more effective countermeasure.

Based on the literature review, the available work zone traffic control approaches can be divided into 3 main groups, including 1) Speed Control Group, 2) Intrusion Prevention and Warning Systems, and 3) Human-Machine Interaction Detection Systems. However, in order to include all the available measures, two additional groups, 4) Smart Work Zone (Advanced Technology) and 5) Traditional Approaches, were included in the report. Additionally, the analysis considered various data collection techniques prevalent in the reviewed literature. These encompassed methods such as interviews with transportation professionals and field data collection for specific time periods within construction work zones.

4.7.1 Speed Control Group

This category primarily focuses on controlling vehicle speeds within construction work zones. The following countermeasures are included:

- Portable changeable message signs (PCMSs) or Variable message signs (VMS): Widely adopted by DOTs due to their portability and adaptability.
- Dynamic speed displays: Effective in reducing speeds, although costlier to implement.
- Portable rumble strips (PRS): Offers speed reduction benefits and is relatively cost-effective.
- Police enforcement: Traditional and known for reducing speeds but comes with a significant cost.
- Radar speed displays or Drone Radar (iCone): These systems provide both speed reduction and less speed variation, making them a subject of considerable research interest.

- Variable Speed Limit (VSL) systems: Studied extensively, with a 0.9 CMF suggesting their effectiveness in reducing crashes.
- Automated Speed Enforcement and other technologies are also explored in the literature but might be less commonly favored by DOTs due to various factors such as cost and public acceptance.

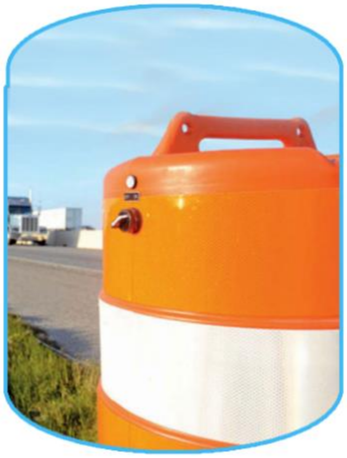


Figure 29. iBarrel from iCone is Used to Provide Real-Time Information on Traffic Patterns in a Work Zone.

4.7.2 Intrusion Prevention and Warning Systems

This category primarily aims to protect workers and prevent unauthorized access to work zones.

- Positive Protection Systems (PPS): Preferred for their significant cost savings in terms of injury and crash costs, such as:
  - Water-Filled Barriers: These barriers are made from plastic and filled with water to provide weight. They are used to absorb impact energy during a collision, reducing the risk of severe injuries. Water-filled barriers are often used where a lighter-weight barrier is preferred or where rapid deployment and removal are needed.
  - Crash Cushions: These are impact attenuators placed at the ends of barriers or hazards to absorb impact energy and reduce the severity of collisions. Crash cushions are designed to be hit and can significantly decrease the damage and injuries resulting from a crash.

- Truck-Mounted Attenuators (TMAs): TMAs are mounted on the back of a truck to protect workers and equipment from errant vehicles. They are designed to absorb impact energy if a vehicle crashes into the truck, reducing the severity of the collision.

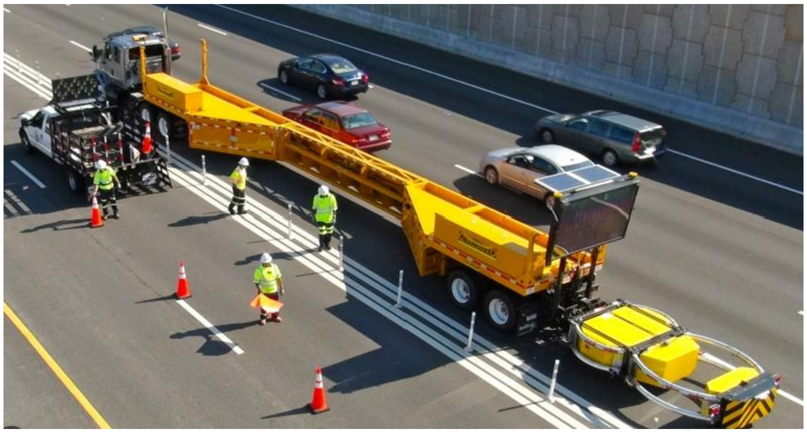


Figure 30. Positive Protection in Work Zones for Protecting Workers

- Intrusion Alert Technologies (IAT) and the use of retroreflective devices are mentioned as additional means to enhance intrusion prevention, such as:
  - Infrared Sensors: Utilize infrared beams to detect motion or intrusion into designated areas. When the beam is broken, an alert is triggered, warning the work crew of the potential danger.
  - Laser Scanners: Employ laser technology to monitor predefined zones for unauthorized intrusions. Upon detection, they can activate warning signals to alert workers.
  - Wearable Alert Devices: These devices can be worn by workers and are activated either manually or automatically in response to an intrusion alert, providing immediate notification through vibrations, sounds, or visual cues.
  - Automated Flagging Assistance Devices (AFADs): While primarily used for traffic control, some AFADs are equipped with intrusion detection capabilities to enhance worker safety by alerting when vehicles mistakenly enter the work zone.

4.7.3 Human-Machine Interaction Detection Systems

- Focuses on improving communication and awareness between workers and drivers.
- Proximity warning systems (PWSs) and visual-based warning systems (VWS) are discussed as potential safety measures, though their adoption might vary.

4.7.4 Smart Work Zone (Advanced Technology)

- Involves the integration of advanced technologies to enhance work zone safety.
- Unmanned Aerial Systems (UAS) and audible warning alarm systems are highlighted as worker safety measures. For example, using UAS, workers and equipment within the work zone could be automatically identified and tracked using object detection algorithms applied to aerial images captured by UAS. Another potential application of UAS is the development of an alarm system to alert workers about an approaching upstream vehicle.
- Queue Warning Systems, ITS countermeasures, and LiDAR technology are explored as ways to reduce crashes and improve traffic flow.

4.7.5 Traditional Approaches

These approaches include standard practices that have been used in work zone traffic control for years.

- Increasing shoulder width, reducing lane widths, and implementing lane closures are common practices, although their effectiveness might be situation-dependent.
- Transition areas are identified as critical and potentially dangerous zones within work zones.

Table 9. Summarizing the Most Common Work Zone Countermeasures and Their Effects

Category	Parameter	Effect	CMF	Implementation	Other
Speed Reduction Systems	Speed-limit signs and work zone signs	-		All States	Drivers glanced at 40% frequency.

	Variable Speed Limit (VSL)	-	0.9	-	-
	Police enforcement	5-10 MPH speed reduction	0.59	All states	-
	Automated Speed Enforcement		0.83		Photo speed enforcement systems
	Radar speed displays or Drone Radar (iCone)	6%-23% speed reduction	-	Florida, Oregon, California, …	Less variation in speeds
	Variable message signs	1-11 MPH speed reduction	-		Most popular in literature
	Portable changeable message signs (PCMSs)		-	Iowa, Oregon, …	Most common between DOTs
	Dynamic speed displays		0.54-0.85	Iowa, Indiana	Cost 9.5K
	Portable rumble strips (PRS)	6-14 MPH Speed reduction	0.4-0.9	Missouri, Georgia, Illinois, Iowa, Kansas, Minnesota, Texas, Washington, Wisconsin, …	Cost 1K
	PRS + Queue Warning System		0.59	Indiana	Cost 250K
	Use of blue LED light trailers in work zones where police detail is not required			Florida	
Intrusion prevention and warning systems (IPWS)	Positive protection systems (PPS), including concrete barriers, ballast-filled barriers, shadow vehicles, vehicle arrestors, guardrails, traffic control barriers, terminal end treatments, impact attenuators, sand barrel arrays, and truck mounted and trailer mounted impact attenuation	-	-	-	Save injury cost savings to DOTs and contractors in the US of up to $1.1 million annually and a crash cost savings of $196,885
	Intrusion alert technologies (IAT), including infrared beams, microwaves, and pneumatic pressured tubes as triggering mechanisms, Sonoblaster, Intellicone, traffic	-	-	Oregon (Research)	-

	worker alert systems, and advanced warning and risk evasion (AWARE)				
	Automated Flagger	-	-	-	-
	Use of retroreflective devices	-	-	-	-
Human-machine-interaction detection systems	Proximity warning systems (PWSs)	-	-	Georgia (Research)	-
	Visual-based warning system (VWS)	-	-	-	-
Smart Work Zone (Advanced Technology)	Using Unmanned Aerial System (UAS) for Active Safety Monitoring	-	-	Georgia (Research)	Worker Safety
	An audible warning alarm system to alert workers	-	-	Research	-
	In-vehicle work zone warning application under the connected vehicle (CV) environment	-	-	Research	-
	Queue Warning System or End-of-Queue Warning System	-	0.3-0.5	Texas (Research)	Reduced Crashes by 44%.
	Intelligent Transportation Systems (ITS) countermeasures, including Variable Speed Limit (VSL), Dynamic Message Sign (DMS)	-	-	Some states	Reduced rear-end collision by 14%
	Alarm device and directional audio system (DAS)	-	-	Missouri (Research)	Reduce Vehicle Merging Speed
	Using LiDAR for Vehicle Detection	-	-	U.S. DOT (Research)	-
Traditional Approaches	Increase Shoulder Width	-	0.9-1	-	Cost 1K
	Reduced lane widths	-	1	-	-
	Shoulder closures	-	-	-	-
	Lane closures	-	-	-	-
	Lane shifts	-	-	-	-

	Retroreflectivity of Pavement Markings	-	-	-	-
	Provision of advance warning areas	-	-	-	-
	Buffer spaces	-	-	-	-
	Transition areas	-	-	-	Most Dangerous Area
	Tapers	-	-	-	-
	Speed Humps	-	-	-	-

While multiple work zone countermeasures are available, the precise effects of certain measures or their combinations remain unstudied. Despite the abundance of reports and literature in this field, portable changeable message signs (PCMSs) emerge as the most frequently employed countermeasure among DOTs, while variable message signs hold this distinction in the literature. Notably, the literature identifies **transition areas as the most hazardous zones within work zones.**

4.8 State-of-the-Practice in DOTs

In this study, a survey was distributed among all DOTs to assess their satisfaction with any of the listed work zone safety countermeasures. This section presents a comprehensive summary of the findings derived from a survey that engaged the active participation of 24 responses collected from 22 states. Each response provided valuable insights into various factors influencing workplace safety and satisfaction. The states that responded to our survey include:

Table 10. List of Engaged States

Kansas	Pennsylvania	California	Illinois
Vermont	Maryland	West Virginia	Delaware
North Carolina	Wisconsin	Georgia	Florida
South Dakota	Minnesota	ARDOT	Missouri
Michigan	Washington, DC	Oklahoma	
Kentucky	Iowa	Colorado	

4.8.1 Factor Analysis

The survey results highlight factors that significantly impact safety and satisfaction, with satisfaction levels ranging from highest to lowest as follows:

Table 11. List of Work Zone Countermeasures and Satisfaction Levels

Factor	Very Satisfied	Satisfied	Neutral	Dissatisfied	Very Dissatisfied	NA	Positive	Negative
Portable Changeable Message Sign (PCMS) or Variable (Dynamic) Message	21%	67%	8%	0%	0%	4%	88%	0%
Lane Closures	8%	79%	4%	0%	4%	4%	87%	4%
Retroreflective devices	29%	58%	8%	0%	0%	4%	87%	0%
Police Enforcement	25%	54%	13%	8%	0%	0%	79%	8%
Shoulder Closures	8%	71%	17%	0%	0%	4%	79%	0%
Positive protection systems (PPS)	25%	38%	13%	0%	0%	25%	63%	0%
Queue Warning System	25%	33%	29%	0%	0%	13%	58%	0%
Portable Rumble Strips (PRS)	13%	42%	17%	4%	13%	13%	55%	17%
Speed Limit and Work zone signs	4%	50%	25%	8%	4%	8%	54%	12%
Automated Flagger	13%	38%	21%	4%	0%	25%	51%	4%
Reduced Lane Width	0%	50%	42%	0%	0%	8%	50%	0%
Radar Speed Display or Drone Radar	0%	50%	21%	0%	0%	29%	50%	0%
Warning Lights (LED light trailers, ...)	4%	42%	25%	0%	0%	29%	46%	0%
Dynamic Speed Display (DSD)	13%	21%	29%	4%	0%	33%	34%	4%
Variable Speed Limit (VSL)	8%	13%	13%	0%	0%	67%	21%	0%

4.8.2 Other Methods

Furthermore, the survey collected responses on additional factors and their corresponding satisfaction levels, including:

Table 12. Non-Listed Work Zone Features and Satisfaction Levels

Factors	Satisfaction
Sequential flashing warning lights on merge tapers	Very Satisfied
Work zone presence lighting	Dissatisfied
Zipper Merge	Satisfied
Full Closures	Very Satisfied
"Obey the flagger" sign placed on the center line across from the "flagger symbol" sign	Satisfied
Sequential flashing warning lights	Satisfied
Automated WZ Speed Enforcement	Very Satisfied

Protection Vehicle	Not mentioned
Maintenance Zone Enhanced Enforcement Program (MAZEEP)	Not mentioned
Solar Advanced Warning Systems (SAWS)	Not mentioned
Speed Photo Enforcement	Satisfied

4.8.3 Challenges

In addition to the satisfaction ratings, the report delves into the challenges associated with implementing these safety measures within work zones. These challenges are thoroughly documented, providing a comprehensive overview of the current landscape and opportunities for improvement in work zone safety and satisfaction.

1. Lack of agency staff and reliance on external resources do not build institutional knowledge within the agency. Staffing issues also make implementation of new/innovative strategies very difficult with current project workloads.
2. Cost, ways to introduce new devices since the traffic control methods are left to the contractor as long as they meet state standards and the MUTCD.
3. Too many devices to set up/takedown each day.
4. Lots of worker exposure.
5. Hard to get contractors to install devices in accordance with standards and specifications.
6. Variable speed limits required legislative approval and were not initially approved but eventually passed.
7. KYTC piloted some temporary rumble strip projects in 2021 and 2022, but feedback from the Districts was not positive. Issues of the strips either sliding or breaking apart were the common complaints. Further research into the products used and where they were installed (i.e., curves or downhill grades) is needed to determine the cause of the issues.
8. Contractor and maintenance force compliance with TTC policies, regulations, and laws when implementing TTC devices.
9. Evaluating the effectiveness of strategies
10. Developing guidelines and specs (measurement and payment).
11. Driver compliance

12. Driver distraction and inattentiveness have been a big issue this season, along with commercial vehicles.
13. I find it hard to install the operation as designed due to contractor installation on a daily basis and constant monitoring of all installations for effectiveness.
14. Blue lights become less effective.
15. Time & Availability. In some instances, getting the needed equipment to use and getting feedback on some new devices takes time. That said, our administration and senior staff are very supportive of cutting-edge technology.
16. Maintenance of devices.
17. Resistance to Change - Technology Integration.
18. Takes time to provide effective results that will influence change allowance as cost/benefit is a difficult balance with all safety and even more challenging when the DOT is not in control of the General Contractor for a project. The changes needed to the overall culture/behavior of the Department, contractors, decision-makers, and the general traveling public is a dynamic target with the many different parts of the state that Delaware has and the roadway network that the DOT is responsible for (subdivision streets through limited-access tolled interstate roads).
19. Availability of law enforcement officers (LEOS), industry resistance to some new methods
20. Driver behavior post-COVID continues to be a challenge with elevated speed.

4.9 Speed Effect

This section analyzes the effect of work zones on drivers' speed. The dataset used for this analysis comprises information from over 200 work zones in Utah using Clearguide, Iteris probe data. We first examined the distribution of work zones across different years to gain insights into the data.

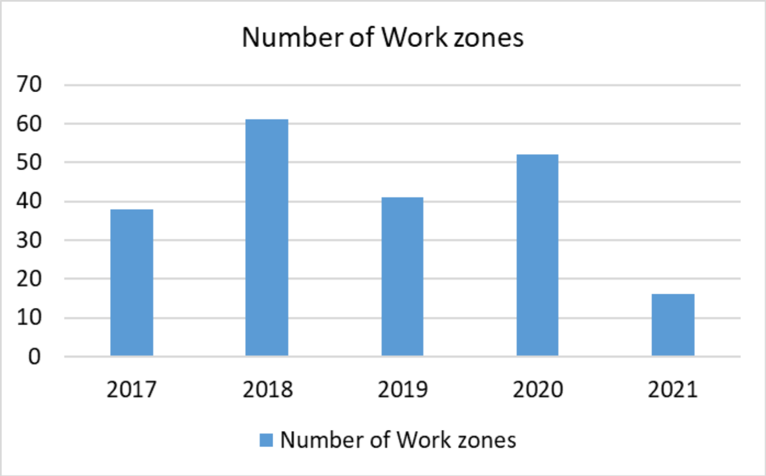


Figure 31. Work Zone Distribution Across Different Years

Using the Clearguide API, we extracted speed information during work zones and compared it with data from one month before implementing work zones. This comprehensive analysis encompassed various speed metrics, including minimum, maximum, average, median, and average travel times. After thoroughly examining these speed metrics within work zones and comparing them to the pre-work zone data, our analysis revealed no significant evidence of an association between work zones and speed reduction. Figure 32 shows the distribution of speed changes in work zones.

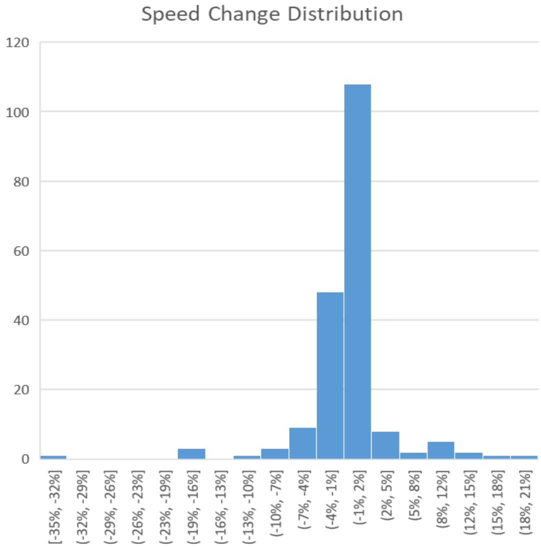
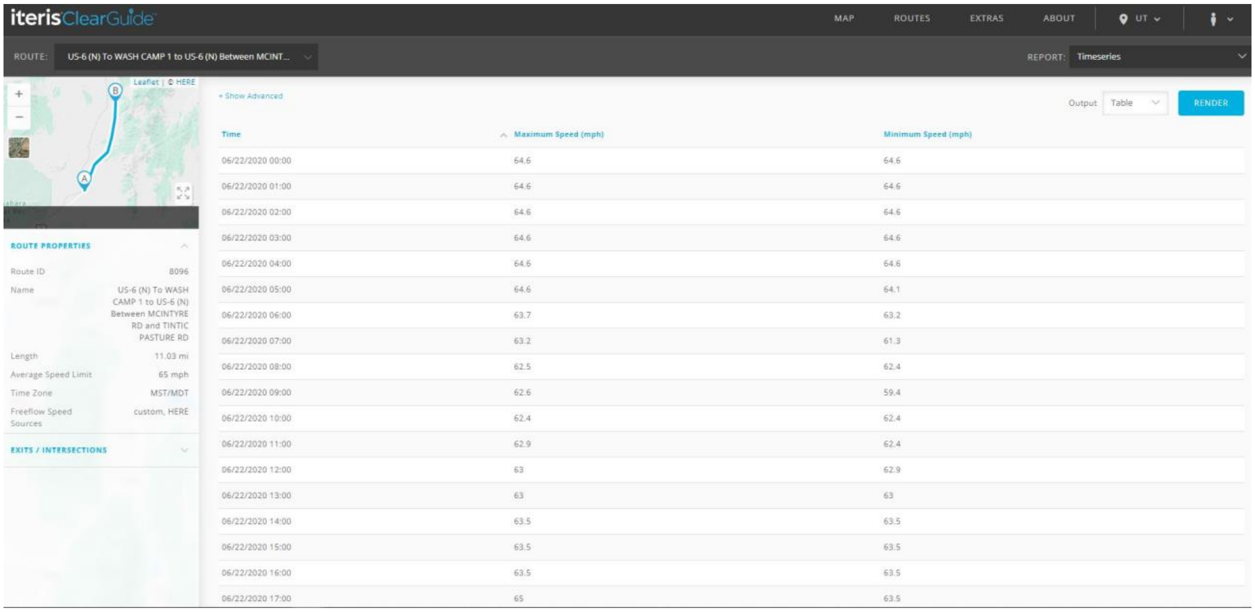


Figure 32. Distribution of Speed Changes in Work Zones

On average, the speed reduction observed was minimal, approximately around 1%. This finding suggests that while slight variations in speed within work zones may exist, it does not translate into a substantial or statistically significant reduction in vehicle speeds. Upon a detailed examination of the data utilized for this analysis, the researchers identified that the scarcity of probe data gathered at work zone sites might account for the minimal differences observed in speeds within work zone areas. Figure 33 displays a screenshot of the Clearguide data for a specific date at a work zone location. The scarcity of probe data, characterized by a limited number or absence of probe data points, has resulted in instances where the minimum and maximum speeds recorded are identical. This uniformity in speed values can be attributed to the insufficient data available for analysis, underscoring the challenge of accurately assessing speed variations within work zones due to the lack of comprehensive data collection.



ROUTE: US-6 (N) To WASH CAMP 1 to US-6 (N) Between MCINT...

REPORT: Timeseries

Output: Table

ROUTE PROPERTIES

Route ID	8096
Name	US-6 (N) To WASH CAMP 1 to US-6 (N) Between MCINTYRE RD and TINTIC PASTURE RD
Length	11.03 mi
Average Speed Limit	65 mph
Time Zone	MST/MDT
Freeflow Speed Sources	custom, HERE

Time	Maximum Speed (mph)	Minimum Speed (mph)
06/22/2020 00:00	64.6	64.6
06/22/2020 01:00	64.6	64.6
06/22/2020 02:00	64.6	64.6
06/22/2020 03:00	64.6	64.6
06/22/2020 04:00	64.6	64.6
06/22/2020 05:00	64.6	64.1
06/22/2020 06:00	63.7	63.2
06/22/2020 07:00	63.2	61.3
06/22/2020 08:00	62.5	62.4
06/22/2020 09:00	62.6	59.4
06/22/2020 10:00	62.4	62.4
06/22/2020 11:00	62.9	62.4
06/22/2020 12:00	63	62.9
06/22/2020 13:00	63	63
06/22/2020 14:00	63.5	63.5
06/22/2020 15:00	63.5	63.5
06/22/2020 16:00	63.5	63.5
06/22/2020 17:00	65	63.5

Figure 33. Clearguide Screenshot Showing the Minimum and Maximum Speeds at a Work Zone Location

4.10 Feature Importance Analysis

Feature importance analysis identifies and ranks the most critical features or variables that contribute to the performance of a predictive model. It helps determine which features have the most significant impact on the model's output and can be used to improve the model's performance by discarding irrelevant or redundant features. By highlighting the relative importance of each feature, it allows data scientists and analysts to focus on the most impactful variables, optimizing the model by potentially discarding irrelevant or minimally influential ones. This process enhances the model's efficiency and accuracy and provides insights into the relationships and dependencies between the features and the target variable. In essence, feature importance ranks the attributes in terms of their significance in predicting the outcome without necessarily specifying their exact values or impact directions. The results of the feature importance analysis depicted that the following features were the most influential factors in crash severity in work zones, listed in order of decreasing importance:

- **Roadway Surface Condition** (Dry, Wet, Snow, …)
- **Crash Type** (Roadway Departure, Rear-end, Mid-block, …)
- **Motorcycle Involved** (Yes/No)
- **Weather Condition** (Clear, Cloudy, Rainy, …)
- **Roadway Junction Type** (Crossover, Intersection, Ramp, …)
- **Type of Project** (Transportation, Rehabilitation, …)
- **Drowsy Driving Involved** (Yes/No)
- **Domestic Animal Involved** (Yes/No)
- **Manner of Collision** (Head On, Front to Rear, Rear to Side, …)
- **Holiday Crash** (Yes/No)
- **Disregard Traffic Control Device Involved** (Yes/No)

4.11 Severity Prediction Models

In order to predict the severity of work zone crashes accurately, we developed two groups of classifiers. The first group comprised traditional machine learning algorithms such as Decision trees, Random forests, and XGBoost. These algorithms were selected for their robustness and ability to handle complex datasets. The second group consisted of probabilistic machine learning models such as Gaussian Naive Bayes (GNB) and Complement Naive Bayes (CNB). By leveraging the strengths of both traditional and deep learning approaches, we aimed to achieve comprehensive and accurate predictions of work zone crash severity.

4.11.1 Deterministic Machine Learning Models

Three popular machine learning algorithms, namely Decision tree, Random forest, and XGBoost, were utilized to train and assess the performance of the work zone crashes dataset. The objective was to assess the effectiveness of these algorithms in predicting and analyzing the severity of work zone crashes, considering five different classes of crash severity. After rigorous training and testing procedures, the results obtained from the experiments have been meticulously summarized in Table 14. This table presents key performance metrics for each algorithm, such as accuracy, precision, recall, and F1-score, providing valuable insights into their predictive capabilities for different severity levels of work zone crashes.

Table 13. Results of Deterministic Machine Learning Models

Model	Classes	Precision	Recall	F1-Score	Accuracy
DT	Fatal	0.57	0.67	0.62	83%
	No Injury/PDO	0.91	0.88	0.89	
	Possible Injury	0.65	0.69	0.67	
	Suspected Minor Injury	0.65	0.69	0.67	
	Suspected Serious Injury	0.69	0.75	0.72	
Total		69.4%	73.5%	73.5%	
RF	Fatal	1	0.67	0.80	89%
	No Injury/PDO	0.89	0.97	0.93	
	Possible Injury	0.84	0.64	0.72	
	Suspected Minor Injury	0.92	0.78	0.84	
	Suspected Serious Injury	1	0.75	0.86	
Total		92.9%	76%	76%	
XGBoost	Fatal	1	0.83	0.91	87%
	No Injury/PDO	0.88	0.96	0.91	
	Possible Injury	0.76	0.62	0.68	
	Suspected Minor Injury	0.93	0.69	0.79	
	Suspected Serious Injury	1	0.75	0.86	
Total		91.25%	76.9%	76.9%	

4.11.2 Probabilistic Machine Learning Models

Two types of Naïve Bayes classifiers have been used in this study, including Gaussian Naive Bayes (GNB) and Complement Naive Bayes (CNB). GNB can be a good choice when dealing with a few classes, as it assumes that each feature is normally distributed within each class. This can make GNB less sensitive to outliers and noise in the data. Additionally, GNB can be computationally efficient and require less training data compared to more complex algorithms (Dimitrijevic et al., 2022). On the other hand, CNB is designed to handle class imbalance, as it estimates the probability that a feature is absent in the other classes. Therefore, this study has chosen CNB and GNB as the two methods to evaluate their performance on the crash dataset.

Moreover, to enhance the performance and simplify the classification process, a revision has been made to the class labels in the system. The original class label "Suspected Minor Injury" has been replaced with the label "Possible Injury," resulting in a reduced number of classes from 5 to 4. This revision brings several advantages to the system. By consolidating the "Suspected Minor Injury" class into the broader category of "Possible Injury," the classification task becomes more streamlined and easier to interpret. The distinction between minor and more severe injuries can be challenging and subjective, often leading to ambiguity in classification. The revised class label helps to alleviate this issue by providing a more inclusive category that covers a wider range of potential injuries.

Table 14. Results of Probabilistic Machine Learning Models

Category	Model	Classes	Precision	Recall	F1-Score	Accuracy
Probabilistic ML	GNB	Fatal	0.80	0.67	0.73	76%
		No Injury/PDO	0.94	0.71	0.81	
		Possible Injury	0.55	0.90	0.68	
		Suspected Serious Injury	0.43	0.50	0.46	
	Weighted Average		82%	76%	77%	
	CNB	Fatal	0.21	0.67	0.32	74%
		No Injury/PDO	0.82	0.83	0.83	
		Possible Injury	0.62	0.52	0.56	
		Suspected Serious Injury	0.27	0.50	0.35	

	Weighted Average		75%	74%	74%	
Non-probabilistic ML	XGBoost	Fatal	1	0.67	0.80	86%
		No Injury/PDO	0.89	0.93	0.91	
		Possible Injury	0.78	0.71	0.74	
		Suspected Serious Injury	0.75	0.50	0.60	
	Weighted Average		86%	86%	86%	

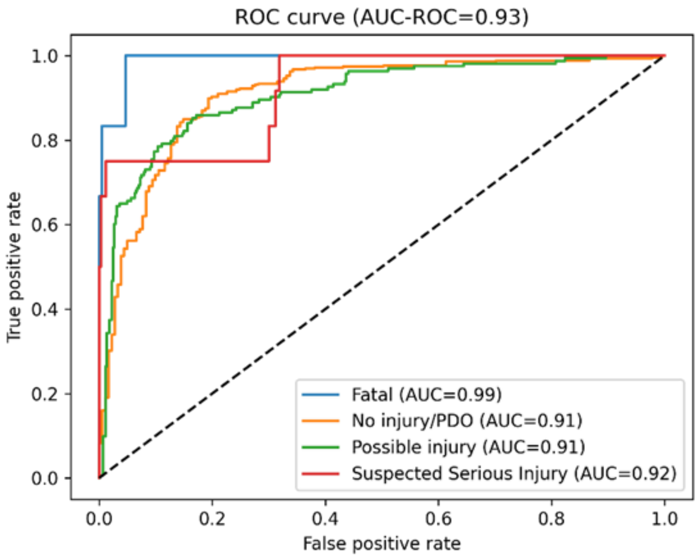


Figure 34. ROC Curve for Random Forest

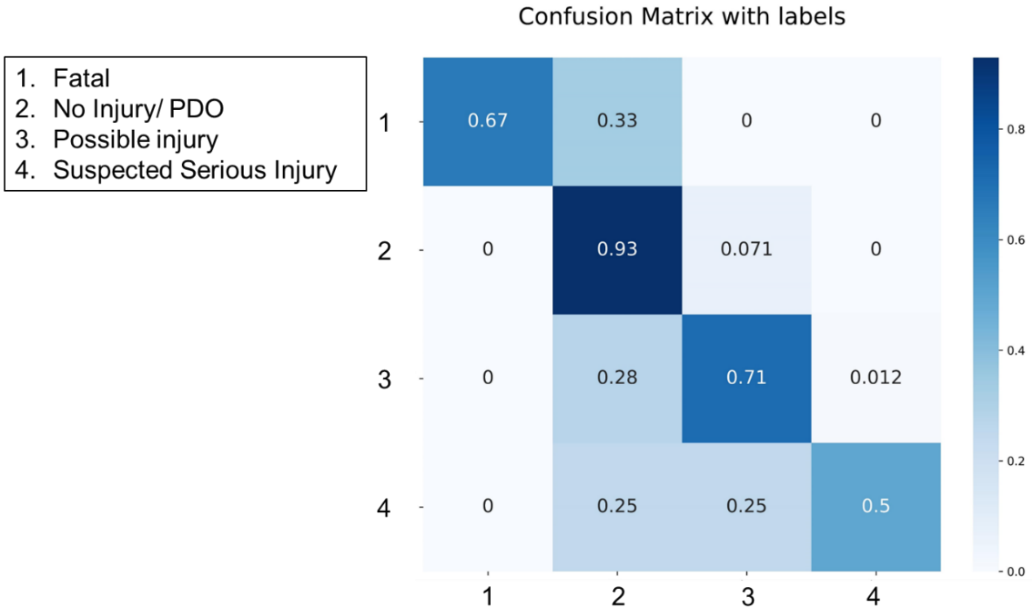


Figure 35. Confusion Matrix for Random Forest

4.12 Summary

The methodology employed in this study encompasses a multifaceted approach to comprehensively analyze work zone safety. Initially, the study gathered relevant data from various sources, including crash reports, speed analyses, and documentation from state DOTs. The study utilized machine learning models to predict crash severity, leveraging features such as location, time of day, weather conditions, and work zone characteristics. The models were trained on historical crash data and evaluated for their predictive accuracy.

Furthermore, the effectiveness of longitudinal rumble strips was assessed through a detailed analysis of roadway departure crashes. This analysis involved comparing crash rates within and outside work zones, shedding light on the overall impact of rumble strips on safety. In addition, the study investigated the influence of different contract types on crash occurrence by analyzing crash data in conjunction with contract specifications. This analysis revealed insights into the relationship between contract mechanisms and work zone safety. Moreover, the study conducted an extensive literature review to identify and evaluate various work zone safety countermeasures. Sources included DOT reports, NCHRP publications, MUTCD guidelines, and academic research. The identified countermeasures were categorized into five groups based on their approach to traffic control.

Additionally, the study surveyed all DOTs to gather insights into factors influencing safety and satisfaction within work zones. The survey responses provided valuable qualitative data, complementing the quantitative analyses conducted in other parts of the study. Overall, this methodology integrates quantitative analysis, machine learning techniques, literature review, and survey research to assess work zone safety and identify effective countermeasures comprehensively.

5.0 CONCLUSIONS

5.1 Summary

In conclusion, this study offers valuable insights into work zone safety through a comprehensive analysis of various factors and the effectiveness of safety countermeasures. The utilization of machine learning models has demonstrated promising results, with 89% accuracy using random forest in predicting crash severity, providing a basis for further research and implementation in work zone management. The analysis of longitudinal rumble strips has revealed their overall impact on reducing roadway departure crashes, albeit with varying effectiveness within work zones. This highlights the need for further investigation and potential modifications to optimize their implementation for enhanced safety. Additionally, the data analysis section reveals that front-to-rear collisions are more common in work zones, attributed to sudden changes in speed.

Moreover, the study has identified the influence of contract types on crash occurrence, emphasizing the importance of considering contract specifications in relation to safety measures within work zones. The analysis revealed that Design-Bid-Build contracts exhibit the lowest crash rates, with 0.57 crashes per 100 million Vehicle Miles Traveled (VMT), while Construction Manager/General Contractor (CMGC) contracts have the highest, with 5.45 crashes per 100 million VMT. This finding underscores the need for collaboration between transportation agencies and contractors to ensure the implementation of appropriate safety measures. Moreover, given the national fatality rate of 1.24 per 100 million Vehicle Miles Traveled (VMT), it is evident that UDOT is performing commendably in managing safety within work zones.

The comprehensive review of safety countermeasures has provided a robust foundation for identifying effective traffic control and intrusion prevention strategies. This study offers practical insights for transportation agencies to enhance work zone safety by categorizing these countermeasures and examining their state of the practice. One of the key insights from the literature review is that transition areas are identified as the most hazardous zones within work zones. Additionally, the survey conducted among DOTs has enriched our understanding of factors influencing safety and satisfaction within work zones, contributing qualitative insights to complement the quantitative analyses conducted in the study. The findings indicate that Portable

Changeable Message Signs, Lane Closures, Retroreflective Devices, and Police Enforcement rank as the most effective methods for traffic control in and around work zones, according to the DOTs surveyed.

Overall, this study underscores the importance of implementing evidence-based safety measures and continuing research efforts to address the complex challenges associated with work zone safety. By adopting a multi-faceted approach and leveraging emerging technologies, we can work towards creating safer work zones, reducing the occurrence and severity of crashes, and ultimately improving overall road safety for all users. In conclusion, our study employed various approaches to analyze work zone safety and explore factors influencing crash occurrence. We utilized machine learning models, such as decision trees, random forests, and extreme gradient boosting, achieving promising accuracy levels. Additionally, we conducted a comprehensive analysis of different aspects related to work zone safety.

5.2 Safety Suggestions

Table 16 summarizes the safety suggestions based on the results of the analysis.

Table 15. Safety Suggestions Based on Analysis Results

Problem	Strategy	Effect
Work Zone Crash Documentation in Police Officer's Report	Adding Work Zone Section to Police Reports	Recording more detailed information about work zones and crashes
Contractor Safety Compliance	Implementing Safety Training and Education, Suggesting Benefits for Implementing Safety Countermeasures, inspection, and penalty	Reduced Frequency and Severity of Crashes, Enhanced Workplace Safety
High Incidence of Rear-End Collisions	Variable Message Signs (VMS) with real-time updates to prepare drivers for changes in traffic patterns and slow-downs ahead.	Expected to reduce sudden braking and rear-end collisions by providing timely information
High Number of Crashes at Locations with No Countermeasures	Having temporary traffic countermeasures	-
Speeding	PCMS	Lowering Speed
	Retroreflective Devices	
	Police Presence	
Manual Traffic Control	Integrating smart traffic control systems with real-time monitoring to adapt to changing conditions.	Reduces human error and the need for manual traffic control while improving the response time to dynamic traffic conditions
Inadequate Hazard Identification for Motorists	Utilization of advanced hazard detection systems coupled with automated warning messages to approaching drivers, such as in-vehicle alerts linked to GPS and traffic apps.	Improve motorists' situational awareness and reduce the likelihood of accidents caused by sudden or unexpected work zone conditions

5.3 Limitations

Despite the comprehensive analysis conducted in this study, certain limitations must be acknowledged. One significant constraint is the lack of accurate and comprehensive data regarding the presence and deployment of work zone countermeasures. This limitation hindered our ability to conduct a thorough investigation and understanding of the effectiveness of these countermeasures. Without precise information on the implementation and usage of various safety measures within work zones, it is challenging to assess their impact accurately. Additionally, the availability of historical crash data, while extensive, may still contain inherent biases or inconsistencies that could influence the study's findings. Thus, future research endeavors should prioritize the collection of precise and detailed data on the deployment and efficacy of work zone safety countermeasures to facilitate more robust analyses and informed decision-making in enhancing work zone safety.

In addition to the aforementioned limitations, it's crucial to acknowledge the dynamic and ever-changing nature of work zones. These environments evolve continuously, with conditions shifting hourly based on ongoing activities within the work zone. Consequently, collecting and maintaining accurate information regarding work zone characteristics, such as the presence and layout of safety countermeasures, can be challenging. The fluidity of work zone conditions introduces complexities in data collection and analysis, as the effectiveness of safety measures may vary throughout the day or in response to specific activities. This dynamic nature underscores the importance of real-time data collection and monitoring to capture the transient nature of work zone safety conditions accurately. Despite efforts to gather comprehensive data, the inherent variability and unpredictability of work zone environments present ongoing challenges in accurately assessing the efficacy of safety countermeasures. Future research endeavors should explore innovative methodologies and technologies to capture and analyze real-time data, enabling a more nuanced understanding of work zone safety dynamics and facilitating proactive safety interventions.

5.4 Future Studies

Here are some future studies that could help better understand work zones:

1. **Real-Time Monitoring and Analysis:** Investigate the feasibility and effectiveness of real-time monitoring systems to continuously assess work zone safety conditions and identify potential hazards. Utilize technologies such as IoT sensors, video analytics, and machine learning algorithms to analyze data and provide timely insights for proactive safety measures.
2. **Impact of Work Zone Layout and Design:** Explore how different layouts and designs of work zones influence driver behavior and crash occurrence. Conduct controlled experiments or simulation studies to assess the effects of factors such as lane configuration, signage placement, and traffic control devices on safety outcomes.
3. **Behavioral Studies:** Investigate driver behavior in work zones and its impact on safety. Use methodologies such as naturalistic driving studies or driving simulators to analyze driver responses to various work zone conditions and interventions. Explore factors such as driver distraction, compliance with traffic control measures, and perception-reaction times.
4. **Evaluation of Emerging Technologies:** Assess the effectiveness of emerging technologies, such as autonomous vehicles, connected vehicle systems, computer vision and machine learning (Farhadmanesh et al., 2021a, 2021b; Hassandokht Mashhadi et al., n.d., 2024; Mashhadi et al., 2024), and advanced driver assistance systems, in improving work zone safety. Conduct field trials or simulation studies to evaluate the potential benefits and challenges associated with integrating these technologies into work zone environments.
5. **Human Factors and Work Zone Safety:** Examine the role of human factors, including driver characteristics, fatigue, workload, and situational awareness, in work zone safety. Investigate strategies to enhance human performance and mitigate error likelihood in work zone driving scenarios.

<u>**6.0 REFERENCES**</u>

Akepati, S. R., & Dissanayake, S. (2011). *Characteristics and contributory factors of work zone crashes*.

Al-Bdairi, N. S. S. (2020). Does time of day matter at highway work zone crashes? *Journal of Safety Research*, *73*, 47–56.

Alkheder, S., Taamneh, M., & Taamneh, S. (2017). Severity prediction of traffic accident using an artificial neural network. *Journal of Forecasting*, *36*(1), 100–108.

Anderson, J., & Hernandez, S. (2017). Roadway classifications and the accident injury severities of heavy-vehicle drivers. *Analytic Methods in Accident Research*, *15*, 17–28.

Chen, E., & Tarko, A. P. (2014). Modeling safety of highway work zones with random parameters and random effects models. *Analytic Methods in Accident Research*, *1*, 86–95.

Coburn, J. S., Bill, A. R., Chitturi, M. V, & Noyce, D. A. (2013). Injury outcomes and costs for work zone crashes. *Transportation Research Record*, *2337*(1), 35–41.

Effati, M., Rajabi, M. A., Hakimpour, F., & Shabani, S. (2015). Prediction of crash severity on two-lane, two-way roads based on fuzzy classification and regression tree using geospatial analysis. *Journal of Computing in Civil Engineering*, *29*(6), 04014099.

Erfani, A., & Tavakolan, M. (2020). Risk Evaluation Model of Wind Energy Investment Projects Using Modified Fuzzy Group Decision-making and Monte Carlo Simulation. *Arthaniti: Journal of Economic Theory and Practice*, 0976747920963222.

Erfani, A., Tavakolan, M., Mashhadi, A. H., & Mohammadi, P. (2021). Heterogeneous or homogeneous? A modified decision-making approach in renewable energy investment projects. *AIMS Energy*, *9*(3), 558–580.

Erfani, A., Zhang, K., & Cui, Q. (2021). TAB Bid Irregularity: Data-Driven Model and Its Application. *Journal of Management in Engineering*, *37*(5), 04021055.

Farhadmanesh, M., Cross, C., Mashhadi, A. H., Rashidi, A., & Wempen, J. (2021a). Highway Asset and Pavement Condition Management using Mobile Photogrammetry. *Transportation Research Record*, 03611981211001855.

Farhadmanesh, M., Cross, C., Mashhadi, A. H., Rashidi, A., & Wempen, J. (2021b). Use of Mobile Photogrammetry Method for Highway Asset Management. *Transportation Research Board 100th Annual MeetingTransportation Research Board*, *TRBAM-21-01864*.

Hassandokht Mashhadi, A., Mohammadi, P., Rashidi, A., Medina, J. C., & Markovic, N. (n.d.). Probabilistic versus Non-Probabilistic Machine Learning Approaches for Estimating the Severity of Crashes in Construction Work Zones. *Construction Research Congress 2024*, 445–454.

Hassandokht Mashhadi, A., Rashidi, A., & Marković, N. (2024). A GAN-Augmented CNN Approach for Automated Roadside Safety Assessment of Rural Roadways. *Journal of Computing in Civil Engineering*, *38*(2), 04023043.

Iranitalab, A., & Khattak, A. (2017). Comparison of four statistical and machine learning methods for crash severity prediction. *Accident Analysis & Prevention*, *108*, 27–36.

Islam, M. (2022). An analysis of motorcyclists' injury severities in work-zone crashes with unobserved heterogeneity. *IATSS Research*.

Jeong, H., Jang, Y., Bowman, P. J., & Masoud, N. (2018). Classification of motor vehicle crash injury severity: A hybrid approach for imbalanced data. *Accident Analysis & Prevention*, *120*, 250–261.

Liu, J., Khattak, A., & Zhang, M. (2016). What role do precrash driver actions play in work zone crashes?: Application of hierarchical models to crash data. *Transportation Research Record*, *2555*(1), 1–11.

Mashhadi, A. H., Farhadmanesh, M., Rashidi, A., & Marković, N. (2021a). Review of methods for estimating construction work zone capacity. *Transportation Research Record*, *2675*(9), 382–397.

Mashhadi, A. H., Farhadmanesh, M., Rashidi, A., & Marković, N. (2021b). State-of-the-Art Methods in Estimating Freeway Work zones Capacity: A Literature Review. *Transportation Research Board 100th Annual MeetingTransportation Research Board*, *TRBAM-21-01863*.

Mashhadi, A. H., & Rashidi, A. (2021). *Evaluating Mobility Impacts Of Construction Workzones On Utah Transportation System Using Machine Learning Techniques*. National Institute for Transportation and Communities (NITC).

Mashhadi, A. H., Rashidi, A., & Markovic, N. (2023). *Automated Safety Assessment of Rural Roadways Using Computer Vision*. Utah. Dept. of Transportation. Research Division.

Mashhadi, A. H., Rashidi, A., Medina, J., & Marković, N. (n.d.). Comparing Performance of Different Machine Learning Methods for Predicting Severity of Construction Work Zone Crashes. In *Computing in Civil Engineering 2023* (pp. 434–442).

Mashhadi, A. H., Rashidi, A., Medina, J., & Marković, N. (2024). Comparing Performance of Different Machine Learning Methods for Predicting Severity of Construction Work Zone Crashes. In *Computing in Civil Engineering 2023* (pp. 434–442).

Mohammadi, P., Rashidi, A., Malekzadeh, M., & Tiwari, S. (2023). Evaluating various machine learning algorithms for automated inspection of culverts. *Engineering Analysis with Boundary Elements*, *148*, 366–375.

Mohammadi, P., Sherafat, B., & Rashidi, A. (2023). *Developing a Culvert Inspection Manual and Estimating Culverts' Deterioration Curve, Inspection Frequency and Service Life for UDOT*. Utah. Department of Transportation.

Mokhtarimousavi, S., Anderson, J. C., Azizinamini, A., & Hadi, M. (2019). Improved support vector machine models for work zone crash injury severity prediction and analysis. *Transportation Research Record*, *2673*(11), 680–692.

Mokhtarimousavi, S., Anderson, J. C., Hadi, M., & Azizinamini, A. (2021). A temporal investigation of crash severity factors in worker-involved work zone crashes: Random parameters and machine learning approaches. *Transportation Research Interdisciplinary Perspectives*, *10*, 100378.

Mokhtarimousavi, S., Azizinamini, A., & Hadi, M. (2020). Severity of worker-involved work zone crashes: A study of contributing factors. *International Conference on Transportation and Development 2020*, 47–59.

Osman, M., Paleti, R., & Mishra, S. (2018). Analysis of passenger-car crash injury severity in different work zone configurations. *Accident Analysis & Prevention*, *111*, 161–172.

Osman, M., Paleti, R., Mishra, S., & Golias, M. M. (2016). Analysis of injury severity of large truck crashes in work zones. *Accident Analysis & Prevention*, *97*, 261–273.

Park, J., Yang, X., Cho, Y. K., & Seo, J. (2017). Improving dynamic proximity sensing and processing for smart work-zone safety. *Automation in Construction*, *84*, 111–120.

Ravani, B., & Wang, C. (2018). Speeding in highway work zone: an evaluation of methods of speed control. *Accident Analysis & Prevention*, *113*, 202–212.

Santos, B., Trindade, V., Polónia, C., & Picado-Santos, L. (2021). Detecting risk factors of road work zone crashes from the information provided in police crash reports: the case study of Portugal. *Safety*, *7*(1), 12.

Work Zone Crashes, Injuries, & Fatalities - Facts & Data | Work Zone Barriers Guide. (n.d.). Retrieved March 8, 2023, from https://www.workzonebarriers.com/work-zone-crash-facts.html

Zhang, K., & Hassan, M. (2019a). Identifying the factors contributing to injury severity in work zone rear-end crashes. *Journal of Advanced Transportation*, *2019*.

Zhang, K., & Hassan, M. (2019b). Injury severity analysis of nighttime work zone crashes. *2019 5th International Conference on Transportation Information and Safety (ICTIS)*, 1301–1308.

Zhang, Z., Akinci, B., & Qian, S. (2022). Inferring the causal effect of work zones on crashes: methodology and a case study. *Analytic Methods in Accident Research*, *33*, 100203.

7.0 Appendix I

In this section, more details about statistical and Machine Learning modeling will be elaborated.

7.1 Statistical Modeling

These models aim to understand the relationship between various factors and the likelihood or severity of crashes. Here are some commonly used statistical modeling approaches for crash severity and frequency:

7.1.1 Generalized Linear Models (GLMs)

GLM is a statistical modeling approach widely used in transportation research to analyze crash severity and frequency. Despite what the name suggests, GLMs can model a wide range of relationships including linear, logistic, Poisson and exponential conditions. The general form of a GLM is expressed by the equation:

$$g(E(Y)) \; = \; \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \qquad (1)$$

where $g()$ is a link function that relates the linear predictor to the expected value of the response variable Y $(E(Y))$. The response variable Y represents crash severity or frequency, and the predictor variables X_1, X_2, ..., X_n correspond to various factors influencing the crash outcome. The β_0, β_1, β_2, ..., β_n are the estimated regression coefficients, which quantify the relationship between the predictors and the response variable.

In the case of crash severity analysis, a GLM can be formulated using a link function which essentially maps a nonlinear relationship to a linear one so that a linear model can be fit. A link function that is appropriate for the outcome variable might include a logit link for binary severity outcomes or a log link for ordinal severity categories. A logit link, also called a logistic regression, takes a linear combination of the covariate values (which could be anything between negative and positive infinity) and converts those to a scale of probability between 0 and 1. A log link, on the other hand, is commonly used when the outcome variable follows a distribution with positive support and exhibits right-skewness. It transforms the linear combination of covariate values to a scale that is directly related to the natural logarithm of the mean of the response variable. This is particularly useful for modeling count data or strictly positive continuous data, where the log link ensures that the predicted values are non-negative.

A Poisson or Negative Binomial distribution is commonly assumed for crash frequency analysis. In crash frequency analysis, the choice of using either a Poisson or negative binomial distribution stems from the nature of the data being analyzed. Crash frequency data often involves counting the number of crashes that occur within a specific time period or at particular locations. This type of data inherently follows a discrete distribution, making the Poisson and negative binomial distributions appropriate choices. The Poisson distribution is commonly utilized due to its ability to model the probability of a certain number of events occurring within a fixed interval, assuming a constant rate of occurrence. However, real-world crash data often exhibits overdispersion, where the variance exceeds the mean, violating the equidispersion assumption of the Poisson distribution. In such cases, the negative binomial distribution provides a better fit by allowing the variance to be larger than the mean, thus accommodating overdispersion.

GLMs offer a flexible and powerful framework for analyzing crash data, enabling researchers to understand the relationships between predictor variables and crash severity or frequency. These models facilitate evidence-based decision-making by identifying significant risk factors and informing the development of targeted safety interventions and policies.

7.1.2 Ordered Probit/Logit Models

Ordered probit, a statistical modeling technique used to analyze ordered categorical outcomes, where the categories have a natural ordering or hierarchy, and ordered logit models, Similar to the ordered probit model, an ordered logit model is a statistical technique used to analyze ordered categorical outcomes are commonly used statistical modeling techniques for analyzing ordered categorical outcomes, such as crash severity levels or injury severity categories, where the variables have natural ordering (e.g., minor, moderate, severe). In an ordered probit model, we use the cumulative distribution function of a standard normal distribution to model the probability of an outcome belonging to a specific category:

$$P(Y \leq j) \ = \ \emptyset(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \ - \ \gamma_j) \qquad (2)$$

where Y represents the outcome variable, X_1, X_2, ..., X_n are the predictor variables, β_0, β_1, β_2, ..., β_n are the estimated coefficients, and γ_j represents the threshold parameter for category j. The cumulative distribution function $\emptyset$() gives the probability that a normally distributed variable takes a value less than or equal to a given threshold. For example, let's say we're using an ordered probit model to analyze crash severity levels (Y), which are categorized as "minor," "moderate," and "severe." We have several predictor variables (X_1, X_2, ..., X_n) such as weather conditions, road type, and vehicle speed. The model aims to predict the probability of a crash falling into each severity category.

$$P(Y \leq moderate) \ = \ \emptyset(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \ - \ \gamma_{moderate})$$

where,

- $P(Y \leq moderate)$ represents the probability of a crash being categorized as "minor" or "moderate."
- $\emptyset$ is the cumulative distribution function of the standard normal distribution.
- β_0, β_1, β_2, ..., β_n are the estimated coefficients obtained from the model.
- $X_1, X_2, ..., X_n$ are the predictor variables, such as weather conditions, road type, and vehicle speed.
- $\gamma_{moderate}$ is the threshold parameter specific to the "moderate" severity category.

In an ordered logit model, the probability of an outcome falling into a particular category is modeled using the logistic cumulative distribution function:

$$\begin{aligned} P(Y = j) \ = \ & exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \ - \ \gamma_j)/(1 + exp(\beta_0 + \beta_1 X_1 \\ & + \beta_2 X_2 + \ldots + \beta_n X_n \ - \ \gamma_j)) \end{aligned} \qquad (3)$$

The coefficients β_0, β_1, β_2, ..., β_n represent the estimated regression coefficients, while γ_j represents the threshold parameter for category j. The logistic function transforms the linear combination of predictors into a probability value between 0 and 1. By looking at the coefficient estimates, researchers can figure out how different things affect whether a car crash or injury is more or less severe. This information is valuable for identifying significant risk factors and informing interventions and policies to reduce crash severity and improve road safety.

7.2 Deterministic Machine Learning

7.2.1 Decision Tree

A decision tree is a supervised learning algorithm that uses a hierarchical structure to make predictions or classify data based on a series of if-else conditions. It can be represented as a flowchart-like structure where each internal node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label or a prediction. The decision tree algorithm builds the tree by repeatedly applying the rule over and over to successive results to group the data based on the values of input features. The goal is to create subsets of data that are as pure as possible regarding the target variable. The purity of a subset is typically measured using metrics such as Gini impurity or entropy. These metrics help us understand how well a subset of data is organized or how mixed its categories are. For example, imagine sorting a bag of marbles by color. If each subset contains only one color, it's considered pure. But if the colors are mixed, the subset is impure. Gini impurity and entropy give us numbers that represent this purity or impurity, helping us make decisions in machine learning algorithms, like decision trees.

Let’s consider the simplest decision tree: A single if-else statement. Say we want to predict someone’s gender, given their height. We have the data for 10 people. It’s naïve to do this, but assume that’s all we have. This is our data (bold is female, italics is male, height in centimeters): **148**, **157**,**158**,*162*,**164**,**168**,*172*,*176*,*180*,*184*. We want to find the threshold value below which we would predict female, or else male. Let’s focus on the group on the left. For any threshold we choose, we want the group to be as *homogeneous* or as *pure* as possible. Let’s say we choose 170 as the threshold. Then, the group on the left would have one “impurity” (162), and the group on the right would have none. If we choose 160 as the threshold, the left group would have no impurities, while the group on the right would have two (164,168).

Gini impurity can be seen as a way to quantify how “good” a group is, so that we can choose the threshold wisely. If a group has all females or all males, the Gini impurity is zero. If it is 50% male and 50% female, then the Gini impurity will be 0.5 (which is the highest value it can hold in this case), and it is the worst-case scenario. Hence, if we go by Gini impurity, a threshold of 182 is terrible (it leads to a group of 5 females and 4 males). And so is 150 (which leads to a group of 5 males and 4 females). So, we would choose something like 170 which intuitively seems to result in a low proportion of impurities in both groups. So, in the bigger picture, when you're deciding a split in the decision tree, you want to maximize the difference between the Gini impurity of the parent and the sum of the Gini impurities of the children nodes.

The decision tree splits the data at each internal node based on a selected feature and a chosen splitting criterion. The splitting criterion determines how well the data is divided into different classes or categories. For example, in a binary classification problem, the Gini impurity is commonly used as the splitting criterion. It measures the probability of misclassifying a randomly chosen element from the subset. The decision tree continues to split the data recursively until a stopping criterion is met. This can be based on various conditions, such as reaching a maximum depth, having a minimum number of samples in a leaf node, or achieving a desired level of purity. Once the decision tree is constructed, it can be used to predict new instances by traversing the tree from the root node to a leaf node based on the values of the input features. The class label or prediction associated with the reached leaf node is then assigned to the instance. Figure 6 shows a decision tree for buying a car.

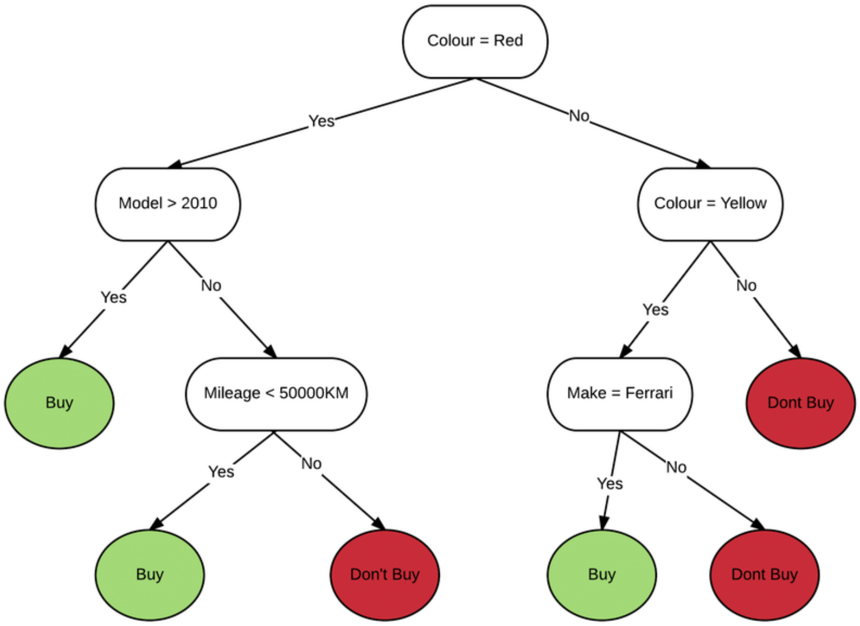


Figure 36. Decision Tree for Buying a Car

7.2.2 Random Forest

Random forest is an ensemble learning method that combines multiple decision trees to make predictions. It is a powerful and popular algorithm known for handling complex problems and producing accurate results. In a random forest, a set of decision trees is trained on different subsets of the original training data. Each decision tree is constructed using a random subset of features at each split. This random feature selection helps reduce the correlation among the trees and increases the diversity of the ensemble. During the training stage, multiple decision trees are grown by repeatedly selecting a random subset of the training data with replacement (known as bootstrapping). For each tree, a random subset of features is selected at each split. The trees are grown until a stopping criterion is reached, such as reaching a maximum depth or having a minimum number of samples in a leaf node.

The prediction stage involves aggregating the predictions of all the individual trees in the forest. The most common aggregation method for classification tasks is voting, where each tree's prediction is counted as a vote, and the class with the majority of votes is assigned as the final prediction. The individual tree predictions are averaged for regression tasks to obtain the final prediction. The strength of random forest lies in its ability to handle high-dimensional data, deal with missing values, and mitigate overfitting. Combining multiple tree predictions, random forest improves the generalization performance and provides robustness against noise and outliers in the data. The prediction of a random forest can be mathematically represented as:

$$\hat{Y} = \frac{1}{N}\sum_{i=1}^{N} f_i(X) \tag{4}$$

where $\hat{Y}$ is the predicted output, N is the number of trees in the forest, and $f_i(X)$ represents the prediction of the i-th tree based on the input features X. Random forest has become a popular choice in various domains, including classification, regression, feature selection, and anomaly detection, due to its versatility, robustness, and ability to handle large datasets. An example of a random forest structure is shown in Figure 37.

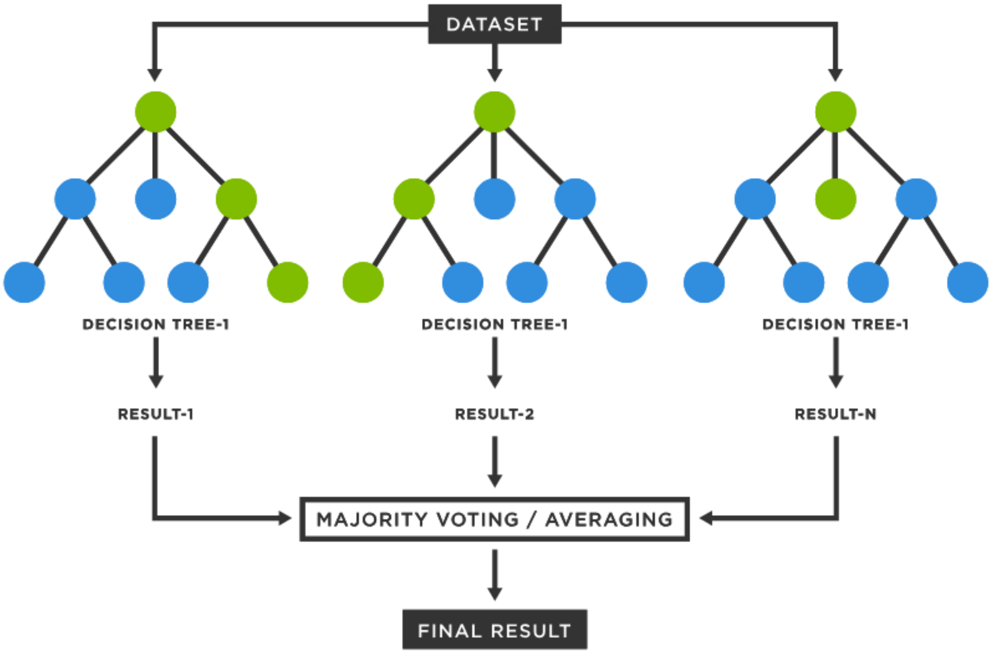


Figure 37. Random Forest Diagram

7.2.3 Support Vector Machines (SVM)

SVM is a popular machine learning algorithm used for both classification and regression tasks. SVM is a classifier that aims to find an optimal hyperplane that separates data points of different classes in a high-dimensional feature space. The main idea behind SVM is to find the hyperplane that maximizes the margin between the nearest data points of different classes. These data points, known as support vectors, play a crucial role in defining the decision boundary. SVMs can handle linearly separable data by using a linear kernel, but they can also handle nonlinear data by utilizing kernel functions that map the data into a higher-dimensional space. Mathematically, SVM can be formulated as an optimization problem:

$$min_{w,b} \frac{1}{2} \|\boldsymbol{w}\|^2 + C \sum_{i=1}^{N} \gamma_i \qquad (5)$$

Subject to: $y_i(w.x_i + b) \geq 1 - \gamma_i$

$$\gamma_i \geq 0$$

where $\boldsymbol{w}$ represents the weight vector, b is the bias term, N is the number of training samples, x_i denotes the feature vector of the i-th sample, y_i is the corresponding class label, and γ_i are slack variables that allow for a certain degree of misclassification. The parameter C controls the trade-off between maximizing the margin and allowing some misclassifications.

SVMs are capable of handling data with complex decision boundaries and have good generalization properties. They can effectively handle high-dimensional data and are less prone to overfitting compared to other models. Additionally, SVMs can handle datasets with a small number of training samples. However, SVMs can be computationally expensive and may require careful selection of kernel functions and tuning of hyperparameters. In addition to binary classification, SVMs can be extended to handle multi-class classification tasks using approaches such as one-vs-one or one-vs-rest. SVMs can also be applied to regression problems by modifying the objective function and incorporating a margin-based loss.

7.2.4 Neural Networks

Neural Networks, also known as Artificial Neural Networks (ANN), are a class of machine learning models inspired by the structure and function of the human brain. Neural networks are composed of interconnected nodes, called neurons, which are organized into layers. Each neuron takes inputs, performs a computation, and produces an output. The basic building block of a neural network is the neuron. The neuron takes a weighted sum of its inputs, applies an activation function to the sum, and produces an output. The weights of the inputs determine the importance of each input in the computation. The activation function introduces non-linearity into the model, enabling the neural network to learn complex patterns and relationships in the data (Figure 38).

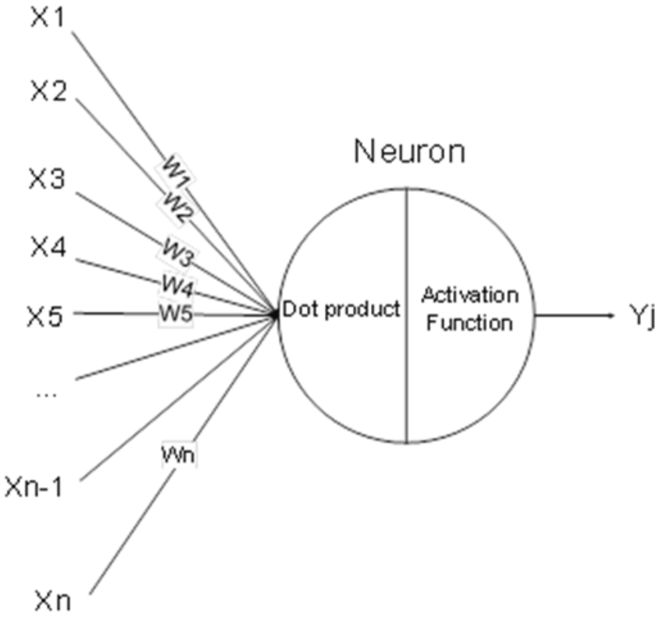


Figure 38. Neuron Structure

Neural networks consist of an input layer, one or more hidden layers, and an output layer. Information flows through the network from the input to the output layer. During training, the network adjusts its weights using an optimization algorithm, such as gradient descent, to minimize a loss function that measures the discrepancy between predicted and true outputs. This process is known as backpropagation, where the error is propagated backward through the network to update the weights. Neural networks are highly flexible and can model complex nonlinear relationships in data. They can learn from large amounts of labeled data and generalize well to unseen examples. However, training neural networks can be computationally intensive and requires careful tuning of hyperparameters, such as the number of layers, number of neurons, and learning rate. Additionally, neural networks are prone to overfitting if the model is too complex, or the training data is limited. Overall, neural networks have revolutionized the field of machine learning and have become a fundamental tool for solving complex problems in diverse domains.

.