Unveiling the Patterns in Traffic Incidents: A Personal Journey Through Data

Ruonan Ji, Mingxuan Yu

https://github.com/ymxnaldo9/DS5500

1. Abstract

This paper presents a comprehensive analysis of road incidents, utilizing advanced data mining techniques to uncover underlying patterns and correlations critical for enhancing road safety measures. By meticulously dissecting the "Crash Reporting - Drivers Data.csv" dataset, we aimed to illuminate the factors contributing to road mishaps. Our approach involved segmenting the dataset for training (80%) and testing (20%) and employing logistic regression, decision trees, random forests, and neural networks. We achieved a notable accuracy of over 80% on the test set, providing valuable insights into time and environmental influences on crash occurrences.

2. Introduction

Road traffic incidents remain a pressing concern worldwide, necessitating an in-depth understanding of their dynamics for effective mitigation. This study embarked on an explorative journey, employing a data-driven methodology to dissect the nuances of traffic incidents. Our primary focus was on predicting the severity of injuries in automobile accidents and identifying key environmental and vehicular factors influencing these events.

3. Methods

3.1 Data preprocessing

The roads, in their silent witnessing of daily commutes, hold untold stories - narratives often lost in the bustling rhythms of everyday life. This paper serves as a conduit, translating those unspoken numbers into narratives that demand attention and action. Our exploration is

driven by a dual purpose: firstly, to forecast the severity of injuries resulting from automobile accidents, and secondly, to identify and understand the critical factors that lead to these incidents, with a particular focus on environmental conditions and vehicle speed.

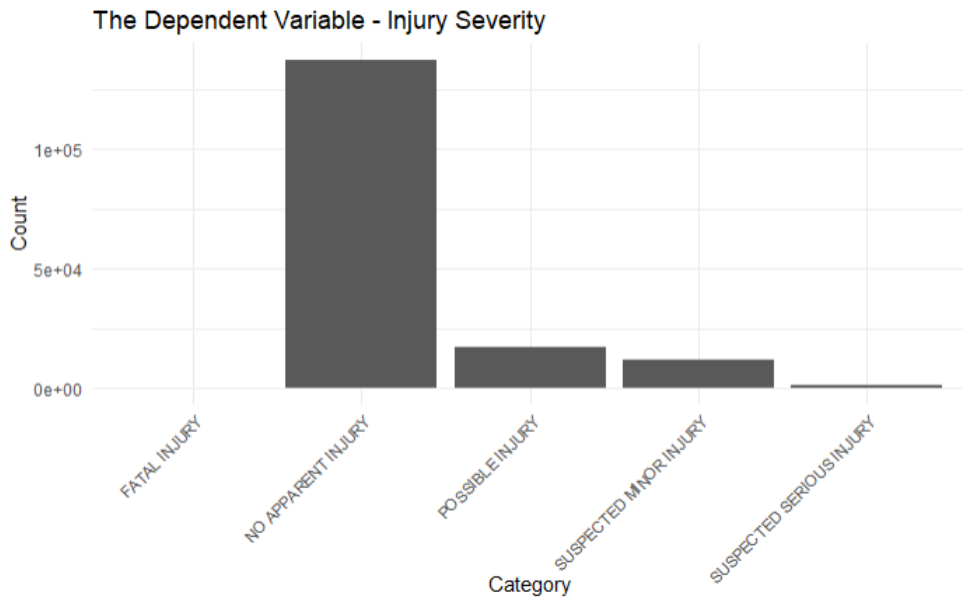
We delve into a dataset that is a rich mosaic of 167,491 observations, encompassing 43 varied numerical and categorical variables. Central to this analysis is the dependent variable, `Injury Severity`, which stands as a key to unlocking the patterns hidden within this extensive collection of raw data.

	Report Number	Local Case Number	Agency Name	ACRS Report Type	Crash Date/Time
1	MCP3040003N	190026050	Montgomery County Police	Property Damage Crash	05/31/2019 03:00:00 PM
2	MCP1307000K	190024786	Montgomery County Police	Property Damage Crash	05/24/2019 05:00:00 PM
3	MCP2846008X	230034260	Montgomery County Police	Property Damage Crash	07/17/2023 10:45:00 AM
4	MCP32610017	230034668	Montgomery County Police	Property Damage Crash	07/20/2023 11:40:00 PM
5	EJ78520081	230033429	Gaithersburg Police Depar	Property Damage Crash	07/13/2023 05:40:00 PM
6	MCP3163005L	230035071	Montgomery County Police	Property Damage Crash	07/23/2023 03:55:00 PM
7	MCP33080022	230032584	Montgomery County Police	Injury Crash	07/08/2023 05:51:00 PM

NO APPARENT INJURY	137253
POSSIBLE INJURY	17011
SUSPECTED MINOR INJURY	11633
SUSPECTED SERIOUS INJURY	1382
FATAL INJURY	151
NONE DETECTED	46
MEDICATION PRESENT	12
UNKNOWN	2
ALCOHOL CONTRIBUTED	1

In our analysis, we primarily focus on the first five categories of injury severity, ranging from 'NO APPARENT INJURY' to 'FATAL INJURY'. This range presents a clear gradation of

injury severity, escalating from minor or non-visible injuries to life-threatening conditions. The other four categories, which do not align with this progressive scale of severity, were excluded from our dataset. Following this refinement, the distribution of data across each of the retained categories is illustrated in the subsequent chart:

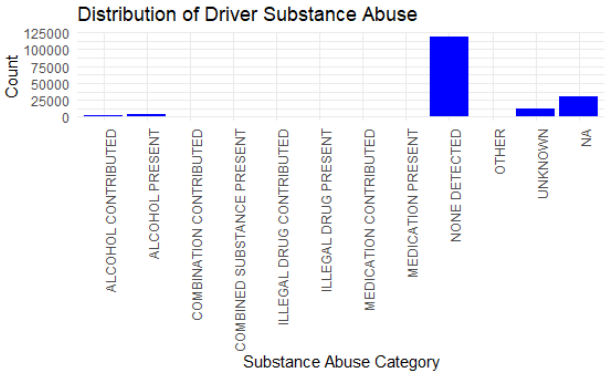


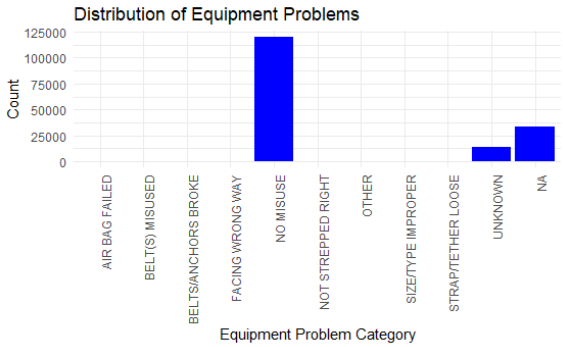
The initial stage of our data mining process involved a detailed examination for missing values and a thorough review of the variables. Initially, we identified and removed columns with an excessive amount of missing data, specifically those with more than a 90% absence of information. This led to the elimination of three particular variables: 'Off-Road Description', which indicates alcohol consumption; 'Related Non-Motorist', detailing the involvement of bicyclists or pedestrians; and 'Non-Motorist Substance Abuse', denoting substance detection. The high volume of missing data in these categories necessitated their removal.

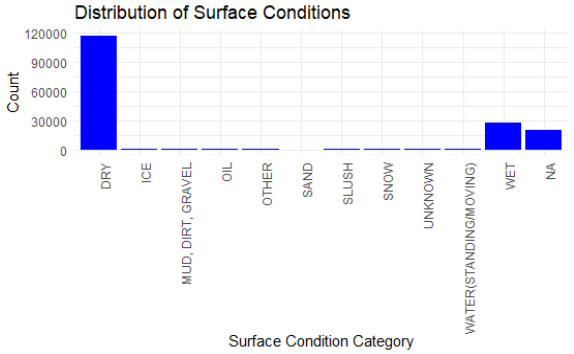
Our scrutiny extended beyond mere empty fields. We treated entries labeled ‘N/A’ or ‘UNKNOWN’ as missing data. This was particularly relevant for the 'Circumstance' and 'Municipality' variables, which we discovered to be excessively incomplete. We opted to exclude

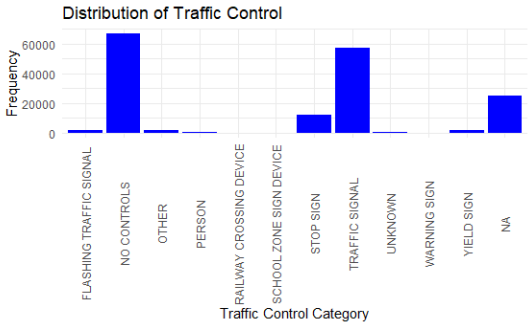
these variables instead of imputing them, as their information largely overlapped with that contained in the 'Weather' and 'Cross-Street Type' variables, leading to redundancy.

Post these eliminations, we utilized the summarise() function in R to reassess the dataset for remaining missing values. Encouragingly, most variables exhibited a low percentage of missing data, affirming that our imputation methods would not significantly alter the dataset's overall integrity. However, four variables still had more than 10% missing data. We decided to remove 'Driver Substance Abuse' due to its predominant 'non-detected' status, rendering it unhelpful for predicting 'Injury Severity', our dependent variable. Similarly, 'Equipment Problems' was dropped, as most of its values indicated 'non-misuse'. On the other hand, 'Surface Condition' and 'Traffic Control' presented a more diverse set of data and were retained for further imputation and distribution analysis.

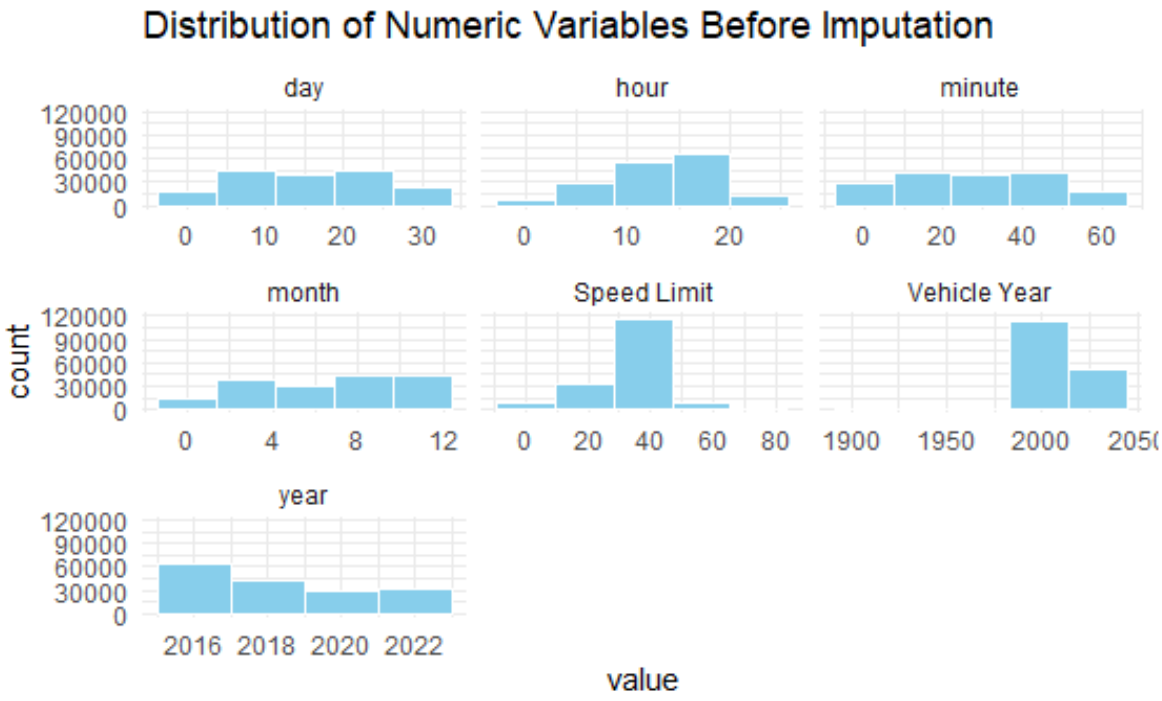


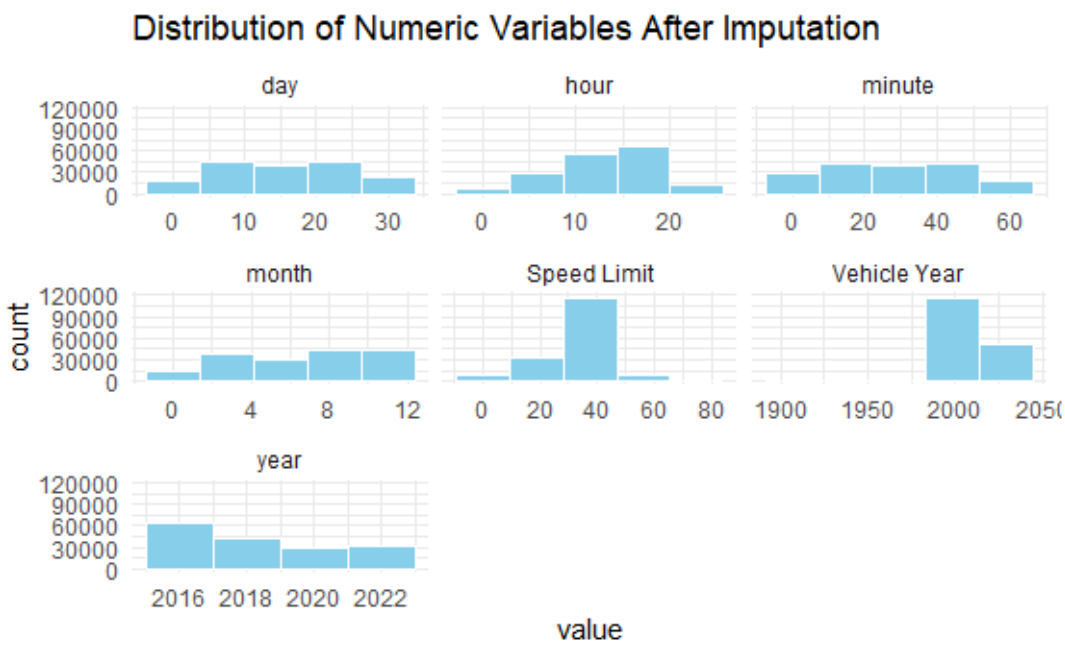






Our approach to imputation was straightforward yet efficacious, employing the mode for categorical data and the median for numerical data. This technique ensured that the inherent characteristics of our dataset were preserved. When examining the plots for numerical variables post-imputation, a remarkable consistency was observed across the board. The only notable variation appeared in the 'Vehicle Year' data, particularly around the year 2000, which aligns with our expectations given that this was the primary numerical variable with missing entries. Despite this slight fluctuation, the overall distribution of the data remained largely unaltered, affirming the reliability and effectiveness of our imputation strategy.





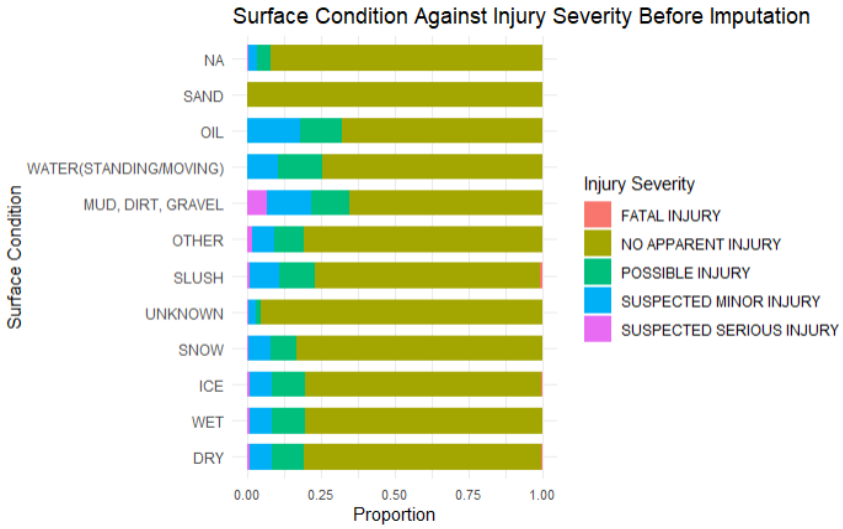
The insightful visualizations derived from our dataset revealed distinct patterns, each telling its own story about the nature of motor vehicle accidents:

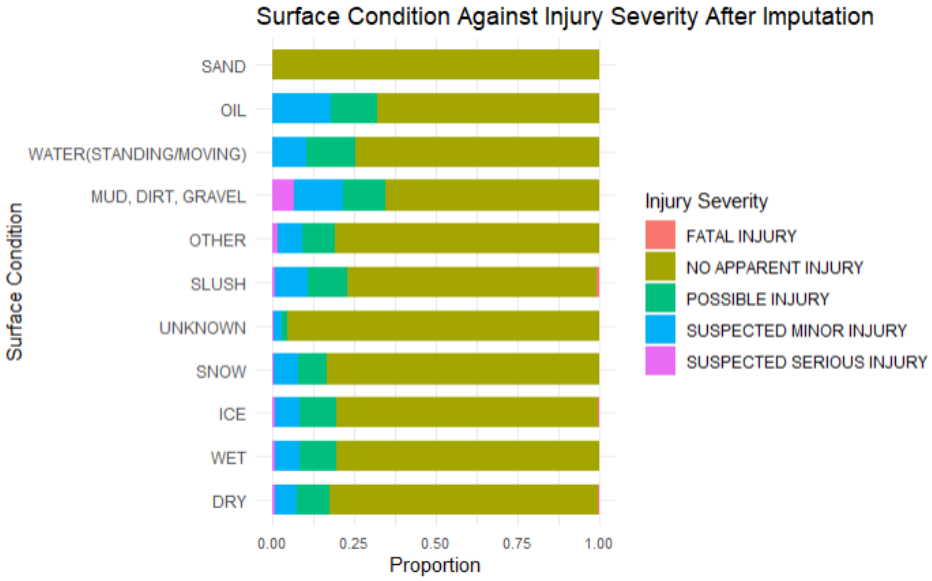
- We observed a significant clustering of accidents during the midday to evening hours, with a predominant occurrence in areas where the speed limit is set at 40 mph. This pattern underscores a correlation between specific times of day and speed limits with the frequency of accidents.
- An intriguing trend emerged in the make of the vehicles involved in these incidents. There was a notable prominence of vehicles manufactured between 1980 and 2010 in the crash statistics. However, a marked decrease in accidents was observed with vehicles manufactured post-2010, suggesting advancements in vehicle safety features or changes in driving behavior with newer models.

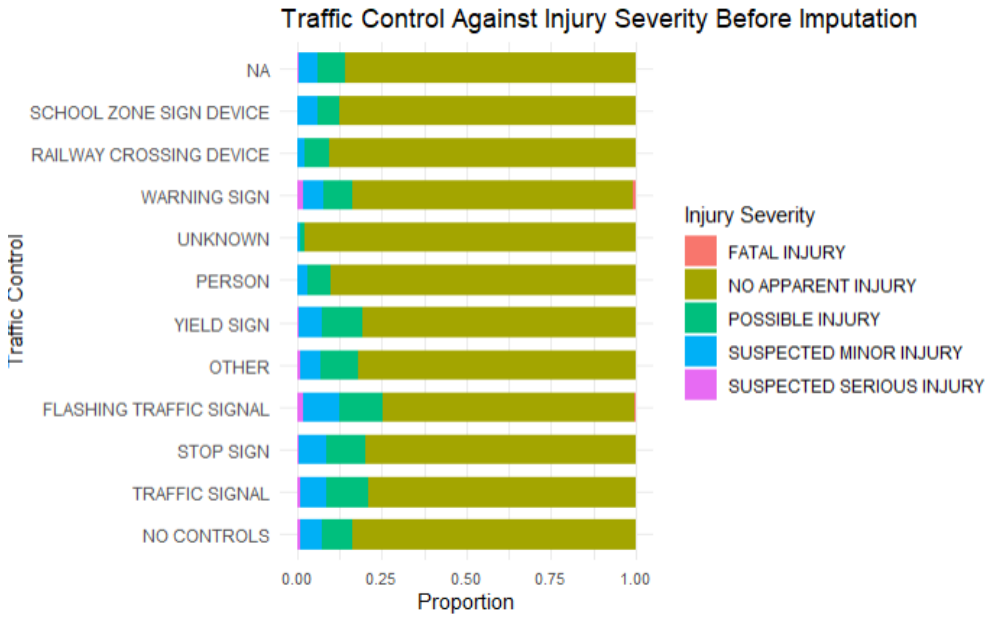
- Encouragingly, the overall trend line for accidents per year exhibited a downward slope. This suggests that, over the years, there have been significant improvements in road safety measures, whether through enhanced vehicle technology, better road conditions, or more effective traffic management and safety campaigns.

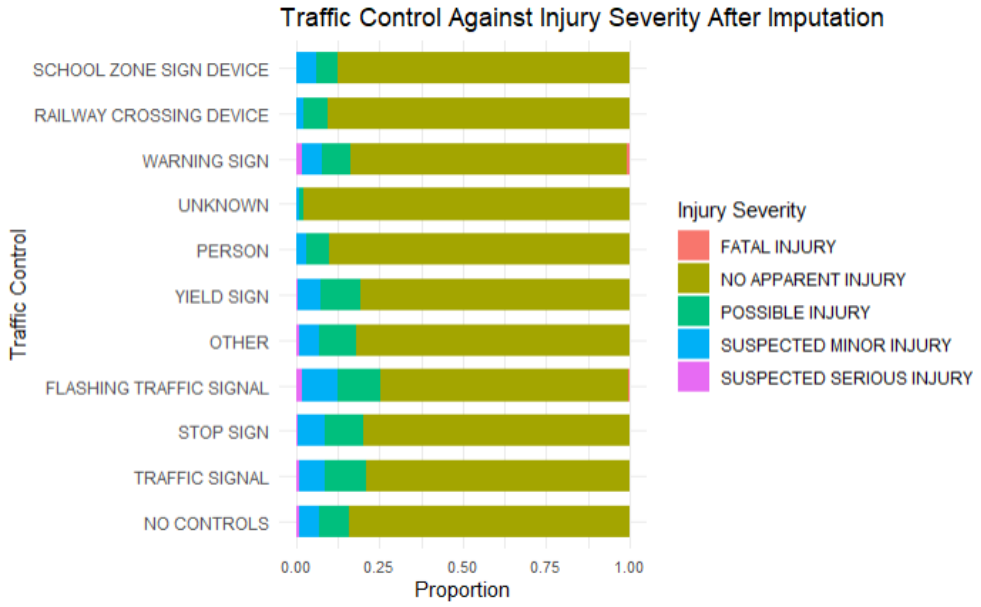
These patterns not only provide a deeper understanding of the factors influencing road safety but also highlight areas for potential improvements and further investigation.

The analysis of the two categorical variables post-imputation revealed that their structural integrity was well-preserved. This consistency in structure indicates that our chosen imputation methods — employing the mode for categorical data and the median for numerical data — were aptly suited for our dataset. By maintaining the original distribution patterns of these variables, we ensured the reliability of our subsequent analyses, affirming the adequacy of our imputation approach in handling missing data effectively.


Surface Condition Against Injury Severity Before Imputation
NA
SAND
OIL
WATER(STANDING/MOVING)
MUD, DIRT, GRAVEL
OTHER
SLUSH
UNKNOWN
SNOW
ICE
WET
DRY
Surface Condition
0.00
0.25
0.50
0.75
1.00
Proportion
Injury Severity
FATAL INJURY
NO APPARENT INJURY
POSSIBLE INJURY
SUSPECTED MINOR INJURY
SUSPECTED SERIOUS INJURY

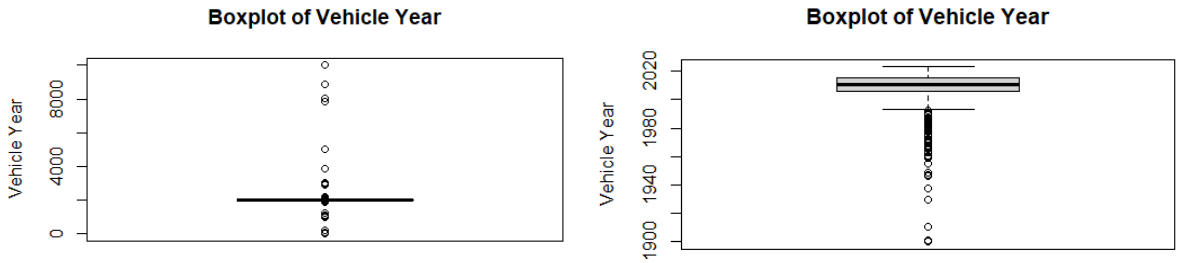

Surface Condition Against Injury Severity After Imputation
SAND
OIL
WATER(STANDING/MOVING)
MUD, DIRT, GRAVEL
OTHER
SLUSH
UNKNOWN
SNOW
ICE
WET
DRY
Surface Condition
0.00
0.25
0.50
0.75
1.00
Proportion
Injury Severity
FATAL INJURY
NO APPARENT INJURY
POSSIBLE INJURY
SUSPECTED MINOR INJURY
SUSPECTED SERIOUS INJURY





Upon closer examination of our dataset, we identified and subsequently removed nine variables, such as `Report Number` and `Local Case Number`, which were deemed non-contributory to the prediction of our primary outcome variable. This decision to streamline the dataset was informed by the realization that these variables lacked predictive relevance or explanatory power regarding the severity of traffic incidents.

Additionally, our detailed outlier analysis revealed certain irregularities, particularly within the 'Vehicle Year' variable. We encountered anomalous values such as 0 and 9999, which fell outside the realm of plausible years for vehicle manufacture. To address this, we meticulously adjusted the 'Vehicle Year' data, restricting it to a more realistic and historically accurate range, spanning from 1900 to 2023. This correction was crucial in preserving the integrity and validity of our analysis, ensuring that our dataset reflected a true-to-life representation of vehicle involvement in traffic incidents.



With a thorough understanding of the dataset now established, we are poised to construct a classification model, with the severity of injuries sustained as the dependent variable. This model serves as a multifaceted prism, designed to refract and reveal the various aspects of the data that potentially influence injury severity.

For this model, we have selectively chosen a set of features that we hypothesize to be significant in predicting injury severity. These features include Cross-Street Type, Collision Type, Weather, Surface Condition, Light, Traffic Control, Driver At Fault, Driver Distracted By, Vehicle Damage Extent, Vehicle First Impact Location, Vehicle Second Impact Location, Vehicle Body Type, Vehicle Movement, and Speed Limit. This selection is based on both their statistical relevance and theoretical implications in the context of traffic incidents.

In our analysis, we focus specifically on incidents where the vehicle is actively driven, thus considering only those data rows where 'Driverless Vehicle' and 'Parked Vehicle' columns are marked 'No'. This distinction is crucial for the accuracy and relevance of our model.

Furthermore, acknowledging the practical challenges drivers face in recognizing certain vehicle-related conditions during an incident, we also endeavor to construct an alternate model. This model excludes variables like Vehicle Damage Extent, Vehicle First Impact Location, Vehicle Second Impact Location, Vehicle Body Type, and Vehicle Movement. The aim is to explore the predictability of injury severity without relying heavily on post-incident vehicle conditions, which may not always be readily accessible by a driver.

To prepare our dataset for modeling, we employ label encoding to transform all textual data into numerical format, thereby facilitating the application of various machine learning algorithms. This encoding step is instrumental in translating categorical variables into a form that can be efficiently processed and interpreted by our classification models.

4. Modeling with vehicle condition

In our comprehensive analysis of the dataset, we will employ three principal methodologies, each tailored to dissect and understand the complexities of traffic incident data in a unique way:

Decision Trees and Random Forests for Discrete Data: We will first apply decision trees and random forests, which are particularly adept at handling discrete attribute data. This approach will allow us to model non-linear relationships and interactions between various predictors and the outcome. Decision trees provide an intuitive, hierarchical structure for decision-making, while random forests, an ensemble of such trees, enhance accuracy and prevent overfitting by averaging multiple decision trees' predictions.

Linear Regression for Injury Severity Scaling: To analyze the different degrees of injury severity, we will assign numerical values to each category of injury severity and utilize linear regression. This method helps in understanding the linear relationship between the predictors and the varying degrees of injury severity. By quantifying injury severity, linear regression will offer insights into how different factors incrementally increase or decrease the severity of injuries in accidents.

Logistic Regression and Neural Networks with One-Hot Encoding: Lastly, we will implement logistic regression and forward neural networks, coupled with one-hot encoding of the injury severity categories. This approach is ideal for multi-classification problems where the outcome variable consists of multiple classes, as is the case with our injury severity variable. The softmax function in neural networks will be particularly useful for handling the multi-class nature of our outcome, providing probabilities for each injury severity category, thereby enabling us to classify each incident into distinct severity levels.

These diverse methodologies will collectively enable a thorough and multifaceted analysis of the data, allowing us to uncover the intricate patterns and relationships within the variables, and ultimately providing a comprehensive understanding of the factors influencing injury severity in traffic incidents.

4.1.1 Decision tree and random forest

Given the presence of numerous discrete features in our dataset, decision trees and random forests naturally emerge as fitting choices for our model training. These methods are particularly effective in handling discrete data, making them ideal for our analysis.

Decision trees, especially those using the ID3 algorithm, are well-suited for processing discrete data. The ID3 algorithm employs information entropy as a key metric to construct the

tree. Information entropy, in this context, measures the degree of uncertainty or impurity in the dataset. It plays a pivotal role in determining the most informative features to split the data at each node of the tree. By using entropy as a criterion, the decision tree aims to increase the homogeneity of the resultant subsets with each split, thereby maximizing the information gain.

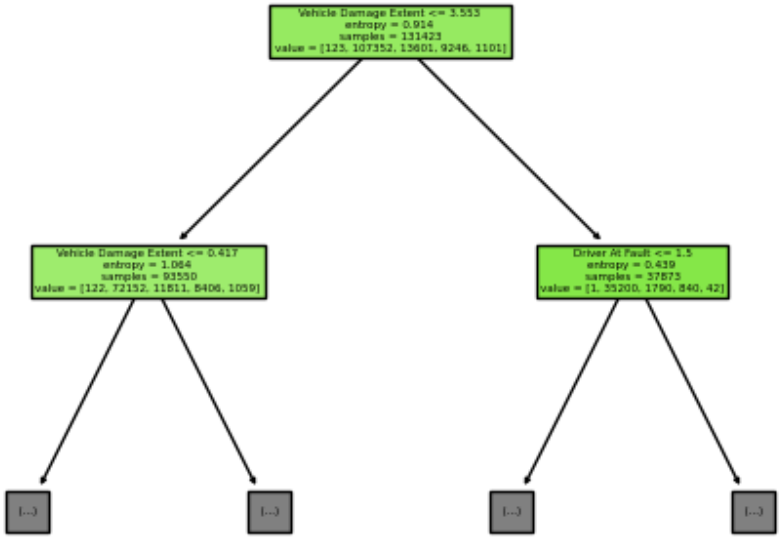
To deepen our understanding of each feature's influence within the model, we begin by calculating and examining the information entropy of each variable. This metric will serve as the primary criterion for splitting the data at the initial node of the decision tree. It allows us to identify which features contribute most significantly to reducing uncertainty about the target variable, in this case, the injury severity. By prioritizing features with higher entropy for earlier splits, the decision tree can more effectively classify the data, leading to a more accurate and interpretable model.

In summary, employing decision trees and random forests, with a focus on information entropy, will enable us to dissect the intricate structure of our dataset. This approach will uncover the most salient features that influence injury severity in traffic incidents, providing valuable insights for our analysis.

Cross-Street Type	0.0107
Collision Type	0.0448
Weather	0.0052
Surface Condition	0.0070
Light	0.0168
Traffic Control	0.0192

Driver At Fault	0.1590
Driver Distracted By	0.0209
Vehicle Damage Extent	0.5361
Vehicle First Impact Location	0.0611
Vehicle Second Impact Location	0.0216
Vehicle Body Type	0.0437
Vehicle Movement	0.0327
Speed Limit	0.0213

In decision tree modeling, high information entropy in a dataset signifies greater chaos and uncertainty. This is particularly evident in nodes with a mixture of diverse categories, making classification more challenging. Analyzing our dataset, we find 'Vehicle Damage Extent' to be highly effective in reducing this entropy, thus aiding in clearer classification. Similarly, 'Driver At Fault' also plays a significant role in diminishing uncertainty. In contrast, features like 'Weather' and 'Surface Condition' exhibit lesser impact on entropy reduction, making them less suitable for initial classification in the decision tree. This understanding helps prioritize the most impactful features for early splits in the model, enhancing the accuracy of our injury severity classification.



We constructed random forest models with ten and one hundred decision trees, similar to the one used initially. Interestingly, their performance closely mirrored that of the single decision tree. This outcome suggests the original decision tree was already effective in capturing the dataset's key patterns. The similarity in results indicates that the additional complexity of a random forest did not substantially improve predictive accuracy in this case. Performance metrics like accuracy, precision, recall, and F1 score are essential for evaluating these models, but the initial decision tree's effectiveness is clear from this experiment.

	Accuracy	F1 Score	Precision	Recall
Decision tree	81.74%	74.06%	70.84%	81.74%

Random forest(10 trees)	81.91%	73.78%	69.99%	81.91%
Random forest(100 trees)	81.91%	73.76%	67.09%	81.91%

4.1.2 Stepwise linear regression

Acknowledging the varying degrees of injury severity, we aim to represent these different levels with distinct values and apply stepwise linear regression for model training. This method involves incrementally adding or removing features, basing each decision on the p-value of the feature.

The p-value serves as a statistical measure, indicating the extent to which the data supports the null hypothesis - the assumption that a specific feature does not influence the model. A p-value below 0.05 suggests that the likelihood of the observed data occurring under the null hypothesis is low, implying that the feature in question has a statistically significant impact. In our analysis, a feature is retained in the model if its p-value is less than 0.05, suggesting its significant role in predicting injury severity.

This stepwise approach allows us to methodically construct a model that includes only the most impactful features, ensuring both efficiency and accuracy in our regression analysis. The process iteratively evaluates the contribution of each feature, reflected by the changes in p-values, and builds a model that best represents the underlying relationships in the data.

New feature added	P value
Vehicle Damage Extent	0.0
Driver At Fault	0.0
Speed Limit	1.65e-49
Vehicle Second Impact Location	7.43e-43
Driver Distracted By	1.21e-38
Vehicle Body Type	1.92e-12
Vehicle Movement	2.71e-7
Traffic Control	5.19e-7
Light	9.36e-5
Cross-Street Type	0.0049

The findings from the linear regression model align with those of the decision tree analysis, confirming that 'Vehicle Damage Extent' and 'Driver At Fault' are the primary and secondary most influential factors affecting Injury Severity. Based on this insight, we constructed a linear regression model incorporating these key features. This model is designed to quantitatively assess how these variables impact the severity of injuries sustained in traffic incidents. The performance of this model is evaluated using standard indicators, which typically include measures like accuracy, R-squared, mean squared error, or others pertinent to regression analysis.

These indicators will help us understand the effectiveness and predictive power of the model in relation to the identified influential factors.

	Accuracy	F1 Score	Precision	Recall
Linear Regression	80.62%	74.75%	70.36%	80.62%

The linear regression model, focusing on 'Vehicle Damage Extent' and 'Driver At Fault', exhibits slightly lower performance compared to the decision tree. However, the difference in their effectiveness is marginal. This outcome suggests that while both methodologies are valid for analyzing injury severity in traffic incidents, the decision tree might be capturing some nuanced patterns or interactions between variables more effectively than the linear model. Nonetheless, the linear regression provides valuable insights, especially in understanding the direct linear relationships between the key variables and injury severity.

4.1.3 Logistic regression and neural network: classification with softmax

In addressing the multi-class classification challenge, we utilized logistic regression and a forward neural network with a softmax function, combined with one-hot encoding. Our neural network featured two configurations: one with two hidden layers of 12 and 8 neurons, and another with 10 and 5 neurons. Despite these variations, both models reached a final loss value of 1.0901, indicating the limit of learning from the dataset. This outcome highlights the balance achieved between underfitting and overfitting, providing valuable insights into classifying injury severity.

	Accuracy	F1 Score	Precision	Recall

Logistic regression	81.91%	73.77%	67.10%	81.91%
Neural network(first model)	81.91%	73.77%	67.10%	81.91%
Neural network(second model)	81.91%	73.77%	67.10%	81.91%

The results across the different algorithms, including logistic regression, neural network with softmax, decision tree, and linear regression, displayed a consistent pattern. This consistency underscores a fundamental alignment in the predictive outcomes, regardless of the varied methodologies applied. Each model, despite its unique approach, essentially converged on similar conclusions regarding the classification and prediction of injury severity.

4.2 Modeling without vehicle condition

We trained models excluding vehicle condition features to assist drivers in assessing injury severity based on crash circumstances, rather than post-accident vehicle conditions. This approach acknowledges the practical challenge for drivers to evaluate vehicle damage accurately immediately after an accident. Therefore, we focused on crash-specific information, aiming to provide insights directly from the incident's cause and context.

The methodology for these models, similar to that in section 5.2, primarily differs in the exclusion of vehicle-specific features. The emphasis here is on the final results, particularly

analyzing the entropy values in decision tree and random forest models to understand the influence of crash-related features on predicting injury severity.

For the decision tree and random forest, the entropy of each features are shown as following:

Cross-Street Type	0.0312
Collision Type	0.4046
Weather	0.0216
Surface Condition	0.0181
Light	0.0304
Traffic Control	0.0621
Driver At Fault	0.1798
Driver Distracted By	0.0813
Speed Limit	0.1908

The corresponding indicators are shown as following:

	Accuracy	F1 Score	Precision	Recall
Decision tree	81.84%	73.76%	69.58%	81.84%
Random forest(10 trees)	81.91%	73.77%	74.17%	81.91%

Random forest(100 trees)	81.91%	73.77%	67.10%	81.91%

For linear regression, the process of stepwise linear regression is shown as following:

New feature added	P value
Speed Limit	0.0
Driver At Fault	2.66e-188
Traffic Control	6.23e-13
Driver Distracted By	2.35e-8
Cross-Street Type	0.0068
Surface Condition	0.0123

The corresponding indicators are as following:

	Accuracy	F1 Score	Precision	Recall
Linear Regression	81.91%	73.77%	67.10%	81.91%

For logistic regression and neural network with softmax. The hidden layers of neural network are respectively 8 and 4:

	Accuracy	F1 Score	Precision	Recall

Logistic regression	81.91%	73.77%	67.10%	81.91%
Neural network	81.91%	73.77%	67.10%	81.91%

5. Results

We trained models on datasets both with and without vehicle condition features. Across these models, the performance metrics were notably similar. We consistently achieved an accuracy of approximately 81.91%, a F1 score around 73.77%, precision about 67.10%, and recall also at 81.91%. In the dataset including vehicle conditions, 'Vehicle Damage Extent' and 'Driver At Fault' emerged as the most influential factors for Injury Severity. Conversely, in the dataset excluding vehicle conditions, different features gained prominence. The decision tree analysis highlighted 'Collision Type' as a primary factor, while stepwise linear regression identified 'Speed Limit' and 'Driver At Fault' as key influencers. This variation presents an intriguing aspect, suggesting further exploration into the dataset might yield additional insights.

6. Discussion

For drivers, minimizing errors is crucial for accident prevention. Additionally, factors like highway speed limits and traffic control play significant roles in traffic engineering, heavily influencing both the frequency of accidents and the severity of injuries involved. Therefore, careful consideration and setting of these parameters are essential to enhance road safety.

In the next phase of our project, we plan to extend our work in two key areas:

1. App Development: Leveraging the findings from our current research, we aim to develop applications that enable drivers to assess their situation more effectively post-accident. These apps, rooted in our data-driven insights, will provide drivers

with practical tools to evaluate the potential severity of injuries and the necessary steps to take following an incident.

2. Integrating GPT-3: We also plan to explore the integration of GPT-3 into our model. By inputting prompts into the GPT-3 model, we intend to assess whether this AI-driven approach can replicate or enhance the effects observed with traditional machine learning algorithms in analyzing daily traffic data sets. This exploration will potentially open new avenues for advanced data interpretation and user interaction within our application framework.

These advancements aim not only to augment the practical application of our research findings but also to explore the innovative fusion of traditional data analysis with cutting-edge AI technologies.

7. Statement of contributions

Ruonan Ji: basic exploration on the dataset, data preprocessing

Mingxuan Yu: data preprocessing and modeling

8. References

National Highway Traffic Safety Administration. (n.d.). Crash Reporting - Drivers Data. Data.gov. Retrieved November 1, 2023, from https://catalog.data.gov/dataset/crash-reporting-drivers-data

9. Appendix

https://github.com/ymxnaldo9/DS5500