Analytical Innovations for Safer Mobility: Redefining Road Safety

Accident severity / frequency prediction | Traffic Design Consulting | Insurance Products

**Group Name:** Big Data Think Tank

**Team Members**: Srishti Agarwal, Anant Bairagi, Domenic Diaz, Fred Xue

# 

# **Table of Contents**

[**Table of Contents**](#_j2162jyqxz9b) **2**

[**Executive Summary**](#_o87t77mnukfv) **3**

[**Introduction**](#_ckw8bahcnt3g) **4**

[**Business Objective**](#_ekg1c8u7jp5s) **4**

[**Business Initiatives**](#_v3bxiu12zn28) **5**

[**Our Clients and Services**](#_bdai1tok02jl) **5**

[**Role of Analytics**](#_6i7p5cnzyyaw) **6**

[**Analytics Methodology**](#_r7k27alffmsq) **6**

[**Data Description**](#_c7eb7ck82wtc) **7**

[**Analysis**](#_htwkxnyomxtx) **8**

[**Results**](#_y2djy0df23z3) **12**

[**Recommendations**](#_41nbynnuinoq) **17**

[**Limitations**](#_np1t60m0oz5x) **18**

# 

# **Executive Summary**

In today's society, traffic accidents pose a significant threat to public safety and impose substantial economic burdens. The Analytical Innovations for Safer Mobility: Redefining Road Safety Project addresses the urgent need for innovative solutions to enhance road safety and minimize accidents. By leveraging historical data and predictive analytics, our project aims to accurately predict accident severity and frequency by analyzing factors such as time of day, location, weather conditions, and nearby infrastructure. These predictions enable the provision of preemptive warnings to local agencies and drivers, empowering them to take proactive measures and prevent accidents.

Our project adopts a data-driven approach, utilizing advanced analytics and predictive modeling techniques to make accurate forecasts regarding accident outcomes. By incorporating various explanatory variables, our model identifies patterns, trends, and correlations, providing valuable insights for accident prevention. These insights enable local agencies to take informed actions, reducing the likelihood of accidents and improving road safety for all.

Alongside accident severity/frequency prediction, our project encompasses a range of strategic initiatives. We offer traffic design consulting services to local agencies, providing expert guidance in planning and constructing urban road and traffic solutions that prioritize safety and minimize accident risks. Through rigorous A/B testing and Propensity Score Matching (PSM), we evaluate the effectiveness of different strategies, facilitating the implementation of optimized traffic solutions.

Furthermore, our project explores the development of insurance products through hotspot analysis. By analyzing and visualizing data, we identify accident hotspots and high-risk areas, serving as a basis for tailoring insurance products that address specific risks associated with these locations. These products provide personalized coverage options and premiums, effectively managing risk for policyholders and enhancing overall insurance protection.

The Analytical Innovations for Safer Mobility Project revolutionizes road safety through the application of advanced analytics techniques. By leveraging historical data, predictive modeling, and data analysis, our project facilitates preemptive warnings, optimizes traffic solutions, and develops tailored insurance products. Our data-driven approach establishes the groundwork for proactive measures, significantly reducing accidents and creating safer driving environments for all. With the collaboration of stakeholders, policymakers, and the wider community, we can transform the landscape of road safety, saving lives, and fostering thriving communities.

# **Introduction**

In the United States, traffic accidents continue to be a pressing issue with profound implications. The statistics surrounding these accidents are both alarming and concerning, highlighting the urgency for innovative solutions. This project report aims to address the critical problem of road safety by presenting a comprehensive business objective and initiative that leverages historical data and predictive analytics to redefine the landscape of traffic safety.

According to the National Highway Traffic Safety Administration (NHTSA), in 2019, motor vehicle crashes claimed the lives of approximately 38,800 individuals in the United States. Additionally, an astounding 4.4 million people suffered injuries requiring medical attention as a result of these accidents. These numbers vividly illustrate the magnitude of the problem and the urgent need for effective intervention. Beyond the human toll, traffic accidents also place a significant economic burden on society. The NHTSA estimates that these accidents cost the United States approximately $242 billion each year in terms of medical expenses, lost productivity, property damage, and other associated costs.

To tackle this multifaceted challenge, our business objective is to leverage advanced data analytics and predictive modeling techniques to accurately forecast accident severity and frequency. By analyzing factors such as time of day, location, weather conditions, and the presence of nearby infrastructure, we aim to generate actionable insights that enable preemptive warnings to be provided to local agencies and drivers. Through proactive measures, we strive to prevent accidents from occurring and ensure the safety of all road users.

In addition to accident prediction, our initiative encompasses traffic design consulting, where we collaborate with local agencies to develop urban road and traffic solutions that prioritize safety and minimize the risk of accidents. By integrating innovative technologies and evidence-based strategies, we aim to enhance the overall safety infrastructure and create a sustainable transportation ecosystem.

This report delves into the methodologies employed, the insights gained, and the recommendations derived from our Analytical Innovations for Safer Mobility project. We explore the challenges encountered during implementation, the potential impact of our initiatives, and the pathways to scalability and sustainability. The findings presented here aim to inspire action, foster collaboration among stakeholders, policymakers, and the wider community, as we collectively strive to create a future where traffic accidents are minimized, lives are saved, and our communities flourish.

# **Business Objective**

Our primary objective is to enhance road safety and prevent accidents through the application of advanced analytics and proactive measures. By accurately predicting accident severity and frequency based on factors such as time of day, location, weather conditions, and nearby infrastructure, we aim to provide preemptive warnings to local agencies and drivers. Our goal is to empower them to take proactive measures and prevent accidents, ultimately ensuring the safety of all road users.

# **Business Initiatives**

To achieve our business objective, we have developed strategic initiatives that harness the power of advanced analytics, innovative strategies, and technology-driven solutions:

**Predictive Accident Severity/Frequency Model**: Develop and deploy a model that accurately forecasts accident severity and frequency. By leveraging historical data and analyzing factors such as time of day, location, weather conditions, and nearby infrastructure, this model will provide valuable insights into potential risks on the road. These insights will enable us to provide preemptive warnings, empowering local agencies and drivers to take proactive measures and prevent accidents.

**Traffic Design Consulting and Implementation:** Provide comprehensive consulting services to local agencies and transportation authorities to optimize traffic flow, reduce accident risks, and enhance overall transportation efficiency. By leveraging data analytics, simulations, and innovative methodologies, we will evaluate and implement effective traffic design strategies. Our aim is to create a well-designed transportation infrastructure that prioritizes safety and minimizes the risk of accidents.

**Assistance Packages for Autonomous Driving Features:** Develop assistance packages tailored specifically for autonomous driving features. By leveraging real-time information on road conditions, weather, and traffic, we will enhance the capabilities of autonomous driving systems. These assistance packages will enable proactive adaptation of driving behavior, ensuring safe and efficient navigation. Our goal is to revolutionize the integration of autonomous vehicles into existing transportation ecosystems.

**Insurance Products with Hotspot Analysis:** Employ hotspot analysis techniques to identify accident-prone areas and high-risk zones. By analyzing comprehensive data and visualizing accident hotspots, we will work closely with insurance companies to develop tailored insurance products. These products will address the specific risks associated with these locations, offering personalized coverage options, effectively managing claims ratios, and providing comprehensive risk protection to policyholders.

# **Our Clients and Services**

We collaborate closely with various stakeholders to achieve our mission of redefining road safety.

***Our clients include:***

**Local Agencies:** We work hand in hand with local government agencies responsible for road and traffic management. By providing data-driven strategies and implementing preemptive measures, we enhance road safety, minimize accidents, and improve transportation infrastructure.

**Transportation/Traffic Design Authorities:** Our partnership with transportation authorities aims to create safer and more efficient transportation systems. By leveraging data analytics and predictive modeling, we help identify high-risk areas and optimize traffic design strategies, ultimately improving traffic flow and reducing accident rates.

**Insurance Companies:** Insurance companies rely on our expertise in hotspot analysis and risk assessment to develop tailored insurance products. By analyzing accident hotspots and high-risk areas, we assist insurance companies in optimizing their risk management strategies, providing personalized coverage options, and enhancing risk protection for policyholders.

**Car Manufacturers:** We work closely with leading car manufacturers to enhance the driving experience and promote road safety. By leveraging advanced analytics and real-time information, we assist in developing autonomous driving features that can adapt to road and weather conditions, ensuring safer and more intelligent vehicles.

***Our services include:***

**Accident Severity/Frequency Prediction:** Using predictive analytics models trained on observational data, we accurately forecast accident severity and frequency. By analyzing factors such as time, location, weather, and infrastructure, we provide valuable insights for proactive accident prevention.

**Traffic Design Consulting:** We offer comprehensive consulting services to optimize traffic flow, reduce accident risks, and enhance transportation efficiency. By leveraging analytics and simulations, we develop data-driven strategies to improve traffic management.

**Autonomous Driving Assistance:** Our assistance packages for autonomous driving systems enhance their capabilities through real-time data integration. By leveraging advanced analytics, we enable autonomous vehicles to adapt to changing road conditions and proactively navigate potential hazards.

**Hotspot Analysis and Insurance Products:** Our hotspot analysis identifies accident-prone areas and high-risk zones. We collaborate with insurance companies to develop tailored insurance products that optimize risk management strategies, ensuring comprehensive coverage for policyholders.

# **Role of Analytics**

Analytics plays a pivotal role in our mission to redefine road safety. Through data exploration, predictive modeling, traffic pattern analysis, risk assessment, and hotspot analysis, analytics enables us to extract meaningful insights. These insights inform decision-making, empower proactive measures, and enable the achievement of our business objectives. By leveraging analytics, we aim to create a safer transportation ecosystem, save lives, and build flourishing communities.

# **Analytics Methodology**

The "Analytical Innovations for Safer Mobility: Redefining Road Safety" project employs a range of analytics methodologies to extract valuable insights, make accurate predictions, and optimize decision-making processes. The key analytics methodologies utilized in this project include:

**Predictive Analytics:**

Predictive analytics plays a crucial role in this project as it allows us to forecast accident severity and duration based on historical data. We employ machine learning models, such as regression analysis and RandomForest decision trees to build predictive models. These models incorporate various explanatory variables, including time of day, location, weather conditions, nearby infrastructure, and more. By analyzing the relationships between these variables and accident outcomes, we can accurately predict the likelihood and severity of accidents.

**Exploratory Data Analysis:**

Exploratory data analysis (EDA) is a critical analytics methodology used to gain insights and identify patterns in the data. By visualizing and examining the data, we uncover hidden relationships, trends, and potential interventions. We utilize techniques such as data visualization, summary statistics, and descriptive analytics to understand the characteristics of accidents, identify hotspot cities and areas, and explore risk factors. EDA helps us generate hypotheses, refine our business initiatives, and develop effective strategies to minimize risks and improve road safety.

**Prescriptive Analysis:**

Our data are observational records of accidents that had already taken place in the past. In order to evaluate the causal relationship between predictors and our target variables (severity, distance, and duration), we need to conduct quasi-experiments. Essentially, we need to recreate the causal patterns of randomized trials within observational data.

Our selected technique is A/B testing with propensity score matching. By matching records with similar characteristics into treatment and control groups, we can compare the effect of the treatment variable on an outcome variable through the lens of causal inference. Essentially, we compare functionally equivalent groups to replicate the rigorousness of randomized controlled trials. However, one important caveat is that given observational data on occurred accidents only, our causal inference through A/B testing with matching can only tell us the effect of the treatment variable on target variables once the accident happens. It cannot offer us insights on whether the treatment variable is effective in preventing accidents from happening.

# **Data Description**

This project utilizes an observational dataset consisting of 2,845,342 accident records. The dataset contains 47 columns, including 16 string, 13 decimal, 13 boolean, and 5 other data types. This data comes through Kaggle, but is originally sourced via several state and federal departments such as The United States Department of Transportation. These sources give the data an extra level of credibility and accuracy. These columns provide valuable information about each accident and its impact on traffic. In order to reveal further insights, we have made an appendage to the data to show populations at both the City and State level. This secondary dataset comes via only one source, the U.S. Census. It takes a snapshot of State/City populations in 2022.

**Accident Details:**

We have columns such as ID, Severity, Start\_Time, End\_Time, Start\_Lat, Start\_Lng, End\_Lat, End\_Lng, Distance (mi), Duration (hr), and Description. These columns provide essential details about each accident, such as its unique identifier, severity level, start and end times, geographical coordinates, the extent of road affected, and the duration of the accident. The Description column offers human-provided descriptions, which further enhance our understanding of the accident circumstances.

**Geographic Location:**

We have columns such as Number, Street, Side, City, County, State, Zip Code, Country, and Timezone. These columns provide information about the specific location of the accidents, including the street number, street name, relative side of the street, city, county, state, zip code, country, and timezone. Understanding the geographic context of accidents helps identify patterns and trends in different areas, contributing to the development of targeted road safety measures.

**Weather Conditions:**

Columns such as Airport\_code, Weather\_timestamp, Temperature (F), Wind\_Chill (F), Humidity (%), Pressure (in), Visibility (mi), Wind\_Direction, Wind\_Speed (mph), Precipitation (in), and Weather\_Condition, capture the prevailing weather conditions at the time of the accidents. These variables offer insights into the temperature, wind chill, humidity, air pressure, visibility, wind direction, wind speed, precipitation, and overall weather condition during each accident. Analyzing weather-related factors helps identify how weather conditions influence road safety and accident occurrences.

**Facilities:**

The columns such as Amenity, Bump, Crossing, Give\_Away, Junction, No\_Exit, Railway, Roundabout, Station, Stop, Traffic\_Calming, Traffic\_Signal, and Turning\_Loop indicate the presence or absence of various facilities or points of interest (POIs) in the vicinity of the

accidents. By understanding the proximity of facilities like amenities, crossings, junctions, railway tracks, and traffic signals, we can assess their impact on accident occurrences and identify areas where infrastructure improvements or safety measures may be necessary.

**Day/Night Classification:**

The columns like Sunrise\_Sunset, Civil\_Twilight, Nautical\_Twilight, and Astronomical\_Twilight provide information about the period of the day based on sunrise, sunset, and different twilight phases. This classification helps analyze the influence of lighting conditions on accidents and supports the identification of specific time periods associated with higher accident risks.

**Duration:**

Duration is a calculated column we created. Its value is the difference between Start\_Time and End\_Time columns. The Duration column tells us the reported amount of time that the accident had an impact on traffic. In other words, it is the amount of time the authorities took to restore natural traffic flow. Duration is one of the three important target variables in our analysis. The longer the duration of an accident, the more traffic was backed up, and thus the more, unmeasured, economic loss to the society.

# **Analysis**

**Cleaning and Data Manipulation:**

We performed inner joins on the two datasets via both zipcode and City. This was necessary to do in python as there are several cities with the same names(Springfield), so analysis on certain cities may have been skewed if we did not perform this fix. In addition to this, we removed all redundant and unnecessary variables to our analysis from both sets. For example, address specific information was excluded.

**Descriptive Analytics:**

Data visualization and descriptive analysis can provide an easy interpretation of the data. By utilizing Tableau, the provided longitudinal data of accidents can be interpreted by the software. We utilized Tableau instead of python for multiple reasons. For one, usability and reliability of tableau over a python API is incredibly important. Second, the visualization and beautification of the data is unparalleled in Tableau. By utilizing these variables: ‘Location’, ‘Latitude’ and ‘Longitude’, we have built a heat density map of accidents located within both the United States and Florida.

Next, We utilized Pyspark and Python to visualize the data through various lenses, such as at the State level and City level. We specifically grouped by the aforementioned variables and visualized the cities and states with the highest count of accidents and the highest per capita of data.

**Predictive Analytics:**

Precise predictions of incident severity and duration can significantly improve decision-making and resource allocation during traffic incidents. Our models focus on the effectiveness of numeric and binary variables in predicting the severity and duration of these incidents using RandomForest Regressor models.

Four RandomForest Regressor models were built and trained on two types of variables: numeric (such as Distance, Precipitation, and Pressure) and binary (including features like Crossing, Junction, Traffic Signal, and Station). Two models predict incident severity, and the other two predict incident duration. All models were trained and tested on distinct splits of the dataset, with a 70-30% train-test ratio.

**Prescriptive Analytics:**

Building off from the predictive analysis, we tested the causal influence of selected treatment variables on target variables. As explained previously, the main methodology was A/B testing with propensity score matching. For each of the three target variables, we matched the data into functionally equivalent groups for the following treatment variables as deemed important in the predictive analysis:

* Precipitation(in)
* Population
* Stop (accident happened near a stop sign)
* Crossing (accident happened near a traffic crossing or intersection)
* Traffic Signal
* Junction
* Station

During matching, we wanted to create treatment and control groups in a way that matched pairs between the two groups are as similar as possible, except for the treatment variable. Only by ensuring this condition we can be certain that any differences in the outcome variable are caused by the difference in treatment.

To achieve this, we first ensured the underlying data is in good shape. We removed rows with missing values, removed outliers (e.g. rows where duration of the traffic was over 7 days), and encoded variables appropriately for the coding component. For categorical variables, we used their natural values for either binary coding or one-hot encoding. For continuous variables, we defined appropriate thresholds for categorizing values into groups by leveraging descriptive statistics of the corresponding treatment variable and online research of relevant domain knowledge. For example, we picked 0.1 inch of precipitation as the threshold to consider whether an accident happened in the rain or not. According to U.S. national standards, less than 0.1 inch of rainfall is considered as light rain and not enough water to cause wet road conditions, which is when driving in the rain becomes truly dangerous due to lower traction and hydroplaning.

Additionally, we carefully picked relevant covariates to benchmark the matching procedure. We selected covariates that are potentially impacting target variables, but are not directly related to the treatment variables. For example, if the treatment variable is Precipitation(in), we included as many relevant variables as possible (such as road condition features) but excluded related weather condition variables (such as Humidity, Temperature, and Visibility) as they are likely to be correlated with each other.

We also checked for data imbalance, which turned out to be a serious issue for our analysis. For several treatment variables such as Precipitation(in), Stop, Crossing, and so on, positive responses (value encoded as 1) are disproportionately smaller than negative responses (value encoded as 0). In many cases, data imbalance was as severe as 1.5% of positive response and 98.5% of negative response. In these situations, we want to achieve acceptable data imbalance (about 30:70 split between the proportion of positive and negative responses). The method we applied to alleviate the issue is a combination of undersampling the majority group (usually the negative responses) and oversampling the minority group (usually the positive responses). For undersampling, we used simple random sampling to filter for a subset of total majority responses. For oversampling, used only in severe cases, we applied Synthetic Minority Oversampling Technique (SMOTE) to generate artificial data with similar characteristics of organic data in the minority group.

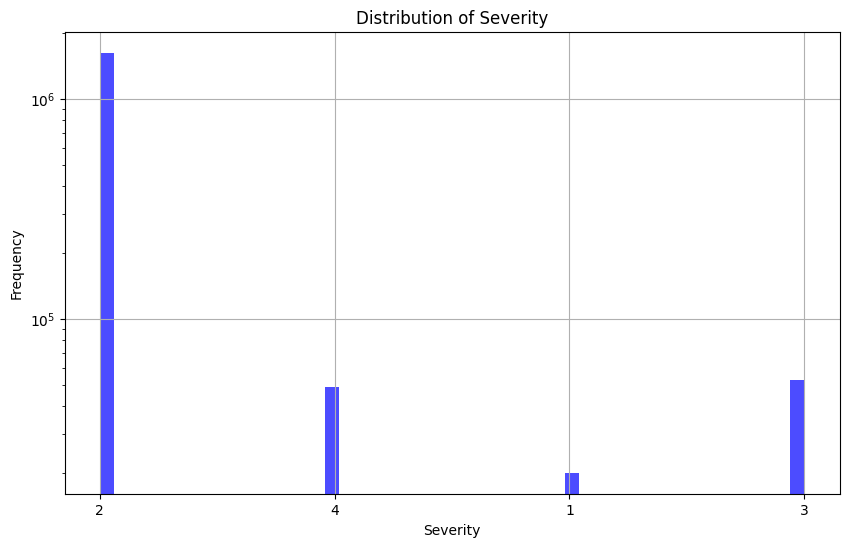
Below we have provided the summary statistics of relevant variables.

|  | **Target Variables** | | | **Continuous Treatment Variables** | |
| --- | --- | --- | --- | --- | --- |
|  | **Severity (1 to 4)\*** | **Duration (hr)** | **Distance (mi)** | **Population** | **Precipitation (in)** |
| Mean | 2.08 | 2.59 | 0.7 | 1,176,674 | 0.005 |
| Median | 2 | 1.7 | 0.21 | 105,087 | 0 |
|  | **Categorical Treatment Variables (Count)** | | | | |
|  | **Crossing** | **Junction** | **Stop\_Sign** | **Traffic\_Signal** | **Station** |
| **Yes** | 128,155 | 149,885 | 37,234 | 165,936 | 38,138 |
| **No** | 1,617,381 | 1,595,651 | 1,708,302 | 1,579,600 | 1,707,398 |

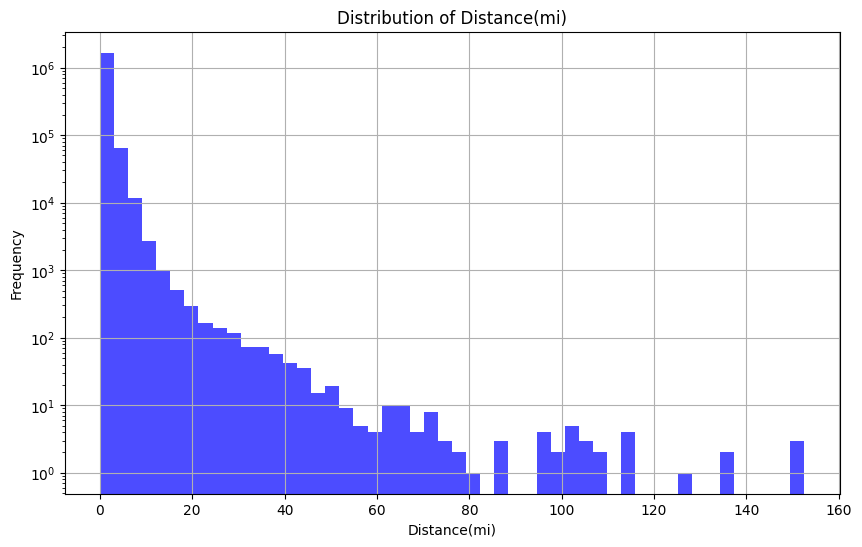
\* Severity levels: 1 = least severe and 4 = most severe

***Table 1****: Summary statistics of target and treatment variables*

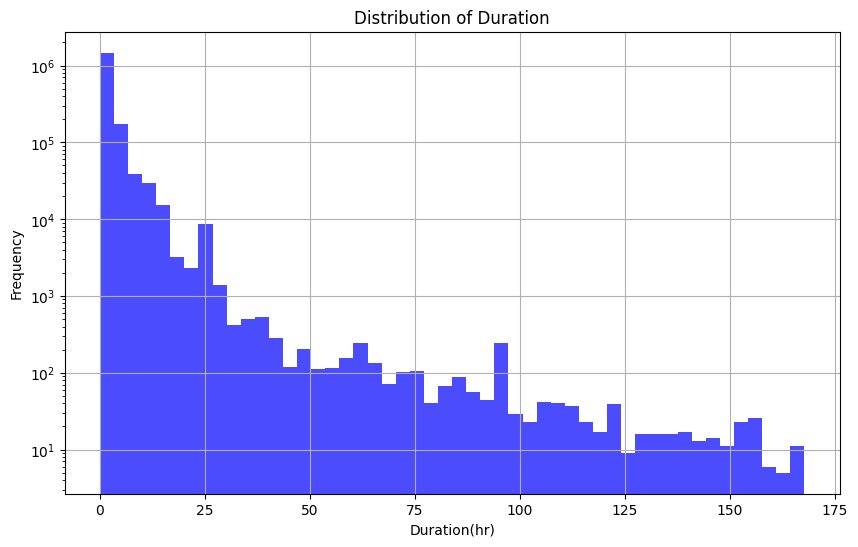
The charts below show the distribution of three main target variables in the predictive analysis. Severity is presumably a self-reported variable. It is indicated by four levels. Most of the accidents are rated as level 2. We consider severity as a loss of life indicator. It mainly represents the bodily harm the accidents have done to the people involved. Distance is the length of traffic built up due to the accident. It is measured in miles. Duration is the amount of time the accident lasted. It is a calculated variable from Start\_Time and End\_Time, measured in hours. Distance and Duration both capture the magnitude of the accident, but in different perspectives. We consider them the indicators of economic loss: the longer the accident in terms of distance and time, the heavier its toll would be on families, businesses, and local communities. Together, these three variables are the main targets of interest we want to gather our comparison upon in the predictive analysis.



***Figure 1****: Distribution of severity level of the accident*



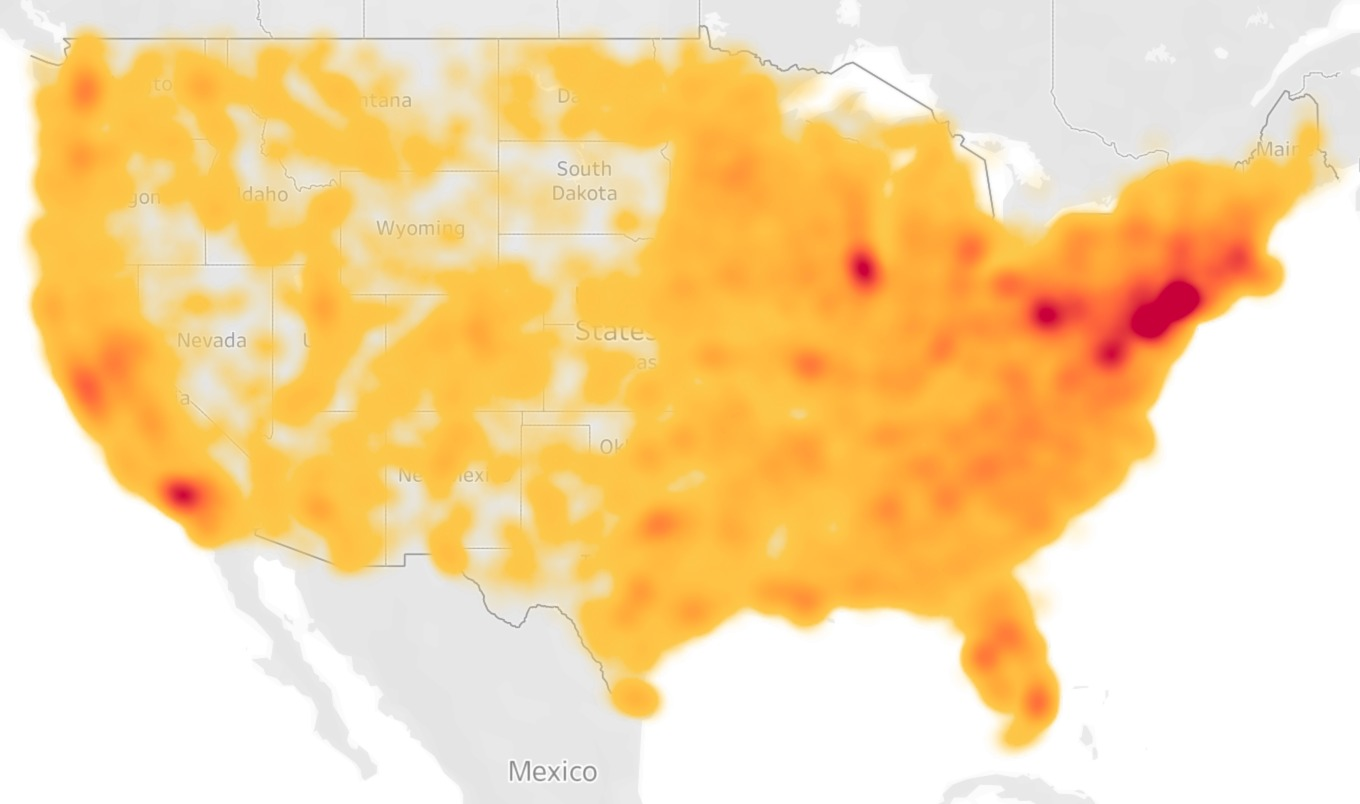
***Figure 2****: Distribution of distance (in miles) affected by the accident*



***Figure 3****: Distribution of duration (in hours) affected by the accident*

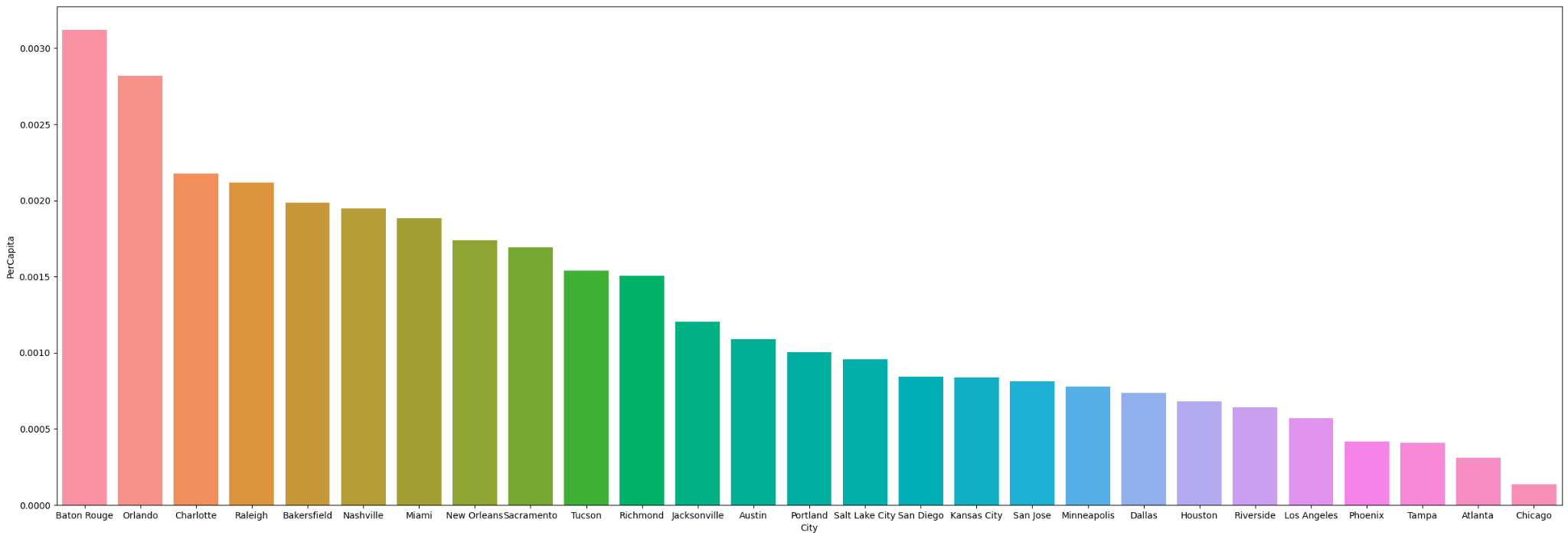
# **Results**

**Descriptive Analytics:**



***Figure 4****: Density Heat Map of Traffic Accidents in U.S. 2016-2023*

We found that some of our results did not match our preconceived notions of traffic accidents. The ‘heat map’ visualizations performed on Tableau showcased high heatmap areas in New York, Florida, and California. While California did have the highest amount of accidents by state, the city level and per capita view revealed a different story. Southern cities and States had the highest amount of traffic accidents through a counting metric and a per capita metric. Cities like Los Angeles fell out of the top 30 given the per capita view. Cities like New York and San Francisco were not even in the top 30 on a counting metric. This reveals to us that the highest indicator of accidents is not simply ‘more people and more cars.’ There are actional and inactionable factors that cause accidents more than just population.

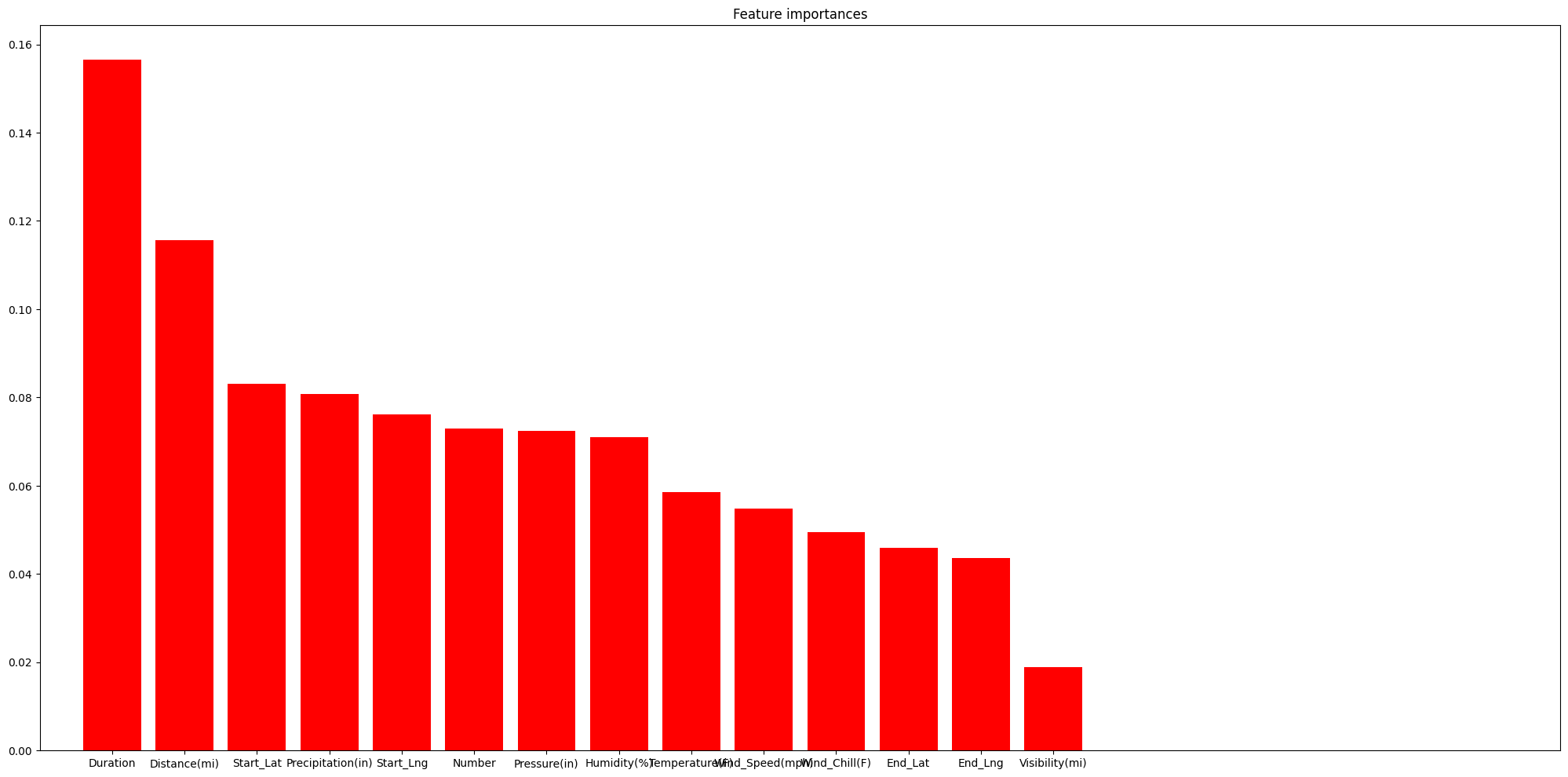
****

***Figure 5****: Traffic Accidents Per City Per Capita in U.S. 2016-2023*

**Predictive Analytics:**

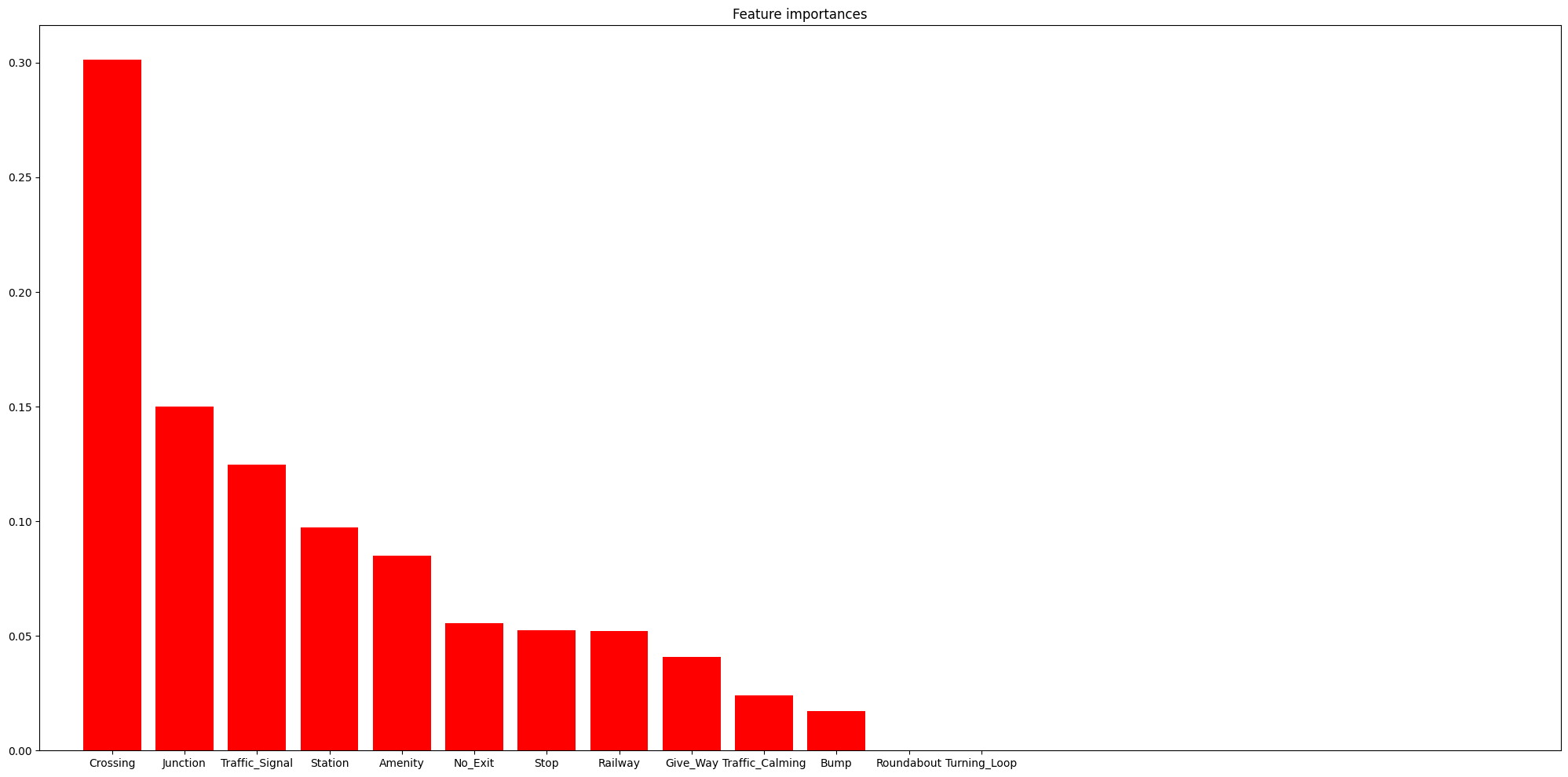
Our results revealed different feature importance rankings across the four models.

For the model predicting incident severity using numeric variables, the most critical features were 'Duration', 'Distance', 'Precipitation', and 'Pressure'.



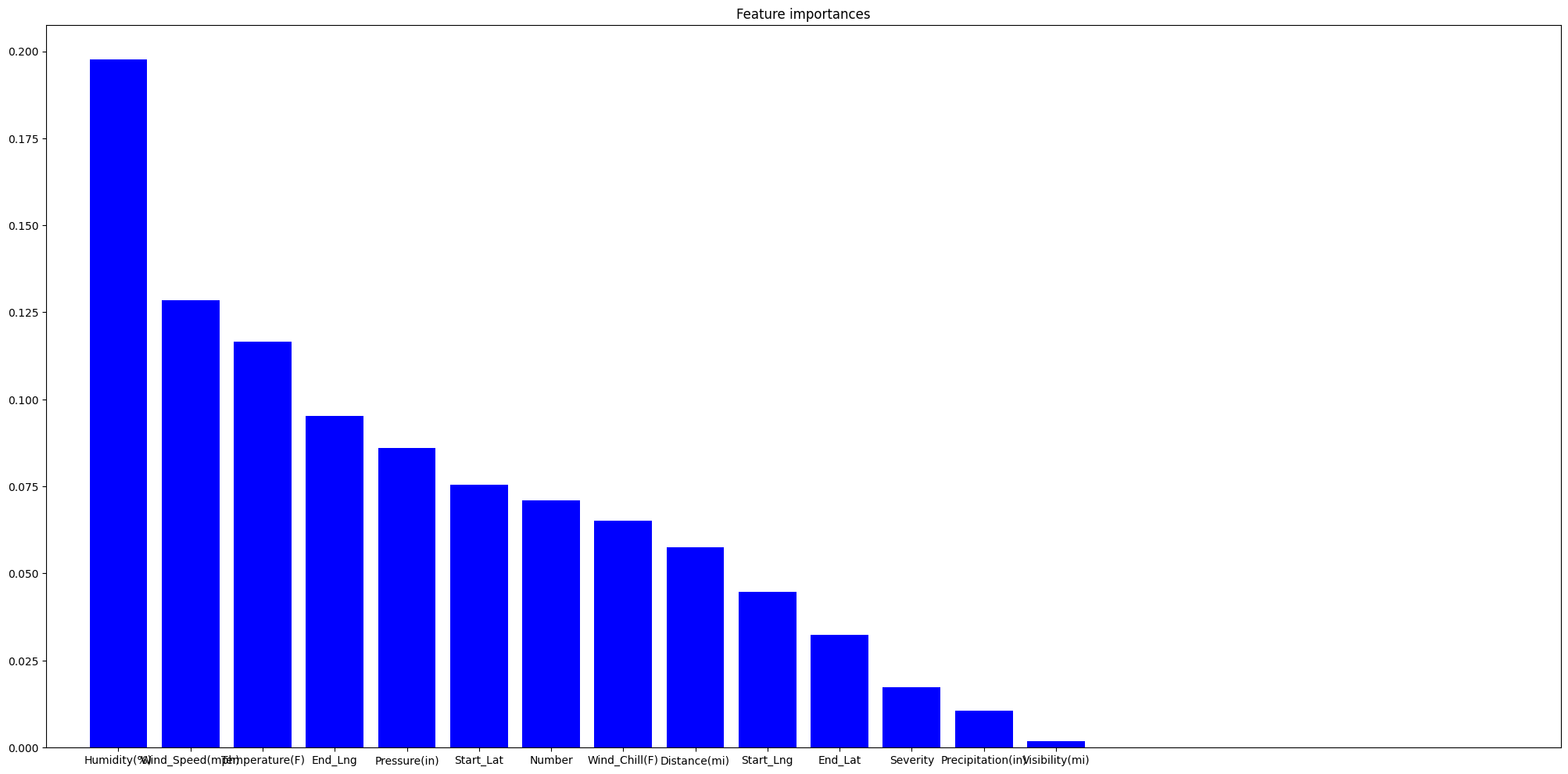
***Figure 6****: RandomForest Model Severity Feature Importance*

In contrast, the model predicting incident severity with binary variables found 'Crossing', 'Junction', 'Traffic Signal', and 'Station' to be the most influential features.



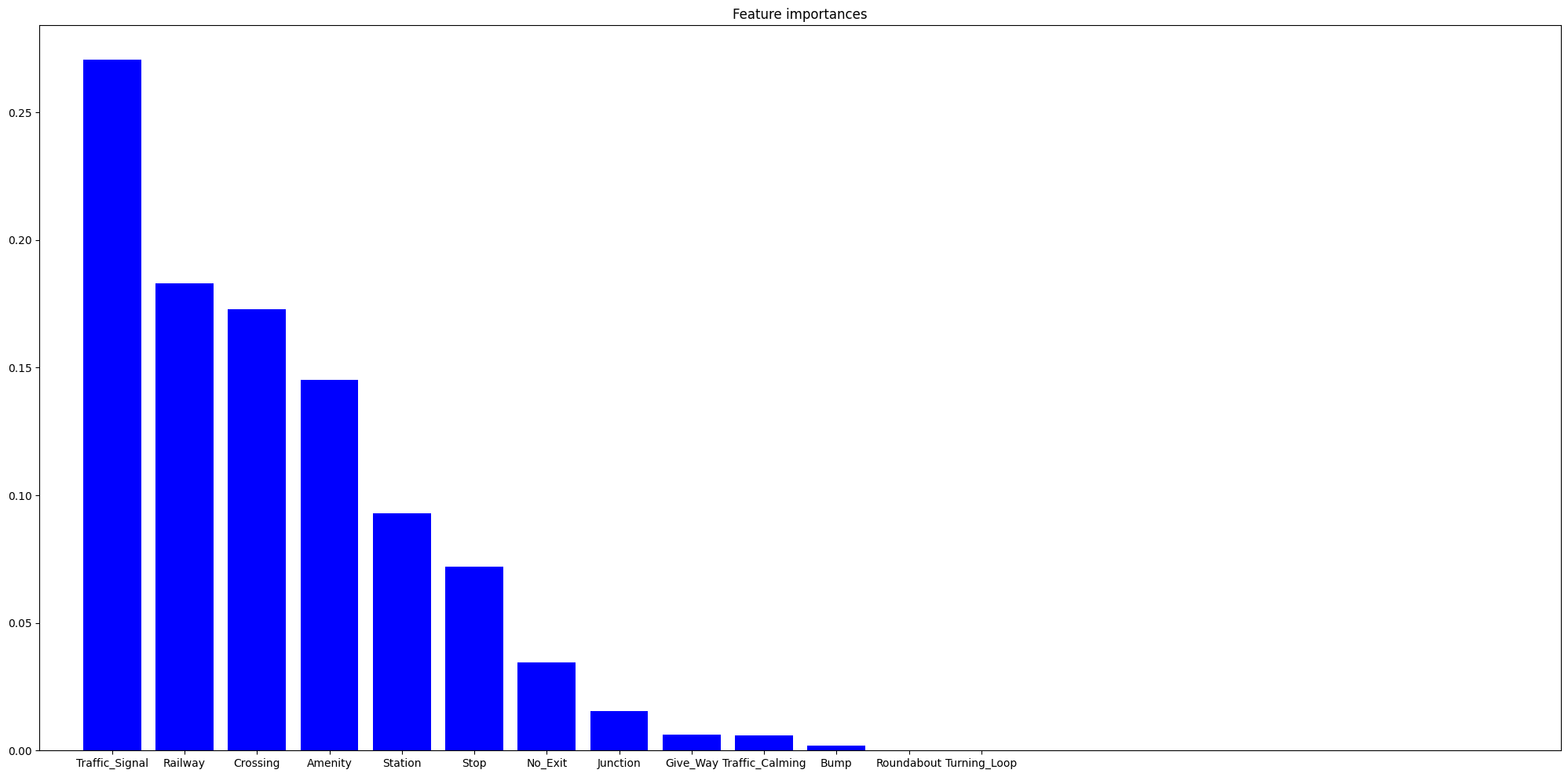
***Figure 7****: RandomForest Model Severity Feature Importance Cont.*

When predicting incident duration, the model utilizing numeric variables highlighted 'Distance', 'Duration', 'Pressure', and 'Precipitation' as the most important features.



***Figure 8****: RandomForest Model Duration Feature Importance*

Lastly, the model predicting incident duration using binary variables determined 'Traffic Signal', 'Railway', 'Crossing', and 'Amenity' as the most significant predictors.



***Figure 9****: RandomForest Model Duration Feature Importance Cont.*

The findings indicate that different types of variables could play critical roles in reducing and predicting traffic. These models are an essential step towards understanding the comparative roles of variables in predicting traffic incident parameters, providing a foundation for A/B testing and matching. One thing to note is that the nature of these models may mean that running the same exact model may result in some different results. However, after running these models several times we found that the most important variables remained the same.

**Prescriptive Analytics:**

Our prescriptive analysis showed some interesting results, which are summarized below. For each treatment variable, we explained the interpretation of the average treatment effect as well.

|  | **Population** | **Precipitation** | **Stop Sign** | **Traffic Signal** | **Junction** | **Crossing** | **Station** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Severity** | 0.0146 | 0.035 | -0.003 | -0.03 | -0.09 | -0.09 | -0.03 |
| **Duration** | 0.41 | -0.49 | 0.51 | -0.62 | -0.58 | 0.33 | 0.76 |
| **Distance** | -0.24 | 0.2 | -0.44 | -0.56 | 0.33 | -0.65 | -0.62 |

***Table 2****: Average treatment effect (ATE) of treatment variables on target variables*

Population

*Severity*: The ATE is 0.0146. This means that on average, accidents in larger cities (population = 1) are slightly more severe than those in smaller cities, with an average increase in severity rating of about 0.015. Considering that the severity scale ranges from 1 to 4, this effect is relatively small.

*Duration*: The ATE of 0.41 indicates that on average, accidents in larger cities last 0.41 hours longer than those in smaller cities. Compared to the mean duration of 2.59 hours, this effect is somewhat significant, and could lead to increased traffic delays and economic loss in larger cities.

*Distance*: The ATE of -0.24 implies that the traffic impact distance of accidents in larger cities is 0.24 miles less on average than in smaller cities. This may suggest that despite having longer-lasting accidents, larger cities might have better traffic management strategies to limit the spatial extent of traffic impact.

Precipitation

*Severity*: An ATE of 0.035 means that on average, accidents that occur during rain are more severe than those in dry conditions by about 0.035 on the severity scale. This effect is relatively small, but still noteworthy, as it indicates that rainy conditions do increase the severity of accidents.

*Duration*: With an ATE of -0.49, this suggests that accidents during rain are resolved about half an hour quicker than those in dry conditions. This might be due to more active management or quicker responses in such conditions, given the weather-related risks.

*Distance*: The ATE of 0.2 indicates that on average, accidents during rain impact about 0.2 miles more of traffic than those in dry conditions. This could be due to slippery road conditions leading to more widespread disruptions.

Stop Sign

*Severity*: The negative ATE of -0.003 indicates that accidents at stop signs are marginally less severe than those not at stop signs. However, the effect is very small and likely not practically significant.

*Duration*: The ATE of 0.51 means that accidents at stop signs last about half an hour longer on average than those not at stop signs. This could be because of complications related to determining right of way in these cases.

*Distance*: The negative ATE of -0.44 implies that accidents at stop signs impact about 0.44 miles less traffic than those not at stop signs. This could be due to the typically lower speeds in areas with stop signs, leading to less severe traffic disruptions.

Traffic Signal

*Severity*: The negative ATE of -0.03 suggests that on average, accidents at traffic signals are less severe than those occurring elsewhere. Given the scale of severity (1-4), this could imply that traffic controls at intersections likely help mitigate the severity of accidents.

*Duration*: The ATE of -0.62 implies that, on average, accidents at traffic signals are resolved approximately 37 minutes (0.62 hours) faster than those not at traffic signals. This could be due to more effective emergency responses at signal-controlled intersections.

*Distance*: The negative ATE of -0.56 suggests that accidents at traffic signals affect about 0.56 miles less traffic than those not at traffic signals. This is understandable as traffic signals generally regulate traffic flow, preventing a wider area from being affected by an accident.

Junction

*Severity*: The negative ATE of -0.09 implies that on average, accidents at junctions are less severe than those not at junctions. This suggests that junction designs might help reduce the severity of accidents, probably due to reduced speeds or better traffic control.

*Duration*: The negative ATE of -0.58 indicates that, on average, accidents at junctions are resolved about 35 minutes faster than those not at junctions. Efficient traffic management at junctions may play a role in faster accident resolution.

*Distance*: An ATE of 0.33 means that accidents at junctions impact about 0.33 miles more of traffic than those not at junctions. This could be due to the converging nature of traffic at junctions, which might extend the traffic impact area when an accident occurs.

Crossing

*Severity*: The negative ATE of -0.09 suggests that accidents at crossings are on average less severe than those not at crossings. This might be attributed to lower vehicle speeds commonly observed at crossings.

*Duration*: The ATE of 0.33 suggests that accidents at crossings last, on average, approximately 20 minutes longer than those not at crossings. This could be due to the need for additional care to accommodate pedestrian traffic during accident resolution.

*Distance*: The negative ATE of -0.65 implies that accidents at crossings impact about 0.65 miles less traffic than those not at crossings. This could be a result of localized traffic disruption typically associated with crossings, preventing a more widespread impact.

Station

*Severity*: The negative ATE of -0.03 suggests that accidents near emergency stations are slightly less severe than those further away. This might be due to quicker response times from emergency services when they are in close proximity.

*Duration*: The ATE of 0.76 indicates that accidents near stations last about 46 minutes longer than those not near stations. This might seem counterintuitive given the proximity of assistance, but it could be due to more thorough on-site investigations or interventions, given the immediate availability of resources.

*Distance*: The negative ATE of -0.62 suggests that accidents near stations affect about 0.62 miles less traffic than those not near stations. This might be due to efficient containment and traffic rerouting facilitated by emergency services present nearby.

# **Recommendations**

**Interpretations**

Based on the results, we want to further discuss a few observations that caught our attention. First of all, it's intriguing that accidents in larger cities tend to have a longer duration but lesser impact distance compared to those in smaller cities. Given that larger cities usually have more resources and potentially more effective traffic management systems, it's unexpected to see longer accident durations. One possibility could be the complexity of traffic accidents in populated areas. This finding may call for an exploration of how urban planning and traffic management strategies differ between small and large cities, and how they might be improved.

Regarding rainy conditions, while it's not surprising that accidents during rain are more severe, it's interesting that they are resolved quicker and have a larger impact distance than those in dry conditions. This may be due to more effective and quick emergency responses during rain, possibly due to increased preparedness for accidents in these conditions. However, the larger impact distance could suggest the need for improved road maintenance and drainage during rainy conditions to limit the spread of disruption.

Furthermore, it's somewhat surprising that accidents at traffic signals are resolved quicker and affect less traffic. Given the frequency of accidents at intersections, this finding suggests that traffic signal control plays a significant role in reducing the impact of accidents. However, it might be beneficial to investigate why these locations are still prone to accidents and explore strategies to reduce the frequency of incidents.

Lastly, the finding that accidents near stations last longer is somewhat counterintuitive since we would expect proximity to emergency services to reduce duration. It could be due to a number of reasons, such as more thorough accident investigations due to immediate availability of resources, or slower traffic flow because of the presence of the station itself. This warrants further exploration to understand the reasons behind this increased duration and to find solutions to expedite accident resolution.

Overall, while some of our variables did not have the ATE that we expected, there are multiple actionable insights to take from our analysis. For example, cars on the road are not the driving force in accident prediction, and being close to an emergency response department actually increases the duration of accidents. Based on our findings, we propose the following recommendations to our clients.

Our analysis indicates that local agencies could concentrate their efforts on enhancing traffic management in larger cities. This could have the potential to minimize the duration of traffic accidents. Strategies for achieving this might involve better planning for accident response, or perhaps increasing resources available for resolving accidents, particularly during peak traffic periods. It's also worth considering road maintenance during rainy conditions. A focus here could potentially lessen the disruption caused by accidents, perhaps through improved drainage systems or swift deployment of road cleaning crews following accidents during rainfall. Furthermore, our data suggests an opportunity for these agencies to scrutinize their current traffic management strategies at junctions, traffic signals, and crossings. By reassessing aspects such as traffic light timings, stop sign placements, and pedestrian crossing designs, there may be potential to further decrease accident severity and duration.

The results also suggested that making modifications to the design of traffic signals, junctions, and crossings could further minimize accident severity and duration. This might involve enhancing the visibility of these areas or making them easier to navigate. In addition, the designs of traffic flow near emergency response stations merit investigation. Accidents near these areas have been found to persist longer on average. Therefore, implementing measures such as dedicated lanes for emergency vehicles might prevent slowdowns in these areas and help reduce accident duration. Also, given the increased severity of accidents during rain, the implementation of safety measures could be beneficial. Using road materials that lessen skidding in wet conditions or increasing the visibility of road markings and signs may help reduce accident severity during these times.

Lastly, the location of an accident could be a crucial factor in its severity, duration, and impact distance. Therefore, insurance companies may want to consider this when pricing insurance products. For instance, our data shows that accidents in larger cities tend to be more severe and last longer but impact a smaller distance. Likewise, accidents during rain have been found to be more severe and impact a larger distance, potentially indicating an increased risk for drivers frequently driving in such conditions. Insurance premiums could reflect this heightened risk. Lastly, the presence of traffic signals, junctions, crossings, and stations can significantly impact the characteristics of an accident. This information might aid insurance companies in developing more accurate risk profiles for policyholders.

# **Limitations**

One shortcoming of this paper, is the lack of data in regards to road infrastructure. Literal conditions of the road, not just weather conditions, may play an important role in predicting accident severity. There is some research done on road conditions available by Congressional Research Service which showcases worse infrastructure in southern states. This may be a factor or variable driving higher accidents per capita in southern states unexplained by our regression tree analysis. There is no available research to gain insights on infrastructure in relation to accident prediction, we recommend further analysis must be done in order to gather the true treatment effect of infrastructure.

Additionally, the severity of traffic accidents in our study may not serve as the most reliable target variable. Since this measure is self-reported, we lack a standardized benchmark for comparison across the dataset. Variances in interpretation of what constitutes a "severe" accident may differ significantly between respondents, and there is a lack of clarity on whether a specific metric is used consistently when rating severity. Therefore, this subjective nature of severity reporting may potentially lead to inconsistencies in our data, thereby affecting the reliability of our analysis. Finally, a major limitation of our study, inherent to many others, is its reliance on observational data. This approach inherently restricts our ability to offer insights on the preventative aspect of future accidents. We can only infer trends and causal effects based on events that have already occurred, not on those that might happen under different circumstances.