Predictive Modeling for Risk-Based Premiums and Real-Time Safety Guidance in Auto Insurance Customer Retention

Marvin Moran
Applied Data Science Master's Program
Shiley Marcos School of Engineering / University of San Diego
marvinmoran@sandiego.edu

Ben Ogle
Applied Data Science Master's Program
Shiley Marcos School of Engineering / University of San Diego
bogle@sandiego.edu

Katie Mears
Applied Data Science Master's Program
Shiley Marcos School of Engineering / University of San Diego
katiemears@sandiego.edu

ABSTRACT

Auto insurance companies increasingly seek innovative strategies to improve customer retention and revenue. Traditional premium pricing models, often reliant on broad demographic data, fail to account for individual driving behaviors and environmental conditions, leading to misaligned premiums and customer dissatisfaction. To address this, BMK Insurance developed a predictive modeling framework to assess individual accident risk and create personalized, dynamic insurance premiums.

This study explored the use of advanced machine learning techniques to predict accident severity based on a comprehensive dataset comprising weather, spatial, and traffic-related features. The primary models evaluated were the Multi-Layer Perceptron (MLP) and CatBoost, which demonstrated strong predictive performance across training, validation, and test datasets. The MLP model, achieving a low Root Mean Squared Error (RMSE) of 0.0436 and an R-squared value of 0.8830, was ultimately selected for deployment due to its ability to capture complex non-linear relationships and scalability. Although CatBoost showed marginally better performance regarding RMSE and R-squared, the MLP′s adaptability and robustness in handling complex interactions made it the optimal choice for real-world applications. Feature importance analysis confirmed the significant role of weather and traffic patterns in predicting accident severity. The findings suggest that integrating machine learning models into risk assessment frameworks can help automobile insurers develop proactive strategies for accident prevention and personalized premium pricing. Future studies could expand this analysis by considering temporal factors, improving dimensionality reduction techniques, and exploring additional machine learning models to further optimize predictive accuracy.

KEYWORDS

traffic accidents, traffic congestion, predictive modeling, data-driven risk assessment, dynamic insurance pricing, accident risk score, weather conditions

1 Introduction

Auto insurance companies are constantly exploring ways to boost customer retention and revenue. Many customers feel distant from their insurers, often viewing insurance as a costly and bothersome expense. Recognizing these frustrations, BMK Insurance has decided to take proactive steps. BMK Insurance believes that safe drivers deserve discounts based on a personalized risk score, exposure to environmental hazards, and the use of alternative, less traffic-dense routes to help keep their insurance rates low.

To further enhance its customer engagement and retention strategies, BMK Insurance is leveraging predictive modeling to assess and forecast individual customer risk levels regarding accident propensity (Welch, 2024). By analyzing historical driving behavior, environmental factors, and route choices, the company can develop sophisticated algorithms that predict how accident-prone a customer may be. These models consider various parameters, such as the types of weather during a commute, traffic density, and the frequency of accidents occurring on their route. By accurately predicting risk, BMK Insurance can tailor its offerings, providing discounts to safe drivers while encouraging risk-reducing behaviors among its customer base. This data-driven approach fosters a more personalized relationship with customers and aligns their pricing models with actual driving behavior, ultimately leading to better customer satisfaction and loyalty.

2 Background

In the evolving landscape of auto insurance, leveraging predictive modeling insights to calculate risk with greater precision presents a transformative opportunity for BMK Insurance. The company can better align premiums with individual risk levels by implementing dynamic, risk-based premium pricing that accounts for each customer's unique driving environment. For instance, policyholders navigating high-risk conditions, such as areas with frequent congestion or adverse weather, may face adjusted premiums that reflect these challenges. Conversely, drivers in safer environments stand to benefit from premium discounts. This personalized approach fosters a sense of fairness in pricing and enhances customer engagement through value-added services, including accident risk alerts and suggestions for safer routes.

These tailored offerings differentiate BMK Insurance in a competitive market and attract safety-conscious customers who value proactive, data-driven insights. By providing real-time guidance to mitigate accident risk, BMK Insurance builds brand trust and encourages customer loyalty, significantly increasing policy renewals (Allen, 2023). This strategy expands the customer base and supports long-term profitability, positioning BMK Insurance as a leader in the industry committed to customer safety and satisfaction.

2.1 Problem Identification and Motivation

The traditional auto insurance model typically employs a one-size-fits-all approach to premium pricing, often relying on broad demographic factors and historical claims data. This method can result in mispricing for many policyholders, particularly placing an undue financial burden on safer drivers whose premiums do not reflect their lower-risk exposure. In contrast, these drivers often face equally high policy premiums as those in riskier conditions, leading to a lack of fairness in the pricing structure.

As various driving environments and individual driving patterns can significantly impact accident probabilities, the limitations of conventional pricing strategies become increasingly evident. This misalignment frustrates conscientious drivers and undermines the incentive for safer driving, as financial responsibility does not correlate with their actual risk level.

The motivation behind this project stemmed from the need to bridge this gap by implementing predictive modeling techniques that harness real-time data to inform premium pricing. By accurately identifying and quantifying risk factors such as traffic patterns, weather conditions, and individual driving behaviors, BMK Insurance can develop a more equitable pricing strategy that reflects each customer's unique circumstances. This approach addresses the shortcomings of traditional methods and enhances customer satisfaction and loyalty through personalized pricing.

Ultimately, implementing risk-based premiums will enable BMK Insurance to remain competitive in a rapidly changing market while fostering a culture of safety and accountability among its policyholders. By prioritizing data-driven insights, the company positions itself as a forward-thinking leader committed to delivering value while mitigating risks.

2.2 Definition of Objectives

The objectives of this study were to quantify the impact of weather conditions on traffic congestion and accident risk, enabling a deeper understanding of how variables such as precipitation, temperature, and visibility affect the number of auto accidents. Using predictive modeling, the study aimed to create a dynamic, risk-based pricing model that adjusts insurance premiums based on each customer's unique driving conditions, supporting a more personalized approach to insurance premium pricing. Additional objectives include developing real-time predictive alerts and alternative route recommendations to improve customer safety and satisfaction.

2.2.1 Data Collection and Integration

The project required the development of a comprehensive dataset that includes integrated data based on weather conditions, traffic patterns, accident history, and driver behaviors. It also required real-time data streaming to capture real-time data regarding weather fluctuations and traffic conditions, ensuring that the analysis reflects the current environment faced by drivers.

2.2.2 Risk Assessment and Modeling

To develop the predictive modeling framework, statistical and machine learning models were developed to assess the impact of various weather conditions on accident likelihood and traffic congestion. For example, the model can quantify how factors such as rain, snow, temperature extremes, and low visibility correlate with accident rates.

Using this predictive model, personalized risk scores were calculated which allowed BMK Insurance to develop a dynamic pricing model capable of adjusting insurance premiums based on real-time risk assessments. This enables BMK Insurance to offer tailored premium pricing that reflects the risk levels associated with individual policyholders' driving environments.

2.2.3 Real-Time Safety Guidance

To accomplish the real-time safety guidance objective, BMK Insurance developed a system that provides real-time alerts to drivers regarding adverse weather conditions and high-risk traffic situations, enabling proactive safety measures. The system also includes a feature that suggests safer or less congested routes to policyholders based on current weather and traffic data.

3 Literature Review (Related Works)

Given the multidimensional focus on weather, traffic congestion, predictive modeling, and personalized insurance premiums, a thematic review relative to existing literature was applied to this study. In doing so, the goal was to position the study in relation to existing research across similar themes. This approach highlights both contributions and gaps across these specific aspects, emphasizing the importance of this particular study. Overall, although several research papers exist for the highlighted themes, they generally only focus on traffic congestion or weather impacts. However, this study aims to combine both in an actionable way for individual drivers and insurance providers. This comprehensive approach provides a holistic view of how traffic and weather jointly impact accident risks.

3.1 **Weather Impacts on Various Types of Road Crashes: A Quantitative Analysis Using Generalized Additive Models**

In 2022, European researchers launched a study investigating the impact of adverse weather conditions, such as rain and snow, on fatal traffic accidents. Using Generalized Additive Models (GAMs), the study identified a non-linear relationship between weather variables (e.g., precipitation) and accident probabilities. The researchers assert that their goal is singular– aiming to only "...quantify the impact of adverse weather on accident risk and to determine which weather conditions (rain, snow, fog) have the most significant influence on accidents" (Becker et al., 2022, p.12). Thus, the influence of climate trends and traffic safety is the emerging theme for the study– confirming facts that are already intuitively suspected. Namely, extreme weather conditions (e.g., heavy rainfall) correlate with increased accident fatalities, especially over extended periods. This suggests that climate factors significantly impact road safety. Although the study includes climate data in long-term safety modeling, confirming findings from other studies that weather conditions are a significant factor in accident risk, it places a greater focus on broader climate patterns rather than specific weather events. Additionally, the scope is limited to fatal accidents only, thereby not fully capturing the impact of weather on non-fatal accidents, which the current study aimed to address.

3.2 **Examining the Effect of Adverse Weather on Road Transportation Using Weather and Traffic Sensors**

The theme of adverse weather conditions in traffic safety is also present in this study. Chinese researchers leverage a more modern approach to analyzing data, focusing on capturing real-time visibility and weather metrics (e.g., fog, rain) via weather and traffic sensors. With this data, the researchers highlight that adverse weather conditions like fog and rain directly influence driver speed, lane-keeping, and spacing between vehicles. Further, "visibility-related issues (fog, rain) cause predictable patterns in driver behavior such as reduced speeds and increased lane deviation, increasing the likelihood of accidents" (Peng et al., 2018, p.15). Although the study reinforces the idea that adverse visibility is a significant factor in traffic safety, its real-time approach only focuses on immediate weather impacts, providing a more instantaneous view of accident risk. It contrasts with studies that examine historical or seasonal data, instead capturing dynamic changes that occur minute-to-minute. Historical trends further enhance the target machine learning models, resulting in greater prediction accuracy for long-term insights, which the current study aimed to accomplish.

3.3 **Understanding the Effect of Traffic Congestion on Accidents Using Big Data**

Although the title of this study only mentions traffic congestion, it also targets adverse weather conditions. Thus, the theme of weather and traffic emerges once more. Explicitly, the study investigates the combined impact of traffic congestion and weather conditions on accident frequency in urban settings with high-density traffic areas across select Latin American cities. Sanchez-Gonzalez et al. (2021) aimed to provide a comprehensive view of how congestion and adverse weather interact to influence accident rates. The study highlights that "...in congested areas, accident risk rises significantly under adverse weather conditions, as congestion compounds the impact of poor weather" (Sanchez-Gonzalez et al., 2021, p.21). The study shows that heavy rain and snow are more likely to result in accidents in high-traffic areas due to reduced mobility and braking capability– reinforcing that congestion is a crucial factor in accident risk. In contrast to the majority of the existing literature, this study differs by including both congestion and weather variables, highlighting the compounded effects of high traffic and adverse weather. Most studies focused solely on climate variables or congestion metrics, but never both. In particular, Sanchez-Gonzalez et al. aimed to leverage their findings to support policymakers in emerging economies by implementing measures to reduce congestion, which largely deviates from the purpose of the current study.

3.4 **Exploring the Impact of Climate and Extreme Weather on Fatal Traffic Accidents**

In this study, Chinese development researchers examine the effects of long-term climate variables (e.g., precipitation) on fatal accidents. Here, the familiar themes of weather and traffic are quickly noted. Zuo et al. (2021) leveraged negative binomial models to examine, analyze, and process over-dispersed historical data on fatal accidents given climate trends. Their goal was to "...understand the macro-level impact of climate variables on accident fatalities…" aimed at providing insight into climate-aware policy and long-term safety planning (Zuo et al., 2021, p.18). Notable patterns in their findings show that extreme weather conditions (e.g., heavy rainfall) correlate with an increase in accident fatalities– especially over extended periods of time. Thus, they conclude that climate factors significantly impact road safety. This conclusion aligns well with other studies on the topic, all validating that adverse weather significantly impacts fatal accidents. Like other studies, this study also focused on broader climate patterns rather than specific weather events, contrasting with studies that look at localized, short-term impacts. Its focus, too, is limited to fatal accidents, which may not fully capture the impact of weather on non-fatal incidents.

3.5 **Machine Learning for Predictions of Road Traffic Accidents and Spatial Network Analysis for Safe Routing on Accident and Congestion-Prone Road Networks**

In this European study, the primary goal is to generate alternative route recommendations given adverse weather conditions. Noting the familiar theme of weather and traffic data analysis once more. Across their research, Berhanu et al. (2024) used historical traffic and weather data to create predictive models for real-time alerts and alternative route recommendations for individual users. They further explore how adverse weather events like heavy rain or snow disrupt traffic flow. The created models support that predictive models based on historical data can effectively inform real-time traffic recommendations, particularly under adverse weather (Berhanu et al., 2024). This conclusion aligns well with existing literature validating the feasibility of predictive modeling in managing traffic flow and congestion. However, it became evident that the purpose of this research contrasted with that of the current study, as it lacked a focus on accident risks and instead concentrated solely on offering alternative routes.

3.6 **Road Car Accident Prediction Using Machine-Learning-Enabled Data Analysis**

In this study, European and Chinese researchers partnered together to analyze and quantify the influence of weather on traffic accidents. Explicitly, they investigated the impact of adverse weather conditions on traffic safety– focusing on how weather patterns such as rain, snow, and temperature changes affect accident risk. According to the authors, they aimed to "...quantify the influence of different weather conditions on traffic accidents [in order to provide] a foundation for proactive traffic safety measures that could inform municipal strategies and improve road safety in adverse weather" (Pourroostaei-Ardakani et al., 2023, p.24). Patterns across the study and its conclusions make it clear that it is deemed feasible to anticipate (i.e., predict) weather-related traffic disruptions. Furthermore, it is then possible to mitigate such occurrences. The study reinforced the significance of incorporating weather data into traffic safety models, supporting the broader view that other literature shares relative to weather conditions significantly impacting accident risk and that therefore should be factored into predictive modeling for safety. However, unlike studies that explore personalized models, this study remains focused on general patterns and aggregate data–providing insights for broad safety measures rather than individualized applications.

4 Methodology

The dataset created for this study combined historical driving data, environmental conditions, and traffic accidents that occurred in San Diego to provide insights into accident risks. A robust exploratory data analysis (EDA) process was conducted to uncover patterns, identify anomalies, and guide the modeling approach. This section outlines the key observations, data preparation steps, and insights gained during EDA.

4.1 **Data Acquisition and Aggregation**

The dataset used in this study was constructed by combining data from three main sources. The US Accidents (2016–2023) dataset, sourced from Kaggle.com, provided comprehensive records of traffic accidents across the United States, including key details such as accident location, severity, and time (US Accidents, 2023). Traffic patterns were captured using the SOC - Local Roads: Speed and Volume Traffic Data, available from opendata.sandag.org, which included metrics on traffic speed and volume on local roadways in San Diego (SANDAG, 2023). Weather data was obtained from OpenWeather Bulk Historical Data through the OpenWeather API, which offered historical environmental information such as temperature, precipitation, and wind speed (OpenWeather, 2023).

To merge these datasets, the Geopandas library was used to filter data relevant to the San Diego Metro area. Geopandas enabled the creation of a visualization to pinpoint the specific geographic region based on longitude and latitude. The U.S. Accidents and Traffic Data (SANDAG) were then synchronized using common latitude and longitude coordinates, integrating traffic data with accident records. Finally, weather data from OpenWeather API was joined using timestamp fields to match the weather conditions with each accident. This method ensured the final dataset accurately represented traffic and weather conditions for each San Diego metro area accident. The resulting dataset contained 91 columns and 96,078 rows of data.

The construction of the dataset for this study required careful consideration of ethical and privacy concerns to ensure responsible use of data. Each dataset used in construction of the final dataset was reviewed for compliance with usage terms. The US Accidents dataset had explicit permission granted for academic purposes and non-commercial research, which was adhered to for this study. The SANTAG Traffic Data was sourced from a publicly available platform under open-data licensing. The OpenWeather API, which sourced historical weather data, was obtained in licensing agreements of the API service, ensuring lawful use of the data.

None of the datasets included personally identifiable information (PII) nor contained specific traffic accident details that could be attributed to individuals. To further safeguard privacy, measures were taken to generalize geographic coordinates and round timestamps, reducing the risk of re-identification. Finally, insights derived from the analysis are responsibly communicated, with a focus on informing public safety and guiding policy decisions while avoiding potential misinterpretation or misuse of the study's findings. This approach underscores the commitment to ethical practices throughout the study.

4.2 **Exploratory Data Analysis**

EDA was performed to gain a deeper understanding of the dataset's characteristics and to uncover insights regarding the relationships between traffic accidents, weather conditions, and traffic patterns.

Before conducting the analysis, missing values in the dataset were addressed to ensure the integrity and reliability of the results. Columns found to have a high percentage of missing values (>69%) were dropped from the dataset. The dropped variables included sea_level, grnd_level, wind_gust, rain_1h, rain_3h, snow_1h, and snow_3h.

Continuous variables such as Temperature(F), Wind_Chill(F), Humidity(%), Pressure(in), Visibility(mi), and Wind_Speed(mph) were interpolated using linear interpolation to fill in missing values. The End_Lat and End_Lng columns were imputed with values from the Start_Lat and Start_Lng columns. Missing values in the Weather_Timestamp column were filled with values from the Start_Time column. Finally, categorical variables such as Wind_Direction, Weather_Condition, Street, Sunrise_Sunset, Civil_Twilight, Street_Name, Astronomical_Twilight, Nautical_Twilight, and PeakPeriod were filled with a default value of Unknown.

4.2.1 Univariate Analysis

Following the handling of missing values, univariate non-graphical analysis was conducted. This analysis included summary statistics for numerical variables, such as the mean, median, and standard deviation, to understand the distribution of continuous variables including Severity, Distance, Temperature(F), Wind_Chill(F), Humidity(%), Pressure(in), Visibility(mi), Wind_Speed(mph), Precipitation(in), Lanes, Speed Limit MPH, Length, timezone, visibility, dew_points, feels_like, temp_min, temp_max, pressure, humidity, wind_speed, wind_deg, clouds_all and weather_id. This analysis uncovered some key insights for the distribution of the numerical variables. The severity levels in the dataset range from 1 (least severe) to 4 (most severe) with a mean of 2.24, indicating that most of the accidents have a severity of 2. The standard deviation of 0.45 suggests a narrow range around the mean severity. Geospatially, the latitudes and longitudes for the start and end points cluster around 32.94°N and -117.16°W, reflecting a specific geographic region. This clustering aligns with expectations, as the analysis was conducted based on locations in San Diego. The average accident distance is 0.43 miles, but the range is wide, with a maximum distance of 22.57 miles, indicating that some accidents span longer distances. Weather conditions show that the average temperature is 64.38°F, with a standard deviation of 10.23°F. Wind chill averages 54.49°F, and wind speed is around 6.38 mph. Humidity is about 64.76% on average, with values ranging from 0% to 100%. Traffic conditions show a Speed Limit range from 12 mph to 84 mph, with an average of 38.02 mph, and most lanes have 1 to 7 lanes, with a mean of 1.18 lanes. Other variables include Pressure, which ranges from 28.13 to 30.54 inches, with a mean around 29.76 inches, and Precipitation, which is mostly 0, with a small mean of 0.01 inches. Visibility is generally high, with a mean of 9.02 miles, and the Year variable spans multiple years, capturing temporal trends. These statistics provide insights into the range, central tendency, and distribution of each feature in the dataset, which will inform decisions for data preprocessing and modeling.

For categorical variables, a count analysis was performed to determine the frequency of each category in variables City, Weather_Condition, Street, Wind_Direction, Sunrise_Sunset, Civil_Twilight, Nautical_Twilight, Street_Name, Highway, Direction, PeakPeriod, weather_main, weather_direction, Bump, Crossing, Give_Way, Junction, No_Exit, Railway, Roundabout, Station, Stop, Traffic_Calming, Traffic_Signal, and Turning_Loop. This analysis helped uncover patterns, such as the most common weather conditions and road features, providing valuable insights into the dataset's structure and trends.

The analysis of categorical variables revealed several key insights into accident patterns in San Diego County. The city of San Diego recorded the highest number of accidents, with 49,361 incidents, followed by Escondido with 5,815, Fallbrook with 4,414, and Carlsbad with 3,361. Weather conditions showed that "Clear" was the most frequently reported condition, with 46,354 observations, followed by "Cloudy" at 39,644 and "Rainy" at 5,767. Regarding street names, the most accident-prone locations included I-5 N with 9,989 accidents, I-5 S with 6,573, Escondido Fwy S with 5,442, and I-805 N with 4,269 accidents. Other frequently recorded streets were Escondido Fwy N, I-805 S, I-8 E, I-8 W, CA-78 W, CA-78 E, and I-15 N. Additionally, the Street Name distribution revealed several high-traffic roads with frequent accidents. Old Highway 395, US 395 HISTORIC recorded the highest number of accidents at 3,788, followed by Camino del Rio North (2,614), Market Street (2,282), and Clairemont Mesa Boulevard (1,836). Other streets with notable accident counts include Home Avenue (1,804), Kearny Villa Road (1,601), and Vista Way (1,549), with many other streets displaying varying accident frequencies. These patterns provide valuable insights into accident distribution, highlighting the cities, weather conditions, and streets most affected by traffic incidents.

Univariate graphical analysis followed, using bar charts and histograms to visualize the distribution, outliers, and counts of numerical and categorical variables.

Figure 4.1 presents a set of charts that visualize the distributions of various numerical variables. These charts provide insights into the frequency distribution of the numerical data, highlighting potential outliers, skewness, and trends across the dataset.

Figure 4.1

Distribution of Severity, Start_Lat, Start_Lng, and End_Lat, End_Lng, Distance(mi), Temperature(F), Wind_Chill, Humidity(%), Pressure(in), Visibility(mi), and Wind_Speed(mph)

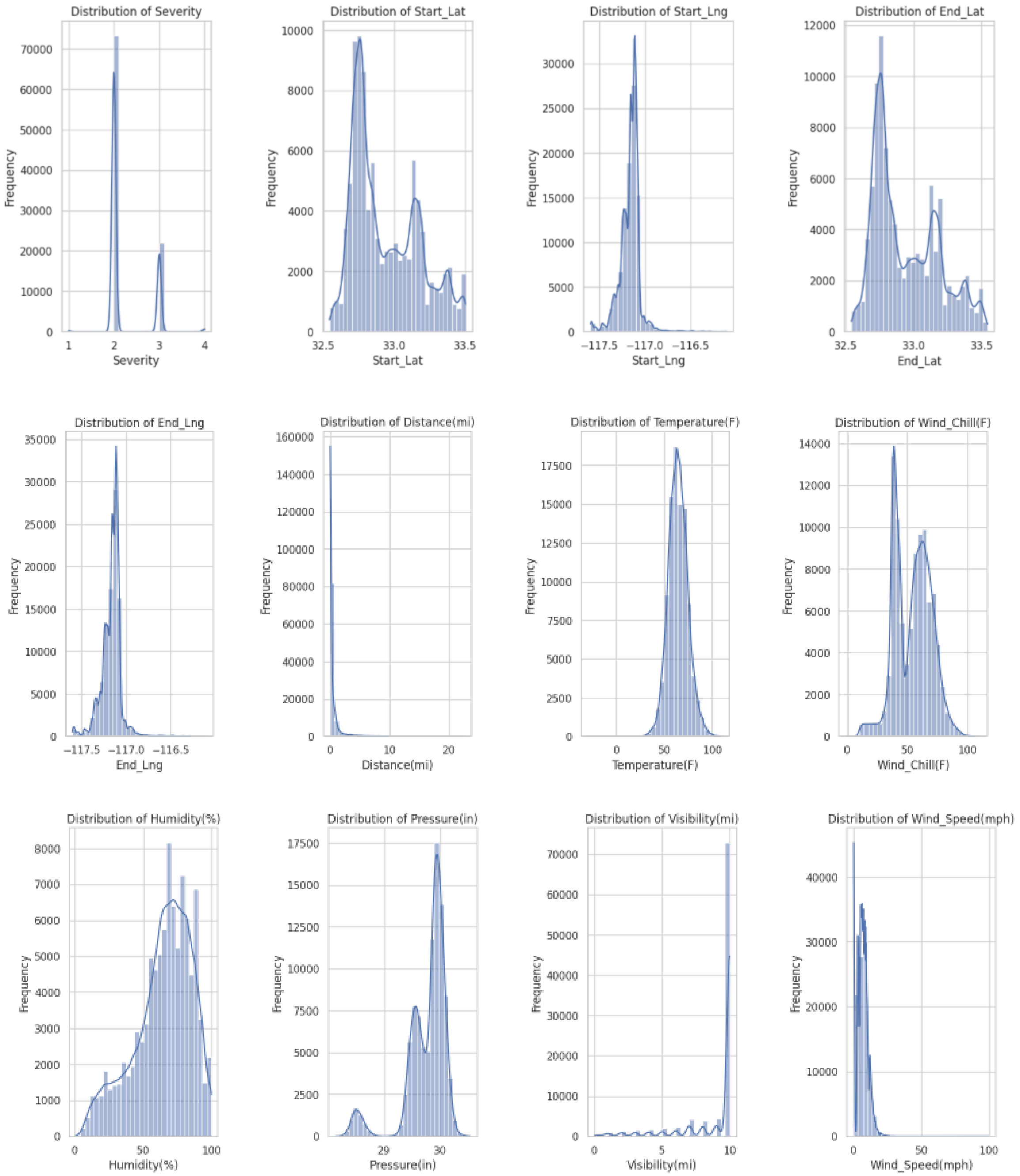


Figure 4.2 illustrates the distribution of accident counts based on the “Direction” variable. The chart visually represents how accidents are distributed across different directions, highlighting the frequency of occurrences in each category. This distribution helps to understand whether certain directions (e.g., North, South, East, West) are more prone to accidents, offering valuable insights for safety measures and traffic management in the region.

Figure 4.2

Distribution of Counts for Direction

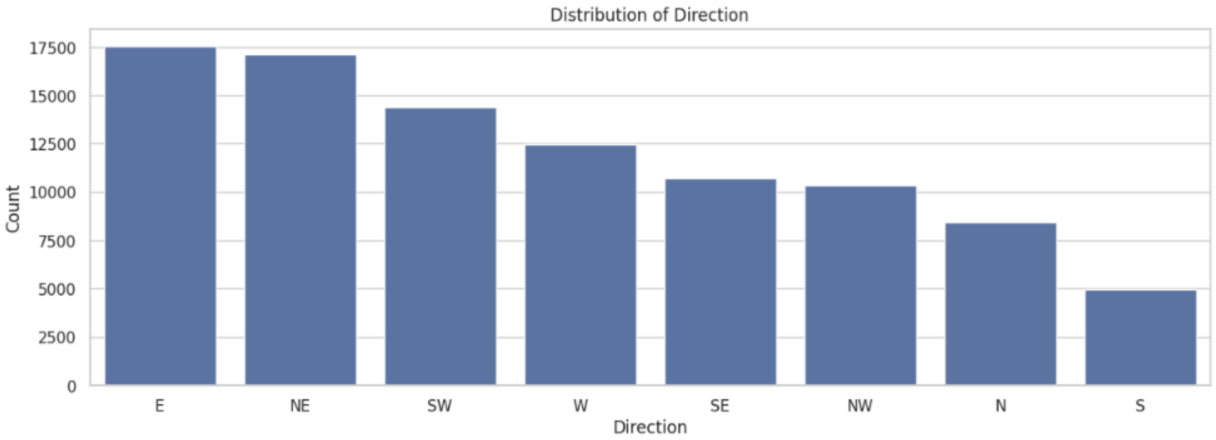


Figure 4.3 displays the distribution of accident occurrences based on various weather conditions, with the majority of incidents reported under “Clear” weather conditions, followed by “Cloudy” and “Rainy” conditions.

Figure 4.3

Distribution of Weather_Conditions

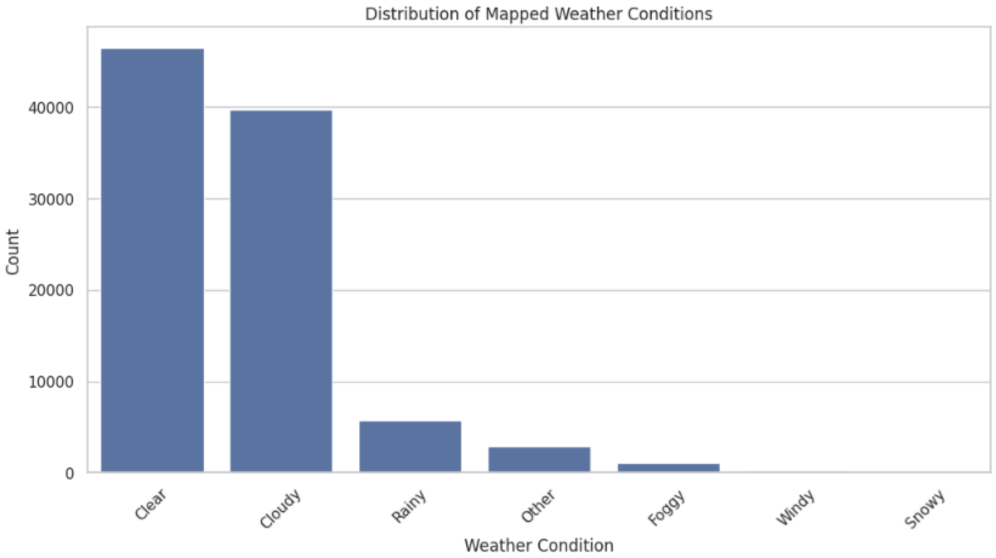


Figure 4.4 illustrates the distribution of accidents across different years. The data reveals trends in the frequency of accidents over time, highlighting any fluctuations in accident occurrences from year to year. Based on this chart, it was identified that the year 2023 is a partial year.

Figure 4.5 depicts the number of accidents per month per year. Upon further inspection of Figure 4.5, it was identified that data for January and February of 2016 is missing from the dataset, and that 2023 only contains accident data for January - March.

Figure 4.4

Distribution of Number of Accidents per Year

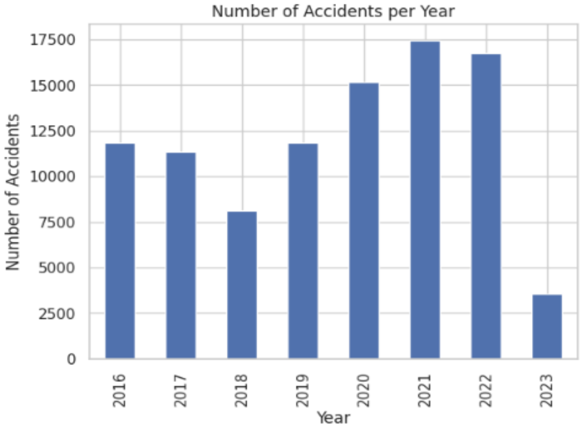
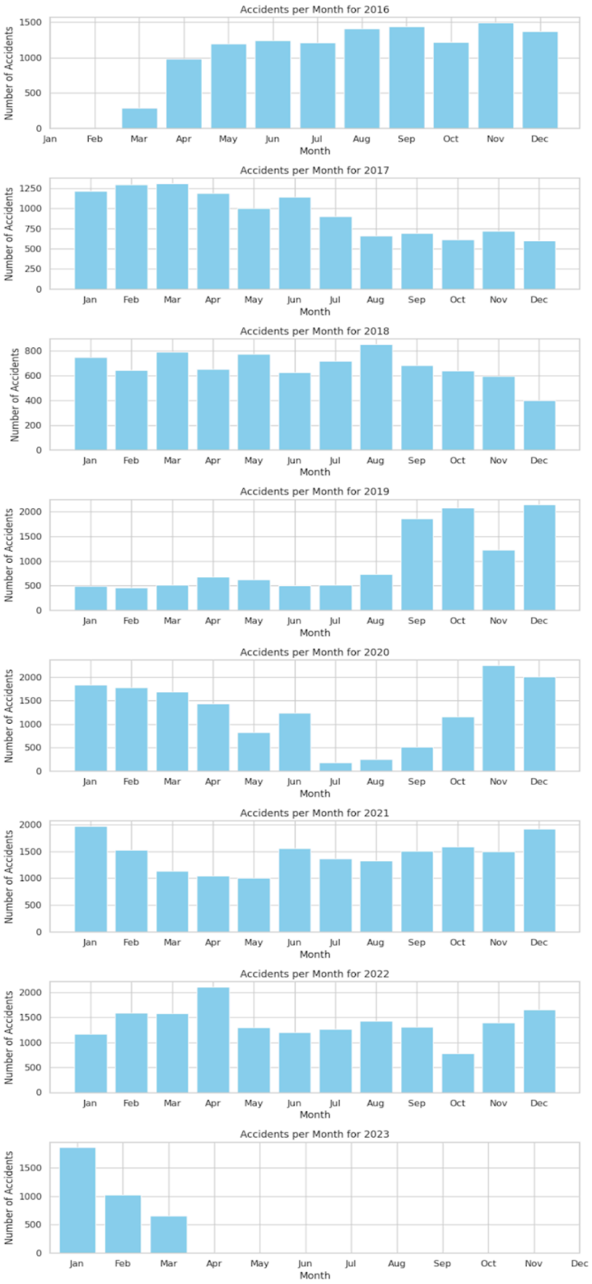


Figure 4.5

Number of Accidents per Month by year



4.2.2 Multivariate Analysis

For multivariate analysis, both non-graphical and graphical techniques were employed. A correlation matrix helped explore relationships between numerical variables, and heatmaps were used to visualize interactions between multiple variables, such as accident severity, weather conditions, and location. These visualizations provided further insights into the interactions between these variables.

Figure 4.6 displays the trend of traffic accidents over time, spanning from 2016 to 2023. This line chart shows the total number of accidents each year, offering a clear visualization of how accident rates have fluctuated over the years. The x-axis represents the years, and the y-axis indicates the total number of accidents reported. By observing the trends, one can identify patterns such as periods of increasing or decreasing accident frequency, which may correlate with various factors like changes in traffic volume, weather conditions, road safety measures, or even socio-economic events. Interestingly, it can be seen in Figure 4.6 that there was a shift in the number of accidents per year starting at the end of 2019 and remained elevated into 2023.

Figure 4.2.6

Accidents Over Time

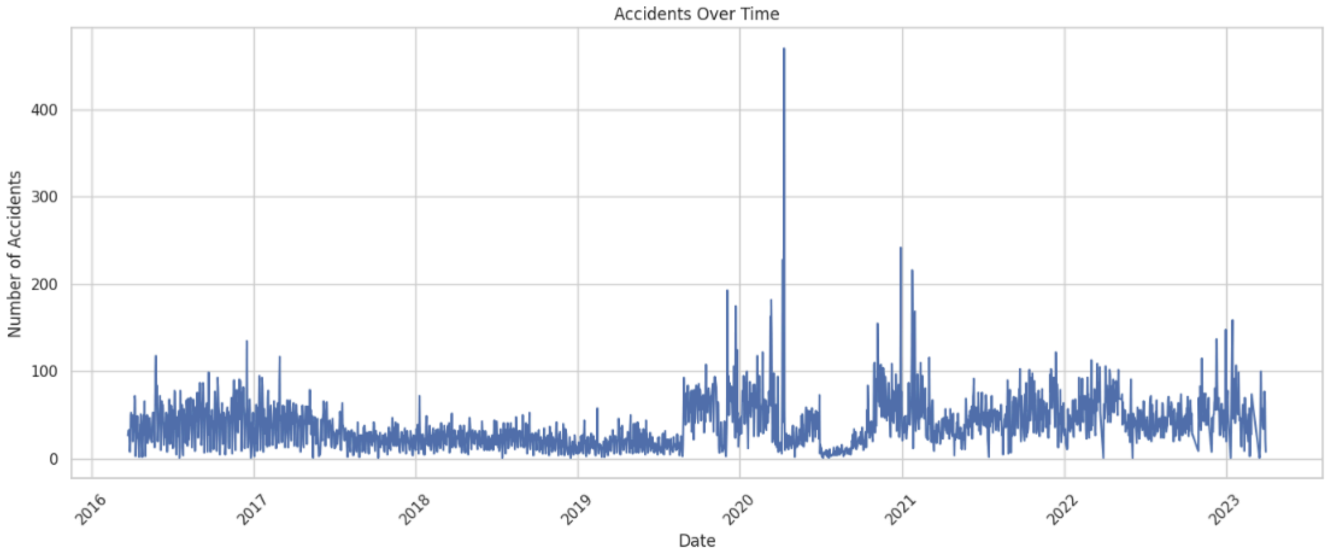


Figure 4.7 depicts a map visualizing the distribution of traffic accidents from 2016 to 2023 based on the Weather Condition at the time of the incident. Each marker on the map represents a single accident, plotted according to its latitude and longitude coordinates. The accidents are color-coded to indicate the different weather conditions present at the time of the accident. A diverse range of weather conditions, such as clear skies, rain, fog, and snow, are represented, providing insights into how these conditions might correlate with accident frequency.

Figure 4.7

Accidents mapped based on Weather_Condition

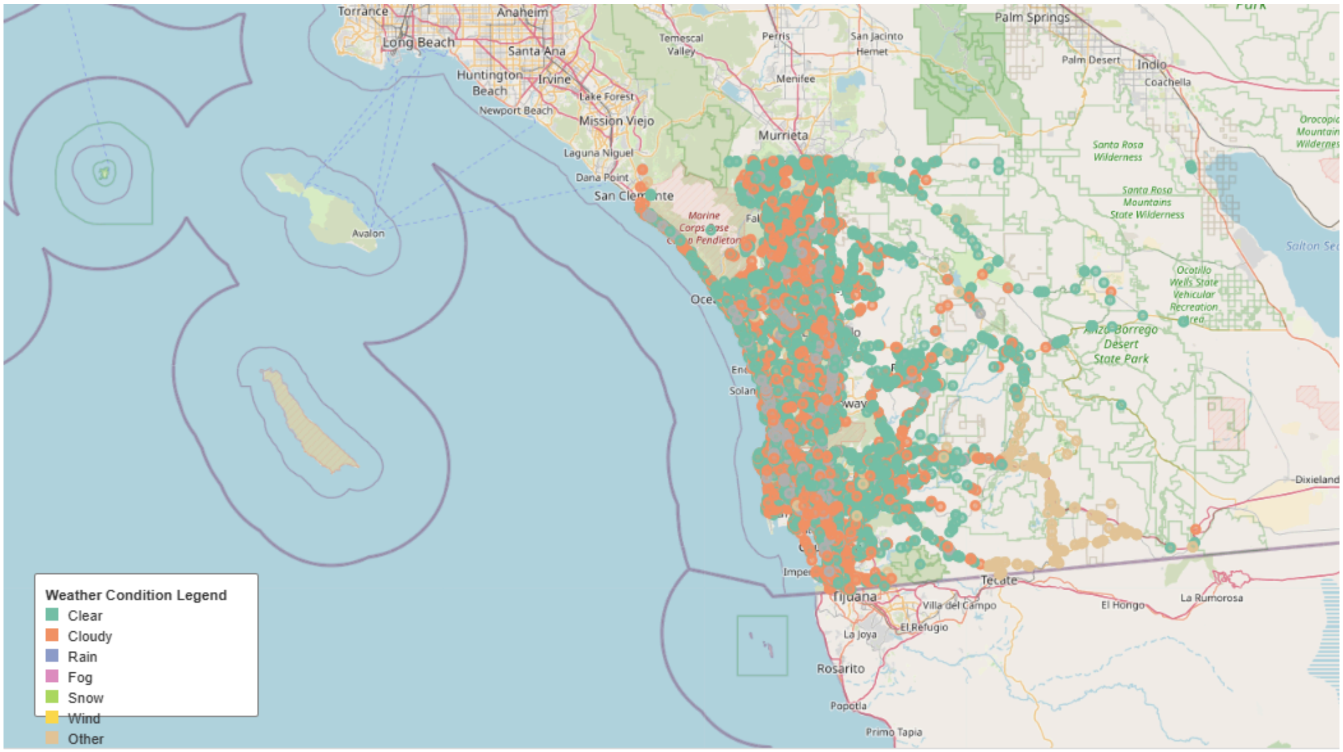


Figure 4.8 visualizes the geographical distribution of traffic accidents from 2016 to 2023, categorized by severity. Each accident is mapped according to its location, with color-coded markers representing different levels of severity, such as minor, moderate, serious, and fatal accidents. These severity levels are denoted by distinct colors, making it easy to identify areas with higher concentrations of accidents based on their severity.

Figure 4.8

Accidents mapped based on Severity

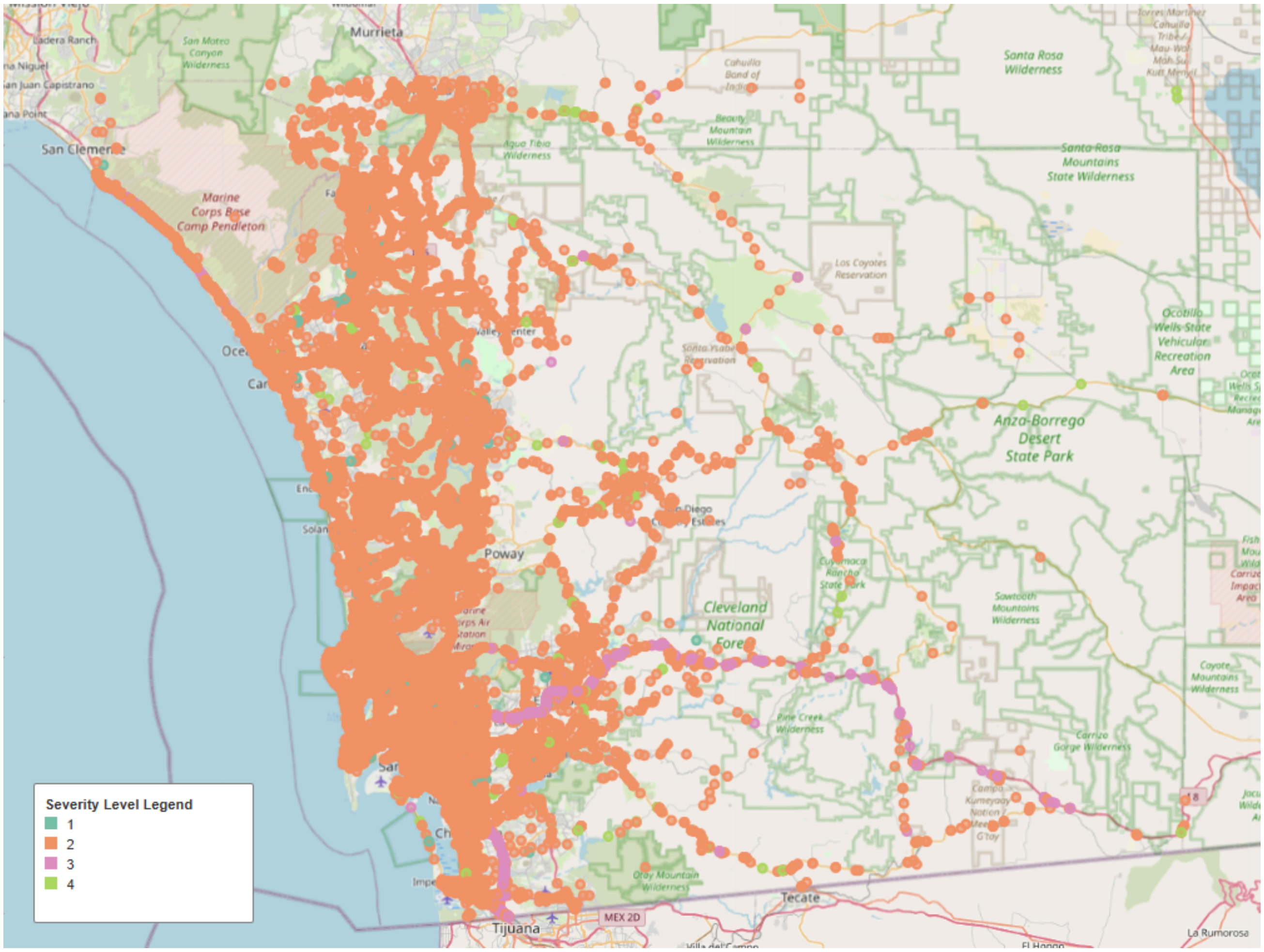
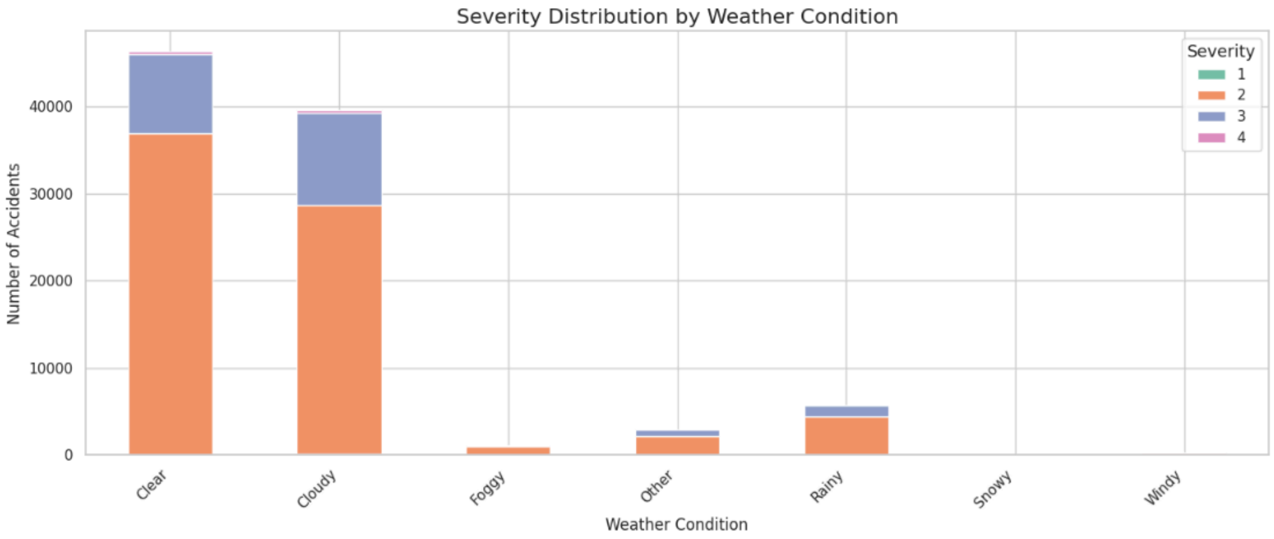

Severity Level Legend
🟩 1
🟧 2
🟪 3
🟩 4
Murrieta
San Clemente
Marine Corps Base Camp Pendleton
Poway
Cleveland National Forest
Anza-Borrego Desert State Park
Santa Rosa Wilderness
Santa Rosa Mountains State Wilderness
Tecate
Tijuana
La Rumorosa

Figure 4.9 illustrates the distribution of traffic accident severity across different weather conditions from 2016 to 2023. The chart breaks down the severity of accidents—ranging from minor to fatal—according to various weather conditions, such as clear, cloudy, fog, rain, and snow. The distribution is presented using a bar chart, with each weather condition represented along the x-axis and the number of accidents for each severity level stacked on top of each other. This visualization enables a comparison of how weather conditions correlate with the severity of traffic accidents, highlighting whether certain conditions, like rain or fog, contribute to more severe accidents.

Figure 4.9

Severity Distribution by Weather Condition



The chart titled "Accidents by Hour of Day and Weather Condition," represented in Figure 4.10 illustrates the distribution of traffic accidents throughout the day, segmented by various weather conditions. The x-axis represents the hours of the day, ranging from 0 (midnight) to 23 (11 PM), and the y-axis shows the number of accidents recorded for each hour. The data is presented as a grouped bar plot, where each group corresponds to a specific hour, and the bars in each group represent different weather conditions, such as clear weather, rain, fog, or snow. For better clarity, these weather conditions are distinguished by color, as indicated in the legend placed to the right of the chart.

The chart effectively highlights trends in accident occurrences, such as peaks between 2 PM and 2 AM, reflecting higher traffic volumes during these times. It also provides insights into how weather conditions influence accident rates. For instance, adverse conditions like rain or snow may increase accident counts compared to clear weather, particularly during high-traffic periods. This makes it possible to observe patterns and correlations between the time of day, weather conditions, and the likelihood of accidents.

Figure 4.10

Accidents by Hour of Day and Weather Condition

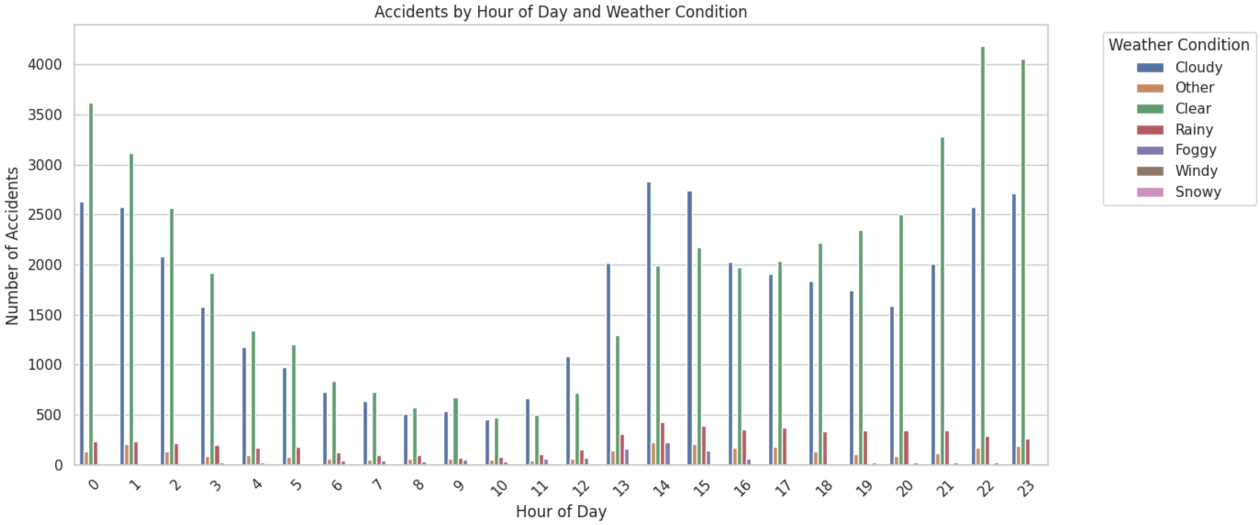
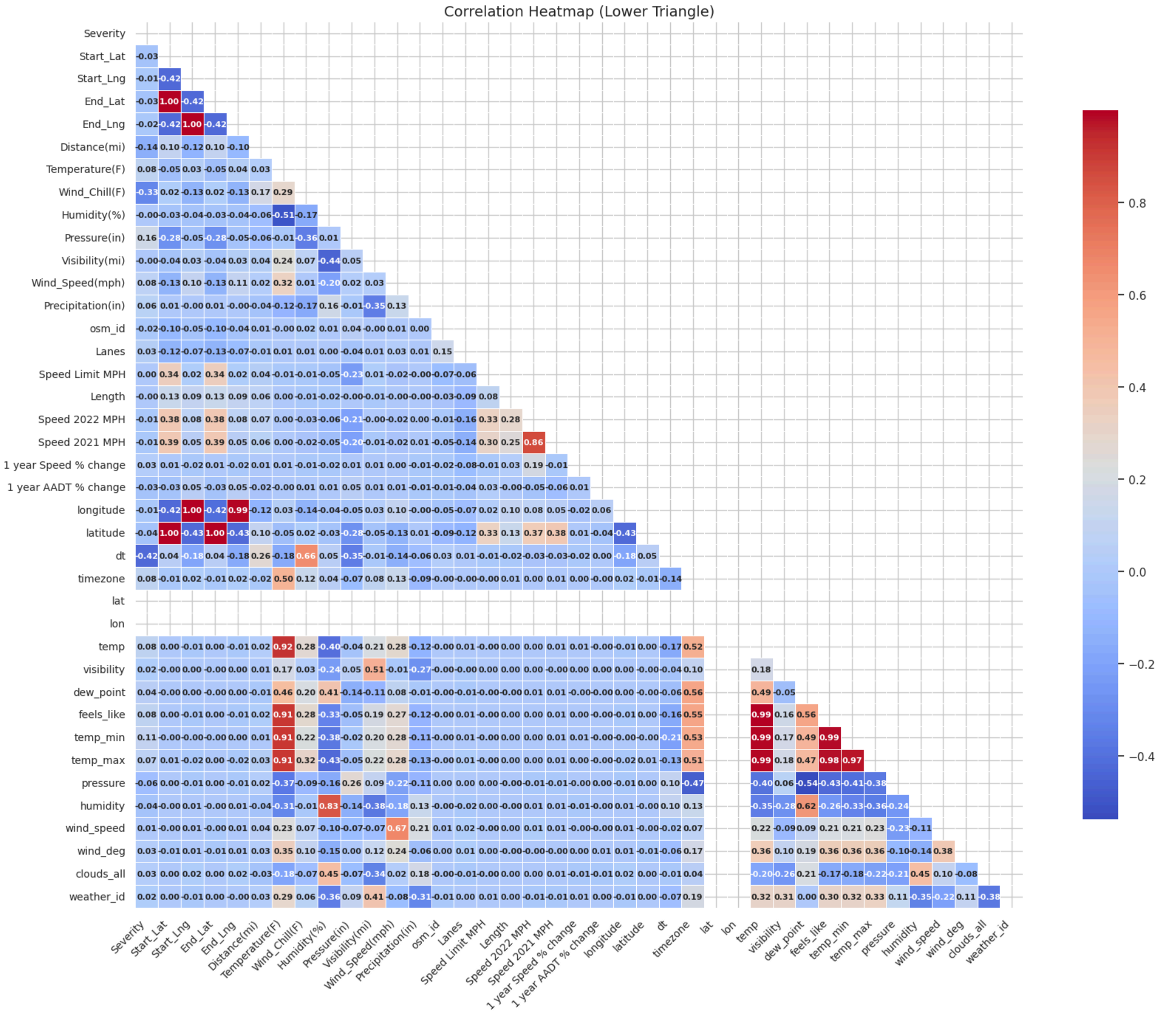


Figure 4.11 presented a correlation heatmap for all of the numerical variables in the dataset. The heatmap visually represented the relationships between the variables, with darker colors indicating stronger correlations. Several pairs of variables showed high correlation with one another, suggesting potential redundancy in the dataset. Upon further inspection, it was noted that these highly correlated variables were duplicates or highly similar measures of the same underlying concept. These duplicated variables were addressed during the feature selection and engineering phase of the project. By removing or consolidating these redundant features, the efficiency of the model was improved, and the risk of multicollinearity, which could negatively affect the performance and interpretability of the model, was reduced.

Figure 4.11

Correlation Heatmap



4.3 Data Quality

Ensuring data quality was a critical aspect of this study to guarantee the validity and reliability of the results. Several dimensions of data quality, including completeness, consistency, accuracy, timeliness and relevance were evaluated. Despite the dataset's comprehensive coverage of traffic accidents, weather conditions and traffic patterns, gaps were identified.

4.3.1 Temporal Coverage and Missing Data

It was identified that the temporal coverage for reporting periods was inconsistent, with 2 months of 2016 and 9 months of 2023 missing. This has the potential to introduce bias in the trend analysis. Additionally, a few variables were found to have a significant amount of missing data. Therefore, these columns were dropped from the analysis. It should be noted that sparse records in specific weather conditions, such as snowfall, could limit the study's ability to assess rare events.

4.3.2 Consistency and Accuracy

Consistency checks ensured geographic and temporal alignment of data. Coordinates from accident records were synchronized, and missing values were imputed where necessary. Similarly, weather data timestamps were matched with accident occurrences for accurate analysis. Categorical variables, including Weather_Condition and Wind_Direction were standardized to address inconsistencies, and missing categories were filled with a default value of "Unknown". Accuracy was another key consideration, though it is influenced by the reliability of the data sources. Accident data, derived from crowd-sourced and public agency reports, may under-represent minor or unreported incidents. Traffic data from SANDAG depends on sensor reliability, and weather data accuracy varies with station proximity to accident sites.

4.4 **Feature Engineering**

Additional feature engineering techniques were carefully applied to enhance data quality and optimize model performance, building on insights uncovered during the EDA phase. To provide clarity and structure, these techniques have been grouped into distinct categories based on their purpose and methodology. This approach ensures a comprehensive yet accessible explanation of the work performed.

4.4.1 Feature Selection

Feature selection involved dropping 50 columns deemed irrelevant or unsuitable for modeling. For these, identifiers, geographic coordinates, and redundant features such as the existence of a nearby railway or freeway exit are targeted. This step streamlined the dataset by focusing only on variables with meaningful predictive power. By reducing the feature space, noise is minimized and computational efficiency is improved, ensuring modeling only considers the most relevant attributes.

4.4.2 Feature Transformation

Feature transformation included converting categorical variables into one-hot encoded columns and transforming boolean columns into integer format for model compatibility. To address skewness, transformations like log1p and inverse transformations were applied to continuous variables, ensuring a more normal-like distribution. Additionally, standardization was performed to scale all features, improving model performance, especially for algorithms sensitive to feature magnitude.

4.4.3 Feature Generation

Polynomial features were generated to introduce non-linear relationships between continuous variables, enriching the dataset with additional predictive power. This step expanded the feature space by creating interaction terms and squared values, capturing complex dependencies that may not be apparent in the original features. These new features had the potential to provide the selected model with more nuanced information to improve predictions.

4.4.4 Encoding

Target encoding was applied to categorical variables by replacing their values with the mean target value for each category. This technique added target-aware information to the dataset improving performance, particularly for models that benefit from capturing direct relationships between features and the target variable. It was especially useful for capturing the influence of specific categorical groups on the target.

4.4.5 Dimensionality Reduction

To reduce the dimensionality of the dataset, incremental PCA was applied. This technique ensured the dataset was transformed into a lower dimensional space while preserving as much of the variance as possible. The results presented a dataset of 10 independent variables, set apart by a defining threshold of 50% or more of explained variance. By selecting components that explain a significant proportion of variance, risks such as the curse of dimensionality and overfitting were addressed alongside improving computational efficiency.

4.5 **Modeling**

The modeling process focused on predicting accident severity using advanced machine-learning techniques. The dataset for this study integrated features derived from weather data, spatial attributes, and traffic-related metrics, offering a rich set of predictors for severity estimation. Weather-related features encompassed continuous variables such as temperature, wind speed, humidity, and precipitation, capturing external conditions influencing accident risk. Spatial attributes include attributes such as latitude, longitude, and road-specific factors, and traffic-related metrics cover elements such as traffic volumes, lane counts, and speed limits, serving as critical indicators of road usage and conditions. The selection of modeling techniques was guided by the context of the final fifty features.

4.5.1 Selection of modeling techniques

The original dataset was designed with factors significantly influencing the modeling approach, including notable feature interactions and high-dimensional nature. Preprocessing involved applying PCA for dimensionality reduction, highlighting underlying patterns and reducing the dataset's complexity. This step, combined with the inclusion of non-linear relationships, was pivotal in determining the appropriate modeling techniques. Consequently, the final models selected were CatBoost, a gradient boosting algorithm, and a Multi-Layer Perceptron (MLP), a neural network algorithm tailored to handle the dataset's refined structure.

Baseline versions of the CatBoost and MLP models were first created using default parameters to establish a performance benchmark. As suspected, these models performed very poorly, failing to capture the intricate relationships within the data. The results underscored the limitations of the baseline configurations and highlighted the necessity of further tuning to enhance predictive accuracy and generalization.

To address these shortcomings, a hyperparameter optimization framework was implemented using RandomizedSearchCV to explore a predefined range of hyperparameters for each model. For CatBoost, the tuning process focused on parameters such as learning rate, depth, L2 leaf regularization, and the number of iterations, leveraging its ability to effectively handle tabular datasets, automatically process categorical variables, and reduce overfitting through L2 regularization (Prokhorenkova et al., 2018). Similarly, the MLP algorithm's optimization targeted hidden layer sizes, activation functions, learning rate, and the regularization parameter (α), enhancing its capacity to learn complex non-linear patterns due to its multiple hidden layers and adaptability (LeCun et al., 2015).

To ensure robust performance, 5-fold cross-validation was employed during the tuning process, minimizing the risk of overfitting by evaluating the models across multiple training-validation splits. The results demonstrated that hyperparameter tuning significantly improved the predictive performance of both models, transforming them into effective tools for capturing complex patterns in the data and achieving optimal results.

4.5.2. Test design, i.e. training and validation datasets

The test design employed a carefully planned dataset-splitting strategy and preprocessing pipeline to provide the machine-learning models with high-quality input data. Considering the substantial size of the final dataset, it was strategically divided into three subsets with stratification of the target variable. This approach resulted in a distribution of 70% for training, 15% for validation, and 15% for testing.

CatBoost and MLP algorithms are data-intensive models that thrive on larger training sets. According to Goodfellow et al. (2016), dedicating a significant portion of the data to the training set enables models to generalize effectively by learning diverse patterns in the input space. Machine learning models, particularly those designed for high-dimensional and complex datasets like CatBoost and MLP, require approximately 70% of the data for training to ensure sufficient learning and to mitigate underfitting. Using a separate validation set for hyperparameter tuning is standard practice, as it allows models to be evaluated on unseen data during training, reducing the risk of overfitting (Bergstra & Bengio, 2012). Allocating 15% of the data for validation provides enough examples for reliable performance monitoring while preserving a robust training set (Hastie et al., 2009). Similarly, reserving 15% of the data for the test set ensures a dependable evaluation of model performance on unseen data, adhering to best practices for balancing test size and training requirements in large datasets (Zhang et al., 2021). Stratified sampling further ensured that the distribution of the target variable remained consistent across all subsets, maintaining the representativeness of the overall dataset.

4.5.3 Model Deployment

The optimal MLP model was effectively deployed using a Streamlit-powered interface that empowers users to input a start and end location for their daily commutes— only zip codes are considered for now. Once the user inputs their commute details, the model identifies the distance to be traveled. These measurements (in miles) are then fed into the model, which uses historical accident data, traffic conditions, and weather patterns to predict the likelihood of an accident occurring along the route. By considering these factors, the model provides an adjusted risk score that reflects the specific conditions of the commute, offering a more personalized assessment of the customer's driving risk.

This adjusted risk score could be seamlessly integrated with the existing baseline system used for calculating insurance rates. Instead of relying solely on general data, the model tailors the risk score based on the individual commute, factoring in variables like road conditions, traffic density, and environmental factors that may influence accident probability. The model's output is used as a multiplier, adjusting the customer's overall accident risk score. This enhanced risk score could then be leveraged to offer more accurate insurance premiums or provide targeted safety recommendations, improving both customer satisfaction and the insurer's ability to manage risk more effectively.

5 Results and Findings

All versions of the CatBoost and Multi-Layer Perceptron (MLP) models were evaluated across training, validation, and test datasets using multiple performance metrics including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), R-Squared, Adjusted R-Squared, and Mean Absolute Percentage Error (MAPE). The results for the baseline versions of both models, as well as their optimized counterparts, are summarized in Tables 5.1, 5.2, 5.3, and 5.4

Table 5.1

Performance Summary for CatBoost Baseline Model

Dataset	Adj. R2	RMSE	MAE	MAPE (%)
Training	0.576	0.292	0.183	0.183
Validation	0.494	0.318	0.198	0.198
Test	0.506	0.313	0.196	0.196

Table 5.2

Performance Summary for CatBoost Optimal Model

Dataset	Adj. R2	RMSE	MAE	MAPE (%)
Training	0.898	0.045	0.017	0.013
Validation	0.867	0.051	0.019	0.015
Test	0.876	0.049	0.018	0.014

Table 5.3

Performance Summary for MLP Baseline Model

Dataset	Adj. R2	RMSE	MAE	MAPE (%)
Training	-14.13	1.75	0.861	0.860
Validation	-4.81	1.08	0.842	0.842
Test	-13.23	1.69	0.859	0.859

Table 5.4

Performance Summary for MLP Optimal Model

Dataset	Adj. R2	RMSE	MAE	MAPE (%)
Training	0.911	0.042	0.015	0.011
Validation	0.884	0.048	0.016	0.012
Test	0.891	0.046	0.016	0.012

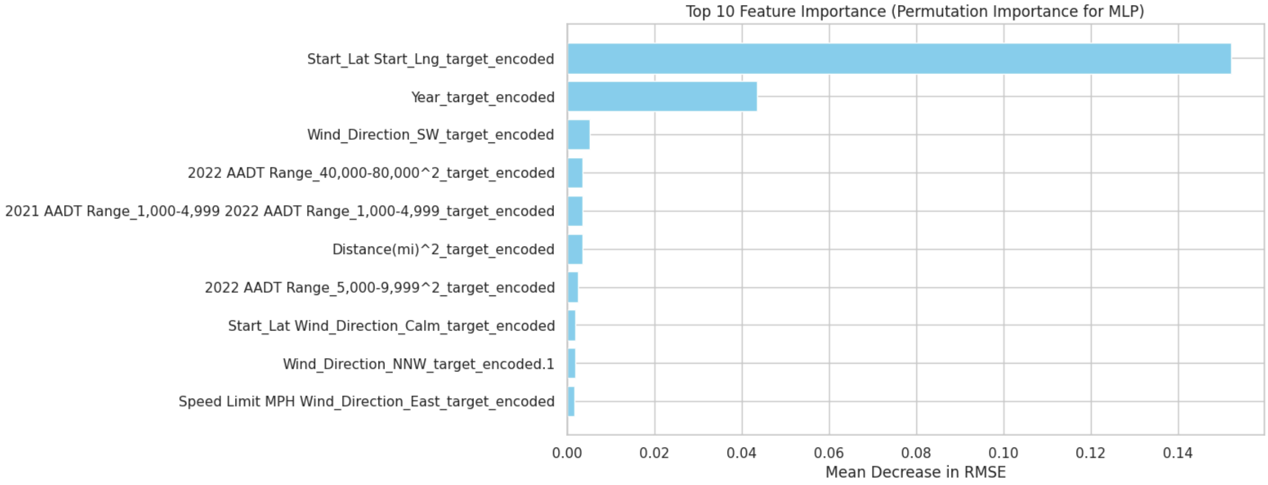
The Multi-Layer Perceptron (MLP) model demonstrated superior performance compared to the CatBoost model across all datasets, as evidenced by its higher Adjusted R² values and lower error metrics (RMSE, MAE, and MAPE). These results highlight the MLP model's ability to effectively learn complex, non-linear relationships in the data. Its neural network architecture, with optimized hyperparameters, enabled it to capture intricate feature interactions, leading to better predictive accuracy and generalization. Although the CatBoost model performed well, especially in capturing patterns in tabular data, it fell short of the MLP model's performance, particularly in terms of error reduction. The MLP model's robustness across training, validation, and test datasets—combined with its lower MAPE values—underscores its suitability for this task, where precise predictions of accident severity are crucial. The MLP model's effectiveness in modeling the dataset is reflected in Equation 1, which represents its capability to generalize complex relationships:

$$\hat{y} = \sum_{k=1}^{32} W_k(3) \bullet Z_k(2) + b^{(3)}. \tag{1}$$

It is also important to note that both models identified similar influential features— confirmed by a feature importance analysis. With little surprise, weather conditions (e.g., wind speed, precipitation) and traffic patterns (e.g., speed ranges) emerged as the top predictors. This, in turn, confirms the original hypothesis that these factors significantly influence the severity of automobile accidents. Below are the top 10 features identified as most important in the modeling process by the optimal model.

Figure 5.1

Top 10 Features for MLP Model



5.1 Evaluation of Results

The choice for the selected performance measures was driven by the desire to evaluate accuracy, reliability, and generalization capability. The optimal MLP model demonstrates low RMSE values across datasets— which in turn demonstrate high precision in predicting accident severity with minimal significant errors. Consistently low MAE values reflect accurate predictions, with small average deviations from actual severity values. Similarly, low MSE values highlight the model's effectiveness in minimizing significant prediction errors, particularly on unseen data. The model's high R-squared and Adjusted R-squared values explain approximately 91% of the variance in accident severity, underscoring the model's strong predictive power. Furthermore, the low MAPE demonstrates that the MLP model's predictions deviate by less than 1% on average from actual severity values, confirming its reliability. The minor differences between training and validation/test scores indicate a well-managed bias-variance tradeoff, showing that the model avoids under and overfitting.

Hypertuning was also a strategic decision applied to the modeling phase. In practice, this technique aids in identifying optimal parameters based on random settings across multiple model iterations. The MLP model, specifically, leveraged a randomized search grid. The respective parameters grid targeted various combinations for hidden layer sizes, learning rates, and activation functions. As expected, the iterative tuning process showed clear improvements in the model's predictive accuracy. For instance, the initial baseline performance metrics exhibited higher RMSE and lower R-squared values compared to the final tuned model. The final model achieved a test RMSE of 0.046, a significant improvement from earlier iterations— confirming the value of tuning.

The alignment of the MLP model's performance on the test dataset and the validation dataset reflects its effectiveness in generalizing unseen data. An RMSE of 0.046, slightly higher than training but consistent with validation, confirms stable generalization. The close alignment of MAE and MAPE errors (~1.2%) across both datasets confirms the model's reliability in predicting accident severity. Finally, the close alignment between the test R-squared value (~0.891) and the validation R-squared value (~0.884) also attests to the model's ability to capture variance across variables— even on unseen data.

6 Discussion

The findings from this study underscore the importance of leveraging machine learning techniques to predict accident severity based on weather, spatial, and traffic-related data. The optimal performance of the Multi-Layer Perceptron (MLP) model highlights its ability to capture complex non-linear relationships among diverse predictors, such as temperature, wind speed, and traffic patterns. The MLP model achieved a low Root Mean Squared Error (RMSE) of 0.042 and an R-squared value of 0.911, demonstrating its precision and reliability in estimating accident severity. These results align with the hypothesis that environmental and traffic conditions play a significant role in determining accident outcomes— emphasizing the utility of these factors in predictive modeling.

Furthermore, the integration of CatBoost as a complementary model showcased its interoperability and robustness, particularly in handling tabular data with mixed feature types. Although CatBoost exhibited marginally better RMSE and R-squared values, the MLP model's scalability and adaptability to non-linear interactions positioned it as the optimal choice for deployment. Overall, this study highlights the potential for using machine learning in dynamic risk assessment, enabling automobile insurers to develop proactive strategies for accident prevention and personalized premium pricing. The consistent performance across datasets and low error metrics validate the reliability of these models, paving the way for future applications in predictive analytics for traffic safety and urban planning.

6.1 **Conclusion**

This study successfully demonstrated the potential of advanced machine learning techniques, specifically the Multi-Layer Perceptron (MLP) and CatBoost models, in predicting accident severity using a rich dataset combining weather, spatial, and traffic-related features. Both models exhibited strong predictive performance across training, validation, and test datasets, with minimal differences in metrics, reflecting the robustness of the modeling approach. Although the CatBoost model marginally outperformed in terms of RMSE and R-squared values, the MLP model was ultimately selected as the optimal model due to its superior capability in capturing complex non-linear relationships and its scalability for deployment.

The study highlights the critical role of weather conditions and traffic patterns as key predictors of accident severity, which is confirmed by the feature importance analysis. These insights validate the original hypothesis and offer actionable value for stakeholders. By integrating predictive models like the MLP into real-world applications, such as personalized risk assessment for insurance or targeted safety interventions, the findings pave the way for data-driven approaches to enhancing traffic safety and reducing accident-related risks.

6.2 **Recommend Next Steps/Future Studies**

This project does an excellent job of analyzing accident severity by considering multiple factors, including accident data, weather conditions, and traffic density. By integrating these elements, it provides a comprehensive view of accident risks. However, there are several opportunities to refine and expand the analysis. For example, a follow-up study could explore which streets and highways are most accident-prone based on additional variables such as time of year, time of day, and seasonal variations in weather and traffic patterns. This type of analysis would enable the identification of specific high-risk areas during different times, providing valuable insights into the impact of temporal and environmental factors on accident severity. By determining which roads are more prone to accidents, a numerical risk rating system could be developed to reflect the level of risk associated with daily commutes on these streets and highways. This risk rating could help BMK Insurance make more informed decisions about driver routes, potentially reducing the likelihood of accidents and improving safety.

Another valuable direction for further study would be to explore how insurance companies can increase profits based on the severity of accidents. By analyzing accident severity alongside insurance payout data, insurers can gain insights into the financial risks associated with different types of accidents and adjust their pricing strategies accordingly. For example, correlating accident severity with insurance claims would enable companies to estimate how much they might need to pay out for various levels of accidents, such as minor collisions, major accidents, or total losses. This understanding would help insurers refine their risk assessment models and set premiums that more accurately reflect the potential payout for different kinds of claims. Additionally, by examining the relationship between accident severity and the likelihood of a large payout, insurance companies could better predict their potential profits and losses, optimizing their pricing structures for greater profitability. This kind of analysis could also help identify high-risk drivers or accident-prone areas, allowing insurers to target their offerings and discounts more effectively. Incorporating this data into dynamic pricing models would enable insurers to align premiums more closely with actual risk, helping them increase profits while still providing fair coverage to their customers.

A promising area for improvement is dimensionality reduction, which could simplify the dataset without losing key predictive information. Currently, some attributes may be complex or difficult to gather, especially if they depend on user location or require real-time data scraping. Automating collecting data based on user inputs is time-consuming and challenging. By reducing the dimensionality of the dataset, focusing on attributes that are either readily available from users or can be easily calculated, we could make the system more efficient and user-friendly. This would save valuable time for users, who are often seeking quick, actionable insights. In today's fast-paced world, users expect instant results, and delays in providing information could lead to frustration. Making the system faster and more intuitive would significantly enhance the user experience.

Lastly, there is room for further refinement in model selection and evaluation. The CatBoost and Neural Network (MLP) models have already shown strong performance based on multiple evaluation metrics such as RMSE, MAE, R-squared, adjusted R-squared, and MAPE. These models have provided a solid foundation for predicting accident severity, but there may be other models that could potentially yield better results. Expanding the analysis to include additional machine learning models could improve predicted accuracy. For example, models such as XGBoost, LightGBM, or Support Vector Machines (SVM) could be tested to see if they offer more precise predictions or better handle specific nuances in the data. A comparative analysis of multiple models would help identify the best-performing one, ensuring that the most accurate and reliable model is chosen for future applications. This iterative process of exploring different models and fine-tuning them can lead to continuous improvement and optimization, ultimately providing users with the most effective predictive tool.

References

Allen, P. (2023). *Your insurance loyalty is costing you money.* Allen and Allen. https://www.allenandallen.com/blog/your-insurance-loyalty-is-costing-you-money/

Becker, N., Rust, H., & Ulbrich, U. (2022). Weather impacts on various types of road crashes: A quantitative analysis using generalized additive models. *European*

Transport Research Review, *14*(37). https://doi.org/10.1186/s12544-022-00561-2

Berhanu, Y., Schroder, D., Wodajo, B., & Alemayehu, E. (2024). Machine learning for predictions of road traffic accidents and spatial network analysis for safe routing on accident and congestion-prone road networks. *Science Direct*, *23*(10). https://www.sciencedirect.com/science/article/pii/S2590123024009927#abs0011

OpenWeather. (2023). *OpenWeather Bulk Historical Data.* OpenWeather API. https://openweathermap.org/api

Peng, Y., Jiang, Y., & Zou, Y. (2018). Examining the effect of adverse weather on road transportation using weather and traffic sensors. *Plos One*, *13*(10), e0205409. https://doi.org/10.1371/journal.pone.0205409

Pourroostaei-Ardakani, S., Liang, X., enna-Mengistu, K., Sugianto-So, R., Wei, X., He, B., & Cheshmehzangi, A. (2023). Road car accident prediction using a machine-learning-enabled data analysis. *Sustainability*, *15*(7). https://www.mdpi.com/2071-1050/15/7/5939

Sanchez-Gonzalez, S., Bedoya-Maya, F., & Calatayud, A. (2021). Understanding the effect of traffic congestion on accidents using big data. *Sustainability*, *13*(13). https://doi.org/10.3390/su13137500

SANDAG. (2023). *SOC - Local roads: Speed and volume traffic data*. https://opendata.sandag.org

US Accidents (2016–2023) Dataset. (2023). Kaggle. https://www.kaggle.com/datasets

Welch, M. (2024). Predictive analytics in insurance. Luxoft. https://www.luxoft.com/blog/benefits-and-use-cases-of-predictive-analytics-in-insurance

Zuo, Y., Zhang, Y., & Cheng, K. (2021). Exploring the impact of climate and extreme weather on fatal traffic accidents. *Sustainability*, *13*. https://doi.org/10.3390/su13010390