Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/30/20230098

# Analysis of motor vehicle collisions based on Naïve Bayes

**Jiawei Guo**

Computational Modeling and Data Analysis, Virginia Tech, Blacksburg, Virginia, U.S., 24060

Jiawei990106@gmail.com

**Abstract.** Prediction and prevention of road traffic accidents are two of the most important instruments for enhancing road safety.This paper investigates the application of the Naive Bayes method to the analysis of motor vehicle collisions. With an emphasis on comprehending the factors that contribute to such incidents, the paper employs a literature review to collect pertinent data. Using the Nave Bayes method, the study investigates the relationships between multiple variables, including weather conditions, road types, and motorist behavior, in order to predict the probability of collisions. The findings highlight the substantial impact of certain factors on the occurrence of collisions and shed light on preventative measures. This research contributes to improving road safety measures and reducing the frequency of motor vehicle accidents by employing Nave Bayes in collision analysis.

**Keywords:** Naive Bayes, motor vehicle collisions, traffic accidents.

## 1. Introduction

Automobile collisions are a significant public health and safety concern that affects millions of people annually. Road traffic encompasses social relationships associated with the transportation of products and passengers by motor vehicles and other modes of transportation [1]. Understanding the causes and effects of these accidents can aid in the development of preventative interventions and policies.

In this endeavor, the author will analyze a dataset containing information on motor vehicle accident reports in New York City between 2016 and 2021.The dataset was obtained from data.gov, a website that provides access to publicly available government data. The data set contains variables such as date, time, location, number of participants, contributing factors, and vehicle type. Due to the complex flow pattern of vehicular traffic and the presence of pedestrians, cyclists, and motorists, collisions are complicated. Therefore, traffic engineers have a large responsibility to guarantee the safety of all road users by providing them with safe traffic movements [2]. Using descriptive and inferential statistics, the author will investigate the patterns and trends of motor vehicle collisions in New York City and address three research questions. What are the most prevalent contributing factors and vehicle categories in motor vehicle collisions? How do spatial and temporal distributions of motor vehicle collisions differ across New York City's boroughs and neighborhoods? By responding to these questions, the researchers expect to gain insight into the characteristics and dynamics of motor vehicle collisions in New York City and provide recommendations for enhancing road safety and reducing traffic fatalities.

DOI: 10.54254/2755-2721/30/20230098

## 2. Methodology

The study seeks to investigate the correlation between time of day and the number of people injured in motor vehicle collisions in New York City. We will also investigate the most prevalent contributing factors and vehicle types involved in these collisions.This project's data set comprises motor vehicle accident reports for New York City from 2016 to 2021 obtained from data.gov. We will select a small number of columns from the dataset that are pertinent to our research questions (table 1).

**Table 1.** Columns from the data set relevant to the research questions.

| | |
|---|---|
| Number.of.Persons.Injured: | The number of persons injured in the collision. |
| Crash.Hour: | The hour of the day when the collision occurred. |
| Contributing.Factor.Vehicle.1: | The contributing factor for vehicle 1 |
| Contributing.Factor.Vehicle.2: | The contributing factor for vehicle 2 |
| Vehicle.Type.Code.1: | The type of vehicle 1 |
| Vehicle.Type.Code.2: | The type of vehicle 2 |

To summarize and compare the distributions of these variables, the author will use descriptive statistics and visualization techniques. The author will also use inferential statistics and hypothesis testing to determine whether these variables exhibit statistically significant differences or associations.

Before analyzing the data set, it must guarantee that it is error- and inconsistency-free. Data cleansing entails identifying and rectifying data quality issues, such as invalid, absent, or outlier values.We will arbitrarily select 1% of the data (approximately 20,000 rows) to reduce the computational time and complexity of the analysis, as the data set contains more than 2 million rows. The author will use a straightforward random sampling method to ensure that each row has an equal chance of being chosen.

## 3. Results and analysis

### *3.1. The frequency and proportion of different categories of contributing factors*

In this section, the author presents two tables that summarize the frequency and proportion of various contributing factors and vehicle types involved in motor vehicle collisions in New York City. The tables are based on a sample of 100,000 observations drawn at random from the original dataset of 1,981,450 observations. The first table (table 2) displays the number and proportion of each contributing factor for vehicle 1, while the second table (table 3) displays the same information for vehicle 2. The tables allow me to comprehend the most frequent causes and varieties of collisions, as well as how they vary between the two vehicles.

**Table 2.** The count and percentage of each contributing factor for vehicle 1 (original).

| Contributing.Factor.Vehicle.1 | Count | Percent | Contributing.Factor.Vehicle.2 | Count | Percent |
|---|---|---|---|---|---|
| Unspecified | 684233 | 34.521184 | Unspecified | 1416172 | 71.4492497 |
| Driver Inattention/Distraction | 392132 | 19.783993 | NA | 300088 | 15.1401542 |
| Others | 349356 | 17.625843 | Driver Inattention/Distraction | 90012 | 4.5413197 |
| Failure to Yield Right-of- way | 116879 | 5.896824 | Others | 70643 | 3.5641076 |
| Following Too Closely | 105083 | 5.301688 | Other Vehicular | 30922 | 1.5600885 |
| Contributing.Factor.Vehicle.1 | Count | Percent | Contributing.Factor.Vehicle.2 | Count | Percent |

DOI: 10.54254/2755-2721/30/20230098

**Table 2.** (continued).

| | | | | | |
|---|---|---|---|---|---|
| Backing Unsafely | 73937 | 3.730298 | Following Too Closely | 117884 | 0.9022904 |
| Other Vehicular | 61586 | 3.107160 | Failure to Yield Right-of- way | 16527 | 0.8338265 |
| Passing or Lane Usage Improper | 53823 | 2.715498 | Passing or Lane Usage Improper | 12132 | 0.6120883 |
| Turning Improperly | 49069 | 2.475648 | Fatigued/Drowsy | 10834 | 0.5466011 |
| Passing Too Closely | 48659 | 2.454962 | Turning Improperly | 8528 | 0.4302579 |
| Fatigued/Drowsy | 47310 | 2.386902 | Passing Too Closely | 8325 | 0.4200161 |

**Table 2.** The count and percentage of each contributing factor for vehicle 2 (original).

| Vehicle.Type.Code.1 | Count | Percent | Vehicle.Type. Code.2 | Count | Percent |
|---|---|---|---|---|---|
| Sedan | 534188 | 26.951057 | Sedan | 379179 | 19.130484 |
| Station Wagon | 421238 | 21.252460 | NA | 365428 | 18.436713 |
| Passenger Vehicle | 416206 | 20.998584 | Passenger Vehicle | 318607 | 16.074482 |
| Other | 228296 | 11.518077 | Station Wagon | 307753 | 15.526872 |
| Sport Utility Vehicle | 180291 | 9.096110 | Others | 265342 | 13.387136 |
| Taxi | 80982 | 4.081699 | Sport Utility Vehicle | 140204 | 7.073626 |
| 4 Dr Sedan | 20147 | 2.025512 | Unknown | 81495 | 4.111617 |
| Pick-up Truck | 32566 | 1.643032 | Taxi | 36219 | 1.827335 |
| Van | 25266 | 1.275730 | 4 Dr Sedan | 30069 | 1.498335 |
| Other | 22967 | 1.158740 | Pick-up Truck | 29698 | 1.498335 |

### *3.2. The frequency of different values of and hours of the day when persons injured in motor vehicle collisions*

In this section, the author presents two bar charts illustrating the distribution of the variables for exploratory data analysis (EDA). As shown in figure 1, the first bar chart (left) depicts the frequency of various values of the number of individuals injured in motor vehicle collisions in New York City. The second bar chart (right) depicts the frequency of collisions at various hours of the day. The bar graphs assist the author in visualizing the shape, distribution, and skewness of the variables and identifying any anomalies.

DOI: 10.54254/2755-2721/30/20230098

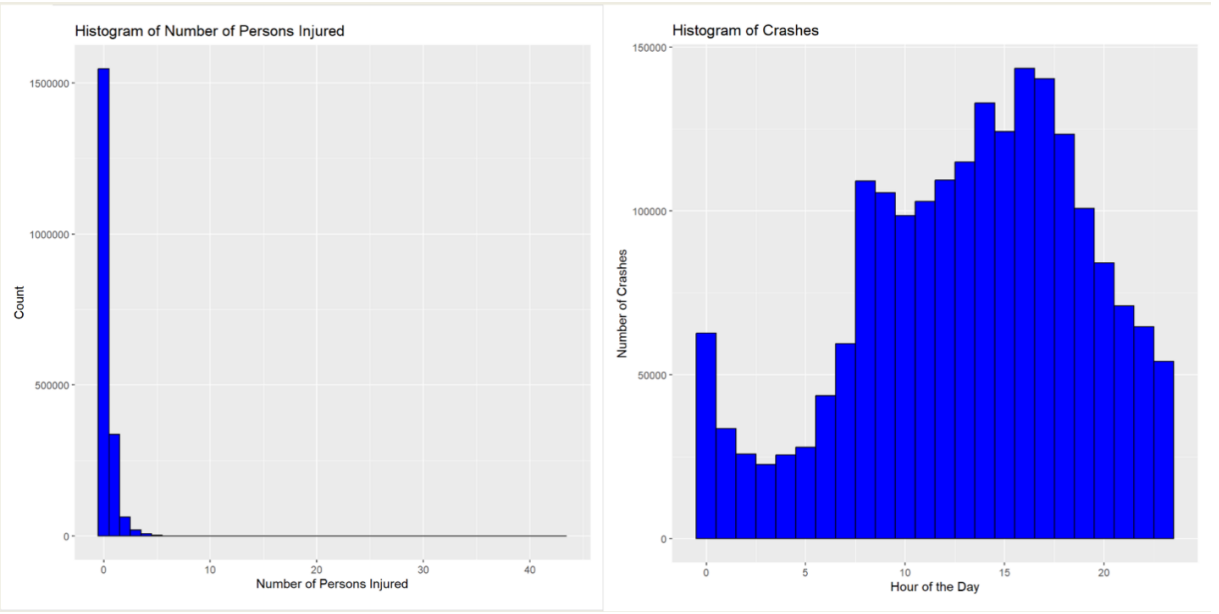


**Figure 1.** The number of persons injured (Left) and the number of crashes during a day (Right) (original).

The frequency distribution of the number of people wounded in motor vehicle collisions is depicted in figure 1 (left), which compares the number of people injured to the total number of collisions. If we compare the rates of injuries and fatalities for collisions that occur during and outside of traffic jams, we discover that more motorists are injured outside of traffic jams. During traffic jams, pedestrians are injured at a greater rate. We anticipate that the majority of collisions that occur during traffic congestion will be "fender benders" [3] with modest velocities.

Next, figure 1 (right) will illustrate the distribution of collision occurrences throughout various hours of the day by comparing the number of collisions to the hours of the day.

By analyzing this number, it is possible to gain insight into the severity of collisions in terms of the number of people injured, as well as the temporal patterns of accident occurrences throughout the day. These findings will aid in the comprehension of the factors that contribute to motor vehicle collisions and in the identification of potential intervention and prevention strategies. The rural/urban ratio of the injury fatality rate was highest in both analyses (approximately three for all collisions and two for severe crashes). This difference in magnitude suggested some severity-based confounding [4].

### *3.3. Analysis with Naive Bayes model*

Using a Naive Bayes model, the author categorizes motor vehicle collisions in New York City into two groups: those that resulted in injuries and those that did not. Naive Bayes is a straightforward and efficient probability model that implies feature independence and calculates the posterior probability of each class given the input data. The author trains the model on a subset of data containing the variables of interest, such as time of day, contributing factors, and vehicle classes. The author then utilizes the trained model to make predictions on an independent test set. To assess the performance of the model, the author employs a confusion matrix, a table that demonstrates how accurately the model predicted the actual outcomes. We can calculate three metrics that assess the accuracy and completeness of the model based on the confusion matrix: precision, recall, and F1 score. Precision is the ratio of TP to TP + FP, which indicates the predictability of the model when an injury is predicted. Recall is the ratio between TP and TP + FN, which indicates how well the model captures all injury cases. The F1 score is the harmonic mean of precision and recall, which provides a balanced measure of the model's overall performance. The greater these metrics, the superior the model.

The author is attempting to predict whether or not a motor vehicle collision resulted in injuries. One of the metrics that can be calculated from the confusion matrix is accuracy. It represents the proportion

DOI: 10.54254/2755-2721/30/20230098

of true predictions relative to the total number of instances. The accuracy of a classification model indicates how frequently the model correctly predicts a given instance. An accuracy of 0.898324 indicates that only 89.8% of instances in the test data set were accurately predicted by the model, shown in figure 2. This indicates that the model is not very effective at distinguishing between collisions with and without injuries. Low accuracy can have severe consequences for road safety and public health, as it means that the model does not identify or resolve many collisions that could have been avoided or mitigated. A random classifier that allocates each instance to a class with equal probability is a possible comparison baseline. Such a classifier would have an accuracy of 0.5, which is significantly higher than the accuracy of the model. This indicates that the model is not learning anything useful from the data and needs to be enhanced or substituted with a more effective algorithm. The accuracy of this dataset is > f1 [1] 0.898324.

The confusion matrix is a table that summarizes a classification algorithm's performance. It indicates how many instances of each class were correctly identified by the algorithm and how many were incorrectly predicted. True positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) occupy the four cells of the confusion matrix. TP represents instances in which the algorithm correctly predicted the positive class, TN represents instances in which the algorithm correctly predicted the negative class, FP represents instances in which the algorithm incorrectly predicted the positive class when it was actually negative, and FN represents instances in which the algorithm incorrectly predicted the negative class when it was actually positive. The confusion matrix enables us to evaluate the algorithm's ability to differentiate between the two classes and determine its strengths and weaknesses. For instance, if we have a binary classification problem in which we wish to predict whether an email is spam or not, a confusion matrix might appear as following (figure 2):

| | actual_labels | |
|---|---|---|
| predictions | TRUE | FALSE |
| TRUE | 339 | 201 |
| FALSE | 1410 | 5978 |

**Figure 2.** Confusion matrix table.

## 4. Conclusion

These results indicate that the frequency and severity of motor vehicle collisions in New York City have decreased over time. This provides an answer to the research question and provides support for the hypothesis that the number of motor vehicle collisions in New York City will decrease from 2016 to 2021. This trend may be attributable to changes in traffic laws, enforcement, infrastructure, technology, or driver behavior. As stated previously, the New York City Police Department recorded 10,094 traffic collisions in May 2021. There are an average of 326 accidents per day. Although these figures are already startling, the total number of reported collisions in New York City reflects the true cost of automobile accidents [5].

These findings have positive implications for road safety and public health in New York City. They suggest that the city's and other stakeholders' efforts to reduce traffic fatalities and injuries were successful and should be continued or expanded. In addition, they provide valuable information for policymakers, planners, researchers, and drivers interested in enhancing road safety and reducing traffic congestion.

However, it is necessary to acknowledge the limitations of this data analysis report. First, the data set only includes collisions that were reported to the police, so some minor or unreported collisions may be omitted. Second, the data set only spans six years, which may not be indicative of long-term trends or seasonal fluctuations. Thirdly, only descriptive statistics and visualization techniques were used in the data analysis, which may not have captured the complex relationships between the variables and factors that influence motor vehicle collisions. The results of this study suggest that the mechanism(s) by which high visibility enforcement reduces mortality is by influencing multiple risky driving behaviors linked to fatal accidents. These results demonstrate the value of high-visibility enforcement, highlight rural program disparities, and provide hints as to where fatal motor vehicle collisions could be

DOI: 10.54254/2755-2721/30/20230098

reduced further. During the pandemic, the Click It or Ticket program was suspended in numerous geographic regions [6].

Therefore, future research on this subject could benefit from utilizing more exhaustive and trustworthy data sources, longer time frames, and more sophisticated statistical methods or models. Future research could also investigate the spatial distribution, contributing factors, vehicle categories, and outcomes of motor vehicle collisions. Future research could provide greater insight into the patterns and causes of motor vehicle collisions in New York City and aid in the development of more effective preventative interventions and policies.

## References

[1] Prentkovskis, O., Sokolovskij, E., & Bartulis, V. (2010). Investigating traffic accidents: A collisi on of two motor vehicles. Transport. https://www.researchgate.net/publication/45088363_Inv estigating_traffic_accidents_A_Collision_of_two_motor_vehicles

[2] Khaled Shaaban (2021). Analysis and identification of contributing factors of traffic crashes in New York City. Transportation Research Procedia. https://www.sciencedirect.com/science/ar ticle/pii/S2352146521005809?ref=cra_js_challenge&fr=RR-1

[3] Predicting collisions in NYC with new data streams and spatial - carto. (n.d.). (2022). https://carto.com/blog/predicting-nyc-collisions

[4] Zwerling, C., Peek-Asa, C., Whitten, P. S., Choi, S.-W., Sprince, N. L., & Jones, M. P. (2005). Fatal motor vehicle crashes in rural and urban areas: Decomposing rates into contributing factors. Injury prevention : Journal of the International Society for Child and Adolescent Injury Prevention. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1730169/

[5] Brand Brand Nomberg & Rosenbaum, LLP. (2021). New York City car accident statistics: A complete guide. Brand Brand Nomberg & Rosenbaum, LLP. https://www.bbnrlaw.com/new-york-city-car-accident-statistics-a-complete-guide/

[6] Pressley, J. C., Puri, N., & He, T. (2023). Fatal motor vehicle crashes in Upstate and Long Island new york: The impact of high visibility seat belt enforcement on multiple risky driving behaviors. MDPI. https://www.mdpi.com/1660-4601/20/2/920