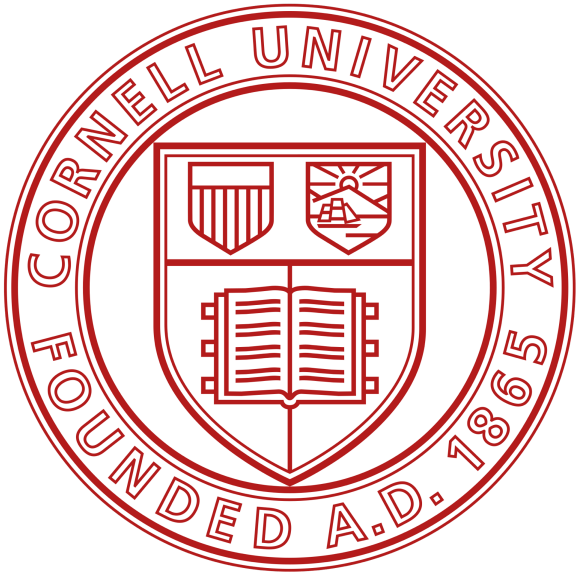


ORIE 4741

Learning with Big Messy Data

Final Report

Recommendations on Safety for Consumer Car Models

Brady Chyla
bmc227

Michael Farkouh
mtf55

May 12, 2023

Table of Contents

List of Figures and Tables

Member Contributions

Brady Chyla: Data cleaning and feature engineering, ran the regressions, helped write report

Michael Farkouh: Determined dataset, did initial analysis, prepared tables, helped write report

Introduction

Everyday, millions of Americans hit the road to run errands, get to work, or joyride. Over the course of a year, thousands of accidents occur which impact the lives of so many trying to drive. For some, these accidents could mean annoying delays and for others it could be a matter of life and death. We believe that examining the vehicles which most greatly reduce the severity of a crash will provide valuable information for drivers when making consumer choices. Also, it could hopefully inspire car manufacturers to see where they place against their competitors in making safe vehicles, and therefore strive to improve their own vehicle safety. Ultimately, our analysis could help people stay safe and reduce injuries.

Dataset Description

The dataset is the Crash Report Sampling System (CRSS) report from 2016 to 2020 produced by the National Highway Traffic Safety Administration (NHTSA). It contains information on 259,007 U.S. traffic accidents across the country, sampled from police-reported crashes of all types. For each year in the time frame, there are three files:

1. Accident Information: This contains information about the accidents, including the location, time, weather conditions, and other contributing factors, as well as how many people were involved, how many vehicles were involved, how many were injured, etc.

2. Individual Information: This contains information about the individuals involved in the crashes, including demographic information as well as whether or not the individual was drinking or was using drugs, went to the hospital, etc.

3. Vehicle Information: This contains information about the vehicles involved in the crashes, including the make and model, the VIN number, whether or not it had to be towed, and the damage occurred to the vehicle.

This dataset is very useful in answering our questions, as it gives a thorough report on all aspects of the crashes, including the make and model of the cars, the injuries incurred, and where the accident happened. All of this will allow us to assess public safety as it relates to traffic accidents and vehicle crashes. It's important to note that for our analysis, we will only be looking at the 2019 data as we want to inquire only on the newest cars and models. Additionally, there are inconsistencies in how the reports are formatted from earlier years, including giving less detailed information. We also worry the pandemic may affect our analysis on 2020 data because of the reduced frequency of driving, so we solely focused on the 2019 data.

Modeling Approach

In order to assess the safety of vehicles, we decided to create two different models that predicted the severity of injuries based on a number of factors, including the type of vehicle that was involved in the crash. To standardize and account for different conditions, we also included the weather conditions, the light levels at the time, and the location of the crash (highway or not). Then, based on the impact of the vehicle types on the model, we could see which ones reduced the severity of the injuries the most. Following that, based on severity rates of each vehicle manufacturer, we could determine the make and models that were the safest, and thus the ones that consumers who are concerned with reducing injuries in the case of a crash should buy.

Data Cleaning and Feature Engineering

Joining Tables and Formatting

Because the dataset was spread across three different files for each year, we first had to join the relevant ones together. Since we were assessing vehicle safety, we joined together the accident information and vehicle information files, which was a simple enough process as the dataset contained a case number key, so that all the information was organized by each accident. Therefore, in joining these two files together, we had information on what vehicles were involved in each accident, as well as details about the crash, such as the weather conditions.

Since there were many additional columns that were not relevant to our assessment, we reduced the dataset to the features that we were using: the vehicle manufacturer, the vehicle model, the severity of the injuries, the weather conditions at the time of the crash (raining, sunny, etc.), the light levels at the time (daylight, dusk, etc.), and the location of the crash (highway or not).

One-Hot Encoding

The next thing we had to do was better engineer features so that we could actually create the models with them. The conditions and the vehicle models were in text and convoluted categorical integer formats, respectively. Therefore, in order to create features we could use, we applied one-hot encoding.

First, for the vehicle models, we used the reference documentation to identify more general categories based on what the integers indicated: standard car (sedan), ATV, bus, light truck, standard truck, motorcycle, motor home, and other. Then, 0-1 encoded vectors were created for the vehicle model (see previous), the weather conditions (blowing particles, blowing snow, clear, cloudy, fog, freezing rain/drizzle, not reported, other, rain, unknown, severe crosswinds, sleet, and snow), the light conditions (dark - lighted, dark - not lighted, unknown, dawn, daylight,

dusk, not reported, other), crash location (highway, not on highway, unknown), and injury severity (0 for no, possible, or minor injury, and 1 for serious or fatal).

Real Encoding

In order to expand the prediction space to better signify the differences between injury severities and change the categorical variable for this feature, we real-encoded a 100 point scale on the injury feature. Originally, the maximum injury severity column that we are using had: 0 to represent no injury sustained, 1 for a possible injury, 2 for a minor injury, 3 for a serious injury, and 4 for a fatality. Obviously, a fatality is much worse than any other of the injury types, and a serious injury is much worse than the minor and possible injuries. Therefore, to account for and factor in these differences, we encoded a new scale from these categorical integer labels: 0 for no injury, 1 for possible injury, 10 for minor injury, 50 for serious injury, and 100 for a fatality. This way, in our model creations, we can more clearly assess vehicle safety in terms of avoiding the much larger numbers indicating serious and fatal injuries.

Model Creation

Linear Regression

After encoding the features we then ran an ordinary least squares regression to predict accident injury severity. From this model we were able to determine the impact of different vehicle types on the injury severity. The results are below.

Table 1: *Linear Regression Feature Relationship to Severity*

Feature	Description	Relationship to Injury Severity
Sedan	Identifies if the vehicle injuries were sustained in was an sedan	+9.229 coefficient
Motorcycle	Identifies if the vehicle injuries were sustained in was a motorcycle	+23.7532 coefficient
Truck	Identifies if the vehicle injuries were sustained in was a truck	+9.7636 coefficient

As seen in the table, the motorcycle vehicle type had the largest increase in injury severity. Also, trucks and sedans had very similar values, trucks only slightly increasing one’s injury severity when compared to sedans. This would mean that, with all other conditions being equal, being in a sedan is only slightly safer than a truck.

Regularized Logistic Regression

To confirm our results in the linear regression, we also created a regularized logistic regression model to predict accident injury severity using the same features. After creating this model, we once again looked at the impact of each vehicle model on predicting the injury severity.

Table 2: *Regularized Logistic Regression Feature Relationship to Severity*

Feature	Description	Relationship to Injury Severity
Sedan	Identifies if the vehicle injuries were sustained in was an sedan	-0.005872 coefficient
Motorcycle	Identifies if the vehicle injuries were sustained in was a motorcycle	+0.153084 coefficient
Truck	Identifies if the vehicle injuries were sustained in was a truck	+0.006263 coefficient

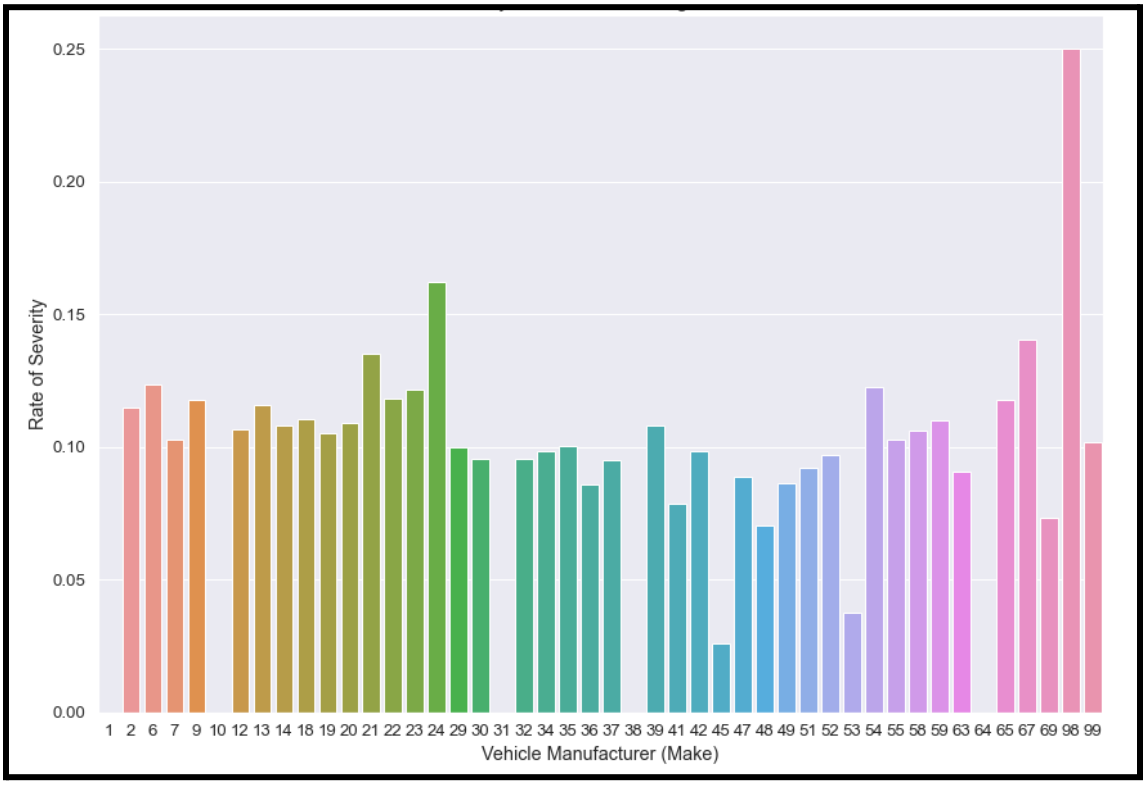
As seen in the table, the motorcycle vehicle type had the largest coefficient, meaning being in a motorcycle was the most likely to cause injuries to be severe. On the other hand, sedans had the smallest coefficient, meaning that they were the least likely to result in severe injuries.

Vehicle Manufacturer Assessment

Based on the two models, the consumer car model type that was most impactful in reducing the injury severity of crashes was a standard sedan car. We ignored buses, motorhomes, and ‘other’ because these are not standard for consumer driving, obviously (and ‘other’ was not clearly defined in the dataset). Notably, motorcycles were much more dangerous in both models, significantly increasing the prediction for a serious or fatal injury.

Since the sedan was the overall safest consumer vehicle, we then focused our attention towards the vehicle manufacturers to see which ones produced the safest sedans so consumers could figure out the best vehicle to choose from. Instead of creating more models to assess this, we simply looked at the overall rates of severe injuries sustained in accidents involving these manufacturers’ sedans.

Figure 1: *Rates of Severity for Crashes Involving Makes of Sedans*



Looking at the bar chart, we identified a few tiers of vehicles based on their sedan safety. First, there were five manufacturers that had no severe injuries in the dataset: American Motors (1), Eagle (10), Alfa Romeo (31), Isuzu (38), and Daewoo (64). American Motors went defunct in 1988, Eagle was discontinued in 1999, Alfa Romeo has only produced 44,000 total units and is an Italian luxury car, Isuzu stopped producing passenger cars in 2002, and Daewoo went defunct in 2002. Therefore, we discount all of these manufacturers, because their low (0) severity rates are clearly resulting from their low incident of being driven in the U.S. (as most are defunct, or are extremely uncommon).

Second, there were two manufacturers that had severity rates below 0.05: Porsche (45) and Suzuki (53). Suzuki is another manufacturer that we discount because of low incident rates, as they stopped selling cars in the U.S. in 2012. Porsche is the first manufacturer that we cannot discount completely, as it is at least somewhat prevalent in the U.S. However, it is a luxury brand, so it may be less accessible for consumers. Therefore, if consumers have a bigger budget to spend money on a car, we recommend Porsche as the safest manufacturer of cars.

In the third tier, we finally get to more standard vehicle manufacturers. These cars had severity rates between 0.05 to 0.08, and were Mazda (41), Subaru (48), and Other Import (69), the last of which was classified as including cars such as Aston Martin, Bentley, Bugatti, and Ferrari. Since

these imported cars are mainly foreign luxury cars that don't have a high prevalence of being driven in the U.S. we also don't fully consider these 'other imports.' However, Mazda and Subaru do have larger presences in the United States and are more oriented towards the common consumer, so for the average consumer looking for the safest car, we recommend Mazda and Subaru.

Next, we identified a fourth tier of car manufacturers, which had severity rates between 0.08 and 0.09. These cars were Fiat (36), Saab (47), Toyota (49), and Kia (63). Fiat sales in the U.S. have dropped dramatically since they peaked in 2014 at 46,000 units, and Saab went defunct in 2016 (and stopped selling cars in the U.S. in 2011), so we ignore these two for being uncommon. Therefore, we identify Toyota and Kia as the next tier of very safe standard car manufacturers oriented towards the common consumer.

Finally, we wanted to identify the bottom tier of car manufacturers, those with severity rates greater than 0.12. The cars in this category were Other Make (98), Saturn (24), Scion (67), Oldsmobile (21), Chrysler (6), Acura (54), and GMC (23). Other make largely includes vehicles that most Americans have likely never heard of, and thus can be deemed irrelevant to our analysis. Also, Saturn, Scion, and Oldsmobile have all shutdown in the last twenty years, which means their brands are also not relevant to most new consumers. However, Chrysler and Acura are prominent luxury brands and GMC is a more affordable brand, all of which many consumers today would consider buying from. Our analysis suggests that consumers who are prioritizing safety though, should rethink buying from these manufacturers.

Final Conclusion & Considerations

Ultimately, we determined that the safest model (type) of car was a sedan. Additionally, there are three tiers of these cars available to consumers that we recommend they purchase if they are looking for the safest car possible. For consumers who have the funds, Porsche was the safest choice by a significant margin, with only about 2.5% of involved crashes resulting in a severe (serious or fatal) injury. Next, the safest choices for more standard consumers were Mazda and Subaru, each with about 7% of involved crashes resulting in a severe injury (for those considering luxury vehicles, foreign imports like Aston Martin, Bentley, Bugatti, and Ferrari also fit in here as being quite safe). Finally, other safe car options for the standard consumer, each with about 8.7% of involved crashes resulting in a severe injury, were Toyota and Kia.

We are fairly confident in our results. Even though the models we created were not very predictive of the severity of injuries, our focus was on the impact of the make and model of car on assessing this. The model of cars had decent statistical significance in the models, and based on the danger revealed for motorcycles in the linear regression and regularized logistic regression matching common sense and other analyses, we think that our assessment that sedans

were the safest consumer car is reliable. Additionally, in looking at the make of the cars, the disparity between the tiers we identified were significant enough that we are also confident in our evaluations on which manufacturers made the safest and least safe cars. Consumers can reliably make changes to their car purchasing habits using our evaluations if they are prioritizing safety, hoping to reduce their chances of a crash resulting in a severe injury.

We do not believe that our results have produced a Weapon of Math Destruction. Our outcomes were very measurable: simply looking at the impact of the type and make of car on the severity of injuries when they are involved in an accident. Our predictions could have negative consequences in the sense that if we are wrong, consumers who use our recommendations would be buying cars that aren't actually as safe as we thought, which might result in more severe injuries. However, the danger of incorrectness exists in every prediction and evaluation task, so we don't think this worry rises to the level of a WMD.

There could be some fairness considerations related to possible bias in the available information. For example, the dataset relies on police reports of the incidents of crashes, no matter the severity. Consumers who drive less expensive cars (from more standard manufacturers) may report smaller accidents (like minor swipes) to the police that don't result in any injuries less frequently than those who drive luxury vehicles. This is because of insurance considerations (don't want their insurance payments to go up from being involved in an accident). Therefore, the severity rates of accidents involving these types of drivers and car types might actually be less because of this possible bias. We aren't greatly concerned with this, because of the overall size of the dataset and the standards of reporting from the Crash Report Sampling System.

Therefore, we conclude that consumers can reliably use our results to aid in their purchasing of safe vehicles, and that they should focus on sedans/cars produced by Porsche, Mazda, Subaru, Toyota, and Kia.

Appendices

Appendix A: Dataset Reference
https://www.kaggle.com/datasets/jonbown/us-2020-traffic-accidents?select=acc_19.csv

Appendix B: CRSS Documentation of the Dataset
https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813436

Appendix C: Further Detailed Documentation of the Dataset
https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813426