Data Modeling and Representation Final Project

Jiayi Zhou, Jiechen Li, Divya Sharma, Dhaval Potdar

Abstract

As of September 2023, a total of 80,000 car accidents have been reported this year, in Chicago alone. In the last five years, an average of 110,000 accidents have been reported annually, leading to 21,000 injuries. About 60% of these crashes cost more than $1500 in damages. With a 0.35% year-on-year increase in car ownership in Chicago[1], these numbers are only expected to go up. One way for cities to combat this situation is to analyze conditions that pose the most risk to life and property, and introduce interventions such as road-side warnings, additional police patrolling, etc. in high risk areas and under high risk conditions. In this project, we analyze data on road crashes maintained by the city of Chicago between January 2022 and December 2022, and predict two quantities - (i) the time taken for the authorities to be notified about a certain crash, and (ii) whether the damage cost would exceed $1500. We use data available at the time of accident such as the distance from downtown, time of day, visibility, precipitation, etc. to make predictions. In the exploratory data analysis phase however, we find it challenging to identify strong indicators for our outcome variables. The diagnostic plots of the models show strong evidence of non-linearity. Although our models would allow authorities to run simulations that would let them identify conditions that pose the most threat to drivers, as it stands, we would need to improve performance before they can be used reliably.

[1] Pearce, R. (n.d.). Forbes Advisor - Smart Financial Decisions Made Simple. Forbes Advisor. https://www.forbes.com/advisor/

Introduction

The primary data source for this project is city of Chicago's official website[2]. The dataset has information on crashes from 2015 to September 2023 and spans across 784K rows and 84 columns. However, for our analysis, we only consider data from 2022. This brings down the number of rows to 76,820. Some sample colums that we use for modeling are - time of day, day of week, and speed limit. We also use an external dataset sourced from an online proprietary weather data service[3]. This dataset gives weather-related information such as precipitation, snow, visibility, etc. for Chicago for every day in our main dataset.

[2] City of Chicago | Data Portal | City of Chicago | Data Portal. (n.d.). Chicago. https://data.cityofchicago.org/

[3] Weather Data & Weather API | Visual Crossing. (n.d.). Www.visualcrossing.com. https://www.visualcrossing.com/

The goal of this analysis is to answer two questions:

1. Given the conditions of a crash, how long does it take for authorities to be officially notified in minutes.
2. Given the conditions of a crash, would the monetary damages be major (>$1500) or minor ($<1500).

Since it is not possible to gather data at the exact instant that a crash occurs, our models gives representative pictures as to what should be expected, given the conditions of a crash. These models can be used by authorities to identify crash conditions that pose the most threat to life and property. This in turn, would let authorities identify infrastructure gaps such as inadequate surveillance or lack of mobile coverage leading to a delay in notifying authorities about crashes. Also, certain conditions that lead to consistently higher monetary damages may signal a problem with road quality or inadequate speed restrictions. With a five-year average of over 100K annual accidents in Chicago, even a 1% average reduction in the time taken to notify authorities about a crash could save lives, and potentially save millions in damages annually.

Methods

Data

For the main dataset from Chicago's official website, we identified the correct data types for each of the variables and modified them accordingly. Then we extracted temporal features such as hour of day, day of week, and month of year from the crash timestamp. We used the crash latitude and longitude to calculate the distance of the crash from downtown Chicago using Haversine distance. We then left-joined the weather data onto the main dataset on the date column, and then did a sanity check to ensure no duplicates were introduced. Lastly, we checked for null and improbable values and removed them from the dataset.

Models

In reference to the first research question[4], our target variable is a continuous variable that indicates the time taken for authorities to be notified about a crash. The target variable is calculated as the time difference between the moment of the incidence, and the moment the police were first notified.

[4] Given the conditions of a crash, how long does it take for authorities to be officially notified in minutes?

The a priori selection of the predictor variables was made by analyzing patterns in the data that indicated a possible relationship with the target variable. We then chose the year 2022, for which we have the complete data. The reason we did this was because we observed significant variation in the data across years, which was difficult to model. We also limited the police notification time to 60 minutes, because the median of the distribution is at 25 minutes, and >60 minutes represents only the trailing end of the distribution. The final model comprised of variables such as month of year, hour of day, traffic way type, distance to downtown and weather-related variables such as temperature, snowfall, and an interaction term between precipitation and month of year.

In reference to the second research question[5], the dependent variable assesses whether monetary damages are categorized as major (>$1500) or minor ($<1500). We employed Logistic regression due to the binary nature of the outcome variable, which manifests as either major damages (>$1500) or minor damages ($<1500).

[5] Given the conditions of a crash, would the monetary damages be major (>$1500) or minor ($<1500)?

We selected the predictor variables basis their association with the outcome variable. The model incorporates explanatory variables such as the time of the crash, speed limit, distance to downtown, traffic way type, roadway surface condition, and weather-related factors like temperature, and snowfall.

Model Assessment

To assess the Linear Regression model, we looked at the residual plots and found a pattern that indicated possible non-linearity. We then dove deeper into each of the predictors in the model and plotted scatter-plots for numeric variables, and

box-plots for categorical variables, with the target variable on the y-axis. Indeed, we found that most of the variables had a mostly non-linear relationship with the target variable. As an alternative, we tried to model the log of the target variable, and still didn't see significant improvements. Upon examining the Q-Q Plot, we saw significant tail-behavior on both ends, which suggests that a linear model may not be the best choice for this problem.

To assess multicollinearity within the Logistic Regression model, we calculated the Variance Inflation Factor (VIF) score. A VIF score below 5 for all predictor variables suggested the absence of multicollinearity. Regarding influential points, the Cook's distance metric was utilized, and points exhibiting high Cook's distances were systematically eliminated. Subsequently, the model was refitted without these influential points, and Cook's distance was reevaluated.

For model performance evaluation, given our predictive objective, the ROC curve was employed to determine the optimal cutoff point. Simultaneously, a confusion matrix was utilized to assess the model, wherein accuracy, kappa, precision, and F1 score were pivotal. Given the larger than $1,500 damage size indicating the severity of the accident, emphasis was placed on the true positive rate (also named sensitivity/recall).

Results

Exploratory Data Analysis

After processing the data using the aforementioned method, the dataset comprises 555,356 observations and 18 variables.

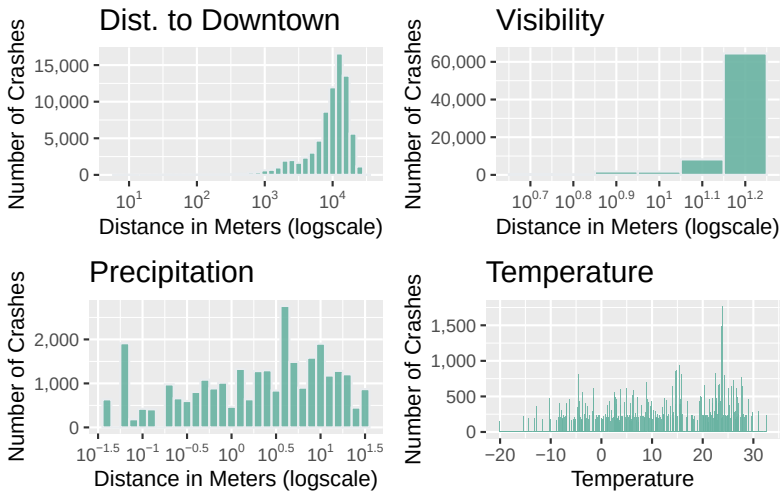
In the context of the research focusing on police response time, we identified that the median notification time is 25 minutes, the mean is 110 minutes (equivalent to 1 hour and 50 minutes), and the maximum recorded time is 1440 minutes (representing 1 day).

Regarding the research question concerning damage size, our analysis revealed that 62% of the damages exceed $\$1,500$, falling into the "Major" category, while 38% of the damages are less than $\$1,500$, falling into the "Minor" category.

For the categorical data, we examined crash hour (ranging from 0 to 23 based on the hour of the day), crash day of the week (ranging from 0 to 7 corresponding to the day of the week), and crash month (spanning from January to December based on the month). Our findings revealed that crashes predominantly occurred during the most common times of the day, specifically between 12 PM and 4 PM, on the most common days of the week, namely Fridays and Saturdays, and during the most common crash months—July, August, and September.

Additionally, we conducted a similar analysis for weather and lighting conditions, as well as road conditions. The results indicate that 65.6% of crashes transpire during daylight, 79.8% occur under clear weather conditions, and 44% take place on non-divided roads.

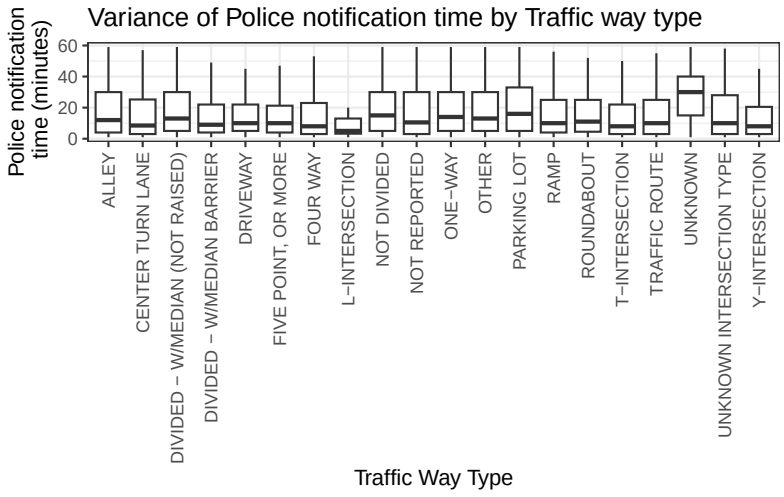
Furthermore, we examined the individual distributions of the numeric variables and observed highly skewed distributions. The accompanying plot illustrates some of these variables.



Linear Regression Model

In developing the Linear Regression model, we hypothesized that the day of week, the hour of day and month of year, would indirectly indicate how much population of Chicago would be outdoors, and thus contain some information that would help predict how long it would take for police to be notified if an accident happened. However, we found no specific relationship between the day of week and police notification time. We did however, find slight variation in the police notification time basis the time of day.

We then looked at the traffic way type, hypothesizing that certain traffic ways such as four-way intersections would have more influx of traffic, thus making accidents more likely to be reported quicker. We did find variations in the distributions of police notification time and traffic-way type.



Finally, since we were unable to find strong indicator variables for our target variable, we sought out additional data. More specifically, we sourced date-level weather data, hypothesizing that lower temperatures, or lower visibility would generally increase the time taken to report accidents. We also added distance from downtown Chicago as a variable and interaction between precipitation and month of year. Despite adding these additional variables, we only observed a slight increase in model performance. Our final model achieved an Adjusted R2 score

of 4.5%. This indicated either that the variables needed to predict the outcome are simply not captured in the data, or that we would need a non-linear model to more strongly predict the outcome. The model summary can be viewed in the appendix section of this document.

Logistic Regression Model

In developing the Logistic Regression model for predicting car accident damage costs, we integrated key factors, notably the roadway surface condition, and visibility, considering the winter conditions of Chicago. The model showed a balanced approach to predicting damages, an important indicator of its unbiased nature, though this did not directly correlate with prediction accuracy.

Our analysis for outliers using Cook's Distance revealed three notable cases, yet their presence did not significantly alter the model's performance. In assessing multicollinearity through the Variance Inflation Factor (VIF), we found that while certain variables like hour of a day had low VIF values, indicating clear independent contributions, others, particularly weather-related factors, showed higher VIF values. This suggested overlapping influences that could affect interpretability.

By examining the The Receiver Operating Characteristic (ROC) curve of the model we identified the threshold value at 0.331. However, the confusion matrix was not indicative of strong performance, and overall accuracy stood at 40%. The model's negative Kappa value of -0.063 was particularly concerning, and indicated that the model was unable to uncover meaningful relationships in the data.

In essence, Logistic Regression's strength lies in its capacity to identify severe accidents, a crucial aspect of our study. Yet, the need for refinement is clear, especially in gathering stronger variables to model the outcome, or to use a more powerful class of models.

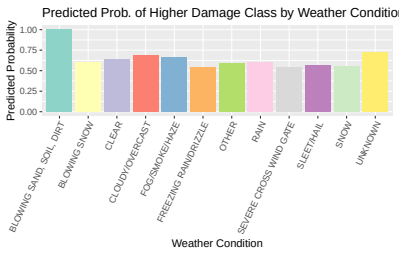


Figure 1: Predicted Probability of Higher Damage Class by Weather Condition

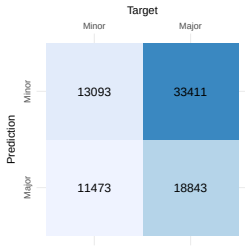


Figure 2: Confusion Matrix for Logistic Regression

Conclusion

The validity of any analysis relies on the accuracy of the assumptions made and the representativeness of the dataset. The strength of this analysis lied in its incorporation of numerous variables, including weather, road conditions, and distance to downtown, all of which were logically relevant to the assessment of damage size and the time it takes for a crash to be reported. However, modeling our outcome variables proved to be very challenging.

For the linear regression model, we found evidence of statistical significance in the traffic way type variable, and also in the interaction between the month of year and precipitation. However, with the Adjusted R2 score of 4.5% we conclude that our model is not sufficiently strong to be used for prediction purposes at this point.

For the logistic regression model, with an F1 score of 40%, it is equally evident that the models is unable to capture patterns in the data to confidently predict the damage class. Although the model has Precision of 62.5%, it struggles with a low recall of 28.5%. Most concerning is the Kappa value of -0.063, which is indicative of the model picking up on mostly noise in the data, rather than meaningful relationships.

Primarily, the presence of unknown values in a majority of the categorical variables made observing meaningful relationships difficult. Additionally, numerous unreasonable values (such as negative time taken for police to be notified) had to be removed from the original dataset since no explanation was found at the source. Furthermore, any relationships found in the dataset were largely counter-intuitive or non-linear.

Enhancements in the model's performance could be achieved with a more comprehensive dataset that accurately captures the nuances of crashes. We would also have to explore a more powerful class of models to capture the non-linearity in the dataset. A dataset of greater scope and accuracy would likely

result in improvements, effectively addressing the limitations inherent in the current analysis.

Appendix

Results for Linear Regression Model

Model Summary

The model achieved a 4.5% R2 score on the dataset. Below is the technical summary of the model.

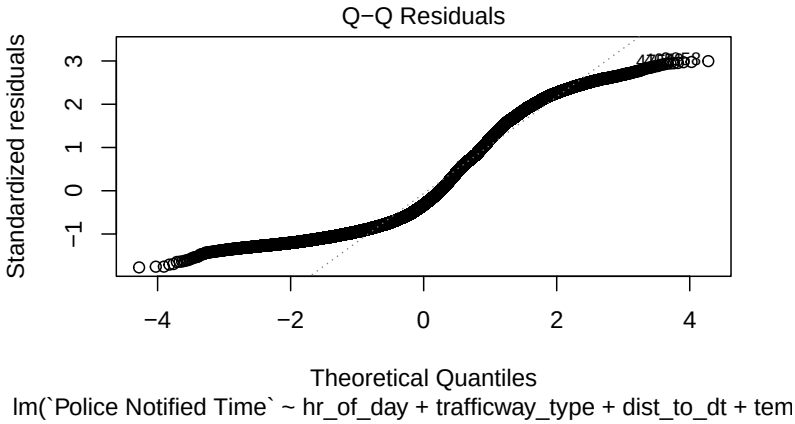
Table 1: Linear Model Summary for Police Notified Time

Predictors	Police Notified Time Estimates	std. Error	CI	p
intercept	15.87	1.03	13.85 – 17.90	**<0.001**
hr of day	0.27	0.01	0.24 – 0.29	**<0.001**
center	-2.40	1.04	-4.43 – -0.36	**0.021**
divided	0.12	0.58	-1.03 – 1.26	0.842
barrier	-3.23	0.62	-4.45 – -2.02	**<0.001**
driveway	-1.97	1.49	-4.88 – 0.94	0.185
five point	-3.31	1.46	-6.17 – -0.44	**0.024**
four way	-3.35	0.59	-4.51 – -2.20	**<0.001**
L-intersect	-8.69	2.93	-14.44 – -2.94	**0.003**
not divided	0.95	0.56	-0.15 – 2.06	0.091
not reported	-0.09	1.87	-3.76 – 3.58	0.962
one way	0.70	0.60	-0.47 – 1.87	0.244
other	0.34	0.71	-1.06 – 1.73	0.636
parking lot	3.03	0.63	1.79 – 4.27	**<0.001**
ramp	-1.45	1.34	-4.07 – 1.18	0.280
roundabout	-1.08	3.09	-7.13 – 4.97	0.726
T-intersect	-3.13	0.70	-4.51 – -1.75	**<0.001**
traffic route	-1.73	1.59	-4.83 – 1.38	0.276
unknown way	8.93	1.00	6.97 – 10.88	**<0.001**
unknown intersect	-1.85	1.15	-4.12 – 0.41	0.108
Y-intersect	-3.96	1.60	-7.09 – -0.83	**0.013**
dist to dt	-0.00	0.00	-0.00 – -0.00	**<0.001**
temp	0.04	0.01	0.01 – 0.07	**0.005**
precipitation	-0.05	0.28	-0.60 – 0.50	0.849
Feb	0.79	0.38	0.04 – 1.53	**0.040**

Mar	1.06	0.43	0.22 – 1.91	**0.014**
Apr	1.00	0.47	0.08 – 1.91	**0.032**
May	1.45	0.53	0.41 – 2.48	**0.006**
Jun	0.60	0.57	-0.53 – 1.72	0.298
Jul	0.82	0.58	-0.32 – 1.96	0.160
Aug	0.90	0.58	-0.23 – 2.04	0.118
Sep	1.26	0.53	0.23 – 2.30	**0.017**
Oct	1.65	0.47	0.73 – 2.57	**<0.001**
Nov	2.12	0.44	1.26 – 2.98	**<0.001**
Dec	1.49	0.41	0.68 – 2.29	**<0.001**
snow	0.03	0.17	-0.31 – 0.37	0.861
snow depth	0.18	0.10	-0.03 – 0.38	0.093
visibility	-0.11	0.05	-0.21 – -0.02	**0.023**
FebXPrecipitation	0.07	0.28	-0.48 – 0.63	0.802
MarXPrecipitation	-0.02	0.28	-0.57 – 0.54	0.957
AprXPrecipitation	0.02	0.28	-0.53 – 0.58	0.942
MayXPrecipitation	0.02	0.28	-0.53 – 0.57	0.950
JunXPrecipitation	0.17	0.29	-0.41 – 0.74	0.570
JulXPrecipitation	0.08	0.28	-0.47 – 0.63	0.779
AugXPrecipitation	0.08	0.29	-0.49 – 0.64	0.792
SepXPrecipitation	0.05	0.28	-0.50 – 0.61	0.856
OctXPrecipitation	0.06	0.28	-0.49 – 0.62	0.827
NovXPrecipitation	-0.15	0.31	-0.75 – 0.45	0.617
DecXPrecipitation	-0.02	0.28	-0.57 – 0.53	0.947

Q-Q Plot

The Q-Q plot shows strong tail-behavior.



Results of Logistic Regression Model

Model Summary

Below is the technical summary of the Logistic Regression Model.

Table 2: Logistic Regression Summary for Damage Class

Predictors	Damage Class Odds Ratios	Std. Error	CI	p
intercept	0.81	0.09	0.66 – 1.00	**0.047**
hr of day	1.01	0.00	1.01 – 1.02	**<0.001**
speed limit	0.99	0.00	0.98 – 0.99	**<0.001**
dist to dt	1.00	0.00	1.00 – 1.00	**<0.001**
visibility	0.99	0.00	0.99 – 1.00	0.302
ice road	0.84	0.07	0.71 – 1.00	0.052
other road	1.14	0.16	0.85 – 1.50	0.368
sand road	1.61	0.79	0.60 – 4.24	0.330
snow road	0.94	0.04	0.86 – 1.03	0.168
unknown road	1.36	0.04	1.29 – 1.43	**<0.001**
wet road	0.96	0.02	0.91 – 1.01	0.118
center	0.79	0.10	0.61 – 1.00	0.053
divided	0.79	0.05	0.70 – 0.90	**<0.001**
barrier	0.65	0.05	0.56 – 0.75	**<0.001**

driveway	1.36	0.21	1.00 – 1.84	0.051
five point	0.63	0.12	0.43 – 0.90	**0.014**
four way	0.54	0.04	0.47 – 0.62	**<0.001**
L-intersect	0.59	0.24	0.25 – 1.27	0.199
not divided	0.91	0.06	0.81 – 1.03	0.148
not reported	0.81	0.18	0.52 – 1.24	0.341
one way	0.82	0.05	0.72 – 0.93	**0.003**
other	0.88	0.07	0.76 – 1.03	0.105
parking lot	1.26	0.08	1.11 – 1.44	**<0.001**
ramp	0.87	0.14	0.64 – 1.18	0.377
roundabout	0.92	0.34	0.44 – 1.85	0.817
T-intersect	0.66	0.05	0.56 – 0.77	**<0.001**
traffic route	0.55	0.12	0.36 – 0.82	**0.004**
unknown way	0.67	0.07	0.55 – 0.81	**<0.001**
unknown intersect	0.48	0.07	0.36 – 0.63	**<0.001**
Y-intersect	0.55	0.11	0.37 – 0.81	**0.003**
temp	1.00	0.00	1.00 – 1.00	**<0.001**
snow	0.99	0.02	0.95 – 1.03	0.663

AUC-ROC

The plot below shows the ROC of the model, along with the threshold value. The AUC is 0.580, which is only slightly better than chance.