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. Using these two models, authorities can run simulations that will let them identify conditions that pose the most threat to drivers, and take appropriate preemptive steps.

[1] Forbes: Car Ownership Statistics 2023 https://www.forbes.com/advisor/car-insurance/car-ownership-statistics/#national_car_ownership_section

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 Traffic Crashes https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if

[3] Visual Crossing Weather Data Services https://www.visualcrossing.com/weather/weather-data-services

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

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).

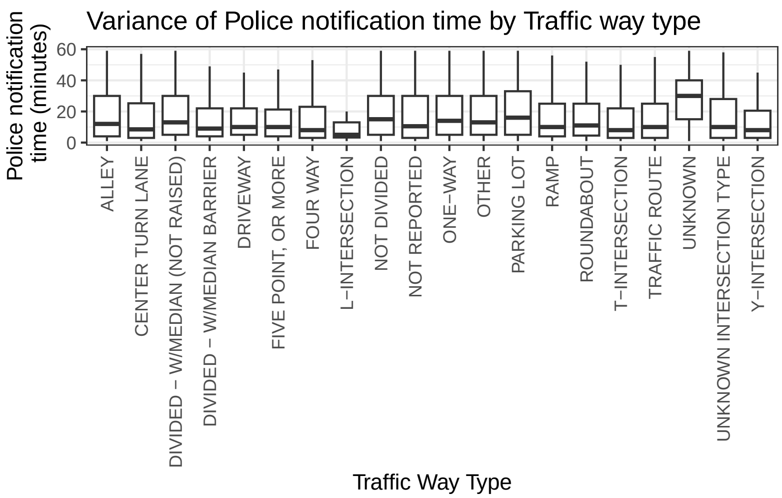
Results

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%.

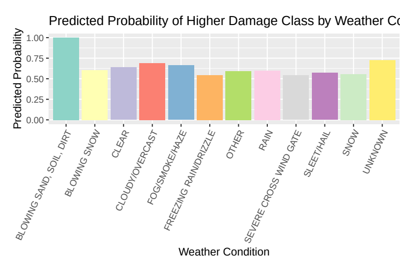


Figure 1: Caption?

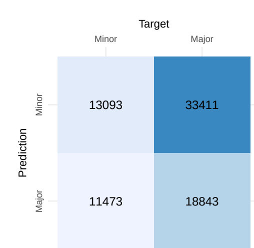


Figure 2: Confusion Matrix for Logistic Regression

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.

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.

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.

		Police Notified Time		
Predictors	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**
trafficway type [CENTER TURN LANE]	-2.40	1.04	-4.43 – -0.36	**0.021**
trafficway type [DIVIDED - W/MEDIAN (NOT RAISED)]	0.12	0.58	-1.03 – 1.26	0.842
trafficway type [DIVIDED - W/MEDIAN BARRIER]	-3.23	0.62	-4.45 – -2.02	**<0.001**
trafficway type [DRIVEWAY]	-1.97	1.49	-4.88 – 0.94	0.185
trafficway type [FIVE POINT, OR MORE]	-3.31	1.46	-6.17 – -0.44	**0.024**

trafficway type [FOUR WAY]	-3.35	0.59	-4.51 – -2.20	**<0.001**
trafficway type [L-INTERSECTION]	-8.69	2.93	-14.44 – -2.94	**0.003**
trafficway type [NOT DIVIDED]	0.95	0.56	-0.15 – 2.06	0.091
trafficway type [NOT REPORTED]	-0.09	1.87	-3.76 – 3.58	0.962
trafficway type [ONE-WAY]	0.70	0.60	-0.47 – 1.87	0.244
trafficway type [OTHER]	0.34	0.71	-1.06 – 1.73	0.636
trafficway type [PARKING LOT]	3.03	0.63	1.79 – 4.27	**<0.001**
trafficway type [RAMP]	-1.45	1.34	-4.07 – 1.18	0.280
trafficway type [ROUNDABOUT]	-1.08	3.09	-7.13 – 4.97	0.726
trafficway type [T-INTERSECTION]	-3.13	0.70	-4.51 – -1.75	**<0.001**

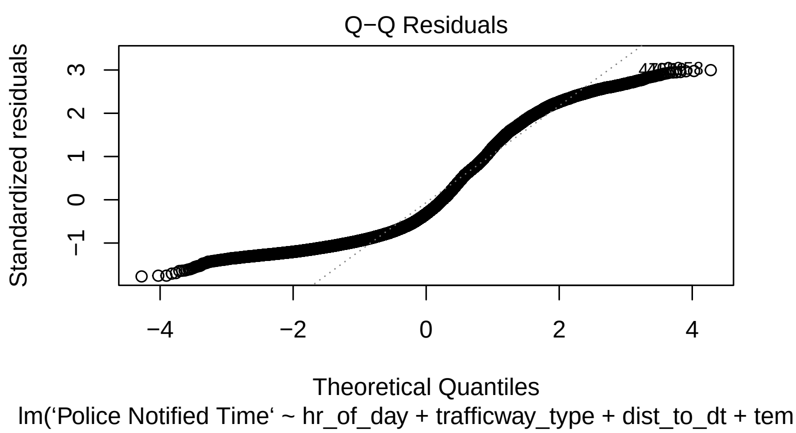
trafficway type [TRAFFIC ROUTE]	-1.73	1.59	-4.83 – 1.38	0.276
trafficway type [UNKNOWN]	8.93	1.00	6.97 – 10.88	**<0.001**
trafficway type [UNKNOWN INTERSECTION TYPE]	-1.85	1.15	-4.12 – 0.41	0.108
trafficway type [Y-INTERSECTION]	-3.96	1.60	-7.09 – -0.83	**0.013**
dist to dt	-0.00	0.00	-0.00 – -0.00	**<0.001**
temp	0.04	0.01	0.01 – 0.07	**0.005**
precipitation	-0.05	0.28	-0.60 – 0.50	0.849
month of year [Feb]	0.79	0.38	0.04 – 1.53	**0.040**
month of year [Mar]	1.06	0.43	0.22 – 1.91	**0.014**
month of year [Apr]	1.00	0.47	0.08 – 1.91	**0.032**
month of year [May]	1.45	0.53	0.41 – 2.48	**0.006**
month of year [Jun]	0.60	0.57	-0.53 – 1.72	0.298
month of year [Jul]	0.82	0.58	-0.32 – 1.96	0.160
month of year [Aug]	0.90	0.58	-0.23 – 2.04	0.118
month of year [Sep]	1.26	0.53	0.23 – 2.30	**0.017**
month of year [Oct]	1.65	0.47	0.73 – 2.57	**<0.001**

month of year [Nov]	2.12	0.44	1.26 – 2.98	**<0.001**
month of year [Dec]	1.49	0.41	0.68 – 2.29	**<0.001**
snow	0.03	0.17	-0.31 – 0.37	0.861
snowdepth	0.18	0.10	-0.03 – 0.38	0.093
visibility	-0.11	0.05	-0.21 – -0.02	**0.023**
precipitation × month of year [Feb]	0.07	0.28	-0.48 – 0.63	0.802
precipitation × month of year [Mar]	-0.02	0.28	-0.57 – 0.54	0.957
precipitation × month of year [Apr]	0.02	0.28	-0.53 – 0.58	0.942
precipitation × month of year [May]	0.02	0.28	-0.53 – 0.57	0.950
precipitation × month of year [Jun]	0.17	0.29	-0.41 – 0.74	0.570
precipitation × month of year [Jul]	0.08	0.28	-0.47 – 0.63	0.779
precipitation × month of year [Aug]	0.08	0.29	-0.49 – 0.64	0.792

precipitation × month of year [Sep]	0.05	0.28	-0.50 – 0.61	0.856
precipitation × month of year [Oct]	0.06	0.28	-0.49 – 0.62	0.827
precipitation × month of year [Nov]	-0.15	0.31	-0.75 – 0.45	0.617
precipitation × month of year [Dec]	-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.

	Damage Class			
Predictors	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
roadway surface cond [ICE]	0.84	0.07	0.71 – 1.00	0.052
roadway surface cond [OTHER]	1.14	0.16	0.85 – 1.50	0.368
roadway surface cond [SAND, MUD, DIRT]	1.61	0.79	0.60 – 4.24	0.330
roadway surface cond [SNOW OR SLUSH]	0.94	0.04	0.86 – 1.03	0.168
roadway surface cond [UNKNOWN]	1.36	0.04	1.29 – 1.43	**<0.001**
roadway surface cond [WET]	0.96	0.02	0.91 – 1.01	0.118

trafficway type [CENTER TURN LANE]	0.79	0.10	0.61 – 1.00	0.053
trafficway type [DIVIDED - W/MEDIAN (NOT RAISED)]	0.79	0.05	0.70 – 0.90	**<0.001**
trafficway type [DIVIDED - W/MEDIAN BARRIER]	0.65	0.05	0.56 – 0.75	**<0.001**
trafficway type [DRIVEWAY]	1.36	0.21	1.00 – 1.84	0.051
trafficway type [FIVE POINT, OR MORE]	0.63	0.12	0.43 – 0.90	**0.014**
trafficway type [FOUR WAY]	0.54	0.04	0.47 – 0.62	**<0.001**
trafficway type [L-INTERSECTION]	0.59	0.24	0.25 – 1.27	0.199
trafficway type [NOT DIVIDED]	0.91	0.06	0.81 – 1.03	0.148

trafficway type [NOT REPORTED]	0.81	0.18	0.52 – 1.24	0.341
trafficway type [ONE-WAY]	0.82	0.05	0.72 – 0.93	**0.003**
trafficway type [OTHER]	0.88	0.07	0.76 – 1.03	0.105
trafficway type [PARKING LOT]	1.26	0.08	1.11 – 1.44	**<0.001**
trafficway type [RAMP]	0.87	0.14	0.64 – 1.18	0.377
trafficway type [ROUNDABOUT]	0.92	0.34	0.44 – 1.85	0.817
trafficway type [T-INTERSECTION]	0.66	0.05	0.56 – 0.77	**<0.001**
trafficway type [TRAFFIC ROUTE]	0.55	0.12	0.36 – 0.82	**0.004**
trafficway type [UNKNOWN]	0.67	0.07	0.55 – 0.81	**<0.001**
trafficway type [UNKNOWN INTERSECTION TYPE]	0.48	0.07	0.36 – 0.63	**<0.001**

trafficway type [Y-INTERSECTION]	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 0.559, which is only slightly better than chance.