Crash Damage Prediction

Yi-Min Yang
Department of Statistics
Purdue University
yang1984@purdue.edu

Yu-Ju Ku
Department of Statistics
Purdue University
ku30@purdue.edu

Yu-Wen Wang
Department of Statistics
Purdue University
wang4862@purdue.edu

Yi-Chen Lin
Department of Electrical and Computer Engineering
Purdue University
lin1513@purdue.edu

I. An overview of the problem

Nowadays, transportation is one of the most essential parts of human society. People now can drive cars, take trains and take planes for moving to places where they want or need to go. The most common way for daily transportation of societies is driving cars. It is quite simple that people can buy cars everywhere and driving cars is also so easy that almost everyone has a driver's license for normal driving. However, there are a myriad of reckless drivers that cause serious accidents that are harmful for other people and societies. Also, except for the human factors that cause accidents, the environment of the roads, weather, time and even types of automobiles can be the significant factors that lead to car accidents.

In this project, we assume we are an insurance company that we want to know under certain condition of a crash accident will cause severe damage or not. If we are able to learn whether such a situation can cause serious damage or not, we can have a proper way to determine a suitable price of the insurance products. Therefore, we aim to build predictive models to estimate the property damage severity regarding an accident in this project.

II. Motivation

A concern over the rapidly rising numbers of traffic accidents, both fatal and non-fatal, raises attention to related property damage. Estimating the value of property damage lets people claim their rights expecting the at-fault driver or their insurance company to pay for damages in car accidents.

Insurance companies often suggest that people could gather purchase receipts, or look for purchase information from your credit card or online shopping account. Demand reimbursement for the total amount of everything you lost in the accident. This way, it makes the insurance company pay the full value of one's lost or damaged personal property.

However, this could take up a lot of time. Plus, some purchases could be made such a long time ago that could barely be tracked. Having a reliable estimation regarding property damage, can help people to have a clearer picture of the amount to claim for insurance cover.

III. Data Source

We obtain the traffic crash data [1] from the Chicago Data Portal [2], a website that provides various kinds of Chicago city data, including but not limited to Education, Environment, Public Safety, Transportation...etc. Furthermore, the data terms of use are provided. [3]

IV. Risk management

- Find another dataset.
- Try to analyze the dataset in other ways. Analyze the same data in a different perspective. If the recommendation of Insurance policy for insurance corporations does not work, we will try to predict the over-rating.

V. Plan of activities and timeline

- 2/2∼3/7: Do simulation of data mining approaches that are applied to the similar problem in three research papers. Summarize the solutions and results presented in those previous works and our simulation.
- 3/8∼3/21: Perform data preprocessing, exploratory data analysis and feature selection. 3/22 3/28:Discussion of the specific data mining task (classification, regression, clustering, pattern discovery, ...) Deciding which features to use on models, and choosing the proper model for accessing the problem.
- 3/28: Second Report Due.
- 3/28∼4/4: How you have solved the task (algorithms used, or describe new ones developed) and discussion of outcomes.
- 4/4∼4/11: Formally analyze outcomes. Using grid search and cross validation method on evaluating the models we used. Discuss the possible metric on evaluating overall performance of the model.
- 4/11∼4/18: What insights have gained to address the original problem? What reason to believe that this could improve the state-of-the-art for that problem?
- 4/18∼4/25: Prepare slides for the final presentation.

VI. Literature Survey

We first take a look at the article *Advancement of Weather-related Crash Prediction Model Using Nonparametric Machine Learning Algorithms* [4]. Weather-related accidents occur in any severe condition such as snow, rain, and winds or on slick pavement. Crash severity prediction due to inclement weather conditions has been the focus of extensive studies to enhance safety and to minimize the economic cost in responding to such emergencies. The weather-related crash prediction models identify the impacts of those contributing factors and enable one to predict the crash severity based on those contributing factors for better decision making and effective strategy implementation. Results presented in this paper demonstrated that the magnitude and dynamics of uncertainty in crash severity depend not only on the weather condition but also on the other crash associated auxiliary variables. This study considered two machine learning applications—random forest (RF) and bayesian additive regression trees (BART) for performance comparison. They found that the RF model produced higher prediction probabilities for determining crash severity compared to BART.

In our project, we also aim at crash severity prediction, with a focus on predicting the damage level. Moreover, by building our prediction models, we want to recognize the contributing factors that impact the crash damage as well. In contrast, the difference between our project and this study would be the factors' type. The factors they focused on were weather conditions, whereas we place emphasis on such conditions as geographic factors, crash types, etc.

Secondly, the article *Exploring The Mechanism of Crashes with Automated Vehicles Using Statistical Modeling Approaches* [5] mainly uses the ordinal Logistic regression and CART on Automobile vehicle crashes analysis. They use these models to verify traffic, roadway and environmental factors that lead to crashes. They use these factors to classify the potential severities and collision types. In their research result, they concluded that crash severity significantly increases if the AV is responsible for the crash by statistical analysis. Also, they conclude that the highway is the location that has a high positive relation with severe injuries. In our case, we also want to know the relation of environmental factors, geographic factors, crash types and the final damage estimation. We may focus more on predictive modeling analysis and try to predict out possible damage level of each observed crash. In modeling, we will try to use Logistic regression and CART as well. We will further use Random Forest and Neural Network on training dataset to see if we can get better predictive results.

Last but not least, we also did survey on the literature *Analysis of Road Traffic Crashes in The State of Qatar* [6]. The difference between our goal and this literature is that, our goal is to predict the cost of the damage and the literature is to find out the reasons why caused the damage.

The methodology of the literature uses Multivariate Analysis of Variance(MANOVA) and considering the time-series, however, we are going to use an artificial neural network, logistic regression to analyze the dataset, besides, predicting the cost that the damage would be.

The literature claims that the main reason for the damage is the weather. The weather has a big difference between Chicago and Qatar. Since the weather at Chicago is between -10°C$\sim$30°C and the weather at Qatar is between 22°C$\sim$43°C. Therefore, we can use the weather as the main idea to analyze, but we cannot claim that the weather is also the main reason that caused damage.

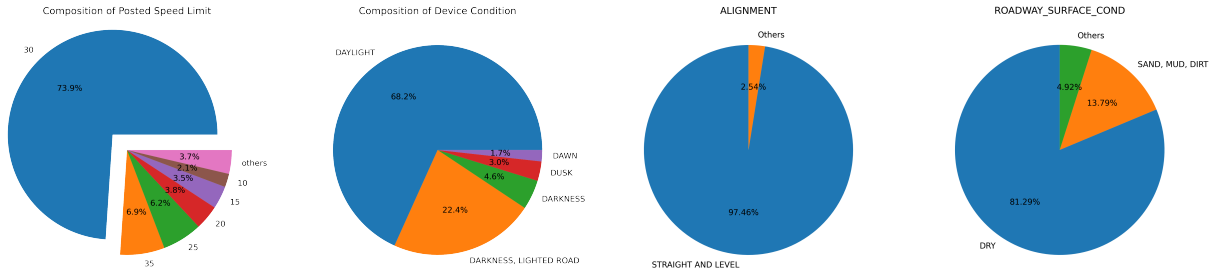
VII. Data Loading and Cleaning

In the part of Data Loading, we obtain our data from the Chicago Data Portal. The data has 598,284 rows and 49 columns with categorical and numerical variables. For data cleaning, first of all, we remove the columns that have more than 100,000 NAs and the rows with NA. This leaves us 574,196 rows and 38 columns.Second, we remove some corrupted data or data with unknown labels. For instance, the data with wrong latitude and longitude. Last but not least, we remove the rows that include 'UNKNOWN'. Some categorical variable like weather condition, it has unknown label in its data. We simply remove it to prevent from the misleading. In summary, we are left with 459,855 rows and 37 columns.

VIII. Exploratory Data Analysis

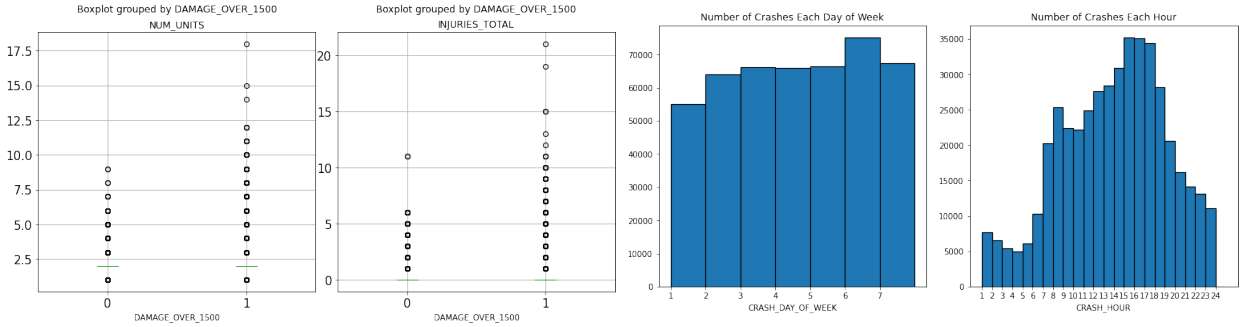
We plotted the composition of each variable. We choose four variable as examples to show the exploratory result of this process. In Figure 1, we have the pie chart of four variables, which are 'POSTED_SPEED_LIMIT', 'LIGHTING_CONDITION', 'ALIGNMENT' and 'ROADWAY_SURFACE_COND'. By observing Figure 1(A), We can see that approximately 74% of posted speed limits are 30. Also, 96.3% of posted speed limits are between 10 to 35, which indicates that the crashes mostly happen at low speed. In Figure 1(B), it is shown that most of the crashes happen while there is sufficient light. For the variable 'ALIGNMENT' plotted in Figure 1(C), the category 'Straight_And_Level' has taken about 97.46%, which is highly unbalance variable for model to learn by this variable. Therefore, we decided to remove this feature since it may not be useful for the model to classify the damage level. Furthermore, as shown in Figure 1(D), the dry condition is 81.29%, and the sand, mud, dirt, is about 13.79%. We expect this will have some effect on the classification. Thus, we will use these different surface conditions to analyze whether it would cause damage.

In terms of Figure 2(A), it shows that both levels of damages mostly have similar number of units that involved in the crashes. Their Q1, mean and Q3 are really close that the box shows a line shape. The plot seems cannot show mean difference of these two level against to this variable. However, we can see that damage over 1500 did have more units involved, which is make sense that the more units involved, the more damage it costs. As shown in 2(B), most injuries are around 0 to 2 no matter what damage they cost, which has similar situation as the previous box plot. In this plot, we can see that the damage over 1500 does have more injuries than



(A) (B) (C) (D)

FIGURE 1



(A) (B) (C) (D)

FIGURE 2

the other case. It also makes sense that the more injuries the more damage it costs. According to Figure 2(C), we can notice that crashes happen the most frequently on Saturdays, whereas Mondays have the least number of crash cases. Although not obvious, the second highest number of crash cases take place on Sundays. We can further conclude that crashes happen the most on weekends and the least on the first day of weekdays. We can find from Figure 2(D) that crashes increasingly happen around 7:00 am to 3:00 pm. In addition, the peak period of crashes is around 3:00 pm to 5:00 pm, which is consistent with the normal rush hour. After 6:00 pm, the number of crash accidents gradually decreases.

As we can see from Figure 3, the crash distribution at Chicago, there are several hotspots that car crashes happen. To name a few, some places are around such tourist spots as John Hancock Building, Millennium Park, Chinatown, etc. Besides, we can also find it obvious that lots of crash accidents take place at some main routes, such as I-90 Expressway, I-55 Stevenson Expressway and I-94 Highway. Plenty of people commute to Chicago via these interstate routes, and thus lots of crashes tend to happen here.

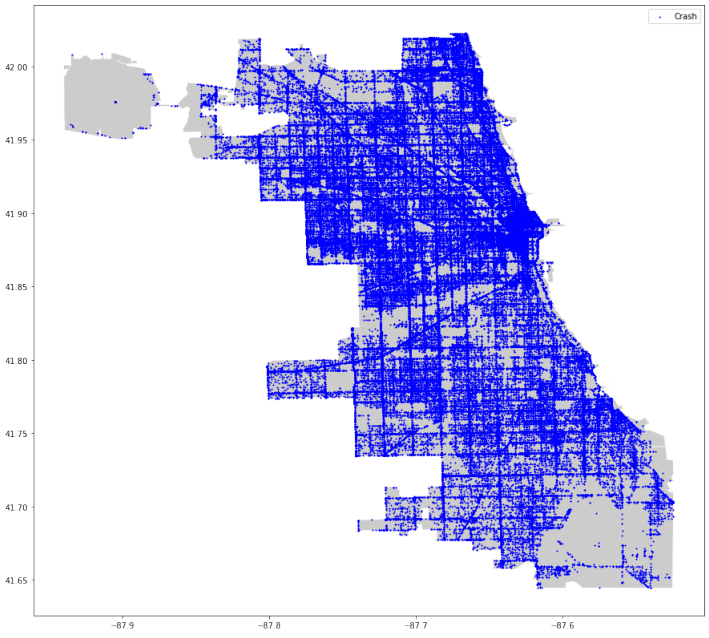


FIGURE 3

IX. Feature Engineering and Discussion

In this Chicago Traffic Crashes data, we want to use environmental, crash characterization, weather, etc. factors to predict possible damage cost for one single specific car crash.

For feature engineering, we removed seven explanatory variables from the data. These variables are either serial numbers or can be represented by other variables. Furthermore, we generate a new variable 'DAMAGE_OVER_1500' from the variable 'DAMAGE' to be our response variable. Since the new response variable is a binary variable, we will mainly

focus on classification models to deal with the goal.

For data mining task, the first model we are going to use is the ordinal logistic regression model, this will be our base model to determine if we can get a better F-1 score and Accuracy score. Second, we will apply random forest model to classify the data. Third, we will use ANN model as the final trial model for classifying the data.

In addition, the input of our model will be the 29 explanatory variables and the output is the Damage_over_1500 with two levels, one is below 1500 and the other is over 1500.

X. Prediction and Performance Analysis

Our task is to predict whether the damage would cost over 1500 or not. It is a Binary classification. We use the algorithm of majority vote to solve this problem and we choose four models to implement it. The first model we choose for the majority vote are Random Forest Classifier, Logistic Regression Classifier and AdaBoost Classifier. The reason why we chose Random Forest Classifier model is because it works well with both categorical and numerical columns. In our dataset, we have several of both categorical and numerical data, therefore we choose the Random Forest Classifier. Why we chose Logistic Regression Classifier, since It can easily extend to multiple classes(multinomial regression) and a natural probabilistic view of class predictions. Furthermore, the reason why we chose AdaBoost Classifier is because it is easier to use with less need for tweaking parameters. The outcome(Figure i) shows that it has the best accuracy and f1-score in contrast to the individual three models above, therefore, we chose the majority vote algorithm as our method.

With over 450k data points, we could identify and remove outliers or missing values more easily. Moreover, we were able to have a clearer view of the underlying data distribution. This allows the "data to tell for itself," instead of relying on assumptions and weak correlations. Presence of more data results in better and accurate models. In addition, as we trained these data on each classifier, we used grid-search and cross-validation, which allows us to run the classifier over a grid of hyperparameters and find the optimal result. While accuracy looks at correctly classified observations both positive and negative, the F1 score is balancing precision and recall on the positive class. Therefore, we used F1 score to evaluate the performance when we tuned hyperparameters. Apparently, the optimal hyperparameters turned out to provide us with robust model performance. This can be proved from the testing results (Figure i). While the labels of the testing dataset would not be seen, it follows the same probability distribution as the training dataset. As we tested our model on the testing dataset, we can find it performed equally effective and reached similar accuracy and F-1 score. Based on this situation, it is believed that the prediction outcomes are robust.

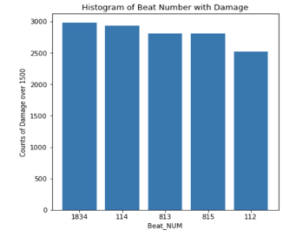
When it comes to generalizing our prediction outcomes to the future, it would be challenging since we will need to test the model on tabular datasets that have identical features. However, although many local governments save car crash datasets in their databases, they keep the data with different features. As a result, if we apply our model on other car crash datasets, we could hardly expect a good performance as long as the data layout is different from the data used in this project.

Based on the outcome of the model, we use features importance (Figure c) from the random forest model and extract out four significant features for finding out useful insights. First, the relative highest importance of the feature is Beat, which is the territory that a police officer patrol. We found out that there are some areas in Chicago where serious damage often happens. Such as the areas of Chicago Police Department Beat Number:1834, 114, 813 and 815.(Figure a) The insurance company might increase the fee for those who drive through or usually driving within these beat areas. The second insight is time-related features. We found out that the data shows higher probability that car accidents with damage over 1500 happen at 4a.m(Figure d). However, most cases happen in daytime, when the hour is between 7 to 19. This is interesting since even more cases happened in daytime, drivers were more likely to suffer from higher damage if they involve in a car accident. Furthermore, according to the weekdays, we can see that serious accidents are more likely to occur on Sunday and Monday while the weekends have more car accident cases recorded(Figure f and h). Take months into consideration, we can also find out that the probability of accidents rises in winter(Figure e). This is grounded since the road condition may be worse due to heavy snowing and cause roads in Chicago harder to drive as usual and then more dangerous than other seasons. As an insurance company, we hope fewer accidents happen, therefore, we can make propaganda to the drivers and also give some advice to the police department that they could deploy more policemen on the road that often causes accidents or implement some regulations in particular time periods in order to prevent serious car accidents. With these insights, we believe that the government or other agencies whose interests are involved in this related problem can pay more attention to specific time and regions for decreasing the rate of serious car accidents. In this way, the overall social benefits will increase and help people keep away from paying for unnecessary costs of damage over people and properties. Insurance company can also have a way for accessing the risk of one driver, they can focus on collecting data on these important features. If people have these precautionary knowledge in mind, then the accidents will surely decrease.

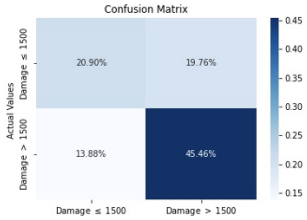
XI. Conclusion and Future Work

The predictability of crash from machine learning models exhibits considerable variation in pattern and magnitude. Results presented in this report demonstrated that the magnitude in crash damage depend not only on the territory that a police officer patrol but also on several time-related features. Knowledge-based crash damage prediction has always been the main focus of insurance companies. As technology is advancing and computing has started becoming more efficient, machine learning-based models for crash damage severity are being brought into the light for more accurate prediction models. To explore different machine learning methodologies

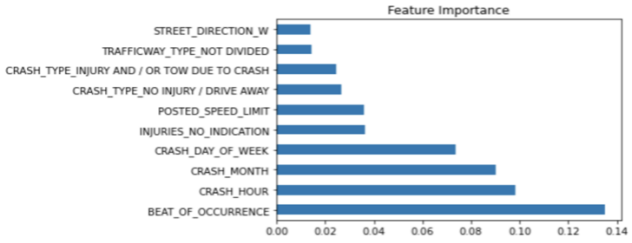
for crash damage prediction, this report considered a combination of three applications—Logistic Regression, Adaboost and Random Forest, which are blended in order to construct a predictive model based on majority vote algorithm. We not only achieved up to 73.75% of F-1 score with the majority vote model in this report, but also sought to study more advanced methodology to obtain a higher prediction performance. Various analyses of prediction probabilities provided useful insights into the models' capability to generate predictions that are consistent with observed data. A feature importance approach confirms that manner of police officer patrol and time conditions are major features that can feed useful information in the modeling estimation process. Further exploration can be useful to compare more machine learning modeling approaches that might have better performance and efficiency. It might also be reasonable to compare their results with statistical approaches to compile a comprehensive comparison, so the stakeholder of insurance companies can make knowledgeable decisions for their operational strategies.



(a)



(b)



(c)

	Hour	Probability
0	4	0.744921
1	2	0.743294
2	3	0.742805
3	1	0.711001
4	0	0.706323

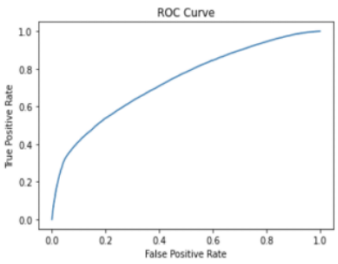
(d)

	Month	Probability
0	11	0.606701
1	1	0.604993
2	12	0.601278
3	10	0.596526
4	9	0.592341

(e)

	Weekdays	Probability
0	1	0.631207
1	7	0.610864
2	6	0.588900
3	4	0.584110
4	2	0.583638

(f)



(g)

	Weekdays	Crashes_Cases
0	6	75062
1	7	67362
2	5	66404
3	3	66273
4	4	65842
5	2	63930
6	1	54982

(h)

	Logistic Regression	Adaboost	Random Forest	Majority Voting
F-1 Score	70.88%	71.72%	74.18%	73.75%
Accuracy	64.17%	64.42%	67.74%	67.25%

(i)

REFERENCES

[1] "Traffic crash data." [Online]. Available: https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if?fbclid=IwAR0RDZUwrB98m__p6-DI7QCXzhPx8nkN3wE3xC_bwZP4XZscg2WteyWaWMI

[2] "Chicago data portal." [Online]. Available: https://data.cityofchicago.org

[3] "Chicago data portal terms of use." [Online]. Available: https://www.chicago.gov/city/en/narr/foia/data_disclaimer.html

[4] A. R. Mondal, M. A. E. Bhuiyan, and F. Yang, "Advancement of weather-related crash prediction model using nonparametric machine learning algorithms," *SN Applied Sciences 2020*. [Online]. Available: https://link.springer.com/content/pdf/10.1007/s42452-020-03196-x.pdf

[5] S. Wang and Z. L. Li, "Exploring the mechanism of crashes with automated vehicles using statistical modeling approaches," *PLOS ONE*. [Online]. Available: https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0214550&type=printable

[6] C. Timmermans, W. Alhajyaseen, A. A. Mamun, T. Wakjira, M. Qasem, M. Almallah, and H. Younis, "Exploring the mechanism of crashes with automated vehicles using statistical modeling approaches," *International Journal of Injury Control and Safety Promotion*. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/17457300.2019.1620289