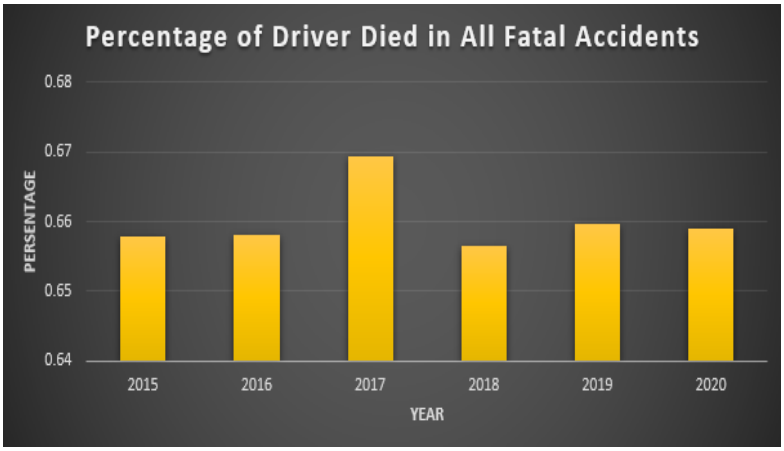
Fatal Accident Driver Causation Prediction Model

Weilan Chen
Ziwen Wang
Kangbo Shi

Objectives

The most important goal of analyzing fatal accident reports is to reduce the chance people die from car accidents. This requires several domains of people to work together. Traffic Control, city planning, police deployment, and civil engineers who work on infrastructures. What was the causation of the accident? What is the driver actually doing when the accident happens? This part of data is often missing because around 65% of drivers die during the accident, these questions remain unanswered.



This project will help any agencies/individuals use other data collected after the accident to predict the factor from which the driver contributed to the fatal accident from external conditions.

Introduction

In modern days, local roads and highways were constructed wherever we could find human activity. As car transportation became prevalent in the last century, investigation of fatal accidents became more and more complicated and more costly. Here is an overview of how many people died per month in fatal accident:

In the year 2020, National Highway Traffic Safety Administration (NHTSA) has added a variable in FARS report named “DRIVERRF”. This variable contains factors related to the drivers of motor vehicles in transport involved in the crash based on a list of driver conditions, unusual situations, and special circumstances. Due to around 65% of the drivers dead from all fatal accidents, it is difficult to investigate the factor of causation of the accident related to the driver. We are curious about what causes a fatal accident. There are both internal issues, such as DUI and fatigue driving, as well as external issues like severe weather and vehicle malfunction. Thus, we are building models to analyze which causations are more likely to cause fatal accidents. We want to use the Fatal accident data from the past 40 years to find out the common features of fatal accidents. Making predictions on the cause of the accident. This project can be used by the government for fatal accident prevention and insurance companies for risk evaluation.

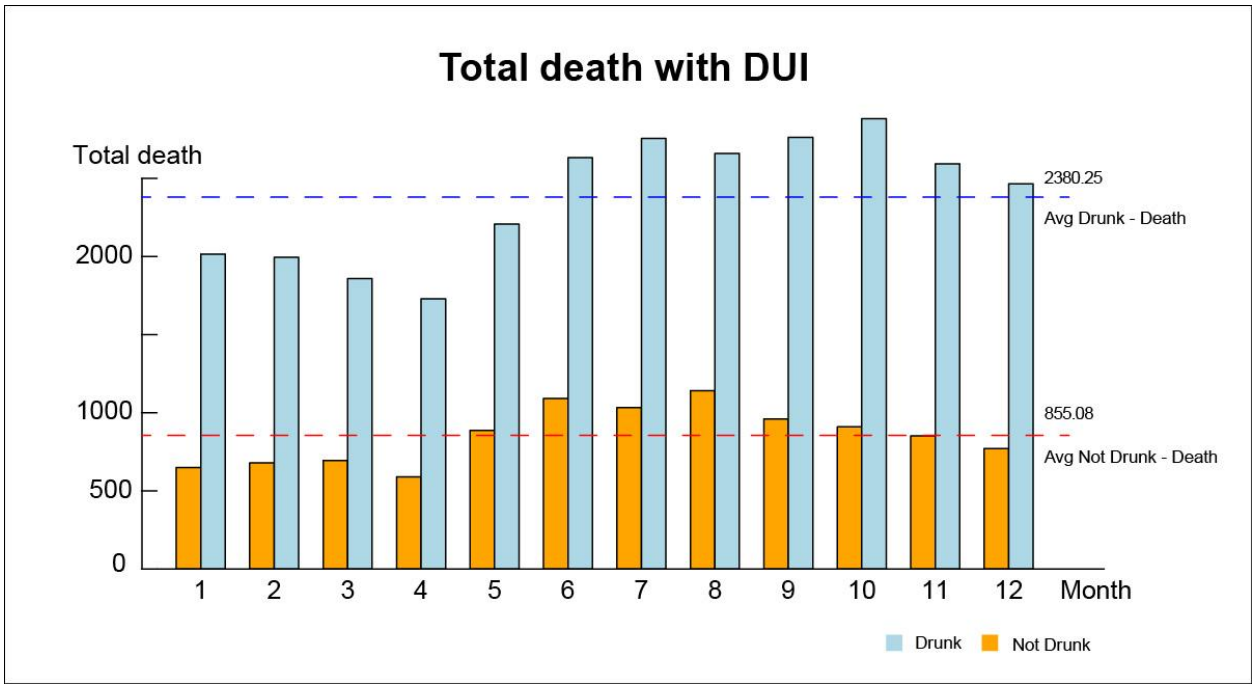
Dataset

We have public data from the U.S. government. The following data are obtained from the official website of NHTSA. This is the link:

https://www.nhtsa.gov/file-downloads?p=nhtsa/downloads/FARS/

In this project we acquire the public FARS data from NHTSA, where FARS stands for Fatality Analysis Reporting System. The data set contains information about all fatal accidents in the

U.S. from 1975-2020. The size of the data set is about 35,000 observations, and 93 variables. A lot of variables are a factor in fatal accidents. We have a data dictionary which describes each variable. The data set contains variables for accidents such as people who died in each accident, weather, DUI driver involved, etc. Here is an example showing DUI rate in total death:



FARS data is using relational tables with the primary key of "ST_CASE" for the overall database. This is the case number of each unique accident. There is one important subclass for each "ST_CASE": "VEH_NO", which means the vehicle number of all involved vehicles in that case. We select conditions from three tables in the database: accident.csv, vehicle.csv, and person.csv. These three tables contain plenty of variables that could determine the factor of the fatal accident. We then selected all variables that could be a causation or relate to causation to the fatal accident.

Method

After we collected the data set described above, we combined the accidents, vehicles and drivers information together from three tables as input, and vehicle number is the reference key for each observation. The driver-related factors of accidents investigated by police (a multi-category variable) is our output. The input variables are selected by us because it is easy to distinguish

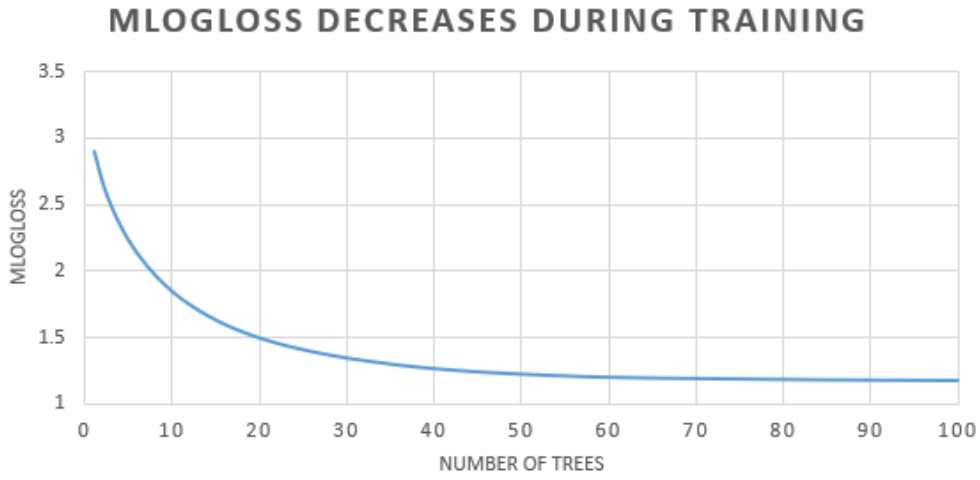
whether a variable is related to the causation of the accident. For missing value, if it's a categorical variable, we treat it as a specific category. If it's a numerical variable, we replace it with the average value of all data. For outliers, we just kept them since we want our model to be able to handle rare conditions. We also encoded all categorical variables for further training. Considering that the variable of driverrf has been only added by NHTSA in the year 2020, information of driver-related factors will be recorded in the reporting system. There is only one year of training data available but there will be much more in future. For this case we randomly choose 20 percent of the data as our testing data and the rest 80 percent observations are our training data.

The methods we used for data processing are numpy and pandas packages of python. We run our program on a jupyter notebook so that it's easy to finetune and show the results.

For the method of data visualization, we use the python library matplotlib and seaborn. We use these packages to draw histograms and math function graphs to display results we have.

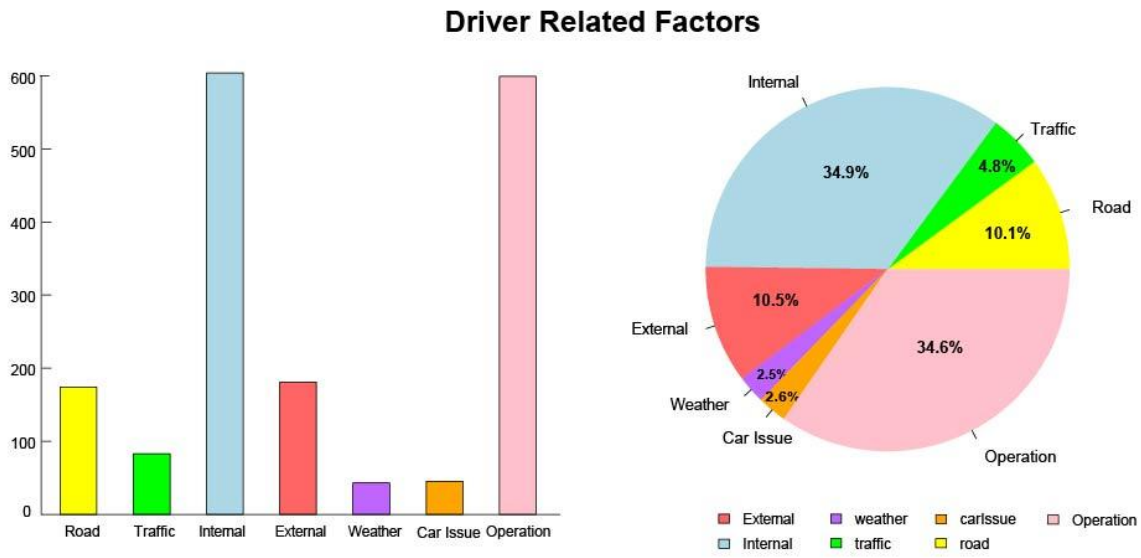
Model

We used a built-in package of decision tree models for this classification task: XGBoost. Decision tree becomes our first choice because almost all of our input variables are categorical, and the decision tree has the natural advantage in dealing with categories. To build this model, we used the xgboost-gpu package of python, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. We finally build a 100-tree boost model and here is how our mlogloss changed during training:



First Stage Result

The test accuracy of our model is 64.7%. Apparently, this is not an ideal result of our model. We will add variable weight and use cross validation for implementing our xgboost model. We did statistics about the known value of driver_rf. We can see that the two major reasons for fatal crashes are internal factors of the driver or improper operation from the driver. Here are two graphs showing the major factors caused fatal accidents (only from the given value of driver_rf 2020).



Improvement of the Model

64.7% accuracy is a very high accuracy model considering that we are predicting a variable that contains 97 values. However, this accuracy can not give us a persuasive conclusion to help city planning or the traffic control department to make any decision. In order to make this model more accurate and applicable, we combined these values into 7 general classes: 1. Road Condition 2. Traffic 3. Personal External 4. Personal Internal 5.Weather 6. Vehicle 7. Improper operation.

We consider a set of given variables (x1, x2,..., xi) and predicting variable y where contains value

yi. In this specific model, we have i $\in$ {1,2,3,4,5,6,7} after the selection of driver rf value. We then construct our optimization model as:

Objective Function = Training Loss + Regularization.

Training Loss:

1. MSE: $L(\theta) = \sum_i (y_i - \hat{y}_i)^2$
2. Logistic Loss: $L(\theta) = \sum_i [y_i \ln(1 + e^{-\hat{y}_i}) + (1 - y_i) \ln(1 + e^{\hat{y}_i})]$

The two ways that we calculate training loss in the last model is using logistic loss for multivalued logistic regression. In fact, neither mean squared error or logistic is a good fit for multi-value prediction. Thus, we modified the logistic regression to suit supervised learning on tree structure:

Modified Logistic Loss:

Obj is the objective function and f_i is the learning function. l is the training loss and $\hat{y_i}^{(t)}$ is the predicted value at step t.

$$obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \omega(f_i)$$

$$\text{obj}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \omega(f_i) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \omega(f_t) + \text{constant}$$

Using Mean Square Error as our loss function, we get:

$$\text{obj}^{(t)} = \sum_{i=1}^{n} (y_i - (\hat{y}_i^{(t-1)} + f_t(x_i)))^2 + \sum_{i=1}^{t} \omega(f_i) = \sum_{i=1}^{n} [2(\hat{y}_i^{(t-1)} - y_i) f_t(x_i) + f_t(x_i)^2] + \omega(f_t) + \text{constant}$$

To get a better approximation of the loss function and make it applicable for computation, we take the second order Taylor expansion of loss function:

$\text{obj}^{(t)} = \sum_{i=1}^{n} [l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \omega(f_t) + \text{constant}$

where h_i and g_i are:

$g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$

$h_i = \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$

The final objective function of this algorithm becomes:

$\sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \omega(f_t)$

Second Stage Result Using Improved Model

We have successfully implemented our program according to the method and model provided above. All Cases are processed with selected variables and defined prediction classes.

Here is the final test accuracy of our model:

```
[98]    validation_0-mlogloss:0.69694
[99]    validation_0-mlogloss:0.69701
test accuracy:  0.8006403318903319

Process finished with exit code 0
```

Here is the confusion matrix of the result of prediction:

real class\prediction	class1	class2	class3	class4	class5	class6	class7
class1	72	0	1	4	0	0	81
class2	2	24	5	6	0	0	72
class3	3	4	105	17	0	0	132
class4	6	1	12	175	0	0	431
class5	2	0	0	0	0	0	14
class6	1	0	2	1	0	12	39
class7	28	6	53	109	0	3	3736

This result can be applied to predicting driver_rf when the driver has passed away and helps the fatal accident investigation. It also indicates that most fatal accidents are caused by driver related factors. Thus. The result can be used to make new policies about traffic rules, or help insurance companies to do price evaluation. It can also be used for city planning or road designing.

Observation

In conclusion, the Fatal Accident Driver Causation Prediction Model developed by us aims to analyze fatal accident reports and predict the factor from which the driver contributed to the fatal accident from external conditions. By using public data from the US government's FARS dataset, the team selected variables related to the causation of the accident and used the XGBoost package for classification. The model achieved a test accuracy of 64.7%. With the improved decision tree model with selected classes, the accuracy has been significantly improved. The results of this project could be useful for government agencies and insurance companies in preventing fatal accidents and evaluating risks.

Conclusion

Overall, this project demonstrates the potential of machine learning in analyzing complex datasets to extract valuable insights and make predictions that could have real-world impacts. Further research and development could lead to more accurate and effective models for preventing fatal accidents and improving road safety.

Future Plan

For the current stage, we successfully implemented our XGBoost algorithm aimed to find the driver related factors. In the future, we will try to build a similar algorithm which will predict the vehicle related factors. Thus, we are able to establish a more comprehensive prediction. In fact,

during this project, the data of the year 2021 was updated to the database on April 1, 2023. We can use data from 2021 to test our model to see if the model fits the data set in general. We will use the same model to enhance the data set collected in future to reduce missing values for data entry. Since our database will keep grouping, it will become a huge database that our algorithm may experience long run time. In this case, we will try horizontal scaling in order to reduce the process time. Besides the data from the United States, we will collect accident data from all the countries and build models to fit their traffic conditions. Therefore, these operations will help reduce the overall fatal accident appearances.

Work division

For this project, we will divide the total workload into three parts: the overall data processing with visualization, building clustering models, and establishing prediction models. Then for the final report, we will finish together. Besides, for each part of the codes and models, we will do a final overview and check at the end.

Kangbo Shi - Overall data processing

Weilan Chen - Building classification models.

Ziwen Wang - graphing and visualization.

All of the team members - final code check and writing the final report.

Attached Files

There are two attached files associated with this report. The first file is selected_factors.pdf. This file contains which variables we have selected for predicting driver_rf. These variables are selected based on the fact that it is directly related to the driver related cause of the accident.

Another attached file is selected_classes.pdf. This file contains classes as the result of a classification model. This file has information about how we have combined the 97 types of values in driver_rf into 8 classes.

References

C.M. Farmer, Apr 2017, *A Projection of United States Traffic Fatality Counts in 2024* Retrieved by A projection of U.S. traffic fatality counts in 2024.docx (iihs.org)

NHTSA Media, Aug 2022, *NHTSA Early Estimates Show Record Increase in Fatalities Nationwide.* Retrieved by NHTSA Early Estimates Show Record Increase in Fatalities | NHTSA

XGBoost Documentary, Retrieved from: https://xgboost.readthedocs.io/en/stable/tutorials/model.html

NHTSA Data Download: https://www.nhtsa.gov/file-downloads?p=nhtsa/downloads/FARS/