ORIE 5741 Project Midterm Report

Sharon Liu, Jingxin Mao

10/31/2021

1 Introduction

Car accidents are common and deadly in modern society. Internationally, road traffic crashes cause the death of around 1.3 million people each year, and cost most countries 3% of their GDP. What are the causes of car accidents? Who are the people that get injured most likely or most severely in the accident? A study on these topics can narrow down the actions to be taken to minimize the risk of future accidents, and can help governments and industry leaders implement new policies or standards, which can save people's lives and increase road safety.

In this project, we will explore this further by answering 'Can you predict the injury severity based on road, user and vehicle information?', which can suggest the protection actions on reducing severe accidents.

The dataset of choice is the 2019 database of road traffic injuries of France, which will be used to identify the key factors that will impact the injury severity, with the hope of reducing future injury severity. The dataset contains over 50 features on general circumstances of the accident, the places where the accidents happened, and the drivers and vehicles involved, for more than 58,000 accidents happened in France in 2019.

2 Data Preprocessing

The data we have are in 4 separate tables, with 59 columns in total, 58,840 accidents, 132,977 users and 100,710 vehicles involved. To clean up each of the tables, we first fixed the 127 accidents with missing values caused by failure in column split in the first dataset (about characteristics of the accidents), then removed 25 columns with irrelevant information, or more than 50% of the data are missing, or more than 80% of that data are labelled under one category, which left us with 34 columns in total for 4 separate tables.

After that, we merged 4 tables into one master sheet with around 133K rows (each row is a unique user) and 24 columns. The dependent variable we want to predict is 'injury_level' which labels the injury severity of each individual. Injury level ranges from 1 to 4, each represents 'Unharmed', 'Killed', 'Injured hospitalized' and 'Slightly injured'.

Then we applied feature engineering on other columns to prepare the date for preliminary model. We first changed all categorical variables into one hot variables. For example, for the security equipment column, we have the following encoding methods:

Table 1: Coding for Safety Equipment

0: No equipment	1: Belt
2: Helmet	3: Children's device
4: reflective vest	5: Airbag (2WD / 3WD)
6: Gloves (2WD / 3WD)	7: Gloves + Airbag (2WD / 3WD)
8: Not determinable	9: Not specified

We changed the safety equipment column into 10 one-hot variables. Each of the 10 new features represent one safety equipment. If the user is wearing that equipment when the accident occurs, the value of its corresponding one-hot variable will be 1; otherwise it will be 0. We then removed the one-hot variable for 'Not Specified' because if all other one-hot variables are with value 0, it is not specified.

We applied this method to other categorical variables, including light, intersection, weather, movable barrier, travel reason, road category, traffic regime, place, user category, initial shock, and maneuver, as shown in Table

2.

Table 2: Feature Summary from the Merged Dataset

feature	type	encoding	feature	type	encoding
light	ordinal	one-hot	road category	categorical	one-hot
agglomeration	categorical	one-hot	traffic regime	categorical	one-hot
age	numerical	NA	safety equipment1	categorical	one-hot
intersection	categorical	one-hot	traffic lanes	numerical	NA
weather	categorical	one-hot	Max Speed	numerical	NA
safety equipment2	categorical	one-hot	safety equipment3	categorical	one-hot
latitude	numerical	NA	place	categorical	one-hot
longitude	numerical	NA	user category	categorical	one-hot
movable barrier	categorical	one-hot	initial shock	categorical	one-hot
hour	numerical	NA	injury level	ordinal	NA^a
holiday	categorical	one-hot	sex	categorical	one-hot
travel reason	categorical	one-hot	maneuver	categorical	one-hot

[a] : The feature is used in prediction

After having all categorical variables transferred to one-hot variables, we ended up with 164 features. In order to reduce the amount of features applied into the preliminary models to avoid overfitting, we plotted a correlation matrix in the form of heatmap (as shown in Figure 1), to show the correlations between features, and dropped the ones with correlation coefficient more than 0.7. However, this action only removed 4 features, and we will try more feature selection methods to narrow down the key ones as one of the next steps.

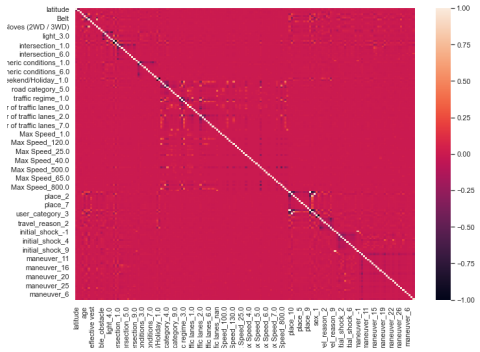


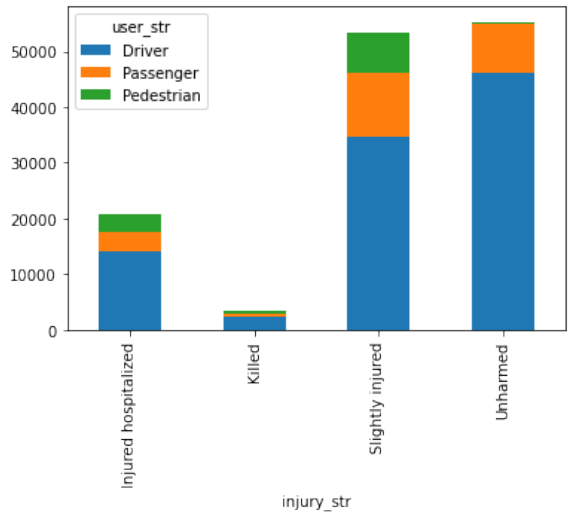
Figure 1: Correlation Heatmap

3 Exploratory Data Analysis

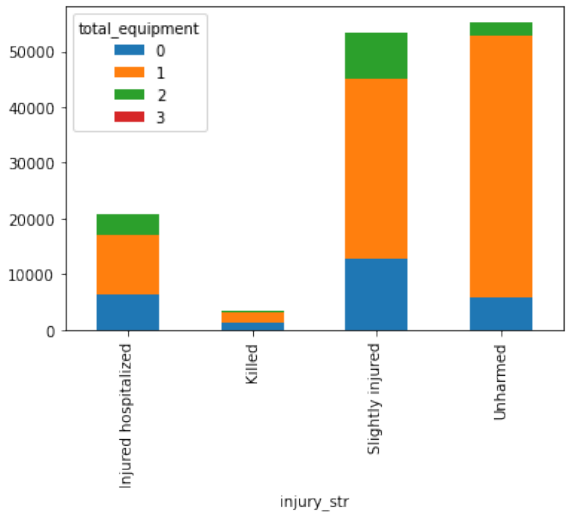
With the merged table, we were able to perform some exploratory data analysis for the accident and injury data. On average, France had 1132 accident per week in 2019, with 67 people killed, 401 people hospitalised, 1025 people slightly injured and 1064 people not injured. Average accidents happened during weekdays is 17% higher than the ones happened during weekends or holidays (169 and 144 accidents per day respectively), with Friday being the day with more accidents, and Sunday with fewer accidents to happen, and 16pm to 20pm being the peak accident time disregard the day of week.

From the injury analysis, we found over 50,000 people experienced accidents in 2019 were are not harmed at all (injury level = 1) and 50,000 were slightly harmed (injury level = 4). Roughly over 20,000 people were injured hospitalized (injury level = 3), and less than 5,000 people were killed (injury level = 2) as shown in Fig2.1. Overall, most people were not harmed or slightly harmed. We also noticed that in every injury level, driver has the most percentage at 70%. Only about 10% to 20% are passengers. Pedestrians has the smallest percentage at around 5% to 10%. Pedestrians occur most frequently in injury level 3 and injury level 4, meaning that most pedestrians involved in traffic accidents were either hospitalized or slightly harmed. Among those who were killed in accidents, almost 90% were drivers. Driver is the main user category in car accidents and they are at extremely high percentages in deadly or heavily injured cases. Using this information, we can explore deeper into the driving causes and impact of drivers' accidents.

We also plotted the number of safety equipment used in each injury level, as shown in Fig 2.2. We saw at least 80% of people involved in accidents have one or more safety equipment. Among the people having safety equipment, over 90% has only 1 safety equipment. 10% has 2 safety equipment; and roughly 1% wears 3 safety equipment. In fact, at injury level 1 when no body gets hurt, only about 10% of people didn't have any safety equipment, but at level 2 when people were killed in the accident, having 0 safety equipment is at 40%. We can see that as injury gets more severe, the percentage of people with 0 safety equipment also gets larger. Wearing a safety equipment indeed helps reduce the severity of injury.



(a) Fig2.1



(b) Fig2.2

Figure 2: Injury Level Analysis

4 Preliminary Model

For the preliminary model, we ran the data on tree classifiers with different ensemble methods - bagging, random forest and gradient boosting, with the hope of selecting the ensemble methods that works best on out data.

We started from splitting the dataset into training and testing sets, with a ratio of 8:2, then apply 3-fold cross validation to average the performance. For all the 3 models, we picked 100 base estimators in the ensemble, and generated the cross validation scores at 0.62, 0.64 and 0.64 for bagging, random forest and gradient boosting ensemble respectively. Since the training for bagging is slightly lower than the other two method, we decided to drop the bagging method and look at the testing error for the random forest and gradient boosting, which are both 0.65, which is close to the training score, so these 2 models are neither overfitted nor underfitted. To show a snapshot for one of the tree in the forest, we plotted a simplified version of decision trees, which limited max depth to be 5, as shown in Figure 3.

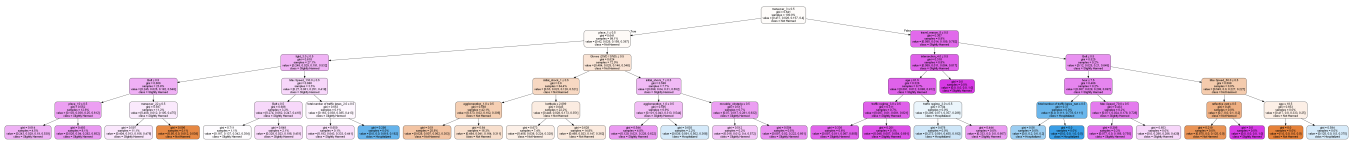


Figure 3: Simplified Sample Tree

5 Future Steps

At current stage, even though the classification tree with random forest and gradient boosting ensemble methods are tie on performance, their performance are not good enough for us to conclude that they are providing good prediction. Therefore, moving forward, we are planning to identify and change to the best parameter for these models, for example the learning rate for gradient boosting, to see whether the performances can be differentiated. In additional, we plan to apply additional classification models like Logistic Regression, to determine the best model that fits our data.

Furthermore, we would like to improve the model accuracy, which consists of 2 parts. The first one is to apply more detailed feature engineering, for example, to normalize numerical data like age, even though they are not significantly large in our dataset, but we would like to see whether the normalization can improve the model accuracy. Secondly, we would like to enhance the feature selection procedure. Currently, we dropped the features that are highly correlated (correlation coefficient > 0.7), but only got rid of 4 features. We would like to apply more feature selection techniques to reduce the number of features feed into the models.