ORIE 5741 Project Midterm Report

Sharon Liu, Jingxin Mao

12/05/2021

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? For most car businesses, they would like to know the answers to these questions; so they can design effective safety equipment and warning systems in their vehicles. A study on these topics can narrow down the actions to be taken to minimize the risk of future accidents, and can also 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 some protective 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 12 columns with irrelevant information, or more than 50% of the data are missing, which left us with 47 columns in total for 4 separate tables. Next, we filled in the missing values in the remaining features. The cause of the missing value was by the definition of the feature - there is one feature labelling number of passengers on the public transportation, but if all vehicles involved are private ones, then the data is missing, so we filled in 0 for the missing information.

After that, we merged 4 tables into one master sheet with around 133K rows (each row is a unique user) and 38 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 classification models. 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. After having all categorical variables transferred to one-hot variables, we ended up with 214 features.

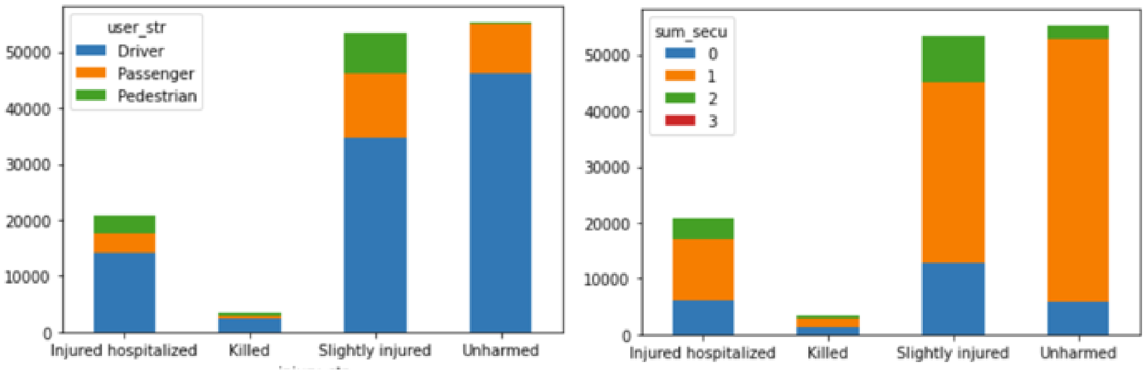
Table 2: Feature Summary from the Merged Dataset

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
collision type	categorical	one-hot	plan layout	categorical	one-hot
surface condition	categorical	one-hot	situation of the accident	categorical	one-hot
motor	categorical	one-hot	number of safety equipment	numerical	NA
Passengers on Public Transportation	numerical	NA			

[a] : The feature is used in prediction

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 disregarding the day of week.



(a) Fig1.1

(b) Fig1.2

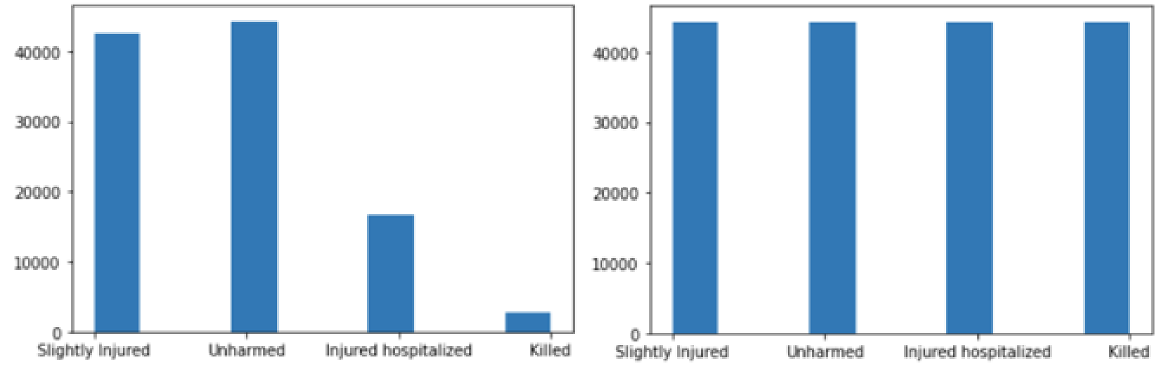
Figure 1: Injury Level Analysis

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 Fig1.1. 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 1.2. 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 where nobody 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.

4 Model Selection

Since we are interested in the injury level of the people involved in the accidents, we decided to use classification models, and we splitted out data into training and testing dataset at a ratio at 8:2. However, one of the observations from the previous Exploratory Data Analysis section is that, our data is highly imbalanced (as shown in Fig2.1): more than 80% of the people involved in accidents are either slightly injured or not injured, whereas 16% of the people are hospitalized and 3% are killed, so if we fit the model on the original imbalanced data, the minority class (eg, killed category) could be completely neglected, but still results in a model with good accuracy. Therefore, in order to prevent this neglect from happening, we decided to rebalance the training dataset with oversampling, which is basically randomly selected samples from the minority class (eg, killed category), with replacement, and added them back into the training dataset, until the number of data points in the four classes are the same (as shwon in Fig2.2). We decided to use oversampling instead of undersampling because the total number of data points after undersampling is around 10% of original dataset, which is missing a lot of information. After this oversampling process, we ended up with a rebalanced training dataset, which was used to train the following 2 classification models.



(a) Fig2.1

(b) Fig2.2

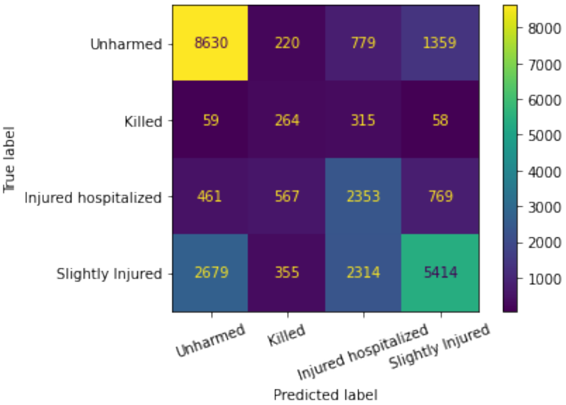
Figure 2: Imbalanced vs Rebalanced Data

4.1 Random Forest

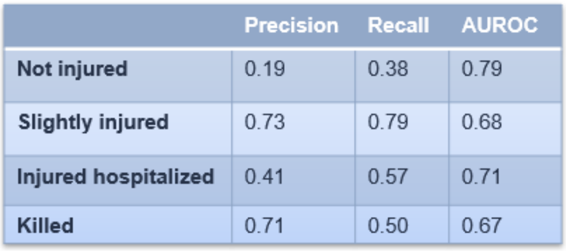
The first model we applied was a tree based model, we have tried all 3 ensemble methods – bagging, random forest and gradient boosting, and found that random forest overperformed the other two, so we decided to use random forest as our final tree based model. We fitted the rebalanced training dataset to the random tree, and used 3-fold cross-validation method to prevent overfitting (controlled the hyperparameters like max_depth and max_features) and apply feature selection (we ranked the features by importance, and fitted the models with different number of most important features, then chose the one with the best accuracy while not overfitted).

The final model we ended up with has the training set accuracy at 0.79 and validation set accuracy at 0.75. We then applied this model on the original imbalanced test set, to measure how good this model will perform on the original real data, and received an accuracy score at 0.63. The test score is 0.12 lower than the validation accuracy, but it is within our expectations, because the oversampling changed the nature of the data, so we expected the testing score to be lower than the training and validation score, and 0.12 is an acceptable decrease.

Since this is a classification model, we are also interested in other measurements for model evaluation, they are precision, recall and AUROC (Area Under The Receiver Operator Characteristics Curve). Precision measures of all people that are labeled as each class, like killed, how many of them were actually in that class. Recall measures of all the people under each class, like killed, how many of them did we label. AUROC measures the ability of a classifier to distinguish between classes. For all of these 3 measurement, the closer to 1, the better the model is. The results are shown in the table 3.2 below. The precision and recall for not injured class, and the precision for injured hospitalized class are low (below 0.5), which means that not injured and hospitalized are the classes that contribute more to the total test error, but both precision and recall for the rest of classes are at least 0.5 or higher. And the AUROC for all of the 4 classes are either around 0.7 or above, which means that the model didn't neglect any of the classes, and we can trust its prediction on the slightly injured, injured hospitalized and killed category; but with some hesitation on the not injured category.



(a) Fig3.1



	Precision	Recall	AUROC
Not injured	0.19	0.38	0.79
Slightly injured	0.73	0.79	0.68
Injured hospitalized	0.41	0.57	0.71
Killed	0.71	0.50	0.67

(b) Fig3.2

Figure 3: Imbalanced vs Rebalanced Data

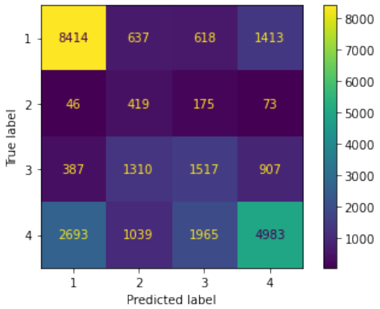
4.2 Multi-class Logistic Regression

Next, we tried a regression model to predict injury levels. Because the dependent variable: injury level has only four unique values, we chose to build a multi-class logistic regression model on the re-sampling data. The other columns in our re-sampling data were used as features in this model.

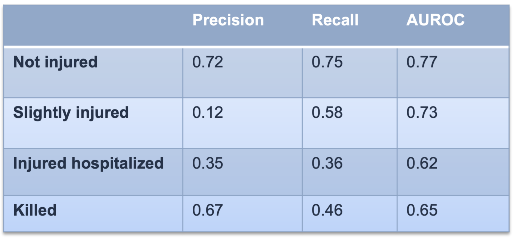
Since logistic regression is a linear model, we also need to deal with multicollinearity among the features. Most of the features in the original data have been one-hotted; so these columns were highly correlated. To prevent multicollinearity, we removed one feature from all the highly-correlated variables. For example, originally we had two boolean variables: Male and Female, and the sum of the two columns is 1 for every row. Thus, they are highly correlated, and we removed the female column to prevent multicollinearity issues in the model. We also removed one variable from other one-hotted variables such as light, agglomeration, intersection, and etc. Another modification we did is to normalize the location and hour columns. This is because latitude and longitude data ranges from -180 to 180, and the variable hour ranges from 0 to 24, whereas other features are boolean variables.

After fitting the model on the re-sampling training set, the model gives a 56.2% accuracy on training and validation accuracy, and the test accuracy is 57.7%. Since training and test accuracy are very close, there's no over fitting issue in this model. However, roughly 50% accuracy is still low compared with the random forest model.

To see how the model performs on test data, we also calculated three metrics: precision, recall and area under ROC curve, shown in Fig 4.2. The AUROC value is similar for all four injury levels, meaning that none of the four levels is neglected in the prediction. However, if we look at precision and recall metrics, the results for slightly injured and injured hospitalized are both below 0.5, meaning that the models' prediction on these two categories are mostly incorrect when compared to test values. The plot of confusion matrix on Fig 4.1 also proves this point; the model prediction on not injured and killed type are mostly correct, and the prediction on slightly injured is the least accurate. Overall, the model gives a good performance on the not injured type, but it does not work well on other types compared with the random forest model. So we will use random forest as our final chosen model.



(a) Fig4.1



	Precision	Recall	AUROC
Not injured	0.72	0.75	0.77
Slightly injured	0.12	0.58	0.73
Injured hospitalized	0.35	0.36	0.62
Killed	0.67	0.46	0.65

(b) Fig4.2

Figure 4: Imbalanced vs Rebalanced Data

5 Fairness and Weapon of Math Destruction (WMD)

In our model, the only protected feature we have is gender. To see if our model has fairness issue, we tested the false positive rates and model performance on the two gender groups to see if the results are similar.

Table 3: False Positive for male and female group

Gender	Not Injured	Slightly Injured	Injured Hospitalized	Killed
Male (RF)	22.7%	5.1%	14.8%	9.8%
Female (RF)	16.2%	3.0%	15.9%	23.8%
Male (LR)	22.1%	14.2%	11.1%	9.9%
Female (LR)	16.3%	5.8%	14.6%	27.8%

From table 3, the false positive rates are similar, with a 3%-15% difference across two gender groups in both models. In general, error rates for the two models are close to each other: for some levels the female group has higher error rates, and for other levels the male group has higher error. There's no clear pattern on which gender group the model performs better, and therefore we can say the two models does not fail the fairness test.

Another type of fairness metrics address whether model performs similarly across two gender groups. Below I show the accuracy scores of two models:

Table 4: Accuracy for two gender groups

Gender	Random Forest	Logistic Regression
Male	63.6%	58.2%
Female	60.6%	56.4%

Accuracy is similar across the two gender groups with a 2% difference, which is acceptable. There's no clear pattern on which gender group the model performs better. However, between the two models, random forest has better accuracy results on both genders, proving that it is a stronger model compared to the logistic regression model.

Our model does not create a Weapon of Math Destruction (WMD) because of the 3 reasons below. Firstly, the outcomes are not hard to measure. The dataset has clear outcomes – the 4 levels of injuries, and for any future data (people involved in future traffic accidents), the injury level will be measured by hospital or health care provider, which will be clear measurements. Secondly, the predictions don't harm people. If the result shows that a certain group of people are more likely to be injured severely, for example, old people are more likely to be injured more severely than young people, people would pay extra attention when aged people are on their transportation or on the road to avoid injuring them, which will actually harm less people. Thirdly, the prediction doesn't create a defeating feedback loop. Similar reasons as above, if a group of people are predicted to be more likely to get injured severely, people would pay additional attention to not hurt them, or the government should push forward regulations on protecting them, which will not get more people in that group injured.

6 Conclusion

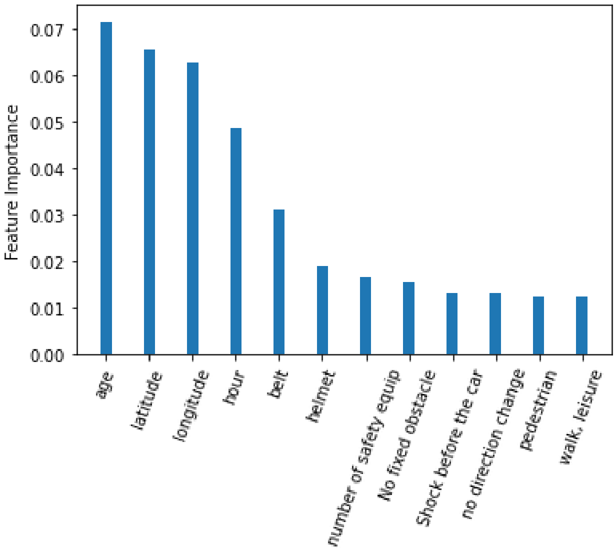


Figure 5: Feature Importance

After building and testing two models, we ended up with the random forest model as our final model. We also selected the most important features from the random forest model. According to Figure 5, the top four features are age, location, time and seat belt. This result shows that old people are more likely to get hurt severely at car accidents. For car manufacturing firms, we recommend to build a automatic break system inside vehicles. Some accidents happen because the pedestrians are in the blind spot of drivers. With the new system, a car can automatically slow down when it detects a pedestrian nearby. In this way, the hitting impact will be minimized or even prevented, and the injury level can be greatly reduced. For location and time factors, we suggest that government place warnings or speed limits at high-accident locations and during high-accident time, which is 16pm to 20pm on Fridays. Finally, we also suggest government to implement policies that will enforce the use of safety equipment, like seat belt when driving. With these policies and systems in place, we believe the accident number and overall injury level will be greatly reduced.