Big Data Project 2

Machine Learning on Car Accident

Jiawang Zhou, Xi Liu

03/11/2019

1.Background

Cars have become an important part of people's daily transportation methods, but the risk of car collision always exists. This report analyzes different features of car accidents and makes predictions on the casualty. The dependent variable of the model is accident severity with three classes: slight, serious and fatal. We first used descriptive statistics to know about data. Then we processed data by excluding all the N.A. and used different methods to build models separately. The best model is the one with the highest accuracy, the Random Forest model. Finally, based on the Random Forest model, we testified our prediction of severity with the original severity.

2.Data Description

1) Variables Selection

Our data was obtained from Kaggle, including 4,287,593 observations and 67 variables. Due to limitation of computer processing capacity, we reduced the dataset to randomly selected 10,000 observations. Then we chose 21 most related variables from the dataset based on logical judgment and machine learning methods. The dependent variable is accident severity and the independent variables include days of the week, time, road condition, urban or rural area, driver profile and casualty profile. The variables include:

Variables Name	Variable Type	Variable Description
Accident Severity	Categorical	1: Fatal, 2: Serious, 3: Slight
Number of Vehicles	Numerical	Vehicles involved in the accident
Number of Casualties	Numerical	Casualties involved in the accident
Year	Numerical	Year of the accident
Day of Week	Categorical	1: Sunday, 2: Monday, 3: Tuesday, 4: Wednesday, 5: Thursday, 6: Friday, 7: Saturday
Hour of Day	Categorical	1-24
1st Road Class	Categorical	1: Motorway, 2: A(M), 3: A, 4: B, 5: C
Light Conditions	Categorical	1: Daylight, 4: Darkness-lights lit, 5: Darkness-lights unlit, 6: Darkness-no lighting
Weather Conditions	Categorical	1: Fine no high winds, 2: Raining no high winds, 3: Snowing no high winds, 4: Fine+high winds, 5: Raining+high winds, 6: Snowing+high winds, 7: Fog or mist
Road Conditions	Categorical	1: Dry, 2: Wet or damp, 3: Snow, 4: Frost or ice, 5: Flood over 3cm, 6: Oil or diesel, 7: Mud
Urban or Rural Area	Categorical	1: Urban, 2: Rural
Vehicle Type	Categorical	1: Pedal cycle, 2: Motorcycle 50cc and under, 3:

		Motorcycle 125cc and under, 4: Motorcycle over 125cc and up to 500cc, 5: Motorcycle over 500cc, 8: Taxi/Private hire car, 9: Car, 10: Minibus (8 - 16 passenger seats), 11: Bus or coach (17 or more pass seats), 16: Horse, 17: Agri vehicle, 18: Tram, 19: Van <3.5t, 20: Van <7.5t, 21: Van>7.5t, 22: Mobility scooter, 23: Electric motorcycle
1st Point of Impact	Categorical	0: Did not impact, 1: Front, 2: Back, 3: Offside, 4: Nearside
Dominate Hand	Categorical	1: Right, 2: Left
Sex of Driver	Categorical	1: Male, 2: Female
Age of Driver	Numerical	Age of driver
Engine Capacity	Numerical	Engine capacity
Age of Vehicle	Numerical	Age of Vehicle
Casualty Class	Categorical	1: Driver or rider, 2: Passenger, 3: Pedestrian
Sex of Casualty	Categorical	1: Male, 2: Female
Age of Casualty	Numerical	Age of casualty

Table 1: Variable Selection

The correlations between accident severity and categorical variables are examined with chi-squared test. The p-value of tests are shown in Table 2. The low p-values show that we have enough confidence to reject the hypothesis that there is no correlation. In other words, there is strong relationship between the chosen variables.

Day of Week	Time	1st Road Class	Light Condition	Weather
$1.05e^{-1}$	$2.1e^{-4}$	$6.25e^{-2}$	$2.62e^{-12}$	$6.66e^{-7}$
Road Surface	Urban/Rural	Vehicle Type	Point of Impact	Dominate Hand
$8.98e^{-7}$	$1.61e^{-33}$	$2.04e^{-19}$	$1.5e^{-14}$	$8.21e^{-1}$
Sex of Driver	Sex of Casualty	Casualty Class		
$1.8e^{-8}$	$2.54e^{-3}$	$2.54e^{-3}$		

Table 2: Correlations between Accident Casualty and Categorical Independent Variables

2) Data Desctiption

Among all the accidents, 3983 accidents resulted in slight casualty, taking up around 83% of total accidents. Then come serious casualty and fatal casualty with 733 and 111 records, taking up 15% and 2%. Therefore, the death rate is 2%. (Shown in Appendix 1)

From 2005 to 2014, there is a clear trend that the accident numbers firstly decreased and maintain stable in the last several years. (Shown in Appendix 2)

Then we studied the distribution of accidents in weekdays. Friday has the most accident records while Sunday has the least. As for fatal casualties, most fatal accidents happened on Friday, Saturday and Monday. Friday is standing out among weekdays when it comes to total accident number. (Shown in Appendix 3)

The number of accidents were fewest and decreasing between midnight and dawn. We have fewest records at 4 o'clock in the morning. Then there comes a sharp increase at 7 o'clock and the number reaches a peak at 8, which are typical time people drive to school or work. There is a decrease at 9 and 10 o'clock and the number starts to rise after noon. 16 and 17 o'clock are the time when most accidents happened. The time with most fatal accidents is 18 o'clock. (Shown in Appendix 4)

Most of the accidents happened on dry road surface, with 3309 accident records out of 4827. The second most dangerous road surface is wet or damp one. Surprisingly, there are only 87 records regarding snow, frost or ice and flood surface in total. This is probably because people just avoid these roads when they are dangerous. (Shown in Appendix 5) Similarly, most accidents happen when the weather is fine. The number decreases sharply in raining days. And there is almost no accident report during extreme weathers such as snowing, fog or high-wind. This may be because people usually don't drive in these weathers. (Shown in Appendix 6)

Most vehicles involved in the accidents are private cars. (Shown in Appendix 7) And most impacts happened in the front of the vehicle, brining most deaths. Second most impacts happened in the back, with no deaths. The impacts on the offside usually caused fatal casualties but those on the nearside didn't. (Shown in Appendix 8)

Most fatal or non-fatal casualties happened on drivers, riders or passengers. Pedestrians had fewest accidents, according to the dataset. The death rate of drivers or riders is 2.19%, passengers 3% and pedestrians 0.53%. Surprisingly, the pedestrians have the least death rate. (Shown in Appendix 9)

Male drivers have more accidents and fatal casualty records than females. The death rate of male involved in the accidents is 2.5% while the rate of female is 2%. The dominate hand of most drivers is right hand. Although there were similar number of accidents in urban and rural places, those happened in rural places have caused much more fatal causalities. (Shown in Appendix 10)

3.Data Processing

The original datasets include both three datasets, accident data (information about accidents) vehicle data (information about vehicles in the accident), casualty data (information about casualties in the accident). We merged these three datasets into one by their shared Accident ID and processed the raw data. The data processing methods include:

1) Transform Categorial Variables to Factor

Many of our variables are labeled as an integer but in fact, those integers just refer to some real-world factor information. For example, the dependent variable in our model is the Accident Severity. Its original data is identified as 1,2,3, which indicate accident severity level slight, serious and fatal. To better analyze and visualize those variables, therefore, we used its real-world meaning as factors to approach the data description.

2) Process NA numbers

There are lots of NAs in the Kaggle dataset, which will influence our analysis. Since we have a massive amount of observation in our dataset, for all variables with a missing value, we just delete the entire rows directly.

3) Process Missing numbers

There are lots of -1 or 0 numbers in our dataset, which means those observations are unknown or not recorded. As we have mentioned, since our data set is huge, we can afford to get rid of those meaningless data meanwhile maintain enough staples for us to do machine learning.

4) Shrink the dataset

Due to the limitation of our computer's calculation power, we fail to test all our data through machine learning models, our R and R studio could only handle 31GB data. To speed up the machine learning modeling and also keep the accuracy as much as possible, we randomly selected 10 thousand rows of data, which shrunk our data set to almost 1/300. After doing data description with both sample data and original data, we found the distribution of variables in the two datasets is similar. Therefore, the sample is representative of our original data.

4.Model

1) Model Training

We used resampling methods named repeated cross-validation to create a set of modified datasets from the training samples. Then we applied lda, cart, knn, svm, random forest, and bagging methods separately to train the model. The algorithm selection standards include accuracy and kappa. Since our dependent variables are not a binary variable, we could not use the ROC Curve. The accuracy and kappa of all the algorithms are shown in Chart 1. The result shows that the random forest algorithm provides the largest accuracy and second highest kappa coefficient. Despite the fact that bagging algorithm has the highest kappa, it keeps the lowest accuracy. Therefore, it is the most accurate and effective method to build a model and make a prediction.

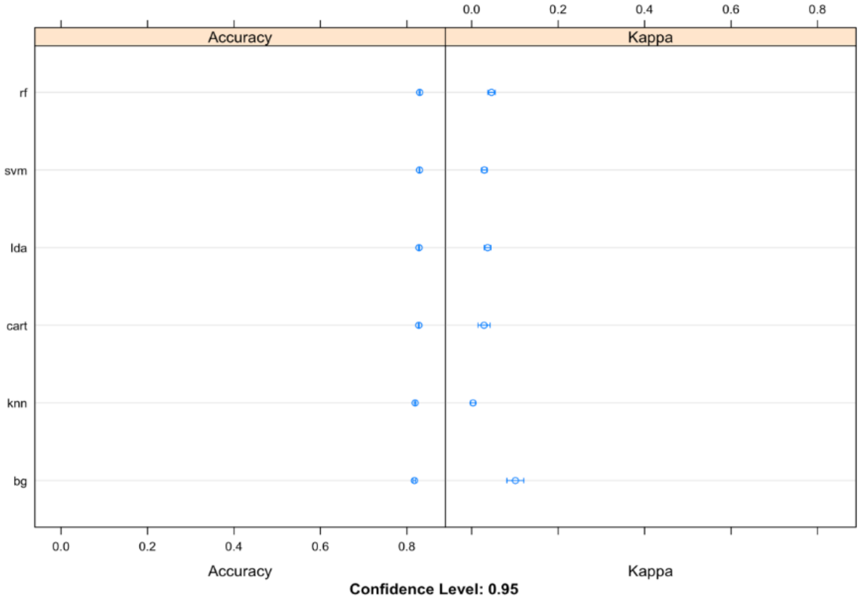
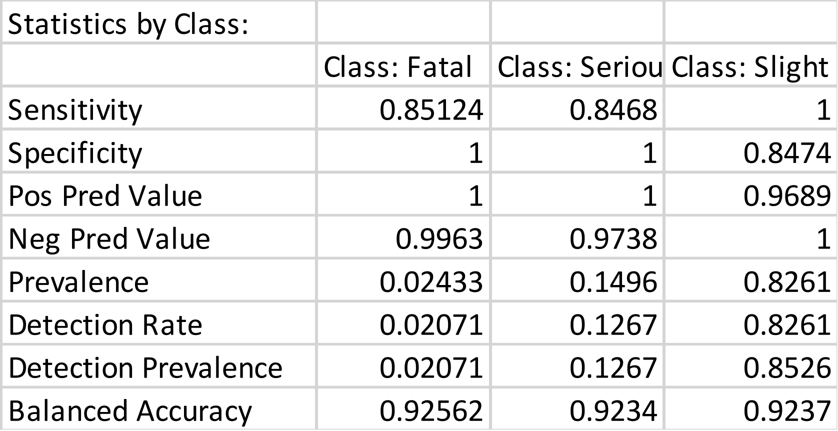


Chart 1: Accuracy of All Algorithms

2) Model Validation

As Table 2 shows, we calculated the confusion matrix information and plot the confusion matrix of our predicted severity and actual severity to reveal the validation of our prediction. This result shows that our model has a strong prediction accuracy.



Statistics by Class:			
	Class: Fatal	Class: Seriou	Class: Slight
Sensitivity	0.85124	0.8468	1
Specificity	1	1	0.8474
Pos Pred Value	1	1	0.9689
Neg Pred Value	0.9963	0.9738	1
Prevalence	0.02433	0.1496	0.8261
Detection Rate	0.02071	0.1267	0.8261
Detection Prevalence	0.02071	0.1267	0.8526
Balanced Accuracy	0.92562	0.9234	0.9237

Table 2: Confusion Matrix Result

5.Conclusion

In conclusion, the random forest algorithm is the best fit to make car accident prediction by providing the highest accuracy. The result shows that the model has a high prediction accuracy and can be used to make predictions.

6.Improvements

There are several improvements that we might promote our machine learning model further.

1) Variable Selection

There are over 65 variables in our dataset, for simplicity we just selected 21 of them. The selection itself has eliminated most part of data which could be useful in machine learning method.

2) Outlier deletion

Our variables are mostly categorical variables. Therefore, it is hard to find outliers merely by looking at the data description chart or make judgment merely based on our common knowledge. However, we intend to investigate this more thoroughly in a later stage (possibly using the 'outliers' package).

3) Model Selection

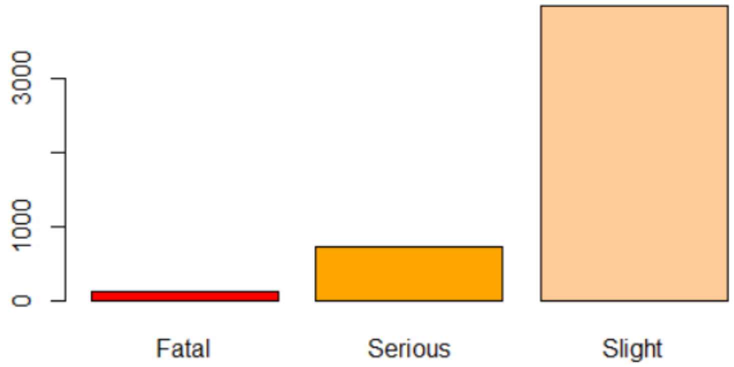
Firstly, though we use the accuracy and kappa coefficient for accuracy estimation, there are other measurements of model accuracy, under which Random Forest might not be the best model.

Secondly, we have used the trial and error method with many different algorithms, unfortunately, there is too many machine learning algorithms out there that may be the better fit for this dataset.

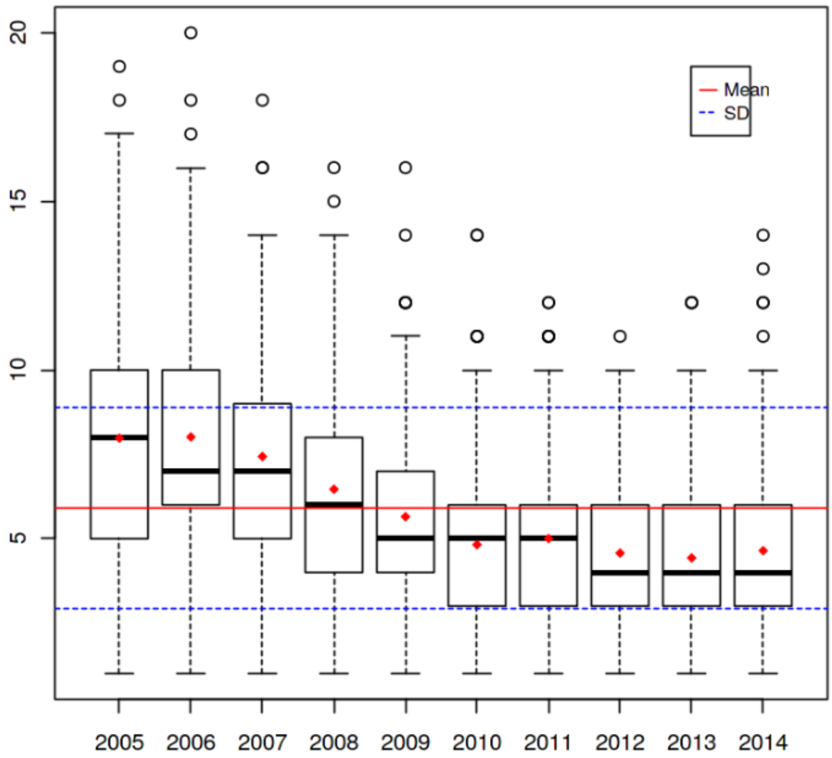
Lastly, instead of using a single Random Forest model for testing, many machine learning papers are using mix models with weighted averaging prediction.

Appendix

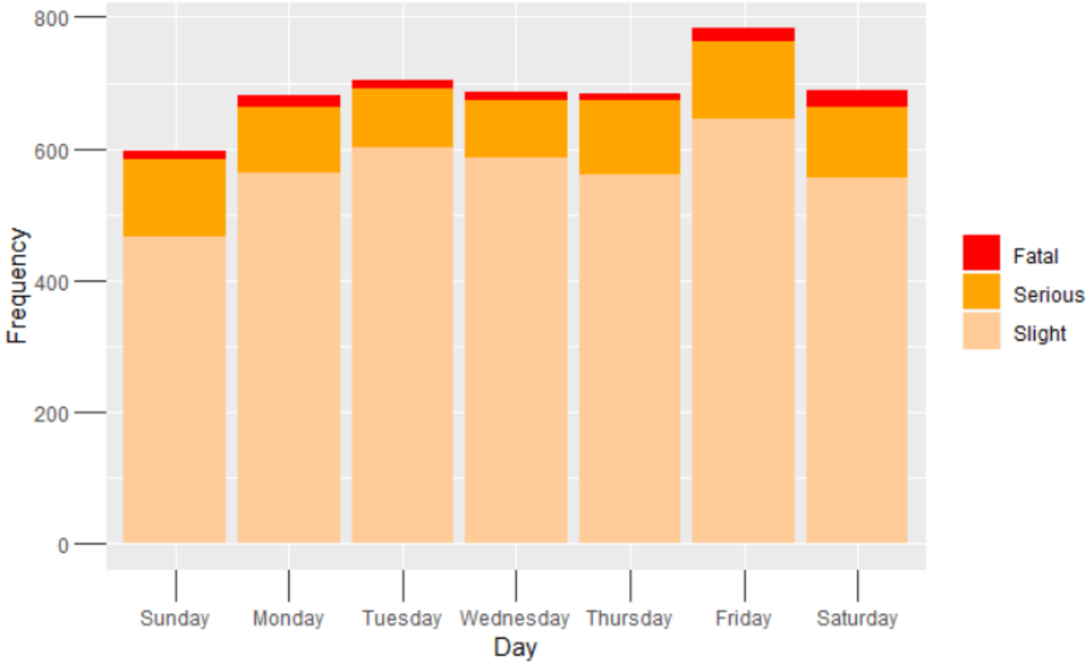
Appendix 1: Three categories of accident causality and distribution



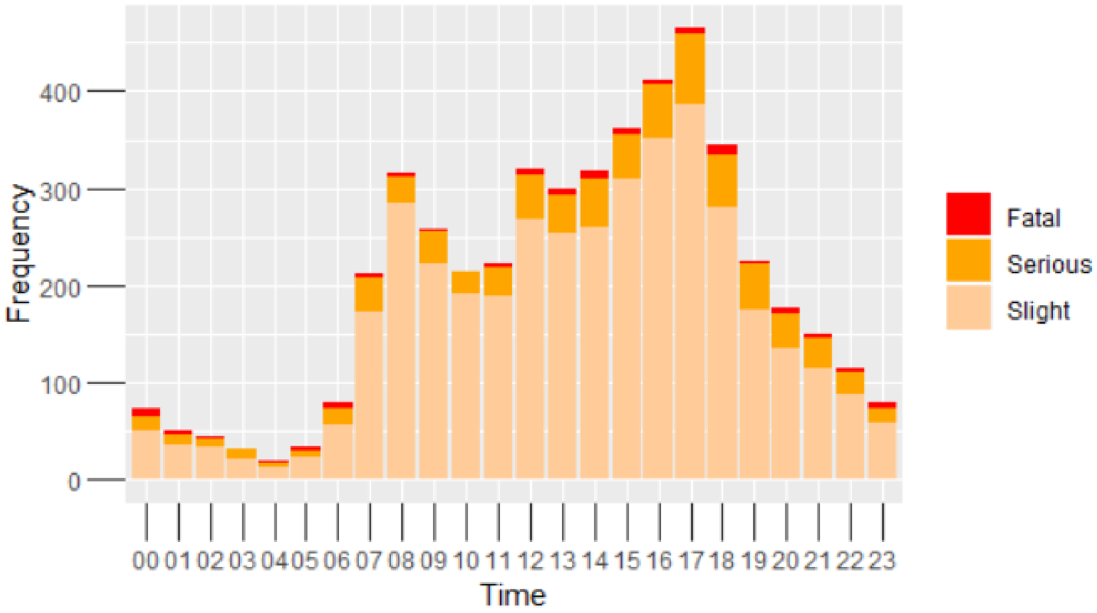
Appendix 2: Number of accidents in different years



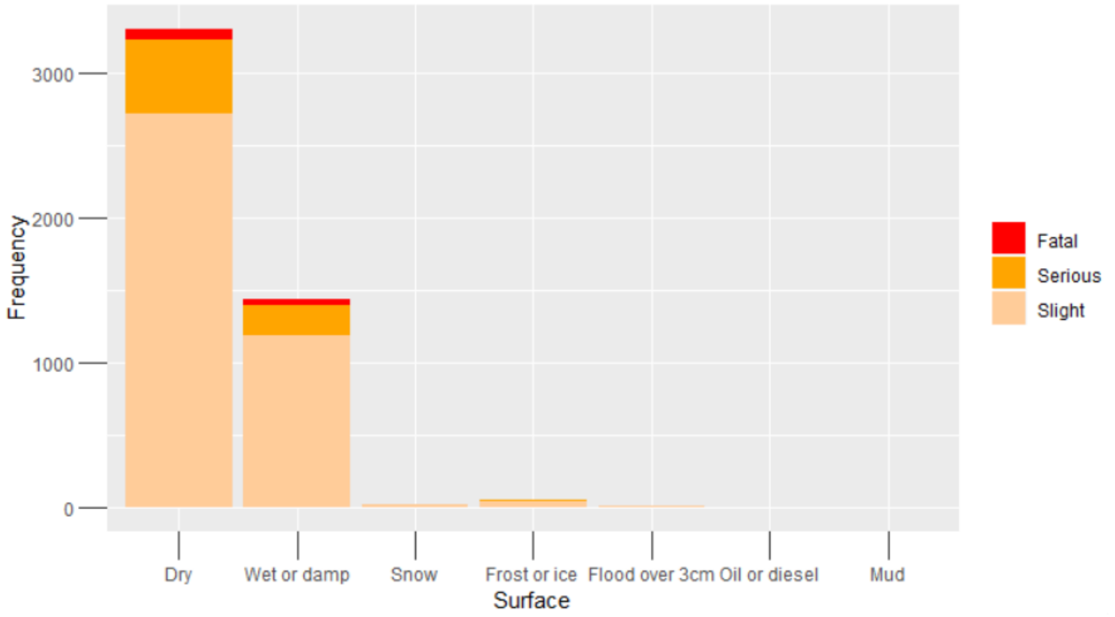
Appendix 3: Distribution of accidents in weekdays



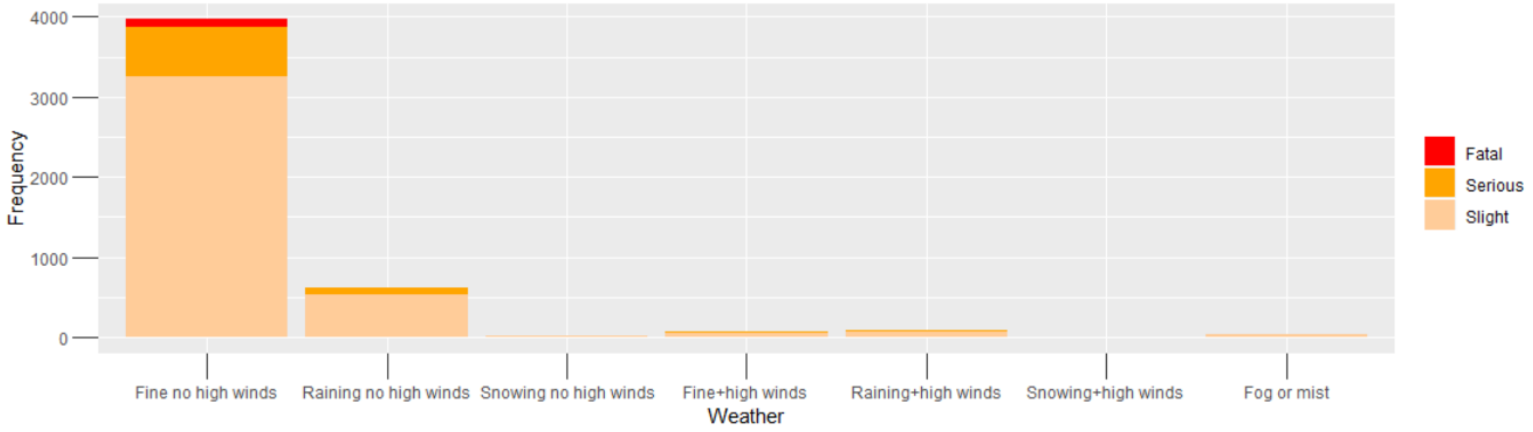
Appendix 4: Time distribution of accidents



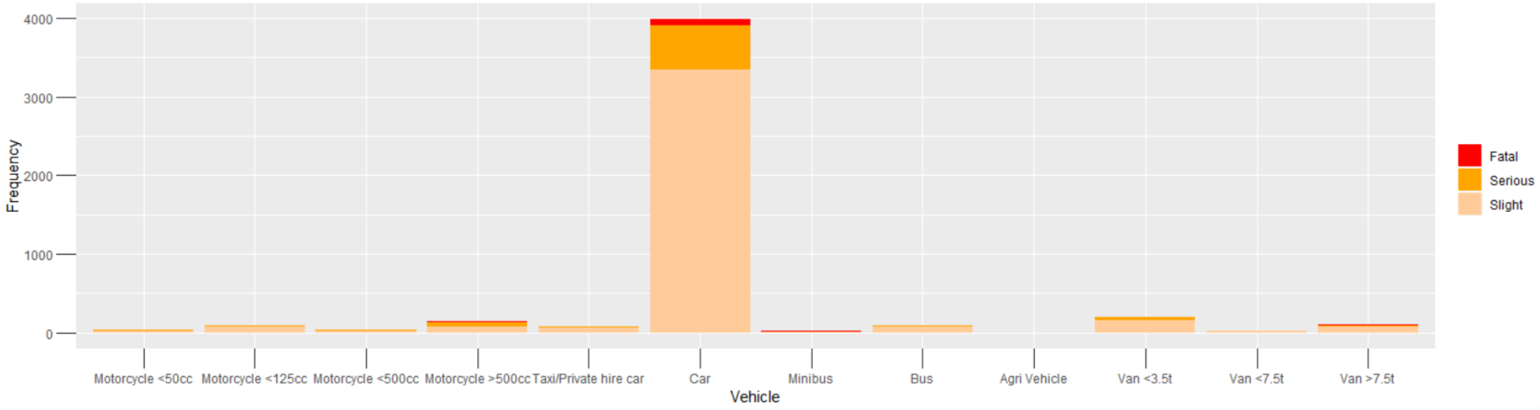
Appendix 5: Type of road surface situations



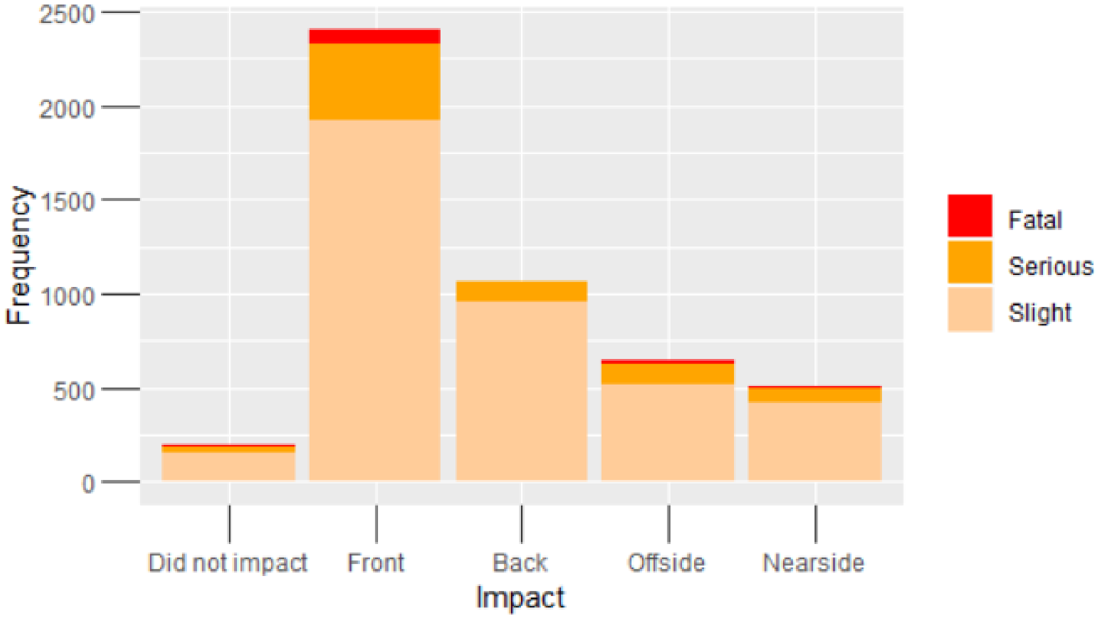
Appendix 6: Type of weather when accidents happened



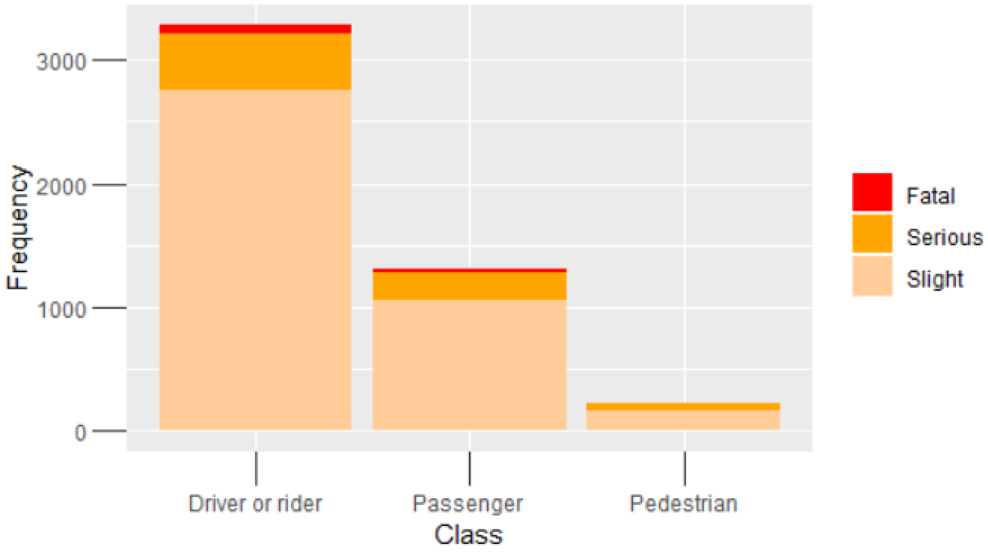
Appendix 7: Type of vehicles



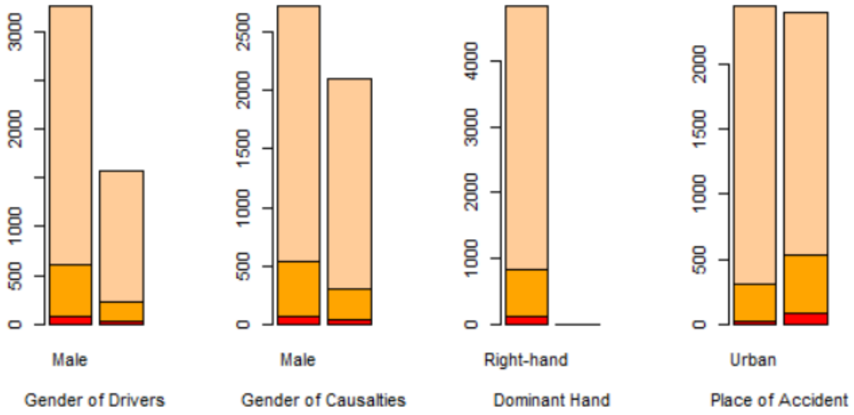
Appendix 8: Where the impact happened



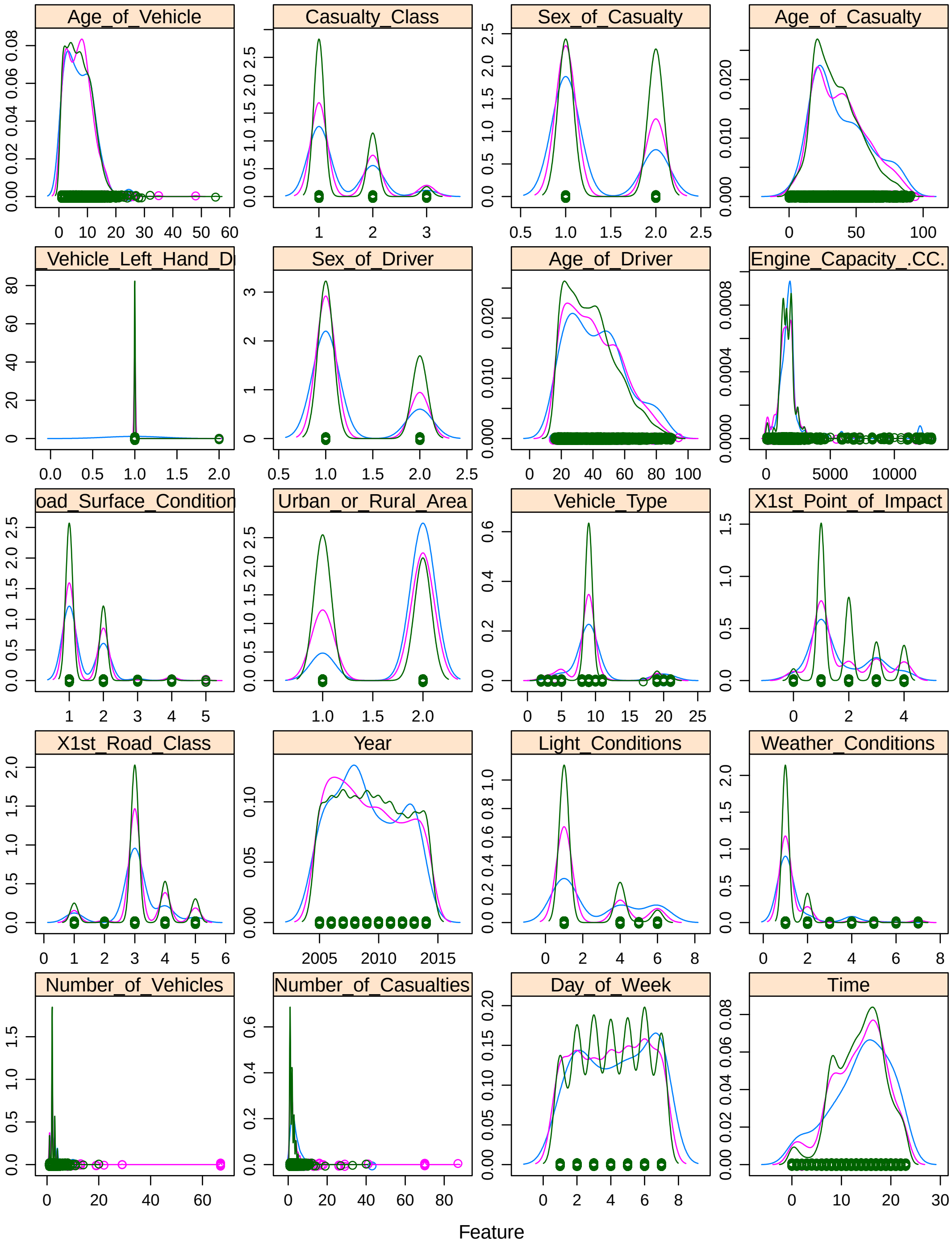
Appendix 9: Casualty class distribution



Appendix 10: Distribution of gender of drivers, gender of casualties, dominant hand and place of accidents



Appendix 11: Variable Destription


Age_of_Vehicle
Casualty_Class
Sex_of_Casualty
Age_of_Casualty
_Vehicle_Left_Hand_D
Sex_of_Driver
Age_of_Driver
Engine_Capacity_.CC.
oad_Surface_Condition
Urban_or_Rural_Area
Vehicle_Type
X1st_Point_of_Impact
X1st_Road_Class
Year
Light_Conditions
Weather_Conditions
Number_of_Vehicles
Number_of_Casualties
Day_of_Week
Time
Feature