ISyE 6416 – Basic Statistical Methods - Fall 2015

Bonus Project: “Big” Data Analytics Final Report

Motor Vehicle Accident Analysis

Team Member

Xueting Wang

Hao Wei

Hongao Yang

1.Problem Statement

Studies show that motor vehicle accident is one of the leading causes of death all over the world. Over 37,000 people in the United States die in road crashes every year. The number of death causing by motor vehicle accident is higher than most of people expected. Especially in United States, it is easy to get the driving license for any people no matter how old he/she is or he/she is just 18 years old. It is a common for an American family to own 2 or more vehicles, which means most family member may have their own vehicles. More and more people get driving licenses and more and more people own their vehicle to drive on road so that it is important to find the factors causing vehicle accident to decrease the number of vehicle accident and death.

However, when we think about the factors causing the vehicle accident, there will be a list of factors that may be related such as weather, road condition and so on. It is impossible to collect so many factors and use them. And most of factors may be neglected. We may select some of the important factors to use in our model. We have found the data of a record of each vehicle involved in a crash as reported to New York State Department of Motor Vehicles for three-year window. We would like to raise suitable regression model to assess the influential causes of the accident. Also, performance of three years' data will be used to check whether the causes of the accident differ from time, and furthermore, whether weights of causes differ from time.

To sum up, we would like to find the most important factors causing the vehicle accident by establishing suitable models, such as logistic regression model and multivariable regression model, and using the official data from New York Department of vehicle.

2.Data Source

Right now we have found a set of data, "Motor Vehicle Crashes – Vehicle Information: Three Year Window", which are attributes about each vehicle involved in a crash as reported to NYS DMV.

Here is the link: https://catalog.data.gov/dataset/motor-vehicle-crashes-vehicle-information-beginning-2009

Here is the sample of this data:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Year	Case Vehicle	Vehicle Body	Registration	Action Prior	Type / Axles	Direction of	Fuel Type	Vehicle Year	State of Regi	Number of C	Engine Cylin	Vehicle Make	Contributing	Contributing	Contributing	Contributing	Event Type
2	2012	9355479	SUBURBAN	PASSENGER	Going Straigh	Not Entered	East	Gas	2004	NY	2	8	DODGE	HUMAN	Not Entered	HUMAN	Not Entered	Not Entered
3	2012	9355480	SUBURBAN	PASSENGER	Going Straigh	Not Entered	East	Gas	2002	NY	1	6	FORD	HUMAN	Not Entered	HUMAN	Not Entered	Not Entered
4	2012	9967254	SUBURBAN	PASSENGER	Going Straigh	Not Entered	East	Gas	2004	NY	1	8	JEEP	HUMAN	Fell Asleep	HUMAN	Not Applicab	Not Applicable
5	2012	9967255	2 DOOR SED	PASSENGER	Parked	Not Entered	East	Gas	1996	NY	N/A	4	HONDA	HUMAN	Not Applicab	HUMAN	Not Applicab	Not Applicable
6	2012	9967294	4 DOOR SED	PASSENGER	Going Straigh	Not Entered	West	Gas	2001	NY	1	4	HONDA	HUMAN	Fell Asleep	HUMAN	Not Applicab	Tree, Collision
7	2012	9967368	4 DOOR SED	PASSENGER	Going Straigh	Not Entered	Northeast	Gas	2011	NY	1	4	CHEVR	HUMAN	Following To	HUMAN	Not Applicab	Not Applicable
8	2012	9967369	4 DOOR SED	PASSENGER	Slowing or St	Not Entered	Northeast	Gas	2011	NY	1	4	FORD	HUMAN	Not Applicab	HUMAN	Not Applicab	Other Motor V
9	2012	9967378	PICKUP TRUC	PASSENGER	Making Left	Not Entered	Northeast	Gas	1995	NY	1	6	FORD	HUMAN	Turning Impr	HUMAN	Unsafe Spee	Fence, Collisio
10	2012	9967389	4 DOOR SED	OMNIBUS - 1	Going Straigh	Not Entered	South	Gas	2003	NY	4	8	LINCO	HUMAN	Reaction to C	HUMAN	Not Applicab	Not Applicable

There are over 1,000,000 rows in this data set and it should be enough for our model.

There are 18 rows in the data and the following are the details.

Original data:

1. Year: {2011, 2012}, Year of accident happened
2. Vehicle Body: {2 Dr Sedan, 4 Dr Sedan, Pickup, …}
3. Registration: {Military, Court, …}
4. Action Prior to Accident: {Avoiding Object, Going straight, Backing, …}
5. Type/Axles of Truck or Bus: Number of Axles of Truck or Bus
6. Direction of Travel: {East, North, …}
7. Fuel Type: {Gas, diesel, electric, …}
8. Vehicle Year: {1996, 2001, 2002, …}
9. State of Registration: {NY, ME, FL, …}
10. Number of Occupants: {N/A, 1, 2, …}
11. Engine Cylinders: {1, 2, 3, 4, …}
12. Vehicle Make: {Dodge, Ford, Jeep, …}
13. Contributing Factor 1: {ENVMT, HUMAN, N/A, VEHICLE}
14. Contributing Factor 1 Description: {Texting, Drug, Eating, …}
15. Contributing Factor 2: {ENVMT, HUMAN, N/A, VEHICLE}
16. Contributing Factor 2 Description: {Texting, Drug, Eating, …}
17. Event Type: {Animal Collision with, Crash Collision, …}

However, some of the data are useless in our model and should be deleted. The remaining data includes too many factors. For examples, there are over 100 brands in the row of make. We need classify and combine several separated data into groups. The following shows the classified data and the details would be shown in Appendix.

1. Vehicle body type:

{2 DOOR SEDAN,4 DOOR SEDAN, ALL TERRAIN VEHICLE, BUS(OMNIBUS), CONVERTIBLE. LIMOUSINE(OMNIBUS), PICKUP TRUCK, SUBURBAN, UTILITY, TAXI, FREIGHT, SPECIAL, MOTORCYCLE}

We combine all kinds of trucks into a group FREIGHT and all other special function vehicle such as police car, ambulance into a group SPECIAL.

2. REGISTRATION:

{BUSINESS registration, PASSENGER registration, OMNIBUS registration, MOTORCYCLE, SPECIAL registration}

We combine the data into the groups divided by the different type of registration, for example, for business, passenger, omnibus and so on.

3. ACTION PIOR TO ACCIDENT:

{park & slow & stop, changing lane & merge & overtake, make turn, BAKCING, OTHER, Avoid, straight}

4. Type / Axis:

{2 axle box, 2 axle single, 3 axle single, not enter}

5. Direction:

{EAST, NORTH, NORTHEAST, NORTHWEST, SOUTH, SOUTHEAST, SOUTHWEST, WEST}

6. Fuel Type:

{NATURAL GAS, DIESEL, ELECTRIC, FLEX, GAS, NONE, NOEN, PROPANE}

VEHICLE YEAR:

{more than 20, 20 YEAR, 10 YEAR, 5 YEAR, 3 YEAR}

We divided the data into different range of times of a vehicle.

7. number of occupants:

{0, 1, 2, 3, 4, 5, 6, 7, >7}

Normal small vehicle includes no more than 7 seats so we combine all the data of over 7 into a group.

8. engine cylinder:

{0, 1, 2, 3, 4, 5, 6, 8, 10}

9. vehicle make"

{korea, japan, USA, German, England, Italy, Other}

We combine the makes into group of different famous car maker countries.

10.Contributor FACTOR

{ENVMT, HUMAN, VEHICLE}

11.Event type:

{Non fatal, Fatal}

There are about 30 different event type and we divide them into 2 groups, FATAL and NON FATAL, according to the possible damage caused by each event type.

Event Type as FATAL and NON-FATAL would be our results of factors and the remaining 10 data would be our factors.

3.Methodology

3.1.Supervised Learning

After preprocessing our data, we use four types of supervised learning algorithm in order to do prediction from our training data. They are Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Random Forest (RF), and Logistic Regression (LR). We split our dataset into training (90%) and testing (10%) data. Moreover, we run 10 times in order to test the robustness of each algorithm. Last

but not least, we also use paired t-test to show that the best method is significantly different from other methods.

3.1.1. Linear Discriminant Analysis (LDA) / Quadratic Discriminant Analysis (QDA)

In this project, our goal is to predict whether a car accident is fatal or not. LDA assume that each class is normally distributed and has same covariance. Under this assumption, the Bayes optimal solution is to predict points as being from the second class if the log of the likelihood ratios is below some threshold (Thres). On the other hand, without the assumption of same covariance of all classes, the discriminant analysis becomes QDA.

3.1.2. Random Forest (RF)

Decision trees are a popular method for various Machine Learning tasks, because it is invariant under scaling and various other transformations of feature values, and is robust to inclusion of irrelevant features. In particular, trees that are grown very deep tend to learn highly irregular patterns. However, they usually overfit their training data with have low bias, but high variance. RF are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance. The training algorithm for random forests applies the general technique of bootstrapping to decision trees, bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples. Besides, for a classification problem only $\sqrt{p}$ (rounded down) features are used in each split, which is 3 in this project.

3.1.3. Logistic Regression (LR)

Logistic Regression is a popular and straightforward learning algorithm in classification. LR has similar concept as LDA, but it does not have any assumption on the distribution of the classes. Therefore, in general LR would perform better on non-normal distributed dataset. However, LDA would have better performance vice versa.

3.2. Regression:

3.2.1. Ridge regression:

For each year, we conducted ridge regression method and then choose the most important factors.

3.2.2. Logistic regression:

We did logistic regression for both years with results in appendix.

4. Evaluation and Final Results

4.1. Supervised Learning Result

(a)

LDA is the best method in this project in terms of mean testing error with 10 runs.

	LDA	QDA	RF	LR
Training error	0.3856	0.3861	0.2396	0.3859
Testing error	0.3920	0.3971	0.4182	0.3930

(b)

We learn that Action Prior to Accident, Contributing Factor, Vehicle Body Type, Fuel Type, and Engine Cylinders are the dominant factors resulting in a fatal car accident. These significant factors are the conclusion from the importance of the tree model, which mainly reduce the error rate.

(c)

Paired t-test:

	LDA vs QDA	LDA vs RF	LDA vs LR
p-value	0.04612	0.0001517	0.009991

Based on the result of paired t-test, LDA is indeed statistically different from other methods

4.2. Regression Result

Ridge regression:

For year 2011, the result of ridge regression is as follows:

```
predictorsVehicle_body_type      predictorsRegistration
                    -0.0115                     -0.0148
          predictorsAction             predictorsType
                     0.0976                     -0.0038
     predictorsDirection              predictorsFuel
                     0.0278                      0.0128
       predictorsYear           predictorsOccupants
                    -0.0081                     -0.0031
     predictorsEngine      predictorsVehicle_Make
                    -0.0204                      0.0360
    predictorsContri1            predictorsContri2
                    -0.0146                     -0.0429
    predictorsContri
                    -0.0326
```

If we choose the cutoff value as 0.03, then it is clearly that the significant factors for vehicle accident in year 2011 is Action Prior to Accident, Vehicle Make, Contributing Factor 2, Contributor Combine. The model will be

$$y = 0.0976predictorsAction + 0.0360predictorsVehicle_{Make} - 0.0429predictorsContri2 - 0.0326\text{predictorsContri}$$

Logistic Regression:

General linear model(glm) in binomial family for logistic regression result for year 2011 and 2012 can be seen in the appendix.

For the year 2011, the influential factors that p-value is less than 0.05 are: Vehicle body type of PICKUP TRUCK and TAXI, most prior actions and directions, engine cylinder's number. What should be treat with strong reasoning is the interaction term of HUMAN-HUMAN contributors.

The stepwise selection of variables was done in backward direction. The smallest AIC is 12253.98 with model: Event ~ Vehicle_body_type + Registration + Action + Fuel_Type + Direction + Engine + Vehicle_Make + Contri1 + Contri2 + Contri1:Contri2. The exponentiated coefficients of this selected model are shown in the appendix.

Prediction and model effectiveness are conducted as shown in appendix. The model is more likely to predict a fatal event as non-fatal.

Same procedures were used to treat the data for year 2012. The influential factors that p-value less than 0.05 are: Vehicle body type of FREIGHT, Registration OMNIBUS, prior actions and directions north facing, 3-year age of Vehicle, engine cylinder's number, USA vehicles and vehicle-vehicle interaction.

For stepwise, the smallest AIC is 12713.58 with model: Event ~ Vehicle_body_type + Registration + Action + Year + Engine + Vehicle_Make + Contri1 + Contri2 + Contri1:Contri2

The comparison of selected model for 2011 and 2012 can be seen in the following table:

Predict variable	2011	2012
Vehicle_body_type	√	√
Registration	√	√
Action	√	√
Fuel_Type	√	X
Direction	√	X
Year	X	√
Engine_Cylinder	√	√
Vehicle_Make	√	√
Contri1	√	√
Contri2	√	√
Contri1:Contri2	√	√

The influential variables for year 2011 and year 2012 are very close, with common ones as: vehicle body type, registration, prior action, engine cylinder, vehicle make, contributing factors and their interactions.

4.3. Conclusion

Combing the results of classify method and regression method, we can find the key factors in Vehicle body type, Actions prior to accidents, Engine Cylinder and Contribution factors. The results of our models are similar with our expected results.

In vehicle body type, the most significant factors are Pickup Truck and Taxi. The coefficient for pickup truck is negative, which indicate it not likely to involve in fatal accident. The pickup truck drivers are better trained than normal drivers, and knows better how to react to an accident on the road. The coefficient of TAXI is positive, indicating a higher number of fatal accident. The reason for taxi to be key factor is mainly because the driving time of taxis are much more than that of other vehicles, which will lead higher possibility to cause accidents.

In Actions prior to accidents, most of factor are significant expecting other and straight. When you go straight, you will have better view in front than that of other directions and you can act faster to the accident. Other factors are in our expectations. When you drive backing, you may have poor view in the back and be easy to ignore others in your front and side. When you are in low speed, such as parking, slowing, starting, drivers would be easy to lose vigilance and attention because they think they are safe in low speed. Making turns, changing lanes, merging and overtaking will cause intersect with other vehicles, which is more dangerous and easier to cause accident.

In Engine Cylinder, the coefficient is positive, which indicates that the larger number of cylinder, the more serious of the car accident. Larger number of cylinder means more power of the engine. More power of the engine will allow higher speed limit of the vehicle. However, higher speed will cause more serious accident.

In Contribution Factor, Human & Human is the most significant factor. This satisfy our common sense. Comparing with factors of environment and vehicle, human is the main factor causing car accident.

No significance on vehicle brand, which may because of the grouping method in terms of country. It can be further study to group vehicle by price or other criteria.

According to our findings in car accidents, we have several suggestions to decrease the number of fatal accidents. The driver need to pay more attention when they drive larger size vehicle than small size vehicle, especially for normal driver. Even with higher power of engine, don’t drive too fast. Don’t lose attention when you think you are safe or you may intersect with other vehicles. Improving driving skills and avoiding drivers’ themselves failure of driving are also very important.

The results of this study can also provide a reference for vehicle insurance amount formulating.

Appendix

Data resource part:

1. Vehicle body type
    - 1.1 2 DOOR SEDAN
    - 1.2 4 DOOR SEDAN
    - 1.3 ALL TERRAIN VEHICLE
    - 1.4 BUS(OMNIBUS)
    - 1.5 CONVERTIBLE
    - 1.6 LIMOUSINE(OMNIBUS)
    - 1.7 PICKUP TRUCK
    - 1.8 SUBURBAN
    - 1.9 UTILITY
    - 1.10 TAXI
    - 1.11 FREIGHT
        - 1.11.1 DELIVERY TRUCK
        - 1.11.2 FLAT BED TRUCK
        - 1.11.3 REFRIGERATOR TRUCK
        - 1.11.4 STAKE TRUCK
        - 1.11.5 TANK TRUCK
        - 1.11.6 TOW TRUCK
        - 1.11.7 VAN TRUCK
    - 1.12 SPECIAL
        - 1.12.1 AMBULANCE
        - 1.12.2 CEMENT MIXTER
        - 1.12.3 DISABLED COMMERCIAL
        - 1.12.4 DUMP
        - 1.12.5 HOUSE ON WHEELS
        - 1.12.6 POWER SHOVEL
        - 1.12.7 SAND OR AGRICULTURAL
        - 1.12.8 TRACTOR
        - 1.12.9 POLICE VEHICLE
    - 1.13 MOTORCYCLE
2. REGISTRATION
    - 2.1 BUSINESS registration
        - 2.1.1 AGRICULTURAL COMMERCIAL
        - 2.1.2 AGRICULTURAL TRUCK
        - 2.1.3 COMMERCIAL

2.1.4 DEALER
2.1.5 FARM
2.1.6 HAM OPERATOR
2.1.7 MEDICAL DOCTOR
2.1.8 LIGHT TRAILER
2.1.9 SEMI-TRAILER
2.1.10 REGIONAL COMMERCIAL
2.1.11 TOW TRUCK
2.1.12 TRACTOR-REGULAR
2.1.13 ORGANIZATIONAL COMMERCIAL
2.1.14 ORGANIZATIONAL

2.2 PASSENGER registration
2.2.1 ALL TERRAIN VEHICLE
2.2.2 PASSENGER OR SUBURBAN
2.2.3 PASSENGER OR SUBURBAN (REGULAR)

2.3 OMNIBUS registration
2.3.1 OMNIBUS-LIVERY
2.3.2 OMNIBUS-PUBLIC SERVICE
2.3.3 OMNIBUS-REGUALR
2.3.4 OMNIBUS-TAXI
2.3.5 OMNIBUS-SPECIAL(PRIVATE RENTAL)

2.4 MOTORCYCLE

2.5 SPECIAL registration
2.5.1 AMBULANCE
2.5.2 BOB BIRTHPLACE OF BASEBALL
2.5.3 FORMER PRISONER OF WAR
2.5.4 FORMER PRISONER OF WAR*
2.5.5 COUNTY LEGISLATURE
2.5.6 INTERNATIONAL REGISTRATION
2.5.7 SPECIAL PASSENGER
2.5.8 SPECIAL PEASSNERGER($15 FEE)
2.5.9 SCHOOL CAR
2.5.10 SPORTS
2.5.11 STATE
2.5.12 POLITICAL SUBDIVISION
2.5.13 REGIONAL
2.5.14 IN TRANSIT PERMIT

3. ACTION PIOR TO ACCIDENT

3.1 park & slow & stop

- 3.1.1 ENTERING PARKED POSITION
- 3.1.2 PARKED
- 3.1.3 SLOWING OR STOPPING
- 3.1.4 STARTING FROM PARKING
- 3.1.5 STARTING IN TRAFFIC
- 3.1.6 STOPPED IN TRAFFIC

3.2 changing lane & merge & overtake

- 3.2.1 CHANGING LANES
- 3.2.2 MERGING
- 3.2.3 OVERTAKING/PASSING

3.3 make turn

- 3.3.1 MAKING LEFT TURN
- 3.3.2 MAKING LEFT TURN ON RED
- 3.3.3 MAKING RIGHT TURN
- 3.3.4 MAKING RIGHT TURN ON RED
- 3.3.5 MAKING U TURN

3.4 BAKCING

3.5 OTHER

- 3.5.1 OTHER
- 3.5.2 POLICE PURSUIT

3.6 Avoid

- 3.6.1 AVOIDING OBJECT IN ROADWAY

3.7 straight

- 3.7.1 GOING STRATIGHT AHEAD

4. Type / Axis
   - 4.1 2 axle box
   - 4.2 2 axle single
   - 4.3 3 axle single
   - 4.4 not enter
5. Direction
   - 5.1 EAST
   - 5.2 NORTH
   - 5.3 NORTHEAST
   - 5.4 NORTHWEST
   - 5.5 SOUTH
   - 5.6 SOUTHEAST
   - 5.7 SOUTHWEST
   - 5.8 WEST
6. Fuel Type

- 6.1 NATURAL GAS
- 6.2 DIESEL
- 6.3 ELECTRIC
- 6.4 FLEX
- 6.5 GAS
- 6.6 NONE
- 6.7 NOEN
- 6.8 PROPANE

7. VEHICLE YEAR
    - 7.1 [~,1993] more than 20
    - 7.2 [1994, 2003] 20 YEAR
    - 7.3 [2004, 2008] 10 YEAR
    - 7.4 [2009, 2010] 5 YEAR
    - 7.5 [2011, 2013] 3 YEAR
8. number of occupants
    - 8.1 0
    - 8.2 1
    - 8.3 2
    - 8.4 3
    - 8.5 4
    - 8.6 5
    - 8.7 6
    - 8.8 7
    - 8.9 >7
9. engine cylinder
    - 9.1 0
    - 9.2 1
    - 9.3 2
    - 9.4 3
    - 9.5 4
    - 9.6 5
    - 9.7 6
    - 9.8 8
    - 9.9 10
10. vehicle make
    - 10.1 korea
        - 10.1.1 DAEWO
        - 10.1.2 HYUND
        - 10.1.3 KIA

10.2 japan

- 10.2.1 Acura
- 10.2.2 HINO
- 10.2.3 HONDA
- 10.2.4 INFIN
- 10.2.5 ISUZU
- 10.2.6 KAWAS
- 10.2.7 LEXUS
- 10.2.8 MAZDA
- 10.2.9 MITSU
- 10.2.10 NISSA
- 10.2.11 Subar
- 10.2.12 Suzuk
- 10.2.13 TOYOT
- 10.2.14 UD
- 10.2.15 YAMAH

10.3 America

- 10.3.1 Ameri
- 10.3.2 BUELL(MOTOR)
- 10.3.3 BUICK
- 10.3.4 CADIL
- 10.3.5 Che
- 10.3.6 Chevr
- 10.3.7 Chrys
- 10.3.8 DODGE
- 10.3.9 EAGLE
- 10.3.10 FORD
- 10.3.11 FREIG
- 10.3.12 FRHT
- 10.3.13 GEO
- 10.3.14 GMC
- 10.3.15 GILLI
- 10.3.16 HA/DA
- 10.3.17 HADA
- 10.3.18 HARLE
- 10.3.19 HUMME
- 10.3.20 IC
- 10.3.21 INTER
- 10.3.22 INTL

- 10.3.23 J&J
- 10.3.24 JAYCO
- 10.3.25 JEEP
- 10.3.26 KENWO
- 10.3.27 LINCO
- 10.3.28 MACK
- 10.3.29 MERCU
- 10.3.30 OLDS
- 10.3.31 OLDSM
- 10.3.32 ORION
- 10.3.33 OSHKO
- 10.3.34 PETER
- 10.3.35 PLYMO
- 10.3.36 PONTI
- 10.3.37 SATUR
- 10.3.38 Sterl
- 10.3.39 THOMA
- 10.3.40 WABAS

10.4 German

- 10.4.1 AUDI
- 10.4.2 BMW
- 10.4.3 ME/BE
- 10.4.4 PORSC
- 10.4.5 Smart
- 10.4.6 VOLK
- 10.4.7 VOLKS
- 10.4.8 VW

10.5 England

- 10.5.1 ADVAN(MOTOR)
- 10.5.2 AS/MA
- 10.5.3 JAGUA
- 10.5.4 LA/RO
- 10.5.5 MINI
- 10.5.6 TRIUM

10.6 Italy

- 10.6.1 April(motor)
- 10.6.2 DUCAT(MOTOR)
- 10.6.3 FERRA
- 10.6.4 PIAGG

- 10.6.5 VESPA
- 10.7 Other
  - 10.7.1 Autoc
  - 10.7.2 BL/BI
  - 10.7.3 BLUEB
  - 10.7.4 CA/AM
  - 10.7.5 Cadli
  - 10.7.6 Carry
  - 10.7.7 Case
  - 10.7.8 Custo
  - 10.7.9 Custo
  - 10.7.10 DIAMO
  - 10.7.11 FOOD
  - 10.7.12 FTL
  - 10.7.13 HOMAD
  - 10.7.14 JEEPO
  - 10.7.15 MATE
  - 10.7.16 OTTAW
  - 10.7.17 PREVO
  - 10.7.18 SAAB
  - 10.7.19 SATAU
  - 10.7.20 Sprin
  - 10.7.21 SYM
  - 10.7.22 TOYOT
  - 10.7.23 UNIVE
  - 10.7.24 VA/NO
  - 10.7.25 VANHO
  - 10.7.26 VOLVO
  - 10.7.27 WE/ST
  - 10.7.28 WINNE

11. Event type
    - 11.1 Non fatal
      - 11.1.1 Animal
      - 11.1.2 Curbing
      - 11.1.3 Deer
      - 11.1.4 Earth embankment/rock cut/ditch
      - 11.1.5 Fence
      - 11.1.6 Fire hydrant
      - 11.1.7 Guide rail

- 11.1.8 Light support/utility pole
- 11.1.9 Other object (non fixed)
- 11.1.10 Other, non-collision
- 11.1.11 Ran off roadway
- 11.1.12 Sign post
- 11.1.13 Submersion
- 11.1.14 Tree

11.2 Fatal

- 11.2.1 Barrier
- 11.2.2 Bicycle
- 11.2.3 Bridge
- 11.2.4 Building/wall
- 11.2.5 Crash cushion
- 11.2.6 Culver/head wall
- 11.2.7 Fire/explosion
- 11.2.8 Median
- 11.2.9 Other fixed object
- 11.2.10 Other motor vehicle
- 11.2.11 Other pedestrian
- 11.2.12 Pedestrian
- 11.2.13 Railroad
- 11.2.14 Snow embankment

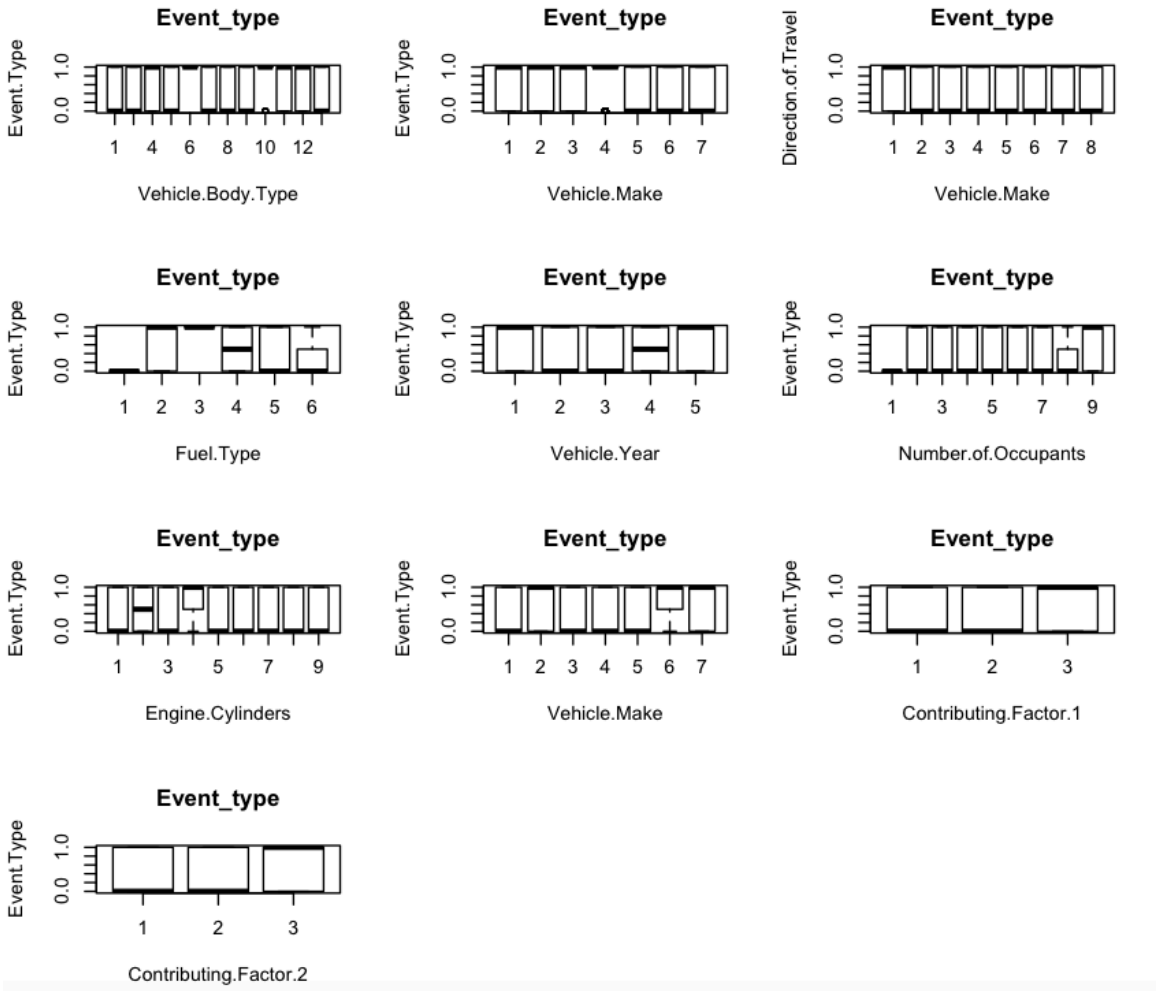
12. Contributor FACTOR

12.1 ENVMT

12.2 HUMAN

12.3 VEHICLE

Distribution Graph



Classify Part:

```
### Read the data
data11 <- read.table(file = "data11.csv", sep = ",", header=T)

head(data11)
data11 = data.frame(data11[2], data11[4], data11[6:13],data11[15])

library(lattice)
splom(data11[,1:15], pscales = 0)
par(mfrow=c(3,3))
colnames <- dimnames(data11)[[2]]
for (i in 2:10) {
    d <- density(data11[,i])
    plot(d, type="n", main=colnames[i])
    polygon(d, col="red", border="gray")
}
par(mfrow=c(4,3))

boxplot(Event.Type ~ Vehicle.Body.Type,data= data11,
main="Event_type",xlab="Vehicle.Body.Type", ylab="Event.Type")
boxplot(Event.Type ~ Action.Prior.to.Accident,data= data11,
main="Event_type",xlab="Vehicle.Make", ylab="Event.Type")
boxplot(Event.Type ~ Direction.of.Travel,data= data11,
main="Event_type",xlab="Vehicle.Make", ylab="Direction.of.Travel")
boxplot(Event.Type ~ Fuel.Type,data= data11,
main="Event_type",xlab="Fuel.Type", ylab="Event.Type")
boxplot(Event.Type ~ Vehicle.Year,data= data11,
main="Event_type",xlab="Vehicle.Year", ylab="Event.Type")
boxplot(Event.Type ~ Number.of.Occupants,data= data11,
main="Event_type",xlab="Number.of.Occupants", ylab="Event.Type")
boxplot(Event.Type ~ Engine.Cylinders,data= data11,
main="Event_type",xlab="Engine.Cylinders", ylab="Event.Type")
boxplot(Event.Type ~ Vehicle.Make,data= data11,
main="Event_type",xlab="Vehicle.Make", ylab="Event.Type")
boxplot(Event.Type ~ Contributing.Factor.1,data= data11,
main="Event_type",xlab="Contributing.Factor.1", ylab="Event.Type")
boxplot(Event.Type ~ Contributing.Factor.2,data= data11,
main="Event_type",xlab="Contributing.Factor.2", ylab="Event.Type")
```

```
###PCA
stdcor <- as.data.frame(scale(data11[,1:10]))
impca<- prcomp(data11[,1:10])
summary(impca)
plot(impca,type="lines")
eigenfaces = data.frame(impca$rotation)
pcomp <- 3
a <- as.matrix(data11[,1:10])%*%as.matrix(eigenfaces[ ,1:pcomp])%*%as.matrix(t(eigenfaces[ ,1:pcomp]))
head(data11)
round(cor(data11),2)
n = dim(data11)[1]; ### total number of observations
n1 = round(n/10);
B = 10;
TEALL=NULL;
TRALL=NULL;
for (b in 1:B){
flag <- sort(sample(1:n,n1));
data11train = data11[-flag,];
data11test = data11[flag,];
data11train$Event.Type <- as.factor(data11train$Event.Type);
##LDA
library(MASS)
fit1 <- lda(data11train[,1:10], data11train[,11])
##training error
pred1 <- predict(fit1, data11train[,1:10])$class
tr1 <- mean( pred1   != data11train$Event.Type)
##testing error
pred11 <- predict(fit1, data11test[,1:10])$class
ttr1 <- mean( pred11   != data11test $Event.Type)
## QDA
fit2 <- qda(data11train[,1:10], data11train[,11])
##training error
pred2 <- predict(fit2,data11train[,1:10])$class
tr2 <- mean( pred2!= data11train$Event.Type)
##testing error
```

```
ttr2 <- mean( predict(fit2, data11test[,1:10])$class != data11test$Event.Type)
##tree-method
library(randomForest)
set.seed(19910822)
tree2 = randomForest(Event.Type ~.,data= data11train,importance=TRUE)
##training error
pred3=predict(tree2,newdata= data11train)
tr3 <- mean( pred3   != data11train$Event.Type)
##testing error
pred33=predict(tree2,newdata= data11test)
ttr3 <- mean( pred33   != data11test$Event.Type)
importance(tree2)
## Logistic Regression
library(nnet)
fit4 <- multinom( Event.Type ~., data= data11train)
step(fit4)
##training error
pred4<- predict(fit4, data11train[,1:10])
tr4 <- mean( pred4 != data11train$Event.Type)
##testing error
ttr4 <- mean( predict(fit4, data11test[,1:10]) != data11test$Event.Type)
TRALL = rbind(TRALL, cbind(tr1,tr2,tr3,tr4));
TEALL = rbind(TEALL, cbind(ttr1,ttr2,ttr3,ttr4));
}
round(apply(TRALL, 2, mean),4)
round(sqrt(apply(TRALL, 2, var)),4)
round(apply(TEALL, 2, mean),4)
round(sqrt(apply(TEALL, 2, var)),4)
## compare LDA with others
t.test(TEALL[,2], TEALL[,1],paired=TRUE)
#p 0.04612
t.test(TEALL[,3], TEALL[,1],paired=TRUE)
# p 0.0001517
t.test(TEALL[,4], TEALL[,1],paired=TRUE)
# p 0.009991
```

Regression part:

R code:

```
data=read.csv(file.choose(),header=TRUE,sep=",")
dim(data)
attach(data)
library(lars)
n=length(data$Event.Type)
Event=rep(0,n)
Event[data$Event.Type=="FATAL"]=1

Vehicle_body_type=data$Vehicle.Body.Type
Registration=data$Registration.Class
Action=data$Action.Prior.to.Accident
Type=data$Type...Axles.of.Truck.or.Bus
Direction=data$Direction.of.Travel
Fuel=data$Fuel.Type
Year=data$Vehicle.Year
Occupants=data$Number.of.Occupants
Engine=data$Engine.Cylinders
Vehicle_Make=data$Vehicle.Make
Contri1=data$Contributing.Factor.1
Contri2=data$Contributing.Factor.2
Contri=data$contributor.combine
Event_Type=data$Event.Type
predictors=cbind(Vehicle_body_type,Registration,Action,Type,
                 Direction,Fuel,Year,Occupants,Engine,Vehicle_Make,
                 Contri1,Contri2,Contri)
predictors=scale(predictors)
y=scale(as.numeric(Event_Type))

#### ridge regression ####
library(MASS)
lambda=seq(0,100,by=0.01)
out=lm.ridge(y~predictors,lambda=lambda)
plot(out)
round(out$GCV,5)
which(out$GCV==min(out$GCV))
```

```
dim(out$coef)
round(out$coef[,10001], 4)

par(mfrow = c(1,1))
plot(lambda, out$coef[1,], type = "l", col = 1, lwd=3,
     xlab = "Lambda", ylab = "Coefficients",
     main = "Plot of Regression Coefficients vs. Lambda Penalty Ridge Regression",
     ylim = c(min(out$coef), max(out$coef)))
abline(h = 0, lty = 2, lwd = 3)
abline(v = 2.25, lty = 2,lwd=3)
for(i in 2:13)
  points(lambda, out$coef[i,], type = "l", col = i,lwd=3)

#### logistic regression ####
out=glm(Event~Vehicle_body_type+Registration+Action+Type+Direction+Fuel+Year+Occupants+Engine+Vehicle_Make+Contri1+Contri2+
          Contri1*Contri2, family = binomial)
summary(out)
step(out,direction = "backward")

## prediction ##
glm.probs<-predict(out2,type='response')
glm.probs[1:10]
contrasts(as.factor(Event))

## model ##
glm.pred<-rep('1',1250)
glm.pred[glm.probs>0.5]<-'0'
table(glm.pred,Event)
```

Logistic regression summary of year 2011

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0556  -1.0814  -0.8572   1.2106   1.7633

Coefficients:
                                     Estimate Std. Error z value Pr(>|z|)
(Intercept)                       -25.862714 459.257962  -0.056 0.955092
Vehicle_body_type4 DOOR SEDAN       0.058599   0.076972   0.761 0.446478
Vehicle_body_typeBUS (OMNIBUS)      0.429765   0.639328   0.672 0.501448
Vehicle_body_typeCONVERTIBLE       -0.301431   0.239972  -1.256 0.209076
Vehicle_body_typeFREIGHT            0.166566   0.219273   0.760 0.447476
Vehicle_body_typeLIMOUSINE (OMNIBUS) 12.490998 227.269630   0.055 0.956169
Vehicle_body_typeMOTORCYCLE        12.104834 324.743781   0.037 0.970266
Vehicle_body_typePICKUP TRUCK      -0.334055   0.116106  -2.877 0.004013 **
Vehicle_body_typeSPECIAL            0.466137   0.314383   1.483 0.138154
Vehicle_body_typeSUBURBAN           0.122396   0.083444   1.467 0.142428
Vehicle_body_typeTAXI               1.207336   0.580345   2.080 0.037491 *
Vehicle_body_typeUTILITY           -0.452740   0.564038  -0.803 0.422162
RegistrationMOTORCYCLE registration -12.444581 324.743730  -0.038 0.969432
RegistrationOMNIBUS registration    0.424580   0.219744   1.932 0.053340 .
RegistrationPASSENGER registration -0.037163   0.128599  -0.289 0.772593
RegistrationSPECIAL registration   -0.162993   0.189656  -0.859 0.390113
ActionBacking                       1.587195   0.289055   5.491 4.00e-08 ***
Actionchang lane & merge & overtake 0.627571   0.213097   2.945 0.003230 **
Actionmake turn                     0.909298   0.210971   4.310 1.63e-05 ***
Actionother                         0.377700   0.264563   1.428 0.153396
Actionpark & slow & stop            1.576577   0.221262   7.125 1.04e-12 ***
Actionstraight                      0.208770   0.197105   1.059 0.289516
Type2 axle single unit box truckN/A 25.604768 397.667997   0.064 0.948662
TypeNot Entered                    13.190102 324.743831   0.041 0.967601
DirectionNorth                     -0.180612   0.063172  -2.859 0.004249 **
DirectionNortheast                 -0.507140   0.134102  -3.782 0.000156 ***
DirectionNorthwest                 -0.442116   0.145165  -3.046 0.002322 **
DirectionSouth                     -0.214913   0.064178  -3.349 0.000812 ***
DirectionSoutheast                 -0.458567   0.144575  -3.172 0.001515 **
DirectionSouthwest                 -0.576922   0.146870  -3.928 8.56e-05 ***
DirectionWest                      -0.136031   0.064953  -2.094 0.036232 *
FuelDiesel                         12.325318 324.743803   0.038 0.969724
```

```
FuelElectric                          25.153347 459.257417   0.055 0.956322
FuelFlex                              12.850992 324.746941   0.040 0.968434
FuelGas                               12.417619 324.743746   0.038 0.969498
FuelNone                              11.839824 324.746205   0.036 0.970917
Year20 YEAR                           -0.080822   0.048339  -1.672 0.094528 .
Year3 YEAR                             0.112850   0.130996   0.861 0.388975
Year5 YEAR                             0.005362   0.082192   0.065 0.947986
YearMORE THAN 20 YEAR                  0.105038   0.118331   0.888 0.374721
Occupants0                           -12.783661 324.744584  -0.039 0.968599
Occupants1                            -0.449650   0.755920  -0.595 0.551952
Occupants2                            -0.383687   0.757186  -0.507 0.612346
Occupants3                            -0.447578   0.760801  -0.588 0.556332
Occupants4                            -0.538527   0.770882  -0.699 0.484811
Occupants5                            -0.429289   0.800006  -0.537 0.591539
Occupants6                            -1.108181   0.925071  -1.198 0.230940
Occupants7                            -1.133913   1.447155  -0.784 0.433306
Engine                                 0.055507   0.019373   2.865 0.004167 **
Vehicle_MakeGERMAN                     0.240610   0.345451   0.697 0.486110
Vehicle_MakeITALY                      1.828805   1.223290   1.495 0.134917
Vehicle_MakeJAPAN                      0.318250   0.332612   0.957 0.338658
Vehicle_MakeKOREA                      0.187596   0.345321   0.543 0.586956
Vehicle_MakeOTHER                      0.421696   0.380414   1.109 0.267637
Vehicle_MakeUSA                        0.133829   0.330699   0.405 0.685708
Contri1HUMAN                          -0.318558   0.165495  -1.925 0.054244 .
Contri1VEHICLE                         0.119861   0.295129   0.406 0.684646
Contri2HUMAN                          -0.336124   0.181395  -1.853 0.063884 .
Contri2VEHICLE                        -0.216562   0.386472  -0.560 0.575236
Contri1HUMAN:Contri2HUMAN              0.647048   0.188604   3.431 0.000602 ***
Contri1VEHICLE:Contri2HUMAN            0.365011   0.362487   1.007 0.313952
Contri1HUMAN:Contri2VEHICLE            0.538224   0.411772   1.307 0.191181
Contri1VEHICLE:Contri2VEHICLE          0.580893   0.546090   1.064 0.287451
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 12671  on 9157  degrees of freedom
Residual deviance: 12148  on 9095  degrees of freedom
AIC: 12274
```

Stepwise selected model: Event ~ Vehicle_body_type + Registration + Action + Type + Direction + Engine + Vehicle_Make + Contri1 + Contri2 + Contri1:Contri2
Exponentiated coefficients:

(Intercept)	Vehicle_body_type4 DOOR SEDAN
8.476727e-07	1.055211e+00
Vehicle_body_typeBUS (OMNIBUS)	Vehicle_body_typeCONVERTIBLE
1.904808e+00	7.470744e-01
Vehicle_body_typeFREIGHT	Vehicle_body_typeLIMOUSINE (OMNIBUS)
1.176082e+00	1.876111e+05
Vehicle_body_typeMOTORCYCLE	Vehicle_body_typePICKUP TRUCK
1.939645e+05	7.215762e-01
Vehicle_body_typeSPECIAL	Vehicle_body_typeSUBURBAN
1.462111e+00	1.127484e+00
Vehicle_body_typeTAXI	Vehicle_body_typeUTILITY
3.434772e+00	6.165203e-01
RegistrationMOTORCYCLE registration	RegistrationOMNIBUS registration
3.758866e-06	1.622724e+00
RegistrationPASSENGER registration	RegistrationSPECIAL registration
9.737593e-01	8.649628e-01
ActionBacking	Actionchang lane & merge & overtake
4.934598e+00	1.892382e+00
Actionmake turn	Actionother
2.526148e+00	1.479002e+00
Actionpark & slow & stop	Actionstraight
4.892905e+00	1.241549e+00
Type2 axle single unit box truckN/A	TypeNot Entered
1.366157e+11	5.758342e+05
DirectionNorth	DirectionNortheast
8.374894e-01	6.069532e-01
DirectionNorthwest	DirectionSouth
6.469950e-01	8.079212e-01
DirectionSoutheast	DirectionSouthwest
6.370037e-01	5.610141e-01
DirectionWest	Engine
8.720806e-01	1.055908e+00
Vehicle_MakeGERMAN	Vehicle_MakeITALY
1.259638e+00	6.170529e+00
Vehicle_MakeJAPAN	Vehicle_MakeKOREA
1.361042e+00	1.212072e+00
Vehicle_MakeOTHER	Vehicle_MakeUSA
1.485525e+00	1.117738e+00
Contri1HUMAN	Contri1VEHICLE
7.199685e-01	1.103588e+00
Contri2HUMAN	Contri2VEHICLE
7.082913e-01	7.951094e-01
Contri1HUMAN:Contri2HUMAN	Contri1VEHICLE:Contri2HUMAN
1.928788e+00	1.441752e+00
Contri1HUMAN:Contri2VEHICLE	Contri1VEHICLE:Contri2VEHICLE
1.734387e+00	1.822058e+00

(Intercept)	Vehicle_body_type4 DOOR SEDAN
4.662499e-01	1.047032e+00
Vehicle_body_typeALL TERRAIN VEHICLE	Vehicle_body_typeBUS (OMNIBUS)
1.230660e-05	1.041441e+00
Vehicle_body_typeCONVERTIBLE	Vehicle_body_typeFREIGHT
1.333053e+00	1.887124e+00
Vehicle_body_typeMOTORCYCLE	Vehicle_body_typePICKUP TRUCK
7.846356e-01	9.019915e-01
Vehicle_body_typeSPECIAL	Vehicle_body_typeSUBURBAN
1.296418e+00	1.168746e+00
Vehicle_body_typeTAXI	Vehicle_body_typeUTILITY
1.952458e+00	1.679164e+00
RegistrationMOTORCYCLE registration	RegistrationOMNIBUS registration
NA	2.394622e+00
RegistrationPASSENGER registration	RegistrationSPECIAL registration
1.048418e+00	2.012538e+05
RegistrationSPECIAL registration	ActionBacking
8.039063e-01	3.490864e+00
Actionchang lane & merge & overtake	Actionmake turn
1.694331e+00	1.712455e+00
Actionother	Actionpark & slow & stop
1.522836e+00	4.332444e+00
Actionstraight	Year20 YEAR
1.016316e+00	9.613421e-01
Year3 YEAR	Year5 YEAR
1.203646e+00	1.105235e+00
YearMORE THAN 20 YEAR	Engine
9.402188e-01	1.071118e+00
Vehicle_MakeGERMAN	Vehicle_MakeITALY
6.133252e-01	2.961509e-01
Vehicle_MakeJAPAN	Vehicle_MakeKOREA
6.078139e-01	5.619218e-01
Vehicle_MakeOTHER	Vehicle_MakeUSA
3.686600e-01	4.613468e-01
Contri1HUMAN	Contri1VEHICLE
1.373169e+00	1.150156e+00
Contri2HUMAN	Contri2VEHICLE
1.263257e+00	1.258930e+00
Contri1HUMAN:Contri2HUMAN	Contri1VEHICLE:Contri2HUMAN
1.140996e+00	1.560906e+00
Contri1HUMAN:Contri2VEHICLE	Contri1VEHICLE:Contri2VEHICLE
1.033082e+00	3.214260e+00

Prediction & Model efficience

```
> glm.probs<-predict(out4,type='response')
> glm.probs[1:10]
        1         2         3         4         5         6         7         8
0.4502724 0.4894694 0.3256510 0.4862976 0.3961110 0.5367180 0.3945527 0.4207833
        9        10
0.3072259 0.4773238
> contrasts(as.factor(Event))
  1
0 0
1 1
> glm.pred<-rep('1',1250)
> glm.pred[glm.probs>0.5]<-'0'
> table(glm.pred,Event)
        Event
glm.pred    0    1
       0 1085 1724
       1  594  396
```