ISyE 6416 – Basic Statistical Methods - Fall 2015

Bonus Project: "Big" Data Analytics Proposal

Motor Vehicle Accident Analysis

Team Member

Xueting Wang

Hao Wei

Hongao Yang

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 causing the vehicle accident factors and their weights by establishing suitable models, such as logistic regression model and multivariable regression model, and using the official data from New York Department of vehicle.

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 t	Type / Axles	Direction of	Fuel Type	Vehicle Year	State of Regi	Number of C	Engine Cylinc	Vehicle Make	Contributing	Contributing	Contributing	Contributing	Event Type
2	2012	9355479	SUBURBAN	PASSENGER (	Going Straigl	Not Entered	East	Gas	2004	NY	2	8	DODGE	HUMAN	Not Entered	HUMAN	Not Entered	Not Entered
3	2012	9355480	SUBURBAN	PASSENGER (	Going Straigl	Not Entered	East	Gas	2002	NY	1	6	FORD	HUMAN	Not Entered	HUMAN	Not Entered	Not Entered
4	2012	9967254	SUBURBAN	PASSENGER (	Going Straigl	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 Straigl	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 Straigl	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, Collisior
10	2012	9967389	4 DOOR SED/	OMNIBUS - T	Going Straigl	Not Entered	South	Gas	2003	NY	4	8	LINCO	HUMAN	Reaction to (	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.

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

Methodology

First, raw data treatment. As the database is really large, we delete the records that contains unclear data and missing data. Divide the data by the year the accident happened. Convert the factors that are not numerical into dummy codes. For example, accident 0=Non-fatal, 1=Fatal; Vehicle Body 1=2 Dr Sedan, 2=4 Dr Sedan etc.

Second, build multiple linear regression and logistic regression model.

Third, model analysis. Test subsets of coefficients, do diagnostics (outliers, influential points, non constant variances, nonlinearity…), and analysis influential factors that significant enough, choose optimal models.

Fourth, further interpretation, like classification and prediction.

Expected Results

We would pick up some significant factors, which mainly account for a severe accident; therefore, delete some unimportant feature factors that would not result in severe car accidents. Besides, we could know the influence of each factor on the accident due to different weight of them. From this report, we could advise drivers to know the probability of occurring a severe car accident by simply evaluating their vehicles' condition and driving habits. In conclusion, people would benefit from avoiding some unnecessary car accidents by carefully examine their car condition and some driving habits. Moreover, auto insurance companies may also benefit from it by learning the likelihood of a car accident of their clients.