**Aegis School of Business & Data Science**

Apr 2017

**Project Report**

**Traffic Accident Analysis and Modeling**

**Johnson Varghese Parag Mehta Ranjeet Singh Rawat**

Table of Contents

[Introduction 3](#_Toc478482411)

[Problem Statement 3](#_Toc478482412)

[Solution Approach 3](#_Toc478482413)

[Visualization 4](#_Toc478482414)

[Modeling 6](#_Toc478482415)

# Introduction

Over 35000 persons were killed in traffic accidents in the US in 2015. This marked a 7% increase in deaths over 2014. This breaks a recent decreasing trend in accident fatalities. The US Department of Transportation has been releasing data related to vehicle accidents every year since 1975. This data is collected and reported using a system known as Fatality Analysis Reporting System (FARS)( <ftp://ftp.nhtsa.dot.gov/fars/> ). The purpose of releasing this data is to seek participation from organizations, individuals, non-profits etc to analyze and come up with recommendations to reduce the number of deaths on the roads.

# Problem Statement

We looked at this problem from two perspectives:

1. Analyze the data for the last 40 years to understand the various mitigating and aggravating factors that could be causing accidents leading to deaths including geographical profiling and time series analysis.
2. Build models that could predict the number of deaths or the severity of injuries in case an accident occurs given a certain set of controllable parameters with a view to use those parameters to reduce the injuries and deaths.

# Solution Approach

The data was downloaded from the FARS site in the form of dbf files. 10 years data was assigned to each team member for analysis. Each year, the number of data files varied, from a low of 4 in 2009 to a max of 26 in 2015. In order to merge the data and do a combined analysis, we chose the 3 files that were common to all the years, i.e. ACCIDENT, VEHICLE and PERSON.

The data files were read into R studio. For each file, the number and names of columns in each year was recorded and compared. For e.g. the number of columns in VEHICLE file ranged from 100 to 127 over the 10 years. Finally the columns that were common to all years were identified. In a few cases, the columns had been renamed, hence it was necessary to change those names to a common one. In most cases, the columns had been reordered over the years, hence they had to be arranged in a common sequence before merging. Finally after all this processing, the files that emerged were:

* ACCIDENT (42 features)
* VEHICLE (68 features)
* PERSON (40 features)

The first step of exploratory analysis was done in Tableau and R. Insights were drawn based on this analysis.

The second step was to create the models. The first model was based on the VEHICLE file which had 450k observations. This was then divided into the training and test sets. The variable ‘DEATHS’ was chosen as the response variable. The predictor variables were chosen from the remaining. Decision tree method was chosen as the technique for modeling due to the limited computing resources at our disposal and the fact that this technique is the most efficient in memory usage as compared to other techniques like SVM, random forest, logistic regression etc. The second model was based on the PERSON file which had 800k observations. This was again divided into the training and test set. In this case, ‘INJ SEV’ was chosen as the response variable and the predictors were chosen from the rest. Here again, Decision tree method was chosen for modeling.

# Visualization

Visualization was done in Tableau and R. The key insights from this exercise were:

* The highest number of accidents and fatalities occur in Texas, California and Florida
* The most number of accidents in a week occur during the nights of Fri-Sat and Sat-Sun
* The highest number of accidents occur in August and the lowest in February
* Most accidents occur on State Highways compared to other type of roads (US Highways, Interstate etc)
* Roads having most accidents: I-10, I-95, I-40
* Close to half (48%) of accidents happen in daylight
* Over 80% accidents happen in clear weather
* In 55% of cases, the driver was the owner of the vehicle leading to 58% of deaths, in 28% of cases, the driver was not the owner of the vehicle leading to 32% of deaths and in 10.65% of cases, the vehicle was company owned leading to 4.65% of deaths. This provides and interesting observation that company owned vehicles were safer for travel.
* Chevrolet, Ford, Honda, Toyota are the brands most involved in accidents. Out of these, Ford had the lowest proportion of fatalities and Honda had the highest. This gives an insight into the safety features incorporated in various brands of vehicles.
* Number of deaths depends on the type of vehicle. *Motorcycles* had the highest mortality rate causing almost 1 death per vehicle while *trucks* had the lowest rate of 0.16 deaths per vehicle. Other types that came in between these two were *standard pickup, compact utility, 4 door sedan and 2 door sedan* in the increasing order of mortality rate. This leads to the conclusion that death rate is inversely proportional to the size of the vehicle.
* No drunk drivers were involved in two thirds of accidents. However, comparing the numbers of drunk and non drunk drivers involved in accidents, it was found that only 32% of drunk drivers survived compared to 62% of sober drivers.
* Most of the vehicles involved in accidents had drivers holding valid permanent licenses. However, here again, the mortality rate was much higher for those with suspended or no licenses.
* Mortality rate was much lower for commercial drivers compared to non commercial drivers. This could mean that either commercial drivers were practicing safer driving OR that most commercial drivers were driving bigger vehicles (trucks, buses, pickups etc)
* A strange relation was shown in terms of airbag deployment and injury severity. Contrary to expectation, it showed that airbag deployment led to more fatal injuries than in cases where no airbags were deployed.
* However, usage of seat belts and shoulder belts were shown to improve safety. When used, they led to lesser deaths
* More than 90% of people involved in accidents were motor vehicle occupants, as expected. Only 6% were pedestrians and 1% were cyclists.
* For motor vehicle occupants, the rear seat was found to be safer as lesser proportion of people occupying the rear seat were killed compared to those occupying the front seat. Also, the right side of the vehicle was found to be safer than the left side (driver side)
* 66% of persons involved in accidents were male. This could mean either that women are safer drivers OR that many more men drive compared to women
* Most of the accidents involve people in the 20-30 age group.
* 69% of those killed in accidents were White, 11% were Black. This could give some insights into vehicle ownership patterns among various communities.
* There is almost no difference in survival rates when victims were evacuated by air as compared to evacuation by ground.

# Modeling

**Model 1: To predict the number of deaths in a vehicle if it were to meet with an accident**

The vehicle data files for 10 years were loaded. Each one had between 100 to 127 features.

> ncol(vehicle2006)

[1] 100

> ncol(vehicle2007)

[1] 106

> ncol(vehicle2008)

[1] 105

> ncol(vehicle2009)

[1] 113

> ncol(vehicle2010)

[1] 116

> ncol(vehicle2011)

[1] 127

> ncol(vehicle2012)

[1] 126

> ncol(vehicle2013)

[1] 102

> ncol(vehicle2014)

[1] 102

> ncol(vehicle2015)

[1] 102

The first step was to identify the common elements across the 10 files and merge them into one. This needed the columns to be rearranged and in some cases, renamed:

> vehicle2015 = vehicle2015[,c(5,19,47,41,66,101,55,102,70,61,71,63,49,60,77, 78,39,10,13,54,35,68,67,69,62,64,65,79,80,57,18,16,11,38,20,17,7,15,72,74,76,75,73,14,52,48,2,1,34,50,51,12,40,4,3,21,22,23,24,25,26,27,28,29,30,31,32,33)]

:

:

:

> vehicle2006 = vehicle2006[,c(16,19,52,29,73,44,35,69,92,68,91,95,31,38,86, 94,53,97,99,32,23,75,74,76,70,72,71,85,93,43,56,17,98,61,57,18,96,21,80,82,84,83,81,20,22,30,55,14,26,24,34,54,27,100,15,1,2,3,4,5,6,7,8,9,10,11,12,13)]

> colnames(vehicle2006)[1] = "NUMOCCS"

> vehicle\_all = rbind(vehicle2006,vehicle2007,vehicle2008,vehicle2009, vehicle2010,vehicle2011,vehicle2012,vehicle2013,vehicle2014,vehicle2015)

68 features were found to be common to all 10 files. These were:

> colnames(vehicle\_all)

[1] "NUMOCCS" "BODY\_TYP" "BUS\_USE" "CARGO\_BT" "CDL\_STAT" "DEATHS" "DEFORMED"

[8] "DR\_DRINK" "DR\_HGT" "DR\_PRES" "DR\_WGT" "DR\_ZIP" "EMER\_USE" "FIRE\_EXP"

[15] "FIRST\_MO" "FIRST\_YR" "GVWR" "HARM\_EV" "HIT\_RUN" "IMPACT1" "J\_KNIFE"

[22] "L\_COMPL" "L\_ENDORS" "L\_RESTRI" "L\_STATE" "L\_STATUS" "L\_TYPE" "LAST\_MO"

[29] "LAST\_YR" "M\_HARM" "MAK\_MOD" "MAKE" "MAN\_COLL" "MCARR\_ID" "MOD\_YEAR"

[36] "MODEL" "MONTH" "OWNER" "PREV\_ACC" "PREV\_DWI" "PREV\_OTH" "PREV\_SPD"

[43] "PREV\_SUS" "REG\_STAT" "ROLLOVER" "SPEC\_USE" "ST\_CASE" "STATE" "TOW\_VEH"

[50] "TRAV\_SP" "UNDERIDE" "UNITTYPE" "V\_CONFIG" "VE\_FORMS" "VEH\_NO" "VIN"

[57] "VIN\_1" "VIN\_2" "VIN\_3" "VIN\_4" "VIN\_5" "VIN\_6" "VIN\_7"

[64] "VIN\_8" "VIN\_9" "VIN\_10" "VIN\_11" "VIN\_12"

Out of these 68, we narrowed down to those columns which were likely to have some predictive value. i.e the number of deaths in the vehicle could be linked to them. Features which were being used for identification (e.g. ST\_CASE, VEH\_NO) and those which were a result of the accident (e.g. HARM\_EV, IMPACT1, M\_HARM, MAN\_COLL, FIRE\_EXP, DEFORMED etc) were removed since we are interested in features that would be known at the beginning of the journey. We identified 31 columns:

> colnames(vehicle\_all)

[1] "NUMOCCS" "BODY\_TYP" "BUS\_USE" "CARGO\_BT" "CDL\_STAT" "DEATHS" "DR\_DRINK"

[8] "DR\_HGT" "DR\_WGT" "DR\_ZIP" "EMER\_USE" "GVWR" "L\_COMPL" "L\_ENDORS"

[15] "L\_RESTRI" "L\_STATE" "L\_STATUS" "L\_TYPE" "MAK\_MOD" "MAKE" "MODEL"

[22] "MONTH" "OWNER" "PREV\_ACC" "PREV\_DWI" "PREV\_OTH" "PREV\_SPD" "PREV\_SUS"

[29] "SPEC\_USE" "TRAV\_SP" "UNITTYPE"

We then identify the highly correlated features using the caret package:

library(caret)

vehcor = cor(vehicle\_all[,1:31])

The correlations between the features are shown below. By removing the yellow highlighted columns, we can eliminate the highly correlated variables.



After removing the correlated features, we divide the data into training and test sets in the ratio 70:30

> vehicle\_all = vehicle\_all[,-c(3,19:21,29)]

> samplesize = floor(0.7\*nrow(vehicle\_all))

> sampleindexes = sample(1:nrow(vehicle\_all),samplesize)

> veh\_train = vehicle\_all[sampleindexes,]

> veh\_test = vehicle\_all[-sampleindexes,]

We then use a decision tree to build our model. The cost complexity factor and minsplit parameter was arrived at after experimenting with a range of values to get the best accuracy.

> library(rpart)

> vehtree = rpart(DEATHS ~ ., data = veh\_train, method = "class", control=rpart.control(minsplit=700,cp=0.00001))

We calculate the accuracy of the model as follows:

predvals = predict (vehtree, veh\_test[,-5],type = "class")

table (predvals, veh\_test[,5])

The accuracy of the model is determined to be 69.18%

We try to improve the model by looking at the variable importance indicated by the model to remove more features.

Variable importance

BODY\_TYP DR\_DRINK DR\_ZIP GVWR CARGO\_BT L\_STATE TRAV\_SP OWNER L\_STATUS

15 13 11 7 7 7 6 6 5

L\_TYPE NUMOCCS DR\_WGT DR\_HGT CDL\_STAT MONTH L\_ENDORS L\_RESTRI PREV\_ACC

4 3 3 2 2 1 1 1 1

L\_COMPL PREV\_OTH PREV\_SPD

1 1 1

We rebuild the model by retaining only the top 10 features and discarding the rest.

> veh\_train = veh\_train[,-c(1,4,7:8,10,12:14,18,20:24,26)]

> veh\_test = veh\_test[,-c(1,4,7:8,10,12:14,18,20:24,26)]

Variable importance

BODY\_TYP DR\_DRINK DR\_ZIP GVWR CARGO\_BT L\_STATE OWNER TRAV\_SP L\_STATUS

18 16 15 9 9 8 8 7 6

L\_TYPE

5

The accuracy of the model is now 69.17%. All the features are now important, so we don’t remove any more features.

**Model 2: To predict the injury severity of a person involved in an accident**

The person data files for 10 years were loaded. Each one had between 68 to 93 features.

> ncol(person2006)

[1] 75

> ncol(person2007)

[1] 75

> ncol(person2008)

[1] 75

> ncol(person2009)

[1] 77

> ncol(person2010)

[1] 82

> ncol(person2011)

[1] 93

> ncol(person2012)

[1] 92

> ncol(person2013)

[1] 68

> ncol(person2014)

[1] 68

> ncol(person2015)

[1] 68

The first step was to identify the common elements across the 10 files and merge them into one. This needed the columns to be rearranged and in some cases, renamed:

> person2006 = person2006[,c(1:3,5:16,18:19,27:35,37:39,41:43,46:48,52,55,56,72,73)]

> colnames(person2006)[c(27,28,29)] = c("P\_SF1","P\_SF2","P\_SF3")

:

:

:

> person2015 = person2015[,c(1,4,5,27:29,31,32,34:37,68,38,39,42,43,30,52,55,54,57,58,60,61,67,62:64,53,2,56,17,19,24:26,18,14,15)]

> person\_all = rbind(person2015,person2014,person2013,person2012,person2011, person2010,person2009,person2008,person2007,person2006)

40 features were found to be common to all 10 files. These were:

> colnames(person\_all)

[1] "STATE" "VEH\_NO" "PER\_NO" "AGE" "SEX" "PER\_TYP" "SEAT\_POS"

[8] "REST\_USE" "AIR\_BAG" "EJECTION" "EJ\_PATH" "EXTRICAT" "LOCATION" "DRINKING"

[15] "ALC\_DET" "ALC\_RES" "DRUGS" "INJ\_SEV" "HOSPITAL" "DEATH\_MO" "DEATH\_DA"

[22] "DEATH\_HR" "DEATH\_MN" "LAG\_HRS" "LAG\_MINS" "RACE" "P\_SF1" "P\_SF2"

[29] "P\_SF3" "DOA" "ST\_CASE" "DEATH\_YR" "MAKE" "BODY\_TYP" "ROLLOVER"

[36] "IMPACT1" "FIRE\_EXP" "MAK\_MOD" "HARM\_EV" "MAN\_COLL"

Out of these 40, we narrowed down to those columns which were likely to have some predictive value. i.e the injury severity could be linked to them. Features which were being used for identification (e.g. ST\_CASE, VEH\_NO, PER\_NO) and those which were a result of the accident (e.g. EJECTION, HOSPITAL, IMPACT1, FIRE\_EXP, HARM\_EV, MAN\_COLL etc) were removed since we are interested in features that would be known at the beginning of the journey. The RACE feature could have been interesting as a predictor, however, on examining the data, it was observed that RACE was only being identified for “fatal” cases i.e. for accident survivors, RACE was shown as “Unknown/Not Reported”. At the time of prediction, RACE would always be “known” , thus the model might be skewed to predict it as a fatality. Hence, RACE was also eliminated from the feature set.

We identified 18 columns:

> colnames(person\_all)

[1] "STATE" "AGE" "SEX" "PER\_TYP" "SEAT\_POS" "REST\_USE" "AIR\_BAG"

[8] "LOCATION" "DRINKING" "ALC\_RES" "DRUGS" "INJ\_SEV" "P\_SF1" "P\_SF2"

[15] "P\_SF3" "MAKE" "BODY\_TYP" "MAK\_MOD"

We then identify the highly correlated features using the caret package:

> library(caret)

> personcor = cor(person\_all[,1:18])

We then eliminate the highly correlated features, viz P\_SF2, P\_SF3, DRINKING and MAK\_MOD

|  |
| --- |
| > person\_all = person\_all[,-c(9,14,15,18)] |
|  |
| |  | | --- | |  | |
| > colnames(person\_all)  [1] "STATE" "AGE" "SEX" "PER\_TYP" "SEAT\_POS" "REST\_USE" "AIR\_BAG"  [8] "LOCATION" "ALC\_RES" "DRUGS" "INJ\_SEV" "P\_SF1" "MAKE" "BODY\_TYP" |
|  |
| |  | | --- | |  | |

After removing the correlated features, we divide the data into training and test sets in the ratio 70:30

> samplesize = floor(0.7\*nrow(person\_all))

> sampleindexes = sample(1:nrow(person\_all),samplesize)

> person\_train = person\_all[sampleindexes,]

> person\_test = person\_all[-sampleindexes,]

We then use a decision tree to build our model. The cost complexity factor and minsplit parameter was arrived at after experimenting with a range of values to get the best accuracy.

> library(rpart)

> persontree = rpart(factor (INJ\_SEV) ~ ., data = person\_train, method = "class", control=rpart.control(minsplit=10,cp=0.00001))

We calculate the accuracy of the model as follows:

> predtree = predict (persontree, person\_test[,-11], type = "class")

> table (predtree, person\_test[,11])

The accuracy of the model is determined to be 61.97%

We try to improve the model by looking at the variable importance indicated by the model to remove more features.

Variable Importance

ALC\_RES REST\_USE BODY\_TYP AIR\_BAG SEAT\_POS PER\_TYP AGE MAKE STATE

25 12 11 10 8 7 7 6 5

LOCATION DRUGS SEX

4 3 3

The feature P\_SF1 is not significant. So, we remove it and rebuild the model.

> person\_train = person\_train[,-12]

> person\_test = person\_test[,-12]

Variable Importance

ALC\_RES REST\_USE BODY\_TYP AIR\_BAG SEAT\_POS PER\_TYP AGE MAKE STATE

25 12 11 10 8 7 7 6 5

LOCATION DRUGS SEX

4 3 3

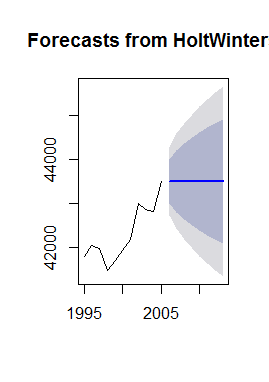
The accuracy of the model is still 61.97%. All the variables are now important. Hence we do not remove any more variables.

**Time series Analysis: Total accidents forecast for next 8 years.**

Original data- 1996-2005. Below are forecasted numbers

The forecasted fatalities for 2006 are about 43499 fatalities, with a 95% prediction interval of (42739, 44260.62).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Forecast year | Forecast | Lo 80 | Hi 80 | Lo 95 | Hi 95 | Actual data |
| 2006 | 43499.95 | 43002.57 | 43997.32 | 42739.27 | 44260.62 | 42708 |
| 2007 | 43499.95 | 42796.58 | 44203.32 | 42424.23 | 44575.66 | 41259 |
| 2008 | 43499.95 | 42638.51 | 44361.39 | 42182.49 | 44817.4 | 37423 |
| 2009 | 43499.95 | 42505.25 | 44494.64 | 41978.69 | 45021.21 | 33883 |
| 2010 | 43499.95 | 42387.85 | 44612.05 | 41799.13 | 45200.76 | 32999 |
| 2011 | 43499.95 | 42281.7 | 44718.19 | 41636.8 | 45363.09 | 32479 |
| 2012 | 43499.95 | 42184.1 | 44815.8 | 41487.53 | 45512.37 | 33782 |
| 2013 | 43499.95 | 42093.25 | 44906.65 | 41348.58 | 45651.31 | 32893 |



**Monthly accidents forecasts taking into account seasonality**

The forecasted fatalities for Jan 2006 are about 3595 fatalities, with a 95% prediction interval of (3011.235, 4176.960).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Forecast Month | Forecast | Lo 80 | Hi 80 | Lo 95 | Hi 95 | Actual data |
| Jan-06 | 3594.097 | 3212.984 | 3975.211 | 3011.235 | 4176.96 | 3216 |
| Feb-06 | 3594.097 | 3074.588 | 4113.607 | 2799.576 | 4388.619 | 2966 |
| Mar-06 | 3594.097 | 2965.978 | 4222.216 | 2633.472 | 4554.722 | 3376 |
| Apr-06 | 3594.097 | 2873.558 | 4314.636 | 2492.128 | 4696.067 | 3498 |
| May-06 | 3594.097 | 2791.714 | 4396.481 | 2366.957 | 4821.237 | 3718 |
| Jun-06 | 3594.097 | 2717.477 | 4470.717 | 2253.423 | 4934.772 | 3726 |
| Jul-06 | 3594.097 | 2649.055 | 4539.14 | 2148.779 | 5039.416 | 3870 |
| Aug-06 | 3594.097 | 2585.262 | 4602.933 | 2051.217 | 5136.978 | 3835 |

