Predicting Drunk Driving

Azka Javaid & Caleb Ki

December 15, 2016

```
## Warning in library(package, lib.loc = lib.loc, character
## logical.return = TRUE, : there is no package called 'ggp
```

Motivation

- Fremont County, Wyoming had 27 fatal traffic accidents involving alcohol per 100,000 people
- Morris County, New Jersey had 0.2 fatal traffic accidents involving alcohol per 100,000 people
- What accounts for the disparity in these traffic fatalities?

Background

- Over 30000 people die in motor vehicle accidents every year
- Alcohol-impaired driving incidents account for about 30% of these deaths
- Cost of alcohol-related crashes generally exceeds the cost of non-alcohol related crashes

Question

- What factors contribute to drunk driving at an individual and socioeconomic county-level?

Data Description

- Primary data comes from the National Highway Traffic Safety Administration (NHTSA) through the Fatality Analysis Reporting System (FARS)
  - Datasets containing information about the vehicle, accident, and people involved
- Supplementary data comes from the U.S. Census Bureau through the American Community Survey (ACS)
  - Provides economic, social, and demographic data at county and state levels

Google BigQuery

- Cloud base serverless analytics data warehouse
- Platform for performing SQL analysis
- Designed to process GB/PB scale data
- Data reading and writing available via Hadoop, Spark and Cloud Dataflow
- Data ingestion abilities available from Google Cloud Storage, Google Cloud Datastore or livestream
- Facilitates collaboration in an infrastructure-less environment

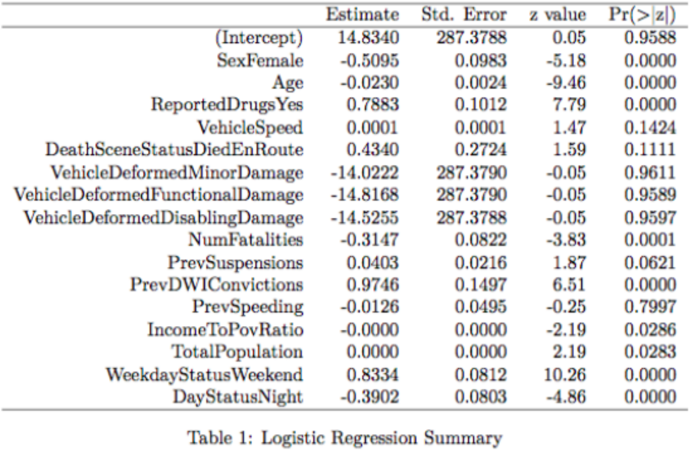
Data Visualization

- Shiny
- Leaflet
- Choroplethr

Variable Description

- Attributes characterized by:
  - Driver: Indicator for drunk driving, Sex, Age, Driver history (past suspensions, DWI and speeding convictions), Indicator for death at scene of accident/en route to a medical facility
  - Vehicle: Vehicle speed prior to crash, Extent of damage
  - Accident: Number of fatalities
  - County-level attributes: Total population, Population by sex, 12-month income to poverty level ratio, Health insurance coverage by sex

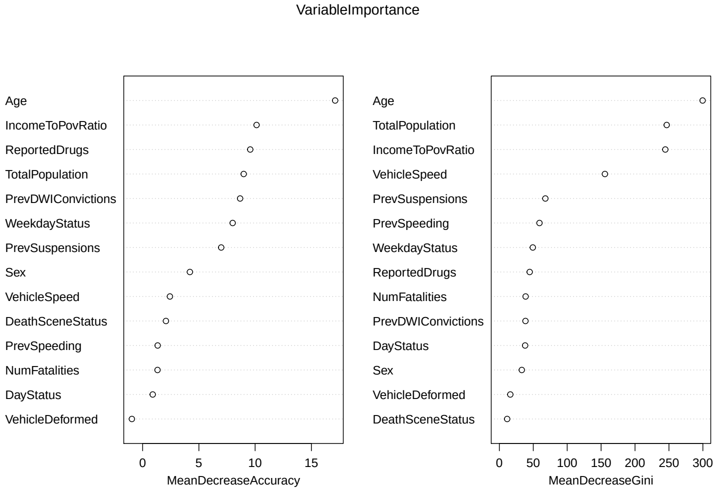
Logistic Regression



	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	14.8340	287.3788	0.05	0.9588
SexFemale	-0.5095	0.0983	-5.18	0.0000
Age	-0.0230	0.0024	-9.46	0.0000
ReportedDrugsYes	0.7883	0.1012	7.79	0.0000
VehicleSpeed	0.0001	0.0001	1.47	0.1424
DeathSceneStatusDiedEnRoute	0.4340	0.2724	1.59	0.1111
VehicleDeformedMinorDamage	-14.0222	287.3790	-0.05	0.9611
VehicleDeformedFunctionalDamage	-14.8168	287.3790	-0.05	0.9589
VehicleDeformedDisablingDamage	-14.5255	287.3788	-0.05	0.9597
NumFatalities	-0.3147	0.0822	-3.83	0.0001
PrevSuspensions	0.0403	0.0216	1.87	0.0621
PrevDWIConvictions	0.9746	0.1497	6.51	0.0000
PrevSpeeding	-0.0126	0.0495	-0.25	0.7997
IncomeToPovRatio	-0.0000	0.0000	-2.19	0.0286
TotalPopulation	0.0000	0.0000	2.19	0.0283
WeekdayStatusWeekend	0.8334	0.0812	10.26	0.0000
DayStatusNight	-0.3902	0.0803	-4.86	0.0000

Table 1: Logistic Regression Summary

Random Forest



Conclusions

Future work

- Extend the study to state level and factor in additional years
- Predict whether drunk driving was involved at an accident level
- Analyze whether time of day and weekday/weekend status affects drunk driving incidence