Predicting Drunk Driving

Azka Javaid & Caleb Ki

December 15, 2016

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)	14.3773	229.5619	0.06	0.9501
SexFemale	-0.4969	0.0986	-5.04	0.0000
Age	-0.0236	0.0024	-9.70	0.0000
ReportedDrugsYes	0.7667	0.1004	7.64	0.0000
VehicleSpeed	0.0000	0.0001	0.59	0.5558
DeathSceneStatusDiedEnRoute	0.3665	0.2659	1.38	0.1682
VehicleDeformedMinorDamage	-13.5596	229.5621	-0.06	0.9529
VehicleDeformedFunctionalDamage	-14.1817	229.5621	-0.06	0.9507
VehicleDeformedDisablingDamage	-13.9746	229.5619	-0.06	0.9515
NumFatalities	-0.3430	0.0826	-4.15	0.0000
PrevSuspensions	0.0376	0.0200	1.88	0.0599
PrevDWIConvictions	1.0307	0.1499	6.87	0.0000
PrevSpeeding	0.0526	0.0490	1.07	0.2826
PrevCrash	-0.0278	0.0623	-0.45	0.6558
IncomeToPovRatio	-0.0000	0.0000	-1.98	0.0481
TotalMale	0.0000	0.0000	2.95	0.0032
TotalFemale	0.0000	0.0000	0.81	0.4188
WeekdayStatusWeekend	0.7486	0.0817	9.16	0.0000
DayStatusNight	-0.3764	0.0801	-4.70	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