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

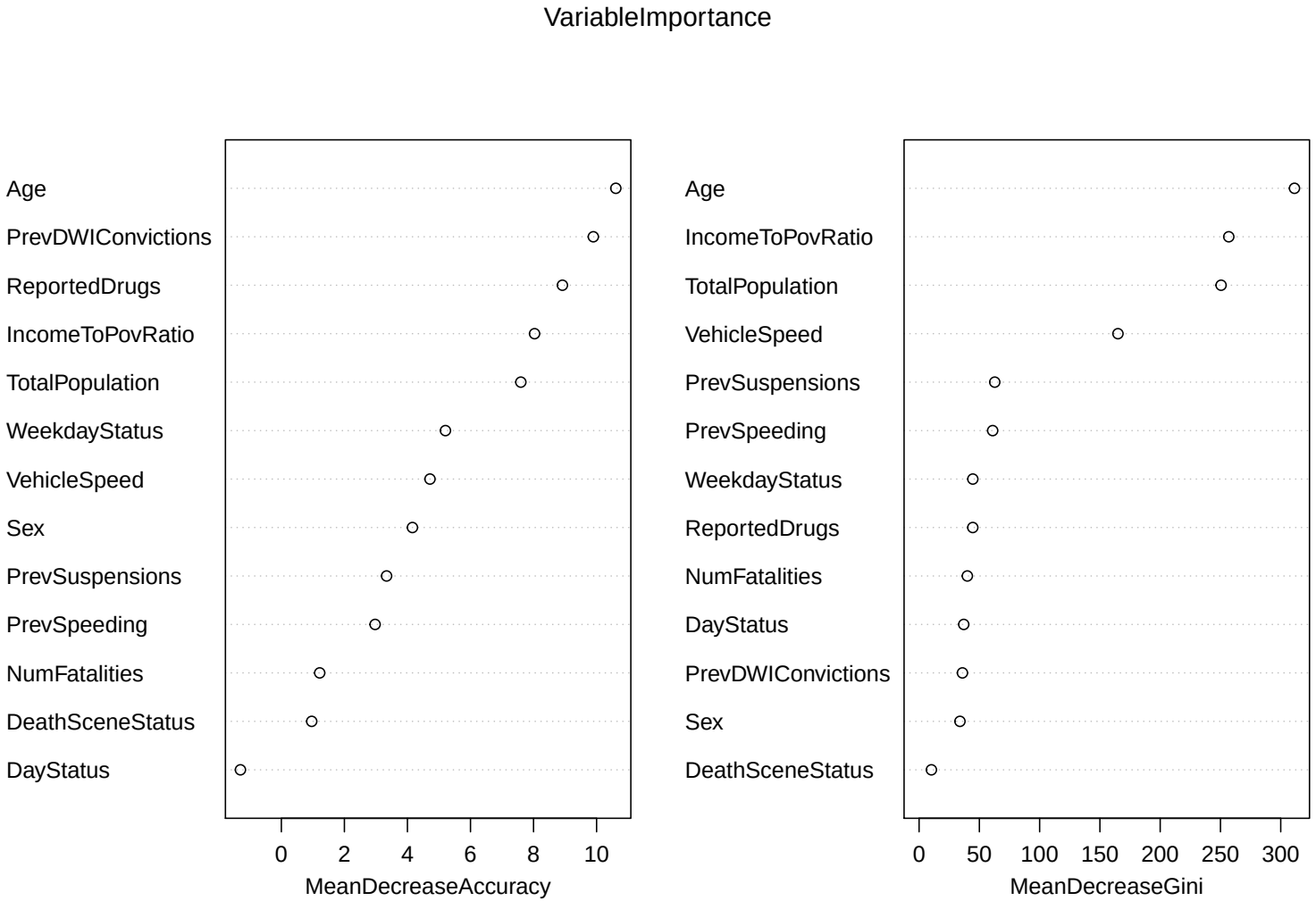
- Attributes characterized by:
  - Driver: Indicator for drunk driving, Sex, Age, Driver history (past suspensions, DWI and speeding convictions), Indicator for death at scene of accident/en route to a medical facility
  - Vehicle: Vehicle speed prior to crash, Extent of damage
  - Accident: Number of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

Logistic Regression

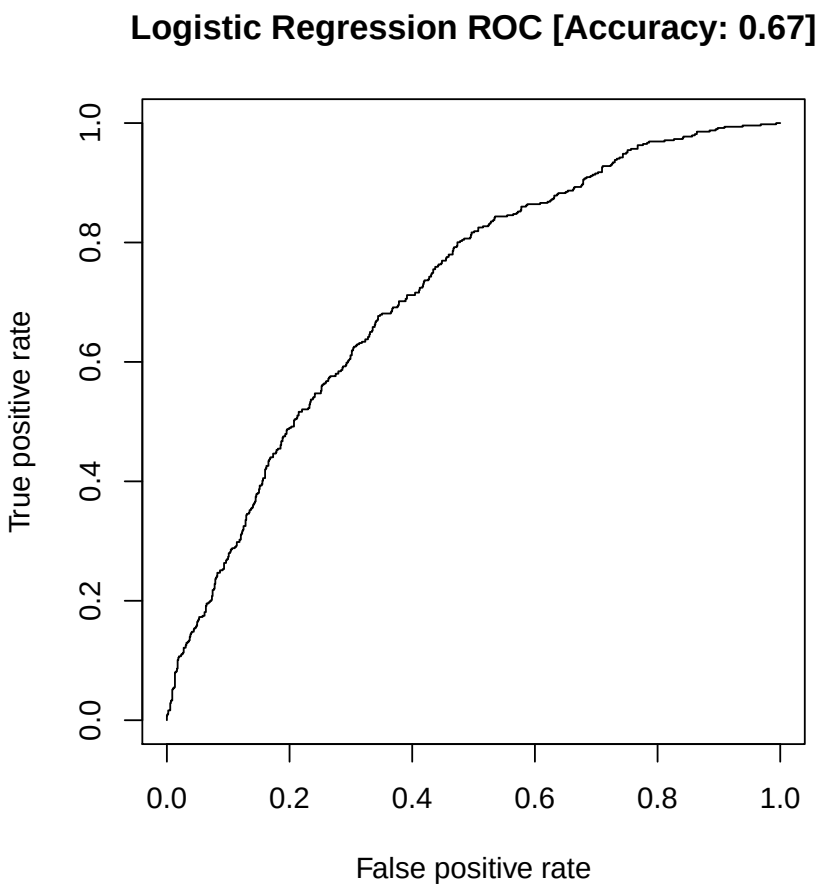
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.1773	0.1669	1.06	0.2879
SexFemale	-0.5373	0.0997	-5.39	0.0000
Age	-0.0217	0.0024	-9.06	0.0000
ReportedDrugsYes	0.7695	0.1002	7.68	0.0000
VehicleSpeed	-0.0000	0.0001	-0.05	0.9615
DeathSceneStatusDiedEnRoute	0.5412	0.2715	1.99	0.0462
NumFatalities	-0.1940	0.0798	-2.43	0.0151
PrevSuspensions	0.0291	0.0200	1.46	0.1450
PrevDWIConvictions	0.7924	0.1384	5.72	0.0000
PrevSpeeding	-0.0037	0.0484	-0.08	0.9392
IncomeToPovRatio	-0.0000	0.0000	-1.55	0.1200
TotalPopulation	0.0000	0.0000	1.56	0.1195
WeekdayStatusWeekend	0.8557	0.0806	10.62	0.0000
DayStatusNight	-0.3342	0.0797	-4.20	0.0000

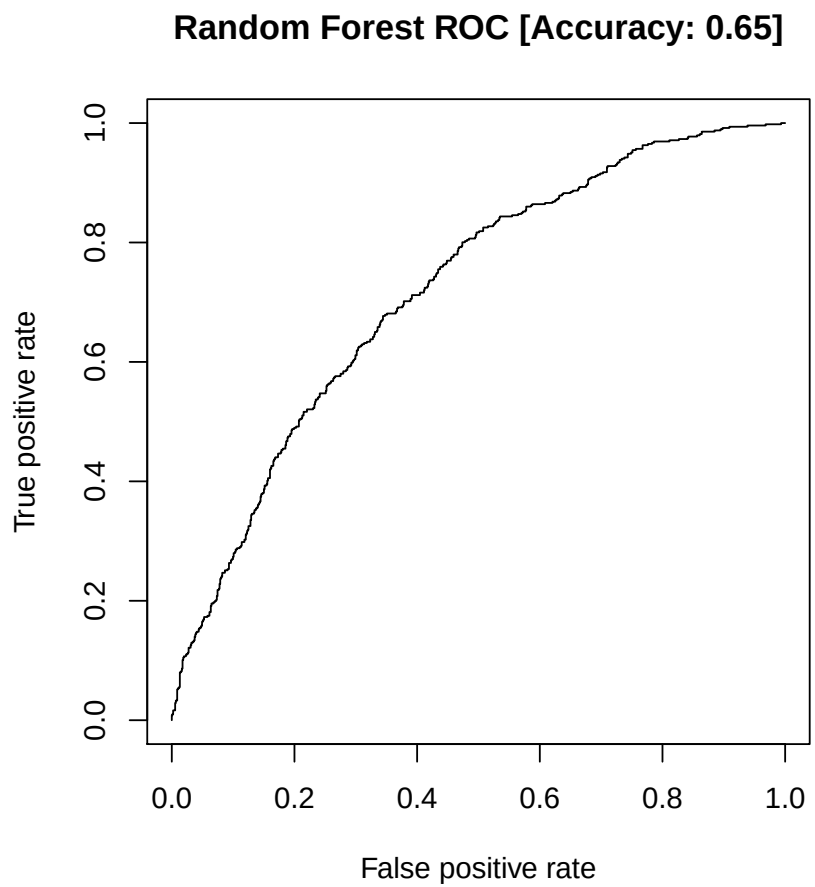
Table 1: Logistic Regression Summary

Random Forest



Model Assessment





Conclusions

- Both models agree that age, previous DWI, reported Drugs, and sex are important predictors
- Total population and income to poverty ratio in a county might help explain the discrepancy in fatal drunk driving incidents
- Whether it was night or not and whether it was the weekend or not can also help predict whether a driver was drunk

Future work and limitations

- Did not account for correlation between observations
  - Use GEE Model
- Extend the study to state level and factor in additional years
- Predict whether drunk driving was involved at an accident level