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, 12-month income to poverty level ratio

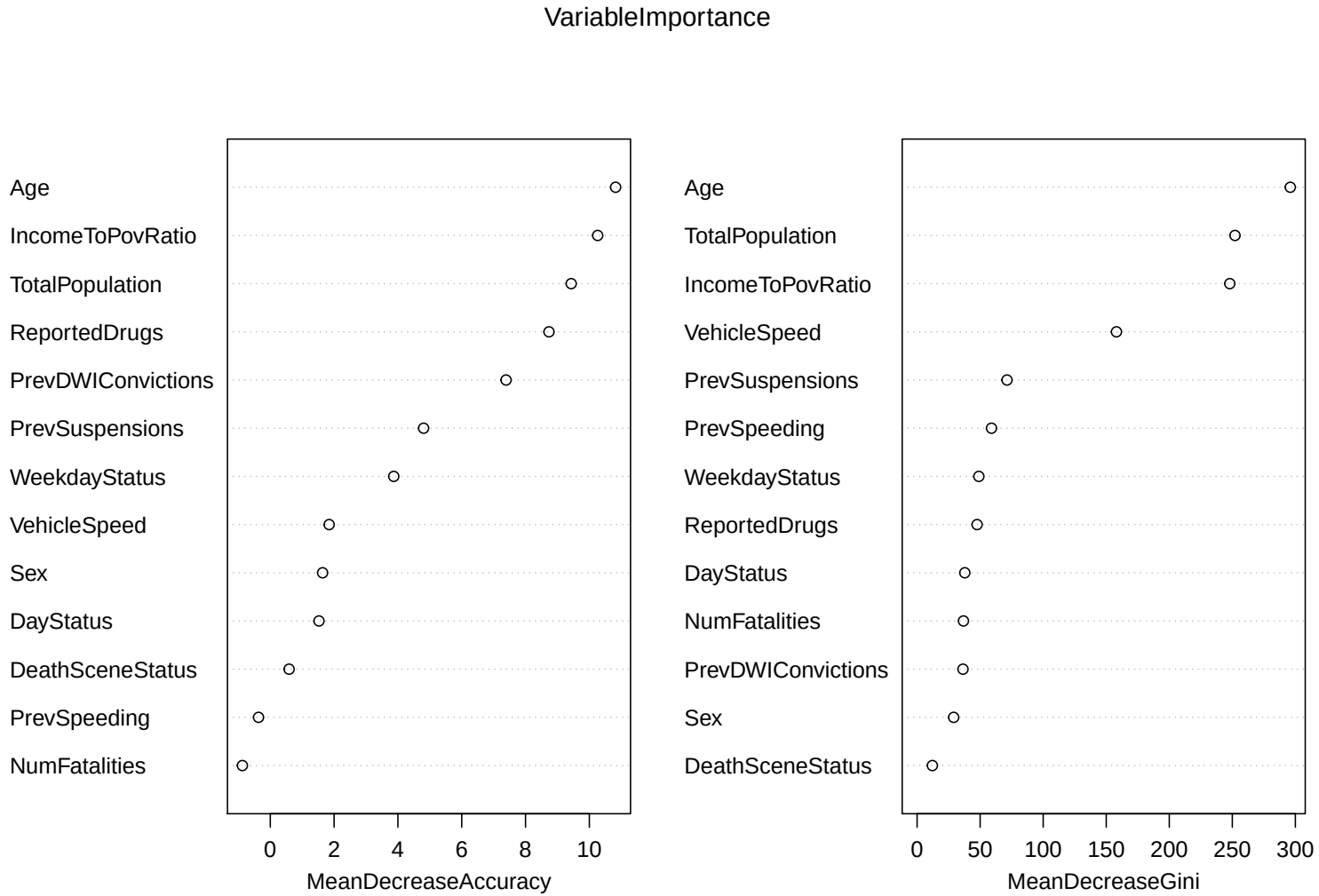
Logistic Regression

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.389221	0.171	2.270	0.023
SexFemale	-0.474642	0.099	-4.810	0.000
Age	-0.023547	0.002	-9.772	0.000
ReportedDrugsYes	0.749694	0.101	7.418	0.000
VehicleSpeed	0.000036	0.000	0.427	0.669
DeathSceneStatusDiedEnRoute	0.178140	0.262	0.680	0.497
NumFatalities	-0.334188	0.085	-3.909	0.000
PrevSuspensions	0.023948	0.022	1.104	0.269
PrevDWIConvictions	1.032822	0.153	6.770	0.000
PrevSpeeding	0.004321	0.048	0.090	0.929
IncomeToPovRatio	-0.000018	0.000	-2.700	0.007
TotalPopulation	0.000017	0.000	2.706	0.007
WeekdayStatusWeekend	0.765346	0.081	9.398	0.000
DayStatusNight	-0.372271	0.080	-4.641	0.000

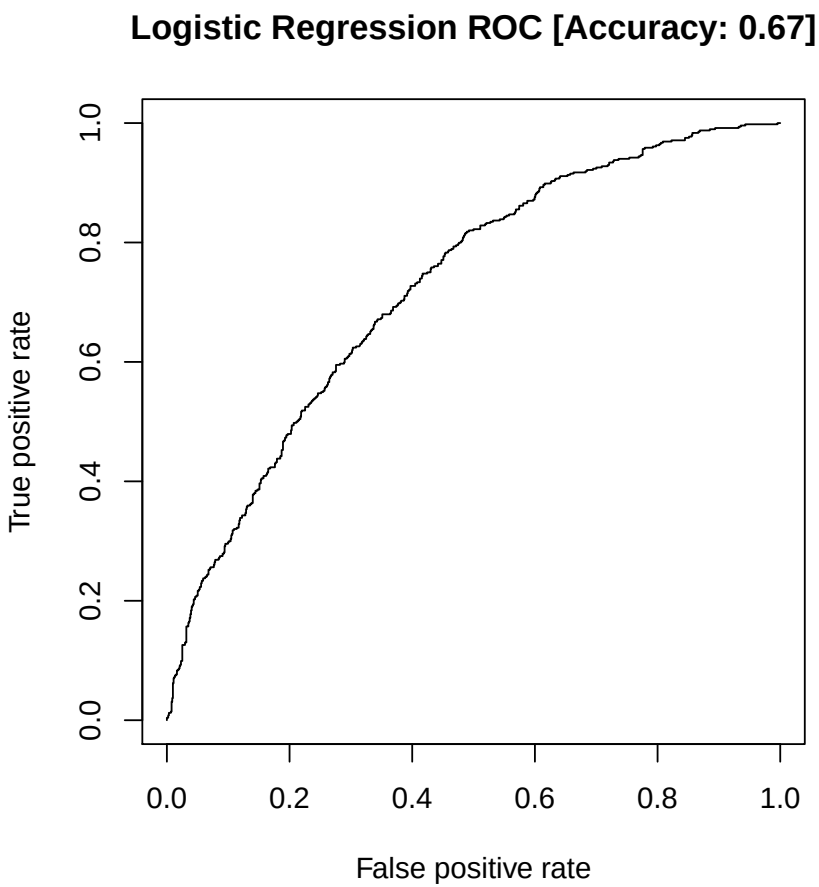
Table 1: Logistic Regression Summary

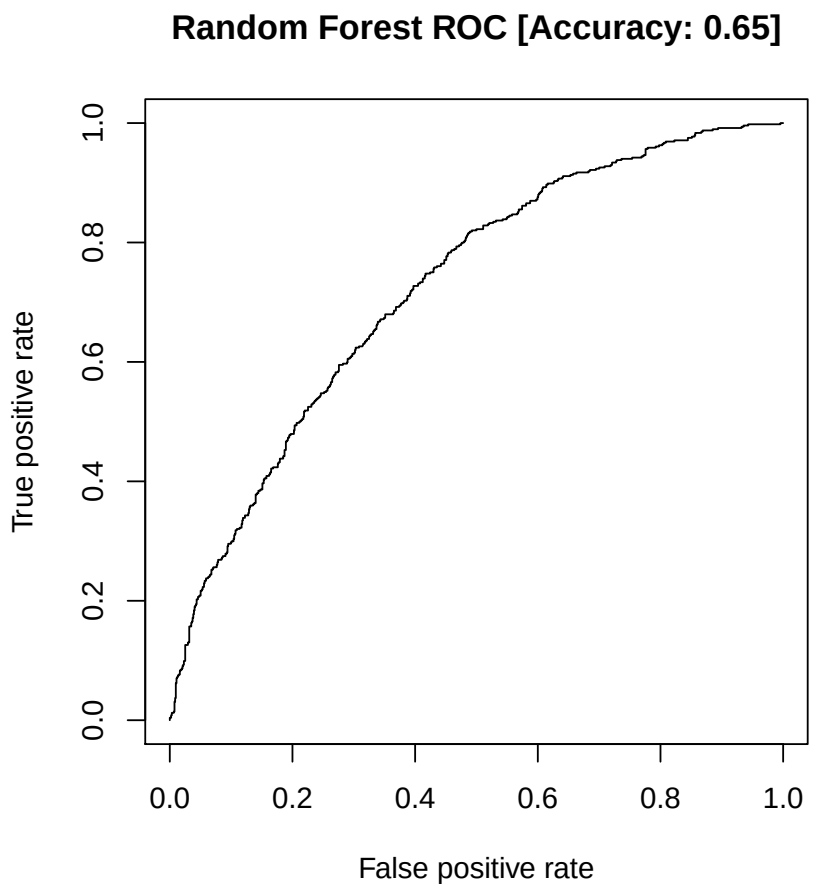
Figure 1:

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
- Weekday status also appears to be an important predictor in predicting drunk driver incidence

Future work and limitations

- Account for correlation between observations by county level
  - Use GEE Model
- Extend the study to state level and factor in additional years
- Predict whether drunk driving was involved at an accident level