Inferring Alcohol Involvement in Fatal Car Accidents with Ensembled Classifiers

Guangsha Shi, Arya Farahi, Chengyu Dai, Cyrus Anderson, Jiachen Huang, Wenbo Shen, Kristjan Greenewald, and Jonathan Stroud
Michigan Data Science Team

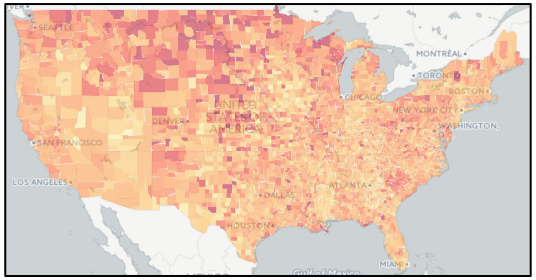
CSE COMPUTER SCIENCE AND ENGINEERING UNIVERSITY OF MICHIGAN
MIDAS MICHIGAN INSTITUTE FOR DATA SCIENCE UNIVERSITY OF MICHIGAN
MDST

The FARS Dataset Challenge

The Fatality Analysis Reporting System (FARS) provides decades of records of fatal car accidents in the United States. The Michigan Data Science Team (MDST) held a competition in which entrants predicted whether or not a drunk driver was involved in each accident. Over 400,000 fatal accident records were provided. Years 2003-10 were used as training data, and 2011-12 and 2013-14 were used for validation and testing, respectively.

Evaluation

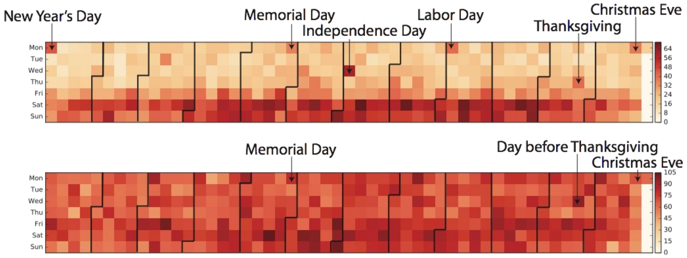
Participants provided probabilistic predictions of drunk driver involvement for each accident in the validation and testing datasets. Submissions were then evaluated using Area Under the Receiver Operating Characteristic (AUROC).



Fraction of fatal accidents that involved a drunk driver by county, 2003-2010.

kaggle in Class

We use Kaggle in Class to distribute datasets and evaluate submissions. First, second, and third place winners were awarded $400, $200, and $100 prizes, respectively. Competition page: `inclass.kaggle.com/c/mdst-fars`



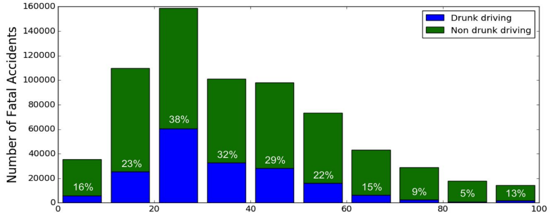
Number of accidents which occurred on each date of the year, (top) involving one or more drunk drivers (bottom) not involving drunk drivers.

Winning Solution

Our top-scoring solution used an ensemble of a neural network and gradient boosted decision trees applied to 94 accident-level features. The neural network contains one hidden layer with 50 sigmoidal neurons and is implemented in Theano. In gradient boosting, we use 500 trees with a maximum depth of 8, implemented using XGBoost. The two methods alone achieved AUC scores of 0.863 and 0.868, respectively. We ensemble using linear weights to improve the performance to 0.869. The ensemble of the two methods effectively prevents overfitting to validation data by combining the outputs of diverse classifiers.

Other Approaches

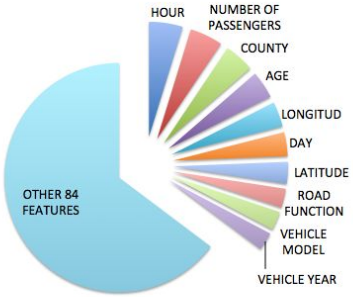
The second and third place teams also employed gradient boosted decision trees in addition to random forest classifiers. Many teams also discovered that including person- and vehicle-level features improved performance significantly.



Fraction of fatal accidents which involved a drunk driver by age group. Drunk driving is most prevalent among adults ages 20-30.

Observations

The most significant predictors of alcohol involvement include crash time, accident location, passenger numbers and ages, and vehicle information. Other features, such as weather and passenger genders were found to be not as informative.



The features which occur with the highest frequency in gradient-boosted decision trees.

Conclusions

It is possible to predict whether or not a drunk driver was involved in a fatal accident using off-the-shelf classifiers applied to data which is readily available at the scene of an accident. Examining these classifiers reveals that there are complex relationships between certain features and the likelihood of alcohol involvement. Ensembles of diverse classifiers improves performance over any single method.

Acknowledgements

We graciously acknowledge the Michigan Institute for Data Science (MIDAS) for funding this competition.