Nationwide: Model Risk Management Assessment/Case Study

Jason Barkeloo, Ph.D.

November 16, 2020

Table of Contents

Code location for further fleshed out examples

All code for these exercises can be found via this hyperlink as a ipython/jupyter notebook located on my github in addition to attachments sent with the presentation:

https://github.com/JTBarkeloo/JupyterNotebooks/blob/master/MRM Assessment.ipynb

Exploratory Data Analysis

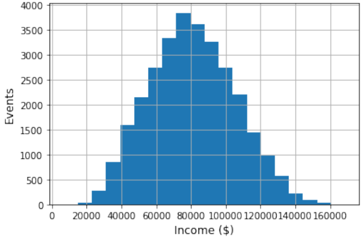
Summary of 30,000 individuals

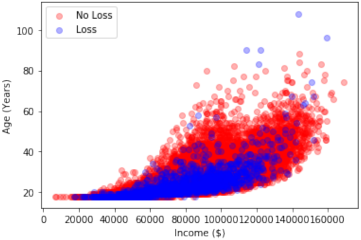
- age - Age of customer
- inc - Income of customer
- car - Category of car: Standard, Luxury, Truck
- edu - Education level of driver (High School or College)
- acc - Number of accidents over the last two years

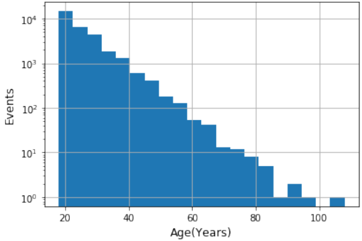
Want to build a model that will optimize recognition of accident events using these values and attempt to categorize riskier drivers from less risky drivers Ideal model would have nice bifurcation between classes, difficult to get without more discriminating variables with separation power

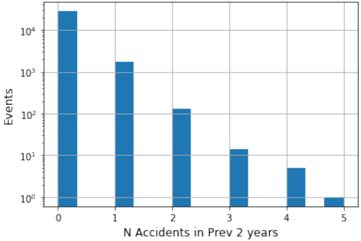
Exploratory Data Analysis

- Start by looking at behavior of noncategorical data, Age is oddly uniform









Statistical Significance: Vehicle and Education Types

Z-test used, conclusions to be drawn depend on how liberal the definition of statistical significance being used is
The use of $p<0.05$ is somewhat arbitrary but is what will be used here as it is a standard choice of convention

- z value for Standard and Luxury: 3.64
- z value for Standard and SUV: 5.31
- z value for Luxury and Truck: 7.11

The null hypothesis can be rejected for all combinations of Vehicle Types

- z value for High School and College: 0.27

The null hypothesis cannot be rejected for Education Types, as such for my models they will be combined and education type not used as a potential discriminator

Categorical variables are changed to numbers using one-hot encoding to prevent ranking errors during model building and testing.

Model Building

A variety of models were employeed to different ends to try and create a binary classification of risk. Multiple collision events are classified with single collision events. With more data a third category of excessive risk could be added to models.

Small minority class, small percentage of those are multiple collision events.

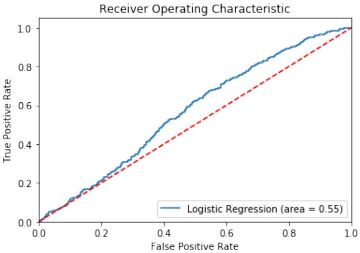
Models presented here (with smote oversampling training data):

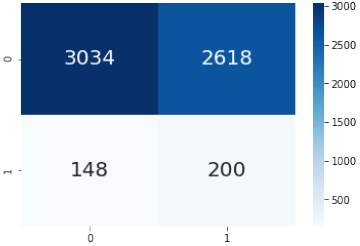
- Logistic Regression
- Boosted Decision Tree (BDT)

A training (80%)/testing(20%) random set split was done to help ensure unbiased results

Multivariate Logistic Regression

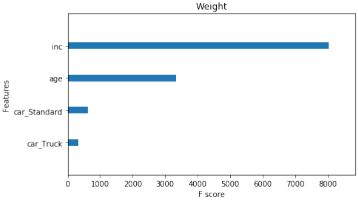
- Naively we could train a model on the data classes as given
- With enough separation power i.e., variables distinct enough in each class, this can be used for event classification
- SMOTE Oversampling used to generate synthetic data that is similar to, but not exactly like the minority class, using a nearest-neighbors approach and fills in space between neighbors
- Balanced Accuracy score: 0.559, Loss Event Accuracy: 0.575

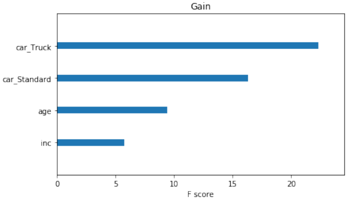




BDT with SMOTE Upsampling

- A BDT was created and trained using SMOTE over-sampling with similar results
- F Scores can give insight into separation powers of training variables
- Weight: How frequent splitting occurs on the variable
- Gain: How useful variable is in terms of separation


Weight


Gain

Loss Events P(Loss Event): mean: 0.510, std: 0.096
NoLoss Events P(LossEvent): mean: 0.480, std: 0.097
Loss Event Accuracy: 55.7%

Model Comments

- Neural networks have been created and trained on a limited set of input variable with success in determination of Loss events
- The addition of further independent input variables would help the separation of the neural network greatly
- A bifurcation of the distributions is starting to occur with the ADASYN network, more input variables and events is likely to cause a major splitting of the distribution into likely Loss events and likely NoLoss events
- Boosted decision tree (BDT) models were also employed in the Jupyter notebook to slightly different ends

Title

Can Multiple Loss Models be Useful Based on Driver Class i.e., Rural Vs. Urban Drivers?

- Rural and Urban drivers face different landscapes of challenges on their daily travels
- Requirement: GPS definition of urban environments
- Expect longer distance/trip for rural drivers while urban drivers have more stop-and-go traffic
- Larger Distances and a larger amount of HardAccelerations are both positvely correlated with loss this seems to be an interesting intersection of these correlations