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

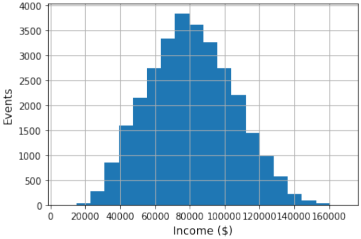
Summary of 30,000 individuals

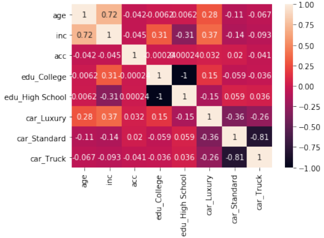
- age - Age of customer
- inc - Income of customer
- car - Category of car: Standard, Luxury, Truck
- edu - Education level of driver (High School or College)
- acc - Number of accidents over the last two years

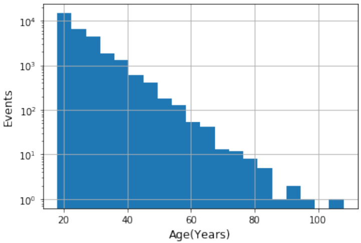
Want to build a model that will optimize recognition of accident events using these values and attempt to categorize riskier drivers from less risky drivers Ideal model would have nice bifurcation between classes, difficult to get without more discriminating variables with separation power Area under ROC curve can be used a a way to classify the model in comparison to other models

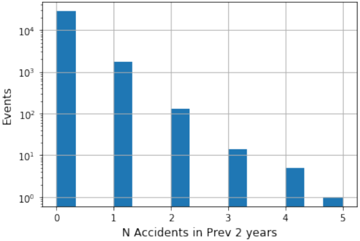
Exploratory Data Analysis

- 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.

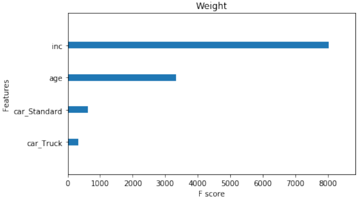
Model presented here (with smote oversampling training data):

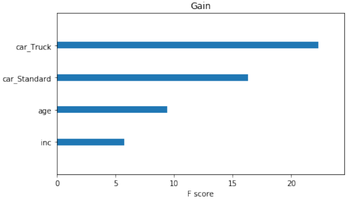
- Boosted Decision Tree (BDT)

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

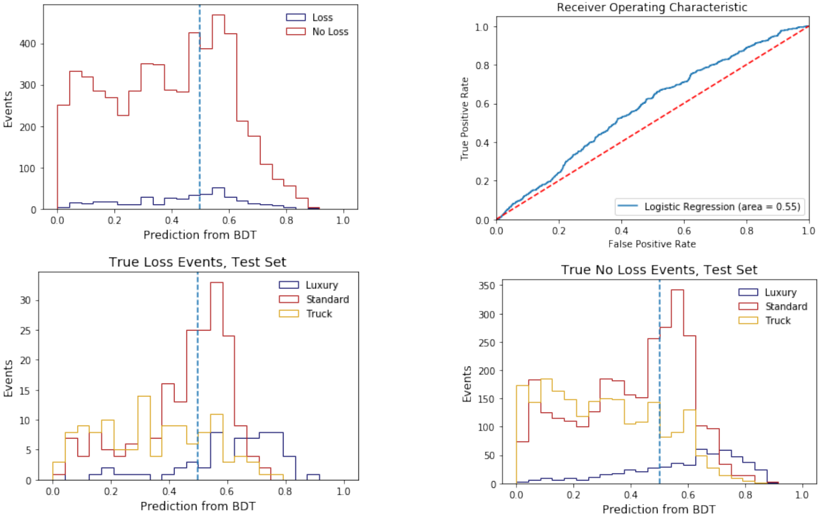
BDT with SMOTE Upsampling

- 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
- 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





BDT with SMOTE Upsampling



Balanced Accuracy Score: 0.553 Area Under ROC: 0.55

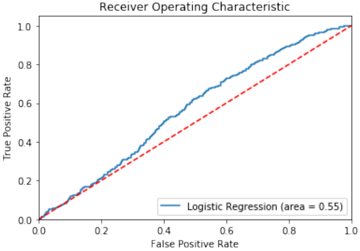
Model Comments and Limitations

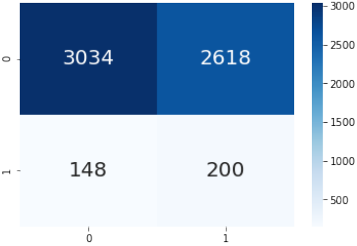
- BDT probably overkill with this amount of data/variables, especially given correlations
- Overtuning difficult to avoid/hyperparameters must be adjusted
- Samilar result i.e., area under ROC curve as Multivariate Logistic Regression (See Backup)
- The BDT model over estimates Standard and Luxury loss from no loss cases. This is most likely do the class makeup that wasnt fully taken into account due to time
- Essentially the model asserts that cars (both types) are significantly riskier than trucks. This seems logical as trucks exist in a separate cargo-loading domain as opposed to cars
- Further data i.e., driving data, would allow this model to be further generalized and the classes further separated

Backup

Multivariate Logistic Regression

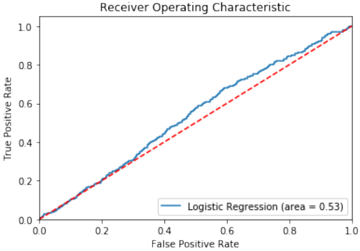
- Naively we could train a model on the data classes as given
- With enough separation power i.e., variables distinct enough in each class, this can be used for event classification
- Balanced Accuracy score: 0.559, Area Under ROC: 0.55





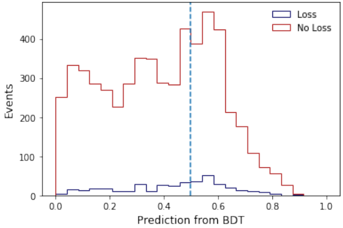
Single Variable Logistic Regression

- The simplest model would be to use the highest correlated variable in a single variable logistic regression to classify events
- Income is highest noncategorical correlated variable to Accidents
- Balanced Accuracy score: 0.529, Area Under ROC:0.53
- There is something to be gained including another highly correlated variable (age) here

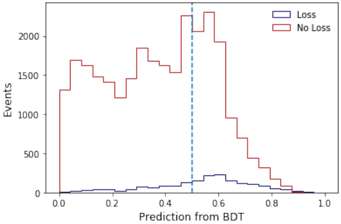


BDT with SMOTE Upsampling

Testing Set



Full Set



Balanced Accuracy Score: 0.553 Area Under ROC: 0.55