Nationwide: Model Risk Management Assessment/Case Study

Jason Barkeloo, Ph.D.

November 16, 2020

Table of Contents

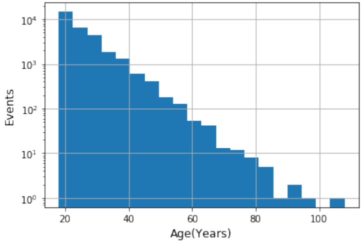
Code location for further fleshed out examples

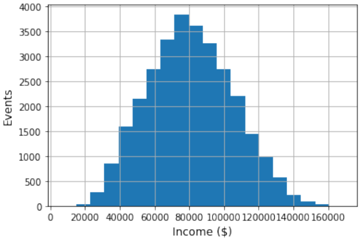
All code for these exercises can be found via this hyperlink as ipython/jupyter notebooks located on my github in addition to attachments sent with the presentation:

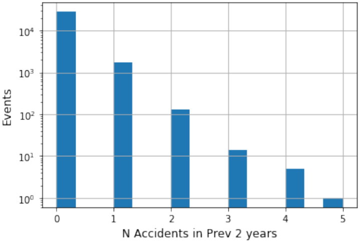
https://github.com/JTBarkeloo/JupyterNotebooks/blob/master/MRM Assessment.ipynb

Exploratory Data Analysis

- Start by looking at behavior of noncategorical data







Model Building

Summary of 30,000 vehicles 1Hz telematics datasets

- Vehicle - Effectively an index on the data
- Days - Number of days data was collected (365 for all)
- Distance - Total number of miles vehicle was driven during data collection
- HardBrakes - Number of hard braking events detected
- HardAccelerations - Number of hard acceleration events detected
- NightTime_Pct - Percentage of total miles driven at night
- VehicleType - str description of type of vehicle
- Loss - Indicator if vehicle has been in a collision

Want to build a model that will optimize recognition of Loss events using these values

Statistical Significance Between Vehicle Types

The 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

- $\boxed{\text{z value for Car and Minivan: 2.48}}$
- z value for Car and SUV: 4.19
- z value for Car and Truck: 2.96
- z value for Minivan and SUV: 4.59
- z value for Minivan and Truck: 3.92
- $\boxed{\text{z value for SUV and Truck: 1.62}}$

The null hypothesis cannot be rejected for the combination of Cars and Minivans and the combination of SUVs and Trucks

The implication then is that there are 2 distributions being sampled for these simulated events. This matches intuition as Trucks/SUVs exist in a cargo-loading domain while Cars/Minivans are what parents may gravitate toward

Model Building

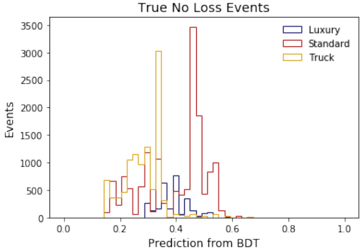
Primarily employing densely connected feed forward neural networks for event classification was chosen as it is the machine learning model I have most experience with for binary classification

- 1 input layer with all potentially useful features (Distance, HardBrakes, HardAccelerations, NightTime_Pct, VehicleType)
- blah

A training (64%)/testing(16%)/validation(16%) random set split was done to help ensure unbiased results

Naive Approach Neural Network

- Naively we could train a neural network 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
- This is not the case for this dataset, only a few variables inputs with a lot of distribution overlap
- This would then be expected to fail with a total accuracy that trends toward the class representation of the majority class, which is seen here



Neural Network with SMOTE Upsampling

- Another network was created and trained using Synthetic Minority Oversampling Technique (SMOTE) over-sampling with similar results
- SMOTE generates synthetic data that is similar to, but not exactly like the minority class, using a nearest-neighbors approach and fills in space between neighbors

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