Capstone Data Inference

The Capstone dataset I chose is from <https://www.kaggle.com/c/porto-seguro-safe-driver-prediction>. The dataset consists of a set of training data and test data. Dataset contains a large number of features, 57 to be exact and a target value that is only present in training data. Target value can be either 1 or 0, 1 means that driver has initiated an insurance claim the next year and 0 means driver did not initiate an insurance claim. For data inference, my goal is to find any correlation between features and target. To do this, I first separate out the training dataset into one set of data where the target is 1 (true set) and another set where target is 0 (false set). Then I setup functions that would compute and plot the ECDF on both the true set and false set of a single feature. Then using the functions setup, I plotted the ECDF of each feature.

|  |  |
| --- | --- |
|  |  |

Next, I went through each individual carefully, and observed that certain features have different ECDF plots between true set and false set.

['ps\_ind\_01','ps\_ind\_03','ps\_ind\_05\_cat','ps\_ind\_06\_bin','ps\_ind\_07\_bin','ps\_ind\_15','ps\_ind\_16\_bin','ps\_ind\_17\_bin','ps\_reg\_01','ps\_reg\_02','ps\_reg\_03',

'ps\_car\_01\_cat','ps\_car\_02\_cat','ps\_car\_03\_cat','ps\_car\_04\_cat','ps\_car\_05\_cat','ps\_car\_06\_cat','ps\_car\_07\_cat','ps\_car\_08\_cat','ps\_car\_09\_cat','ps\_car\_11\_cat','ps\_car\_12','ps\_car\_13','ps\_car\_15']

Here is the set of features that have different ECDFs. This helps reduces the features which I need to worry about down to 24.

Next, I ran Pearson correlation coefficient between quantitative features and the target, but the resulting Pearson r values are all relatively small (largest magnitude is 0.053). Then I ran chi-square test on nominal features (binary and categorical). To conduct the test, I first tallied up the number of occurrences of each of the unique combination between the feature and target. Then I used the set where target is 0 as the expected value and set where target is 1 as the observed value to compute chi square. There is a noticeable difference between features that had similar ECDF to those that didn’t. For example ps\_ind\_11\_bin has similar ECDF plot:

chi2: 2.44839164851

p: 0.117645355928

dof: 1

expected: [[ 2.16572973e+04 3.67026505e+01]

[ 5.72547703e+05 9.70297350e+02]]

whereas ps\_ind\_06\_bin which has a different ECDF plot has much higher chi square and lower p value.

chi2: 688.748032799

p: 8.36655285187e-152

dof: 1

expected: [[ 13152.15971452 8541.84028548]

[ 347699.84028548 225818.15971452]]