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 the features and the target, but the resulting Pearson r values are all relatively small (largest magnitude is 0.053). This could be the nature of the features, some features are binary or categorical, which means that it’s value are labels rather than actual value, so that we can’t rely on Pearson correlation coefficient to draw correlation.