Statistical Write Up

My question in this assignment is to see which features best predict/explain my target variable. The dataset in particular is extremely imbalanced (348/9822 cases of a Caravan policy ownership), so recognizing this statistic before moving to the more granular categorical relations is vital to better understanding the dataset. In this project, I focused on the categorical correlations Theil's U (also known as the uncertainty coefficient) to understand the relationships between the features. Theil's U (utilized in SweetViz function) reals much more information than a simple correlation. Theil's U which is measured from 0-1, measures conditional relationships, so instead of blinding seeing a correlation, we can see the direction of the relation.

In the dataset, there are 28 features that are correlated with the target variable (owning a Caravan policy). Those that provided information on the target variable were:

'Contri_car_pol', 'Contri_fire_pol', 'Num_car_pol', 'subtype_L0' ,'maintype_L2', 'Avg_inc', 'PP_cls', 'Inc_u_30' , 'L_lvl_edu', 'Contri_prv_3p_insur', 'N_car', 'Num_prv_3p_insur', 'R_house','O_house'

Meanwhile, the target variable provided information on the following:

'Num_boat_pol', 'Contri_boat_pol', 'Contri_surfb_pol', 'Num_surfb_pol', 'num_ss_insur_pol', 'Contri_ss_insur_polo', 'Contri_car_pol', 'Num_disabl_insur_pol', 'Contri_disabl_insur_pol', 'Contri_fam_accid_insur_pol', 'Num_car_pol', 'Num_fam_ccid_insur_pol', 'Num_ag_machines_pol', 'Contri_fire_pol'

Given the sheer number of features and weak correlations with the target variable (less than 0.20), breaking down the subcategories within each feature was my next step. I also wanted to ensure that the features I would be researching were statistically relevant by using LassoCV.

I used LassoCV because it does a good job of picking out the interesting features in a dataset. For context, I kept anything whose weights were not converging to zero. This allowed me to keep variables that were negatively and positively related to the target variable. The columns I kept I then encoded them to single out each subcategory and ran it through another LassoCV to reach my final features in the model.

The metrics I used for measuring the model were the accuracy score, balanced_accuacy_score, F1 score, confusion matrix, precision and recall, and ROC. I deemed these most important because they allow me to analyze the performance of the algorithm and where a weakness might be. Being that the dataset was so imbalanced, ensuring the balanced_accuracy_score was closed to the accuracy score was imperative in ensuring the model was not just ignoring the target feature and justified the results in the ROC. Because of the imbalance, I ran my models on the dataset before and after I rebalance the dataset 5-1 to show the difference in results improved the different model performances.