Stat512 Research Proposal

Dataset: Communities and Crime

Description of datset: 1994 observations, 128 attributes. Out of these attributes, 5 of which are information and provide nothing to the inference and prediction of data. Hence, we have 122 useful attributes and 1 prediction of interest: total number of violent crimes per 100K population. All numeric data was normalized into the decimal range 0.00-1.00 using an Unsupervised, equal-interval binning method.

Research Question of Interest: At this stage, due to the sheer numbers of covariates, it’s difficult to pinpoint significant contributors to the model.

We understand the crime rate is influenced by many factors, most obvious of which are socioeconomic factors, education level, employment/unemployment ratio, demographical compositions etc. And many of which are intercorrelated as well, such as a poor neighborhood is more likely to have high crime rate because of low education level, which can contribute to high unemployment rate, which lead to lower income. Therefore, it’s critical for us to understand the relationship between the most important factors and to develop a multivariate regression model for prime prediction.

However, with all statistics, we are interested in both the inference provided by our model as well as the predictive power of the model. Inference allows us to understand the relations in the model, while the predictive power can help us predict and therefore develop measures to prevent or abate crime incidences.

Methods: From what we’ve learned in Stat 512, we will execute the following procedures (as of today, not including procedures and tools yet to be learned, and not necessarily in that order, some procedures may need to be done multiple times)

1. Clean, and organize data
2. Look at descriptive statistic (which is categorical, which is real and continuous)
3. Split randomly into test and training datasets
4. Run an overall model, initial filtering of significant Xs.
5. Use VIF(>10) to delete related covariates.
6. Diagnostics, diagnostics, and diagnostics
7. Transformation, different types of fitting
8. Inference vs. Prediction ( Using test set)