Data Cleaning

Tuesday, 25 May 2021

12:28 PM

* Performed in R
* Both training and test datasets appear to be completely balanced with no missing variables. No further imputation required
* Train data has creditability column while test does not
* Structure of data says all variables are numeric. Cross referencing with the data dictionary shows that most of the variables are categorical. Feature engineering can be applied to revert numbering to character values
  + Numeric attributes
    - Duration of Credit
    - Credit Amount
    - Age

* Univariate Analysis:
  + Majority of creditors are good
  + Sex/Marital status
    - Messy attribute, should be split into 'sex' and 'marital status' attributes
    - There appears to be no single females in either training nor test datasets which may affect the accuracy of predictive models
    - Can remove mixed attribute
  + Purpose
    - Appears as though there are many variables
    - Might not be necessary to combine them
* Bivariate Analysis:
  + Likely skipping since correlations will likely be revealed during principal components analysis
  + May return to bivariate analysis in the future if there is time

* PCA
  + Not particularly helpful
  + Cumulative eigenvalues go up to around 60% of explained variance (measured up to 22 principal components)
  + whole dataset is not explained even using the same number of principal components as there are attributes
  + Principal components