## Data reduction: Deletion of highly correlated data

In the first iteration, for the reasons of simplicity, our team has decided to forgo the sophisticated data reduction techniques like principle component analysis. Instead, we decided to simply remove highly correlated variables in order to reduce the number of variables in order to reduce the number of variables.

Using the “caret” package we have first calculated the correlation matrix (function “cor”). The correlation matrix is a square matrix with the correlation factors for every combination of two variables. The visualization of the correlation matrix (corplot) is presented in the Figure … of the section exploratory data analysis.

The „findCorrelation(correlationMatrix, cutoff=0.8)“ function of the package “caret” searches through the correlation matrix and returns the vector of column indexes that have to be removed to reduce the pair-wise correlation. The “cutoff” parameter specifies the maximal correlation coefficient, above which the variables should be deleted.

Applying the “findCorrelations” function with “cutoff “ parameter of 0.99, ten variables could be identifies as almost perfectly correlated: "attempt\_Mean", "complete\_Mean", "complete\_Range", "attempt\_Range", "totmou", "ovrrev\_Mean", "totcalls" ,"ovrrev\_Range", "totrev", "Customer\_ID". The variable "Customer\_ID" can not be deleted. According to Rud (2001) the correlation limits of .7 and above are frequently mentioned by the benchmarks. In the first iteration we have removed from our dataset the variables that correlated to any other variables by the correlation factor of higher then .75.

Sources

Rud, O. (2001). *Data mining cookbook*. New York: Wiley.