Filter

Nick Wawee

3/22/2021

## Filter

df = read.csv('../data/raw/BankChurners.csv', stringsAsFactors = T)

The original data set contains 23 variables and 10127 customers. The code below will filter the customer churn dataset for all unknown or missing values.

### Remove Naive Bayes Columns

The Naive Bayes columns will be removed because they are not a part of the analysis.

df = df[,c(-22,-23)]

### Categorical Variables

Next the categorical variables will be filtered to not have any unknowns.

cvars = c('Attrition\_Flag', 'Gender', 'Education\_Level', 'Marital\_Status', 'Income\_Category',  
 'Card\_Category')  
  
for (c in cvars){  
 numunk = length(which(df[, c] == 'Unknown'))  
 cat('The', c, 'variable has',numunk, 'unknowns.\n')  
}

## The Attrition\_Flag variable has 0 unknowns.  
## The Gender variable has 0 unknowns.  
## The Education\_Level variable has 1519 unknowns.  
## The Marital\_Status variable has 749 unknowns.  
## The Income\_Category variable has 1112 unknowns.  
## The Card\_Category variable has 0 unknowns.

The education level, marital status, and income category variables with unkown observations will be filtered out.

df = df[-which(df[, 'Education\_Level'] == 'Unknown'), ]  
df = df[-which(df[, 'Marital\_Status'] == 'Unknown'), ]  
df = df[-which(df[, 'Income\_Category'] == 'Unknown'), ]

### Numerical Variables

`%notin%` <- Negate(`%in%`)  
numvars = colnames(df)[colnames(df) %notin% cvars]  
  
for (v in numvars){  
 numunk = length(which(is.na(df[, v])))  
 cat('The', v, 'variable has',numunk, 'NAs.\n')  
}

## The CLIENTNUM variable has 0 NAs.  
## The Customer\_Age variable has 0 NAs.  
## The Dependent\_count variable has 0 NAs.  
## The Months\_on\_book variable has 0 NAs.  
## The Total\_Relationship\_Count variable has 0 NAs.  
## The Months\_Inactive\_12\_mon variable has 0 NAs.  
## The Contacts\_Count\_12\_mon variable has 0 NAs.  
## The Credit\_Limit variable has 0 NAs.  
## The Total\_Revolving\_Bal variable has 0 NAs.  
## The Avg\_Open\_To\_Buy variable has 0 NAs.  
## The Total\_Amt\_Chng\_Q4\_Q1 variable has 0 NAs.  
## The Total\_Trans\_Amt variable has 0 NAs.  
## The Total\_Trans\_Ct variable has 0 NAs.  
## The Total\_Ct\_Chng\_Q4\_Q1 variable has 0 NAs.  
## The Avg\_Utilization\_Ratio variable has 0 NAs.

write.csv(x = df, file = '../data/processed/BankChurners\_filtered.csv')