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]
```

```r
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.
```

```r
write.csv(x = df, file = '../data/processed/BankChurners_filtered.csv')
```