Data details

The dataset contains 1000 observations and 45 features, 1 target variable and 1 index. Each datapoint is a US public stock, each feature represents a financial matrix of that stock in 2014, and the target variable is the annual stock return in the next year (2015).

This dataset is a subset of the a much larger dataset originates from https://www.kaggle.com/cnic92/200-financial-indicators-of-us-stocks-20142018?select=2015_Financial_Data.csv. The original dataset has 35000 datapoints and 230 features. Because of computer power limit I random sampled 1000 datapoints and handpicked 45 features, such as "PB.ratio", "EBIT.Growth", "EV.to.Sales", "Debt.to.Equity", using the code below.

```
data= read.csv('financial-indicators/2014_Financial_Data.csv')
data = data[sample(1:nrow(data),1000),]
growths = c((221-34):221) #34 growths features
ratios = c(147:156) #10 ratios features
data = data[,c(ratios,growths,224)]
write.csv(data,'financial.csv')
```

Data Cleaning

Missing data is filled with column means and all features are standardized.

```
data= read.csv('financial.csv') #
data=data.matrix(data)
data=data[,-1] # The first column is removed as it is the sample index of the original data.
y      =   data[,46]
X      =   data.matrix(data[,-46])

### filling NAs with col means
for(i in 1:ncol(X)){
  X[is.na(X[,i]), i] = mean(X[,i], na.rm = TRUE)
  }

### standardizing ratio features
standardize =function(x){x/sqrt(mean((x-mean(x))^2))}
X = apply(X,2,standardize)
```