Predicting Stock Returns with Cluster-Then-Predict

When selecting which stocks to invest in, investors seek to obtain good future returns. In this problem, we will first use clustering to identify clusters of stocks that have similar returns over time. Then, we'll use logistic regression to predict whether or not the stocks will have positive future returns.

For this problem, we'll use StocksCluster.csv, which contains monthly stock returns from the NASDAQ stock exchange. The NASDAQ is the second-largest stock exchange in the world, and it lists many technology companies. The stock price data used in this problem was obtained from infochimps, a website providing access to many datasets.

Each observation in the dataset is the monthly returns of a particular company in a particular year. The years included are 2000-2009. The companies are limited to tickers that were listed on the exchange for the entire period 2000-2009, and whose stock price never fell below $1. So, for example, one observation is for Yahoo in 2000, and another observation is for Yahoo in 2001. Our goal will be to predict whether or not the stock return in December will be positive, using the stock returns for the first 11 months of the year.

This dataset contains the following variables:

- **ReturnJan** = the return for the company's stock during January (in the year of the observation).
- **ReturnFeb** = the return for the company's stock during February (in the year of the observation).
- **ReturnMar** = the return for the company's stock during March (in the year of the observation).
- **ReturnApr** = the return for the company's stock during April (in the year of the observation).
- **ReturnMay** = the return for the company's stock during May (in the year of the observation).
- **ReturnJune** = the return for the company's stock during June (in the year of the observation).
- **ReturnJuly** = the return for the company's stock during July (in the year of the observation).
- **ReturnAug** = the return for the company's stock during August (in the year of the observation).
- **ReturnSep** = the return for the company's stock during September (in the year of the observation).
- **ReturnOct** = the return for the company's stock during October (in the year of the observation).
- **ReturnNov** = the return for the company's stock during November (in the year of the observation).
- **PositiveDec** = whether or not the company's stock had a positive return in December (in the year of the observation). This variable takes value 1 if the return was positive, and value 0 if the return was not positive.

For the first 11 variables, the value stored is a proportional change in stock value during that month. For instance, a value of 0.05 means the stock increased in value 5% during the month, while a value of -0.02 means the stock decreased in value 2% during the month.

Q1. Load StocksCluster.csv into a data frame called "stocks". How many observations are in the dataset?

Q2. What proportion of the observations have positive returns in December?

Q3. What is the maximum correlation between any two return variables in the dataset? You should look at the pairwise correlations between ReturnJan, ReturnFeb, ReturnMar, ReturnApr, ReturnMay, ReturnJune, ReturnJuly, ReturnAug, ReturnSep, ReturnOct, and ReturnNov.

Q4.Train a logistic regression model (name it StocksModel) to predict PositiveDec using all the other variables as independent variables.

Q5. Post Clustering, which cluster has the largest number of observations?.