MSBA Python Bootcamp

July 2022
Practical Exercise 7 – Classification and Clustering

This practical exercise is based on the Forbes2000.csv dataset and you will be building some simple models to classify marketvalue and cluster sales, profits and assets. In particular, perform each of the following tasks, and report your results and observations.

PE7-01 – Data Modeling – Classification

Binary classification of marketvalue (low and high) – Build a decision tree classifier to predict marketvalue as a binary variable (see PE04-01-g) using sales, profits and assets. You should use Scikit Learn.

We would be using Graphviz to generate the decision tree visually. Graphviz can be installed with the following command using pip:

```
python -m pip install graphviz
```

You also need to download the Graphviz executable package from here –
https://graphviz.org/download/

How does this classification model compare with the multiple linear regression models in PE5-02?

PE7-02 – Data Modeling – Clustering

Clustering of sales, profits and assets – Build two k-means clustering models to cluster the world's leading companies in the 2004 Forbes list using $k = 2$ and $k = 3$. Which clustering model is the better one?

Calculate the within-cluster mean and standard deviation for each of the independent variables – sales, profits and assets – based on the better clustering model.

How do your findings compare with the supervised learning models that we have created earlier using regression and classification analysis?