Submission Number: 2

Group Number: 08

Group Members:

Full Legal Name	Location (Country)	E-Mail Address	Non-Contributing Member (X)
Monika Singh	India	Ms19626@gmail.com	
Ananta Narayana Satapathy	India	ansatapathy@yahoo.com	
Harshil Sumra	India	Harshilsumra1997@gmail.com	

Statement of integrity: By typing the names of all group members in the text box below, you confirm that the assignment submitted is original work produced by the group (*excluding any non-contributing members identified with an "X" above*).

Monika Singh
Ananta Narayana Satapathy
Harshil Sumra

11. Which model - CART classification, CART regression, or SVM -- provides the best fit to the data? Why?

When we compared CART classification, CART regression and SVM for our model evaluation, we found that CART classification tree performed much better with accuracy around 63% and F1-score of 81%, whereas the regression tree generally gives very low or negative R2. We also transformed the output from the regression tree to a binary signal variable to compare accuracy directly with the classification tree. We observed that that the predicted signals from the regression tree give an F1-score of 74% as against 81% for the classification tree. Even for SVM, root mean squared error value is quite high at 47.99. CART performs better than SVM here as it handles the collinearity in the log returns data better than SVM.

12. Which model provides better interpretation of the results?

The CART classification model provides better interpretation of results compared to CART regression model and SVM. This is based on accuracy, RSquared and RMSE observed for the three models.

13. How did your group divide the work?

We have divided the work into three parts since we are three members. First part is Data importing, dimensional data and volatility summary preparation and data exploration. Second part consists of model building such as CART and SVM. Third part consists of generating technical and non-technical report to senior management.

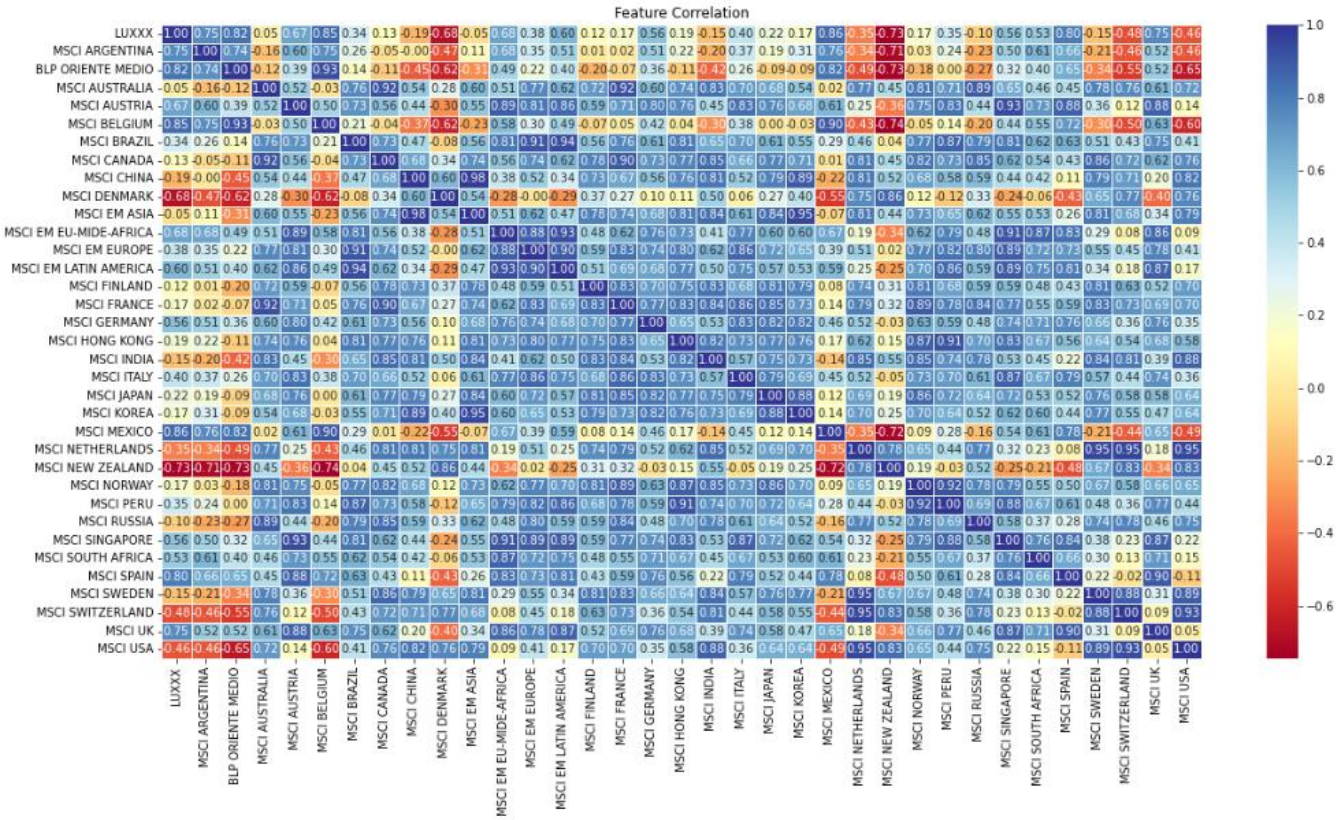
14. Write a TECHNICAL 1-page report of your findings to your FE boss.

Problem Overview

Our objective is to predict the performance of the stock market, in terms of both actual returns as well as the sign of daily returns. We have used LUX Index (Stock market index of the Luxembourg stock Exchange) as dependent variable and MSCI Indexes as independent variables from various areas of the world as the independent variables.

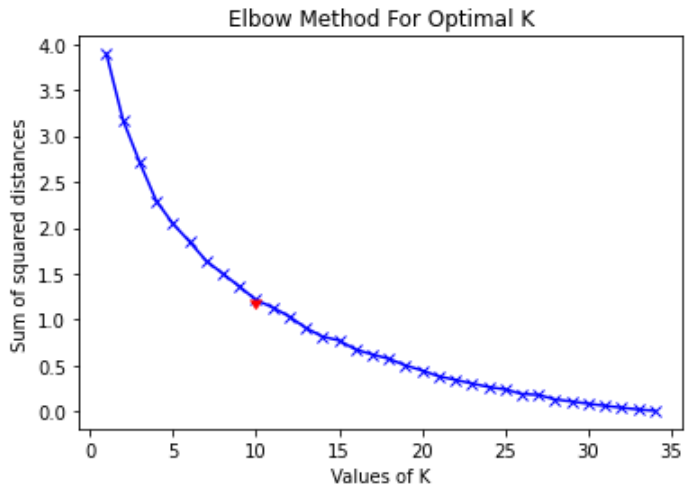
Correlation between Variables

We have checked the correlation between the variables using a heat map. As seen in the figure below and as expected, several of the indices show a high correlation with each other. Thus, the data exhibits high collinearity.

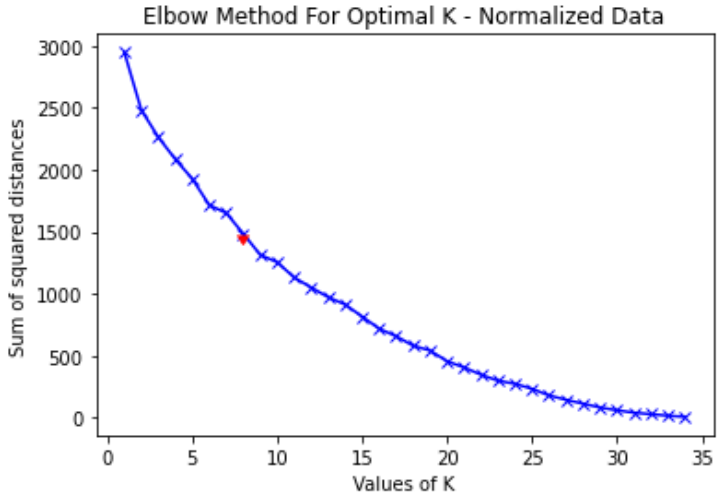


Clustering

We have run a k-means clustering algorithm on the log returns and then on the normalized log returns. For the raw data, using 10 clusters appears to provide the best balance between fit and number of clusters, as it satisfies the 'elbow method' approach visually.



For the raw data, using 5 clusters appears to provide the best balance between fit and number of clusters, as it satisfies the 'elbow method' approach visually.

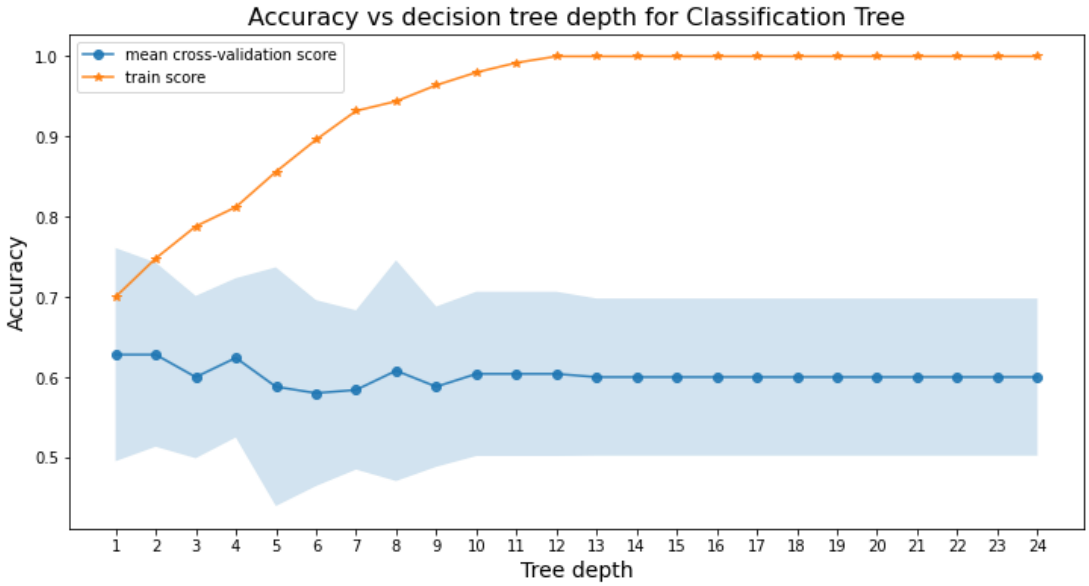


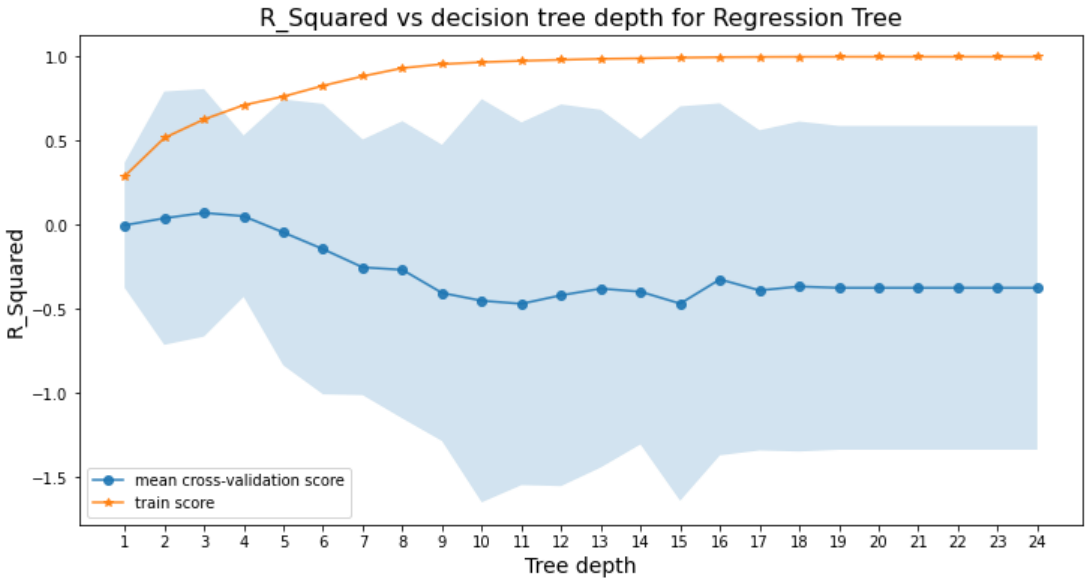
Normalized data gives a lower numbers of clusters as it gives improved clustering. The grouping data also confirms that the grouping of the series has changed after the data was normalized.

Algorithm Explanation

We have used various supervised learning algorithms to evaluate the relationship between independent and dependent variables. The three approaches used are classification decision tree, regression decision tree and SVM. For classification decision tree, we have used the sign of weekly returns as a binary response variable, whereas for the regression decision tree and SVM, we have used the weekly log returns. For classification and regression trees (CART), we have calibrated the depth of the tree using K-fold cross-validation approach and picking the depth that gives maximum prediction accuracy and RSquared respectively.

As seen from the figures below, we choose depth of 5 for classification tree and 4 for the regression tree.

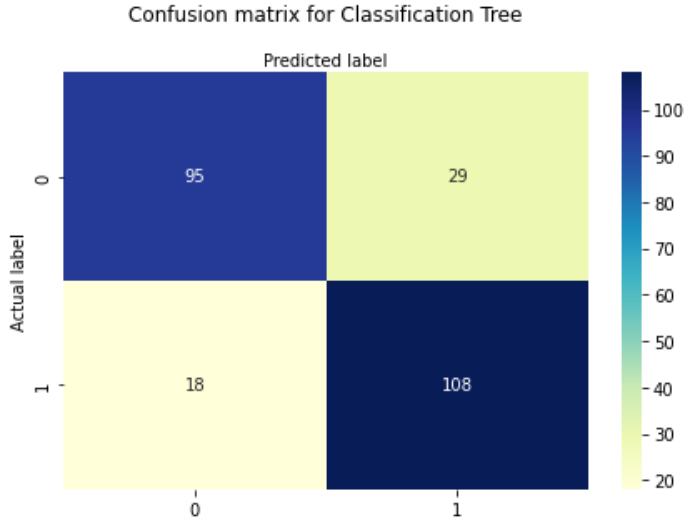


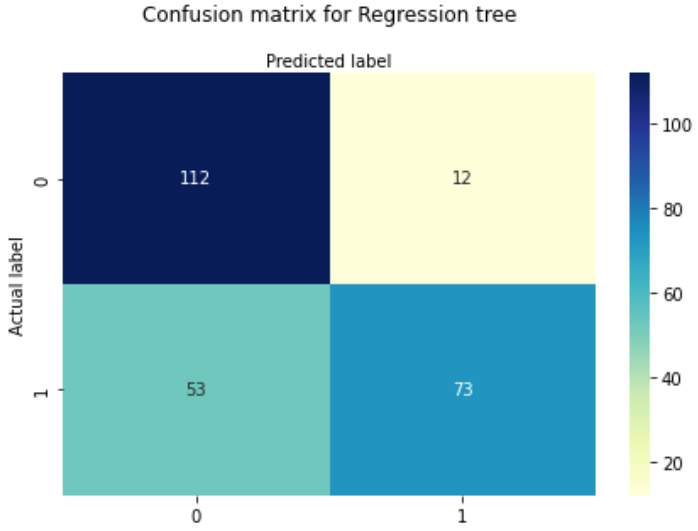


Experimental Evaluation

To evaluate our model we have used R-squares and MSE metrics for regression models and accuracy for classification models. We have divided our data training data into 2/3rd and 1/3rd to our testing data. R-squared help us to check goodness of fit. Higher the value of fitness better the model has performed. MSE is a mean squared error which is the calculation of the mean of error which is difference of actual value and predicted value. For classification models, accuracy gives the percentage of correctly predicted outcomes.

In order to compare the classification and regression trees, we also transformed the output from the regression tree to a binary signal variable to compare accuracy directly with the classification tree. We observed that that the predicted signals from the regression tree give an F1-score of 74% as against 81% for the classification tree. We present the confusion matrix for the two cases below, which clearly shows that the classification tree performs better than the regression tree.





Results and Conclusion

When we compared the prediction from the three approaches for our model evaluation, we found that CART classification tree performed much better with accuracy around 63%, whereas the regression tree generally gives very low or negative R2. We observed that that the predicted signals from the regression tree give an F1-score of 74% as against 81% for the classification tree. Even for SVM, root mean squared error value is quite high at 47.99. CART performs better than SVM here as it handles the collinearity in the log returns data better than SVM.

15. Write a non-technical 1 paragraph email of your findings for senior management

We have performed an analysis to check the relationship between the stock market indices in various parts of the world. Our investigation shows that there is a relationship between various MSCI indexes. We have used LUX index as our dependent variable and the others left as independent variable. We have used classification and regression decision tress (CART) and support vector machines (SVM) regression for the analysis. To evaluate the model we have used metrics which basically capture the difference between the actual value and predicted value. We have concluded that for this dataset, the classification decision tree gives best results compared to a regression decision tree and SVM.

References

https://www.thebalance.com/msci-index-what-is-it-and-what-does-it-measure-330594 8

https://towardsdatascience.com/explainable-artificial-intelligence-part-2-model-interpretation-strategies-75d4afa6b739

https://www.digitalvidya.com/blog/classification-and-regression-trees

https://zenodo.org/record/4170989/files/Kaggle%20-%20DDAS%20Implementation.pdf