Submission Number: 2

Group Number: 15

Group Members:

Full Legal Name	Location (Country)	E-Mail Address	Non-Contributing Member (X)
Ewurama Fordjour	Ghana	ewurama.fordjour@gmail.com	
Deven N. Valecha	India	devenvalecha@gmail.com	
Ishaan Narula	India	ishaan.narula@outlook.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*).

[Ewurama Fordjour?], Deven N. Valecha, Ishaan Narula

Use the box below to explain any attempts to reach out to a non-contributing member. Type (N/A) if all members contributed.

N/A

** Note, you may be required to provide proof of your outreach to non-contributing members upon request.*

Answer 11: CART Classification vs CART Regression vs SVM w.r.t Fit

As demonstrated in the Jupyter notebook, SVM with a linear kernel (i.e. no kernel) provides the best fit to the data. By fit, we refer to the performance of the trained model (based on the training set) on unseen data, i.e. the test set.

The below table provides a summary of the prediction accuracy (model score in sklearn) of the various models on the test set which were fitted to the training set:

CART			SVM		
Depth	Classification	Regression	**Regularisation (C)**	Linear Kernel	RBF Kernel
2	69.05%	31.31%	**1.5**	--	70.24%
3	64.29%	--	**2.0**	79.76%	66.67%
4	--	25.48%	**2.5**	77.38%	--

The SVM with a linear kernel and C = 2.0 delivers the best performance on the test set, with a roughly 80% accuracy.

Two reasons why decision trees did not deliver a high performance on the test set:

- They are prone to overfitting if they are too deep and to underfitting when classes are imbalanced
- The presence of multicollinearity in the dataset causes decision trees to greedily choose the best variable among 2 or more which explain the same thing

The SVM with Linear Kernel delivered high performance due to the high degree of linear separability of the data. Upon training the SVM on the entire dataset in answer 10.3 and evaluating the resulting model's prediction accuracy on that dataset itself, we see that it predicts the return direction of 73% of the data accurately. That said, the data has a high degree of linear separability, which makes an SVM with a Linear Kernel deliver high performance on the test set

Answer 12: Comparison of Model Interpretability

In terms of interpretation, we think that results from CART models are easier to interpret relative to SVMs.

Although SVMs can be excellent in learning both linear and more complex non-linear decision boundaries (through use of kernels), professionals not familiar with ML algorithms may have a hard time in understanding the mechanics of the model, its parameters and the ideas of large margin classification and kernels.

In contrast, decision trees are much more intuitive to understand by non-technical professionals since they are non-parametric and a combination of multiple if-then-else statements and can be displayed diagrammatically. They can also enable a better understanding of the dataset and how each of the predictors contributes to the results produced by the model.

Answer 13: Work Split Report

- Deven worked on questions 1-4, 6-8
- Ishaan worked on questions 9-14
- Ewurama worked on 5, 15
- Together we all had a look at each others' parts and made necessary suggestions for the completion of the assignment

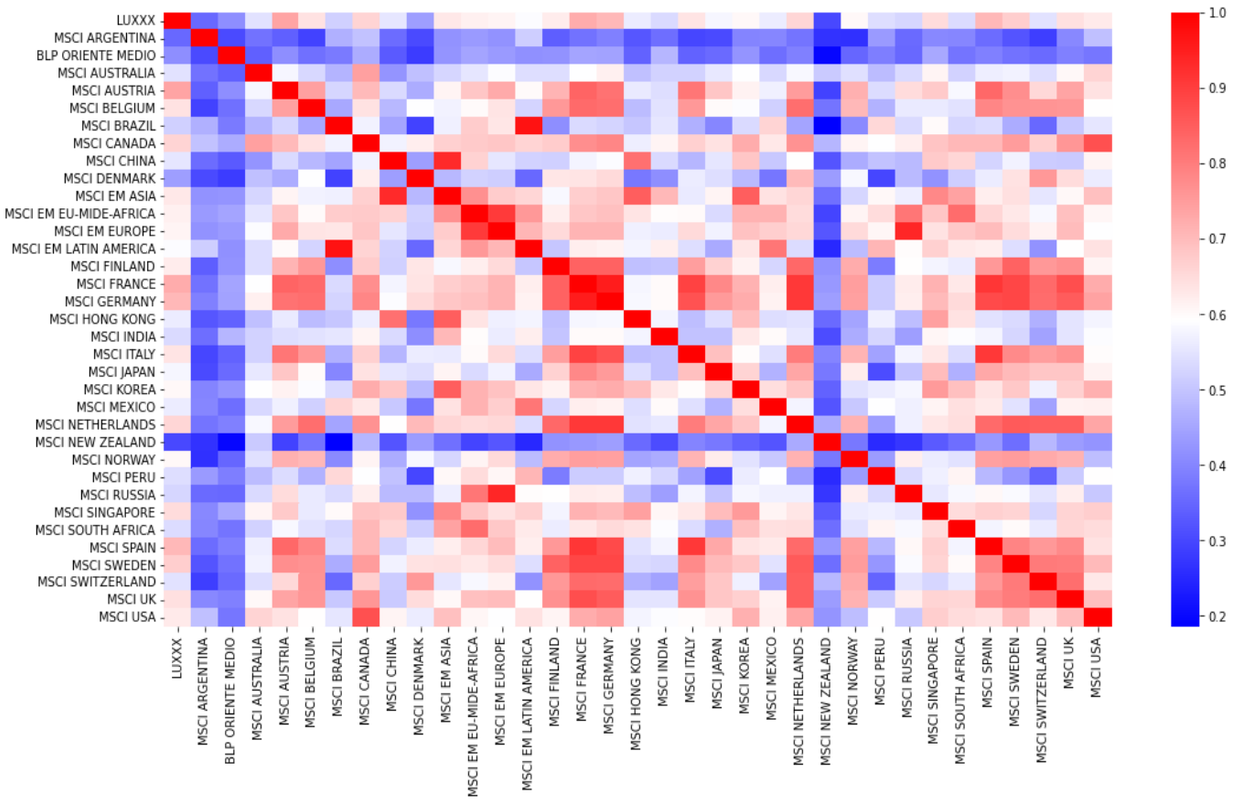
Answer 14: Technical Report

Introduction

In this project, we apply unsupervised and supervised learning techniques on weekly price data from 2016 to 2020 for ETFs capturing stock market performance of 35 countries. The dataset provided in the file *MScFE 650 MLF GWP Data.csv*. After analysing averages, dispersions and correlations across the 35 series, we run the K-means clustering algorithm on this dataset. This is followed by CART and SVM models which are trained taking weekly LUXXX ETF returns as the dependent variable.

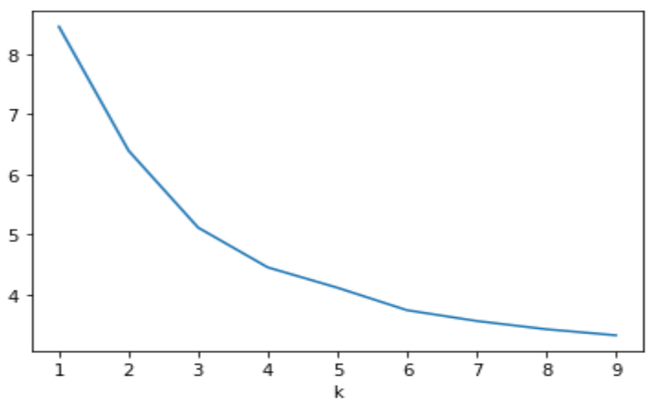
Data Summaries

We estimate the expected returns and volatilities of the 35 time series, and create a heat map of the price correlations across these. The expected return is estimated as a weighted average of the mean, 5% trimmed mean and median using 30%, 50% and 20% as the respective weights. For the expected volatility, we weight the average EWMA, the unconditional variance from a GARCH(1,1) model and the average 1-month range for each time series using 35%, 25% and 40% as the respective weights.



Unsupervised Learning: Clustering

K-means clustering offered us the opportunity to segregate the data into clusters with each dataset belonging to its nearest mean. This by far helped in the image segmentation of our datasets. Further, a K-means cluster algorithm was run for non-standardized and standardized data.



Supervised Learning: CART vs. SVM

We then take LUXXX log returns to be the dependent variable and run Classification and Regression trees and Support Vector Machines (SVMs) to evaluate the remaining series' explanatory power to predict returns in LUXXX. As an exception for the Classification Tree and the SVM, LUXXX for a given week takes a value of 1 if the return for that week was greater than that of the previous week and a value of 0 otherwise.

Before we run the models, we split the relevant dataset into training, cross-validation and test sets. We keep the test set to 1/3 (i.e. 84 examples) of the total no. of training examples. We also perform 5-fold cross-validation on the training set to choose model hyperparameters (depth for CART and C for SVM) before we run the final models on the test set (more details on the model training and testing methodology in the Jupyter notebook).

Upon training the models and testing their performance on the for various hyperparameters, we find that the SVM with a linear kernel and C = 2.0 delivers the best performance on the test set, with a roughly 80% accuracy (more details in Answer 11).

Conclusion

K-means was used to segregate the data into clusters with each dataset belonging to its nearest mean. This by far helped in the image segmentation of our datasets. For supervised learning, while the SVM with no kernel provides a better fit to this dataset due the data's high degree of linear separability, the CART could be used to provide better explainability of this study's approach to a non-technical audience.

Answer 15: Non-technical E-mail For Senior Management

Dear Management,

Kindly note that per our findings, though both CART and SVM are supervised learning, it can clearly be deduced that the results from CART are self-explanatory as compared to that of SVM. To put into perspective non-technical people can easily relate to the findings from CART. Upon categorizing the 35 series into clusters we realized that the SVM equally helped in better explanation of the variables hence providing an overall understanding of the response and the variables.

To conclude, SVM provided a better fit for data with higher dimensions and Regression trees provided the best interpretability of the coefficients.