Submission Number: 1

Group Number: 3

Group Members:

Full Legal Name	Location (Country)	E-Mail Address	Non-Contributing Member (X)
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*).

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 9.1: Non-Technical Report to Portfolio Manager

Introduction

Monthly price and volume data for the period 2014-2019 is imported for ETFs covering 11 US sectors. These serve as the response variables. 19 leading, coincident and lagging economic indicators provided by The Conference Board have been used as predictors, monthly data for which has been imported from https://fred.stlouisfed.org for the period 2014-19. Once monthly log returns based on Adjusted Closing price are calculated for each of the ETFs, we run linear and lasso regressions, cluster analysis and regression trees on our dataset.

Linear and Lasso Regressions

For each ETF, we run 4 linear and lasso regressions, one each using Leading Economic Indicators (LEI), Lagging Economic Indicators (LAG), Coincident Economic Indicators (CEI) and all economic indicators combined (ALL). Each set of predictors is normalised before running these models, since the predictors have very different ranges. The lasso regressions are run using 10 different penalty values ranging from 0.1-1.0 (110 lasso models in total).

For the linear model, Table 1 in the appendix shows that the R-squared obtained from the regression of all indicators on the various responses is the highest, followed by the regression of the LEIs on the responses. So, based on R-squared values, the linear model categorises all ETFs into the LEI bucket.

Comparing the lasso model with a penalty of 0.5 to the linear model, we note that:

- introducing the penalty term reduces some coefficients to zero and the explained variation in returns (Table 3)
- ranking of the ETFs' R-squared produced by the lasso model is different from the one produced by the linear model
- lasso model places all ETFs into the LEI bucket except IYZ and XLI which are placed into the CEI category (Table 2)

Cluster Analysis

We then start by computing the pairwise distances among the various datapoints for each ETF, which is followed by a computation of the distance correlation which, unlike Pearson's correlation, captures both linear and non-linear association between two variables.

As the matrix in Table 4 shows, the distance correlation between a majority of the ETFs is greater than 0.5, implying a strong relationship between them. Because of this, one can imagine that returns across the 11 ETFs are governed by a common set of underlying factors.

However, the linear model shows that the 19 economic indicators combined explain only a limited part of the total variability (13-30%) in the ETFs' returns. These findings could imply the following:

- either the selected economic indicators are not a good choice of predictors to explain the ETFs' returns
- or if they are, the true underlying relationship between the ETFs' returns and the economic indicators is non-linear

We then run a K-means clustering with K=3 and find that 9 of the ETFs are placed in one cluster whereas XLE (Energy) and XLV (Healthcare) are in two separate clusters.

Regression Trees

We run 4 regression trees for each response with a maximum depth of 5 (to avoid too much overfitting), in the same manner as we did for linear and lasso regressions.

The regression tree does a much better job of fitting the datasets compared with linear and lasso regressions (Table 5). This is because linear and lasso regressions are parametric approaches which suffer from the risk of model misspecification, i.e. the functional form of the chosen model does not match the true unknown function. But in general, fitting a more flexible and complex model like a regression tree can lead to a phenomenon known as overfitting the data, which essentially means they follow the errors, or noise, too closely.

Conclusion *(ref. to Table 7)*

Linear regression places all ETFs into the LEI category. Lasso regression places all ETFs into LEI category except IYZ and XLI which it places in the CEI category. Regression tree places all ETFs into LEI category except XLK and XLU which it places in the LAG category.

Appendix

	R-squared LEI	R-squared LAG	R-squared CEI	R-squared ALL
IYR	0.114582	0.040511	0.021508	0.188660
IYZ	0.046510	0.020704	0.004206	0.137117
XLB	0.103479	0.028873	0.034730	0.193065
XLE	0.104510	0.031102	0.083671	0.249276
XLF	0.170083	0.077047	0.041653	0.316068
XLI	0.139086	0.015905	0.037666	0.211946
XLK	0.143782	0.031851	0.019776	0.293204
XLP	0.119721	0.017474	0.003617	0.195829
XLU	0.148679	0.045125	0.029058	0.264932
XLV	0.149585	0.031734	0.010836	0.193789
XLY	0.085052	0.018730	0.020184	0.200484

Table-1: R-Squared Values from Linear Regression

ETF	R-squared LEI	R-squared LAG	R-squared CEI
IYR	0.073172	0.002555	0.036283
IYZ	0.023793	0.000914	0.069930
XLB	0.030541	0.001194	0.011287
XLE	0.006931	0.000000	0.000100
XLF	0.008274	0.000000	0.002266
XLI	0.011122	0.000000	0.015968
XLK	0.073675	0.000000	0.000947
XLP	0.064817	0.000000	0.008954
XLU	0.021283	0.000000	0.001749
XLV	0.013651	0.000000	0.007341
XLY	0.019168	0.000000	0.002947

Table-2: R-Squared Values from Lasso Regression

	R-Squared Linear	R-Squared Lasso
IYR	0.188660	0.135550
IYZ	0.137117	0.149100
XLB	0.193065	0.049317
XLE	0.249276	0.008757
XLF	0.316068	0.013005
XLI	0.211946	0.023132
XLK	0.293204	0.084415
XLP	0.195829	0.081718
XLU	0.264932	0.024268
XLV	0.193789	0.020461
XLY	0.200484	0.042013

Table 3: R-Squared Comparison – All Predictors: Linear vs. Lasso

	iyr log return	iyz log return	xlb log return	xle log return	xlf log return	xli log return	xlk log return	xlp log return	xlu log return	xlv log return	xly log return
iyr log return	1.0	0.364468	0.327421	0.27978	0.299049	0.366434	0.375585	0.544865	0.629949	0.501191	0.434098
iyz log return	0.364468	1.0	0.614839	0.492231	0.472404	0.546916	0.548503	0.412458	0.260671	0.499889	0.57113
xlb log return	0.327421	0.614839	1.0	0.685077	0.697103	0.823083	0.675631	0.405641	0.162361	0.596323	0.73264
xle log return	0.27978	0.492231	0.685077	1.0	0.549166	0.595779	0.452804	0.301993	0.153785	0.359273	0.537837
xlf log return	0.299049	0.472404	0.697103	0.549166	1.0	0.774958	0.611338	0.431736	0.170647	0.605111	0.672079
xli log return	0.366434	0.546916	0.823083	0.595779	0.774958	1.0	0.681413	0.539493	0.179869	0.655141	0.746988
xlk log return	0.375585	0.548503	0.675631	0.452804	0.611338	0.681413	1.0	0.493577	0.242756	0.539397	0.748096
xlp log return	0.544865	0.412458	0.405641	0.301993	0.431736	0.539493	0.493577	1.0	0.533488	0.478115	0.577819
xlu log return	0.629949	0.260671	0.162361	0.153785	0.170647	0.179869	0.242756	0.533488	1.0	0.212626	0.212824
xlv log return	0.501191	0.499889	0.596323	0.359273	0.605111	0.655141	0.539397	0.478115	0.212626	1.0	0.657767
xly log return	0.434098	0.57113	0.73264	0.537837	0.672079	0.746988	0.748096	0.577819	0.212824	0.657767	1.0

Table 4: Distance Correlation Matrix

	ETF	Optimal Depth(s)	Maximum Prediction Score
0	IYR	5.0	0.647173
1	IYZ	5.0	0.696546
2	XLB	5.0	0.787266
3	XLE	5.0	0.670153
4	XLF	5.0	0.678176
5	XLI	5.0	0.712876
6	XLK	5.0	0.585768
7	XLP	5.0	0.827359
8	XLU	5.0	0.500362
9	XLV	5.0	0.504710
10	XLY	5.0	0.500029

Table 5: Maximised Prediction Score from Regression Tree modelling with all indicators

ETF	Max. Score LEI	Max. Score LAG	Max. Score CEI
IYR	0.717892	0.581063	0.573891
IYZ	0.749770	0.476207	0.310148
XLB	0.740543	0.637279	0.548508
XLE	0.668151	0.602693	0.405547
XLF	0.786410	0.602404	0.511610
XLI	0.800414	0.709075	0.587755
XLK	0.580392	0.667930	0.568089
XLP	0.802224	0.430558	0.383161
XLU	0.501709	0.542921	0.287469
XLV	0.504710	0.418976	0.327995
XLY	0.497861	0.510396	0.400938

Table 6: Maximised Prediction Score from Regression Tree modelling with LEI, LAG and CEI

ETF	Linear Model	Lasso Model	Regression Tree
IYR	LEI	LEI	LEI
IYZ	LEI	CEI	LEI
XLB	LEI	LEI	LEI
XLE	LEI	LEI	LEI
XLF	LEI	LEI	LEI
XLI	LEI	CEI	LEI
XLK	LEI	LEI	LAG
XLP	LEI	LEI	LEI
XLU	LEI	LEI	LAG
XLV	LEI	LEI	LEI
XLY	LEI	LEI	LAG

Table 7: Categorisation of the Responses into LEI, LAG and CEI by the Linear, Lasso and Tree models

Answer 9.2: Work Split Report

I am contributing individually for this submission, so all of the tasks required in the assignment have been carried out by me.