FORECAST EARNING FOR SHARE INDEX- EPS

1. Analysis objective

The purpose of the following paper is to provide an estimate as accurate as possible about the value that the Earning Per Share (EPS) index will assume in twelve months, with the aim of understanding what growth will be generated in the following year. The index measures the net profit expressed in monetary terms with reference to each share, also known as “earnings per share”.

2. Description of the dataset

The dataset analyzed contains 2493 records and a total of thirteen variables, one of which will be the variable that we will estimate in relation to a set of carefully chosen features.

The “Russel_3000_foundamental_enlarged” dataset contains a set of data concerning the largest companies, both in the United States and beyond, by market capitalization.

The process that led to providing a more or less accurate estimate of the index in question, starts from a preliminary exploratory analysis of the data through an examination of the correlation of the features, then proceeds with the implementation of regularized regression models, Lasso and Elastic Net specifically, up to the actual implementation of the linear regression model.

In the choice of features, the considerations are strengthened with possible further models, stepwise selection, backward and forward selection, to justify the final choice.

3. Analysis process

3.1. Correlation index

Firstly, the dataset was downsized by eliminating those qualitative and quantitative features but irrelevant for the forecast. Specifically, the following variables were not taken into consideration: 'Record_ID', 'Isin', 'Record_Name', 'Industry' and 'Subindustry'. The rest were analyzed to understand the degree of information redundancy, using a special correlation matrix, as shown in figure 1.

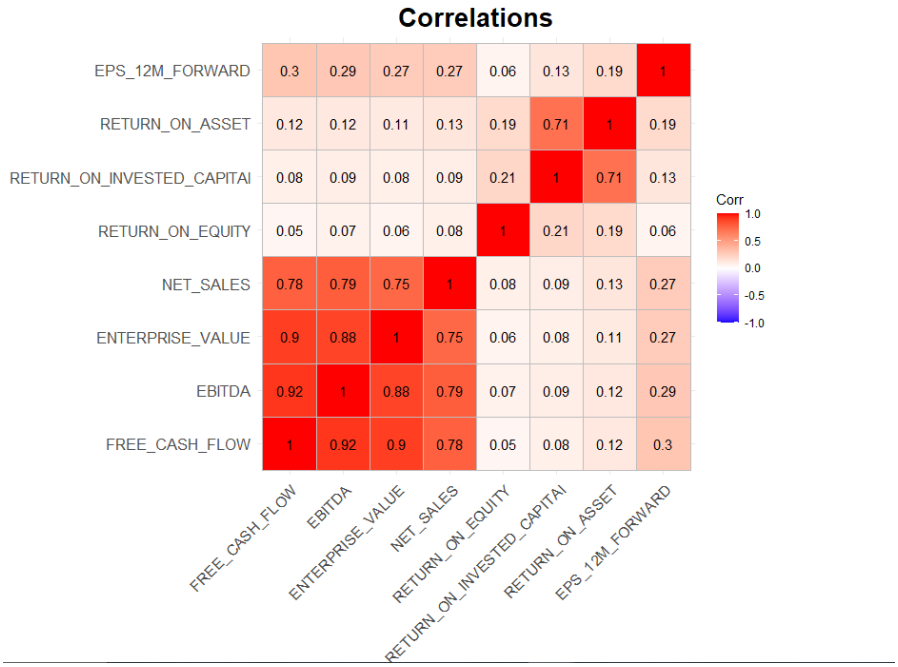

Figure 1- Correlation index of the variables

It is immediately evident that the 'Return On Asset' variable is positively correlated with the 'Return on invested capital' features, which is understandable given the intrinsic definition of the index itself. The 'Return on asset', ROA, indicates the ability of the company to generate a stream of income deriving from the conduct of its business, while the ROIC ('Return on invested capital') represents the profitability of investments, and it is logical to think that as the profitability of investments increases, the income flow of the business consequently increases.

Furthermore, a further group of related features emerged from the correlation analysis: 'Free cash flow', 'Ebitda', 'Enterprise_value' and 'Net_Sales' are all positively correlated with each other.

3.2. Elastic Net & Lasso Regression

The choice of which features to use for the estimation of the EPS index is then evaluated on the regularized regression model, in particular Elastic Net and Lasso.

In the case of Elastic Net, the coefficients that survive the selection are those relating to the 'Free cash flow', 'Net_Sales', 'Return On Asset' variables and, to a very residual extent, also the 'Ebitda' variable, as can be seen in figure 2.

```
9 x 1 sparse Matrix of class "dgCMatrix"
                                      s0
(Intercept)                -0.032463710
(Intercept)                 .
FREE_CASH_FLOW              0.063558117
EBITDA                      0.044793869
ENTERPRISE_VALUE            0.005927138
NET_SALES                   0.094996292
RETURN_ON_EQUITY            .
RETURN_ON_INVESTED_CAPITAl  0.024014479
RETURN_ON_ASSET             0.066985889
>
```

Figure 2- Elastic Net coefficients

The Lasso instead, being more shrink, offers me the same set of variables already selected with the Elastic Net, but with the expected elimination of the coefficient of the variable 'Ebitda' (0.00681, it was considered as 0 in the analysis).

```
9 x 1 sparse Matrix of class "dgCMatrix"
                                      s0
(Intercept)                -0.031805020
(Intercept)                 .
FREE_CASH_FLOW              0.096174424
EBITDA                      0.006817173
ENTERPRISE_VALUE            .
NET_SALES                   0.141724567
RETURN_ON_EQUITY            .
RETURN_ON_INVESTED_CAPITAl  0.007580732
RETURN_ON_ASSET             0.102115099
```

Figure 3- Lasso coefficients

3.3. Stepwise Selection

Finally, additional models were tested to reinforce the choice of variables to be used in the final linear regression.

The stepwise model, forward stepwise and backward stepwise suggest a model consisting of three, maximum four predictors in terms of best R-squared, adjusted R-squared, AIC and C (p). In line with the Lasso regression, the regressors that survived the various features selection tests are 'Free cash flow', 'Net_Sales' and 'Return On Asset'. In figure 4, the variables considered by the 'forward' method can be observed, in figure 5 the variables eliminated by the 'backward' method and finally, in figure 6 the values assumed by the index R2, AIC and C (p).

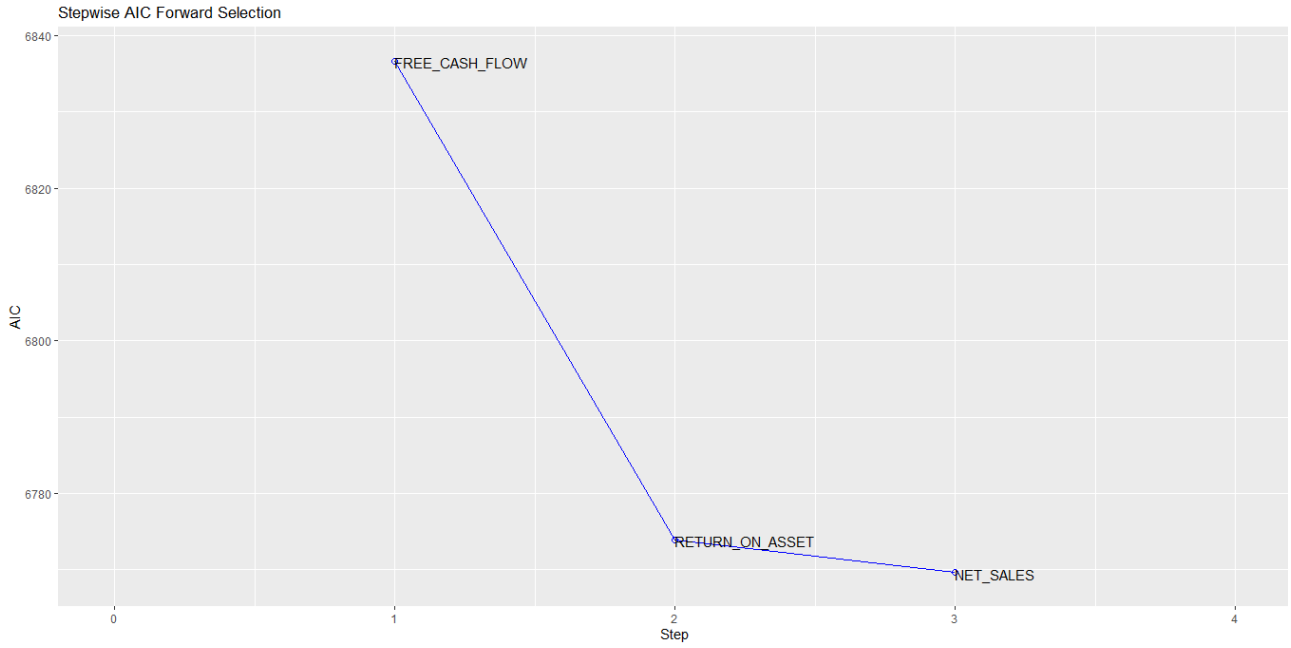

Figure 4- Stepwise AIC Forward Selection

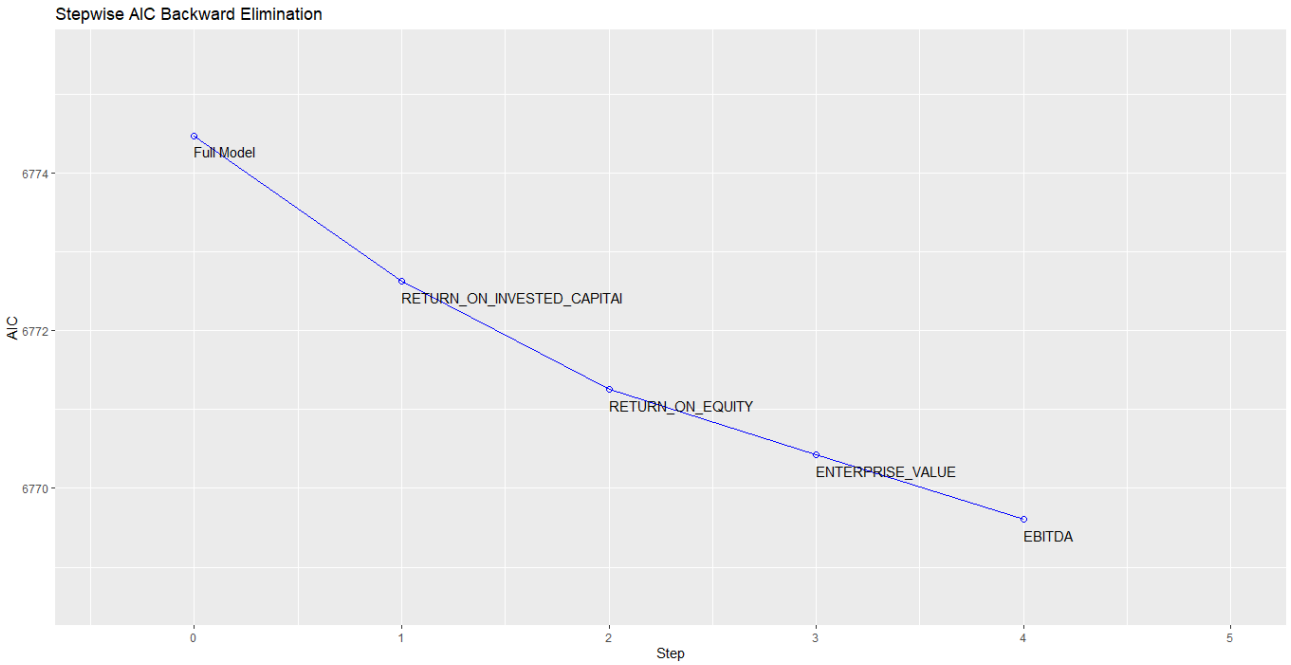

Figure 5- Stepwise AIC Backward Elimination

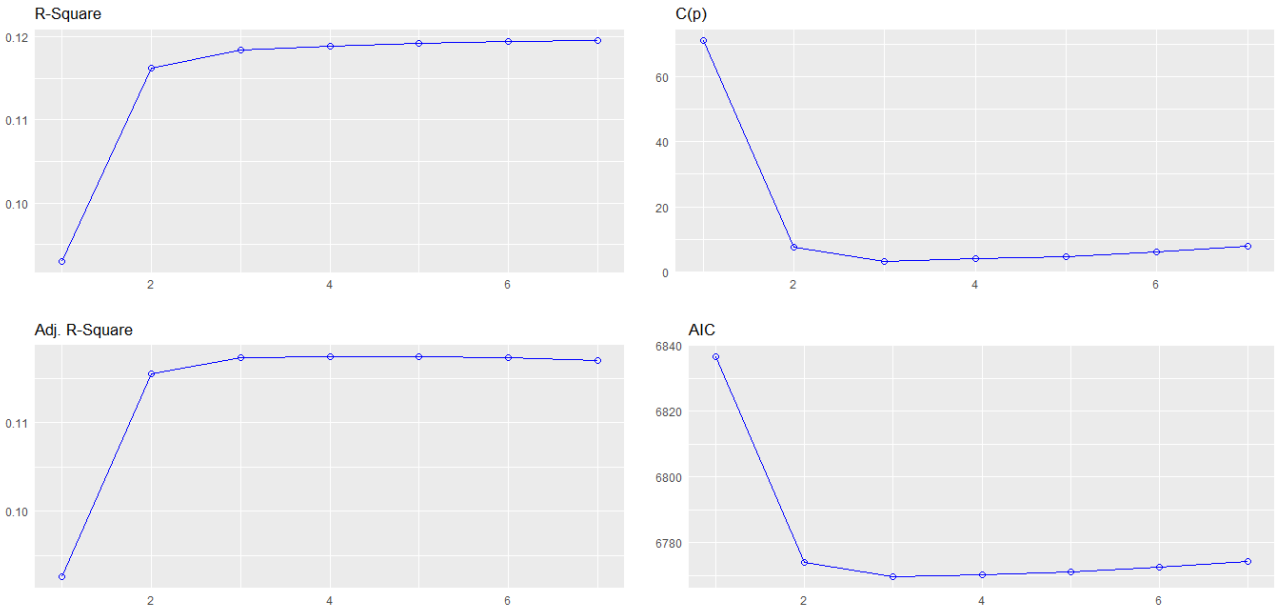

Figure 6- R-Square, Adjusted R-Squared, AIC and C (p) values

On the basis of the results obtained, it is suggested to conduct a linear regression analysis using the three variables mentioned above.

Once the model was created, the results obtained were subjected to an appropriate statistical analysis.

```
Call:
lm(formula = EPS_12M_FORWARD ~ ., data = train2)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8556 -0.2173 -0.1097  0.0753  6.8408

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.03165    0.01324  -2.391   0.0169 *
FREE_CASH_FLOW   0.10784    0.02097   5.142 3.02e-07 ***
NET_SALES        0.14966    0.02351   6.365 2.50e-10 ***
RETURN_ON_ASSET  0.11726    0.01256   9.337  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5528 on 1741 degrees of freedom
Multiple R-squared:  0.2091,    Adjusted R-squared:  0.2077
F-statistic: 153.4 on 3 and 1741 DF,  p-value: < 2.2e-16

> t.test(models$coefficients)

        One Sample t-test

data:  models$coefficients
t = 2.1362, df = 3, p-value = 0.1223
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.04200928  0.21356273
sample estimates:
 mean of x
0.08577673
```

Figure 7- Linear Regression and t-test

As can be seen in figure 7, the R-square is equal to 0.2091 (20%) and indicates how well the model is able to explain the Earning Per Share index with the chosen regressors. It is not an optimal result but it is the best R-framework that can be obtained with the current data available.

The regressors are all three significant, with positive coefficients and suggest that as cash-flow, year-end book value (Net_Sales) and 'Return On Asset (ROA)' increase, the EPS index increases positively. The standard-errors are small, which means that the fluctuations of the coefficients are also contained, the values of the t-value are large enough so that the t-statistic does not fall into the core of the distribution.

The p-values are all less than .05, so it is possible to reject the null hypothesis and validate the alternative hypothesis that there is at least one coefficient significantly other than zero.

The F-statistic is equal to 153.4, very high, while the p-value is sensibly small, confirming the fact that the regressors used make sense and are significant.

the Variance Inflaction Factor (VIF) suggests that the variables do not present collinearity problems, as the value of this index is contained for all three regressors: 2.72, 2.74 and 1.01 for respectively 'Free cash flow', 'Net_Sales' and' Return On Asset '.

Finally, an Anova analysis was conducted to verify the robustness of the model.

```
> anova(models)
Analysis of Variance Table

Response: EPS_12M_FORWARD
                 Df Sum Sq Mean Sq F value    Pr(>F)
FREE_CASH_FLOW    1  98.53  98.530 322.464 < 2.2e-16 ***
NET_SALES         1  15.45  15.447  50.555 1.685e-12 ***
RETURN_ON_ASSET   1  26.64  26.635  87.170 < 2.2e-16 ***
Residuals      1741 531.97   0.306
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 8- Anova Analysis

Given the results obtained, it is possible to confirm with a good probability that the 'Free cash flow' regressor has a higher F-value of all and is therefore the most powerful feature able to explain the dependent variable.

4. Interpretation of the results obtained

Once the model results had been analyzed, we proceeded by estimating the EPS index on the test-set, creating a confidence interval where the values can be expected to fluctuate.

In light of the results obtained from the predictions, we observe that for positive values of cash flows, the expected EPS index tends on average to always be positive while, for negative values of the aforementioned predictor, the index is in most cases assuming values negatives.

In the case of ALPHABET, for example, a leading company in the technology sector, the cash flow is positive and equal to 13.25, the ROA index is 1.08 and for this reason, with good probability we can say that the estimated EPS index will be positive and understood. within a confidence interval ranging between 2.5 and 3.2.

In the same way, the estimate of the EPS index turns out to be negative for the company ROCKET PHARMACEUTICALS. Value justified by as many negative values of the cash-flow equal to -0.21 and of the ROA index which turns out to be -1.2 (-27.87, not standardized real figure). Therefore, according to the analyzes carried out so far, it is possible to state with a good approximation that the EPS index will be negative in twelve months, included in an estimated confidence interval between -0.28 and -0.20. this result could be taken into consideration if the company decides to increase or not future investments, given the not exactly rosy estimate of the index in question.

Work done by Ridolfi Lorenzo.