Enhanced Value-Investing Stock Screening Model

Kobena Amoah

2/15/2024

Enhanced Stock Screening Model using Decision Trees and Data Envelopment Analysis (DEA)

Introduction:

The stock screening model begins with a Decision Tree Regression which aims to select the stocks that offered the best returns and identify the similarities among them. It does so by asking a series of questions of a stock to infer its total returns (cumulative returns over a year). Of course, the questions are predominantly based on accounting values where, a binary question, like "Is the Price/Sales ratio > 5?" is used to define a cut-off value. The algorithm then proceeds to apply Data Envelopment Analysis (DEA) on the stocks who currently share similar characteristics as those with the best returns in the past year that were identified in the Decision Tree. The application of Data Envelopment Analysis (DEA) aims to identify potentially undervalued stocks by comparing their financial performance metrics to their market valuation. DEA, a method derived from operational research, allows for a comprehensive comparison and ranking of stocks without assuming the importance or weights of individual metrics.

Methodology:

Decision Tree Regression:

Decision Tree Regressions involves a stratificaiton or segmentation of the predictor into a number of simple regions. It involves two basic steps.

1. The predictor space – that is, the set of possible values for $X_1, X_2, ..., X_P$ – into J distinct and non-overlapping regions, $R_1, R_2, ..., R_J$. Of course, the regions are constructed by finding the predictor space that minimizes the Residual Sum of Squares (RSS) which is given by

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j}^{J} (y_i - \hat{y}_{R_j})^2$$

Where:

$\hat{y}_{R_j}$ is the mean response for the training observations within the jth box. Given the computational infeasibility of considering every possible partition, the approach used is top-down as it starts at the top of the tree (at which point all observations belong to a single region) and each predictor space is successively split.

2. The process is repeated in order to split the data further so as to minimize the RSS of each of the resulting regions and continues until a stopping criterion is reached. For every observation that falls into the region R_j, we make the same prediction which is simply the mean response value for the training observations in R_j.

The process described above may produce good predictions, but is likely to overfit the data, as the resulting tree may be too complex. To combat this overfitting, the tree is pruned, in that the RSS that is used to split has to achive some (high) threshold. This pruning is achieved via a Grid Search Cross-Validation which performs an exhaustive search over all combinations of hyperparameters specified in the parameter grid, using cross-validation to evaluate the performance of each combination.

DEA:

After the completion Decision Tree Regression, the algorithm proceed to Data Envelopment Analysis which uses a model to calculate the efficiency E_i of each stock i using the following formula:

$$E_i = \frac{\sum_{r=1}^{N} u_{r,i} y_{r,i}}{\sum_{s=1}^{M} v_{s,i} x_{r,i}}$$

Where: E_i is the efficiency of stock i, u and v are the weights assigned to each output and input of the stock.

The problem of finding the best weights for a stock is formulated as an optimization problem:

$$\begin{aligned} \text{maximize} \quad h &= \frac{\sum_{r=1}^{N} u_{r,i} y_{r,i}}{\sum_{s=1}^{M} v_{s,i} x_{r,i}} \\ \text{subject to} \quad & \frac{\sum_{r=1}^{N} u_{r,i} y_{r,j}}{\sum_{s=1}^{M} v_{s,i} x_{r,j}} \leq 1 \, \text{for every record} \, j \\ & \text{and} \, u_{r,i}, v_{s,i} \geq 0 \end{aligned}$$

This optimization problem is solved using Linear Dynamic Programming.

Inputs and Outputs:

DTR:

- Inputs: Beta, operating margin, profit margin, revenue per share, return on assets, return on equity, EPS, revenue growth, leverage ratio, Trailing P/E, forward P/E, EV/Sales, EV/EBIT, P/BV, PEG, P/sales
- Outputs: Total Returns

DEA:

- Inputs: Beta, operating margin, profit margin, revenue per share, return on assets, return on equity, EPS, revenue growth, leverage ratio.
- Outputs: Trailing P/E, forward P/E, EV/Sales, EV/EBIT, P/BV, PEG, P/sales.

Application:

- The model is applied to all stocks, and those with the lowest efficiency are considered potentially undervalued.
- Further analysis is conducted on the most undervalued stocks to determine if their low valuation is justified by fundamental factors.
- This approach provides valuable insights and leads to deeper investigation into potential investment opportunities.

Conclusion:

By leveraging Decision Tree Regression and DEA, the stock screening model offers a robust systematic approach to identify undervalued stocks based on their financial performance and market valuation. This enables investors to uncover opportunities that may have been overlooked and make informed decisions in the realm of fundamental investing.

References

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. (2013). An introduction to statistical learning : with applications in R. New York :Springer,

Tong, Hogan (2020). Stock Screening Model Based on Data Envelopment Analysis (Version 2.0.4). https://github.com/HoagieT/Stock-Screening-Model-Based-On-Data-Envelopment-Analysis