

Calculation of RFE and Correlation Scores

RFE Score

Recursive Feature Elimination (RFE) is a Feature Selection Technique often used alongside machine learning models to identify which features contribute the most to predicting a target variable. RFE works as follows:

1. **Model Fit:** RFE starts by fitting a model (“RandomForestRegressor” in this study) to the entire set of features.
2. **Feature Importance:** it uses the model's built-in feature importance measure. For “RandomForest”, this is often based on how much each feature decreases the impurity of the split.
3. **Elimination:** the least important feature (or features, depending on the step size) is removed from the current set of features.
4. **Iteration:** the model is refitted with the reduced set of features, and the next least important feature is eliminated. This process is repeated.
5. **Ranking:** Each feature is given a rank based on the iteration in which it was removed. Features removed early receive a higher rank (considered less important), and features removed later (or those never removed if the process stops early) are assigned a lower rank (considered more important).

Calculation of RFE Scores

- Each feature is ranked, where a lower number means a more important feature, i.e. the feature ranked as 1 is the most important.

- **max_rank:** This is the maximum rank among all features. In a situation where all features are ranked (none are left out), this would be equal to the number of features.
- The formula adjusts the rank into a score on a 0 to 1 scale. Subtracting the rank from max_rank reverses the order. Therefore, higher original ranks, which are less important, yield lower scores.
- Dividing by (max_rank - 1) normalizes this value to range between 0 and 1. A feature with a rank of 1 (most important) would receive a score of 1 (as the top rank translates to the highest score), and the least important feature (highest rank, such as max_rank) would get a score close to 0.

Range of RFE Scores

- **Minimum:** The minimum value of an RFE score is close to 0 (but never exactly 0 unless the number of features is 1, which is a special case), which indicates the least important features according to the model used.
- **Maximum:** The maximum value of an RFE score is 1, indicating the most important feature or features in the model.

This scoring system effectively highlights the relative importance of features as determined by their contribution to the model's performance, taking into account the model's ability to predict the target variable accurately with or without each feature.

Correlation Scores

In the current code, the correlation scores are calculated using Spearman's rank correlation method. This method assesses how well the relationship between two variables can be described using a monotonic function; it doesn't assume a linear relationship between variables, nor does it require the variables to be on interval scales.

Calculation of Correlation Scores:

The formula for Spearman's rank correlation coefficient, denoted here as *SRCC*, between two variables x and y with n observations each, is given by:

$$SRCC = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i is the difference between the ranks of the corresponding elements x_i and y_i , and n is the number of observations (data points here).

Range of Correlation Scores

The Spearman correlation coefficient ranges from -1 to 1 , where:

- **-1:** Indicates a perfect negative monotonic relationship (as one variable increases, the other decreases in a perfectly predictable manner).
- **0:** Indicates no monotonic correlation (the variables do not change with respect to each other in any specific monotonic pattern).
- **1:** Indicates a perfect positive monotonic relationship (as one variable increases, the other also increases in a perfectly predictable manner).

By taking the absolute values, the range of the correlation scores in the current script changes to 0 to 1 , where:

- **0:** No correlation. The variables do not share a monotonic relationship.
- **1:** Perfect correlation. The variables move together either always increasing or always decreasing in a perfectly predictable monotonic pattern.