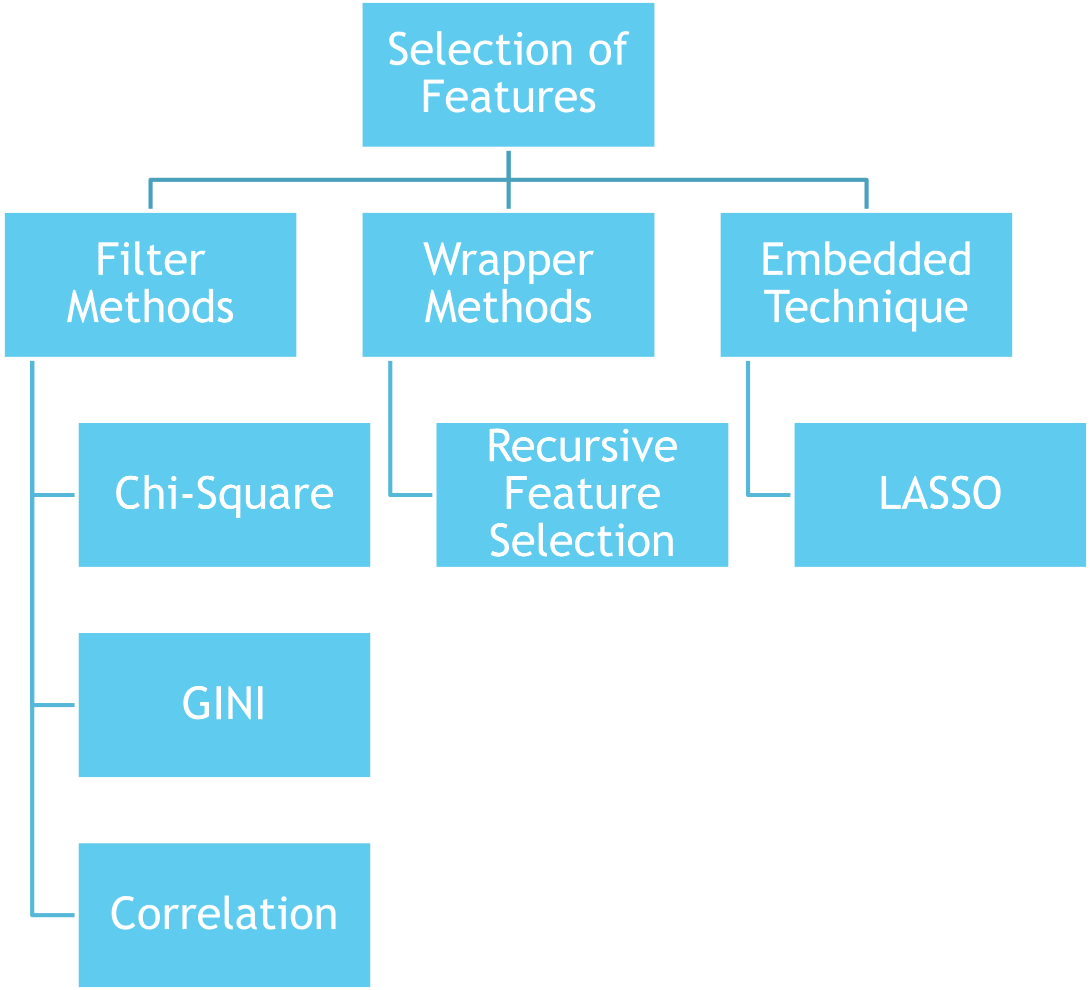
Feature Selection

DataDNA

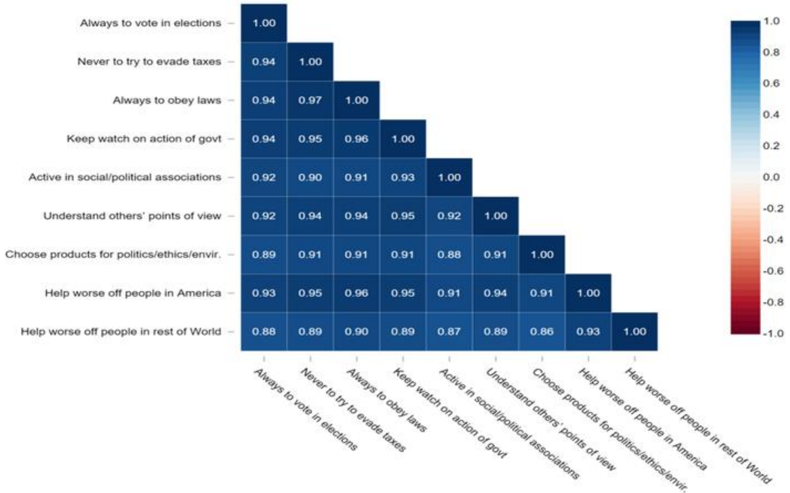
Advanced Feature Selection Techniques



Correlation Coefficients

- Correlation is a bivariate analysis to choose variables/understand variables
- It’s value is between -1 to +1 (Highly Negative to Highly Positive Correlation)
- Since it is bivariate , the method does not consider multiple variable interactions.
- It can be used in the initial stages of variable selection

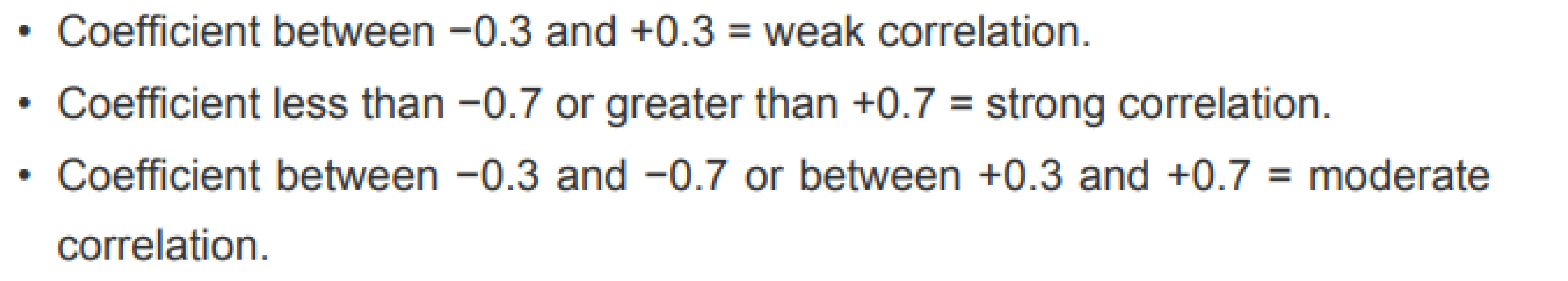
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$



	Always to vote in elections	Never to try to evade taxes	Always to obey laws	Keep watch on action of govt	Active in social/political associations	Understand others' points of view	Choose products for politics/ethics/envir.	Help worse off people in America	Help worse off people in rest of World
Always to vote in elections	1.00								
Never to try to evade taxes	0.94	1.00							
Always to obey laws	0.94	0.97	1.00						
Keep watch on action of govt	0.94	0.95	0.96	1.00					
Active in social/political associations	0.92	0.90	0.91	0.93	1.00				
Understand others' points of view	0.92	0.94	0.94	0.95	0.92	1.00			
Choose products for politics/ethics/envir.	0.89	0.91	0.91	0.91	0.88	0.91	1.00		
Help worse off people in America	0.93	0.95	0.96	0.95	0.91	0.94	0.91	1.00	
Help worse off people in rest of World	0.88	0.89	0.90	0.89	0.87	0.89	0.86	0.93	1.00

Correlation Coefficients – Rule of Thumb

Reference: Source Link



- Coefficient between −0.3 and +0.3 = weak correlation.
- Coefficient less than −0.7 or greater than +0.7 = strong correlation.
- Coefficient between −0.3 and −0.7 or between +0.3 and +0.7 = moderate correlation.

GINI Index

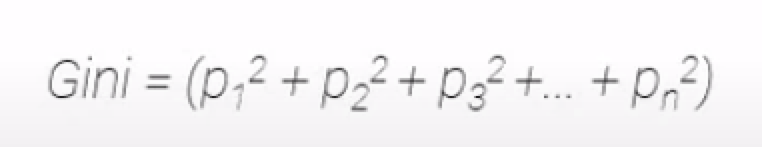
- One of the most common methods to do feature selection.
- It gives a mathematical number to identify top features
- The Gini coefficient measures the inequality among values of a frequency distribution (for example, levels of income).
- **A Gini coefficient of zero** expresses **perfect equality**, where all values are the same (for example, where everyone has the same income).
- **A Gini coefficient of one** (or 100%) expresses **maximal inequality** among values (e.g., for a large number of people, where only one person has all the income or consumption, and all others have none, the Gini coefficient will be very nearly one).
- If it's continuous, it is intuitive that you have subset A with value <= some threshold and subset B with value > that threshold.
- If it's categorical, to make things simpler, say the variable has 2 categories. Then subset A will be a subset of original dataset with this variable equals category 1 and subset B will be the subset with this variable equals category 2.

Steps to calculate Gini

- Calculate the Gini index for sub nodes

  Gini Index = 1- Gini

- Gini = Sum of square of probabilities for each class/category



$$Gini = (p_1^2 + p_2^2 + p_3^2 + ... + p_n^2)$$

- To calculate the Gini index for equality, take weighted Gini impurity of sub nodes of the split

Higher the Gini Index, Lesser is the homogeneity

Recursive Feature Elimination

- ***Recursive*** – Repetitive
- ***Feature*** – All IVs
- ***Elimination*** – Removing it from further phases of Modelling

- Recursively eliminate the insignificant variables(no relationships variables) from the analysis

RFE = > Wrapper Style Feature Selection Algorithm

- Parameters:
  - Algorithm (Eg:- Decision Tree Clasifier, SVM, Linear Regression)
  - No of Features to be selected

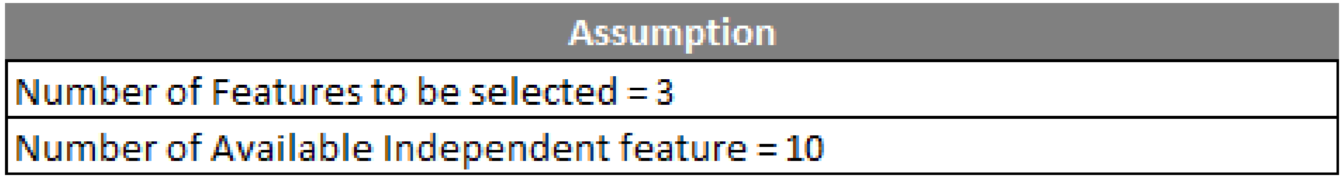
- ***Why Should we do this ?***
  - Too many variables can lead to slowness in execution
  - Helps in reducing the complexity of model
  - Increases the capability of model to find the better underlying patterns

Recursive Feature Elimination

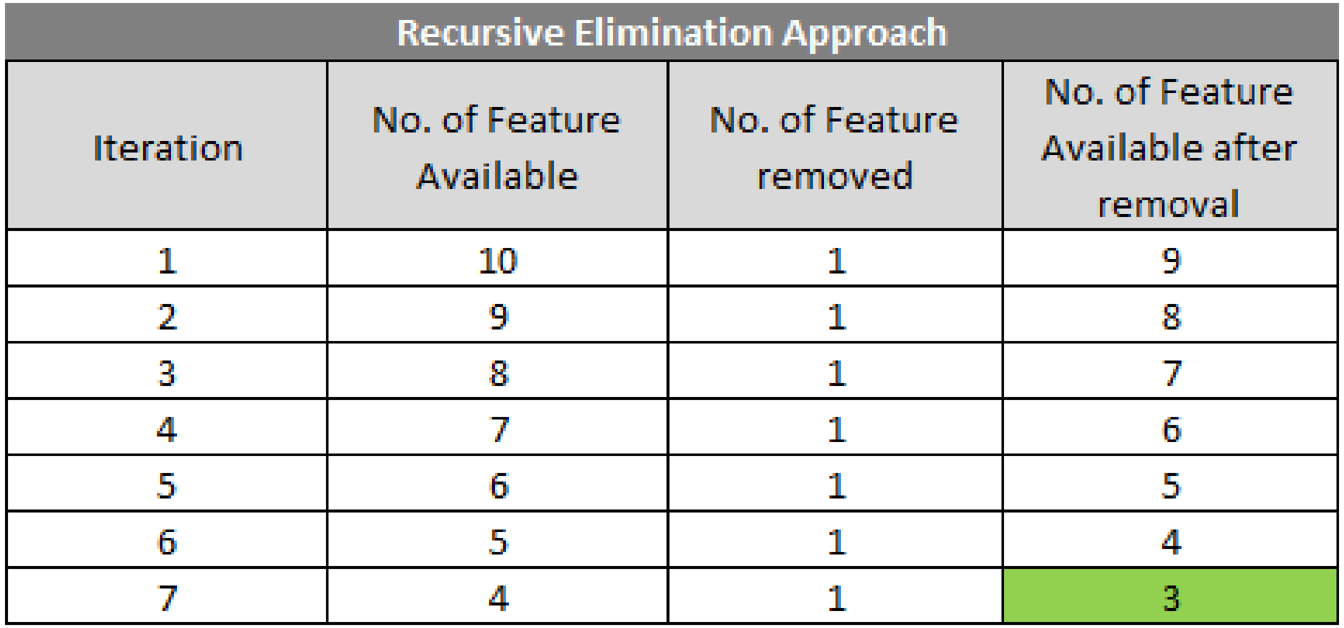
- Recursive Feature Elimination (RFE) recursively removes features, builds a model using the remaining attributes and calculates model accuracy.
- It has forward, backward mechanism where it adds & removes variables from the model built
- Supports regression models & classification algorithms
- we have coefficients of each feature or feature importance.
- We drop the feature with least coefficient or importance. Then the model is fit on the remaining features.
- The process is repeated until we have a necessary number of features (or some other criteria is fulfilled).

Recursive Feature Elimination

Strategy: One Feature Removal at a time



Assumption
Number of Features to be selected = 3
Number of Available Independent feature = 10

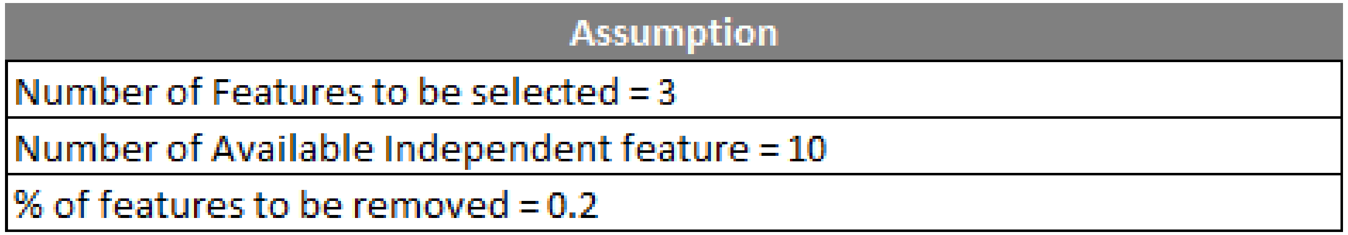


Recursive Elimination Approach			
Iteration	No. of Feature Available	No. of Feature removed	No. of Feature Available after removal
1	10	1	9
2	9	1	8
3	8	1	7
4	7	1	6
5	6	1	5
6	5	1	4
7	4	1	3

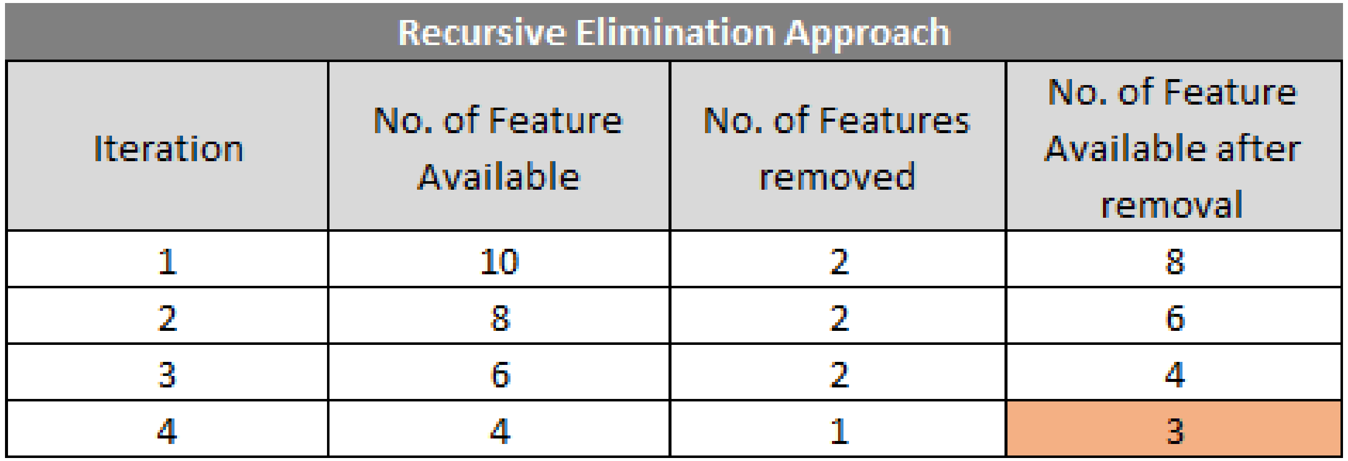
☞ So once desired number is reached, it will rank the selected variable based on the coefficient or importance which is dependent on algorithm.

Recursive Feature Elimination

Strategy: % of Features Removal in an iteration



Assumption
Number of Features to be selected = 3
Number of Available Independent feature = 10
% of features to be removed = 0.2



Recursive Elimination Approach			
Iteration	No. of Feature Available	No. of Features removed	No. of Feature Available after removal
1	10	2	8
2	8	2	6
3	6	2	4
4	4	1	3

So once desired number is reached, it will rank the selected variable based on the coefficient or importance which is dependent on algorithm.