# Feature Engineering

Feature selection methods are intended to reduce the number of input variables to those that are believed to be most useful to a model in order to predict the target variable. We have performed the feature selection techniques highlighted in bold below.

We can summarize feature selection as follows.

* Feature Selection: Select a subset of input features from the dataset.
  + Unsupervised: Do not use the target variable (e.g. remove redundant variables).
    - * Correlation – **Pearson Correlation Coefficient, Variance Threshold**
  + Supervised: Use the target variable (e.g. remove irrelevant variables).
    - Wrapper: Search for well-performing subsets of features.
      * RFE
    - Filter: Select subsets of features based on their relationship with the target.
      * Statistical Methods – **Univariate Selection**
      * Feature Importance Methods
    - Intrinsic: Algorithms that perform automatic feature selection during training.
      * Decision Trees
* Dimensionality Reduction: Project input data into a lower-dimensional feature space.

## Filter Selection methods

* Filter methods: The features are ranked by the score and either selected to be kept or removed from the dataset.
* Basic filter methods, correlation coefficient scores, Chi-squared test, mutual information
* Wrapper methods: The selection of a set of features as a search problem, where different combinations are prepared, evaluated and compared to other combinations.
  + Forward selection, backward selection, exhaustive feature selection
* Embedded methods–Combine the advantages of both methods–Include the feature selection process in the machine learning model training
  + Regularization methods such as LASSO, elastic net and ridge regression, decision tree algorithms (giniindex, information gain)
* Wrapper/Embedded methods
  + Recursive feature elimination–Recursive feature elimination with CV

## Filter Methods

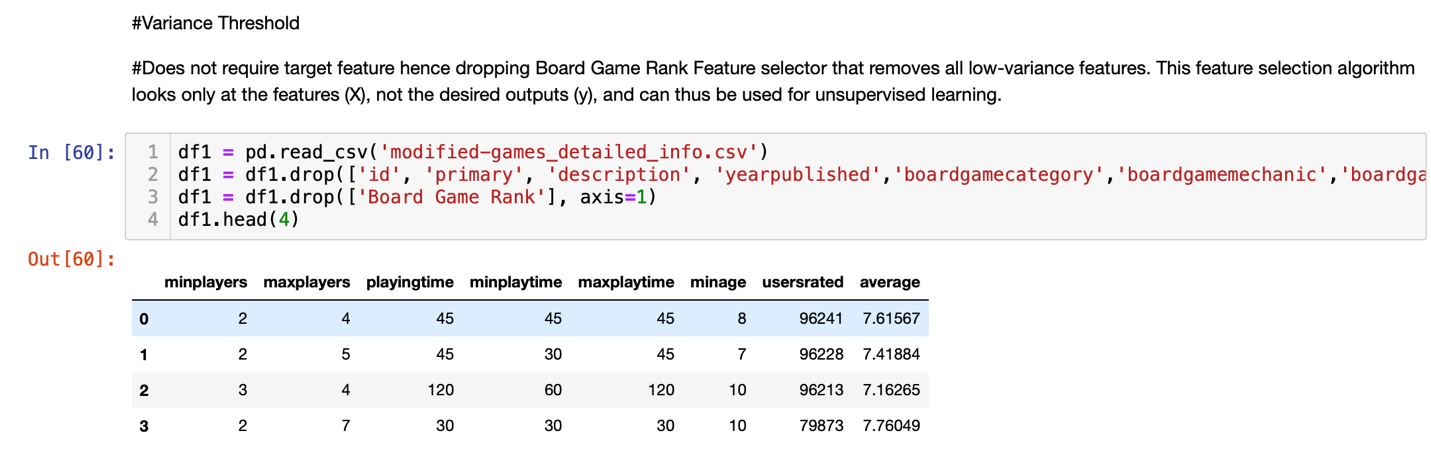
Select features independently of the machine learning algorithm model

* Basic filter methods
  + Remove constant features
  + Remove quasi-constant features
  + Remove duplicated features
* Correlation methods
  + Features are selected on the basis of their scores between their correlations
  + correlation coefficient scores
* Statistical ranking filter methods
  + Features are statistically ranked by the score and either selected to be kept or removed from the dataset
  + Chi-squared test, mutual information

## Variance Threshold

Variance Threshold removes features with a variance less than the specified threshold. Consider a feature that takes the same value for all the observations (rows) in the dataset. It would not add any informative power to a model. Using this feature also adds an unnecessary computation burden. Thus, we should just eliminate it from the dataset. Similarly, features with a very small variance can also be omitted.

[sklearn.feature\_selection](https://scikit-learn.org/stable/modules/classes.html#module-sklearn.feature_selection).VarianceThreshold is a feature selector that removes all low-variance features. This feature selection algorithm looks only at the features (X), not the desired outputs (y), and can thus be used for unsupervised learning.



Table

Description automatically generated

Graphical user interface, application, table

Description automatically generated

Table

Description automatically generated

* After processing, we can conclude that column ‘maxplaytime’ had the least variance and hence was dropped from the dataset.

## Pearson Correlation Coefficient

A Pearson correlation is a number between -1 and 1 that indicates the extent to which two variables are linearly related. The Pearson correlation is also known as the “product moment correlation coefficient” (PMCC) or simply “correlation”. Pearson correlations are suitable only for metric variables.

The correlation coefficient has values between -1 to 1

* A value closer to 0 implies weaker correlation (exact 0 implying no correlation)
* A value closer to 1 implies stronger positive correlation
* A value closer to -1 implies stronger negative correlation

In Pearson correlation coefficient:

* Both variables are normally distributed (Gaussian distributions)
* A straight-line relationship between the two variables
* Data is distributed around the regression line
* Pearson correlation coefficient and linear regression are highly correlated

Table

Description automatically generated

Chart

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

From the above results , we can observe that the feature ‘Board Game Rank’ and ‘maxplaytime’ are somewhat corelated.

## Univariate Selection

Univariate feature selection works by selecting the best features based on univariate statistical tests. We compare each feature to the target variable, to see whether there is any statistically significant relationship between them. It is also called analysis of variance (ANOVA). When we analyze the relationship between one feature and the target variable, we ignore the other features. That is why it is called ‘univariate’.

Each feature has its test score. Finally, all the test scores are compared, and the features with top scores will be selected. Statistical tests can be used to select those features that have the strongest relationship with the output variable. The scikit-learn library provides the SelectKBest class that can be used with a suite of different statistical tests to select a specific number of features.

The example below uses the chi-squared (chi²) statistical test for non-negative features to select 10 of the best features from the Mobile Price Range Prediction Dataset. It is very similar to Spearman’s rank correlation coefficient.

Table

Description automatically generated

A picture containing table

Description automatically generated

Text

Description automatically generated

* Since we have 7 features we will try to select just 7 best features.
* From the above score we can see that ‘usersrated’ and ‘playingtime’ are directly corelated to Board Game Rank and so on in that order. So more the number of users rated and more the playing time, higher will be the Game rank.
* A good board game rank will specify a smaller value Ex. Game at rank 1 will be better than game at rank 78. So less the playing time of game, less or better will be the board game rank.
* We can make some inference from these results if we wish to design a game with good Board Game rank
  + Keep the playing time and maximum playing time of game short since it directly affects the board game rank.
  + Minimum players and minimum age would not have much impact on the board game rank.