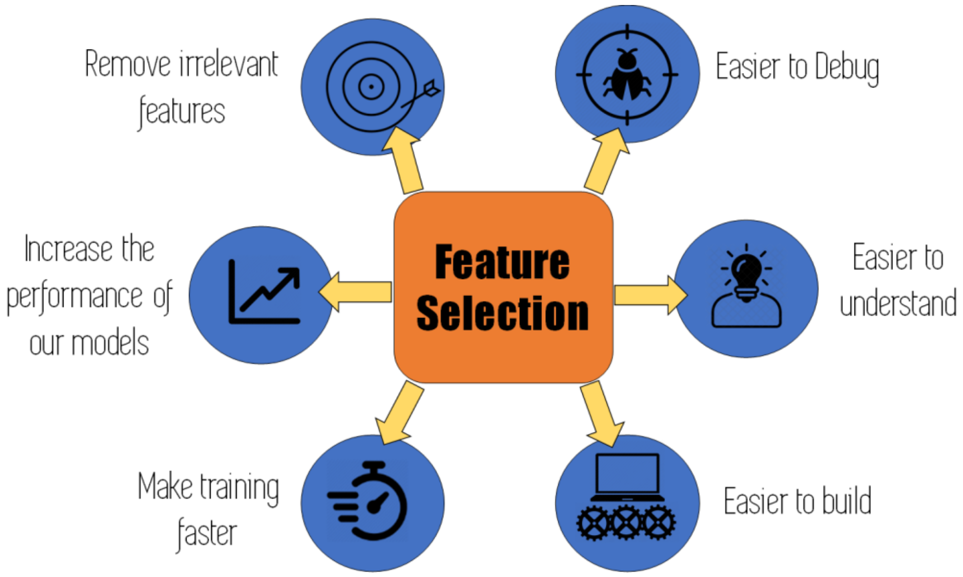
Chapter 6
Feature Selection

Machine learning works on a simple rule – if you put garbage in, you will only get garbage to come out. By garbage here mean noise in data. This becomes even more important when the number of features are very large. You need not use every feature at your disposal for creating an algorithm. You can assist your algorithm by feeding in only those features that are really important. It is proven that feature subsets giving better results than complete set of feature for the same algorithm.

Feature Selection is one of the core concepts in machine learning which hugely impacts the performance of your model. The data features that you use to train your machine learning models have a huge influence on the performance you can achieve. Irrelevant or partially relevant features can negatively impact model performance. Feature selection and Data cleaning should be the first and most important step of your model designing.

6.1 Why should we do Feature selection?

Alright, now that we know what feature selection is, let's see why we should use it when training a Machine Learning model:



Source: https://towardsdatascience.com/an-introduction-to-feature-selection-dd72535ecf2b

- Using feature selection, we can remove irrelevant features that would not be affecting or changing the output of our model. If we try to predict the price of a house in Spain, using variables that include the weather conditions in China, these variables will probably not be very useful.

- These kind of irrelevant features can actually decrease the performance of your model by introducing noise.
- Less features usually means faster training models: for parametric models like linear or logistic regression, it means there are less weights to calculate, and for non-parametric models like Random Forest of Decision trees, it means there are less features to evaluate at each split.
- When putting models into production, less features means less work for the team building the application that will use the model. Using feature selection, we can reduce the integration time for the application.
- When we keep the most important features, discarding the ones that our feature selection methods advise us to remove, our model becomes simpler, and easier to understand. A model with 25 features is a lot simpler than a model with 200 features.
- Once the application has been finished, and is being used periodically, a model with fewer features is a lot easier to debug in case of abnormal behaviour than a model with a lot of features.

6.2 What is Feature Selection?

Feature selection is also called variable selection or attribute selection. It is the automatic selection of attributes in your data (such as columns in tabular data) that are most relevant to the predictive modeling problem you are working on.

"*Feature selection is the process of selecting a subset of relevant features for use in model construction*."

6.3. Feature selection Vs Dimensionality Reduction

Feature selection is different from dimensionality reduction. Both methods seek to reduce the number of attributes in the dataset, but a dimensionality reduction method do so by creating new combinations of attributes, whereas feature selection methods include and exclude attributes present in the data without changing them.

Examples of dimensionality reduction methods include Principal Component Analysis, Singular Value Decomposition and Sammon’s Mapping.

6.4 Feature Selection Algorithms

There are two general classes of feature selection algorithms: filter methods and wrapper methods

Filter Methods

Filter feature selection methods apply a statistical measure to assign a score to each feature. The features are ranked by the score and either selected to be kept or removed from the dataset. The methods are often univariate and consider the feature independently, or with regard to the dependent variable. Some examples of filter methods include the Chi squared test, information gain and correlation coefficient scores.

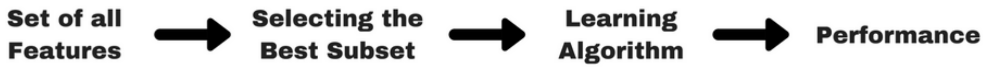
Wrapper Methods

Wrapper methods consider the selection of a set of features as a search problem, where different combinations are prepared, evaluated and compared to other combinations. A predictive model used to evaluate a combination of features and assign a score based on model accuracy.

The search process may be methodical such as a best-first search, it may stochastic such as a random hill-climbing algorithm, or it may use heuristics, like forward and backward passes to add and remove features. An example if a wrapper method is the recursive feature elimination algorithm.

6.5 Filter Approach

Filter methods are generally used as a preprocessing step. The selection of features is independent of any machine learning algorithms. Instead, features are selected on the basis of their scores in various statistical tests (such as Pearson's Correlation, LDA, ANOVA and Chi-Square) for their correlation with the outcome variable.



Pearson's Correlation: It is used as a measure for quantifying linear dependence between two continuous variables X and Y. Its value varies from -1 to +1. Pearson's correlation is given as:

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

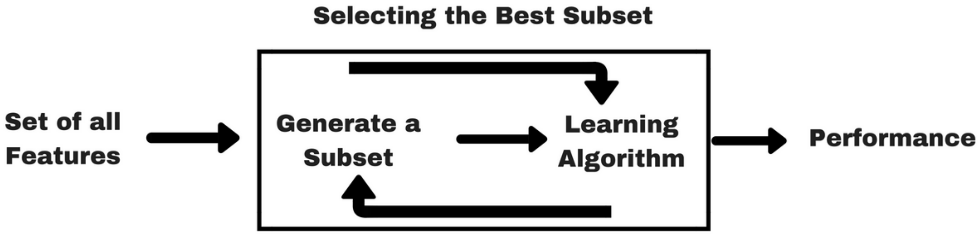
LDA: Linear discriminant analysis is used to find a linear combination of features that characterizes or separates two or more classes (or levels) of a categorical variable.

ANOVA: ANOVA stands for *Analysis of variance*. It is similar to LDA except for the fact that it is operated using one or more categorical independent features and one continuous dependent feature. It provides a statistical test of whether the means of several groups are equal or not.

Chi-Square: It is a is a statistical test applied to the groups of categorical features to evaluate the likelihood of correlation or association between them using their frequency distribution.

6.6 Wrappers Approach

- It Optimizes for a specific learning algorithm
- The feature subset selection algorithm is a "wrapper" around the learning algorithm
  - ✓ Pick a feature subset and pass it in to learning algorithm
  - ✓ Create training/test set based on the feature subset
  - ✓ Train the learning algorithm with the training set
  - ✓ Find accuracy (objective) with validation set
  - ✓ Repeat for all feature subsets and pick the feature subset which led to the highest predictive accuracy (or other objective)
- Basic approach is simple
- Variations are based on how to select the feature subsets, since there are an exponential number of subsets
- Evaluation uses criteria related to the classification algorithm.
- The objective function is a pattern classifier, which evaluates feature subsets by their predictive accuracy (recognition rate on test data) by statistical resampling or cross-validation.



Forward Search

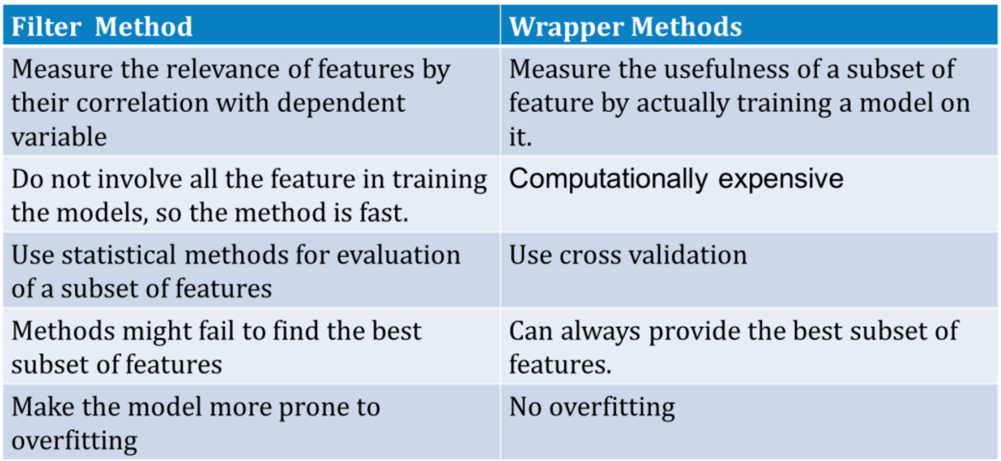
- Score each feature by itself and add the best feature to the initially empty set FS (FS will be our final Feature Set)
- Try each subset consisting of the current FS plus one remaining feature and add the best feature to FS
- Continue until stop getting significant improvement (over a window)

Backward Search

- Score the initial complete set FS (FS will be our final Feature Set)

- Try each subset consisting of the current FS minus one feature in FS and drop the feature from FS causing least decrease in accuracy
- Continue until begin to get significant decreases in accuracy

6.7 Difference between Filter and Wrapper Methods



Filter Method	Wrapper Methods
Measure the relevance of features by their correlation with dependent variable	Measure the usefulness of a subset of feature by actually training a model on it.
Do not involve all the feature in training the models, so the method is fast.	Computationally expensive
Use statistical methods for evaluation of a subset of features	Use cross validation
Methods might fail to find the best subset of features	Can always provide the best subset of features.
Make the model more prone to overfitting	No overfitting