# DATA CLEANING

Data cleaning is one of the most important steps in machine learning workflow cycle.

* Outlier treatment (treat / remove outliers).
* Imputing missing data (replace None or NaN or NULL).
* Identifying Malicious data (Magical values).
* Identifying erroneous data.
* Irrelevant data (remove irrelevant data to business query).
* Inconsistent data (same data present in different ways).
* Formatting (same formatting across observations in a feature).

# FEATURES SELECTION METHODS

Feature Selection is a very critical component in a Data Scientist’s workflow. When presented data with very high dimensionality, models usually choke because

* Training time increases exponentially with number of features.
* Models have increasing risk of overfitting with increasing number of features.
* Model complexity increases and interpretability decreases with increase in features

Feature Selection methods helps with these problems by reducing the dimensions without much loss of the total information. It also helps to make sense of the features and its importance.

The process of finding and selecting the most useful features in a dataset, is a crucial step while making a model using machine learning. Unnecessary features decrease training speed, decrease model interpretability and most importantly decrease the performance and accuracy of the model.

The main aim behind feature selection is to maintain a balance between model accuracy and model interpretability.

Some of the basic rules while selecting features from a data set are below:

* Missing Values: If there are features in the data set with fraction of missing values above 60%, drop those columns.
* Collinear Features: Drop all the features with high correlation coefficient i.e. features dependent on each other.
* Single Unique Value Features: A feature with only one unique value cannot be useful for machine learning because this feature has zero variance so we can drop that variable.
* Identity values: A feature with more unique values will not be useful as it will not convey any information about the dependent variable.

Feature selection can be done using:

* Filter methods
* Wrapper methods
* Embedded methods

## FILTER METHODS

Filter methods are generally used as a pre-processing 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 for their correlation with the outcome variable. The correlation is a subjective term here. For basic guidance, you can refer to the following table for defining correlation coefficients.

These are also known as single factor analysis or univariate analysis. Using these methods, we analyse the importance of an independent variable with the dependent variable.

Below are some techniques used:

* Correction with target variable (correlation matrix).
* Information value in the variable (mostly used in logistic regression)
* Chi square value (categorical variable IV and DV)

A variable is said to have more information value if the variable is correlated with the dependent variable and has large variation. A variable with less variation is said to have less information about dependent variable.

Weight of evidence (WOE) and Information value (IV) are simple, yet powerful techniques to perform variable transformation and selection. These concepts have huge connection with the logistic regression modeling technique. It is widely used in credit scoring to measure the separation of good vs bad customers.

## WRAPPER METHODS

In wrapper methods, we try to use a subset of features and train a model using them. Based on the inferences that we draw from the previous model, we decide to add or remove features from your subset.

* **Forward Selection:** Forward selection is an iterative method in which we start with having no feature in the model. In each iteration, we keep adding the feature which best improves our model till an addition of a new variable does not improve the performance of the model.
* **Backward Elimination:** In backward elimination, we start with all the features and removes the least significant feature at each iteration which improves the performance of the model. We repeat this until no improvement is observed on removal of features.
* **Recursive Feature elimination:** It is a greedy optimization algorithm which aims to find the best performing feature subset. It repeatedly creates models and keeps aside the best or the worst performing feature at each iteration. It constructs the next model with the left features until all the features are exhausted. It then ranks the features based on the order of their elimination.

## AUTOMATED / EMBEDDED METHODS

Feature selection is implemented by algorithms which have their own built-in feature selection methods.

Some of the most popular examples of these methods are LASSO and RIDGE regression which have inbuilt penalization functions to reduce overfitting.

* Lasso regression performs L1 regularization which adds penalty equivalent to absolute value of the magnitude of coefficients.
* Ridge regression performs L2 regularization which adds penalty equivalent to square of the magnitude of coefficients.

## DIFFERENCE BETWEEN FILTER AND WRAPPER METHODS

The main differences between the filter and wrapper methods for feature selection are:

* Filter methods measure the relevance of features by their correlation with dependent variable while wrapper methods measure the usefulness of a subset of feature by actually training a model on it.
* Filter methods are much faster compared to wrapper methods as they do not involve training the models. On the other hand, wrapper methods are computationally very expensive as well.
* Filter methods use statistical methods for evaluation of a subset of features while wrapper methods use cross validation.
* Filter methods might fail to find the best subset of features in many occasions but wrapper methods can always provide the best subset of features.

The importance of feature selection for you:

* It enables the machine learning algorithm to train faster.
* It reduces the complexity of a model and makes it easier to interpret.
* It improves the accuracy of a model if the right subset is chosen.
* It reduces Overfitting.

## F-TEST

F Test is a statistical test used to compare between models and check if the difference is significant between the model.

F-Test does a hypothesis testing model X and Y where X is a model created by just a constant and Y is the model created by a constant and a feature.

The least square errors in both the models are compared and checks if the difference in errors between model X and Y are significant or introduced by chance.

F-Test is useful in feature selection as we get to know the significance of each feature in improving the model.

Scikit learn provides the Selecting K best features using F-Test.



For Classification tasks



DRAWBACK

There are some drawbacks of using F-Test to select your features. F-Test checks for and only captures linear relationships between features and labels. A highly correlated feature is given higher score and less correlated features are given lower score.

Correlation is highly deceptive as it doesn’t capture strong non-linear relationships.

## MUTUAL INFORMATION

Mutual Information between two variables measures the dependence of one variable to another. If X and Y are two variables, and

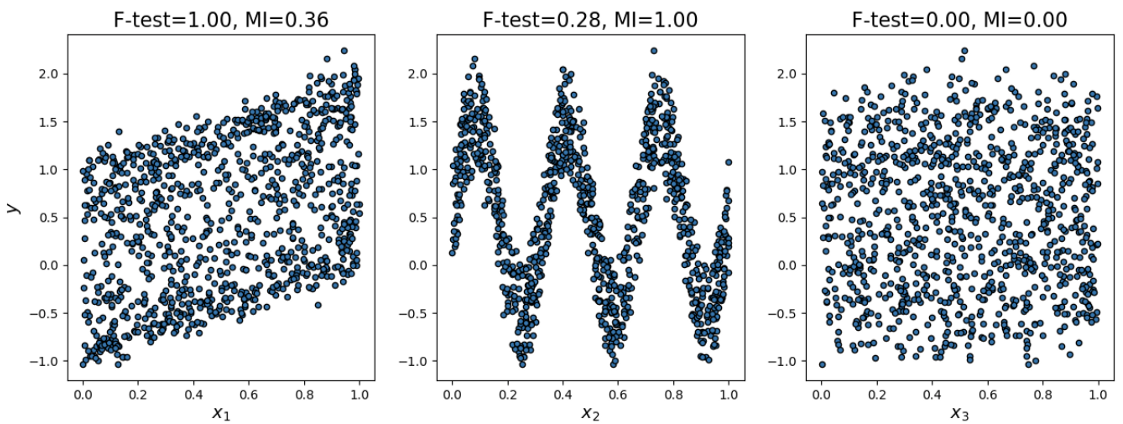
* If X and Y are independent, then no information about Y can be obtained by knowing X or vice versa. Hence their mutual information is 0.
* If X is a deterministic function of Y, then we can determine X from Y and Y from X with mutual information 1.
* When we have Y = f (X, Z, M, N) then 0 < mutual information < 1

We can select our features from feature space by ranking their mutual information with the target variable.

Advantage of using mutual information over F-Test is, it does well with the non-linear relationship between feature and target variable.

Sklearn offers feature selection with Mutual Information for regression and classification tasks.





F-Test captures the linear relationship well. Mutual Information captures any kind of relationship between two variables.

## VARIANCE THRESHOLD

This method removes features with variation below a certain cutoff.

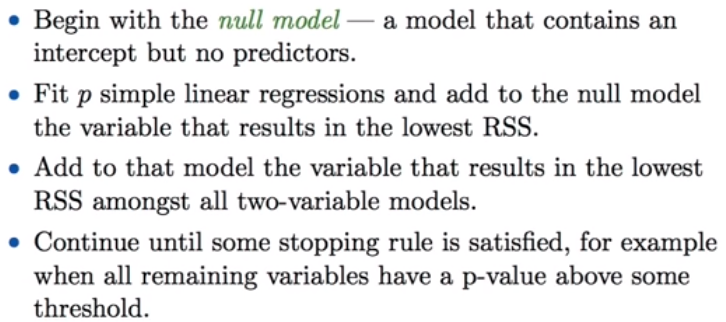
The idea is when a feature doesn’t vary much within itself, it generally has very little predictive power.



Variance Threshold doesn’t consider the relationship of features with the target variable.

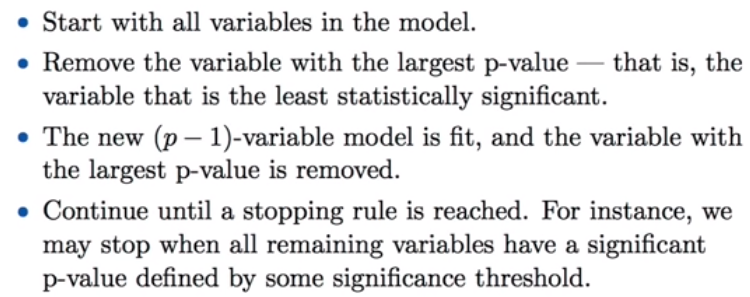
## FORWARD SELECTION

See below



## BACKWARD SELECTION

See below



## MACHINE LEARNING ALGORITHMS

Feature selection can also be achieved by the insights provided by some Machine Learning models.

LASSO Linear Regression can be used for feature selections. Lasso Regression is performed by adding an extra term to the cost function of Linear Regression. This apart from preventing overfitting also reduces the coefficients of less important features to zero.

Other than Lasso, we also have ensemble algorithms which have important feature attributes by which we can get importance score for each feature.

# OUTLIER DETECTION

Data exploration consists of many things, such as variable identification, treating missing values, feature engineering, etc. Detecting and treating outliers is also a major cog in the data exploration stage. The quality of your inputs decides the quality of your output.

# DIMENSION REDUCTION

Sometimes, feature selection is mistaken with dimensionality reduction. But they are different. Feature selection is different from dimensionality reduction. Both methods tend to reduce the number of attributes in the dataset, but a dimensionality reduction method does so by creating new combinations of attributes (sometimes known as feature transformation), whereas feature selection methods include and exclude attributes present in the data without changing them.

Some examples of dimensionality reduction methods are:

* Factor analysis.
* Principal Component Analysis.
* Linear Discriminant Analysis.