1. What is feature engineering, and how does it work? Explain the various aspects of feature engineering in depth.

Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.

Feature engineering in ML consists of four main steps: Feature Creation, Transformations, Feature Extraction, and Feature Selection. Feature engineering consists of creation, transformation, extraction, and selection of features, also known as variables, that are most conducive to creating an accurate ML algorithm.

Feature engineering is the process that takes raw data and transforms it into features that can be used to create a predictive model using machine learning or statistical modelling

2. What is feature selection, and how does it work? What is the aim of it? What are the various methods of function selection?

Feature selection is the process of reducing the number of input variables when developing a predictive model. It is desirable to reduce the number of input variables to both reduce the computational cost of modelling and, in some cases, to improve the performance of the model.

The objective of feature selection is to remove irrelevant and/or redundant features and retain only relevant features. Irrelevant features can be removed without affecting learning performance. Redundant features are a type of irrelevant features.

There are three types of feature selection: Wrapper methods (forward, backward, and stepwise selection), Filter methods (ANOVA, Pearson correlation, variance thresholding), and Embedded methods (Lasso, Ridge, Decision Tree).

3. Describe the function selection filter and wrapper approaches. State the pros and cons of each approach?

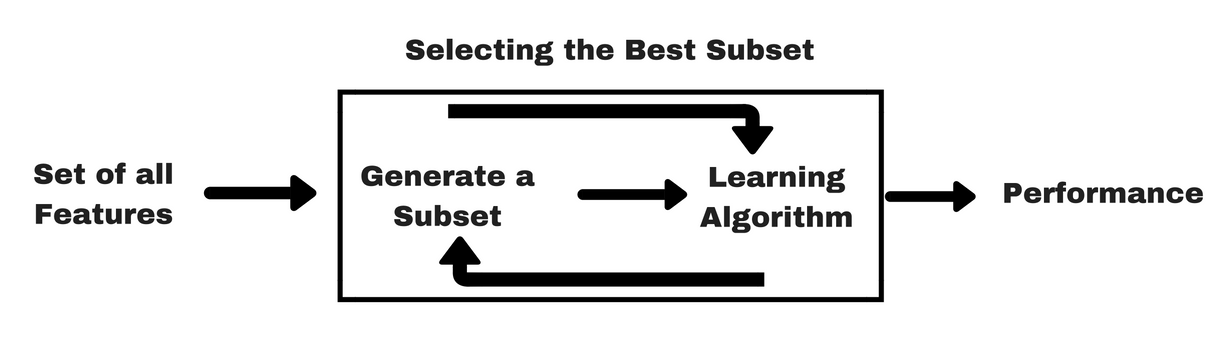
The main differences between the filter and wrapper methods for feature selection are:

* 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.
* Using the subset of features from the wrapper methods make the model more prone to overfitting as compared to using subset of features from the filter methods.

**A.Filter approaches:**



**B. Wrapper approaches:**



The filter method has the fastest running time; however, it does not consider feature dependencies and tends to each feature separately when univariate techniques are used . The wrapper method has the advantages of better generalization and robust interaction with the classifier used for feature selection

4.

i. Describe the overall feature selection process.

Feature selection is also related to [dimensionally reduction](https://machinelearningmastery.com/dimensionality-reduction-for-machine-learning/) techniques in that both methods seek fewer input variables to a predictive model. The difference is that feature selection select features to keep or remove from the dataset, whereas dimensionality reduction create a projection of the data resulting in entirely new input features. As such, dimensionality reduction is an alternate to feature selection rather than a type of feature selection.

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
  + **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
      * 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.

ii. Explain the key underlying principle of feature extraction using an example. What are the most widely used function extraction algorithms?

Feature extraction is a process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing. A characteristic of these large data sets is a large number of variables that require a lot of computing resources to process.

PCA is the optimal procedure for feature selection. However, there are several procedures for feature selection and different procedures may give different solution for the Same data set

5. Describe the feature engineering process in the sense of a text categorization issue.

Text classification is the problem of assigning categories to text data according to its content. The most important part of text classification is feature engineering: the process of creating features for a machine learning model from raw text data.

Feature extraction identifies those product aspects which are being commented by customers, sentiment prediction identifies the text containing sentiment or opinion by deciding sentiment polarity as positive, negative or neutral and finally summarization module aggregates the results obtained from previous two steps.

6. What makes cosine similarity a good metric for text categorization? A document-term matrix has two rows with values of (2, 3, 2, 0, 2, 3, 3, 0, 1) and (2, 1, 0, 0, 3, 2, 1, 3, 1). Find the resemblance in cosine.

7.

i. What is the formula for calculating Hamming distance? Between 10001011 and 11001111, calculate the Hamming gap.

Hamming Distance is the number of differing characters between two strings of equal lengths or the number of differing bits between two numbers.

Between 10001011 and 11001111, calculate the Hamming gap: **2**

ii. Compare the Jaccard index and similarity matching coefficient of two features with values (1, 1, 0, 0, 1, 0, 1, 1) and (1, 1, 0, 0, 0, 1, 1, 1), respectively (1, 0, 0, 1, 1, 0, 0, 1).

8. State what is meant by "high-dimensional data set"? Could you offer a few real-life examples? What are the difficulties in using machine learning techniques on a data set with many dimensions? What can be done about it?

High dimensional data are data characterized by few dozen to many thousands of dimensions (see the definition of high dimensional data in the CHDD 2012 international conference https://sites.google.com/site/chdd12naples/). Estimating the embedding dimension of a data set is a hard problem

In today's big data world it can also refer to several other potential issues that arise when your data has a huge number of dimensions: If we have more features than observations than we run the risk of massively overfitting our model — this would generally result in terrible out of sample performance.

When dealing with high dimensional data, it is often useful to reduce the dimensionality by projecting the data to a lower dimensional subspace which captures the “essence” of the data. This is called dimensionality reduction.

9. Make a few quick notes on:

PCA is an acronym for Personal Computer Analysis.

NO

PCA means Principal Component Analysis. It is an unsupervised algorithm technique to reduce the dimensions of data set

2. Use of vectors

Vectors are commonly used in machine learning as they lend a convenient way to organize data. Often one of the very first steps in making a machine learning model is vectorizing the data.

A support vector machine analyses vectors across an n-dimensional space to find the optimal hyperplane for a given data set.

3. Embedded technique

An embedding is a relatively low-dimensional space into which you can translate high-dimensional vectors. Embeddings make it easier to do machine learning on large inputs like sparse vectors representing words. An embedding can be learned and reused across models.

10. Make a comparison between:

1. Sequential backward exclusion vs. sequential forward selection

Sequential floating forward selection (SFFS) starts from the empty set. After each forward step, SFFS performs backward steps as long as the objective function increases. Sequential floating backward selection (SFBS) starts from the full set.

2. Function selection methods: filter vs. wrapper

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.
* Using the subset of features from the wrapper methods make the model more prone to overfitting as compared to using subset of features from the filter methods.

3. SMC vs. Jaccard coefficient