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

Ans: Feature engineering is the process of selecting and transforming variables (also known as “features”) in a dataset to improve the performance of a machine learning model. it involves creating new features from existing ones, selecting relevant features, and scaling or transforming features to make them more useful for a particular task.

Various aspects of feature engineering involves:

**1. Feature Creation:** Feature Creation is the process of generating new features based on domain knowledge or by observing patterns in the data.**Types of Feature Creation:**

1. **Domain-Specific:**Creating new features based on domain knowledge, such as creating features based on business rules or industry standards.
2. **Data-Driven:**Creating new features by observing patterns in the data, such as calculating aggregations or creating interaction features.
3. **Synthetic:**Generating new features by combining existing features or synthesizing new data points.

**2. Feature Transformation:**Feature Transformation is the process of transforming the features into a more suitable representation for the machine learning model. This is done to ensure that the model can effectively learn from the data.**Types of Feature Transformation:**

1. **Normalization:**Rescaling the features to have a similar range, such as between 0 and 1, to prevent some features from dominating others.
2. **Scaling:**Rescaling the features to have a similar scale, such as having a standard deviation of 1, to make sure the model considers all features equally.
3. **Encoding:**Transforming categorical features into a numerical representation. Examples are one-hot encoding and label encoding.
4. **Transformation:**Transforming the features using mathematical operations to change the distribution or scale of the features. Examples are logarithmic, square root, and reciprocal transformations.

**3. Feature Extraction** is the process of creating new features from existing ones to provide more relevant information to the machine learning model. This is done by transforming, combining, or aggregating existing features.**Types of Feature Extraction:**

1. **Dimensionality Reduction:** Reducing the number of features by transforming the data into a lower-dimensional space while retaining important information. Examples are PCA and t-SNE.
2. **Feature Combination:**Combining two or more existing features to create a new one. For example, the interaction between two features.
3. **Feature Aggregation:**Aggregating features to create a new one. For example, calculating the mean, sum, or count of a set of features.
4. **Feature Transformation:**Transforming existing features into a new representation. For example, log transformation of a feature with a skewed distribution.

**4. Feature Selection** is the process of selecting a subset of relevant features from the dataset to be used in a machine-learning model. It is an important step in the feature engineering process as it can have a significant impact on the model’s performance. **Types of Feature Selection:**

1. **Filter Method:**Based on the statistical measure of the relationship between the feature and the target variable. Features with a high correlation are selected.
2. **Wrapper Method:**Based on the evaluation of the feature subset using a specific machine learning algorithm. The feature subset that results in the best performance is selected.
3. **Embedded Method:**Based on the feature selection as part of the training process of the machine learning algorithm.

**5. Feature Scaling** is the process of transforming the features so that they have a similar scale. This is important in machine learning because the scale of the features can affect the performance of the model.**Types of Feature Scaling:**

1. **Min-Max Scaling:**Rescaling the features to a specific range, such as between 0 and 1, by subtracting the minimum value and dividing by the range.
2. **Standard Scaling:** Rescaling the features to have a mean of 0 and a standard deviation of 1 by subtracting the mean and dividing by the standard deviation.
3. **Robust Scaling:** Rescaling the features to be robust to outliers by dividing them by the interquartile range.
4. What is feature selection, and how does it work? What is the aim of it? What are the various methods of function selection?

Ans: Feature Selection is the process of selecting a subset of relevant features from the dataset to be used in a machine-learning model. It is an important step in the feature engineering process as it can have a significant impact on the model’s performance. This can include techniques like correlation analysis, mutual information, and stepwise regression.

**Aim of Feature Selection:**

1. To Reduce Overfitting: By using only the most relevant features, the model can generalize better to new data.
2. To Improve Model Performance: Selecting the right features can improve the accuracy, precision, and recall of the model.
3. To Decrease Computational Costs: A smaller number of features requires less computation and storage resources.
4. To Improve Interpretability: By reducing the number of features, it is easier to understand and interpret the results of the model.

**Types of Feature Selection:**

1. **Filter Method:**Based on the statistical measure of the relationship between the feature and the target variable. Features with a high correlation are selected.
2. **Wrapper Method:**Based on the evaluation of the feature subset using a specific machine learning algorithm. The feature subset that results in the best performance is selected.
3. **Embedded Method:**Based on the feature selection as part of the training process of the machine learning algorithm.

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

Ans: Selection Filter : These methods select features from the dataset irrespective of the use of any machine learning algorithm.Selection of feature is evaluated individually which can sometimes help when features are in isolation (don’t have a dependency on other features) but will lag when a combination of features can lead to increase in the overall performance of the model.



Filter Methods Implementation

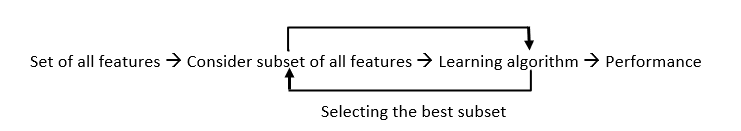
Some techniques used are:

* Information Gain – It is defined as the amount of information provided by the feature for identifying the target value and measures reduction in the entropy values. Information gain of each attribute is calculated considering the target values for feature selection.
* Chi-square test — Chi-square method (X2) is generally used to test the relationship between categorical variables. It compares the observed values from different attributes of the dataset to its expected value.
* Fisher’s Score – Fisher’s Score selects each feature independently according to their scores under Fisher criterion leading to a suboptimal set of features. The larger the Fisher’s score is, the better is the selected feature.
* Correlation Coefficient – Pearson’s Correlation Coefficient is a measure of quantifying the association between the two continuous variables and the direction of the relationship with its values ranging from -1 to 1.

**Pros and cons** :In terms of computation, they are very fast and inexpensive and are very good for removing duplicated, correlated, redundant features but these methods do not remove multicollinearity

Wrapper methods:

* Wrapper methods, also referred to as greedy algorithms train the algorithm by using a subset of features in an iterative manner. Based on the conclusions made from training in prior to the model, addition and removal of features takes place. Stopping criteria for selecting the best subset are usually pre-defined by the person training the model such as when the performance of the model decreases or a specific number of features has been achieved.



* Wrapper Methods Implementation
* Some techniques used are:
* Forward selection – This method is an iterative approach where we initially start with an empty set of features and keep adding a feature which best improves our model after each iteration. The stopping criterion is till the addition of a new variable does not improve the performance of the model.
* Backward elimination – This method is also an iterative approach where we initially start with all features and after each iteration, we remove the least significant feature. The stopping criterion is till no improvement in the performance of the model is observed after the feature is removed.
* Bi-directional elimination – This method uses both forward selection and backward elimination technique simultaneously to reach one unique solution.
* Exhaustive selection – This technique is considered as the brute force approach for the evaluation of feature subsets. It creates all possible subsets and builds a learning algorithm for each subset and selects the subset whose model’s performance is best.

Pros and cons:The main advantage of wrapper methods over the filter methods is that they provide an optimal set of features for training the model, thus resulting in better accuracy than the filter methods but are computationally more expensive.

4

* 1. Describe the overall feature selection process.

**Feature Selection** is the process of selecting a subset of relevant features from the dataset to be used in a machine-learning model. It is an important step in the feature engineering process as it can have a significant impact on the model’s performance. Types of Feature Selection:

Filter Method: Based on the statistical measure of the relationship between the feature and the target variable. Features with a high correlation are selected.Some techniques such as Information gain,Chi-square test,Fisherman’s score,correlation coefficient etc.are used

Wrapper Method: Based on the evaluation of the feature subset using a specific machine learning algorithm. The feature subset that results in the best performance is selected.some techniques like forward,backward ,bi-direction eliminination etc are used.

Embedded Method: Based on the feature selection as part of the training process of the machine learning algorithm.e.g regularisation,tree based are used.

Features Selection Algorithms are as follows:

1. Instance based approaches: There is no explicit procedure for feature subset generation. Many small data samples are sampled from the data. Features are weighted according to their roles in differentiating instances of different classes for a data sample. Features with higher weights can be selected.

2. Nondeterministic approaches: Genetic algorithms and simulated annealing are also used in feature selection.

3. Exhaustive complete approaches:Branch and Bound evaluates estimated accuracy and ABB checks an inconsistency measure that is monotonic. Both start with a full feature set until the preset bound cannot be maintained.

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

Feature Extraction is basically a process of dimensionality reduction where the raw data obtained is separated into related manageable groups. A distinctive feature of these large datasets is that they contain a large number of variables and additionally these variables require a lot of computing resources in order to process them. Hence Feature Extraction can be useful in this case in selecting particular variables and also combining some of the related variables which in a way would reduce the amount of data. The results obtained would be evaluated with the help of precision and recall measures.Algorithm:

One Hot Encoding  
Bag of Word (BOW)  
 n-grams  
Tf-Idf  
Custom features  
Word2Vec(Word Embedding)

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

Ans: **Feature selection** is the process of reducing the number of input variables when developing a predictive mode. As text data mostly have high dimensionality problems. To reduce the curse of high dimensionality, feature selection techniques are used. Text Preprocessing is the phase of preparing raw data to make it suitable for a machine learning model. [transform text into something a machine can understand](https://monkeylearn.com/blog/beginners-guide-text-vectorization/). This is often carried out using a [bag of words](https://machinelearningmastery.com/gentle-introduction-bag-words-model/), where a vector represents the frequency of a word within a predefined list of words.It includes text cleaning, stopwords removal, stemming and lemmatization.  The string is converted to lowercase and punctuation is removed before text gets tokenized. After this **Stemming**and **Lemmatization**  are applied to generate the root form of words. After this word count,character count,sentence count,average word length,average sentence length will be calculated .

Text classification is a technique, where extracted documents are classified into predefined classes. Text Classification technique can be broadly classified into two types: Supervised Document Classification and Unsupervised Classification.

The [**Naive Bayes**](https://monkeylearn.com/text-classification-naive-bayes/)**algorithm** is a probabilistic classifier that makes use of [Bayes' Theorem](https://en.wikipedia.org/wiki/Bayes%27_theorem) – a rule that uses probability to predict the tag of a text based on prior knowledge of conditions that might be related. It calculates the probability of each tag for a given text, and then predicts the tag with the highest probability. The [Naive Bayes](https://monkeylearn.com/text-classification-naive-bayes/)  uses probability to predict the tag of a text based on prior knowledge of conditions that might be related. It calculates the probability of each tag for a given text, and then predicts the tag with the highest probability.

### Support Vector Machines:[Support Vector Machines](https://monkeylearn.com/text-classification-support-vector-machines-svm/) (SVM) is a classification algorithm that performs at its best when handling a limited amount of data. It determines the best result between vectors that belong to a given group or category, as well as the vectors that don’t belong to the group. FE also includes Sentiment Analysis,Topic Analysis,Intent Detection

For text categorization some of the popular similarity measures are –

1. Euclidean Distance.
2. Manhattan Distance.
3. Jaccard Similarity.
4. Minkowski Distance.
5. Cosine Similarity.

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.

Ans: Cosine similarity is : Cos(x, y) = x . y / ||x|| \* ||y||.

The cosine similarity is beneficial because even if the two similar data objects are far apart by the Euclidean distance because of the size, they could still have a smaller angle between them. Smaller the angle, higher the similarity. When plotted on a multi-dimensional space, the cosine similarity captures the orientation (the angle) of the data objects and not the magnitude.

x = { 2, 3, 2, 0, 2, 3, 3, 0, 1}  
y = { 2, 1, 0, 0, 3, 2, 1, 3, 1}

The formula for calculating the cosine similarity is : Cos(x, y) = x . y / ||x|| \* ||y||

x . y = 2\*2 + 3\*1 + 2\*0 + 0\*0+2\*3+3\*2+3\*1+0\*3+1\*1 = 4+3+6+6+3+1=23

||x|| = √ (2)^2 + (3)^2 + (2)^2 + (0)^2+(2)^2+3^2+3^2+(0)^2+(1)^2 =6.324

||y|| = √ 2^2+ 1^2+ 0^2+ 0^2+ 3^2+ 2^2+ 1^2+ 3^2+ 1^2=5.385

∴ resemblance of Cos(x, y) = 23 / (6.324 \*5.385) = 0.6757

7.

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

Ans: In order to calculate the Hamming distance between two strings, and , we perform their XOR operation, (a⊕ b), and then count the total number of 1s in the resultant string.

Hamming Distance: 10001011 ⊕11001111=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).

Ans: A=(1, 1, 0, 0, 1, 0, 1, 1),B=(1, 1, 0, 0, 0, 1, 1, 1),C=(1, 0, 0, 1, 1, 0, 0, 1)

Jaccard Indexfor A and B=  |A intersect B|/  |A| + |B| — |A intersect B|

each are having 8 coe.,but 6 are common =6/ 8+8-6=0.6

Jaccard Index for B and C=3/8+8-3 =3/13=0.2307

Jaccard Index for A and C =5/8+8-5=5/11=0.4545

[**Simple Matching Coefficient**](https://www.vcalc.com/wiki/vcalc/simple-matching-coefficient) is a coefficient that indicates the degree of similarity of two communities based on the number of species that they have in common.  The formula for the Simple Matching Coefficient is: SMC = (M00 + M11) / (M00 + M11 + M10 + M01)

SMC (A&B) =5/8=0.625 ,SMC(B&C)=6/8=0.75 ,SMC(A&C)=3/8=0.375

The Jaccard Index and SMC for above coefficients are having variations

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?

Ans: **High dimensional data** refers to a dataset in which the number of features *p* is larger than the number of [observations](https://www.statology.org/observation-in-statistics/) *N*, often written as *p* >> *N*.

For example, a dataset that has *p* = 6 features and only *N* = 3 observations would be considered high dimensional data because the number of features is larger than the number of observations.For example High dimensional data is common in healthcare datasets where the number of features for a given individual can be massive (i.e. blood pressure, resting heart rate, immune system status, surgery history, height, weight, existing conditions, etc.).High dimensional data is common in healthcare datasets where the number of features for a given individual can be massive (i.e. blood pressure, resting heart rate, immune system status, surgery history, height, weight, existing conditions, etc.).When the number of features in a dataset exceeds the number of observations, we will never have a deterministic answer.

The difficulties in using machine learning techniques on a data set with many dimensions are:

* + Overfitting
  + Model Performance not good
  + Increase in Computational Costs

1.Choose to include fewer features**.**There are several ways to decide which features to drop from a dataset, including:

* Drop features with many missing values: If a given column in a dataset has a lot of missing values, you may be able to drop it completely without losing much information.
* Drop features with low variance: If a given column in a dataset has values that change very little, you may be able to drop it since it’s unlikely to offer as much useful information about a response variable compared to other features.
* Drop features with low correlation with the response variable: If a certain feature is not highly correlated with the response variable of interest, you can likely drop it from the dataset since it’s unlikely to be a useful feature in a model.

2.Use a regularization method.

Another way to handle high dimensional data without dropping features from the dataset is to use a regularization technique such as:[Principal Components Analysis](https://www.statology.org/principal-components-analysis-in-r/),[Principal Components Regression](https://www.statology.org/principal-components-regression/),[Ridge Regression](https://www.statology.org/ridge-regression/),[Lasso Regression](https://www.statology.org/lasso-regression/)

9. Make a few quick notes on:

1.PCA:PCA is an acronym for Personal Computer Analysis,is a statistical technique used to reduce the dimensionality of a large dataset. It is a commonly used method in machine learning, data science, and other fields that deal with large datasets.

PCA works by identifying patterns in the data and then creating new variables that capture as much of the variation in the data as possible. These new variables, known as principal components, are linear combinations of the original variables in the dataset.

The first principal component captures the most variation in the data, the second captures the second most, and so on. The number of principal components created is equal to the number of original variables in the dataset.

PCA can be used for a variety of purposes, including data visualization, feature selection, and data compression. In data visualization, PCA can be used to plot high-dimensional data in two or three dimensions, making it easier to interpret. In feature selection, PCA can be used to identify the most important variables in a dataset. In data compression, PCA can be used to reduce the size of a dataset without losing important information.It involves the following steps:

1. Construct the covariance matrix of the data.

 2. Compute the eigenvectors of this matrix.

3. Eigenvectors corresponding to the largest eigen values are used to reconstruct a large fraction of variance of the original data.

The data instances are projected onto a lower dimensional space where the new features best represent the entire data in the least squares sense.

It can be shown that the optimal approximation, in the least square error sense, of a d-dimensional random vector x2< d by a linear combination of independent vectors is obtained by projecting the vector x onto the eigenvectors e1 corresponding to the largest eigen values l1 of the covariance matrix (or the scatter matrix) of the data from which x is drawn.

The eigenvectors of the covariance matrix of the data are referred to as principal axes of the data, and the projection of the data instances on to these principal axes are called the principal components. Dimensionality reduction is then obtained by only retaining those axes (dimensions) that account for most of the variance, and discarding all others.

2. Use of vectors: Vectors are (sometimes huge) lists of numbers which represent a set of coordinates in some space.Support Vector Machine can be applied to any kind of vectors which encode any kind of data. SVM can be used for text classification for that texts have to be transformed into vectors.

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

The blue circles in the graph below are representations of training texts which talk about the *Pricing* of a SaaS Product and the red triangles are representations of training texts which do not talk about that. the algorithm has determined the decision boundary for the category you want to analyze, you only have to obtain the representations of all of the texts you would like to classify and check what side of the boundary those representations fall into.

3. Embedded technique

Embedded methods selects the important features while the model is being trained, You can say few model training algorithms already implements a feature selection process while getting trained with the data.

Some of embedded techniques are Lasso Regression , Ridge regression , decision tree.

Lasso regression is a L1 regularized regression when there is a penalty for more complicated coefficients.

### Decision Tree:During the construction of a decision tree the structure of the decision tree is such that the more important features are higher up , are closer to the root

10. Make a comparison between:

1. Sequential backward exclusion vs. sequential forward selection

* Forward selection – This method is an iterative approach where we initially start with an empty set of features and keep adding a feature which best improves our model after each iteration. The stopping criterion is till the addition of a new variable does not improve the performance of the model.
* Backward elimination – This method is also an iterative approach where we initially start with all features and after each iteration, we remove the least significant feature. The stopping criterion is till no improvement in the performance of the model is observed after the feature is removed.

2. Function selection methods: filter vs. wrapper

Filter method:These methods select features from the dataset irrespective of the use of any machine learning algorithm.Selection of feature is evaluated individually which can sometimes help when features are in isolation (don’t have a dependency on other features) but will lag when a combination of features can lead to increase in the overall performance of the model.

Wrapper methods, also referred to as greedy algorithms train the algorithm by using a subset of features in an iterative manner. Based on the conclusions made from training in prior to the model, addition and removal of features takes place. Stopping criteria for selecting the best subset are usually pre-defined by the person training the model such as when the performance of the model decreases or a specific number of features has been achieved

Filter method ,in terms of computation, they are very fast and inexpensive and are very good for removing duplicated, correlated, redundant features but these methods do not remove multicollinearity

The main advantage of wrapper methods over the filter methods is that they provide an optimal set of features for training the model, thus resulting in better accuracy than the filter methods but are computationally more expensive.

1. SMC vs. Jaccard coefficient

Jaccard similarity index (sometimes called the Jaccard similarity *coefficient*) compares members for two sets to see which members are shared and which are distinct. It’s a measure of similarity for the two sets of data, with a range from 0% to 100%. The higher the percentage, the more similar the two populations. Although it’s easy to interpret, it is extremely sensitive to small samples sizes and may give erroneous results, especially with very small samples or data sets with missing observations.The formula to find the Index is: **J(X,Y) = |X∩Y| / |X∪Y**

**SMC : simple matching coefficient :**

Each attribute of A and B can either be 0 or 1. The total number of each combination of attributes for both A and B are specified as follows:

M11=represents the total number of attributes where A and B both have a value of

 M01=represents the total number of attributes where the attribute of A is 0 and the attribute of B is 1.

M10= represents the total number of attributes where the attribute of A is 1 and the attribute of B is 0.

 M00=represents the total number of attributes where A and B both have a value of 0. Each attribute must fall into one of these four categories, meaning that

SMC=M11+M00/M11+M00+M01+M10

The main difference is that the SMC has the term M00 in its numerator and denominator, whereas the Jaccard index does not. Thus, the SMC counts both mutual presences (when an attribute is present in both sets) and mutual absence (when an attribute is absent in both sets) as matches and compares it to the total number of attributes in the universe, whereas the Jaccard index only counts mutual presence as matches and compares it to the number of attributes that have been chosen by at least one of the two sets.