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

engineering in depth.

Feature engineering is a machine learning technique that leverages data to create new variables that aren't in the training set. It can produce new features for both supervised and unsupervised learning, with the goal of simplifying and speeding up data transformations while also enhancing model accuracy.

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. These processes entail:

Feature Creation: Creating features involves identifying the variables that will be most useful in the predictive model. This is a subjective process that requires human intervention and creativity. Existing features are mixed via addition, subtraction, multiplication, and ratio to create new derived features that have greater predictive power.

Transformations: Transformation involves manipulating the predictor variables to improve model performance; e.g. ensuring the model is flexible in the variety of data it can ingest; ensuring variables are on the same scale, making the model easier to understand; improving accuracy; and avoiding computational errors by ensuring all features are within an acceptable range for the model.

Feature Extraction: Feature extraction is the automatic creation of new variables by extracting them from raw data. The purpose of this step is to automatically reduce the volume of data into a more manageable set for modeling. Some feature extraction methods include cluster analysis, text analytics, edge detection algorithms, and principal components analysis.

Feature Selection: Feature selection algorithms essentially analyze, judge, and rank various features to determine which features are irrelevant and should be removed, which features are redundant and should be removed, and which features are most useful for the model and should be prioritized.

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, one of the main components of feature engineering, is the process of selecting the most important features to input in machine learning algorithms. Feature selection techniques are employed to reduce the number of input variables by eliminating redundant or irrelevant features and narrowing down the set of features to those most relevant to the machine learning model.

The main benefits of performing feature selection in advance, rather than letting the machine learning model figure out which features are most important, include:

simpler models: simple models are easy to explain - a model that is too complex and unexplainable is not valuable

shorter training times: a more precise subset of features decreases the amount of time needed to train a model

variance reduction: increase the precision of the estimates that can be obtained for a given simulation avoid the curse of high dimensionality: dimensionally cursed phenomena states that, as dimensionality and the number of features increases, the volume of space increases so fast that the available data become limited - PCA feature selection may be used to reduce dimensionality

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

approach?

In wrapper methods, the feature selection process is based on a specific machine learning algorithm that we are trying to fit on a given dataset.

It follows a greedy search approach by evaluating all the possible combinations of features against the evaluation criterion. The evaluation criterion is simply the performance measure which depends on the type of problem, for e.g. For regression evaluation criterion can be p-values, R-squared, Adjusted R-squared, similarly for classification the evaluation criterion can be accuracy, precision, recall, f1-score, etc. Finally, it selects the combination of features that gives the optimal results for the specified machine learning algorithm.

1. Forward selection

In forward selection, we start with a null model and then start fitting the model with each individual feature one at a time and select the feature with the minimum p-value. Now fit a model with two features by trying combinations of the earlier selected feature with all other remaining features. Again select the feature with the minimum p-value. Now fit a model with three features by trying combinations of two previously selected features with other remaining features. Repeat this process until we have a set of selected features with a p-value of individual features less than the significance level.

2. Backward elimination

In backward elimination, we start with the full model (including all the independent variables) and then remove the insignificant feature with the highest p-value(> significance level). This process repeats again and again until we have the final set of significant features.

3. Bi-directional elimination(Step-wise Selection)

It is similar to forward selection but the difference is while adding a new feature it also checks the significance of already added features and if it finds any of the already selected features insignificant then it simply removes that particular feature through backward elimination.

4.

i. Describe the overall feature selection process.

Feature Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data.

It is the process of automatically choosing relevant features for your machine learning model based on the type of problem you are trying to solve. We do this by including or excluding important features without changing them. It helps in cutting down the noise in our data and reducing the size of our input data.

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 part of the dimensionality reduction process, in which, an initial set of the raw data is divided and reduced to more manageable groups. So when you want to process it will be easier. The most important characteristic of these large data sets is that they have a large number of variables. These variables require a lot of computing resources to process. So Feature extraction helps to get the best feature from those big data sets by selecting and combining variables into features, thus, effectively reducing the amount of data. These features are easy to process, but still able to describe the actual data set with accuracy and originality.

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

NLP is a subfield of artificial intelligence where we understand human interaction with machines using natural languages. To understand a natural language, you need to understand how we write a sentence, how we express our thoughts using different words, signs, special characters, etc basically we should understand the context of the sentence to interpret its meaning.

If we can use these contexts as features and feed them to our model then the model will be able to understand the sentence better. Some of the common features that we can extract from a sentence are the number of words, number of capital words, number of punctuation, number of unique words, number of stopwords, average sentence length, etc. We can define these features based on our data set we are using. In this blog, we will use a Twitter data set so we can add some others features like the number of hashtags, number of mentions, etc. We will discuss them in detail in the coming sections.

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.

Cosine similarity measures the similarity between two vectors of an inner product space. It is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction. It is often used to measure document similarity in text analysis.

94.30% - SIMILARITY SCORE

7.

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

calculate the Hamming gap.

Hamming distance is a metric for comparing two binary data strings. While comparing two binary strings of equal length

The hamming distance 10001011 and 11001111 is 3

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).

The 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.

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 defined as data in which the number of features (variables observed), p, are close to or larger than the number of observations (or data points). 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.)

Dimensionally cursed phenomena occur in domains such as numerical analysis, sampling, combinatorics, machine learning, data mining and databases. The common theme of these problems is that when the dimensionality increases, the volume of the space increases so fast that the available data become sparse.

9. Make a few quick notes on:

PCA is an acronym for Personal Computer Analysis.

Principal component analysis, or PCA, is a statistical procedure that allows you to summarize the information content in large data tables by means of a smaller set of “summary indices” that can be more easily visualized and analyzed. The underlying data can be measurements describing properties of production samples, chemical compounds or reactions, process time points of a continuous process, batches from a batch process, biological individuals or trials of a DOE-protocol, for example.

2. Use of vectors

Vectors can be used to represent physical quantities. Most commonly in physics, vectors are used to represent displacement, velocity, and acceleration. Vectors are a combination of magnitude and direction, and are drawn as arrows

3. Embedded technique

Embedded methods combine the qualities' of filter and wrapper methods. It's implemented by algorithms that 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

10. Make a comparison between:

1. Sequential backward exclusion vs. sequential forward selection

Forward selection starts with a (usually empty) set of variables and adds variables to it, until some stop- ping criterion is met. Similarly, backward selection starts with a (usually complete) set of variables and then excludes variables from that set, again, until some stopping criterion is met.

2. Function selection methods: filter vs. wrapper

Wrapper methods measure the “usefulness” of features based on the classifier performance. In contrast, the filter methods pick up the intrinsic properties of the features (i.e., the “relevance” of the features) measured via univariate statistics instead of cross-validation performance. So, wrapper methods are essentially solving the “real” problem (optimizing the classifier performance), but they are also computationally more expensive compared to filter methods due to the repeated learning steps and cross-validation. The third class, embedded methods, are quite similar to wrapper methods since they are also used to optimize the objective function or performance of a learning algorithm or model. The difference to wrapper methods is that an intrinsic model building metric is used during learning. Let me give you a – off the top of my head – list of examples from these three categories.

Filter methods:

* information gain
* chi-square test
* fisher score
* correlation coefficient
* variance threshold

Wrapper methods:

* recursive feature elimination
* sequential feature selection algorithms
* genetic algorithms

3. SMC vs. Jaccard coefficient

The SMC is very similar to the more popular Jaccard index. The main difference is that the SMC has the term M\_{00} 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.