**1) Explain the difference between Data Selection and Data Extraction?**

A) Feature extraction is a quite complex concept concerning the translation of raw data into the inputs that a particular Machine Learning algorithm requires. The model is the motor, but it needs fuel to work. Features must represent the information of the data in a format that will best fit the needs of the algorithm that is going to be used to solve the problem.

While some inherent features can be obtained directly from raw data, we usually need derived features from these inherent features that are actually relevant to attack the underlying problem. A poor model fed with meaningful features will surely perform better than an amazing algorithm fed with low-quality features – “garbage in, garbage out”.

Feature extraction fills this requirement: it builds valuable information from raw data – the features – by reformatting, combining, transforming primary features into new ones… until it yields a new set of data that can be consumed by the Machine Learning models to achieve their goals.

The response is pretty well defined regarding feature selection, which is enclosed:

Wrappers: a wrapper evaluates a specific model sequentially using different potential subsets of features to get the subset that best works in the end. They are highly costly and have a high chance of overfitting, but also a high chance of success, on the other hand. Learn more here.

Filters: for a much faster alternative, filters do not test any particular algorithm, but rank the original features according to their relationship with the problem (labels) and just select the top of them. Correlation and mutual information are the most widespread criteria. There are many easy to use tools, like the feature selection sklearn package.

Embedded: this group is made up of all the Machine Learning techniques that include feature selection during their training stage. LASSO is an example.

**2) What is Regression analysis?**

a) Regression analysis is a reliable method of identifying which variables have impact on a topic of interest. The process of performing a regression allows you to confidently determine which factors matter most, which factors can be ignored, and how these factors influence each other.

In order to understand regression analysis fully, it’s essential to comprehend the following terms:

Dependent Variable: This is the main factor that you’re trying to understand or predict.

Independent Variables: These are the factors that you hypothesize have an impact on your dependent variable.

In our application training example above, attendees’ satisfaction with the event is our dependent variable. The topics covered, length of sessions, food provided, and the cost of a ticket are our independent variables.

How does regression analysis work?

In order to conduct a regression analysis, you’ll need to define a dependent variable that you hypothesize is being influenced by one or several independent variables.

You’ll then need to establish a comprehensive dataset to work with. Administering surveys to your audiences of interest is a terrific way to establish this dataset. Your survey should include questions addressing all of the independent variables that you are interested in.

Let’s continue using our application training example. In this case, we’d want to measure the historical levels of satisfaction with the events from the past three years or so (or however long you deem statistically significant), as well as any information possible in regards to the independent variables.

Perhaps we’re particularly curious about how the price of a ticket to the event has impacted levels of satisfaction.

To begin investigating whether or not there is a relationship between these two variables, we would begin by plotting these data points on a chart, which would look like the following theoretical example.

Regression Analysis: Plotting data is the first step in figuring out if there is a relationship between independent and dependent variables

(Plotting your data is the first step in figuring out if there is a relationship between your independent and dependent variables)

Our dependent variable (in this case, the level of event satisfaction) should be plotted on the y-axis, while our independent variable (the price of the event ticket) should be plotted on the x-axis.

Once your data is plotted, you may begin to see correlations. If the theoretical chart above did indeed represent the impact of ticket prices on event satisfaction, then we’d be able to confidently say that the higher the ticket price, the higher the levels of event satisfaction.

**3) What is Pearson correlation coefficient?**

a) The Pearson correlation coefficient (r) is the most widely used correlation coefficient and is known by many names:

Pearson’s r

Bivariate correlation

Pearson product-moment correlation coefficient (PPMCC)

The correlation coefficient

The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

Although interpretations of the relationship strength (also known as effect size) vary between disciplines, the table below gives general rules of thumb:

Pearson correlation coefficient (r) value Strength Direction

Greater than .5 Strong Positive

Between .3 and .5 Moderate Positive

Between 0 and .3 Weak Positive

0 None None

Between 0 and –.3 Weak Negative

Between –.3 and –.5 Moderate Negative

Less than –.5 Strong Negative

The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables.

Visualizing the Pearson correlation coefficient

Another way to think of the Pearson correlation coefficient (r) is as a measure of how close the observations are to a line of best fit.

The Pearson correlation coefficient also tells you whether the slope of the line of best fit is negative or positive. When the slope is negative, r is negative. When the slope is positive, r is positive.

**4) Describe Factor Analysis in detail.**

a) Factor analysis is a technique that is used to reduce a large number of variables into fewer numbers of factors. This technique extracts maximum common variance from all variables and puts them into a common score. As an index of all variables, we can use this score for further analysis. Factor analysis is part of general linear model (GLM) and this method also assumes several assumptions: there is linear relationship, there is no multicollinearity, it includes relevant variables into analysis, and there is true correlation between variables and factors. Several methods are available, but principal component analysis is used most commonly.

Types of factoring:

There are different types of methods used to extract the factor from the data set:

1. Principal component analysis: This is the most common method used by researchers. PCA starts extracting the maximum variance and puts them into the first factor. After that, it removes that variance explained by the first factors and then starts extracting maximum variance for the second factor. This process goes to the last factor.

2. Common factor analysis: The second most preferred method by researchers, it extracts the common variance and puts them into factors. This method does not include the unique variance of all variables. This method is used in SEM.

3. Image factoring: This method is based on correlation matrix. OLS Regression method is used to predict the factor in image factoring.

4. Maximum likelihood method: This method also works on correlation metric but it uses maximum likelihood method to factor.

5. Other methods of factor analysis: Alfa factoring outweighs least squares. Weight square is another regression based method which is used for factoring.