1. What are the key tasks that machine learning entails? What does data pre-processing imply?

Machine learning entails the following key tasks:

* Data Gathering
* Data Pre-processing
* Model Selection
* Model Training
* Model Evaluation

Data Pre-processing is the stage where data is prepared before it is fed to the model for training. For building an optimal model, the data fed to it needs to be good. Raw data can contain many issues such as missing values, incorrect or invalid values, outliers, skewness etc. It is crucial that the data fed to the algorithm is free from such issues as much as possible, if not, the predictions made by it will not be as accurate as they could have been. The bad data will negatively affect the model performance.

2. Describe quantitative and qualitative data in depth. Make a distinction between the two.

Quantitative data refers to the data that can be represented in numerical values and arithmetic operation on such data can be performed and give meaningful results. These can be data like age of a person, price of property, number of people in a country etc

Qualitative data, as the name suggests, describes the quality of something that cannot be represented numerical values where arithmetic operations will present a meaningful outcome. These can be values like colour of the car, educational degree of a person, true or false etc.

4. What are the various causes of machine learning data issues? What are the ramifications?

The raw data collected for creating a ML model can have various issues with it that can cause suboptimal performance or issues with the creation of model itself.

Here are some of the problems that one may face with the data used:

* Data may contain missing values in random observations. In instances such as sensor data collection, a sensor may malfunction and stop recording the data. In such cases the data will contain missing values
* NaN or invalid values are when the data may contain non numeric values where it is not supposed to or values that do not make sense in observation. For example, a prices column containing text values or weight of an object.
* Noise in the data is also a problem that can lead to creation of bad models. Such data can be a result of sensor malfunctions, programming errors or bad user input.
* Mislabelled data, this can lead to incorrect decisions in feature selection.
* Skewness, i.e. data may not follow a normal distribution but has a rather skewed distribution

All of these issues cause the data to be unusable or may lead to formation of a model that yields suboptimal performance. Various techniques are employed during the pre-processing stage to eliminate or reduce these issues for creating an optimally performing model.

5. Demonstrate various approaches to categorical data exploration with appropriate examples.

Categorical data has values limited and fixed number of possible values. Examples of such data include Gender, grades, blood groups, vehicle model etc. For such data, we cannot calculate meaningful mean or median as the values are distinct and do not represent any clear order.

For analysing such data, we need to use methods suited to it. This can include using concepts like:

* Mode: This is simply the value occurring the most frequently in the dataset. For example in the dataset ["Car", "Bat", "Bat", "Car", "Bat", "Bat", "Bat", "Bike"] the mode would be “Bat” since it occurs most frequently.
* Expected value: For ML this is the probability of the occurrence of the value
* Using graphs: Visual representation of categorical data can help tremendously in its analysis. We can easily visualize the type of data like determining if it is multimodal or not using bar graphs, how the data is distributed using pie charts, determining the presence of outliers using boxplots etc.

6. How would the learning activity be affected if certain variables have missing values? Having said that, what can be done about it?

And

7. Describe the various methods for dealing with missing data values in depth.

If the input data contains missing values, it can result in the model building process outright throwing an error, hence prevent in the creation of the model.

For handling missing values, various different approaches can be taken:

* Removing the observations with missing values. This approach can be useful if we have enough values. However, this will reduce the amount of training data which can lead to a less robust model. Also, the model may miss out on important insights present in the rest of the columns.
* Filling with zeros. This approach is a quick way to handle the missing values without any complex calculations. This will allow the model to be created but since the zeros were not present in the original data and are likely not close to the original values, it may result in wrong or less accurate predictions
* Filling with mean, median or mode. This approach is slightly better as these values may better represent the values that should have been. This can result in better model than the previous approach but is still not optimal
* Approximating values using KNN Imputer. This approach is much better as it will be able to consider the surrounding values and make a more informed choice about the value being filled in place of the missing values. This is much better than of the previous approaches as it is not replacing with a singular value but is rather using the dataset itself to approximate the original values

8. What are the various data pre-processing techniques? Explain dimensionality reduction and function selection in a few words.

Data pre-processing involves several steps that require various techniques for dealing with different issues:

1. Data cleaning:

* Missing values are dealt by either removing the row or columns with me missing values or filling them with zeros, mean, median, mode or more complex methods like KNN imputation
* Nan or invalid values are dealt in a similar fashion to missing values
* Noisy data can lead to creation of suboptimal models and hence needs to be eliminated

1. Data transformation: Raw data can contain useless features, huge variations in scaling of the different features or the nature of data may be skewed. For resolving this, techniques like Feature selection, normalization, standardization etc are used to prepare it for creation of an optimal model
2. Data reduction: Data may contain many attributes, all of which do not contribute equally to the predictions of the output. Hence, for creating a simpler and more generalized model, we can remove certain features that either do not or have very low impact on the final outcome. Another way to simplify and eliminate the effects of multicollinearity is Dimensionality reduction which can take several feature and combine them into lesser number of features for the model.

Feature selection is simply determining what features are necessary for the creation of the model. We determine using statistical analysis about the impact of the feature in prediction of the final output. This can be done using the ols analysis that can give us an idea about what features are worth keeping and which can be eliminated. Apart from this, the domain expert can also guide us in determining this.

Dimensionality Reduction is a technique that combines various features of the dataset and combine the, into lesser number of features by creating a relationship between the various feature columns. This is preferred in cases where there is a lot of data and some of the attributes may be correlated. While this certainly helps in eliminating the negative effects of multicollinearity and creation of a simpler model, it can also result in loss of information from the original data as the number of features get reduced. Hence, it is essential that factors like VIF are considered to determine if the operation is necessary and if it is worth using the technique.

9. i. What is the IQR? What criteria are used to assess it?

IQR is a measure of statistical dispersion, the spread of the data. It represents the middle 50% of the data. It is calculated as the difference of 75th and 25th percentiles of the data.

For calculation of IQR, the data is divided into quartiles or four rink ordered even parts via linear interpolation the quartiles are denoted as Q1 - the lower quartile(25th percentile), Q2 – the median, and Q3 – the upper quartile(75th percentile). The IQR is then described as the difference of Q3 and Q1, representing the middle 50% of the data distribution

This measure can be used to calculate a BOXPLOT, a simple graphical representation of the data distribution, determining outliers, data cleaning etc

ii. Describe the various components of a box plot in detail? When will the lower whisker surpass the upper whisker in length? How can box plots be used to identify outliers?

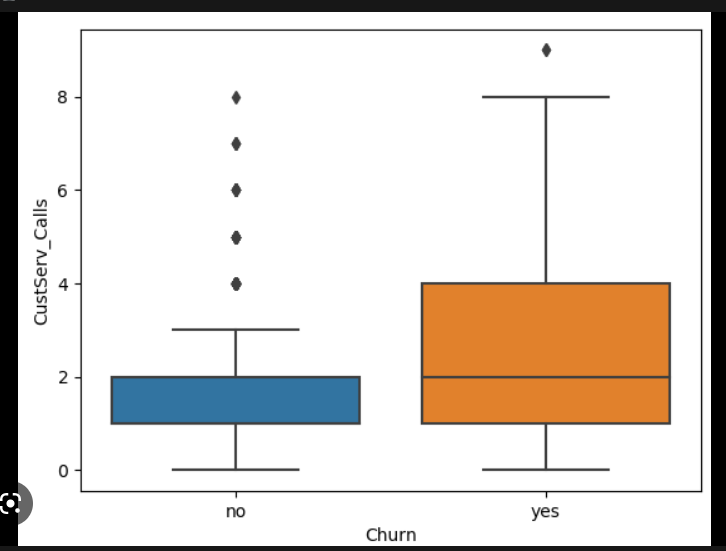
A Boxplot is a graphical representation of the data that depicts the distribution of values of a continuous distribution of data. It can be calculated based on the quartile values of the distribution and the IQR.

A boxplot displays a five summary of the distribution, depicting the minimum, maximum, median, 25th and 75th percentile values of the data. The box represents the data distribution between the first and the third quartile with the line within the box being the median value. The lines (or whiskers) emerging from either side of the box go from each of the quartiles to the maximum and minimum.

In case of a lower whisker being bigger in length than the upper whisker, we can summarize that the data distribution is right skewed depicting that most values lie near to the maximum of the data.

Outliers are depicted as dots outside of the minimum and maximum value lines of the plot.

The below image can be used to visualize the depiction of the Q1, Q2 and Q3 values, the maximum and minimum, outliers and the skewness of the data distribution



10. Make brief notes on any two of the following:

i. The gap between the quartiles:

For a data distribution, quartiles represent the 25th, 50th and 75th percentile values of the distribution. The gap between the quartiles, as can be seen in a boxplot is the space where 50% values of that distribution lie. This is beneficial in gaining several insights about the data like skewness, median etc.

ii. Use a cross-tab

Cross-tabs are multidimensional comparisons between multiple variables. It can be used to gain information about the relationships between the variables. Cross tabs can be useful in understanding categorical data. For example, a customer reviews data with different groups separated by region. A cross tab can be used to create a quick way to gain information like how many people from a particular region gave a positive, negative or an average review for the service