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

Data ingestion, EDA, pre-processing and preparing data, feature selection, Train the models, model evaluation, model selection, hyperparameter tuning, performance calculation, monitoring the model, model retraining

Pre-processing involve cleaning data, filling missing values, solving outlier problems if needed, transforming data if it is needed, removing duplicate data

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

Quantitative data is numbers-based, countable, or measurable. It tells us how many, how much, or how often in calculations while qualitative data is interpretation-based, descriptive, and relating to language. It can help us to understand why, how, or what happened behind certain behaviours.

3. Create a basic data collection that includes some sample records. Have at least one attribute from each of the machine learning data types.

Data= pd.DataFrame({‘age’: [54, 23, 30, 65, 42],

‘education’: [‘diploma’, ‘graduated’, ‘PHD’, ‘graduated’, ‘diploma’],

‘income’: [5000, 10000, 30000, 15000, 8000]})

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

Poor quality of data, overfitting, underfitting, lake of data, slow implementation, imperfection in algorithm when data increase, process complexity, irrelevant features

the ramification problem is concerned with the indirect consequences of an action. It might also be posed as how to represent what happens implicitly due to an action or how to control the secondary and tertiary effects of an action. It is strongly connected to, and is opposite the qualification side of, the frame problem

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

- Bar plot which shows number of samples in each category for example number of male and female in a survey

- pie chart to show portion of each category. For example, number of people in each education category

- print describe of them which include number of categories, what is the most frequent category,

- in bivariate presenting data of a column when categorize then based on another categorical column. For example, number of people in each education level per gender

- presenting histogram of numerical column per categories on another categorical field, for example histogram of age per gender

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

Although there are some algorithm such as Knn that can handle missing values but most of the algorithms fail during training if we have missing value in features, even in those algorithm it can bias the results of the machine learning models or reduce the accuracy of the model then handling them is vital for having good performance.

There are variety of methods regard to type of data, such as dropping rows with missing value if there are less number of them are less as compare to total data, drop columns with more than 50% missing, filling them with fix number. Mean or median of related column, using imputing technique such as knn to predict missing value based on the other columns data

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

There are three main approach:

* Dropping: dropping rows if number of missing are less as compare to whole data or columns if number of missing values are high even more than 50% of whole data.
* Imputation: it depend on data type and also problem, but some common methods are filling with mean, median r mode of column, fill with most frequent value or grouping data based on other column and fill missing of each group based on mean, median or most frequent on that group, using knn or other imputing methods
* Predictive method: using predictive models for filling missing, for this divide data into two part with and without missing and train model with data that does not have missing and then use model to predict missing value

8. What are the various data pre-processing techniques? Explain dimensionality reduction and

function selection in a few words.

Data Cleaning, Dimensionality Reduction, Feature Engineering, Sampling Data, Data Transformation, Imbalanced Data are some of the most common technique of pre-processing.

In real problems, there are many attributes which does not help during training to improve accuracy of model and sometimes if we don’t reduce them, it affects the model’s performance too. The main reasons of this process are less computational resources, increasing the overall performance of the model, preventing overfitting, avoiding multicollinearity and reducing the noise data.

Feature selection refers to the process of selecting the most important variables related to prediction variable, in other words, selecting the attributes which contribute most to your model.

9.

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

The interquartile range (IQR) measures the spread of the middle half of your data and shows central tendency of data. the formula is IQR = Q3 – Q1 , we need to calculate first quintile and third quantile which mostly imply to 25% and 75% and by calculating IQR we can find the range of 50% of data.

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?

Box plot shows these values:

Minimum value, Q1 which shows 25 quartile, average of data or Q2 or mean, 75 quartile or Q3 and maximum value. Which is a method to show distribution of data, the values less than Q1 and higher than Q3 presented as outliers in box plot.

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

1. Data collected at regular intervals

It is used mostly in collecting data for time series problems when data collected every fix period of time for example every 5 minutes or 1-hour data will collect.

2. The gap between the quartiles

It shows the distance between two quartiles, for example Q1 = 25 and Q3 = 75 then the gap between these two values shows the frequency of data between these two quartiles in also called as IQR or internal quartile range.

3. Use a cross-tab