1. What are the key tasks involved in getting ready to work with machine learning modelling?

Answer: Following are the key tasks involved in getting ready to work with machine learning modelling:

1. Data collection
2. EDA
3. Feature engineering and data transformation and Step b again
4. If step c is fine, we can go for ML modelling and based on performance metrics chosen for the use case we will try to improve performance metrics and even consider changing models suiting the use case.
5. Once we are done till step d, we consider deploying the model on cloud services though which everyone can interact it with the help of front end design.

2. What are the different forms of data used in machine learning? Give a specific example for each of them.

Answer: Most often the different types of data used in machine learning are text, time series, categorical data and numerical

1. Text – used for NLP tasks
2. Time series: Order time details for ecommerce websites
3. Categorical: only a list of allowed values in a within a feature
4. Numerical : any integer/float continuous value

3. Distinguish:

1. Numeric vs. categorical attributes

Answer:

1. Categorical data: Categorical data can be stored into groups or categories with respect to name or labels and has 2 data types, mainly nominal and ordinal data.
2. Numerical data: Numerical data is a type of data that can mostly be expressed in terms of numbers rather than natural language expressions

2. Feature selection vs. dimensionality reduction

Answer: Feature selection is simply selecting and excluding given features without changing them while dimensionality reduction transforms features into a lower dimension. Feature selection includes removing missing values, removing features with lesser variances, removing highly correlated features and recursive feature elimination.

Dimensionality reduction is transforming a feature into lesser dimensions such that we retain the max variance/information after conversion. Dimensionality reduction includes PCA and t-SNE.

4. Make quick notes on any two of the following:

1. The histogram

2. Use a scatter plot

3.PCA (Personal Computer Aid)

Answer:

1. Histogram:

Histogram is a graphical representation of distribution of numerical data. To construct this histogram, first we need to divide histograms into bins, i.e a range of values into intervals and then count how many values fall in each interval. Bins can be adjacent and not of equal size.

1. Scatter plot:

It helps in finding out the relation between 2 variables in a 2D Cartesian plot where we can also accommodate a 3rd variable if we decide to color the points based on 3rd variable/feature.

5. Why is it necessary to investigate data? Is there a discrepancy in how qualitative and quantitative data are explored?

Answer: It is very necessary to investigate data as whatever simplest model to complex model we design it needs to have a meaningful understanding of the pattern in the data and then it can make prediction accordingly. If data is garbage, then output from the ML model will be garbage too if we do not investigate the data properly.

There is difference in qualitative and quantitative data is explored as the 1st one is all about phenomenon that can be observed but not captured in numbers , whereas in the 2nd approach, it is all about quantities and numbers.

6. What are the various histogram shapes? What exactly are ‘bins'?

Answer: Different histogram shapes are bell shaped, left or right skewed, double peaked, one edge peaked and dog food distribution. Bins are a range of values into intervals and then count how many values fall in each interval. Bins can be adjacent and not of equal size.

7. How do we deal with data outliers?

Answer: We deal with outliers by dropping them if they are in small amounts, by using Boolean marking and by rescaling of features.

8. What are the various central inclination measures? Why does mean vary too much from median in certain data sets?

Answer: Different central tendency measures are mean, median and mode. Mean varies from median as mean is affected by outliers during it’s calculation while for median first we sort the values accordingly and there can be cases that mean and median will be separated at a distance.

9. Describe how a scatter plot can be used to investigate bivariate relationships. Is it possible to find outliers using a scatter plot?

Answer: Yes scatter plot can be used to investigate the bivariate relationships as we can take one of the feature as x axis values and 2nd feature as yaxis values and then, we plot the graph for every data point and see the direction, relation. If there are some points which are very far from normal points, we consider those points as outliers.

10. Describe how cross-tabs can be used to figure out how two variables are related.

Answer: We use cross tabs to check if the 2 features/variables are related or not at nominal or ordinal levels with a set of supporting statistics. We look at the distribution of 2 variables simultaneously.