Common Problems in Model and Solutions

1) **What are outliers?**
   Data points that are far away from the fitted regression line are called outliers.
2) **How to detect outliers?**
   - plotting scatter plot
   - plotting box plot
3) **How to deal the outliers?**
   - Remove outliers
   - Replace with mean, median
   - Normalize variables
4) **What is multi-collinearity & Variation inflation factor?**
   The dependent variable should have a strong relationship with independent variable; however, any independent variable should not have strong correlation among other independent variables.

   And multi-collinearity is an incident where one or more of the independent variables are strongly correlated with each other. In such incidents, we should use only one among correlated independent variable.

   VIF is an indicator of the existence of multi-collinearity. A value of greater than 10 is the rule of thumb for possible existence of high multi-collinearity. The standard guide line is as follows.
   VIF = 1 → No correlation existence.
   VIF > 1 & VIF < 5 → moderate correlation.
   $VIF = 1/(1-R^2)$

5) **What is Homoscedasticity?**
   The error should be constant which is known as homoscedasticity and error should be normally distributed.
6) **What is Heteroscedasticity?**
   Heteroscedasticity happens when the standard errors of variables are non-constant.
7) **How to detect Heteroscedasticity?**
   - using Durbin-watson Test
   - Breach Godffry test
   - NCV test
8) **How to fix Heteroscedasticity?**
   It can be fixed by using Box-Cox transformation of the variables
9) **What is Auto-Correlation?**
   Autocorrelation is a mathematical representation of the degree of similarity between a given time series and lagged version of itself over successive time intervals. It is the same as calculating the correlation between two different time series, except autocorrelation uses the same time series twice: once in its original form and once lagged one or more time periods.
10) **What is Durbin-Watson?**
    It is one of the common statistics used to determine the existence of the Autocorrelation. It is always between the number of 0 & 4. A value around 2 is ideal (range of 1.5 to 2.5 relatively normal.)

11) **How to fix the Auto-correlation issue?**
Autocorrelation issue can be fixed by applying PCA on variables.
12) **What is non-linearity?**
The relationship between the predictors and the outcome variables should be linear. If the relationship isn't linear that is called non-linearity.
13) **How to detect non-linearity problem?**
If straight line not fitting well, then we can conclude there is non-linearity issue in the data.
14) **How to fix non-linearity issue?**
It can be fixed by applying appropriate transformations to the independent variables such as log, square, root, higher order polynomials.
15) **What is Polynomials?**
It is a form of higher-order regression model between dependent and independent variables as an n^{th} degree polynomials. Although it is linear it can fit curves better.
16) **What is overfitting?**
Overfitting Occurs when the model fits the data too well, capturing all noises. In this case, you can notice a high accuracy on the training dataset, whereas the same model will result a low accuracy on the test dataset. This means the model has fitted the line so well to the train dataset and that it failed to generalize it to fit well on an unseen dataset.
17) **How to detect Overfitting?**
By segmenting the dataset, we can examine the performance of the model on each set of data to spot overfitting when it occurs, as well as see how the training process works.
18) **How model performance is measured?**
The performance can be measured using the percentage of accuracy observed in both data sets to conclude on the presence of overfitting. If the model performs better on the training set than on the test set, it means that the model is likely overfitting.
19) **What is under fitting?**
Under fitting Occurs when the model does not fit the data well and unable to capture the underlying trend in it. In this case, we can notice a low accuracy in training & testing dataset.
20) **What is Bias?**
If the model accuracy is low on a training dataset as well as the test dataset, the model is said to be under fitting or that the model has high bias. This means the model is not fitting the training dataset points well in regression or Decision boundary is not separating the classes well in classification; and two key reasons for bias are
• Not including the right features
• Not picking the correct orders of polynomial degrees for model fitting.
21) **How to fix the bias or under fitting issues?**
• Try to include more meaningful features.
• try to increase model complexity by trying higher order polynomial fittings.
22) **What is variance?**
If model is giving high accuracy on a training dataset, however on a test dataset the accuracy drops drastically. The key reasons for variance or over fitting
• Using higher order polynomial degree (it may not be required which will fit Decision Boundary too well to all data points including noise of train dataset, instead of the underlying relationship.

This will lead to a high accuracy (actual vs predicted) in the train dataset, and when applied to the test dataset, the prediction error will be high.

23) **How to solve overfitting or variance issue?**

• Try to reduce number of features that is keep only meaningful features.

• Try regularization methods that will keep all the features however, reduce the magnitude of the feature parameter.

• Dimensionality Reduction can eliminate noisy features, in turn reducing the model variables

• Brining more data points to make training dataset large will also reduce variance.

• Choosing right model parameters can help to reduce the bias and variance.

For example:

• using right regularization parameters can decrease variance in regression based models.

• For decision tree, reducing the depth of the Decision Tree will reduce the variance.

24) **What is Regularization?**

With increase of number of variables and increase in model complexity, the probability of over-fitting also increases. The Regularization is a technique to avoid overfitting with an increase in model complexity the size of the coefficients increase exponentially, so, the ridge & lasso regression apply penalty to the magnitude of the coefficients to handle the issue.

25) **What is LASSO Regression?**

This provides a sparse solution also known as (L1 Regularization) it guides parameter value to be zero that is the coefficient of the variables that add minor value to the model will be zero, and it adds a penalty equivalent to absolute value of the magnitude of coefficients.

26) **What is Ridge Regression?**

Ridge Regression also known as (L2 Regularization) it guides parameters to be close to zero, but not zero. You use this when you have many variables that add minor value to the model accuracy individually; however it improves overall the model accuracy and cannot be excluded from the model.

Ridge Regression apply a penalty to reduce the magnitude of the coefficient of all variables that add minor value to the model accuracy, and which adds penalty equivalent to square of the magnitude of coefficients.

Alpha is the regularization strength & must be positive float.

27) **What is K-Fold cross validation?**

K-fold cross validation splits the training dataset into k-folds without replacement, that is, any given data point will only be part of one of the subset, where k-1 folds are used for the model training and fold is used for testing. The procedure is repeated k-times so that we obtain k-models & performance estimates.

We then calculate the average performance of the models based on individual fold to obtain a performance estimate that is less sensitive to the sub partitioning of the training data compared to the holdout or single fold method.

28) **What is stratified k-fold cross validation?**

An extended cross-validation is the stratified k-fold cross-validation where the class proportions are preserved in each fold, leading to better bias and variance estimate.

29) **What is Selection Bias?**

Selection Bias is a result of the sample group not representing the entire target population.

30) **What is Eigen Vector?**

A vector that under goes pure scaling without any relation is known as the Eigen vector.

31) **What is Eigen Value?**

The scaling factor (stretch ratio) is known as the Eigen value.

32) **What is the Dimensionality Reduction?**

Dimensionality Reduction is the process of reducing the number of random variables under consideration by obtaining a set of principle.