**LAB 02**

**Harshitha S A**

**21BDA56**

**2. Answer the following questions:**

**i. What are the assumptions of linear regression?**

There are four assumptions associated with a linear regression model:

Linearity: The relationship between X and the mean of Y is linear.

Homoscedasticity: The variance of residual is the same for any value of X.

Independence: Observations are independent of each other.

**Normality:**The residuals of the model are normally distributed.

**ii. How can we evaluate a Regression model? Define each metric and its interpretation.**

There are 3 main metrics for model evaluation in regression:

1. R Square/Adjusted R Square.

2. Mean Square Error(MSE)/Root Mean Square Error(RMSE)

3. Mean Absolute Error(MAE)

Matric used to evaluate a regression model:

1. Metrics like correlation coefficient and coefficient of determination. Root mean square error (RMSE) is a common metric for comparisons
2. goodness of fit plots which include Pred vs. Observed, Pred vs Residuals (which include plots like residuals and weighted residuals). In many instances you may also want to look at Pred vs Independent Variable as well as residuals vs. independent variable.
3. When comparing models you may look for all of the above plus use statistical metrics and tests for selection of a better fit. These depend on whether the two models that are being compared are nested or non-nested. For nested models, comparison can be made with a chi-squared test. We can compare two models (nested or non-nested) using goodness of fit criteria like the Akaike information criteria (AIC) or the Bayesian information criterion (BIC). In general for non-nested models, the lower the AIC value the better is the fit to the data.
4. Other diagnostic plots are also used to evaluate the underlying distribution assumptions (assumptions of normality) which include plots like Q-Q plots.

**iii. Can R squared be negative?**

 it is possible to get a negative R-square for equations that do not contain a constant term. Because R-square is defined as the proportion of variance explained by the fit, if the fit is actually worse than just fitting a horizontal line then R-square is negative.

**iv. What is dummy variable trap?**

The Dummy variable trap is a scenario where there are attributes that are highly correlated (Multicollinear) and one variable predicts the value of others. When we use one-hot encoding for handling the categorical data, then one dummy variable (attribute) can be predicted with the help of other dummy variables.

**v. Is One Hot Encoding different from Dummy Variables?**

Yes, One Hot Encoding is different from Dummy Variables.

**vi. How is polynomial regression different from linear regression?**

Polynomial regression is a form of Linear regression where only due to the Non-linear relationship between dependent and independent variables we add some polynomial terms to linear regression to convert it into Polynomial regression.

**vii. Interpret the screenshot below from the notebook we discussed in class today:**

The model.score (x\_test, y\_test) gives the accuracy of test set and model.score ( x\_train, y\_train) gives accuracy of training set.