Regression Techniques

Regression is a method of modelling a target value based on independent predictors. This method is mostly used for forecasting and finding out cause and effect relationship between variables.

Regression Types :

- Linear Regression
- Logistic Regression
- Polynomial Regression
- Stepwise Regression
- Ridge Regression
- Lasso Regression
- ElasticNet Regression

Linear Regression :

- Linear Regression establishes arelationship between dependent variable (Y) and one or more independent variables (X) using a best fit straight line.
- Equation :- Y = a + bX
- The best fit straightline can obtained by Least Square method - $\min_w ||Xw - y||_2^2$
- Application :-Financial portfolio prediction, salary forecasting, real estate predictions and in traffic in arriving at ETAs.
- Advantages :-
  - Linear Regression performs well when the dataset is linearly separable.
  - Linear Regression is easier to implement, interpret and very efficient to train.
- Disadvatages :-
  - Linear Regression Is Limited to Linear Relationships
  - Linear Regression Only Looks at the Mean of the Dependent Variable
  - Linear Regression Is Sensitive to Outliers
  - More prone to overfitting
  - Data Must Be Independent

Logistic Regression :-

- Logistic regression is used to find the probability of event=Success and event=Failure. We should use logistic regression when the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature.

- Logistic regression is widely used for classification problems.
- Application - Today enterprises deploy Logistic Regression to predict house values in real estate business, customer lifetime value in the insurance sector and are leveraged to produce a continuous outcome such as whether a customer can buy/will buy scenario.
- Advantages :-
  - Doesnt require linear relationship between dependent and independent variables.
  - Logistic Regression not only gives a measure of how relevant a predictor (coefficient size) is, but also its direction of association (positive or negative).
  - Logistic regression is less prone to over-fitting but it can overfit in high dimensional datasets.
- Disadvantages :-
  - Main limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables. In the real world, the data is rarely linearly separable. Most of the time data would be a jumbled mess.
  - If the number of observations are lesser than the number of features, Logistic Regression should not be used, otherwise it may lead to overfit.
  - Logistic Regression can only be used to predict discrete functions.

Polynomial Regression :-

- Is used for curvilinear data.

  $Y = a\_1*X\_1 + (a\_2)^2*X\_2 + (a\_3)^4*X\_3 \ldots\ldots a\_n*X\_n + b$
- Suitable for handling non-linearly separable data
- For a polynomial regression, the power of some independent variables is more than 1.
- Advantages :-
  - Able to model non-linearly separable data; linear regression can’t do this.
  - It is much more flexible in general and can model some fairly complex relationships.
  - Full control over the modelling of feature variables .
- Disadvantages :-
  - The presence of one or two outliers in the data can seriously affect the results of the nonlinear analysis.
  - These are too sensitive to the outliers.
  - In addition, there are unfortunately fewer model validation tools for the detection of outliers in nonlinear regression than there are for linear regression.

Stepwise Regression :-

- Stepwise regression is used for fitting regression models with predictive models.
- It is carried out automatically. With each step, the variable is added or subtracted from the set of explanatory variables.
- The approaches for stepwise regression are forward selection, backward elimination, and bidirectional elimination.
- Advantages :-
  - The ability to manage large amounts of potential predictor variables, fine-tuning the model to choose the best predictor variables from the available options.
  - It's faster than other automatic model-selection methods.
  - Watching the order in which variables are removed or added can provide valuable information about the quality of the predictor variables.
- Disadvantages :-
  - Stepwise regression often has many potential predictor variables but too little data to estimate coefficients meaningfully. Adding more data does not help much, if at all.
  - Collinearity is usually a major issue. Excessive collinearity may cause the program to dump predictor variables into the model.

Ridge Regression :-

- Ridge Regression is a technique used when the data suffers from multicollinearity (independent variables are highly correlated).
- In multicollinearity, even though the least squares estimates (OLS) are unbiased, their variances are large which deviates the observed value far from the true value.By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors.
- It shrinks the value of coefficients but doesn't reaches zero, which suggests no feature selection feature
- Advantages :-
  - Least squares regression doesn't differentiate "important" from "less-important" predictors in a model, so it includes all of them. This leads to overfitting a model and failure to find unique solutions. Ridge regression avoids these problems.
  - Ridge regression adds just enough bias to make the estimates resonably reliable approximation to true population values.
  - Ridge regression performs well, compared to the ordinary least square method in a situation where you have a large multivariate data with the number of predictors (p) larger than the number of observations (n).

- The ridge estimator is especially good at improving the least-squares estimate when multicollinearity is present.

- Disadvantages :-
  - Firstly ridge regression includes all the predictors in the final model, unlike the stepwise regression methods which will generally select models that involve a reduced set of variables.
  - A ridge model does not perform feature selection. If a greater interpretation is necessary where we need to reduce the signal in our data to a smaller subset then a lasso model may be preferable.
  - Ridge regression shrinks the coefficients towards zero, but it will not set any of them exactly to zero. The lasso regression is an alternative that overcomes this drawback.

LASSO Regression :-

- LASSO stands for Least Absolute Selection Shrinkage Operator wherein shrinkage is defined as a constraint on parameters.
- Is a regression analysis method that performs both variable selection and regularization.
- Lasso regression uses soft thresholding.
- Lasso regression selects only a subset of the provided covariates for use in the final model.
- Applications : - Lasso-type regressions are also used to perform stress test platforms to analyze multiple stress scenarios. Lasso regression algorithms have been widely used in financial networks and economics.
- Advantages :-
  - As any regularization method, it can avoid overfitting. It can be applied even when number of features is larger than number of data.
  - It can do feature selection.
  - It is fast in terms of inference and fitting.
- Diadvantages :-
  - The model selected by lasso is not stable. For example, on different bootstrapped data, the feature selected can be very different.
  - The model selection result is not intuitive to interpret.
  - When there are highly correlated features, lasso may randomly select one of them of part of them. The result depends on the implementation. To overcome this drawback the elastic net was intoduced.

Elastic Net Regression :-

- ElasticNet regression is a regularized regression method that linearly combines the penalties of the lasso and ridge methods.
- ElasticNet regression is used for support vector machines, metric learning, and portfolio optimization.
- This is preferred over ridge or lasso regression.