**Regression model**:

We are going to start our modeling with linear regression because of its wide usability, fast run time, easy to use and high interpretability. It is basis for many other methods.

It is a type of predictive modeling that is used to find the relationship between dependent and independent variables. Regression is widely used for analyzing data by looking at the fit of a curve/line. The fit of the curve is a line connecting to the data points in such a way that reduces the distances between the data points from the fitting line.

In the regression analysis, we can predict the value of an unknown variable by looking at its relationship with the known variable. In the linear regression method, the dependent variable is continuous, and the independent variable can be continuous or discrete and regression lines come linear.

It is represented by an equation Y=a+b\*X + c,

where a is the intercept,

b is the slope of the line and

c is the error term.

A regression line can be obtained by Least Square Method. Its calculation is based on finding the best-fit line of observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line. between positive and negative values. The regression model can be evaluated by using the metric R-square.

#Model Validation

**##Lasso method: l1 regularization (**[**link**](https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b)**)**

We will be starting model validation with Lasso. Lasso regression are some of the simple techniques to reduce model complexity and prevent over-fitting which may result from simple linear regression. So, Lasso regression not only helps in reducing over-fitting, but it can help us in feature selection.

Lasso is standing for Least Absolute Shrinkage and Selection Operator. It can penalize the absolute size of the regression coefficients and reduce the variability which can improve the accuracy of linear regression models.

The larger penalty can further shrink the estimates towards zero which is important for the variable selections. If variable groups are highly correlated, lasso shrinks the others to zero and picks only one from them.

**##XGBoost Machine Learning**

The main benefits of using XGBoost are high execution speed and model performance. In both classification and regression predictive modelling, XGBoost dominates structured or tabular datasets.

XG boost (Extreme Gradient Boosting) is widely used for classification and regression problems and gives better performance than other algorithms. It is the execution of gradient boosted decision trees. It is good for the small to medium tabular or structured data.

It is used for supervised machine learning algorithms. It handles overfitting by using techniques of regularization. It is enabled with the inbuilt Cross-Validation (CV) function. It can handle the missing values by finding the trends and catching them. It has the power to save the data matrix and reload.

XGboost carries out the gradient boosting decision tree algorithm. Boosting is an ensemble technique that enables the model to make a prediction based on resolving the errors in the new model that has come from the old model. Model performance can be improved by tuning parameters.

These default metrics for the classification type of problem is an error and for regression metric is RMSE.

**## Linear Regression Vs XGBoost**

Linear regression is a parametric model: it assumes the target variable can be expressed as a linear combination of the independent variables (plus error). Gradient boosted trees are nonparametric: they will approximate any function.

xgboost deprecated the objective Reg; Linear. It has been replaced by reg: squared error, and has always meant minimizing the squared error, just as in linear regression.

So, boost will generally fit training data much better than linear regression, but that also means it is prone to overfitting, and it is less easily interpreted. Either one may end up being better, depending on your data and your needs. (link)

We have started from Regression model and later moved from lasso to XGboost Classifier.