Algorithms Used:

1. Decision Tree:

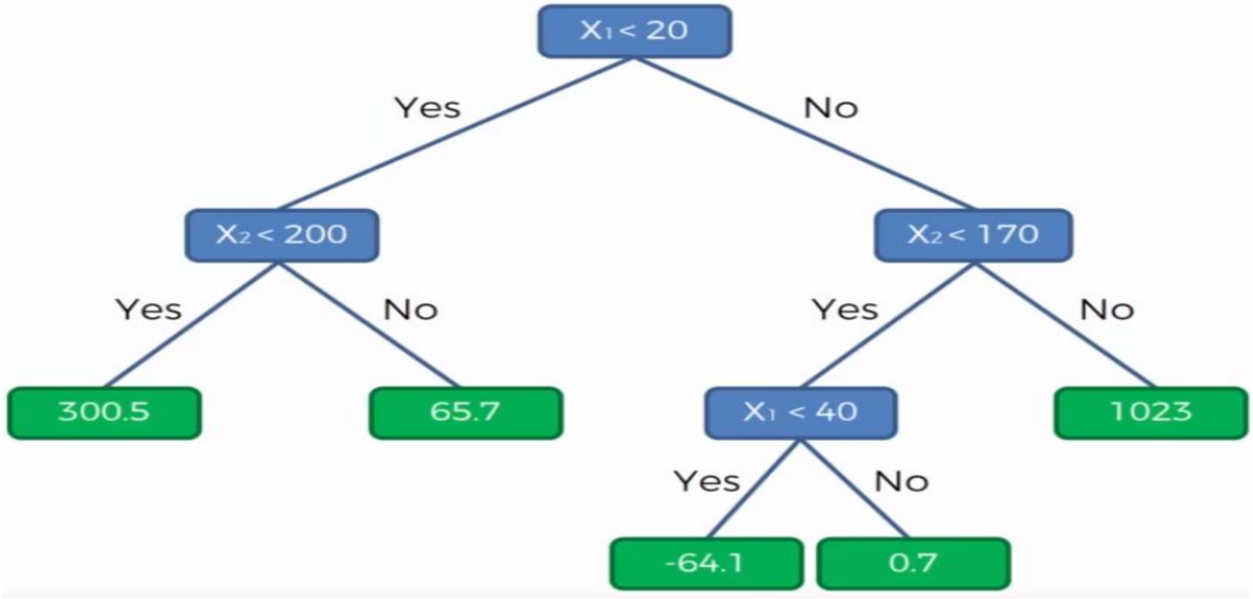
Dependent variable(Y) : Categorical
Independent variable(X): Categorical and Continous
Used for classification and prediction.
Supervised Learning.

Decision Tree is a classification technique used to classify records in a pictorial format.
In Decision Trees, for predicting a class label for a record we start from the **root** of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

Decision trees use multiple algorithms to decide to split a node into two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. In other words, we can say that the purity of the node increases with respect to the target variable.

Attribute Selection Measures:
Gini Index, Entropy
Gini index: a cost function used to evaluate splits in the dataset. It is calculated by subtracting the sum of the squared probabilities of each class from one.
(Relative Probability of class j at node t: summation over j p(j/t))

$$Gini=1-\sum_{i=1}^{C}(p_i)^2$$



Feature selection using Random forest
OOB Out of Bag Score: This metric is the accuracy of examples xi using all the trees in the random forest ensemble for which it was omitted during training.
Use of forests of trees to evaluate the importance of features on an artificial classification task.

Xgboost:
It works on gradient boosting algorithm
Gradient boosting algorithm works on the basic principle gradient descent.
This model is built using tree-based learners(Decision Trees)
Boosting works on the principle of ensemble techniques where errors from earlier models are reduced by the new models
XGradient boosting Algorithm:

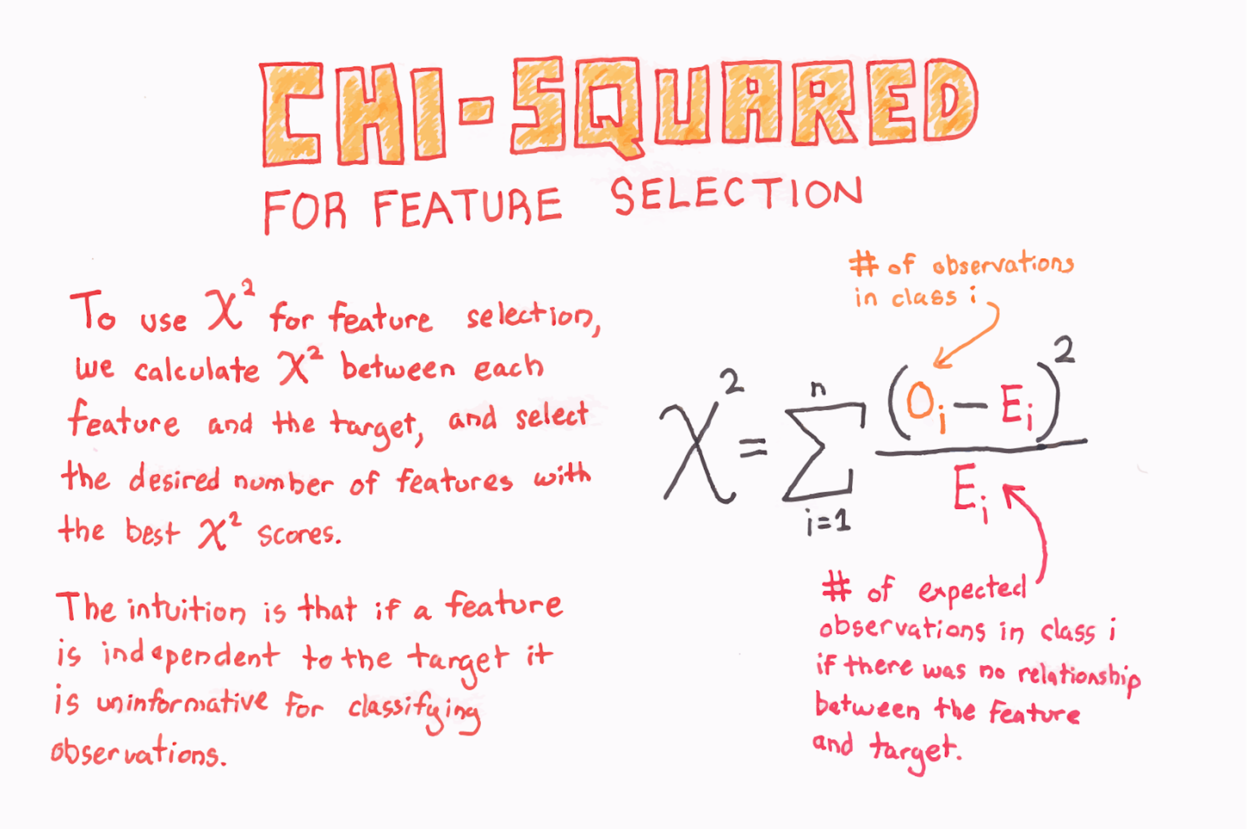
Final prediction=Base value(the starting prediction from basic decision tree)+LR*w1+LR*w2+..+LR*wn
Where LR= learning rate=eta
w1=residual predicted value by 1st residual model
wn=residual predicted value by nth residual model
Xgboost is different from other gradient boost is because of its tuning parameters
The main tuning parameters are
1)regularisation parameter(Lambda)

2)threshold that defines auto pruning (Gamma)
3)Learning rate(eta)

We are predicted based on similarity score, gain of branch
similarity score=(sum of residuals)^2/(Lambda+no of residuals)
Gain=(total SS)afternode split-(SS)before node split
If Gain>Gamma then only further splitting occurs else it stops and gives predictions(Wi)
After calculating Wi from the residual model

Feature Selection Techniques USed :

Chi-Squared Test



CHI-SQUARED

FOR FEATURE SELECTION

To use χ^2 for feature selection, we calculate χ^2 between each feature and the target, and select the desired number of features with the best χ^2 scores.

The intuition is that if a feature is independent to the target it is uninformative for classifying observations.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

# of observations in class i
# of expected observations in class i if there was no relationship between the feature and target.

- Quantifies the independence of pairs of categorical variables.
- The test calculates a statistic that has a chi-squared distribution, named for the Greek capital letter Chi (X) pronounced “ki” as in kite.
- Null hypothesis: that the observed frequencies for a categorical variable match the expected frequencies for the categorical variable.

- The result of the test is a test statistic that has a chi-squared distribution and can be interpreted to reject or fail to reject the assumption or null hypothesis that the observed and expected frequencies are the same.

Mutual Information Feature Selection

- Mutual information is calculated between two variables and measures the reduction in uncertainty for one variable given a known value of the other variable.
- Entropy is a measure of the randomness in the information being processed. The higher the entropy, the harder it is to draw any conclusions from that information.
- It is commonly used in the construction of decision trees from a training dataset, by evaluating the information gain for each variable, and selecting the variable that maximizes the information gain, which in turn minimizes the entropy and best splits the dataset into groups for effective classification.

- Information gain can also be used for feature selection, by evaluating the gain of each variable in the context of the target variable. In this slightly different usage, the calculation is referred to as mutual information between the two random variables.
- Information Gain, or IG for short, measures the reduction in entropy or surprise by splitting a dataset according to a given value of a random variable.A larger information gain suggests a lower entropy group or groups of samples, and hence less surprise.
- We can think about the entropy of a dataset in terms of the probability distribution of observations in the dataset belonging to one class or another, e.g. two classes in the case of a binary classification dataset.

For example, in a binary classification problem (two classes), we can calculate the entropy of the data sample as follows:

- Entropy = -(p(0) * log(P(0)) + p(1) * log(P(1)))

Information gain can be calculated as follows:

- IG(S, a) = H(S) – H(S | a)

Where *IG(S, a)* is the information for the dataset *S* for the variable a for a random variable, *H(S)* is the entropy for the dataset before any change (described above) and *H(S | a)* is the conditional entropy for the dataset given the variable *a*.

The conditional entropy can be calculated by splitting the dataset into groups for each observed value of a and calculating the sum of the ratio of examples in each group out of the entire dataset multiplied by the entropy of each group.

- H(S | a) = sum v in a Sa(v)/S * H(Sa(v))

Where *Sa(v)/S* is the ratio of the number of examples in the dataset with variable a has the value *v*, and *H(Sa(v))* is the entropy of a group of samples where variable a has the value *v*.