Extreme Gradient Boosting (XGBoost)

- XGBoost regression is short form for extreme gradient boost regression.

- XGBoost is a high-performance machine learning algorithm based on gradient boosting, where multiple weak prediction models, typically decision trees, are sequentially trained to correct preceding errors and create a strong predictive model.

- XGBoost introduces regularization into the objective function to control complexity, reducing overfitting.

- It also has built-in handling of missing values, learning the best imputation during training. An advantage over other boosting methods is its tree pruning techniqe which enhances efficiency by eliminating non-beneficial splits.

- Additionally, XGBoost leverages a Column Block structure for efficient sparse data handling, and employs parallel computing for speed.

- It works well compared to others machine learning algorithms.XGBoost is one of the best supervised learning algorithms which can inferred by the way it flows, it consists of objective function and base learners.

- Loss function is present in objective function which shows the difference between actual values and predicted values whereas regularisation term is used for showing how far is actual value away from predicted value.

- Ensemble learning used in XGBoost considers many models which are known as base learners for predicting a single value.

- Not all base learners are expected to have bad prediction so that after summing up all of them bad prediction cancelled out by good prediction.

- A regressor is the one that fits a model using given features and predicts the unknown output value.

XG Boost Regression Algorithm for House Price Prediction

Input: House attributes dataset.

Output: Price of house.

1. Check input dataset for missing values and calculate d mean is replaced in place of missing value.

2. Divide attributes based on values in data fields as categorical and non-categorical rows.

3. Check Non categorical rows for outliers using outlier detection techniques and remove all outliers.

4. Convert categorical rows into binary vectors using one hot encoding.

5. Divide dataset for cross validation using train test split.

6. Apply Ensemble learning through training and combining individual models termed as base learners in order to derive a single prediction.

a) Calculate Mean Squared Error (MSE) with true values to predicted values.

b) Classify independent models as weak-learners and strong-learners using error detection.

c) Total mean cancels bad prediction with good prediction.

7. Objective function contains the loss function and regularisation term to calculate difference between

actual value and predicted value.

Data pre-processing is a process used for refining data before fed into model.

Data pre-processing is vaguely divided into four stages called

- Data cleaning
- Data Editing
- Data reduction
- Data wrangling

Data cleaning is process where inaccurate data or if a data field is empty, then value is filled using mean or median or entire record is deleted from data. If data is recorded manually these problems tend to happen. Calculate the mean value considering the value of attributes and number of records in the data.

Data editing is process where outliers are picked from data and eradicated. Outliers are mainly recorded in data mainly due to experimental errors produced by machined due to malfunctioning or due to some other parameters.

Data reduction is termed as the process of reducing data using some kind of normalisation for easy process of data. Z score is one of processes used for normalisation.

Data wrangling is termed as a process where data is transformed or mapped. Data munging, data visualisation and data aggregation comes under this process.

Data visualisation is process where statistics are used for producing graphs. Data aggregation is process where data is filtered before fed into model.

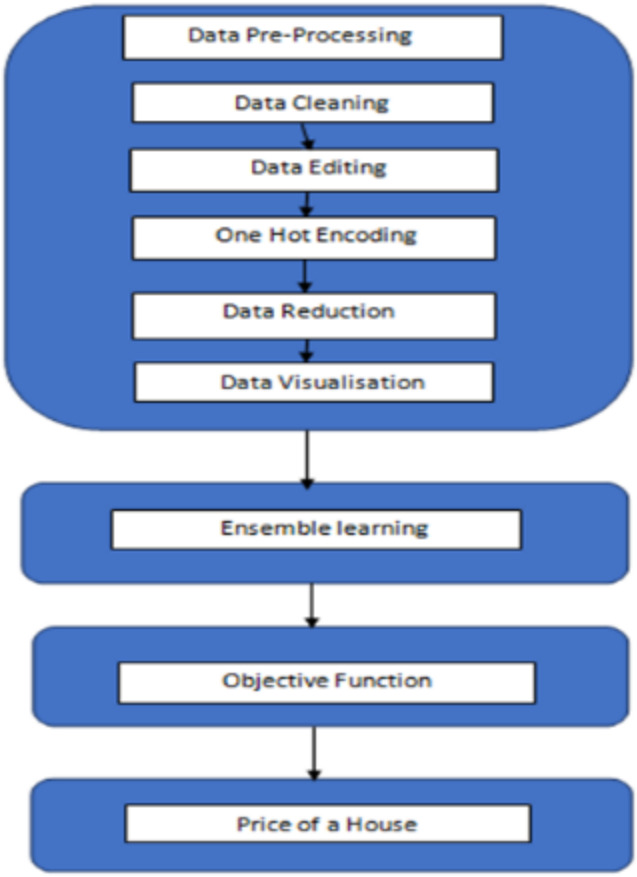

Fig.1.Flow Diagram for House Price Prediction

Implementation of XGBoost

- Extreme Gradient Boosting (XGBoost) is an enhanced gradient boosting machine using the tree ensemble boosting process.
- This process ends in the sum of the outputs from all the trees.
- The XGBoost algorithm used the XGBoost package in Python to evaluate house prices.
- The sample data allocation scheme used in this model is the same as the previous model which is LightGBM algorithm.
- Analysis of the feature importance with XGBoost model shows the rank of features (in sequence, highest to lowest): size of the house,

the number of parking lots provided for a house, and the nearest distance to the public transport feature.

- The selection of the features is quite different compared to the previous two models (Multiple Linear Regression and Ridge Regression).

- The proposed XGBoost model is the first application of XGBoost to the study of the Kuala Lumpur housing market.

- The model used in this analysis was able to tackle the problems of the housing market in Kuala Lumpur as the XGBoost model has a better fitting and predictive abilities.

- The XGBoost model was able to generate results that were more consistent and justifiable than other models used for housing market data.

- The XGBoost model achieved better predictive ability, with the lowest mean absolute error (MAE) and root mean squared error (RMSE), and adjusted R-squared value closest to 1, which indicates the most accurate model.

- In addition, consistent model performance was found in the XGBoost model as XGBoost outperformed other models in the training and testing R-squared value.

- The proposed XGBoost model is, therefore, effective in predicting housing prices, which favor not only future house buyers but also investors and policymakers in the real estate industry.

Conclusion

A model for house price prediction that assists both buyer and seller had been proposed. The proposed house prediction system helps seller to sell house at best price and it also helps buyer to buy house at best price. Price of said land costs, material varies from place to place.

Prediction of price of house is difficult as it varies from place to place as all attributes doesn't have same proportion in all places. Deep learning algorithms may enhance the prediction of price of house and decrease test error rate percentage.

The XGBoost regression algorithm helps to satisfy the needs of customers by increasing accuracy of the choice of estates and decreasing the risk for customers to invest in real estate. The system can be made widely acceptable by including more number of features. Future scope is to create estate database including more cities in order to help customers explore more number of estates and get accurate decision.