Documentation for house price prediction

The data set consists of 21 variables and 21597 observations.

(Note: For some of the variables that are self explanatory, no definition has been provided)

- **Id**: Unique identification number for the property.

- **date**: date the house was sold.

- **price**: price of the house.

- **waterfront**: house which has a view to a waterfront.

- **condition**: How good the condition is (overall). **1** indicates worn out property and **5** excellent.

- **grade**: Overall grade given to the housing unit, based on King County grading system. 1 poor ,13 excellent.

- **Sqft_above**: square footage of house apart from basement.

- **Sqft_living15**: Living room area in 2015(implies - some renovations). This might or might not have affected the lotsize area.

- **Sqft_lot15**: lotSize area in 2015(implies - some renovations).

In this project following method were applied for price prediction and multilayer perceptron and gradient boost regressors perform better for price predition for this dataset.

- Linear Regression (variables are linear in the predictor and it could be used after fixing the target and skewness of features)
- Ridge Regression (It is most suitable when a data set contains a higher number of predictor variables than the number of observations or when multicollinearity is experienced in a set)
- Bayesian Ridge Regression (Usually more accurate then Lreg.)
- Selective Vector Regression (High-dimension data/a lot of features)
- Decision Tree Regression (Categorical and continuous variables)
- Random Forest Regression (Many decision Tree)

- Kneighbors Regression (simple and easy)
- Gradient Boost Regression (Fast , flexible, no data preprocessing required)
- MLP Regression (suitable algorithm where a real-valued quantity is predicted given a set of inputs)