**Assignment 1**

***Description of the forecasting problem***

This forecasting problem deals with predicting the housing prices based on the attributes of a house like number of bedrooms, number of bathrooms, number of floors and the total area of the house in square feet.

Y (Variable of interest) = Price of the house

X1 = number of bedrooms

X2 = number of bathrooms

X3 = number of floors

X4 = Total area in square feet

Where X1, X2, X3 and X4 are the independent variables/attributes, based on the values of which the dependent variable Y can be predicted.

***Description of the available data***

|  |  |  |
| --- | --- | --- |
| **Column name** | **Type** | **Description** |
| Id | Numeric | a notation for a house |
| date | String | Date on which the house was sold |
| price | Numeric | Price of the house |
| bedrooms | Numeric | Number of bedrooms in the house |
| bathrooms | Numeric | Number of bathrooms in the house |
| sqft\_living | Numeric | Square feet area of the living room |
| sqft\_lot | Numeric | Square feet area of the whole house |
| floors | Numeric | Number of floors/levels in the house |
| waterfront | Numeric | House which has the view of a waterfront |
| view | Numeric | Has been viewed |
| condition | Numeric | How good the condition is (Over all) |
| grade | Numeric | Overall grade given to the housing unit,  based on King County grading system |
| sqft\_above | Numeric | Square feet of the house apart from basement |
| sqft\_basement | Numeric | Square feet of the basement |
| yr\_built | Numeric | Year in which the house was built |
| yr\_renovated | Numeric | Year in which the house was last renovated |
| zipcode | Numeric | Zip Or Postal code of the area |
| lat | Numeric | Latitude Co-ordinate |
| long | Numeric | Longitude Co-ordinate |
| sqft\_living15 | Numeric | Living room area in 2015 (Implies some renovations) |
| sqft\_lot15 | Numeric | Lot size area in 2015 (Implies some renovations) |

**Attributes used:**

Y (Variable of interest) = Price of the house

X1 = number of bedrooms

X2 = number of bathrooms

X3 = number of floors

X4 = Total area in square feet

***Short overview of the selected algorithms***

**Linear Regression**

Regression – an approach for modeling the relationship between a dependent variable and independent variables

Linear regression – linear relationship between a dependent variable and independent variables

In simple linear regression, we predict scores on one variable from the scores on a second variable. The variable we are predicting is called the *criterion variable* and is referred to as Y. The variable we are basing our predictions on is called the *predictor variable* and is referred to as X. When there is only one predictor variable, the prediction method is called *simple regression*. If there are more than one predictor variables, the prediction is called *multivariate linear regression.*

**ARIMA (Time series)**

In statistics and econometrics and in particular in time series analysis an **autoregressive integrated moving average (ARIMA)** model is a generalization of an autoregressive moving average (ARMA) model. Both of these models are fitted to time series data either to better understand the data or to predict future points in the series (forecasting) ARIMA models are applied in some cases where data show evidence of non-stationarity, where an initial differencing step (corresponding to the "integrated" part of the model) can be applied one or more times to eliminate the non-stationarity.

The AR part of ARIMA indicates that the evolving variable of interest is regressed on its own lagged (i.e., prior) values. The MA part indicates that the regression error is actually a linear combination of error terms whose values occurred contemporaneously and at various times in the past. The I (for "integrated") indicates that the data values have been replaced with the difference between their values and the previous values (and this differencing process may have been performed more than once). The purpose of each of these features is to make the model fit the data as well as possible.