Domain Background

With the growing demand and exponentially increasing prices in the real estate market, it has become imperative to study in detail the various factors affecting the increase in the prices. This field of study is crucial to both the parties involved - the owner/user/renter on the demand side and the developer/renovator on the supply side. Here, we will try to predict the prices of residential properties using data from Ames, Iowa, USA.

The dataset, known as the Ames Housing dataset and compiled by Dean De Cook, can be seen as an alternative to the famous Boston Housing dataset.

Having worked on the Boston Housing dataset as part of the Supervised Learning project, this dataset would allow me to explore a similar problem in much more detail and depth.

Problem Statement

To predict the sale price of residential homes on the basis of 79 variables which describe the quality and quantity of the different attributes of the individual properties.

Datasets and Inputs

The data has been taken from a Kaggle competition - House Prices: Advanced Regression Techniques.

The dataset has 79 input variables that covers almost all aspects of a residential home. The input contains both numeric and categorical variables.

We need to predict the sale prices of the individual properties. The data has been divided into a training (1460 data points) and a testing data (1459 data points).

Having a continuous target variables, this is a typical example of a regression problem with-

- input variables that are both numeric and categorical,
- many independent variables which would require advanced feature selection/engineering, and
- a lot of missing data.

Solution Statement

Due to the presence of a large number of independent variables, intelligent feature engineering and the identification of key variables is of utmost importance.

There are also a lot of missing values (NA) in the data. Appropriate imputation measures will need to be taken for effective missing value treatment.

After detailed exploratory data analysis (EDA) and training different models, the predicted values of the prices for the individual houses in the test dataset would be obtained.

Benchmark Model

A benchmark model here would be a very simple linear regression model taking into account few variables that would intuitively be the most useful ones.

I will then use various other machine learning techniques to improve upon this benchmark model. Different techniques such as Linear Regression, Lasso/Ridge Regression, Elastic Nets Regression, XGBoost, Random Forest and Neural Networks.

Evaluation Metrics

This being a regression problem with a continuous target variable, the following metrics can be checked for -

- R Square/ Adjusted R Square (in case of linear regression)
- Root Mean Squared Error (RMSE)
- Mean Absolute Percentage Error (MAPE)

Project Design

Steps to be followed would include -

- imputation of the missing values
- identification of the most significant variables in the data
- building various regression models on the training dataset
- predicting on the testing dataset and checking for various metrics
- choosing the model that best predicts the target prices in the testing data
- cross validation of the results to ensure there is not overfitting

Visualizations would include -

- scatter plots: to study the correlations between the continuous variables
- box plots: to study the distributions of the different input variables
- bar plots: to study the distribution of the different categorical variables with respect to the prices
- line plots: to compare the predicted prices of the properties to their actual prices