House Price prediction

Introduction:

The housing and real estate markets influence people's basic necessities all throughout the world and play a significant role in the global economy. By using data science technologies, businesses can improve revenue, hone their marketing tactics, and adjust to changing patterns in real estate deals. Housing companies are able to accomplish their goals through the use of market analysis and predictive modeling approaches.

Problem Statement:

Surprise Housing, a US-based housing company, plans to enter the Australian market by purchasing properties below their actual market value and reselling them at a profit. To achieve this, the company needs to accurately predict the market value of prospective properties using data analytics. The objective is to develop a machine learning model that can:

1. Identify the key variables that are important in predicting house prices.
2. Explain how these variables influence the house prices.

The model will enable Surprise Housing to make informed investment decisions and strategically enter the Australian real estate market.

Data Collection and Preprocessing:

Overview

We have two datasets: `train.csv` (1460 entries, 81 variables) and `test.csv`, along with "Data file.csv" and "Datadescription.txt". The target feature is `Salesprice`, indicating a regression problem. The dataset contains significant missing values and outliers.

Data Collection:

1. Training Data (`train.csv`): Used to train the model, including `Salesprice` and 80 other features.

2. Test Data (`test.csv`): Used for model evaluation, containing similar features excluding `Salesprice`.

Data Preprocessing Steps:

1. Handle Missing Values: Impute or remove missing data.

2. Outlier Treatment: Identify and mitigate outliers.

3. Feature Engineering: Create and select relevant features.

4. Data Transformation: Normalize/standardize numerical features and encode categorical variables.

5. Data Splitting: Split training data into training and validation sets.

Exploratory Data Analysis (EDA):

To prepare the dataset for modeling, a thorough exploratory data analysis (EDA) was carried out, with special attention paid to managing missing values, multicollinearity, outliers, and comprehending feature associations with the target variable, {Salesprice}. To guarantee a complete dataset, missing values were first found and either imputed or eliminated. The variance inflation factor (VIF) was used to measure multicollinearity, and in order to enhance model performance, columns with significant multicollinearity were eliminated. Skewness analysis was used to identify outliers. Rather of deleting them, which would have caused a 60% loss in data, the outliers were imputed using upper and lower quartile values derived from the interquartile range (IQR).
To determine which variables have a favorable or negative impact on {Salesprice}, a correlation matrix was constructed. For focused analysis, features were then divided into numerical and category types. To provide insights into distribution, countplots were utilized for categorical variables, and boxplots were employed to illustrate the influence of each category variable on {Salesprice}. Plotting probability distribution functions (PDFs) allowed users to see the distributions of numerical features. To control dimensionality and prevent an unwarranted increase from one-hot encoding, categorical data were encoded using label encoding. In order to guarantee consistency and dependability, the training and testing datasets underwent the same preprocessing procedures. This comprehensive EDA helped Surprise Housing make wise decisions by offering vital insights and preparing a solid dataset for the creation of an accurate prediction model.

Model Selection and Training:

Ideally, target variables should be present in both training and test datasets. The target variable is utilized to train the model in the training dataset, and in the test dataset, the

target variable is used to gauge the model's effectiveness using the dataset.
The following notation is typically used to represent these:
x_train is the training dataset's independent variable.

y_train is the training dataset's target variable.
x_test is the test dataset's independent variable.
y_test is the test dataset's target variable.
We are unable to test our dataset on y_test since it is not provided in our case for the test dataset. As stated in the issue statement, our objective is to forecast the test dataset's target feature, or "salesprice".

ML models used & Model evaluation:

For this problem we applied Linear Regression, Random Forest, Gradient Boasting , KNN regressor and Xgboast and evaluated our performance for each of these algorithym by the help of evaluation metrics like Mean Absolute error, Root Mean Square Error and R2 score (co-efficient of determination) The model which has less Root mean square error and a higher R2 value is the best fitting model for this dataset. I find out that Random Forest gives the best result compared to the others .

Conclusion:

After finding the best model I use that model to predict the target feature of our already preprocessing test dataset.

Business Implications:

Significant business ramifications stem from the created machine learning methodology for Surprise Housing's entry into the Australian market. The organization benefits from the strategic advantage of being able to make well-informed property investment decisions by precisely projecting house prices based on important independent variables. This gives management the ability to deploy resources wisely, concentrate on possibilities with high returns, and customize their market entry plan to achieve the best results. In addition, the model's understanding of pricing dynamics offers useful market knowledge, empowering management to accurately and confidently traverse the new market environment.