CoachX

House Price Prediction

Data Analysis using Python

House Price Prediction

House Price Prediction Project Overview:

House price prediction aims to estimate the selling price of a property using various factors like **physical conditions, concept, and location**. These factors influence the sale price and help developers, real estate agents, and buyers make better decisions. For instance, a seller may adjust the listing price based on these estimates, and buyers can gauge the right time to make a purchase based on market conditions.

This project involves performing **feature engineering** and analyzing these features to create a machine learning model capable of predicting **house prices**. The main objective is to find the relationship between the features (both categorical and numerical) and the sale price, as well as handle missing data, outliers, and temporal variables.

Datasets is given in with this Project

The dataset provided consists of various features related to real estate properties, with the goal of predicting the **house prices** (SalePrice). It has the following columns:

1. **Id**: Unique identifier for each house.
2. **MSSubClass**: Type of dwelling involved in the sale (e.g., 20 for 1-story 1946 and newer, 60 for 2-story 1946 and newer).
3. **MSZoning**: General zoning classification of the sale (e.g., RL for Residential Low Density, RM for Residential Medium Density).

4. **LotArea**: Lot size in square feet.
5. **LotConfig**: Configuration of the lot (e.g., Inside, Corner, CulDSac).
6. **BldgType**: Type of dwelling (e.g., 1Fam for Single-family detached, TwnhsE for Townhouse end unit).
7. **OverallCond**: Rates the overall condition of the house (on a scale of 1 to 10).
8. **YearBuilt**: Original construction year of the house.
9. **YearRemodAdd**: Year when the house was remodeled.
10. **Exterior1st**: Exterior covering on the house (e.g., VinylSd, Wd Sdng).
11. **BsmtFinSF2**: Type 2 finished square feet in basement.
12. **TotalBsmtSF**: Total square feet of basement area.
13. **SalePrice**: The property's sale price

Task 1: Basic Data Exploration

1. **Import necessary and essential libraries**
   Before any analysis, we need to import libraries like **pandas**, **numpy**, **matplotlib**, and **seaborn** for data manipulation and visualization, along with **scikit-learn** for machine learning.
2. **Display all the columns of DataFrame**
   We will display all the columns of the dataset using Pandas to get a better understanding of the data structure.
3. **Read the data and display the first 100 rows**
   We'll read the dataset from a file (e.g., CSV) and display the first 100 rows to get a quick overview of the data.
4. **Give column insights**
   For each column, we'll analyze what the data represents, whether it's numerical, categorical, or temporal, and its potential impact on the house prices.

Task 2: Data Cleaning and Analysis

Q1) **Checking for Missing Values**
It's important to identify missing values in the dataset because they can affect model performance. We'll use functions like isnull() and sum() to check for missing values in each column.

Q2) **Features with NAN Values**
We will identify which columns contain missing values and how many rows are affected. This will help us decide whether to drop or impute these values.

Q3) **Calculate Mean Sales Price for Missing/Present Information**
For each feature with missing data, we will calculate the mean sale price based on whether information is present or missing to understand its impact.

Q4) **Count of Numerical Features**
We'll count the number of numerical columns in the dataset, which are important for statistical analysis.

Q5) **Print the First Five Rows of Numerical Values**
Displaying the first five rows of numerical features will help us understand the range and structure of the numerical data.

Q6) **Compare the Difference Between Year Features and SalePrice**

We will compare features like YearBuilt and YearRemodAdd with the sale price to see how the age or renovation of a house affects its value.

Q7) **Relationship Between Discrete Variables and Sales Price**

Discrete variables (e.g., OverallCond) will be analyzed to see how they influence the sale price using statistical techniques and visualizations.

Q8) **Relationship Between Continuous Variables and Sales Price**

We will analyze continuous variables like LotArea and TotalBsmtSF to check their correlation with the sale price.

Q9) **Histogram Analysis for Continuous Variables**

A histogram will be created for continuous variables to visualize their distribution and see if any transformation (like log transformation) is needed.

Q10) **Logarithmic Transformation**

Apply logarithmic transformation to skewed data to make it more normally distributed and improve model performance.

Task 3: Feature Engineering and Advanced Analysis

Q1) **Find Outliers**
We will use methods like boxplots and statistical techniques to detect outliers, which can skew predictions and need to be handled carefully.

Q2) **Relationship Between Categorical Features and SalePrice**
We will explore how categorical features like MSZoning and BldgType impact the sale price, using group-by operations and visualizations.

Q3) **Correlation Between Numerical Features and SalePrice**
Correlation analysis (like Pearson correlation) will be used to identify which numerical features have the strongest correlation with the sale price.

Q4) **Continuous Features vs. SalePrice**
We will examine continuous variables' impact on house prices using scatter plots and correlation matrices.

Q5) **Feature Engineering**

- **Handle Missing Values**: Techniques like mean/mode imputation or using predictive models to fill in missing data.

- **Handle Categorical Variables**: Convert categorical variables into numerical ones using techniques like **One-Hot Encoding** or **Label Encoding**.
- **Handle Numerical Variables**: Scale or transform numerical variables to ensure that they fit the model's assumptions.
- **Handling Temporal Variables**: Extract relevant information from temporal features like age of the house (e.g., difference between YearBuilt and the current year).

Advanced Questions for the Project:

- **Q11**: How does the location (based on zoning and lot configuration) influence the house price?
- **Q12**: What is the impact of the overall condition of the house on the sale price?
- **Q13**: Does the presence of a basement (based on TotalBsmtSF) significantly affect the house price?
- **Q14**: How do remodeling and renovations (YearRemodAdd) influence the property's value over time?
- **Q15**: Can we predict house prices using just categorical features (like building type, exterior material) without numerical features?

DOCUMENTATION

After Completeion of the projects you have to Create one Docx file in that you have to Make Report of above Projects and include some Key Factors like

1. Introduction
2. Methodology
3. Requirement Analysis
4. Other Parameters depending upon the Projects
5. All Visualization like All Charts which is there in the Dashboards
6. Insights from the Charts as well as Dashboards
7. Conclusion

After creating the Reports for above project upload that docx file or pdf file in Assignment links

Tip :- For your Reference One Sample report is associated with this Project . You can refer that project for your use-case .