

**CSE 180 – Advanced Python – Continuous Assessment 4**

**Project report**

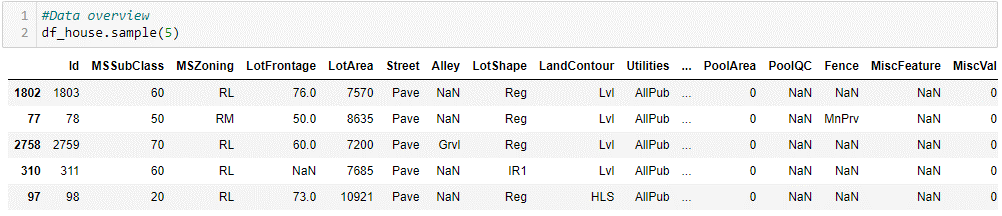
**House Price Prediction**

*Arjun- E0222054*

*Sujan- E0222005*

|  |  |  |  |
| --- | --- | --- | --- |
| **Sno** | **Split up** | **Maximum Mark** | **Mark obtained** |
| 1 | Dataset Preprocessing | 25 |  |
| 2 | Use of ML algorithms | 10 |  |
| 3 | Result Visualization | 15 |  |
|  | **Total** | **50** |  |

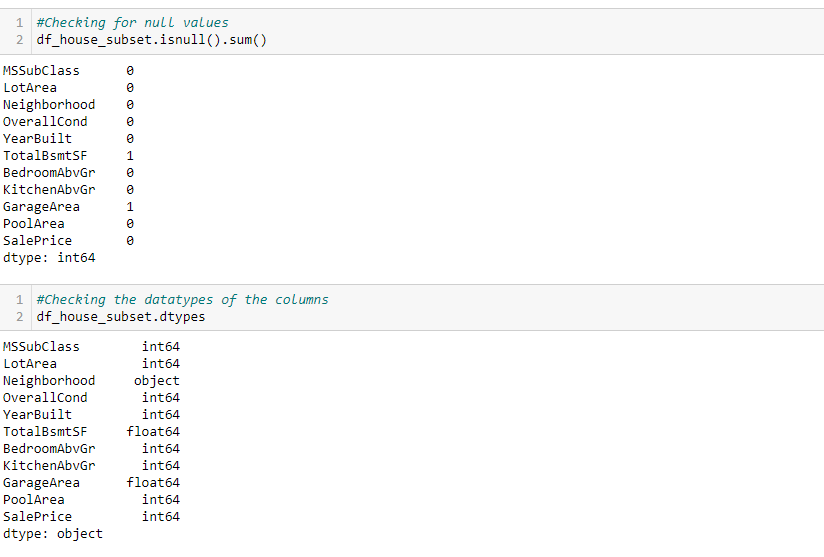
1. **Paste Dataset here (top 5 rows)**

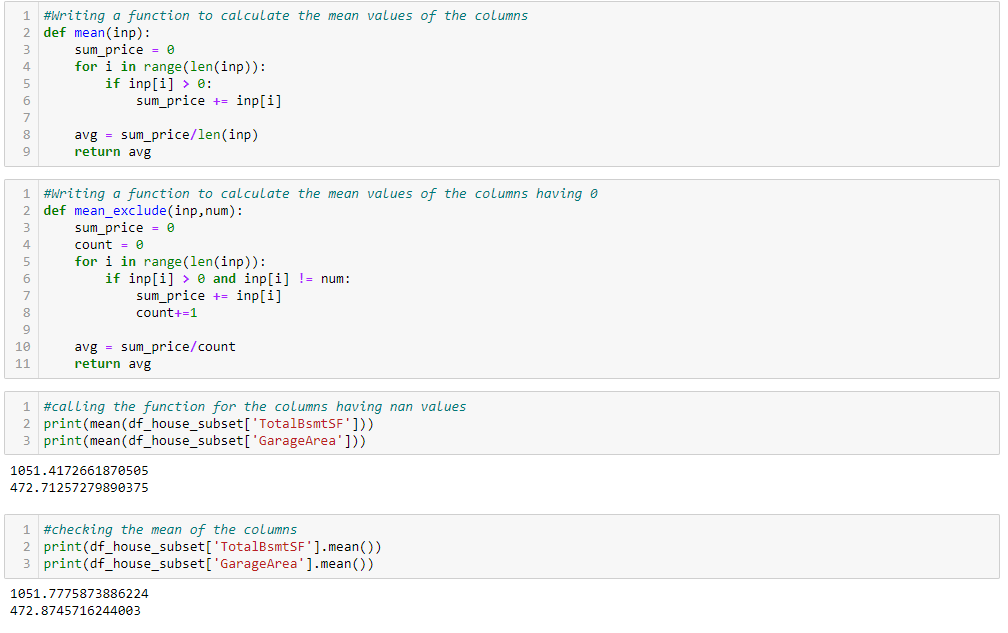


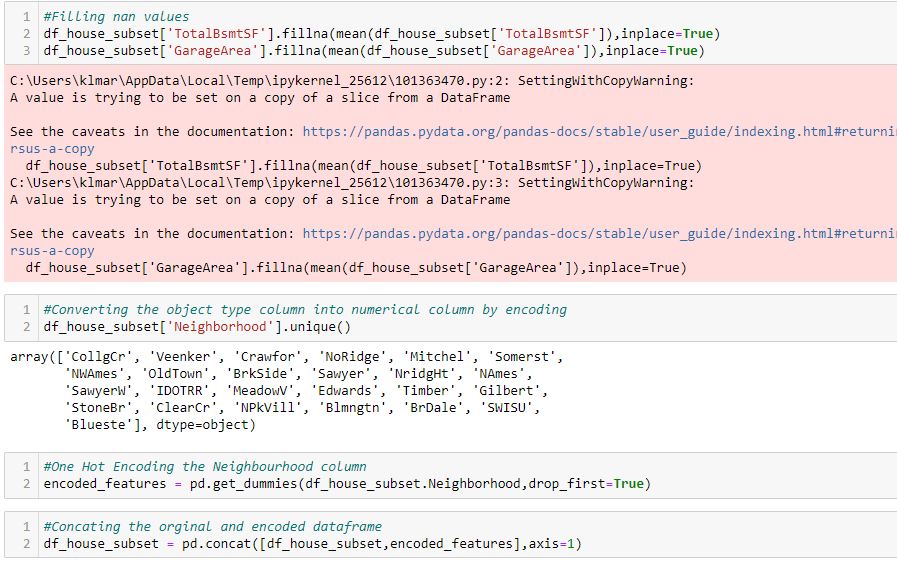
1. **Explain features (their need in classification, independent and dependent)**

* **MSSubClass: This column refers to the type of dwelling involved in the sale. The value is a categorical variable indicating the building class. This variable could provide some insights into the type of home being sold, such as whether it is a single-family home, townhouse, or condominium. Different home types may have different values depending on the location and demand, so this column could help capture some of that variation.**
* **LotArea: This column refers to the size of the lot in square feet. Lot size is often a significant factor in home values, as larger lots may be more desirable for some buyers. This column could capture some of the variation in home prices due to differences in lot size.**
* **Neighborhood: This column refers to the location of the home within a specific neighborhood. Neighborhood can be an important factor in home values, as different areas may have different amenities, schools, or other features that affect demand. This column could help capture some of the variation in home prices due to location.**
* **OverallCond: This column refers to the overall condition of the home, as rated by the seller. The condition of a home is often an important factor in its value, as buyers may be willing to pay more for a home that is in better condition. This column could help capture some of the variation in home prices due to differences in condition.**
* **YearBuilt: This column refers to the year the home was built. Age can be an important factor in home values, as newer homes may be more desirable to some buyers. This column could help capture some of the variation in home prices due to differences in age.**
* **TotalBsmtSF: This column refers to the total square footage of the basement. Basements can be a valuable feature in some homes, especially if they are finished and can be used as additional living space. This column could help capture some of the variation in home prices due to differences in basement size.**
* **BedroomAbvGr: This column refers to the number of bedrooms above ground level. The number of bedrooms can be an important factor in home values, as larger homes with more bedrooms may be more desirable to some buyers. This column could help capture some of the variation in home prices due to differences in bedroom count.**
* **KitchenAbvGr: This column refers to the number of kitchens above ground level. The number of kitchens may not always be relevant, but in some cases, homes with more than one kitchen (such as a separate apartment or mother-in-law suite) may be more valuable. This column could capture some of that variation.**
* **GarageArea: This column refers to the size of the garage in square feet. Garages can be a valuable feature in some homes, especially if they are large enough to accommodate multiple vehicles or offer additional storage space. This column could help capture some of the variation in home prices due to differences in garage size.**
* **PoolArea: This column refers to the size of the pool in square feet. Pools can be a valuable feature in some homes, especially in warm climates or for families with children. This column could help capture some of the variation in home prices due to differences in pool size.**
* **SalePrice: This column is the target variable - the value we are trying to predict. It provides information on the actual sale price of the home, which we can use to train and test our predictive models. By using the other columns as predictors, we hope to build a model that can accurately predict sale price for new homes based on their features.**

1. **Implement pre-processing**

****

****

****

1. **Explain about the ML model (upto 3 pages)**

**Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It is a powerful tool for predicting values of a dependent variable based on the values of independent variables. In this article, we will provide an in-depth explanation of linear regression, including its assumptions, types, and applications.**

**Linear Regression Basics:**

**Linear regression is a mathematical technique used to find the best linear relationship between a dependent variable and one or more independent variables. It can be used to predict the value of the dependent variable for a given value of the independent variable. The relationship between the dependent variable and independent variables can be represented by the following equation:**

**y = b0 + b1x1 + b2x2 + ... + bn\*xn + e**

**where y is the dependent variable, x1, x2, ..., xn are the independent variables, b0, b1, b2, ..., bn are the regression coefficients, and e is the error term. The regression coefficients represent the change in the dependent variable for a unit change in the corresponding independent variable, holding all other independent variables constant.**

**The goal of linear regression is to estimate the regression coefficients that minimize the sum of the squared residuals between the predicted and actual values of the dependent variable. The squared residuals are the difference between the predicted value and the actual value of the dependent variable, squared to ensure that negative and positive errors do not cancel each other out.**

**Assumptions of Linear Regression:**

**Linear regression relies on several assumptions about the data:**

**Linearity: The relationship between the dependent variable and independent variables should be linear.**

**Independence: The observations should be independent of each other.**

**Homoscedasticity: The variance of the errors should be constant across all values of the independent variables.**

**Normality: The errors should be normally distributed.**

**No Multicollinearity: The independent variables should not be highly correlated with each other.**

**Types of Linear Regression:**

**There are two main types of linear regression:**

**Simple linear regression: Simple linear regression involves a single independent variable and a single dependent variable. The equation for simple linear regression is:**

**y = b0 + b1\*x + e**

**where y is the dependent variable, x is the independent variable, b0 is the intercept, b1 is the slope, and e is the error term.**

**Multiple linear regression: Multiple linear regression involves two or more independent variables and a single dependent variable. The equation for multiple linear regression is:**

**y = b0 + b1x1 + b2x2 + ... + bn\*xn + e**

**where y is the dependent variable, x1, x2, ..., xn are the independent variables, b0, b1, b2, ..., bn are the regression coefficients, and e is the error term.**

**Applications of Linear Regression:**

**Linear regression has a wide range of applications in various fields, including:**

**Economics: Linear regression is used to model the relationship between variables such as demand and price, or GDP and population.**

**Finance: Linear regression is used to model the relationship between variables such as stock prices and economic indicators.**

**Marketing: Linear regression is used to model the relationship between variables such as advertising spend and sales.**

**Health: Linear regression is used to model the relationship between variables such as age, weight, and blood pressure.**

**Science: Linear regression is used to model the relationship between variables such as temperature and reaction rate.**

**Conclusion:**

**In summary, linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It is a powerful tool for predicting values of a dependent variable based on the values of independent variables.**

1. **Explain about the performance metrics (any 3)**

**Pearson Correlation Coefficient**

**The Pearson correlation coefficient (r) measures the linear relationship between two variables. It is a value between -1 and 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation. The formula for calculating the Pearson correlation coefficient is:**

**r = (NΣXY - ΣXΣY) / sqrt((NΣX^2 - (ΣX)^2) \* (NΣY^2 - (ΣY)^2))**

**where N is the number of data points, ΣXY is the sum of the products of the corresponding values of the two variables, ΣX and ΣY are the sums of the values of the two variables, and ΣX^2 and ΣY^2 are the sums of the squares of the values of the two variables.**

**The Pearson correlation coefficient is useful for identifying the direction and strength of the relationship between two variables. For example, if the correlation coefficient between the price of a house and its size is 0.8, this indicates a strong positive relationship between the two variables. However, the Pearson correlation coefficient only measures the linear relationship between variables and cannot capture nonlinear relationships.**

**R-squared Score**

**The R-squared score (R^2) is a commonly used metric to evaluate the performance of a regression model. It measures the proportion of the variance in the dependent variable that is explained by the independent variables. R-squared values range from 0 to 1, where 0 indicates that the model explains none of the variability in the dependent variable, and 1 indicates that the model explains all of the variability.**

**The formula for calculating R-squared is:**

**R^2 = 1 - (SSres / SStot)**

**where SSres is the sum of the squared residuals and SStot is the total sum of squares. The residual is the difference between the predicted value and the actual value, and the total sum of squares is the sum of the squared differences between each data point and the mean of the dependent variable.**

**R-squared is useful for comparing different models and determining which model explains the variance in the dependent variable better. However, R-squared can be misleading if the model includes too many independent variables, as the model may overfit the data and perform poorly on new data.**

**Adjusted R-squared**

**The adjusted R-squared metric is a modified version of R-squared that takes into account the number of independent variables in the model. It penalizes the R-squared value for including too many independent variables that may not contribute significantly to the prediction of the dependent variable.**

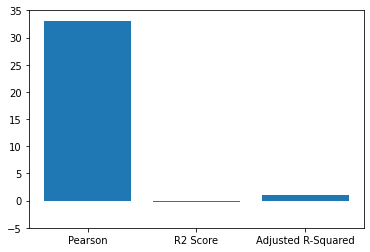
**The formula for calculating the adjusted R-squared is:**

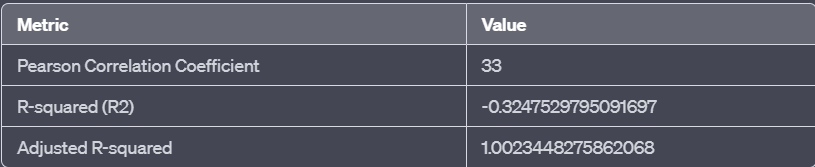
**Adjusted R^2 = 1 - [(1 - R^2) \* (n - 1) / (n - p - 1)]**

**where n is the number of observations, and p is the number of independent variables.**

**The adjusted R-squared is useful for comparing models with different numbers of independent variables. The adjusted R-squared will always be less than or equal to the R-squared, and the adjusted R-squared value increases only if the addition of an independent variable improves the model more than would be expected by chance.**

1. **Present the results of all performance metrics in tabular and graph form**

****

****