CONVERSATIONAL AI-DATA SCIENCE
ASSIGNMENT-1- HOUSE PRICES- ADVANCED REGRESSION TECHNIQUES

VATSAL NANDA
101916047, 3CSE10

1. DOCUMENTATION

- **IMPORTING THE LIBRARIES/DEPENDENCIES**
  - cuDF- It is a python GPU DataFrame library for loading, joining, aggregating, filtering and mainpulating data. It provides a pandas like API also.
  - cuPYy It is a Numpy/Scipy compatible array for a GPU accelerated computing with Python.
  - cuML- It is a fast, GPU-accelerated Machine Learning algorithms designed for data science and analytical tasks
  - Other libraries like 'from cuml.linear_model import LinearRegression' ,etc. are required to load the Linear Regression and other models.

- **READING THE DATA**
  - Reading the train and test datasets in train_df and test_df.
  - We check their shapes, info, description as well.

- **DATA PREPROCESSING**
  - We store the SalePrice column in Y and then drop it from the training dataset
  - We check for null values in the dataset
  - To visualise the null values, we plot a heat map for training and testing data.
  - Calculate the numerical and categorical features.
  - To handle null/ missing values, we drop all features having more than 50% null values, for categorical features, we take the mode for all the missing values and for the numerical features, we take the mean.
  - We again check for missing values by plotting a heatmap.
  - In training dataset, 'Electrical' Column, we had a null values instead of dropping it, we filled it with the previous value.
  - Concatenate the training and testing data as both of them have categorical features which can be converted into dummy/indicator variables using .get_dummies().
  - We split the final data into training and testing again.
  - We store the final training dataset in X

- **MODEL CREATION**
  - We first split X,Y into training and testing datasets in the ratio of 71:29 (training:testing).
  - We call the linear regression algorithm with the solver 'svd-jacobi' and make predictions as well.
  - Then we use the remaining algorithms, 'svd','eig','svd-qr and 'qr' in a loop and calculate evaluation metrics like MSE,MAE and R2 score.
  - 'svd'-alias for sad-jacobi
  - 'eig'-use an eigendecomposition of the covariance matrix
  - 'qr'-use qr decomposition algorithm
  - 'svd-qr'-compute SVD decomposition using QR algorithm
  - 'svd-jacobi'-compute SVD decomposition using Jacobi iterations
  - 'Mean Squared Error(MSE)'- Difference between model's predictions and the ground truth, square it and average out across the dataset.
  - Mean Absolute Error(MAE)'- Difference between model's predictions and the ground truth, apply the absolute value to that difference and average out across the dataset.
  - 'R2 Score'-Used to evaluate the performance of a regression based model, also known as the coefficient of determination and is measured by the amount of variance in the predictions explained by the dataset. Lies in the range of 0 to 1.
  - **LINEAR REGRESSION**-LinearRegression is a simple machine learning model where the response y is modelled by a linear combination of the predictors in X.
  - **RESULTS**-

ALGORITHM	MSE	MAE	R2 SCORE
SVD	918930744.330969	19170.1891252955	0.838884830474854
EIG	4870708400.37825	26288.6713947991	0.146022319793701
SVD-QR	889070439.115839	18962.9314420804	0.844120144844055
SVD-JACOBI	918930744.330969	19170.1891252955	0.838884830474854

  - (There is another algorithm 'qr' but when running this algorithm, it shows cannot handle missing values but there are no missing values in the dataset as it was checked in the previous cells and all the other algorithms were running)
  - Then a submission file with the predictions was created .
  - **OPTIONAL-**
  - **1) RIDGE REGRESSION**-Ridge extends LinearRegression by providing L2 regularization on the coefficients when predicting response y with a linear combination of the predictors in X. It can reduce the variance of the predictors, and improves the conditioning of the problem.

- 3 algorithms-‘eig’,’svd’,’cvd’
- ‘Eig’-uses a eigendecomposition of the covariance matrix, and is much faster
- ‘Svd’-SVD is slower, but guaranteed to be stable
- ‘Cd’-CD or Coordinate Descent is very fast and is suitable for large problems
- **2) LASSO REGRESSION**-Lasso extends LinearRegression by providing L1 regularization on the coefficients when predicting response y with a linear combination of the predictors in X. It can zero some of the coefficients for feature selection and improves the conditioning of the problem.
- **3) RANDOM FOREST REGRESSOR**(Not a part of the assignment but only to improve Kaggle score for submission)-Implements a Random Forest classifier model which fits multiple decision tree classifiers in an ensemble.
- Predictions were made and submitted.

- **FINAL RESULTS-**

MODEL	KAGGLE SCORE (RMSE)
LINEAR REGRESSION (SVD-JACOBI)	0.19678
RIDGE REGRESSION (EIG)	0.19634
RIDGE REGRESSION (SVD)	0.19645
LASSO REGRESSION	0.19399
RANDOM FOREST	0.14565

2. **GITHUB LINK OF THE CODE**- https://github.com/VatsalNanda/DATA-SCIENCE-UCS663_ASSIGNMENT-1/blob/main/datascience-lab-assignment-1.ipynb

3. **KAGGLE NOTEBOOK LINK**- https://www.kaggle.com/vatsalnanda/datascience-lab-assignment-1?scriptVersionId=88704360