Linear Regression to Predict House Prices

Using OLS & Sklearn

11:11PM

Project Pipeline

Data Collection

EDA

Data Preprocessing

Data Modelling

1. Removing Non-significant Data:
- Drop features that are more then 40% nans

2. Fixing Field Values:
- correct any inconsistencies

3.Handling Missing Values:
- replace NaNs with the mean of the column
- replace NaNs with the most frequently occurring value
- replace NaNs with zero

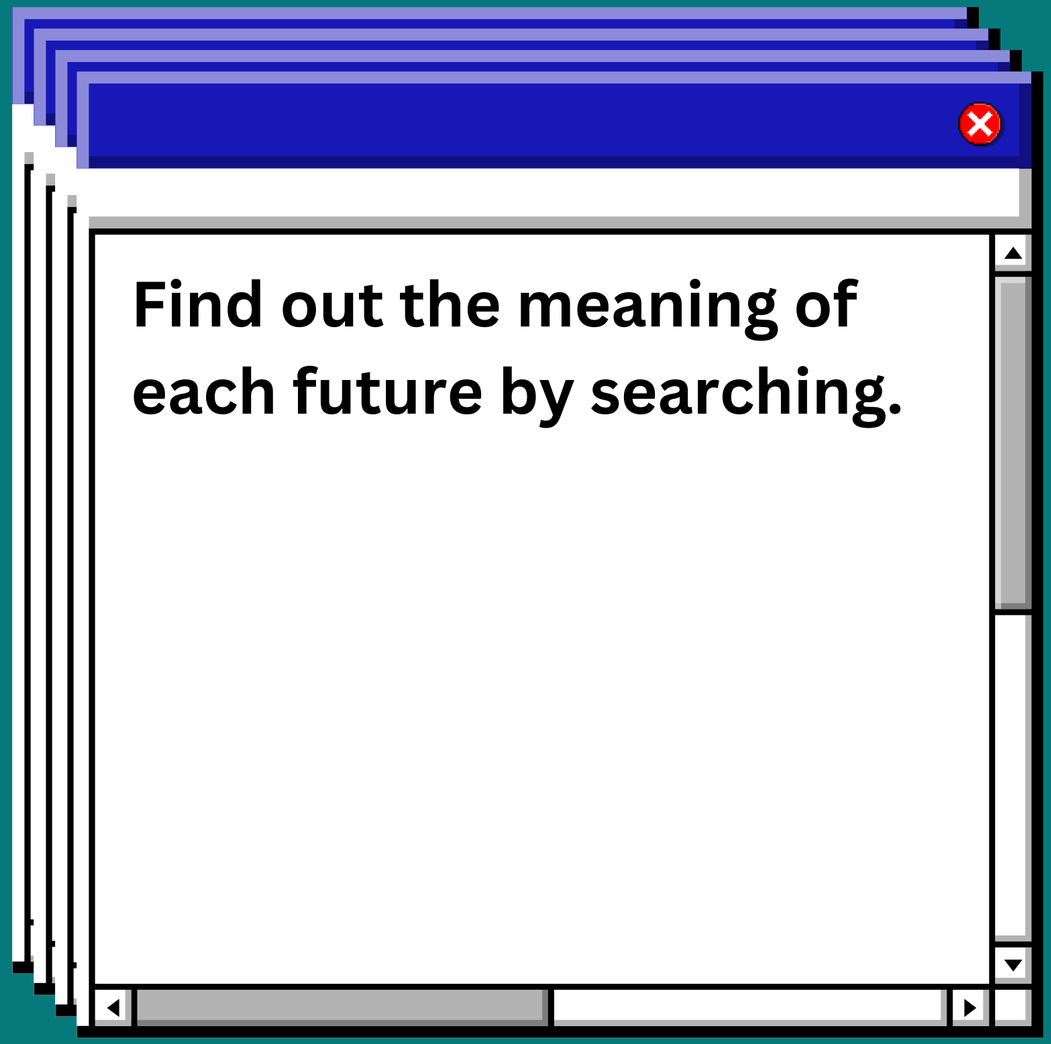
4. Rewrite features names

1. **Label Encoding:** Convert categories into integer values.
2. **Min-Max Scaling:** Scale numerical features to a range of [0, 1].
3. **One-Hot Encoding (Dummy):** Convert categorical variables into a set of binary variables.
4. **Feature Selection:** Identify and keep only the most relevant features for the model.

1. **OLS**
2. **Linear Regression using”Sklearn” Package:**
   - Normal Trine Validation and Test
   - Cross-Validation

11:11PM

Data Collection



EDA

1. Removing Non-significant Data: We will assess the number of NaN (missing) values in each feature and remove any feature that has more than 40% missing data. The features identified for removal are:

- Alley
- Pool QC
- Fence
- Mas Vnr Type
- Mas Vnr Area
- BsmtFin SF 2
- 2nd Flr SF
- Low Qual Fin SF
- Bsmt Full Bath
- Bsmt Half Bath
- Fireplace Qu
- Wood Deck SF
- Open Porch SF
- Enclosed Porch
- 3Ssn Porch
- Screen Porch
- Pool Area
- Misc Feature
- Misc Val

EDA

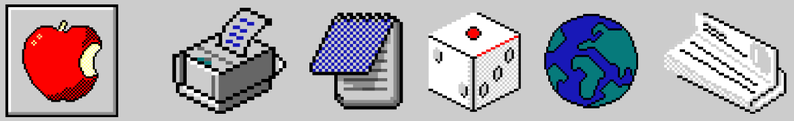
2. Fixing Field Values:

- There are some wrong values is 'MS Zoning' feature
- - A (agr)
- - C (all)
- - I (all)
- we don't need (all) and (agr) so we have to replace them withe the correct value (A, C, I) only

EDA

3. Handling Missing Values: Next, we will deal with the NaN values using different strategies:

- **Replace NaNs with the Most used Value:** For columns such as
- 'Bsmt Qual'
- 'Bsmt Cond'
- 'Bsmt Exposure'
- 'BsmtFin Type 1'
- 'BsmtFin Type 2'
- 'Electrical'
- 'Garage Type'
- 'Garage Finish'
- 'Garage Cars'
- 'Garage Qual'
- 'Garage Yr Blt'
- 'Garage Cond'



EDA

- **Replace NaNs with the Mean:** For the columns
- 'Lot Frontage'
- 'Garage Area'

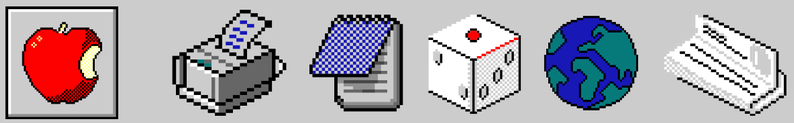
we will calculate the mean and replace NaNs with this value.

EDA

- **Replace NaNs with Zero:** For 'Bsmt Unf SF', we will replace NaNs with zero, as it is reasonable to assume that there was no unfinished space in the basement.

EDA
4. Rewrite Features Names

- Replace spaces, dots and ‘/’ with underscores in the column names (features)
- Replace 1st with ‘First’ in the column names (features)
- Replace ‘&’ with ‘and’ in the column names (features)



Data Preprocessing

1.Label Encoding: Convert the float features into integers to ensure consistency and facilitate further analysis.

- 'Lot Frontage'
- 'Total Bsmt SF'
- 'Garage Area'
- 'BsmtFin SF 1'
- 'Bsmt Unf SF'
- 'Garage Yr Blt'
- 'Garage Cars'

Data Preprocessing

2.Min-Max Scaling: We will scale the ‘Lot_Area’ f feature to a range of [0, 1] using Min-Max scaling and then multiply the scaled values by 100 to maintain the same data type (integer).

Data Preprocessing

3.One-Hot Encoding (Dummy): It transforms each category into a new binary column, where a value of 1 indicates the presence of that category and a value of 0 indicates its absence.

Data Preprocessing

4.Feature Selection: Analyzing the correlation between features and the target variable to select relevant features that are most predictive of the target.

11:11PM

Data Modelling

1. **Ordinary Least Squares (OLS):** The OLS regression analysis successfully identified significant predictors of `SalePrice`, with the final model achieving an ***R-squared*** value of **0.879** and **Adjusted R-squared 0.877**.

Data Modelling

2. Linear Regression using ”Sklearn” Package:
Implementing a robust regression analysis using Scikit-Learn after addressing multicollinearity in our features

Normal Trine Val and Test

The dataset is split into training, validation, and test sets. The results from the model evaluation are as follows:

- R-squared of val: 0.898
- R-squared of test: 0.821

Cross-Validation

We used 5-fold cross-validation to evaluate the model by training on different subsets and validating on others, ensuring robust results.

- Mean Cross-Validation R^2: 0.864