ISLAMIC UNIVERSITY OF TECHNOLOGY

Organization of Islamic Cooperation

Board Bazar, Gazipur

Lab 02

CSE 4622

CSE'18

23rd December, 2021

# Dataset Pre-processing

The original data consisted of many 0 values. These were replaced with null values to make processing easier.

Features which were determined to contain too many null values (more than 50%) were dropped. It would have been possible to perform further pre-processing here to keep some of these features. For example, the ‘yr\_renovated’ feature was dropped since a large number of samples had a 0 value, meaning they had not been renovated at all. It would perhaps be acceptable to set the renovation year to the year at which these houses were built, thus allowing us to keep this feature. However, such detailed pre-processing was avoided for now.

Features that had non-numerical datatypes were also dropped. Again, there are ways in which these features could have been kept, which were avoided for now. For example, the ‘statezip’ feature may have been kept by assigning a numerical value to each state.

After all of this, the few remaining samples which still had null values were dropped. Feature scaling was also applied. The pre-processing stage left 8 features and 4549 samples.

## Model Training

The data was split randomly into training and testing sets in the ratio 80-20. The model was trained for 200 iterations at a learning rate of 0.01. A vectorized implementation of the gradient descent algorithm was used since it provides better performance than an iterative approach.

The model was set to halt once the difference in the mean squared error values falls below a pre-determined threshold (0.05). Unfortunately, due to the huge mean squared error values, this threshold was never reached. The model runs for whatever number of iterations is given. Tests were performed for up to 20,000 iterations, with no change to this behaviour. Using the SciKit Learn library’s built-in linear regression model seemed to give similar results.

The model was later trained at various learning rates to observe the effect on the overall error for both the training and testing sets.

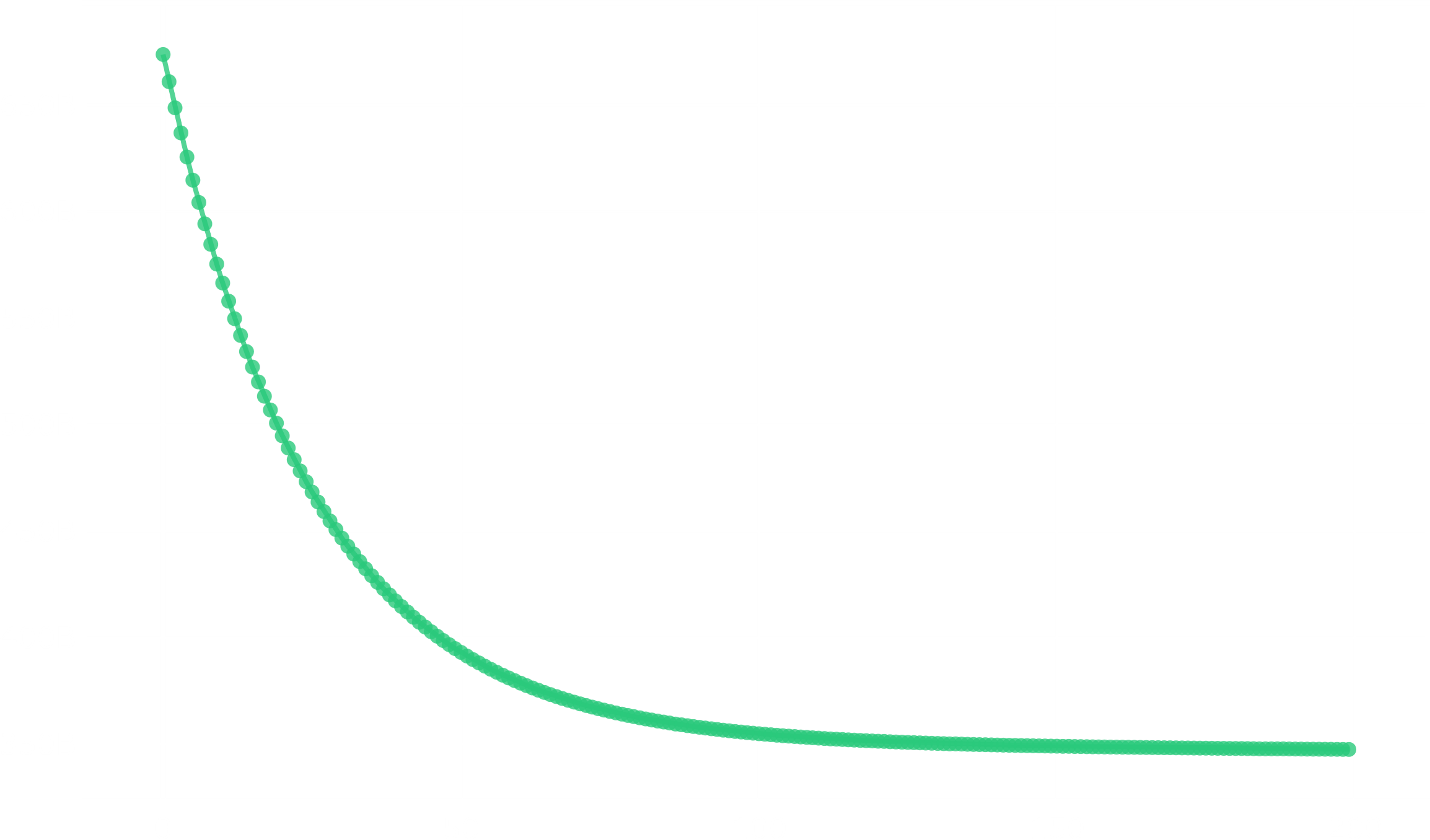
## Feature Selection Mechanism

The feature selection done during the initial stage is described in the data pre-processing section above.

For the last part of the task, the model was trained using the SciKit Learn library. One part of this involved hand-picking features. This was done based on a correlation matrix for the training data. Features that were determined to be highly correlated with the output (price) but had low correlation with each other were retained, while the others were dropped.

## Conclusion

Analysis of the mean squared error shows that the curve flattens around the 150-200 iteration mark.



However, the error itself remains quite high. This seems odd, but since the built-in linear regression library provided by SciKit Learn agrees with the results, the conclusion is that the results are correct.