PREDICTION OF HOME-OWNERS ASSOCIATION FEES USING HOUSE DATA

Padma Vasudevan

AGENDA

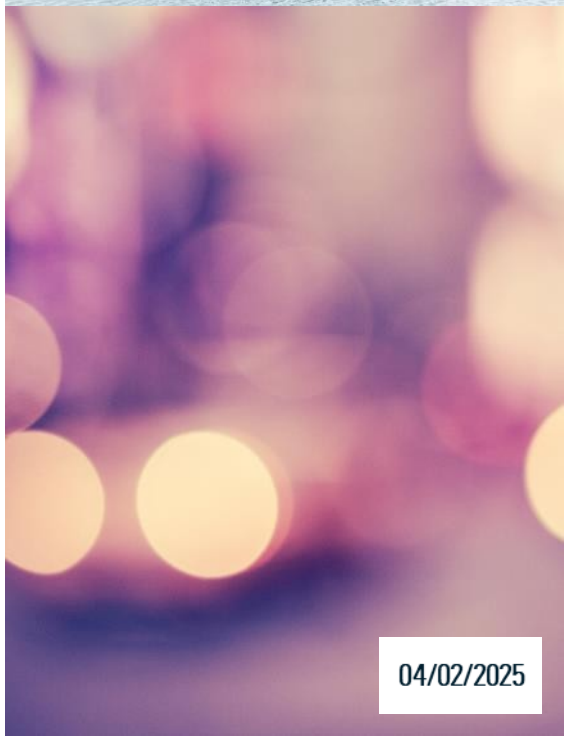
Introduction

Problem Statement

Dataset Overview

Exploratory Data Analysis

Modelling & Performance



INTRODUCTION

Objective

The goal of this project is to develop a predictive model for Home-Owners Association (HOA) fees, helping real estate professionals, buyers, and sellers estimate HOA costs based on property characteristics and location data.

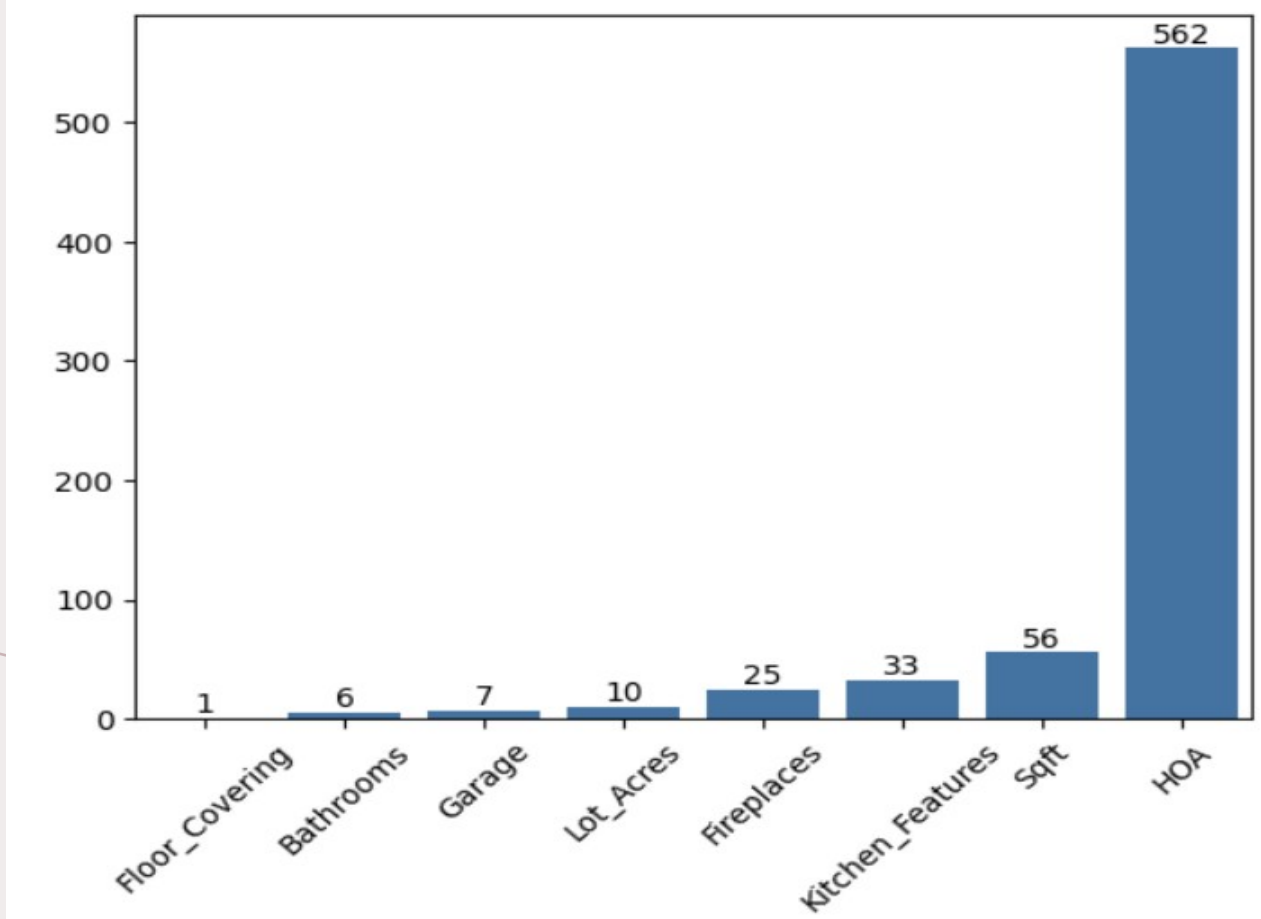
Business Impact

- **Homebuyers & Investors**: Helps buyers factor in HOA fees when assessing the affordability of a property.
- **Real Estate Agents**: Allows agents to set better price expectations for listings and guide clients on cost-effective properties.
- **HOA & Property Managers**: Enables HOAs to compare their fees with similar properties and adjust them based on market trends.

DATASET OVERVIEW

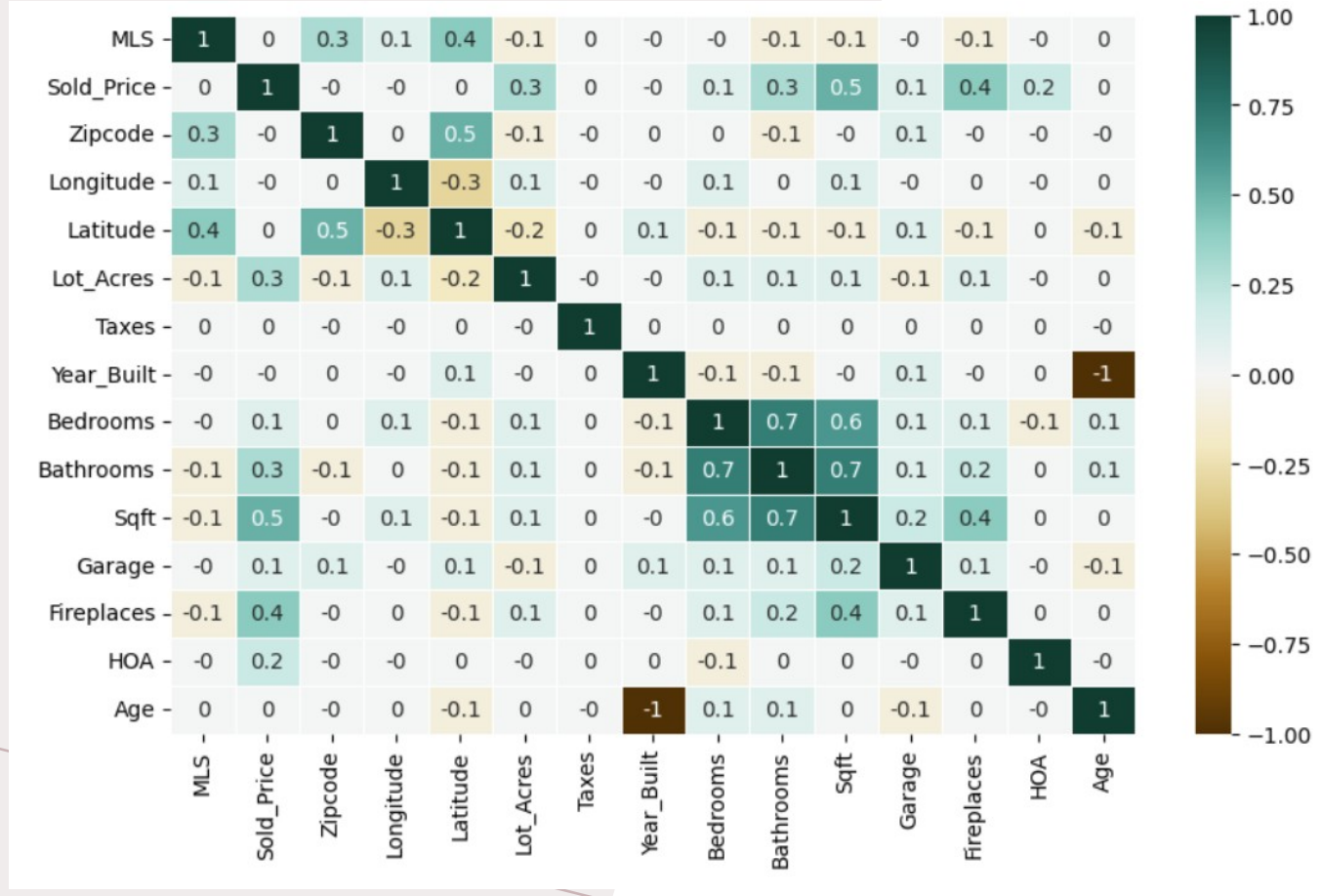
- Description of the Housing dataset:

Shape of the dataset: (5000,16)




```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
cell output actions  total 16 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   MLS               5000 non-null   int64
 1   sold_price        5000 non-null   float64
 2   zipcode           5000 non-null   int64
 3   longitude         5000 non-null   float64
 4   latitude          5000 non-null   float64
 5   lot_acres         4990 non-null   float64
 6   taxes             5000 non-null   float64
 7   year_built        5000 non-null   int64
 8   bedrooms          5000 non-null   int64
 9   bathrooms         4994 non-null   float64
 10  sqrt_ft           4944 non-null   float64
 11  garage            4993 non-null   float64
 12  kitchen_features  4967 non-null   object
 13  fireplaces        4975 non-null   float64
 14  floor_covering    4999 non-null   object
 15  HOA               4438 non-null   float64
```

EXPLORATORY DATA ANALYSIS



From this heatmap, I found that Bedrooms and Bathrooms are multicollinear, HOA is correlated only with Sold_Price.

MODELLING AND PERFORMANCE

KNN Classification
KNN Regression
OLS Regression

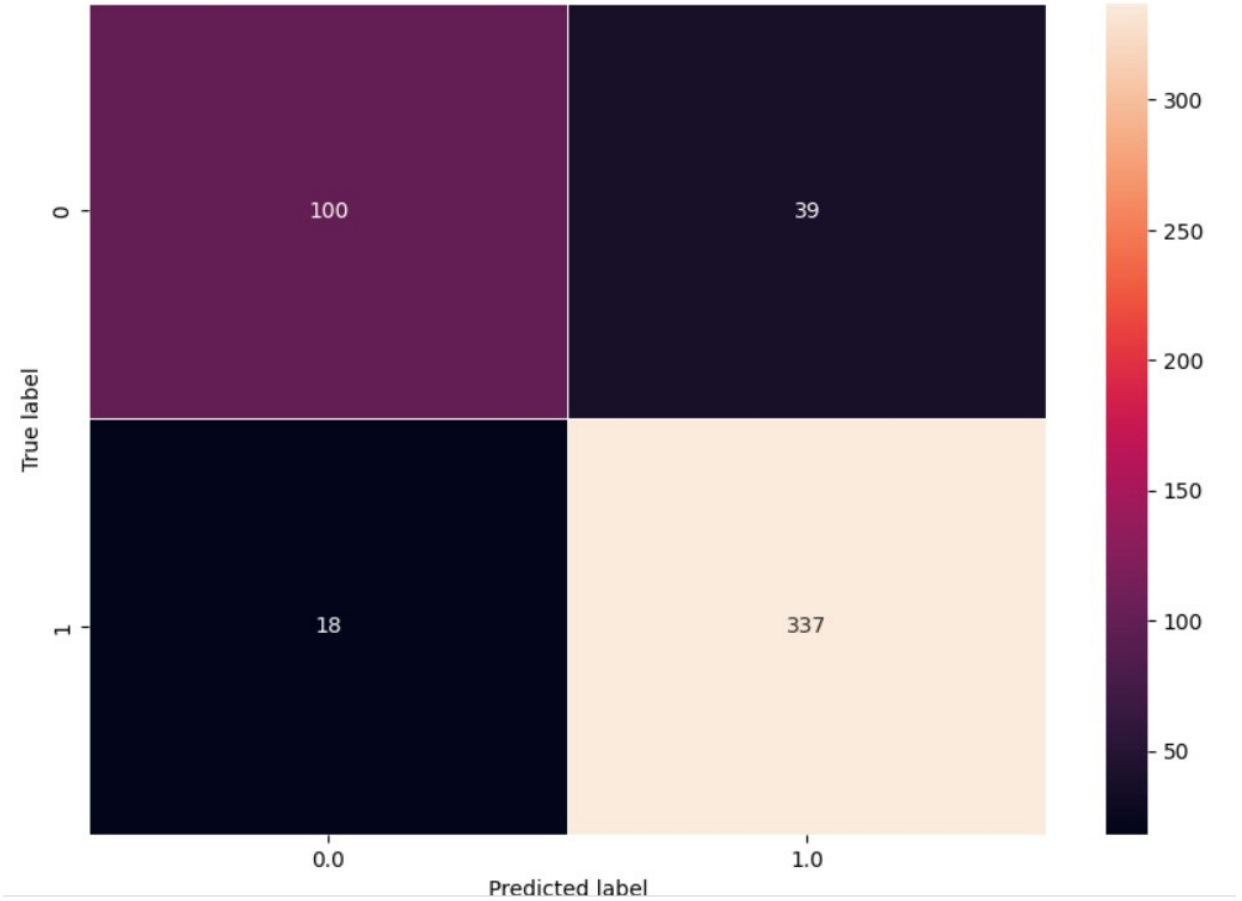
KNN CLASSIFIER

data_Classifier = data[['HOA_bins', 'Latitude', 'Longitude']]

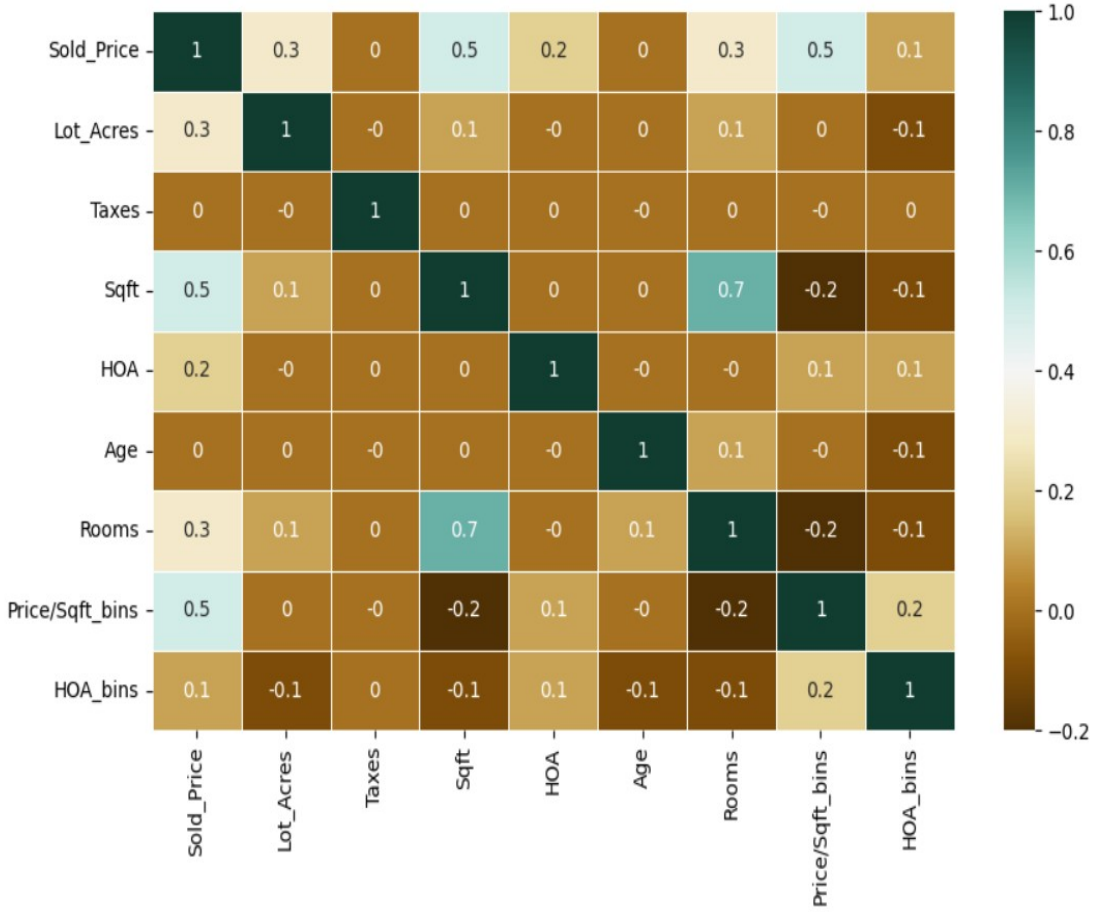
Training data Accuracy: 0.91
Test data Accuracy : 0.88

Using Longitude and Latitude, I classified HOA values as classes in a new HOA_bins column.

True 0s: 100
True 1s: 337
False 0s: 18
False 1s: 39



KNN REGRESSION



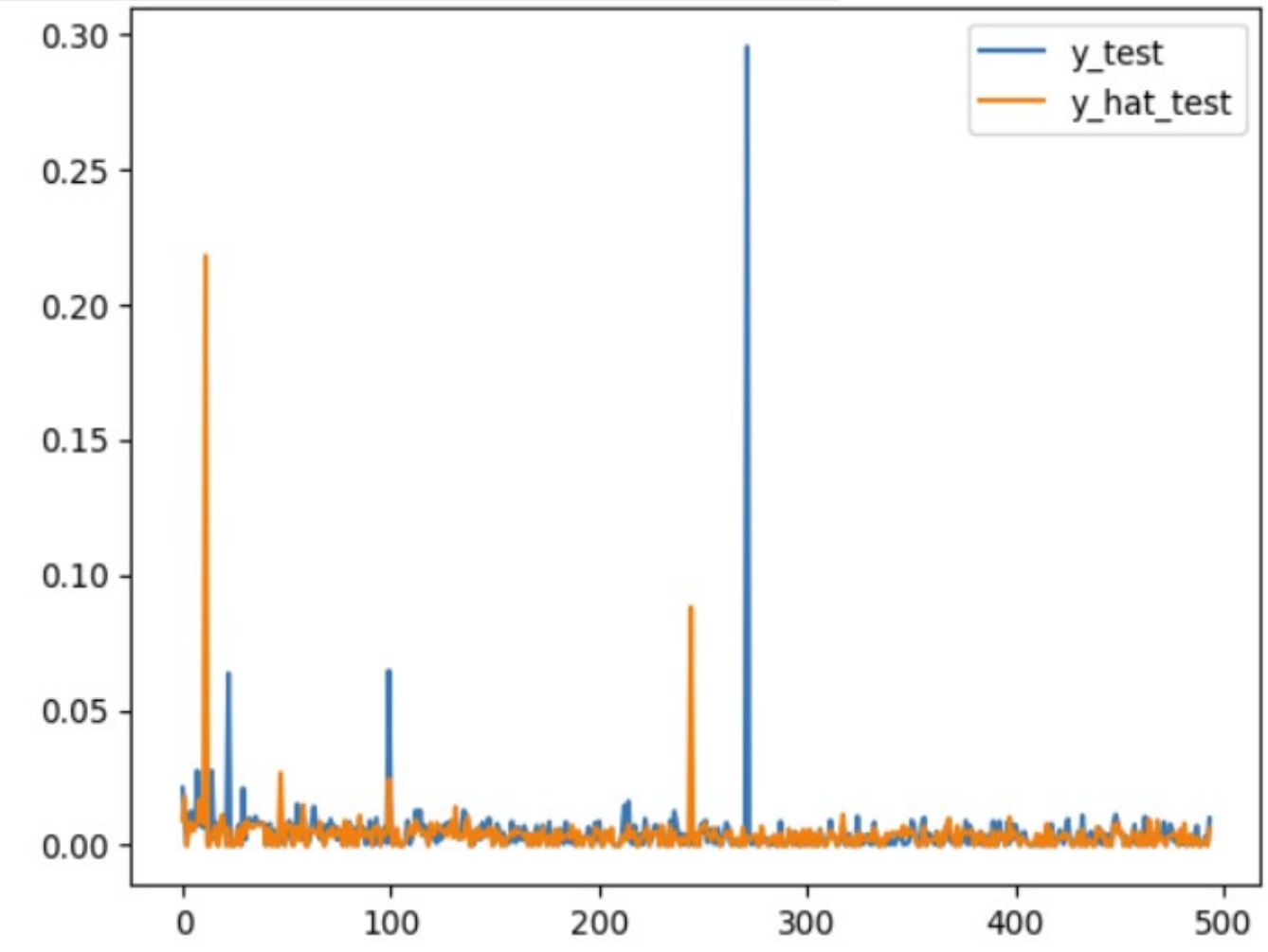
From this Correlation matrix, I found that HOA_bins are correlated moderately with other features.

Features: Sold_Price, Lot_Acres, Taxes, Sqft, Age, Rooms, Price/Sqft_bins, HOA_bins

Target: HOA

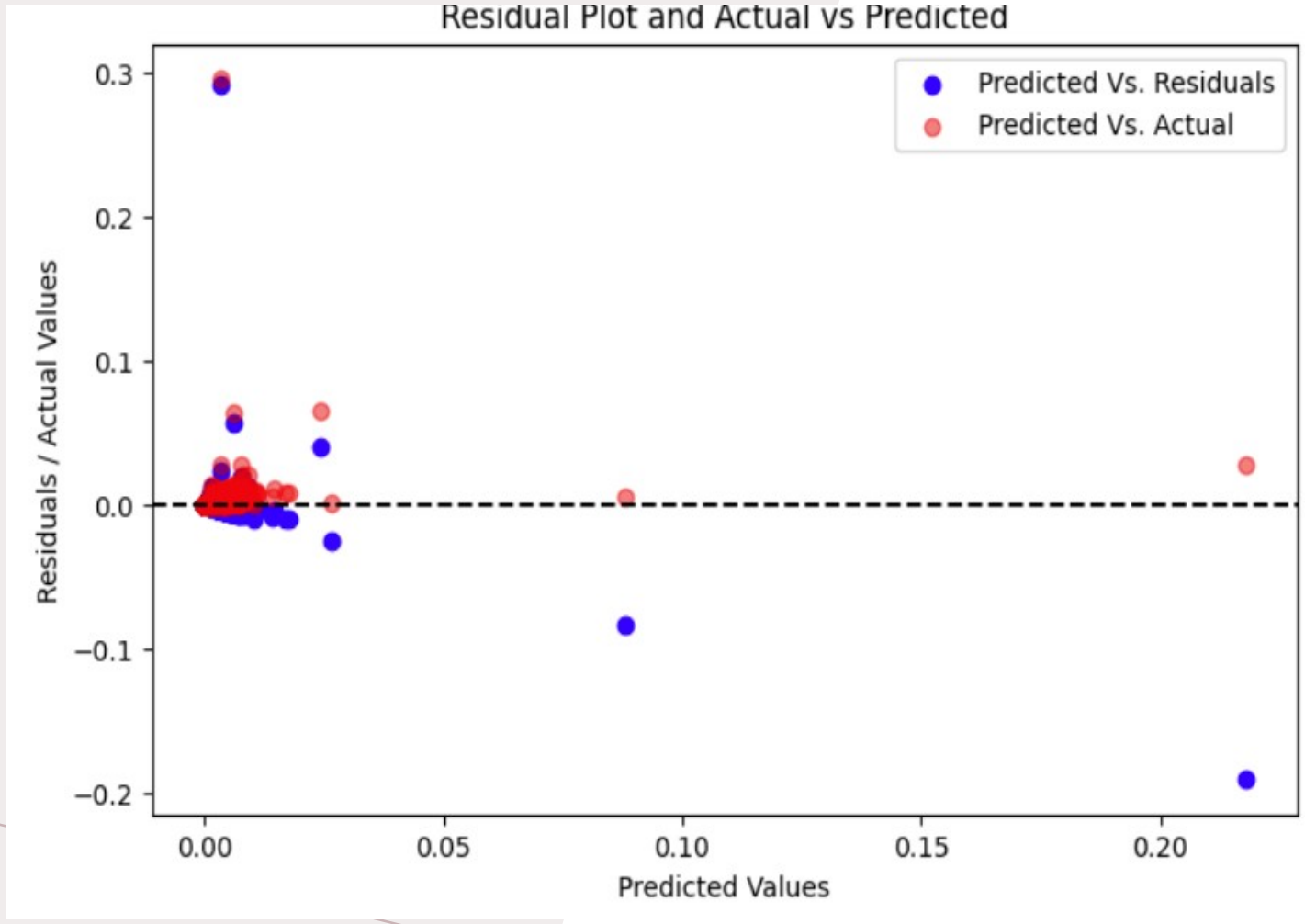
Mean Absolute Error: 703.2039
R^2 Score: -1.3828

KNN REGRESSION



The plot compares the actual values (y_test) with the predicted values (y_hat_test) from the KNN regression model. The two lines are almost closely aligned, so I concluded that the model is performing good.

KNN REGRESSION

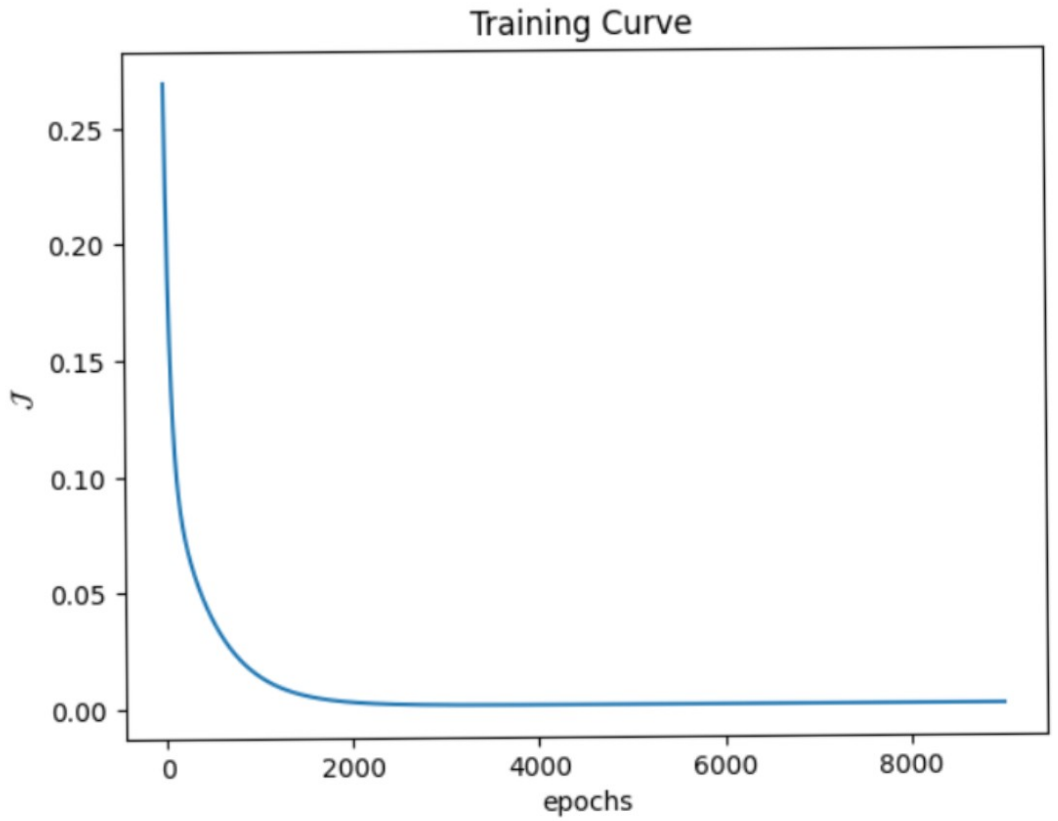


It looks like residuals are randomly scattered around zero. So, the model is well fitted.

Mean Absolute Error: 703.2039

R^2 Score: -1.3828

OLS REGRESSION

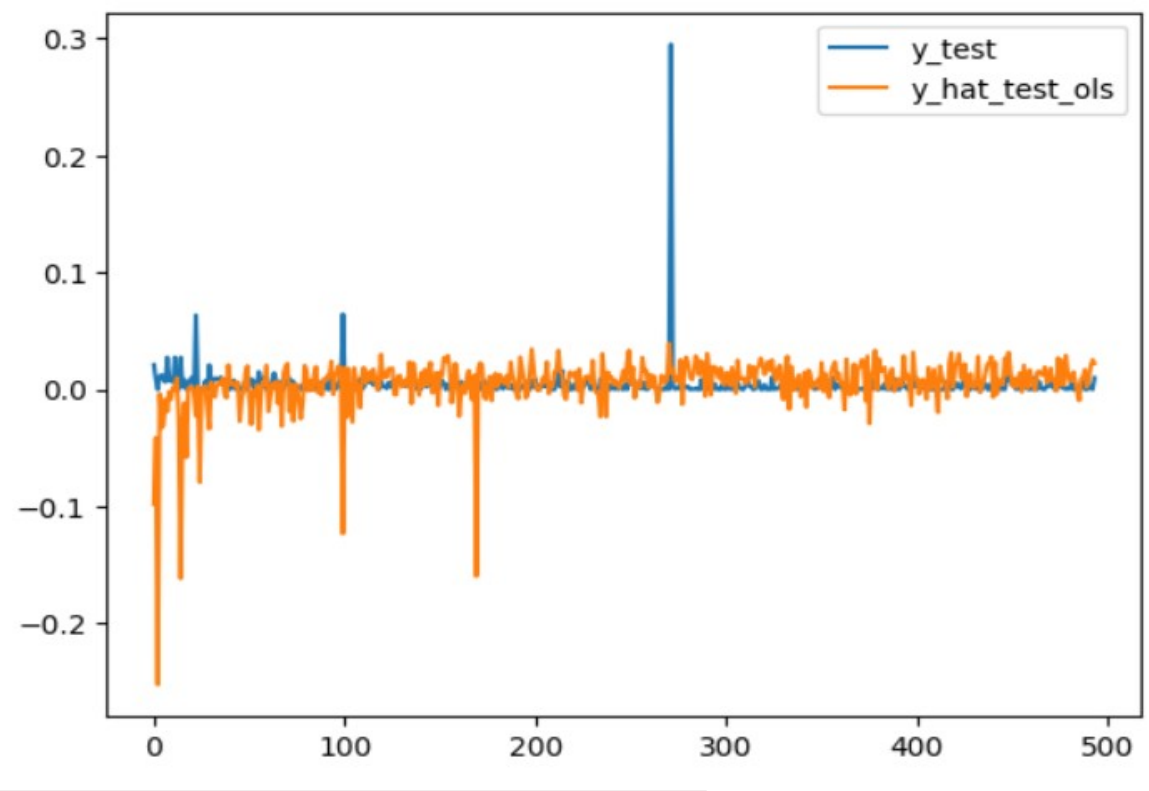


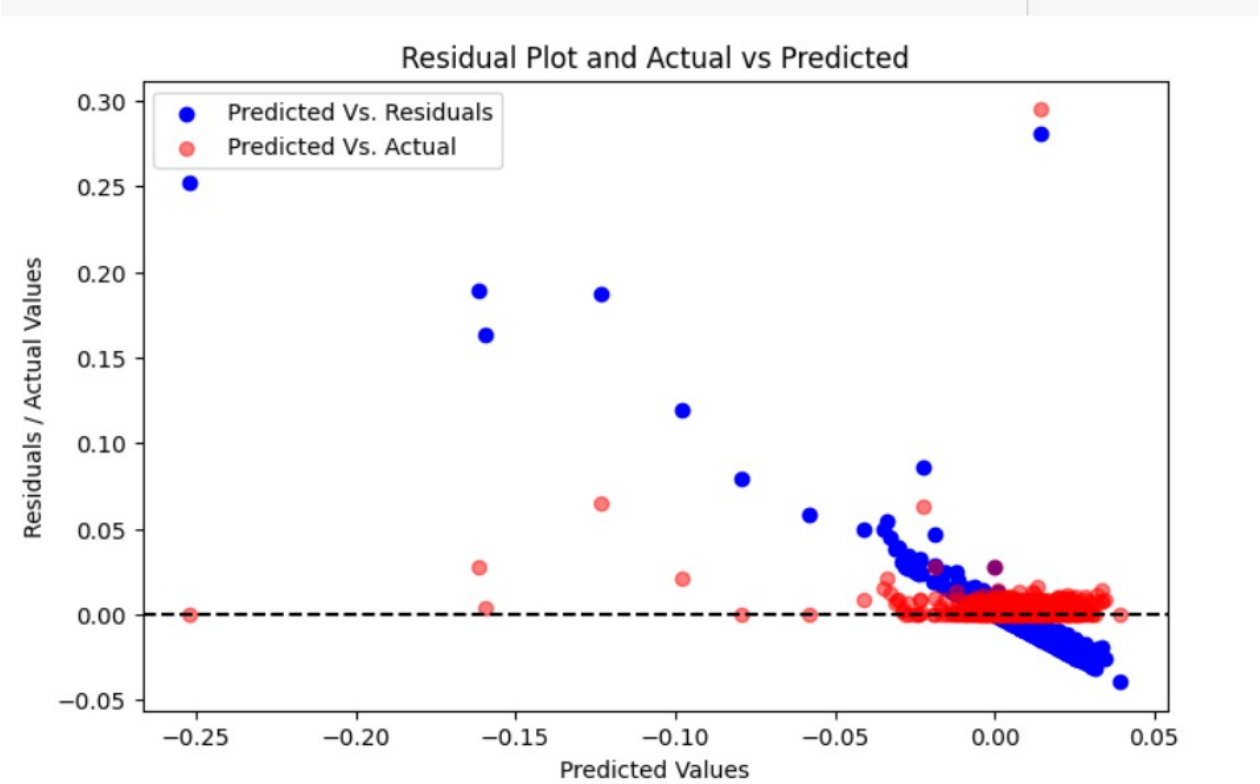
This curve shows the training loss (cost function) over epochs for the MVLinearRegression model. The model uses Gradient Descent to minimize the error over 9,000 epochs with a learning rate (eta) of 0.01.
eta = 1e-2, epochs = 9e3

MAE: 20672042693.5798

R^2 Score: -0.5187

OLS REGRESSION





The plot compares the actual values (y_test) with the predicted values (y_hat_test_ols) from the OLS regression model.
The two lines are not closely aligned, so I concluded the model is not performing well.

It looks like residuals are not scattered around zero. So, the model is not well fitted.

SUMMARY

Predicting HOA fees is more important using Machine Learning algorithms as this fees is invisible for the buyers to check while they decide to fix a house. Accurate HOA prediction is more important in this case.

I chose KNN Regressor Algorithm for this prediction problem as this model gave better performance metrics compared to OLS regression.

THANK YOU