

Report

House Price Prediction

Pranava Kadiyala (pskadiya)

Data and Description

A dataset containing information on house prices and features such as bedrooms, bathrooms, house area, main road connection, etc., was utilized for the house price prediction task. The dataset is accessible on [Kaggle](#). Minimal cleaning and preprocessing were necessary as it didn't contain any missing values or outliers. Categorical variables were transformed into dummy variables to facilitate regression analysis. For the purposes of the analysis, price is the target variable and all other housing related variables are the features.

[Link to Datasheet](#)

Exploratory data analysis

Initial visualizations and summary statistics were generated to get a sense of the spread of the data and most importantly identify any outliers. The histograms in figure 1 show the skewness and spread of continuous variables. There seems to be a left skewness to the data but no apparent outliers.

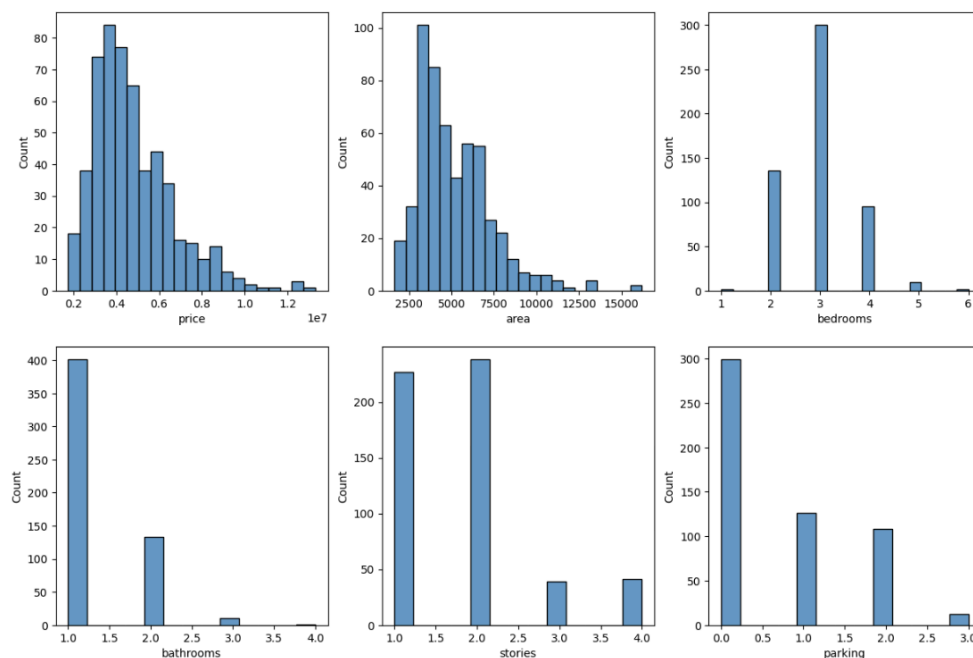


Figure: Histograms of numerical variables

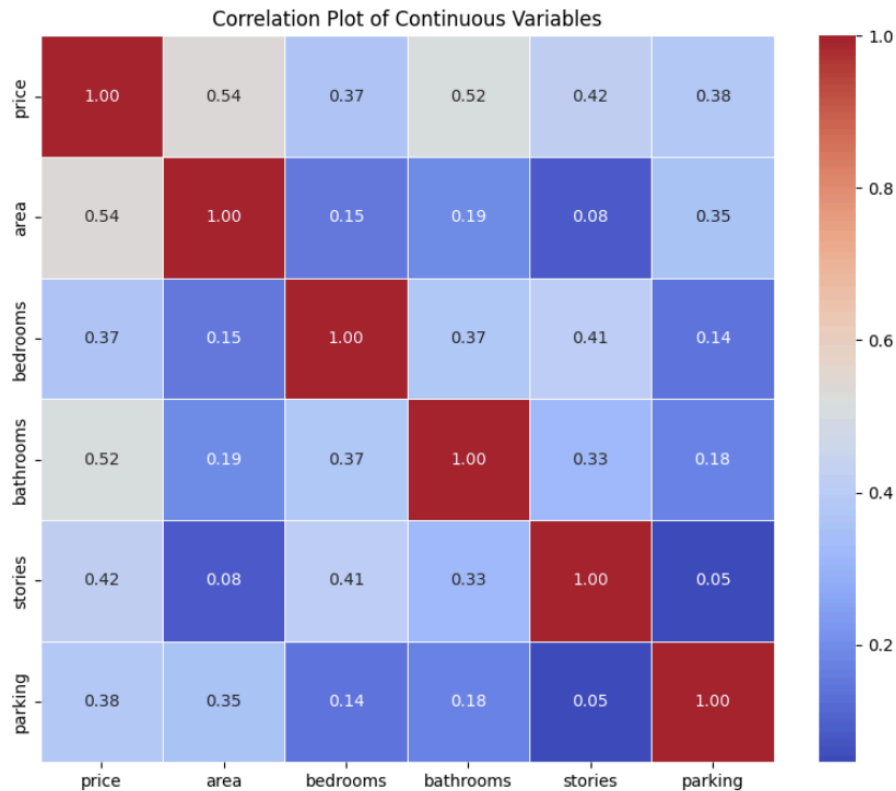


Figure 2: Correlation heatmap

Figure 2 shows the correlation between the numerical variables. Area and bathrooms have the highest positive correlation with the target variable.

[Link to visualizations](#)

Model Building

[Notebook link](#)

Approach

For regression, I decided to use a **Random Forest Regressor**. Post pre-processing of the data, a random forest regressor model was fitted into the data. The metrics were logged using MLflow for tracking and logging. For MLflow set-up instructions outlined in lab1 'MLflow Tracking Quickstart' notebook were followed. The final parameters used for random forest are:

n_estimators: 50, max_depth: 25, max_features: sqrt

Model Comparison

The model was compared with two other models, one from the Zillow's Home Value Prediction (Zestimate) Kaggle Competition and another from the GitHub repository on house price prediction models.

The specific parameters and details of the chosen models are:

1. **Lasso**(alpha=1e-6, normalize=True).
<https://www.kaggle.com/code/fletcherhag/ml-ensemble-scikit-learn-style-ensemble-learning>
2. **XGBRegressor**(n_estimators=900, learning_rate=0.04, n_jobs=4)
<https://github.com/vaibhavvikas/housing-price-predictor/tree/main>

All three models were trained and the experiments were logged in a [registry](#). Learning to work with MLflow was a steep learning curve. Having never worked with it before, getting started was challenging, but having worked on this project alone with three models, I appreciate its utility.

Model Performance and Evaluation

The chosen performance metric or loss function used to compare across models was the mean absolute error (MAE). MAE represents the average of the absolute differences between the predicted values and the actual values. Selecting MAE as the loss function emphasizes minimizing this difference as much as possible to identify the best model.

An interesting insight was the relatively poorer performance of XGBoost. In theory, XGBoost should outperform Lasso, which is a linear model. However, Lasso exhibited the lowest MAE. This underscores the importance for data scientists to begin with simpler models as potential baselines, as they can sometimes perform exceptionally well for specific datasets.

MAE for the three models:

Lasso: 820399.213

Random Forest Regressor: 821864.124

XGBoost: 927498.900

Docker

Unfortunately, I couldn't use Docker to containerize. However, I highlight my steps below and provide context for my errors.

First, I generated a python script with the best model and the best parameters. Similar to the lab, I defined the `generate_plot_metrics` function which returns the metrics of choice - MAE in the [house_prediction.py](#) file. For the model, another [app.py](#) script created for displaying the metrics and generating a graph plotting the actual price against the predicted price. A [Dockerfile](#) and [requirement](#) files were also generated, and the necessary commands from the lab instructions were executed. However, during this process, the following errors were encountered:

```
((FOAI) (base) Pranavas-MacBook-Air:OAI-lab1 pranavakadiyala$ docker build -t myapp .
[+] Building 0.0s (1/1) FINISHED
=> [internal] load build definition from Dockerfile
=> => transferring dockerfile: 2B
ERROR: failed to solve: failed to read dockerfile: open Dockerfile: no such file or directory
((FOAI) (base) Pranavas-MacBook-Air:OAI-lab1 pranavakadiyala$ ls
Dockerfile.txt      README.md           house_price_prediction copy.ipynb
FOAI                Untitled.ipynb      house_price_prediction.ipynb
Housing.csv         __pycache__         mlruns
MLflow Automatic Logging.ipynb  app.py              model.py
MLflow Run Comparison.ipynb    house_prediction.py  requirements.txt
MLflow Tracking Quickstart.ipynb house_price.zip      requirements_docker.txt
((FOAI) (base) Pranavas-MacBook-Air:OAI-lab1 pranavakadiyala$ docker build -t myapp .
[+] Building 0.0s (1/1) FINISHED
=> [internal] load build definition from Dockerfile
=> => transferring dockerfile: 2B
ERROR: failed to solve: failed to read dockerfile: open Dockerfile: no such file or directory
((FOAI) (base) Pranavas-MacBook-Air:OAI-lab1 pranavakadiyala$ docker run -p 80:80 myapp
Unable to find image 'myapp:latest' locally
docker: Error response from daemon: pull access denied for myapp, repository does not exist or may require 'docker login': denied: requested access to the resource is denied.
See 'docker run --help'.
((FOAI) (base) Pranavas-MacBook-Air:OAI-lab1 pranavakadiyala$

((FOAI) (base) Pranavas-MacBook-Air:OAI-lab1 pranavakadiyala$ docker tag myapp pskadiya/myapp:latest
Error response from daemon: No such image: myapp:latest
((FOAI) (base) Pranavas-MacBook-Air:OAI-lab1 pranavakadiyala$ docker pull pskadiya/myapp:latest
Error response from daemon: manifest for pskadiya/myapp:latest not found: manifest unknown: manifest unknown
((FOAI) (base) Pranavas-MacBook-Air:OAI-lab1 pranavakadiyala$ docker run -p 80:80 username/myapp:latest
Unable to find image 'username/myapp:latest' locally
docker: Error response from daemon: pull access denied for username/myapp, repository does not exist or may require 'docker login': denied: requested access to the resource is denied.
See 'docker run --help'.
((FOAI) (base) Pranavas-MacBook-Air:OAI-lab1 pranavakadiyala$
```

I also get this error when trying to run `app.py`.

