Taif University

Predicting House Prices using Machine Learning: A Case Study of Saudi Arabia

Prepared by: *Nasser R. AlSwaidi, Abdulkarim S. Sultan, Khaled S. AlKhamash, Hesham D. AlShehri, Ibrahim S. AlSulimani, Majed A. AlDossary, and Hassan S. AlZahrani* **Supervised by:** *Dr. Abdullah M. Baqasah*

Department of Information Technology, College of Computers & Information Technology (2022/2023)

Abstract

One of the problems that many people face is the search for a home that suits their requirements and financial ability. Among the things that people are looking for are location, number of rooms, area size, etc. Thus, we need a system that can solve the previous issues and make good prediction of house pricing. This project solve house price predictions by designing a machine learning (ML) technology. We first gather a dataset from a real website used for real-estate (aqar.com) and perform exploratory data analysis (EDA). Then, we use the datasets to train and test the model to enhance the accuracy of predictions.

Project Scope

The target audience is people who wish to view real-estate prices or who are ready to purchase an apartment, and the scope is based on anticipating real-estate prices in Saudi Arabia.

Tools

Python Language - for data analysis and model creation.

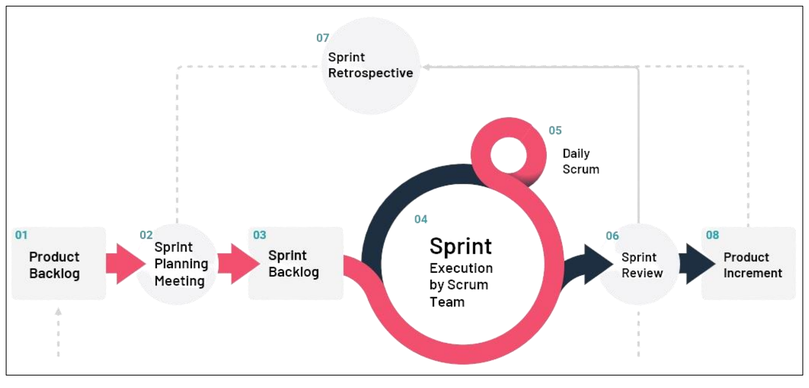
Jupyter Notebook - for combining codes, texts, and visualizations.

Streamlit – a web-based tool for visualizing python outputs and figures.

VSCode – for code editing

Methodology

Agile is firmly rooted in helping developers with cycle testing and rapid prototyping. We chose it for this project due to its flexibility and rapid training and testing of ML model. Team members will be actively involved in their responsibilities thanks to project management. Agile improves communication within a machine learning project, enhancing relationships between team members.



Data Collection

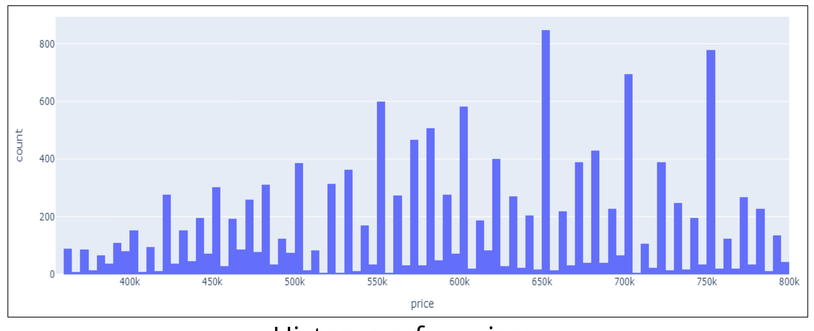
Our dataset is real data that is collected from data available on aqar.fm and scraped using selenium library. Selenium is deployed primarily for web scraping. It's the Python library that makes it easy for dynamic web scraping. We use the scrapper to extract about 20k apartments.

Exploratory Data Analysis (EDA)

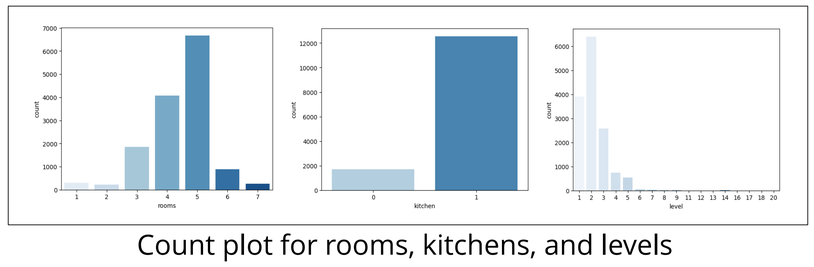
EDA is an approach for data analysis that employs a variety of techniques to:

- Maximize insight into a data set.
- Uncover underlying structure.
- Extract important variables.
- Detect outliers and anomalies.
- Test underlying assumptions.
- Develop parsimonious models.
- Determine optimal factor settings.

The following figures shows EDA for some categorical and numerical features.

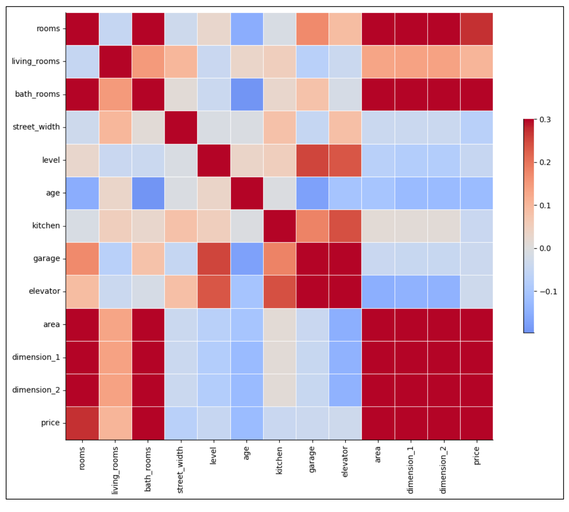


Histogram for prices



Count plot for rooms, kitchens, and levels

And the following Heatmap figure shows the correlation between dataset features.



Model Selection

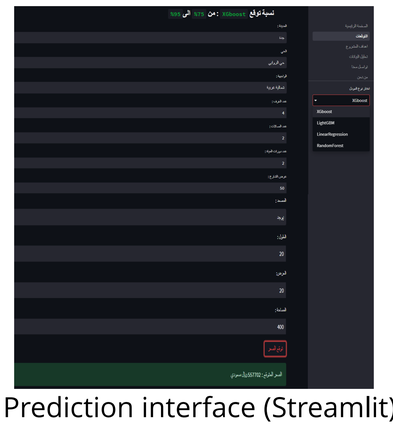
To choose the best home price forecasting model, we consider the following factors: 1) Dataset size, 2) Complexity of the problem, 3) Data quality, and 4) Improving the rating scale. To choose the best model, you should try to train and evaluate several different models using cross validation and compare their performance.

Model Trainings

First, after feeding the models with a large set of training data along with the correct labels and adjusting the model parameters so that it correctly predicts the labels for the majority of the training data. The goal is to find the best prediction ratio. The following figure shows the ratio of each model.

Interfaces

Main user interface (Streamlit)



Prediction interface (Streamlit)

Conclusion

In this project, a price forecasting system was developed from scratch. This technique was developed to assist customers in estimating prices based on details, such as the number of rooms, the location of the area, etc. As a future work we will make use of the “description” attribute to extract useful information and develop in model that predict prices not only for apartments but also for houses and villas.