PREDICTING HOUSE PRICES WITH LINEAR REGRESSION

by Ruba Alnashwan

PROPLEM UNDERSTANDING 1

BACKGROUND

In this project, we will use data from the (Aqar) website, which provides villas for sale and their prices. Our goal from this project is to build a linear regression model that predict house prices.

DATA DESCRIPTION

In this project,We used data for villas offered for sale in the city of Riyadh that were scraped from a real estate website that contains the specifications and location of the villa in addition to its price.

Features	Description	Type
Area	The area of the villa land in square meters	object
Price	villa price	object
Direction	The front of the villa, if the villa is east and west, means that it is on two streets	object
bedrooms	Number of bedrooms in the villa	Int
pathrooms	Number of pathrooms in the villa	float
living_rooms	The number of living rooms in the villa	object
streetwidth	Width of the street on which the villa is located	Int
oldness	New or used villa and how many years of use	Int
views	The number of users who viewed the ad of the villa	object
NP	The location of the villa in the city of Riyadh (North, South, West, East, Central)	Int

2

SCOPE

The scope of this project was the villas for sale in the areas of Riyadh, whether old or new villas. The dataset contains 2099 rows × 10 columns.

TECHNOLOGIES AND LIBRARIES

- Python
- Jupyter Notebook
- LinearRegression
- PCA
- scale
- Lasso
- Ridge
- StandardScaler
- MinMaxScaler
- Pandas
- NumPy
- Seaborn
- Matplotlib
- Requests
- sqlalchemy
- sklearn
- bs4
- PolynomialFeatures

ALGORITHMS 3

FEATURE ENGINEERING

- Principal component analysis (PCA) to reduce the data
- Polynomial Features But the results were not good
- Dummies for Neighborhood Feature ,the model has been greatly improved
- Scaling: MinMaxScaler, StandardScaler
- Feature selection: lasso, ridge

MODELS

The linear regression model was worked on the training data set, where the data was divided into 60% training data ، 20% test data and 20% validation data.

Main Metric used for evaluation : R2, RMSE
Model used: linear Regression

MODEL EVALUATION AND SELECTION

basline

TRAINING SCORE: 0.37746

VALIDATION SCORE: 0.5955

one feature power of 2

TRAINING SCORE: 0.37963

VALIDATION SCORE: 0.59218

dummies

TRAINING SCORE: 0.60967

VALIDATION SCORE: 0.56537

log target

TRAINING SCORE: 0.60571

VALIDATION SCORE: 0.59849

Polynomial Features

TRAINING SCORE: 0.56041

VALIDATION SCORE: 0.63017

PCA

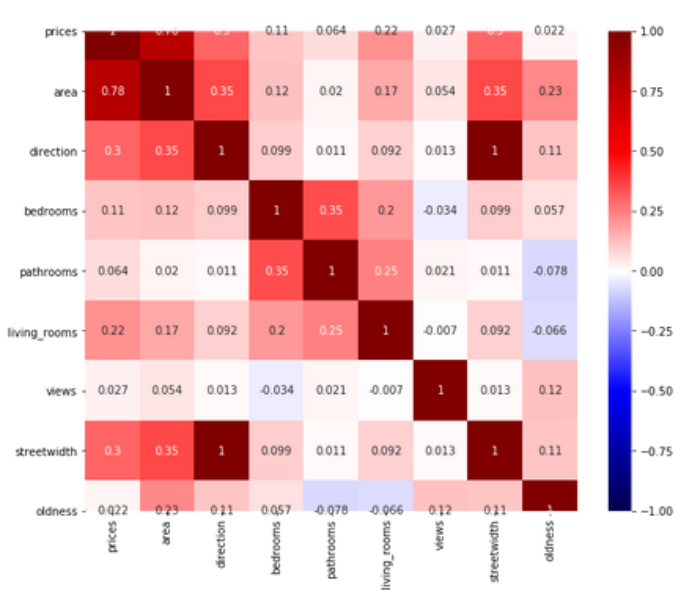
TRAINING SCORE: 0.6097

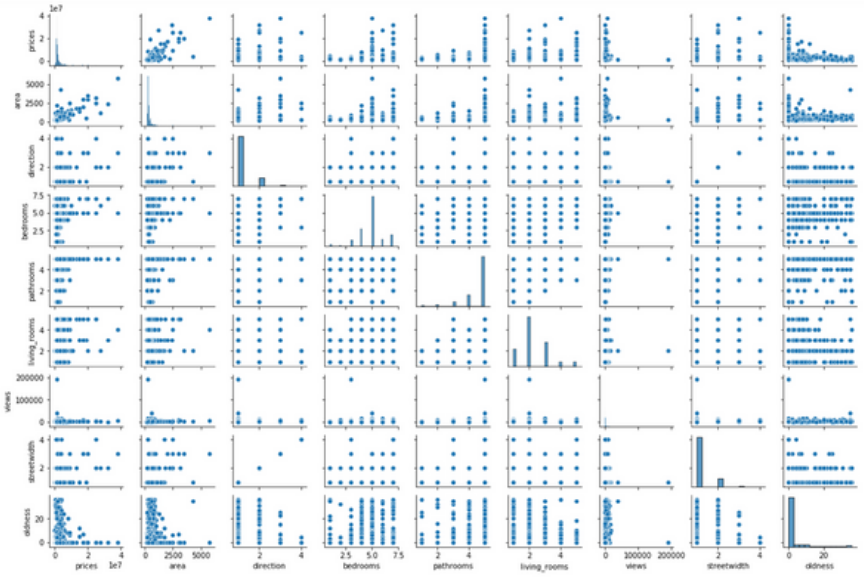
VALIDATION SCORE: -0.10154

Result for the best model:
Training score: 0.60555
Validation score: 0.59771

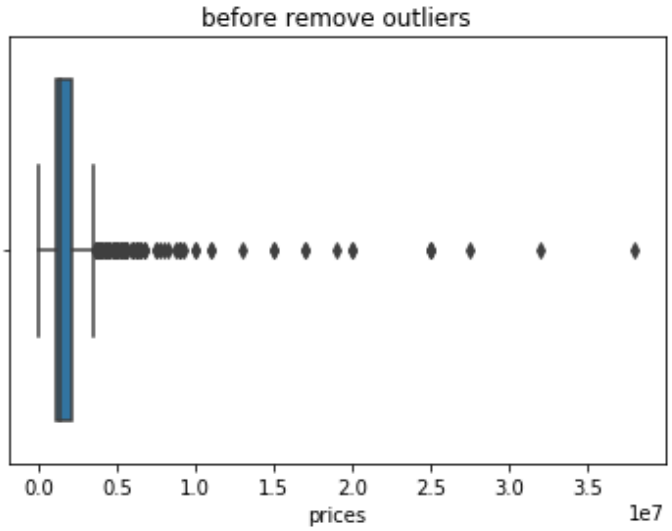
COMMUNICATION 4

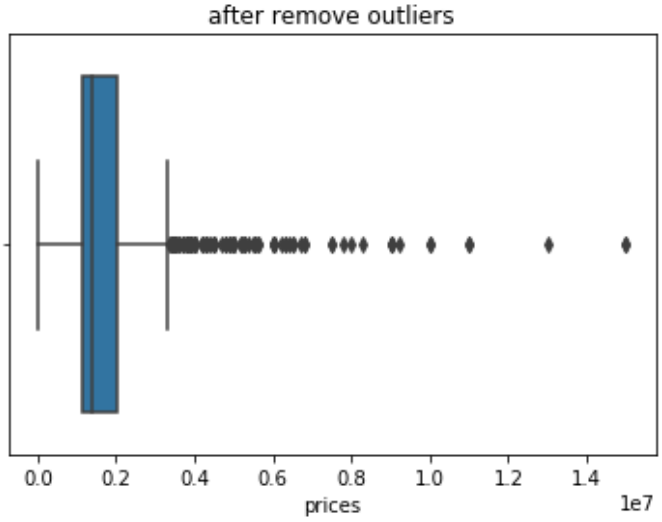
This heat map shows that there is no strong relationship between the dependent variable and features except for the area.





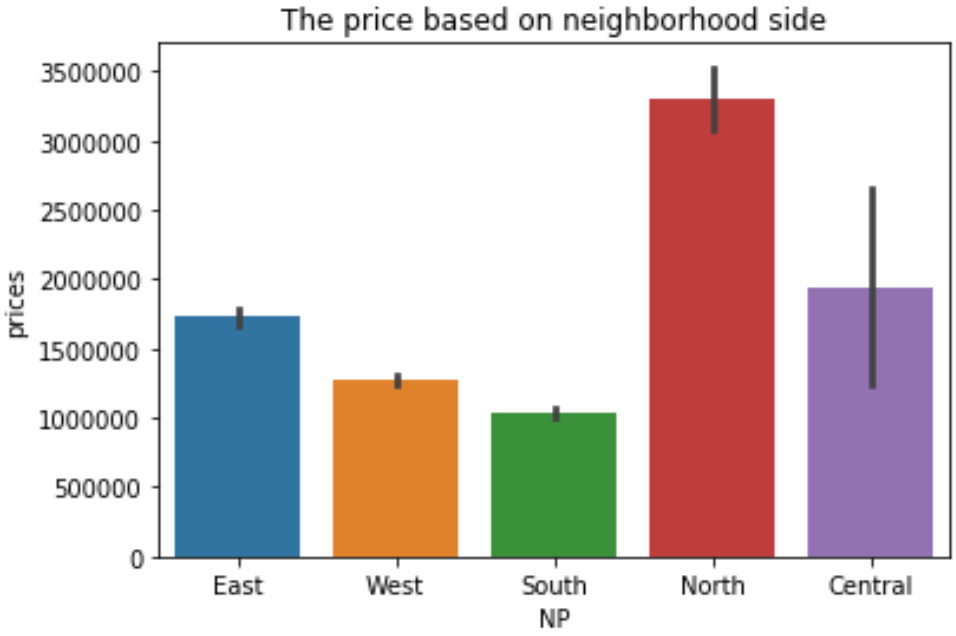
check outliers using EDA and remove





COMMUNICATION 5

The par plot shows the highest prices are located in the north



In this histogram shows the most of the density lies between 1M and 2M, but there appears to be a lot of outliers on the pricier side.