T5 Boot camp

Data science

Project written description

Estate’s predict model

mohammad alajmi

mohammad almalki

Instructor:

Mejdal Al-Qahtani

abstract:

An Estate agent asked to make a system that predict the prices of estates to help the agent to distinguish between the overpriced estate and other way around, as the estate agent said that the data could be collected from different websites, , the estate agent specified some neighborhood is not desired so it should be disqualified from the model due to the low accuracy that it has, the agent also asked to study the data and visualize it to help him to have better understanding about estate market.

In this project we are willing to answer these following questions:

1. What is the best model to build a model that predict the estate market?

2. To study the market and to train the model predict the prices

3. To study the market and to test the model predict the prices and to check for it is accuracy

The main goal of this project:

The main goal of this project is to study the Estates market and to train a model to predict the prices of the estates of and testing it using linear regression models.

Design:

The used data set in this project is derived from AQAR website, the data collected using web scraping library called beautiful soap, it is about the prices of estates and other information about the estate, the data has about 40 thousand rows before the cleaning and about 7 columns.

Data:

The columns describe a name, area in Square and the neighborhood within Riyadh. The columns also describe the number of bedrooms, the number of bathrooms and the price of the Estate.

Algorithms:

in this project we used different models to achieve this goal mentioned as follows:

1. linear regression
2. ridge regression
3. lasso regression
4. polynomial regression
5. random forest

The error values measured in five methods:

1. MAE
2. MSE
3. RMSE
4. R square
5. Cross validation

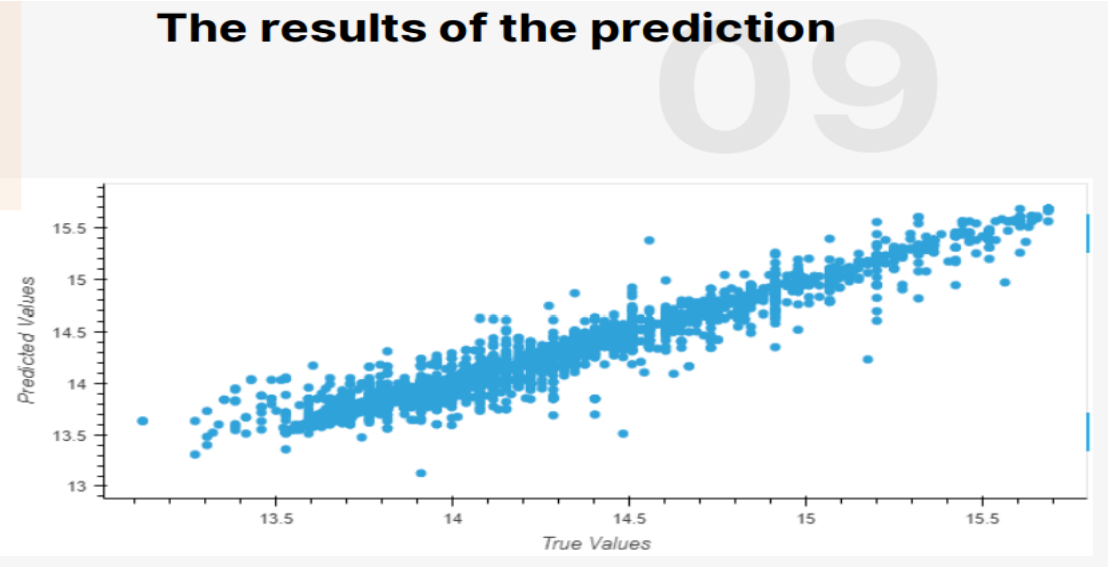
The used tools:

I have used different tools to achieve this goal such as:

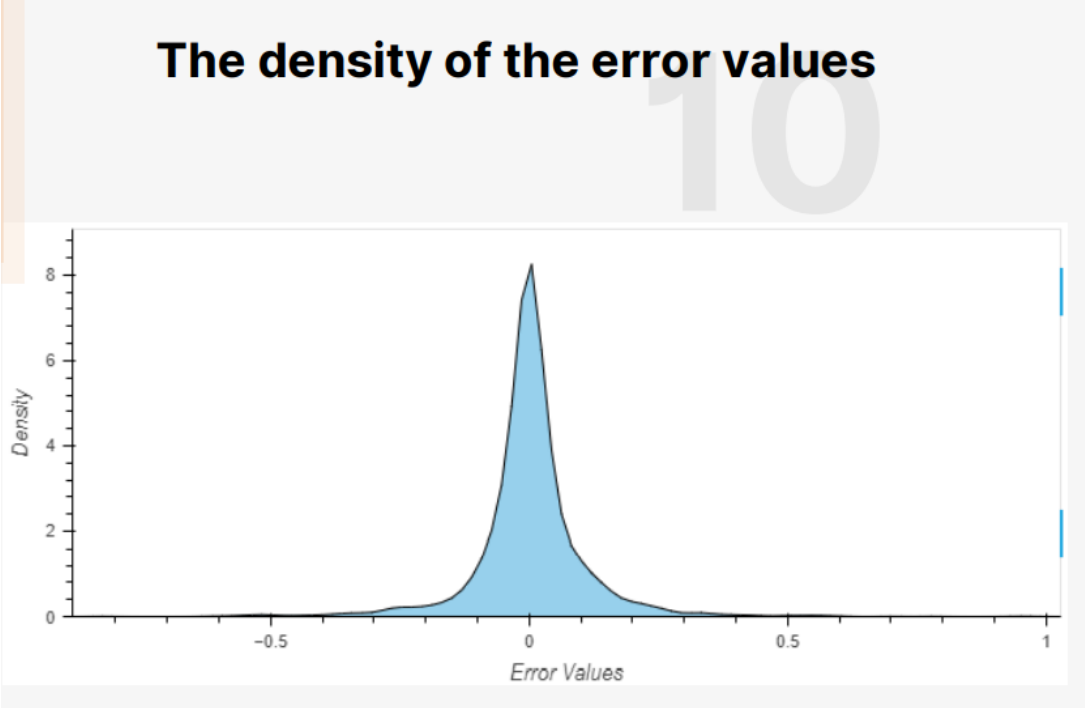
1- Python programming language
2- Numby library
3- Pandas' library
4- Matplotlib library
5- SeaBorn
6- Sklearn
7- BeautifulSoup
8- Request library

Communication:

In figure 1 the used plot is called scatter plot; it represents the results of the prediction.



In figure 2 the used plot is called Hv plot, it represents the densities of the error values.



In figure 3 the table summarizes the results of the regression models with every measurement method.

	Model	MAE	MSE	RMSE	R2 Square	Cross Validation
0	Linear Regression	0.098681	0.020143	0.141925	0.918249	0.912443
1	Ridge Regression	0.098787	0.020157	0.141975	0.918192	0.912469
2	Lasso Regression	0.261338	0.110437	0.332320	0.551785	-0.023750
3	Polynomail Regression	0.096210	0.019465	0.139515	0.921002	0.000000
4	Random Forest	0.061577	0.011478	0.107138	0.953414	0.952802