Model Deployment

Data Cleaning and Building a model

dφ

Democratizing Data Science Learning

DPhi

Learning Objectives

Data Cleaning

Feature Engineering

Outlier Detection and Removal

Building a ML Model

Getting Started

Dataset

We'll be working with the house price data of a city in India - Bengaluru (also known as Bangalore).

https://www.kaggle.com/amitabhajoy/bengaluru-house-price-data

Instead of downloading it, you can access it directly with this link:

https://raw.githubusercontent.com/dphi-official/Datasets/master/Bengaluru_House_Data.csv

This dataset is unstructured, messy and somewhat resembles the kind of data you'll get in a real-world setup.

DPhi

Notebook Link

We'll be following along the series of the whole Data Science pipeline by the instructor. It is interesting to note the kind of functions he is creating to tackle Data Cleaning.

The code from all the videos of the series can be found in a single notebook:

https://github.com/codebasics/py/blob/master/DataScience/BangloreHomePrices/model/banglore_home_prices_final.ipynb

Make sure you spend enough time on each step to understand what is happening.

Data Cleaning

Data Cleaning

In the video, you'll observe that the instructor has removed some columns(mostly categorical or text-based).

Instead of doing that, it's always better to perform feature selection techniques to determine the important as well as non-important features.

Also, instead of removing the rows with missing values from the DataFrame, they can instead be substituted using the handling missing values techniques we discussed earlier.

You'll also find some functions to handle data cleaning of different features and that's quite a feasible way to process all rows of the DataFrame.

Data Cleaning cntd. + Feature Engineering

DPhi

Outlier Detection and Removal

DPhi

Building a ML Model

DPhi

At the end of building a model, he saved it using pickle.
He also saved the location and column information to a JSON file

Let’s understand what it is in-depth along with its other alternatives in the next unit.

Slide Download Link

You can download this unit from the below link:

https://docs.google.com/presentation/d/1pSsaUc7y99g1-kuahTNs38ZsQG6YPuFMF0KpzJjk-00/edit?usp=sharing

That’s it for this unit. Thank you!

Feel free to post any queries on Discuss.

DPhi