Recommender Systems Final Lab

Due Date: 23:59 June22, 2024

You have a dataset about housing prices in various regions of Taiwan. We need to observe some important information from the data, learn how to handle missing values, and finally build a machine learning model using decision trees.

About data

T-Brain Machine Learning Competition

歡迎您使用「T-Brain AI實戰吧平台服務」，一旦您進入「T-Brain AI實戰吧平台服務」即表示您同意遵守下列條款及細則，與任何不定時提供給您的政策、準則及更新條款及相關比賽規範，包括 (但不限於) 服務政策和法律聲明 (下稱「條款」)如下：

https://tbrain.trendmicro.com.tw/Competitions/Details/30

- train.csv

ID	縣市	鄉鎮市區	路名	土地面積	使用分區	移轉層次	總樓層數	主要用途	主要建材	...	建物面積	車位面積	車位個數	橫坐標	縱坐標	備註	主建物面積	陽台面積	附屬建物面積	單價
TR-1	台北市	大安區	敦化南路二段	-0.256716	None	11	11	住家用	鋼筋混凝土造	...	-0.174154	-0.819326	0.0	305266	2768378	NaN	0.393926	0.183700	-0.438452	4.627714
TR-2	台北市	萬華區	水源路	0.100134	None	7	12	住家用	鋼筋混凝土造	...	0.314204	-0.819326	0.0	300677	2767990	NaN	-0.316131	0.608577	-0.438452	1.887258
TR-3	高雄市	鳳山區	北忠街	0.181921	None	10	15	集合住宅	其他	...	0.423366	0.161624	1.0	184815	2504666	NaN	-0.098871	-0.360620	1.525881	1.489072
TR-4	新北市	新莊區	福前街	0.085594	None	9	14	集合住宅	鋼筋混凝土造	...	0.164249	0.524653	1.0	296653	2772355	NaN	-0.071147	0.315088	0.231984	2.051217
TR-5	新北市	板橋區	文化路一段	-0.938116	None	41	43	住家用	鋼骨造	...	0.985839	0.532377	1.0	297377	2768472	NaN	0.791954	1.719400	-0.438452	3.269198

- test.csv

ID	縣市	鄉鎮市區	路名	土地面積	使用分區	移轉層次	總樓層數	主要用途	主要建材	...	屋齡	建物面積	車位面積	車位個數	橫坐標	縱坐標	備註	主建物面積	陽台面積	附屬建物面積
TR-10001	高雄市	三民區	金山路	-0.678289	None	2	14	集合住宅	鋼筋混凝土造	...	6.500000	-0.665958	0.540101	1.0	180114	2507034	NaN	-0.937361	-0.937361	-0.124586
TR-10002	新北市	蘆洲區	長安街	-0.705230	None	6	6	住家用	鋼筋混凝土造	...	25.416667	-0.938290	-0.819326	0.0	296364	2775200	NaN	-1.007235	-0.445937	-0.288992
TR-10003	新北市	鶯歌區	龍七路	0.359937	None	4	9	集合住宅	鋼筋混凝土造	...	0.666667	0.188954	0.663685	1.0	283889	2762718	NaN	-0.133101	-0.754784	0.795662
TR-10004	高雄市	楠梓區	大學十一街	-0.206101	None	5	15	集合住宅	鋼筋混凝土造	...	0.416667	0.122308	0.614252	1.0	176785	2514469	NaN	-0.297460	-0.065425	-0.039180
TR-10005	台北市	中山區	民權東路一段	0.140943	None	12	16	住家用	鋼骨造	...	6.083333	2.839874	3.620439	2.0	302818	2772898	NaN	2.269778	1.675036	0.594958

- test_submit_template.csv

ID	predicted_price
TR-10001	
TR-10002	
TR-10003	

Q1: Analyze data by matplotlib

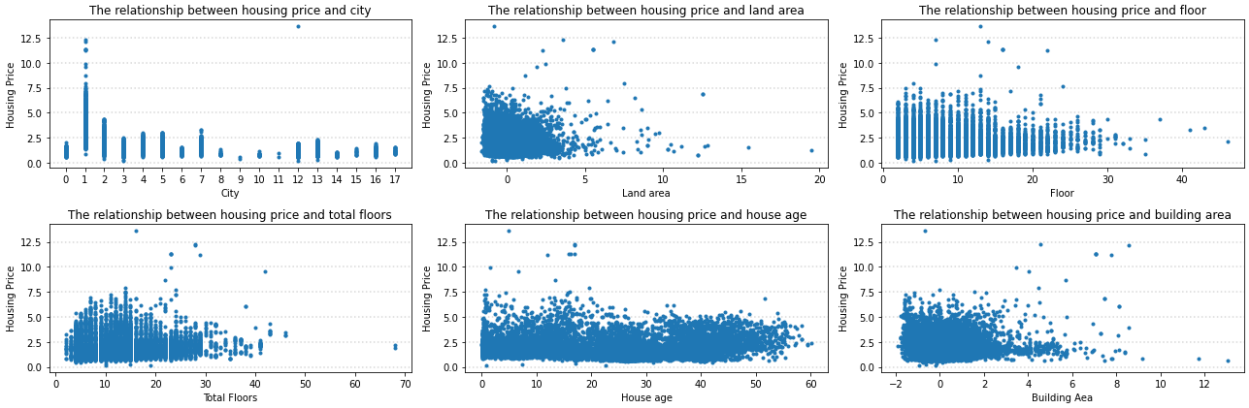
> ⚠ You need to user at least two type of charts

Q-1: Show the correlation between '縣市' and '單價'.

Q-2: Show the correlation between '主要用途' and '單價'.

Q-3: Choose two variable by yourself.

Example



Q2: Build tree-based ML model by scikit-learn

Q-1: How to preprocess and deal with missing value

Q-2: Build a tree-based ML model

Q-3: Cross-Validation Results for [cv = 5]

Q-4: Calculate results for test.csv and make conclusion