哈尔滨工业大学（深圳）

大数据导论大作业报告

题目：基于随机森林回归预测的房价预测

姓　　名　　陈致远

学　　号　　200111521

报告日期　　2022-10-28

一、 实验目的

本次实验主要在题目中给出的各种类型的房价数据进行预处理，然后利用机器学习中的一种预测算法对其进行预测，对模型进行评估，在训练好的模型上对其进行预测房价。

二、 实验内容分析

①数据预处理

在查看 train.csv 文件时，发现“公寓号”一列存在着大量的空缺，认为其对模型没有直接的帮助，可以直接删除。

并且房价跟出售日期没有直接的关系，考虑直接删去，不再保留。

可以认为修建年份对房价有比较大的影响，在查看修建年份中，发现有 3444 条为 0 的扰动数据，将其所在行全部删除，避免对训练模型造成干扰。

查看到地役权几乎全部空白，予以删除。

删除当前税收级别为空白的两行数据。

删除出售价格小于 1000 的行。

考虑到测试集里面土地平方英尺和总平方英尺同样存在着空缺数据，所以将训练集里面的两个指标所在列全部删除。

考虑到邮箱编码可以由所属区域和社区确定，所以删除邮箱编码一列。

删除税收级别为空白的 439 条数据。

②进行相关编码操作，

对社区的编码：

AIRPORT LA GUARDIA	1	DONGAN HILLS-OLD TOWN	61	INWOOD	121	PELHAM PARKWAY NORTH	181	WHITESTONE	241
ALPHABET CITY	2	DOUGLASTON	62	JACKSON HEIGHTS	122	PELHAM PARKWAY SOUTH	182	WILLIAMSBRIDGE	242
ANNADALE	3	DOWNTOWN-FULTON FERRY	63	JAMAICA	123	PLEASANT PLAINS	183	WILLIAMSBURG-CENTRAL	243
ARDEN HEIGHTS	4	DOWNTOWN-FULTON MALL	64	JAMAICA BAY	124	PORT IVORY	184	WILLIAMSBURG-EAST	244
ARROCHAR	5	DOWNTOWN-METROTECH	65	JAMAICA ESTATES	125	PORT RICHMOND	185	WILLIAMSBURG-NORTH	245
ARROCHAR-SHORE ACRES	6	DYKER HEIGHTS	66	JAMAICA HILLS	126	PRINCES BAY	186	WILLIAMSBURG-SOUTH	246
ARVERNE	7	EAST ELMHURST	67	JAVITS CENTER	127	PROSPECT HEIGHTS	187	WILLOWBROOK	247
ASTORIA	8	EAST NEW YORK	68	KENSINGTON	128	QUEENS VILLAGE	188	WINDSOR TERRACE	248
BATH BEACH	9	EAST RIVER	69	KEW GARDENS	129	RED HOOK	189	WOODHAVEN	249

BATHGATE	10	EAST TREMONT	70	KINGSBRIDGE HTS/UNIV HTS	130	REGO PARK	190	WOODLAWN	250
BAY RIDGE	11	EAST VILLAGE	71	KINGSBRIDGE /JEROME PARK	131	RICHMOND HILL	191	WOODROW	251
BAYCHESTER	12	ELMHURST	72	KIPS BAY	132	RICHMONDTOWN	192	WOODSIDE	252
BAYSIDE	13	ELTINGVILLE	73	LAURELTON	133	RICHMONDTOWN-LIGHTHS HILL	193	WYCKOFF HEIGHTS	253
BEDFORD PARK/NORWOOD	14	EMERSON HILL	74	LITTLE ITALY	134	RIDGEWOOD	194		
BEDFORD STUYVESANT	15	FAR ROCKAWAY	75	LITTLE NECK	135	RIVERDALE	195		
BEECHHURST	16	FASHION	76	LIVINGSTON	136	ROCKAWAY PARK	196		
BELLE HARBOR	17	FIELDSTON	77	LONG ISLAND CITY	137	ROOSEVELT ISLAND	197		
BELLEROSE	18	FINANCIAL	78	LOWER EAST SIDE	138	ROSEBANK	198		
BELMONT	19	FLATBUSH-CENTRAL	79	MADISON	139	ROSEDALE	199		
BENSONHURST	20	FLATBUSH-EAST	80	MANHATTAN BEACH	140	ROSSVILLE	200		
BERGEN BEACH	21	FLATBUSH-LEFFERTS GARDEN	81	MANHATTAN VALLEY	141	ROSSVILLE-CHARLESTON	201		
BLOOMFIELD	22	FLATBUSH-NORTH	82	MANOR HEIGHTS	142	ROSSVILLE-PORT MOBIL	202		
BOERUM HILL	23	FLATIRON	83	MARINE PARK	143	ROSSVILLE-RICHMOND VALLEY	203		
BOROUGH PARK	24	FLATLANDS	84	MARINERS HARBOR	144	SCHUYLERVILLE/PELHAM BAY	204		
BRIARWOOD	25	FLORAL PARK	85	MASPETH	145	SEAGATE	205		
BRIGHTON BEACH	26	FLUSHING MEADOW PARK	86	MELROSE/CONCOURSE	146	SHEEPSHEAD BAY	206		
BROAD CHANNEL	27	FLUSHING-NORTH	87	MIDDLE VILLAGE	147	SILVER LAKE	207		
BRONXDALE	28	FLUSHING-SOUTH	88	MIDLAND BEACH	148	SO. JAMAICA-BAISLEY PARK	208		
BROOKLYN HEIGHTS	29	FORDHAM	89	MIDTOWN CBD	149	SOHO	209		

BROWNSVILLE	30	FOREST HILLS	90	MIDTOWN EAST	150	SOUNDVIEW	210
BULLS HEAD	31	FORT GREENE	91	MIDTOWN WEST	151	SOUTH BEACH	211
BUSH TERMINAL	32	FRESH KILLS	92	MIDWOOD	152	SOUTH JAMAICA	212
BUSHWICK	33	FRESH MEADOWS	93	MILL BASIN	153	SOUTH OZONE PARK	213
CAMBRIA HEIGHTS	34	GERRITSEN BEACH	94	MORNINGSIDE HEIGHTS	154	SOUTHBRIDGE	214
CANARSIE	35	GLEN OAKS	95	MORRIS PARK/VAN NEST	155	SPRING CREEK	215
CARROLL GARDENS	36	GLENDALE	96	MORRISANIA /LONGWOOD	156	SPRINGFIELD GARDENS	216
CASTLE HILL/UNIONPORT	37	GOWANUS	97	MOTT HAVEN/PORT MORRIS	157	ST. ALBANS	217
CASTLETON CORNERS	38	GRAMERCY	98	MOUNT HOPE/MOUNT EDEN	158	STAPLETON	218
CHELSEA	39	GRANT CITY	99	MURRAY HILL	159	STAPLETON-CLIFTON	219
CHINATOWN	40	GRASMERE	100	NAVY YARD	160	SUNNYSIDE	220
CITY ISLAND	41	GRAVESEND	101	NEPONSIT	161	SUNSET PARK	221
CITY ISLAND-PELHAM STRIP	42	GREAT KILLS	102	NEW BRIGHTON	162	THROGS NECK	222
CIVIC CENTER	43	GREAT KILLS-BAY TERRACE	103	NEW BRIGHTON-ST. GEORGE	163	TODT HILL	223
CLINTON	44	GREENPOINT	104	NEW DORP	164	TOMPKINSVILLE	224
CLINTON HILL	45	GREENWICH VILLAGE-CENTRAL	105	NEW DORP-BEACH	165	TOTTENVILLE	225

CLOVE LAKES	46	GREENWICH VILLAGE-WEST	106	NEW DORP-HEIGHTS	166	TRAVIS	226
COBBLE HILL	47	GRYMES HILL	107	NEW SPRINGVILLE	167	TRIBECA	227
COBBLE HILL-WEST	48	HAMMELS	108	OAKLAND GARDENS	168	UPPER EAST SIDE (59-79)	228
COLLEGE POINT	49	HARLEM-CENTRAL	109	OAKWOOD	169	UPPER EAST SIDE (79-96)	229
CONCORD	50	HARLEM-EAST	110	OAKWOOD-BEACH	170	UPPER EAST SIDE (96-110)	230
CONCORD-FOX HILLS	51	HARLEM-UPPER	111	OCEAN HILL	171	UPPER WEST SIDE (59-79)	231
CONEY ISLAND	52	HARLEM-WEST	112	OCEAN PARKWAY-NORTH	172	UPPER WEST SIDE (79-96)	232
CO-OP CITY	53	HIGHBRIDGE/MORRIS HEIGHTS	113	OCEAN PARKWAY-SOUTH	173	UPPER WEST SIDE (96-116)	233
CORONA	54	HILLCREST	114	OLD MILL BASIN	174	VAN CORTLANDT PARK	234
COUNTRY CLUB	55	HOLLIS	115	OZONE PARK	175	WAKEFIELD	235
CROTONA PARK	56	HOLLIS HILLS	116	PARK SLOPE	176	WASHINGTON HEIGHTS LOWER	236
CROWN HEIGHTS	57	HOLLISWOOD	117	PARK SLOPE SOUTH	177	WASHINGTON HEIGHTS UPPER	237
CYPRESS HILLS	58	HOWARD BEACH	118	PARKCHESTER	178	WEST NEW BRIGHTON	238
DONGAN HILLS	59	HUGUENOT	119	PELHAM BAY	179	WESTCHESTER	239

DONGAN HILLS-COLONY	60	HUNTS POINT	120	PELHAM GARDENS	180	WESTERLEIGH	240

对建筑类型的编码：

01 ONE FAMILY DWELLINGS	1
02 TWO FAMILY DWELLINGS	2
03 THREE FAMILY DWELLINGS	3
04 TAX CLASS 1 CONDOS	4
05 TAX CLASS 1 VACANT LAND	5
06 TAX CLASS 1 - OTHER	6
07 RENTALS - WALKUP APARTMENTS	7
08 RENTALS - ELEVATOR APARTMENTS	8
09 COOPS - WALKUP APARTMENTS	9
10 COOPS - ELEVATOR APARTMENTS	10
11A CONDO-RENTALS	11
12 CONDOS - WALKUP APARTMENTS	12
13 CONDOS - ELEVATOR APARTMENTS	13
14 RENTALS - 4-10 UNIT	14
15 CONDOS - 2-10 UNIT RESIDENTIAL	15
16 CONDOS - 2-10 UNIT WITH COMMERCIAL UNIT	16
17 CONDO COOPS	17
21 OFFICE BUILDINGS	18
22 STORE BUILDINGS	19
23 LOFT BUILDINGS	20
25 LUXURY HOTELS	21
26 OTHER HOTELS	22
27 FACTORIES	23
28 COMMERCIAL CONDOS	24
29 COMMERCIAL GARAGES	25
30 WAREHOUSES	26
31 COMMERCIAL VACANT LAND	27
32 HOSPITAL AND HEALTH FACILITIES	28
33 EDUCATIONAL FACILITIES	29
34 THEATRES	30
35 INDOOR PUBLIC AND CULTURAL FACILITIES	31
36 OUTDOOR RECREATIONAL FACILITIES	32
37 RELIGIOUS FACILITIES	33
38 ASYLUMS AND HOMES	34
39 TRANSPORTATION FACILITIES	35
40 SELECTED GOVERNMENTAL FACILITIES	36
41 TAX CLASS 4 - OTHER	37

42 CONDO CULTURAL/MEDICAL/EDUCATIONAL/ETC	38
43 CONDO OFFICE BUILDINGS	39
44 CONDO PARKING	40
45 CONDO HOTELS	41
46 CONDO STORE BUILDINGS	42
47 CONDO NON-BUSINESS STORAGE	43
48 CONDO TERRACES/GARDENS/CABANAS	44
49 CONDO WAREHOUSES/FACTORY/INDUS	45

建筑类别的编码：

A0	1	D6	31	H8	61	M4	91	RG	121
A1	2	D7	32	HB	62	M9	92	RH	122
A2	3	D8	33	HH	63	N2	93	RK	123
A3	4	D9	34	HR	64	N9	94	RP	124
A4	5	E1	35	HS	65	O1	95	RR	125
A5	6	E2	36	I1	66	O2	96	RS	126
A6	7	E7	37	I3	67	O3	97	RT	127
A7	8	E9	38	I4	68	O4	98	RW	128
A9	9	F1	39	I5	69	O5	99	S0	129
B1	10	F2	40	I6	70	O6	100	S1	130
B2	11	F4	41	I7	71	O7	101	S2	131
B3	12	F5	42	I9	72	O8	102	S3	132
B9	13	F9	43	J1	73	P2	103	S4	133
C0	14	G0	44	J8	74	P5	104	S5	134
C1	15	G1	45	J9	75	P6	105	S9	135
C2	16	G2	46	K1	76	P8	106	T2	136
C3	17	G3	47	K2	77	P9	107	V0	137
C4	18	G4	48	K3	78	Q1	108	V1	138
C5	19	G5	49	K4	79	Q8	109	V3	139
C6	20	G6	50	K5	80	Q9	110	V6	140
C7	21	G7	51	K6	81	R1	111	V9	141
C8	22	G8	52	K7	82	R2	112	W1	142
C9	23	G9	53	K9	83	R3	113	W2	143
CM	24	GU	54	L1	84	R4	114	W3	144
D0	25	GW	55	L3	85	R5	115	W4	145
D1	26	H1	56	L8	86	R6	116	W8	146
D2	27	H2	57	L9	87	R8	117	W9	147
D3	28	H3	58	M1	88	R9	118	Y1	148
D4	29	H4	59	M2	89	RA	119	Y3	149
D5	30	H6	60	M3	90	RB	120	Z0	150
								Z2	151
								Z9	152

③模型训练

本次作业第一次试探使用线性回归模型，但是测试结果的拟合优度R^2一直小于 0.1，效果很差，所以决定采用随机森林回归模型，原理如下：

随机森林回归模型由多棵回归树构成，且森林中的每一棵决策树之间没有关联，模型的最终输出由森林中的每一棵决策树共同决定。

随机森林的随机性体现在两个方面：

1、样本的随机性，从训练集中随机抽取一定数量的样本，作为每颗回归树的根节点样本；

2、特征的随机性，在建立每颗回归树时，随机抽取一定数量的候选特征，从中选择最合适的特征作为分裂节点。

算法原理如下：

（a）从训练样本集 S 中随机的抽取 m 个样本点，得到一个新的 S1…Sn 个子训练集;

（b）用子训练集，训练一个 CART 回归树(决策树)，这里在训练的过程中，对每个节点的切分规则是先从所有特征中随机的选择 k 个特征，然后在从这 k 个特征中选择最优的切分点在做左右子树的划分。(这里的得到决策树都是二叉树)

（c）通过第二步，可以生成很多个 CART 回归树模型。

（d）每一个 CART 回归树最终的预测结果为该样本点所到叶节点的均值。

（e）随机森林最终的预测结果为所有 CART 回归树预测结果的均值。

三、 实验过程及结果

在最开始进行实验时，使用随机森林回归模型进行训练测试得到的 RMSE 虽然比较大，但是查看R^2已经达到 0.8 多，所以决定采用此算法，在数据基础上进行多次训练。由于房价数额普遍比较大，所以在查看 MSE、RMSE 两个指标之外，另外新建一个百分比指标，如下：

$$y\_error\_percentage = \frac{|Y\_pred - Y\_test|}{Y\_test}$$

该指标反映出预测的差错与原始房价对比误差的占比，下面给出训练集上的前部分指标数据：

	y_true	y_pred	y_error	y_error_percentage
0	900000	815800.0	84200.0	0.093556
1	1440000	1762000.0	322000.0	0.223611
2	499500	504600.0	5100.0	0.010210
3	11015000	12573172.9	1558172.9	0.141459
4	160000	253280.0	93280.0	0.583000

测试集上的前部分指标数据：

	y_true	y_pred	y_error	y_error_percentage
0	610000	6.450000e+05	35000.000000	0.057377
1	1185000	1.070858e+06	114141.666667	0.096322
2	960000	1.248000e+06	288000.000000	0.300000
3	110000	2.582667e+05	148266.666667	1.347879
4	438000	4.376570e+05	343.000000	0.000783

可以看出大部分 y_error_percentage 都很小，在容忍范围之内。

使用 train_test_split 方法，按测试数据集占 0.1 的比例对数据进行分割,计算训练集上的平均绝对误差(MSE)、R2 决定系数、均方根误差 RMSE 三个指标，最开始时查看到训练集上的 R2 决定系数高达 0.845，但是测试集上的决定系数仅仅有 0.345，合理怀疑是训练集上的脏数据导致。于是每轮对训练集上的 y_error_percentage 最大的 100 行数据进行删除，并且每删除一轮就测试在训练集和测试集上的表现效果。

在经过多轮数据筛选、删除敏感值之后，在训练集上得到的三个指标数值如下：

```
{'平均绝对误差(MSE)': [3461354204184.418], 'R2 决定系数': [0.881], '均方根误差RMSE': [1860471.501]}
```

在测试集上的三个指标如下：

```
{'平均绝对误差(MSE)': [3983560516191.027], 'R2 决定系数': [0.596], '均方根误差RMSE': [1995885.898]}
```

其中，由于房价的差距比较悬殊，所以这里只关注R^2.在训练集上达到 0.881 说明预测效果十分良好，可以接受，在测试集上的R^2也达到了 0.596，说明预测精度可以接受（继续删除发现测试集的决定系数开始下降了），在接下来使用这个训练好的模型去预测测试集里面的房价.

对测试集采用与问题一同样的方法进行编码。

查看到测试集的税收级别里面有空白值，使用税收级别里面的中位数 5 进行填充。

下面给出预测结果的部分展示图，详细数据可以参考附件。

支撑材料列表：

数值化后的 train.xlsx:处理后的训练集数据

数值化后的 test.xlsx:处理和预测后的测试集数据

大数据导论作业.py:使用上面两个数据集进行模型训练的过程

Excel 代码.txt:数据预处理过程中使用到的宏代码

Submit.csv:预测结果表格

四、 实验心得

这次大作业完成之后我对于机器学习的预测算法有了更深层次的理解，对机器学习预测模型的一些指标如MSE、$RMSE$、R^2有了更为深刻的理解，同时也对于在 python 中利用 numpy、pandas、sklearn 库处理数据更加的熟练掌握，感谢这次大作业提供的机会。