基于多模型融合的房价预测分析

摘要

本文以拥有多因子的房价数据集为研究对象，旨在通过该数据集实现对数据集中房屋的售卖价格情况的预测分析。

对于数据集，首先进行了初步的属性观察。通过观察得到价格分布情况和各列为空的情况等信息。随后去除无用的信息列，进行缺失值填补工作，在不同情况下使用中位数，None，0，均值等方法进行填补。随后使用 Corr 函数观察各特征与价格之间的相关系数，并对非数值特征进行独热编码和 Labelencoder 编码操作。最后对于处理完成的数据进行归一化，完成数据预处理环节。处理数据之后进入模型的建立环节。其中分别使用了 XGBoost 回归器，LGBM 回归器，KernelRidge 三个模型。最后根据三个模型的交叉验证成绩决定模型融合时的结果权重，随后按比将模型的预测值进行融合，至此完成整个房价预测问题的求解。

关键词：多因子，数据预处理，相关系数，独热编码，XGBoost，LGBM 回归，KRR 回归，模型融合，交叉验证。

Abstract

This paper takes the house price data set with multiple factors as the research object, and aims to realize the analysis and prediction of the sale price of houses in the data set through the data set.

For the data set, we first observe the attributes. Through the observation, we can get the information of price distribution and empty columns. Then remove the useless information column. Then the missing values were filled, and the median, none, 0 and mean values were used in different cases. Then the corr function is used to observe the correlation coefficient between each feature and price, and the unique hot coding and labelencoder coding are used for non numerical features. Finally, the processed data is normalized to complete the data preprocessing. After processing the data, the model is established. Xgboost regressor, lgbm regressor and kernel ridge models are used respectively. Finally, according to the cross validation results of the three models, the weight of the model fusion is determined, and then the predicted values of the model are fused according to the ratio. So far, the whole problem is solved.

Key words: multi factor, data preprocessing, correlation coefficient, exclusive hot coding, xgboost, lgbm regression, KRR regression, model fusion, cross validation.

目录

1. 数据初步分析

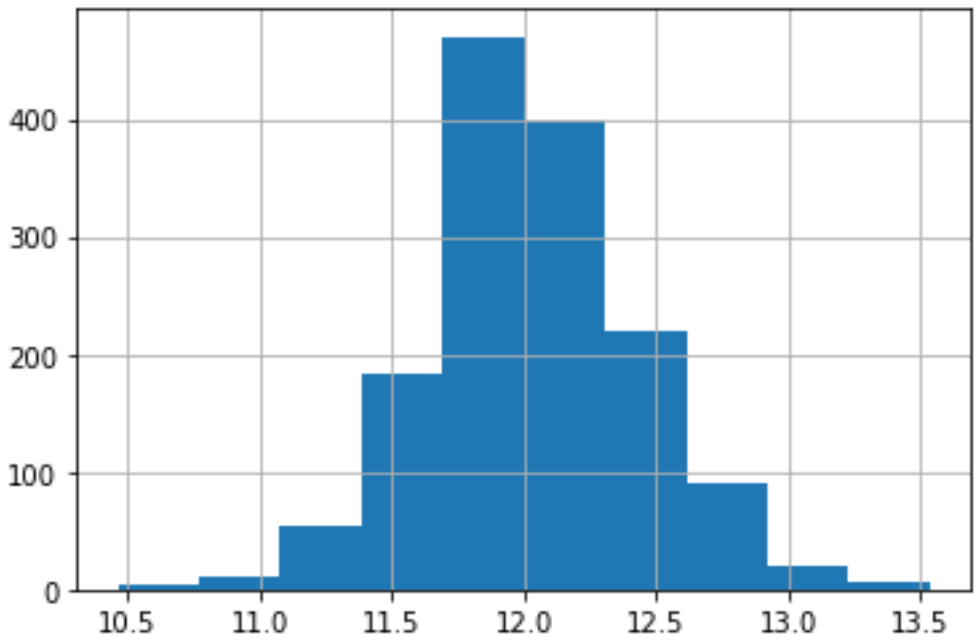
1.1 读取数据

读取数据均采用 pandas 库的 read_csv()方法进行

1.2 售价情况分析

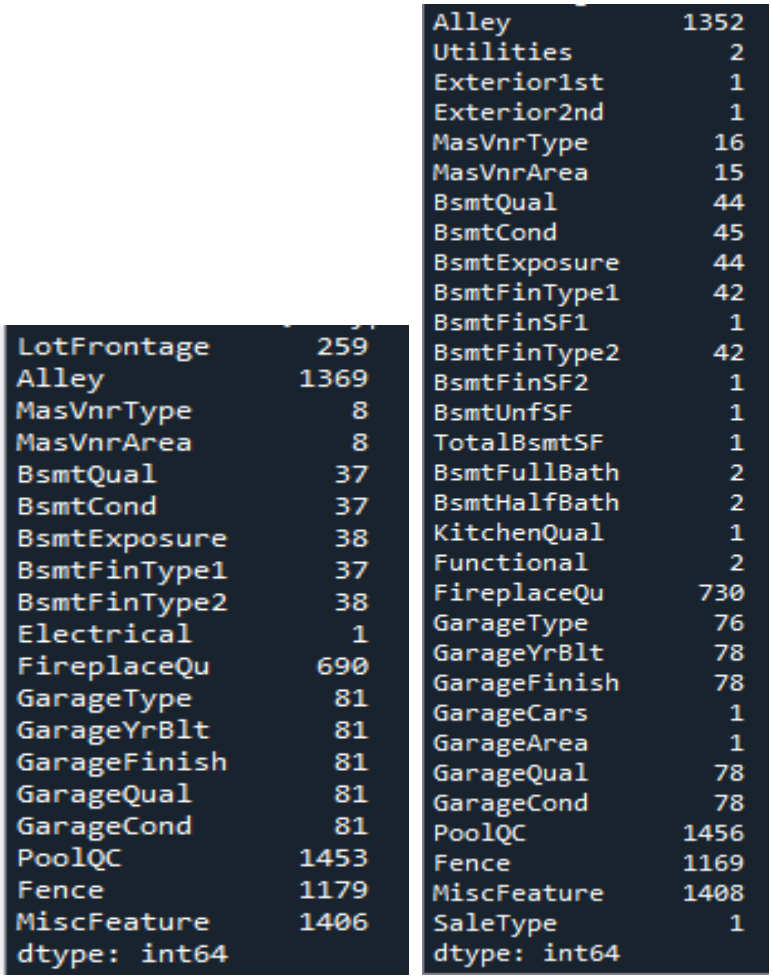
观察数据集中数据大致的分布情况。对特征进行初步的认识。首先观察售价的情况并画出价格的分布直方图（如下图）。

```
   0_1  0_a  1_3  1_b  2_4  2_c
0    1    0    1    0    1    0
1    0    1    0    1    0    1
count    1460.000000
mean       12.024051
std         0.399452
min        10.460242
25%        11.775097
50%        12.001505
75%        12.273731
max        13.534473
Name: SalePrice, dtype: float64
```



1.3 空值情况分析

对 train,test 数据集中特征为空的数据进行观察。



LotFrontage	259
Alley	1369
MasVnrType	8
MasVnrArea	8
BsmtQual	37
BsmtCond	37
BsmtExposure	38
BsmtFinType1	37
BsmtFinType2	38
Electrical	1
FireplaceQu	690
GarageType	81
GarageYrBlt	81
GarageFinish	81
GarageQual	81
GarageCond	81
PoolQC	1453
Fence	1179
MiscFeature	1406
dtype: int64	

Alley	1352
Utilities	2
Exterior1st	1
Exterior2nd	1
MasVnrType	16
MasVnrArea	15
BsmtQual	44
BsmtCond	45
BsmtExposure	44
BsmtFinType1	42
BsmtFinSF1	1
BsmtFinType2	42
BsmtFinSF2	1
BsmtUnfSF	1
TotalBsmtSF	1
BsmtFullBath	2
BsmtHalfBath	2
KitchenQual	1
Functional	2
FireplaceQu	730
GarageType	76
GarageYrBlt	78
GarageFinish	78
GarageCars	1
GarageArea	1
GarageQual	78
GarageCond	78
PoolQC	1456
Fence	1169
MiscFeature	1408
SaleType	1
dtype: int64	

2. 数据预处理

2.1 LotFrontage 字段处理

对该字段使用所有相同邻居的住宅的距离中位数来填充

2.2 Electrical 字段处理

对该字段使用众数填充

2.3 MasVnrType 字段处理

地板类型和面积是一致的，所以使用相互对照填充

2.4 车库，地下室，FirePlaceQu 字段处理

对以上字段均使用 None 进行填充，表示没有该部分结构

2.5 剩余字段处理

Exterior1st、Exterior2nd，MSZoning、utilities、KitchenQual、Functional、SaleType，GarageCars、GarageArea 字段均为数值类型，由于没有该部分，于是直接填充 0。

3. 特征工程

3.1 数据集合并

将训练数据和测试数据进行拼接,方便做相同的类型转换和编码处理

3.2 独热编码

使用提前编写好的 category2num 函数进行独热编码转换，也可以使用 pandas 库自带的 get_dummies 函数进行操作。

3.3LabelEncoder 编码

对存在明显大小关系的属性采用LabelEncoder进行编码

3.4 归一化处理

采用 numericStandard 函数对所有数据进行归一化处理

4. 模型构建

4.1 XGBoost

对 XGB 模型进行验证，得分如下

```
Xgboost score: 0.8993 (0.0115)
```

4.2 LGBM 回归

对 XGB 模型进行验证，得分如下

```
Lightgbm score: 0.8920 (0.0154)
```

4.3 KRR 回归

对 XGB 模型进行验证，得分如下

```
KRR score: 0.8566 (0.0216)
```

4.4 模型融合

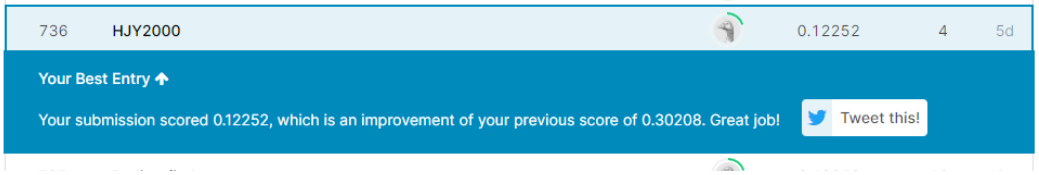
采用线性加权求和的方式对上述模型进行融合，其中权重分别为 50%，30%，20%。将加权的结果保存于 CSV 中。

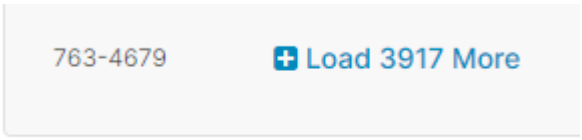
5. 附录

5.1 程序源代码

房价预测.py

5.2 成绩证明





736/4679=15.87%

5.3 运行设计说明

5.3.1 输入说明

输入文件为 train.csv,test.csv。分别为训练数据和待 测试数据，都存放于房价预测文件夹中

5.3.2 输出说明

输出结果为一个 CSV 文件，结果为最终的预测结果，文件名为 submition_2 .csv,存放于房价预测文件夹中

5.3.3 注意事项

房价预测.py 文件在引用路径时使用了默认当前路径，故 需要保持程序文件和输入文件在同一目录。