房屋销售分析 - 日本房屋价格预测

郭牧豪

2020 年 10 月

1 摘要

“居者有其屋” 是人们的基本生活需求，房屋作为人们的主要财产，其价格一直以来是人们关注的焦点，尤其是人口密集型国家和地区，像中国，日本，美国等，人们对房屋价格的价格尤为关注。本报告根据 2005 年到 2019 年日本国土交通省（MLIT）调查的日本房地产交易价格的记录，建立房屋价格预测模型，并根据地区的房屋价格对地区的发展提出可行性建议。本模型由数据预处理，数据降维分析，分类预测，精准预测等模块组成。模型以 LDA 分类，逻辑回归，随机森林，支持向量机分类器和线性回归为基础，优化出一种双次训练的机器学习模型，其中采用 PCA 和 LDA 进行降维分析，预测部分分为分类预测和精准预测。

2 介绍

2.1 题目分析

提供的数据包括 2005 年到 2019 年日本国土交通省（MLIT）调查的日本房地产交易价格的记录。数据分为县级代码和房价数据两部分。房价数据有 47 个文件，分别代表 47 个县级地区的房价记录，每个文件中，有 38 列数据，包括 1 个序号列，一个房价标签数据列，和 36 个 “特征” 列。

2.2 模型概论

对每个县的数据进行单独分析。分别得出每个县市地区的房屋价格预测。分为数据预处理阶段和模型分类预测和模型精准预测 3 个阶段。

2.2.1 数据预处理

- 数据读取：以县级地区为单位，读取每一个数据文件
- 特征编码：对于字符串特征，进行直接编码转化为浮点型特征，对于特殊性浮点型特征如 “TimeToNearestStation”，由于其内部有些数据由字符串表示，但其本质是浮点型特征，因此进行人工赋值处理。
- 标签分类：根据每个数据集的 25% 分位数、50% 分位数、75% 分位数，将数据集分为 4 类并赋予对应的标签。初步转化为分类问题。与此同时，根据房屋价格由低到高，将数据分为 4 部分，每一部分代表一个价格区间的数据，以便之后精准预测使用。

- 训练集测试集分离：将所有数据分为测试集和训练集。另外，上一步中的用于精准预测的四部分小数据集作为第二阶段的训练集，根据不同的策略选取不同的小训练集进行训练。
- 降维分析：包括 PCA 主成分分析和 LDA 线性判别分析，得到新的特征或选择最好的特征，并且提高计算速度。

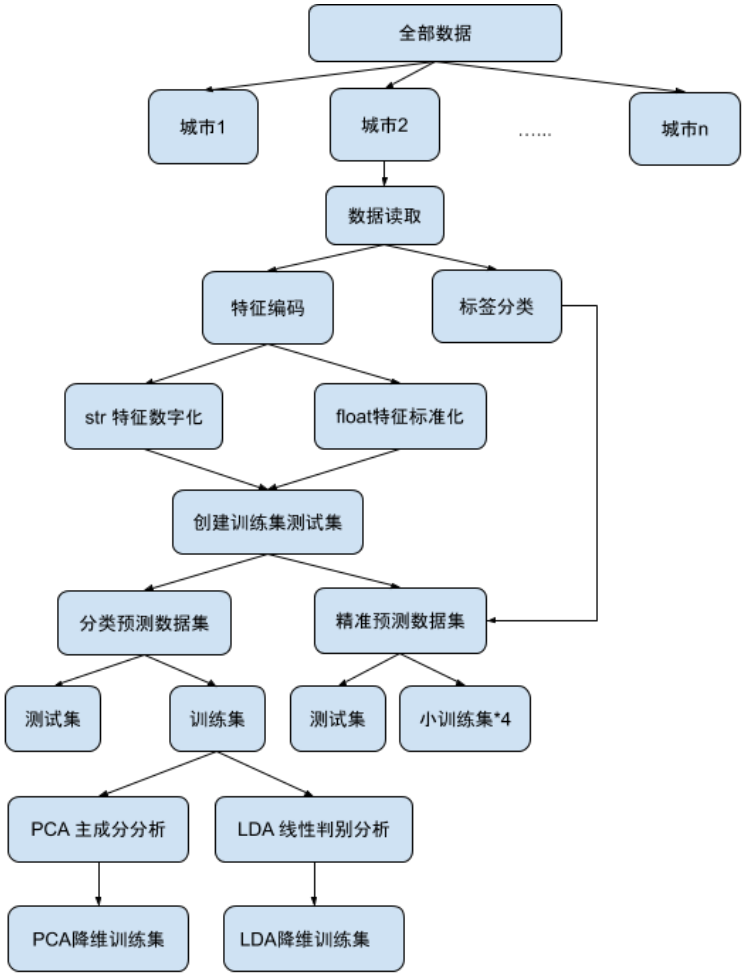

图 1: 数据预处理

2.2.2 模型分类预测

比较不同的分类器，选取预测成功率最高的分类器。以该分类器最为精准预测的前驱分类器。使用的分类器如下：

- LDA
- Logistic Regression with L1 penalty
- Logistic Regression with L2 penalty
- Random Forest Tree

- Support Vector Classifier

2.2.3 模型精准预测

根据模型分类预测的结果，得到预测结果类别，根据此类别，运用对应的精准预测训练集作为精准预测模型的训练集，通过线性回归模型进行房屋价格的精准预测。

2.2.4 模型选择、预测流程图

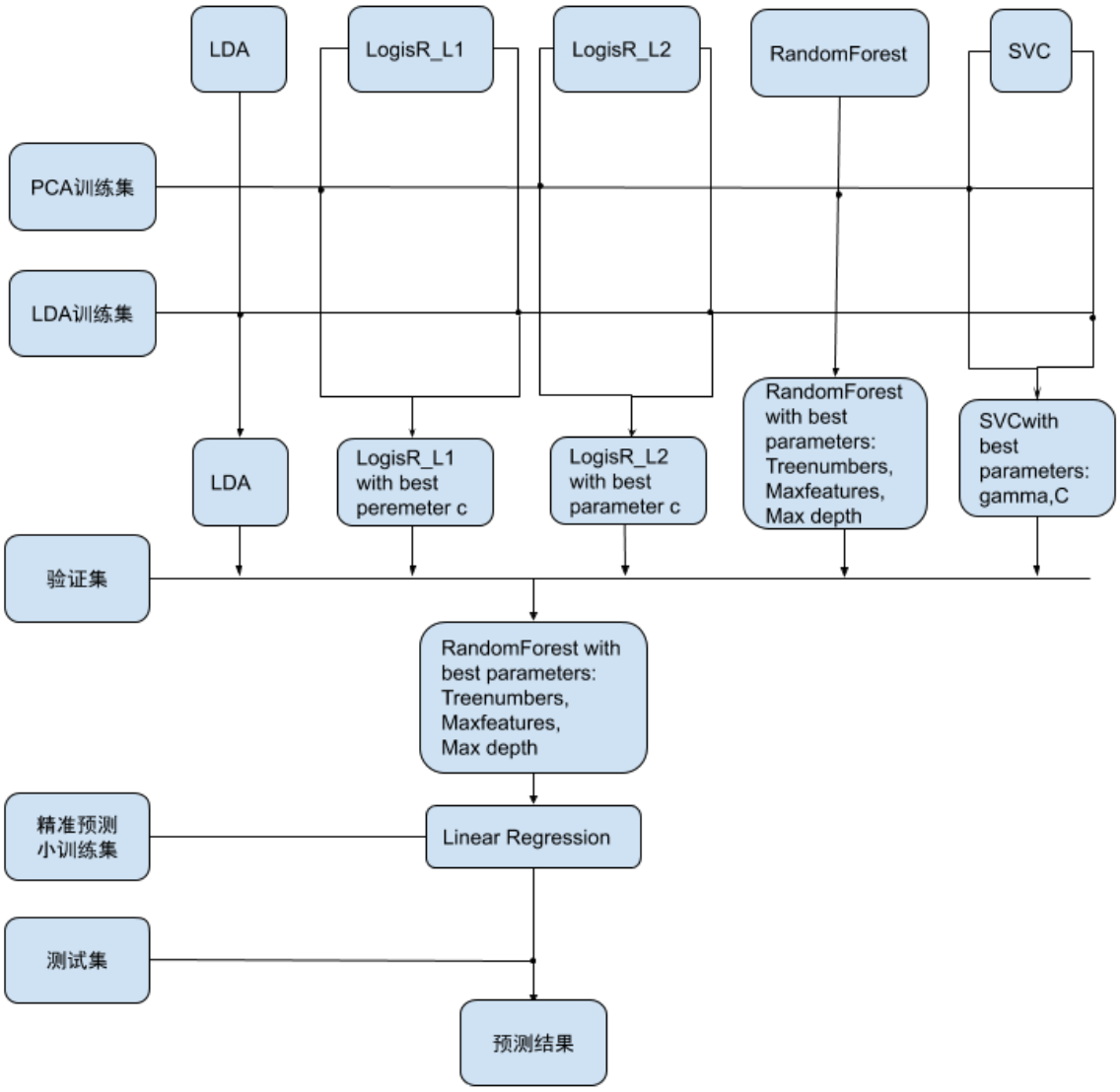

图 2: 模型选择、预测流程图

3 实验执行

3.1 数据预处理

3.1.1 数据集特征选取

原始数据共 36 列，将 TradePrice 作为标签列，No 是序号列，去掉重复列后，为最大程度的保留原始数据，共选择 30 列作为特征列，其中 10 列为 Float 型数据，3 列为 Boolean 型数据，17 列为 String 型数据。选择的特征列如表1 所示。

3.1.2 价格标签分析

分别选取 “TradePrice” 列的 25% 分位数（A）、50% 分位数 (B)、75% 分位数 (C) 作为分类边界，将数据分为 4 类，在原始数据中添加 “TradePrice_class” 列，作为分类标签列，类标签分别为 “0 - A”，”A - B”，”B - C” 和”C - +”。

3.1.3 特征编码

- String 特征编码：使用 LabelEncoder() 函数将 String 特征转化为 Int 特征。
- 特殊 Float 特征：对于”TimeToNearestStation” 特征，取中间值作为数值特征” 30-60minutes “取 45 min，“1H-1H30” 取值 75 min，“1H30-2H” 取值 105 min
- 空值 Float 类型特征：舍弃该样例
- 常规 Float 特征：不变

3.1.4 训练集测试集分离

按照 7：3 的比例，随机选取样本，构成训练集和测试集。训练集用于训练模型。对于总训练集，又将其按照不同的 “TradePrice_class” 类别，分为 4 个小训练集，分别记为 “df_0_A” ，“df_A_B” ，“df_B_C” 和 “df_C_D” 便于之后的精准预测使用。

3.1.5 PCA 主成分分析

使用 PCA 主成分分析，选取 10 个影响最大的特征组合，作为训练特征。

3.1.6 LDA 线性判别分析

使用 LDA，由于一共分为 4 类，所以共选取 3 个特征，作为训练特征。

3.2 模型分类预测

3.2.1 LDA

LDA 不仅可以用来降维数据，也可以用来作为一个分类器，直接用在降维步骤中训练好的 LDA 模型作为分类器进行预测。

表 1: 特征变量及类型

特征	类型
Type	String
Region	String
MunicipalityCode	String
Prefecture	String
DistrictName	String
NearestStation	String
TimeToNearestStation	Float
MinTimeToNearestStation	Float
MaxTimeToNearestStation	Float
FloorPlan	String
Area	Float
AreaIsGreaterFlag	Boolean
LandShape	String
Frontage	Float
FrontageIsGreaterFlag	Boolean
TotalFloorArea	Float
TotalFloorAreaIsGreaterFlag	Boolean
BuildingYear	String
PrewarBuilding	String
Structure	String
Use	String
Purpose	String
Direction	String
Classification	String
Breadth	Float
CityPlanning	String
CoverageRatio	Float
FloorAreaRatio	Float
Year	String
Quarter	String

3.2.2 Logistic Regression with L2 penalty

我们希望最小化损失函数，增加 L2 惩戒避免过拟合。使用交叉验证选择最佳参数 C，C 作为正则化系数的倒数，共选取 11 个值，分别为 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{2}, 10^{3}, 10^{4}, 10^{5}。使用 10 折交叉验证。

3.2.3 Logistic Regression with L1 penalty

和 Logistic Regression with l2 penalty 相似，我们希望最小化损失函数，增加 L1 惩戒避免过拟合。使用交叉验证选择最佳参数 C，C 作为正则化系数的倒数，共选取 11 个值，分别为 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{2}, 10^{3}, 10^{4}, 10^{5}。使用 10 折交叉验证。

3.2.4 Random Forest

随机森林的预测结果基于森林中每棵树的投票。使用随机森林分类器是因为它可以很好地处理具有特征空间的数据集，特别是多特征的数据。我使用交叉验证来选择森林中的树数，每次迭代的特征数以及树的最大深度。树个数从 100 到 1500，间隔为 50。由于特征的数量应少于特征总数，因此我还使用交叉验证来选择每次迭代的要素数量，即从 6 开始到 10，间隔为 1。对于树的深度，为树的深度设置了三个值：无，50 和 100。使用 10 折交叉验证。

使用交叉验证选择最佳参数 n_estimators”，“max_features”，和” max_depth。使用 10 折交叉验证。最大特征数从 5 到 10，树个数从 100 到 1500，间隔为 50

3.2.5 SVC

假如数据集不是线性可分离的，那么支持向量分类器可以在这种情况下较好地工作。SVC 的目标是拟合提供的数据并返回最合适的超平面，该超平面可以将数据集划分为不同的类。我选择通用内核函数“RBF” 内核，并使用交叉验证来获得最佳参数 “gamma” 和 “C”，它们分别代表 “RBF” 核系数和正则化参数。“C” 的作用是避免过度拟合。

使用交叉验证选择最佳参数 “gamma” 和 “C”。我选择 gamma 值分别为: 0.1, 0.01, 0.001, 0.0001, C 的值分别为: 0.1, 1, 10, 100。并使用 10 折交叉验证。

3.3 模型精准预测

基于模型分类预测的结果，选择不同的小训练集进行线性回归模型的训练以实现精准预测。

4 实验结果

4.1 分类模型实验结果

我们以青森🄵（Aomori）为例，实验结果如下：

表 2: 实验结果–分类模型预测

算法	预测正确率
LDA	0.679176
Logistic Regression_L2 with LDA	0.677183
Logistic Regression_L2 with PCA	0.661242
Logistic Regression_L1 with LDA	0.542677
Logistic Regression_L1 with PCA	0.538359
Random Forest with PCA	0.715044
SVC with LDA	0.688143
SVC with PCA	0.669212

4.2 结论

模型分类预测正确率会由不同的数据集产生不同的正确率。在前期分类过程中，效果最好的是 Random Forest 模型，以青森🄵（Aomori）为例，Random Forest 预测正确率为 0.715，再结后期的线性回归模型，可以预测出 2020 年日本某地区的房价。

4.3 分类模型实验结果截图

```
LDA : the predict Score is 0.6791763533709732
LDA : the predict Class is ['C ~ +' 'C ~ +' 'C ~ +' ... 'C ~ +' 'C ~ +' 'C ~ +']
Logistic Regression_L2 : the predict Score is 0.6771836599136499
Logistic Regression_L2 : the predict Class is ['C ~ +' 'C ~ +' 'C ~ +' ... 'C ~ +' 'C ~ +' 'C ~ +']
Logistic Regression_L1 : the predict Score is 0.5426768515443374
Logistic Regression_L1 : the predict Class is ['C ~ +' 'C ~ +' 'C ~ +' ... 'C ~ +' '0 ~ A' 'C ~ +']
selected_model_index is15
the best maxfeature for the best model is 5
the best treenumber for the best model is 800
Random Forest : the predict Score is 0.7150448356027898
Random Forest : the predict Class is ['B ~ C' 'C ~ +' 'C ~ +' ... 'C ~ +' 'B ~ C' 'C ~ +']
the parameters for the best model are {'C': 10, 'gamma': 0.1}
the cross validation score for the best model is 0.6865480427046263
SVC : the predict Score is 0.6881434739289273
SVC : the predict Class is ['C ~ +' 'C ~ +' 'C ~ +' ... 'C ~ +' 'B ~ C' 'C ~ +']
```

图 3: 实验截图 1

```
LDA : the predict Score is 0.6761873131849884
LDA : the predict Class is ['C ~ +' 'C ~ +' 'C ~ +' ... 'C ~ +' 'C ~ +' 'C ~ +']
Logistic Regression_L2 : the predict Score is 0.6612421122550648
Logistic Regression_L2 : the predict Class is ['C ~ +' 'C ~ +' 'C ~ +' ... 'C ~ +' 'C ~ +' 'C ~ +']
Logistic Regression_L1 : the predict Score is 0.5383593490534706
Logistic Regression_L1 : the predict Class is ['C ~ +' 'C ~ +' 'C ~ +' ... 'C ~ +' 'C ~ +' 'C ~ +']
selected_model_index is66
the best maxfeature for the best model is 7
the best treenumber for the best model is 550
Random Forest : the predict Score is 0.6874792427764862
Random Forest : the predict Class is ['C ~ +' 'C ~ +' 'C ~ +' ... 'C ~ +' 'C ~ +' 'C ~ +']
the parameters for the best model are {'C': 100, 'gamma': 0.1}
the cross validation score for the best model is 0.6872597864768684
```

图 4: 实验截图 2

```
the parameters for the best model are {'C': 1, 'gamma': 0.0001}
the cross validation score for the best model is 0.6511032028469751
SVC : the predict Score is 0.6692128860843574
SVC : the predict Class is ['C ~ +' 'C ~ +' 'C ~ +' ... 'C ~ +' 'C ~ +' 'C ~ +']
```

图 5: 实验截图 3

5 总结

本模型分为分类预测和精准预测，选出了最优分类模型，即随机森林模型，之后再次将分类模型的结果应用于精准模型中，即线性回归模型，最终可以预测出房价。

5.1 优势与创新

- 最大程度了运用了所有特征
- 降维分析提高运算速度
- 两次训练结合，既可以得到大概的预测区间，又可以得到精准预测结果
- 选用多个分类器，确保最优分类器

5.2 不足与改进

- Float 型特征遇到空值时直接舍弃，有些样本浪费掉。之后可以用均值或近似值策略代替
- LDA 和 PCA 在提高运算速度的同时，降低了预测准确率。之后应该使用更优化的方法兼顾准确率和效率。

5.3 代码

Github: https://github.com/MuhaoGuo/–Japanese-house-price-prediction
Github-Japanese-house-price-prediction