|  |
| --- |
| **机器学习** |
| **Sberbank俄罗斯住房市场** |
| 主 研 人：杜思君  参 研 人： |
| 审 核 人：  方 向：工业算法案例研究  版 本 号：A |
|  |

**二〇一九年四月**

**版本更新**

|  |  |  |  |
| --- | --- | --- | --- |
| **日期** | **版本** | **更新描述** | **作者** |
| 2019/02/15 | A | 初稿 | 杜思君 |

**选题表格**

|  |  |  |  |
| --- | --- | --- | --- |
| **时间** | **竞赛名** | **竞赛背景描述（50字以内）** | **类型（分类/回归）** |
| 2019/4/3 | Sberbank Russian Housing Market | 为俄罗斯著名银行Sberbank提供住房市场预测 | 回归 |

目录

[1. 背景描述 4](#_Toc25575_WPSOffice_Level1)

[1.1 竞赛赛题描述 4](#_Toc5940_WPSOffice_Level2)

[1.2 评估指标描述 5](#_Toc12357_WPSOffice_Level2)

[2. 数据来源及描述性统计分析 5](#_Toc5940_WPSOffice_Level1)

[2.1 大赛数据来源 5](#_Toc22683_WPSOffice_Level2)

[2.2 数据的描述性统计 5](#_Toc16174_WPSOffice_Level2)

[2.2.1 数据基本情况描述： 5](#_Toc5940_WPSOffice_Level3)

[2.2.2 数据字段介绍： 5](#_Toc12357_WPSOffice_Level3)

[2.2.3 数据描述性统计 6](#_Toc22683_WPSOffice_Level3)

[3. 优秀算法思路 11](#_Toc12357_WPSOffice_Level1)

[3.1 方案一 1](#_Toc32293_WPSOffice_Level2)1

[3.1.1 方案一数据预处理及特征工程部分方案 11](#_Toc16174_WPSOffice_Level3)

[3.1.2 方案一模型设计、建立部分方案 11](#_Toc32293_WPSOffice_Level3)

[3.1.3 方案一结果、排名等 14](#_Toc29922_WPSOffice_Level3)

[3.1.4 方案一算法流程图 14](#_Toc4787_WPSOffice_Level3)

[3.2 方案二 1](#_Toc29922_WPSOffice_Level2)5

[4. 算法比较 17](#_Toc22683_WPSOffice_Level1)

[表4-1 算法比较 17](#_Toc7253_WPSOffice_Level2)

[5. 总结与展望 17](#_Toc16174_WPSOffice_Level1)

[5.1 总结 17](#_Toc23198_WPSOffice_Level2)

[5.2 建模思路 17](#_Toc16619_WPSOffice_Level2)

1. 背景描述

住房成本需要消费者和开发商的大量投资。当涉及到预算时，无论是个人还是公司 ，任何人都需要思考的最后一件事就是他们的大笔费用之一的不确定性。 Sberbank是俄罗斯历史最悠久，规模最大的银行，通过预测房地产价格来帮助客户，因此租房者，开发商和贷方在签订租约或购买建筑物时更有信心。尽管俄罗斯的住房市场相对稳定，但该国经济的动荡使得预测价格成为一个独特且极具困难挑战。以及住房特征（如卧室数量和位置）之间复杂的相互作用会值得使定价预测变得复杂。增加不稳定的经济效应意味着Sberbank及其客户需要的不仅仅是简单的回归模型。

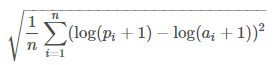
**1.1 竞赛赛题描述**

在这次竞赛中，Sberbank希望参赛者们能开发一些性能卓越的算法，这些算法可以通过使用广泛的特征数据来预测房地产价格。参赛者们将依赖丰富的数据集，其中包括住房数据和宏观经济模式。准确的预测模型将使Sberbank在不确定的经济环境中为其客户提供更多的确定性的帮助。

**1.2 评估指标描述**

竞赛中评估模型优劣的指标。

在选手提交的预测价格和实际数据之间进行评估，以对数均方根误差(Root Mean Squared Logarithmic Error(*RMSLE*))为评价指标，其计算公式如下：



其中，*n*为预测总数，*pi*与*ai*为对应的预测值和真实值，*log*为对数符号。

1. 数据来源及描述性统计分析

**2.1 大赛数据来源**

赛题的数据由Sberbank提供，是源自俄罗斯房价市场的真实数据，数据集中还提供俄罗斯经济和金融部门的整体状况信息。

以下是数据的超链接：

<test.csv>（测试集数据）

<train.csv>（训练集数据）

<sample_submission.csv>（提交样例）

<macro.csv> （宏观经济及金融数据）

<data_dictionary.txt> （数据字段解释）

**2.2 数据的描述性统计**

**2.2.1 数据基本情况描述：**

训练集数据来自2011年8月至2015年6月，测试集来自2015年7月至2016年5月的俄罗斯房价数据。目标变量在train.csv中称为price\_doc。

数据文件描述：

* train.csv，test.csv：有关个别交易的信息。这些行由“id”字段索引，该字段指的是单个事务（特定属性可能在单独的事务中出现多次）。这些文件还包括有关每个房产的当地区域的补充信息。
* macro.csv：关于俄罗斯宏观经济和金融部门的数据（可以加入“时间戳”栏目中的列车和测试集）
* sample\_submission.csv：格式正确的示例提交文件
* data\_dictionary.txt：解释其他数据文件中可用的字段

**2.2.2 数据字段介绍：**

由于训练集和测试集中有多达292（1）维的变量在宏观经济以及金融数据里也有100个变量，由于篇幅有限在此就不一一介绍。在官方提供的[data\_dictionary.txt](file:///C:\Users\DSJ的暗影精灵\Desktop\房价\data_dictionary.txt) 数据集里有各个变量的名词解释，在<https://www.kaggle.com/c/sberbank-russian-housing-market/data> 也能找到数据的数据类型等。

* + 1. **数据描述性统计**

数据集中有大量且非匿名的特征，为数据可视化，分析提供了很多资源。

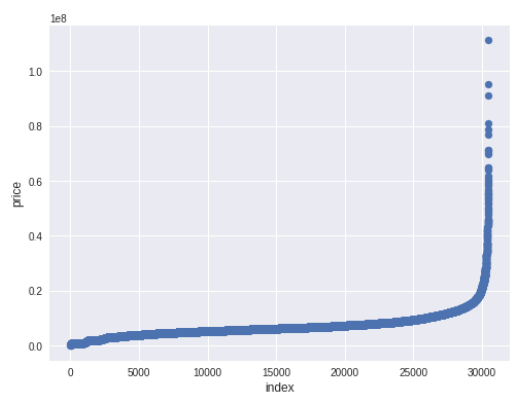
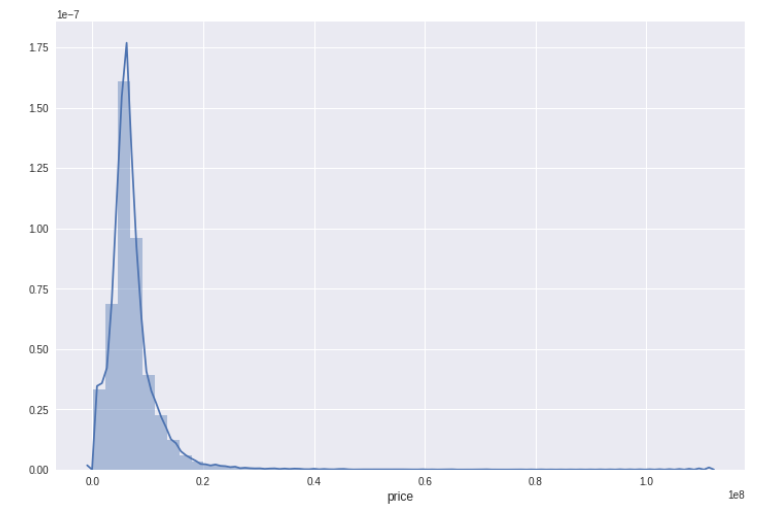
1) 对训练集中目标值（房价）的分析

Figure 2.2.3-1： 目标变量训练集中的散点分布图

从上图我们不难看出，数据中有一些值偏高，但个别的超高房价也许是合理的。

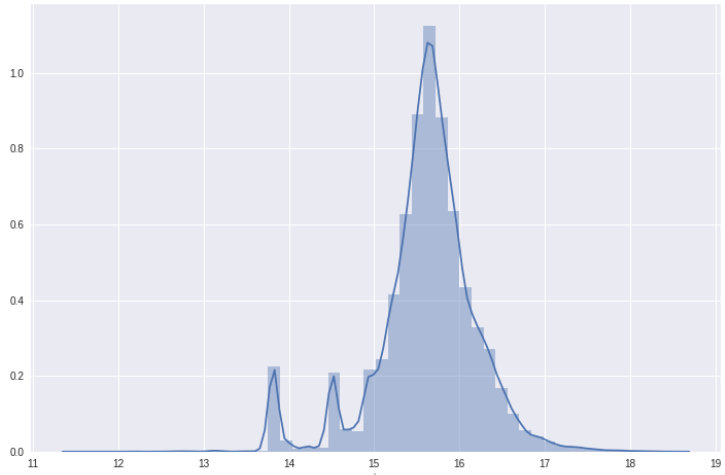


Figure 2.2.3-2: 目标值在训练集中的直方分布图（右图为取对数后的分布）

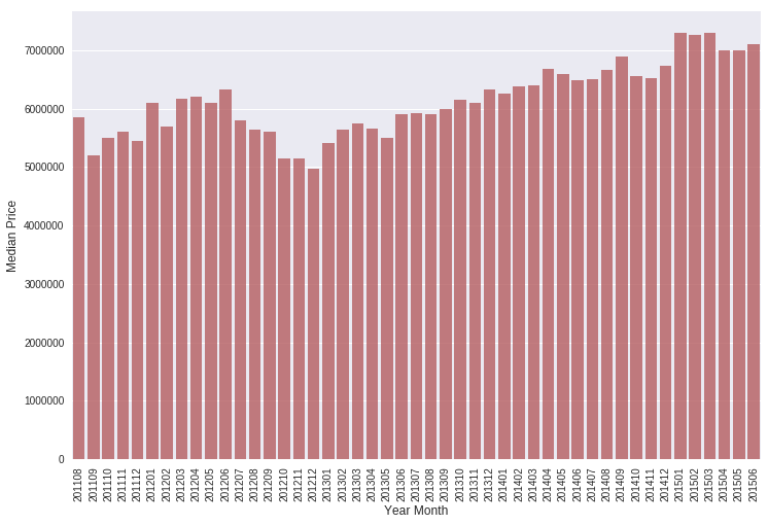


Figure 2.2.3-2: 房价的中位数随时间的变化（训练集中）

Figure 2.2.3-3: y在训练集中的正态分布图

从图中可以看出中间价格与时间有一些差异。 接近尾声时，价格似乎呈线性上升。

2) 对特征变量的分析

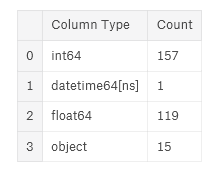
1. 数据类型统计

表2.2.3-1： 变量数据类型统计

由此可见绝大多数变量都是数值型变量，仅有15个特征是字符串型和一个时间形变量。

1. 缺失值占比

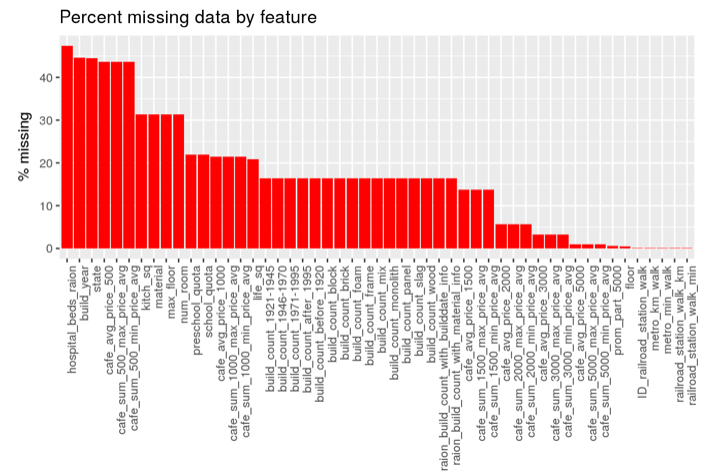


Figure 2.2.3-4:变量中缺失值占比情况

3) 房屋内部特征对房价的影响

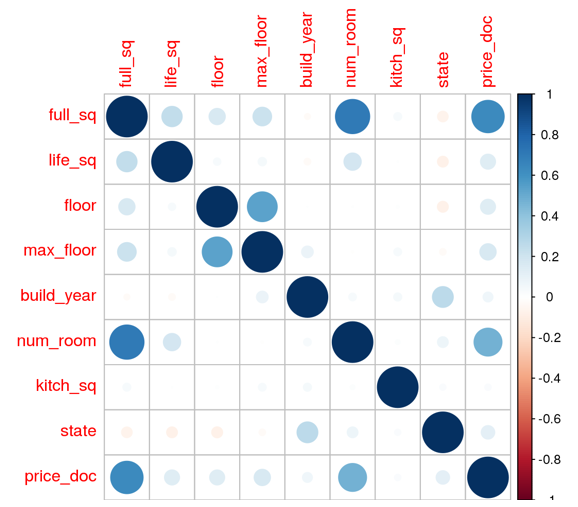
* 房屋特征总览

Figure 2.2.3-4: 住房特征

* 房屋面积以及房屋居室数与房价的统计

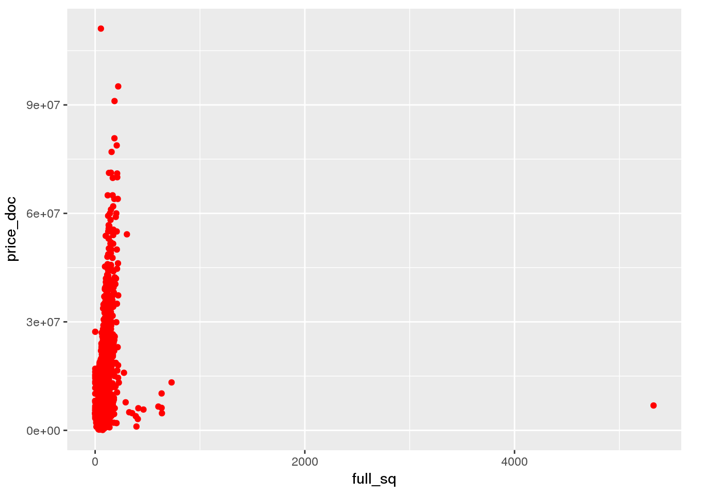


Figure 2.2.3-5:full\_sq与房价的分布关系

从图中看出有一个明显的离群值，将其去除。

* 住房面积与房屋价格

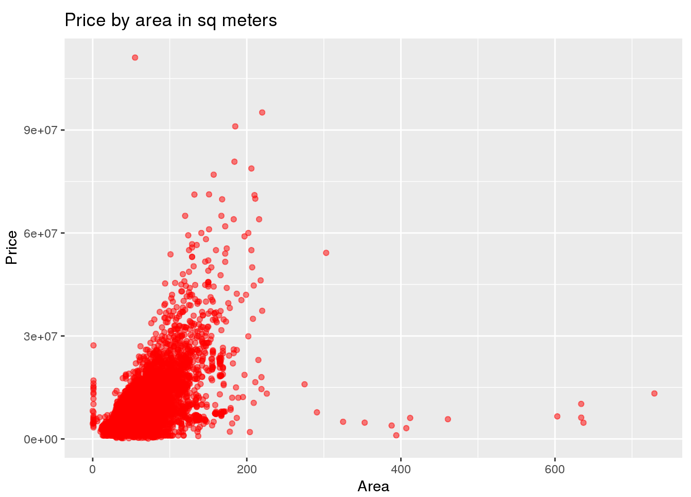


Figure 2.2.3-6:房屋面积与房价的关系

* 房屋居室统计

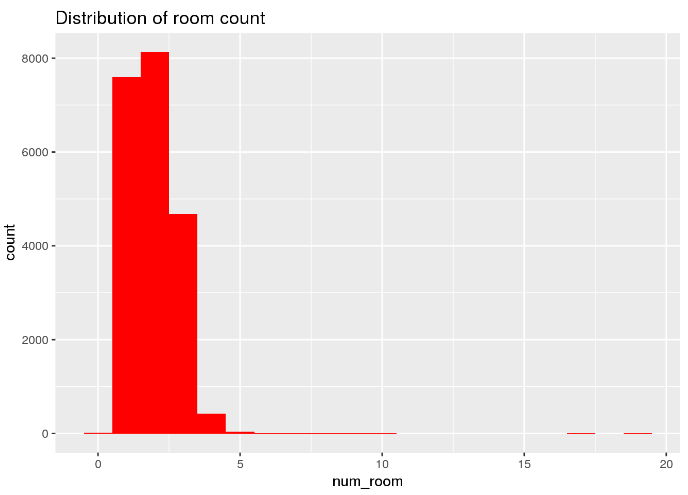


Figure 2.2.3-7:房屋房间数统计

大多数房价只有3个及以下的房间。

* 房屋建造年限统计与房价的分布关系

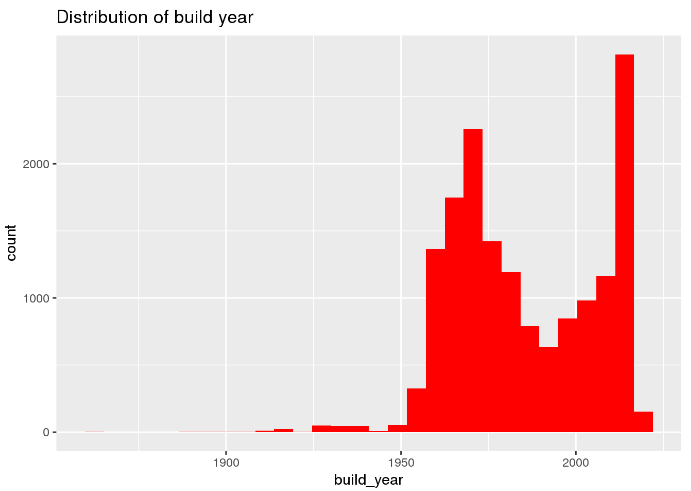


Figure 2.2.3-8:房屋建造时间统计

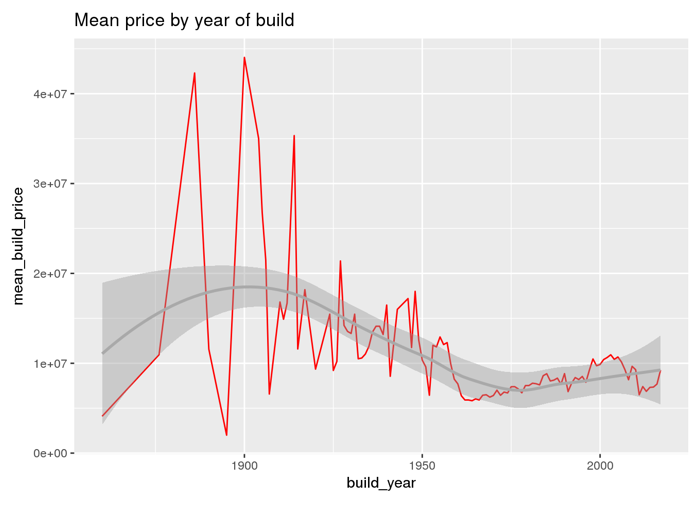


Figure 2.2.3-8:房屋建造时间与房屋均价统计

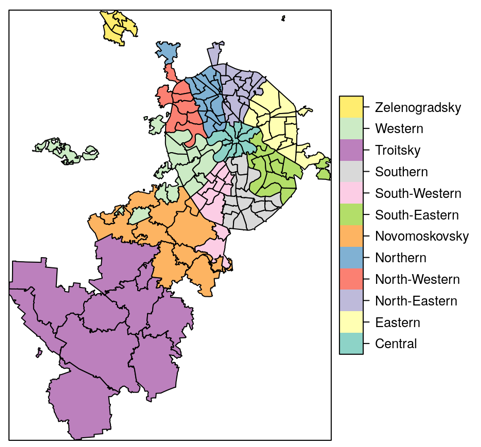
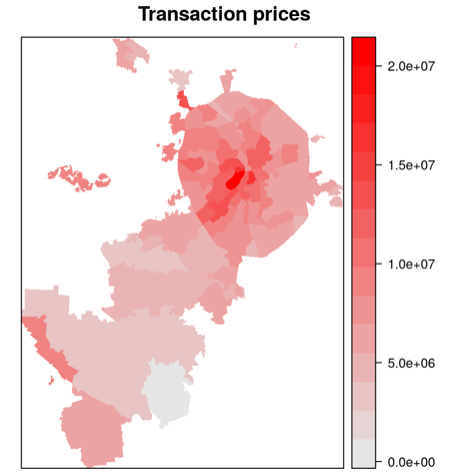
4) 房屋所在地对房价的影响

Figure 2.2.3-9 房屋所在地区与房价分布

3. 优秀算法思路

* 1. **方案一**

**3.1.1方案一数据预处理及特征工程部分方案**

方案一的队伍中有两个选手，分别针对自己的模型进行了的数据清洗。

* **选手一的模型中的数据清理:**

在选手一的LightGBM模型中，full\_sq这个特征具有非常重要的作用，但full\_sq中包含了太多的错误数据，难以直接使用。所以选手以full\_sq和其他一些特征一同构造了一些特征，构造公式如下：

其中，*X*是不同的分类特征，如房间数量、建造时间、楼层等

* **选手二的模型中的数据清理:**

第二个选手的模型同样基于LightGBM模型构建，不过他的思路与选手一侧重的方面不同。

选手二修正了数据集中有关房屋面积、建造年份和楼层等特征中的一些错误。此外，选手二还对具有相同地址的公寓增加了如建筑年份，房间数目等特征。

选手二还对数据集中有很多低价数据进行了去除，由于它们与避免卖方的税收目的有关，而且没有足够的信息来检测它们。这些低价的假价格不仅仅是1~3百万的价格，在一些地区，房屋的均价是1000万，但有些房屋的价格却在2~3百万之间，远低于市场价格。选手二使用了他的投资模型来找到它们（选手二猜测大多数非投资价格都是真实的）而且下降数据是预测和目标有很大不同（2倍或更多）。并在第一次迭代后再重复一次以清理更多。总的来说，由于数据价格不合理，选手将近10％的训练集中的数量下降。如果在训练期间保留它们，可能会遇到两种类型的麻烦：在本地验证期间，低价格数据的更好或更低的准确性可能掩盖了良好数据准确性的变化，当微调模型时，可能会错过一些有用的东西。

* **对时间型数据进行分割**

将时间型数据按照年月日分割，提取出更细的特征。从之前的分析中看，房价与时间的关系较为密切。

* **对字符串型数据进行编码**

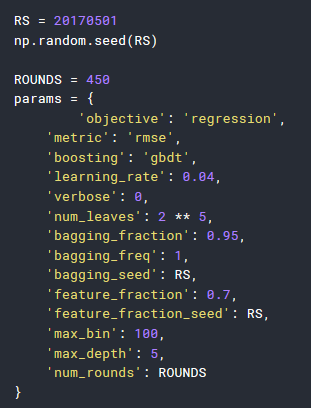
将所有的字符串型数据用one Hot进行编码。

**3.1.2** 方案一模型设计、建立部分方案

1. **选手一构建的模型**

选手一基于LightGBM构建了两个模型，分别针对投资型住房和居住（所有）型住房。选手发现这两类住房差异比较明显，用不同的模型能获得更好的效果。

接下来选手对参数和随机数种子进行了调节，获取最优参数。下图是选手一模型的参数（源自kaggle的kernel）



**2. 选手二构建的模型**

选手二没有之间预测整个房屋的售价，而是考虑预测房屋每平米的价格。选手二将此次比赛分为两个任务

1. 确定宏观经济情况对价格的影响

II ) 确定具体位置，建筑和公寓条件的价格

**对附加数据——宏观经济的使用**

选手二构建的LightGBM很好的解决了一个任务，他使用基于投资模型来查找训练集期间的宏观经济情况，然后根据这个将所有时间内的经济情况缩放到一个级别（这不是精确的，但比没有它更好）。选手二还试图构建模型以获得系数的趋势，但是发现数据集中没有足够的宏观数据。

同样的，选手二也对投资型住房和非投资型住房进行了分别建模

1. **投资型**

对于投资部分，选手二使用所有数据进行训练，并基于自己构建的验证方法重新筛选和构造了特征。

1. **非投资型**

对于非投资部分，选手二使用了两阶段方法 - 首先使用所有数据进行训练并预测。之后，选手使用了另一个基于非投资数据和第一阶段预测的模型。  
 对于非投资部分，我在训练集时段使用了额外的价格缩放。新房屋大部分的价格增长速度远高于一般水平。

**3. 模型融合**

该队伍最好的提交是来自这两个选手解决方案的几个模型的集合， 他们分别合并了不同的预测子集 - 例如每种产品类型（Investment和OwnerOccupier）都是由不同的模型预测的。该队伍通过分析他们预测的残差，发现并修复了模型的薄弱环节以及数据中的错误 ，由此取得了第一名的成绩。

**模型融合中出现的困难：**

在对Basmannoe-Savelki和Kuncevo这两个地区进行预测的时候，两个选手的模型相差较大。该队伍通过分析周边地区的情况，调整模型进行尝时，最终觉得对这两个地区针对性建模，最终分数提高较为明显。

**3.1.3** 方案一结果、排名等

得分：*RMSLE*: 0.30087

排名：*rank*: 1/3835

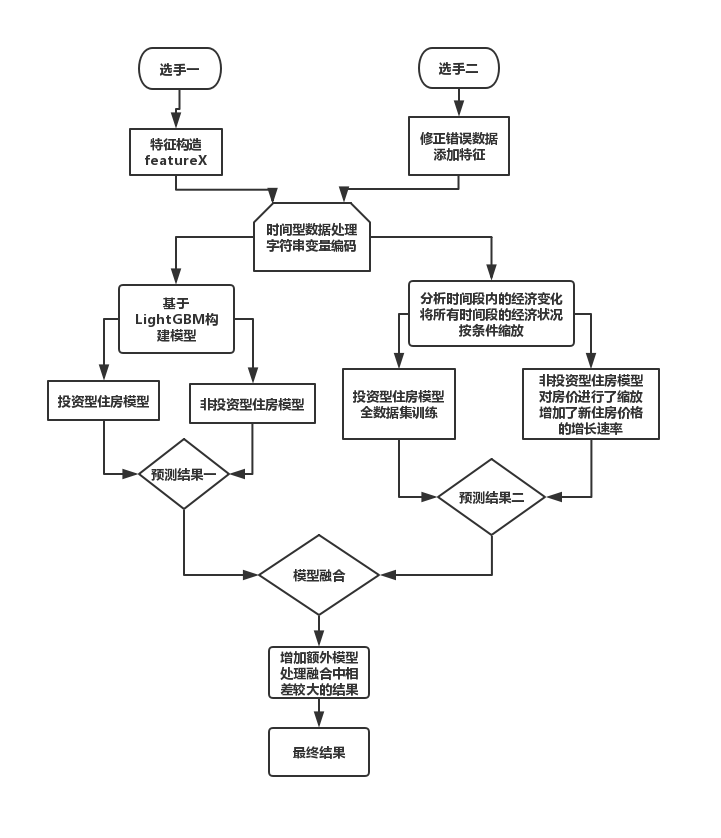
**3.1.4** 方案一算法流程图

Figure3.1.4-1:算法流程图

4. 算法比较

|  | **评估指标** | **特征工程** | **基础算法** | **基本库** |
| --- | --- | --- | --- | --- |
| **算法1** | **RMSLE值** | **特征构造** | **LightGBM** | **Lightgbm** |
| **算法2** | **RMSLE值** | **数据清洗修复等** | **LightGBM** | **Lightgbm** |

我们直接就队伍中两名选手的算法进行比较，两位选手都采用了最新的LightGBM算法来构建模型，但两者的切入点和思考方式不一致，各自的模型对某一类房价会有偏向性，但两者模型的融合确起到了互补的效果。

5. 总结与展望

**5.1 总结**

本次比赛有大量且非匿名的数据，为选手们提供了数据分析和特征探索的资源，对数据的探索和分析成为了制胜的关键。在算法上很多选手都使用了LightGBM，这种算法相较XGBoost更具效率，更加节省内存。

在本次收集中学习到了很多分析数据的方法，尤其是这种带有含义的变量，此外也对LightGBM算法有了一些了解，会在今后的建模中尝试使用。

**5.2 建模思路**

我在考虑建模时，首先还是会用XGBoost算法做一个尝试，XGBoost在性能上也能达到一个很不错的效果，关键还是在于对特征的处理。

大致思路如下：

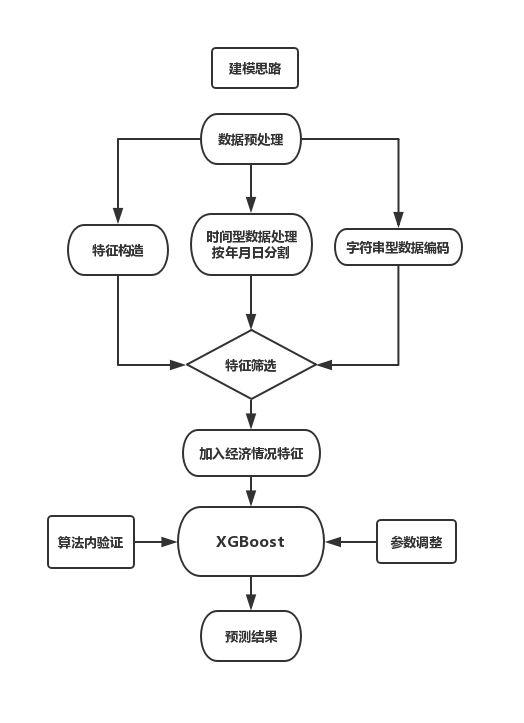


Figure5.2-1大致的建模思路