

# 某市政府重点区域人群密度预测

## 背景描述及业务需求

2020年，新型冠状病毒（COVID-19）感染的肺炎疫情牵动着全国人民的心，全国同舟共济、众志成城，打响了一场没有硝烟的疫情阻击战。习近平总书记指出：要鼓励运用大数据、人工智能、云计算等数字技术，在疫情监测分析、病毒溯源、防控救治、资源调配等方面更好发挥支撑作用。

而人口的流动聚集，客观上加大了疫情传播的风险和防控的难度。出于对公共卫生、重大公共利益相关影响的为研究目的，进一步掌握人员流动聚集动向，做好紧急疫情的防控工作，针对疫情相关的重点区域开展人群聚集密度预测的任务。

本文对重点区域的人群密度进行回归预测；数据挖掘建模的目标如下：

- 1、借助重点区域历史的人群密度，统计人流量指数和迁徙指数特征；
- 2、建立模型预测重点区域未来的人群密度，掌握人员流动聚集的动向；
- 3、针对人群密度较大的区域，做好紧急疫情防控工作。

本案例重点区域人群密度预测的总体流程如图1-1所示。

- 背景描述及业务需求
- 实施过程
  - 数据接入
  - 数据探索
  - 数据预处理
    - 特征衍生
    - 数据变换
    - 预处理流程图
  - 建立模型
- 总结

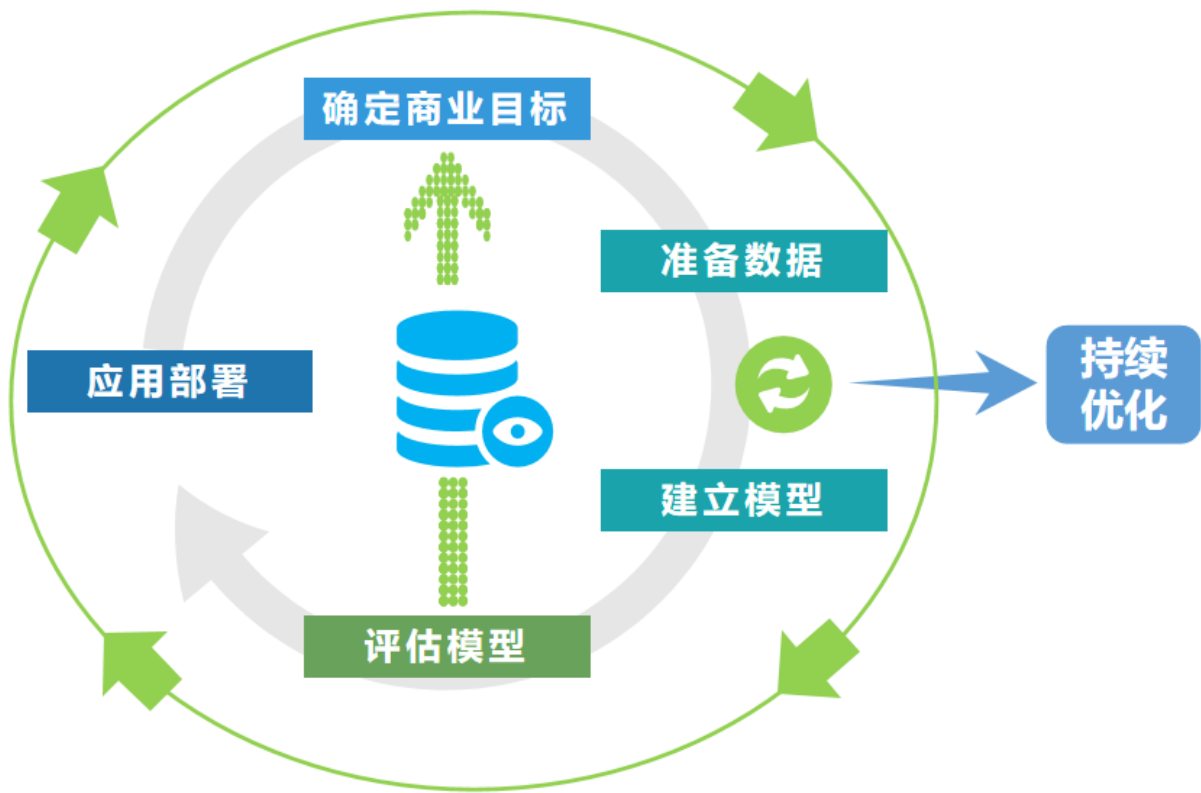


图1-1

- 1) 获取数据，数据来源于赛题重点区域人群密度预测数据；
- 2) 对获取的数据进行基本的处理操作，分组统计人流量指数和迁徙指数，作为模型的输入特征；
- 3) 根据统计特征数据建立重点区域人群密度预测模型；
- 4) 对模型结果进行评估。

## 实施过程

本案例共有3个数据集，为疫情期间20200117-20200215人流相关的数据。

以下是每个数据集的字段说明：

表2-1 重点区域人流量情况表

| 字段名称 | 字段说明  | 示例         |
|------|-------|------------|
| _c0  | 区域ID  | 10         |
| _c1  | 日期、小时 | 2020011703 |
| _c2  | 人流量指数 | 2.9        |

表2-2 重点区域信息表

| 字段名称 | 字段说明                     | 示例         |
|------|--------------------------|------------|
| _c0  | 区域ID                     | 10         |
| _c1  | 区域名称                     | 北京市石景山体育场  |
| _c2  | 区域类型                     | 运动健身；体育场馆  |
| _c3  | 区域中心点经度，国测局02坐标系         | 116.201709 |
| _c4  | 区域中心单纬度，国测局02坐标系         | 39.906121  |
| _c5  | 区域中心点所在网格的中心点经度，国测局02坐标系 | 116.200936 |
| _c6  | 区域中心点所在网格的中心点纬度，国测局02坐标系 | 39.906260  |
| _c7  | 区域面积，单位为平方米              | 90062.3125 |

表2-3 北京市迁徙指数表

| 字段名称 | 字段说明   | 示例       |
|------|--------|----------|
| _c0  | 迁徙日期   | 20200201 |
| _c1  | 迁徙出发省份 | 广东省      |
| _c2  | 迁徙出发城市 | 深圳市      |
| _c3  | 迁徙到达省份 | 北京市      |
| _c4  | 迁徙到达城市 | 北京市      |
| _c5  | 迁徙指数   | 1.83     |

数据说明：

- 重点区域人流量情况表中，人流量指数指与某天某小时内该区域内出现的人数成正比。A区域的人流量指数越大，表示A区域出现的人越多，反之越少。
- 北京市迁徙指数表中，迁徙指数指与某天北京市与其他城市之间的人群流动量成正比。A城市到B城市的迁徙指数越大，表示从A城市迁徙到B城市的人数越多，反之越少。

数据接入

我们在实验中添加 [数据源](#) 节点，将上述3张表的数据读取进来，部分数据如图2-1所示。

ⓘ 当前显示 100 条 / 总共有 717840 条数据 提示: 点击单元格可查看超出的内容

| #_c0 | #_c1       | #_c2 |
|------|------------|------|
| 1    | 2020011700 | 1.8  |
| 1    | 2020011701 | 1.5  |
| 1    | 2020011702 | 1.3  |
| 1    | 2020011703 | 1.3  |
| 1    | 2020011704 | 1.7  |
| 1    | 2020011705 | 1.2  |
| 1    | 2020011706 | 2.9  |
| 1    | 2020011707 | 4.7  |
| 1    | 2020011708 | 6.1  |
| 1    | 2020011709 | 5.2  |
| 1    | 2020011710 | 5.5  |
| 1    | 2020011711 | 6.2  |
| 1    | 2020011712 | 7.1  |
| 1    | 2020011713 | 7.9  |

注意: 表头中\*表示特征列, \*表示标签列

表头真名 表头别名

图2-1

为了便于理解字段含义，让字段看起来更加直观，我们使用 元数据编辑 节点，添加中文字段别名。更改后的输出如图2-2所示，流程图如图2-3所示。

ⓘ 当前显示 100 条 / 总共有 717840 条数据 提示: 点击单元格可查看超出的内容

| # 区域ID | # 日期小时     | # 人流量指数 |
|--------|------------|---------|
| 1      | 2020011700 | 1.8     |
| 1      | 2020011701 | 1.5     |
| 1      | 2020011702 | 1.3     |
| 1      | 2020011703 | 1.3     |
| 1      | 2020011704 | 1.7     |
| 1      | 2020011705 | 1.2     |
| 1      | 2020011706 | 2.9     |
| 1      | 2020011707 | 4.7     |
| 1      | 2020011708 | 6.1     |
| 1      | 2020011709 | 5.2     |
| 1      | 2020011710 | 5.5     |
| 1      | 2020011711 | 6.2     |
| 1      | 2020011712 | 7.1     |
| 1      | 2020011713 | 7.9     |

注意: 表头中\*表示特征列, \*表示标签列

表头真名 表头别名

图2-2

数据接入

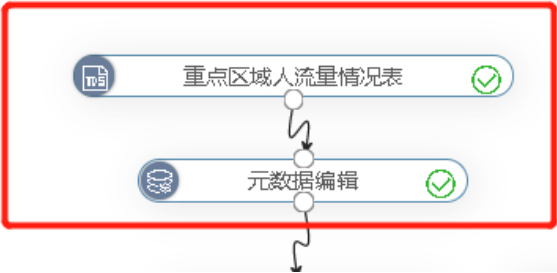


图2-3

数据探索

本案例的探索分析是对数据进行缺失值分析与数据分布分析，分析出数据的缺失和分布情况。通过数据观察之后，发现重点区域人流量情况表和北京市迁徙指数表中的日期时间字段格式并不一致（如图2-4和图2-5），会对表合并有影响，因此需先统一两张表的日期时间格式。

⌂ 当前显示 100 条 / 总共有 717840 条数据 [提示:点击单元格可查看超出的内容](#)

| # 区域ID | # 日期小时     | # 人流量指数 |
|--------|------------|---------|
| 1      | 2020011700 | 1.8     |
| 1      | 2020011701 | 1.5     |
| 1      | 2020011702 | 1.3     |
| 1      | 2020011703 | 1.3     |
| 1      | 2020011704 | 1.7     |
| 1      | 2020011705 | 1.2     |
| 1      | 2020011706 | 2.9     |
| 1      | 2020011707 | 4.7     |
| 1      | 2020011708 | 6.1     |
| 1      | 2020011709 | 5.2     |
| 1      | 2020011710 | 5.5     |
| 1      | 2020011711 | 6.2     |
| 1      | 2020011712 | 7.1     |
| 1      | 2020011713 | 7.9     |

注意：表头中\*表示特征列，\*表示标签列

表头真名 ☒ 表头别名

图2-4 重点区域人流量情况表

当前显示 100 条 / 总共有 21459 条数据 提示:点击单元格可查看弹出的内容

| # 日期     | 迁徙出发省份 | 迁徙出发城市 | 迁徙到达省份 | 迁徙到达城市 | # 迁徙指数  |
|----------|--------|--------|--------|--------|---------|
| 20200117 | 陕西省    | 商洛市    | 北京市    | 北京市    | 9.4E-4  |
| 20200117 | 北京市    | 北京市    | 江西省    | 上饶市    | 0.01468 |
| 20200117 | 北京市    | 北京市    | 山东省    | 淄博市    | 0.03619 |
| 20200117 | 湖南省    | 怀化市    | 北京市    | 北京市    | 0.00288 |
| 20200117 | 山东省    | 淄博市    | 北京市    | 北京市    | 0.01698 |
| 20200117 | 福建省    | 宁德市    | 北京市    | 北京市    | 0.00249 |
| 20200117 | 江西省    | 新余市    | 北京市    | 北京市    | 0.00143 |
| 20200117 | 辽宁省    | 鞍山市    | 北京市    | 北京市    | 0.01053 |
| 20200117 | 辽宁省    | 丹东市    | 北京市    | 北京市    | 0.00748 |
| 20200117 | 北京市    | 北京市    | 湖北省    | 咸宁市    | 0.01186 |
| 20200117 | 贵州省    | 贵阳市    | 北京市    | 北京市    | 0.01892 |
| 20200117 | 北京市    | 北京市    | 山西省    | 晋城市    | 0.01717 |
| 20200117 | 甘肃省    | 平凉市    | 北京市    | 北京市    | 0.0012  |
| 20200117 | 北京市    | 北京市    | 山东省    | 日照市    | 0.01717 |

注意: 表头中\*表示特征列, \*表示标签列

表头真名 表头别名

图2-5 北京市迁徙指数表

然后我们接入一个 全表统计 节点，用于查看整个数据集数值型数据的情况。选中所有数值型字段如图2-6，输出结果如图2-7所示，可以看到所有数据均不存在缺失值。

选择列

源数据列表0/0

请输入搜索内容

无数据

已选字段列表0/3

请输入搜索内容

☐ #区域ID

☐ #日期小时

☐ #人流量指数

< 到左边

到右边 >

注意: 非数值型变量已被禁用

确定

取消

图2-6 选取所有数值型字段

🕒 查看全量统计

| 行数     | 唯一值  | 缺失值 | 偏度         | 峰度         | 中位数        | 下四分位       | 上四分位       |
|--------|------|-----|------------|------------|------------|------------|------------|
| 717840 | 997  | 0   | -3.0871... | -1.2000... | 499        | 250        | 748        |
| 717840 | 720  | 0   | -1.0535... | -1.9584... | 2020020... | 2020012... | 2020020... |
| 717840 | 7196 | 0   | 18.5337... | 553.291... | 6.2        | 2.1        | 20.5       |

图2-7 数据缺失情况

## 数据预处理

本案例主要采用特征衍生和数据变换的预处理方法。

### 特征衍生

通过数据探索分析，我们发现两张数据表中的字段格式不统一，无法进行合并。因此需统一字段格式，具体处理方法：对两张表均接入一个 [派生列](#) 节点，截取出日期时间字段的年月日信息，再统一字段格式。所以我们接入一个派生列节点，派生列配置如图2-8所示。

🔍 派生列配置 ×

输入关键字搜索列

DOUBLE

人流量指数

INTEGER

区域ID

日期小时

函数

累积

字符拼接(拼接符)

字符拼接

字符截取

小数点保留

内容提取(年)

内容提取(月)

添加/编辑表达式

1 concat\_ws("-",substring([日期小时],1,4),substring([日期小时],5,2),substring([日期小时],7,2))

派生列名: 日期 操作: 确定 取消

日期

concat\_ws("-",substring([日期小时],5,2),substring([日期小时],7,2))

小时

substring([日期小时],9,2)

确定 取消

图2-8 派生列配置

派生列后结果如图2-9。

当前显示 100 条 / 总共有 717840 条数据 提示:点击单元格可查看超出的内容

| # 区域D | # 日期小时     | # 人流量指数 | A <sub>0</sub> 日期 | A <sub>0</sub> 小时 |
|-------|------------|---------|-------------------|-------------------|
| 1     | 2020011700 | 1.8     | 2020-01-17        | 00                |
| 1     | 2020011701 | 1.5     | 2020-01-17        | 01                |
| 1     | 2020011702 | 1.3     | 2020-01-17        | 02                |
| 1     | 2020011703 | 1.3     | 2020-01-17        | 03                |
| 1     | 2020011704 | 1.7     | 2020-01-17        | 04                |
| 1     | 2020011705 | 1.2     | 2020-01-17        | 05                |
| 1     | 2020011706 | 2.9     | 2020-01-17        | 06                |
| 1     | 2020011707 | 4.7     | 2020-01-17        | 07                |
| 1     | 2020011708 | 6.1     | 2020-01-17        | 08                |
| 1     | 2020011709 | 5.2     | 2020-01-17        | 09                |
| 1     | 2020011710 | 5.5     | 2020-01-17        | 10                |
| 1     | 2020011711 | 6.2     | 2020-01-17        | 11                |
| 1     | 2020011712 | 7.1     | 2020-01-17        | 12                |
| 1     | 2020011713 | 7.9     | 2020-01-17        | 13                |

注意: 表头中#表示特征列, \*表示标签列

表头真名 表头别名

图2-9 转换后的日期时间格式

根据转换后的日期时间格式可以衍生字段“weekday”，表示当天属于一周中的第几天。我们再接入一个 派生列 节点，派生列配置如图2-10所示。

派生列配置

输入关键字搜索列

STRING

日期

小时

DOUBLE

人流量指数

INTEGER

区域D

日期小时

函数

累积

字符拼接(拼接符)

字符拼接

字符截取

小数点保留

内容提取(年)

内容提取(月)

添加/编辑表达式

1 dayofweek([日期])-1

派生列名: weekday 操作: 确定 取消

weekday

dayofweek([日期])-1

确定 取消

图2-10 派生列配置

“weekday” 字段衍生后结果如图2-11。

⌚当前显示 100 条 / 总共有 717840 条数据 [提示:点击单元格可查看超出的内容](#)

| # 区域ID | # 日期小时     | # 人流量指数 | 日期         | 小时 | # weekday |
|--------|------------|---------|------------|----|-----------|
| 1      | 2020011700 | 1.8     | 2020-01-17 | 00 | 5         |
| 1      | 2020011701 | 1.5     | 2020-01-17 | 01 | 5         |
| 1      | 2020011702 | 1.3     | 2020-01-17 | 02 | 5         |
| 1      | 2020011703 | 1.3     | 2020-01-17 | 03 | 5         |
| 1      | 2020011704 | 1.7     | 2020-01-17 | 04 | 5         |
| 1      | 2020011705 | 1.2     | 2020-01-17 | 05 | 5         |
| 1      | 2020011706 | 2.9     | 2020-01-17 | 06 | 5         |
| 1      | 2020011707 | 4.7     | 2020-01-17 | 07 | 5         |
| 1      | 2020011708 | 6.1     | 2020-01-17 | 08 | 5         |
| 1      | 2020011709 | 5.2     | 2020-01-17 | 09 | 5         |
| 1      | 2020011710 | 5.5     | 2020-01-17 | 10 | 5         |
| 1      | 2020011711 | 6.2     | 2020-01-17 | 11 | 5         |
| 1      | 2020011712 | 7.1     | 2020-01-17 | 12 | 5         |
| 1      | 2020011713 | 7.9     | 2020-01-17 | 13 | 5         |

注意: 表头中\*表示特征列, \*表示标签列

表头真名 ☒ 表头别名

图2-11 weekday字段

数据变换

由于原始的重点区域人群密度表只提供了历史20200117-20200215的每天分小时人流量。要做回归预测，需给该目标值构造特征，具体构造特征方法为：利用日期、小时、weekday、区域和区域类型的人流量指数和迁徙指数统计值聚合特征，如最小值、最大值、均值、总和等；

我们接入 聚合 节点，对日期、小时、weekday、区域、区域类型分别作Group操作，人流量指数、迁徙指数作Min、Max、Avg、Sum操作，如图2-12、2-13、2-14、2-15、2-16。

⌚聚合配置

添加聚合:

选择字段

结果列名(可选)

操作(可选)

+

| 已选字段(别名) | 结果列名               | 操作    |   |
|----------|--------------------|-------|---|
| # 区域     | Group_区域           | Group | 🗑 |
| # 人流量指数  | Group_区域_Min_人流量指数 | Min   | 🗑 |
| # 人流量指数  | Group_区域_Max_人流量指数 | Max   | 🗑 |
| # 人流量指数  | Group_区域_Avg_人流量指数 | Avg   | 🗑 |
| # 人流量指数  | Group_区域_Sum_人流量指数 | Sum   | 🗑 |

确定

取消

图2-12 根据区域聚合人流量指数

聚合配置

×

添加聚合:

选择字段

▼

结果列名(可选)

操作(可选)

▼

+

| 已选字段(别名) | 结果列名                | 操作    |  |
|----------|---------------------|-------|--|
| Ab_区域类型  | Group_区域类型          | Group |  |
| #人流量指数   | Group_区域类型_Min_人流量指 | Min   |  |
| #人流量指数   | Group_区域类型_Max_人流量指 | Max   |  |
| #人流量指数   | Group_区域类型_Avg_人流量指 | Avg   |  |
| #人流量指数   | Group_区域类型_Sum_人流量指 | Sum   |  |

确定

取消

图2-13 根据区域类型聚合人流量指数

聚合配置

×

添加聚合:

选择字段

▼

结果列名(可选)

操作(可选)

▼

+

| 已选字段(别名) | 结果列名               | 操作    |  |
|----------|--------------------|-------|--|
| Ab_小时    | Group_小时           | Group |  |
| #人流量指数   | Group_小时_Min_人流量指数 | Min   |  |
| #人流量指数   | Group_小时_Max_人流量指数 | Max   |  |
| #人流量指数   | Group_小时_Avg_人流量指数 | Avg   |  |
| #人流量指数   | Group_小时_Sum_人流量指数 | Sum   |  |

确定

取消

图2-14 根据小时聚合人流量指数

聚合配置

×

添加聚合：

选择字段

结果列名(可选)

操作(可选)

+

| 已选字段(别名)  | 结果列名                   | 操作    |   |
|-----------|------------------------|-------|---|
| # weekday | Group_weekday          | Group | 🗑 |
| # 人流量指数   | Group_weekday_Min_人流量指 | Min   | 🗑 |
| # 人流量指数   | Group_weekday_Max_人流量指 | Max   | 🗑 |
| # 人流量指数   | Group_weekday_Avg_人流量指 | Avg   | 🗑 |
| # 人流量指数   | Group_weekday_Sum_人流量指 | Sum   | 🗑 |

确定

取消

图2-15 根据weekday聚合人流量指数

聚合配置

×

添加聚合：

选择字段

结果列名(可选)

操作(可选)

+

| 已选字段(别名) | 结果列名              | 操作    |   |
|----------|-------------------|-------|---|
| # 日期     | Group_日期          | Group | 🗑 |
| # 迁徙指数   | Group_日期_Min_迁徙指数 | Min   | 🗑 |
| # 迁徙指数   | Group_日期_Max_迁徙指数 | Max   | 🗑 |
| # 迁徙指数   | Group_日期_Avg_迁徙指数 | Avg   | 🗑 |
| # 迁徙指数   | Group_日期_Sum_迁徙指数 | Sum   | 🗑 |

确定

取消

图2-16 根据日期聚合迁徙指数

接入聚合节点之后，我们对聚合后的特征使用 JOIN 节点进行合并，合并后可接入 全表统计 节点查看所有特征字段的分布情况，如图2-17。

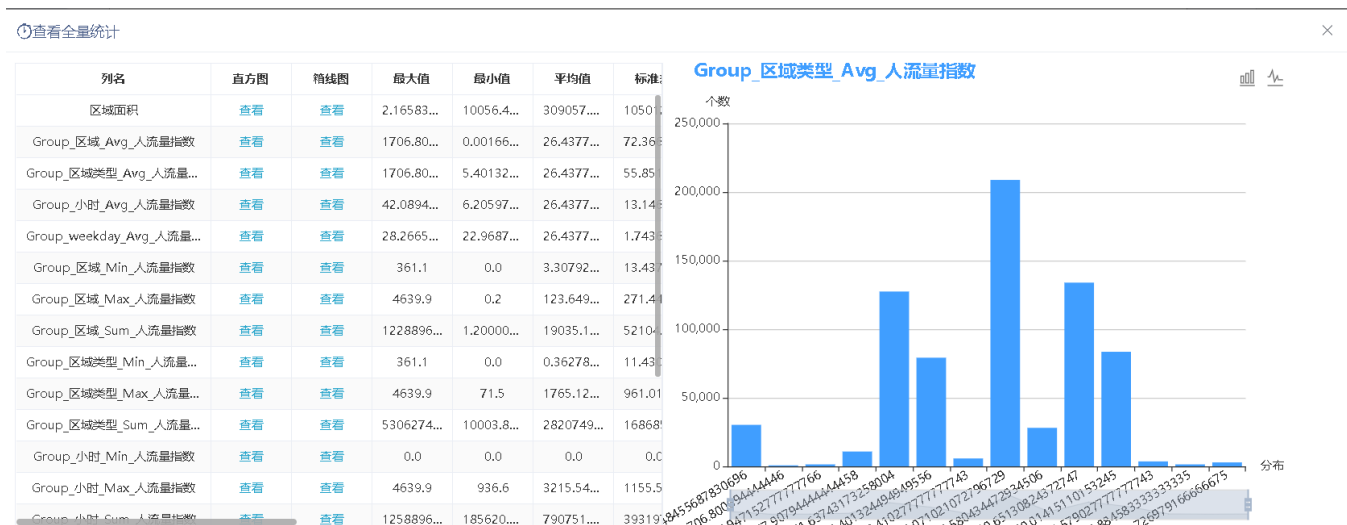


图2-17 指标数值分布情况

预处理流程图

整个预处理流程图如图2-18所示。

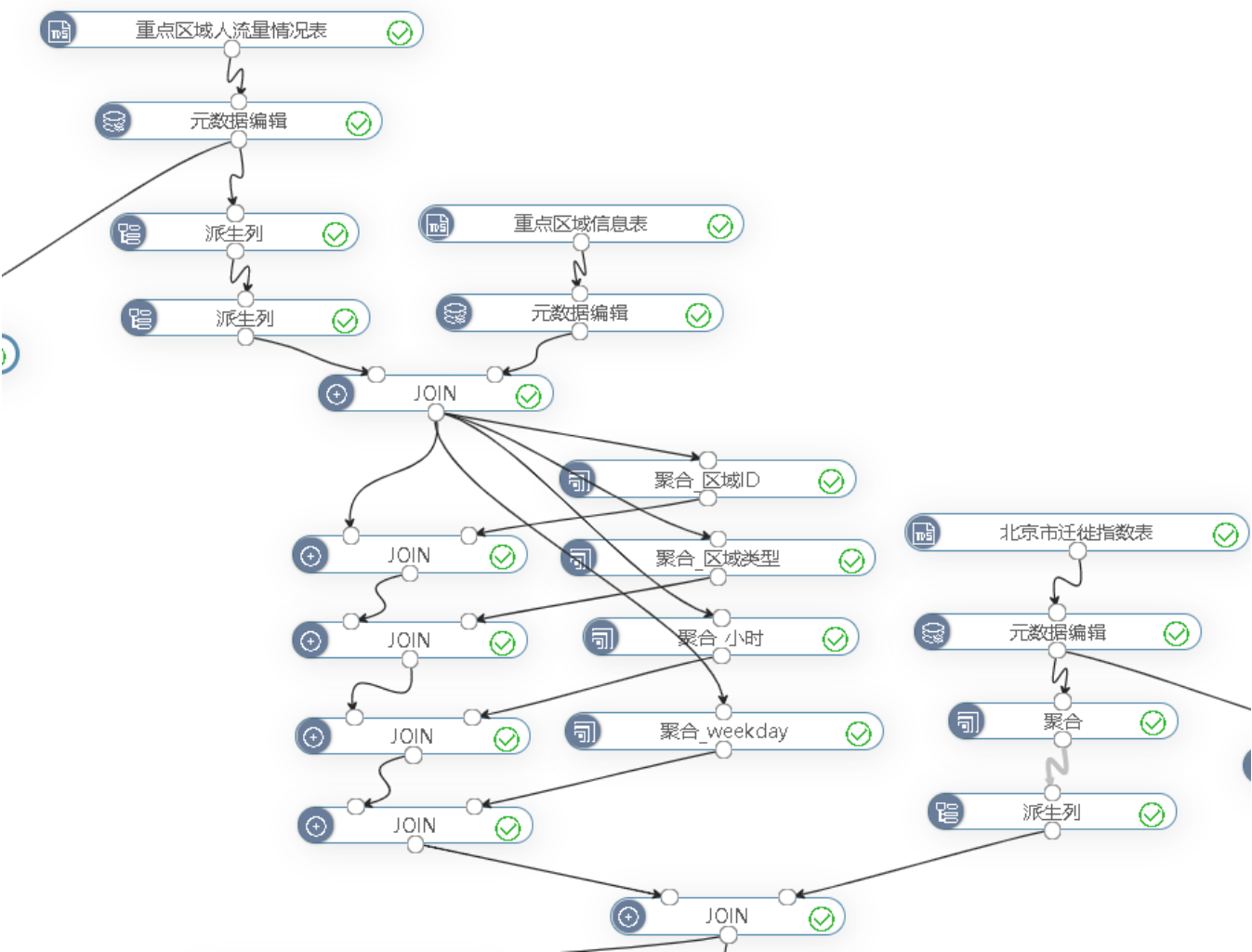


图2-18

建立模型

我们使用一个回归算法，这里选用 [梯度提升回归树算法](#)。整体的实验流程如图2-19所示。

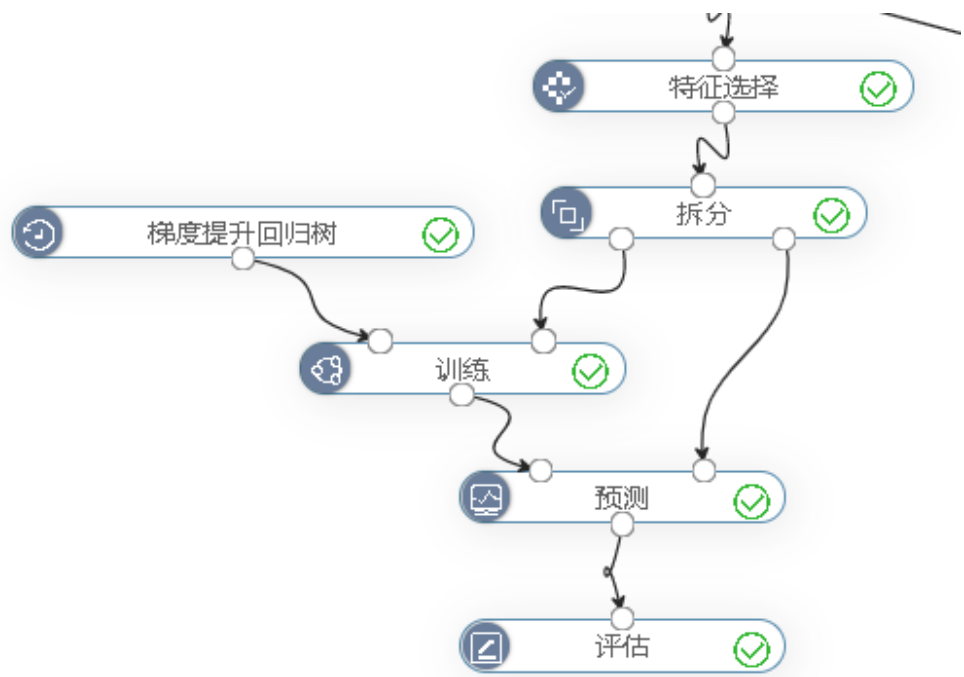


图2-19 人群密度回归预测模型

[特征选择](#) 节点，特征列选择数据变换步骤输出的聚合特征，如图2-20所示；



图2-20 特征选择特征列

特征选择的目标列选择人流量指数，如图2-21；

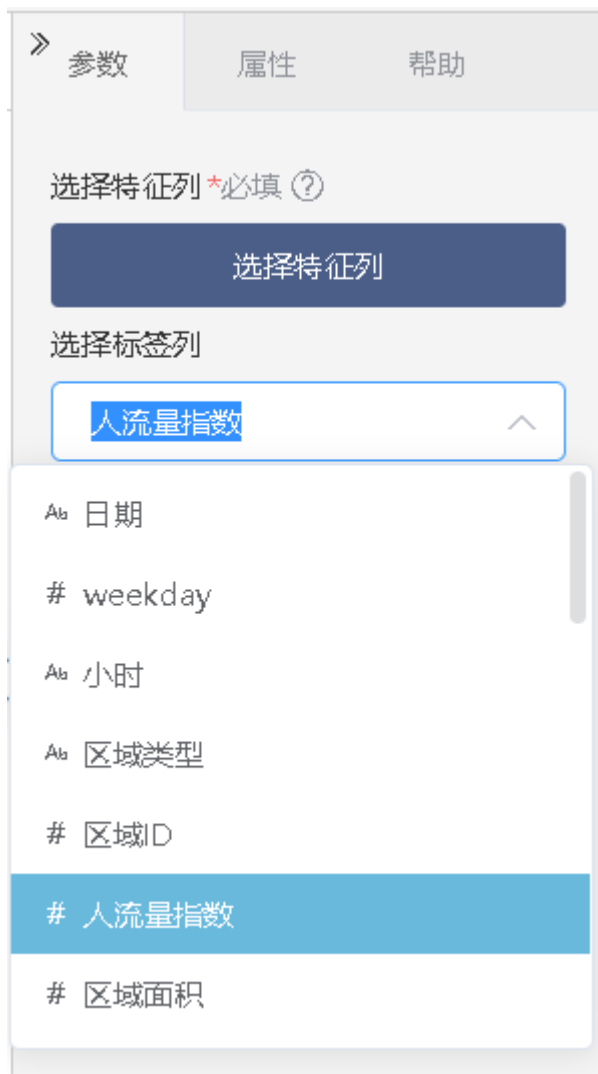


图2-21 选择目标列

[拆分](#) 节点使用默认参数配置，训练集与测试集的占比为7：3；

[梯度提升回归树](#) 的参数配置如图2-21所示：

>>

参数

属性

帮助

归一化

方法选择

标准化

参数

☐ 单位标准差归一化

☐ 平均数据中心化

分裂特征的数量 范围是>=2的整数 ?

128

树的深度 范围是[1,30]的整数 ?

6

计算信息增益的方式 ?

gini

图2-22 梯度提升回归树的参数配置

评估 节点的输出结果如图2-22所示，R2大概为0.96。

查看分析结果

×

| 指标                  | 值                  |
|---------------------|--------------------|
| rmse(均方根误差)         | 18.04585648334939  |
| mse(均方误差)           | 325.6529362176432  |
| r2(决定系数)            | 0.9625961381582396 |
| mae(平均绝对误差)         | 6.187756964387478  |
| adjusted r2(校正决定系数) | 0.9625924855222858 |

图2-23 模型评估结果

## 总结

本案例结合疫情期间重点区域人流量密度预测案例，重点介绍了回归预测分析在实际案例中的应用。案例还借助重点区域历史的人群密度，统计人流量指数和迁徙指数特征；建立模型预测重点区域未来的人群密度，掌握人员流动聚集的动向；针对人群密度较大的区域，做好紧急疫情防控工作。