|  |
| --- |
| **机器学习** |
| Zillow的房屋价值预测 |
| 主 研 人：赵翰宇  参 研 人： |
| 审 核 人：  方 向：工业算法案例研究  版 本 号：A |
|  |

**版本更新**

|  |  |  |  |
| --- | --- | --- | --- |
| **日期** | **版本** | **更新描述** | **作者** |
| 2019/4/19 | A | 初稿 | 赵翰宇 |

**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） |
| 2019/4/17 | Zillow’s Home Value Prediction (kaggle) | Zillow Prize是一项拥有100万美元大奖的竞赛，参与者将开发一种算法，对未来的房屋销售价格进行预测。 | 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 4](#_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 7](#_Toc22683_WPSOffice_Level3)

[3. 优秀算法思路 8](#_Toc12357_WPSOffice_Level1)

[3.1 方案一 8](#_Toc32293_WPSOffice_Level2)

[3.1.1 方案一特征提取 8](#_Toc16174_WPSOffice_Level3)

[3.1.2 方案一模型设计与建立 9](#_Toc32293_WPSOffice_Level3)

[3.1.3 方案一结果、排名等 10](#_Toc29922_WPSOffice_Level3)

[3.1.4 方案一算法流程图 10](#_Toc4787_WPSOffice_Level3)

[4. 算法比较 10](#_Toc22683_WPSOffice_Level1)

[表4-1 算法比较 10](#_Toc7253_WPSOffice_Level2)

[5. 总结与展望 11](#_Toc16174_WPSOffice_Level1)

[5.1 总结 11](#_Toc23198_WPSOffice_Level2)

[5.2 建模思路 11](#_Toc16619_WPSOffice_Level2)

1. 背景描述

自11年前首次发布以来，Zillow的Zestimate房屋估值已经震撼了美国房地产行业。

家庭通常是一个人在其一生中所做的最大和最昂贵的购买。确保房主拥有可信赖的方式来监控这项资产非常重要。Zestimate旨在为消费者提供尽可能多的有关房屋和住房市场的信息，标志着消费者首次免费获得此类房屋价值信息.

“Zestimates”是根据750万个统计和机器学习模型估算的房屋价值，这些模型分析每个房产的数百个数据点。而且，通过不断提高误差中位数（从开始时的14％到今天的5％），Zillow已成为美国最大，最值得信赖的房地产信息市场之一，也是影响力的领先典范。机器学习。

Zillow Prize是一项拥有100万美元大奖的竞赛，它正在挑战数据科学界，以帮助进一步提高Zestimate的准确性。获胜算法将影响美国各地110M家庭的房屋价值

**1.1 竞赛赛题描述**

在这场价值百万美元的竞赛中，参与者将开发一种算法，对未来的房屋销售价格进行预测。比赛分为两轮，2017年5月24日开幕的资格赛和2018年2月1日开幕的100个顶级资格赛的私人赛。在资格赛中，你将建立一个模型来改善Zestimate残差。在最后一轮中，您将从头开始构建房屋估价算法，使用外部数据源来帮助设计新功能，使您的模型在竞争中占据优势。

由于房地产交易数据是公共信息，因此在每轮竞赛结束后将有三个月的销售跟踪期，您的预测将根据房屋的实际销售价格进行评估。最终排行榜在销售跟踪期结束前不会公布。

**1.2 评估指标描述**

本次竞赛的评估指标为预测的log误差值与实际log误差值的MAE。log误差值定义如下：

将logerror记录在交易培训数据中。如果在该段时间内没有对某个属性进行交易，则忽略该行，并且不计入MAE的计算中。

1. 数据来源及描述性统计分析

**2.1 大赛数据来源**

数据来源于房地产交易数据公共信息

这里是数据下载的超链接：https://www.kaggle.com/c/6649/download-all

**2.2 数据的描述性统计**

**2.2.1 数据基本情况描述：**

properties\_2016.csv - 具有2016年主页功能的所有属性。注意：某些2017年新属性除了parcelid之外没有任何数据。当properties\_2017.csv可用时，应填充这些数据点。

properties\_2017.csv - 所有具有2017年主页功能的房产（2017年2月10日发布）

train\_2016.csv - 2016年1月1日至2016年12月31日期间交易的训练集

train\_2017.csv - 训练活动时间为2017年1月1日至9月15日（2017年2月10日发布）

sample\_submission.csv - 格式正确的示例提交文件

**2.2.2 数据字段介绍：**

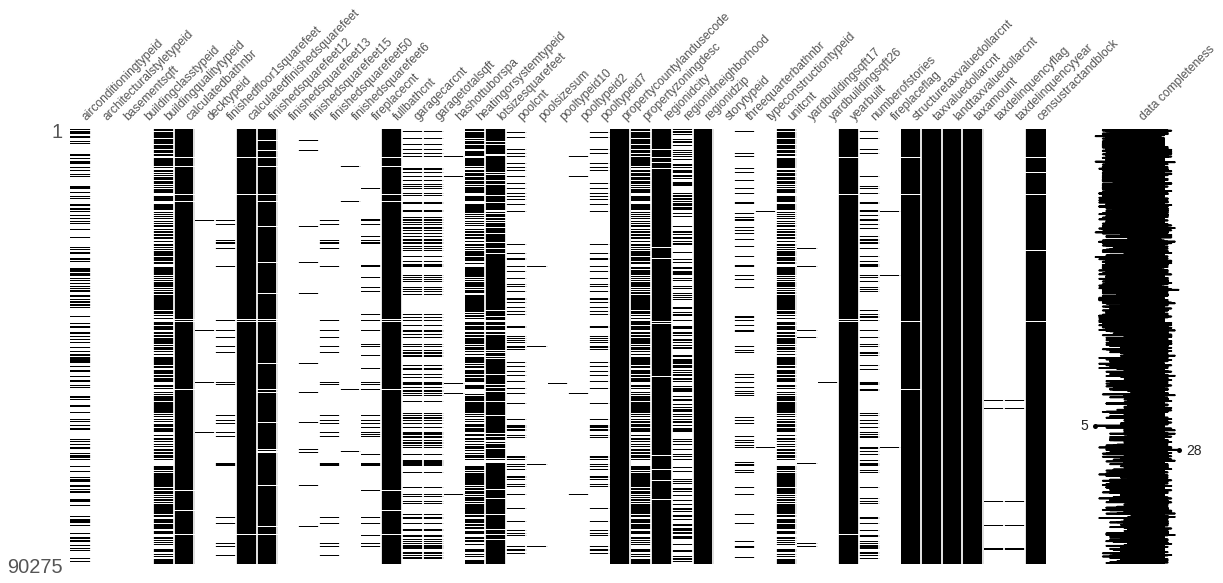
表2-1 train\_2016.csv数据表字段介绍

|  |  |  |  |
| --- | --- | --- | --- |
|  | **变量含义** | **变量类型** | **缺失率** |
| **parcelid** | 唯一标志符 | 离散 | 0% |
| **logerror** | 误差 | 离散 | 0% |
| **transactiondate** | 交易日期 | 日期 | 0% |

表2-2 properties\_2016.csv数据表字段介绍

|  |  |
| --- | --- |
| **变量名称** | **变量描述** |
| 'airconditioningtypeid' | 家用冷却系统类型 |
| 'architecturalstyletypeid' | 住宅的建筑风格（即牧场、殖民地、分层等） |
| 'basementsqft' | 地面以下或部分以下的完工生活区 |
| 'bathroomcnt' | 家庭浴室数量，包括部分浴室 |
| 'bedroomcnt' | 家中卧室数量 |
| 'buildingqualitytypeid' | 从最佳到最差的建筑物整体状况评估 |
| 'buildingclasstypeid' | 建筑框架类型（钢框架、木框架、混凝土/砖） |
| 'calculatedbathnbr' | 家庭浴室数量，包括部分浴室 |
| 'decktypeid' | 包裹上的甲板类型 |
| 'threequarterbathnbr' | 室内3/4卫生间数量（淋浴+水槽+卫生间） |
| 'finishedfloor1squarefeet' | 家庭一（入口）层已完工生活区的大小 |
| 'calculatedfinishedsquarefeet' | 住宅竣工总面积计算 |
| 'finishedsquarefeet6' | 基底未完工和完工区域 |
| 'finishedsquarefeet12' | 成品生活区 |
| 'finishedsquarefeet13' | 周边生活区 |
| 'finishedsquarefeet15' | 总面积 |
| 'finishedsquarefeet50' | 家庭一（入口）层已完工生活区的大小 |
| 'fips' | 联邦信息标准码 |
| 'fireplacecnt' | 家中的壁炉数量 |
| 'fireplaceflag' | 这个家里有壁炉吗 |
| 'fullbathcnt' | 家中有完整浴室（水槽、淋浴+浴缸和卫生间）的数量 |
| 'garagecarcnt' | 地上所有车库（包括附属车库） |
| 'garagetotalsqft' | 地上所有车库（包括附属车库）的总面积 |
| 'hashottuborspa' | 家里有热水浴缸或温泉吗 |
| 'heatingorsystemtypeid' | 家庭供暖系统类型 |
| 'latitude' | 包裹中间的纬度乘以10e6 |
| 'longitude' | 包裹中间的经度乘以10e6 |
| 'lotsizesquarefeet' | 地块面积（平方英尺） |
| 'numberofstories' | 家庭拥有的故事或层次数 |
| 'parcelid' | 包裹（批次）的唯一标识符 |
| 'poolcnt' | 地块上的水池数量 |
| 'poolsizesum' | 地产上所有游泳池的总面积 |
| 'pooltypeid10' | 温泉浴缸 |
| 'pooltypeid2' | 带水疗/热水浴缸的游泳池 |
| 'pooltypeid7' | 不带热水浴缸的游泳池 |
| 'propertycountylandusecode' | 县土地使用规范，即县级分区 |
| 'propertylandusetypeid' | 物业分区的土地用途类型 |
| 'propertyzoningdesc' | 该物业的许可土地用途（分区）说明 |
| 'rawcensustractandblock' | 人口普查区域和块ID组合-还包含扩展的块组分配 |
| 'censustractandblock' | 人口普查区域和块ID组合-还包含扩展的块组分配 |
| 'regionidcounty' | 物业所在县 |
| 'regionidcity' | 物业所在城市 |
| 'regionidzip' | 属性所在的邮政编码 |
| 'regionidneighborhood' | 物业所在的社区 |
| 'roomcnt' | 主要住宅的房间总数 |
| 'storytypeid' | 多层房屋的楼层类型（即地下室和主楼层、分层、阁楼等） |
| 'typeconstructiontypeid' | 用什么建筑材料建造房子 |
| 'unitcnt' | 结构内置的单元数（即2=双面，3=三层等） |
| 'yardbuildingsqft17' | 院子里的露台 |
| 'yardbuildingsqft26' | 堆场仓库/建筑物 |
| 'yearbuilt' | 主要住宅建造年份 |
| 'taxvaluedollarcnt' | 包裹的纳税总价值 |
| 'structuretaxvaluedollarcnt' | 地上已建结构的评估价值 |
| 'landtaxvaluedollarcnt' | 宗地土地面积评估值 |
| 'taxamount' | 该课税年度所评定的物业税总额 |
| 'assessmentyear' | 物业税评税年度 |
| 'taxdelinquencyflag' | 该地块的物业税已于2015年到期。 |
| 'taxdelinquencyyear' | 未缴物业税的到期年份 |

**2.2.3 数据描述性统计**

 该竞赛数据维度较高，半数特征数据缺失率极高，其余非连续数值型特征的分布可看为正态分布。

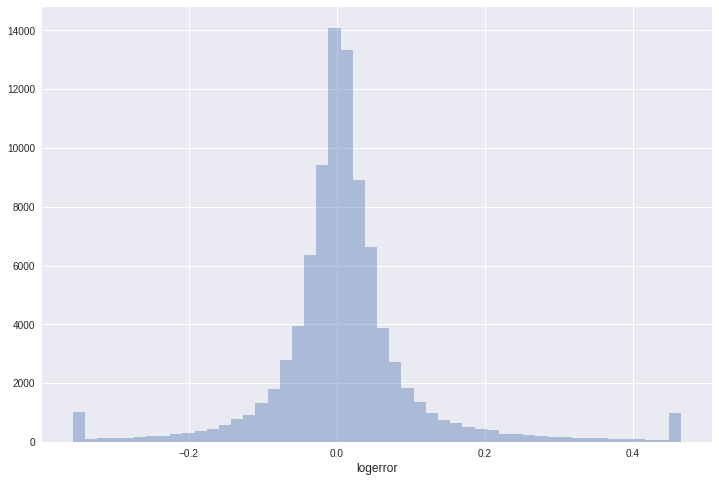
图2-1 部分特征缺失率

图2-2 筛选相关性大的特征后的数据分布

3. 优秀算法思路

**3.1 方案一**

**3.1.1 方案一特征工程**

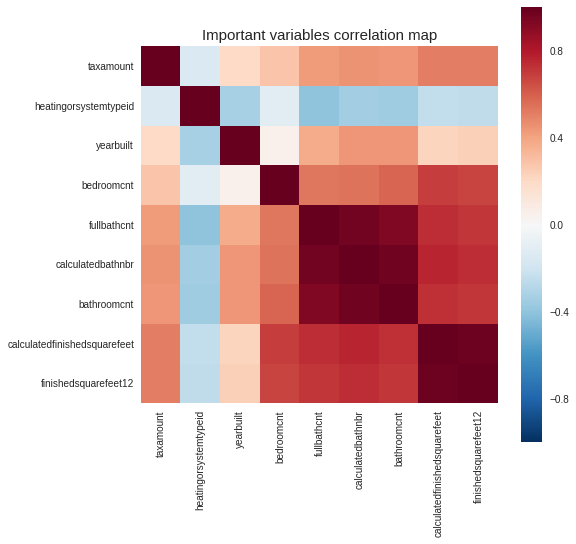
 方案一先对每个特征的数据进行了缺失值统计，然后进行了相关性分析。

图3-1 筛选后的特征相关性

方案建立了3个新特征：

1，每平方英尺结构值

2，平均房间大小

3，结构相对于土地的价值

方案还使用一些直觉对变量进行分类处理处理：如果它们具有自然顺序就分为一组。例如：空调类型-根据质量的基本评估生成整数值。如果它们相似就分为一组。如：四联、联排别墅

方案一用两种方法从房子提取的信息。

1，邻域平均：平均特征值、家园，差离平均值与特征、房屋

2，每个邻域：这是一个很多的变量。需要限制两两的关系。

方案同时考虑到有限的数据集，所以对拟合施加约束可能是有意义的，以便获得更可靠的结果。于是创建使用在其整个价值范围内都很重要的功能的偏好，使其在经济上更重要（与噪音相比）。权重由整个数据集确定，而不是数据集的一部分。

**3.1.2 方案一模型设计与建立**

方案一建立了GBDT模型，将评估值与其他功能匹配。将剩余部分用作新功能。

街道宽度（交通代理）。

家庭所面对的方向。

附近的密度。

靠近“公园”（空置空间）

最近出售的附近房屋的比例。

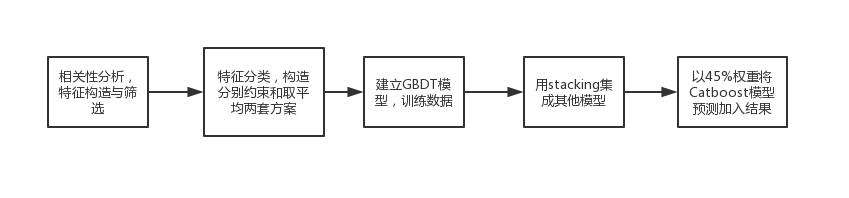
先前同一房屋销售

但方案一后来发现两类特征互相冲突，权重不好分配，训练效果不是很好，于是用了集成的stacking。结合了LightGBM, XGBoost, ANN (MLP), OLS, Constrained OLS等学习方法，建立三层模型，在第一层建模中分离的特征集，来处理特征的不同强度的优先级。训练多种模型后取平均再以45%的权重加入Catboost模型预测的结果。最终得到预测值。

**3.1.3 方案一结果、排名等**

此方案获得了第17名，公共排行榜上为0.06398，排名第20位，私人排行榜上为0.07479，排名第17位。

**3.1.4 方案一算法流程图**



**图3-2**

4. 算法比较

**表4-1 算法比较**

|  | **评估指标** | **特征工程** | **基础算法** | **基本库** |
| --- | --- | --- | --- | --- |
| **算法1** | **Logerror** | **构造新特征，删除相关性小的特征** | **Lgb，Xgboost，Catboost，GBDT，stacking** | **Python的Sklearn** |

算法一在特征建立上，先进行了相关性分析，选取了相关性大的特征。然后定义了特征的分类，将分类后的特性进行平均和分开约束两种操作。同时建立了一些特征，使用有关邻近房产的信息，包括与销售无关的~3M观测资料。因此，特征对预测值更有针对性

在模型建立上，算法一对第一个问题先进行了GBDT模型训练，选出了两类权重，加上了限制条件。在第二问上，算法一用了stacking集成算法，将XGboost，lgb等模型预测结果加权平均，再融合Catboost，这样使得结果更稳健，模型效果更好。

5. 总结与展望

**5.1 总结**

此次房屋预测竞赛的关键在于特征的建立筛选，用集成算法能得到更好地结果。难点就在于对特征的处理和各模型的权重分配上。

**5.2 建模思路**

首先进行数据预处理，进行相关性分析，删除特征，以均值填充缺失值。其次将前一天的特征加入到后一天的特征中，以便分析时间序列对数据的影响。在算法上，用lgb算法，得到最终预测结果。