华中师范大学大学生

创新创业训练计划项目进展报告书

项目名称

基于网络搜索行为对洪山区商品房价格的短期预测

项目类别 创新训练

起止时间 2017 年 3 月—2017 年 12 月

项目负责人 蒲东齐

专　　业 软件工程 年级 2015

华中师范大学教务处制

阶段主要工作及成果：

（一） 项目研究进展情况:

（1） 算法学习阶段

时间：2017.03.01—2017.07.20

方式：老师指导；书籍材料，视频材料，同类论文材料自学

内容：R 语言的简单使用（语法及 R 包），机器学习有关的算法（逐步线性回归，岭回归，Lasso，决策树， bagging ，adaboost，随机森林）

（2） 选取变量和数据收集阶段

时间：2017.07.20-2017.08.01

范围：40 个关键词（分别涉及二手房，房贷利率，公积金，户型，建材，房地产，物业，装修，租房，相关网站等 10 个方面）

数据来源：百度指数，可以将百度用户的搜索行为记录下来。

数据时间跨度：2011 年 1 月 1 日——2017 年 8 月 23 日

数据处理：由于文本研究的是月度数据，要将关键词的天数据转化为月度数据，即按照日历中各月的天数进行简单汇总即可。

变量的选取：

1.简单手动筛选。首先先将数据为零的关键词，变化趋势不明显的关键词剔除。

2.线性回归筛选。将前两步选出的全部变量作为自变量对因变量进行简单线性回归，对模型进行多重共线性诊断，最后利用逐步回归及 AIC 准则选择最终的关键词。

（3） 误差评价函数:

$$MSE = \frac{1}{n}\sum (Y - \hat{Y})^2$$

$$NMSE = \frac{MSE}{\mathrm{var}(Y)} = \frac{n-1}{n}\frac{\sum (Y - \hat{Y})^2}{\sum (Y - \bar{Y})^2}$$

其中：MSE 是指参数估计值与参数真值之差平方的期望值。

NMSE是指均方误差与方差的比值

（4） 划分训练集和测试集

按照训练集占 75%，测试集占 25%的原则划分（随机抽样），来避免有系统性差异

（5） 模型选择阶段

线性回归模型：一元一次的回归曲线是用最小二乘法确定各个变量的系数，推广到多元其实也是一样的，最小二乘法（又称最小平方法）是一种数学优

化技术。它通过最小化误差的平方和寻找数据的最佳函数匹配。利用最小二乘法可以简便地求得未知的数据，并使得这些求得的数据与实际数据之间误差的平方和为最小。简单说就是使得方程真实值和预测值的残差平方和最小。逐步线性回归就是选取使得 AIC 最小的模型。

套索模型：普通线性模型遇到多重共线性的情况下，可以通过 Lasso 构建一个一阶的惩罚函数获得一个精炼的模型；通过最终确定一些指标（变量）的系数为零， 然后就直接淘汰掉，从而达到特征选择或者压缩变量的目的。擅长处理具有多重共线性的数据。 LASSO 回归复杂度调整的程度由参数λ来控制，λ越大对变量较多的线性模型的惩罚力度就越大，从而最终获得一个变量较少的模型。

回归树模型：训练模型的时候根据对回归树用平方误差最小化准则，得到能够使得两分枝的反应变量的变异最大的预测变量的某个值进行特征选择，生成二叉树，节点的每次分裂都把原样本空间划分为互不相交的两个子集。每次都根据某个局部标准，选择最好的划分。以此不断分叉，分到不能分为止，这样就得到了一个训练好的回归树，在进行预测的时候，变量的数值代入到回归树中，很据不同的叶子规则得到不同的预测值，然后返回所有叶子节点的均值。

袋装模型：是一种有放回的抽样方法（可能抽到重复的样本），从原始样本集中抽取训练集。每一次随机地从大小为 n 的训练集中抽取 n 个样本作为此

次的训练样本集（在训练集中，有些样本可能被多次抽取到，而有些样本可能一次都没有被抽中）。共进行 k 轮抽取，得到 k 个训练集。（k 个训练集之间是相互独立的），每次使用一个训练集得到一个模型，k 个训练集共得到 k 个模型。对回归问题，计算上述模型的均值作为最后的结果。（所有模型的重要性相同）

随机森林模型：由很多个决策树组合而成，单个决策树用随机方法构成；

学习集是原训练集中通过有放回抽样得到的自助样本，参与构建决策树的变量也是随机抽出，用 CART 训练决策树但是不剪枝，最后预测结果取决于各个决策树的平均值。随机性主要体现在两个方面：数据的随机性选取，以及待选特征的随机选取。数据的随机选取：从原始的数据集中采取有放回的抽样，构造子数据集；利用子数据集来构建子决策树。特征的随机选取，随机森林中的子树的每一个分裂过程并未用到所有的待选特征，而是从所有的待选特征中随机选取一定的特征，之后再在随机选取的特征中选取最优的特征。

（二）项目研究已取得的阶段性成果和收获:

（1）对洪山区商品房的房价可以进行一个较为有效的预测。

（2）论文的初稿已经完成。

（三）项目研究存在的主要问题及应对思路与措施:

1.关键词是根据实际经验选取的，带有一定的主观性。

2.关键词的选取量较少，如果更多的加入其他与房价相关的经济变量，模型的预测精度会更高一些。

3.特征的值只能精确到武汉市，不能很好的反映洪山区的情况。

4.百度搜索数据不能完全代表购房者和房地产投资者的整体行为信息，因为存在一部分人没有利用百度搜索相关信息，而是通过实体广告、朋友亲戚介绍、其他的搜索引擎等等 。

项目负责人（签字）_____

年　　月　　日

下阶段工作计划及预计成果：

（一） 主要任务及时间进程安排：

（1）学习支持向量机算法，进一步完善，验证是否可以得到一个更好的预测模型。

（2）进一步修改论文以及发表。

（3）准备结项材料。

（二）项目经费使用情况：

劳务费:1800 元整

数据收集费用：200 元整

书籍费用：500 元整

交通费：500 元整

项目负责人（签字）_____

年 月 日

指导教师意见：

指导教师（签字）_____

指导教师职称：_____

年　　月　　日

项目所在学院意见：

院、系（盖章）　　负责人（签字）_____ **年　　月　　日**
教务处意见：   **教务处（盖章）** **年　　月　　日**