|  |  |
| --- | --- |
| 设计报告成绩 |  |

河北工程技术学院

人工智能与大数据学院

设计报告题目：基于数据挖掘的房价预测实现

课程名称： 数据挖掘

学 期： 2022-2023学年第二学期

专业班级：2020级数据科学与大数据技术班

学生姓名：

学 号：

任课教师：

目录

[一、任务单 3](#_Toc137062892)

[二、正文 4](#_Toc137062908)

[一、研究目的及其意义 4](#_Toc137062909)

[（一） 研究目的 4](#_Toc137062910)

[（二）研究意义 4](#_Toc137062911)

[二、主要算法 5](#_Toc137062912)

[三、过程实现 5](#_Toc137062913)

[四、预测结果 8](#_Toc137062914)

[三、课程设计报告评分标准 9](#_Toc137062915)

# 一、任务单

|  |  |
| --- | --- |
| 题目 | 基于数据挖掘的房价预测实现 |
| 目标 | 利用数据挖掘算法实现房价预测 |
| 任  务  描  述 | 总体任务描述：针对房价预测热点，利用数据挖掘算法实现房价预测分析。  具体任务描述：  1、房价预测研究目的和意义；  2、房价预测的研究现状；  3、讲明具体采用某种数据挖掘算法实现，过程需详细描述。 |
| 任务下达教师 | 宋燕妮 |

# 二、正文

**基于数据挖掘的房价预测实现**

**学生姓名（宋体，四号字，加粗，居中）**

## 研究目的及其意义

### 研究目的

房地产市场是一个动态变化的市场，房价的波动受到多种因素的影响，如经济环境、政策变化、供需关系等。通过对房价的预测研究，可以帮助我们更好地理解市场的走势和趋势，为投资者、开发商、政府等提供决策参考。

房地产投资是一项重要的资产配置活动，准确的房价预测可以为投资者提供重要的信息和指导。投资者可以根据预测结果来评估投资回报、选择适宜的投资时机、分散风险等，提高投资决策的准确性和效果。

房价的预测研究可以增加市场的透明度，使市场参与者更清晰地了解市场的动态和价格趋势。这有助于消除信息不对称，减少市场操纵和不正当行为，促进市场的公平竞争和健康发展。

对房价进行预测研究还可以为政府提供重要的决策参考。政府可以根据预测结果来制定适当的房地产政策，调控市场供需关系，稳定市场价格，防范潜在的风险。

### （二）研究意义

房地产市场是一个重要的经济领域，房价的波动对整个经济体系有深远影响。通过房价预测，可以提供给投资者、开发商和政府决策者等相关方面的重要参考，帮助他们做出明智的决策和规划。

房产投资是人们的重要资产配置方式之一，而房价的波动对投资回报有着重要影响。通过房价预测，投资者可以更好地评估和预测未来的房地产市场走势，以便做出更明智的投资决策，降低风险，提高投资回报率。

对于购房者来说，房价预测可以提供有关房价趋势和价格变动的信息，帮助他们做出购房决策。购房者可以根据预测结果来选择适当的购房时机，避免高峰期购房或者错过低价购房的机会。

政府和监管机构可以利用房价预测研究结果，监测房地产市场的运行情况，及时发现异常波动和潜在风险，采取相应的监管措施，维护市场的稳定和健康发展。

## 主要算法

XGBoost模型是多个基础模型的线性拟合。确定第t棵树最优，先确定第t-1棵树最优，以此类推。所以其目标函数为：

表示经过第t轮迭代后的模型预测值， 表示已知t-1个基础模型得到预测值; 表示第t个基础模型), 寻找一个使目标函数尽可能最大化降低的。因为前面的t-1颗树这时候已经固定了,找到之后，模型主要求解部分基本结束，再把正则项化简，即可求解出预测房价。

求解出训练数据集和测试数据集的平均绝对误差。

## 过程实现

读取数据。

|  |
| --- |
| df = pd.read\_csv('./MELBOURNE\_HOUSE\_PRICES\_LESS.csv', encoding='utf-8')  df |

选取特征值。选取和房价预测相关的特征值为Rooms, Type, Postcode, Propertycount, Distance, Price。

|  |
| --- |
| dd = df[['Rooms', 'Type', 'Postcode', 'Propertycount', 'Distance', 'Price']] |

查看数据类型。

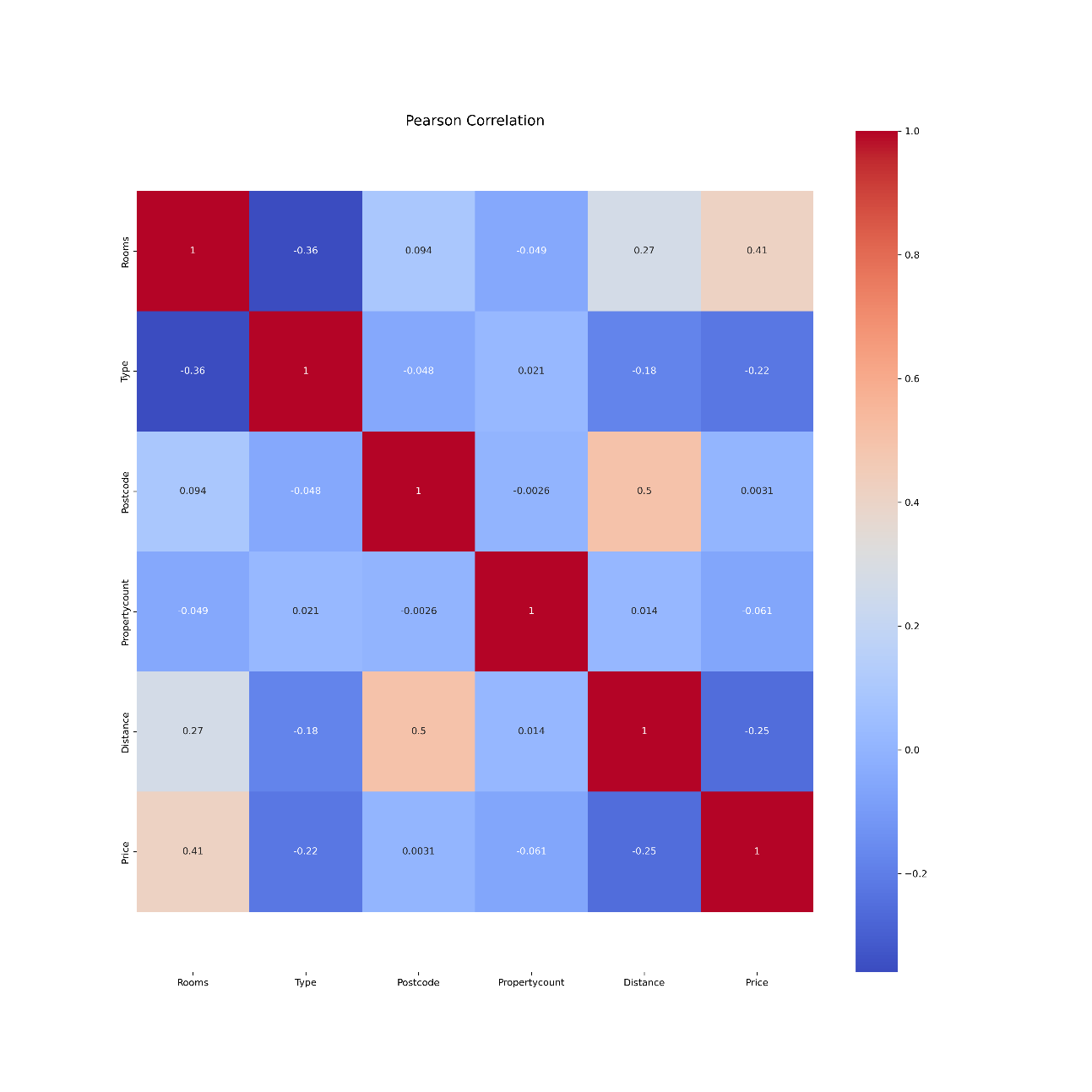
|  |
| --- |
| dd.info() |

将数据格式转换。

|  |
| --- |
| dd['Type'].value\_counts()  dd['Type'] = dd['Type'].map( {'h': 1, 'u': 2, 't' :3} ).astype(int)  dd.head() |

查看数据。

|  |
| --- |
| dd.describe() |

对数据进行相关性分析，并画出相关性热力图如图3-1所示。

|  |
| --- |
| colormap = 'coolwarm'  plt.figure(figsize=(16,16))  plt.title('Pearson Correlation', size=15)  ax = sns.heatmap(dd.corr(), cmap=colormap, annot=True)  bottom, top = ax.get\_ylim()  ax.set\_ylim(bottom + 0.5, top - 0.5)  plt.savefig("Pearson Correlation.png",dpi=300)  plt.show() |

图3-1 相关性热力图

检查数据是否含有空缺值。

|  |
| --- |
| dd.isnull().sum() |

删除数据所含空缺值并检验。

|  |
| --- |
| dd.dropna(axis=0, how='any', inplace=True)  dd.isnull().sum() |

选取Rooms、Type和Distance为特征。

|  |
| --- |
| X = dd[['Rooms', 'Type', 'Distance']]  y = dd['Price']  x = X.tail(100)  X = X.drop(X.tail(100).index)  y = y.drop(y.tail(100).index) |

分割数据集。

|  |
| --- |
| X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42) |

将数据进行标准化处理。

|  |
| --- |
| mean = X\_train.mean()  std = X\_train.std()  X\_train = (X\_train - mean) / std  X\_test = (X\_test - mean) / std  x = (x - mean) / std |

导入XGboost模型。

|  |
| --- |
| import xgboost  xgb\_reg = xgboost.XGBRegressor(n\_estimators=300, eta=0.1, max\_depth=5)  xgb\_reg.fit(X\_train, y\_train) |

求解出训练数据集的mae为205010.15182776932。

|  |
| --- |
| mean\_absolute\_error(xgb\_reg.predict(X\_train), y\_train) |

求解出测试数据集的mae为211398.52655459638。

|  |
| --- |
| mean\_absolute\_error(xgb\_reg.predict(X\_test), y\_test) |

求解出训练数据集的mse为119245457821.34677。

|  |
| --- |
| mean\_squared\_error(xgb\_reg.predict(X\_train), y\_train) |

求解出测试数据集的mse为122573553395.51479。

|  |
| --- |
| mean\_squared\_error(xgb\_reg.predict(X\_test), y\_test) |

求解出训练数据集的R2为0.4682533356016555。

|  |
| --- |
| r2\_score(xgb\_reg.predict(X\_train), y\_train) |

求解出测试数据集的R2为0.44210149453343206。

|  |
| --- |
| r2\_score(xgb\_reg.predict(X\_test), y\_test) |

## 预测结果

根据建立好的XGboost模型进行预测。

预测结果如表4-1所示。

表4-1 房价预测表

|  |  |
| --- | --- |
| 序号 | Price\_pre |
| 0 | 1.891248 |
| 1 | 1.891248 |
| 2 | 1.814422 |
| 3 | 2.014212 |
| 4 | 1.814422 |
| … | …. |
| 95 | 1.891248 |
| 96 | 1.814422 |
| 97 | 2.028729 |
| 98 | 2.028729 |
| 99 | 1.891248 |

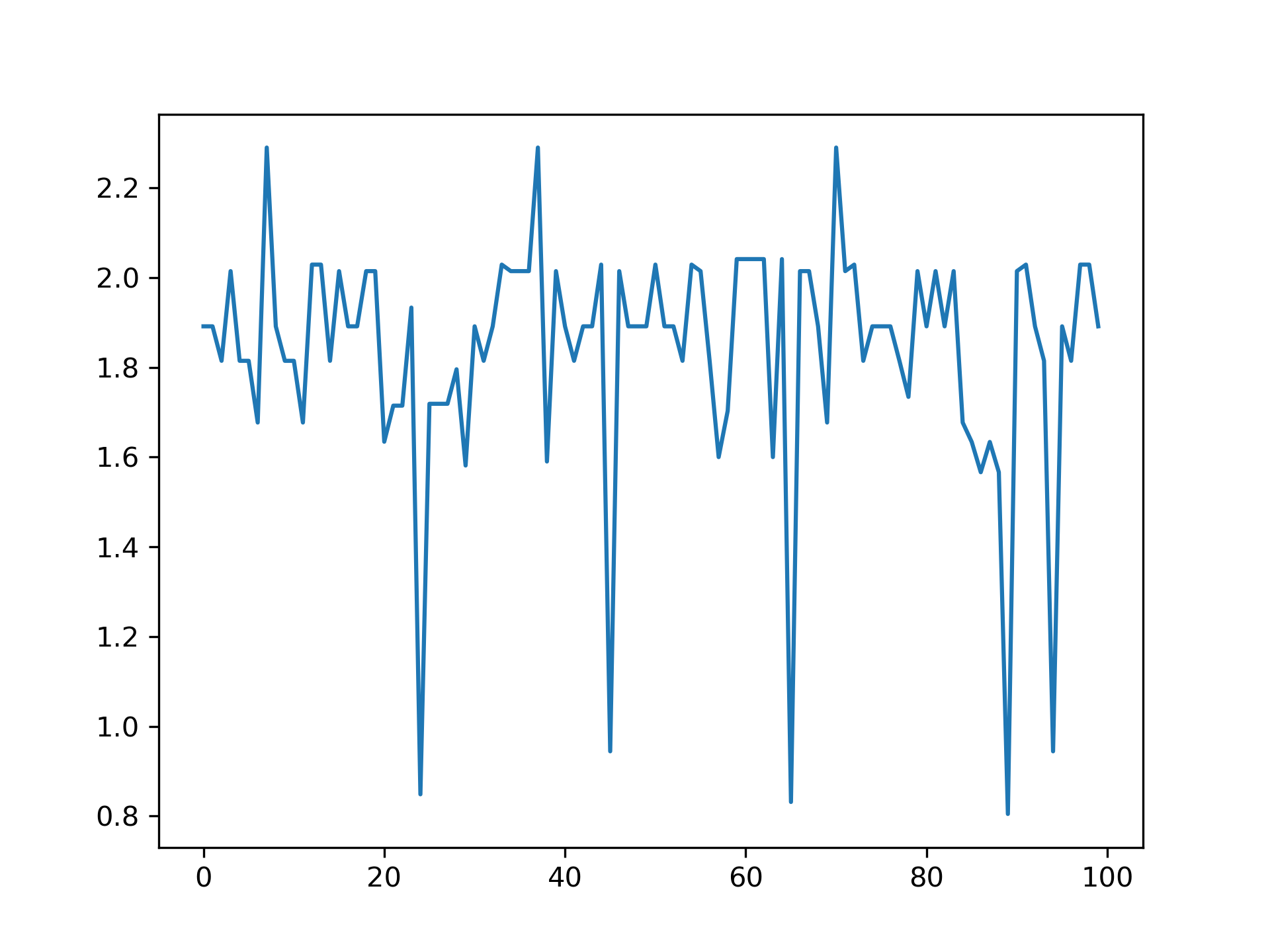
根据预测得到的数据画出房价预测折线图，如图4-1所示。

图4-1 房价预测

# 三、课程设计报告评分标准

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **指标** | **评价内容** | **评价等级（分值）** | | | | **得分** |
| **A** | **B** | **C** | **D** |
| **知识要求** | 知识运用熟练程度 | 10-7 | 6-4 | 3-1 | 0 |  |
| **内容**  **要求** | 内容分析是否深刻、透彻；是否详略得当；条理是否清晰；是否能够很好实现目标 | 40-32 | 31-24 | 23-12 | 11-0 |  |
| **结构**  **要求** | 结构是否完整、是否严谨；逻辑性是否体现；层次是否分明； | 20-16 | 15-11 | 10-6 | 5-0 |  |
| **思路创新能力** | 思路是否新颖，是否有独特的视角；创新性有否明显体现；是否能独立运用本门课程的有关理论知识分析和解决问题 | 20-16 | 15-11 | 10-6 | 5-0 |  |
| **规范**  **要求** | 书写、格式、体例是否规范；文字表达是否准确、流畅；是否符合学术道德规范；字数是否符合要求 | 10-8 | 7-6 | 5-3 | 2-0 |  |
| **其他说明** | **方案自己独立完成，严禁抄袭，出现雷同，一律以零分记录。** | | | | | |
| **评阅老师：**  **评阅时间： 年 月 日** | | | | | | |