上海大学
Shanghai University

shanghai university

利用神经网络预测上海租房价格

范舒舰
23121677

从数据清洗到
机器学习建模

上海大学
Shanghai University

实验流程

1 业务理解

明确问题：根据房屋属性预测月租金

→

2 数据理解

识别字段、样本规模与数据质量风险

→

3 数据获取

分析爬虫结构与已有数据快照

→

4 数据清洗

去重、解析数值、筛除异常、构建建模样本

5 特征工程

数值标准化 + 类别 one-hot + 标签 TF-IDF

→

6 模型训练

神经网络回归、验证集早停、学习率衰减

→

7 模型评估

独立测试集一次性评估 MAE

→

8 结果总结

解释误差、归纳创新点与后续优化方向

上海大学
Shanghai University

项目背景与研究目标

研究背景

上海租房市场样本规模大、房源异质性强，租金同时受区位、面积、户型、楼层、朝向、设施与文本描述等因素影响，单纯依赖经验难以统一刻画多维影响。

研究目标

输入房屋属性与文本信息，输出月租金。任务类型属于监督学习中的回归问题，核心指标为 MAE（平均绝对误差）。本次实验学习了完成环境搭建、清洗、建模与独立测试评估的规范。

原始样本	建模样本	输入特征	最终测试 MAE
29,982	15,621	120 维	1254.24 元
CSV 原始记录数	筛选 3000–12000 元区间	数值 + 类别 + 文本	新版规范流程结果

上海大学
Shanghai University

数据理解与样本概况

原始数据	去重后	建模样本	矩阵维度
29,982	18,931	15,621	120
CSV 快照行数	按 标题+地点+价格 去重	价格筛选后	最终输入特征数

原始字段的特点

面积、价格、房型、楼层等字段包含单位与自然语言，例如“72.38㎡”“2室1厅1卫”“高楼层”。地点字段是“区-商圈-小区”的层级描述，配套设施是以逗号分隔的标签列表，房源标签和描述则属于短文本。

为什么不能直接训练

神经网络只能接受数值矩阵，无法直接理解网页文本。因此必须完成：

① 数值字段抽取；

② 类别字段编码；

③ 文本字段向量化；

④ 缺失值与异常值处理。

上海大学
Shanghai University

数据获取与爬虫分析

原项目爬虫结构

Scrapy 工程主体为：parse → parse_region → parse_overview → parse_info。

1) 首页获取各区链接；

2) 进入每个区分页遍历；

3) 列表页提取标题、地点、房型；

4) 详情页补充价格、朝向、楼层、配套设施、描述等字段。

关键抓取字段

标题、地点、房屋类型、房源编号、价格、房源标签、租赁方式、面积、朝向、楼层、电梯、车位、用水、用电、燃气、采暖、配套设施、房源描述。

本次实验的取舍

考虑到当前网站反爬更严格、现网状态变化更快，本次实验重点放在“核心复现”：使用已有数据快照完成数据清洗、特征工程与模型训练。仅保留对爬虫进行分析，认识原项目数据结构，体现原项目完整的数据来源链路。

上海大学
Shanghai University

实验平台与环境搭建

平台

AutoDL

远程算力平台

显卡

RTX 2080 Ti

约 9.8 GB 显存

Python

3.10

conda 独立环境

深度学习框架

TensorFlow GPU

核心环境命令

```
conda create -n rent_gpu python=3.10 -y
source /root/miniconda3/bin/activate rent_gpu
pip install "numpy<2" pandas scikit-learn matplotlib seaborn
openpyxl jieba requests
pip install "tensorflow[and-cuda]"
```

工程性问题与处理

曾遇到 NumPy 2 与 matplotlib / TensorFlow 的兼容问题，最终通过新建环境并固定 numpy<2 解决。

激活 conda 环境时使用 source /root/miniconda3/bin/activate rent_gpu，避免 base 环境污染。

4

上海大学
Shanghai University

数据清洗与样本筛选

清洗规则

1) 统一中英文字段名；
2) 删除表头误入数据行；
3) 按“标题+地点+价格”去重；
4) 用正则提取面积、楼层、价格数值；
5) 从房屋类型中提取室/厅/卫；
6) 从地点字段中提取所属区；
7) 从配套设施文本中展开二值特征。

建模样本筛选

本次核心复现将租金限制在 3000–12000 元/月区间，以减少极端样本干扰并聚焦更常见的租赁市场主体样本。筛选后建模样本数为 15621 条。

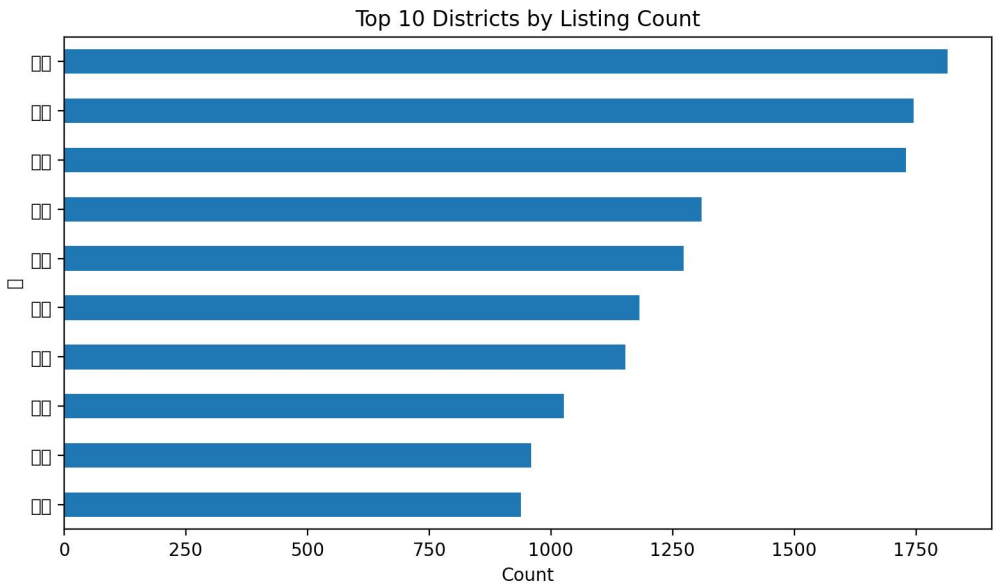
原始 CSV 29982 → **去重后 18931** → **价格筛选后 15621** → **训练集 12496** → **测试集 3124**

清洗后的核心结果：X_train = (12496, 120)，X_test = (3124, 120)。

上海大学
Shanghai University

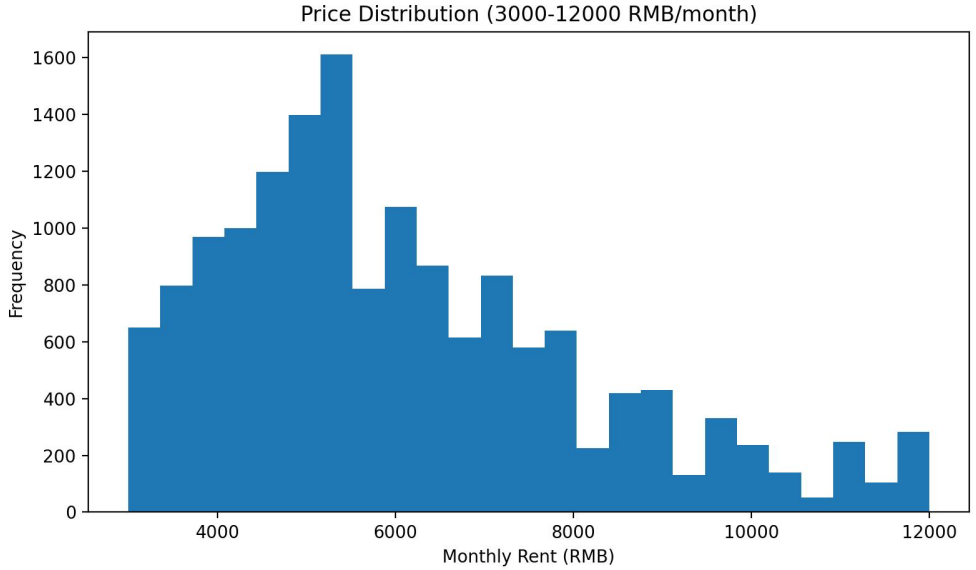
EDA发现

区域房源数量 Top 10



- 浦东、闵行、宝山等区房源数量最多，说明样本主要集中在租赁活跃区域。
- 价格分布明显右偏，说明大部分样本集中在中低至中等价位区间。
- 这意味着后续模型需要重点学习“主流价位样本”的规律。

价格分布（3000–12000 元）



- 房租并非均匀分布，极端高价样本较少。
- 因此选择 MAE 作为指标更直观，也更适合描述平均偏差。
- 在建模前限制价格区间，有助于提高训练稳定性。

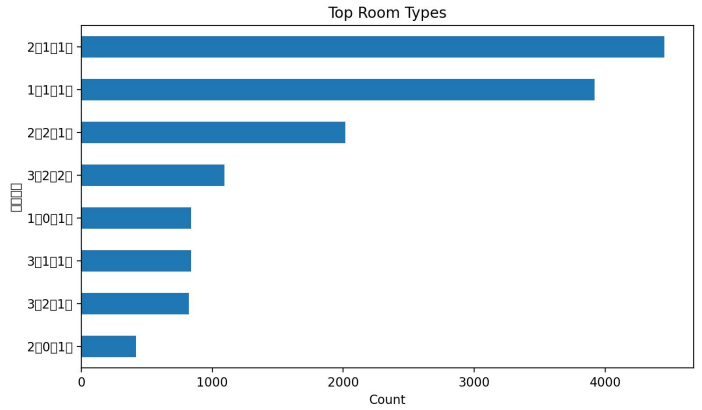
上海大学
Shanghai University

EDA发现

面积与租金关系

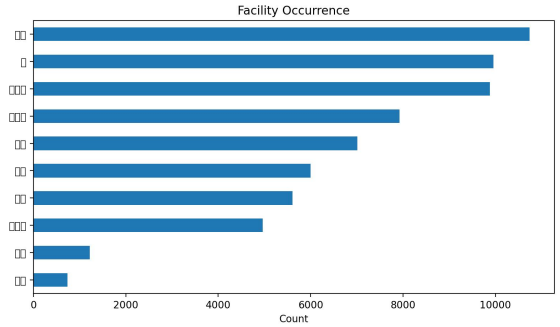
- 面积与租金总体呈正相关，但并非简单线性关系。
- 同等面积的房源因区位与装修不同，价格差异仍然明显。

主流户型



- 1室1厅1卫、2室1厅1卫等户型占比高。
- 户型信息适合拆解为室/厅/卫数值特征。

配套设施观察



- 空调、床、冰箱等设施出现频繁，可转化为二值特征。

上海大学
Shanghai University

特性工程设计

数值特征

面积、室、厅、卫、楼层。

处理方式：StandardScaler 标准化，使不同量纲的特征处于可比较尺度。

类别特征

区、朝向、租赁方式、燃气、采暖、用水、用电、电梯、车位，以及展开后的设施开关。

处理方式：One-Hot 编码。

文本特征

房源标签。

处理方式：先按逗号切分，再将标签串联为词序列，使用 TF-IDF 向量化，补充短文本信息。

原始字段 网页文本 → **数值/类别/文本 特征编码** → **120 维输入矩阵 X_train / X_test**

上海大学
Shanghai University

模型设计

网络结构

输入层：120 维特征

隐藏层：256 → 128 → 64

激活函数：ReLU

正则化：Dropout=0.2，L2=1e-5

输出层：1 个租金回归值

训练参数

batch_size = 32

epochs = 200

optimizer = Adam (lr = 1e-4)

loss = MAE

seed = 42

训练策略升级

1) 训练集内部划分 10% 验证集；

2) EarlyStopping 根据 val_mae 自动早停；

3) ReduceLROnPlateau 在验证集停滞时降低学习率；

4) ModelCheckpoint 自动保存最佳模型。

训练命令

```
python train_rent_nn_v2.py --data rent_prepared.npz --epochs 200 --batch-size 32 --save-model
rent_nn_best.keras
```

相较旧版“直接把测试集作为每轮验证集”的做法，改进训练流程更符合机器学习实验规范

上海大学
Shanghai University

训练结果

最佳验证轮次

56

best val mae

0.0972 万元

final test mae

0.1254 万元

结果如何理解

训练中最优的是验证集表现，因此保存的是“best model”。最终测试误差并不是最后一轮训练误差，也不等于最佳验证误差；它是在独立测试集上进行的一次最终泛化评估。

与旧版的区别

旧版 final test mae 约为 1198.74 元，但旧版训练过程中反复使用测试集做验证。新版虽然测试误差略高，但方法上更规范：测试集只使用一次，因此结果更可信。

上海大学
Shanghai University

模型分析

为什么 final test 不是“最佳值”

best val mae 和 final test mae 分别来自不同数据集。验证集参与了模型选择，因此更“乐观”；测试集从不参与选择，因此更能代表真实泛化能力。这也是规范实验中验证集与测试集分离的价值。

当前误差说明了什么

1254 元左右的 MAE 说明模型已经具备实际预测能力，但仍存在继续优化空间。

误差来源可能包括：

① 现有快照与原始项目数据版本不同；

② 地理信息未做经纬度增强；

③ 文本特征仅使用房源标签，未深入利用描述字段。

如何进一步提升

1) 扩充地理特征（经纬度、商圈、地铁）；

2) 引入更多文本字段；

3) 继续做误差分析与特征选择；

4) 尝试树模型或集成学习与神经网络对比。

上海大学
Shanghai University

结论

实验结论

1) 成功在 AutoDL GPU 环境下复现上海租房价格预测任务；

2) 完成了从 29982 条原始记录到 15621 条建模样本的清洗；

3) 成功构造 120 维输入特征；

4) 改进训练流程下 best val mae = 971.99 元，final test mae = 1254.24 元；

核心样本	特征维度	最优验证误差	最终测试误差
15621	120	971.99 元	1254.24 元
最终建模数据量	输入矩阵维度	第 56 轮	独立测试集

shanghai university

上海大学
Shanghai University

谢谢！

上善若水 海纳百川
大道明德 学用济世