

《人工智能概论》课程大作业（美国房价预测）

1 作业导读

本作业适用于《人工智能概论》课程结课后的课程大作业，以 Ames Housing（美国房价）数据集为基础，聚焦“基于多维度房屋属性预测房价 SalePrice”的回归任务，综合考察学生的数据分析与可视化、回归建模、模型评估与参数优化能力。

学生需结合房地产估值场景，理解地段、面积、质量、建造年代、配套设施等因素与房价之间的关系，运用课程所学回归算法解决真实问题，形成完整的 AI 建模闭环。

作业核心任务及分值分布如下：

任务 1	数据分析与特征工程	40 分
任务 2	回归模型搭建	30 分
任务 3	模型评估与优化	30 分

第 3 章为开放性课题作业，需在完成前三个任务后开展，旨在探索更先进的建模方法（如深度学习回归），完成后可作为加分项。

2 美国房价预测模型

2.1 作业背景

在房地产市场中，准确预测房屋价格对购房决策、资产评估、金融风控、城市规划均具有重要意义。传统人工估价方式依赖经验，主观性较强，难以在大规模样本中保持稳定精度。

通过机器学习方法，可利用房屋结构、面积、装修质量、地段等多源特征建立价格预测模型，为房价评估提供可解释、可量化的决策支持。

在本次作业中，需要建立回归模型回答以下问题：

“如何基于房屋的结构、面积、年代、配套等特征，预测其最终成交价格 SalePrice？”

2.1.1 数据集说明

- 训练集：train.csv（包含 SalePrice 标签）
- 测试集：test.csv（不包含 SalePrice）
- 数据集字段：20 个核心特征

数据集字段说明如下：

- OverallQual：整体材料和完工程度评分
- GrLivArea：地上居住面积
- GarageCars：车库容量（车位数）
- GarageArea：车库面积
- TotalBsmtSF：地下室总面积
- 1stFlrSF：一层面积

- FullBath: 全卫数量
- TotRmsAbvGrd: 地上总房间数
- YearBuilt: 建造年份
- YearRemodAdd: 翻修年份
- GarageYrBltn: 车库建造年份
- MasVnrArea: 砖石贴面面积
- Fireplaces: 壁炉数量
- BsmtFinSF1: 地下室已装修面积
- LotFrontage: 街道临接长度
- WoodDeckSF: 木平台面积
- 2ndFlrSF: 二层面积
- OpenPorchSF: 开放门廊面积
- HalfBath: 半卫数量
- LotArea: 地块面积

标签字段:

- SalePrice: 房屋成交价格 (回归目标)

2.2 作业要求

本次作业明确为回归任务，目标是预测 SalePrice 连续数值。可选模型包括:

- 线性回归 (基线)
- 随机森林回归 (非线性)
- 梯度提升回归 (XGBoost/LightGBM)
- 支持向量回归 (SVR)

评价指标建议以 MAE 为核心，同时报告 RMSE、 R^2 。

具体要求如下:

1. 技术要求: Python、scikit-learn、numpy、pandas、matplotlib、seaborn
2. 数据分析与特征工程:
 - 统计 SalePrice 分布特征 (均值、标准差、中位数、最值、偏度)
 - 分析核心特征与 SalePrice 的相关性并可视化 (热力图/条形图)
 - 缺失值处理 (均值/中位数/众数填充, 或删除缺失率过高字段)
 - 异常值检测 (箱线图、IQR、z-score 等)
 - 标准化/归一化 (适配线性回归、SVR)
3. 模型搭建: 至少构建 2 种回归模型, 测试集比例不低于 20%
4. 模型评估: 在测试集上报告 MAE、RMSE、 R^2 , 并对误差分布进行分析
5. 模型优化: 使用交叉验证、网格搜索/随机搜索进行参数优化
6. 提交材料: 代码、完整设计文档、方案说明 PPT

2.3 作业任务

2.3.1 任务 1 (40 分)：数据分析与特征工程

数据探索与理解 (15 分)

- 统计目标变量 SalePrice 的分布特征
- 分析 20 个核心特征与 SalePrice 的相关性，识别强关联特征

数据预处理 (15 分)

- 缺失值处理并说明策略
- 对需要的模型进行标准化/归一化处理

特征工程 (10 分)

- 基于相关性或模型重要度进行二次筛选
- 构造新特征 (示例)：
- $\text{HouseAge} = \text{YrSold} - \text{YearBuilt}$
- $\text{RemodAge} = \text{YrSold} - \text{YearRemodAdd}$
- $\text{TotalSF} = \text{TotalBsmtSF} + 1\text{stFlrSF} + 2\text{ndFlrSF}$

2.3.2 任务 2 (30 分)：回归模型搭建

建模思路设计 (10 分)

- 明确特征与房价关系假设
- 训练集/测试集划分并说明随机种子设置

回归模型搭建 (20 分)

- 至少完成 2 种模型训练与预测
- 对比模型泛化性能与训练时间

2.3.3 任务 3 (30 分)：模型评估与优化

多维度模型评估 (15 分)

- 比较 MAE、RMSE、 R^2 指标
- 绘制真实值-预测值散点图、残差分布图并分析误差来源

模型优化与验证 (15 分)

- 对主模型进行参数搜索并二次建模
- 说明优化前后性能变化与原因

3 开放性课题作业 (加分项)

在完成传统机器学习回归任务后，可采用 PyTorch、TensorFlow、Keras 或 MindSpore 搭建神经网络回归模型，对比线性回归/随机森林/梯度提升模型的效果并分析原因。

作业具体要求如下：

1. 考察维度：创意、完成度、完成质量、实用价值
2. 提交内容：调研分析报告、方案完整设计文档、方案说明 PPT、代码