Linear Model For House Pricing

Hongming Liang
2022201480

2025/04/03

Scaling

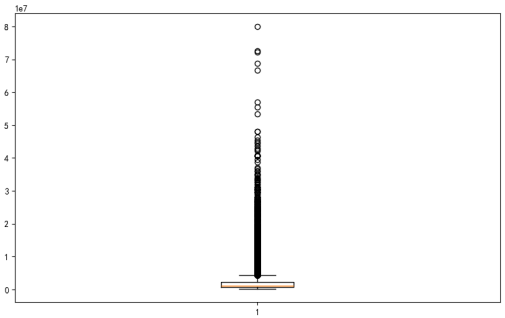

图: BoxPlot for *price*

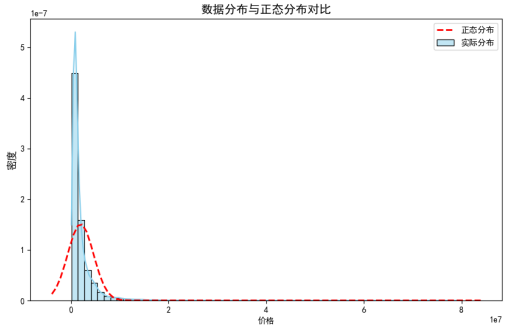

图: Distribution for *price*

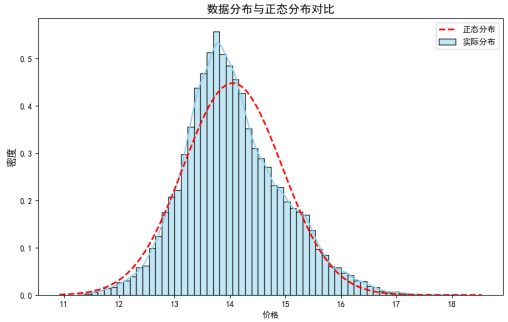

图: Distribution for $\ln$ *price*

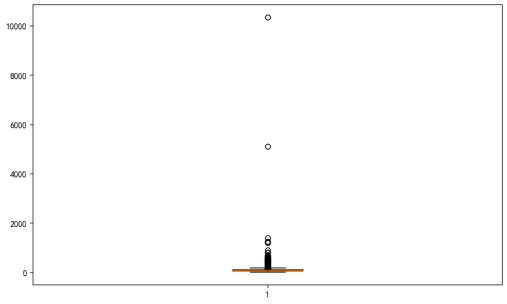

图: BoxPlot for *area*

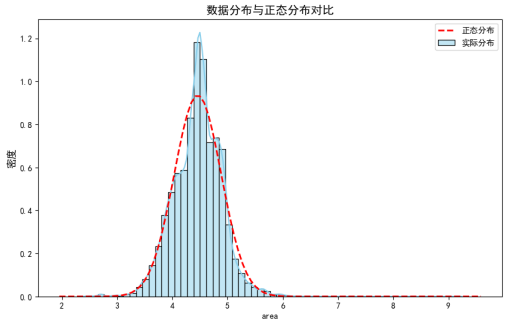

图: Distribution for $\ln$ *area*

Varible	*price*, *area*	$\ln()$
MAE	176341	156847
R^2	0.8559	0.9671
Score	58.12	81.00

表: Comparison Using OLS (Out-of-sample)

Features

- Numeric features: ' 建筑面积', ' 梯户比例', ' 绿化率', ' 停车位'。其中' 梯户比例' 提取字符串中的数值计算得到
- Categorical features: ' 板块', ' 城市', ' 环线', ' 房屋户型', ' 所在楼层', ' 小区名称', ' 建筑结构', ' 装修情况',' 房屋朝向', ' 配备电梯', ' 别墅类型', ' 交易权属', ' 房屋用途', ' 房屋年限', ' 产权所属', ' 区域'
- Special Care:
  - ' 环线': 创建城市与环线的交乘项。不同城市的环线存在异质性
  - ' 房屋朝向': 提取对应的字符串，创建了八个虚拟变量，即是否朝向' 东',' 南',' 西',' 北',' 东北',' 东南',' 西南',' 西北' 八个方向
  - ' 房屋户型': 提取字符串中的房间个数用于创建虚拟变量'n_ 室','n_ 厅','n_ 厨','n_ 卫'
  - 为了充分利用信息，类别特征的缺失值统一用' 未知' 来填补，例如' 配备电梯' 这一特征。

Scores

- Model:

$$\ln \textit{price}_i = \beta_0 + \beta_1 \ln \textit{area}_i + \gamma X_i + \delta_i + \epsilon_i$$

表: MAE for OLS, Lasso and Ridge

Metrics	In sample	Out of sample	Cross-validation	Datahub Score
OLS	143464	156847	1766221	81.00
Lasso	238477	236534	1733797	79.55
Ridge	143683	156802	1766258	80.99
Best	143464	156847	1766221	81.00

Note: The best model is OLS. Lasso with parameter $\alpha = 10^{-5}$. Ridge with parameter $\alpha = 0.01$. 1284 outliers were removed.