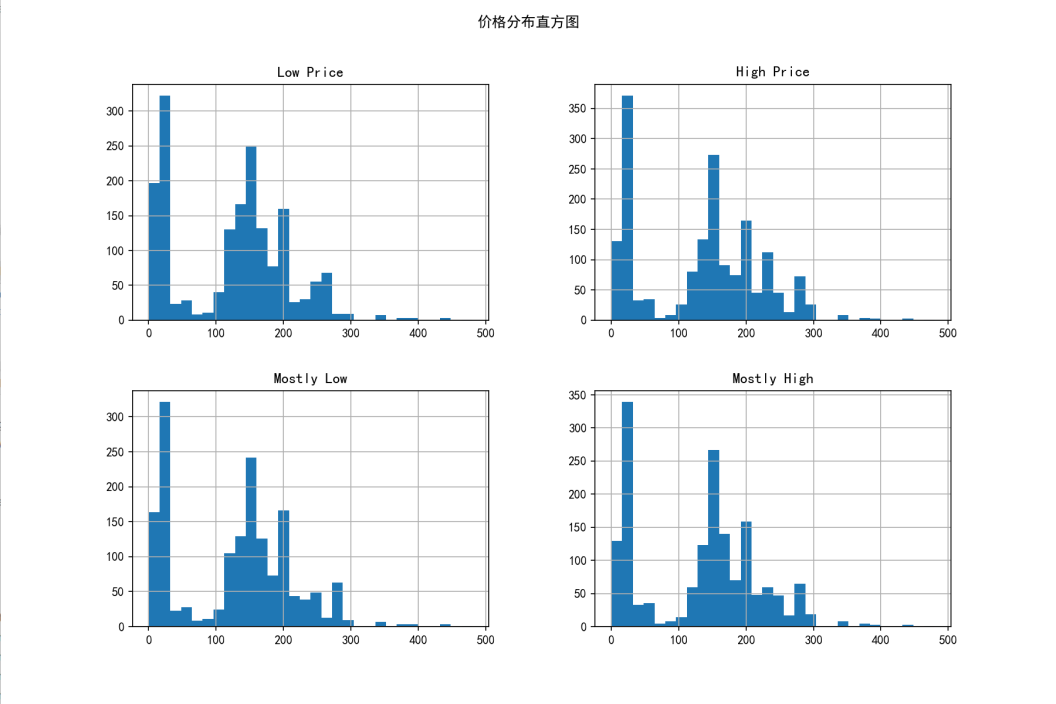
主要处理了以下缺失值

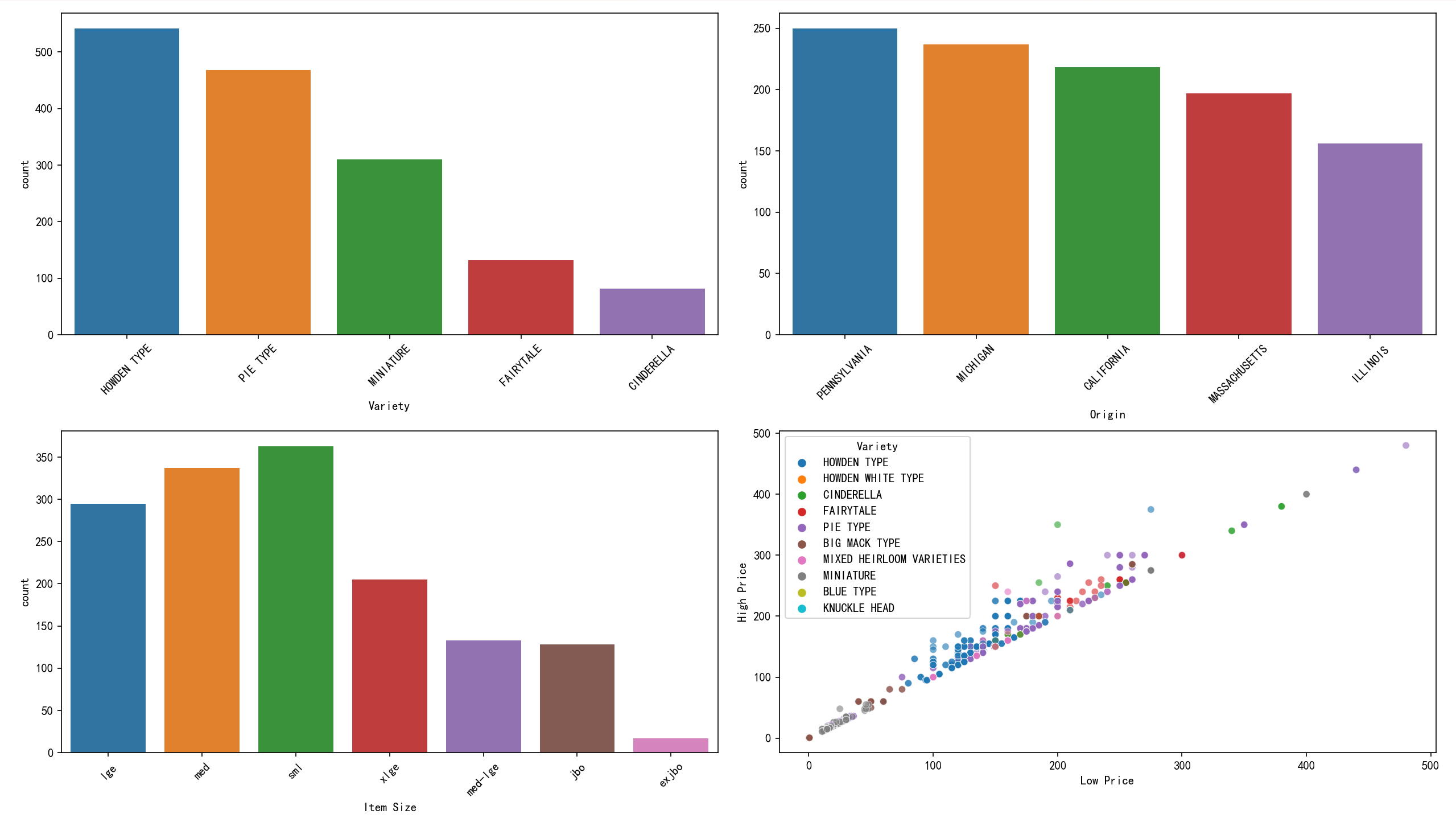
|  |  |  |  |
| --- | --- | --- | --- |
| **特征** | **缺失数** | **缺失率** | **处理方式** |
| **Color** | 616 | 35 % | 填充 `"UNKNOWN"` |
| **Item Size** | 279 | 16 % | 填充 `"MEDIUM"` |
| **Unit of Sale** | 1595 | 91 % | 填充 `"BIN" |
| **Variety / Origin** | < 5 条 | < 1 % | 保留原值 |
| **Mostly High / Low** | 103 | 6 % | 直接丢弃对应行 |



****Low Price 和 High Price****：这两个价格的分布较为相似，都呈现出右偏分布，即大部分价格集中在较低的区间，随着价格的增加，频数逐渐减少。大多数 Low Price 和 High Price 都集中在 0 到 200 之间，且在 0 到 100 区间内的频数最高，分别达到了 300 左右。这表明在市场上，南瓜的价格大多处于中低水平。

****Mostly Low 和 Mostly High****：它们的分布与 Low Price 和 High Price 类似，同样呈现出右偏分布，且大部分值集中在较低的区间。这进一步印证了南瓜价格普遍不高，且价格波动在一定范围内。

****整体分析****：四种价格指标的分布都呈现出明显的右偏态，说明南瓜价格市场整体价格水平较低，高价南瓜相对较少。

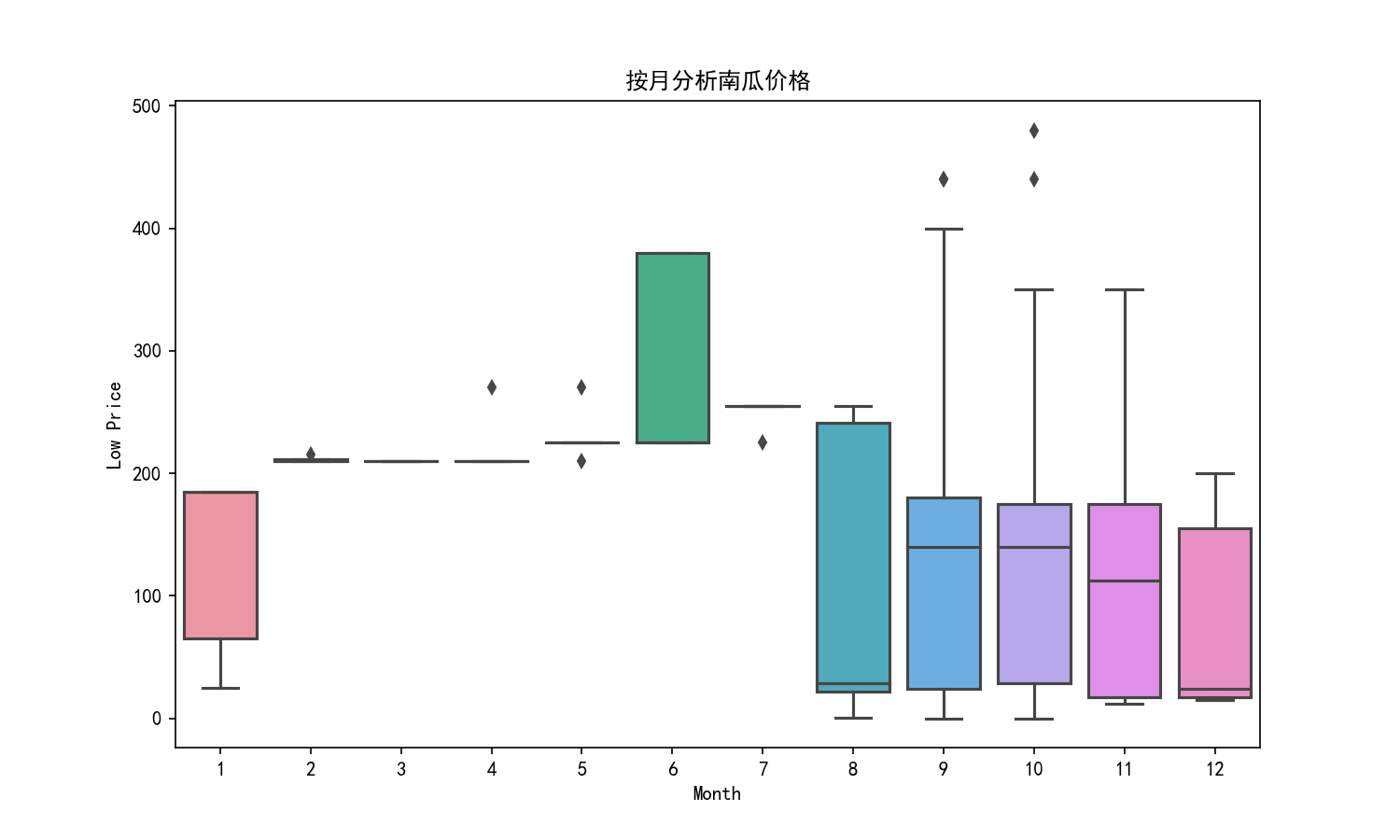


****Item Type 计数分布****：不同 Item Type 的数量差异较大。其中，“pie type” 和 “pumpkin type” 的数量最多，分别超过了 300 和 250；而“squash type” 和 “gourd type” 相对较少，尤其是“gourd type”，数量最少。这表明在数据集中，某些类型的南瓜或相关产品更为常见，可能与市场需求、种植面积等因素有关。

****Variety 计数分布****：从图中可以看出，“Howden” 是最常见的品种，数量接近 250；其次是“Connecticut Field”等品种，而像“Early Sweet” 等品种的数量相对较少。这反映了不同南瓜品种在市场上的流行程度或种植规模存在较大差异。

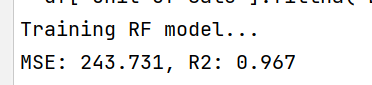
****Item Size 计数分布****：Item Size 的分布相对集中，“1 1/9 bushel” 和 “1/2 bushel” 是最主要两种包装规格，数量都在 150 以上；而“1/4 bushel” 等较小规格以及“bag” 等其他包装方式的数量较少。这说明在南瓜的销售中，中等大小的包装更为普遍，可能与消费者的购买习惯或商家的包装策略有关。

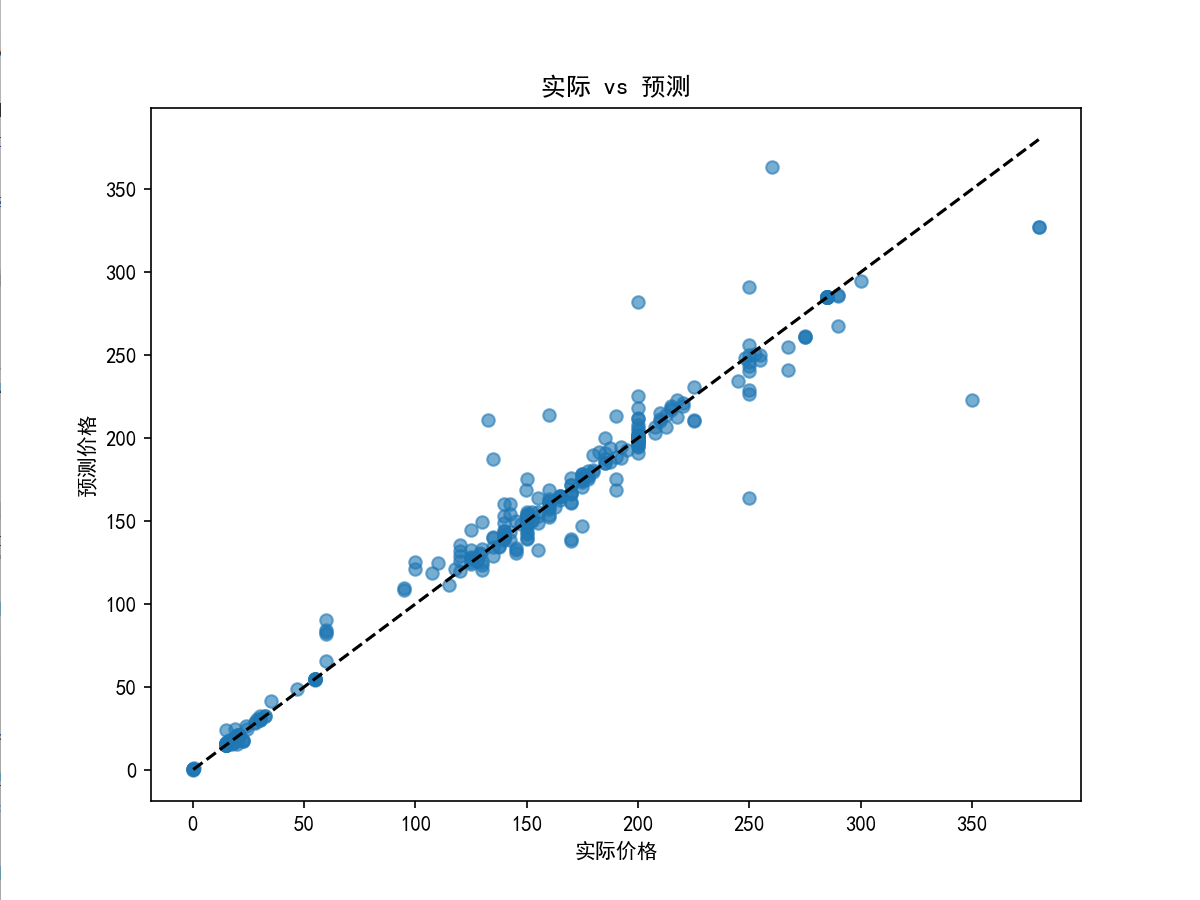
****Low Price 和 Score 的关系****：散点图显示了 Low Price 和 Score 之间的关系，但整体上相关性不明显。价格在 0 到 200 之间的样本分布较为分散，不同 Score 对应的价格差异不大。这可能意味着南瓜的价格并不完全依赖于 Score 这一指标，或者 Score 的影响在其他因素的干扰下被削弱了。

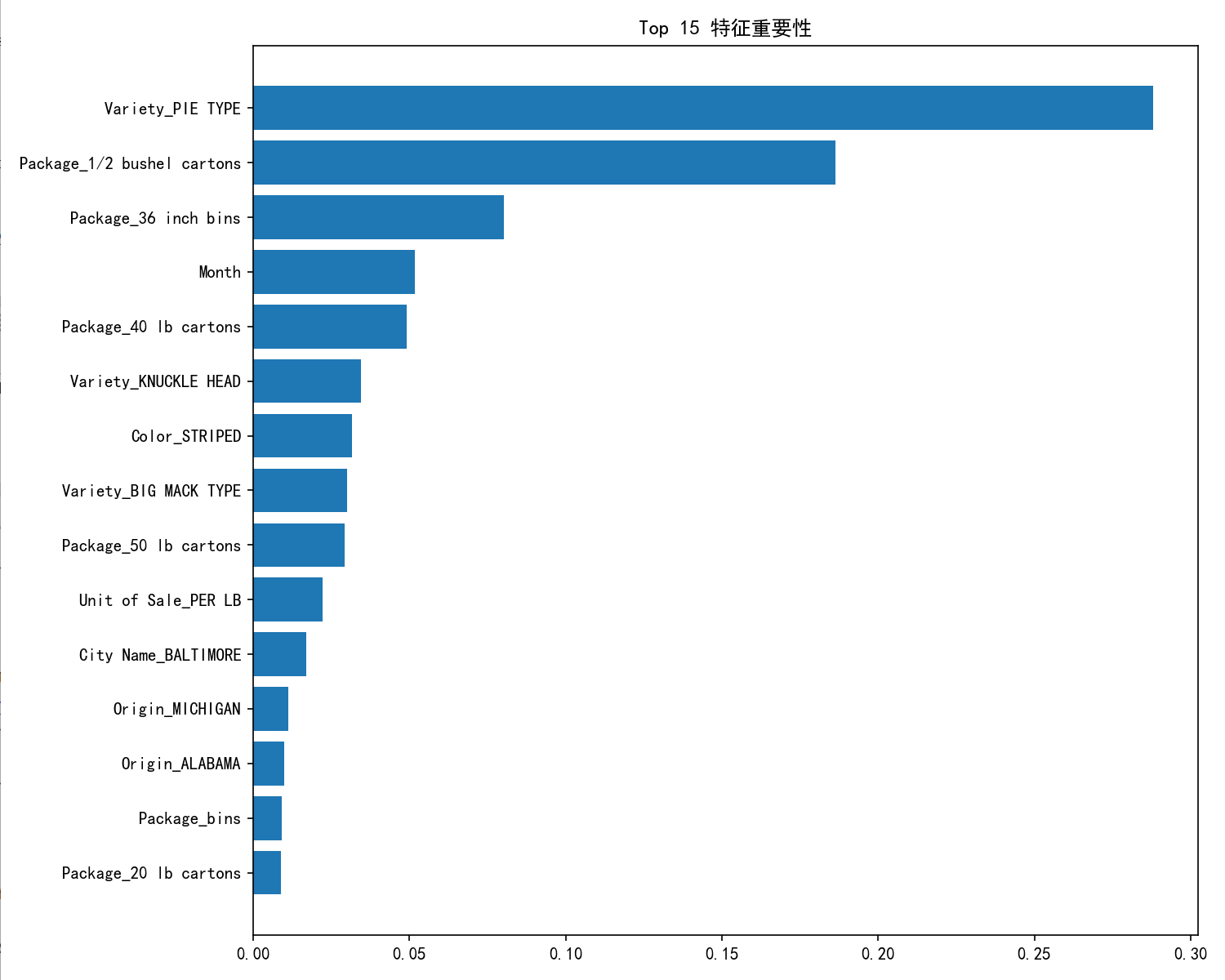
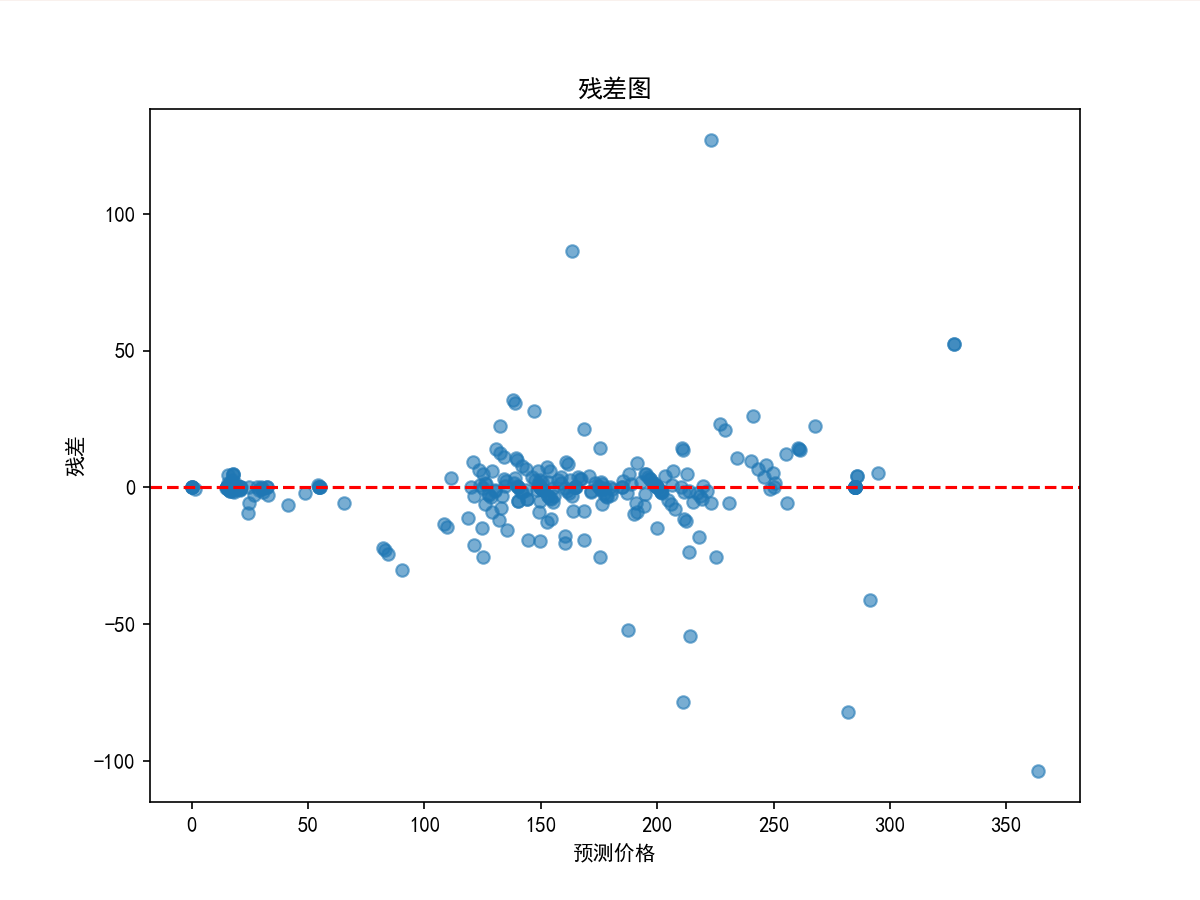


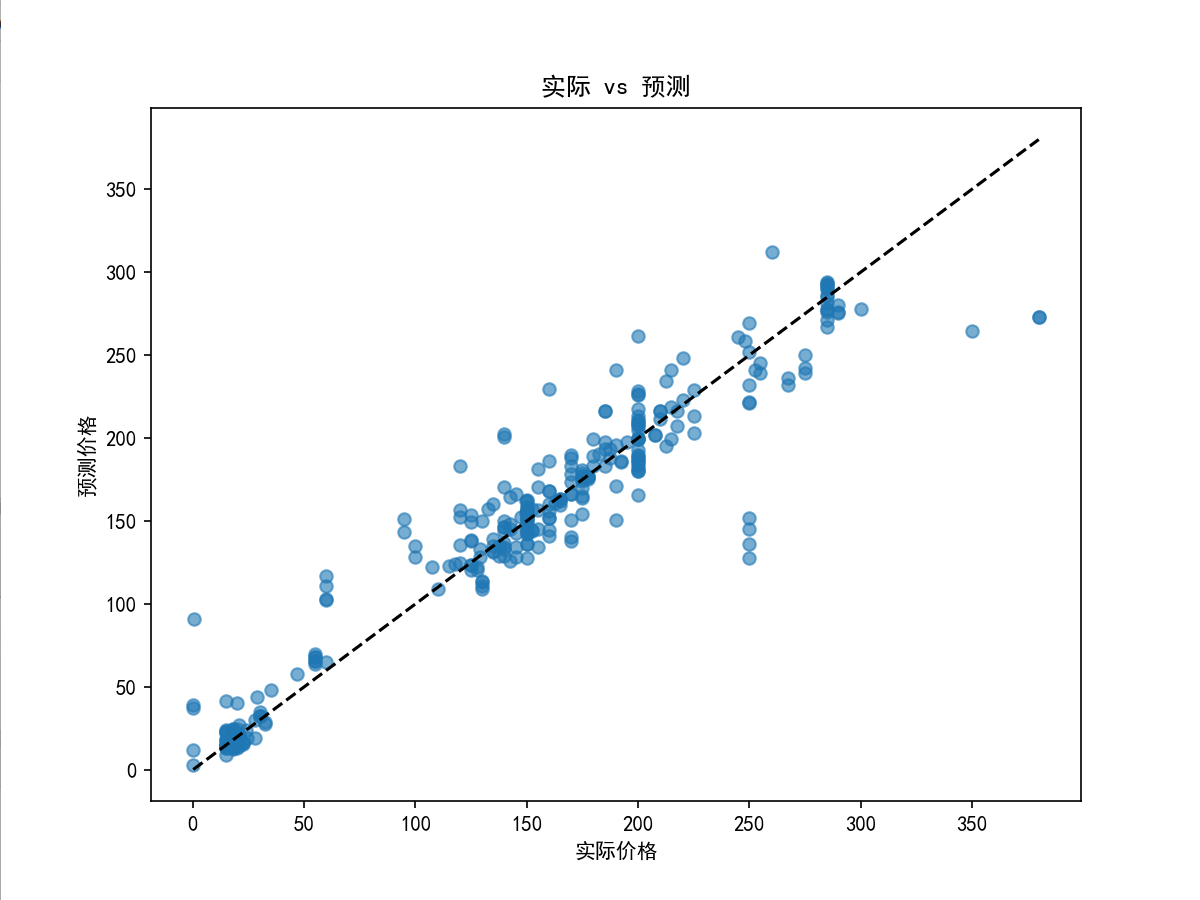
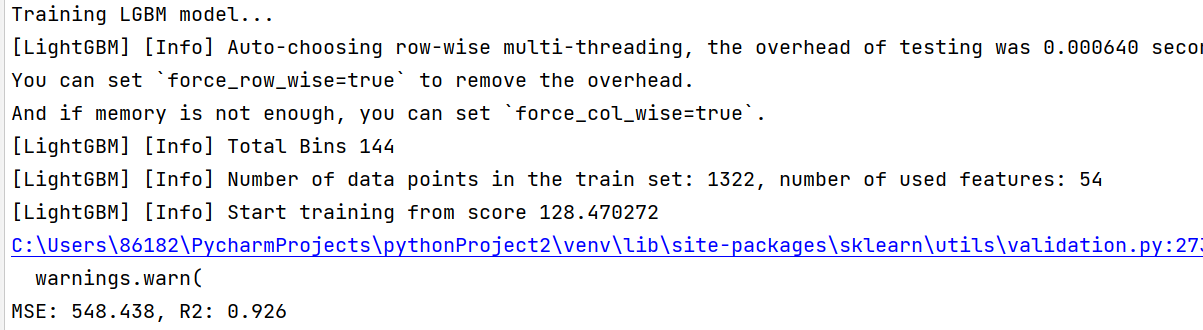
南瓜价格表现出明显的季节性波动，夏季和秋季的价格较高，尤其在 6 月达到全年最高点。这可能与南瓜的生长周期、收获季节以及特定节日的需求增加有关。此外，6 月和 9 月的价格波动较大，可能存在供应不稳定或市场需求变化较大的情况。

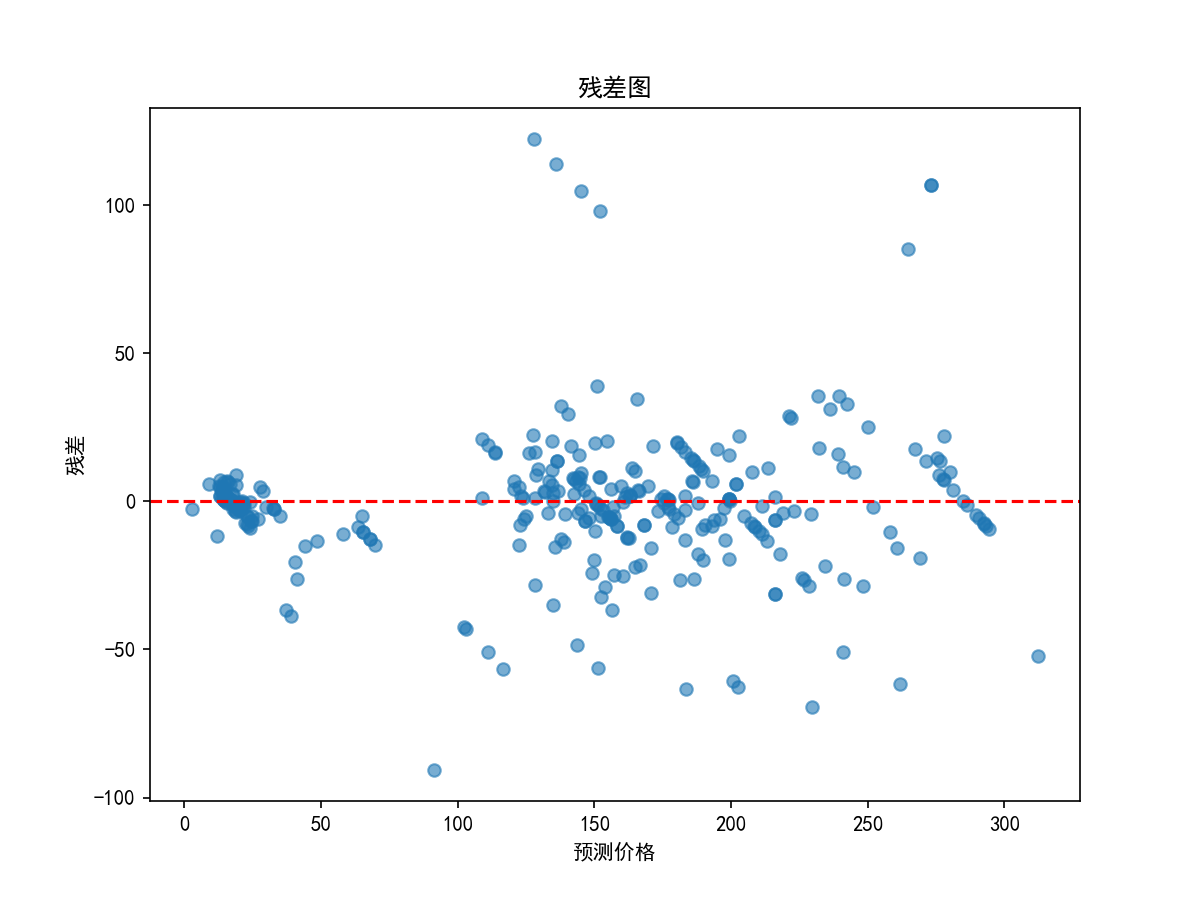
|  |  |  |
| --- | --- | --- |
| **模型** | **MSE** | **R²** |
| **RandomForest** | 243.731 | 0.967 |
| **LightGBM** | 548.438 | 0.926 |
| **XGBoost** | 240.533 | 0.968 |

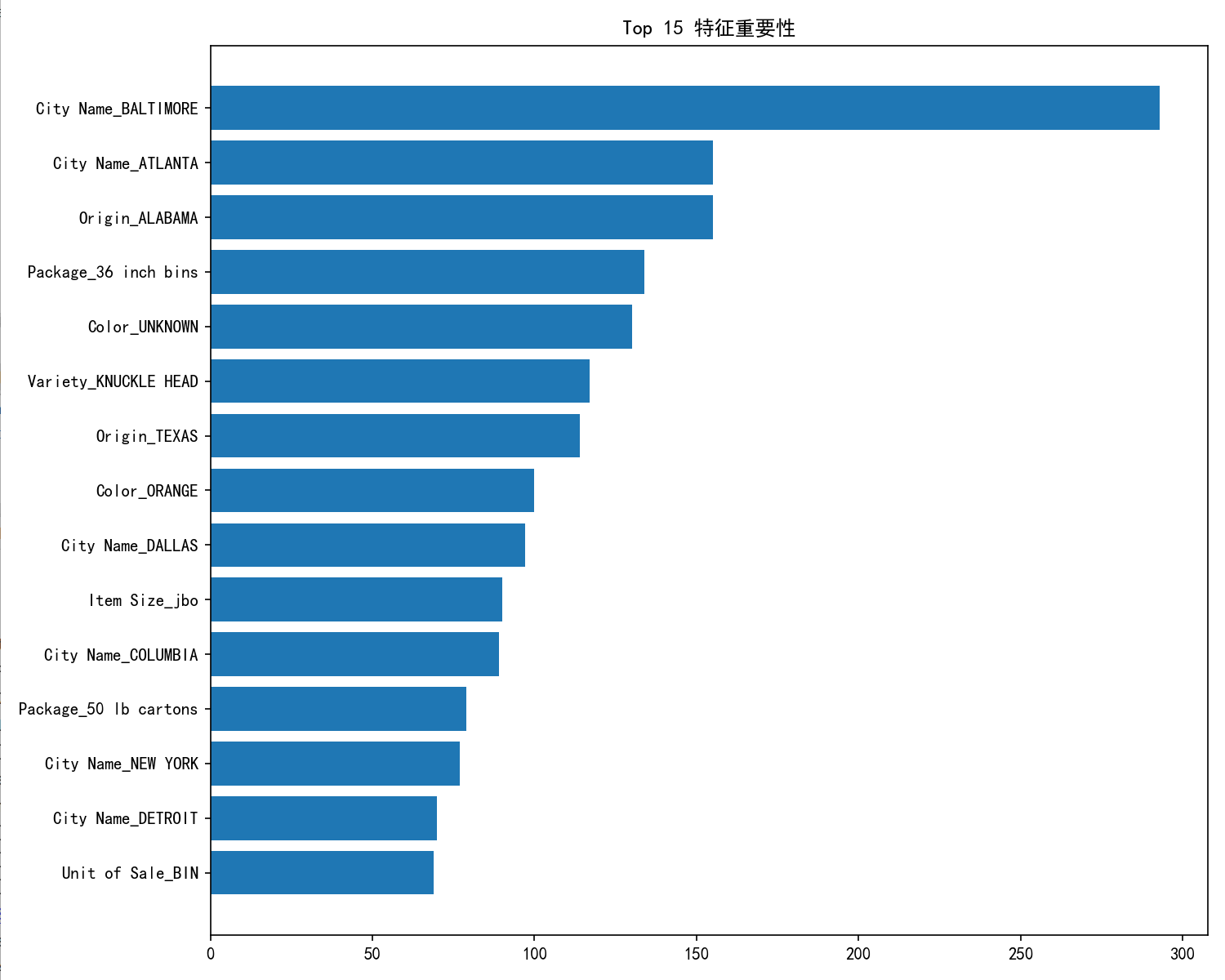


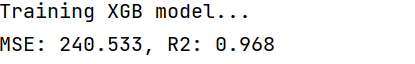


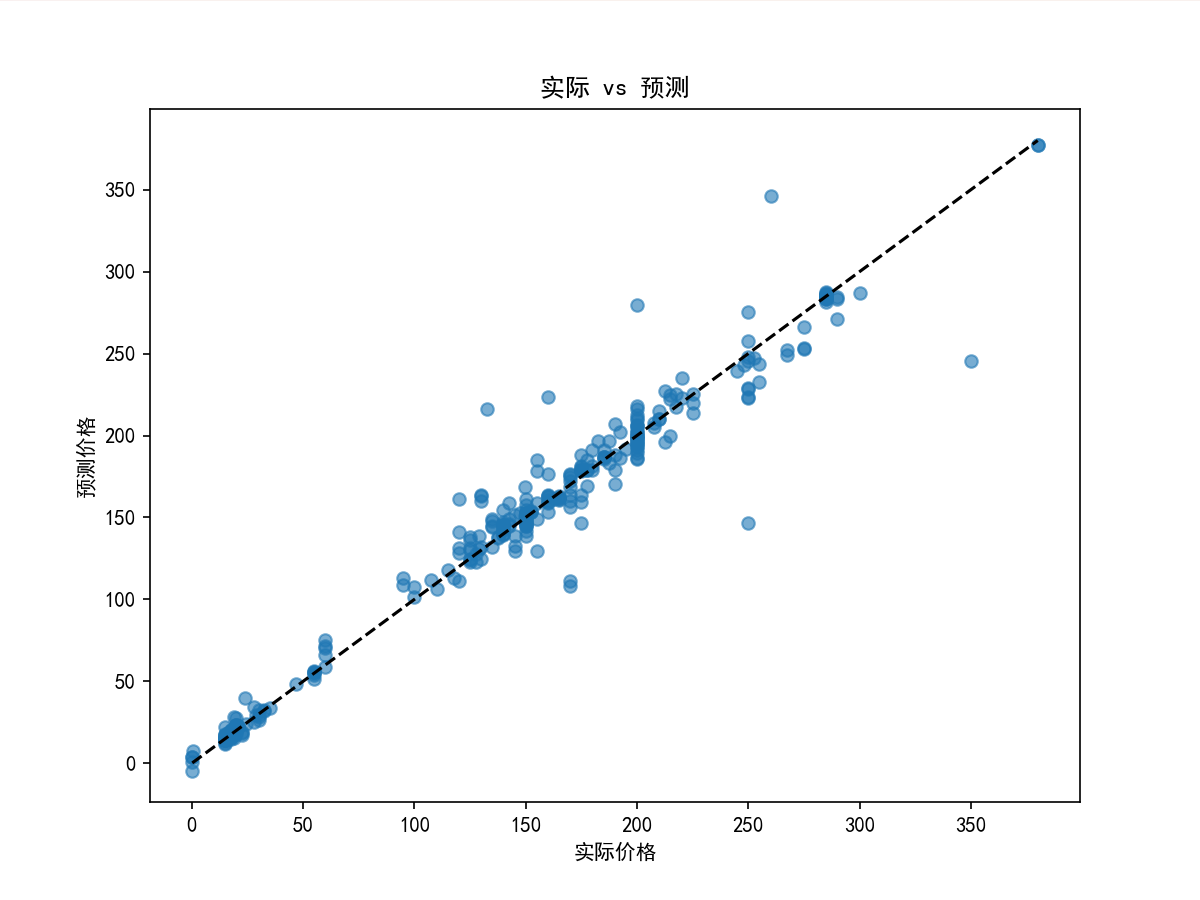


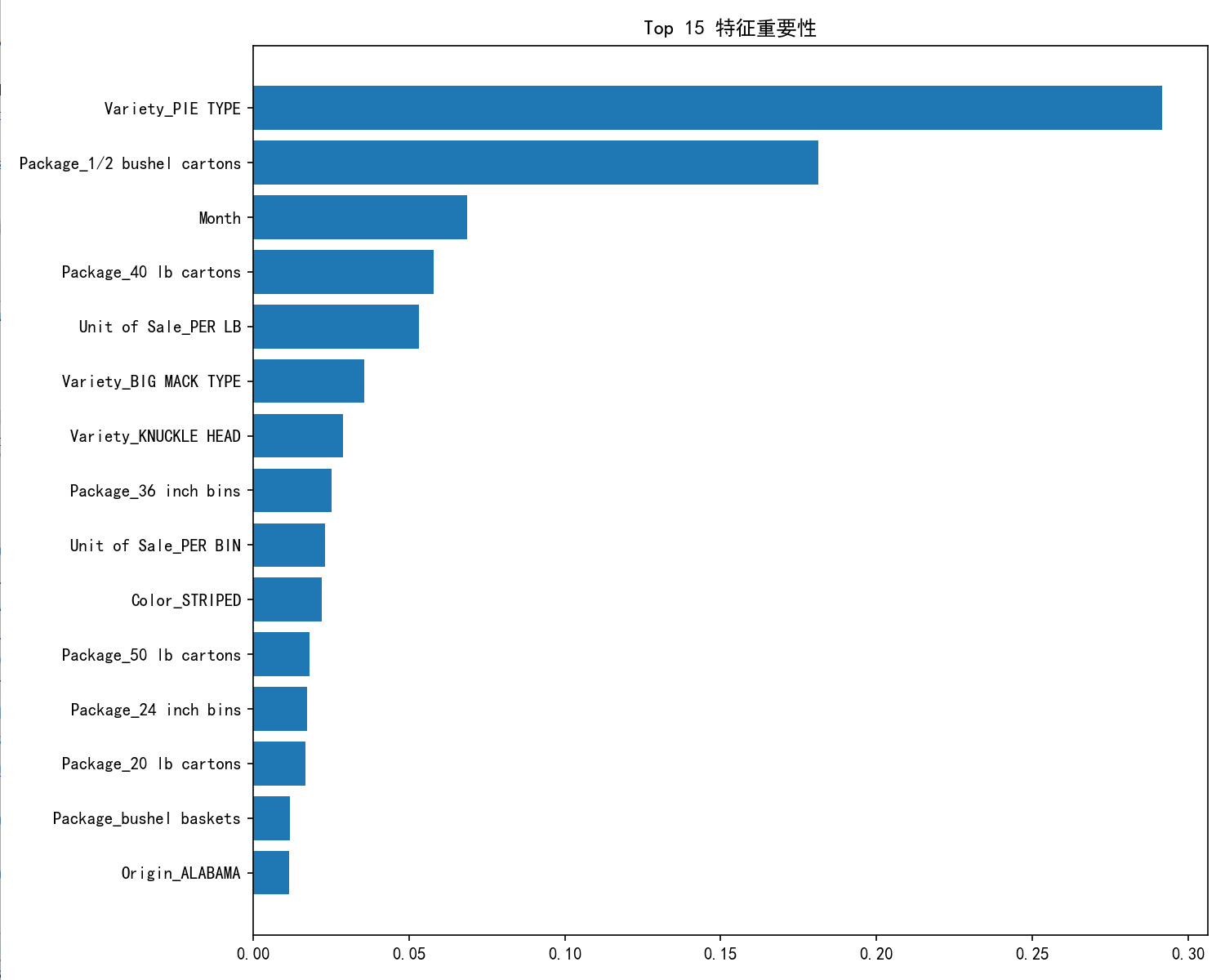
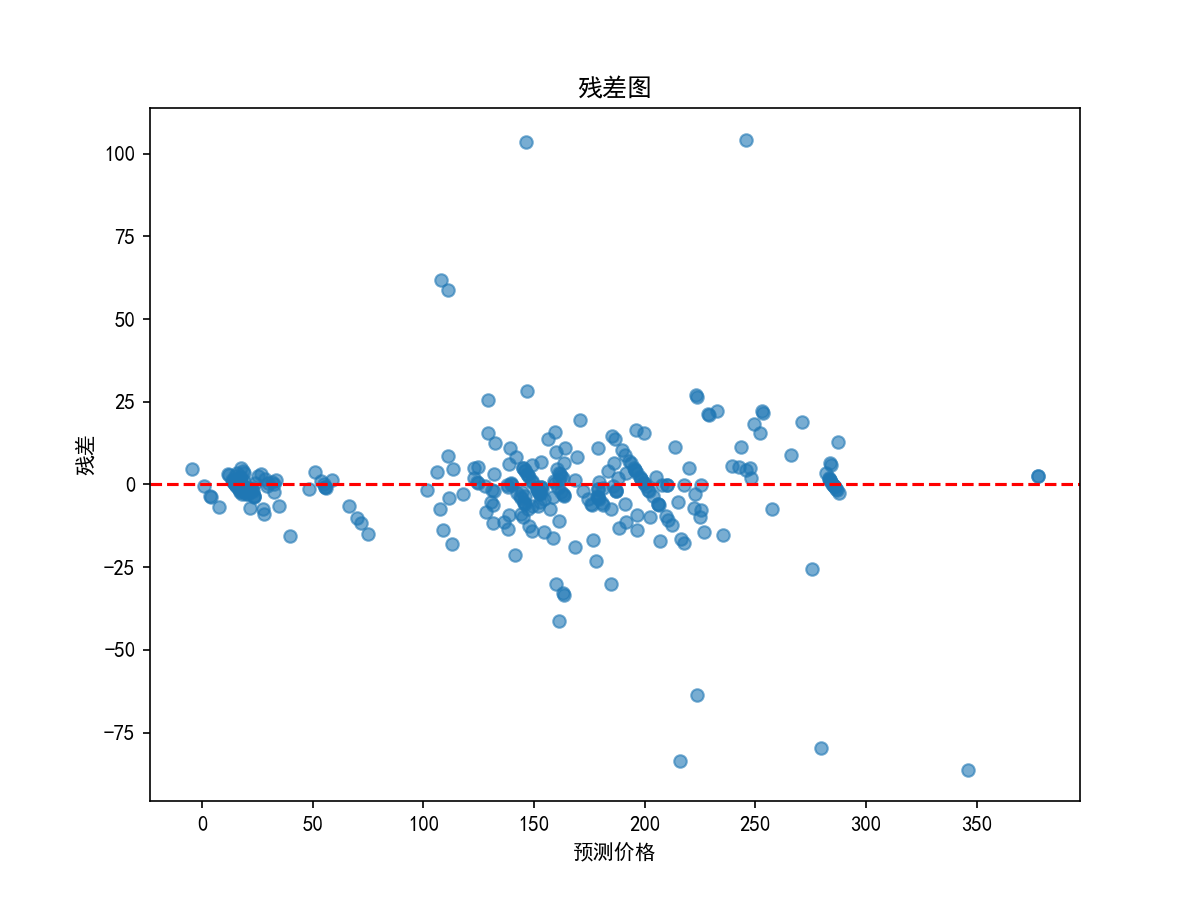












****特征重要性分析****：

****随机森林和 XGBoost****：City Name、Variety、Month 是主要特征

****LightGBM****：City Name、Month、Package 是主要特征。

****共同点****：所有模型都显示 City Name 和 Month 是影响价格的关键特征。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 模型 | 训练数据规模 | 测试数据规模 | 平均训练性能 | 平均测试性能 |
| RF | (1102, 10) | (551, 10) | RMSE:8.73, MAE: 3.15, R²: 0.99 | RMSE:21.04,MAE: 7.96, R²: 0.94 |
| LGBM | (1102, 10) | (551, 10) | RMSE:19.24, MAE: 10.26, R²: 0.95 | RMSE:26.42,MAE: 13.79, R²: 0.91 |
| XGB | (1102, 10) | (551, 10) | RMSE:6.71, MAE: 3.00, R²: 0.99 | RMSE:21.11,MAE: 8.25, R²: 0.94 |