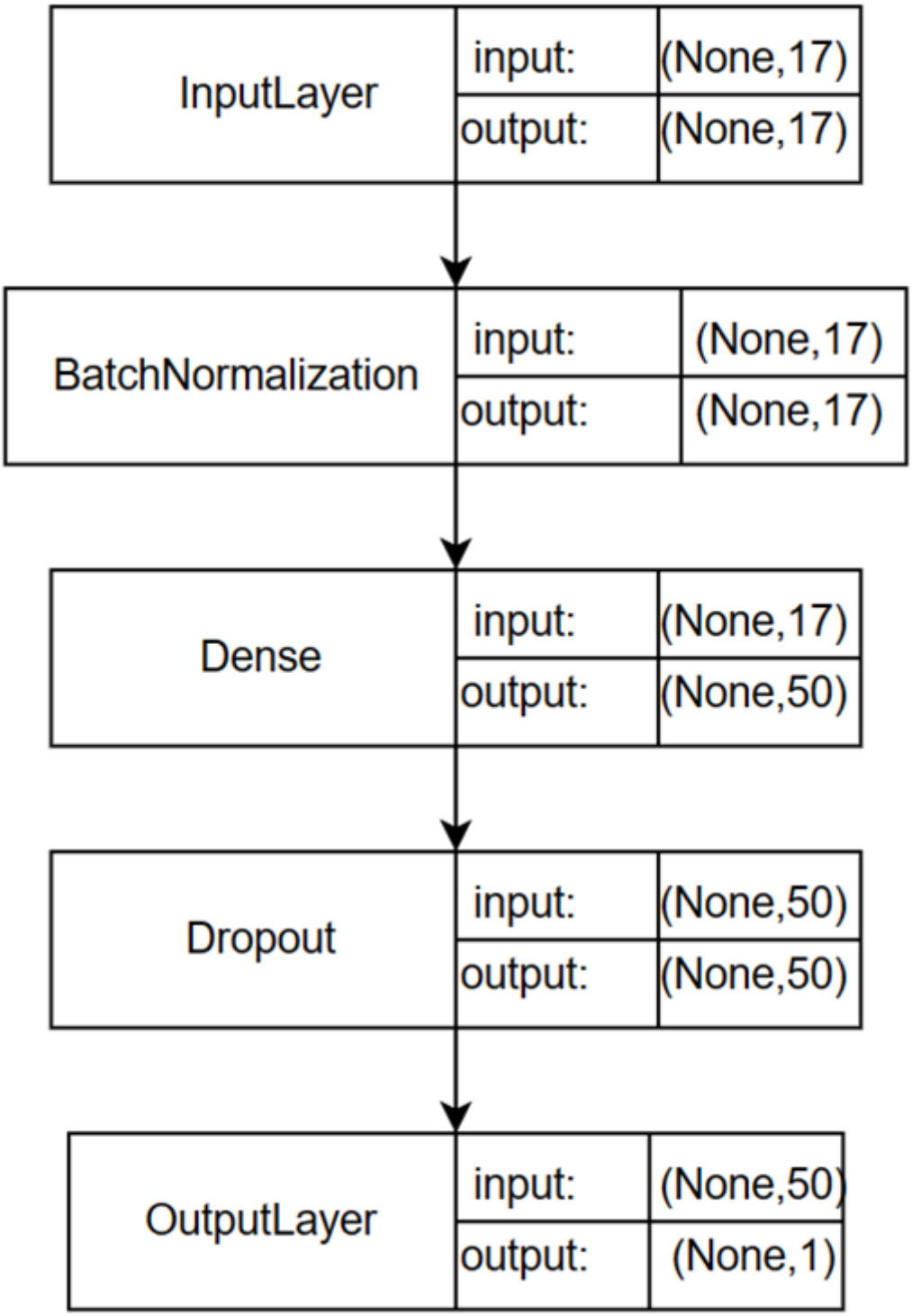
电影票房预测深度学习建模以及模型解释报告

1.深度学习模型

由于本任务是一个经典的表格数据的回归预测任务，因此我们采用了基础的全连接神经网络对数据集进行预测。

该神经网络模型包括一个带有批标准化的输入层；两层带有relu激活函数的k个神经元的隐藏层，并对隐藏层添加dropout、L1、L2正则的功能；由于本任务为回归预测任务，输出层为不添加激活函数的线性输出层。



在数据分析过程中可以看到，预算“budget”属性中，高预算数据与低预算数据在数据分布上存在较为明显的差异，因此我们使用模型分别对整体数据（processed_bop_train.csv）、高预算数据（processed_high_train.csv）、低预算数据（processed_low_train.csv）进行训练建模并测试。

1.1以整体数据为训练集

注：该部分代码及结果保存在文件"deep_learning_model.ipynb"中。

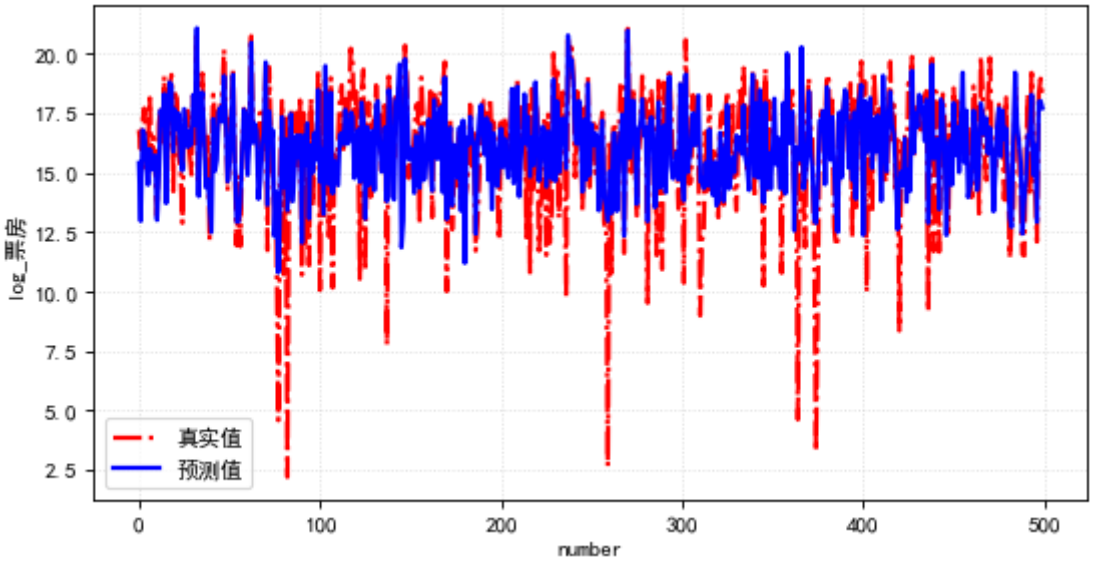
首先以整体数据processed_bop_train.csv为训练集训练模型，将测试集、高预算部分测试集、低预算部分数据集分别进行测试，结果如下：

1.1.1整体测试集

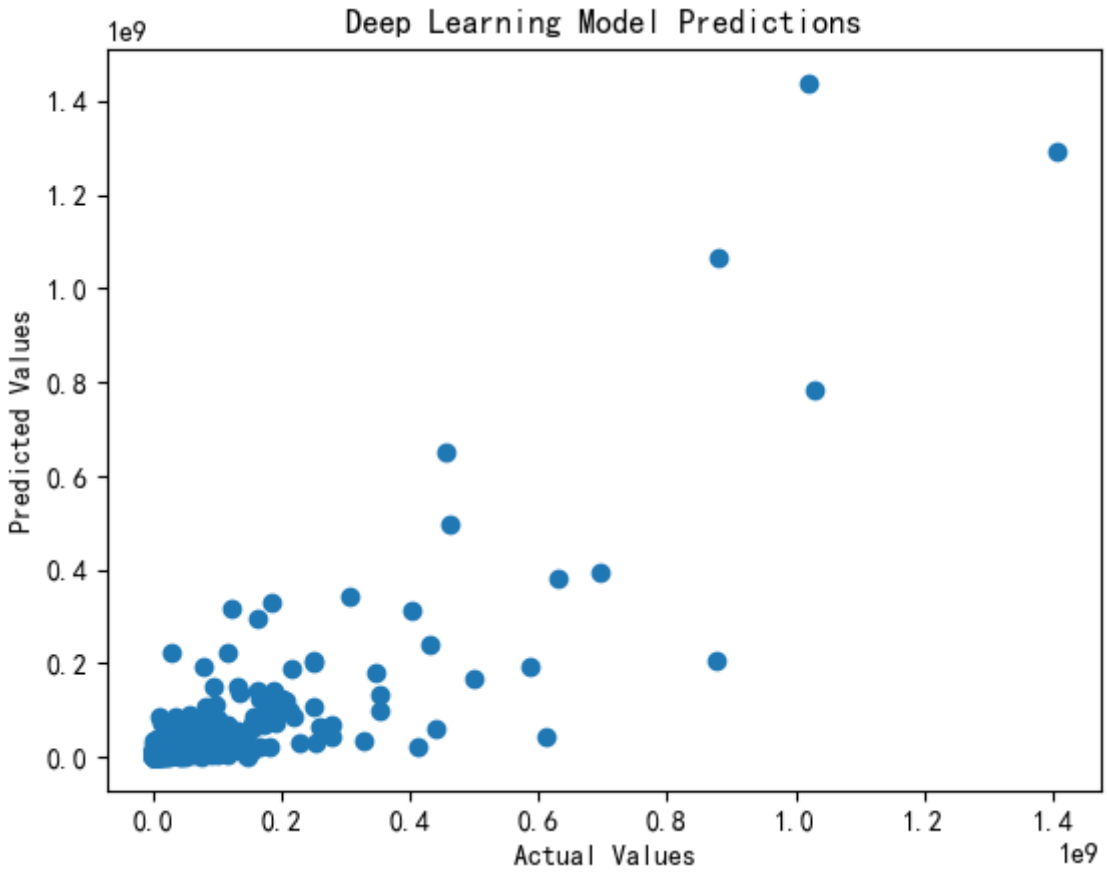
整体测试集的rmsle值为：1.9494792295173162

实际值与测试值的mae为：1.362346242531215

实际log_票房与预测log_票房的对比折线图：



实际票房与预测票房的对比散点图：



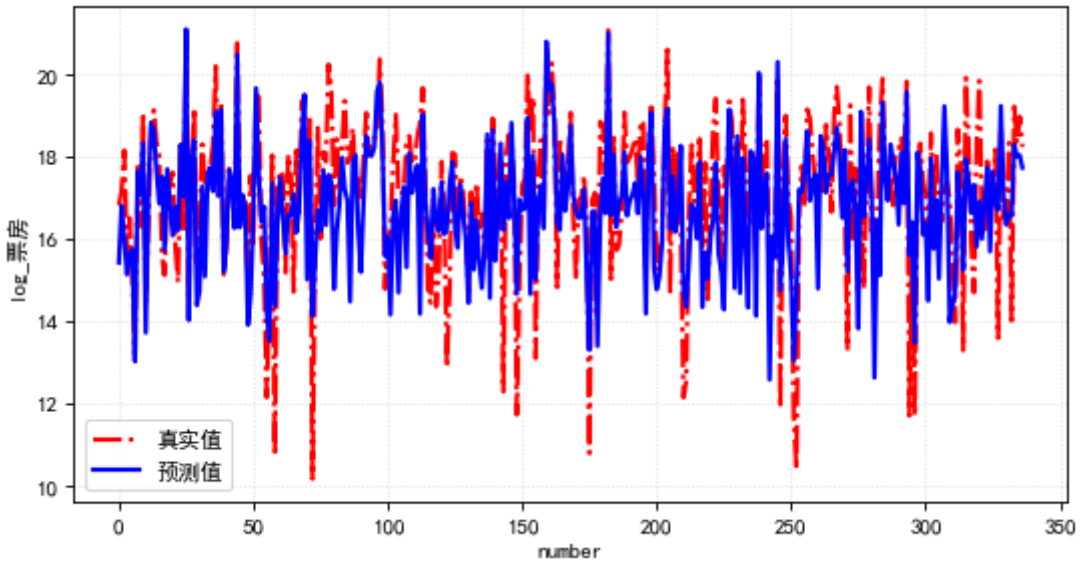
由上图可以看到，实际值与预测值总体趋势基本一致，但在部分极端数据上仍出现了较大的偏差。

1.1.2高预算测试集

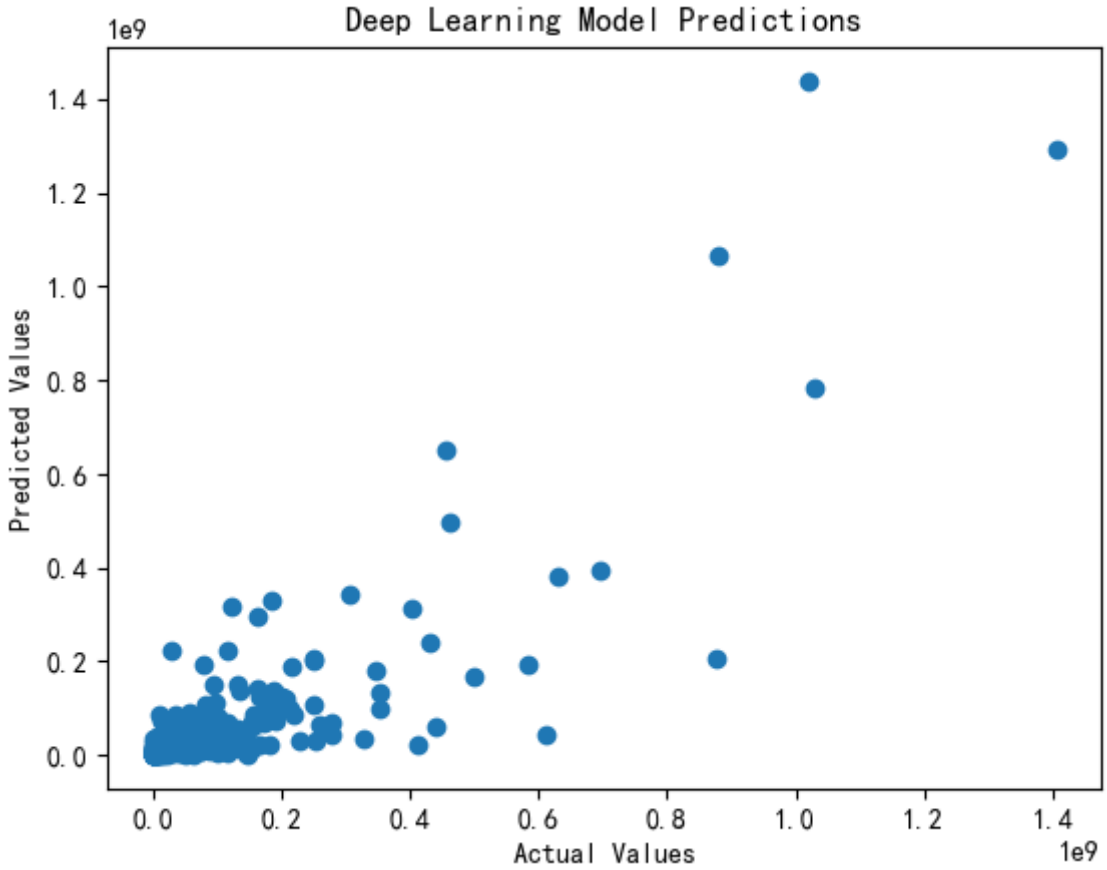
高预算测试集的rmsle值为：1.4042929922719325

实际值与测试值的mae为：1.0831982412993402

实际log_票房与预测log_票房的对比折线图：



实际票房与预测票房的对比散点图：



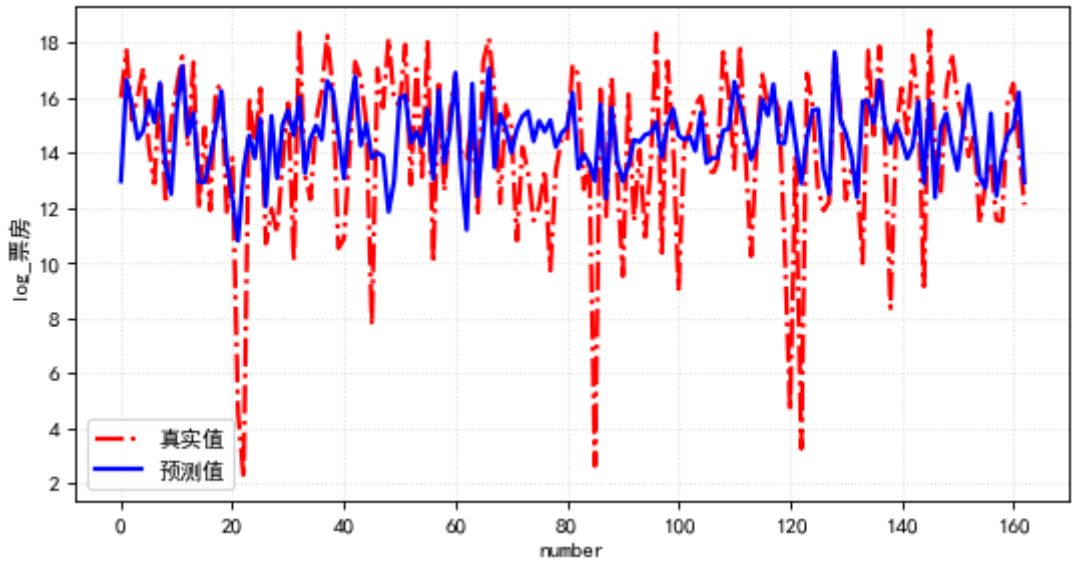
可以明显观察到，该模型在高预算测试集中的表现要由于在整体测试集上的表现，这一结果也证实了我们之前的结论：高预算数据具有较好的数据分布。

1.1.3低预算测试集

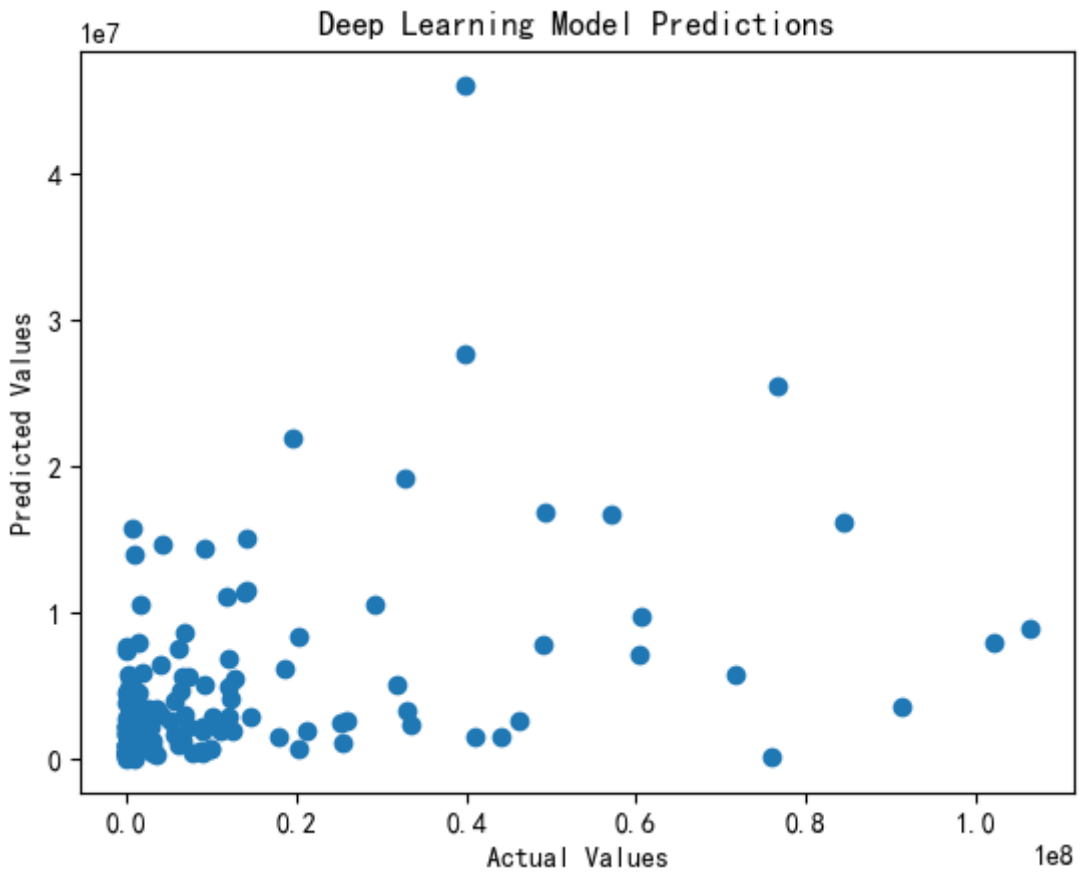
低预算测试集的rmsle值为：2.7533109480137123

实际值与测试值的mae为：1.9394804536670536

实际log_票房与预测log_票房的对比折线图：



实际票房与预测票房的对比散点图：



该模型在低预算测试集上出现了明显的性能下滑，推测原因可能是数据在记录过程中出现了差错，或票房表现出了“超出预期”的表现。

1.2以高预算数据为训练集

注：该部分代码及结果保存在文件“deep_learning_model_high.ipynb“中。

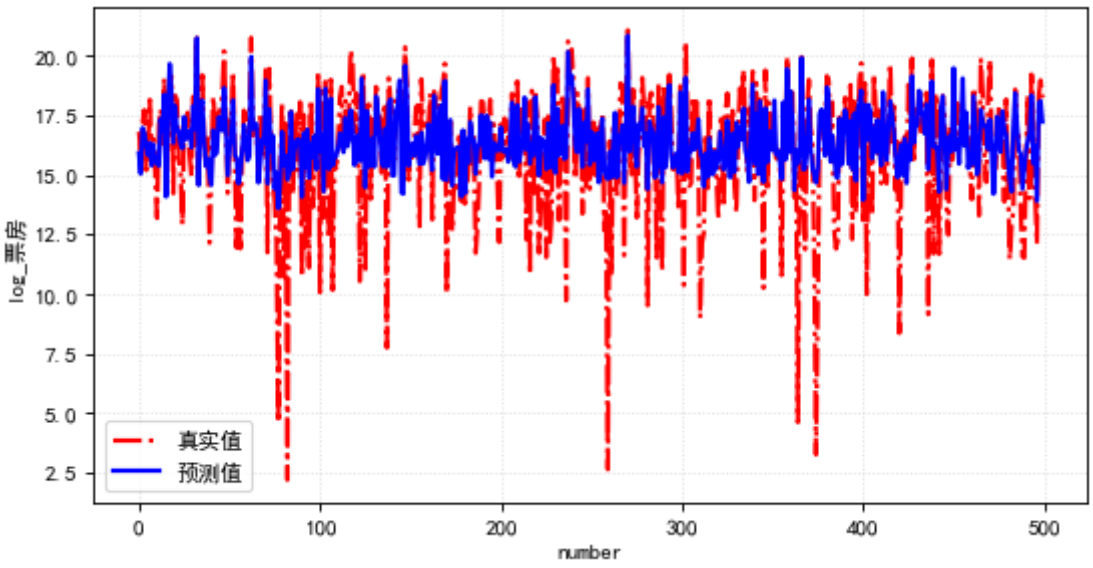
由上述模型可以看出，高预算数据具有更优的数据分布，因此我尝试了以高预算数据为训练集的模型，结果如下：

1.2.1整体测试集

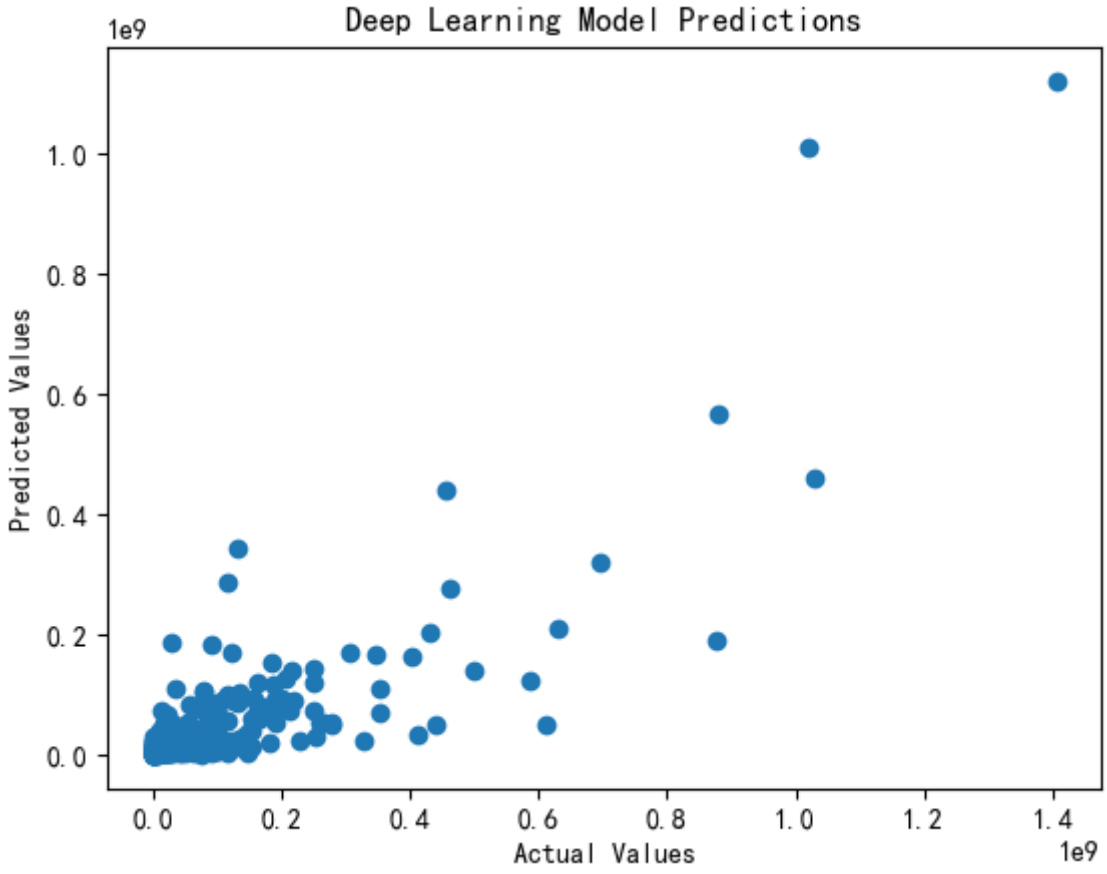
整体测试集的rmsle值为：2.1408517589202973

实际值与测试值的mae为：1.440722899297683

实际log_票房与预测log_票房的对比折线图：



实际票房与预测票房的对比散点图：



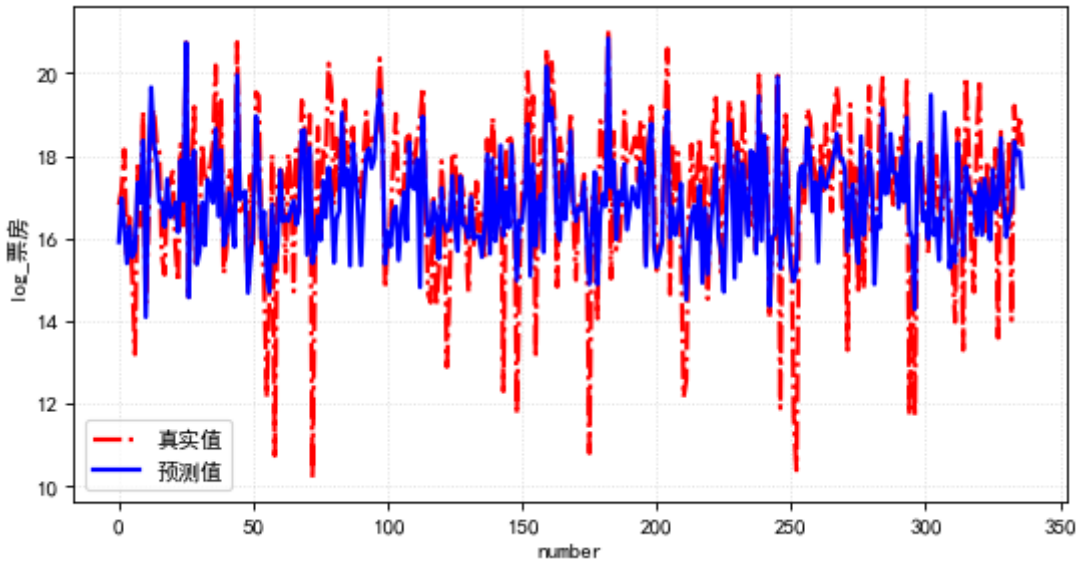
该测试效果比1.1中所训练的模型效果稍差，从折线图中可以看出，模型倾向于不给出低票房的预测，推测原因可能是由于训练集中缺少低预算（因此可能导致的低票房）数据，导致模型拟合不够好。

1.2.2高预算测试集

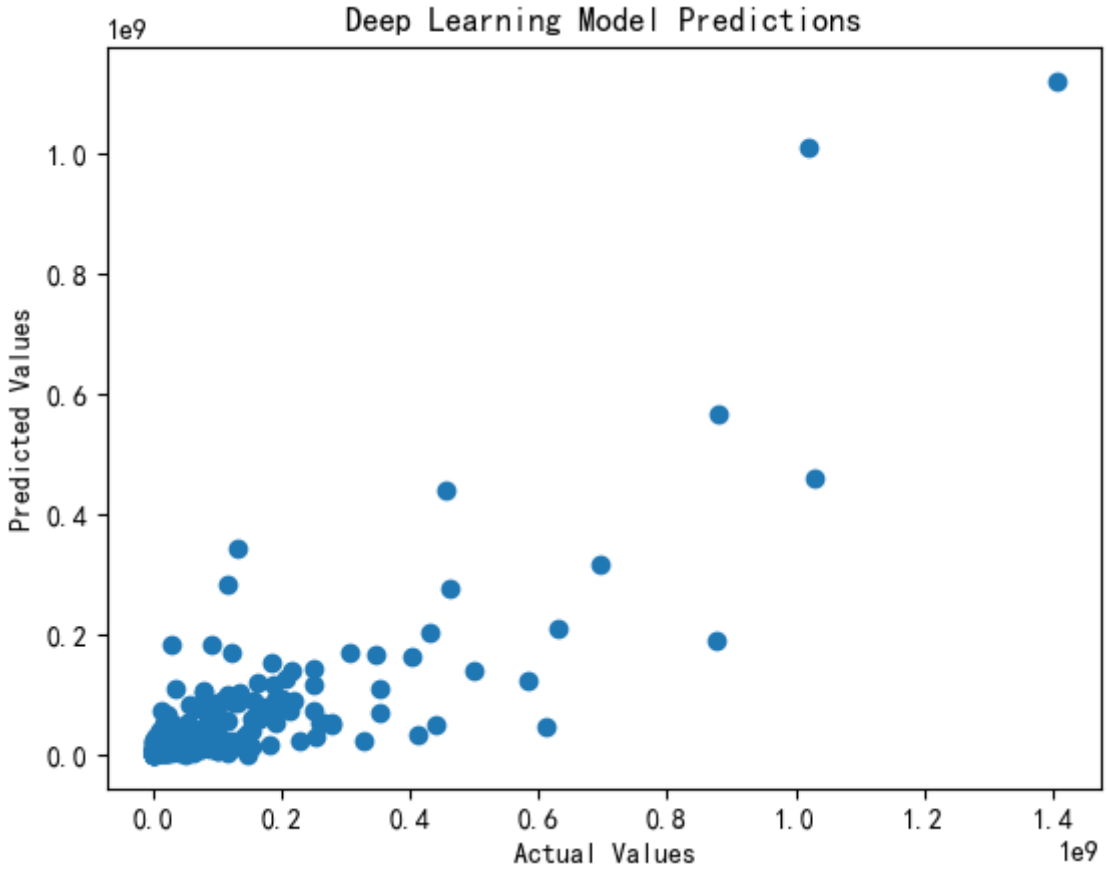
高预算测试集的rmsle值为：1.3907028902776242

实际值与测试值的mae为：1.0578636871373521

实际log_票房与预测log_票房的对比折线图：



实际票房与预测票房的对比散点图：



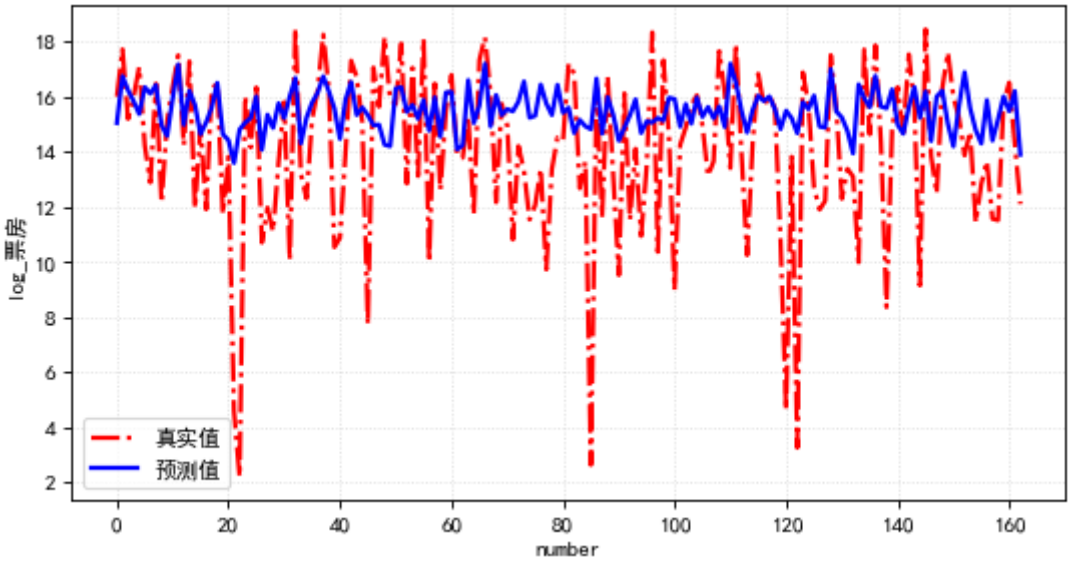
通过对比可以发现一个有趣的现象：在高预算测试中，整体训练模型的表现优于高预算训练模型。因此，虽然高预算的数据分布较好，但是仅使用高预算数据集进行训练是不可取的。

1.2.3低预算测试集

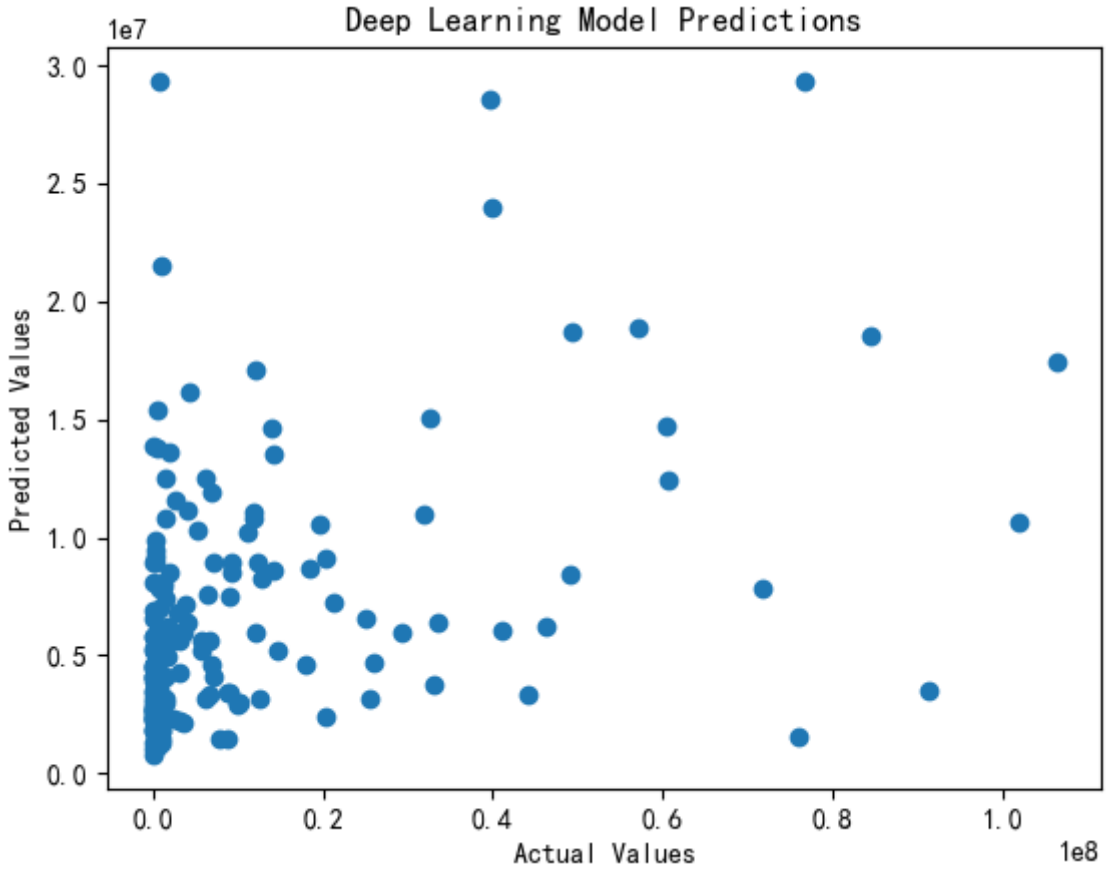
低预算测试集的rmsle值为：3.1718148372637955

实际值与测试值的mae为：2.232278407410022

实际log_票房与预测log_票房的对比折线图：



实际票房与预测票房的对比散点图：



从上图可以看出，模型出现了明显的高估现象，也可以看出高预算与低预算数据的数据分布确实存在较大差异。

1.3以低预算数据为训练集

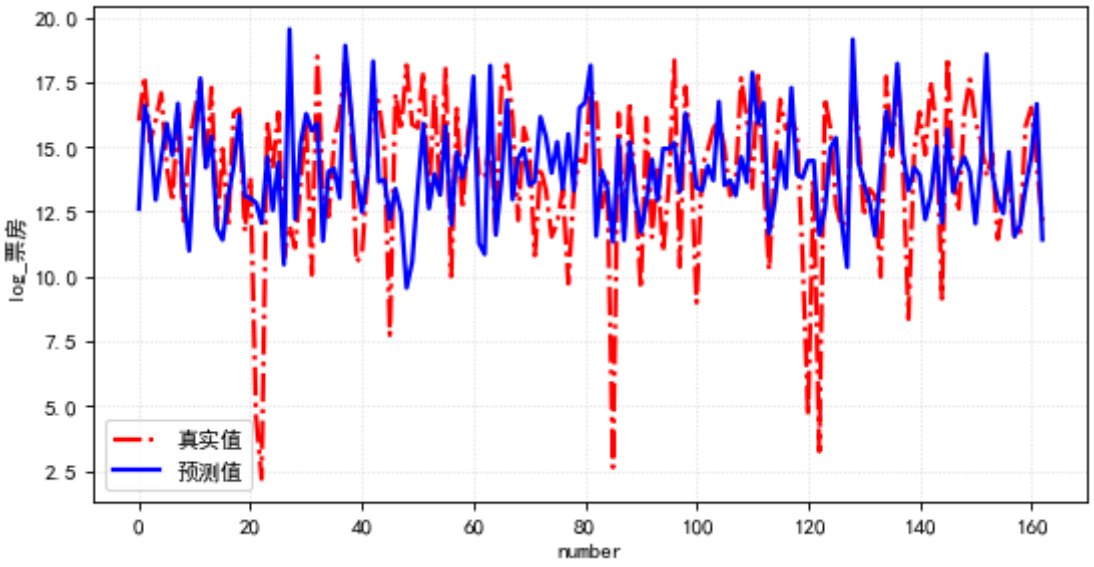
注：该部分代码及结果保存在文件"deep_learning_model_low.ipynb"中。

该部分模型训练不用于模型对比，仅用于数据探索，因此只使用了低预算部分的测试集进行测试：

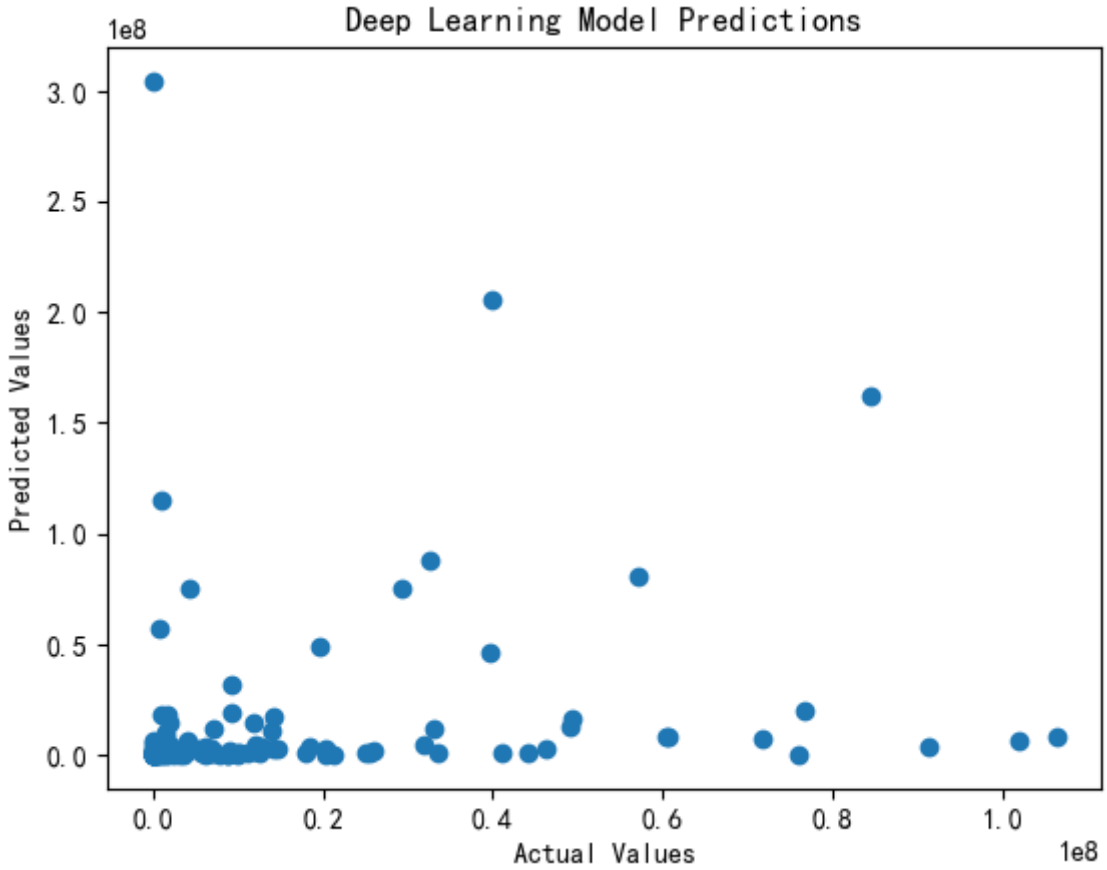
低预算测试集的rmsle值为：2.880698052951494

实际值与测试值的mae为：2.1501461067387373

实际log_票房与预测log_票房的对比折线图：



实际票房与预测票房的对比散点图：

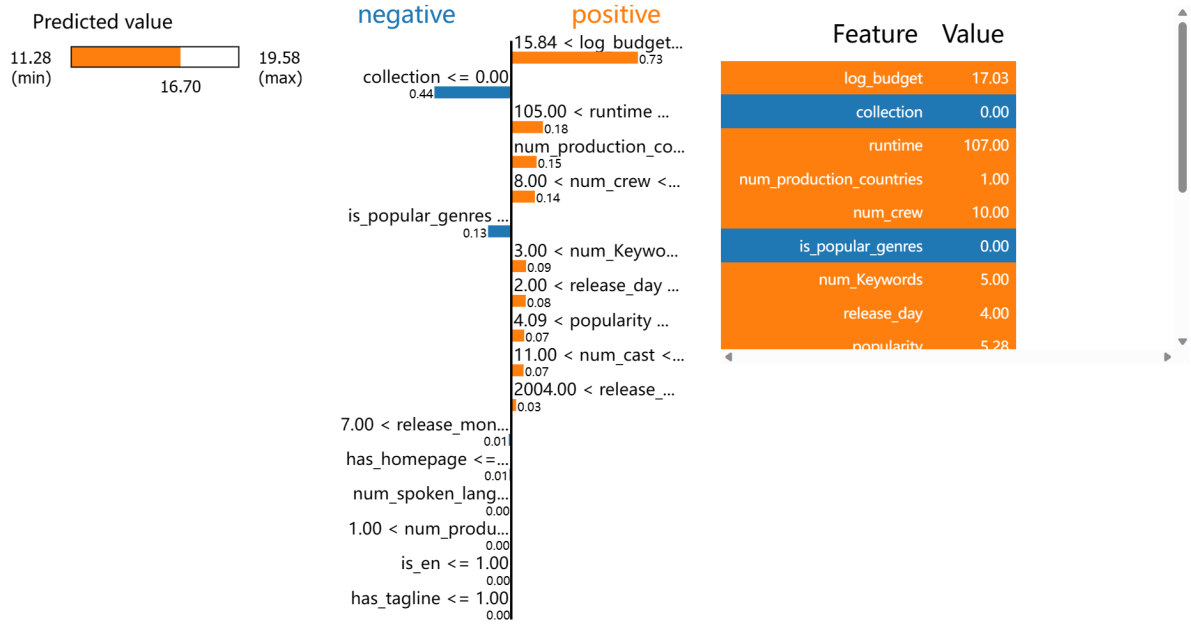


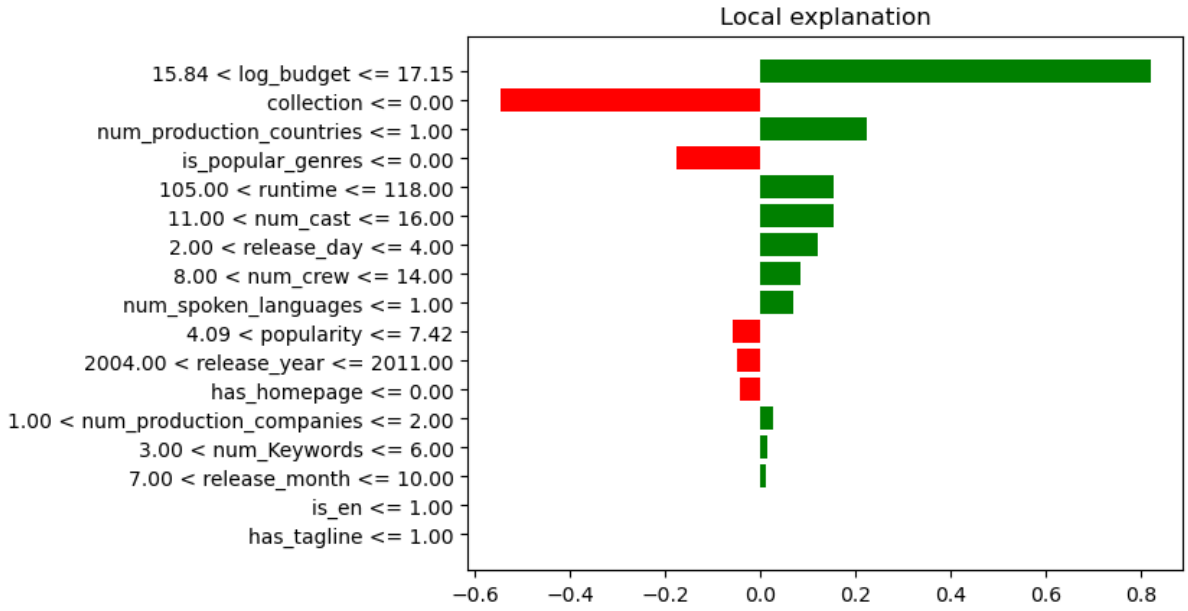
可以看到，在仅使用低预算数据作为训练集与测试集的情况下，模型拟合效果依然很差。低预算部分的数据表现出了一定的无规律性。

2.模型解释

注：该部分代码及结果保存在文件"model_interpretability/LIME.ipynb"中。

经过多种模型的训练与评价，我们最终选择随机森林模型作为预测模型，并使用LIME包对该模型中的单个样本进行解释，以整体测试集中的第一组数据为例：





以列表形式进行解释：

```
[('15.84 < log_budget <= 17.15', 0.8193400748840702),
 ('collection <= 0.00', -0.5461356646561762),
 ('num_production_countries <= 1.00', 0.22371015549664078),
```

```
('is_popular_genres <= 0.00', -0.1772645630170876),
('105.00 < runtime <= 118.00', 0.15320681607703196),
('11.00 < num_cast <= 16.00', 0.15304649897066216),
('2.00 < release_day <= 4.00', 0.12177498245593275),
('8.00 < num_crew <= 14.00', 0.08405003935288727),
('num_spoken_languages <= 1.00', 0.06859849971010011),
('4.09 < popularity <= 7.42', -0.058122844098940495),
('2004.00 < release_year <= 2011.00', -0.0479906739524461),
('has_homepage <= 0.00', -0.04418507664149942),
('1.00 < num_production_companies <= 2.00', 0.026365487151517347),
('3.00 < num_Keywords <= 6.00', 0.015858373038446213),
('7.00 < release_month <= 10.00', 0.01072674271406627),
('is_en <= 1.00', 0.0),
('has_tagline <= 1.00', 0.0)]
```

可以看出，在该样本的预测过程中，起正面影响的属性分别有："log_budget"、"num_production_countries"、"runtime""num_cast"、"release_day"、"num_crew"、"num_spoken_languages"、"num_production_companies"、"num_Keywords""release_month"，影响依次降低。起负面影响的属性分别有"collection"、"is_popular_genres"、"popularity"、"release_year"、"has_homepage"，影响依次降低。