# **金融科技创新大赛报告**

基于风险控制的保险产品推荐

# 问题引入

保险产品本身的作用是帮助人们以当前可承受的价格去规避未来难以承受的风险。因此购买保险对于客户来说是利用确定性的金钱将未来的不确定损失转移给保险公司。顾客在投保时，有一些复杂的流程，还需要经过一定的等待期。

现在，保险公司现需要一个基于风控条件下的产品推荐模型，可以根据申请人的属性数据预测未来风险，此模型可以降低保险公司承担的风险，简化客户购买保险产品的流程并缩短流程时间，这样可以提高成单几率，增加客户满意度，改善公众对保险行业的看法。

# 数据处理

## 数据降维：

观察中我们发现某些列中存在数据丢失的现象，所以我们先对测试集和训练级中的空数据进行清理。若某列NAN占该列总数据的90%以上，就除去该列数据。相关列为Medical\_History\_10、Medical\_History\_24、Medical\_History\_32。

对测试集和训练集中的重复数据进行清理，若某个数字在本列存在比例在99%以上就去掉该列，由此我们去掉了24列数据。

通过对空数据和重复数据的清理，我们实现了数据的降维。这样的好处在于：减少了数据维度和需要的存储空间；节约模型训练计算时间；去掉了冗余变量，提高了算法的准确度；有利于数据的可视化。

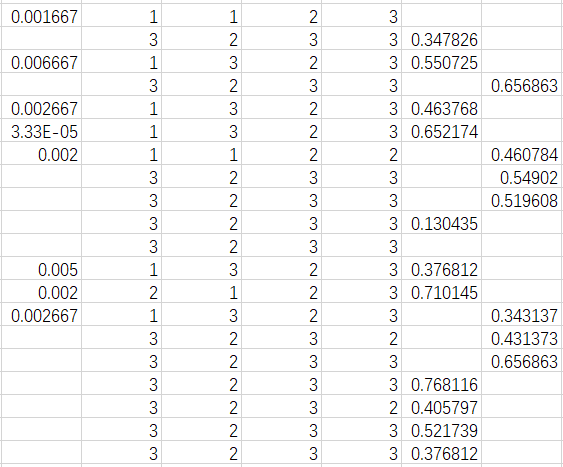


图2.1 部分数据丢失

## 数据类型转换

数据中共存在三种数据类型，分别为float64，int64和object。浮点类型的数据已完成标准化，整数类型的数据均为分类数据，便于数据的处理和模型性能的提升，我们将object类数据转化为int类型。

## 训练特征与标签

在经过上述处理的数据文件中，Response作为标签，将去掉Id的其余数据作为训练特征。Response分为8个级别，为便于模型训练，将1-8级改成0-7。

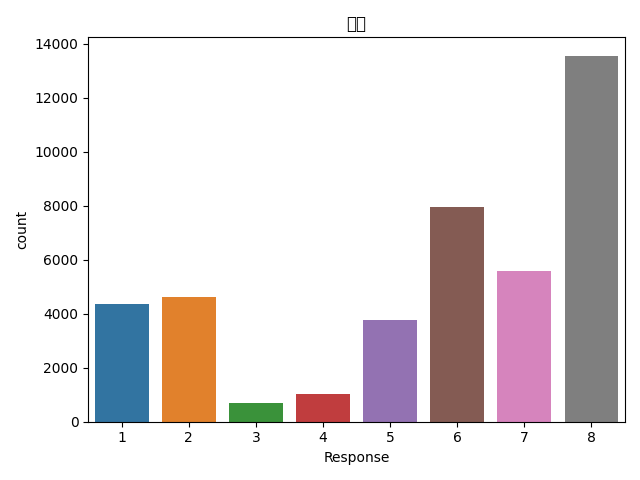


图2.2 Response的分布

# 建模思路与方法

## 模型选取

### 决策树模型：

该问题需要基于多个关于申请人及产品的属性数据并需要对申请人风险来进预测并推荐保险产品。目前使用较多的预测模型有神经网络（Neural Network），支持向量机（SVM），k-NN，决策树（Decision Tree）。神经网络模型适合的是图像、序列等有明显模式的数据，对于本题中的表格数据，只能用低效的全连接网络，在实践中效果不太好。SVM属于非参数方法，拥有很强的理论基础和统计保障，其损失函数拥有全局最优解，而且当数据量不大的时候，收敛速度很快，但本题所给的数据大、种类多，SVM效率并不是很高，有时候很难找到一个合适的核函数。决策树有直观的决策规则，可以处理非线性特征，考虑了变量之间的相互作用等优点，能很好地适配我们这次题目的数据。

### LightGBM算法:

目前广泛使用的算法有AdaBoost、GBDT、RF、XGBoost、LightGBM等。XGBoost算法虽然非常热门，是一种优秀的拉动框架，但是在使用过程中，其训练耗时很长，内存占用比较大。LightGBM是一个实现GBDT算法的分布式高效框架。它通过leaf-wise分裂方法进行决策树的生成，基于直方图的算法寻找特征分割点，支持并行学习，能够更高效的处理大数据。

直方图算法的基本思想是先把连续的浮点特征值离散化成k个整数，同时构造一个宽度为k的直方图。在遍历数据的时候，根据离散化后的值作为索引在直方图中累积统计量，当遍历一次数据后，直方图累积了数据的统计量，然后根据直方图的离散值，遍历寻找最优的分割点。

相比于常规的预排序决策树算法，直方图算法能够极大地降低内存消耗与计算代价。在实际应用中发现，**模型中分割点是否足够精确并不是很重要；较粗的分割点也有正则化的效果，可以有效地防止过拟合**；即使单棵树的训练误差比精确分割算法稍大，但在梯度提升（Gradient Boosting）的框架下没有太大的影响。

LightGBM优化了对类别特征的支持，可以直接输入类别特征，不需要额外的0/1展开。并在决策树算法上增加了类别特征的决策规则。相比0/1展开的方法，训练速度可以加速8倍，并且精度一致。据我们所知，LightGBM是第一个直接支持类别特征的GBDT工具，本题中分类数据特征超过100个，LightGBM更合适。

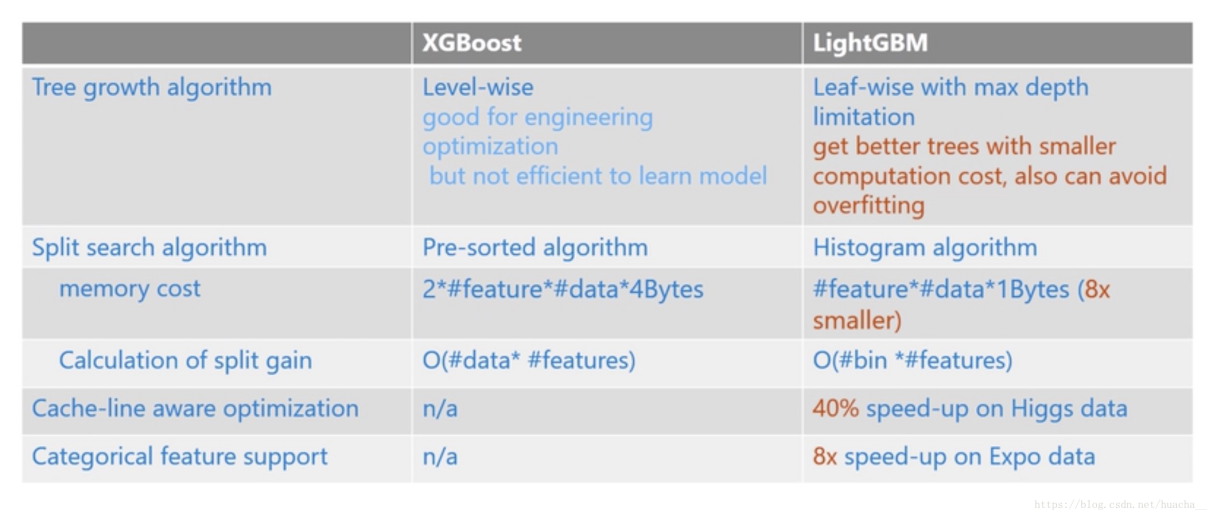


图3.1 XGBoost算法与LightGBM算法的比较

## 模型训练

首先按照4：1划分训练数据与验证数据，使用lightgbm包中的Dataset()方法为模型构建训练集和验证集。

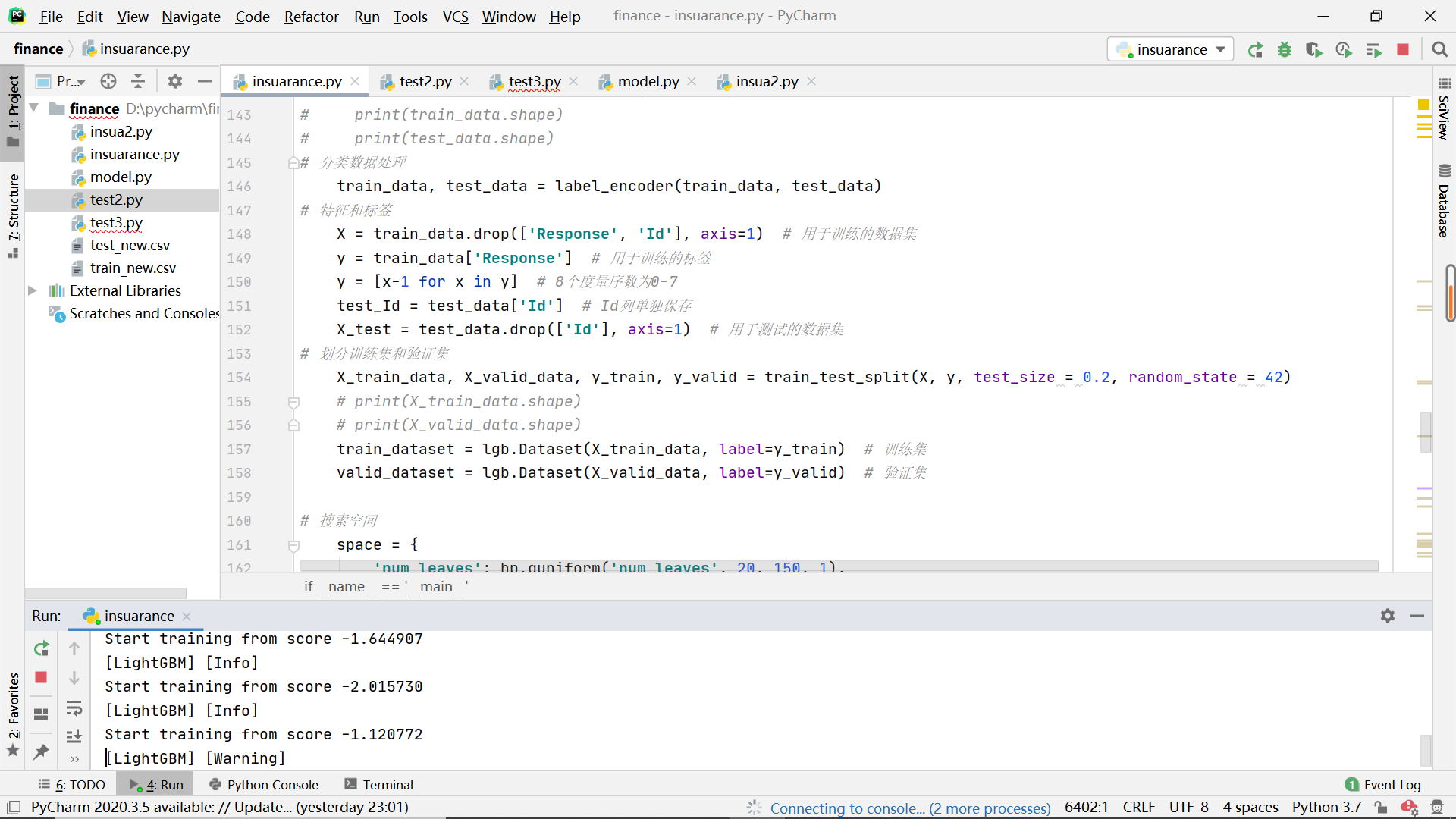


图3.2 分割训练集和验证集

由于该问题为多分类问题，分为8类，使用LGBMClassifier方法，并且使用训练好的最优参数进行建模，使用fit()方法训练模型。

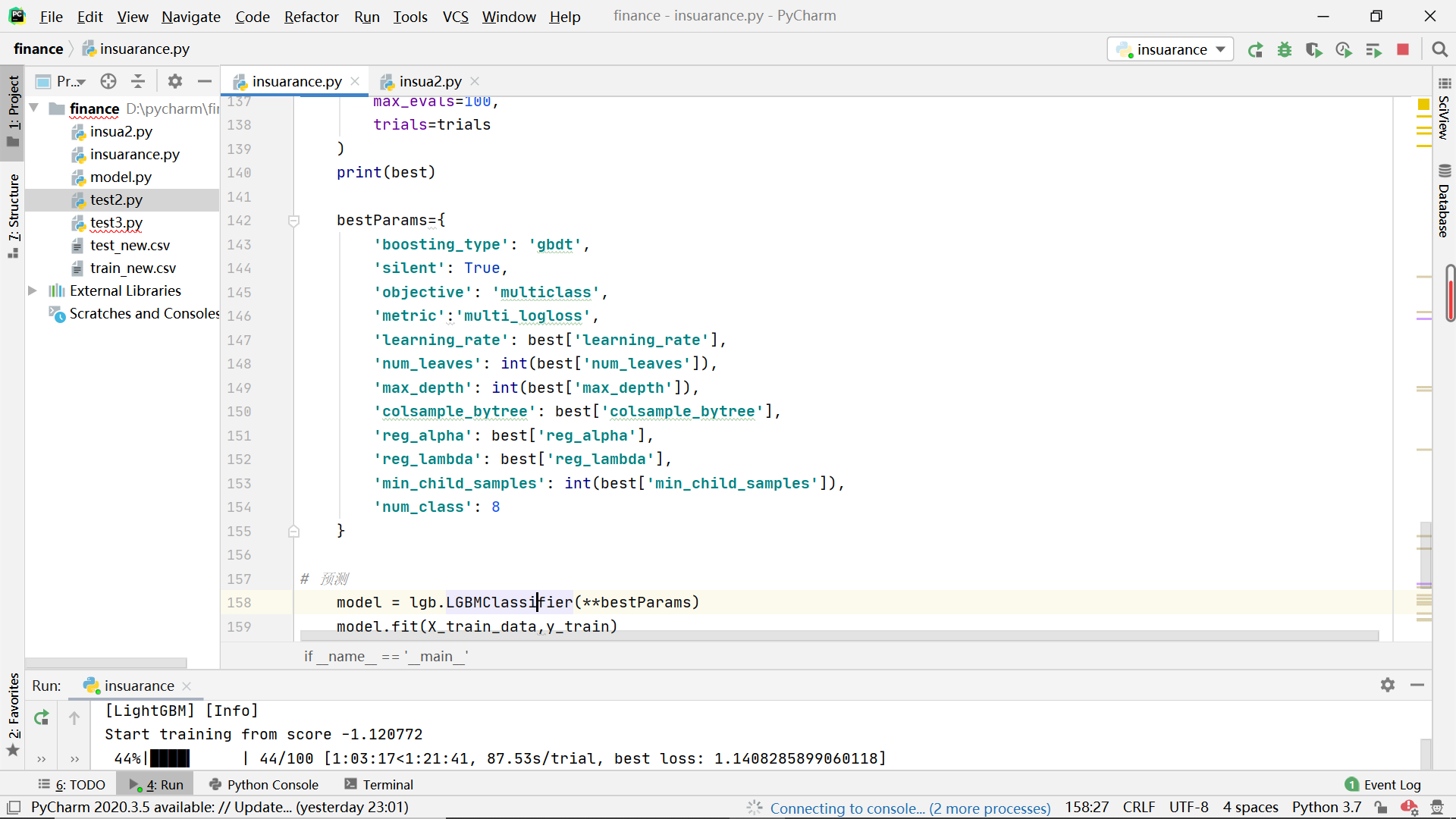


图3.3 最优参数

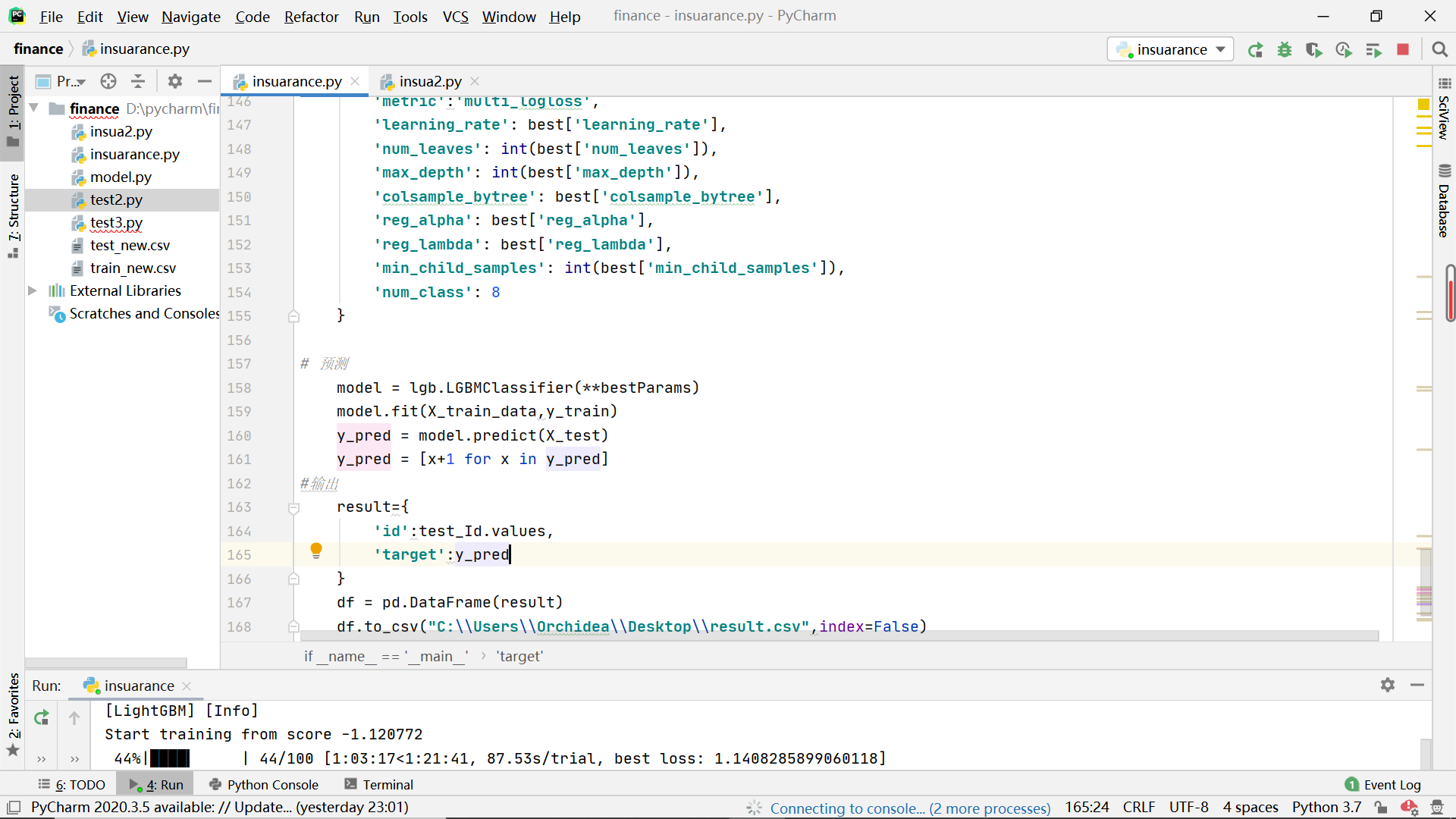


图3.4 模型构建

## 参数调优

由于手动调参需要大量的经验，且具有一定的偶然性，我们选择使用hyperopt贝叶斯搜索自动调参。

hyperopt自动调参需要四个参数，分别是目标函数、搜索空间、优化算法以及调整次数。

### 目标函数（objective）：

使用LGBMClassifier构建基线模型，每次迭代的树的深度（max\_depth）都在新一轮参数的基础上增加10，学习率（learning\_rate）使用参数的学习率，其他参数则通过随机种子（random\_state）来默认构建。

使用5折交叉验证对模型训练，lightgbm包中提供cv()方法来实现交叉验证，每次验证10000轮(num\_boost\_round)，如果200轮时无明显提升则停止(early\_stopping\_rounds)，评价标准则使用“多分类对数损失（multi\_logloss）”，将其中最小的平均损失（multi\_logloss\_mean）返回。

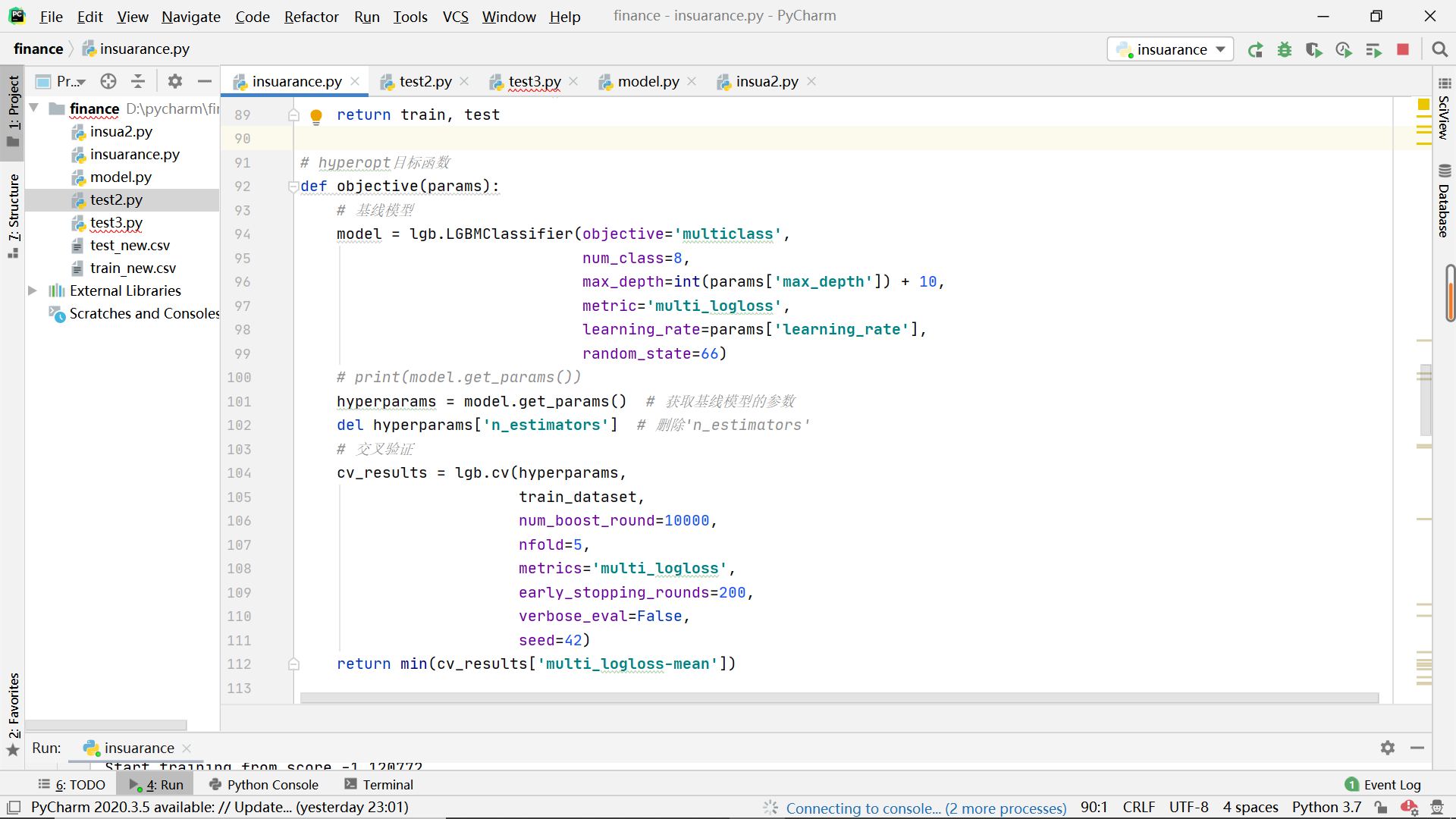


图3.5 目标函数（objective）

### 搜索空间（space）：

搜索空间规定了参数调整的范围，提高了搜索速度，能够在合理范围内更快地得到最好的训练参数。本模型中使用的搜索空间如图3.6。

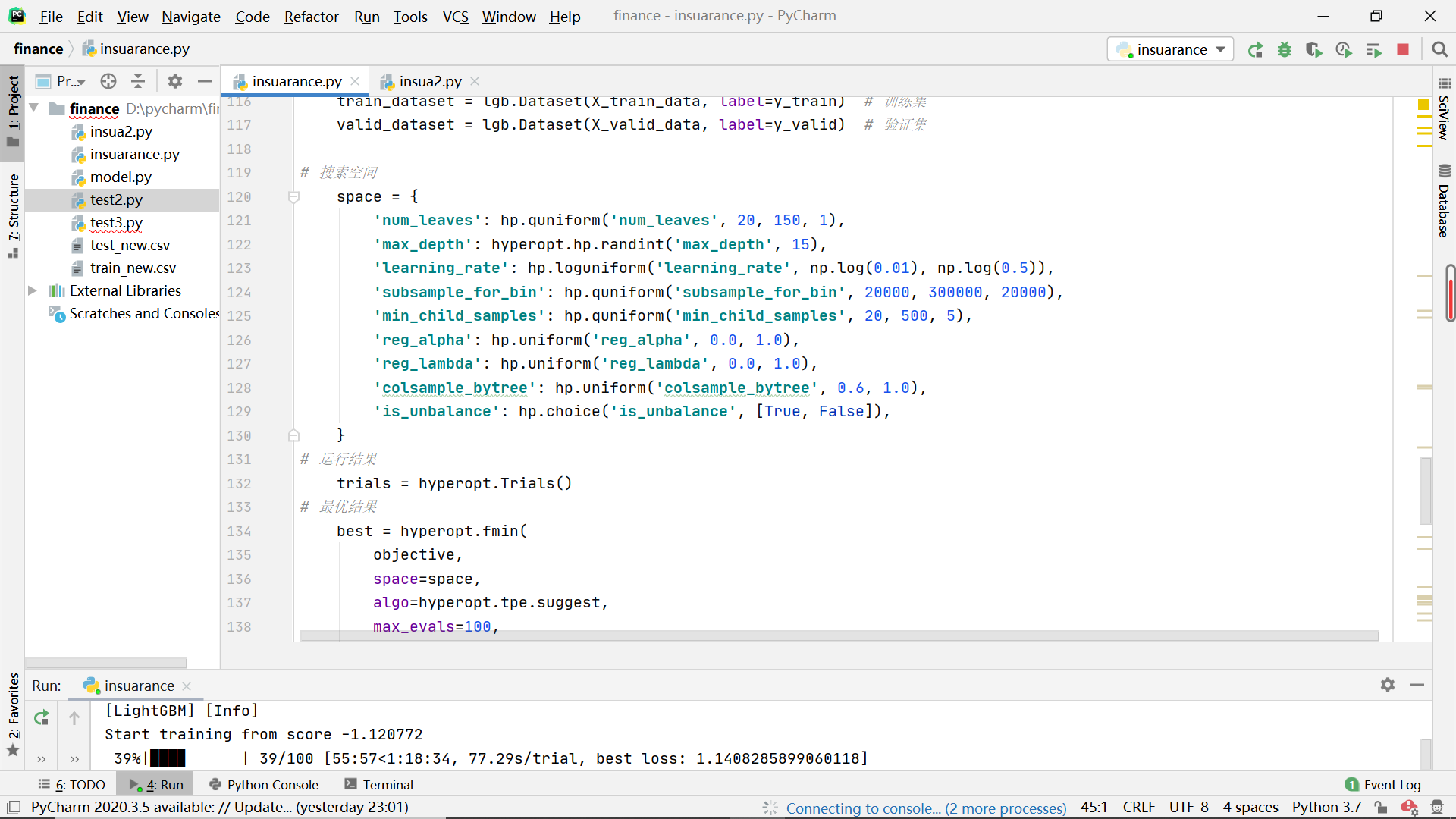


图3.6 搜索空间

# 模型评价

模型采用“多分类对数损失”来评价。最终模型得到的评分范围在1.13~1.15之间，主要取决于搜索的深度和调优的次数。损失随着搜索树的深度增加和调优次数的增加而减少。模型速度较快，在1分钟左右可以完成1次调优迭代，即10000次5折交叉检验。最终结果以csv文件格式输出，文件名为result.csv。

附件：result.csv

