统计机器学习课程论文：地震后建筑物损毁程度预测

郜源善 张其乐 廖乐乐 李春子

摘要：地震是一种破坏性十分严重的自然灾害，因此，运用震前数据进行地震预测具有十分重要的现实和理论意义。本文中，我们使用 Driven Data 网站提供的尼泊尔某次地震的房屋震前震后情况数据集，并依次按照数据集探索性分析、样本采样均衡、特征工程、基学习器构建、模型融合的环节构建了较为准确的房屋在地震中受灾的程度预测模型，最终集成模型的准确率能够达到 80.51%。

关键词：特征工程 XGBoost LightGBM SVC MLP 集成模型 定序分类数据

一、 问题背景及研究计划

地震作为一种自然灾害，对社会财产及人民生命安全的危害是不可估量的，其中首要的是对房屋和基础设施的破坏。在地震发生后，对房屋破坏程度进行合理的评估和预测是十分重要的。只有正确地预测地震灾区的房屋受损情况，才有可能在地震发生后及时将物资空投到生存率较高的地方、将空降兵战士投放到相对安全的地方，从而保证救援的正常开展。最后，我们还能通过变量重要性，明确地震局等机构在地震前对地区的观测目标，以便在发生地震后能进行更准确的评估进而展开针对性地救援。

研究计划：首先为验证模型效果留出验证集，并做必需的重复留出训练集划分。通过对训练集的数据做基础变量分析，同时根据变量的特性进行数据预处理及特征工程。在建模阶段，建立多种树模型、SVC 支持向量机、MLP 神经网络进行预测，并选择效果最好的树模型与其他两种模型进行模型融合得出最终的模型。

二、 数据集的基本情况

本数据集来源于 Driven Data 的数据竞赛，此次数据竞赛的目的是根据建筑位置和建筑结构的不同特征，预测 2015 年尼泊尔地震对建筑造成的破坏程度。数据集的样本量为 260602。

数据竞赛的验证集一般由官方保留，并对参赛者的模型进行评价排序。但由于我们在课程任务中需要对整体模型的效果进行评价，所以需要提前划分出验证集用于模型评价。**留出测试样本的方法：**利用分层抽样的办法，保证验证集的目标变量与训练集中的各类的比例基本一致，这也考虑了现实可能存在的类别不平衡问题。具体而言，我们留出了全部数据集的 5%，得到 247572 个训练样本与 13130 个测试样本。

首先**对目标变量进行分析**，确定模型的建模方向。本文的数据集中，目标预测变量是分类变量，取值为 1、2、3，表示地震对建筑造成的破坏程度。因此，我们将问题确定为分类问题，建立多种不同的分类器预测类别。进一步发现，预测的目标不同于普通的多分类问题，因为就地震对建筑造成的破坏程度而言，分类变量存在**顺序关系**，数值越大表示破坏程度越大。后续我们会建立针对本文定序

变量预测问题的损失函数及评价标准，使模型的结果更加符合竞赛问题的要求。

在对目标变量 Y 的分布进行分析后，我们发现其中存在比较明显的**类别不平衡**问题，需要利用重抽样的方式解决。利用三次的 Bootstrap 方法得出三个不同子模型的训练集，并同步得出对应的袋外样本，用于三个子模型的评价。图 1 呈现了三类损毁程度的房屋的分布情况，其中，损毁程度中等的房屋最多，损毁程度最轻的房屋最少，三类房屋的个数明显不均衡。

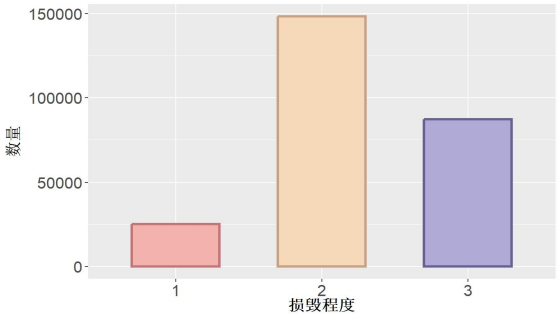


图 1 三类损毁程度的房屋的分布情况

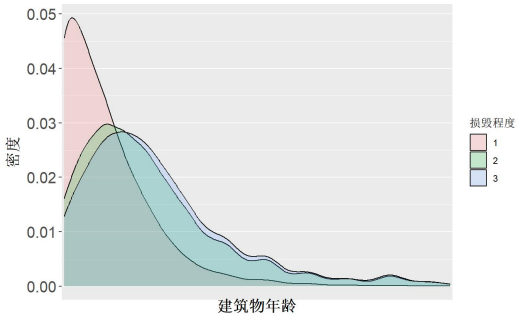


图 2 三类损毁程度的房屋的年龄分布

表 1 为协变量的数据说明表。在竞赛提供的数据集中，出现了明显的“定性变量多定量变量少”的情况，因此在后文的特征构造中，需要更多考虑定量变量的构造。值得注意的是，由于竞赛数据的隐私保护，将小部分的分类变量赋值为未知含义的标签，所以我们可能无法对此类变量进行有效的特征处理。与此同时，我们也发现部分变量会体现出天然的影响效果，例如建筑物年龄、建筑物高度都能相对直接地影响房屋在地震灾害中的破损情况。一般来说，建筑物年龄越大、建筑物高度越高，在地震中都会倾向于更严重的受损程度。针对这部分相对重要的变量，会在后续特征工程重点考虑。

表 1 数据说明表

变量名称	变量含义	变量类型	变量取值
count_floors_pre_eq	层数	定量	1~9
age	建筑物的年龄	定量	0, 5, 10, …, 200, 995
area_percentage	正则化面积	定量	1~100
height_percentage	正则化高度	定量	2~32
count_families	住在该建筑中的家庭数量	定量	1~9
geo_level_1_id, geo_level_2_id, geo_level_3_id	所处的地理区域，从最大的（1 级）到最具体的次区域（3 级）	定性	1 级：0-30，2 级：0-1427，3 级：0-12567
land_surface_condition	建筑物所在土地的表面状况	定性	3 类
foundation_type	地基类型	定性	5 类
roof_type	屋顶类型	定性	3 类
ground_floor_type	底层的类型	定性	3 类
other_floor_type	比地面层更高的建筑类型（屋顶除外）	定性	4 类
position	建筑物的位置	定性	4 类
plan_configuration	建筑物平面构设	定性	10 类
legal_ownership_status	所在土地的合法所有权状况	定性	4 类
has_superstructure (several types)	有无某种上层建筑材料（如粘土泥浆、混凝土等）	定性	0：无该材料 1：有该材料
has_secondary_use	建筑物是否被用于任何次要目的	定性	0：无次要目的 1：有次要目的
has_secondary_use (several types)	是否被用于某种次要目的（如学校、医院等）	定性	0：无该次要目的 1：有该次要目的

三、 数据探索及清洗

一、变量信息挖掘

图 2 和图 3 显示，不同分类变量下，房屋损毁程度占比的差别很大。以地基材料类型为例，第*i*种地基材料的房屋的抗震能力较强，表现在第 3 种损毁程度的房屋占比极小，第 1 种损毁程度的房屋占比很大。综合图表，各种分类变量与损毁程度之间有很大的相关性，将在后续模型中起到重要作用。

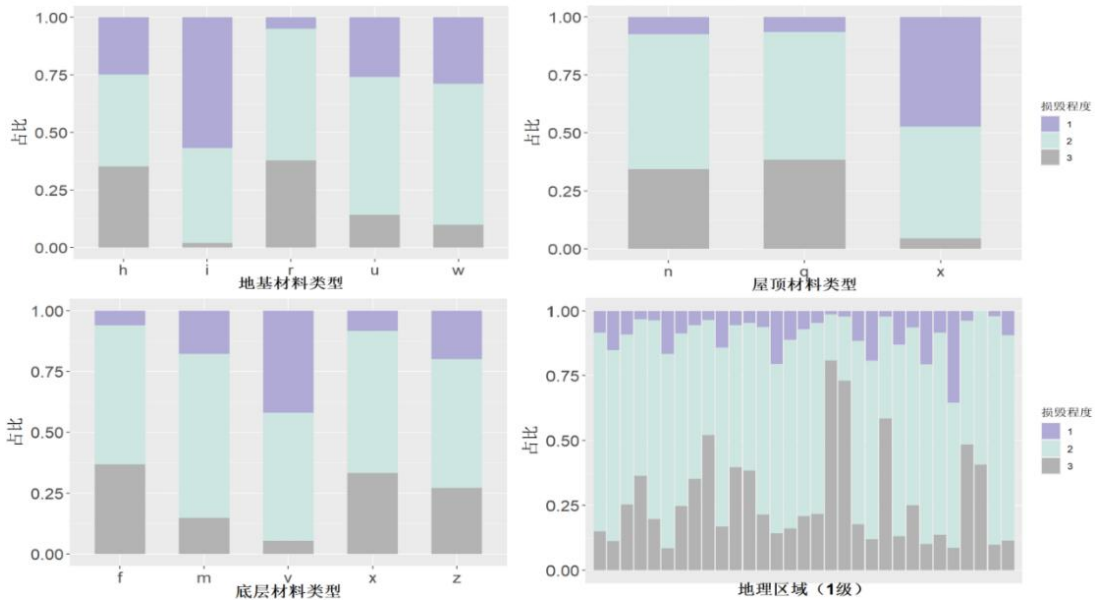


图 3 不同分类变量下，房屋损毁程度的占比

二、缺失值、异常值

竞赛提供的数据集相对完整，未出现缺失情况，无需进行缺失值处理。同时我们对分类变量进行独热编码。数据集中的异常值主要出现在以下三个不同部分：

- 首先，是目标变量 Y，存在个别样本的 Y 被赋值为 0，不符合数据集对目标变量的设定，所以我们对其中目标变量 Y 为 0 的异常值删除处理。
- 其次，变量之间存在矛盾的问题，例如“是否某种上层建筑材料”是 0-1 变量。同时，后面列举了多种不同的上层建筑材料。后面数列出现取值为 1 但汇总列取值为 0 为异常值，做删除处理。
- 最后，建筑年龄普遍分布在 1-200，但一部分样本的建筑年龄为 995，显著偏离大部分值。这可能是因为尼泊尔地区存在古建筑，会有 995 年的年龄，从而保留这部分的异常值，不做额外处理。

四、 特征工程

一、初步特征重要性评价与特征选择

我们首先利用整体训练集完成了一个简易的随机森林，得到初始的分类准确率 baseline 为 57%，并得到对于变量重要性的刻画。由于总变量个数为 65 个，相对于目前大多数机器学习的算法能力而言并不算高维数据，我们的目标算法能够比较快速的计算出结果，因此不需要在最开始进行特征选择。

但变量重要性可以为后续的特征构造提供参照的依据，例如，我们可以更加重视相对重要性强的特征。

二、 特征构造

构造 1 数据中包含建筑的地理位置信息，从最大的区域（1 级）到最具体的次区域（3 级）。由于区域类别过多，需要进行特征处理。不同区域与震中的距离不同，距离越近，受到地震的影响程度越大，房屋损毁程度越高。同时，根据房屋的损毁程度可以推断区域距离震中的距离。

在交叉验证时，根据每次的训练集，计算每个区域的震中距离，用震中距离指标代替三级区域指标纳入模型，进行训练。震中距离指标计算过程如下：

Step1：1、2、3 级区域的编号三元组对应一个具体区域。分别计算出 1、2、3 级区域的房屋损毁程度均值，代表每一区域平均的房屋损毁程度。三个级别区域均值的合理加权平均可以作为该具体区域的损毁程度，从而推断震中距离。

Step2：分别计算出 1、2、3 级区域的房屋损毁程度的标准差。1 级不同区域的标准差之间具有可比性，标准差越大，说明该区域各次级区域的损毁程度差距较大，因此该区域的损毁程度均值对于判断具体区域的损毁程度的作用较小。$\sigma_{r,max}$代表第 r 级各区域的损毁程度标准差的最大值，$\sigma_{r,i}$代表第 r 级第 i 个区域的损毁程度标准差。$\theta_{r,i} = \sigma_{r,max}/\sigma_{r,i}$代表第 r 级第 i 个区域的标准差的相对大小，$\theta_{r,i}$越大，代表该区域相对标准差较大，对于判断的作用越大。

Step3：加权平均得到每个具体区域的距震中距离：

$$\xi_j = (\theta_{1,i}\mu_{1,i} + \theta_{2,i}\mu_{2,i} + \theta_{3,i}\mu_{3,i})/(\theta_{1,i} + \theta_{2,i} + \theta_{3,i}),$$

其中$\mu_{r,i}$代表第 r 级第 i 个区域的损毁程度的均值，$\theta_{r,i}/(\theta_{1,i} + \theta_{2,i} + \theta_{3,i})$代表第 r 级区域损毁程度对于震中距离推断的作用大小。加权得到的损毁程度越大，则很大可能距离震中越近。

构造 2 层高：层高等于建筑物高度除以层数，层高将影响建筑物的稳定性，比如平层的抗震性好于跃层、复式和错层户型的房子。

构造 3 人口密度：人口密度等于家庭数除以建筑物面积，我们推测人口密度可能影响建筑物的使用情况从而影响在地震中的受损程度。

构造 4 二次用途总数：经推断，房屋的二次用途越多，使用程度越大，受损程度可能更高。

构造 5 上层建筑材料：根据上层建筑材料的牢固程度，得出建筑材料坚固性的加权均值。

五、 模型训练和验证

1. 模型选择、调参及模型验证

在上文解决无验证集问题后，我们将剩余的样本作为综合训练集，并作为下文集成模型的训练集。在集成模型的构造中，我们主要使用了 Bagging 方法，完成集成的工作。而在集成学习的基学习器选取上，我们在树模型中比较出一个效果最好的模型，还尝试拟合支持向量机 SVM 和神经网络 MLP 模型，在分别调参以保证基学习器有较好的预测表现后，进行集成。

在 Bagging 模型中，为保证基学习器的多样性，我们利用 Bootstrap 抽样法得到训练集。考虑到样本不平衡问题，我们在自助抽样法中进一步利用欠采样和过采样的方式得到子训练集，三个子训练集的样本量为 21 万。由于 Bootstrap 采样方法会天然得到袋外样本，我们利用每个子训练集对应的袋外样本作为对应模型的验证集。而在每一个基学习器中，我们考虑九折交叉验证的方式进行调参。我们将训练集划分为九份，每次利用其中一份作为验证集其余作为训练集，并将每次得到的评价进行平均作为当次参数的效果评估。通过这种方式选出最优参数。确定超参数后，利用整个子训练集对模型进行训练，并利用袋外样本进行评价。

2. XGBOOST

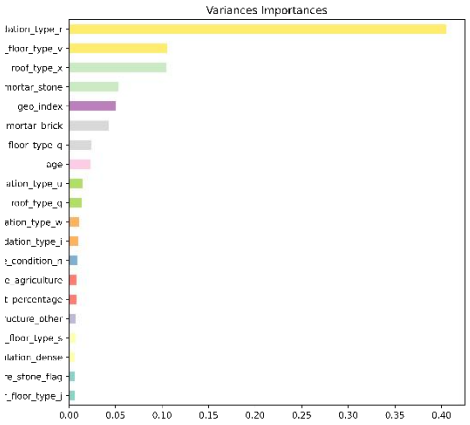
在 XGBOOST 模型的构建中，我们主要介绍模型调参的问题，其中利用网格搜索与联合调参的方式确定最优参数组合。除了模型基本参数的调整外，我们注重正则化系数与影响模型运行时间的参数。在后文其他树模型中不重复叙述模型调参的具体细节。

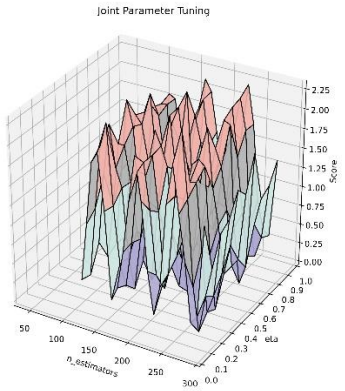
首先进行第一组联合调参，是学习率 learning rate 与基学习器个数 n_estimator。在这组参数中，基学习器的个数越少，会导致模型对样本学习的程度不足，所以需要较大的学习率增强模型的学习能力。而图中对应的结果验证上述说法，为了节省运行时间，我们选择了较小的基学习器个数 n_estimator，同时保证了不错的拟合程度。第二组的联合调参是最大树深度 max_depth 与分枝停止规则阈值 gamma，在这组模型中，但最大树深度限制了树的结构不会过于复杂，所以应该进行放宽分枝停止规则阈值 gamma 的要求，选择较小的值，让有效的分枝情况进可能出现。在完成两步联合调参后，我们对模型的正则化系数的调整，保证模型不会出现严重的过拟合现象。我们调整的正则化参数包括：subsample、col_sample_bytree、alpha。其中，系数 alpha 能在调整 L1 惩罚项的系数，同时完成特征选择，减少了特征维度提升模型的泛化能力。

在 XGBOOST 模型中，优化得到参数取值与变量重要性如下表 2 所示。

表 2 XGBoost 调参结果

参数名称	n_estimator	eta	subsample	gamma
调参范围	[50,300]	[0,1]	[0,1]	[1,45]
最优参数	200	0.40	0.60	35





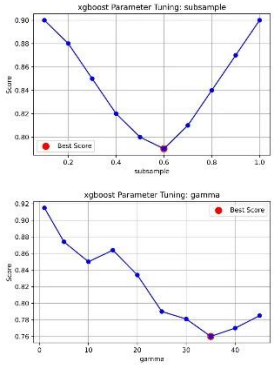
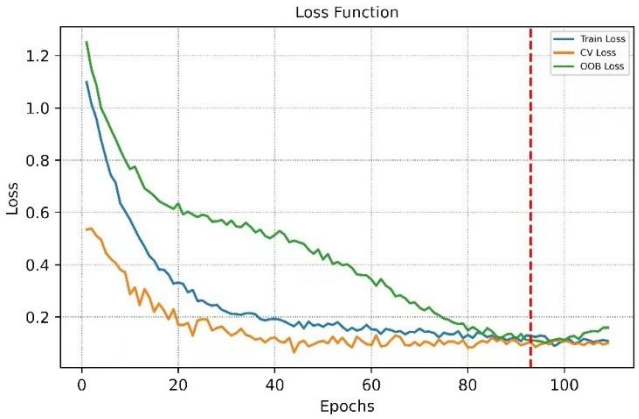


图 4 变量重要性图形　　图 5 XGBoost 调参效果图

在确定最优参数取值后，利用全部的子训练集进行训练，得到模型效果评价与训练轮数的关系如下图，最优的迭代次数为 93。



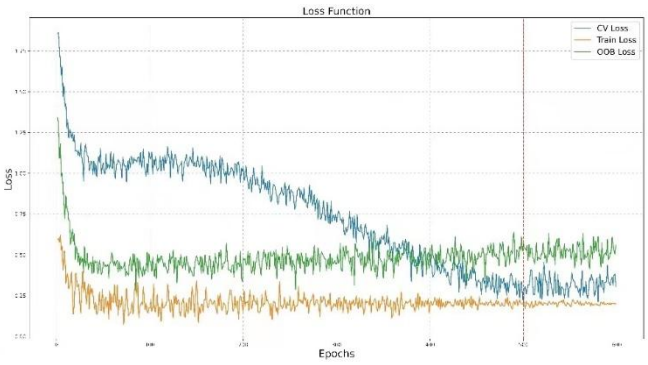


图 6 XGBoost 损失函数图　　图 7 LightGBM 损失函数图

3. LIGHTGBM

尝试建立 LightGBM 模型，并利用算法中互斥特征捆绑 EFB 的方法对我们训练集中稀疏变量较多的情况进行处理。

由于在训练集特征工程中，我们对定性变量进行了独热编码的处理，我们在 LightGBM 中需要提前整合回到分类变量。因为在 LightGBM 的 EFB 算法中，会将稀疏的变量整合起来，而独热编码可能得到多个稀疏的变量。为了避免合并稀疏变量时的盲目性，我们加入变量的信息，也就是不做独热编码，将定性变量直接加入模型中。而针对后续部分的非独热的系数变量，我们还是采用 EFB 的算法简化模型的特征，提升模型的运行速度。

在 LightGBM 模型中，优化得到参数取值如下表所示。LightGBM 模型最终确定迭代轮数为 500。

表 3 LightGBM 调参结果

参数名称	n_estimator	eta	num_leaves	max_depth
调参范围	[10,200]	[0,1]	[10,100]	[1,100]
最优参数	155	0.35	50.00	20

4. 树模型部分小结

对于其余两种树模型 AdaBoost 和随机森林，我们进行了适当的参数优化，例如联合调参和正则化参数的优化，得出了模型拟合效果较好的模型。而四种树模型的具体评价与表现如所示。

表 4 三种不同树模型的建模结果

树模型	XGBoost	LightGBM	Adaboost	Random Forest
预测准确率	69.56%	68.12%	64.39%	65.16%
加权累计错分率	29.58%	31.69%	34.46%	33.64%

5. 支持向量机

在支持向量机建模中，我们希望尽量简化模型中的特征。因此，我们建立了一个小样本的模型，目的是在时间耗费少的同时得到变量重要性的刻画，为整体模型的构建提供参照。在 4 万小样本模型的构建中，我们首先利用网格搜索的办法确定了核函数的类型和代价参数 C，然后进行核函数中参数 gamma 的优化，得到模型的准确率达到 0.62。利用置换法确定的变量重要性如下图所示，将重要性系数在 0.08 以下的变量剔除出去，剩下 16 个影响力较强的变量，缓解了可能存在的维度过高的问题。

在使用全部子训练集的大模型中，我们使用 10 折交叉验证和网格搜索结合的方式确定了核函数的类型为高斯径向基函数和代价参数 C，然后对高斯径向基函数中的参数 gamma 优化，确定的参数如下图所示。最终，支持向量机函数得到的模型准确率为 66.59%。

表 5 支持向量机调参结果

参数名称	核函数	C	gamma
调参范围	多项式、Sigmoid、RBF	[0.00001,1000]	[0.00001,100]
最优参数	RBF	32	0.03125

6. 深度学习

为了检验深度学习模型在表格数据上的预测效果，我们构建了 MLP 模型作为分类器模型，这里我们主要考虑经过设计的损失函数是否会对于定序分类数据起到更加优良的分类效果。

我们的 baseline 模型由 4 层全连接的线性层组成，各层神经元数量依次为 140，70，35，17，损失函数为 RELU，最后用 softmax 函数归一化输出三类对应的三元概率向量组，以结果的期望与某类别的最近距离作为最终的分类类别，设定的损失函数为普通的交叉熵损失。经 20 轮训练之后，所得的测试集分类准确率约为 65.57% 。

由于我们的数据集的标签表示了受灾的严重程度，但标签之间的绝对距离与受灾程度的差异并不对应，因此为定序分类数据。传统的交叉熵损失为$-\sum_{i=1}^{n}\sum_{k=1}^{c} y_{ik} * log(p_{ik})$，假如使用该损失函数作为模型效果的评价指标，无法关注到定序数据的序列特征，因此我们重新设计损失函数为

$$\sum_{i=1}^{n}\sum_{k=1}^{c} y_{ik} \times log\left(1-\sum_{j=1,j\neq k}^{c}\frac{w_j}{\sum_{j=1,j\neq k}^{c} w_j}\times p_{ik}\right),$$

其中，w_i为权重参数且$\sum_{i=1}^{c} w_i = 1$。由于将标签为 1 的样本分类到标签为 3 的样本相对于将标签为 1 的样本分类到标签为 2 的样本更为严重，因此我们对于w_1和w_3赋予更大的值以获取区分惩罚效果。仅调整损失函数对 baseline 模型进行 20 轮重新训练之后，我们在测试集上的分类准确率上升到 73.08%。

在对其余超参数进行网格搜索调参之前，我们考虑到对于定序分类数据的模型分类效果评价标准也应当考虑到值标签的定序特征，因此我们也对评价指标做了新的思考和修改。我们首先计算模型输出结果的混淆矩阵，并以对应位置的样本量在测试集中所占比例代替该元素。随后，我们将i,j错分率以$\frac{w_i}{\sum_{j=1,j\neq i}^{c} w_j}$加权后进行累和作为“加权累计错分率”，对于正确分类的样本量所占比例则不计入该指标。

综上，MLP 模型的参数选择结果如下表所示，加权累计错分率是 19.55%，模型拟合效果较好

表 6 MLP 调参结果

参数名称	网络层数	损失函数
调参范围	3, 7, 10, 15	["RELU","Tanh","Sigmoid"]
最优参数	7	"RELU"

。

六、 模型融合

在得到基学习器在验证集上的预测结果后，为了保证模型结果的合理性，我们对同一样本对应的三个预测结果进行比较。发现三个模型得到的预测结果一致性较强，说明基学习器预测效果基本统一，具有较强的可信度，下面进行模型融合。

在前文中已经提到，我们所获得的三个基学习器是基于不同的 Bootstrap 样本训练的，因此天然具有较大的样本独立性。

在文章的最后，我们决定采取 Bagging 的方法集成三个模型，期望取得更好的分类预测效果。我们主要采取了软投票策略。具体策略为直接取三个基学习器的预测概率三元向量组进行平均，计算期望之后取最近整数作为预测类别。进行 Bagging 的集成后，我们的预测准确率达到了 80.51%。

表 6 集成模型结果

集成状态	XGBoost	SVC	MLP	Bagging 模型
预测准确率	69.09%	67.43%	73.08%	80.51%
加权累计错分率	34.91%	37.87%	19.55%	23.73%

七、 结论与反思

关于尼泊尔地震中建筑损毁程度的预测问题，我们决定采用集成学习的方法。首先预留出测试集，接着运用 bootstrap 法解决样本不平衡的问题。建模前，我们对变量的特征进行充分探索，实施特征工程，找到部分关键变量，比如将地理位置转化为与震中的距离。

在模型构建中，构建了多个树模型、SVC 和 MLP 作为备选的基学习器，且通过调整参数得到较好效果的基学习器。其中，树模型中 XGBOOST 的模型效果最好，地理位置等特征是其中相对最重要的特征。SVC 模型经过联合调参和优化正则化参数，正确率达到 66%。MLP 模型利用了针对定序变量新定义的损失函数，累计加权错分率为 19.55%。在此基础上，通过 Bagging 进行模型融合，构建了针对于定序分类变量的交叉熵损失和错分率，最终正确率达到 80.51%。

根据变量重要性得出，地基种类、与震中的距离是最重要的影响变量，对于处在地震高发区的建筑物，需要规范建筑材料、建筑规格，考虑周围地理环境，从而防患于未然，减轻地震损失。

对于我们的建模过程，我们也有以下可以继续改进的空间：Bootstrap 方法浪费了部分样本，SVC 模型由于训练复杂度问题未能得到在未降维前的数据集上的最优参数组合，对 MLP 模型针对性设计的损失函数的权重设定存在主观因素影响问题。