基于随机森林模型的地震后建筑受灾等级预测

蔡梓铭, 陈霖, 林元莘, 温作凯

2024 年 1 月 11 日

摘要

本次报告根据 2015 年 4 月尼泊尔大地震的震后数据集，对地震引起的建筑损害数据集进行了深入分析，囊括数据集概述、探索性数据分析、数据预处理及建筑受灾等级预测模型的构建。其中，探索性数据分析深入探讨了受灾等级分布、基于地区的聚类，以及连续型和分类型变量的分析。在数据预处理阶段，讨论了去除未来信息、生成虚拟变量和交互项的必要性。模型建立部分，对比了逻辑回归、朴素贝叶斯、决策树和随机森林等不同模型在预测建筑受灾等级方面的有效性。最后根据数据分析以及模型，给出关于地震后建筑受灾等级的分析以及尼泊尔当地情况的分析。

关键词：层次聚类; 特征工程; 随机森林

目录

1 数据集简介

1.1 数据集背景

本次报告所使用数据集来自 Kaggle 竞赛网站上的数据集Earthquake Magnitude, Damage and Impact。

在 2015 年 4 月，7.8 级的格尔克地震发生在尼泊尔甘达基省的格尔克地区附近。这场地震造成了近 9,000 人丧生，数百万人瞬间无家可归，损失高达 100 亿美元——约占尼泊尔名义 GDP 的一半。在此后的几年中，尼泊尔政府积极重建受影响地区的基础设施。在这个过程中，国家规划委员会、加德满都生活实验室以及中央统计局共同生成了有史以来最大的一份灾后数据集，包含了关于地震影响、家庭状况以及社会经济人口统计的宝贵信息。

1.2 数据集特征介绍

本次报告主要针对数据集中的“csv_building_structure.csv”文件进行数据分析。并通过建筑地震前的特征变量，来对响应变量受灾等级”damage_grade”进行预测。

该数据集含有 762106 行 *30 列数据。这个数据集主要包含了来自 11 个区县的建筑物结构的信息。数据集中的每一行代表了地震袭击区域的一栋特定建筑。其中，建筑物结构特征是连续型和离散型变量混合的。

表 1: 数据集特征介绍

特征名称	特征意义	数据类型
building_id	地震袭击区域建筑标识符	整数
district_id	该栋建筑所在县区编号	整数
vdcmun_id	该栋建筑所在乡村/城市行政单位编号	整数
ward_id	该栋建筑所在街区编号	整数

表 1 – 续上页

特征名称	特征意义	数据类型
count_floors_pre_eq	地震前该建筑楼层数	整数
count_floors_post_eq	地震后该建筑楼层数	整数
plinth_area_sq_ft	建筑基座面积（平方英尺）	整数
height_ft_pre_eq	地震前该建筑高度（平方英尺）	整数
height_ft_post_eq	地震后该建筑高度（平方英尺）	整数
age_building	建筑年龄	整数
land_surface_condition	陆地表面条件	类别
foundation_type	建筑基础材料类型	类别
roof_type	屋顶建筑材料类型	类别
ground_floor_type	地面楼层建筑材料类型	类别
other_floor_type	其他楼层材料类型	类别
position	该建筑和其他建筑相连情况	类别
plan_configuration	该建筑规划结构形状	类别
has_superstructure_ X	该建筑上部结构是否由 X 材料构成	布尔
condition_post_eq	地震后该建筑受损情况	类别
damage_grade	该建筑受灾等级 (等级 1 至等级 5)	类别
technical_solution_proposed	计划修复该建筑的技术方案	类别

2 探索性数据特征分析

在深入到预测模型和复杂的统计分析之前，我们首先尽量对数据集进行了详尽的探索性数据分析。这一步骤对于揭示数据的内在结构、识别关键变量以及检测任何异常或异常模式至关重要。EDA 不仅帮助我们更好地理解数据，还为之后特征工程以及选择合适的预测模型指明了方向。

2.1 受灾等级分布

针对响应变量受灾等级，考虑所有建筑物的受灾等级分布：

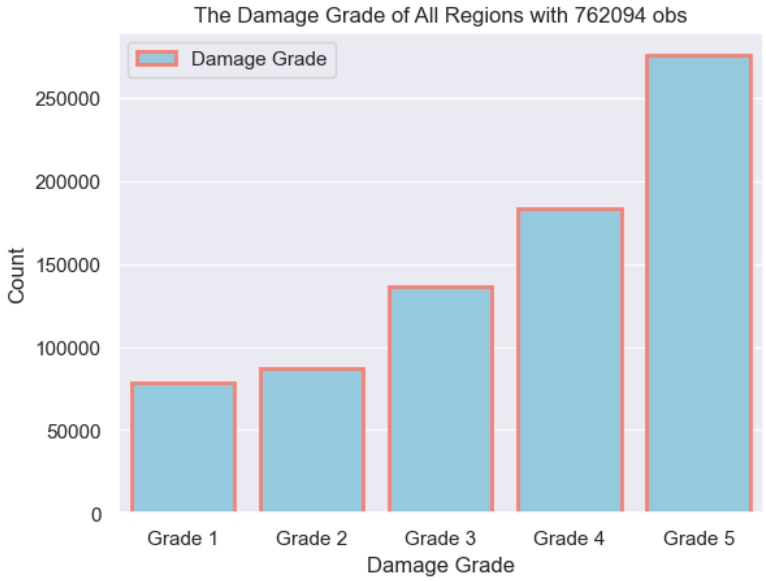


图 1: 所有建筑物的受灾等级分布

可以看到，在 762094 个建筑物的受灾等级中，受灾等级为 5 的占据了大部分，并且数量从等级 1 到等级 5 依次递增。

由于建筑物来自 11 个不同的区县，不同区县的地理差异较大。并且由于地震的性质，震源中心向外辐射强度逐渐减少。这对于整体预测会有较大影响。所以我们考虑分区县受灾等级分布4。

可以看到，每个区县的受灾等级分布十分不均匀。例如 Sindhupalchock（也就是最后一张图），受灾等级为 5 的建筑非常多，数量远远大于其他等级的建筑；而例如 Okhaldhunga（第二行第三张图），受灾等级相对均匀。地理位置的不同，很大程度影响了每个地区的受灾等级分布，从而对整体进行预测，会产生很大的偏差。

同时，我们尝试将受灾等级分布映射到真实的尼泊尔地图上3，以此观察受灾等级分布和地理位置的关系，从而可以进行关于区县的聚类。

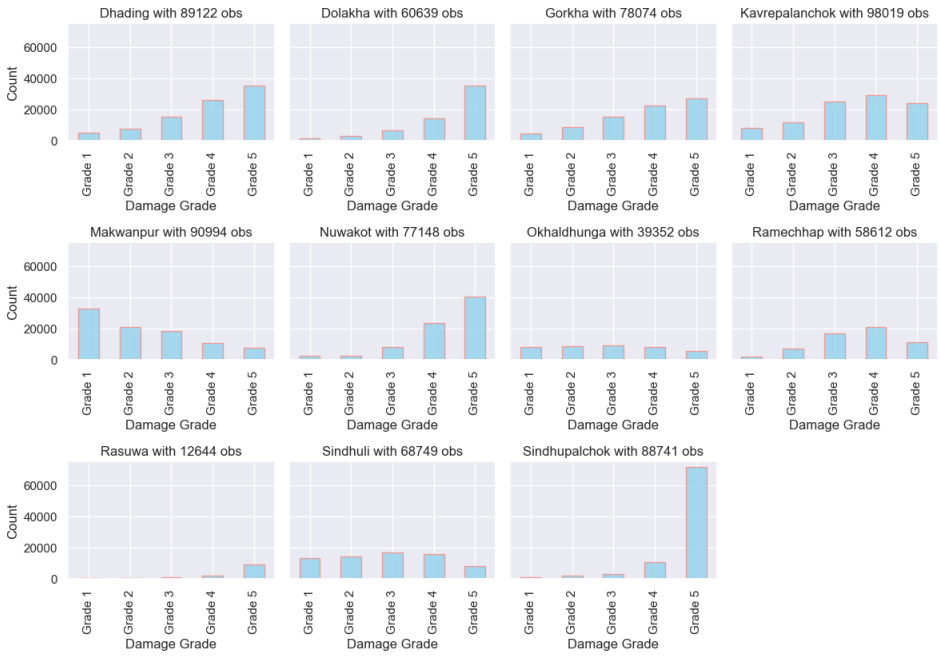


图 2: 分区县的受灾等级分布

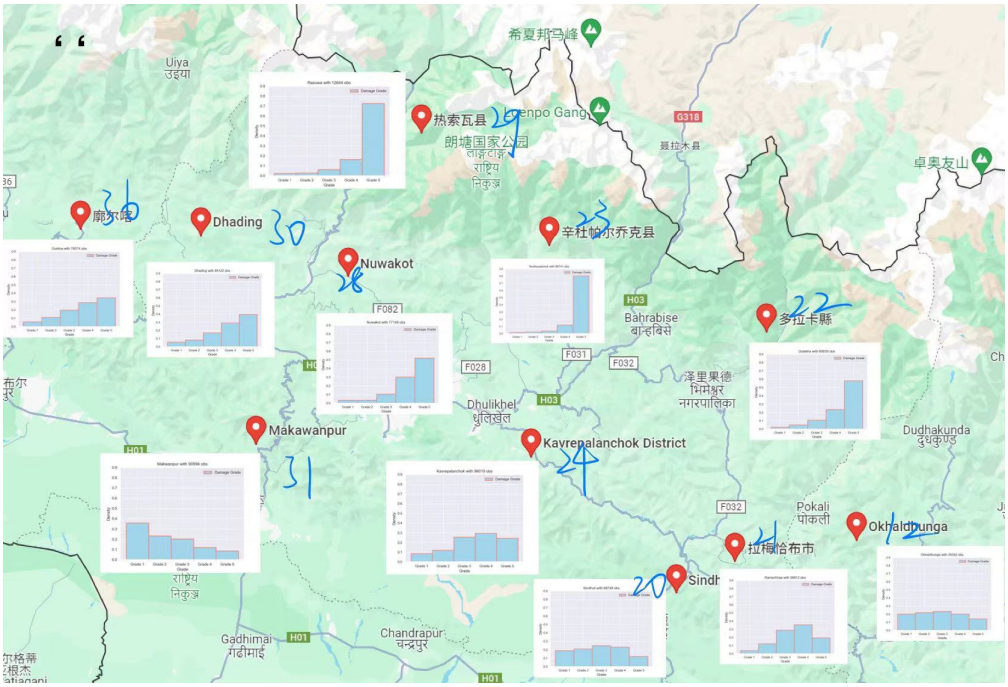


图 3: 分区县的受灾等级分布在地图上的呈现

2.2 地区聚类

由于受灾等级与地理信息可能是强相关的，我们在这里通过将地区聚类以弥补地理信息与其他非建筑信息的缺失。假设一个地区的不同受灾等级建筑个数为 $(n_1, n_2, n_3, n_4, n_5)$，我们根据计算不同受灾等级建筑的比例：$r_i = \frac{n_i}{\sum_{i=1}^{5} n_i}$，$i = 1, 2, 3, 4, 5$ 生成一个地区的受灾等级分布 $(r_1, r_2, r_3, r_4, r_5)$。根据每个地区的受灾等级分布，并综合考虑地理要素（如各地区地形、海拔、距大地震震中的位置），我们使用 $3 - cluster$ 的层次聚类4将各个地区分类，以弥补地理信息与其他非建筑信息的缺失，并在分类后的三个大区内分别进行建模。下表为聚类后的结果：

地区编号	地区名称	所属聚类区
22	Dolakha	2
23	Sindhupalchok	2
28	Nuwakot	2
29	Rasuwa	2
21	Ramechhap	1
24	Kavrepalanchok	1
30	Dhading	1
36	Gorkha	1
12	Okhaldhunga	0
20	Sindhuli	0
31	Makwanpur	0

在核对地图后，我们进一步发现，聚类区 0 的地区多分布于海拔较高的山地，聚类区 1 的地区多分布于丘陵地带，聚类区 2 的地区地势总体较为平缓，这也一定程度上印证了我们关于受灾等级分布与地理因素有较强相关性的猜想。

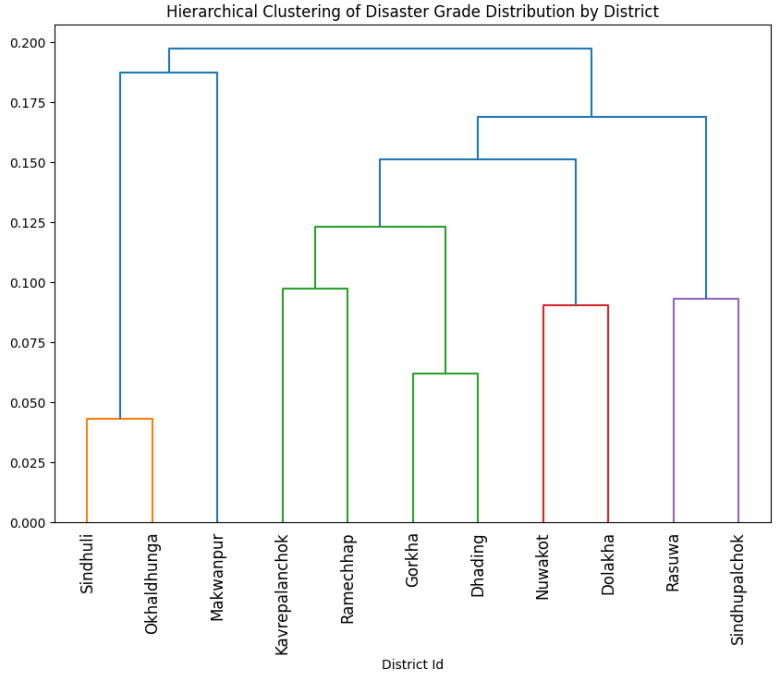


图 4: 区县关于受灾等级的层次聚类树状图

2.3 连续变量分析

针对我们将要在模型中可能要使用的连续变量，分析其与受灾等级的关系。

2.3.1 建筑基座面积（平方英尺）(plinth_area_sq_ft)

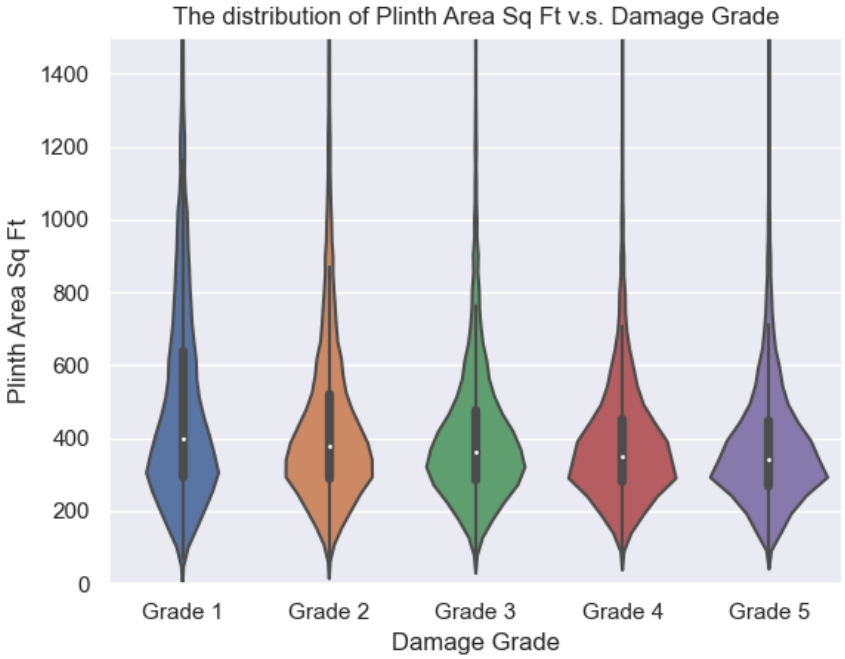


图 5: 建筑基座面积（平方英尺）的受灾等级分布

图中省略了一些 outlier。总体来看，plinth_area_sq_ft 的分布与 damage_grade 的分布相关程度较低。

2.3.2 地震前该建筑高度（平方英尺）(height_ft_pre_eq)

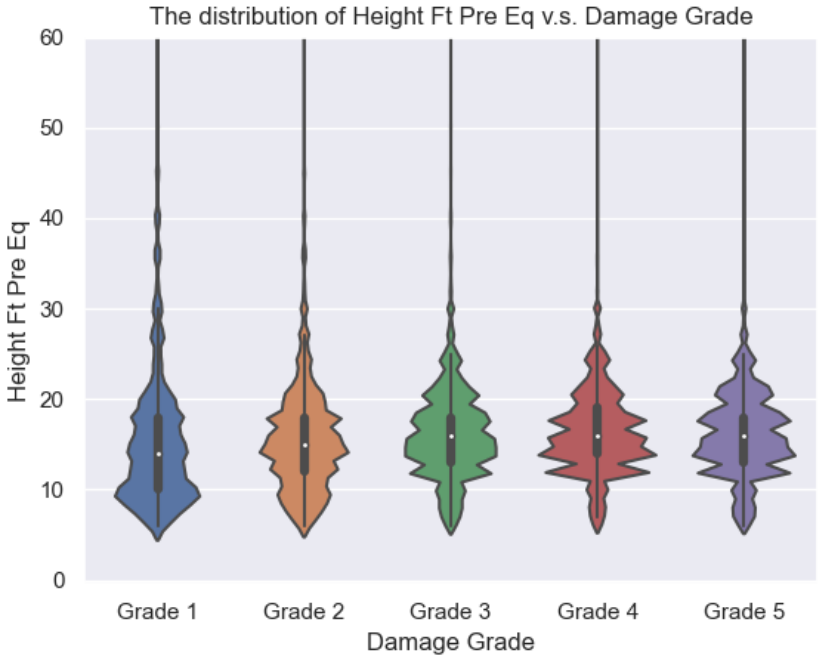


图 6: 地震前该建筑高度（平方英尺）的受灾等级分布

图中省略了一些 outlier。从样本中位数的变化趋势与数据的分布情况来看，height_ft_pre_eq 的分布与 damage_grade 的分布可能存在一定的正相关。这也与“建筑物越高，在大地震中可能受灾情况越严重”这一猜测相符合。

2.3.3 建筑年龄 (age_building)

图 7: 建筑年龄的受灾等级分布

图中省略了一些 outlier。从样本中位数的变化趋势与数据的分布情况来看，age_building 的分布与 damage_grade 的分布存在一定的正相关。这也与建筑物越老，抗灾能力越差这一常识相符合。

2.4 分类变量分析

2.4.1 地震前该建筑楼层数 (count_floors_pre_eq)

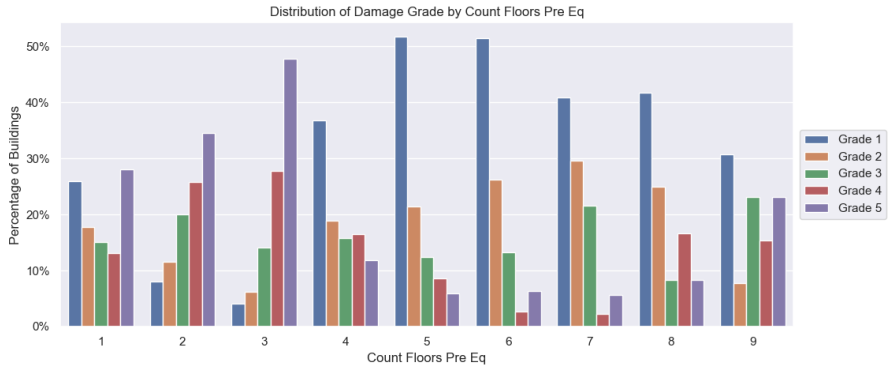


图 8: 地震前该建筑楼层数的受灾等级分布

从结果上来看，1-3 层的建筑受灾情况反而比 4-9 层的建筑受灾情况要严重。这是出人意料的。在查看数据后发现，4-9 层的建筑只占样本的 1% 左右。这个柱状图给我们带来了一些启发：建筑的受灾情况很可能与其建筑材质有较强关系。

2.4.2 陆地表面条件 (land_surface_condition)

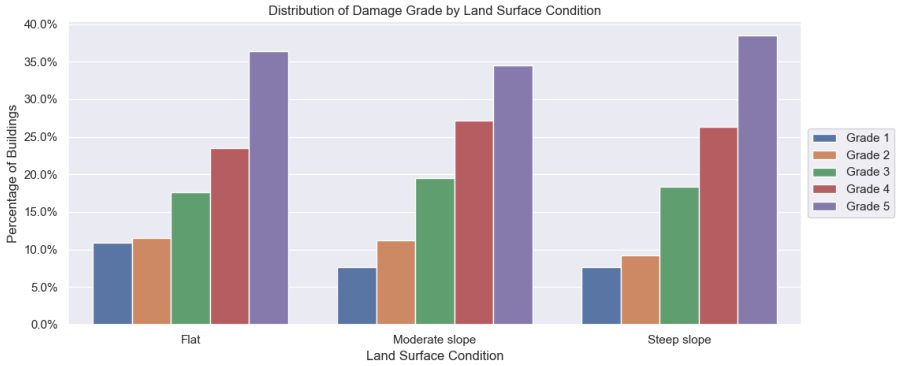


图 9: 陆地表面条件的受灾等级分布

从样本的统计来看,我们认为 land_surface_condition 的分布与 damage_grade 可能没有什么关系。

2.4.3 建筑基础材料 (foundation_type)

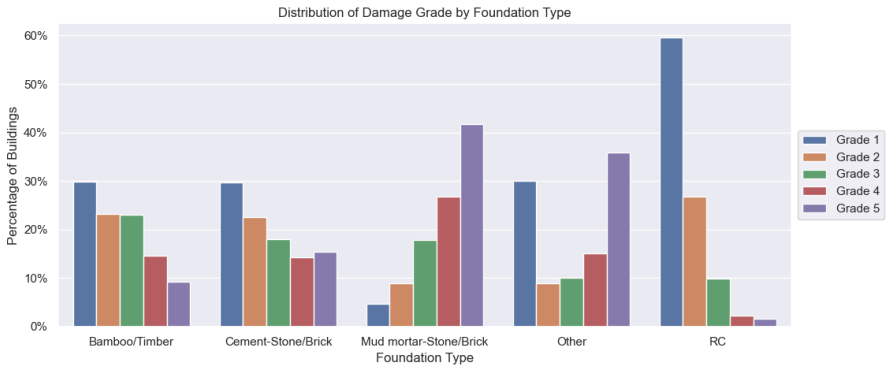


图 10: 建筑基础材料的受灾等级分布

可以发现，建筑基础材料与受灾等级分布有关。特别地，材料为 RC 时，建筑的受灾程度普遍较低。

进一步地，我们画出建筑基础材料与地震前该建筑楼层数的分布，验证了“楼越高，越安全”这一状况与建筑材质确实有较强相关性。

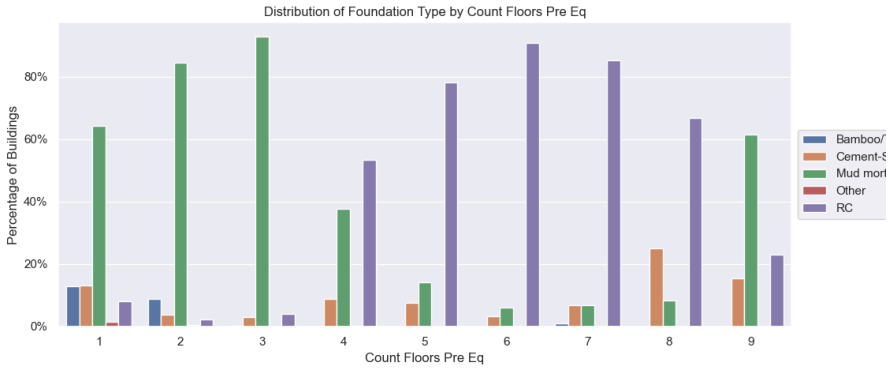


图 11: 建筑基础材料与震前该建筑楼层数的分布

2.4.4 屋顶建筑材料类型 (roof_type)

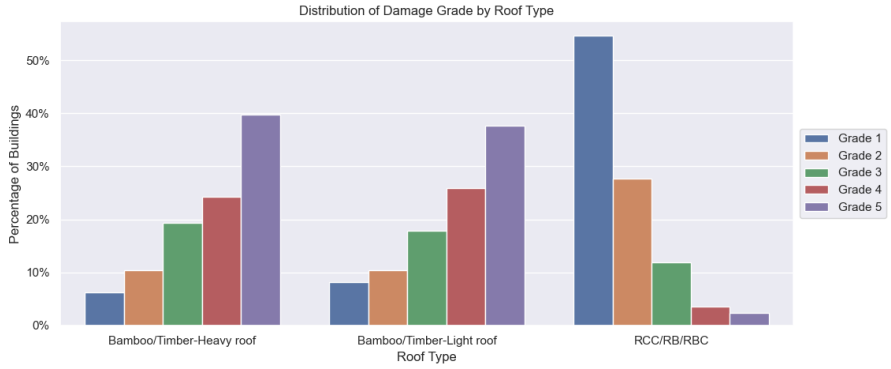


图 12: 屋顶建筑材料类型的受灾等级分布

可以发现，当屋顶材质为 RCC/RB/RBC 时，建筑物的受灾情况最低，但是这类情况仅占样本的 5%。其余两种情况对建筑受灾情况无明显的影响。

2.4.5 地面楼层建筑材料类型 (ground_floor_type)

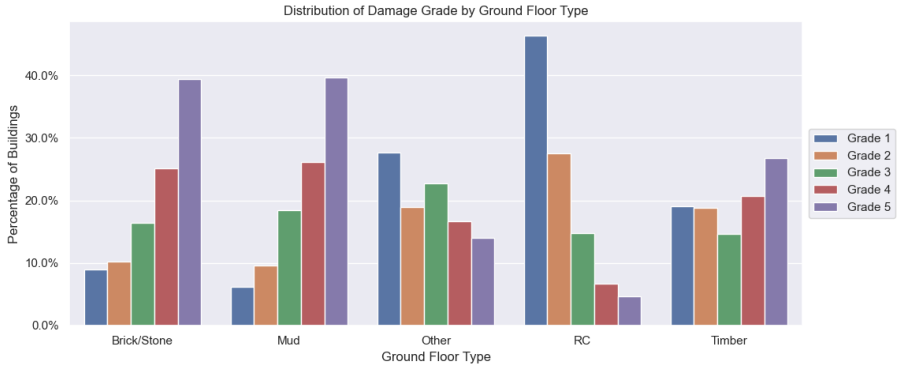


图 13: 地面楼层建筑材料类型的受灾等级分布

与建筑基础材料类似地，地面楼层建筑材料类型为 RC 时，建筑的受灾程度最低，此类样本占比约为 10%。

2.4.6 其他楼层材料类型 (other_floor_type)

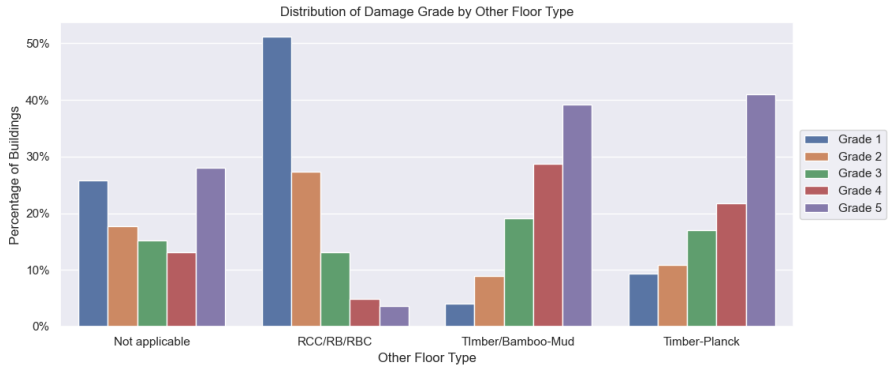


图 14: 其他楼层材料类型的受灾等级分布

与屋顶材料类型类似地，其他楼层材料类型是 RCC/RB/RBC 时，建筑的受灾程度最低，但这类情况只占 5% 不到。数据中，Not Applicable 的出现是因为建筑只有一层，没有其他楼层。

2.4.7 该建筑和其他建筑相连情况 (position)

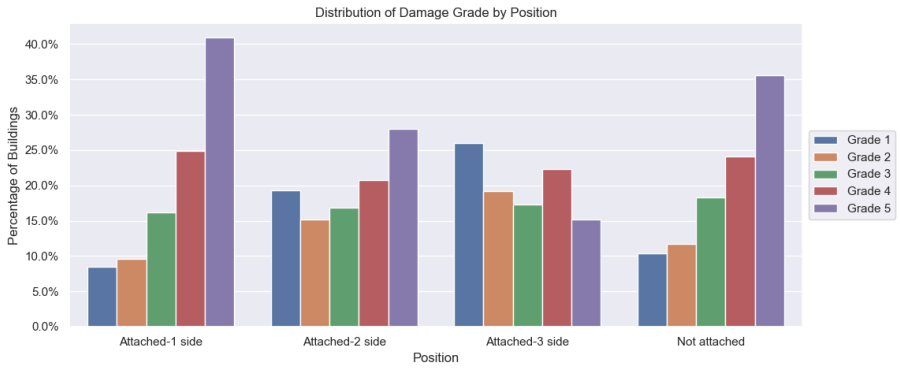


图 15: 该建筑和其他建筑相连情况的受灾等级分布

从其中可以看到，似乎建筑物聚集程度越高，受灾程度越低，这与常识相悖，我们推测这可能与建筑物的其他信息有关。

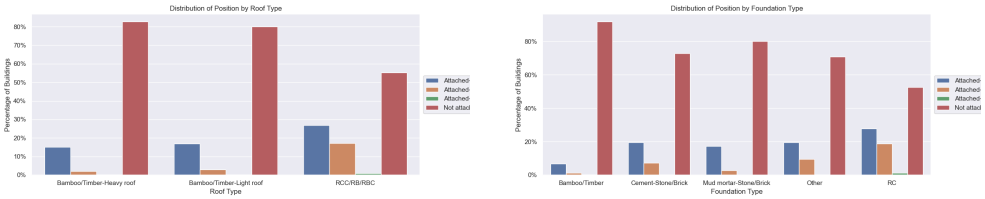


图 16: 该建筑和其他建筑相连情况与部分变量的关系

可以发现，材质较好的建筑聚集程度更高，而建筑材质与受灾程度相关，这为这个现象提供了一种解释。

2.4.8 该建筑规划结构形状 (plan_configuration)

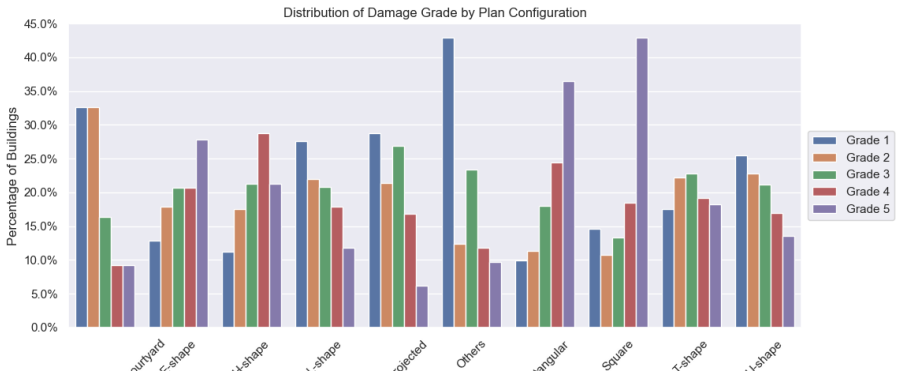


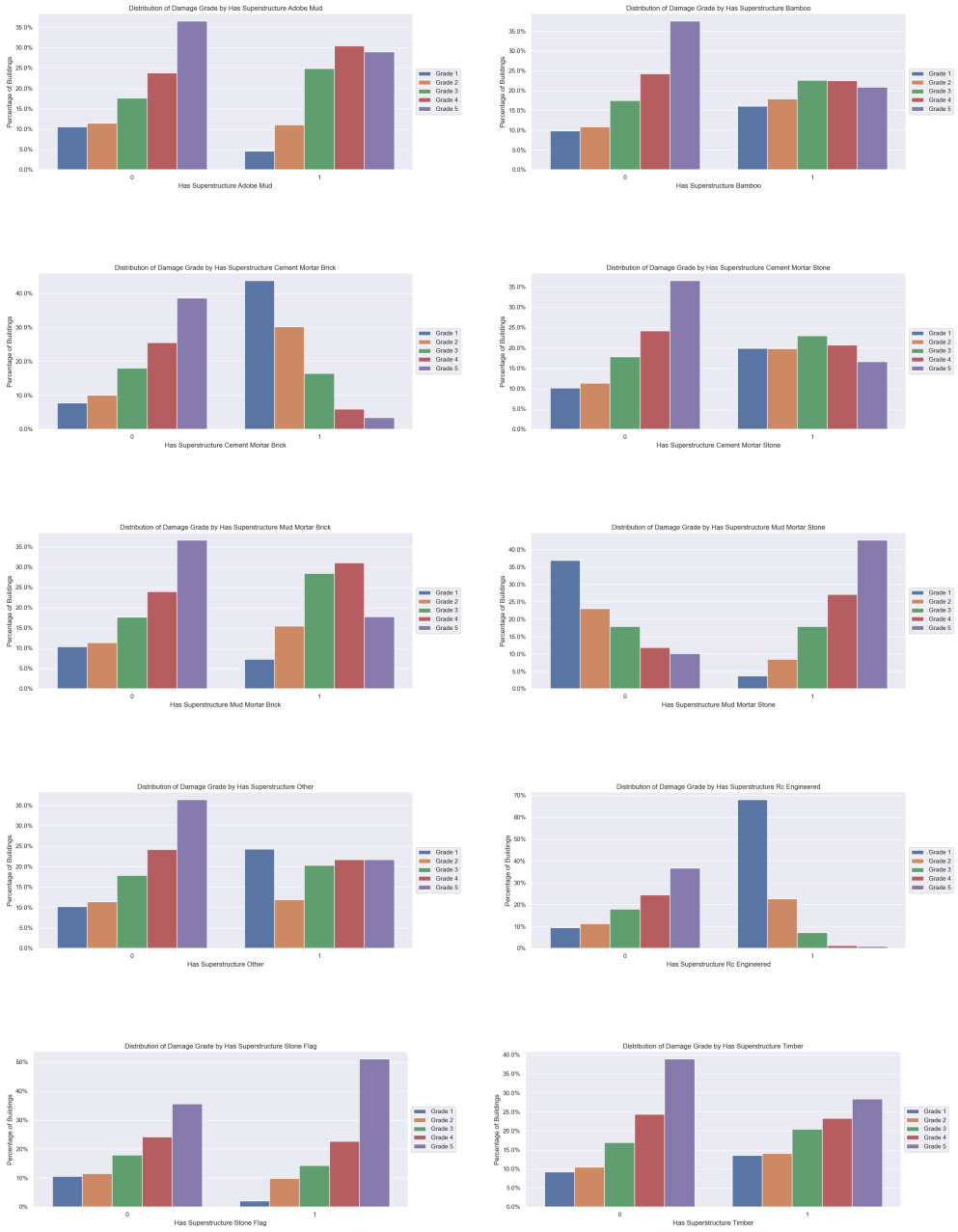
图 17: 该建筑规划结构形状的受灾等级分布

这些规划结构形状中，Rectangular,Square,L-shape 占到了样本的 99%，而其中占比最小的 L-shape 受灾程度最轻，但受制于样本个数，我们很难断定建筑物规划结构形状是否对受灾程度有影响。

2.4.9 该建筑上部结构是否由 X 材料构成 (has_superstructure_X)

这部分的数据较为复杂，但总体来看，一些材料能够减轻建筑物受灾程度，另一些则会加重建筑物受灾程度。

图 18: 该建筑上部结构是否由 X 材料构成



3 数据预处理

3.1 剔除未来信息

由表1信息知，数据集中有少量地震后的信息。为了防止模型构建中出现，我们将以下特征剔除。

特征名称	特征意义
count_floors_post_eq	地震后该建筑楼层数
height_ft_post_eq	地震后该建筑高度（平方英尺）
condition_post_eq	地震后该建筑受损情况
technical_solution_proposed	计划修复该建筑的技术方案

3.2 生成哑变量

由表1信息知,数据集所涉及的数据包含类别变量,我们采用了**哑变量(Dummy Variables)** 的方法进行处理。类别变量是指那些反映种类或分组的变量，例如该数据集中'foundation_type' 特征，它就含有各种非数值变量，例如‘泥浆砂浆石’等。这些类别变量在数据分析中通常是非数值的，不能直接用于统计模型中的数值计算。

为了克服这一限制，我们将每个类别变量转换成一个或多个哑变量。哑变量是指用 0 和 1 表示的二元变量，用以反映某个类别的有无或属否，这样就可以实现在多种统计分类模型上进行数据集处理。

同时，一些有趋势的分类变量将其加权使其连续化。即为了将地面条件‘land_surface_condition’数值化，将‘Flat’记为 0，‘Moderate Slope’记为 1，‘Steep slope’记为 2。

3.3 生成交叉项

特征交叉是一种特征工程常用技术，旨在融合两个及以上特征的信息，生成新的特征，以提高机器学习模型的性能。

我们根据对于建筑物的理解以及相关性质的探究，依据已有经验生成以下特征交叉项。

特征名称	特征意义
age_mul_hight	建筑物年龄与高度乘积
age_mul_floors	建筑物年龄与楼层数乘积
slope_mul_hight	建筑物地形斜率和高度乘积
slope_mul_age	建筑物地形斜率和年龄乘积
slpoe_mul_area	建筑物地形斜率和地基面积乘积
height_mul_cement	建筑物高度和是否使用水泥建造的乘积
height_mul_mortar	建筑物高度和是否使用砂浆石建造的乘积
height_mul_RC	建筑物高度和是否使用钢筋混凝土建造的乘积
height_mul_area	建筑物高度和地基面积的乘积
height_div_floor	建筑物高度和楼层层数的商
height_div_area	建筑物高度和地基面积的商

再进行特征分析知，一些交叉特征项与受灾等级有较强的相关性。例如

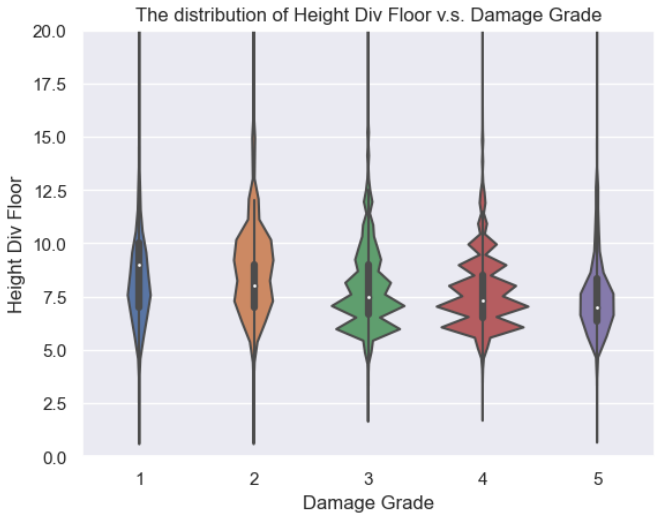


图 19: 高度与楼层的商与受灾等级小提琴图

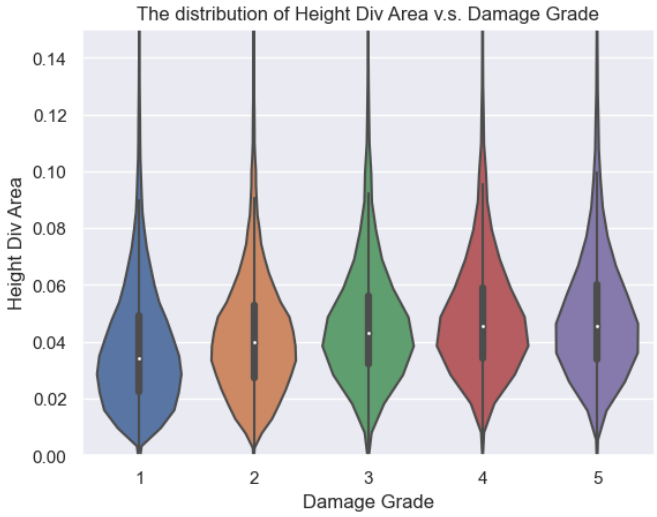


图 20: 高度与地基面积的商与受灾等级小提琴图

图 21: 建筑物年龄与楼层数的积与受灾等级小提琴图

4 建筑物受灾等级预测模型构建

4.1 五分类模型，未使用地区聚类结果

由于不使用地区聚类，受灾等级分布1较为均衡，则不考虑使用数据增强。将处理过的数据以 4:1 的比例分为训练集与测试集，并通过对比四种模型来决定使用哪一种模型作为主要模型。

4.1.1 多分类逻辑回归

逻辑回归（Logistic Regression）是一种用于解决二分类问题的统计学习方法。逻辑回归是一种分类算法，用于估计输入特征与离散输出（类别标签）之间的关系。而逻辑回归二分类算法可以通过增加分类器的方法扩展到多分类算法。

在这种方法中，对于具有多个类别的多分类问题，我们为每个类别训练一个独立的二分类逻辑回归模型。对于每个类别，将其余类别视为负类别，而将该类别视为正类别。在预测时，对于每个模型，选择具有最高预测概率的类别作为最终的分类结果。

我们将五分类逻辑回归模型运用到该数据集上，得到如下结果。

受灾等级	精确度	召回率	F1 分数	数据量
1	0.52	0.07	0.13	15923
2	0.23	0.01	0.02	17537
3	0.15	0.00	0.00	27305
4	0.27	0.00	0.00	36732
5	0.37	0.99	0.54	54922
总体准确率			0.37	152419
宏平均	0.31	0.22	0.14	152419
加权平均	0.30	0.37	0.21	152419

混淆矩阵热力图如下

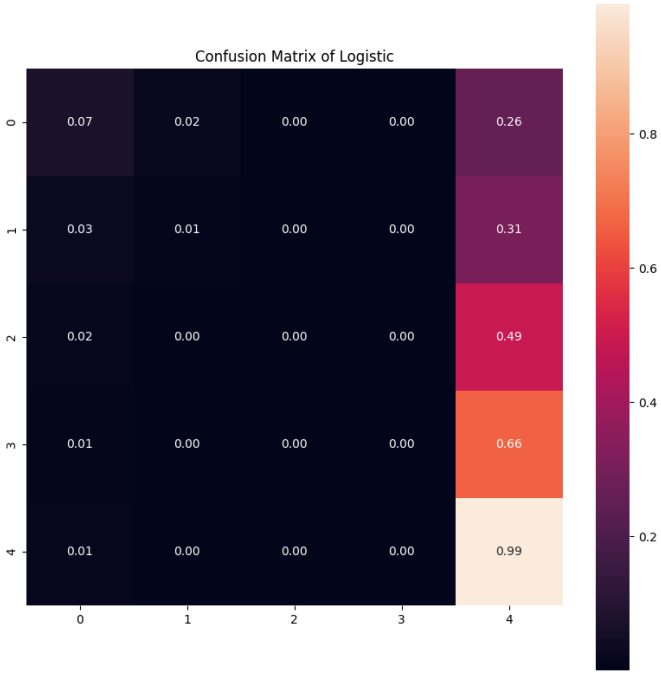


图 22: 多分类逻辑回归混淆矩阵

4.1.2 朴素贝叶斯

朴素贝叶斯（Naive Bayes）是一种基于概率统计的机器学习算法，用于解决分类和文本分类问题。它基于贝叶斯定理，假设特征之间相互独立，然后使用贝叶斯定理来估计输入特征与响应类别变量之间的关系。

我们将朴素贝叶斯模型运用到该数据上，得到如下结果

受灾等级	精确度	召回率	F1 分数	数据量
1	0.51	0.45	0.48	15923
2	0.21	0.11	0.14	17537
3	0.25	0.09	0.13	27305
4	0.32	0.09	0.14	36732
5	0.44	0.88	0.59	54922
总体准确率			0.42	152419
宏平均	0.35	0.32	0.30	152419
加权平均	0.36	0.42	0.34	152419

混淆矩阵热力图如下

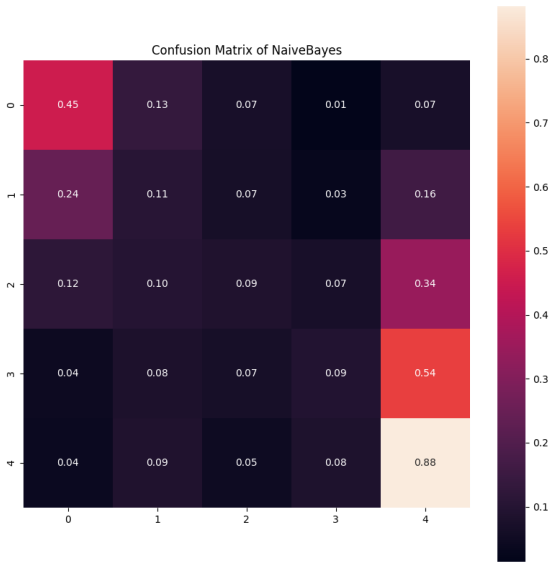


图 23: 朴素贝叶斯混淆矩阵

4.1.3 决策树

决策树算法的算法流程是通过递归分裂数据集，选择信息熵增益最大的最佳特征进行分裂，构建树状结构，直到达到停止条件（如达到最大深度或分裂不再提高纯度），然后剪枝以防止过拟合，最终得到一个可用于分类的树模型。

我们将决策树模型运用到该数据集上，得到

受灾等级	精确度	召回率	F1 分数	数据量
1	0.44	0.46	0.45	15923
2	0.22	0.23	0.22	17537
3	0.25	0.25	0.25	27305
4	0.33	0.33	0.33	36732
5	0.53	0.51	0.52	54922
总体准确率			0.38	152419
宏平均	0.35	0.35	0.35	152419
加权平均	0.38	0.38	0.38	152419

混淆矩阵热力图如下：

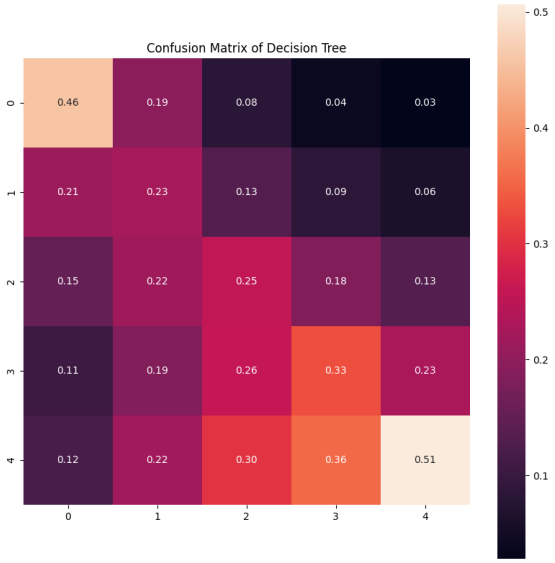


图 24: 决策树混淆矩阵

4.1.4 随机森林

随机森林算法从训练数据中随机采样出多个子数据集分别构建决策树模型，但在构建每个树的过程中，也会随机选择特征的子集进行分裂节点的选择。最后，

通过投票或平均每个树的预测结果，随机森林将得到一个集成的模型，具有更好的鲁棒性和泛化能力，适用于分类和回归任务。

受灾等级	精确度	召回率	F1 分数	数据量
1	0.51	0.56	0.54	15923
2	0.27	0.20	0.23	17537
3	0.29	0.24	0.26	27305
4	0.36	0.34	0.35	36732
5	0.53	0.62	0.57	54922
总体准确率			0.43	152419
宏平均	0.39	0.39	0.39	152419
加权平均	0.41	0.43	0.42	152419

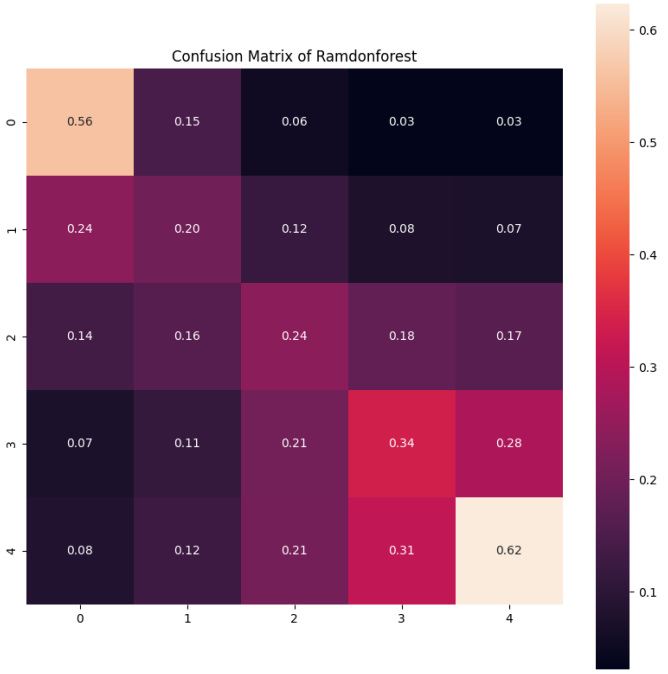


图 25: 随机森林混淆矩阵

重要性前 15 的特征：

表 2: 特征重要性

特征	重要性
height_div_area	0.139680
height_mul_area	0.139019
plinth_area_sq_ft	0.136930
age_mul_height	0.085994
age_mul_floors	0.077948
age_building	0.075696
height_mul_mortar	0.030518
height_div_floor	0.027002
height_ft_pre_eq	0.025584
slope_mul_area	0.021775
has_superstructure_mud_mortar_stone	0.020851
slope_mul_age	0.015850
has_superstructure_timber	0.013514
foundation_type_Mud_mortar-Stone/Brick	0.011082
slope_mul_height	0.009640

4.1.5 总结

通过比较上述模型，经过交叉验证，我们认为随机森林对于该数据集的表现较好。在后续的继续处理上，我们将继续使用随机森林模型作为主要的方法进行预测。

其次，在目前的模型表现上，对于受灾等级等级 1 和等级 5 预测较为准确。而在较中间的部分会有较大程度的混淆。而在决策树和随机森林模型上，发现了较强的三对角特征。这些都说明相邻等级之间并没有分得很开，这为我们之后做三分类、二分类做了指引。

此外，模型的特征庞大，大部分特征重要性较低，并且影响了模型训练时间。所以删除部分特征以简化模型是必要的。

4.2 三分类模型，使用地区聚类结果

由2.2得到的聚类结果，我们从三类地区出发，分别进行三分类预测。我们将受灾等级等级 1 归为一类‘1’，等级 2、等级 3 归为一类‘2’，等级 4、等级 5 归为一类‘3’。

值得注意的是，每一区域是根据地震的强弱程度划分的，某些区域的受灾等级分布极度不均匀。此时可以采用数据增强技术来平衡数据集，以提高模型的性能。

数据增强的主要思想是通过生成新的训练样本来增加少数类别的样本数量，从而使数据集更加平衡。这里我们采用 SMOTE 方法

4.2.1 随机森林

我们再次应用随机森林，在 0 号聚类区中得到如下结果

受灾等级	精确度	召回率	F1 分数
1	0.66	0.64	0.65
2	0.53	0.57	0.55
3	0.61	0.58	0.60
总体准确率	0.60		
宏平均	0.60	0.60	0.60

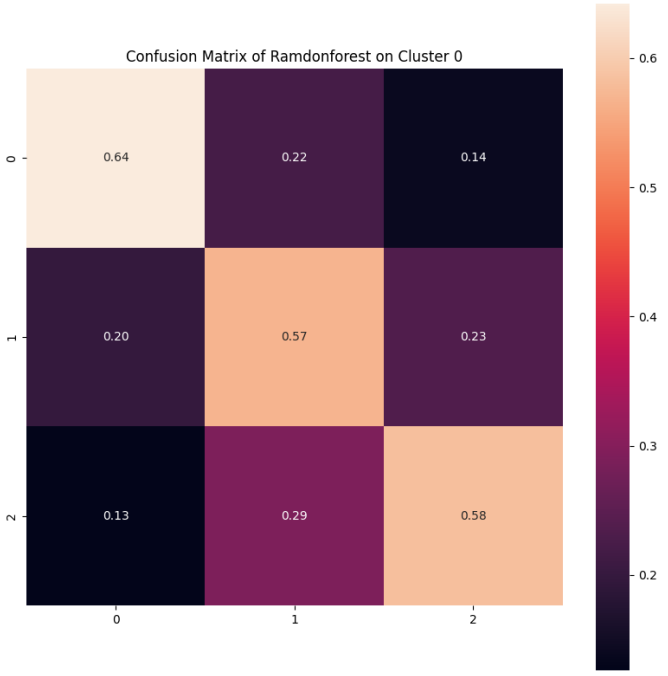


图 26: 0 号区随机森林混淆矩阵

重要性前 15 的特征:

表 3: 0 号区的特征重要性

特征	重要性
height_div_area	0.110549
height_mul_area	0.109680
plinth_area_sq_ft	0.106868
age_mul_height	0.082865
age_mul_floors	0.069926
age_building	0.064415
has_superstructure_mud_mortar_stone	0.045384
height_div_floor	0.042056

Continued on next page

Table 3 continued from previous page

Feature	Importance
height_mul_mortar	0.035860
height_ft_pre_eq	0.034361
slope_mul_area	0.018003
ground_floor_type_RC	0.017552
other_floor_type_Timber/Bamboo-Mud	0.015478
roof_type_Bamboo/Timber-Light roof	0.015394
has_superstructure_timber	0.015348

由精确度可以看到，0 号聚类区使用随机森林的精确度有待提高，总体准确率不够理想。从混淆矩阵来看分类并没有太大误差，区分错误的也集中在对角线附近。

从特征重要性来看，0 号聚类区的受灾等级对建筑的高度、地基面积、建筑年龄有较强的相关性。

在 **1 号聚类区**中得到如下结果

受灾等级	精确度	召回率	F1 分数	数据量
1	0.88	0.82	0.85	39114
2	0.61	0.57	0.59	39114
3	0.69	0.78	0.73	39114
总体准确率			0.72	117342
宏平均	0.73	0.72	0.72	117342
加权平均	0.73	0.72	0.72	117342

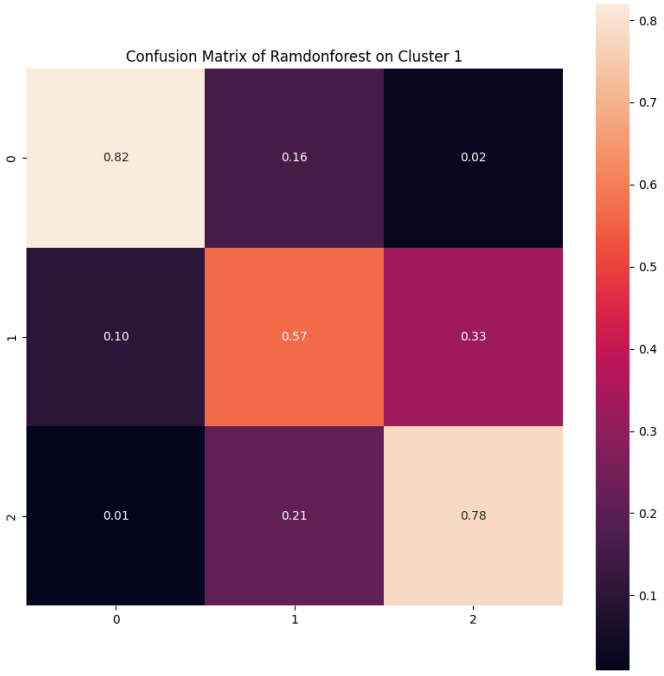


图 27: 1 号区随机森林混淆矩阵

表 4: 1 号区的特征重要性

特征	重要性
height_mul_area	0.086033
plinth_area_sq_ft	0.081784
height_div_area	0.080458
age_mul_height	0.065007
has_superstructure_mud_mortar_stone	0.064966
age_mul_floors	0.059074
height_mul_mortar	0.057560
age_building	0.052779
height_div_floor	0.038593

Continued on next page

Table 4 continued from previous page

特征	重要性
roof_type_RCC/RB/RBC	0.038228
ground_floor_type_RC	0.035167
foundation_type_RC	0.027013
height_ft_pre_eq	0.024670
height_mul_RC	0.022867
other_floor_type_RCC/RB/RBC	0.020821

从精确度可以看到，1 号聚类区对于受灾等级为 1 和受灾等级为 4,5 的预测较准确，而对受灾等级为 2,3 的却预测不准。从混淆矩阵来看，模型预测不容易有较大混淆，模型在测试集上表现不错。

从特征重要性看，相较于 0 号聚类区，1 号聚类区有了更多建筑材料相关的特征，也正说明了 1 号聚类区内，需要一定强度的建筑材料才能预防地震侵害。

在 **2 号聚类区**得到如下结果

受灾等级	精确度	召回率	F1 分数	数据量
1	0.86	0.73	0.79	41372
2	0.71	0.75	0.73	41372
3	0.87	0.95	0.91	41372
总体准确率			0.81	124116
宏平均	0.81	0.81	0.81	124116
加权平均	0.81	0.81	0.81	124116

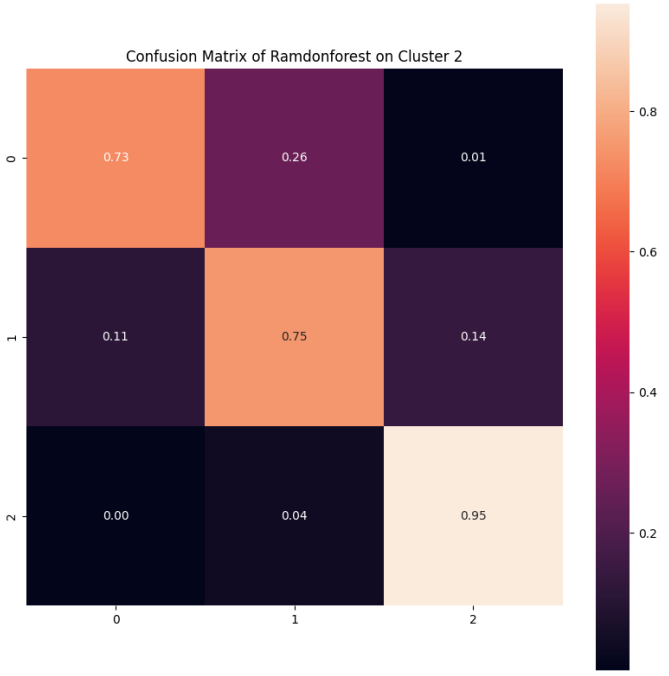


图 28: 2 号区随机森林混淆矩阵

表 5: 2 号区的特征重要性

特征	重要性
has_superstructure_mud_mortar_stone	0.075442
height_mul_mortar	0.068238
height_mul_area	0.066726
plinth_area_sq_ft	0.056419
height_div_area	0.055599
height_div_floor	0.052499
age_mul_height	0.050489
age_mul_floors	0.047364
age_building	0.039129

Continued on next page

Table 5 continued from previous page

特征	重要性
ground_floor_type_RC	0.032063
roof_type_Bamboo/Timber-Heavy roof	0.031208
roof_type_Bamboo/Timber-Light roof	0.029147
roof_type_RCC/RB/RBC	0.028989
other_floor_type_TImber/Bamboo-Mud	0.028633
position_Attached-1 side	0.028531

从精确度和混淆矩阵来看，2 号聚类区的预测能力最强，在受灾等级 4,5 上表现最好。这其实也对应了 2 号聚类区较为震中的性质。

2 号聚类区在特征上，除了和其他都相似的高度、地基面积、建筑年龄，还有就是建筑材料的影响比重增大了，地基材质和屋顶所用的材质也对此处的受灾等级预测有较大的影响。

综上，0 号、1 号、2 号聚类区，建筑的高度、楼层数、地基面积、年龄都对建筑物受灾等级影响很大。而建筑材料的影响从 0 号到 2 号逐渐增大。

5 总结与反思

1. 本次大作业中，我们在对特征实际意义的分析、数据挖掘、数据处理后，构建了以随机森林为核心的分类模型以预测不同建筑物在地震后的受灾等级，并尝试了地区聚类、重采样、将五分类降为三分类等方法以提高预测的准确率。

2. 由于不同地区的数据存在受灾等级分布不均的情况，我们通过重采样来平衡受灾等级的分布，提高模型准确率；在分析了原始数据中的连续变量与分类变量的具体含义与对受灾程度分布可能的影响后，我们对不同的特征进行非线性的组合以产生新特征，提高了模型中的特征重要性与模型的预测精；为了避免未来信息提高模型预测精度，我们删除了所有原数据中含有的未来信息，以使我们的

模型更加真实可靠。

3. 我们本次的分析与建模仍有许多不足之处：地区聚类的方法只是地理信息的一种替代，我们仍需要更多的地理信息来使建模的结果更加准确；连续型特征中，一部分观测值被认为是 outlier，但由于数据来源的权威性我们并未处理这一部分数据，或许需要处理 outlier；我们通过特征工程生成的新特征并不多，可能需要更多的数据挖掘；在使用 RandomForest 时，我们尝试过剪枝等方法，但是结果并不理想，故最后使用的模型是未剪枝的版本，我们在模型调参方面还有很大的进步空间；由于数据集较大，五分类问题若使用交叉验证则较为耗时，故我们未加入交叉验证；此外，使用地区聚类后同一个聚类区中受灾程度分布相似，是否仍需要使用重采样也是值得讨论的。

4. 一些有趣的发现与猜测：在我们的地区聚类中，各个聚类区建筑的高度、地基面积、建筑年龄、楼层数的分布是相似，但是建筑材料从 0 号聚类区到 2 号聚类区变化较大。受限于尼泊尔的城市化水平，建筑材料大多比较原始，针对尼泊尔地震受灾程度的预测模型可能泛化性较差。

5. 本次大作业在文字表述、代码实现方面使用了 ChatGPT 作为辅助。所使用数据集、代码均在同一目录下。