首先从lending club下载了2007-2018所有的借贷数据，此时拥有两百多万个用户的借贷数据，借贷标签共有105个。

从105个标签中选中zip\_code（邮编） addr\_state（所在州）作为地域特征标签，

再选取loan\_amnt（借贷金额） purpose（目的） emp\_length（工作时长）等特征作为其它标签。

由于数据量太大机器性能不足，决定先选取贷款成功与失败各十万条数据。

由于数据中存在缺失值，异常值等，首先进行数据清洗和准备工作，确保数据格式的一致性，并对需要进行编码的分类变量进行编码。

处理之后剩余贷款成功与失败案例各为100000条和99966条。

将两个数据集分别加一列label特征，accepted为1，rejected为0.

合并两个数据集，划分测试集和训练集，训练逻辑回归模型。最后查看系数。

报错，十万条数据太多了，需要53.9G内存。

解决方法：降低数据集的维度

降低数据集的维度是为了减少模型训练所需的内存空间以及提高模型的训练效率。以下是一些常用的方法来降低数据集的维度：

1. 特征选择（Feature Selection）：

- 使用统计方法（如方差分析、卡方检验等）或基于模型的方法（如逻辑回归、随机森林等）选择最具预测能力的特征。

- 使用基于特征重要性的方法，如随机森林的特征重要性，选择最重要的特征。

2. 特征提取（Feature Extraction）：

- 使用降维技术，如主成分分析（PCA）、线性判别分析（LDA）等，将原始特征投影到一个低维度的子空间中。

- 使用自编码器等深度学习模型进行特征提取，学习数据的低维表示。

3. 特征组合（Feature Combination）：

- 将多个相关性较高的特征进行组合，构造出新的特征，以减少数据集的维度。

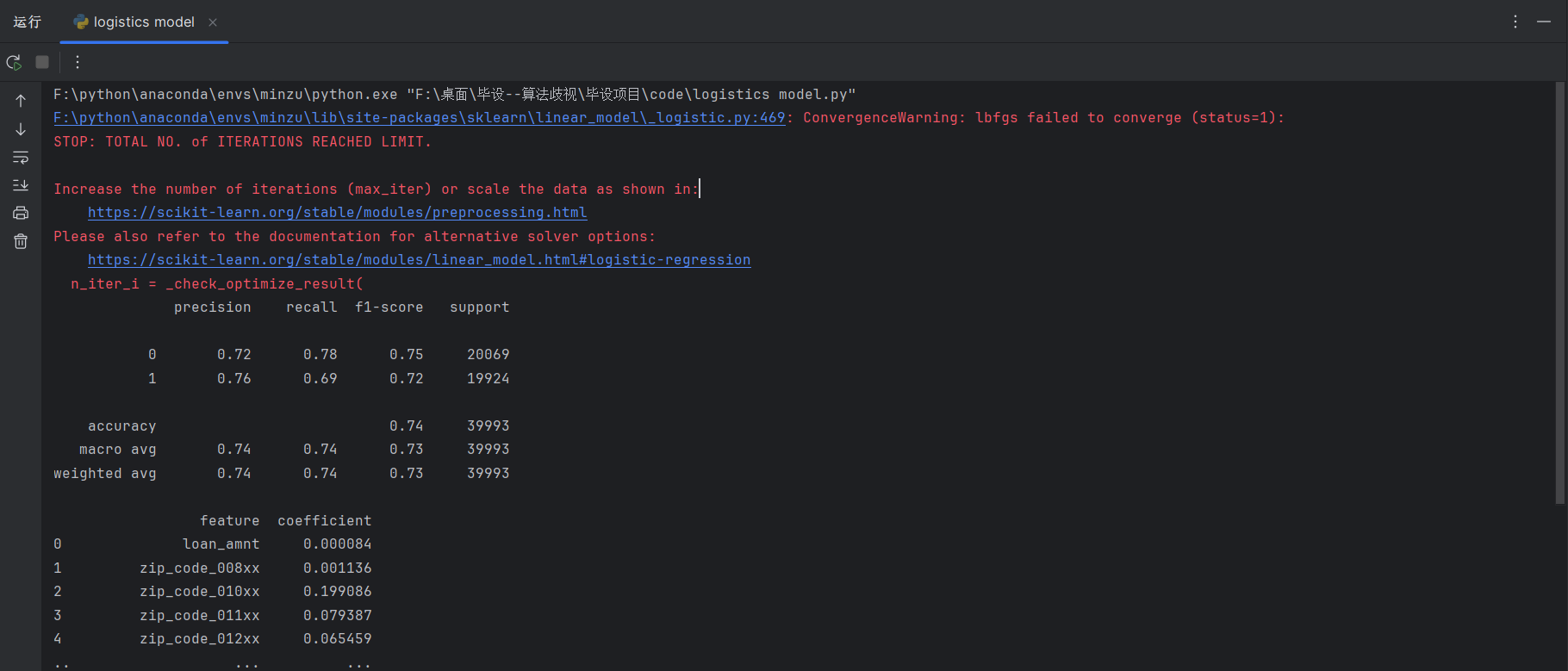
4. 过滤低方差特征（Low Variance Filter）：

- 移除方差较低的特征，因为它们可能不包含足够的信息来区分样本，且在大多数情况下不会对模型的性能产生显著影响。

5. 特征降采样（Feature Downsampling）：

- 对于高维稀疏数据，可以使用降采样技术来减少特征的数量，如随机抽样、聚类抽样等。

决定删除purpose这个标签。

模型未能收敛，重新处理数据集。

将zip\_code (邮编)特征删去，将50个州根据GDP分为5个级别。

