类别不平衡即在机器学习分类任务过程中，不同类别的训练样本样本数有较大差别的情况。对类别不平衡数据使用传统学习机分类，往往不能充分利用少数类样本的信息，结果分类精度不高、查准率、查全率、AUC值等指标较差，并且样本相对较少的一类会有更高的错分代价。Japkowicz与Stephen于2002年提出类别不平衡对传统学习机的负面影响因素可概括为由类别不平衡比率（IR）、重叠区域大小、训练样本的绝对数量、类内子聚集想想的严重程度、样本噪声比率以及样本维度等方面。以朴素贝叶斯分类器为例，类别不平衡样本的IR、重叠区域越大，其决策结果的分类也将越偏离实际结果。本研究旨在预测P2P网贷借款者是否会违约，将借款者分为不会违约和会违约两类。从实际收集的数据来看，违约样本的样本量相对较小，出现类别不平衡现象。因此直接对样本数据使用传统学习机进行分类得到的结果并不令人满意。

不平衡数据的处理方法主要分为两大类，这两类方法分别有不同的侧重点，一类方法侧重数据，因为样本是不平衡的，所以可以通过某种策略对样本抽样，以减小样本的不均衡性；另一类方法侧重算法，通过赋予不同错误分类情况不同的代价，对算法进行优化，使得算法能够在不平衡数据下也能有较为满意的结果。

其中从数据角度处理不平衡样本的方法又可以细分为以下几种：（1）随机下采样，从多数类中随机选取一定量的数据与少数类组合在一起，构成新的样本。这种方法操作简便，但缺点也尤其明显，在随机抽取的过程中，容易把样本中的有用的信息遗失，尤其是初始样本总量不太大的情况下，样本信息丢失更为严重。（2）随机过采样，从少数类中随机选取一定类的样本进行复制与多数类组合在一起，构成新的样本。同样，这种方法操作简便，而且新生成的样本数量也较多，但是其也有不小的缺点，首先其完全复制少数类样本，容易造成少数类数据分布的偏误，使得少数类数据的分布逼近初始少数类样本点的离散分布；其次在复制过程中，其难免也会复制其中的噪声，使得噪声信息被放大。（3）SMOTE算法生成样本，这是一种基于随机过采样方法的改进方案，其基本思想是对少数类样本进行分析并根据少数类样本进行人工合成添加到数据集中，这种方法能够有效减轻随机过采样所带来的过拟合问题。

从算法角度解决不平衡样本的方法主要为代价敏感学习，代价敏感学习可以从构造代价敏感学习模型、基于分类结果的后处理和改变原始样本的初始分布等方面减小不平衡样本的影响，提高模型的效果。