**连续特征离散化（分箱）**



当把数据进行分箱处理后，**数据会变得更加稳定，之前取值范围不定的数据经过分箱后，变成了取值固定的数据；模型的泛化能力增强，过拟合可能性也随之降低。**

除了上面这两个比较大的好处，分箱另一个优势是会加快模型的训练速度，对数据进行离散化后，模型的复杂性降低了，计算速度会比之前快，并能够加快模型的训练速度。所以对使用数据集规模大、对模型运行效率或响应速度有要求的以及需要部署上线的模型比较合适。

此外，分箱的实际操作中，一般会将空值、缺失值单独处理，将他们在分箱操作时作为一个特殊类别存在，这样既能很好的处理空值，又能保证其存在被模型训练考虑在内。

最后，特别说一下，分箱的过程在很多分类问题上都会进行，但应用最广泛的是逻辑回归算法。这是因为在逻辑回归模型中，这主要是因为逻辑回归属于广义线性模型，它与树模型的建模原理本质是不同的。因此把一个特征离散为N个哑变量（Dummy Variable）之后，每一个哑变量就会产生一个单独的权重，相当于为模型引入非线性，进而提升模型表达能力，加大模型拟合能力。

**分箱的好处有下面5点：**

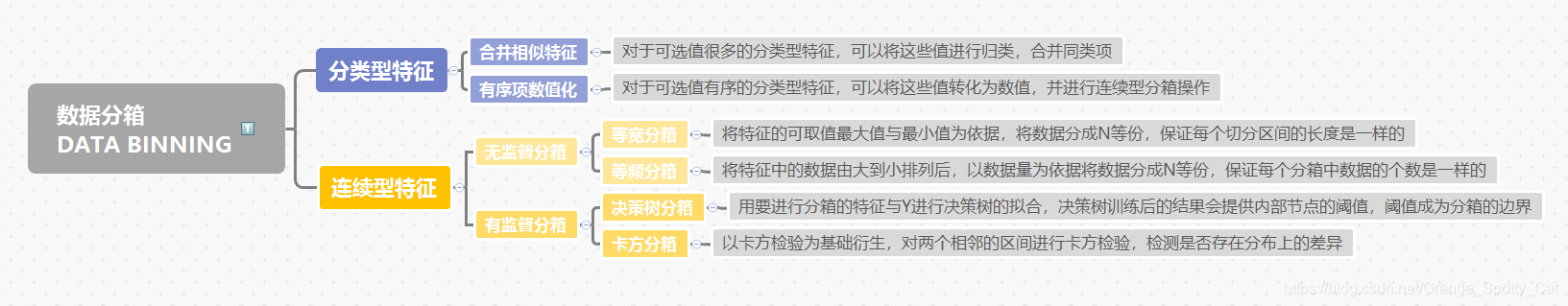
**1、提高模型的稳定性与鲁棒性**

**2、防止过拟合问题**

**3、加快模型训练速度**

**4、很好的处理空值与缺失值**

**5、增强逻辑回归的拟合力**



**连续型特征的分箱**

**连续型特征的分箱分为无监督分箱与有监督分箱两类**，分别各有2种应用的比较广的方法：

* 无监督分箱：不需要提供Y，仅凭借特征就能实现分箱
  + 等宽分箱，实现方法：**pd.cut()**
  + 等频分箱， 实现方法：**pd.qcut()**

**等频和等宽都不能很好的反应“尖头”的数据，除非人工手动干涉，聚类本身对于这类问题的表现也并不稳定，经常可能出现的情况就是“尖头”数据有一部分分到平缓分布的数据里去，导致最终的分箱结果没有代表性。**

sklearn的0.21版中的KBinsDiscretizer 目前支持等频、等宽和聚类分箱。可参考https://scikit-learn.org.cn/view/324.html

* 有监督分箱：需要结合Y的值，通过算法实现分箱
  + 决策树分箱
  + 卡方分箱

可以直接通过Python中的sklearn拟合决策树，并进行分箱结果提取。但是在Github中，有一个很好用的分箱包scorecardpy，可以直接实现决策树分箱

卡方分箱原理

