问题2：

淘宝上有几亿的买家和千万级的商家，我们需要在每个月初给头部的商家（已知商家的名单）推荐一定数量的用户，便于商家有针对性的进行营销活动。

要求：

a.     请给出具体的方案和实现方法，包括所用的模型、数据、训练和预测的方法。

b.     已有的数据是过去半年所有买家的行为数据，即过去半年的每一天里，单个买家在每个商品上的浏览、收藏、推荐，和购买数量。

1. 对题目的思考

此题是根据用户的历史特征来向商家推荐合适的用户。思考后有两种思路：

1. 采用基于内容的推荐算法，即根据用户的历史信息归纳出用户对商家的“喜好特征”，然后使用规则对这些特征进行组合计算得到用户对商家的总体喜好程度，排序后进行推荐。
2. 由于已知的历史信息有半年之久，所以可以使用这些历史数据划分训练集和测试集，然后使用监督学习的方式进行二分类建模。例如可以使用3个月的数据作为训练集（从前2个月的历史数据中抽取特征X，第3个月抽取用户在指定商家的购买信息作为目标Y），训练好模型后然后再抽取最近2个月的数据作为测试集进行预测排序就可以得到推荐列表了。

第1种思路主要是基于人工规则，由于有很强的先验知识，所以预测结果可能可以满足预期，但是由于人工规则得到的模型比较”生硬”，没法灵活地推荐，而且需要较多的人工，于长期地发展是不利的。第2种建模的思路通过已知的信息使用模型计算较优解，只要特征和模型选取得好，能够得到较好的预期。如下介绍第2种建模思路。

1. 数据预处理
2. 去除数据中可能存在的一些噪声数据。例如某些用户的访问量异常（考虑爬虫因素），或者某些用户从来不购买商品等。
3. 填补缺失数据。源数据中在某些时间段用户的数据可能为空，如果缺失量较大可以删除掉，否则需要考虑填补数据，填补的方法主要是均值平滑、插值滤波等。

三、特征提取

特征提取是非常重要的一个工作，主要包括以下几种特征：

1. 商家特征，可以根据已知商家的名单使用爬虫抽取。商家特征主要包括商家类别特征（对于头部商家，一般来说应该都对应于某种品牌，所以可以提取出销售品牌特征，以及该品牌所属类别特征等），商家销售信息特征，商家热门指数特征等。
2. 用户特征。包括用户总购买，总点击，浏览、收藏等特征。
3. 用户在对应商家的浏览、购买、收藏等特征。
4. 交叉特征。例如用户对品牌的偏好、对商家的偏好等，可以通过已知的特征计算得到。

对特征的一些特殊处理：

1. 考虑到提供的数据包括最近半年每一天的数据，数据量是比较大的，所以抽取特征的时候可以考虑在时间尺度上进行处理，一方面可以通过分析数据，设定抽取特征的时间长度，另一方面可以在时间维度上对数据分片，例如间隔较久的数据用处较小，分片间隔可以放大一些。
2. 特征降维，如上可以提取到的特征还是非常多，使用一些非线性分类器的话需要非常多的内容和训练时间，所以可以去掉一些冗余特征。同时这样做也可以降低模型的过拟合程度以及减少特征之间的共线性（特征共线性会给优化带来困难）。

四、模型的选取

原问题已经变成了一个二分类问题，能够使用的模型有很多，从最简单的感知机到深层的人工神经网络都适用。综合考虑后，打算结合各种模型的优点组合得到一个融合模型。

具体实现：LR+GBDT+RF+DNN, 这四种模型单独训练好后进行线性组合即可。

上述的实现方式能够学习到数据中线性和非线性的特点，在非线性这一方面更是能结合当下流行的baggging和boosting思想以及深层网络，预期能够有较好的结果。

五、参数优化

参数优化在机器学习里面一直是比较“玄学”的东西，但是也有一些先验的调参知识可以借鉴来优化结果。

1. 时间充裕的话，可以针对参数进行stacking，即设定一个参数集合，然后训练不同参数的模型融合得到一个结果。
2. 时间不够的话，可以根据参数的重要程度进行“阶梯式调参“。

六、一些其他的优化思路

仔细分析业务场景，往往能够得到一些先验知识来改善结果，毕竟模型总是有欠缺的，因为特征是人工构建的，一定有缺陷。例如很容易想到的：

1. 有些商家回头客很高，一些商家回头客几乎没有。
2. 已经购买过的用户很可能无需再推荐