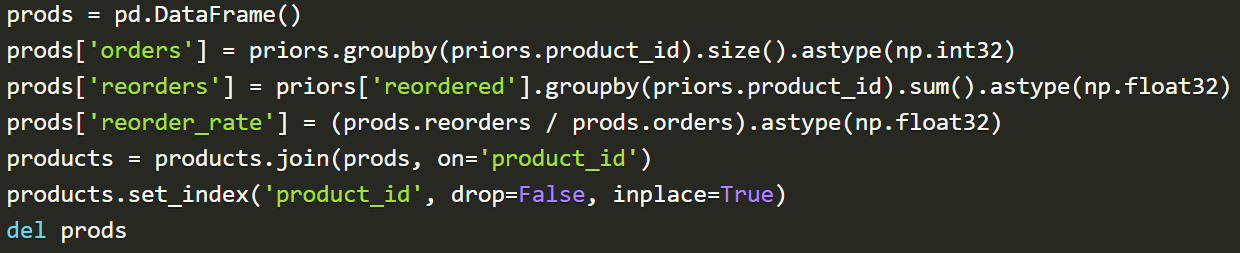
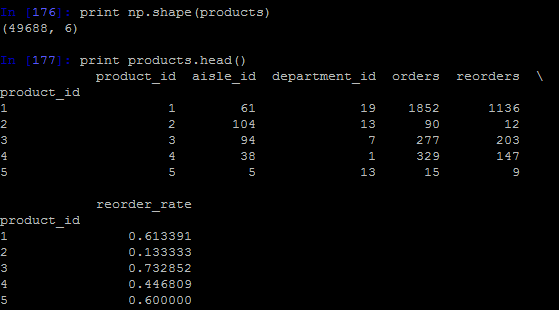
# Data\_Explaination\_and\_extraction

1. From <https://www.kaggle.com/frednavruzov/dumb-and-the-dumber-baselines-plb-0-3276826>

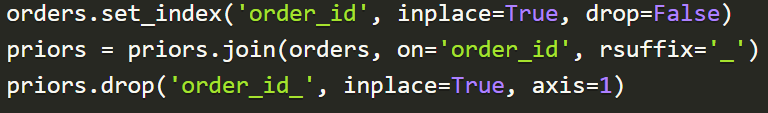
# Mode\_1\_Using\_LightGBM

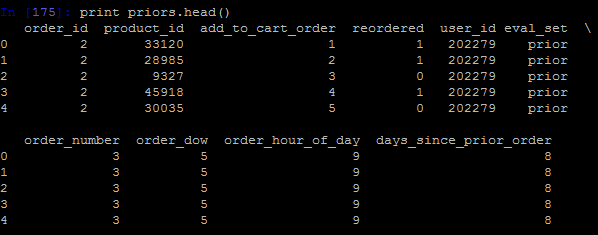
1. From https://www.kaggle.com/paulantoine/light-gbm-benchmark-0-3692
2. Introduction about LightGBM <http://www.jianshu.com/p/48e82dbb142b>
3. 算法思路：
4. Train-order\_products\_train.csv; orders-orders.csv; products-products.csv
5. 依product\_id 在products中加入产品被买的次数，被再购买的次数及被再购买率。并将index设为product\_id



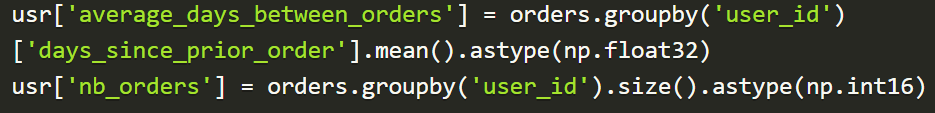


1. 将orders的index设置为order\_id，并将其并入priors

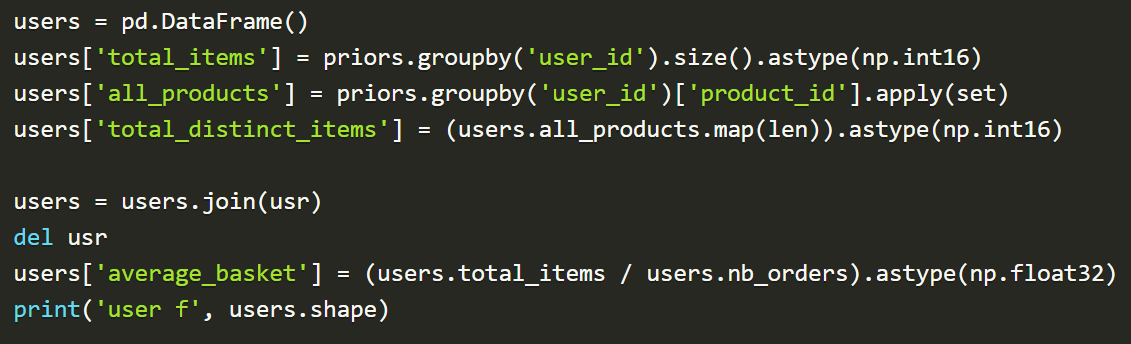


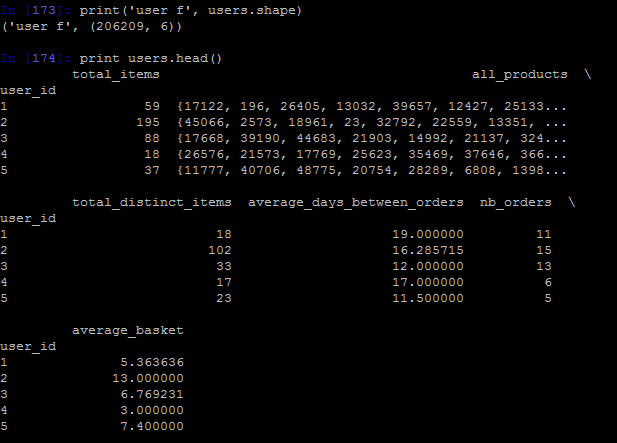


1. 构建DataFrame usr. 在orders中计算每个顾客的下单间隔、下单总数



1. 构建DataFrame users。在priors中计算每个客户购买的商品总数，并将购买的商品表示成数组的形式，并计算购买的不同的商品的总数。将usr合并到users，并计算顾客每单平均购买多少商品



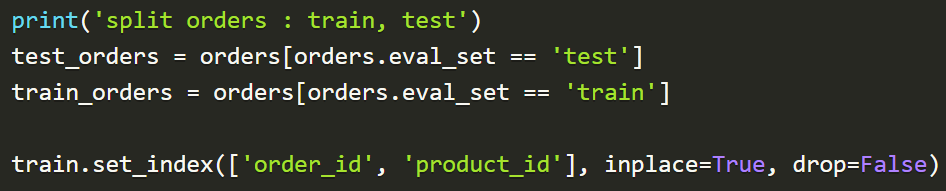


1. 构建userXproduct feature 暂时不用先

在priors中增加列user\_product. Len(products)=49688,可以保证user\_product是唯一的。



1. 分离训练集合数据集

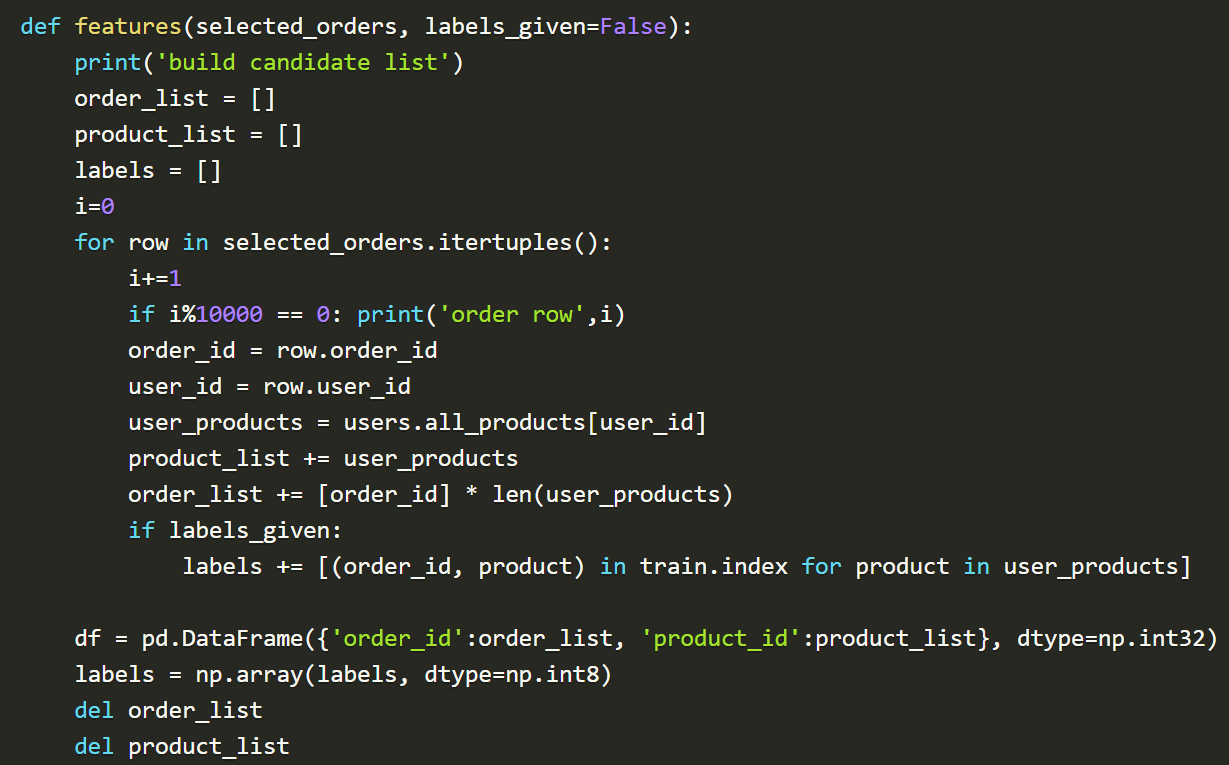


1. 提取特征

从train\_orders中构造送入GBM的数据。Order\_list=[], product\_list=[], labels=[].

从users中提取user.all\_products（购买的所有商品列表），将order\_id扩展成相同长度，就得到了两个一一对应的order和product列表。

如果给定labels\_given=True, 注意users中使用的是priors数据，所以会有富余，在train\_orders中可能没有，此时标定为负样本，否则标定为正样本。至此训练样本构建基本完成。



1. 构建测试样本及预测



预测思想：预测一条待测样本的概率，如果大于阈值，则认为是正样本。训练的样本特征不涉及order\_id 和product\_id。

# Mode\_2\_Using\_Word2Vec

1. From <https://www.kaggle.com/omarito/word2vec-for-products-analysis-0-01-lb>