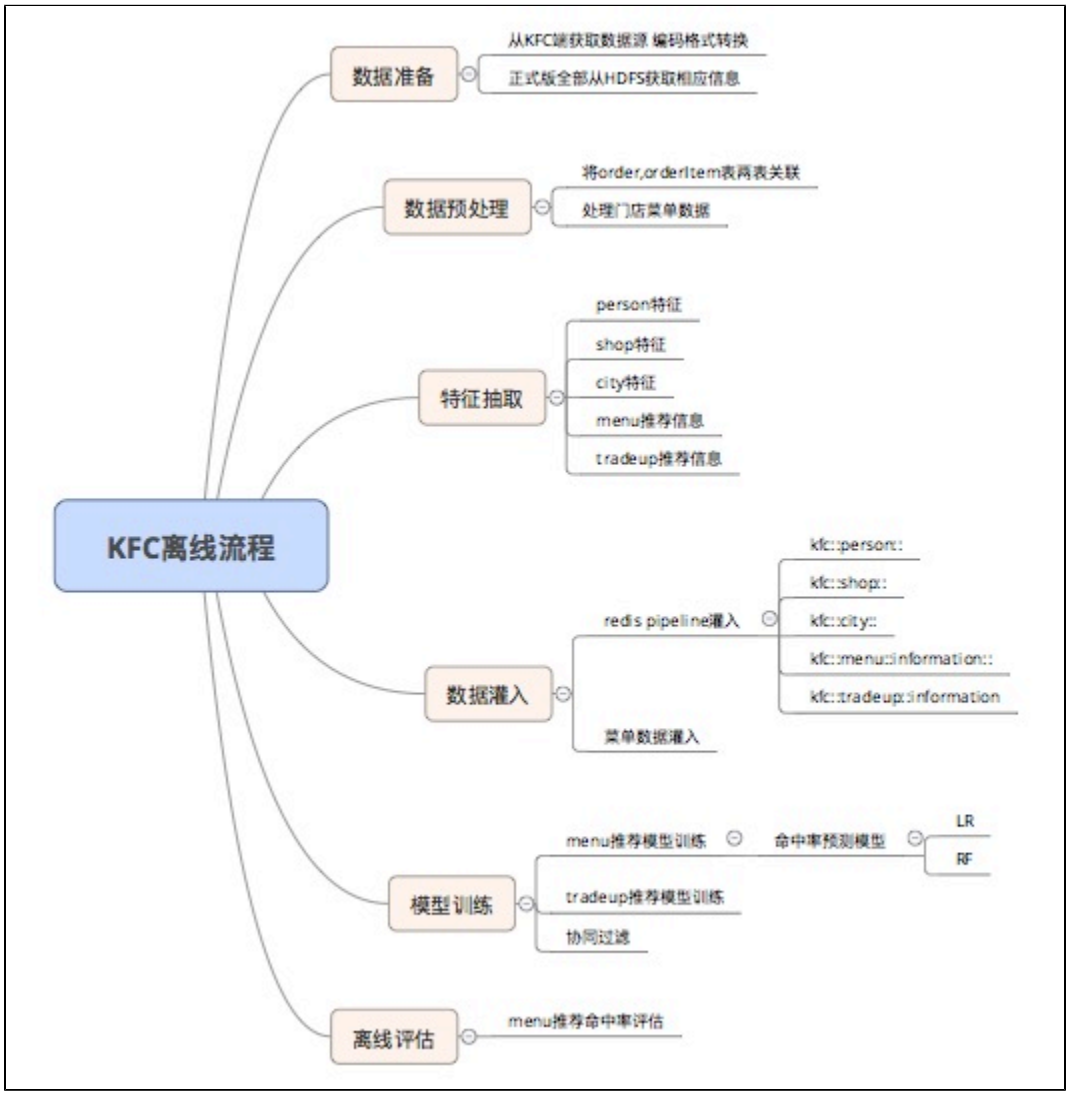
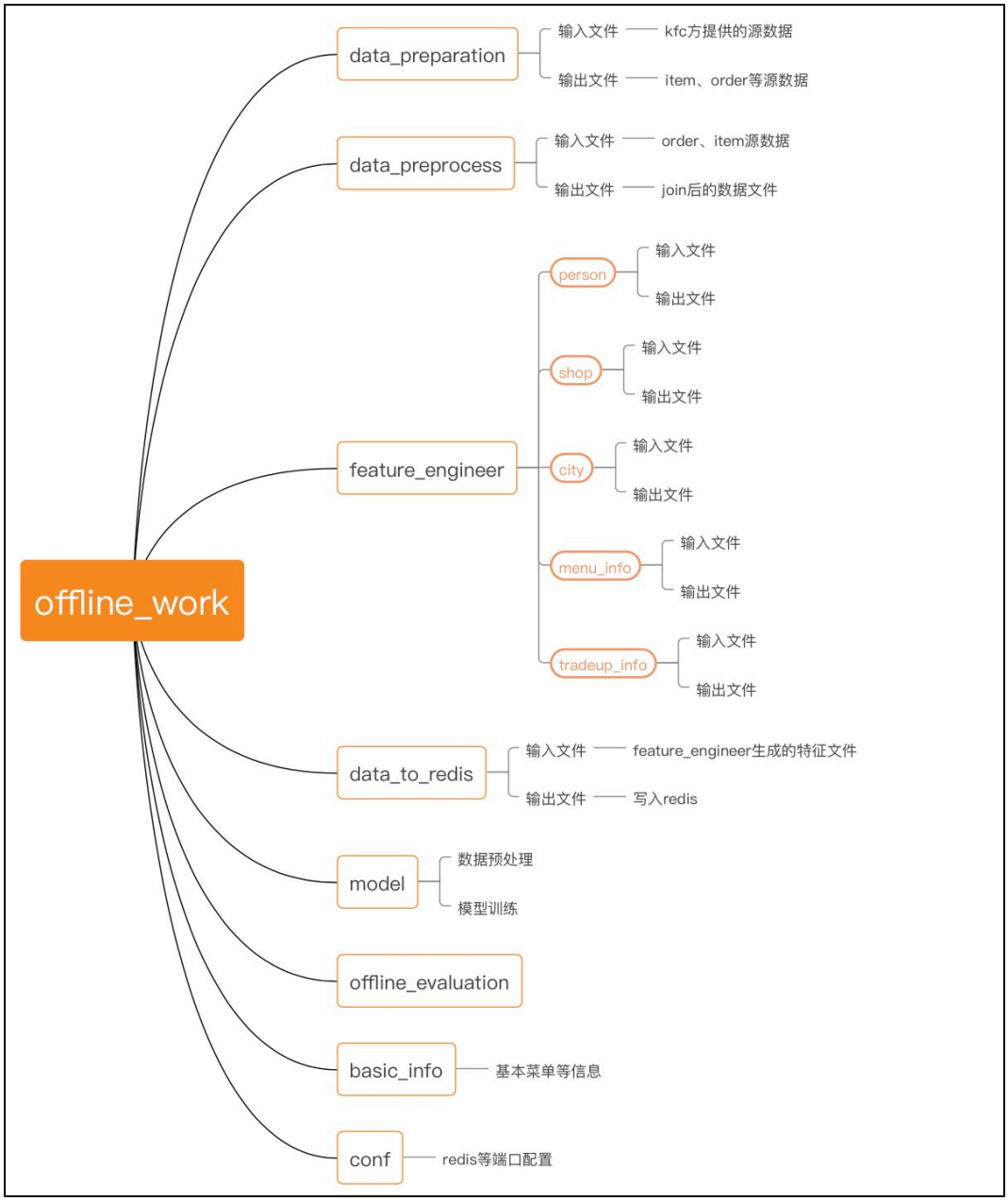
离线处理流程说明

主要说明离线脚本流程及相应子模块输入输出形式

1. 离线总体流程



2. 离线目录说明



a. 子模块说明

(1) 数据准备模块

功能：将kfc提供的源数据，导入至本地hadoop，并转换相应格式至可读入状态
输入文件：kfc方源数据
输出文件：/data/offline_data/origin_data/order_info/日期；（订单表）
/data/offline_data/origin_data/extra_info/日期；（补充信息表）
/data/offline_data/origin_data/store/store_code.txt； （门店信息表）
/data/offline_data/origin_data/menu_info/日期；（菜单推荐日志）
/data/offline_data/origin_data/tradeup_info/日期；（tradeup推荐日志）
/data/offline_data/origin_data/tradeup_list；（tradeup物品列表）

(2) 数据预处理模块

功能：将订单表与补充信息表连接，并添加SFTAG（根据规则等清洗后的唯一表示符）
输入文件：/data/offline_data/origin_data/order_info/日期；
/data/offline_data/origin_data/extra_info/日期；

输出文件: /data/offline_data/data_preprocess/join_data/日期;

编号	字段	备注
0	transation_guid	订单id
1	user_code	用户id
2	transaction_start_time	下单时间
3	transaction_amount	订单总金额
4	store_code	餐厅编码
5	channel	订单渠道 superapp/other
6	unit_sold	商品个数
7	product_sell_price	商品单价
8	is_combo_flag	单品0 可选套餐头1 套餐子项2 优惠3 固定套餐20
9	nameCN	套餐名
10	subclass	单品名
11	category	Drink / Snack
12	class	Tea / fried wing
13	subcategory	Tea /wing
14	linkid	同样的产品在不同的餐厅linkid不同
15	shunfeng_tag	添加的商品标识

其中is_combo_flag 单品：0，套餐头：1，套餐子项：2，卡包：90，tradeup产品：100， 其他优惠：3，4

数据预处理功能补充：1. 产品升级redis写入（未使用）

2. menu强推redis写入（未使用）
3. single菜单写入
4. item菜单写入
5. inkid-subclass映射表写入
6. tradeup菜品菜单 （暂未更新）

（3）特征处理模块

功能：根据前(1)，(2)模块产出的数据；生成person、shop、city、menu_info、tradeup_info五类特征；
并以子key命名产出的不同子特征文件名，以保证redis自动化读入
输入文件：/data/offline_data/data_preprocess/join_data/日期;
/data/offline_data/origin_data/store/store_code.txt;
输出文件：/data/offline_data/feature_data/下person、shop、city、info、extra等五类特征文件

子特征说明：

(a) person类特征说明：（single: 单品； item:套餐头；subitem_with_price 套餐子项 > 0; subitem_without_price 套餐子项 = 0）

kfc_origin_history_buy_times_single : 用户历史购买次数(单品)
kfc_origin_history_buy_times_item : 用户历史购买次数(套餐)
kfc_origin_history_buy_times_subitem_with_price : 用户历史购买次数(套餐子项>0)
kfc_origin_history_buy_times_subitem_without_price : 用户历史购买次数(套餐子项=0)
kfc_origin_food_price : 用户不同购买金额占比(包括price = 0的子项)
kfc_food_association_subclass : 用户基于subclass维度的关联规则
kfc_food_association_category : 用户基于category维度的关联规则
kfc_hotSale_userid_recentQuarter_single : 最近3个月的用户热销(单品)
kfc_hotSale_userid_recentQuarter_item : 最近3个月的用户热销(套餐)
kfc_hotSale_userid_recentQuarter_subitem_with_price : 最近3个月的用户热销(套餐子项>0)
kfc_hotSale_userid_recentQuarter_subitem_without_price : 最近3个月的用户热销(套餐子项=0)
kfc_hotSale_userid_allTime_single : 用户最近一年的热销(单品)
kfc_hotSale_userid_allTime_item : 用户最近一年的热销(套餐)

kfc_hotSale_userid_allTime_subitem_with_price ：用户最近一年的热销(套餐子项 > 0)

kfc_hotSale_userid_allTime_subitem_without_price ：用户最近一年的热销(套餐子项 = 0)

kfc_hotSale_userid_recentQuarter_tradeup ： 用户最近3个月tradeup推荐热销

kfc_times_userid_recentQuarter_category ：用户最近3个月购买各类category的次数

kfc_times_userid_recentQuarter_price_single ：用户最近3个月购买单品的各类价格区间的次数

kfc_times_userid_recentQuarter_price_item ：用户最近3个月购买套餐的各类价格区间的次数

kfc_times_userid_recentQuarter_single :用户最近3个月的购买单品次数（下单里面包含单品的次数）

kfc_times_userid_recentQuarter_item:用户最近3个月的购买套餐次数

kfc_price_userid_recentQuarter_single :用户最近3个月的单品购买金额

kfc_price_userid_recentQuarter_item :用户最近3个月的套餐购买金额

kfc_hit_recentMonth_tradeup:用户最近一个月的tradeup命中情况【单独写入redis及特征更新】

(b) shop类特征说明：

kfc_hotSale_shop_recentQuarter_single: 商户最近3个月的热销 (单品)

kfc_hotSale_shop_recentQuarter_item: 商户最近3个月的热销 (套餐)

kfc_hotSale_shop_recentQuarter_subitem_with_price: 商户最近3个月的热销(套餐子项>0)

kfc_hotSale_shop_recentQuarter_subitem_without_price: 商户最近3个月的热销(套餐子项=0)

kfc_hotSale_shop_allTime_single ： 商户最近一年的热销(单品)

kfc_hotSale_shop_allTime_item ： 商户最近一年的热销(套餐)

kfc_hotSale_shop_allTime_subitem_with_price ： 商户最近一年的热销(套餐子项>0)

kfc_hotSale_shop_allTime_subitem_without_price ： 商户最近一年的热销(套餐子项=0)

(d) city类的特征说明：

kfc_hotSale_city_recentQuarter_single : 城市最近3个月的热销(单品)

kfc_hotSale_city_recentQuarter_item : 城市最近3个月的热销(套餐)

kfc_hotSale_city_recentQuarter_subitem_with_price ：城市最近3个月的热销(套餐子项>0)

kfc_hotSale_city_recentQuarter_subitem_without_price ：城市最近3个月的热销(套餐子项=0)

kfc_hotSale_city_allTime_single : 城市最近一年的热销(单品)

kfc_hotSale_city_allTime_item : 城市最近一年的热销(套餐)

kfc_hotSale_city_allTime_subitem_with_price ：城市最近一年的热销(套餐子项>0)

kfc_hotSale_city_allTime_subitem_without_price ：城市最近一年的热销(套餐子项=0)

(e) info类的特征说明 (具体分类待进一步确认)

kfc_recentHotSale_single : 最近一周热销top30(单品)

kfc_recentHotSale_item : 最近一周热销top30(套餐)

kfc_recentHotSale_subitem_with_price ：最近一周热销top30(套餐子项>0)

kfc_recentHotSale_subitem_without_price ：最近一周热销top30(套餐子项=0)

kfc_recentHotSale_weekend_single : 最近一周非工作日单品热销

kfc_recentHotSale_weekend_fix : 最近一周非工作日套餐热销

(4) 数据写入

功能：将特征数据写入redis；采用自动化读入形式，自动扫描person、shop、city、menu_info、tradeup_info下的所有子文件，并根据文件名生成子key名，

写入redis； 包含两种写入方式：单条hmset 以及pipe批量写入；person采用pipe批量写入的形式，shop、city采用hmset方式，menu_info、tradeup_info待定

输入文件：/data/offline_data/feature_data/person、shop、city、menu_info、tradeup_info

输出文件：写至redis

(5) 离线模型 & 离线评估

功能：包括离线数据预处理 及 模型训练； 数据预处理方式 需保持与线上一致；离线评估指标包括ta、命中率等；

输入文件：/data/offline_data/feature_data/person、shop、city、menu_info、tradeup_info

输出文件：中间数据 及结果数据，存至 /offline_data/model目录下

1） 数据样本收集

包含样本标注、特征文件收集两个过程

a. 源数据采集

a) 样本标注

输入文件：/data/offline_data/origin_data/recommend_info/日期
输出文件：trade_up
标注：/data/offline_data/model/trade_up/samples_data/annotation_tradeup/日期
menu 标注:
/data/offline_data/model/menu/samples_data/annotation_menu/日期

b）购物车信息采集（只有tradeup）
输入文件：kfc方提供源日志

输出文件：/data/offline_data/model/trade_up/samples_data/shopcart_tradeup/日期
输出文件格式：json形式存储：{transaction_id
:[{"systemId":"100020427"},{}]}

c) 基本信息采集
输入文件：样本标注文件；join数据文件
输出文件：trade_up
标注：/data/offline_data/model/trade_up/samples_data/annotation_map_tradeup/日期
menu 标注:
/data/offline_data/model/menu/samples_data/annotation_map_menu/single_日期

/data/offline_data/model/menu/samples_data/annotation_map_menu/item_日期
输出文件格式：

	字段	备注
0	transation_guid	订单id
1	transaction_start_time	下单时间
2	user_code	用户id
3	store_code	餐厅编码
4	city_code	城市编码
5	product_sell_price	商品单价
6	is_combo_flag	单品0 可选套餐头1 套餐子项2 优惠3 固定套餐20
7	nameCN	套餐名
8	subclass	单品名
9	category	Drink / Snack
10	class	Tea / fried wing
11	subcategory	Tea /wing
12	linkid	同样的产品在不同的餐厅linkid不同
13	shunfeng_tag	添加的商品标识
14	class_flag	是否购买

输入文件：样本标注文件；join数据文件
输出文件：
detail表：/data/offline_data/model/common_data/detail_日期
link_id表: /data/offline_data/model/common_data/single_link_id_日期

/data/offline_data/model/common_data/item_link_id_日期

b. 特征提取

输入文件：tradeup标注：/data/offline_data/model/trade_up/samples_data/annotation_tradeup/日期

menu标注：/data/offline_data/model/menu/samples_data/annotation_menu/日期
相关特征数据文件

输出文件：tradeup特征：/data/offline_data/model/trade_up/samples_data/feature_tradeup/日期_日期.txt

menu特征： /data/offline_data/model/menu/samples_data/feature_menu/single_日期_日期.txt

/data/offline_data/model/menu/samples_data/feature_menu/item_日期_日期.txt

2） 数据处理：归一化 & 数据平衡

输入文件：tradeup特征：/data/offline_data/model/trade_up/samples_data/feature_tradeup/日期_日期.txt

menu特征： /data/offline_data/model/menu/samples_data/feature_menu/日期_日期.txt

输出文件：

tradeup归一化：/data/offline_data/model/trade_up/samples_data/normalize_tradeup/日期_日期.txt

menu归一化： /data/offline_data/model/menu/samples_data/normalize_menu/日期_日期.txt

tradeup数据平衡：/data/offline_data/model/trade_up/samples_data/balance_tradeup/日期_日期.txt

menu数据平衡：/data/offline_data/model/menu/samples_data/balance_menu/日期_日期.txt

2.0版本归一化说明：归一化采用Min-Max Normalization方法

demo数据20181211

min,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0

max,1.0,217433.0,20141.0,356936.0,137465.0,4717.0,3914.0,195.0,168.0,1.0,1.0,1.0,1.0,1.0,0.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,0.0,0.0

3） 模型训练

输入文件：tradeup特征：/data/offline_data/model/trade_up/samples_data/feature_tradeup/日期_日期.txt

menu特征： /data/offline_data/model/menu/samples_data/feature_menu/日期_日期.txt

输出文件：tradeup模型：/data/offline_data/model/trade_up/offline_model/model/tradeup.m

menu模型：单品：/data/offline_data/model/menu/offline_model/model/single_menu.m

套餐：/data/offline_data/model/menu/offline_model/model/item_menu.m

配置文件： tradeup：/data/offline_data/model/trade_up/offline_model/conf/...

menu：/data/offline_data/model/menu/offline_model/conf/...

4） 模型效果评估： 离线 & 在线

评估文件：(根据具体指标调整)

tradeup：/data/offline_data/model/trade_up/evaluation/日期.txt

menu: /data/offline_data/model/menu/evaluation/日期.txt

5） 特征说明（迭代中...）

初版： 根据现有的poc阶段数据，训练初版模型

物品信息特征：用户商品与购物车商品的category关联值；商品店铺维度全年销量；商品城市维度全年销量；

商品店铺维度三个月销量；商品城市维度三个月销量； 商品一周单品销量；

用户信息特征：

用户-物品关联特征：商品（映射后）是否在用户购物车内；商品用户维度全年销量；商品用户维度三个月销量；商品用户维度一周销量；

其他特征：

2.0版本特征说明（按顺序依次，Tradeup而言）

【1】是否在购物车内、【2】最近一周热销量、【3】商品与购物车内商品关联规则值（如果购物车category存在该category，对应category关联值在累加时置0）、

~~【4】城市全年热销量、【5】城市最近三个月热销、【6】商户全年热销、【7】 商户最近三个月热销、~~【8】用户全年热销、【9】用户最近三个月热销、【10】category是否是Drink、

【11】category是否是dessert、【12】category是否是side_item、【13】category是否是snack、【14】category是否是main、

【15】时段是否是breakfast、【16】时段是否是morning、【17】时段是否是lunch、【18】时段是否是afternoonTea、

【19】时段是否是dinner、【20】时段是否是midnightSnack、【21】价格区间是否在（0-6）、【22】价格区间是否在（6-12）、【23】价格区间是否在（12-）、【24】推荐品的category和购物车内的category是否一致、【25】购物车是否有drink

【26】购物车是否有Main　　【27】购物车内是否有drink * 推荐产品是否是drink【没有且推荐则为1】

【28】购物车是否有main * 推荐产品是否是main【没有且推荐则为1】

【29】购物车category是否一致 * 推荐产品是否是Snack

b. 【30】用户购买过的category与推荐商品category相同的订单次数（如果购物车category有该类，则置0）、

i. 【31】用户购买过的最高价格区间（顶峰35-45）是否高于购物车金额 + 推荐产品金额（购物车大于50的，直接置为1，+5元缓冲，最小值30）、

【32】购物车金额是否小于25　　【33】购物车金额是否大于50　　【34】购物车金额（50以内保留原值，大于50置0？）

【35】0-15价格区间购买次数

【36】15-25价格区间购买次数【37】25-35价格区间购买次数【38】35-45价格区间购买次数

【39】45-购买次数　　【40】是否属于高额用户（最近三个月在区间35-45有购买记录）【41】用户下单频次

ii. 【42】用户关联规则值max　【43】用户关联规则值min　【44】用户关联规则值mean

【45】购物车单品个数　　【46】购物车套餐个数

【47】购物车是否既有单品又有套餐

【48】 用户最近3个月tradeup推荐热销　　【49】推荐命中次数

【50】推荐未命中次数

【51】 推荐命中率（36 /（36 + 37））【52】tradeup产品命中率（hit_num - 10 / recom_num - 10；若数值出现负值，置0）

iii. 【53】用户推荐命中次数 / 该tradeup产品总命中次数（46/hit_num）** 注意顺序，采用下标　【54】最近一周tradeup热销

【55】用户偏好产品购买次数　　【56】 用户偏好购买率

【57】 用户偏好率：48 * 47　　【58】商品价格

iv.

v. （todo ： 商品在tradeup一周的热销）

思考：缺失值填充方式；todo：不同购物车金额区间，命中产品分布分析；考虑是否是某category * 该category是否在购物车，尤其针对category为drink】

【新奥尔良烤翅和鸡翅。。共同出现； dessert 和 drink】

主要特征对应的key(正式版)说明：

最近一周热销量：key值：kfc:info；field值：kfc_recentHotSale_single

商品与购物车内商品关联规则值：key值：kfc::person::用户id；field值：kfc_food_association_category

城市全年热销量：key值：kfc:city:城市id；field值：kfc_hotSale_city_allTime_single

城市最近三个月热销：　key值：kfc:city:城市id；field值：kfc_hotSale_city_recentQuarter_single

商户全年热销：key值：kfc:shop:商户id；field值：kfc_hotSale_shop_allTime_single

商户最近三个月热销：key值：kfc:shop:商户id；field值：kfc_hotSale_shop_recentQuarter_single

用户全年热销：key值：kfc:person:用户id；field值：kfc_hotSale_userid_allTime_single

用户最近三个月热销：key值：kfc:person:用户id；field值：kfc_hotSale_userid_recentQuarter_single

时段说明：

“breakfast”：6:00-9:30,
“morning”:9:30-11:00,
“lunch”:11:00-14:00
“afternoonTea”:14:00-17:00,
“dinner”:17:00-20:00:
“midnightSnack”:20:00-23:00

用户购买过的category与推荐商品category相同的订单次数：key值：kfc::person::用户id；field值：kfc_times_userid_recentQuarter_category

用户购买过的价格区间与推荐商品+购物车商品价格区间相同的订单次数占比：key值：kfc:person:用户id；field值：　kfc_origin_food_price

用户购买过的价格区间定义：（具体待商量）

2.0版本特征说明 （按顺序依次，Menu—single而言）

1）最近一周热销；2）城市全年热销；3）城市最近3个月热销；4）商户全年热销；5）商户最近三个月热销；6）用户全年热销；

7）用户最近三个月热销； 8）category是否是Drink；9）category是否是dessert；10）category是否是side_item；11）category是否是snack；

12） category是否是main； 13）时段是否是breakfast；14）时段是否是morning；15）时段是否是lunch；

16）时段是否是afternoonTea；17）时段是否是dinner；18）时段是否是midnightSnack；19）价格区间是否在(0-10)；

20）价格区间是否(10-20)； 21）价格区间是否(20-)；22）
用户购买过的category与推荐商品category相同的订单次数； 23）用户购买单品区间与该用户单品区间相同的次数；

24）用户该时段的购买单品次数
；25）用户购买总金额区间在(0-15)的次数； 26）用户购买总金额区间在(15-25)的次数；

27）用户购买总金额区间在(25-35)的次数；28）用户购买总金额区间在(35-45)的次数；29）用户购买过的总金额区间在（45-）的次数

主要特征对应的key(正式版)说明：

最近一周热销量：key值：kfc:info；field值：kfc_recentHotSale_single

城市全年热销量：key值：kfc:city:城市id；field值：kfc_hotSale_city_allTime_single

城市最近三个月热销： key值：kfc:city:城市id；field值：kfc_hotSale_city_recentQuarter_single

商户全年热销：key值：kfc:shop:商户id；field值：kfc_hotSale_shop_allTime_single

商户最近三个月热销：key值：kfc:shop:商户id；field值：kfc_hotSale_shop_recentQuarter_single

用户全年热销：key值：kfc:person:用户id；field值：kfc_hotSale_userid_allTime_single

用户最近三个月热销：key值：kfc:person:用户id；field值：kfc_hotSale_userid_recentQuarter_single

时段说明：

“breakfast”: 6:00-9:30,

“morning”:9:30-11:00,

“lunch”:11:00-14:00

“afternoonTea”:14:00-17:00,

“dinner”:17:00-20:00:

“midnightSnack”:20:00-23:00

用户购买过的category与推荐商品category相同的订单次数:key值：kfc:person； field值:kfc_times_userid_recentQuarter_category

用户购买单品区间与该用户单品区间相同的次数:key值：kfc:person;
field值：kfc_times_userid_recentQuarter_price_single

价格区间：（0-5）；（5-10）；(10-15)；.....(55-60)；(60-70)； (70-90)；(90-)

用户购买过的单品次数：key值：kfc:person;
field值：kfc_times_userid_recentQuarter_single

用户购买金额次数： key值：kfc:person；field值:kfc_origin_food_price；

格式说明：

```
{
    “0-15”: 5,
    “15-25”:4,
      “25-35”:3,
      “35-45”:2,

    “45-”:2

 }
```

2.0版本特征说明 （按顺序依次，Menu—item而言）

1）最近一周热销；2）城市全年热销；3）城市最近3个月热销；4）商户全年热销；5）商户最近三个月热销；6）用户全年热销；

7）用户最近三个月热销； 8）时段是否是breakfast；9）时段是否是morning；10）时段是否是lunch；

11）时段是否是afternoonTea；12）时段是否是dinner；13）时段是否是midnightSnack；14）价格区间是否在(0-30)；

15）价格区间是否(30-60)；16）价格区间是否(60-)；

17）用户购买套餐区间与该用户套餐区间相同的次数 ；18）用户该时段的购买套餐次数；19）用户购买总金额区间在(0-15)的次数； 20）用户购买总金额区间在(15-25)的次数；21）用户购买总金额区间在(25-35)的次数；22）用户购买总金额区间在(35-45)的次数；23）用户购买过的总金额区间在（45-）的次数

主要特征对应key说明:

最近一周热销量：key值：kfc:info；field值：kfc_recentHotSale_item

城市全年热销量：key值：kfc:city:城市id；field值：kfc_hotSale_city_allTime_item

城市最近三个月热销： key值：kfc:city:城市id；field值：kfc_hotSale_city_recentQuarter_item

商户全年热销：key值：kfc:shop:商户id；field值：kfc_hotSale_shop_allTime_item

商户最近三个月热销：key值：kfc:shop:商户id；field值：kfc_hotSale_shop_recentQuarter_item

用户全年热销：key值：kfc:person:用户id；field值：kfc_hotSale_userid_allTime_item

用户最近三个月热销：key值：kfc:person:用户id；field值：kfc_hotSale_userid_recentQuarter_item

时段说明：

“breakfast”: 6:00-9:30,
“morning”:9:30-11:00,
“lunch”:11:00-14:00
“afternoonTea”:14:00-17:00,
“dinner”:17:00-20:00:
“midnightSnack”:20:00-23:00

用户购买套餐区间与该用户套餐区间相同的次数:key值：kfc:person;
field值：kfc_times_userid_recentQuarter_price_item

价格区间：（0-5）；（5-10）；(10-15); .....(55-60); (60-70); (70-90); (90-)

用户购买过的套餐次数：key值：kfc:person;
field值：kfc_times_userid_recentQuarter_item

用户购买金额次数： key值：kfc:person；field值:kfc_origin_food_price；

格式说明：

```
{
    “0-15”: 5,
    “15-25”:4,
     “25-35”:3,
     “35-45”:2,

    “45-”:2

 }
```

(6) 基本信息 & 配置文件

导入如菜单信息等基本——信息tradeup菜品信息

4. 正式版 调整需求：

所有的热销特征，都分单品 和 套餐 和 套餐子项 （> 0 的套餐子项， = 0 的套餐子项）；

5. 正式版 menu 模型

menu分别对单品、套餐头和套餐子项排序，并按餐期分开预估，并选择top_k存redis，具体格式详见正式版特征key格式说明

其中single，top_k=3

item, top_k = 2

subitem , top_k = 5

具体流程：

（1）根据city 和 shop获取具体某家店的分时段的产品作为召回候选集1

（2） 根据person的分时段热销产品作为召回候选集2；候选集1与候选集2的并集作为用户最终候选集（注意去重）

（3） 针对召回集预测模型，单品选取top3，套餐头选取top2，

（4）将数据写入redis

kfc_model_menu_single

kfc_model_menu_item

6. 线上特征自动化写入注意的点

【1】 写入前确认join_data相应日期数据都存在

【2】 feature产出后，相应特征对对应存在(检查对应特征都有产出、特征大小合理、特征格式ok)

【3】redis确认已写入（之前存在分片号调整，导致未写入的情况）

redis访问命令：redis-cli -h ip地址 -p 端口号（对应ip地址及端口号 可至conf文件查询）

具体处理流程参见该wiki 3.(3)、3.(4)流程