最终报告

组员：徐晶，王鹏，唐正，申长虹

一、 简介

随着移动定位服务的流行，阿里巴巴和蚂蚁金服逐渐积累了来自用户和商家的海量线上线下交易数据。蚂蚁金服的 O2O 平台“口碑”用这些数据为商家提供了包括交易统计，销售分析和销售建议等定制的后端商业智能服务。

该项目主要想解决的问题：预测客户流量对商家的经营管理至关重要。在口碑平台上，将客户流量定义为“单位时间内在商家使用支付宝消费的用户人次”。在这个问题中，将提供用户的浏览和支付历史，以及商家相关信息，并希望可以以此预测所有商家在接下来 14 天内，每天的客户流量。

二、 问题陈述

2.1 问题场景

在本问题中，需要预测测试集中所有商家在未来 14 天（2016.11.01-2016.11.14）内各自每天（00:00:00-23:59:59）的客户流量。预测结果为非负整数。

c_{it}：第 t 天，商家 i 的客户流量预测值（由参赛选手提供）

c_{it}^{g}：第 t 天，商家 i 的客户流量实际值

$$L = \frac{1}{nT}\sum_{i}^{n}\sum_{t}^{T}\left|\frac{c_{it} - c_{it}^{g}}{c_{it} + c_{it}^{g}}\right|$$

2.2 数据集

数据集为从 2015.07.01 到 2016.10.31（除去 2015.12.12）的商家数据，用户支付行为数据以及用户浏览行为数据。提供数据的类型统一为 string 类型，提交预测的类型为整形。文件统一为 utf-8 编码，没有标题行，并以“,”分隔的 csv 格式。

1. user_pay：用户支付行为

Field	Sample	Description
user_id	0000000001	用户id
shop_id	000001	商家id，与shop_info对应
time_stamp	2015-10-10 11:00:00	支付时间

2. shop_info：商家特征数据

Field	Sample	Description
shop_id	000001	商家id
city_name	北京	市名
location_id	001	所在位置编号，位置接近的商家具有相同的编号
per_pay	3	人均消费（数值越大消费越高）
score	1	评分（数值越大评分越高）
comment_cnt	2	评论数（数值越大评论数越多）
shop_level	1	门店等级（数值越大门店等级越高）
cate_1_name	美食	一级品类名称
cate_2_name	小吃	二级分类名称
cate_3_name	其他小吃	三级分类名称

3. user_view：用户浏览行为

Field	Sample	Description
user_id	0000000001	用户id
shop_id	000001	商家id，与shop_info对应
time_stamp	2015-10-10 10:00:00	浏览时间

4. prediction：测试集与提交格式

Field	Sample	Description
shop_id	000001	商家id
day_1	25	第1天的预测值（需要选手提供)
day_2	3	第2天的预测值（需要选手提供)
.....		
day_14	1024	第14天的预测值（需要选手提供)

2.3 预期结果

在本问题中，需要预测测试集中所有商家在未来 14 天（2016.11.01-2016.11.14）内各自每天（00:00:00-23:59:59）的客户流量。预测结果为非负整数。

三、 技术方案

3.1 外部数据及数据预处理

3.1.1 外部数据

外部数据分为机场天气数据和节假日信息两部分，具体如下：

（1）天气数据

天气数据来源：https://www.wunderground.com， 提供了世界各地在机场附近检测到的气象信息，包含气温，露点，湿度，气压，能见度，风速，瞬时风速，降水量，天气状况等信息。历史气象信息的采样间隔为 30 分钟。测试集首日，北京首都国际机场 2016 年 11 月 1 日气象条件的时间序列分布如图 1。

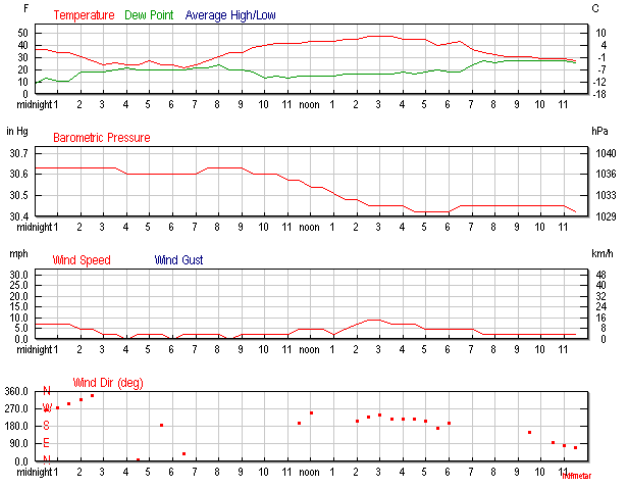

图 1 北京首都国际机场 2016 年 11 月 1 日气象条件的时间序列分布

数据说明：

日降水量：采样间隔为天，爬取程序为 Weather_underground_day.py，生成的降水量表格为 PRECIP.csv

气象条件时间序列：采样间隔为 30 分钟，爬取程序为 Weather_underground_hour.py，生成的气象条件时序表格为 WEATHER_raw.csv，

降水指数和天晴指数：天气状况给出天气条件纷繁复杂，通过经验将天气状况列简单转换为降水指数和天晴指数两个指标。分别为 feature/WEATHER_CON_LEVEL.csv 中 RAIN_IND 及 CLEAR_IND 对应列。

由于人体对于气象参数的感受不为线性关系，生成人体舒适度指数 SSD SSD=(1.818t+18.18)(0.88+0.002f)+(t-32)/(45-t)-3.2v+18.2 其中：温度 t，湿度 f，风速 v

城市天气确定：通过城市经纬度计算城市到各机场距离，城市对应天气采用与之最近的机场信息。

（2）节假日信息

节假日信息 HOLI.csv，将日期类型简单分为三个类别，其中工作日标签为 0，周末标签为 1，假期标签为 2。表格来源为比赛官方论坛。

3.1.2 数据预处理

数据清洗包含三部分，通过规则清除，通过模型预训练清除及仅保留销量统计信息：

（1） 规则清除

原始数据中，存在单用户某小时内大量购买的现象，如 userID 为 9594359 用户在 2016 年 1 月 30 日在 shopID 为 878 的商家累计购买了 209 次。针对此类现象，对于单个用户单小时内的购买数量 x，采用以下公式处理消除异常消费，如图 2 所示：

$$f(x) = 1 + log_2 x$$

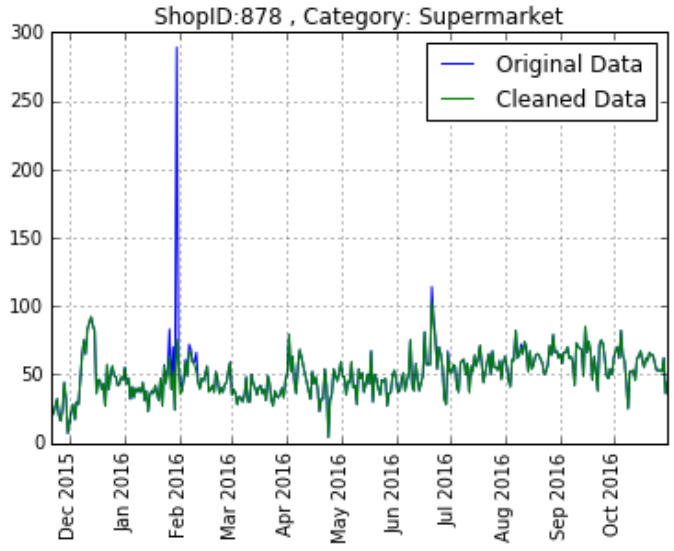

图 2 单个用户随时间购买量

商家初始入驻口碑平台存在一定的启动时间，同时销售过程中会在销量中断的现象，如下图 shopID 为 1072 的商家所示。针对此类现象，开业前 7 天数据不用于训练集，销量间断前后 3 天数据不用于训练集，如图 3 所示。

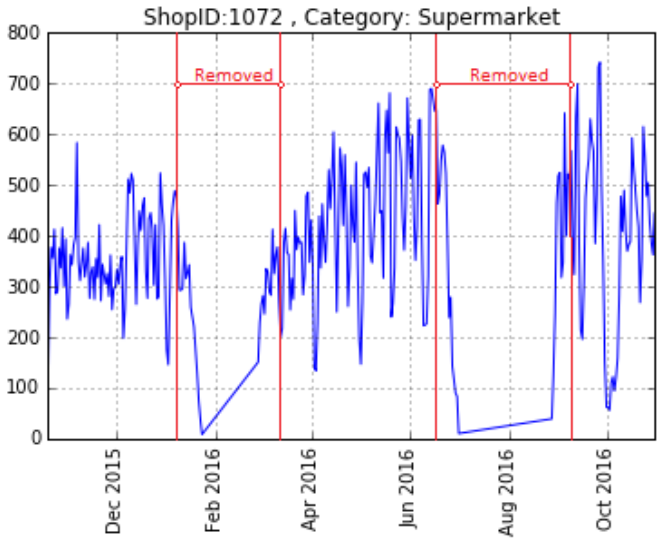

图 3 剔除商家开业前 7 天数据图

销量以历史过去 14 天销量的 μ±2σ 为限制，其中 μ 为均值，σ 为均方根，以排除异常的销量数，如图 4 所示。

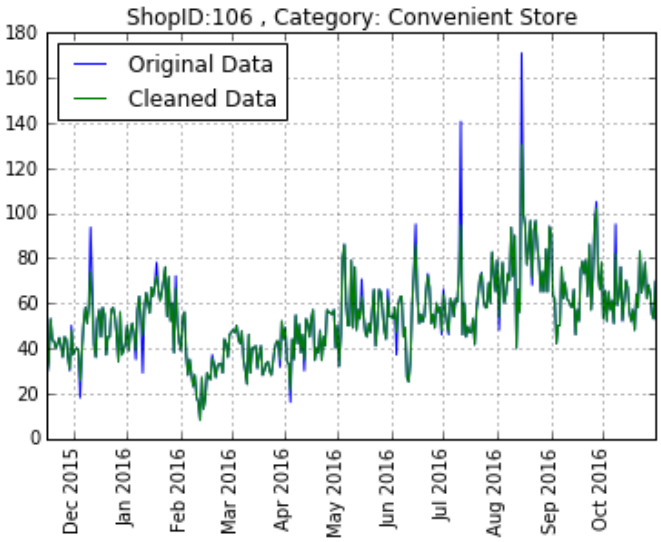

图 4 剔除异常销量数据图

（2） 模型预训练清除

详见第三部分。商家日销量，可能存在一些难以预计的大幅波动，如促销，商家停业等。对于这些规则难以清除的异常值，采用预训练的方式清除。模型训练中首先采用欠拟合的算法模型预训练，并清除残差最大的 10%(xgboost1,GBDT)和 25%(xgboost2)的样本。

（3） 仅保留销量统计信息

由于只需要预测商家的日销量，无需识别单个用户的行为，按照大数定理，可以只针对分时段的浏览与购买总数进行预测。因而在数据清洗后，保留的数据仅按小时统计商户总销量，在这一步剔除了用户 ID，使得数据量仅为原始的约 1/10。

3.2 预测模型

最终销量预测结果由未来 14 天常规销量预测及双 11 修正系数预测两步两部分组成。

通过双 11 修正系数，分别对于 2016-11-11，2016-11-12，2016-11-13 三天的销量按照 1.0，0.2，0.1 倍的系数作乘法修正。双 11 修正部分训练采用 xgboost 单模型，特征为商家特征信息，标签为上一年（2015 年）双 11 当天的销量增量百分比。常规销量预测部分，采用基本模型有 4 套，分别为 2 套 xgboost 模型(特征处理及数据清洗程度不同)，GBDT 模型和均值模型，整体模型框架如图 5 所示。

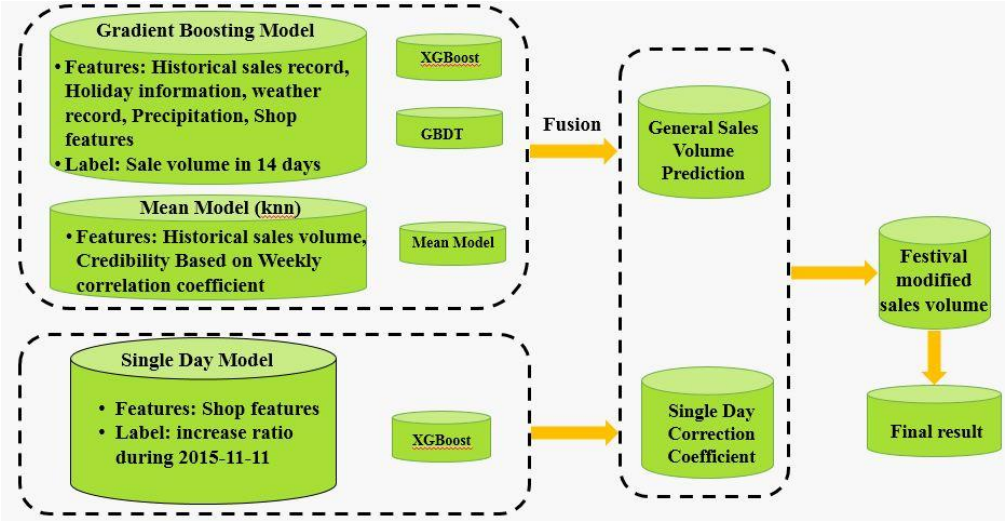

图 5 整体模型框架

对于模型训练的具体说明如下：

3.2.1 常规销量预测模型

特征与标签见表 1：

表 1 特征与标签

特征与标签	说明
历史销量特征	过去21天的历史销量
节假日特征	过去21天及预测14天的节假日标注
天气特征	过去21天及预测当天附近4天(之前两天，当天，之后一天)的降水量，人体舒适度SSD值，降水指数，天晴指数
商家特征	平均View/Pay比值，平均每天开店时间，关店时间，开店总时长；首次营业日期，非节假日销量中位数，节假日销量中位数，节假日/非节假日销量比值；商家类别，人均消费，评分，评论数，门店等级
标签	未来14天日销量

训练方式：

（1）采用滑窗对于 2000 个商家日销量的时间序列生成 481143 条有效训练样本，清除间断前后及异常值后保留 468535 条样本。

（2）采用 2 次训练的方法，第一次采用最大深度为 3 欠拟合模型进一步清洗脏数据。采用了 xgboost 与 sklearn 的 GBDT 模型训练，具体参数如下： XGBoost-Round_1: 日销量仅作 log 处理，预训练后样本保留量为 90%。 XGBoost-Round_2: 日销量仅作 log 处理后，采用

过去三周的中位数作无量纲，预训练后样本保留量为 75%。

XGBoost	objective	max_depth	learning_rate	n_estimators	reg_alpha	reg_lambda
Round_1	reg:linear	3	0.1	500	0	1
Round_2	reg:linear	5	0.03	1600	1	0

GBDT：第一次训练样本保留量为 90%。

GBDT	loss	max_depth	learning_rate	n_estimators	alpha
Round_1	lad	3	0.1	500	0.95
Round_2	lad	5	0.1	500	0.95

3.2.2 历史均值模型

输入:过去 21 天的历史销量，过去三周的销量相关度矩阵。

输出：未来 2 周的销量及其对应在模型融合中置信度。

方法：过去 21 天的按工作日平均，得到按工作日平均的均值销量。通过过去三周按周统计的销量中位数及平均值，做线性拟合得到销量增量。将历史均值销量叠加销量增量即得到未来 2 周预测销量。

由于方法本质上寻找历史上相似的(过去三周相关度较高)销量曲线作为未来预测，本质上为均值模型与 KNN 方法的结合。

置信度即为融合系数，仅当三周相关系数或后两周相关系数的最小值大于 0.7 时有效。均值模型的融合比例最大为 0.75。融合系数计算方法为：

$$Cred = 0.75 \times \frac{\min(\overline{Cor}, Cor_{last}) - 0.7}{0.3}$$

3.2.3 双 11 销量修正模型

模型概述：需要预测的时间段（11 月 1 日到 11 月 14 日范围内）包含双 11 节日。从诸多商家的销量图上能明显看到在双 11 当天存在较大波动，可能的原因为网商促销对实体店的冲击，双 11 作为光棍节对于餐饮业的促进。然而仅有约 1/3 的商家存在 2015 年双 11 的销量记录，需要通过这部分商家去年双 11 信息，预测其余商家双 11 销量表现。

特征描述：仅包含商家特征，包含平均 View/Pay 比值，平均每天开店时间，关店时间，开店总时长；首次营业日期，非节假日销量中位数，节假日销量中位数，节假日/非节假日销量比值；商家类别，人均消费，评分，评论数，门店等级。

双 11 销量增量，计算方法为上一年(2015 年)11-11 当天销量 V1111 与其前后两周对应工作日 V1028，V1104，V1118，V1125 的加权销量的比值，权重系数分别为 $0.15,0.35,0.35,0.15$.

$$M_{11} = \frac{V_{20151111}}{0.35 \times (V_{20151104} + V_{20151118}) + 0.15 \times (V_{20151028} + V_{20151125})}$$

训练方法：采用 xgboost 单模型训练，由于双 11 当天对应的工作日不同，2015 年数据并不能很好反映出 2016 年双 11 节假日情况，且超市便利店类商店存在大量的数据缺失。为防止过拟合，参数设置均较为保守，最大深度为 2，且加了较大的L_1正则项，具体如下:

max_depth = 2, learning_rate=0.01, n_estimators=500, reg_alpha=10, gamma = 1。

3.2.4 模型融合

（1）多套 gradient boosting 的结果间的融合 xgboost1，xgboost2, GBDT 三份结果按 0.47, 0.34, 0.19 比例融合。

（2）gradient boosting 与均值模型融合 将均值模型结果与步骤 1 gradient boosting 的结果融合，均值模型的融合系数为通过相关度得到的置信度。

（3）双 11 系数进行销量调制 双 11 当天销量乘以双 11 销量修正模型得到的销量增量，11-12, 11-13 由于为周六周日，有理由相信其销量与 11-11(周五)的表现存在相似性， 因而乘以 0.2 及 0.1 倍的销量增量系数。

四、 实现和实验结果

4.1 实现步骤

步骤 1：生成精简版本 user_pay, user_view 表格

由于只需要预测商家的日销量，无需识别单个用户的行为，按照大数定理，可以只针对分时段的浏览与购买总数进行预测。此程序按小时统计商户销量，并进行用户异常刷单清理，生成精简版本的 pay 和 view 表格分别为，user_pay_new.csv 和 user_view_new.csv，文件大小减小至原始数据的 1/10，以便后续访问及特征提取。

步骤 2：外部数据爬取

从 https://www.wunderground.com 按天读取机场所在地信息，爬取信息包含 7 列分别为[Port, Date, Precip , Rise_act, Rise_cil, Set_act, Set_cil]，对应内容为[机场代号，日期，降水量，真实日出时间，修正日出时间，真实日落时间，修正日落时间]。

步骤 3：特征生成

（1）生成天气特征表 WEATHER_FEATURES.csv, 包含四项, 分别为人体舒适度 SSH 值，SSH 值日增量，降水指数，天晴指数。

（2）生成商家特征表 SHOP_FEATURES.csv，包含平均 View/Pay 比值，平均每天开店时间，关店时间，开店总时长；首次营业日期，非节假日销量中位数，节假日销量中位数，节假日/非节假日销量比值；商家类别，人均消费，评分，评论数，门店等级。

（3）生成测试集历史过去三周销量表格，修正异常销量，以历史过去 14 天销量的 μ±2σ 为限制，其中 μ 为均值，σ 为均方根

整合所有特征，生成方便训练模型读取的 X.csv, Y.csv, Xtest.csv 三个表格

步骤 4：常规销量模型训练

包括 GBDT 模型和均值模型。

步骤 5：双 11 修正系数训练

双 11 修正模型，获得双 11 当天销量增加百分比，以此针对常规销量在 11-11, 11-12, 11-13 三天进行乘法修正。

步骤 6：模型融合

各模型融合并作双 11 修正生成最终提交结果

4.2 实验结果

由测评获得的 Loss 提升表格如表 2，按特征的重要性排序，分别为：历史销量特征，节假日特征，降水天气特征，商家特征。

表 2　Loss 提升结果

方案	Loss
最后三周按工作日平均	0.895
最后六周按工作日平均，乘1.05系数	0.860
Xgboost+最后三周销量特征	0.824
上述增加，节假日特征	0.813
上述，增加天气，商家特征	0.798
上述，预训练剔除10%脏数据	0.791
上述，测试集10月缺失数据填补	0.788
增加双11节假日修正模型	0.780
Xgboost + GBDT模型融合	0.774
上述按相关系数融合平均销量	0.772

从实验结果看，我们的模型对于 Loss 的提升结果均大于 0.77，说明我们的模型能够有效地预测测试集中所有商家在未来 14 天（2016.11.01-2016.11.14）内各自每天（00:00:00-23:59:59）的客户流量，证明了该模型的可行性。

但模型依然存在一些不足：

(1)赛题关键在于各个商店总体销量预测，需要对于历史整体趋势有很好把握。预测日期 11 月 1 日到 11 月 14 日区间，由于距离国庆假期较近，容易受到脏数据干扰，因而特征提取过程中仅采用过去三周的信息提取特征。现有的解决方案中，对于周期更长的趋势仍把握不足。

(2)预测日期包含了双 11，给预测增加了难度。现有方案采用少量商家在 2015 年双 11 当天销量增量百分比预测其余所有商家在 2016 年的销量增量，存在着诸多不足。比如这两年的双 11 并不在同一个工作日，2015 年缺少大量超市类数据等，导致训练集与预测集分布差异大。后续可能需要通过更多的相似节假日(圣诞，七夕)挖掘商家销量规律。

这些问题都将在我们未来的工作中去考虑。