

第十一届“泰迪杯”数据挖掘挑战赛——

B 题：产品订单的数据分析与需求预测

一. 问题背景

近年来企业外部环境越来越不确定,复杂多变的外部环境,让企业供应链面临较多难题。需求预测作为企业供应链的第一道防线,重要程度不言而喻,然而需求预测受多种因素的影响,导致预测准确率普遍较低,因此需要更加优秀的算法来解决这个问题。需求预测是基于历史数据和未来的预判得出的有理论依据的结论,有利于公司管理层对未来的销售及运营计划、目标,资金预算做决策参考;其次,需求预测有助于采购计划和安排生产计划的制定,减少受业务波动的影响。如果没有需求预测或者预测不准,公司内部很多关于销售、采购、财务预算等决策都只能根据经验而来了,会导致对市场预测不足,产生库存和资金的积压或不足等问题,增加企业库存成本。

二. 数据说明

附件中的训练数据 (order_train1.csv) 提供了国内某大型制造企业在 2015 年 9 月 1 日至 2018 年 12 月 20 日面向经销商的出货数据 (格式见表 1), 反应了该企业产品在不同销售区域的价格和需求等信息, 包括: order_date (订单日期)、sales_region_code (销售区域编码)、item_code (产品编码)、first_cate_code (产品大类编码)、second_cate_code (产品细类编码)、sales_chan_name (销售渠道名称)、item_price (产品价格) 和 ord_qty (订单需求量)。

表 1: 训练数量 (历史数据) 的数据格式

order_date	sales_region_code	item_code	first_cate_code	second_cate_code	sales_chan_name	item_price	ord_qty
2015/9/1	104	22069	307	403	offline	1114	19
2015/9/1	104	20028	301	405	offline	1012	12
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

其中“订单日期”为某个需求量的日期; 一个“产品大类编码”会对应多个“产品细类编码”; “销售渠道名称”分为 online (线上) 和 offline (线下), “线上”是指淘宝和京东等电商平台, “线下”是指线下实体经销商。

附件中的预测数据 (predict_skul.csv) 提供了需要预测产品的销售区域编码、产品编码、产品品类和产品细品类 (格式见表 2)。

表 2: 需要预测的产品的数据样例

sales_region_code	item_code	first_cate_code	second_cate_code
101	20002	303	406
101	20003	301	405
⋮	⋮	⋮	⋮

三. 需要解决的问题

1. 请对附件中的训练数据 (order_train1.csv) 进行深入地分析, 可参照但不限于下述主题。

- (1) 产品的不同价格对需求量的影响;
- (2) 产品所在区域对需求量的影响, 以及不同区域的产品需求量有何特性;

- (3) 不同销售方式（线上和线下）的产品需求量的特性；
 - (4) 不同品类之间的产品需求量有何不同点和共同点；
 - (5) 不同时间段（例如月头、月中、月末等）产品需求量有何特性；
 - (6) 节假日对产品需求量的影响；
 - (7) 促销（如 618、双十一等）对产品需求量的影响；
 - (8) 季节因素对产品需求量的影响。
2. 基于上述分析，建立数学模型，对附件预测数据（predict_sku1.csv）中给出的产品，预测未来 3 月（即 2019 年 1 月、2 月、3 月）的月需求量，将预测结果按照表 3 的格式保存为文件 result1.xlsx，与论文一起提交。请分别按天、周、月的时间粒度进行预测，试分析不同的预测粒度对预测精度会产生什么样的影响。

表 3：预测结果的数据样例

sales_region_code	item_code	2019 年 1 月预测需求量	2019 年 2 月预测需求量	2019 年 3 月预测需求量
101	20002	1	2	3
101	20003	2	3	4
⋮	⋮	⋮	⋮	⋮