2021 年 MathorCup 高校数学建模挑战赛——大数据竞赛

赛道 A：二手车估价问题

随着我国的机动车数量不断增长，人均保有量也随之增加，机动车以“二手车”形式在流通环节，包括二手车收车、二手车拍卖、二手车零售、二手车置换等环节的流通需求越来越大。二手车作为一种特殊的“电商商品”，因为其“一车一况”的特性比一般电商商品的交易要复杂得多，究其原因是二手车价格难于准确估计和设定，不但受到车本身基础配置，如品牌、车系、动力等的影响，还受到车况如行驶里程、车身受损和维修情况等的影响，甚至新车价格的变化也会对二手车价格带来作用。目前国家并没有出台一个评判二手车资产价值的标准。一些二手车交易平台和二手车第三方估价平台都从自身的角度建立了一系列估价方法用于评估二手车资产的价值。

在一个典型的二手车零售场景，二手车一般通过互联网等线上渠道获取用户线索，线下实体门店对外展销和售卖，俗称 O2O 门店模式。门店通过“买手”从个人或其他渠道收购二手车，然后由门店定价师定价销售，二手车商品和其他商品一样，如果定价太高滞销也会打折促销，甚至直接以较低的价格打包批发，直至商品最终卖出。

基于以上背景，请你们团队根据附件给出的数据，通过数据分析与建模的方法帮助二手车交易平台解决下面的问题：

初赛问题

问题 1：基于给定的二手车交易样本数据（附件 1：估价训练数据），选用合适的估价方法，构建模型，预测二手车的零售交易价格，数据中会对 id 类，主要特征类等信息进行脱敏。主要数据包括车辆基础信息、交易时间信息、价格信息等，包含 36 列变量信息，其中 15 列为匿名变量。字段如下：

序号	Features	Description
1	carid	车辆 id
2	tradeTime	展销时间
3	brand	品牌 id
4	serial	车系 id
5	model	车型 id
6	mileage	里程
7	color	车辆颜色
8	cityId	车辆所在城市 id
9	carCode	国标码
10	transferCount	过户次数
11	seatings	载客人数
12	registerDate	注册日期
13	licenseDate	上牌日期
14	country	国别
15	maketype	厂商类型
16	modelyear	年款
17	displacement	排量
18	gearbox	变速箱
19	oiltype	燃油类型
20	newprice	新车价
21	anonymousFeature	15 个匿名特征
22	price	二手车交易价格（预测目标）

请采用附件 1 中的“估价训练数据”（带标签）训练模型和测试模型，自行设置测试集，使用训练完成后的模型对附件 2 中的“估价验证数据”（不带标签）进行预测，并将预测结果保存在附件 3“估价模型结果”文件中，注意不要修改格式，单独上传到竞赛平台。

其中附件 1“估价训练数据”和附件 2“估价验证数据”只相差最后 1 列数据（二手车交易价格（预测目标）），附件 3“估价模型结果”文件字段如下：

车辆 id	预测价格
id1	预测价格 1
id2	预测价格 2

附件 1、附件 2、附件 3 中各字段间采用“\t”分隔符分割，不包含表头。

模型评测标准：

$$0.2*(1-Mape)+0.8*Accuracy_5$$

Ape(相对误差)：

$$Ape=\left|\hat{y}-y\right|\Big/y$$

Mape(平均相对误差)：

$$Mape=\frac{1}{m}\sum_{i=1}^{m}Ape_i$$

其中，真实值 $y=(y_1,y_2,\cdots,y_m)$，模型预测值为 $\hat{y}=(\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_m)$；

Accuracy$_5$(5%误差准确率):

$$Accuracy_5=count(Ape<=0.05)/count(total)$$

其中，$count(Ape<=0.05)$为相对误差 *Ape* 在 5%以内的样本数量，$count(total)$为样本总数量。

问题 2：在门店模式中，车辆在被“买手”收车以后，会进入门店进行售卖，车辆能否成功交易，除了取决于销售的谈判技巧，更重要的是车辆本身是否受消费者青睐，价格是否公道。假设你们是门店的定价师，请你们结合附件 4“门店交易训练数据”对车辆的成交周期（从车辆上架到成交的时间长度，单位：天）进行分析，挖掘影响车辆成交周期的关键因素。假如需要加快门店在库车辆的销售速度，你们可以结合这些关键因素采取哪些行之有效的手段，并进一步说明这些手段的适用条件和预期效果。

附件 4“门店交易训练数据”包括 6 个字段，如下表所示，其中所有 carid 等相关信息包含在附件 1“估价训练数据”中。各字段间采用“\t”分隔符分割，不包含表头。

序号	Features	Description
1	carid	车辆 id
2	pushDate	上架时间
3	pushPrice	上架价格
4	updatePriceTimeJson	{价格调整时间：调整后价格}
5	pullDate	下架时间(成交车辆下架时间和成交时间相同)
6	withdrawDate	成交时间

问题 3：依据给出的样本数据集，你们觉得还有哪些问题值得研究，并给出你们的思路？

将问题 1、2、3 的解决过程写成一篇论文，明确你们的思路、模型、方法和结果。

附件

附件 1：估价训练数据.txt；

附件 2：估价验证数据.txt；

附件 3：估价模型结果.txt；

附件 4：门店交易训练数据.txt。