种子杯竞赛结题报告

——京东物流次日达

一.数据和特征处理：

1.通过基本的数据观察，我们剔除了收货时间过短和收货时间比送货时间早的数据。

2.由于问题的特殊性，仓库、发货省份、物流公司在对测试数据的实际预测中均为不可知变量，我们将测试数据进行了数据的增广。具体方法为，在已知的训练数据中，综合对现实情况的考量，我们利用树形结构，构造了以下三种索引查值表：

①商家id-收货地址-物流仓库

②商家id-商品目类-发货省份

③商家id-商品目类-物流公司

在三种对应关系中，绝大多数的所以满足前两个因素有且仅对应一种第三个因素，仅有少数存在二对多的情况，在这里我们采用了概率进行统计，选择概率最大的作为此处取值。

3.训练数据中有多种时间的考量（包括但不限于预售时间，下单时间，出仓时间，送件时间）。在观察数据的过程中，发现了大量预售时间为1970年，在后续操作中我们默认此类物品不存在预售时间。也存在相当数量的发货时间比预售时间早的物件。在后续操作中，我们进行了两种情况的考量训练——与预售时间相关（只将发货时间不早于预售时间的物件的预售时间记为有效预售时间，对于其他的物件我们则将其预售时间设置为无，i.e.与订单签单时间同）和与预售时间无关（不考虑预售时间而只考虑下单时间和签收时间）。

二.规则决策

在数据预处理过程中，我们对签收时间进行了概率统计，并储存为pkl文件详见相应文件。在凌晨至8am间几乎无物件签收，于是我们制定以下规则：

①对于凌晨签收的问题数据我们依据训练数据概率分布随机修改为签收时间值。

②对于每一条数据我们都有一定的概率阈值会按照训练数据的概率分布进行修改签收时间的小时数进行调整。

③我们设置一个较大的概率阈值进行天数的修改，从减两天至加两天不等。概率分布依据训练数据的预测与真实值之差的概率分布而定。

以上概率的产生均符合正太分布。

三.模型选择：

赛程期间，我们主要采取了三条模型选择主线：树形结构，Baseline 全连接扩展，Baseline With Convolution。

树形结构：(机器学习)(初期)

1.基于树模型的Adaboost，RandomForest，Xgboost，Extraboost,经过适当的一定时间的调参过程，Extaboost效果最佳，榜单提交了有效成绩为rankScore 64，OnTimePercent 99.8%。随后基于时间的规则调整处理，达到了rankScore 60。

2.Baseline全连接扩展(中期)

赛方发布Baseline后我们针对Baseline进行了调参，rankScore可以降低到60以下，但OnTimePercent始终无法达到98%的提交要求。针对于此，我们进行了Embedding引入我们数据增广后的特征，全连接层加深，在Embedding层后加入9层全连接层帮助拟合。同时辅以以下操作：

①损失函数调整：

α.调整日期惩罚与小时预测间的惩罚比例（修改天数惩罚的惩罚系数），但仍然尽量往预测日期提前的方向去调整。

β.修改损失函数的计算方式，baseline对每个epoch指定的损失函数的计算规则为：如果当前epoch对时间的预测平均大于真实值，则加大惩罚能力。针对于此，我们认为会存在提前与延迟近似对半的情况，这对于模型的训练是没有收益的。我们对此进行了修正，对于每个epoch内的每条训练数据，如果当前数据存在延迟则针对当前数据的残差进行放大处理，正常或提前数据则仍按正常流程操作，i.e.换平均为个体。对每个epoch内的所有数据进行同样的处理。

γ.修改损失函数:分别利用平方损失函数，均方根损失函数，交叉熵损失函数训练。

②学习率调整：

在调参过程中，我们对多种初始的学习率和学习率衰减因子进行了实验。

③预测方式的调整试验：

将原本统一整体的网络设置为一个单独预测天和小时的二分支网络。并将day的loss权重乘以24

3.Baseline With Convolution(后期)

我们希望网络可以挖掘出更多的数据的深层特征，于是自然引入了卷积算子。我们在全连接模型中将特征转化为13\*100的Embedding特征图。为了方便卷积处理，我们利用全连接层进行特征信息融合并reshape成30\*30的特征图（i.e.13,100->reshape to 13\*100->FC 900->reshape30,30）。在卷积层提取特征图后继续辅以多层FC进行拟合，损失函数和学习了吧的调整同上。模型rankScore稳定50以下OnTImePercent常在98%附近波动，rankScore一度达到38左右。但OnTimePercent始终在97~98之间，无法达到准入门槛。

针对于此，我们引入了模型融合的策略，将全连接扩展后表现较好的模型（准时率极优，rank欠佳）与引入卷积的模型进行答案的加权平均决策，权衡天数与小时，辅以规则决策，模型达到了rank 44 左右的成绩。

赛程后期我们考虑引入残差网络和多尺度卷积模块，限于时间和计算机算力的原因，代码未能在比赛日期之前训练多次,因此综合下来效果没有现有模型好.