阿里天池智慧交通预测挑战赛 赛后总结（Top3%）

先前有做过几次阿里天池的比赛，不过成绩都不是很好，无意中翻看自己先前做过的比赛，想着继续参与一个比赛来练手，因此在临近初赛结束的前一周报名参加了该比赛。由于时间仓促，开始并没有什么思路，尝试寻找一些统计特征，可是效果均不是很好，一直到初赛结束前一天，排名依然无法进入前100。这时，我感觉已是无望，不过，心想还有一天，只要尽人事听天命就好。抱着这样的心态，我整整花了一天的时间，不断修改特征，提取特征，最终总算跑了一遍模型，并顺利提交了成绩。结果也没有令我失望，顺利排进了前100名，进入的复赛。

在复赛过程，我的模型调整不是很大，主要是因为复赛要求在odps和阿里云机器学习平台上完成，而此时阿里天池多个比赛同时开赛，部分比赛占用大量的计算资源，导致提交的执行代码长时间无法运行，比赛自然也无法顺利进行。最后，经过多轮讨论后决定在线下进行。而此时复赛赛程已过半，我一个人单枪匹马，自然心有余而力不足。下面，我从特征工程、模型选择和结束语三个方面总结本次比赛。

1. 特征工程
2. 赛题数据及目标

比赛提供了三份原始数据，分别是路段（link）属性表、 link上下游关系表和link历史通行时间表。具体数据表格式参看官方链接。

本次比赛的目标即预测2017年7月1日到2017年7月31日上午8点至9点、下午3点至4点以及晚上6点到7点，每两分钟平均车辆通行时间。

1. 特征

我建立的模型特征主要包括三个方面：当天已有数据统计特征、历史统计特征、路段特征和日期特征。

1. 当天已有数据统计特征

构建这类特征时，我主要选择的预测时段前的一个小时的数据，通过统计过去一个小时内、20分钟内该路段的平均通行时间、众数、中位数、标准差、最大值和最小值。另外还有使用one-hot编码的时间相对于每个预测时间段起点的时间间隔数。

1. 历史统计特征

历史的统计特征，主要选取可获得的数据中同一时间间隔内的平均时间、众数、中位数、标准差、最大值和最小值，以及7天前同一时段和15天前同一时段的通行时间（其实选择14天前的数据更加合适）。

1. 路段特征

官方提供的数据集不仅仅含有各路段全天的通行时间记录，还提供每条路段的信息（路段的车道数量、入度路段以及出度路段）。因此可以构建起每一条路段的车道数量、上游汇入的路段数量、下游分出的路段数量。

1. 日期特征

这个特征主要使用one-hot编码表示每个日期所处的星期信息，用以区分工作日和非工作日的特性。

在进行特征提取过程中，由于收集的数据存在不完整性，但数据记录的频率很高（每两分钟采集一次数据）， 因此在处理缺失值时，主要是将数据进行按路段、时间排序后，采用后向填充的方法补全缺失数值。

1. 模型选择

基于已有的特征，我将这个题目看成一个回归问题来处理。有的人可能会使用时间序列的方式进行处理。但是，考虑到路段数量大，历史数据多，因此如果对每一条路径都进行单独的时间序列分析，不仅计算量特别大，而且每条路径的预测都是独立的。在模型选择上，当然首选xgboost。前期，我在本地使用多gbdt，但是速度极慢。

1. 结束语

本次的比赛，我没有使用其他更加复杂的模型。我觉得并给模型越复杂越好，相反，模型效果的好坏更多的取决于特征工程的质量。当然，在交流中，有选手使用深度学习的seq2seq方法来构建模型也是一个很好的思路。