提分Trick

交叉验证

事实证明，交叉验证往往能取得更好的效果，我们是怎么进行交叉验证的呢？

- 交叉验证不需要额外划分验证集和训练集，直接拿整个原始数据集来做就好
- 将训练集依次划分为五份，每次取其中的四份来训练，一份作为验证，并将测试集模型直接通过该次模型输出
- 取多次输出平均值作为测试集模型输出

特征选择

某些特征可能导致过拟合，根据特征的重要度、相似性或者一些方法去除非必要的特征。简单的方法比如Null Import就是将特征与正确的标签训练，特征与打乱的标签训练，看看他们之间的重要度有没有变化。

对数变换

可以看到，我们的`distance`，`distance_have`，`duration`差距过大，可通过对数变化方便其他方法运行。对于树模型其实并不太需要对数变化就是。

我们采用交叉验证的LGBM得分达到了0.36654，挺不错的了

模型融合

线性加权

模式	得分
1111	0.36581
3322	0.36563
4321	0.36546
5221	0.36560

最好得分为 37/1206

Stacking

模型一：全量特征训练，采用特征为：

```
columns=[ 'vendor_id','distance','duration_log',
'distance_have','ave_speed',
         'distance_type',
         'motorway', 'trunk', 'primary',
       'secondary', 'tertiary', 'unclassified', 'residential',
       'nTrafficSignals', 'nCrossing', 'nStop', 'nIntersection',
       'pc', 'pickup_longitude', 'pickup_latitude',
       'dropoff_longitude', 'dropoff_latitude', 'store_and_fwd_flag',
       'month', 'weekday',
          'hour', 'minute',
       'real_hour', 'rush_hour', 'is_weekend', 'year', 'day',
          'humidity', 'pressure', 'temperature',
        'wind_speed',
       'direction_bins', 'maximum temperature', 'minimum temperature',
       'average temperature', 'precipitation', 'snow fall',
         'snow depth',
         'in_Cluster','out_Cluster',
         ]
```

模型二：构造特征训练，采用特征为：

```
columns=[ 'distance','duration_log','ave_speed',
         'distance_type',
         'motorway', 'trunk', 'primary',
       'secondary', 'tertiary', 'unclassified', 'residential',
       'nTrafficSignals', 'nCrossing', 'nStop', 'nIntersection',
       'pc',
       'day_of_year','weekday',
          'hour', 'minute',
       'real_hour', 'rush_hour', 'is_weekend',
          'humidity', 'pressure', 'temperature',
        'wind_speed',
       'direction_bins',  'precipitation', 'snow fall',
         'snow depth','month','day',
         'speed','trip_duration',
         ]
```

这两个模型均采用LightGBM，其得分分别为：

模型	得分
1	0.36597
2	0.39763

第二个模型比较蠢，我们不能在剔除聚类特征的同时剔除经纬度信息，因而还是保留经纬度信息。

```
columns=[ 'distance','duration_log','ave_speed',
          'distance_type',
          'motorway', 'trunk', 'primary',
        'secondary', 'tertiary', 'unclassified', 'residential',
        'nTrafficSignals', 'nCrossing', 'nStop', 'nIntersection',
        'pc',
        'day_of_year','weekday',
           'hour', 'minute',
        'real_hour', 'rush_hour', 'is_weekend',
           'humidity', 'pressure', 'temperature',
         'wind_speed',
        'direction_bins',  'precipitation', 'snow fall',
          'snow depth','month','day','pickup_longitude', 'pickup_latitude',
        'dropoff_longitude', 'dropoff_latitude',
'store_and_fwd_flag','vendor_id',
          'speed','trip_duration',
          ]
```

结果表明，聚类特征确实有些用处，当然也不知道是不是这个温度数据的影响。

模型	得分
1	0.36597
2	0.37052

模型三：针对不同路程分级构造模型

我们注意到不同路程的数据差异巨大，所以这里采用不同的方法处理。
针对正常路段(0.2,+inf)，采用一个模型训练

针对异常路段，调整参数采用模型训练

事实证明，在0,0.2上怎么也训练不好，差距太大了，所以这部分直接用前面比较不错的模型。

该模型的得分为：

模型	得分
3	0.36597

模型四： 全量参数的XGBoost训练

需要很久的时间，所以最后做。

该模型的得分为：

Trick

注意到有很多异常值，所以我们再构建一些模型：

模型五：全量模型LightGBM
模型六：去除异常值的LightGBM
模型七：二分类模型，预测是否是异常值

最终的结果就可以是：

$$k1 = p * 异常值均值 + (1-p) * 非异常值模型结果$$

$$k2 = 0.5 * k1 + 0.5 * 全量模型$$

同时，对比之前的Stacking模型：

$$k3 = 0.5 * k1 + 0.5 * stacking$$

或者是：

$$k4 = p * 异常值均值 + (1-p) * stacking$$

最终的结果为：

模型	得分
k1	
k2	
k3	
k4	

至此，我们的技巧部分就说完了，想要再提分可能就要更多的构建特征工程了，这部分就不做了，已经花费两个星期多了。