|  |
| --- |
| **机器学习** |
| **ECML/PKDD 15: 出租车行程时间预测** |
| 主 研 人：杜思君  参 研 人： |
| 审 核 人：  方 向：工业算法案例研究  版 本 号：A |
|  |

**版本更新**

|  |  |  |  |
| --- | --- | --- | --- |
| **日期** | **版本** | **更新描述** | **作者** |
| 2018/10/31 | A | 初稿 | 杜思君 |

**选题表格**

|  |  |  |  |
| --- | --- | --- | --- |
| **时间** | **竞赛名** | **竞赛背景描述（50字以内）** | **类型（分类/回归）** |
| 2019/3/15 | ECML/PKDD 15: Taxi Trip Time Prediction (II) | 为了提高城市内电子出租车调度系统的效率，能够预测驾驶员将其出租车占用多长时间是很重要的。根据出租车当前（初始部分的）运行轨迹，预测在葡萄牙波尔图市的出租车行程时间。 | 回归 |

目录

[1. 背景描述 4](#_Toc25575_WPSOffice_Level1)

[1.1 竞赛赛题描述 4](#_Toc5940_WPSOffice_Level2)

[1.2 评估指标描述 5](#_Toc12357_WPSOffice_Level2)

[2. 数据来源及描述性统计分析 5](#_Toc5940_WPSOffice_Level1)

[2.1 大赛数据来源 5](#_Toc22683_WPSOffice_Level2)

[2.2 数据的描述性统计 5](#_Toc16174_WPSOffice_Level2)

[2.2.1 数据基本情况描述： 5](#_Toc5940_WPSOffice_Level3)

[2.2.2 数据字段介绍： 5](#_Toc12357_WPSOffice_Level3)

[2.2.3 数据描述性统计 6](#_Toc22683_WPSOffice_Level3)

[3. 优秀算法思路 6](#_Toc12357_WPSOffice_Level1)

[3.1 方案一 6](#_Toc32293_WPSOffice_Level2)

[3.1.1 方案一数据预处理及特征工程部分方案 6](#_Toc16174_WPSOffice_Level3)

[3.1.2 方案一模型设计、建立部分方案 6](#_Toc32293_WPSOffice_Level3)

[3.1.3 方案一结果、排名等 6](#_Toc29922_WPSOffice_Level3)

[3.1.4 方案一算法流程图 6](#_Toc4787_WPSOffice_Level3)

[3.2 方案二 7](#_Toc29922_WPSOffice_Level2)

[… 8](#_Toc4787_WPSOffice_Level2)

[4. 算法比较 8](#_Toc22683_WPSOffice_Level1)

[表4-1 算法比较 8](#_Toc7253_WPSOffice_Level2)

[5. 总结与展望 8](#_Toc16174_WPSOffice_Level1)

[5.1 总结 8](#_Toc23198_WPSOffice_Level2)

[5.2 建模思路 8](#_Toc16619_WPSOffice_Level2)

1. 背景描述

本次比赛由ECML / PKDD 2015组织

为了提高电子出租车调度系统的效率，能够预测出租车驾驶员该次行程将花费多长时间是很重要的。如果系统调度员能够大致知道出租车司机能在何时将结束他们当前的行程，就能够更好地识别乘客的乘车请求并合理的分配给每个司机，从而减少双方等待的时间，也能更好的利用交通资源。

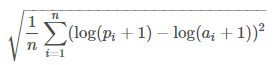
**1.1 竞赛赛题描述**

在本次比赛中，主办方要求参赛者建立一个预测框架，能够根据出租车一次行程的（初始）部分轨迹推断出该车在葡萄牙波尔图的出行时间。该框架的输出必须是特定出租车行程所消耗的时间。

**1.2 评估指标描述**

竞赛中评估模型优劣的指标。

选手提交结果与实际出租车出行时间作对比，以对数均方根误差(Root Mean Squared Logarithmic Error(RMSLE))为评价指标，结果越小越好，计算公式如下：



其中，n为预测总数，pi与ai 为对应的预测值和真实值，log为对数符号。

1. 数据来源及描述性统计分析

**2.1 大赛数据来源**

本次比赛的数据源自葡萄牙波尔图市的出租车调度系统，是该市出租车运行的真实纪录。

（数据源自kaggle：<https://www.kaggle.com/c/pkdd-15-taxi-trip-> time-prediction-ii/data）

以下是数据的超链接：

<train.csv>训练集 <test.csv> 测试集 <sampleSubmission.csv>提交样例

附加文件：

<metaData_taxistandsID_name_GPSlocation.csv>出租车站台名和位置。

<evaluation_script.r> 评估指标的计算文件（R语言）

**2.2 数据的描述性统计**

**2.2.1 数据基本情况描述：**

赛事主办方提供了一个准确的数据集（训练集），描述了葡萄牙波尔图市所有442辆出租车的轨迹（从2013年7月1日至2014年6月30日）。 这些出租车从出租车调度中心运行，使用安装在车辆中的移动数据终端进行纪录。 我们将每次乘车分为三类：A）通过在出租车调度中心乘车B）在出租车站台乘车C）在非出租车站台乘车。 第一类（A）是指调度中心提供匿名身份证，以便通过电话提供此类信息给乘车的乘客。 最后两类是指在 B）在出租车站台上，或者是C）随机在街道上直接向出租车司机提出的乘车请求。  
 每个数据样本对应一个完整的出租车出行旅程。

此外，赛事主办方还提供了五个测试集来评估参赛者的预测框架，这些数据同样是来着葡萄牙波尔图市的出租车调度中心（从2014年7月1日至2017 年12月31日）对出租车行程纪录。这些数据集中的每一个都将提供给定时间戳上当前网络状态的快照。 它将为该特定时刻的每个正在进行的行程提供部分轨迹。

**2.2.2 数据字段介绍：**

|  |  |  |  |
| --- | --- | --- | --- |
|  | 变量含义 | 变量类型 | 缺失率 |
| **TRIP\_ID** | **行程的唯一标识符** | 分类（String） |  |
| **CALL\_TYPE** | **乘客乘车方式（A,B,C）** | 分类（char） |  |
| **ORIGIN\_CALL** | **电话号码的标识符**  **当CALL\_TYPE=A才生效，其余为NULL值** | 分类（integer） |  |
| **ORIGIN\_STAND** | **出租车站的标识符**  **当CALL\_TYPE=B才生效，其余为NULL值** | 分类（integer） |  |
| **TAXI\_ID** | **每次行程出租车的唯一标识符** | 分类（integer） |  |
| **TIMSTAMP** | **行程开始的时间** | 离散（integer） |  |
| **DAYTYPE** | **日期类型（A,B,C）**  **B为假期或特殊日期，C为B型日期前一天，A为正常日期（即工作日）** | 分类（char） |  |
| **MISSING\_DATA** | **GPS数据完整时为True，有一个或多个缺失值为False** | 分类（Bool） |  |
| **POLYLINE** | **GPS坐标列表**  **每个坐标点对应15s（time=（points - 1）\*15）** | （String） |  |

**Table 2.2.2-1: 训练集字段描述**

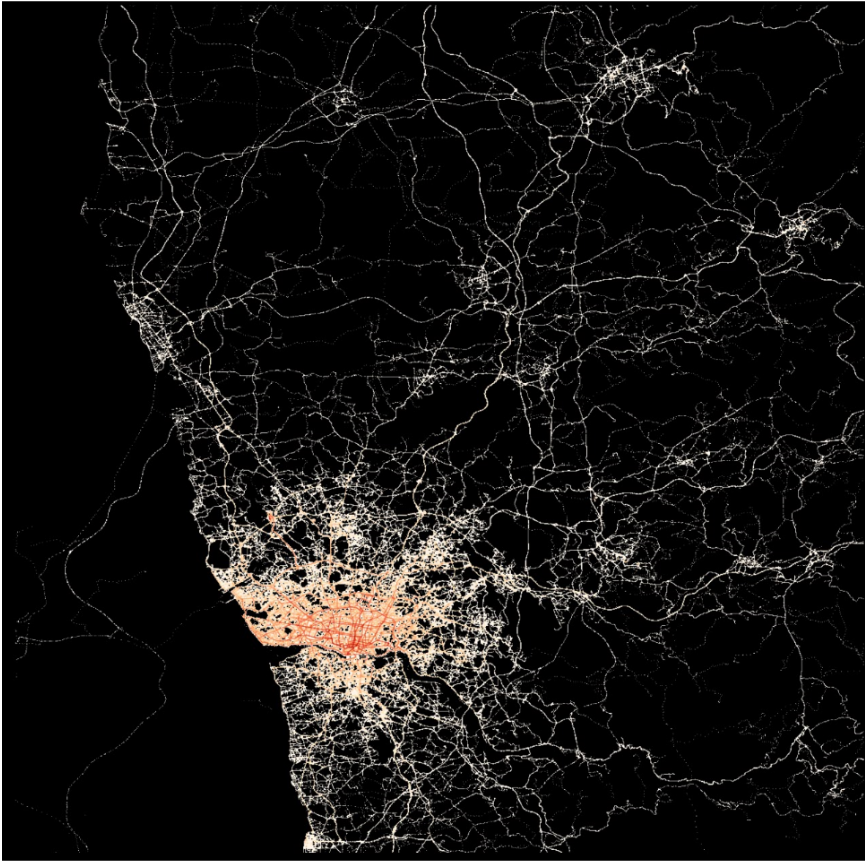
|  |  |  |  |
| --- | --- | --- | --- |
|  | 变量含义 | 变量类型 | 缺失率 |
| **TRIP\_ID** | **行程的唯一标识符** | 分类（String） | 0% |
| **CALL\_TYPE** | **乘客乘车方式（A,B,C）** | 分类（char） | 1个 |
| **ORIGIN\_CALL** | **电话号码的标识符**  **当CALL\_TYPE=A才生效，其余为NULL值** | 分类（integer） | 1个 |
| **ORIGIN\_STAND** | **出租车站的标识符**  **当CALL\_TYPE=B才生效，其余为NULL值** | 分类（integer） | 1个 |
| **TAXI\_ID** | **每次行程出租车的唯一标识符** | 分类（integer） | 1个 |
| **TIMSTAMP** | **行程开始的时间** | 离散（integer） | 1个 |
| **DAYTYPE** | **日期类型（A,B,C）**  **B为假期或特殊日期，C为B型日期前一天，A为正常日期（即工作日）** | 分类（char） | 1个 |
| **MISSING\_DATA** | **GPS数据完整时为True，有一个或多个缺失值为False** | 分类（Bool） | 1个 |
| **POLYLINE** | **GPS坐标列表**  **每个坐标点对应15s（time=（points - 1）\*15）** | （String） | 1个 |

**Table 2.2.2-2: 测试集字段描述**

|  |  |  |  |
| --- | --- | --- | --- |
|  | 变量含义 | 变量类型 | 缺失率 |
| **ID** | **ID列** | 连续（Numeric） | 0% |
| **Descricao** | **出租车站名（地点）** | 分类（String） | 0% |
| **Latitude** | **纬度** | 离散（String） | 0% |
| **Longitude** | **经度** | 离散（Numeric） | 2% |

**Table 2.2.2-3 : 附加文件-出租车站名和位置的字段描述**

**2.2.3 数据描述性统计**

1.波尔图市出租车行程轨迹可视化

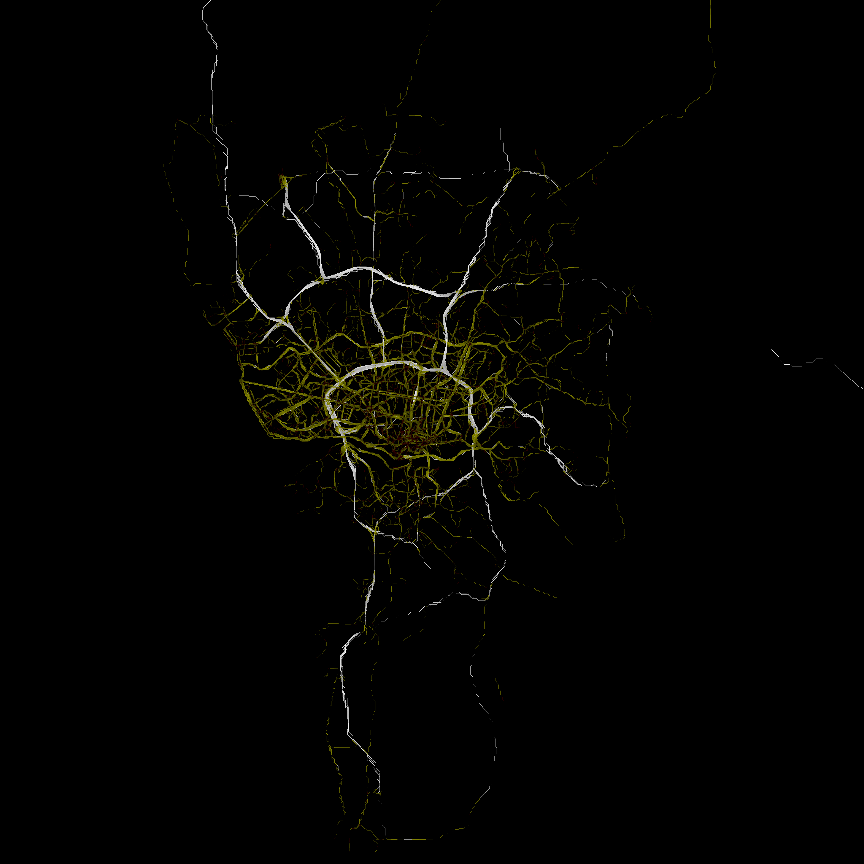
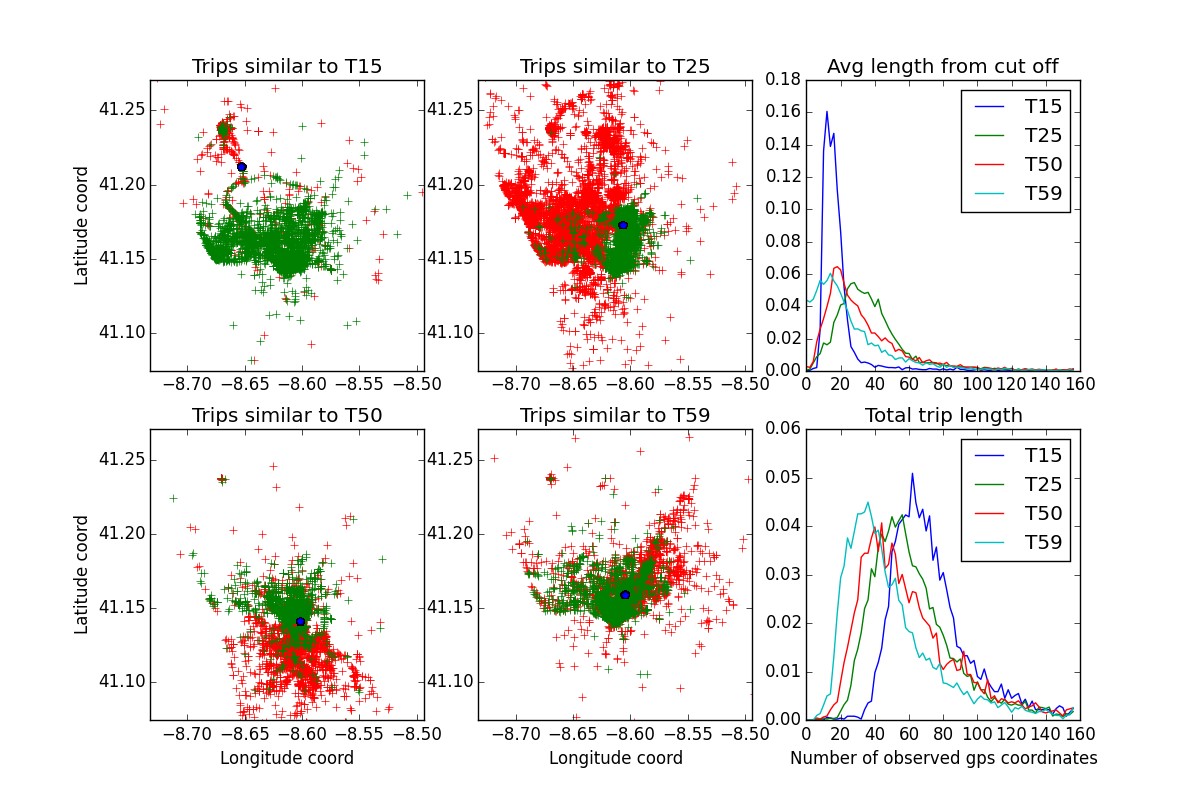


Figure2.2.3-1: 波尔图市出租车轨迹可视化（训练集中）

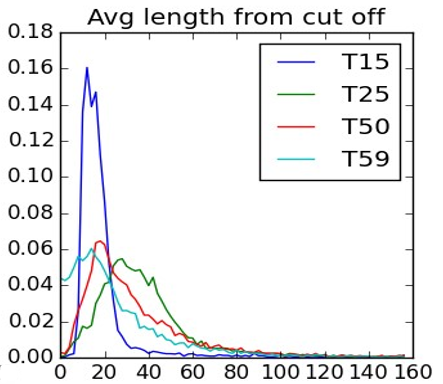
2. 训练集中行程起始位置的大致分布



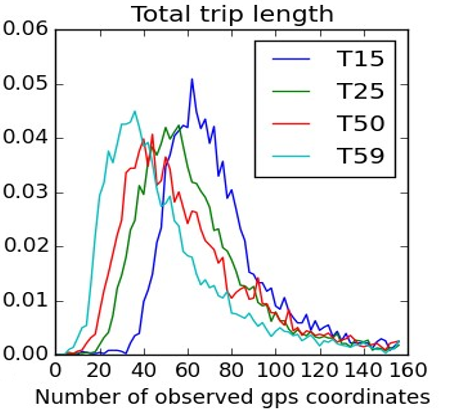
**Figure2.2.3-2：训练集中行程起始-终点位置的分布**

**起点：绿色；终点：红色（PS:行程相近）**

1. 行程的长度分布



**Figure2.2.3-3:测试行程的平均长度（从截止点）**



**Figure2.2.3-4: 行程总长度与剩余坐标点个数分布图**

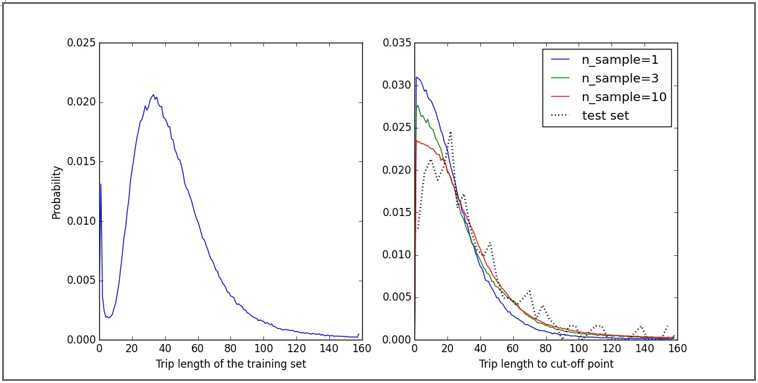
3. 优秀算法思路

**3.1 方案一**

**3.1.1 方案一数据预处理及特征工程部分方案**

训练集中包含了许多非常短的行程，如***figure3.1.1-1***的左图所示，其显示了总行程长度的概率曲线图。不难发现，低于4的高比例行程不遵循分布的一般类型, 因此选手将这些行程排除在分析之外。此外，还发现另一种错误是误读了GPS坐标，选手通过切断非常长距离的行程来排除这些错误。对于最终位置的预测，选手将阈值设置为99％的行程长度；而对于行程时间预测，阈值设置为99.9％。

剩余的行程用于​​生成模型的特定训练集。只有基本模型的训练集包含所有行程。对于这个数据集，每次行程都是随机切断的。由于测试集是在特定时间点收集的，因此更有可能包括更长的行程。如***Figure3.1.1-1***所示，对于较长的行程，采样频率随着行程的长度线性增加。这样可以减少短行程的频率，而且得到的分布更接近测试集的分布（黑色虚线）。



**Figure3.1.1-1:**

**Left: 训练集中行程长度的分布 Right：采样后训练集的行程长度**

**（对较长时间的行程进行更高频的采样减少了训练集中短途行程的数量，并且使得得到的分布根类似于测试集的分布（黑色虚线））**

最终的训练集包括以下特征：

* 工作日（周一至周日）
* 行程开始的时间
* 行程的长度
* 行程起点的纬度坐标
* 行程起点的经度坐标
* 纬度坐标截止点
* 经度坐标截止点
* 起始点到市中心的距离
* （从起点）前往市中心
* 从市中心到行程截止点的距离
* 从市中心前往（行程截止点）
* 到截止点的净距离（直线距离）
* 出租车在整个行程的中间位置的速度
* 出租车在行程截止点的速度
* 在截止点出租车的驶向

有趣的是，根据该选手的测试，大多数原始数据几乎没有预测能力

最终该选手较多的使用了时间戳数据（Time stamp）

**3.1.2 方案一模型设计、建立部分方案**

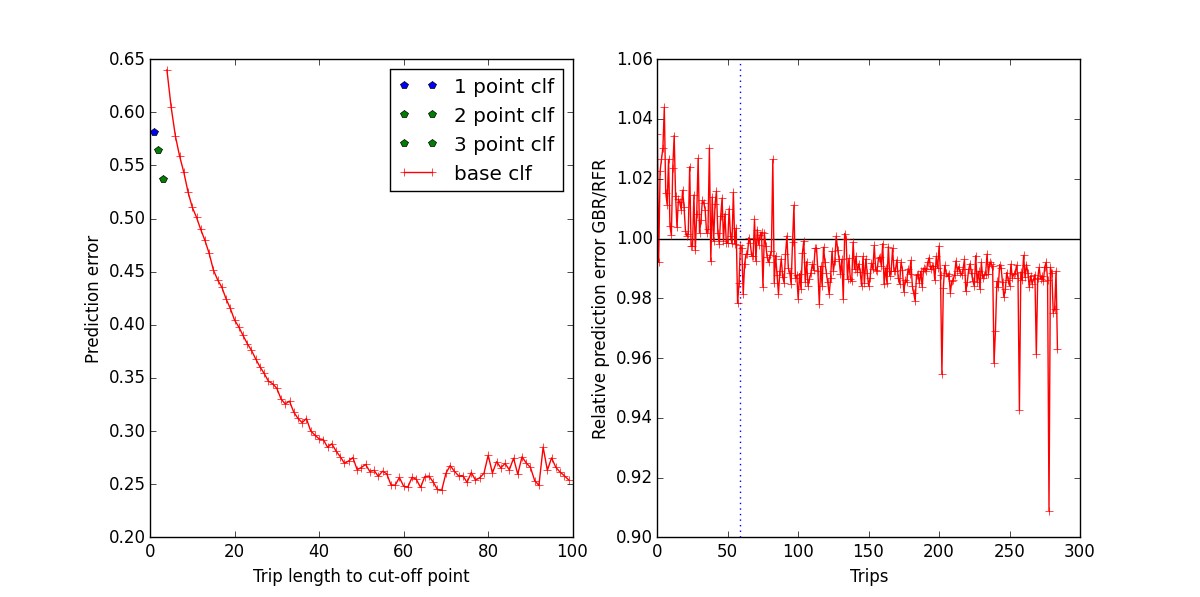
1） 模型建立

该选手最终决定为测试数据集中的每个行程训练一个单独的模型，并回归到更一般的模型。由此训练了以下三个模型：

* + - **基础模型**：基于数据集，其中从训练集中的所有轨迹中提取特征，并且比较短轨迹更频繁地采样较长轨迹。
    - **短途行程专家模型：**（例如：已知1,2或3个位置点）。
    - **针对每个测试行程的专家模型**（例如：对在行程截止位置穿过测试行程的道路进行训练）

2） 模型训练

该选手在在所有的模型中都使用了5折交叉折叠方法进行了训练。选手使用了sklearn库里的随机森林回归（RandomForestRegressor（RFR））以及梯度提升回归（GrandientBoostingRegressor（GBR））两种算法，除将数的数量设置为200以外，其他参数均采用默认参数。



**Figure3.1.2-1:**

**Left: 基本模型与行程长度的交叉验证预测误差 Right: GBR与RFR的预测误差之比**

如上左图显示了基本模型与行程长度的交叉验证预测误差，而短途旅行专家模型（蓝色/绿色）的预测误差相比基本模型要低得多。右图展示了随着训练集的增加，GBR与RFR两种算法的预测误差之比的变化；由图不难看出，当训练集的大小超过4000（如图中蓝线所示）时，GBR的预测效果略好于RFR。

***Figure3.1.12-1***左侧显示了RandomForestRegressor算法的交叉折叠预测误差与基本模型的行程长度（红色）的关系曲线。 由于预测行程时间中使用了对数变换，使得预测误差快速降低到0.3以下。 因此，在预测出租车行程时间中，优化预测误差尤其是短途行程的误差至关重要。 训练一个较好的短途行程专家（一个（蓝色），两个或三个（绿色）如图所示）明显优于基本模型。 使用RandomForestRegressor和GradientboostingRegressor训练行程截止位置的短途行程专家模型。 训练集大小从几十到十万不等。模型仅针对那些训练集大小超过1000的测试行程进行训练。有趣的是，随着训练集大小的增加，GradientBoostingRegressor算法的CV预测误差略好于 RandomForest,如***figure3.1.2-1***右图所示。

（PS:文中所说的短途行程专家模型，是指该选手针对较短的行程而设计的模型，选手起名为短途行程专家）

3） 模型融和

由于大赛允许参赛者可以选择两个结果提交进行评估，因此该选手提交了两种结果，如下：

1. 提交1：

提交结果1的模型是基础模型与短途行程专家模型的混合，其中短途旅行专家仅基于行程截至位置，将行程长度的阈值设置为15。

2 ) 提交2：

提交2的模型基于提交1的模型。在此之外，预测被替换为具有足够训练集大小的所有行程的行程专家模型，使用1000作为阈值。

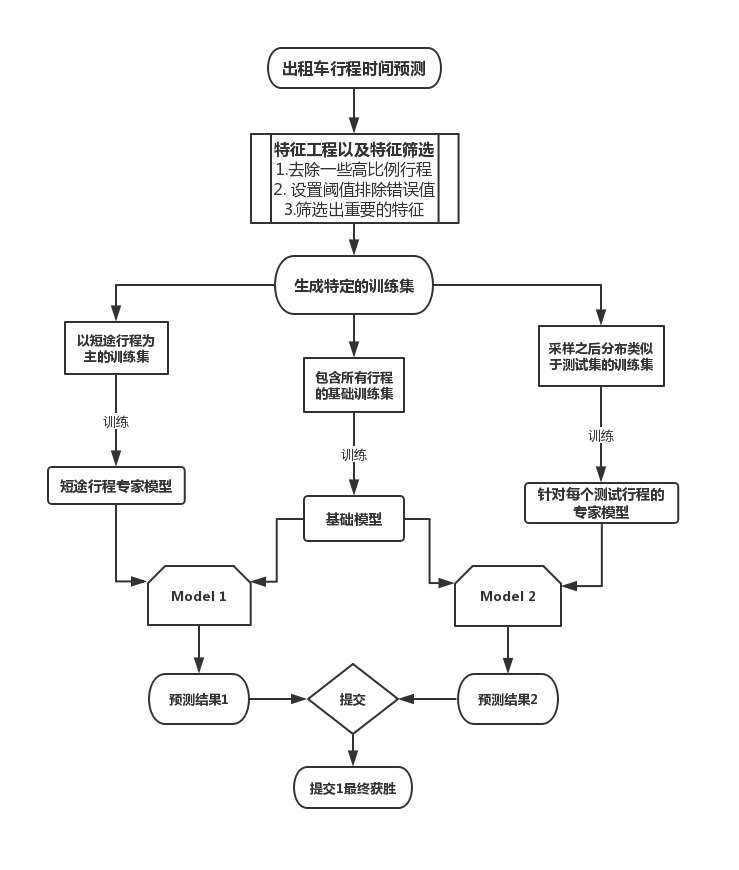
最终，提交1在私人排行榜（Private Leaderboard）中的得分高于提交2（0.5092对0.5045），而对于公共排行榜(Public Leaderboard)则相反（0.5253对0.5354）。最终获胜的结果是提交1（0.5253）

**3.1.3 方案一结果、排名等**

**结果：(RMSLE:0.52528)**

**排名：（Rank：1/345）**

**3.1.4 方案一算法流程图**



**3.2 方案二**

由于此比赛参赛人数较少，分享内容不多，仅仅找到了top1选手的简单总结，没有其他top10的选手分享。

4. 算法比较

|  | **评估指标** | **特征工程** | **基础算法** | **基本库** |
| --- | --- | --- | --- | --- |
| **算法1** | **RMSLE** |  | **RFR, GBR** | **Sklearn** |

**Table4-1:算法比较**

5. 总结与展望

**5.1 总结**

本题是一道涉及到图运算和拓扑学的动态规划（DP）问题；回归问题。给人的第一感觉较为复杂，比一般的回归问题要难分析和处理，而Top1选手仔细的筛选和探索，去除掉了许多无用特征，而且根据需求构造了三个不同的训练集用于训练三个不同但各有特点的模型，构思巧妙。我也在在本次收集中学习到了一些筛选特征，构造特征的办法，以及训练更加具有针对性的模型的模型，受益匪浅。

**5.2 建模思路**

由于之前没有处理过这一类复杂的题目，大致只能在该选手的方法的基础上上提供一些其他思路。

1. 在该选手的特征选择基础上，构建神经网络进行训练。
2. 使用XGB算法来构建其中一个模型，在对几个模型找到最优参数，比较几个模型间的误差，选择更好的模型或者做适当的stacking融合。