

**项目开发报告**

**题目： \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**课程名称： \_\_\_机器学习\_\_\_**

**专业班级： \_\_\_ \_\_\_\_**

**学 号： \_ \_\_**

**姓 名： \_\_ \_\_\_**

**指导教师： \_\_李\_玉\_华\_\_\_\_**

**报告日期： \_\_\_\_\_\_**

**计算机科学与技术学院**

# 项目开发报告

## 1.1项目目的

本项目是作为2020中国高校计算机大赛·华为云大数据挑战赛的参赛项目，目的是使用基于机器学习的方法，解决实际生产中的问题（船舶到达时间预测）。截止到撰写该报告为止，我的排名是复赛56名（复赛总共100支队伍）。希望在完成本项目的同时，最终也可以在比赛中取得优秀的名次。

## 1.2问题分析

要解决的问题是根据船只行驶过程中的gps记录，预测到港时间。船运公司会和数据供应公司进行合作，对运输用的船通过GPS进行定位以监控船的位置；在运输管理的过程中，货物到达目的港的时间是非常重要的一项数据，那么需要通过船运的历史数据构建模型，对目的港到达时间进行预测，预测时间简称为ETA（estimated time of arrival），目的港到达时间预测为ARRIVAL\_ETA。

比赛提供历史运单GPS数据、历史运单事件数据、港口坐标数据，预测货物运单的到达时间，对应“历史运单事件”数据中EVENT\_CODE字段值为ARRIVAL AT PORT时EVENT\_CONVOLUTION\_DATE的时间值。

## 1.3设计与分析

### 1.3.1 数据分析

首先根据题目给出的数据设计算法要用到的数据中，最主要的就是gps数据和港口坐标数据。运单gps数据的各个字段为：

### 1.3.2 算法流程设计

由于本项目要解决的问题的数据都来自于真实的生产环境，因此往往面临着数据噪声大，难以处理，脏数据多，有许多缺失数据等特征，所以在设计算法的过程中对数据的预处理就显得尤其重要。同时由于比赛提供的数据量极大，全部的GPS记录有1亿多条，在本地的电脑上很难同时读取，因此在选取训练数据的过程中，也必须对完整的训练及进行初步筛选，不能使用全部数据进行训练，而只采用其中一小部分，基于上述考虑，整个算法的流程分为三步：数据筛选，数据清洗以及最终的学习算法。

我使用的编程环境是Python，主要用到的第三方库有pandas，numpy和xgboost。总体的算法流程图描述如下：

。。。。。。。

### 1.3.3 机器学习算法设计

具体的机器学习算法在设计过程中也充满了挑战性，主要的难题是，比赛给出的数据及数据类型十分独特。比如测试集中要求预测的是每条订单预计的到达目的地的时间，但是训练集中实际给出的却是每个订单中若干GPS点的数据，每条数据仅仅代表船在某一时刻的状态，而并不能反映这个订单在整个航行过程中的某个特征。因此在设计算法时，如何对数据进行组织就至关重要。此外，即使是不同的订单，数据也可能出现相同路由或者相同船只的情况，因此数据的组合方式十分复杂，为算法的设计带来了诸多困难。

。。。。。。。

## 1.4结果分析

程序训练过程的运行截图如下图所示：

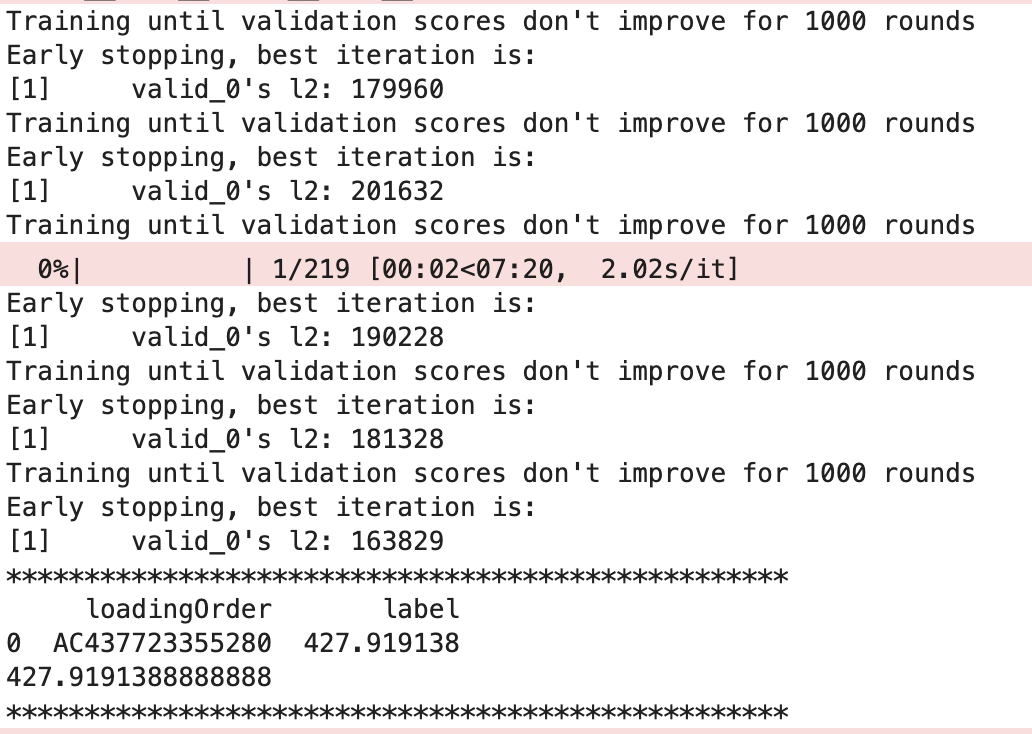


图1.3 程序运行截图

## 1.5思考与总结

本次项目我选择了自选题目的方式，主要是出于效率上的考虑。因为我现在正好在参加这个比赛，题目也与机器学习十分契合，因此就将其作为了机器学习课程的结课项目。同时，机器学习课程中学到的内容，也帮助我在算法的选择上提供了经验和判断。而且这个项目的数据量极大，并且由于是真实数据，数据本身的分布也极其复杂，因此总体来说难度较大，具有一定的挑战性。总的来说，这个项目也让我学到了很多，后续我也会继续优化算法，争取取得更好的效果。