中山大學
SUN YAT-SEN UNIVERSITY

数据挖掘导论大作业

题目 Title：回归任务下的多种数据挖掘算法的性能研究

院 系
School (Department)：数据科学与计算机学院

专 业
Major：软件工程

学生姓名
Student Name：郑先淇

学 号
Student No.：16340305

时间： 2019 年 6 月 25 日

Date: April Month 25th Day 2019 Year

【摘　要】

回归任务是机器学习领域和数据挖掘领域中一种非常常见的问题分类。回归任务的已知条件通常为一个已知的样本数据集，回归任务的目标是通过对已知数据集的分析来对未知数据的一些属性进行尽可能准确地预测。随着近些来机器学习领域的逐渐流行，很多有关于回归任务的算法被提出。这些算法自着不同的学派，采用不同的方式、从不同的方面对数据集进行处理。在面对具体问题的时候，这些算法各有优劣，在面对不同的数据集的情况下有着不同的性能表现。基于此，我和我的队友们以 Kaggle 网站上的一个回归问题作为研究对象，分别尝试了最小二乘线性回归、最小二乘多项式回归、随机森林、DART、GBDT 等多种算法，对其结果进行对比和分析。

【关键词】数据挖掘；回归任务；算法；性能对比；

[ABSTRACT]

Regression tasks are a very common classification of problems in the field of machine learning and data mining. The known condition of a regression task is usually a known sample data set. The goal of the regression task is to predict some of the attributes of the unknown data as accurately as possible by analyzing the known data sets. With the recent popularity of machine learning, many algorithms for regression tasks have been proposed. These algorithms use different methods to process data sets from different aspects. With specific problems, these algorithms have their own advantages and disadvantages, and have different performances . Therefore, our team use a regression problem on the Kaggle website as the research object, and try various algorithms such as least squares linear regression, least squares polynomial regression, random forest, DART, GBDT...etc. And we summarize the results of all these algorithms to compare their performance and analyse their advantages and disadvantages.

[Keywords]: data mining; linear regression; algorithm; performance;

目　录

第一章 概述/引言

1.1 回归任务下多种机器学习算法性能对比的意义

随着机器学习领域近些年来的流行和计算机计算能力的快速提升，越来越多更加复杂的学习算法被提出，这些算法看起来要比经典的学习算法“智能”得多。但是，很多这些新提出的智能算法却是在那些我们都熟知的经典算法上改进而来的，通过对这些算法的性能进行研究和对比，我们可能可以找到这些算法各自的适用场景。另外，通过对这些算法进行整体的对比和分析，也有助于我们更好地掌握本学期以来学习的课程内容。

1.2 问题的描述

我们的研究问题来自于 Kaggle 网站上的一个开放性问题，以下为问题描述：

“It is a regression task: Given 10m data samples as train set, each of 13 features, please predict the label (range unlimited) for the whole test set containing 10915121 data samples. This project expects you to design and implement a parallel decision tree algorithm, i.e. GDBT or Random Forest.”

1.3 本文的工作

我和我的队友们分别尝试了最小二乘线性回归、最小二乘多项式回归、随机森林、DART、GBDT 等多种算法对该回归问题进行了讨论了分析，并最终把各个算法的结果整合到一起进行比对分析。我个人完成的工作是对给定的数据进行分析，即所谓的“特征工程”，并使用最小二乘线性回归和最小二乘多项式回归算法对进行预测。

1.4 论文结构简介

本文第一章主要是阐述研究基于回归任务下的多种智能算法的性能对比和分析的意义以及本文所要大致工作内容；第二章主要讲述本次实验中所使用的一些知识的原理；第三章主要是对数据集的特征进行分析以及相关的数据预处理过程；第四章是模型训练的内容；第五章是汇总整个小组所用的算法并进行性能比较；第六章是小组成员贡献度说明。

第二章 相关知识介绍

2.1 多元最小二乘线性回归模型介绍

给定由 d 个属性描述的实例 x = (x1;x2;..;xd),其中 xi 是 x 在第 i 个属性上的取值，线性模型试图学得一个通过属性的线性组合来进行预测的函数，即

$$f(\boldsymbol{x}) = w_1 x_1 + w_2 x_2 + \ldots + w_d x_d + b$$

一般用向量形式写成

$$f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x} + b$$

其中 w = (w1;w2; ... ; wd)。w 和 b 学得之后，模型就得以确定。

2.2 多元最小二乘多项式回归原理

多项式回归模型是线性回归模型的一种，此时回归函数关于回归系数是线性的。多项式回归问题可以通过变量转换化为多元线性回归问题来解决。对于一元 m 次多项式回归方程，令

$$x_1 = x, x_2 = x^2, \cdots, x_m = x^m$$

，则

$$\hat{y} = b_0 + b_1 x + b_2 x^2 + \cdots + b_m x^m$$

就转化为 m 元线性回归方程

$$\hat{y} = b_0 + b_1 x + b_2 x_2 + \cdots + b_m x_m$$

因此用多元线性函数的回归方法就可解决多项式回归问题。需要指出的是，在多项式回归分析中，检验回归系数b_1是否显著，实质上就是判断自变量 x 的 i 次方项对因变量 y 的影响是否显著。

2.3 交叉验证的原理

“交叉验证法”将数据集 D 划分成 k 个大小相似的互斥子集，即

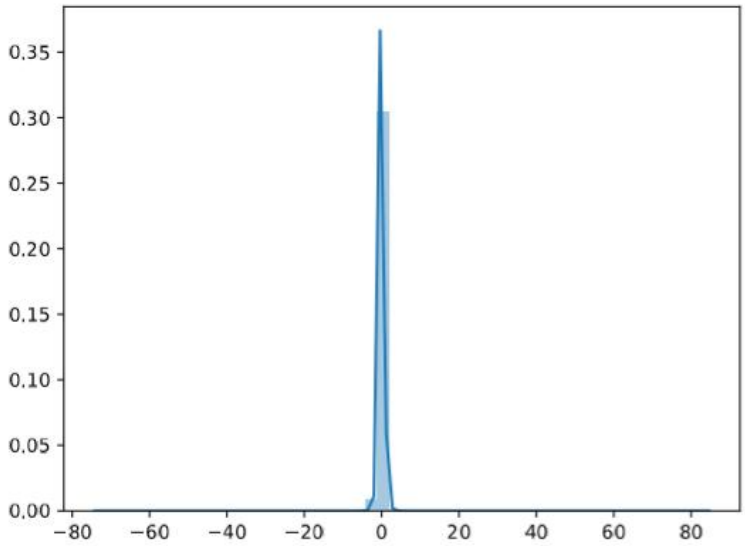
$$D = D_1 \cup D_2 \cup \ldots \cup D_k,\ D_i \cap D_j = \varnothing\ (i \neq j)$$

每个子集 D_i 都尽可能保持数据分布的一致性，即从 D 中通过分层采样得到。然后，每次用 k-1 个子集的并集作为训练集，余下的那个子集作为测试集；这样就可以获得 k 组训练/测试集，从而可进行 k 次训练和测试，最终返回的是这 k 个测试结果的均值。交叉验证法评估结果的稳定性和保真性很大程度上取决于 k 的取值，为强调这一点，通常把交叉验证法称为“k 折交叉验证”。

第三章 特征分析与数据预处理

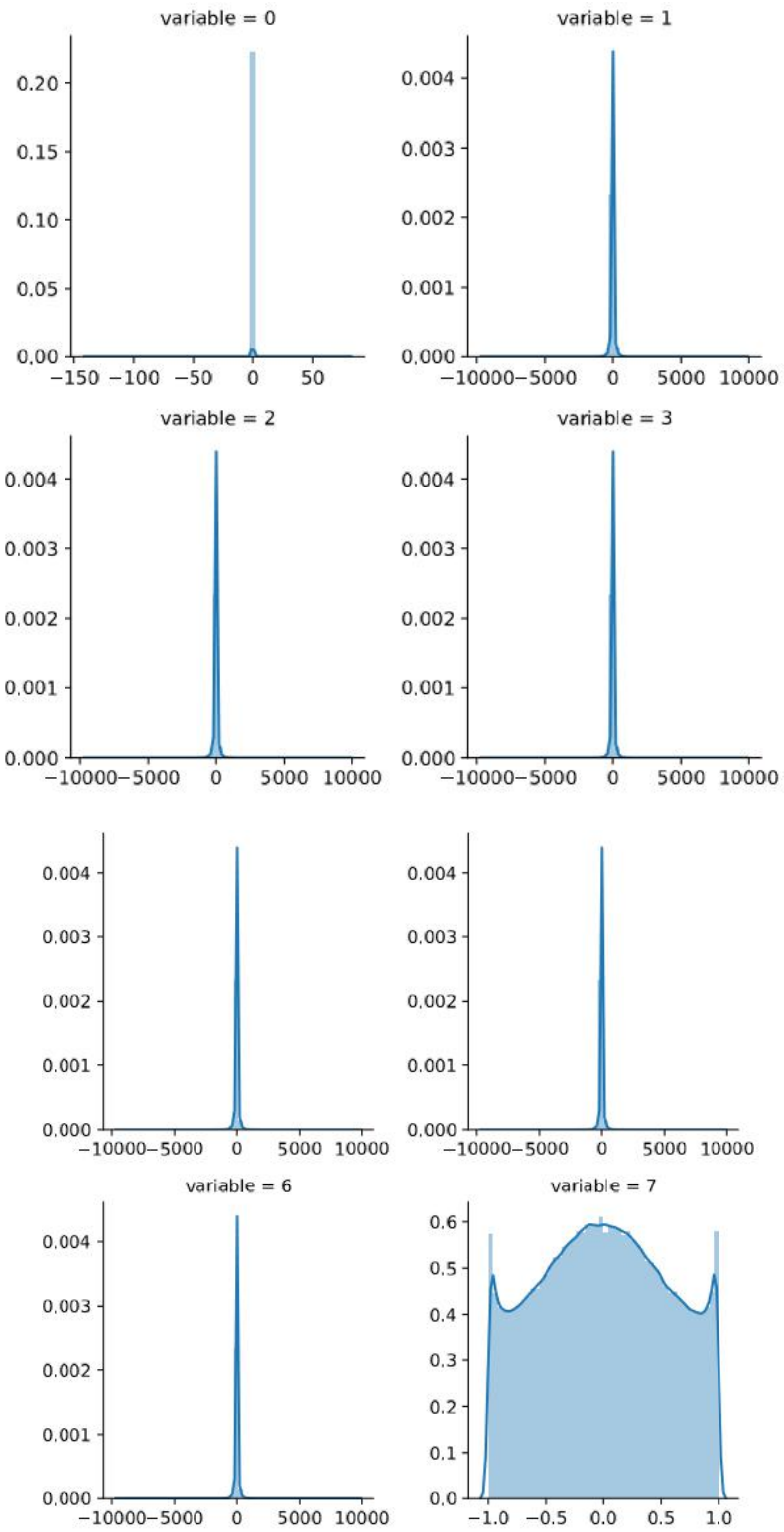
3.1 特征分析

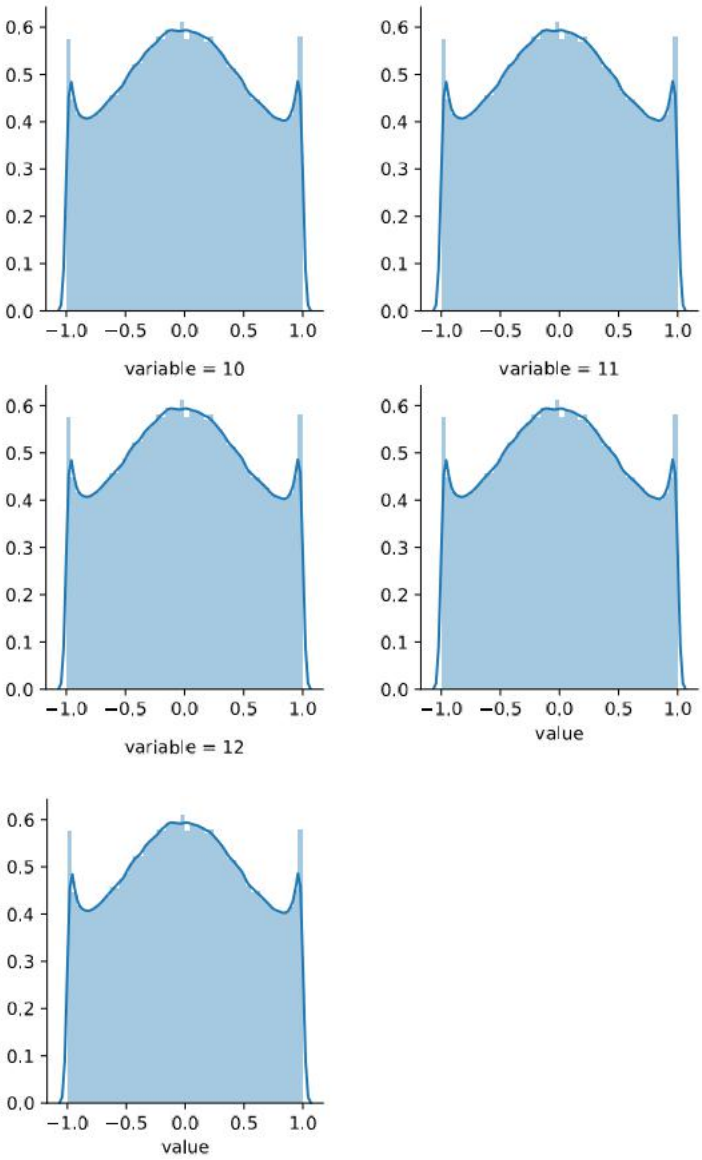
首先绘制 label 的分布情况如下：



由上图可知，Label 的分布比较均匀，基本服从正态分布，故不需要对 label 进行预处理。

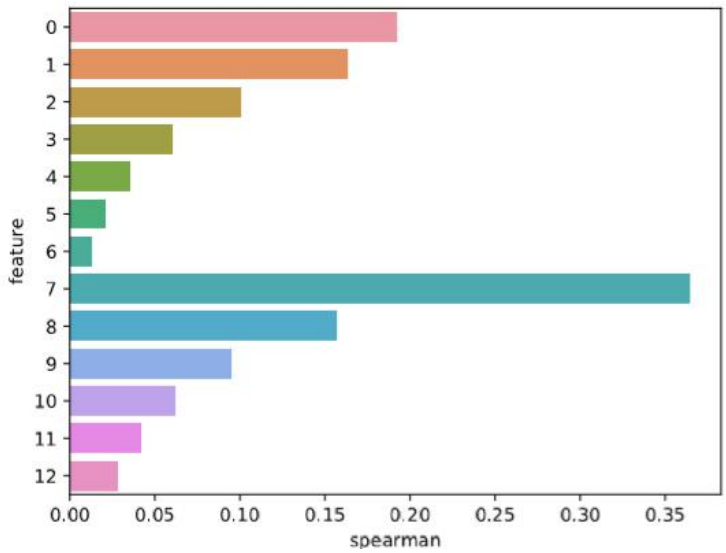
然后研究每个 Feature 的分布情况，情况如下：


variable = 0
variable = 1
variable = 2
variable = 3
variable = 6
variable = 7



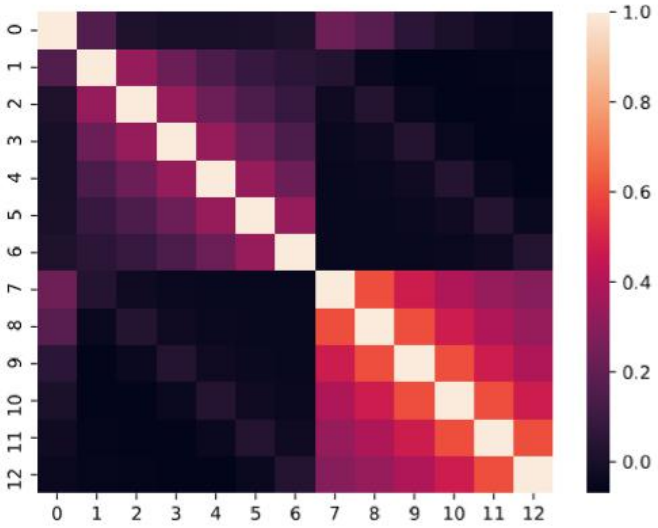
由上图可知，13 个 feature 的分布情况并不是很好，可通过对 feature 进行正则化处理来改善 feature 的分布情况。

接下来我们看一下每个 feature 和 label 的相关性，情况如下：



观察上图，不难发现与 label 相关性最高的相关系数也仅仅只有 0.35 左右，因此我们可以粗略认为 feature 和 label 之间没有明显相关性的结论。

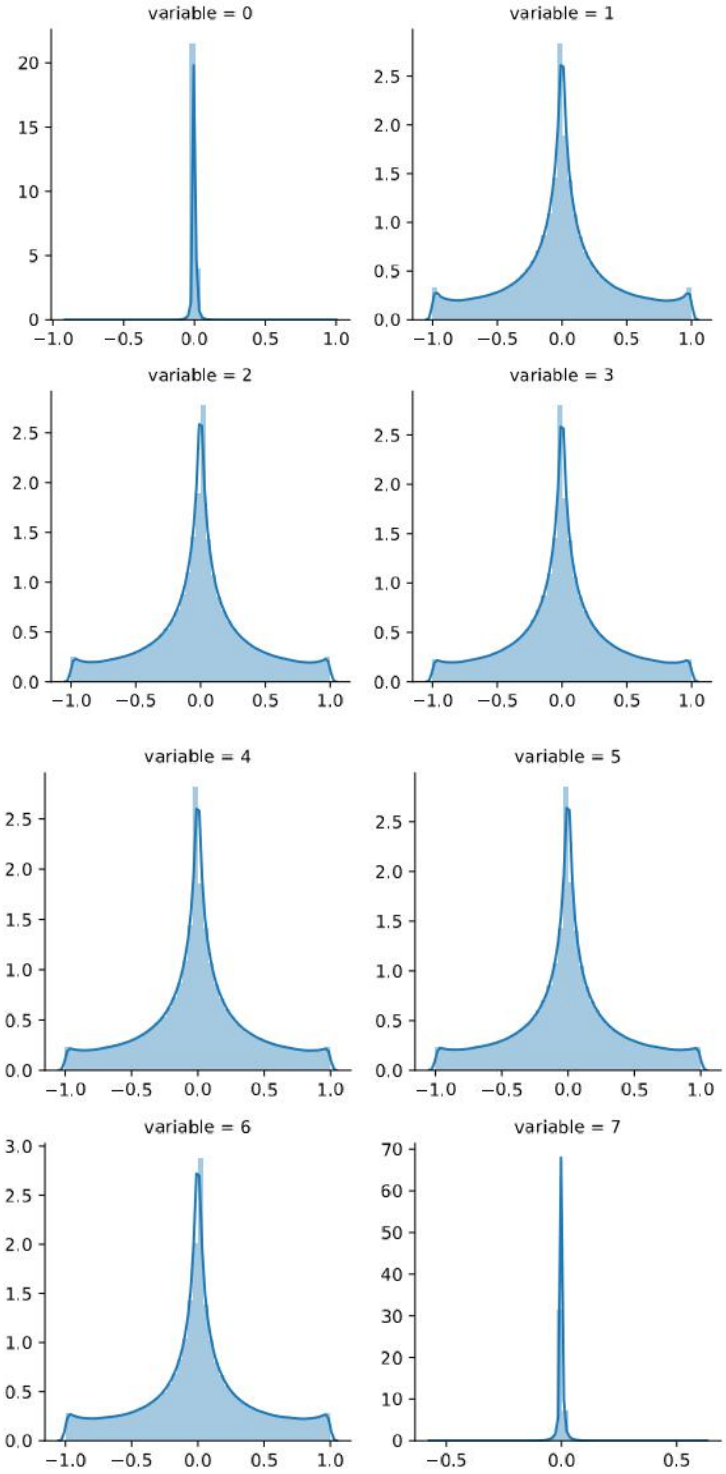
然后研究每个 feature 之间的相关性，情况如下：

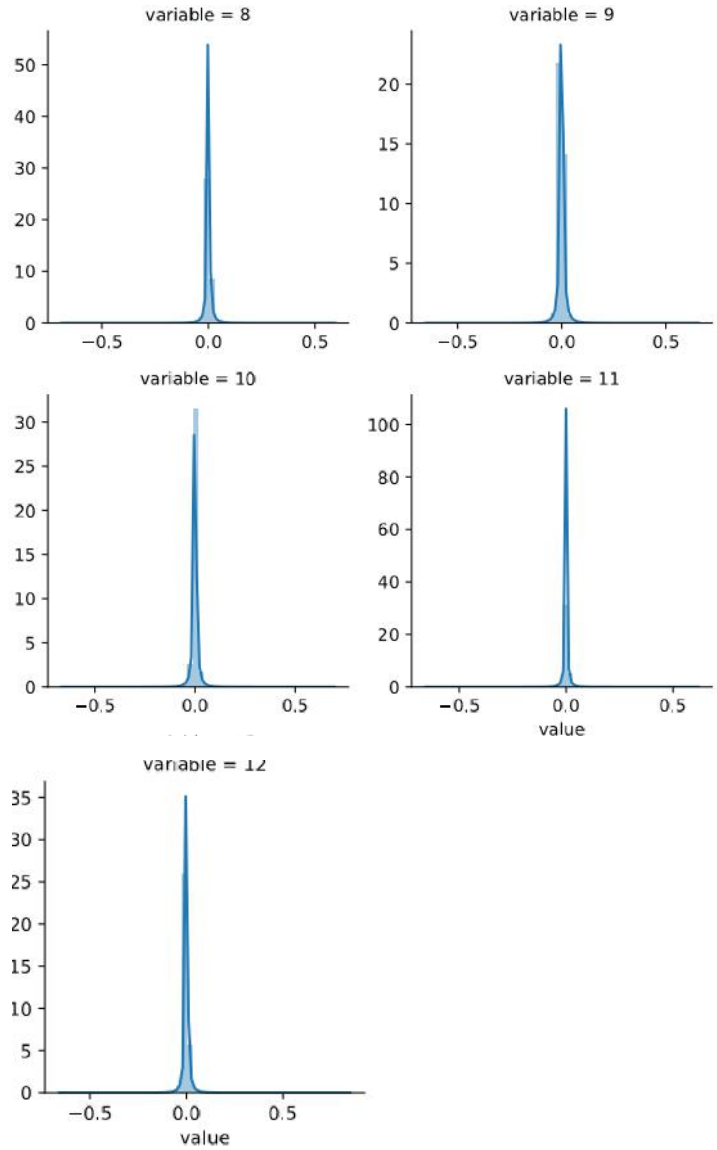


上图中，颜色越浅的部分代表对应的两个 feature 的相关性越高，可以看到 feature7、8、9、10 的相关性相对较高，但是其他 feature 之间没有明显的相关性。

3.2 数据预处理

由 3.1，我们知道 feature 的分布情况并不是很好，因此我们对每个 feature 进行正则化处理，处理后的结果如下：





由上图可得，经过正则化处理之后，feature 的分布情况明显看起来规则了很多，这样更有利于我们实验的进行。

第四章 模型训练

4.1 判别模型设定

在这个判别模型里，我把评判指标设定为“r2”，交叉验证的 k 设定为 5。

4.2 模型训练

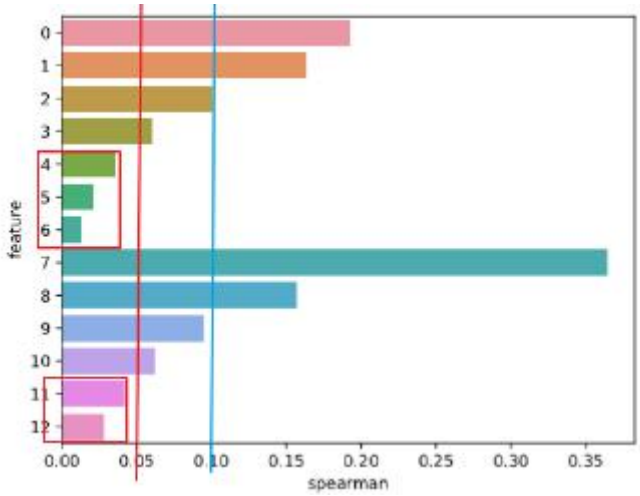
在本次实验中，我使用 Python 的 sklearn 库作为辅助建模的工具。

4.2.1 最小二乘线性回归

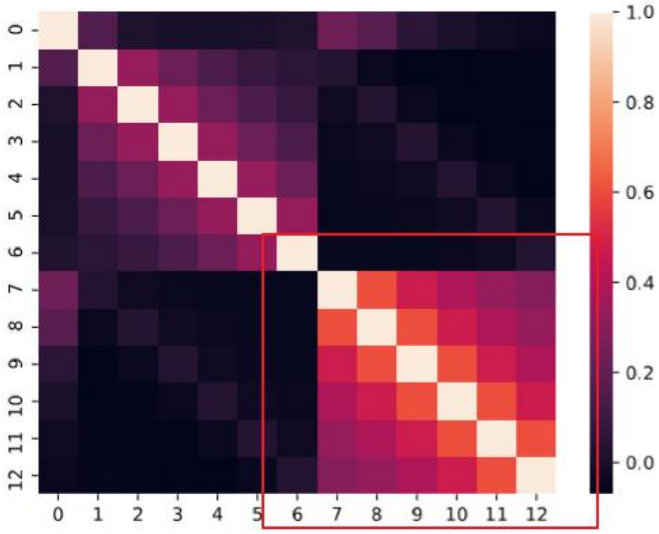
使用 sklearn 的 linear_model.LinearRegression()进行模型训练，并将 test 数据集预测得到的结果提交到 kaggle 网站进行测评。

4.2.2 最小二乘多项式回归

在本次研究中，我设定最小二乘多项式回归的最高次数为 2。首次尝试输入所有的数据进行训练，结果出现内存溢出的错误。探究其原因，是因为当设定最高次为 2 时，原本的训练集的 feature 由 13 种变成了 C_{13}^{2} + 13 种，数据量骤增，由于我的电脑内存仅为 4G，无法处理这样的数据量，因此根据以上的 feature 的相关性分析，我将一些与 label 关联性很小的 feature 删掉，具体如下：



由上图可知，feature4、5、6、11、12 与 label 的相关性小于 0.05，feature 3、9、10 与 label 的相关性小于 0.1。再由 feature 之间的相关性情况：



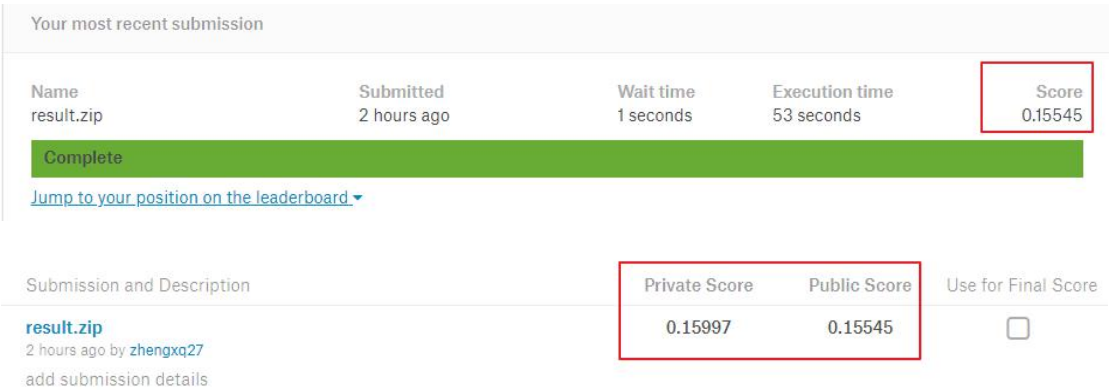
上图中红色方框圈出来的部分，即 feature9、feature10、feature11、feature12 之间的关联性很高。综合以上考虑，我们对 feature 数据处理如下：删除 feature4、feature5、feature6、feature9、feature10 的数据，对剩下的 feature 采用二项式回归的方法进行训练，将预测结果提交 kaggle 网站。

4.3 结果

4.3.1 最小二乘线性回归

验证集的 r2 分数为：0.1393

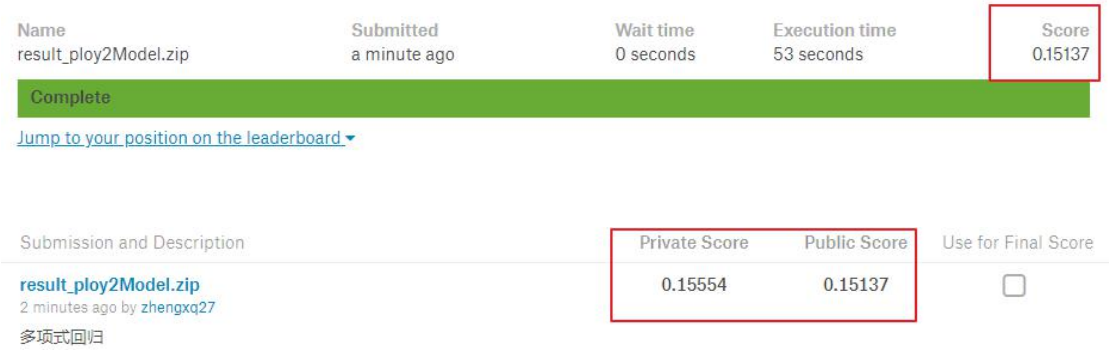
Kaggle 的评分结果如下：0.1599



4.3.2 最小二乘多项式回归

验证集的 r2 分数为：0.1364

Kaggle 的评分结果如下：



第五章 小组成果对比和算法性能分析

我们小组分别实验了随机森林、Dart、决策树、岭回归、多项式回归等多种算法，各种算法的实验结果统计如下：

算法	R2 验证集分数	kaggle 测试分数
最小二乘回归	0.139	0.1555
多项式回归	0.136	0.1513
随机森林	0.151	0.1684
Dart	0.157	0.1741
决策树	0.155	0.1734
岭回归	0.139	0.15545

可以看到，在这次实验里面，最小二乘回归、多项式回归和岭回归等回归模型取得的分数都稍微要差一些，说明在这个数据集下回归模型不能取得很好的效果。相反，决策树模型和 Dart 取得的分数就要相对乐观一点，当然以上的两种的算法在调参方面要付出比上面的回归模型更多的努力。

第六章 小组成员贡献表

组员	贡献
张吉祺	使用了岭回归以及决策树算法进行实验分析
郑国林	使用了随机森林以及 Dart 算法进行实验分析
郑先淇	使用了最小二乘回归、多项式回归进行实验分析