**机器学习作业1-PUBG排名预测**

**郭远帆，唐荣俊，罗乙然**

目录

[任务介绍 1](#_Toc527654737)

[实现方案 1](#_Toc527654738)

[实验结果 1](#_Toc527654739)

[分析与后续 1](#_Toc527654740)

# 任务介绍

## 任务说明

本次任务来源于Kaggle数据科学竞赛网站提供的比赛项目:**PUBG Finish Placement Prediction （绝地求生:大逃杀玩家排名预测）**

比赛说明网站: <https://www.kaggle.com/c/pubg-finish-placement-prediction> 网站提供了该任务的详细说明，输出要求以及数据（包括训练集和测试集）

这是一个非常有趣的项目，竞赛网站与PUBG制作商蓝洞公司合作，获取了超过65，000场游戏的玩家数据，要求竞赛参加者根据这些数据预测玩家在比赛中的排名。

## 数据集的使用

本项目原数据集相当庞大，训练集与测试集数据量一共有超过650万条数据，每一条数据拥有25个特征。在实际使用中，我们对其进行采样，以缩短训练时间，从而在短时间内了解多个模型的特点以及调参方法。

## 任务目标与规划

我们小组以此项目作为机器学习第一次作业，该问题属于回归问题。而由于该项目本身是数据科学竞赛项目，其中必然存在数据分析与特征提取的过程（但并非主要的），我们将任务划分为几个阶段:

第一阶段: 数据缺失处理，进行简单的数据可视化分析

第二阶段: 参考竞赛网站其它竞赛参加者提供的特征工程方案，提取简单特征

第三阶段: 构建不同的机器学习模型训练，并且进行模型的评估

第四阶段: 采用集成学习技术将数个模型整合，构成最终模型并进行分析比较

其中第三第四阶段为主要的工作

# 实现方案

## 编程环境

## 基本特征工程（线性模型）

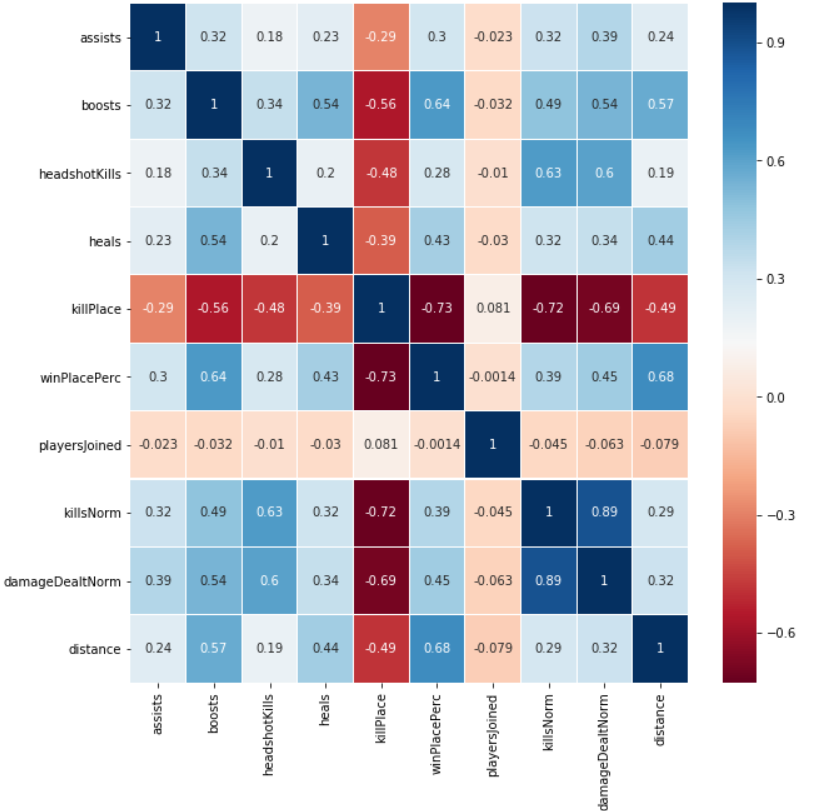
线性回归模型以模型简单、速度快而在工业界广泛应用。但是由于模型过于简单，线性回归模型需要更多的特征提取工作，否则非常容易欠拟合而得到不满意的结果。

# 实验结果

## 线性模型

在完成一轮特征工程后，数据集中留下来的特征拥有较好的独立性，并且对于初步模型来说降低了数据的复杂度，因此也降低了拟合的难度。我们采用不同的线性模型进行拟合并观察实验结果，并尝试着对实验结果进行解释。

### LinearRegression



调用sklearn 中的最小二乘法线性模型LinearRegression进行训练以及预测。

**实验结果**:

M = 0.095，即线性回归模型平均将引入9.5%的误差。

鉴于该模型的低成本性，9.5%误差的结果可以接受。线性模型的准确性在相当程度上依赖于特征工程，因此在进行线性模型的特征工程时进行了多组实验。得到实验表格如下:

## SVR, LinearSVR, NuSVR

# 分析与后续