三、预测科比投篮是否命中

本案例的数据来源于著名数据建模和数据分析竞赛平台Kaggle上的一个名为[”Kobe Brayant Shot Selection“](https://www.kaggle.com/c/kobe-bryant-shot-selection)的一个竞赛，该竞赛非常适合用于分类训练、特征工程和时间序列分析。本案例需要分析科比在20年职业生涯中投中和投失球的各项数据，并根据他投球的位置和投球的方式等信息预测该球能否投中，这是一个典型的二分类问题。

数据集包含30697个样本，共有25个变量，其中有3个浮点型（float）变量、11个整型（Int）变量和11个对象型（object）变量,除了shot\_made\_flag变量外都不存在缺失值，标签为shot\_made\_flag，每个变量对应的含义如下所示：





步骤要求：

1. 投篮区域可视化，即在二维平面画出样本点的[loc\_x, loc\_y]，并指定分类变量为shot\_zone\_area。类似地，指定分类变量分别为shot\_zone\_basic及shot\_zone\_range
2. 特征game\_event\_id，game\_id，team\_name以及team\_id与投篮命中预测无关，可以直接舍去
3. 查看数值型变量以及非数值型变量的分布，正负类样本数据是否平衡
4. 画出正负类样本数据关于各种数值型变量的箱线图，并分析哪些变量可以有效区分正负类
5. 基于散点图分析数值型变量两两之间的相关性
6. 画出正负类样本数据关于各种非数值型变量的统计图（柱状图）
7. 数据清洗：shot\_made\_flag未知的数据作为测试集，shot\_made\_flag已知的数据作为训练集。去掉lat变量（与loc\_x相关），去掉lon变量（与loc\_y相关）
8. 特征转换：将剩余分钟数（minutes\_remaining）以及秒数(seconds\_remaining)统一转换为剩余时间，新建二进制变量（last\_5\_sec\_in\_period）以定义结束时间是否小于5秒，将Matchup转换为二进制变量home\_play，将game\_date分解为game\_year以及game\_month，将loc\_x及loc\_y利用pd.cut离散化
9. 对非数值型变量即名义型特征，进行One-Hot编码
10. 分别利用方差阈值、随机森林、卡方检验以及RFE方法进行特征选择，并将4种方法各自的top20合并，形成最终的特征
11. 利用PCA来分析前两个主成分
12. 利用AdaBoost，KNN，CART，Naïve Bayes，Random forest在训练集上进行3折交叉验证，并基于AUC来比较不同模型的预测效果
13. 对测试集进行预测