Kaggle NYC Taxi Fare Prediction - Step 1

First step: Download all the data from the website

Second step: Load and explore all the data

* 用 pandas 读取前 10 万行数据
* 查看数据结构 .info()，统计信息 .describe()
* 检查缺失值 & 异常值

Third step: Data cleaning

* 删除负数车费、极端经纬度、不合理乘客数

Fourth step: Feature engineering

* 从 pickup\_datetime 提取 year, month, day, hour, weekday
* 计算 Haversine 距离 (distance\_km)

data processing, training, validation 三步骤

overfit underfit，训练数据 training80 validation20 for beginning

Kaggle NC Taxi Fare Prediction – Step 2

## Step 2: Model Training and Validation 第二步：模型训练与验证

### First: Split data into training and validation sets 第一步：拆分数据集

- 使用 train\_test\_split 将特征（X）与标签（y）按照 80% 训练集、20% 验证集 的比例分开。

### Second: Train Linear Regression model 第二步：训练线性回归模型

- 模型：LinearRegression()  
- 原理：假设车费与各个特征（经纬度、时间、乘客数、距离）是线性关系。  
- 结果：  
 - 验证集 RMSE（均方根误差）：5.22

### Third: Train Random Forest Regression model 第三步：训练随机森林回归模型

- 模型：RandomForestRegressor(n\_estimators=100)  
- 原理：通过多棵决策树组合预测，提升非线性拟合能力。  
- 结果：  
 - 验证集 RMSE：4.00

### Fourth: Train LightGBM model 第四步：训练 LightGBM 模型

- 模型：LGBMRegressor  
- 原理：基于梯度提升决策树（GBDT）  
- 参数：500 棵树、学习率 0.05、采样比例 0.8  
- 结果：  
 - 验证集 RMSE：3.92

### Fifth: Summary of model performance 第五步：模型表现总结

| Model 模型 | RMSE |
| --- | --- |
| Linear Regression | 5.22 |
| Random Forest | 4.00 |
| LightGBM | 3.92 |

LightGBM > Random Forest > Linear Regression