# 一、项目简介

本项目是天河上的一个比赛，旨在通过对二手车数据进行建模，预测二手车的交易价格。该比赛使用的数据集来自某交易平台的二手车交易记录，包含31列变量信息，其中15列为匿名变量。为保证比赛的公平性，将从中抽取15万条作为训练集，5万条作为测试集A，5万条作为测试集B，同时会对name、model、brand和regionCode等信息进行脱敏。

本项目的任务是使用数据集中的变量，建立模型，预测二手车的交易价格。在模型建立过程中，需要进行数据预处理、特征选择和模型调优等操作，并通过交叉验证等方法评估所建模型的性能。

# 二、数据预处理

本项目中，需要进行的数据预处理主要包括缺失值填充、异常值处理和日期数据处理等。

在进行数据预处理之前，首先需要读取训练数据和测试数据，并使用 Pandas 库将它们合并到一起。合并后的数据集总共有 20 万条数据，并包含 31 列变量信息，其中 15 列为匿名变量。

* 读取训练数据和测试数据。
* 使用 Pandas 库将训练数据和测试数据合并到一起。
* 对 'price' 做对数变换。
* 用众数填充缺失值。
* 处理异常值:将功率大于 600 的标为 600。
* 将 ‘notRepairedDamage’ 中的缺失值替换为 None。
* 对可分类的连续特征进行分桶。
* 对日期数据进行处理，提取年、月、日等信息和计算二手车使用时间。

# 三、特征工程

在数据预处理之后，需要进行特征工程，以提取更多有用的特征。本项目中包括对数据进行缺失值填充、数据统计和特征组合等操作。

具体的操作包括：

1. 对数据进行缺失值填充。
2. 对行驶路程和功率数据进行统计，例如计算行驶路程与功率的最大值、最小值、中位数和均值等。
3. 为部分属性列的数据生成新的特征，主要是通过对 V0、V3、V8 和 V12 四个特征进行组合生成新的二元和三元特征。

# 四、模型训练与预测

在完成特征工程之后，需要使用机器学习模型对数据进行训练，并预测测试集中二手车的交易价格。本项目中使用的模型是 LightGBMRegressor，使用了四个参数（'n\_estimators'、'learning\_rate'、'num\_leaves'和'lambda\_l2'）进行调参。在训练模型的过程中，使用了 K-Fold 交叉验证，最终得到的结果是五折交叉验证的平均结果。

## 五、结果评估

在使用模型对测试集进行预测之后，需要对结果进行评估。本项目中使用的评估指标是 MAE（Mean Absolute Error），即预测值与实际值之间的平均绝对误差。

## 六、总结

本项目通过对二手车数据进行建模，预测了二手车的交易价格。在完成项目的过程中，主要涉及到数据预处理、特征工程、模型训练与评估等多个方面。其中特征工程的部分使用了数据的统计和组合等方法，提取了更多有用的特征，从而提升了模型的准确性和稳定性。

在实际应用中，可以通过对数据集进行更加深入的探索和分析，以进一步提高建模的准确性和稳定性，并应用更加高级的机器学习模型和技术来进行建模和预测。