1. **实验介绍**

题目链接：https://tianchi.aliyun.com/competition/entrance/231784/submission

本实验以预测二手车的交易价格为任务，本次实验的数据集包括15万条训练数据集，5万条测试集用于模型测试。作为大作业前的最后一次实验，目的是让大家对数据挖掘的完整流程有一个基本的了解。以下是实验相关介绍：

train.csv：训练集，包含31列变量信息，其中15列为匿名变量，每辆车都有其二手交易价格。

test.csv：测试集，包含31列变量特征，其中15列为匿名变量，每辆车的二手价格是模型需要预测输出的，最后将预测结果以submit.csv文件格式输出，在提交结果处上传submit.csv文件，如下图所示，网站会对你的提交进行验证，计算模型的性能指标，每天可以提交三次测试。

图形用户界面, 文本, 应用程序

描述已自动生成

部分变量信息介绍如下：

| **Field** | **Description** |
| --- | --- |
| SaleID | 交易ID，唯一编码 |
| name | 汽车交易名称，已脱敏 |
| regDate | 汽车注册日期，例如20160101，2016年01月01日 |
| model | 车型编码，已脱敏 |
| brand | 汽车品牌，已脱敏 |
| bodyType | 车身类型：豪华轿车：0，微型车：1，厢型车：2，大巴车：3，敞篷车：4，双门汽车：5，商务车：6，搅拌车：7 |
| fuelType | 燃油类型：汽油：0，柴油：1，液化石油气：2，天然气：3，混合动力：4，其他：5，电动：6 |
| gearbox | 变速箱：手动：0，自动：1 |
| power | 发动机功率：范围 [ 0, 600 ] |
| kilometer | 汽车已行驶公里，单位万km |
| notRepairedDamage | 汽车有尚未修复的损坏：是：0，否：1 |
| regionCode | 地区编码，已脱敏 |
| seller | 销售方：个体：0，非个体：1 |
| offerType | 报价类型：提供：0，请求：1 |
| creatDate | 汽车上线时间，即开始售卖时间 |
| price | 二手车交易价格（预测目标） |
| v系列特征 | 匿名特征，包含v0-14在内15个匿名特征 |

1. **实验评估指标**

采用MAE来评估模型的预测精准度，若真实值为，模型的预测值为，则MAE的计算公式为

MAE值越小，则说明模型预测得越准确。在本实验中即计算预测价格和真实价格之间的差别。

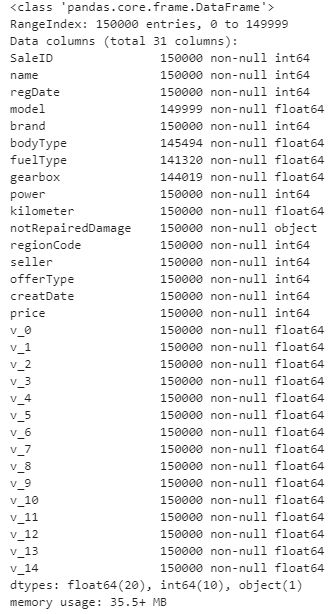
1. **实验内容**

（1）赛题调研  
针对问题性质，列举三种可以使用的预测算法模型，并简要说明该算法的原理和优缺点。

（2）数据分析与预处理【可参考以下操作，也可自由发挥】

* 载入训练和测试数据集，输出其规模，以及各特征的数据类型。

示例：



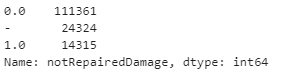
* 检查各特征是否存在缺失值，以及使用descibe()函数输出每个特征的的统计指标，查看是否存在异常值。

示例：

表格

描述已自动生成

* 针对object类型特征进行检测，查看是否存在异常值。



* 结合自身理解，对发现的缺失值和异常值选择合适方法进行处理。
* 计算各数值特征的偏度和峰值，针对高偏度的数值特征，采用Box Cox变换。
* 针对price属性，绘制分布曲线以及分位图，根据绘制的图进行分析，看其是否符合正态分布，若不符合，其对其进行变换。

1. 特征工程【可参考以下操作，也可自由发挥】

* 将所有数值特征与price进行相关行分析计算，并绘制热力图，找出哪些特征与price相关性较高，删除与目标值无关的特征。
* 对原始时间属性特征【regDate、createDate】进行处理，构造像使用年长、天数等有效的时间特征。
* 对可分类的连续特征的进行分桶操作，使用分类特征进行特征交叉得到一系列特征，基于lightgbm进行筛选。
* 结合自身理解组合和构造新特征，比如不同品牌售卖价格的平均值，最小值，最大值等，不同行驶公里售卖价格的平均值等。
* 课外调研匿名特征【v\_0-v\_15】的处理方法
* 对特征数据进行降维操作，以优化后续模型计算量和计算时间。

（4）学习模型

使用单一模型进行预测或采用多模型融合来进行预测

（5）模型评估与分析

说明数据集的划分方法，模型训练完毕后将submit文件上传到<https://tianchi.aliyun.com/competition/entrance/231784/information>进行测试，记录分数和排名。