**Topic: Predict prices of used cars**

训练集： used\_car\_train\_20200313.csv

测试集： used\_car\_testA\_20200313.csv

具体步骤以及展示结果：

1. 理解数据：
2. 读取数据以后，看到数据的label如下：

['SaleID', 'name', 'regDate', 'model', 'brand', 'bodyType', 'fuelType', 'gearbox', 'power', 'kilometer', 'notRepairedDamage', 'regionCode', 'seller', 'offerType', 'creatDate', 'price', 'v\_0', 'v\_1', 'v\_2', 'v\_3','v\_4', 'v\_5', 'v\_6', 'v\_7', 'v\_8', 'v\_9', 'v\_10', 'v\_11', 'v\_12', 'v\_13', 'v\_14'].

B. 观察每个特征的分布，可以看到price并不是高斯分布，对其log处理使其高斯分布。

C.统计每个维度的缺失值，发现fuelType的nan值占比不小，对其用0进行填充，因为fuelType是类别特征，只有0/1

1. 特征工程：
2. 删除异常值：

通过箱线图来去除异常值，只需要去除“power”列的异常值即可，因为其余的都是固定的，不存在异常值

B.基于现有的特征挖掘潜在特征，然后删掉不需要的特征：

通过regdate和createdate获得车龄, 以及对于每个品牌，价格的中位数，平均数

C.为随机森林和神经网络/线性回归分别准备数据集，因为后者需要**归一化**，而前者一些feature需要**编码**。

D.训练前对数据进行处理，去掉“-”等特殊符号的样本，不然报错

1. 建模（我们选4个模型）

随机森林：结果很好

线性回归：mae很高，不用

神经网络：结果很好

决策树：结果很好

1. 模型评估

使用交叉验证法分别对三个模型进行评估

交叉验证法：将数据集分成k份，其中一份作为测试集（不重复选取），其余作为训练集，计算该模型在测试集上的MSE，最后的模型得分为k次平均的MSE，aka MAE（mean average error）

交叉验证的优点：

1. 可以看出模型泛化能力如何，因为每次的测试集是不一样的，可以通过观察结果来避免overfit/underfit.
2. 可以评估不同模型质量，从而选取最优的模型或者加权进行模型融合
3. k值没有固定说法，一般当模型不稳定时，k越大越好，但是k越大对机器的开销越大，所以我们选5
4. 模型融合

基于上一步，不同模型得到不同的得分，根据此得分进行加权融合，然后得到预测结果。

1. 可以优化的地方：

可以进行数据分桶

用算法进行调参优化模型