**Machine Learning Part Exploration Process**

**1 整体构思**

本项目使用机器学习的目的是对相机价格进行预测。但是由于相机种类繁多，且价格曲线各不相同，因此我们决定先对相机种类进行划分，再对不同种类相机的价格分别进行预测。

* 1. **聚类**

对于相机的分类，我们首先想到的就是使用聚类模型，因此我们针对每个相机都找了一些主要的参数来进行分类（例如：ISO，Continuous Shot，Effective Pixel 等）。找到参数之后，对于我们这个案例，应该选择一个非监督学习的模型，因为聚类是为了找到不同相机内部的潜在关系。

最后我们在K-Means，DBSCAN和Hierarchical Clustering这三个模型中进行决定，考虑到K-Means需要预先确定簇的数量，而DBSCAN会处理噪声点，而我们的相机应该始终属于一个类中，所以最终我们决定使用Hierarchical Clustering模型，其优点在于可以根据树状结构图理解数据，在运算后决定需要划分类的数目。

**1.2 价格预测**

分好类之后，我们需要对每个类中的相机价格统一进行处理，用一个类中所有的相机价格来训练其对应的模型，每个类使用的模型相同，但是需要根据具体情况调整模型的参数。我们在众多模型中选择，一开始我们想直接使用time Series模型预测，但是这个模型没有办法捕捉长时间内的价格变化，同时考虑到即使是同一个类中的相机，价格相差也不是相等的，甚至相差很多，因此此时决定要对相机价格进行标准化，统一除以该相机的官方发布价格，这样无论是增长还是降低，都有了一个参考标准，所有相机的价格都转化为了相对价格。其次，对于日期这个参数，不同相机在市场上出现的时间也不尽相同，因此采用相机的发布日期到取样日期所经过的天数作为日期的值。

在这样的变换下，价格的预测就可以采用函数拟合的方式进行，我们最终决定在Multiple Linear Regression，Vetor Auto Regression和Support Vector Regression中进行选择，但由于聚类还没有完成，因此，还无法测试哪个模型效果更佳。

**2 代码测试**

**2.1 数据处理**

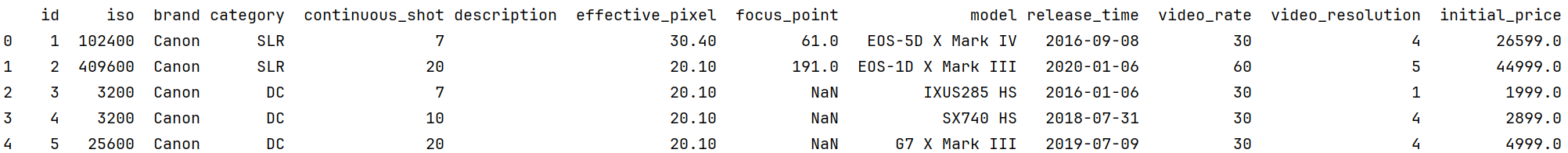
此次的训练数据全部是在上传到数据库之后，在通过pymysql对数据库访问得到的。首先在camera表中，如图2-1所示，展示了前五行的数据，一共是35个不同的相机。首先可以看到focus point列里有缺失值，我们选择将缺失值设置为0，其次brand，category这两个属性是分类参数，需要对其进行One Hot编码，而iso，video resolution和video rate是离散的数值参数，我们决定根据区间进行分类后再进行One Hot编码，如图2-2所示。

图2-1 xxx

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

图2-2 xxx

图片包含 形状

描述已自动生成 对数据进行One Hot编码后，需要对数值列进行归一化，我们首先想到的是Standard Scaler，但是考虑到数据量小且数据不符合正态分布，因此选用更合适的Min Max Scaler，对四列进行归一化，如图2-3所示。

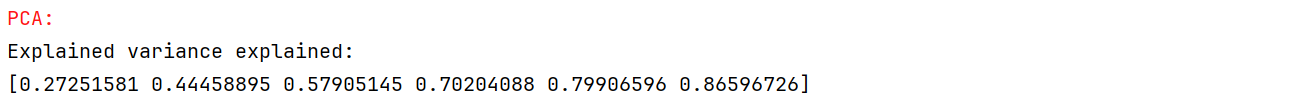
图2-3 xxx

图2-4 xxx

进行过One Hot编码后数据维度过大，因此需要使用PCA进行降维操作，经过我们多次尝试，最终选取n\_components=6，可以得到86.6%的成分占比，如图2-4所示。

**2.2 聚类**

图表

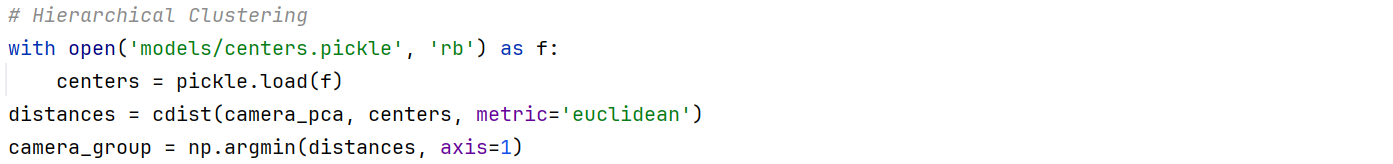
描述已自动生成数据处理完成后，将其放入Hierarchical Clustering模型中，得到结果如图2-5和图2-6所示。

图表, 散点图

描述已自动生成图2-5 xxx

图2-6 xxx

在我们对Hierarchical tree进行观察后，我们决定将相机分为4类是最合理的。再预测新加入的相机属于哪个类的时候，我们遇到了问题，Agglomerative Clustering的模型是没有predict方法的，因此我们查询了资料，是需要将聚类中心点的值计算出来，再根据欧式距离的计算公式计算新加入相机距离哪一个中心点更近来决定相机属于哪个类，如图2-7所示。

图2-7 xxx

**2.3 价格预测**

在价格预测的模型选择上，我们决定将三种模型都测试一遍，最终选取表现最好的一种模型。在将日期转换成相对发布日期的天数，价格转换成相对发布价格的比值之后，把同一组相机的价格放在一个表格内，处理成模型可以接受的数据格式，开始测试。

2.3.1 Vetor Auto Regression

VAR是一种多元时间序列模型，用于捕捉多个时间序列变量之间的动态关系。每个变量不仅与自身的过去值相关，还与其他变量的过去值相关。由于相机的发布时间不同，而我们收集的数据又都是近几年的数据，就会导致相对发布日期的天数是不同的，数据合并起来后，数据表格内存在过多缺失值，VAR模型不能接受有缺失值的数据，所以需要在训练之前处理缺失值。有两种处理缺失值的方法，一种是将有缺失值的数据行删除，这显然不行，有些相机时间跨度很大，会导致很大一部分相机在这一段时间内都没有数据，如果删除后，那么剩余的数据不够训练；第二种是填入缺失值，我们选择了这种方法，使用interpolate方法填入线性数据。VAR模型是一个很好的预测模型，能够综合多个时间序列的关系，但是由于我们是人为填入过多数据，导致数据重复，训练的模型是不可靠的，如图2-8所示。

表格

描述已自动生成

图2-8 xxx

2.3.2 Multiple Linear Regression

图形用户界面, 图表, 直方图

描述已自动生成使用多项式回归进行的很顺利，为了减少数据波动对拟合的影响，我们将数据进行平滑处理，以30行数据为窗口进行平均处理，处理后将数据放入模型训练，得到以下结果，如图2-9，2-10所示。

图2-9 xxx

图形用户界面, 文本, 应用程序

描述已自动生成图2-10 xxx

可以看出对于每组数据，使用多项式拟合的结果是非常准确的，得到的R2也是很接近1的。但是单纯多项式回归拟合有一个很严重的问题，就是拟合完成后结尾是发散的，不能达到预测的效果，所以舍弃了。

2.3.3 Support Vector Regression