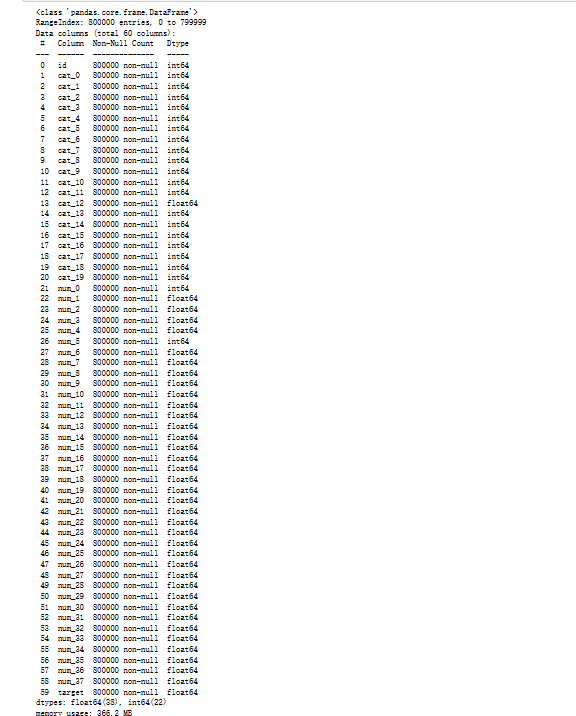
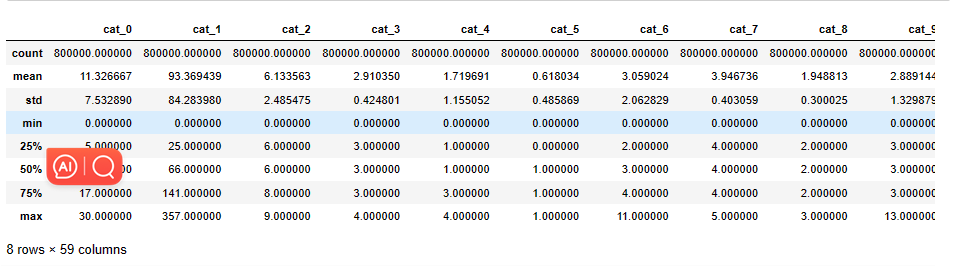
1. 对任务的思考

任务是基于训练集的样本数据，构建一个模型来预测测试集中用户的新增情况。是一类数据挖掘任务，需要人工提取特征并构建模型。用户预测问题可抽象为二分类问题或0~1的回归问题。其中目标是离散型特征和数值型特征，预测该用户是否属于新增用户。需要利用给定的数据集进行特征工程、模型选择和训练，然后使用训练好的模型对测试集中的用户进行预测，并生成相应的预测结果。

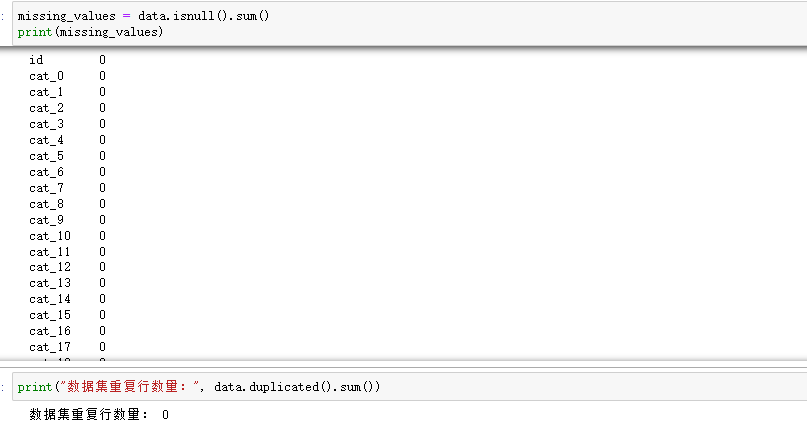
1. 数据的分析和特征工程



根据提供的数据，共有800,000个样本。每个样本包括58个特征，其中前20个是分类特征(cat\_0至cat\_19)，后40个是数值特征(num\_0至num\_37)。目标变量命名为"target"。

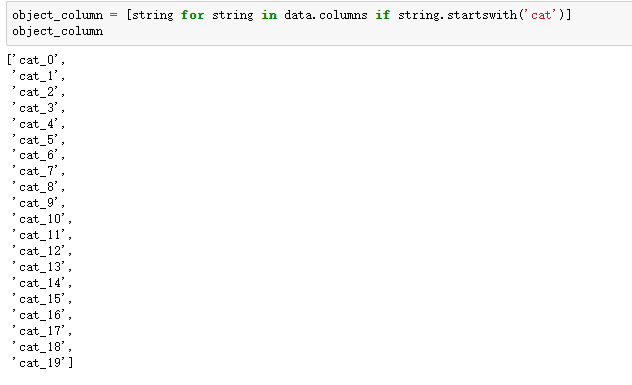


包含了59个特征和目标变量的数据集，特征包括了cat\_0到cat\_9和num\_0到num\_37。这些特征的统计信息显示了它们的数量、均值、标准差、最小值、25%分位数、中位数、75%分位数和最大值



IMG_256

首先查看数据重复值和重复值，均为0，因此不做处理



筛选全部类别变量

IMPORT sklearn.preprocessing.LabelEncoder AS label\_encoder

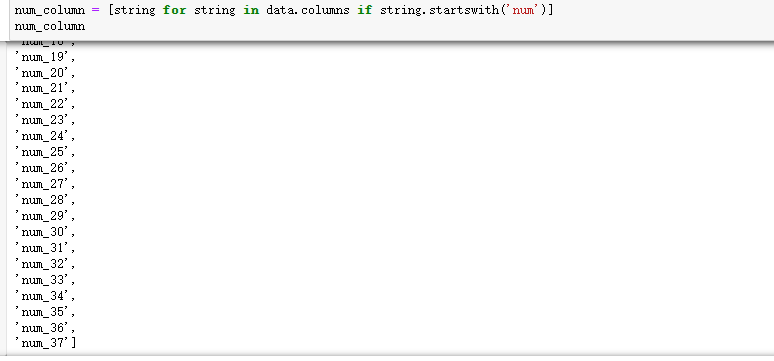
label\_encoder = NEW LabelEncoder()

FOR EACH i IN object\_column DO

data[[i]] = label\_encoder.fit\_transform(data[[i]])

END FOR

使用sklearn中的LabelEncoder类对数据集中的object类型的列进行标签编码。标签编码是将分类变量转换为数值变量的一种方法，它将每个不同的类别映射到一个整数值，从而方便后续模型的处理。在循环中，我们对每个object类型的列进行了标签编码，并将编码后的结果更新到原始数据集中。这样，数据集中的所有列都会被转换为数值类型，以便于后续的特征工程和建模。



筛选全部数值变量

CREATE scaler AS MinMaxScaler()

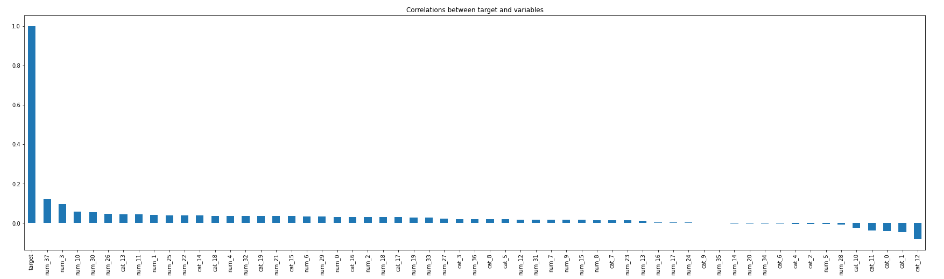
FOR EACH i IN num\_column DO

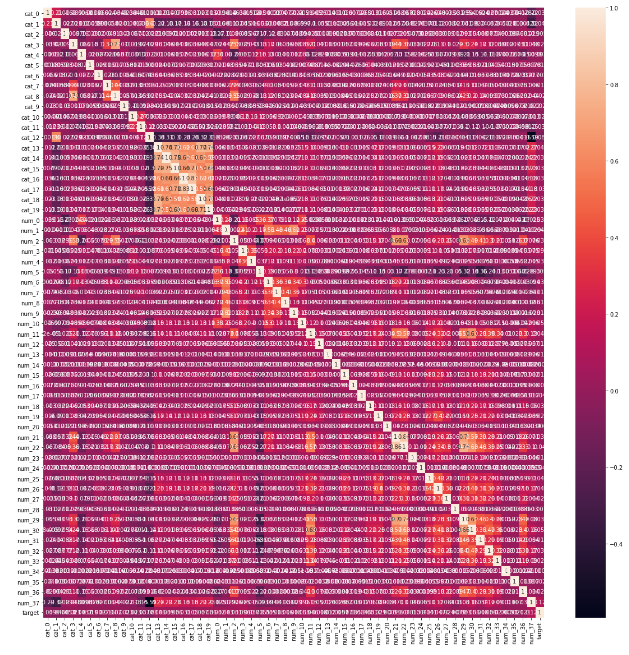
data[[i]] = scaler.fit\_transform(data[[i]])

END FOR

使用MinMaxScaler对数据集中的数值型列进行最小-最大标准化。最小-最大标准化是一种常见的数据预处理方法，可以将数值特征缩放到一个指定的范围，通常是[0, 1]或者[-1, 1]。在循环中，我们对每个数值型的列进行了最小-最大标准化，并将标准化后的结果更新到原始数据集中。通过这种方式，我们可以确保数据集中的数值特征都落在了相似的数值范围内，有利于模型的训练和收敛。这种预处理方法可以帮助提高模型的性能并加快模型的训练速度。

相关性分析





这份相关性分析结果给出了不同特征与目标变量之间的相关性大小，可以用于特征选择和模型建立。其中，num\_37、num\_3、num\_10、num\_30、num\_26等特征与目标变量之间的相关性较高，说明它们对目标变量的预测可能有较大的影响。而cat\_6、cat\_4、cat\_2、num\_5、num\_28等特征与目标变量之间的相关性较低，可能对预测结果贡献较小。

三．模型原理及选择原因

1.Logistic Regression（逻辑回归）是一种广泛使用的线性分类模型，它使用逻辑函数将输入变量与二分类输出变量之间建立关系。适用于处理线性可分或近似线性可分的问题。

用户预测可抽象为二分类问题，逻辑回归模型可处理二分类问题，模型简单而高效，计算复杂度低较低，面对用户预测的大规模数据其也能高效地给出预测结果。

1. Naive Bayes（朴素贝叶斯）是基于贝叶斯定理和特征条件独立性假设的概率分类模型。它在训练阶段通过计算特征的类条件概率和类先验概率来构建模型，然后在预测阶段根据后验概率选择最可能的类别。朴素贝叶斯模型简单、快速，并且在处理高维数据和文本分类等任务中表现良好。

用户预测问题可抽象为分类问题，朴素贝叶斯模型简单高效，训练时间复杂度低，适用于用户预测的大规模数据集和实时分类任务。其基于概率的决策方式，通常能给出较为稳健的分类结果。

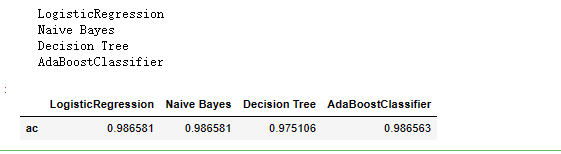
1. Decision Tree（决策树）是一种基于树状结构的分类模型，通过在特征空间中划分数据集来进行分类决策。决策树模型简单直观，可以自动进行特征选择，并且对于非线性可分的问题具有较好的拟合能力。然而，决策树容易出现过拟合的问题，需要通过剪枝等方法进行调优。

用户预测问题可抽象为分类问题，决策树模型对于数据集的分类问题，通常具有较高的准确率。且决策树模型只需要一次构建，预测效率高。同时决策树模型在处理异常值和噪声数据时，仍能保持较好的分类性能。

1. AdaBoost Classifier（自适应增强分类器）是一种集成学习方法，通过组合多个弱分类器来构建一个强分类器。它采用迭代的方式，每一轮根据前一轮分类结果调整样本权重，使得错误分类的样本在下一轮中被更加关注，从而提高整体分类性能。

用户预测问题可抽象为分类问题，AdaBoost可以结合多种不同的分类算法和特征提取方法，根据具体问题选择最优的模型组合，从而更好地处理复杂的用户预测问题。AdaBoost在处理复杂问题和处理噪声数据时表现出色。

四．结果分析及不同模型的比较



根据提供的准确率结果，可以得出以下总结：

1.Logistic Regression（逻辑回归）和 Naive Bayes（朴素贝叶斯）模型在测试集上表现非常接近，准确率均为0.986581。这意味着它们在处理该数据集时能够准确地进行分类预测，且具有相似的性能。

2.Decision Tree（决策树）模型在测试集上的准确率为0.975106。虽然准确率仍然很高，但相对于 Logistic Regression 和 Naive Bayes 模型而言，Decision Tree 模型略有下降。这可能是因为决策树在处理复杂问题时容易出现过拟合，需要进行一定的调优操作。

3.AdaBoost Classifier（自适应增强分类器）模型在测试集上的准确率为0.986563，接近于 Logistic Regression 和 Naive Bayes 模型。AdaBoost Classifier 是一种集成学习模型，通过组合多个弱分类器来构建强分类器，其能力在处理复杂问题和噪声数据方面非常出色。

综上所述，Logistic Regression、Naive Bayes 和 AdaBoost Classifier 模型在这个数据集上表现较好，准确率都达到了很高的水平。决策树模型的准确率稍低，可能需要进行一些调优操作来提高性能。在决策树模型中可以考虑将cat\_6、cat\_4、cat\_2、num\_5、num\_28等特征与目标变量之间的相关性较低的变量删除，做剪枝处理。