结果展示——面向客户（投资人）

1. 动机与目标

近年来，移动支付 APP 在我们的生活中越来越广泛，与此相伴的是，基于移动支付的欺诈交易也越来越多，这些欺诈交易不仅给相关的公司造成了利益损失，也使用户的个人财产遭受损害。因此，对于许多的移动支付产品（如阿里巴巴的支付宝）而言，及时检测并阻止欺诈交易已经变得至关重要。

现在的欺诈交易主要有两种方向，第一种是采用人工抽样检测的方法，但这种方法需要雇佣大量人力而且往往不够及时和准确，效率较低；第二种是基于交易金额为依据进行判定，如 PaySim 数据集中即是自动标记金额大于 200,000 的交易为欺诈交易并阻止。但这个方法主要有两个缺点：1. 无法检测交易金额为 200,000 以下的欺诈交易，这个机制很容易被不法分子发现并规避。2. 这个方法可能会影响用户正常的交易需求。

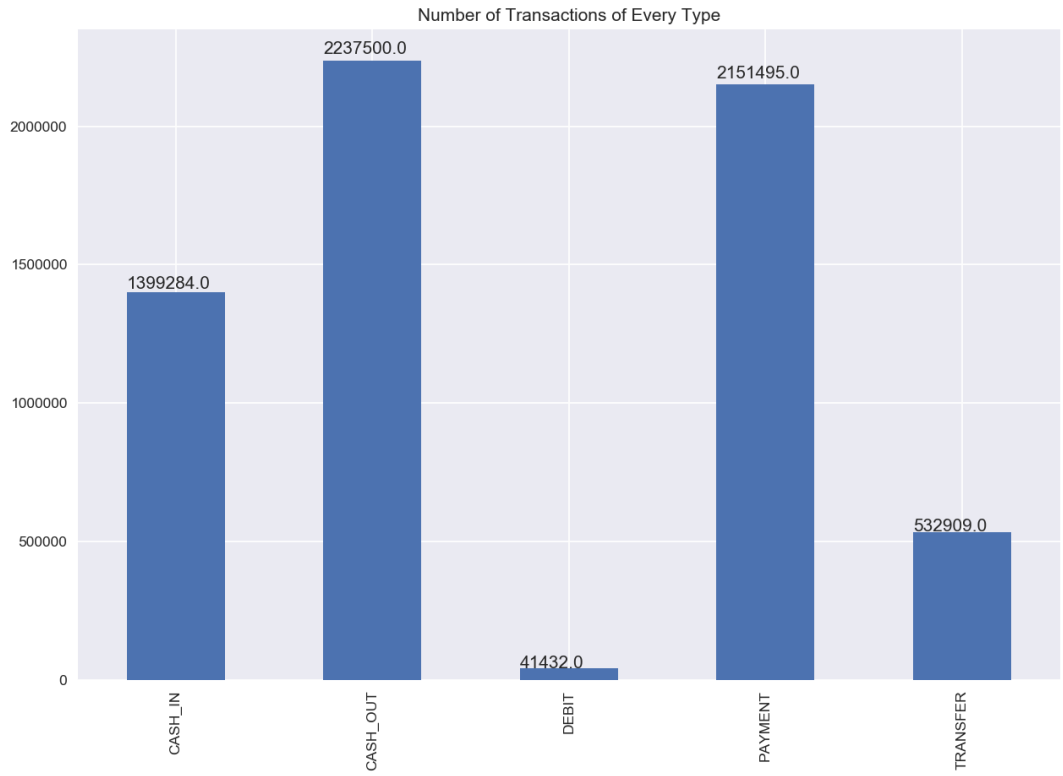
基于这个背景，我们决定基于相关数据集，对欺诈交易的监测进行分析和研究，并利用机器学习的方法设计初步的欺诈交易检测模型。为将来开发和部署具有一定可靠性和准确性的智能欺诈交易检测系统做准备。

PaySim 数据集是一个基于从非洲国家实施的移动货币服务的一个月财务日志中提取的真实交易样本来模拟移动货币交易的数据集。该数据集采用的日志最初是由一家跨国公司提供的，该公司是目前在全球十四多个国家运行的移动金融服务提供商。

2. 以 PaySim 数据集为例的移动货币交易分析

首先，我们对 PaySim 数据集进行了数据分析，以获取较为准确的交易特征信息，为构建机器学习模型做准备。

1. 交易类别分析：

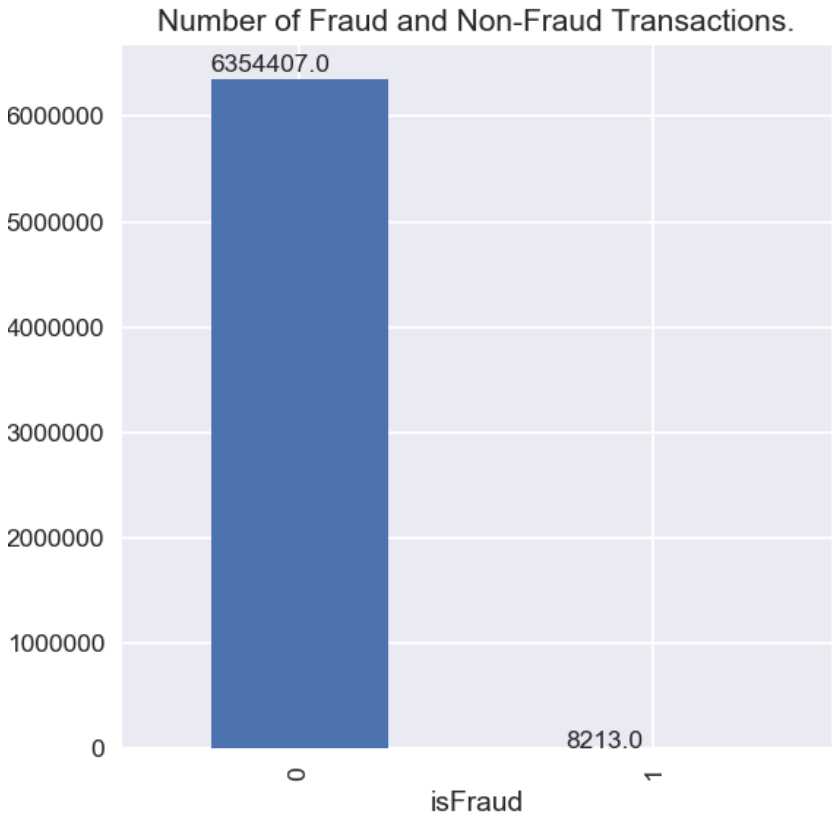


这个图反映了各个类别的交易数量，其中：

- CASH_IN：存钱
- CASH_OUT：提现
- DEBIT：借钱（参考支付宝的蚂蚁花呗）
- PAYMENT：支付
- TRANSFER：转账

由图可知，PAYMENT（支付）和 CASH_OUT（提现）、CASH_IN（存入）是该 APP 最常被使用的服务。

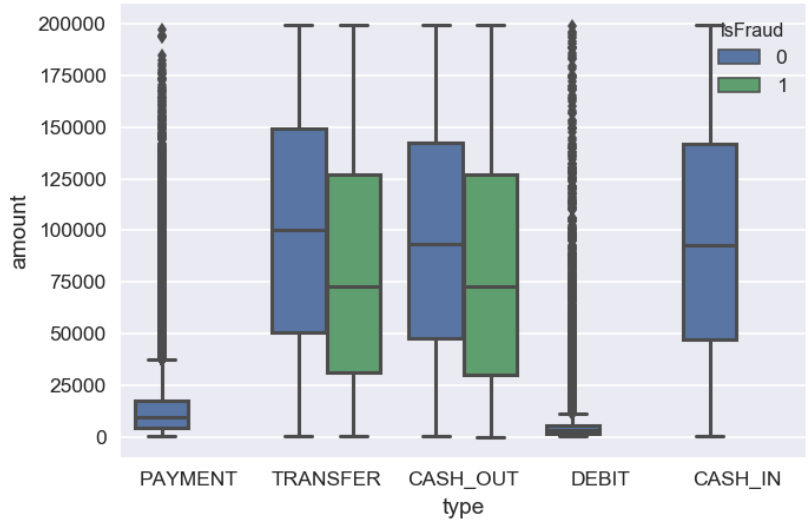
2. 欺诈交易和诚实交易数量（isFraud 为 1 代表欺诈交易）：



由图可见，欺诈交易的数量相对于诚实交易的数量很少。

计算得到欺诈交易数占总交易数的比例大约为 0.129%。

3. 按类别和是否为欺诈交易的的金额（amount）箱型图

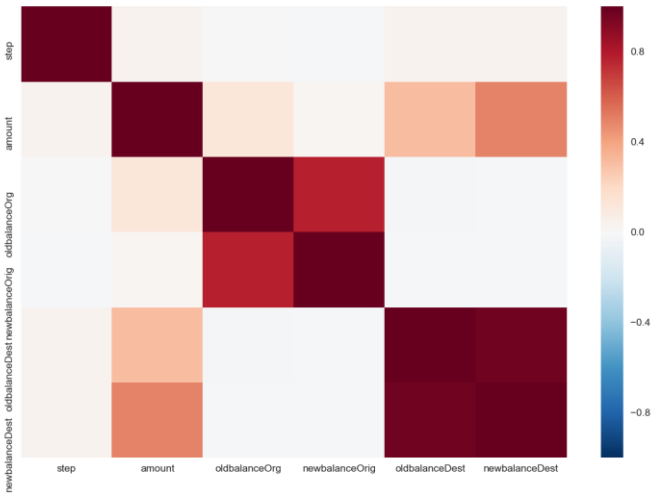


该图体现了该数据集中的欺诈交易只出现 TRANSFER 和 CASH_OUT 两类数据中。

令人意外的是，按类别之后，发现 TRANSFER 和 CASH_OUT 两个类都是诚实交易的涉及的金额更大一些。

很有意思的一点是，绝大部分的 DEBIT 和 PAYMENT 涉及的金额都是很小的，因此有一些离群点（outlier），而相对来说 TRANSFER，CASH_IN，CASH_OUT 就比较大，这一点和人们的直观认知是相符合的。

4. 基于热图（heatmap）的特征关联度分析

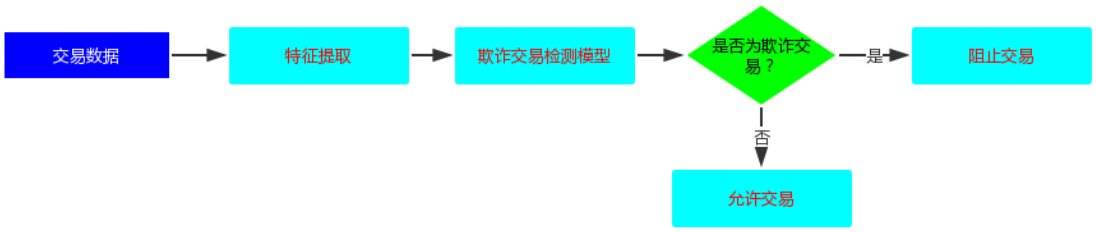


交易之前的目标账户余额（oldbalanceDest）和交易之后的目标账户余额（newbalanceDest）相关度非常高。交易之前的发起交易账户余额（oldbalanceOrg）和交易之后的发起交易账户余额（newbalanceOrig）的相关度也比较高。其中，和交易金额（amount）相关度最高的两个特征为交易前后的目标账户余额（newbalanceDest、oldbalanceDest）。交易时间（step）和任意一个其他特征相关性都不大，一定程度上说明这些特征相对于时间是独立的。

总的来说，这些特征是符合认知也是符合预期的。

3. 欺诈交易检测模型

为了实现自动欺诈交易检测，我们尝试构建一个基于机器学习模型的系统来有效代替人工和现有的其他自动检测机制。该系统的流程如下所示：



我们采用了两类机器学习算法作为基础模型进行比较，最终决定选取决策树模型作为我们开发系统的原型模型。模型细节这里就不详细阐述，我们直接来看模型在测试集上的结果：

* 由于该数据集中的欺诈交易只出现 TRANSFER 和 CASH_OUT 两类数据中，所以下面的评价标准都是建立在这两类数据上面。例如：准确度（Accuracy）$=\frac{\text{预测正确的交易数}}{\text{TRANFER 和 CASH\_OUT 两类交易总数}}$

1. 基准模型：逻辑回归模型的欺诈交易检测表现：

逻辑回归模型的预测准确率为： 0.997406521056

	precision	recall	f1-score	support
0	1.00	1.00	1.00	552446
1	0.55	0.64	0.59	1636
avg / total	1.00	1.00	1.00	554082

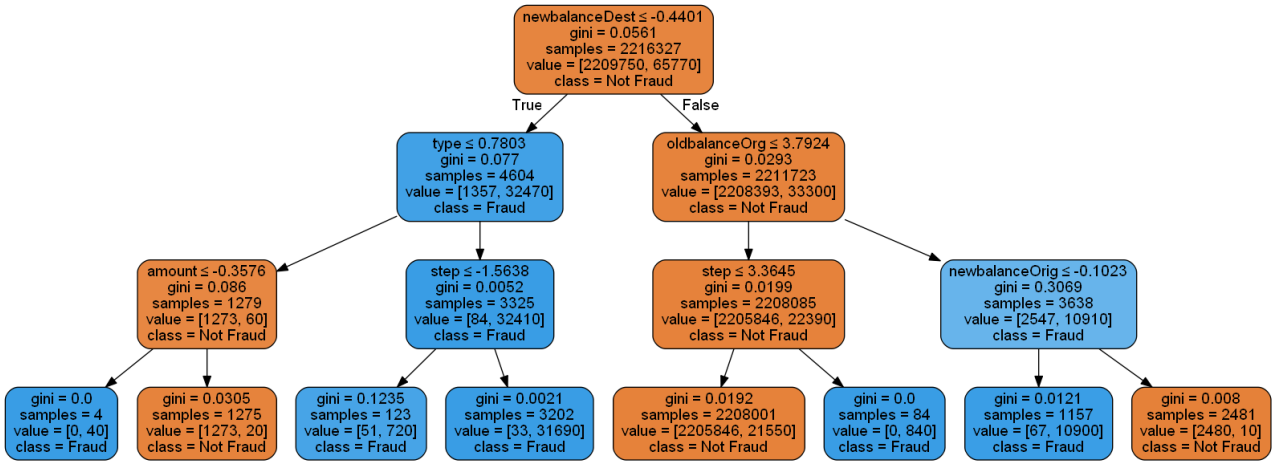
逻辑回归模型作为比较常用的分类方法，总体的预测准确率达到了 99.74%，但是很明显这个数据对于欺诈交易（表中标为 1）的精确率和召回率都比较低，但这正是我们的目标所在。因此显然，逻辑回归模型的表现并不十分尽人意。

2. 主要模型：决策树模型的欺诈交易检测表现：

决策树模型的预测准确率： 0.998980295335

	precision	recall	f1-score	support
0	1.00	1.00	1.00	552446
1	0.97	0.67	0.80	1636
avg / total	1.00	1.00	1.00	554082

我们采用的决策树模型，效果明显，总体准确率达到了 99.89%，并且对于欺诈交易（表中标为 1）的精确度也达到了 97%，召回率相比较逻辑回归模型也有了一定的提升。因此，我们认为决策树模型对于欺诈交易监测更加有效。接下来简单展示一下我们训练得到的决策树：



以上只是我们提出的初步原型模型，实际开发系统的时候我们将会尝试集成方法，即同时训练多种或多个模型来提升模型们的表现。

3. 初步提升：基于集成方法的随机森林模型的欺诈交易检测表现：

随机森林模型的预测准确率为： 0.999025415011

	precision	recall	f1-score	support
0	1.00	1.00	1.00	552446
1	0.98	0.68	0.81	1636
avg / total	1.00	1.00	1.00	554082

为了供大家参考，我们简单地训练一个基于 30 个决策树的随机森林，并对其进行了评估。

结果显示：该随机森林的预测总体准确度已经达到了 99.90%，对欺诈交易的精确率和召回率也都有一定的提升。说明集成模型对于提升模型表现还是十分有效的。

4. 结果与发现

本次工作中我们基于 PaySim 数据集对移动货币支付的交易数据进行了一次全面的分析，并主要利用了两个机器学习的算法初步构建了欺诈交易检测的模型。

我们得到的决策树模型，在最可能出现欺诈交易的两类交易上预测的准确率达到了 99.89%，并且预测欺诈交易的精确度也达到了 97%，这个模型相对于现有的单纯以交易金额判断是否是欺诈交易的机制已经有了十分大的进步，

能够更有效地检测并阻止欺诈交易，减少或避免用户和公司的经济损失。

综上所述，我们提出的基于机器学习的方法为自动欺诈交易检测提供了光明的前景。

5. 未来可能的工作

未来我们还有两个需要攻克的难点：

1.更有效地从交易信息中提取特征。数据的特征对于机器学习模型是十分重要的，一个好的特征对机器学习模型表现地提升有着巨大的帮助。本工作中我们仅仅是简单提取和增强了一些特征，未来我们将会尝试用更多特征工程的方法，优化我们的特征。

2.降低模型的召回率，本工作中得到的决策树模型和随机森林模型已在总体准确率和欺诈交易的精确率上得到了十分可观的结果，但是欺诈交易检测的召回率还有很大的提升空间，这也是我们未来的工作方向。

3. 构建完整的欺诈交易检测流水线。我们未来将会在开发表现更好的模型的同时，构建完整的流水线系统，做到高度自动化，高度实时性和高度准确性。