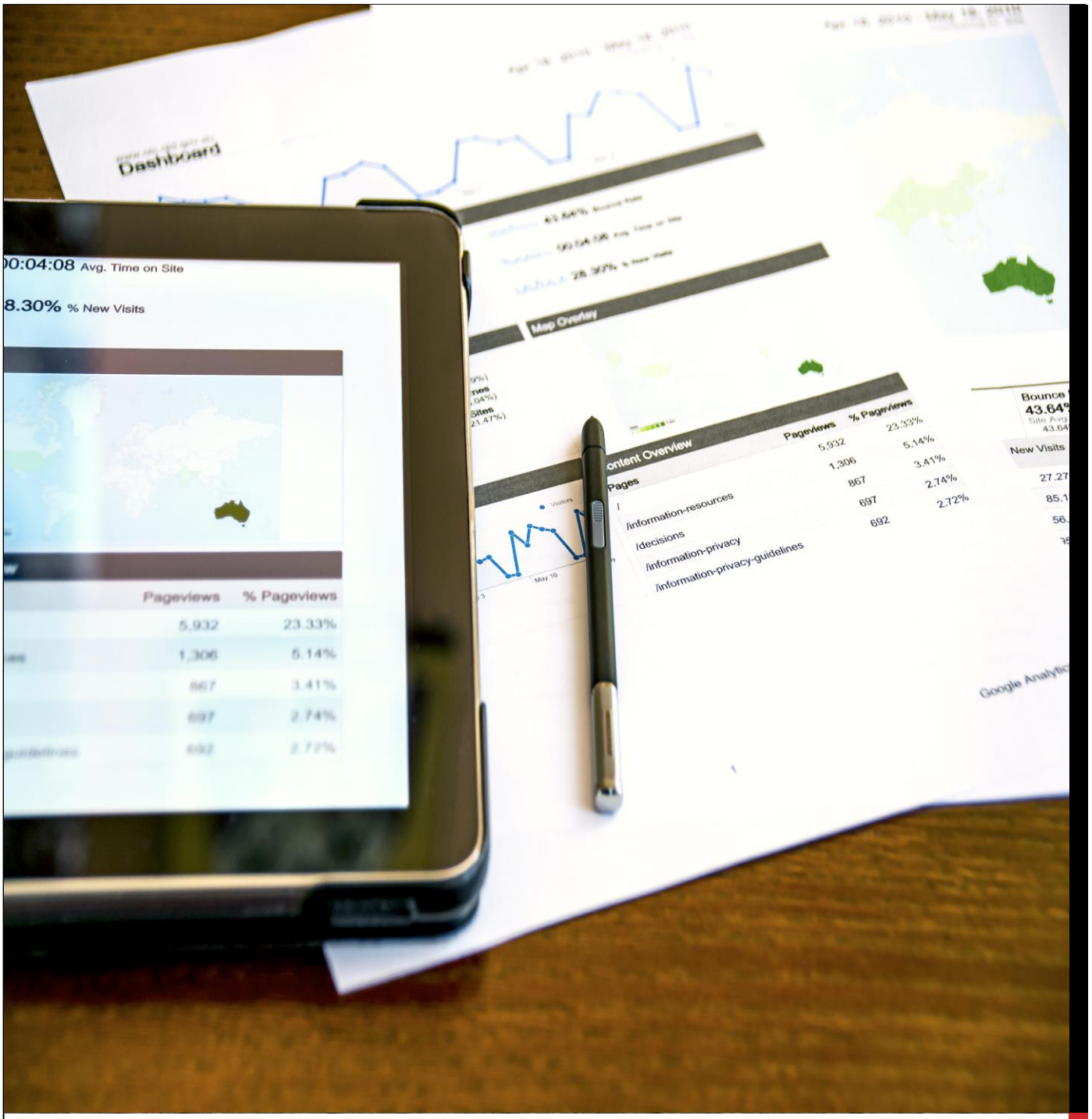

2016 中国金融科技“创新与融合”发展论坛

暨《互联网金融行业分析与评估（2016~2017）》金融蓝皮书发布会

国家金融与发展实验室

2016 年 12 月 20 日

2016 中国金融科技“创新与融合”发展论坛

暨《互联网金融行业分析与评估（2016~2017）》金融蓝皮书发布会

李丹枫

阿里巴巴（友盟+）首席数据官

李丹枫：

我先大概介绍一下。友盟加是今年年初有阿里巴巴收购的3家公司组成的，分别是友盟，CNZZ和DN。这3家公司实际上是中国业界做互联网统计和移动互联网统计领先的3家公司，所以这3个数据体量加起来也非常大。一会我也会介绍到。那阿里收购这3家公司主要是为了丰富她自己的数据内容。因为阿里是强电商数据，但是对于用户在电商场景之外的一些数据，并不是特别清楚。所以我们是帮助阿里去知道，在阿里买东西的这些人，他们平时去哪些网站？看什么内容用什么app？我原来是在CNZZ，然后成立以后我在这个公司做CDO，整体负责数据业务，为什么做风控呢，跟我个人的经历相关，因为我原来在美国工作了5年。所以我想有这些数据的话，我们可以在风控上做一些尝试，我们在年初做了些尝试。后来效果，我觉得还是不错的。所以今天为大家分享一下。

主要是一些结果，因为实际上我们用的用户行为数据。这个可能跟一般的风控的不太一样。因为你讲整个风控的时候，我们说风控数据的金字塔，那最上面肯定是信用相关的数据，你借还款的记录。那下面是一些消费数据。芝麻信用，是跟消费数据非常相关的。那再下面是通讯跟社交，在很多p2p公司自己做风控的时候。他们会去查查你的通话记录，实际上，他们会把这些数据叫做风控里头。最底下是行为数据。这个实际上是我

们拥有的，我们应该是中国体量最大的行为数据。那这个金字塔越往上，跟风控的相关性是越强的。那越往下其实它的覆盖率越高。尤其是在中国金字塔顶端的数据，它的覆盖率还是不够的。它覆盖的很多人群其实也不是普惠金融去借钱的这些人群，所以说在这个群体中我们认为下面的这个数据起的作用会更大。我今天讲的结果基本上完全基于行为数据，所有的结果都是基于行为数据的，我们现在是监测中国110万app，那基本上头5000名app，超过60%的市场占有率。我们监测超过600万家的网站，每天收集的线上线下数据一共是260亿。在双11当天，我们有一个广告监测的产品。双当天我们监测的广告投放的资金是34亿，实际上，按投放量来说的话，可能还需要乘10到20。那么每天可以触达的我们称之为独立活跃的设备，是14亿设备。基本上中国的一个正常的手机，我们在这里面都能看到数据。如果看不到数据。其实还是说明问题的。我们的数据其实包括这几方面，一个是app的使用数据。app的启用关闭，访问时长，等等。网站使用数据，网站浏览内容，设备相关的信息。基于这些信息。我们打了行为标签，另外还包括地理位置信息。这些数据其实收集起来，最重要的是要把他相互连接。连接起来之后，我们会通过人口学的信息进行一个判断，因为这些行为学的数据，你不太可能去判断他是男是女，或者年龄段，但这个好处是我们后面是阿里，所以阿里有很多真实的人的数据，我们可以拿的这些数据做模型去训练。用模型，我们可以去做一个预测判断，因为我们的训练量非常大，所以预测结果还是比较可信的。行为数据包括线上和线下的行为，我们现在已经开始收集一些wifi的数据，这样更丰富我们的数据场景。另外我们会做一些人物的关联，人与物，人与世界，人与空间，人与时间，人与人之间的一些关联。这些实际上是通过我们的数据挖掘。把这些关联起来。关联起来之后我们有很多场景。当然这个其实用的更多的是广告场景。现在我们是用的风控的场景。还有一些其他的特征，这都不说了。

刚才我们提到，我们的覆盖量非常大。这是一个案例。这个实际上是一个p2p公司，他们给了我们一批数据。这批数据有一些是没有逾期的，按时还款的，有一些是逾期的，我们用了最近一星期的数据去进行匹配，发现总体可以匹配到71%左右。比较有意思的发现，如果说对于这些没有逾期的，发现他们的匹配率要比逾期的高很多。这个反过来说明了什么问题？如果你这个手机在我们这

找不到匹配，说明它的这个风险要高。在这个特殊案例里，如果你这个设备在我这找不着，它的风险率要高28%。其实这已经是一个很好的做模型的变量。当然现在我们如果用一个月的数据的话匹配率现在会超过90%，那有的时候会达到99%这种状况。所以这个增强了我们的信息，让我们去做这件事情。包括我们现在跟网上银行，蚂蚁花呗都有这个合作。这个他们也是主要看重我们的数据覆盖率。尽管阿里自身拥有的数据非常强，但是从覆盖率的角度来说，我们还是要比阿里自有的这些数据要大得多。可以看一下，应该非常好理解，苹果的风险要比安卓的风险要低。这张图是不同手机品牌的风险率。可以看到这个案例，里面苹果比较低，htc比较高。

当然了，可能不同的数据源，对它这个数据是从中国哪个区域来的，可能会有差别。但是这些信息相对来说还是比较明显的，所以这些数据都可以用到建模。另外一个这个也非常好理解。价格越低的设备，它的风险越高。因为他这种便宜的设备专门去装很多p2p的app，到处去借钱。因为现在也没有一个机制来知道你在这台借钱，在那台也借钱，其实没有人知道，那它都用一些非常便宜的设备。我们其实自己内部，还有一个设备质量评级。因为我知道你的设备上面很多的行为，我们判断你这个设备是不是一个正常使用的设备。那如果他是一个正常使用的设备我们就定义为1，如果说你这个设备非常不正常。可能你这个是用来刷app的，那他就是4。所以它设备评级本身，跟风险的相关性也非常高了。这个其实我们做，最开始是渠道评级用的，因为一个app要推广，他可能要通过不同的渠道去推广，那可能这种刷量设备就非常多，他们就说你没有办法去判断哪个渠道的刷量设备多。我们其实内部就用了一个比较简单的统计方法，做了一个设备评级。然后用到这个场景，我们发现其实他跟这个风险是非常相关的。刚才介绍一些非常简单的模型，非常简单的指标，就是说基本上拿来原始数据，我们就做一下这种直方图，就可以直接看出来了。那我们真正做模型，实际上是我们，模型本身就开始，应该是6000多维度，然后从6000多维度里选出来的。那这个特征，是包括几个方面，一方面包括app的安装和使用，你这个设备上装了什么app你经常使用什么app？因为我们是SDK植入这个app，所以我们对这个app的使用是非常清晰的。实际上。很多p2p公司它们实际上也在扫设备，也就是说，你的p2p的app它本身的SDK，他会去扫描你这个设备

上装过什么其他的app，但是他不可能特别频繁的去扫描，只能在这个设备开启的时候去扫描，他对app的使用其实信息量并不多。你像我的手机，平时可能装了100个app，但是我用的可能只有10个。所以这个使用信息在我们这边是非常有用的。那另外当然是设备本身了，这刚才其实有很多维度都是设备。另外还有一个就是用户的兴趣。用户的兴趣在我们这边一个比较大的优势，我们可以做这个数据打通，也就是说你移动端的设备和你的这个PC端的设备我们可以打通。那这样的话我们其实可以对人的兴趣有一个更好的判断。那下面是一个地理位置信息，模型用的其实比较简单，都是一些比较标准的模型。

我们其实做了一些特征，看哪些特征最重要，在这里我们发现刚才我们显示了一些直接相关的特征，但是你发现。这些特征没有基本上没有进到最后，这里面还是跟人的兴趣标签以及应用的使用偏好，都是排在最前面的。包括这些聚类ID，也是应用的聚类，这些都是排在前列的。那些信息反而在这个模型里显得不是很重要。我们有两个今年比较早期的时候做的。第一个是个人借贷互联网金融机构，5000到10万。我们这里有20万的样本，用的是这个IMEI，用的是安卓的设备ID。第二个贷款额是从3000到20万，大概10万的样本。用的也是IMEI。第一个结果是一个曲线，当然这个曲线越靠近这个角，说明你这个模型越好。然后那边我们叫KS，这个KS可以达到0.3，给大家一个概念，芝麻信用分，大概KS可以得到0.6。那你想我们这个模型只用了用户行为数据，其他的什么数据都没有加，已经可以达到0.3。所以这个，还是比我想象的要好的。我们最近在跟花呗做的案例，数据量比较大，KS可以得到0.4。0.4的KS其实基本上已经可以独立的去判断借贷的指标。这个是另外的一个例子，但是稍微低一点，KS是0.28。最后我们是把分数归到从300到800，大家可以看到违约率的不同，其实谈这个我要说一点，因为我们用的数据，大家肯定认为是，一个是合规不合规，另外一个是用户隐私的问题。关于合规不合规，我是跟我们的客户说，我是不希望用我们这个分数你去拒绝谁，但是你可以准入。因为这些人里面，他本身金融的属性非常少。我本来对你就没法判断，那我就没办法给你贷款。但是他如果没有那个属性，但是有我们的属性，发现它的分很高，那它就可以作为一个你可以贷款的理由。这样的话在合规上我觉得，不是存在太大的问题。他并没有说，因为您使用htc手机，我就拒绝给你贷款。当然我不知道中国现在

目前规范的制定是怎么样了，如果是在美国，对信用的规则是非常严的。你不能用性别，你不能用年龄，你不能用邮政编码，这些都不可以用。在中国，现在相对属于一个野蛮生长的状态。你什么信息能用，这都可以用。另外分数的稳定性，这个其实是主要给运营者用的，因为他不希望我一个模型到另外一个模型，或者说这个模型本身的，它的分数差太大，所以关系其实不是特别大。我基本上今天就分享那么多。如果大家有兴趣进一步探讨的话，大家可以通过EMAIL联系我。好，就这样谢谢大家。