招商银行沈阳分行信息技术部 总经理 温更

从业务思路上优化

“以业务为核心，以思路为重点，以挖掘技术为辅佐”是数据挖掘实践成功的基础。从业务思路上优化模型是最重要的模型优化措施，对于模型效果的提升是根本性、源头上的突破。从业务思路上优化主要可以从以下几个方向进行考虑。

1.特征标签优化

在数据挖掘建模的工作中特征工程耗时最长，而特征标签的选取是特征工程中的重要环节。例如从海量数据中提取有意义的特征，需要和预测目标呈潜在的强关联关系才能对建模有益，而这些完全依赖于对业务的理解，需要业务专家根据多年的业务经验给出提取思路。

以沈阳分行的金卡客群流失预警模型为例，最初选取的特征标签都是资产类和转账类特征。通过调试模型AUC达到了0.85，准确率60%，从评价指标上看模型的效果优异。但从业务人员应用模型的视角出发，资产类和交易类都是大而全的宏观特征，无法直接通过这些特征总结出有效防止客户流失的策略。评价模型的优劣需要基于模型业务落地应用后的实际效果和业务反馈，进而意识到在原有标签的选取上缺少业务事件背后的真实原因。因此，需要结合业务专家对事件的分析，优化特征指标。

结合业务专家的分析，优化做法是在原有的资产类和转账类特征基础上，加入事件类、APP行为类和经营类等能够指导业务方向的特征，这样根据模型的特征重要性就可以清楚知道如何去挽回客户，会大幅提升模型的效果和模型落地的有效性和准确性。

2.特征标签重构

特征标签重构指在原始特征标签的基础上生成衍生变量，生成衍生变量的目的很直观，即通过对原始数据进行简单、适当的数学公式推导，产生更加有业务意义的新变量，从而来捕捉重要的业务关系。衍生变量允许数据挖掘模型把人类的见解融入到建模过程中，并允许模型利用客户、产品和市场等已知的重要特征。创建衍生变量是数据挖掘过程中最有创意的部分之一。精心重构的衍生变量可以增强模型预测的准确性和可解释性。

以沈阳分行的金卡客群流失预警模型为例，在原始的特征标签基础上通过组合和数学公式推导的方式加入了大量重构的衍生变量，例如将交易金额、交易笔数和交易摘要、交易方向等特征分别进行组合，得到更加细分的交易特征；运用数学公式推导的方式，通过标准差除以均值得到近30天支出笔数的波动系数，通过计算斜率得到近7天交易金额的变化率，还有将原始特征标签减去平均值的去中心化操作等。加入衍生变量后，AUC和准确率都有明显提升，AUC达到了0.87，准确率提升到65%。中心化处理等操作因将绝对值转换成为相对值，并且新加入的衍生变量的特征重要性排名也都非常靠前，进一步增强了模型的解释能力。

3.特征标签前移

数据挖掘利用高于人类的计算能力，帮助人类洞察隐性或潜在价值，甚至得出人类无法感知的结论。因此在做数据挖掘前，必须考虑处理同一业务场景时，业务专家的处理方式。参考业务专家的观点，认为金卡客群是一个庞大的群体，细分客群后再分析金卡客群流失问题效果会更好。

特征标签前移指在建模之前先根据特征标签进行细分客群，其建模思路和作用是对分析对象的一次筛选。细分后的各个群体相比之前的整体对象多了精细化的分割，群里多了共性特征。对精细化的群体分别建模，能显著提升模型的效果。

针对金卡客群流失预警模型，建模之前将金卡客群细分为代发客群、理财产品到期客群、信贷客群等分别进行建模，AUC平均提升4个百分点。

4.特征标签后移

建模结论的业务可解释性也是一个好的数据挖掘项目的重点要求。特征标签后移指在建模结束得到预测结果后，通过分析客户的特征标签来解释业务，制定下一步的营销策略，即将模型结果高效精准落地。数据挖掘的输出是预测的结果，而业务的期望是模型的精准落地。因此得到预测结果后对客户特征标签的分析必不可少。

以金卡客群流失预警模型为例，用XGBoost算法得到两个输出：一是潜力流失的客户；二是模型的特征重要性排名，排名靠前的特征标签即是筛选出的高价值的特征标签。为了实现模型的业务可解释性，下一步即可对这些特征标签进行传统统计学描述性分析，从而得出将模型落地的实施方案。

从技术思路上优化

从建模的技术思路上优化是指在建模的算法和建模技巧方向上进行比较、权衡。建模的算法优化是指不同建模算法的选择过程；建模技巧的优化是指在特定的建模算法基础上，模型特征标签不同抽样方法或不同预处理方法的选择过程。

1.建模算法的优化

大财富管理最核心的能力是在洞察客户、洞察市场基础上为客户创造价值的能力。而“洞察”对科技而言，就是基于数据和算法形成的判断。数据是燃料，算法是引擎。算法和数据都很重要，数据挖掘的研究与实践其实就是在这两个领域发挥能力。针对数据，我行数据的体量、种类、价值等多维度的丰富程度完全能够支撑算法充分的发挥。针对算法，要掌握多种不同的建模算法，同时还要善用集成学习技术。通常，在不同的随机数据集上学习多个分类器能够建立更强大的模型。在算法的选择上，推荐奥卡姆剃刀原则：用能够满足需求的最简单的算法，如非绝对地必要，不要增加复杂性。按照从简单到复杂排序，可以选择的算法包括逻辑回归、决策树、支持向量机、深度神经网络等。不同的建模算法针对不同的具体业务场景会有不同的表现，针对同一个业务需求，可以多尝试不同的建模算法，从中比较、权衡，择其优者而用之。

2.建模技巧的优化

该方向包括参数调整、不同的抽样方式、不同的特征标签预处理方式等，这部分优化在于多做实验，最后选取最优的方式。以抽样方式为例详细介绍，采取抽样措施，主要原因在于如果数据全集的规模太大，针对数据全集进行分析运算不但会消耗更多的运算资源，还会明显增加运算分析的时间。而采用抽样措施，可以显著降低这些负面的影响，在很多小概率事件、稀有事件的预测建模过程中，如果按照原始的数据全集、原始的稀有占比来进行分析挖掘，很难通过分析挖掘得到有意义的预测和结论的，所以对此类稀有事件的分析建模，通常会采取上抽样或者下抽样的措施，即人为的增加样本中的“稀有事件”的浓度和在样本中的占比。对抽样后得到的分析样本进行分析挖掘，可以比较容易地发现稀有事件与分析变量之间有价值、有意义的一些关联性和逻辑性。

建模优化的限度

数据化运营实践中的数据分析和数据挖掘非常强调时效性，在业务需求给出的有限时间里完成优化并投入应用。因此，时间因素是思考适度的主要维度。数据挖掘模型的每一次优化和提升都需要有资源的投入，且满足特定的业务需求。在模型优化和资源投入之间，在投入数据分析资源和满足特定业务需求之间，又有一个微妙的平衡点——性价比，其决定了模型的优化和完善既需要持续

探索又是有限度的。

总 结

大数据蕴含的潜在可能性和海量机会仍有待持续开发。本文结合银行数据分析工作，分析了数据挖掘建模持续优化过程的重要性，通过案例探讨了从业务思路上优化和从技术思路上优化的实践路径。在建模过程中，业务思路上的优化比建模算法上的优化更重要，而建模算法上的优化又比单纯的建模技巧的优化更重要。最后我们也要深刻地认识到，数据挖掘建模的优化绝不仅仅是技术问题，更多需要从业务视角去实现数据驱动和价值交付。

（栏目编辑：杨昆桦）