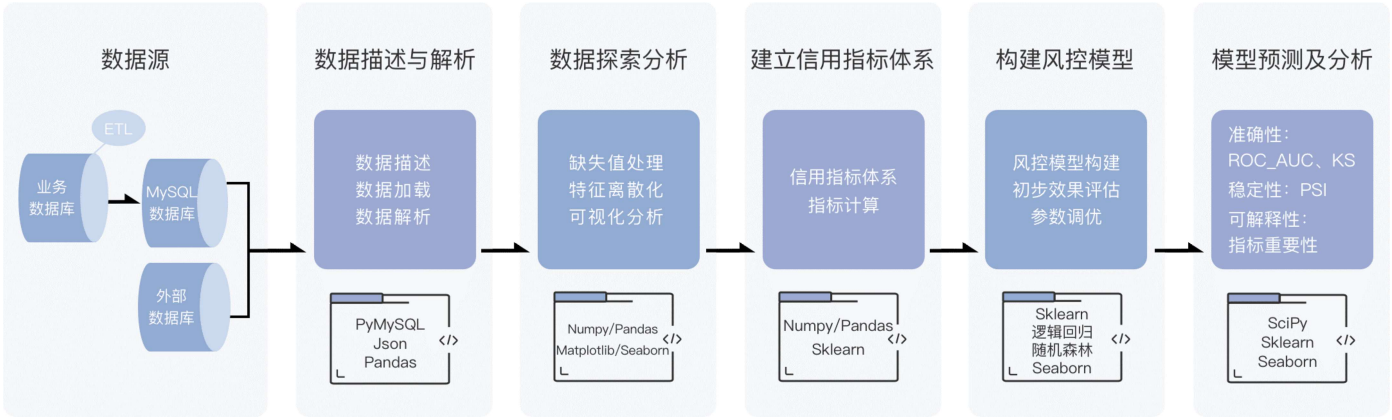
1. 项目背景

信用风险是商业银行长期以来面临的主要风险。个人消费信贷业务成为我国商业银行新的利润增长点，而个人信用风险管理手段的落后成为制约个人消费信贷产业发展的瓶颈。在传统方法中，大部分银行主要依靠信贷审批人员经验来决定，个人喜好对评估结果影响很大，而且随着业务量的大幅上升，人员相对不足，造成审批时间长，效率低，成本高。

随着移动互联网技术、智能设备的普及、大数据技术的提高正改变这一状况。消费分期信贷业务经过长期的开展，积累了大量的用户以及数据。但是面临业务规模难以扩大、坏账率较高等问题。通常的特点包括"授信额度小"、"授信时间要求快"、"人群分散"等，其评估指标多且冗杂，对风险控制要求较高。需要在风控流程简单、风控成本控制在较低水平的同时，保证风控效果。大数据分析能够基于大量多种来源的数据，自动构建个人信用风险评估体系，对我国商业银行发展，特别是消费信贷业务的发展意义重大。



2. 项目目标

银行通过分析客户的个人基本信息、信用历史、消费与偿还能力等指标，利用有监督学习等方法建立个人信用风险评估模型，预测申请贷款的客户是否有违约风险，提供有效信息给决策者，帮助其决定是否向贷款申请人放贷，从而在减少运营成本的前提下，降低坏账风险，对银行进一步扩展个人信贷业务具有十分重要的意义。

3. 数据描述与解析

本项目包含一份银行客户的脱敏数据集，每个客户记录了其公安、银行、互金、运营商和法院等多种来源的信息，一共包含了32个字段。

个人基本信息	三要素验证、城市级别、文化程度、婚姻状况、身份验证、性别、民族、年龄、开卡时长、在网时长
偿债能力	年取现笔数均值、年有取现笔数记录月数、年取现金额均值、年有取现金额记录月数、总取现金额、总取现笔数
信用历史	有无逾期记录、黑名单信息查询记录、法院失信传唤记录、有无犯罪记录
消费能力	年消费笔数均值、年有消费笔数记录月数、年消费金额均值、年有消费金额记录月数、月最大消费金额、网上消费金额、网上消费笔数、公共事业缴费金额、公共事业缴费笔数、年无消费周数占比、总消费金额、总消费笔数

4. 数据探索分析

原始数据中，有些评价指标存在5%-20%的缺失率，如“性别”、“年龄”、“文化程度”等；部分特征存在数据缺失严重的问题（达到60%-80%左右的缺失率），如“身份验证”等。根据特征类型、特征意义我们需要采用不同的缺失值处理方法，缺失过于严重的特征直接删去，其他根据意义进行填充，缺失值填补好之后我们还需对入模特征进行进一步处理如离散化、数字编码等，同时对异常值进行排查和处理。

- 缺失值处理：依据不同特征属性值含义进行不同的缺失值填补处理，如使用默认值；和使用众数填补等。
- 特征编码：对于某些离散型的特征如“婚姻”、“性别”、“三要素信息是否一致”等，由于特征中各个取值之间没有大小关系，我们将其进行One-Hot编码。
- 异常值处理：包括检验离散型特征的唯一取值，看是否为特殊标记值；检验连续型特征的最大最小值，以及通过盒图排除异常值；检验数据取值与业务逻辑不符合的数据等。

同时还将利用Python中绘制图形的库(如Seaborn、Matplotlib)进行一些可视化操作，辅助对数据的预处理工作，同时探究各个特征各自的统计信息及其与预测目标之间的关系。

5. 建立信用评估指标体系

在对数据进行预处理后，我们参照国际信用评分指标体系的建立方式，开始建立信用评估指标体系：分为个人信息、信用历史、偿债能力、消费能力四大指标类。

6. 构建客户风险评估模型

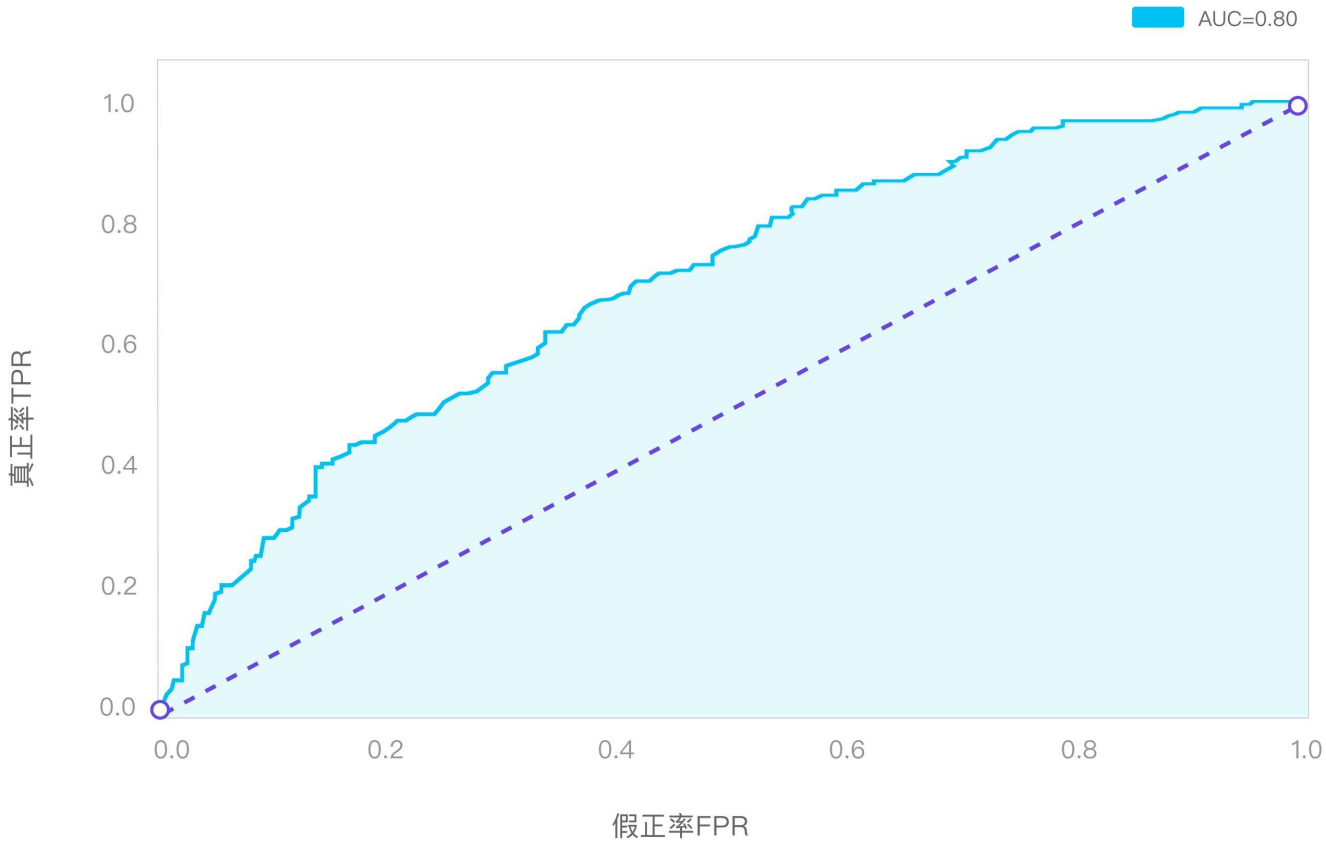
在对数据进行探索分析（包括缺失值填补、异常值检测，以及通过可视化查看各特征分布以及特征之间关系）之后，我们基于信用评估指标体系，开始构建“银行客户信用风险评估”模型。构建风险评估模型的基本流程为：

这里我们使用逻辑回归和随机森林两种算法来进行客户风险评估模型的构建，并分别对两种算法进行了调参优化工作。

7. 模型预测及分析

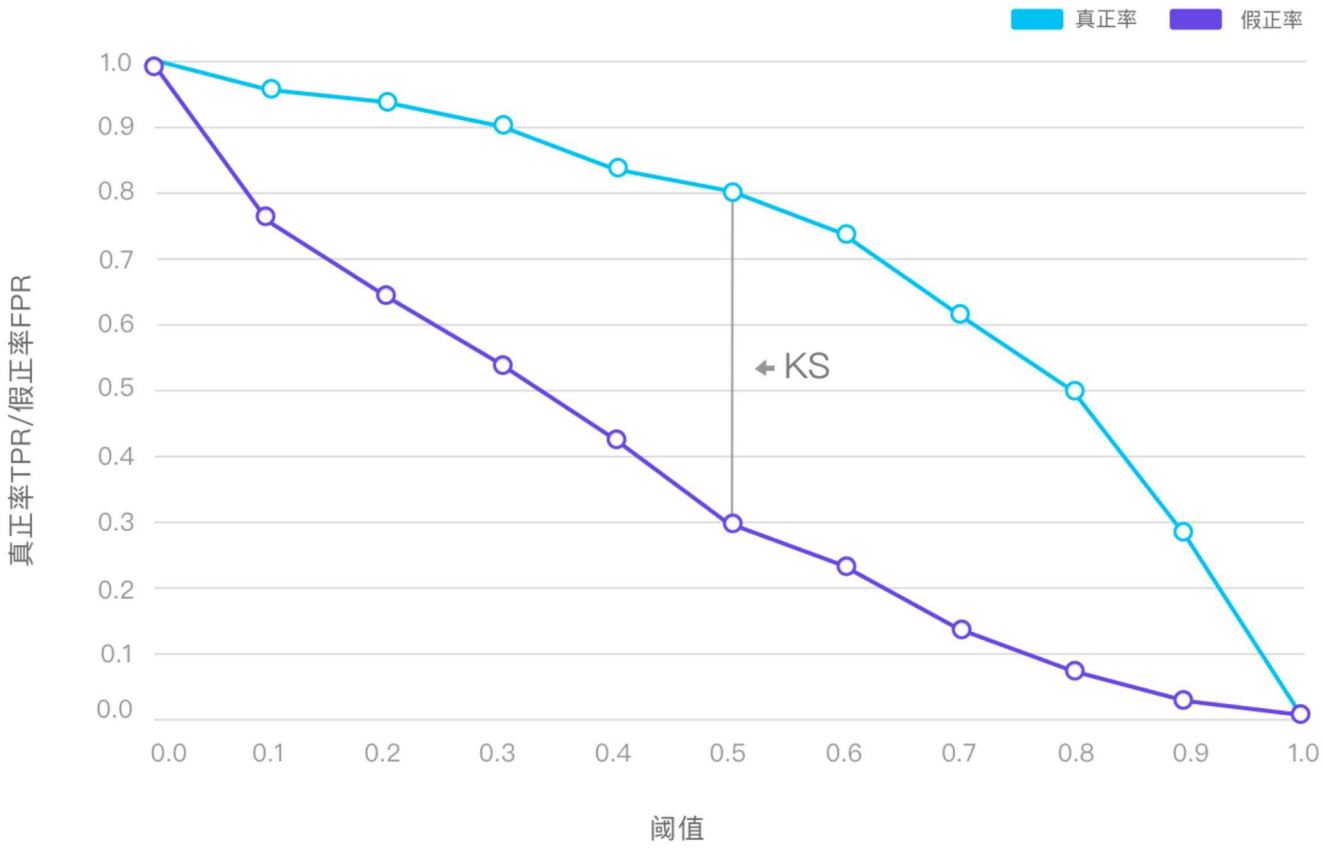
在模型有效的前提下，需要用外推样本对模型进行评估。本项目主要从准确性、稳定性和可解释性三个方面来评估模型。其中准确性指标包括感受性曲线下面积(ROC_AUC)和区分度指标(KS)。

ROC曲线示例如图所示：

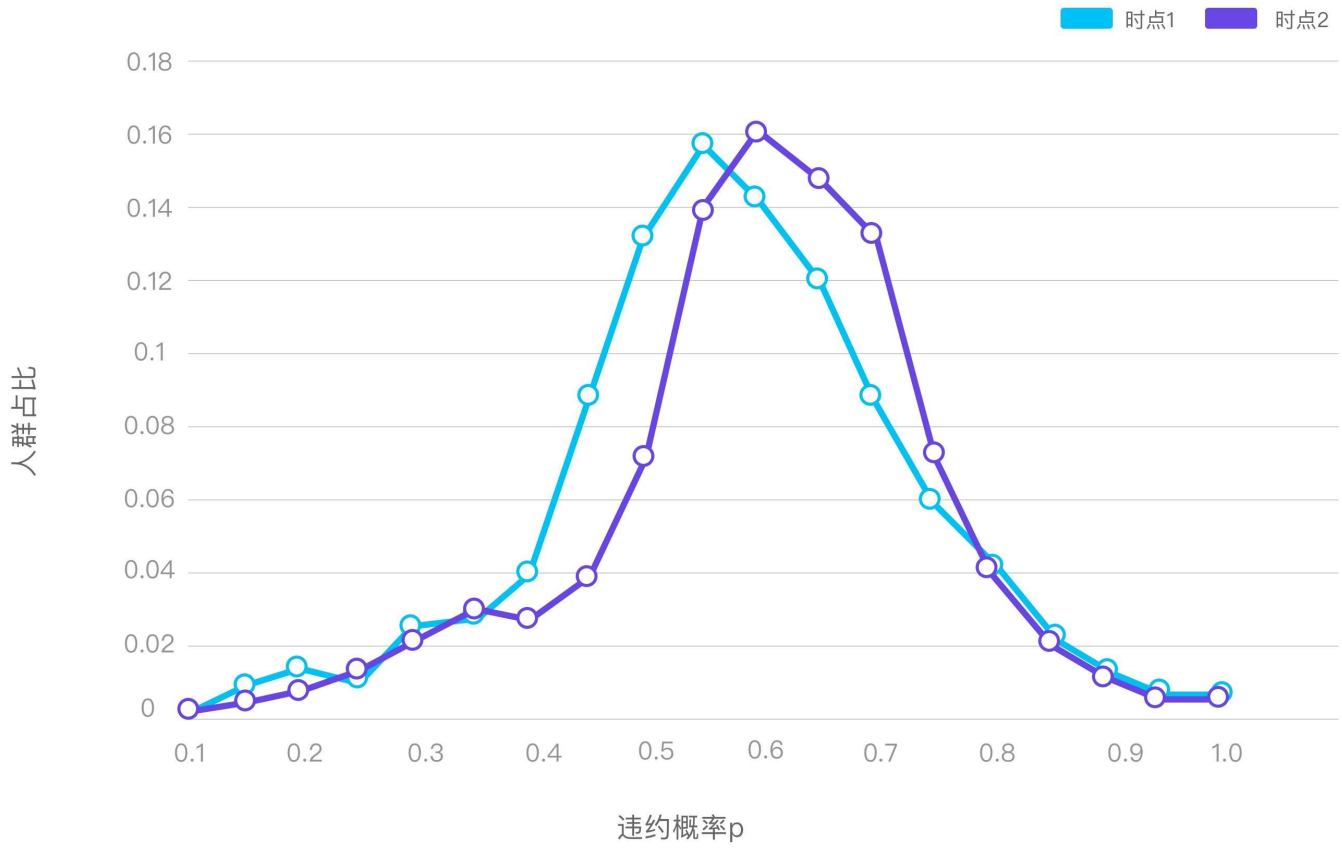


测试集的AUC值代表ROC曲线下方区域的面积。AUC数值范围为[0.5,1]，值越大表明模型相对越好。

KS值通过测量模型对违约和未违约客户的区分能力来评估模型的准确性，示意图如下



稳定性指标主要参考群体稳定指数(PSI)，衡量测试样本及模型开发样本评分的分布差异，其值越小模型越稳定。两个分布的差异示意图如下：



可解释性可通过指标重要度来进行评估，其中指标重要度用于衡量各个解释变量对算法预测结果影响的程度。

8. 项目总结

本项目基于银行、公安、运营商、互联网金融和法院五大数据源的脱敏客户信用数据集。

经过MySQL数据库数据加载、数据预处理和可视化分析等步骤，建立包括个人信息、历史信用、偿债能力和消费能力四大类别38个指标的信用评估指标体系。

基于Python的Sklearn工具，采用逻辑回归和随机森林两种大数据算法构建客户信用风险评估模型，并使用正则化、交叉验证和网格搜索等技术优化模型效果。

从准确性(ROC曲线下面积和KS值)、稳定性(PSI)和可解释性(指标重要度分析)三个方面对客户信用风险评估模型进行综合评估和分析。

本项目开发的客户信用风险评估模型可以支持银行消费分期产品的全流程风控服务，帮助个人信贷以及消费分期业务的进一步开展，扩大业务规模、降低运营成本、降低坏账率。