基于机器学习的银行客户流失预测

2025-11-02

这份 PPT 会把代码路线、实验设计、论文结构、以及从今天到 11 月 23 日的每日安排一次讲清楚。

目标：能按时交付代码、写一篇 80 页英文论文。

当前进度

你现在的仓库里有两个核心文件：

- `EDA.ipynb`：做了初步探索。
- `Train_test_baseline.ipynb`：有一个入门 baseline。

你的论文草稿主要包含：数据集来自 Kaggle 的 Bank Customer Churn，约 1 万行、14 列，目标列是 `Exited`，大约 20% 流失、80% 留存；处理上去掉了 `RowNumber`、`CustomerID`、`Surname`，未见缺失值，做了箱线图看离群，做了基本的类型转换。这些信息我会在我们的正式论文里保留，但需要更系统地写。

总体设计

- 模型这边先把“经典方法”跑全、把评测做足；然后实现一个自己的方法作为创新点；
- 论文这边给出一个清晰的英文章节框架，边做实验边写；最后一周合并结果、补图、润色与排版，确保 80 页足量且逻辑顺。

Baseline 模型清单

先把这些方法稳定、可复现的实验跑出来：

- Logistic Regression（L1/L2 正则，类权重平衡）
- Decision Tree（限制深度，防过拟合）
- Random Forest（含类别重采样或类权重）
- Gradient Boosting：XGBoost / LightGBM（含 AUC 优化、早停）
- CatBoost（原生处理类别特征，强力 baseline）
- SVM（RBF 核，配合标准化与类权重，作为对比）

每个模型都产出：ROC-AUC、PR-AUC、F1、Balanced Accuracy、Brier Score、Top-K 指标 (Precision@K/Recall@K)。

对比用数据集

为了保证论文的可泛化性，建议至少用 3 套公开数据做对比：

- Kaggle *Churn for Bank Customers*（也叫 `Churn_Modelling.csv`）：银行场景，10k 行、若干财务 + 账户活跃特征，是你当前用的这套。链接：kaggle.com/datasets/mathchi/churn-for-bank-customers。数据平衡度约 20/80。
- Kaggle *Telco Customer Churn*：电信场景， 7k 行、20+ 特征，包含资费、合同、付费方式等，流失率约 26%；常用于分类与解释（SHAP）示例。链接：kaggle.com/datasets/blastchar/telco-customer-churn
- KDD Cup 2009 *Customer Churn (Orange)*：工业级别、特征维度大、噪声多，适合做鲁棒性与特征选择展示。链接：kdd.org/kdd-cup/view/kdd-cup-2009

统一的评测指标与选择阈值的“业务化”方案

我们统一三类指标，避免只看 AUC 不做决策：

1. 分类与排序类：ROC-AUC、PR-AUC、F1、Balanced Acc、KS、Top-K（如 Top10% 的 Precision/Recall、Lift）。
2. 概率质量与可用性：Brier Score、校准曲线（Platt/Isotonic/Beta 校准）。
3. 业务收益（重点）：给出成本矩阵，直接用“期望收益”挑阈值。

期望收益的简化公式（样本 i 的客户价值权重 v_i、召回成本 c_{contact}）：

$$\text{Profit}(\tau) = \sum_{i=1}^{N} \Big(\mathbb{1}[\hat{p}_i \geq \tau] \cdot \big(y_i \cdot \underbrace{\text{RetainGain} \cdot v_i}_{\text{挽回的价值}} - (1 - y_i) \cdot c_{\text{contact}} \big) \Big)$$

我们用验证集在 $\tau \in [0, 1]$ 上搜索最大化利润的阈值，并报告该阈值下的各项指标与混淆矩阵。

创新点：AutoCost-Stack（成本敏感 + 校准 + 堆叠的轻量级方法）

目标是“有新意、写得清、可复现、赶得上”。我们设计一个不依赖重模型调参的组合：

1. **价值加权训练**：用客户“价值分”（由 Balance、EstimatedSalary、Tenure、IsActiveMember 归一化后线性加权）作为样本权重，提升对高价值流失的敏感度。
2. **焦点损失的梯度提升**：在 XGBoost/LightGBM 里用自定义 focal loss（γ, α）对难例更敏感，缓解类不平衡。
3. **三路基模型**：LR（可解释）、CatBoost（类别强项）、LightGBM（速度/效果折中），用 5 折 OOF 产生二级特征。
4. **校准与堆叠**：对每一路做 Isotonic 校准，再用带类权重的 Logistic 作为元学习器做 stacking。
5. **双约束阈值**：在验证集上同时满足“预算占比不超过 $B\%$”和“收益最大化/Recall@TopK 达标”，输出最终阈值。
6. **解释与合规**：全流程产出 SHAP、全局/局部解释、分组公平性（Gender/Geography 的差异）与稳定性（bootstrap）。

这套方法实现成本不到 300 行代码，论文可写出一整章方法与一章实验 + 消融。

AutoCost-Stack：详细步骤

1. 数据切分：Stratified 5 折，Geography 分层占比尽量保持。
2. 预处理：数值缺失（如有）用中位数，类别用目标编码（CatBoost 可原生处理）；LR 分支加标准化。
3. 价值分 v_i：Min-Max 归一化后
   $v_i = \alpha\,\text{Balance} + \beta\,\text{EstimatedSalary} + \gamma\,\text{Tenure} + \delta\,\text{IsActiveMember}$，默认
   $\alpha{=}0.4, \beta{=}0.3, \gamma{=}0.2, \delta{=}0.1$。
4. 训练三路基模型：类权重或样本权重 $w_i = v_i \times w_{\text{class}}$；提升模型用 focal loss。
5. 校准：每一路对 OOF 概率做 Isotonic 校准，得到 $\tilde{p}^{(k)}$。
6. 堆叠：元学习器输入 $[\tilde{p}^{(1)}, \tilde{p}^{(2)}, \tilde{p}^{(3)}]$ 与部分强特征交互项，目标是流失概率。
7. 阈值搜索：在验证集上最大化 Profit，同时约束 TopK 的联系人数不超过预算。
8. 解释与报告：特征重要性、PDP/ICE、群体差异、稳定性区间（bootstrap 1000 次）。

实验设计与消融的写作

- **主实验**：三套数据（银行、Telco、KDD 2009），6 个经典基线 + AutoCost-Stack；统一 5 折，统一指标；报告平均与标准差。
- **消融**：去掉价值加权；去掉 focal loss；不用校准；不用堆叠；单阈值而非双约束；逐项剥离看收益变化。
- **稳健性**：不同随机种子、不同 TopK（5%、10%、20%）；不同预算比；不同类不平衡采样策略（SMOTE/None/ClassWeight）。
- **解释性**：全局 SHAP 排名、典型个案解释 3 例；性别/国家两个敏感分组的误差对比与可能的校正建议。

结果表格示例（论文里建议用 `booktabs` 样式）：

Model	ROC-AUC	PR-AUC	F1	Brier	Precision @Top10%	Profit (norm.)
LR	0.84	0.53	0.54	0.17	0.38	0.00
CatBoost	0.89	0.60	0.58	0.15	0.45	+0.12
LightGBM	0.90	0.62	0.60	0.14	0.47	+0.15
AutoCost-Stack	**0.92**	**0.66**	**0.62**	**0.13**	**0.51**	**+0.22**

论文目录

下面是直接可用的章节题目，页数是建议占比，最后会自然超过 80 页：

1. Introduction（6–8 pp）
2. Background and Related Work（12–15 pp）
3. Datasets and Problem Formulation（8–10 pp）
4. Proposed Method: AutoCost-Stack（10–12 pp）
5. Experimental Setup（6–8 pp）
6. Results and Discussion（12–15 pp）
7. Interpretability and Fairness Analysis（6–8 pp）
8. Limitations and Threats to Validity（3–4 pp）
9. Conclusion and Future Work（3–4 pp）
10. Appendices（code/env、更多结果等）（10+ pp）

每章写什么

Introduction：讲痛点、业务目标、数据与约束、方法概览、贡献点、结构安排。
Background and Related Work：定义 churn 场景；银行 vs 电信对比；类不平衡与代价敏感学习；校准/阈值/TopK；解释性与公平性；小结研究空白。
Datasets and Problem Formulation：三套数据的来历、特征表、统计分布；训练/验证/测试切分；问题形式化（二分类 + 排序 + 收益最大化）。
Proposed Method：AutoCost-Stack 的动机、价值分构造、损失函数、OOF 堆叠、双约束阈值、时间复杂度、可复现细节。
Experimental Setup：指标、评价协议、调参范围、计算环境、随机种子策略。
Results and Discussion：主实验表、曲线图、显著性检验、误差分析、与文献对比。
Interpretability and Fairness：全局/个体解释、群体差异、可能的补救措施。
Limitations：数据可得性、可迁移性、预算/阈值假设、外部干预缺失。
Conclusion：结论、业务落地建议、下一步工作。
Appendices：更多图表、参数表、代码与环境

第一章（Introduction）逐段怎么写

按段落来落笔，写起来不会跑题：

1. 场景与痛点：银行留存为什么要紧，行业背景一两段真实的业务化表述。
2. 研究问题：把“预测流失概率并基于预算做排序推荐”的目标说清楚。
3. 数据来源与限制：我们用到的 3 套公开数据，现实里可能没有干预信息，这意味着只能做“流失倾向”而非“增益”（uplift）。
4. 方法概要：一句话概括 AutoCost-Stack 的 5 个要点（价值加权、focal、三路基模型、校准堆叠、双约束阈值）。
5. 贡献列表：条目化写 3–4 点，分别对应“方法”“指标体系”“解释性/公平性”“复现实践”。
6. 论文结构：对应 2–9 章各讲一句话，给读者导航。

第二章（Background and Related Work）逐段怎么写

推荐 6–8 个段落，逻辑从“问题—方法—评价—合规”展开：

1. 流失预测的定义与银行场景特征（与电信/订阅类的异同）。
2. 类不平衡、代价敏感学习与阈值选择（含 TopK/预算约束的动机）。
3. 经典分类器在 tabular 上的表现概览：LR/RF/GBDT/SVM/CatBoost 等。
4. 概率校准的重要性：为什么只看 AUC 不够，用 Brier/校准曲线来约束。
5. 解释性与可审计：SHAP、PDP/ICE，业务落地需要什么样的证据。
6. 公平性与群体差异：按 Geography/Gender 的误差差异与可能的缓解思路。
7. 现有工作的不足：很少把“收益最大化 + 阈值约束 + 校准 + 价值加权”成体系地放在一起。
8. 本文定位：我们在这一空白上做出可复现、可落地的组合方案。

你现在论文里哪些信息能直接沿用

- 数据基本情况描述（表格变量、目标定义、20/80 不平衡）可以保留，但要规范化重写到第 3 章的数据部分，并补上统计图与切分策略。
- “无缺失值、剔除纯 ID 特征、做了基本类型转换与箱线图”这些操作可以写进 3.2 的“预处理流程”，配一两张更清晰的图。
- 目前论文里的 EDA 叙述太粗，我们会加上分布对比、群体差异、相关性热力图、以及流失与价值分的交叉分析。

代码目录与复现规范

- src/：data.py（加载/切分/价值分）、models.py（六个基线 +AutoCost-Stack）、train.py（CV 主循环）、metrics.py（所有指标 +Profit）、plots.py（曲线/解释图）、utils.py（日志/随机种子/配置）。
- notebooks/：EDA.ipynb、Baselines.ipynb、AutoCost_Stack.ipynb、Interpretability.ipynb。
- configs/：YAML 超参与数据路径；reports/：图表与表格导出；paper/figs：论文配图统一来源。
- 随机种子固定：2025，2026，2027 三个；报告平均与标准差。

从今天到 11 月 23 日：代码与论文并行的日程表

今天是 11 月 2 日，DDL 是 11 月 23 日。我们把每一天的目标定清楚：

- 11/02（日）：整理仓库结构；把 `EDA.ipynb` 改成可复用的 `data.py`；定好价值分 v_i 的参数；论文第 1 章起笔（1–3 段）。
- 11/03：跑通银行数据的 6 个 baseline（固定随机种子）；导出首批表格；论文第 1 章完稿。
- 11/04：加入校准与 TopK 指标；完成银行数据的结果图；论文第 2 章开写（相关工作 1–3 段）。
- 11/05：实现 AutoCost-Stack 的 OOF 管道与 focal loss；在银行数据上首次跑通；论文第 2 章 4–6 段。
- 11/06：做银行数据的消融实验；论文第 2 章完稿、润色并加引用占位。
- 11/07：接入 Telco 数据；统一预处理；论文第 3 章数据与问题形式化初稿。
- 11/08：跑 Telco 的 baseline 与 AutoCost-Stack；导出对比表；论文第 3 章完稿。
- 11/09：解释性与公平性分析（银行 +Telco）；论文第 4 章方法起笔（1–3 节）。
- 11/10：完成方法章节（包含算法图、伪代码、复杂度与实现细节）。
- 11/11：写第 5 章实验设定；统一所有表格与图的编号与 caption。
- 11/12：写第 6 章结果与讨论；开始合并 PDF 图表。

答辩 PPT 建议结构

- 背景与问题（1–2 页）：行业动因、研究目标、数据来源。
- 方法（3–4 页）：AutoCost-Stack 的图与要点。
- 实验与结果（4–5 页）：主表、曲线、TopK、Profit、显著性。
- 解释与公平（2–3 页）：全局 SHAP、个案解释、分组误差。
- 业务着陆（1–2 页）：阈值选择与预算、潜在收益区间。
- 总结与展望（1 页）：要点回顾、未来工作。

写作上的注意事项

- 每一章都先给 1–2 段“为什么要看这一章”的动机，再展开细节。
- 表述时多用主动语态和短句，动词优先，少用名词化堆叠。
- 图表优先：任何结论都尽量配一张图或一张表，方便快速审阅。
- 统一术语表：如 churn、retention、propensity、calibration、lift、budget 等。
- 引用占位先打好，最后统一换成 BibTeX。