

# 大数据分析模型设计

——【教学赛】金融数据分析赛题 1：银行客户认购产品预测



武汉大学

学院：工业科学研究所

专业：机械电子工程

年级：2021 级

姓名：朱文康

学号：2021206520006

二〇二二年十二月

## 一、业务调研

用户购买预测是数字化营销领域中的重要应用场景，通过高精度大数据模型分析可高效预测用户购买产品的倾向，从而实现精准营销，最终节约人力资源、提高资源投送效率。本赛题以银行产品认购预测为背景，目的是预测下客户是否会购买银行的产品。

在和客户沟通的过程中，银行会记录和客户联系的次数，上一次联系的时长，上一次联系的时间间隔，同时在银行系统中保存了客户的基本信息，包括：年龄、职业、婚姻、之前是否有违约、是否有房贷等信息，此外还统计了当前市场的情况：就业、消费信息、银行同业拆解率等。

## 二、准备数据

数据从“阿里天池-天池大赛-赛题与数据”板块获得，具体包含 `train.csv`，`test.csv`，`submission.csv` 三个文件，使用阿里云账号登陆并报名参加该比赛后可在赛题与数据板块中下载，获取链接为 <https://tianchi.aliyun.com/competition/entrance/531993/information>。

其中 `train.csv` 包含 22500 条记录用于模型训练，每条记录除 `id` 外有 20 个字段可用于预测，另有 1 个字段是用户是否购买的结果；`test.csv` 中包含 7500 条记录用于大数据模型的测试，每条记录除 `id` 外有 20 个字段用于预测；`submission.csv` 是测试结果提交模板，含有 7500 条记录，每条记录仅有 `id` 和 `subscribe` 两个字段。

### 三、浏览数据

银行系统在和客户沟通的过程中记录了和客户联系的次数、上一次联系的时长、上一次联系的时间间隔，加上系统中保存的客户基本信息如年龄、职业、婚姻等，共有用户相关信息 21 类，其中包括最终是否订阅银行产品结果。数据各个字段含义解释如表 3-1。

表 3-1 各字段含义解释

字段	说明
age	年龄
job	职业： admin, unknown, unemployed, management...
marital	婚姻： married, divorced, single
education	受教育程度
default	信用卡是否有违约: yes or no
housing	是否有房贷: yes or no
contact	联系方式： unknown, telephone, cellular
month	上一次联系的月份： jan, feb, mar, ...
day_of_week	上一次联系的星期几： mon, tue, wed, thu, fri
duration	上一次联系的时长（秒）
campaign	活动期间联系客户的次数
pdays	上一次与客户联系后的间隔天数
previous	在本次营销活动前，与客户联系的次数
poutcome	之前营销活动的结果： unknown, other, failure, success

字段	说明
emp_var_rate	就业变动率（季度指标）
cons_price_index	消费者价格指数（月度指标）
cons_conf_index	消费者信心指数（月度指标）
lending_rate3m	银行同业拆借率 3 个月利率（每日指标）
nr_employed	雇员人数（季度指标）
subscribe	客户是否进行购买：yes 或 no

在统计数据中很多字段的类型都是字符串，如职业 `job`、婚姻 `marital`、受教育程度 `education` 等，为了方便处理，我们将这些字符串使用两种方法编码供后续使用：（1）统一编码为 0-1 之间的数值；（2）使用独热编码将字段向量化。表 3-2 和表 3-3 分别展示了两种不同编码方法下字段值和编码的对应关系。

表 3-2 字符串字段统一编码为 0-1 之间的数值对应表

字段	字段值	编码
job	admin.	0.00
	services	0.09
	blue-collar	0.18
	entrepreneur	0.27
	management	0.36
	technician	0.45
	housemaid	0.55
	self-employed	0.64
	unemployed	0.73
	retired	0.82
	student	0.91
	unknown	1.00

字段	字段值	编码
marital	divorced	0.00
	married	0.33
	single	0.67
	unknown	1.00
education	professional.course	0.00
	high.school	0.14
	basic.9y	0.29
	university.degree	0.43
	unknown	0.57
	basic.4y	0.71
	basic.6y	0.86
	illiterate	1.00
default	no	0.00
	unknown	0.50
	yes	1.00
housing	yes	0.00
	no	0.50
	unknown	1.00
loan	yes	0.00
	no	0.50
	unknown	1.00
contact	cellular	0.00
	telephone	1.00
month	aug	0.00
	may	0.11
	apr	0.22
	nov	0.33
	jul	0.44
	jun	0.56
	oct	0.67
	dec	0.78
	sep	0.89
	mar	1.00
day_of_week	mon	0.00
	wed	0.25
	fri	0.50

字段	字段值	编码
	tue thu	0.75 1.00
poutcome	failure nonexistent success	0.00 0.50 1.00
subscribe	no yes	0.00 1.00

表 3-3 字符串字段独热编码对应表

字段	字段值	编码
job	admin.	[1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
	services	[0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
	blue-collar	[0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
	entrepreneur	[0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0.]
	management	[0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0.]
	technician	[0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0.]
	housemaid	[0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0.]
	self-employed	[0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0.]
	unemployed	[0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0.]
	retired	[0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0.]
	student	[0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1.]
	unknown	[0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1.]
marital	divorced	[1. 0. 0. 0.]
	married	[0. 1. 0. 0.]
	single	[0. 0. 1. 0.]
	unknown	[0. 0. 0. 1.]
education	professional.course	[1. 0. 0. 0. 0. 0. 0. 0.]
	high.school	[0. 1. 0. 0. 0. 0. 0. 0.]
	basic.9y	[0. 0. 1. 0. 0. 0. 0. 0.]
	university.degree	[0. 0. 0. 1. 0. 0. 0. 0.]
	unknown	[0. 0. 0. 0. 1. 0. 0. 0.]
	basic.4y	[0. 0. 0. 0. 0. 1. 0. 0.]
	basic.6y	[0. 0. 0. 0. 0. 0. 1. 0.]
	illiterate	[0. 0. 0. 0. 0. 0. 0. 1.]

字段	字段值	编码
default	no	[1. 0. 0.]
	unknown	[0. 1. 0.]
	yes	[0. 0. 1.]
housing	yes	[1. 0. 0.]
	no	[0. 1. 0.]
	unknown	[0. 0. 1.]
loan	yes	[1. 0. 0.]
	no	[0. 1. 0.]
	unknown	[0. 0. 1.]
contact	cellular	[1. 0.]
	telephone	[0. 1.]
month	aug	[1. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
	may	[0. 1. 0. 0. 0. 0. 0. 0. 0. 0.]
	apr	[0. 0. 1. 0. 0. 0. 0. 0. 0. 0.]
	nov	[0. 0. 0. 1. 0. 0. 0. 0. 0. 0.]
	jul	[0. 0. 0. 0. 1. 0. 0. 0. 0. 0.]
	jun	[0. 0. 0. 0. 0. 1. 0. 0. 0. 0.]
	oct	[0. 0. 0. 0. 0. 0. 1. 0. 0. 0.]
	dec	[0. 0. 0. 0. 0. 0. 0. 1. 0. 0.]
	sep	[0. 0. 0. 0. 0. 0. 0. 0. 1. 0.]
	mar	[0. 0. 0. 0. 0. 0. 0. 0. 0. 1.]
day_of_week	mon	[1. 0. 0. 0. 0.]
	wed	[0. 1. 0. 0. 0.]
	fri	[0. 0. 1. 0. 0.]
	tue	[0. 0. 0. 1. 0.]
	thu	[0. 0. 0. 0. 1.]
poutcome	failure	[1. 0. 0.]
	nonexistent	[0. 1. 0.]
	success	[0. 0. 1.]
subscribe	no	[1. 0.]
	yes	[0. 1.]

图 3-1 使用第一种编码方式将字符串统一编码为 0-1 之间的数值，  
将其他字段全部除以训练集中该字段的绝对值最大值归一化后绘制

频率直方图，每个数字字段的绝对值最大值如下表 3-4：

表 3-4 训练集中数字字段绝对值最大值

字段	训练集绝对值最大值
age	101
duration	5149
campaign	57
pdays	1048
previous	6
emp_var_rate	3.4
cons_price_index	99.46
cons_conf_index	53.28
lending_rate3m	5.27
nr_employed	5489.50

通过数据可视化并结合表 3-1 发现，客户的大部分情况是已知的，未知数据占比较少，因此可以较为准确地分析银行客户在不同维度的分布情况：

（1）年龄（age）：银行服务客户的年龄呈正态分布，主要客户群体年龄范围是 25-40 岁；

（2）工作（job）：客户最多从事管理角色，其次是服务业、经理，客户从事最少的职业是技术工人和个体户；

（3）婚姻（marital）：在婚姻角度可视化发现，已结婚的客户数量最多，其次是单身客户和离婚客户；



(4) 受教育程度 (education)：从客户的受教育角度分析，客户群体的教育水准集中在大学毕业、高中、9年义务教育和职业教育，文盲 (illiterate) 的比例最少；

(5) 信用卡违约 (default)：从客户的信用卡违约角度看，大部分人没有违约记录，确认有违约记录的客户占比小于 1%；

(6) 住房情况 (housing)：在已知情况下，有住房的客户数量略多于没有住房的客户数量；

(7) 贷款 (loan)：据可视化分析，大部分银行产品客户没有贷款，数量约为 17500 人，无贷款的客户在数量上占绝对优势；

(8) 联系方式 (concat)：约 14000 人选择了通过蜂窝网络联系，而剩余约 8000 人选择了电话联系；

(9) 上一次联系月份 (month)：上一次联系的月份多为 5 月份、6 月份、7 月份、8 月份，联系最少的月份为 12 月份；

(10) 上一次联系星期几 (day\_of\_week)：周一到周五几乎是平均分布，没有明显区别；

(11) 上一次联系的时长 (duration)：训练集中最大联系时长为 5149 秒，超过 12000 人的联系时长小于 515 秒；

(12) 活动期间联系客户的次数 (campaign)：结合表 3-4 和图 3-1 分析，有超过 20000 个客户活动期间被联系次数小于 6 次，有客户最多在活动期间被联系 57 次，

(13) 上一次与客户联系后的间隔天数 (pdays)：最大间隔天数为 1048 天，大部分客户的联系间隔天数为 940 天-1048 天之间；

(14) 在本次营销活动前，与客户联系的次数（previous）：最多的联系次数为 6 次，大部分客户在本次营销活动前被联系的次数为 1 次或更少，其他联系次数的客户数量几乎平均分布；

(15) 之前营销活动的结果（poutcome）：超过 14000 个客户在之前营销活动时未存在于银行数据库，而存在的用户是否参与之前营销活动的数量几乎一致；

(16) 就业变动率（emp\_var\_rate）：最大就业变动率为 3.4，相关客户数量为大于 12000 个，占绝对多数；此外几乎所有客户对应时间的就业变动率均为负数；

(17) 消费者价格指数（cons\_price\_index）：最大消费者价格指数为 53.28，最小为 46.88，其分布类似正态分布；

(18) 消费者信心指数（cons\_conf\_index）：消费者信心指数最低为-53.28，所有客户对应的消费者信心指数都为负数；

(19) 银行同业拆借率 3 个月利率（lending\_rate3m）：拆借率最大为 5.27，高于 6000 个客户对应拆借率大于 4.24；

(20) 雇员人数（nr\_employed）：雇员人数最多为 5489.50，最低约为 4720.97，其分布形似正态分布，呈中间多两边少的态势；

(21) 客户是否进行购买（subscribe）：客户购买率总体较低，最终有 19548 个客户没有购买银行产品，未购买率为 86.88%。

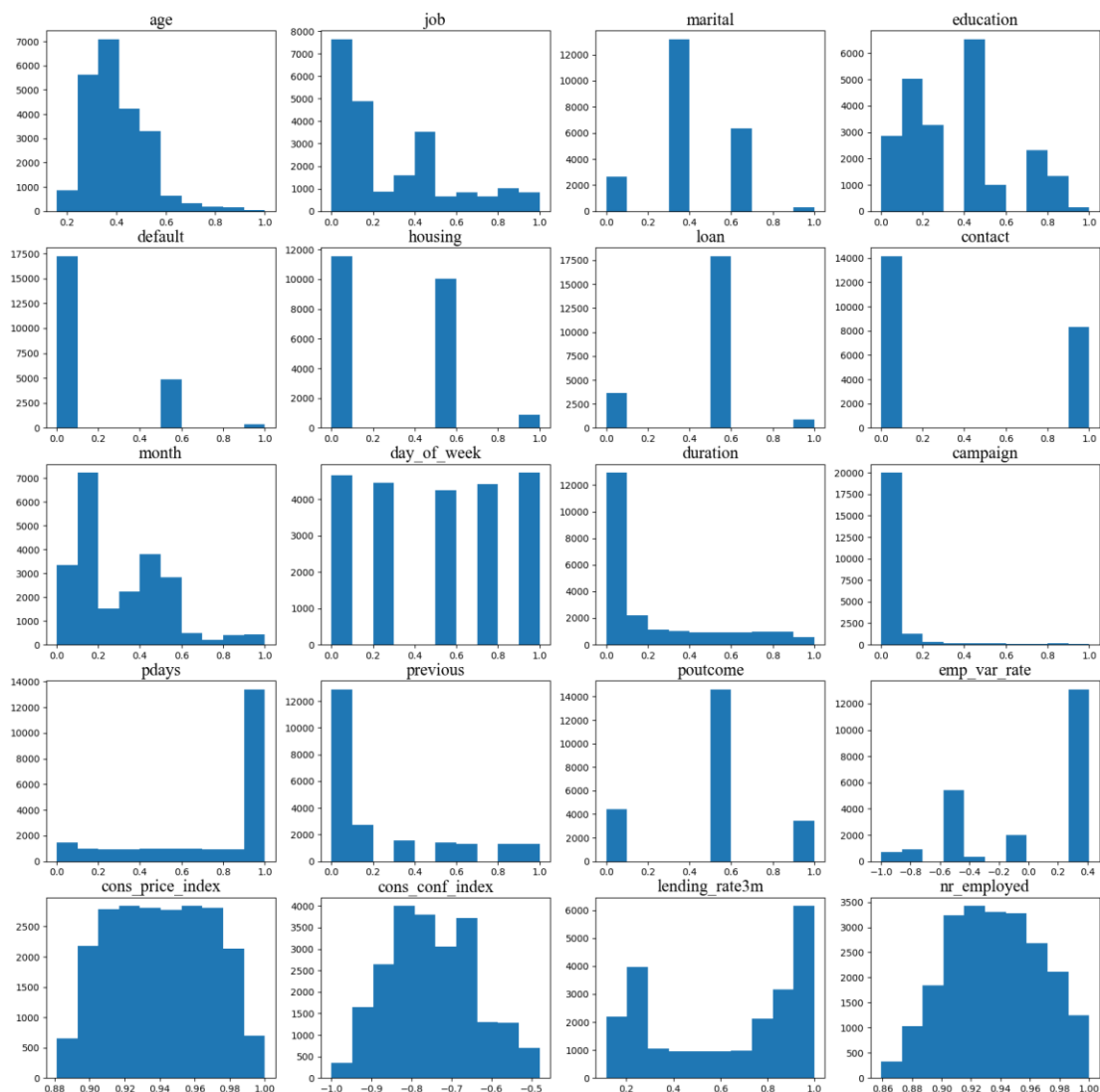


图 3-1 数据归一化频率分布直方图

## 四、变量选择

阿里天池提供了一个 **baseline** 模型，在该模型中使用了 **LightGBM** 利用决策树迭代训练得到最优模型，并使用 **lightgbm** 自带的函数展示各输入字段对结果的重要性可视化图，如图 4-1 所示。可见对预测结果最低的字段是 **poutcome**，由上文分析可知超过 14000 个客户当时在银行系统中尚未存在，其影响力低是可预见的，但对

于训练集中约 8500 个之前在银行有存档的用户而言，该字段仍对其最终决策有一定影响力，因此应该保留。其余具有较低水平的影响力的字段如联系方式、是否有住房、是否贷款、是否信用卡违约等都对其最终决策有一定影响，因此本文将保留所有的 20 个字段作为后续模型的输入。

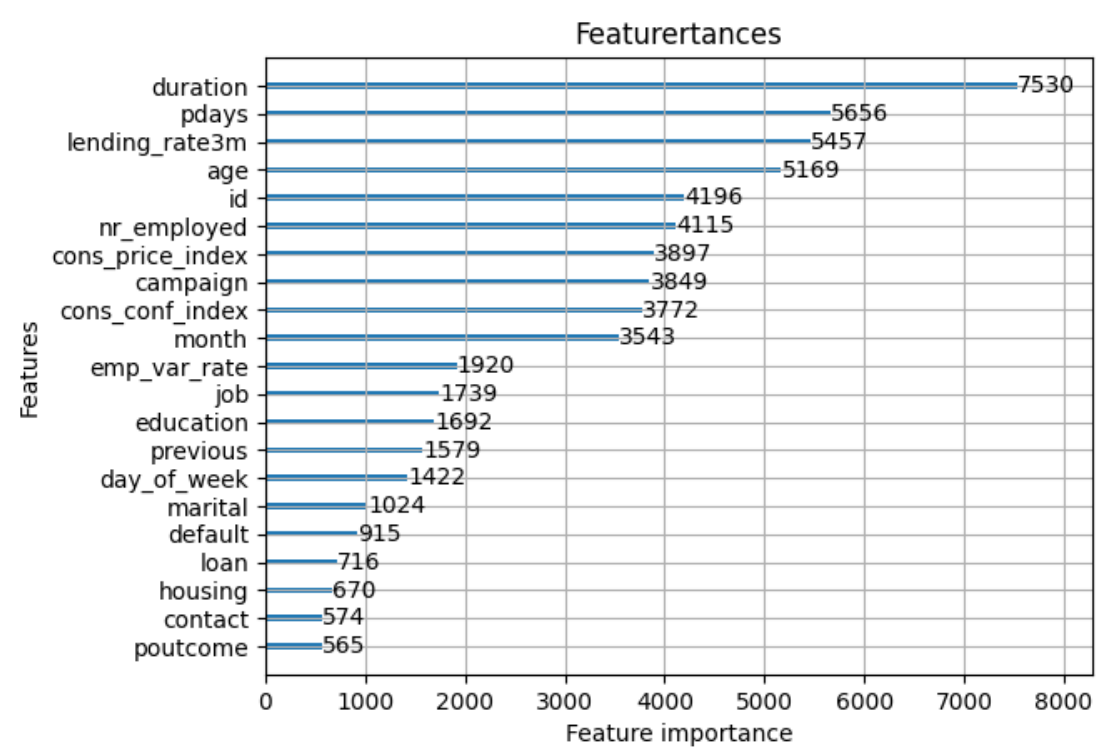


图 4-1 各字段对预测结果的重要性可视化

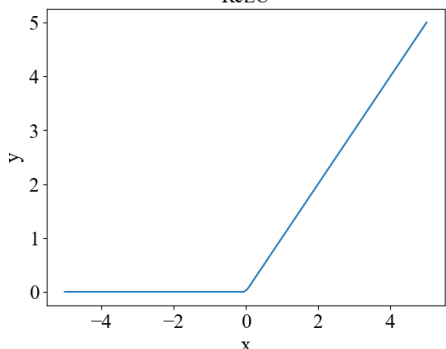
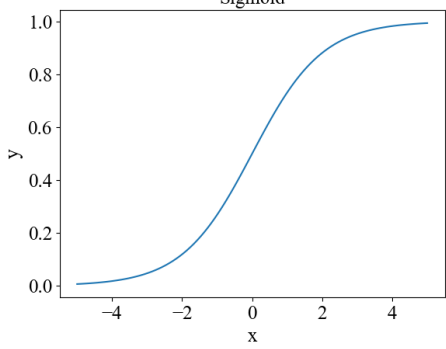
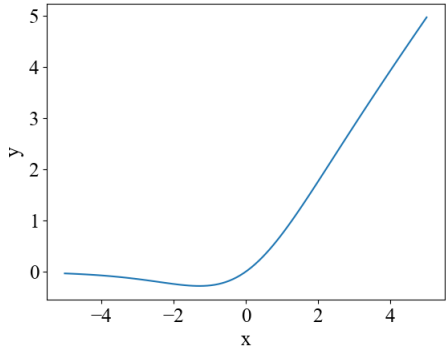
## 五、定义或发现模式

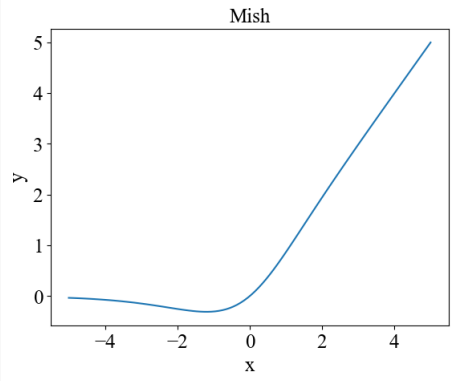
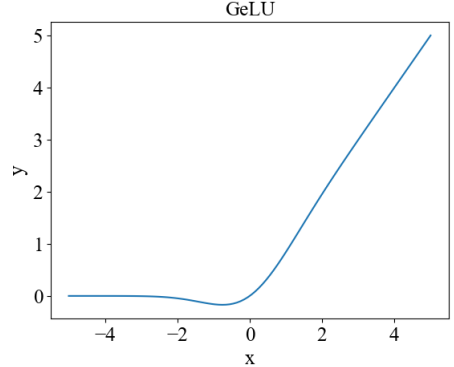
本文采用多个全连接神经网络综合作为预测模型，每个全连接神经网络均为 5 层，每层有 2048 个节点。输入的数据分别使用第三

章中提出的两种方法进行编码，第一种方法将所有数据归一化编码后输入向量元素数量是 20，输出向量元素数量是 1；第二种方法将所有字符串字段独热编码后输入向量元素数量是 63，输出向量元素数量是 2。

激活函数的目标是将神经网络非线性化，通常是连续且可导的函数。本文分别采用了 **ReLU**、**Sigmoid**、**SiLU**、**Mish**、**GeLU** 作为激活函数。**ReLU** 函数使用广泛，指代数学中的斜坡函数，相比于传统的神经网络激活函数，其计算复杂度极低，但其定义域为负数时等于零的特性使神经元在传播中无法被激活，导致神经网络部分失活；**Sigmoid** 函数是使用范围最广的一类激活函数，具有指数函数形状，它在物理意义上最为接近生物神经元，是一个在生物学中常见的 S 型的函数，也称为 S 型生长曲线，其值域为(0, 1)在本文中被用于结果的归一化；**SiLU** 函数就是 **Sigmoid** 加权线性组合，与前述激活函数不同，**SiLU** 的激活不是单调递增的，在[-2, 0]范围内有导数为 0，在神经网络训练过程中起到了隐式正则化的作用；**Mish** 具有平滑、非单调、上无界、有下界等特性，与 **SiLU** 函数类似，**Mish** 函数有下界且导数在下界为零的特点有助于正则化的实现；**GeLU** 函数在 **Transformer** 模型（谷歌的 **BERT** 和 **OpenAI** 的 **GPT-2**）中得到了应用，可以认为是某些函数（比如双曲正切函数 **tanh**）与近似数值的组合，**GeLU** 函数的非线性变化是一种符合预期的随机正则变换方式，能避免梯度消失问题。

表 5-1 各激活函数公式及图形

激活函数	公式、图形
ReLU	$f(x) = \max(0, x)$ <p>ReLU</p> 
Sigmoid	$f(x) = \frac{1}{1 + e^{-x}}$ <p>Sigmoid</p> 
SiLU	$f(x) = x * \text{Sigmoid}(x)$ <p>SiLU</p> 
Mish	$f(x) = x * \tanh(\ln(1 + e^x))$

激活函数	公式、图形
	
GeLU	$f(x) = 0.5x(1 + \tanh(\sqrt{2/\pi}(x + 0.044715x^3)))$ 

本文提出的模型综合了 5 个全连接神经网络模型的结果，如图 5-1，通过继承 PaddlePaddle 的 `paddle.io.Dataset` 类读取 `test.csv` 中的数据，并可选用归一化编码或独热编码方式对表格信息编码，将编号后的输入数据分别输入到使用不同激活函数的 5 种模型中。图 5-1 上部分展示了使用 ReLU 函数的模型结果，其他模型的结构与其相似。输入数据为  $x$ ，经过 5 层全连接神经网络，每层神经网络的节点数量为 2048，每层隐藏层计算完成后将结果先送至 ReLU 函数激活，然后经过一维批归一化将分散的数据重新聚合，再投入到下一层网络计算，在最后一层全连接层输出时，根据数据编码方式选用 Sigmoid 或 Softmax 函数对结果进行归一化或概率化。每个模型获得

test.csv 文件的预测结果后，以 no 为 0，yes 为 1 进行逻辑与操作，获取多个模型对本次会订阅银行产品客户的交集，达成预测共识。

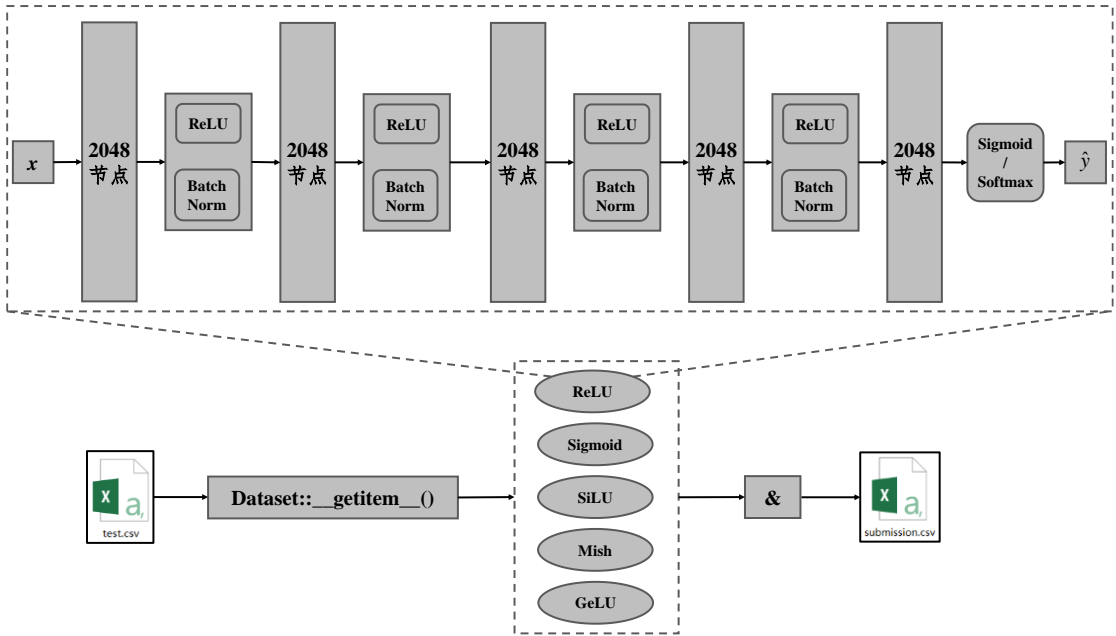


图 5-1 本文模型结构

## 六、计算模型参数

模型使用百度飞桨 PaddlePaddle 定义、训练并测试，模型总参数为 37902339 个，约 144.59 MB。训练过程中设置 batch size 为 2500，epoch 设置为 200，则总共有  $22500/2500 \times 200 = 1800$  步。使用的代码编写 IDE 为 VS Code + Jupyter Notebook，其中 VS Code 用于编写被上层模块调用的底层.py 文件（如网络类定义、数据集类定义），Jupyter Notebook 用于编写训练、测试、可视化等顶层功能。

训练过程中选用 MSE 作为损失函数，MSE 的计算方法如下：



$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2$$

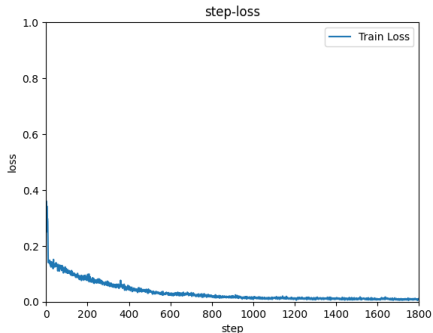
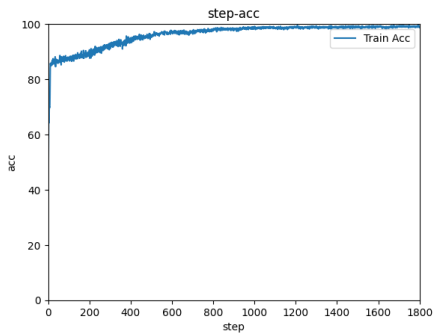
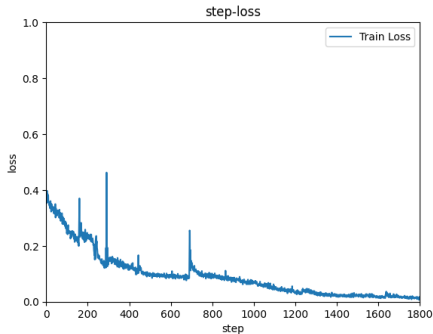
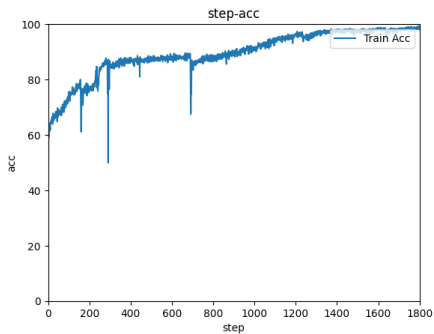
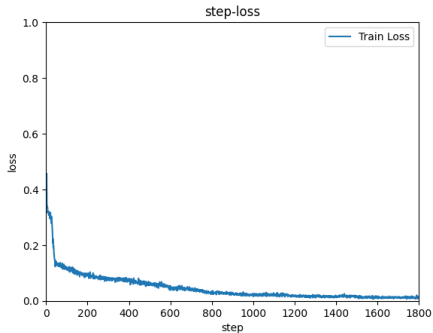
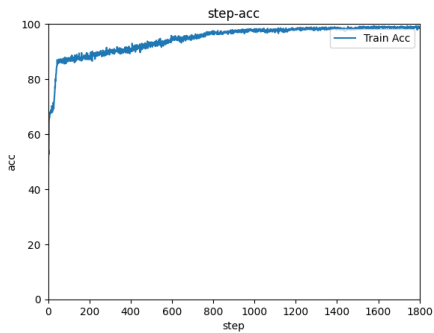
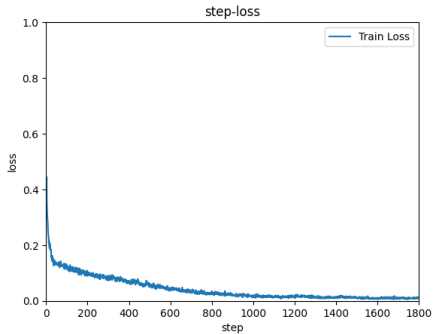
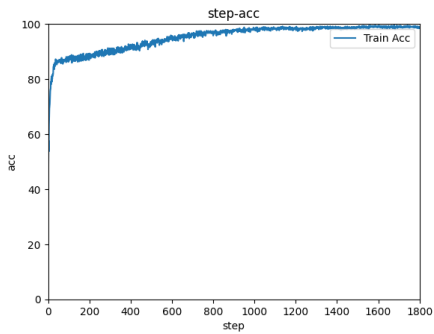
此处， $n$  为向量  $y$  中的结果数目， $\hat{y}_i$  和  $y_i$  分别为模型预测值和模型标签， $MSE$  即对向量中所有元素对位相减后平方，最后取平均值。

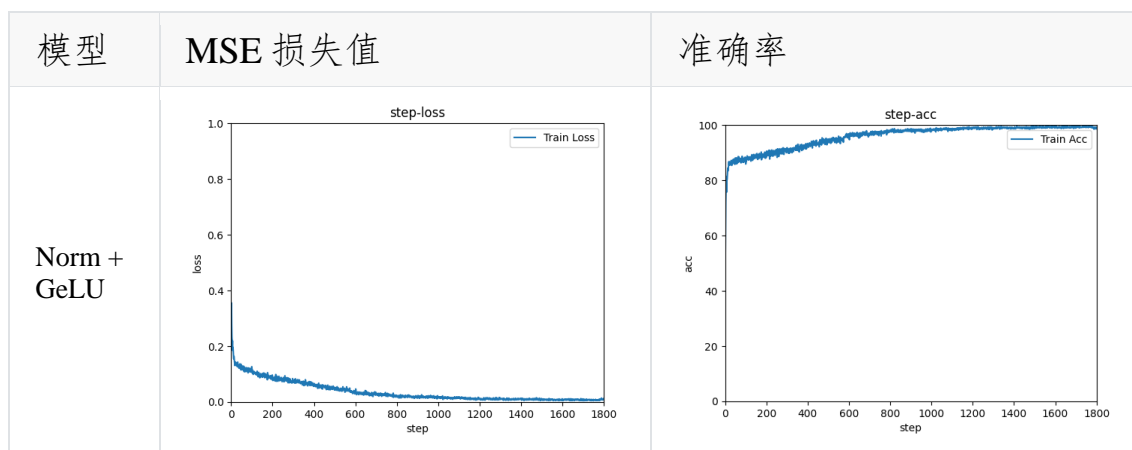
训练过程使用 Adam 优化器优化网络参数，Adam 优化器由 Kingma 和 Lei Ba 两位学者在 2014 年提出，结合 AdaGrad 和 RMSProp 两种优化算法的优点。对梯度的一阶矩估计（First Moment Estimation，即梯度的均值）和二阶矩估计（Second Moment Estimation，即梯度的未中心化的方差）进行综合考虑，计算出更新步长。使用 Adam 优化器训练网络参数可以提高收敛速度，提高模型性能。

由于使用的百度飞桨 PaddlePaddle 框架没有合适的 Metric，本文通过继承 paddle.metric.Metric 类封装了 Accuracy 函数。该函数秉承的思想是将输出取整后与标签对比，统计元素对应相等的个数，除上元素总数即可得到准确率。

使用表 3-2 编码方式对所有数据归一化处理，使用不同激活函数的各模型收敛图分别如表 6-1。在表 6-1 中，可以看到各个模型在训练集上的 MSE 损失值基本降为 0，同时在训练集上的准确率基本达到 100%。其中，Sigmoid 函数作为激活函数的模型收敛不太稳定，可能是网络在传播过程中出现了梯度消失；GeLU 函数作为激活函数的模型收敛最快，最终对训练集的拟合也最好；除 Sigmoid 的其他激活函数在 1000 个训练步之后基本收敛且波动很小。

表 6-1 各模型训练过程中 MSE 损失值和准确率变化

模型	MSE 损失值	准确率
Norm + ReLU		
Norm + Sigmoid		
Norm + SiLU		
Norm + Mish		



## 七、模型的解释与评估

本文提出的模型最终在测试集上取得了 0.93 的准确率，排名 495。如图 7-1、图 7-2，作者昵称是“猪老大”，组建的单人队伍名称是“finishthelesion”。

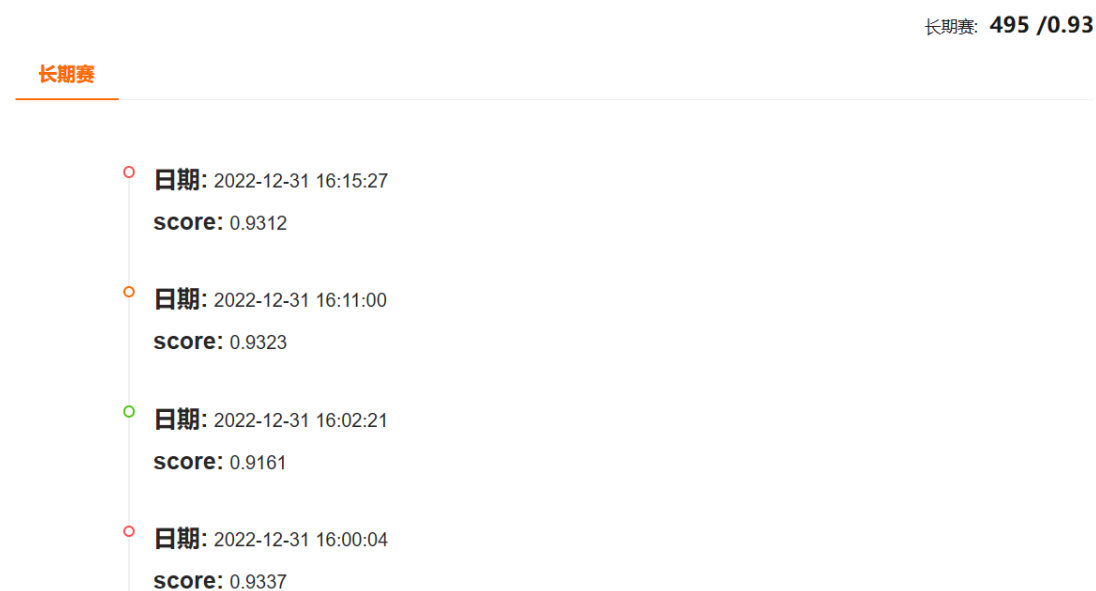


图 7-1 阿里天池后台个人成绩

492	aliyun2146121292	自由自由	0.94	2022-09-13
494	猪老大 1234	电子科技大学	0.93	2022-11-03
495 ↑4	finishthelesson	武汉大学	0.93	2022-12-31
496 ↓-1	1473409592621...	杭州银行	0.93	2022-12-13
497 ↓-1	游客bdwzwlgr5...	中国农业大学	0.93	2022-11-03

图 7-2 阿里天池排行榜

本文中单个模型在训练集取得的准确率基本都是 0.99，多模型联合推理准确率将近 100%，但是在测试集上的误差较大，仅取得了 0.93 的准确率。本文的模型必然存在一些问题，如：

- (1) 过拟合：训练集和测试集准确率相差较大，可通过减少训练轮次、在网络中增加 dropout 层等方法提升模型的鲁棒性；
- (2) 编码紊乱：以表 3-2 中有代表性的 month 为例，结合图 3-1，客户上一次被联系最多的月份是 5、6、7、8 月份，这几个月份具有连续性，然而这四个月份编码却分别是 0.11、0.56、0.44、0.00，中间穿插了其他几个月份，不构成常识上的连续性。具有相似编码问题的字段还有 education、day\_of\_week 等；
- (3) 联合推理粗暴：将几个模型的结果以 no 为 0、yes 为 1，使用逻辑与叠加得到最终结果，事实上会掩盖一些想要购买银行产品的客户群体，造成准确率的瓶颈。

本文代码地址：<http://git.wenyuanhome.top/lankning/2022Bigdata>