天猫用户复购预测

> 2025年12月30日 - 基于Python的问题解析 - 期末大作业

1 问题背景

在电商领域，大型促销活动（如“黑色星期五”、“双十一”或“节礼日大促”）已成为商家吸引流量和提升短期销量的重要手段。然而，大量研究表明，此类促销往往吸引的是价格敏感型的一次性消费者——他们在享受折扣后便不再复购，导致商家投入的营销成本难以转化为长期客户价值。这种“高获客、低留存”的现象不仅削弱了促销活动的实际效益，也对商家的客户生命周期价值（CLV）管理提出了严峻挑战。

因此，如何从海量新买家中精准识别出具有高复购潜力的用户，成为提升营销效率与投资回报率（ROI）的关键。若能提前锁定那些未来可能转化为忠诚客户的群体，并对其实施差异化、精细化的运营策略（如定向优惠、个性化推荐或会员激励），商家将有望显著降低无效营销支出，同时增强用户粘性与品牌忠诚度。

基于这一现实需求，本项目依托天猫平台长期积累的丰富用户行为日志（包括浏览、收藏、加购、点击等多维度交互数据），尝试构建更全面的用户表征，从而挖掘其潜在的消费意图与忠诚倾向。本项目的训练数据集包含约20万新买家，这些用户均在“双十一”期间首次在指定商家完成购买。任务目标是：预测每位新买家在未来六个月内是否会再次在同一家商家发生购买行为。通过构建预测模型，商家可优先对高潜力用户进行资源倾斜，实现从“广撒网”到“精耕细作”的营销范式升级。

这是一个典型的二分类预测问题，基于用户在"双十一"促销期间的行为数据，预测其在未来六个月内再次购买的概率。项目总体的解决思路为：首先，利用用户行为日志数据构建足够丰富的特征体系。通过特征工程和模型调优提升预测准确性，重点关注用户-商家交互行为，挖掘潜在的复购信号。

2 数据说明

- 本数据集为阿里云天池大赛-天猫复购预测赛的公开数据集，具体网址如下：算法大赛-天池大赛-阿里云的赛制。数据集包含了匿名用户在 "双十一 "前6个月和"双十一 "当天的购物记录，标签为是否是重复购买者。
- 数据规模：测试集的规模与训练集相当。
- 字段说明：原始数据共分为四张表

  a. 用户行为日志表

字段名称	描述
user_id	购物者的唯一ID编码
item_id	商品的唯一编码
cat_id	商品所属品类的唯一编码
merchant_id	商家的唯一ID编码
brand_id	商品品牌的唯一编码
time_tamp	行为时间（格式：mmdd）
action_type	包含{0, 1, 2, 3}，0表示单击，1表示添加到购物车，2表示购买，3表示添加到收藏夹

b. 用户画像表

字段名称	描述
user_id	购物者的唯一ID编码
age_range	用户年龄范围
gender	用户性别。0表示女性，1表示男性，2和NULL表示未知

c. 训练数据表和测试数据表

字段名称	描述
user_id	购物者的唯一ID编码
merchant_id	商家的唯一ID编码
label	包含{0, 1}，1表示重复买家，0表示非重复买家。测试集这一部分需要预测，因此为空。

3 数据探索

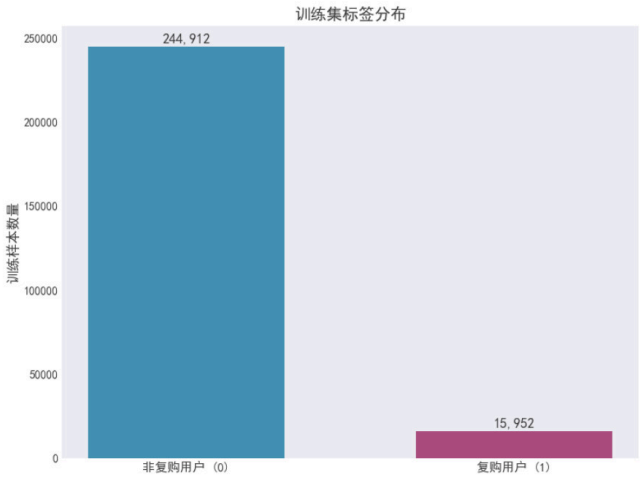
3.1 数据集概览

数据集总共包含四个表，其中训练集和测试集数据量均为26余万，规模相近。用户行为日志数量为5492余万，用户画像数量为42余万。

训练集和测试集均无空值。用户行为日志中，仅字段brand_id存在空值，空值比例为0.17%。用户画像中，字段age_range和gender均存在空值，空值比例分别为0.52%和1.52%。

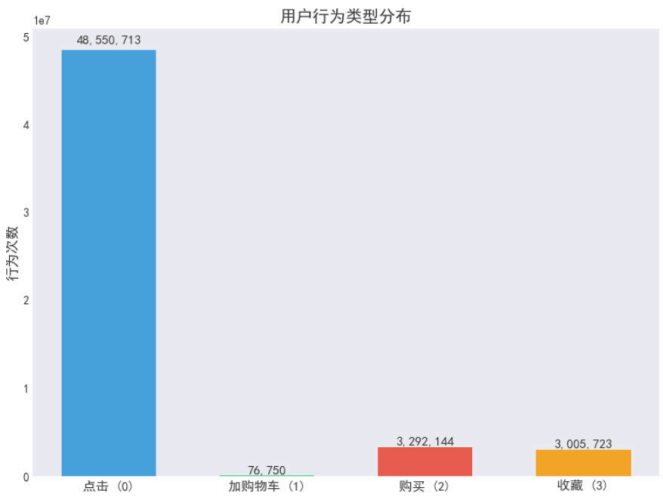
3.2 训练集标签分布

如下图所示，训练集中，非复购用户占比93.88%，复购用户仅占比6.12%。正负样本的比例极端不平衡，非复购用户远远少于复购用户。由于标签分布问题，在后续训练中，可能要考虑欠采样、过采样或调整权重等方法，平衡两种标签对训练模型的影响。



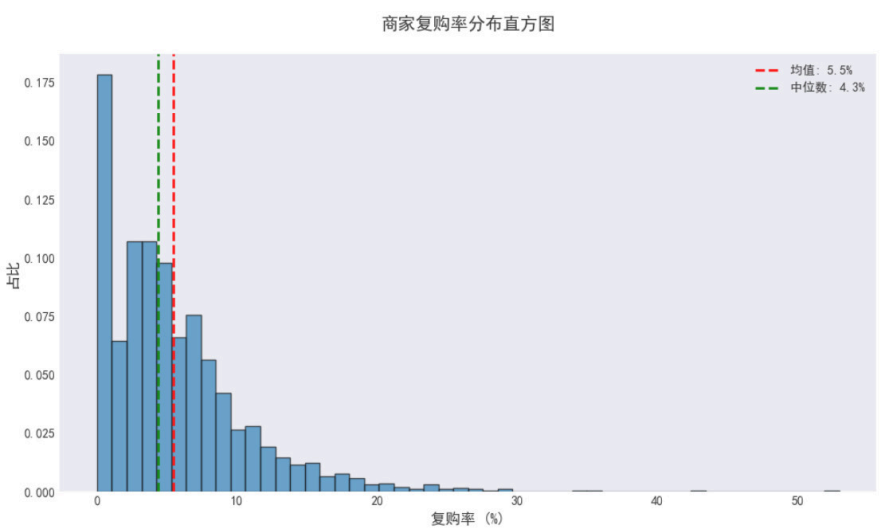
3.3 用户行为类型分布

如下图所示，点击行为占比最大，约占总行为次数的88.39%，购买和收藏行为占比相当，分别为5.99%, 5.47%，加购行为最少。



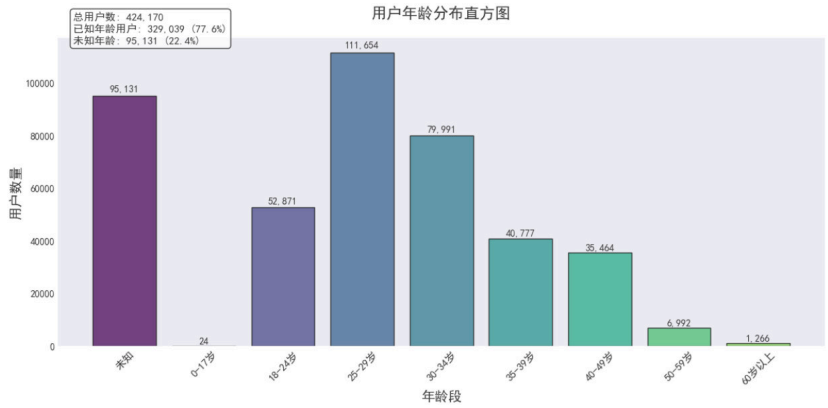
3.4 商家复购率分布

如下图所示，商家复购率平均为5.5%，没有复购的商家占比最大，复购率总体分布呈现为长尾分布，复购率可以达到20%以上的商家非常少。



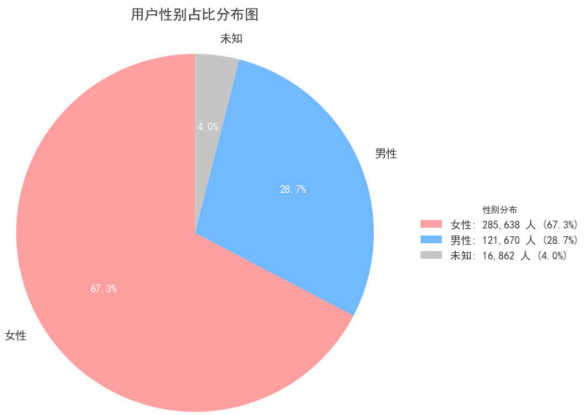
3.5 用户年龄分布

如下图所示，网购用户的年龄集中在18-34岁，其中25-29岁的用户是网购最为频繁的群体。



3.5 用户性别分布

如下图所示，女性占比67.3%，男性占比28.7%。



4 特征工程

4.1 提取特征

项目构建了多层次的特征体系，涵盖用户、商家和用户-商家交互三个维度。

1. 用户级别特征

特征类型	具体内容
基础属性	用户年龄等人口统计特征
行为特征	用户点击、加购、购买、收藏等行为模式特征
多样性特征	商品多样性、品类集中度、品牌覆盖密度
行为时间模式	最早/最晚活跃月份、行为一致性、近期活跃信号等

用户行为特征的构建基于用户行为表（user_log），首先按 user_id 对行为记录进行聚合，统计每位用户的总行为次数，并分别计算点击、加购、购买和收藏四类核心操

作的频次。这些基础计数不仅刻画了用户的整体活跃度，也为后续衍生特征提供了关键分子与分母。同时，我们还统计了用户交互过的商品数、品类数以及有效品牌接触次数，用以衡量其兴趣覆盖的广度。

在此基础上，进一步构造了多组高阶行为特征。通过将各类行为频次除以总行为数，得到点击、加购、购买和收藏的比例，其中购买占比（即购买次数占总行为的比例）尤为重要——它直接反映了行为的“转化效率”。此外，我们模拟典型购物决策路径，计算了从点击到加购、从加购到购买的两阶段转化率，尤其是加购到购买的转化率，被证明是预测复购意愿的强信号，因为它体现了用户将购物意向转化为实际交易的能力与意愿。

为捕捉用户兴趣的集中程度，设计了商品多样性、品类集中度和品牌覆盖密度等指标：例如，商品多样性定义为交互商品数与总行为数之比，值越低说明用户更倾向于反复关注或购买少数商品，往往表现出更高的商家忠诚度，从而更可能产生复购行为。

除此之外，本项目还挖掘了用户在时间维度上的行为模式，以刻画其活跃周期、行为稳定性及近期参与度。这些时间特征对于复购预测至关重要，因为用户的购买行为往往具有明显的时序规律——例如季节性消费、近期活跃用户更可能复购、行为时间分散的用户可能兴趣不稳定等。具体而言，本项目统计每位用户在数据观测窗口内的时间边界与分布特性：包括最早和最晚活跃的月份、活跃过的月份数量等，这些基础时间特征反映了用户整体的时间跨度和日常活跃节奏。在此基础上，还进一步衍生出三类高阶时间特征：(1) 活跃周期特征：用户从首次到末次活跃所跨越的月份数，体现其生命周期长度；(2) 行为一致性特征：衡量用户是否倾向于在固定时间段（如每月月初或周末）进行操作；(3) 近期活跃信号：通过判断用户最后一次活跃是否发生在11月，直接捕捉“近30天内是否活跃”这一强复购先验。挖掘用户时间维度的行为模式可以有效增强模型对用户行为动态趋势的感知能力，尤其有助于区分长期沉睡用户与高潜力回流用户。

所有上述行为特征均通过 user_id 与主特征表进行左连接，并对未出现任何行为的用户对应字段统一填充为0，确保特征完整性。该部分共引入二十余个特征，从多个维度全面刻画用户行为模式。

2. 用户-商家交互特征

特征类型	具体内容
交互行为统计	用户在特定商家的总行为次数,以及点击,加购,购买,收藏的具体频次
交互多样性特征	用户在特定商家交互过的商品数,品类数,品牌数
时间相关特征	用户首次和最近一次与特定商家互动的日期,两者之间的天数跨度

该部分提取用户与特定商家（seller）之间的交互特征。这些特征聚焦于“用户是否曾与该商家互动、如何互动、互动频率及时间跨度”，是复购预测任务中极具判别力的信息——因为复购本质上就是用户与同一商家的重复交易行为。

针对数据集中的每一条用户-商家对，筛选出用户在特定商家下的所有历史行为，并构建交互强度特征和交互结构特征。首先，在交互存在的情况下，统计了用户在特定商家的总行为次数，以及点击、加购、购买和收藏四类行为的具体频次。其中，历史购买次数是最直接的复购先验信号——若用户过去曾在该商家多次购买，则再次购买的可能性显著更高。此外，还计算了用户在该商家交互过的商品数、品类数和品牌数，用以衡

量其在该商家内的兴趣广度；通常，交互商品集中（多样性低）但购买频次高的用户，表现出更强的忠诚度。

其次，还引入了时间维度：记录用户首次和最近一次与该商家互动的日期，并计算两者之间的天数跨度。这一信息有助于判断用户与商家关系的持续性与新鲜度。例如，近期仍有互动且历史跨度长的用户，更可能是稳定客户；而仅有单次远古交互的用户则复购概率较低。

在上述特征的基础上，还进一步衍生出两个关键比率特征：

- 交互强度：总交互次数归一化处理，用于衡量用户对该商家的关注程度；
- 购买强度：即在该商家的购买次数占其总交互次数的比例，反映用户在该商家的“成交效率”。高购买强度意味着用户一旦接触该商家，就容易完成交易，是高质量客户的重要标志。

综上，共构建了十余个用户–商家对级别的交互特征，从频次、结构、多样性、时间动态和转化效率等多个角度刻画了用户与目标商家的历史关系。这些特征直接捕捉了“用户是否熟悉该商家”“是否信任该商家”“是否曾复购”等核心信号，对提升复购预测模型的准确性具有决定性作用。

3. 商家特征

特征类型	具体内容
基础指标	服务过的独立用户数、总购买次数、涉及商品及品类数等
活跃度特征	人气指数、活跃指数
转化效率特征	总购买次数/总行为次数
商品多样性特征	品类数/商品数
规模特征	小型、中型、大型

商家层面的特征，包括商家的规模、活跃度与商品多样性等。这些特征反映了商家在平台生态中的定位和吸引力，对复购预测具有重要价值——例如，高转化率或高复购友好型商家更可能促成用户再次购买。

首先，项目统计了每个商家的核心基础指标：服务过的独立用户数、总行为次数、总购买次数、涉及的商品数及品类数。这些原始统计量构成了商家画像的基础。
在此基础上，函数进一步构建四类衍生特征：

- 活跃度与人气特征：通过将独立用户数和总行为数分别除以其全局最大值，得到归一化的人气指数和活跃指数，用于横向比较商家的受关注程度；
- 转化效率特征：计算“总购买次数 / 总行为次数”，衡量商家将用户浏览、加购等行为转化为实际交易的能力，是反映商品质量和服务水平的关键指标；
- 商品多样性特征：通过品类数与商品数的比值评估商家商品结构的集中或分散程度——高多样性可能吸引广泛兴趣用户，而低多样性可能代表垂直领域专业商家；
- 规模分层特征：根据服务用户数量将商家划分为“小型”（<100人）、“中型”（100–999人）和“大型”（≥1000人）三类，并以独热编码形式表示。这一设计有助于捕捉不同规模商家在用户忠诚度和复购模式上的系统性差异（例如，小商家可能依赖熟客，大商家依赖流量）。

最终，该部分构建了涵盖人气、效率、结构与规模的多维商家特征体系，有效补充了商家侧信息，使模型能够综合判断“用户是否可能在该类型商家处复购”，从而提升整体预测性能。

4.2 特征预处理

对训练集和测试集进行统一的特征预处理，确保输入模型的数值稳定性、分布一致性等，是机器学习流程中保障模型性能的关键环节。具体预处理流程包括：

- 对数据中的无穷值进行清洗，将其统一替换为 NaN。这类异常值可能源于特征构造过程中的除零操作（如计算转化率时分母为0后未完全处理），若不处理，会导致模型训练失败或结果失真。
- 对所有缺失值（包括原始缺失和由无穷值转换而来的 NaN）被填充为 0。多数特征的缺失本质上表示“无此行为”，填 0 符合业务逻辑。
- 对数值型特征进行标准化。将每个数值特征转换为均值为 0、标准差为 1 的分布，有效消除不同特征间的量纲差异，加速模型收敛，并提升基于距离或梯度的算法（如逻辑回归、神经网络、SVM 等）的稳定性与性能。标准化仅应用于数值列，且严格遵循“在训练集上拟合 scaler，在测试集上仅 transform”的原则，避免信息泄露。

5 模型训练

本项目采用统一的交叉验证框架对三种模型：逻辑回归、随机森林、深度神经网络进行系统性训练与调参。所有模型均基于 5 折分层交叉验证进行超参数优化，并在确定最佳配置后于全量训练集上重新训练最终模型用于测试集预测。

5.1 逻辑回归

采用网格搜索对关键超参数进行全面扫描，搜索空间包括：

- 正则强度 C ∈ {0.001, 0.01, 0.1, 1, 10, 100}
- 正则类型 penalty ∈ {’l1’, ’l2’}
- 求解器 solver ∈ {’liblinear’, ’saga’}
- 最大迭代次数 max_iter ∈ {100, 200, 500}
- 类别权重 class_weight ∈ {None, ’balanced’}

优化目标为ROC-AUC，以匹配复购预测任务对排序能力的核心需求。最终模型使用网格搜索返回的最佳参数组合在全训练集上拟合，并输出概率预测。

5.2 随机森林

鉴于随机森林参数空间较大且训练成本较高，采用 随机搜索策略，在有限计算预算下高效探索高维超参空间。搜索范围包括：

- 树的数量 n_estimators ∈ {100, 200, 300}
- 树深度 max_depth ∈ {10, 20, 30, None}
- 节点分裂最小样本数 min_samples_split ∈ {2, 5, 10}
- 叶节点最小样本数 min_samples_leaf ∈ {1, 2, 4}
- 特征采样策略 max_features ∈ {’sqrt’, ’log2’}
- 类别平衡策略 class_weight ∈ {None, ’balanced’, ’balanced_subsample’}

随机搜索执行 20 次迭代，每次从上述分布中随机采样一组参数进行 CV 评估。该策略在保证探索广度的同时显著降低计算开销。

5.3 深度神经网络

针对 DNN 训练耗时长、超参敏感的特点，设计 两阶段调参流程：

- 快速调参阶段：在 20,000 样本子集 上对学习率（[0.001, 0.0005]）、Dropout 率（[0.2, 0.3]）和批量大小（[64, 128]）进行粗粒度网格搜索，利用早停（EarlyStopping）和学习率衰减（ReduceLROnPlateau）控制训练。
- 交叉验证评估阶段：将筛选出的最佳超参组合应用于完整训练集的 5 折 CV，每折独立训练并记录验证 AUC，以评估泛化稳定性。

网络架构采用 4 层全连接 + BatchNorm + Dropout 的标准结构，并引入 L2 权重正则化防止过拟合。最终模型在全训练集（含验证集划分用于早停）上重新训练。

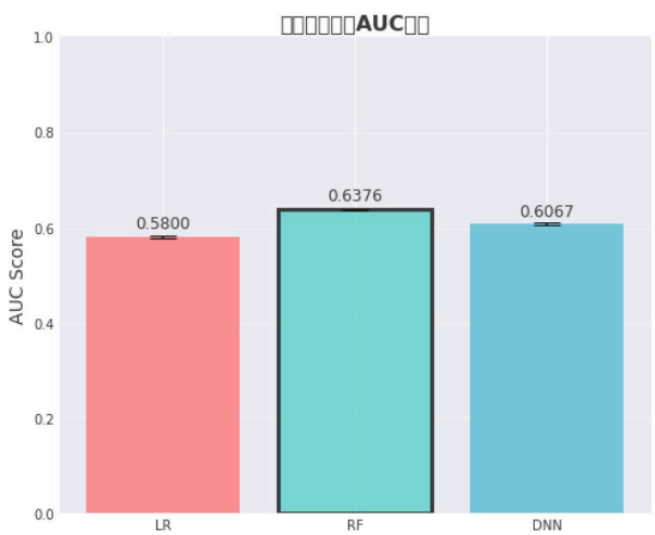
6 结果分析

模型	交叉验证平均AUC	标准差
随机森林（RF）	0.6376	±0.0062
DNN	0.6067	±0.0091
逻辑回归（LR）	0.5800	±0.0087

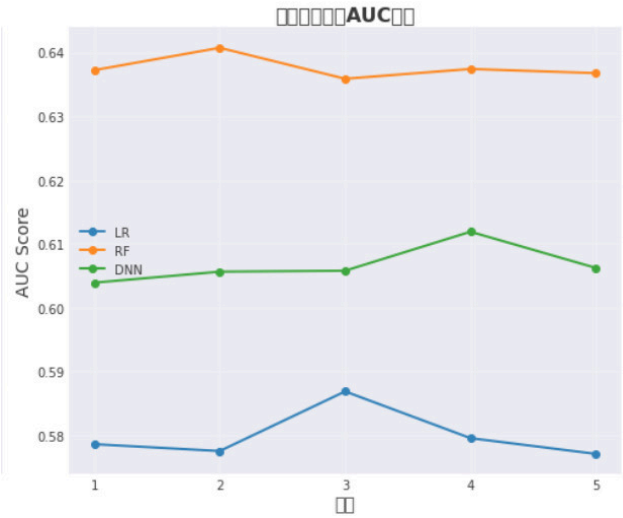
结果显示，随机森林显著优于其他模型，不仅 AUC 最高，且折间标准差最小，表明其在不同数据划分下表现稳定。DNN 虽具备建模非线性关系的能力，但在当前特征体系与调参策略下未能超越树模型。逻辑回归因模型表达能力受限，仅略优于随机猜测（AUC=0.5），反映出线性假设难以捕捉用户复购行为的复杂模式。

- RF 与 DNN 的预测概率均值约为 0.061，与典型电商复购率（通常 <10%）相符，说明模型输出具有合理校准性。
- LR 预测均值高达 0.491，接近 0.5，表明其对正负样本区分能力极弱，概率输出严重偏离真实分布，存在明显校准偏差。

下图为模型交叉验证平均AUC对比：



下图为模型各折交叉验证AUC得分：



7 总结

本项目围绕天猫用户复购预测任务，完成了从数据预处理、特征工程到多模型训练、调参与评估的完整机器学习流程。通过系统性构建用户、商家以及用户–商家交互特征，并对逻辑回归、随机森林和深度神经网络进行交叉验证与超参数优化，最终以随机森林模型取得最优性能（AUC = 0.6376）。

尽管当前模型已具备弱判别能力，但整体 AUC 仍处于较低水平，反映出现有特征体系对复购行为的刻画尚不充分。整体 AUC 未突破 0.65，表明当前模型上限受限于特征表达能力不足，而非模型选择或调参问题。未来工作可从以下三方面深入优化：

- 深化特征工程：引入用户–商家交互序列特征（如最近购买间隔、行为衰减权重）、群体先验（如同龄用户在该商家的复购率）及跨品类行为迁移信号；
- 探索更高效模型：尝试XGBoost 等梯度提升树模型，或结合 Embedding 技术的浅层神经网络，以更好融合稀疏ID类特征与稠密统计特征；
- 采用负采样策略：本项目由于算力限制，对于正负样本极端失衡的情况并未进行处理，后续可以在模型训练前先实行负采样策略。

综上，本项目验证了基础建模流程的有效性，并明确了下一步技术突破的关键路径。随着特征表达能力的持续增强和训练策略的改善，模型性能有望显著提升。

> 项目GitHub链接：hp761/tmall-repurchase-2025: 基于Python的问题解析-期末大作业