# 大数据技术及应用期末个人作业报告

221870001 马梓豪

2025 年 8 月 8 日

## 目录

**一：数据获取与描述性统计** **3**

1. 数据统计性描述 . . . . . 3
2. 数据爬取 . . . . . 6
    2.1 数据爬取实现 . . . . . 6
    2.2 反爬虫应对探索 . . . . . 7

**二：数据的探索性分析** **8**

1. 数据清洗 . . . . . 8
    1.1 缺失值处理 . . . . . 9
    2.1 缺失值处理 . . . . . 9
    3.1 缺失值处理 . . . . . 9
    4.1 表关联与字段筛选 . . . . . 9
2.k-means 聚类分析（无监督学习） . . . . . 9
    2.1 问题研究简介 . . . . . 9
    2.2 算法实现（调参） . . . . . 9
    2.3 可视化与结论 . . . . . 10
3.LDA 主题建模（无监督学习） . . . . . 14
    3.1 问题研究简介 . . . . . 14
    3.2 算法实现（调参） . . . . . 15
    3.3 可视化与结论 . . . . . 16
4. 随机森林 +SHAP（集成学习） . . . . . 18
    4.1 问题研究简介 . . . . . 18

4.2 算法实现 . . . . . 18
4.3 可视化与结论 . . . . . 18
5. 逻辑回归 . . . . . 20
5.1 问题研究简介 . . . . . 20
5.2 算法实现 . . . . . 20
5.3 可视化与结论 . . . . . 21
6.XGboost 预测模型 . . . . . 22
6.1 问题研究简介 . . . . . 22
6.2 算法实现 . . . . . 23
6.3 可视化与结论 . . . . . 24

**三：总结与展望** **26**

# 一：数据获取与描述性统计

## 1. 数据统计性描述

我选择的是京东健康平台，爬取了皮肤科医生的相关信息与用户评论数据。选择的地区包括：上海、江苏、浙江、四川、重庆、山西、山东、河南、河北。共爬取医生 1399 位，有效评论 17298 条，其中评论数大于 10 条的医生 610 位，爬取到的数据量如下图所示：

表 1: 各地区皮肤科医生与评论数量汇总

| 城市 | 医生数量 | 评论数 |
|---|---|---|
| 上海 | 202 | 1065 |
| 江苏 | 140 | 1002 |
| 浙江 | 90 | 642 |
| 四川 | 153 | 1378 |
| 重庆 | 30 | 435 |
| 山西 | 141 | 2359 |
| 山东 | 280 | 4977 |
| 河南 | 214 | 3187 |
| 河北 | 149 | 2253 |
| 共计 | 1399 | 17298 |

数据存储在 MySQL 的三个表中，参考字段表如下：
在 MySQL 中，表之间用外键关联，关系如图：

表 2: doctor_info 表字段说明

| 列名 | 数据类型 | 含义 |
|---|---|---|
| id | int (auto increment) | 医生唯一编号 |
| name | varchar(50) | 医生姓名 |
| title | varchar(50) | 医生职称 |
| hospital | varchar(100) | 所属医院 |
| department | varchar(100) | 科室 |
| city | varchar(50) | 所在城市 |
| good_rating | float | 好评率 |
| num_consults | int | 咨询人数 |
| num_flags | int | 获得锦旗数 |
| speciality | text | 专业擅长 |
| bio | text | 个人简介 |
| visit_total | int | 访问量 |
| article_count | int | 发布文章数 |
| online_patients | int | 在线患者数 |
| num_reviews | int | 患者评价数 |
| entry_date | datetime | 入驻时间 |
| reply_quality_score | float | 回复质量评分 |
| service_attitude_score | float | 服务态度评分 |
| reply_speed_score | float | 回复速度评分 |

表 3: doctor_price_info 表字段说明

| 列名 | 数据类型 | 含义 |
|---|---|---|
| id | int (auto increment) | 主键，自增编号 |
| doctor_id | int | 医生编号（外键） |
| type | varchar(20) | 问诊类型（图文、电话、视频等） |
| price | decimal(10,2) | 总价格（单位：元） |
| unit_duration | varchar(20) | 服务时长单位（如“15 分钟”、“48 小时”） |
| price_per_min | decimal(10,2) | 每分钟价格（统一换算后） |

表 4: patient_comments 表字段说明

| 列名 | 数据类型 | 含义 |
|---|---|---|
| id | int (auto increment) | 主键，自增编号 |
| doctor_id | int | 医生编号（外键） |
| username | varchar(20) | 患者昵称 |
| comment_text | text | 评论内容 |
| consult_type | varchar(20) | 问诊类型（图文、电话、视频等） |
| comment_date | date | 评论日期 |
| reply_quality | varchar(20) | 回复质量评价 |
| service_attitude | varchar(20) | 服务态度评价 |
| reply_speed | varchar(20) | 回复速度评价 |
| overall_star | int | 综合评分（星级） |

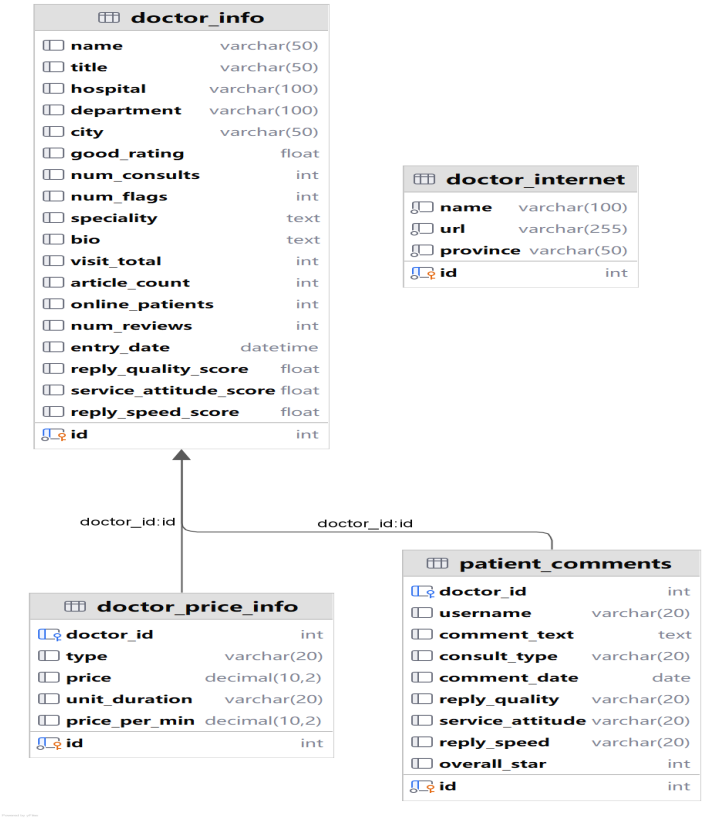


图 1: MySQL 表结构

## 2. 数据爬取

### 2.1 数据爬取实现

在前期尝试过程中，我首先通过人工分析京东健康皮肤科医生列表的网页结构，发现不同地区的 URL 格式中嵌入了一组代表省份的数字 ID 作为路径前缀，例如：

https://cont.jd.com/department/2_0-4-1?secondDepartmentId=0&doctorTitleId=0&sortItem=1&doctorServiceTypeId=0

其中 `2_0-4-1` 表示“上海”的编码。通过进一步归纳，我成功提取出上海、江苏、浙江、四川、重庆、山西、山东、河南、河北共九个省份的编码规律，构造出对应的翻页 URL 模板。

爬虫脚本 get_url.py 对上述 URL 模板进行自动迭代，利用 requests + BeautifulSoup 对网页结构进行解析，逐页提取出医生姓名与其主页链接，并标记所属省份，最终写入数据库表 doctor_internet 中作为爬虫下游任务的基础数据来源。

在获取医生主页链接后，爬虫 get_info.py 从 doctor_internet 表中读取所有医生 URL，并利用正则表达式提取出医生主页中所携带的唯一标识 doctor_id。该编号是京东健康平台为每位医生分配的唯一标识，用于构造医生信息页与评分页的目标请求地址：

https://cont.jd.com/doctor/jianjie/{doctor_id}

https://cont.jd.com/dianping/{doctor_id}

每个医生主页 URL 中均包含该唯一编号 doctor_id，爬虫程序通过该编号定位目标页面，进而提取出医生的结构化信息与评分指标。

在 get_comments_info.py 脚本中，程序同样以 doctor_internet 表为起点，逐条读取医生主页链接，并从中提取出医生唯一编号 doctor_id。该编号用于构造医生的评论页 URL，拼接规则如下：

https://cont.jd.com/dianping/doctor_id?diseaseLabelActive

其中，doctor_id 是医生主页链接中提取的真实编号。通过构造这样的 URL，程序能够精准定位每位医生对应的患者评价页面，从而进行批量信息抓取。

### 2.2 反爬虫应对探索

对于反爬措施，我一开始使用的和爬取豆瓣评论一样的反爬机制，即初始阶段通过检测网页返回的 HTTP 状态码是否为 418 作为反爬拦截的判据。但是，在文件执行后人为检查数据发现，京东平台即使返回 200 OK，将请求跳转至无实际数据内容的空白页面，仍构成反爬行为，导致该反爬机制失效。因此，我的反爬策略修改为基于内容解析的反爬判定机制，即在爬取目标页面前，尝试解析页面中关键字段（如医生姓名、医院名称等）。若关键字段无法提取，则视为遭遇反爬。

而且，我使用时间延迟控制策略（sleep time），但数据量较大，即使设置较长时间间隔仍会被反爬，然后，我尝试引入代理池机制，可惜网上价格便宜的代理 IP 质量太差，连接成功率与稳定性均较差，而高质量代理价格又太贵。

于是我想把开始被反爬的医生 id 记录下来，**多轮爬取**作为主要反爬策略：**所有爬虫脚本都内嵌失败日志记录功能**，程序在每轮运行时首先读取 `failed_doctor_ids.csv` 文件（初始化时人为设置为 1），用 if 判断，只开始爬取 id>=csv 文件读入的 `failed_id` 记录当前轮爬取失败的医生 ID，并在每轮运行结束后自动更新该文件。随后就能在间隔一段时间后，重新运行原始脚本以逐步补全数据，直至所有医生信息与评论数据均成功采集完成。

其中最难以应对的是评论爬取部分，我在手动翻页找规律过程中发现，京东健康网站医生下面的评论翻页每一页数据一模一样，在多次确认后发现这并不是反爬机制，因为这个网站的数据是从小程序搬过来的，所以我的结论是：这就是京东健康网页版的 bug。然后我发现小程序上的评论是正常的，于是我研究如何爬取小程序数据，用 fiddler 抓包，拿到小程序的请求接口去请求，把所有的接口全拿下来再当网页一样请求就可以了，但是京东小程序的反抓包机制太强，有加密参数，需要解包，这个技术我在网上找不到教程，于是就放弃。在我后期的尝试中，我意外发现，网页 url 本身的结构是

`https://cont.jd.com/dianping/{doctor_id}?diseaseLabelActive=all`

只需要手动把”=all” 去掉一页就可以显示 50 条评论，经过爬取后发现，本身评论数大于等于 50 的医生仅有 27 位，所以这样就可惜牺牲部分数据量来绕过小程序反抓包问题，直接在网页端爬取数据。

# 二：数据的探索性分析

## 1. 数据清洗

原始数据总计涵盖 1399 位皮肤科医生及其对应的 17298 条有效评论，数据存储于 MySQL 数据库中，由文件 `data_clean.py` 统一读取至 pandas 中进行处理与分析。清洗后的数据保存为以下两类文件，

- **cleaned_doctor_data.csv**：医生维度汇总表
- **cleaned_comments_data.csv**：患者评论清洗结果

供后续数据分析使用。在数据清洗过程中主要做了以下处理：

### 1.1 缺失值处理

针对少量缺失的数据字段，结合具体情况进行填充或删除。例如，个别医生可能缺少部分收费项，采用 0 填补；评论表中若存在空白评价内容则视为无效数据剔除。对于缺失较严重且无分析意义的字段予以舍弃。

### 2.1 缺失值处理

异常值处理：检查数值型特征的分布，识别明显异常值。例如，极端高的服务价格或异常的评分值。对于确定为录入错误或非常不合理的异常值，进行了纠正或剔除；

### 3.1 缺失值处理

数据类型转换：将字符串格式的数值字段转换为数值型，将日期时间字符串转换为标准日期类型。确保各字段类型正确，以便后续计算和分析。

### 4.1 表关联与字段筛选

通过医生 ID 将医生信息、价格信息与评论数据进行关联，构建综合分析表。根据分析需求筛选保留关键字段，如医生的地区、省份、医院级别、职称、平均评分、服务价格、问诊量，以及评论的内容、评分和日期等。一些与分析无关或冗余的字段（如医生头像 URL 等）被剔除。

## 2.k-means 聚类分析（无监督学习）

### 2.1 问题研究简介

在这一部分中，我想首先通过 k-means 聚类挖掘皮肤科医生的分层结构，可视化不同簇中的医生画像，了解他们的特征，再将地区分为三部分，山河四省、江浙沪和川渝，研究各地域不同簇医生的占比，观察是否有明显的地域特征。

### 2.2 算法实现（调参）

我选用 **K-Means 聚类算法**对医生个人信息数据无监督聚类分析，在 `k_means.py` 文件中读入文件 `cleaned_doctor_data.csv`, 目标是基于医生的服务质量、定价、问诊活跃度等多维特征，挖掘医生群体的潜在分层结构。

为了确定最佳的聚类数 $K$，我采用以下两种方法进行评估：

- **肘部法**：通过绘制不同 $K$ 值对应的簇内平方误差和（Inertia）变化趋势图，观察曲线拐点位置，以判断聚类效果的边际收益递减点。
- **轮廓系数**：衡量样本的聚类紧致度与类间分离度，其值范围为 $[-1, 1]$，越接近 1 表明聚类效果越好。通过比较不同 $K$ 值下的平均轮廓系数，辅助判断最优聚类数。

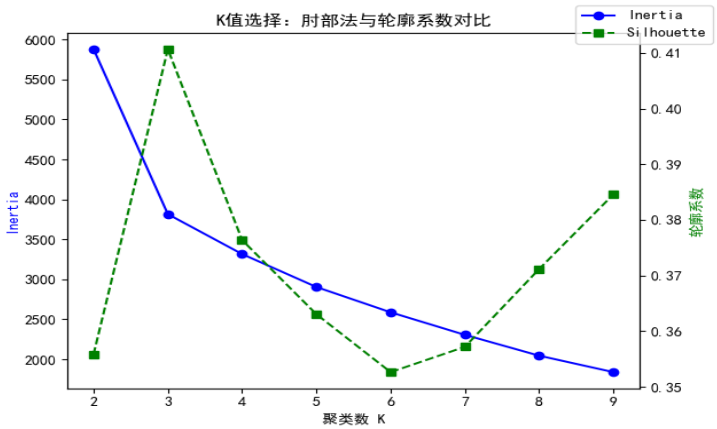


图 2: 肘部法与轮廓系数图

由图可得，k=3 是图像的拐点，且在 k=3 时，轮廓系数为 0.4107，经查阅资料，轮廓系数越接近 1 效果越好，但很少有能超过 0.7 的，这里 0.4107 说明效果不错。最终选定聚类数为 $K = 3$。随后使用标准的 K-Means 算法对标准化特征矩阵进行聚类建模，并将聚类标签附加至原始数据中。

### 2.3 可视化与结论

为便于对聚类结构的理解与解释，我采用多种可视化手段对 K-Means 聚类结果进行分析，主要包括：

- **PCA 降维可视化**：将高维特征通过主成分分析（PCA）降维至二维平面，绘制聚类散点图。从图中可以观察到医生群体在二维主成分空间中呈现较好的聚类效果，不同聚类之间具有一定的可分性，说明聚

类结构在原始高维特征空间中具有合理性。而且每簇样本量相近，不存在假分层。

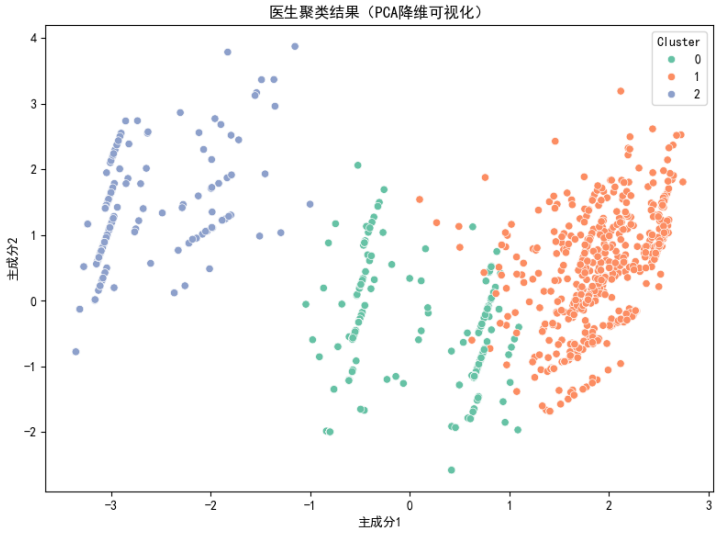


图 3: 聚类散点图

- **雷达图**：展示各聚类簇在各项标准化特征上的平均水平，形成“能力画像”。其中，簇 1 在所有维度上均接近最高水平，表现出全面领先的特征；簇 0 在服务评分高但价格低、评论数少，具有较强的亲民性；簇 2 在各项指标上均明显低于其他簇，表现出低活跃或新入驻医生的特征。

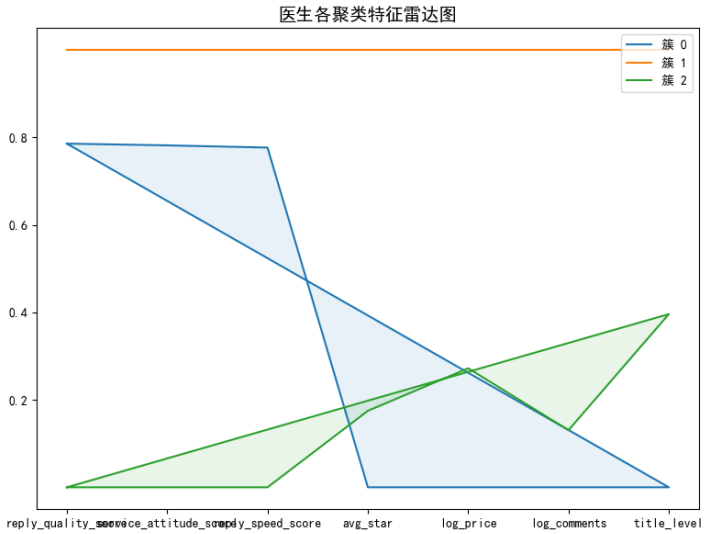


图 4: 特征雷达图

- **热力图**：展示各簇在原始特征下的均值。热力图清晰显示簇 1 在评分、评论数、价格、职称等方面显著高于其他簇，簇 2 的评分和问诊量处于最低水平，进一步印证了雷达图中的分布特征。

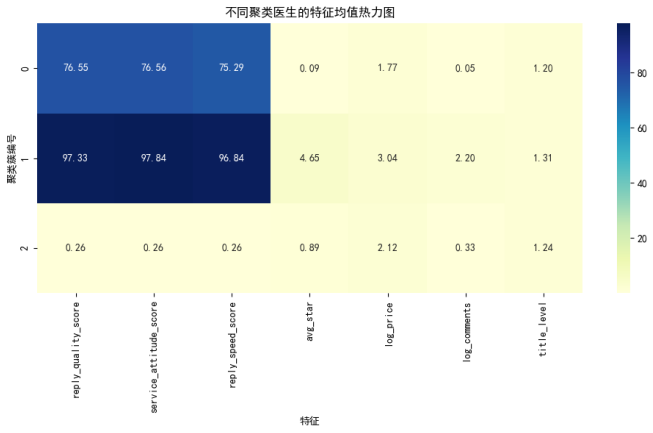


图 5: 特征热力图

基于可视化分析结果，可以得到医生有三类群体的结论：

- **簇 1：高价值专家组**，在所有维度上表现突出，服务评分、平均星级、价格、评论数量和职称等级均为最高，是平台上的顶级专家医生群体，建议重点推广；
- **簇 0：亲民优质组**，在服务质量方面得分较高，但价格较低、评论数量较少，可能为年轻医生或新晋平台医生，具有发展潜力；
- **簇 2：低活跃边缘组**，在所有特征维度上均处于劣势，可能为不活跃医生或服务质量较差者，平台应评估其服务质量或加强服务引导。

同时，在地区选取时，我就想考虑**地域因素**的影响，分为江浙沪、川渝、山河四省，于是我首先对聚类结果进行了卡方检验，程序输出 Chi2 = 18.51, 自由度 = 4, p 值 = 0.0010，说明区域分布与医生聚类之间具有统计学显著关联 ($p < 0.05$)，于是基于此发现，我绘制不同地域的分组堆叠柱状图和聚类分布比例热力图，以及不同区域医生特征均值热力图。

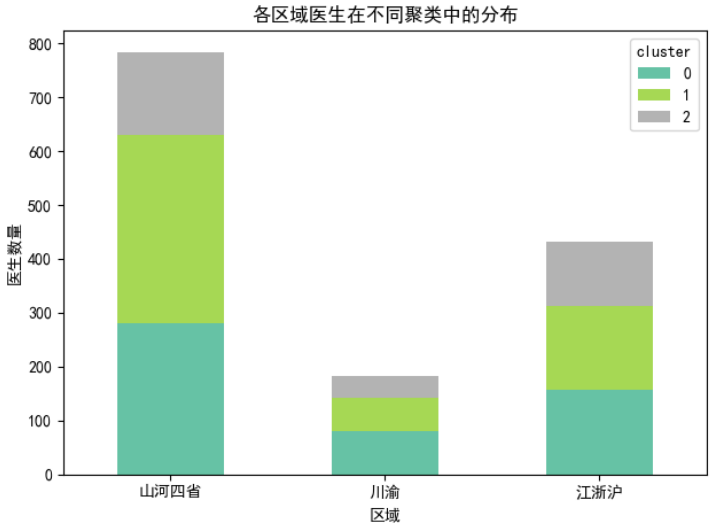


图 6: 各区域医生在不同聚类簇中的数量分布（堆叠柱状图）

由图可得，山河四省医生数量明显最多，聚类 0 和聚类 1 占主导, 川渝医生较少，但聚类 0 占比突出, 江浙沪医生较平均地分布于三类聚类，说明群体多样性强。

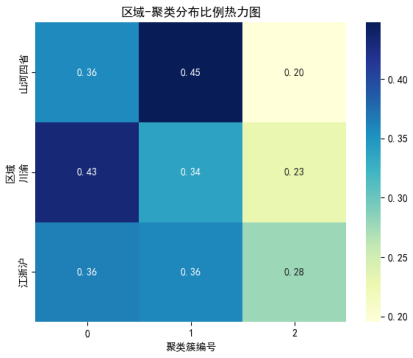


图 7: 区域-聚类比例热力图

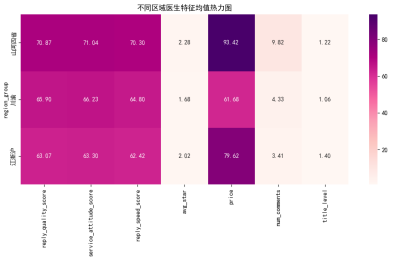


图 8: 各区域医生特征均值热力图

由此可以看出，三地在医生画像聚类上有明显差异，山河四省医生评分、问诊量均最高，可能平台服务供需最活跃；川渝地区医生价格和评分偏低，整体资源水平或活跃度略低；江浙沪医生职称水平最高（`title_level` = 1.40），说明高职称医生集中；差异说明区域间用户活跃度 + 医生资源结构有显著不同。

但是这与我在进行数据分析前的主观感受不同：在完成作业之前我根据生活经验等推断应该是江浙沪地区的皮肤科医生问诊量更高，因为近些年医美行业火热，而绝大多数明星、网红聚集在江浙沪地区，他们是医美项目的主力军。而江浙沪地区医生职称水平最高符合我们的主观感受，因为该地区经济发达，可以吸引更多名医，至于问诊量的数据分析结果的进一步解释可能需要更多细致的数据（比如具体挂号的项目，可能是因为京东健康上的问诊大部分是传统的皮肤病，而这部分数据网站无法爬取，所以后期需要再学习其他方法）。

## 3.LDA 主题建模（无监督学习）

### 3.1 问题研究简介

在这部分研究中，我想要以数据为基础，提取医生评价的“潜在维度”，研究患者评论中普遍关注哪些服务维度，各主题在评论中所占比例是否均衡，是否存在主导性主题，评论的情感倾向（如正面、一般）在不同主题下分布有何差异，通过这方面的研究，了解患者需求声音，有助于更好理解应该在哪些方面改进，精准为患者服务。

### 3.2 算法实现（调参）

我采用了 Latent Dirichlet Allocation（LDA）主题模型对患者评论文本进行建模，旨在挖掘评论中的隐含语义主题，揭示患者在就诊过程中关注的核心维度。整个建模过程分为三个阶段，分别对应于三个 Python 脚本：`lda_evaluation.py`、`lda.py` 和 `lda_analysis.py`。

首先是 `lda_evaluation.py` 文件，该脚本旨在辅助确定 LDA 模型中最优的主题数 $K$ 值，通过计算一致性得分与困惑度两项指标，在 2 到 15 个主题数量区间内评估模型表现，并最终通过可视化图像辅助选择最合适的主题数。

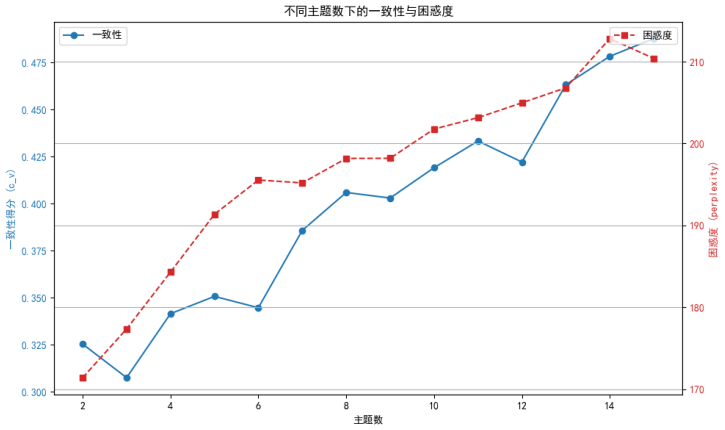


图 9: 不同主题下一致性与困惑度

由图像可得一致性得分从 k=2 开始逐步上升，在 k=14 达到最大值（约 0.475），之后略有下降。困惑度持续上升，在 k=14 达到最高点后略有回落。最终选择 k=13，一致性已非常高，困惑度也尚可。

之后是 `lda.py` 文件，设定主题数为 13，调用 LatentDirichletAllocation 构建主题模型，打印展示每个主题的关键词组成，将最终的主题编号与原始评论一一对应，保存为 `comment_topics_k13.csv` 文件，供后续分析使用.

再用 `lda_analysis.py` 文件，本脚本的主要目标是对 `lda.py` 脚本输出的主题分布结果（即 `comment_topics_k13.csv`）进行进一步的可视化与分析，尤其聚焦于患者评论的情感倾向与主题偏好之间的关系。整体思路如下：

通过构建一套正面情感关键词词典（如“感谢”、“有效”、“医术高明”

等)，对每条评论进行粗略情感判定，划分为“正面评论”与“一般评论”两类，以支持后续对比分析。分别统计“正面评论”和“一般评论”中主主题的分布情况，并绘制双柱状图，直观展示不同情感倾向下的主题偏好差异。进一步计算各主题在“正面评论”与“一般评论”中的占比差异，并输出差异最大的主题及其关键词，辅助识别影响患者满意度的潜在主题因素。

### 3.3 可视化与结论

最终得到的 13 个主题关键词分别为：

表 5: LDA 主题含义解释与关键词示例

| 主 题 编号 | 核心关键词示例 | 主题含义解释 |
|---|---|---|
| Topic1 | 服务态度、医德、热心、问诊、病人 | 服务态度与职业素养 |
| Topic2 | 迅速、诊断、准确、快捷、判断 | 诊断速度与专业判断 |
| Topic3 | 仔细、负责、温柔、感激、精准 | 医生的责任心与人文关怀 |
| Topic4 | 非常感谢、敬业、接诊、经验丰富、快捷 | 敬业精神与就诊便捷性 |
| Topic5 | 认真、帮助、效果、开药、希望 | 治疗效果与方案满意度 |
| Topic6 | 满意、医院、服务、太慢、一次 | 就诊过程体验与反馈（含负面情绪） |
| Topic7 | 解释、耐心、透明、疑问、责任心 | 病情解释与沟通质量 |
| Topic8 | 细心、讲解、解释、清楚、到位 | 沟通清晰与讲解细致 |
| Topic9 | 态度、及时、点赞、说话、好评 | 总体认可与态度称赞 |
| Topic10 | 很快、耐心、医生、精湛、解决问题 | 医术水平与处理效率 |
| Topic11 | 感谢、用药、治疗、效果、方案 | 药物与治疗建议有效性 |
| Topic12 | 细致、喜欢、焦点、缓解、真爱 | 情绪安抚与心理支持 |
| Topic13 | 热情、答疑、解惑、亲切、和谐可亲 | 亲和力与沟通态度 |

绘制每个主题在总体评论中的占比柱状图：

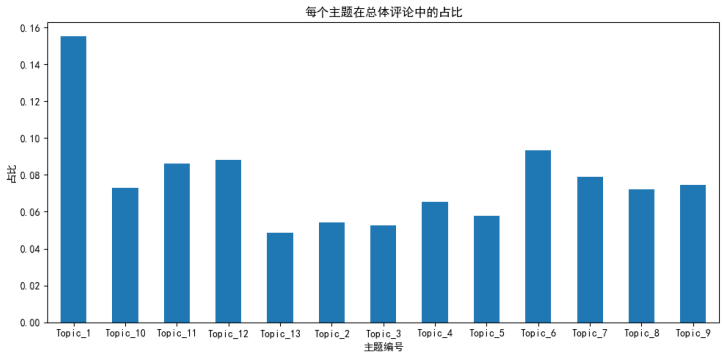


图 10: 每个主题在总体评论中的占比

正面评论 vs 一般评论主题对比的柱状图：

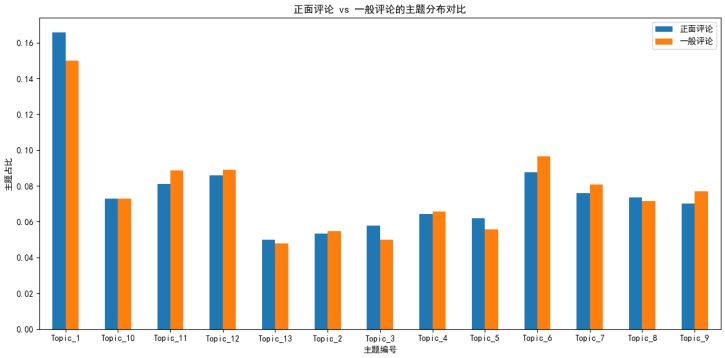


图 11: 正面评论 vs 一般评论

通过 LDA 主题分析可以提炼“潜在语义主题”，**理解评论中好评多在哪些主题、差评多在哪些点，想知道是否有隐藏的不满风险点**，可以了解患者对医生的核心需求，提供有力的数据支持，通过分析，我得到如下结论：

结合图 10 与表格分析可知，Topic1（服务态度）占比最高，超过 15%，说明患者最关注医生态度与职业素养,Topic12、Topic11、Topic6 也占比较高，说明情绪安抚、用药指导与就诊体验是患者评论中的高频关注点。Topic2、Topic3、Topic5、Topic13 等主题占比居中，说明医生责任感、人文关怀、治疗效果与亲和力是次要但也不可忽视的关注点。而负面或一般评论更容易集中在“流程体验”、“沟通解释”等方面（Topic6、Topic7、Topic8）。

而结合图 11，我主要关注一般评论占比更高的主题，Topic6（就诊体验/慢):一般评论占比高于正面,说明存在对速度或流程不满的情况。Topic7、8（沟通解释类)：在一般评论中占比略高，反映出一些沟通不到位的问题。提示平台与医生可以从提升流程流畅性和沟通能力入手改善满意度。

## 4. 随机森林 +SHAP（集成学习）

### 4.1 问题研究简介

注意到各医生之间线上问诊价格差异较大，所以这部分我通过机器学习中的集成学习方法（随机森林）和解释性工具（SHAP)，系统挖掘影响医生服务定价的关键变量。主要想得知哪些医生个人信息或行为特征对其服务价格具有显著影响。

### 4.2 算法实现

因为研究目标是识别“影响医生价格的关键特征”，并不追求预测精度最优，因此使用默认参数构建随机森林回归模型。其中，entry_years 为医生入驻平台的年份差异变量，由 entry_date 计算得到。对于 title、hospital 和 city 等类别变量，采用 LabelEncoder 进行数值化处理，以满足模型输入要求。为获得更具解释性的特征贡献效果，进一步引入 SHAP 方法，基于博弈论思想分配每个特征对预测值的边际贡献。

### 4.3 可视化与结论

我通过两类图示结果呈现变量对价格的贡献程度与方向性影响,SHAP 重要性条形图（图 12）揭示了各特征对医生定价模型预测结果的重要性大小；而 SHAP 分布图（图 13）进一步展示了特征取值的变化如何影响模型输出的方向与幅度。可视化结果不仅提升了模型的可解释性，还提供了定量化的证据支持上述研究发现，为平台优化价格策略与患者就医选择提供了重要参考。

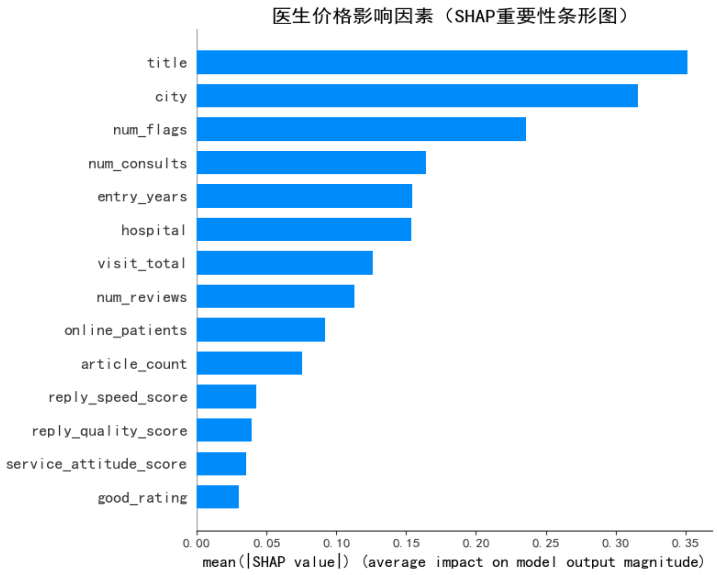


图 12: SHAP 重要性条形图

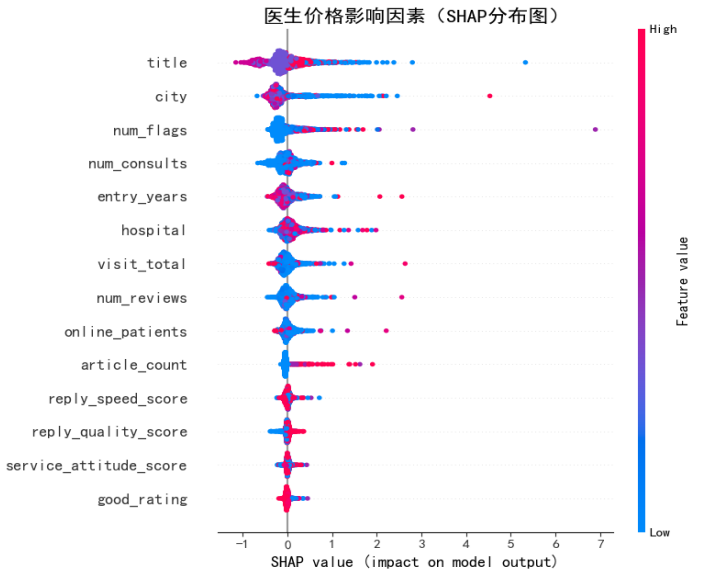


图 13: SHAP 分布图

基于这两张图，得出以下关键结论:

医生职称是影响价格的首要因素。高职称（如主任医师）显著对应更高的预测价格，且其 SHAP 值远超其他变量，说明医生等级在定价中起决定性作用。医生所在城市对价格有显著影响。部分一线城市的医生价格普遍更高，地域性经济差异与医疗资源分布可能是关键驱动。而这一点与之前 2.3 中由聚类分析得到的结果一致，江浙沪经济发达地区医生资源更好，职称较高，且区域总体消费水平相对高，所以价格较高。

其次，被“加旗”次数越多，价格越高。平台认可度反映医生受欢迎程度，对定价有正向推动效应。医生入驻平台的年限具有一定影响力。经验更丰富或平台资历较长的医生价格略高，但作用边际有限。用户反馈类变量（如服务态度评分、好评率）影响较弱。虽然服务质量可能影响患者满意度，但在价格形成中权重较低。这一结论又说明线上医疗平台不同于电商平台，用户还是更看重硬指标。

## 5. 逻辑回归

### 5.1 问题研究简介

在这部分中，我主要聚焦于患者评论部分，想通过患者评论的数据解释影响医生评分的因素，主要影响因素可能是“回复质量”、“服务态度”、“回复速度”、“咨询类型”，通过构建逻辑回归模型可以进行这方面的数据探索性分析。

### 5.2 算法实现

首先，原始数据中，三项服务指标（reply_quality、service_attitude、reply_speed）为中文描述，如“非常满意”、“满意”等，将其统一映射为评分 5 ~ 1。同时将 consult_type 编码为类别变量. 最终，构造了一个包含 4 个解释变量和 1 个被解释变量（评分）的数据集，供模型使用.

因为被解释变量即评分为 1-5 是离散的，所以用一般的线性回归不合适，所以我最开始尝试使用 softmax 分类器，在 multiple_regression.py 脚本中将评分视为无序分类变量，使用 MNLogit 模型进行建模，但是效果并不理想，终端输出 Warning，而且可视化后的图形显然是错误的，使用 softmax 得到的回归系数分布图如下：

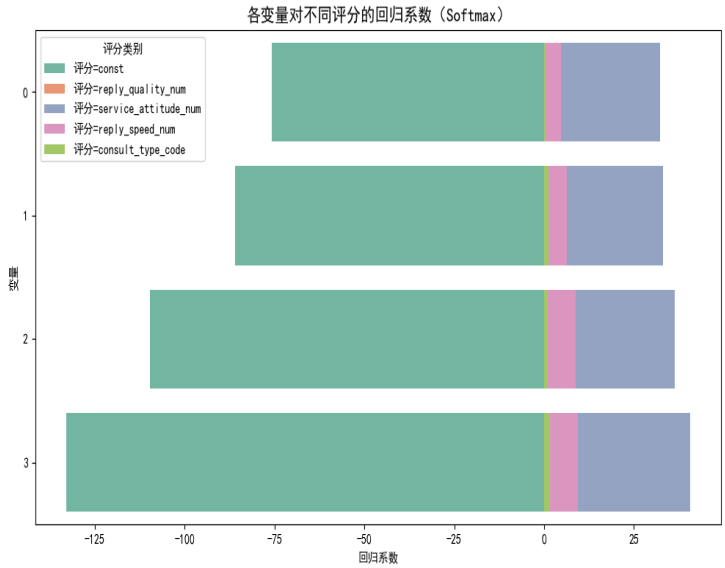


图 14: softmax 分类器建模

显然不具有可解释性，且终端提示模型有不收敛的问题，于是在我查阅资料后发现是因为数据集好评（5 分）的占比过于大，超过百分之 95，某些评分等级在数据中样本过少，可能造成某些类别参数估计不稳定，这表明，将评分当作无序多分类变量处理并不合理。

意识到评分变量本质上具备有序等级属性，我又采用有序逻辑回归建模，该模型可在解释变量的基础上，对因变量等级进行排序建模，适用于有自然顺序的评分数据。

此模型成功拟合，解决了非收敛问题，并可清晰输出每个变量对应的回归系数及置信区间。

### 5.3 可视化与结论

基于模型输出结果，绘制了如下回归系数图

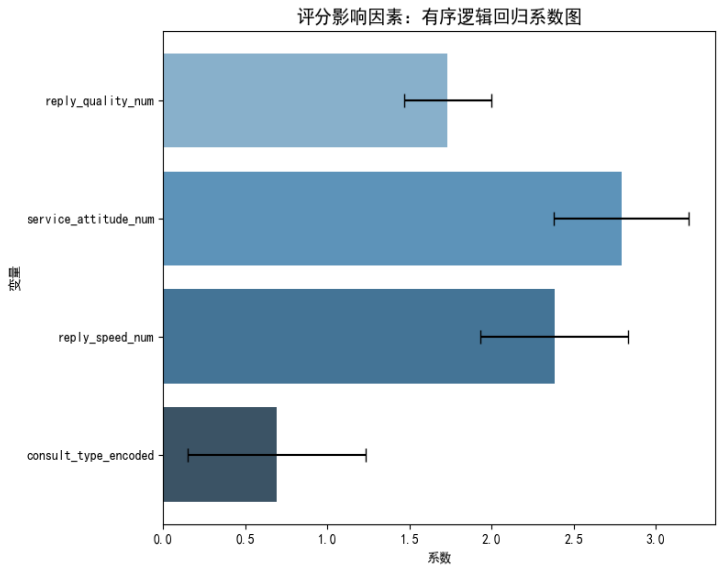


图 15: 有序逻辑回归系数图

通过这个可视化的图像，我得到以下结论：服务态度影响最大，回归系数约为 2.8，置信区间不跨 0，说明具有显著正向影响。回复速度影响较大，回归系数约为 2.4，且置信区间也完全为正，表明其对评分也具有显著影响。回复质量影响居中，回归系数约为 1.6，置信区间也显著为正。咨询类型影响最小，回归系数约为 0.7，且置信区间相对较宽，接近于包含 0。

这个结论也具有说服力，影响医生评分的关键不在于医生资历背景，而在于其提供服务的质量——主要态度和响应速度。模型结果为这一结论提供了量化支持，服务维度的重要性远超其他因素，说明只要医生在问诊过程中让患者感觉“耐心、温和”，就几乎保证了好评。而问诊类型并不会对评分造成影响，说明各种类型的问诊各有优势和劣势，患者并不在意问诊类型。

## 6.XGboost 预测模型

### 6.1 问题研究简介

在医生个人信息内有一栏为 good_rating，即好评率，在这个模块，我主要通过 XGboost 建立预测模型，即将 good_rating, 作为被解释变量其余数据作为解释变量。

### 6.2 算法实现

一开始我直接使用各种回归方法和深度学习算法发现无论如何调整参数，模型总是存在过拟合的问题，在探索后发现主要是因为数据集中，爬取到的数据极端分布较为明显，good_rating 为 100 的医生有 600 多位，占到数据总量的一半多，而他们好评率为 100 是因为绝大多数只有一到两个患者评分，导致数据不具有说服力，还有很大一部分原因是绝大多数医生的好评率都分布在 95 以上，只有极少数医生的好评率低于 95。

于是，我先整体剔除好评 =100 的天花板样本，再判断剩余医生是处于 [95,99) 还是 [99,100) 两个高分段。为避免极端好评样本对模型造成偏置，剔除好评率为 100 的样本，并按照如下规则筛选样本：

表 6: 建模使用的结构化特征清单

| 特征名 | 类型 | 说明 / 建议预处理 |
|---|---|---|
| `title` | 类别型 | 独热编码。 |
| `city` | 类别型 | 独热编码。 |
| `num_consults` | 数值型（非负整数） | 取 $\log(1+x)$。 |
| `num_flags` | 数值型（非负整数） | 稀疏二值化。 |
| `visit_total` | 数值型（非负整数） | 标准化。 |
| `article_count` | 数值型（非负整数） | 标准化。 |
| `online_patients` | 数值型（非负整数） | 与 `num_consults` 协同使用。 |
| `num_reviews` | 数值型（非负整数） | 标准化 |
| `entry_date` | 日期/时间 | 转为 “已入驻天数”。 |

出于避免目标泄漏与提升可解释性的考虑，以下特征不纳入建模：

`reply_quality_score`, `service_attitude_score`, `reply_speed_score`.

GridSearchCV 采用 5 折交叉验证，网格参数包括 `n_estimators`、`max_depth`、`learning_rate`、`subsample`、`colsample_bytree`。默认采用 `StratifiedKFold`，脚本支持按城市分组和时间滚动的可切换模式。

### 6.3 可视化与结论

**(1) 交叉验证表现与最优参数** 最优参数为：`n_estimators` = 300, `max_depth` = 4, `learning_rate` = 0.05, `subsample` = 0.8, `colsample_bytree` = 0.8；五折平均验证分数 F1 = 0.97796。不同参数组合的验证分数分布集中，模型对超参不敏感且稳定性强（见图16）。

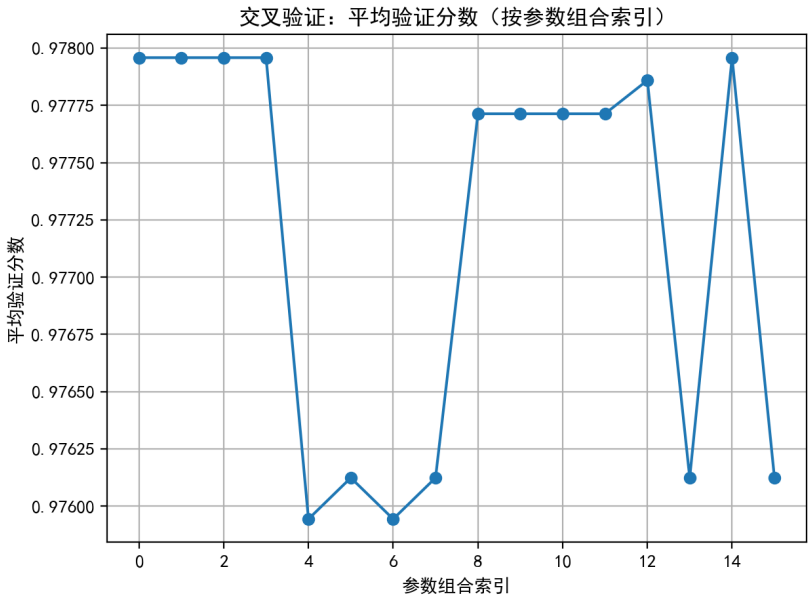


图 16: 交叉验证：平均验证分数

**(2) 测试集整体精度与误差结构** 测试集结果：`accuracy` = 0.9506、`macro-F1` = 0.9022。分类别 F1：[95, 99) 为 0.83，[99, 100) 为 0.97。混淆矩阵（图17）显示：真实 [95, 99) 判对 10 例、误判为高分 2 例；真实 [99, 100) 判对 67 例、误判为低分 2 例。可见高分段识别更为稳定，极少数中高分被判为高分。

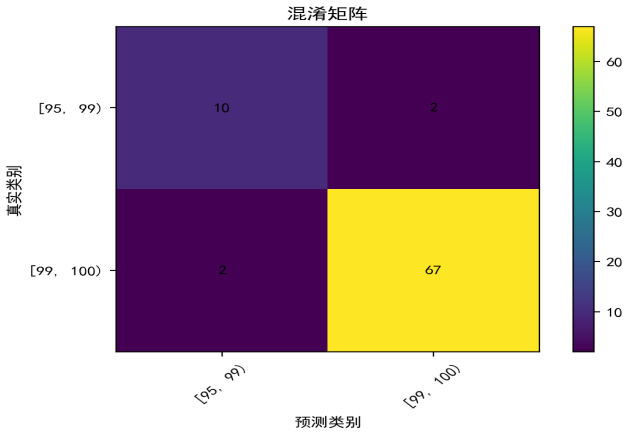


图 17: 混淆矩阵（测试集）

**(3) 特征贡献与解释** 图18 为模型自带的重要性，`num_reviews` 贡献最大城市与职称的若干虚拟变量居中；时间衍生特征也有一定作用。置换重要性（图19）进一步佐证 `entry_date_month` 与 `entry_date_dow` 在验证分数上的平均影响。

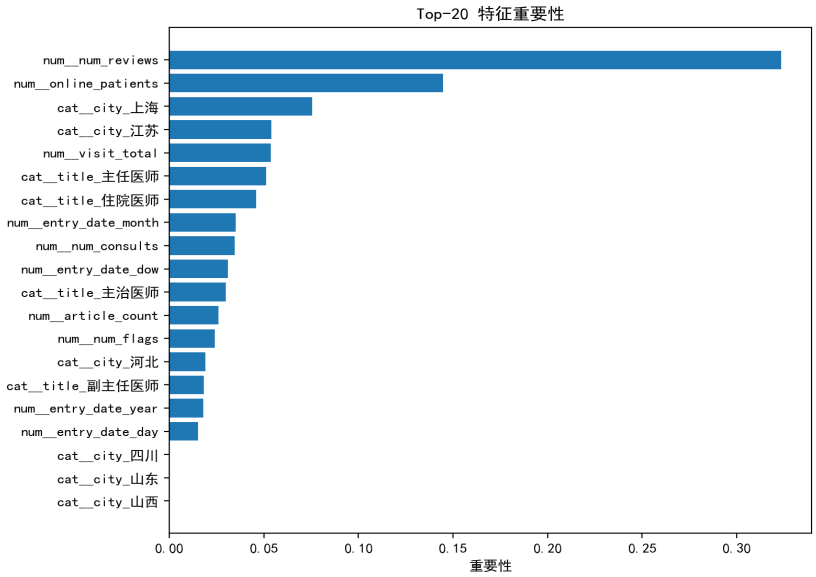


图 18: Top-20 特征重要性

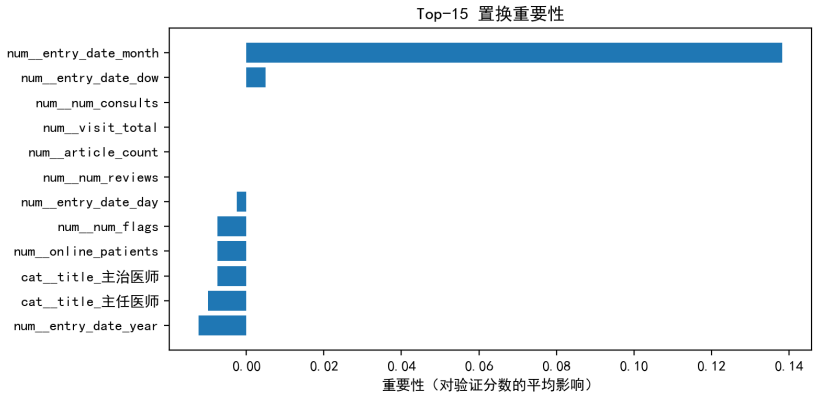


图 19: Top-15 置换重要性

# 三：总结与展望

我的课程作业以京东健康平台皮肤科医生为研究对象，构建了多维度、多角度的数据分析框架，综合应用了数据挖掘、无监督学习、监督学习与可解释性建模等方法，系统探讨了医生画像、评论内容、服务质量与患者评分之间的关联关系。

首先，在数据获取层面，我通过对网页结构的深入分析与多轮反爬机制测试，成功爬取了包括医生基本信息、服务价格、患者评论在内的完整数据集，并构建了规范的 MySQL 数据表结构，为后续分析奠定了坚实基础。

在探索性分析中，我分别应用了如下方法：

- **K-Means 聚类分析**：医生群体可划分为三类：
  - 高价值专家组：服务评分与价格显著更高；
  - 亲民优质组：服务质量高但价格较低；
  - 低活跃边缘组：整体指标偏低。

  聚类结果与地域分布具有统计显著相关性，不同区域呈现明显结构特征差异，如江浙沪地区职称水平最高，山河四省问诊活跃度最高，川渝医生则偏向亲民。

- **LDA 主题建模**：提取出 13 个患者关注的服务主题，关注度最高的包括“服务态度”、“治疗效果”、“就诊体验”等；负面或中性评论多集

中在“流程体验”和“沟通解释”类主题，提示平台应优化服务效率与沟通质量。

- **随机森林 + SHAP 分析**：揭示影响医生定价的核心因素为职称、城市、加旗数和平台资历，用户评分等反馈类变量影响较弱，表明线上医疗平台的定价机制更依赖医生的“硬指标”。
- **有序逻辑回归分析**：发现“服务态度”和“回复速度”对评分具有最显著的正向影响，远高于“回复质量”和“问诊类型”，说明提升医生服务体验是影响患者满意度的关键。
- **XGboost 预测模型**：通过使用 XGboost 模型并通过交叉验证调整参数，得到将好评率作为被解释变量，其余作为解释变量的预测模型，模型具有较高的 F1 分数，在测试集上表现良好，并有较强的鲁棒性，有很好的泛化能力。

最后，我的项目仍然具有一定的局限性，主要原因还是来自京东健康网站本身数据量不够丰富，或者有很多无意义的数据。进一步探索，一方面是获取更加详细的数据信息，或者扩大数据量的爬取，也许可以得到一些更加有意思的结论。另外，一些模型还可以再进一步进行完善探索，比如 LDA 模型可以结合词云和情感分析得到更加全面的结论，或者对某一个具体的研究方向使用多种数据分析方法，使得结论更有说服力。而在预测模型方面，我一开始想要引入神经网络构造模型，但是由于数据量不足，导致神经网络总是出现过拟合的问题，方差较大，对于 XGBoost 模型中也只能预测 90-99 好评率的医生，所以如果可以拥有更加丰富的训练集，就可以使用其他强大的机器学习算法进一步提升模型的性能。