Hugging Face 万模大战

原创 蓝色北极星 Blue Polaris 2024年08月01日 17:15 广东

Blue Polaris推荐搜索

huggingface | 大模型 | 爬虫 | 数据分析

“本文爬取Hugging Face 30000个开源模型，并进行了一些简单的数据分析，本项目旨在自娱自乐😼，也欢迎看到推文的各位观众大佬评论 + 指点我后续改进。”

🤗01：Hugging Face介绍!

“Hugging Face是一个开源的人工智能社区和平台。它提供了丰富的预训练模型、数据集和工具。”——来自某AI的介绍

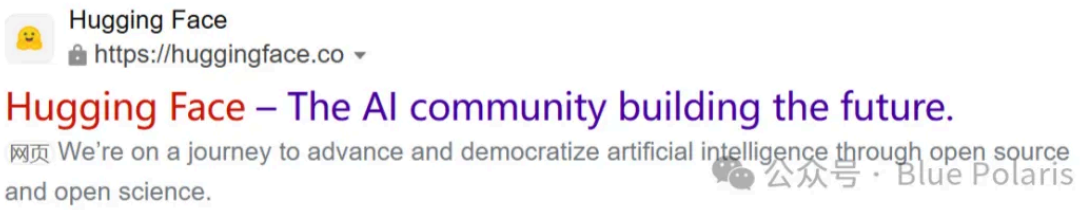

图1 Hugging Face官方网址

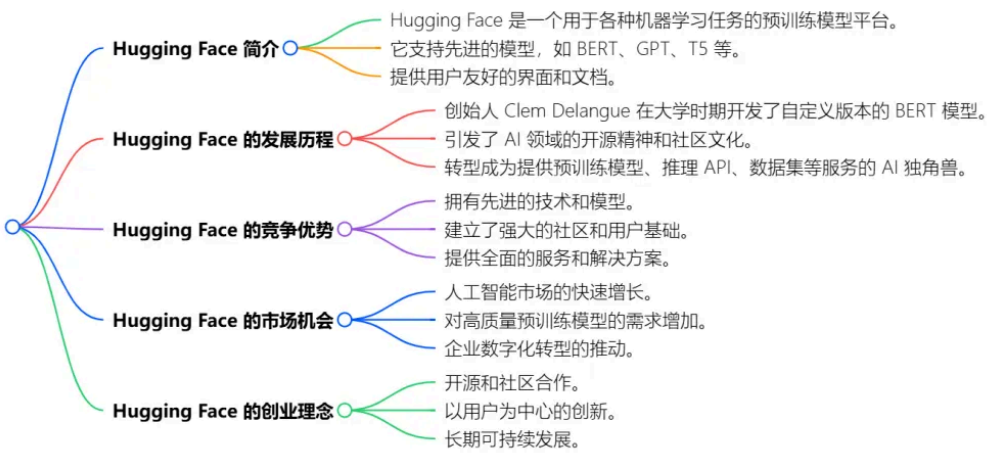

图2 Hugging Face介绍（豆包大模型生成，仅作参考）

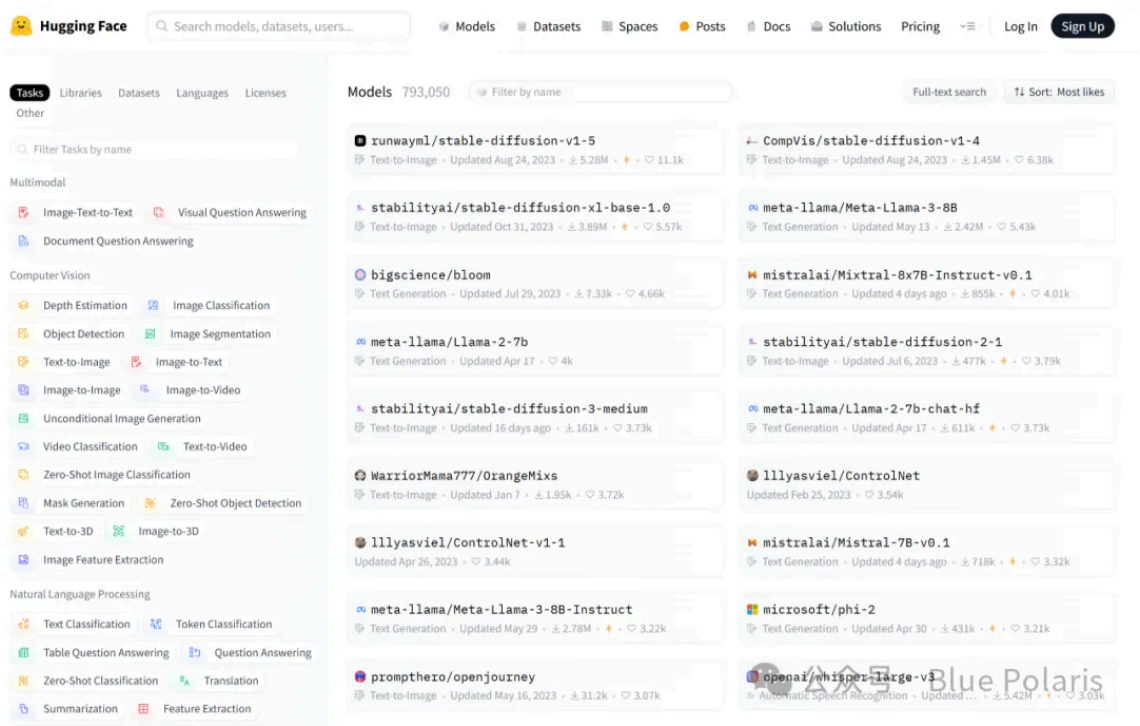

图3 Hugging Face页面

感谢开源！

02：数据爬取

- **爬取过程**

(1)按照收藏量从多到少排序，爬取了1000页模型信息，最终得到29017个模型数据。

(2)爬取指标：

organization	模型发布组织
name	模型名称
task	模型任务，如文本生成/文生图
downloads	下载量
likes	收藏量
url	模型地址
model_size	模型参数大小
tensor_type	张量类型如float64

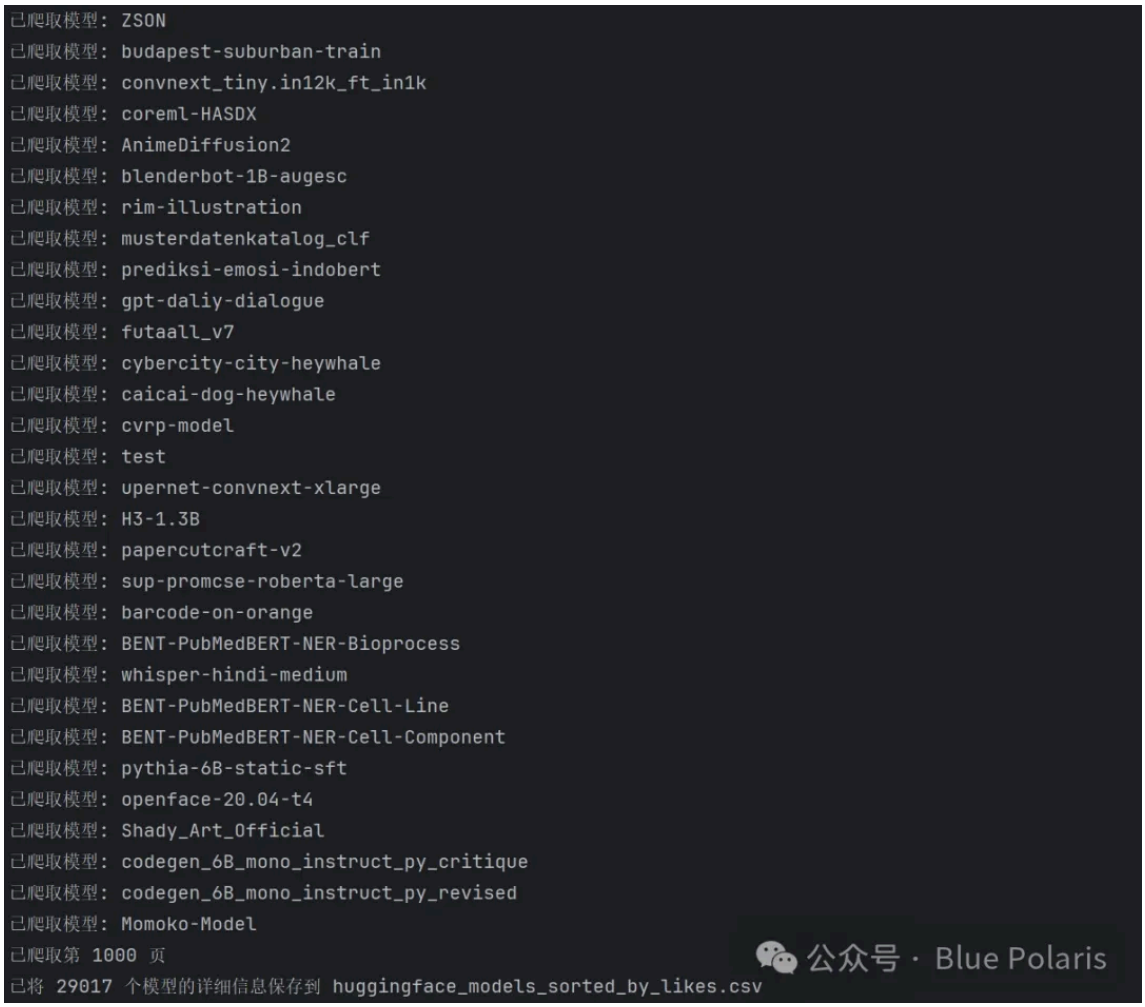

图4 爬虫过程截图

🤯03：数据分析

- **数据缺失情况：**

(1)超6成模型未公开model_size

(2)约1/4的模型缺失task说明

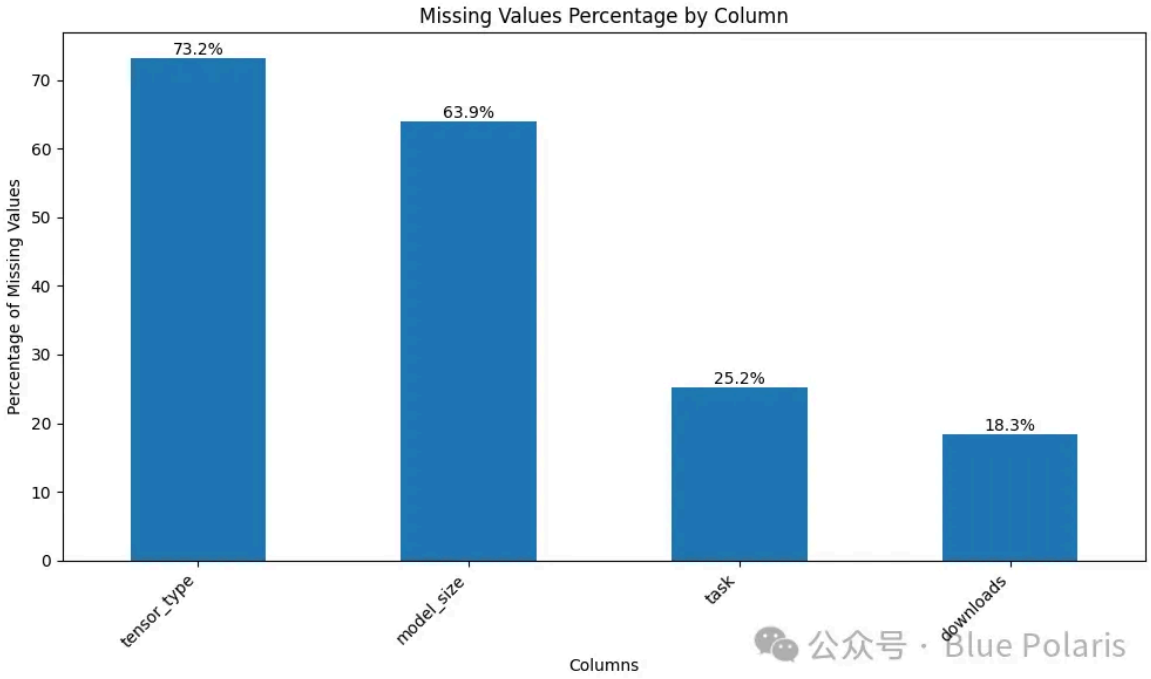

图5 数据缺失程度

- **Task前五名:**

🔥 Text Generation (文本生成)

🐉 Text-to-Image (文本生成图像)

🤿 Text Classification (文本分类)

🦈 Text2Text Generation (文本到文本生成)

🐂 Fill-Mask (填充掩码)

"其中,Text Generation任务的模型数量远远超过其他类型,显示了在自然语言处理领域,生成任务受到了极大的关注。Text-to-Image位列第二,反映了多模态AI的快速发展趋势。Text Classification和Text2Text Generation分别位居第三和第四,表明这些基础NLP任务仍然是研究和应用的重点。Fill-Mask任务虽然排名第五,但其模型数量相对较少,可能是因为它是一个相对特定的预训练任务。"

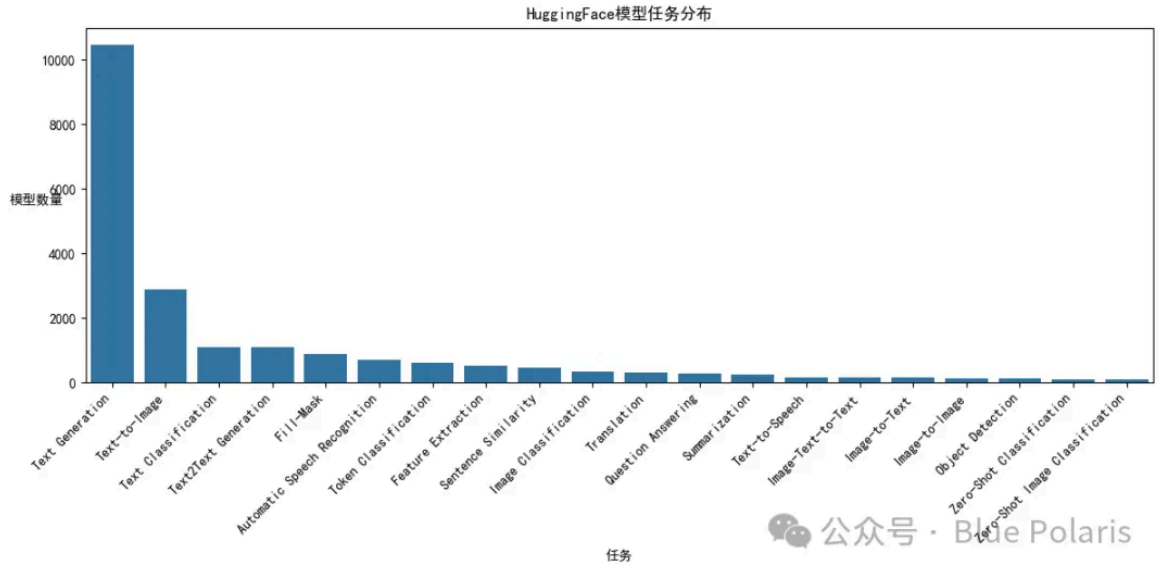

图6 模型数量最多的前20个任务类型

- **Organization前五名:**

🦅 TheBloke (个人)

🦁 facebook

🐘 sd-concepts-library

🦊 google

🐼 bartowski

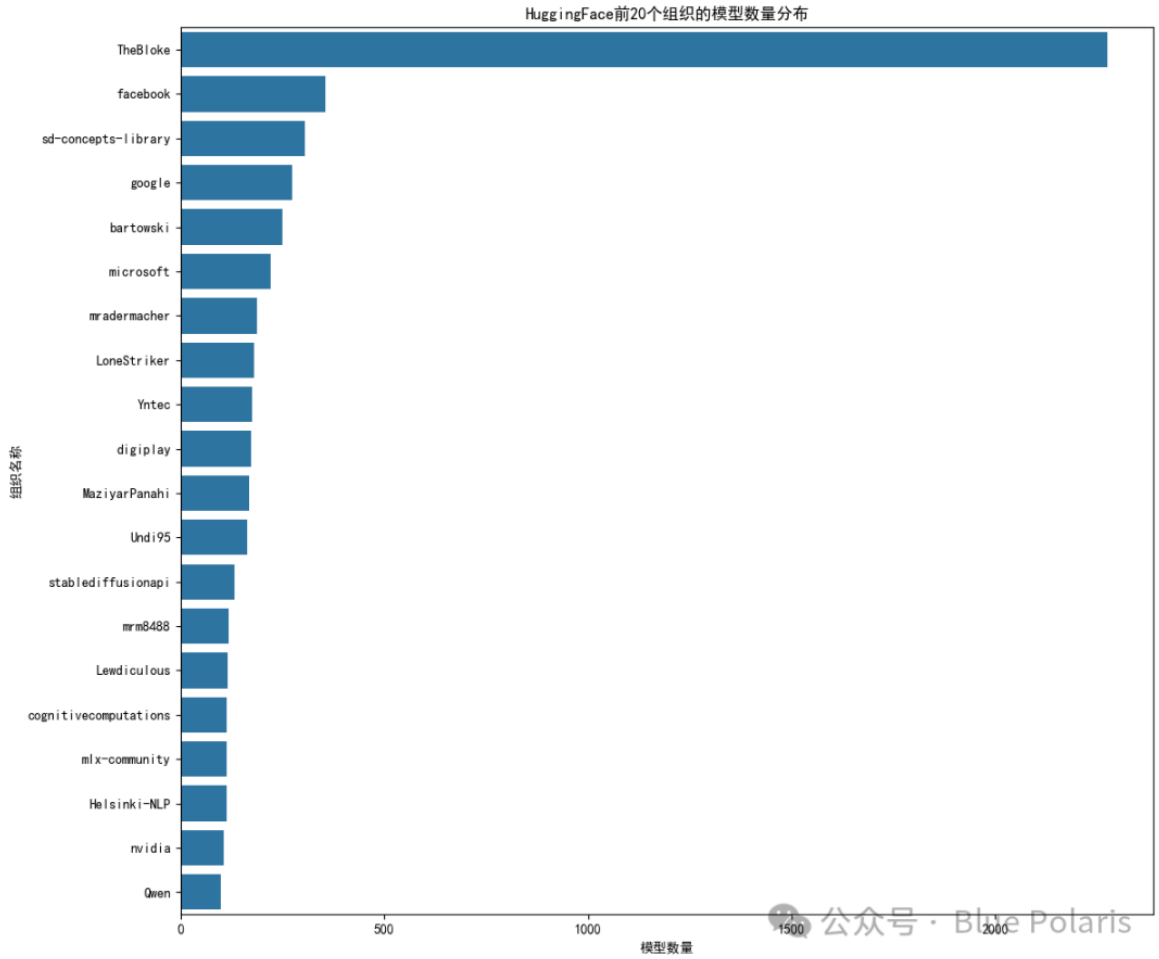

图7 贡献模型数量最多的前20个组织/个人

"其中,TheBloke贡献的模型数量遥遥领先,远超其他组织。facebook和sd-concepts-library紧随其后,分别位列第二和第三。google和bartowski的模型数量相对较少,但仍进入了前五名。"

"这个分布反映了在AI模型开发和共享领域,个人贡献者(如TheBloke)和大型科技公司(如facebook和google)都在发挥重要作用。同时,专注于特定领域的库(如sd-concepts-library)也在模型生态系统中占有重要地位。"

*国产Qwen 开发团队模型贡献数量在第20名位置.

- **model_size聚类+数量分布:**

(此处删除缺失值,因此总数为9022个模型)

🦊 104K - 145M prams(1377个模型)

🐼 146M - 725M prams(1130个模型)

🦁 728M - 3.7B grams (1807个模型) **

🦅 3.7B - 20.9B prams (4708个模型) ***

🐘 21.1B - 703B prams(1390个模型) **

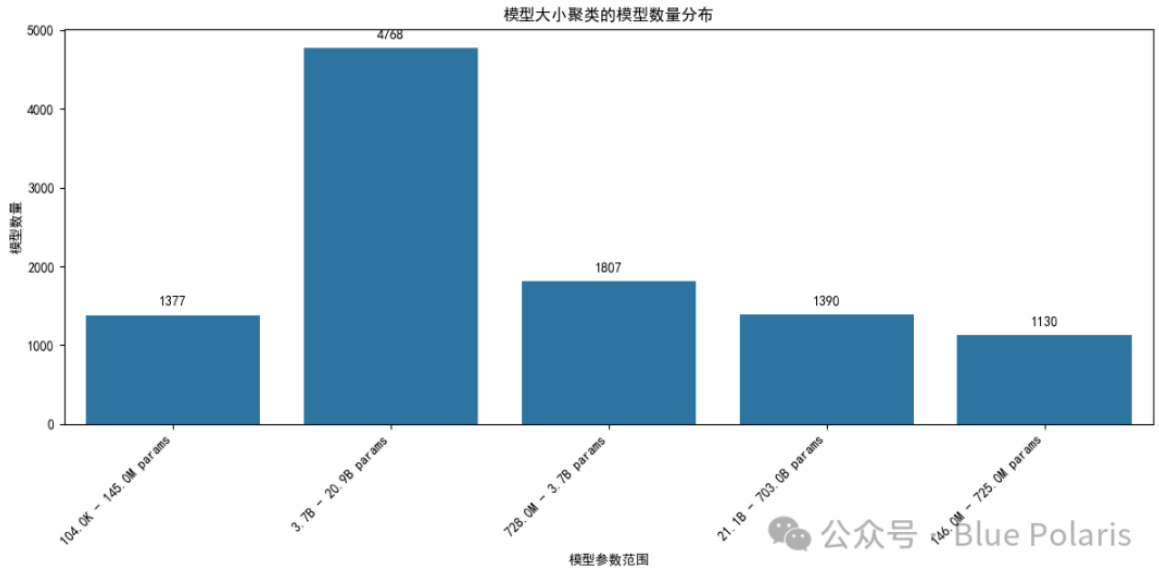

图8 模型大小聚类--模型数量分布

"这个分布显示,中等规模的模型数量最多,远超其他规模区间。较大规模和较小规模的模型数量紧随其后。该分布反映了在模型开发中,研究者和开发者倾向于平衡模型性能和计算资源需求,同时也表明了不同规模模型在各种应用场景中的需求。"

- **model_size-downloads-likes分布**

(1)Top 20 Task-平均下载量-平均收藏量

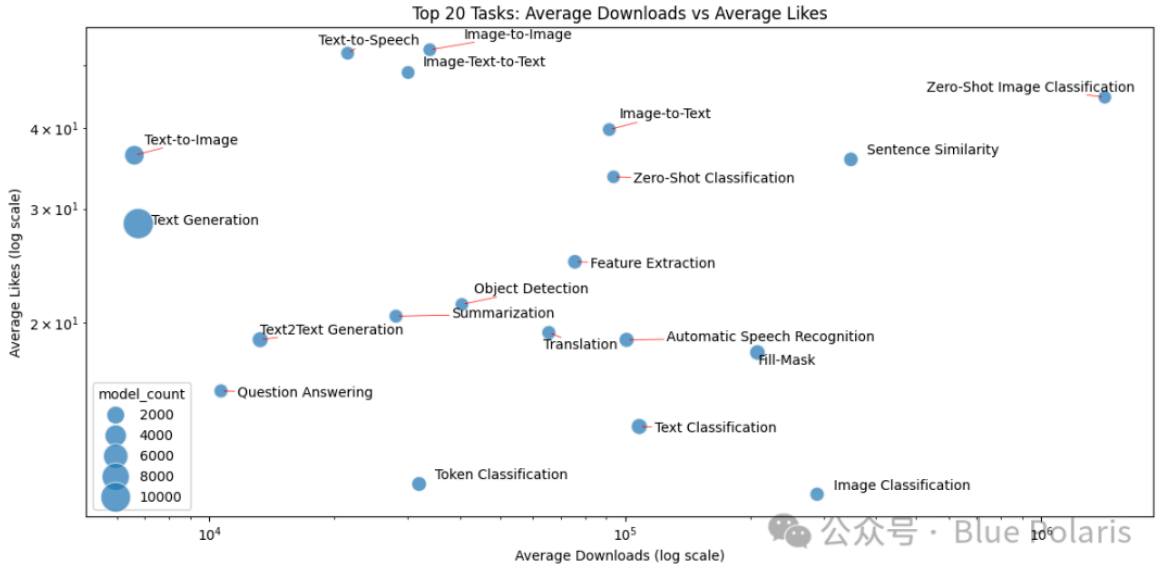

图9 Top 20 Task-平均下载量-平均收藏量

(由于原始数据存在类别极度不平衡,因此平均数仅作参考)

🔍 主要观察:

1. 受欢迎的模型:
   - Zero-Shot Image Classification（零样本图像分类）和Sentence Similarity（句子相似度）位于右上角，表明这些任务的模型十分受欢迎。
2. 新兴任务:
   - Image-to-Image（图生图）和Text-to-Speech（文本生语音）位于左上角，虽然下载量不高，但点赞数较高。
3. 基础任务:
   - Text Classification（文本分类）、Translation（翻译）等位于中下部，代表了一些基础但广泛使用的AI任务。

(2)Top 20 Organization-平均下载量-平均收藏量

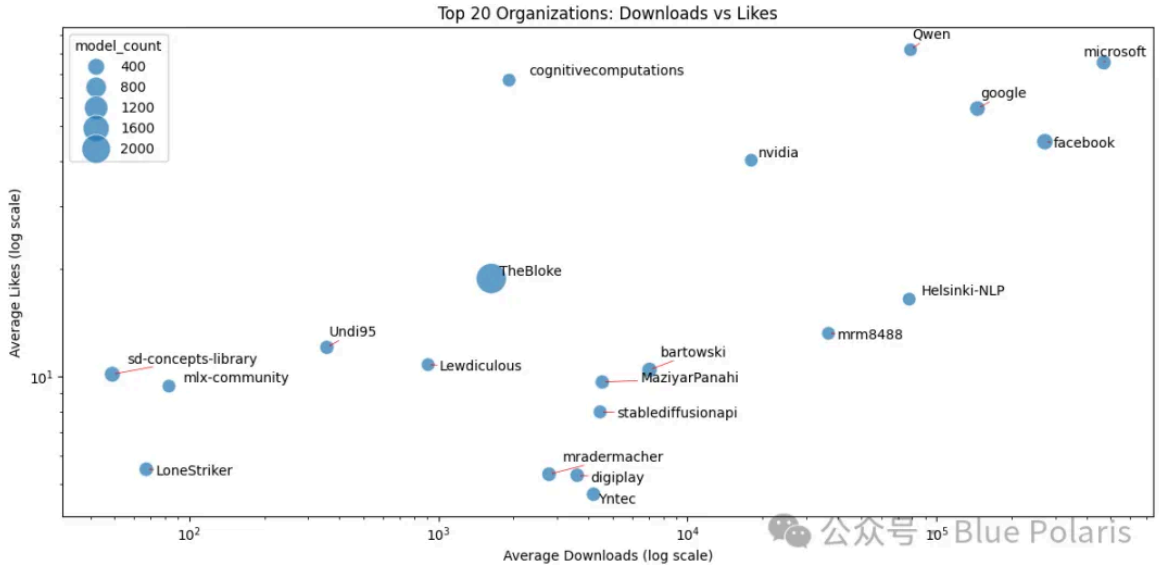

图10 贡献模型数量前20的组织/个人-平均下载量-平均收藏量

(由于原始数据存在类别极度不平衡,因此平均数仅作参考)

🔍 主要观察：

1. 大型科技公司主导：Microsoft、Google和Facebook位于图表右上角，表明它们的模型既受欢迎（高下载量）又高质量（高点赞数）。
2. TheBloke现象：作为个人贡献者，TheBloke的表现尤为突出，贡献模型数量多，下载量和点赞数都相当可观。
3. 长尾效应：图表左下角聚集了许多较小的组织或个人贡献者，形成"长尾"，体现了开源社区的多样性。

📊 趣味数据：

- 最高下载量：Microsoft
- 最多点赞：Qwen
- 模型数量最多：TheBloke

🤿 04：ML实验

最后本项目进行了一个简单的实验环节,以Task为标签，以其他指标为特征，进行机器学习训练和评估一个多分类任务。探究各个机器学习模型在该数据集上面的效果,

结果如下图所展示:

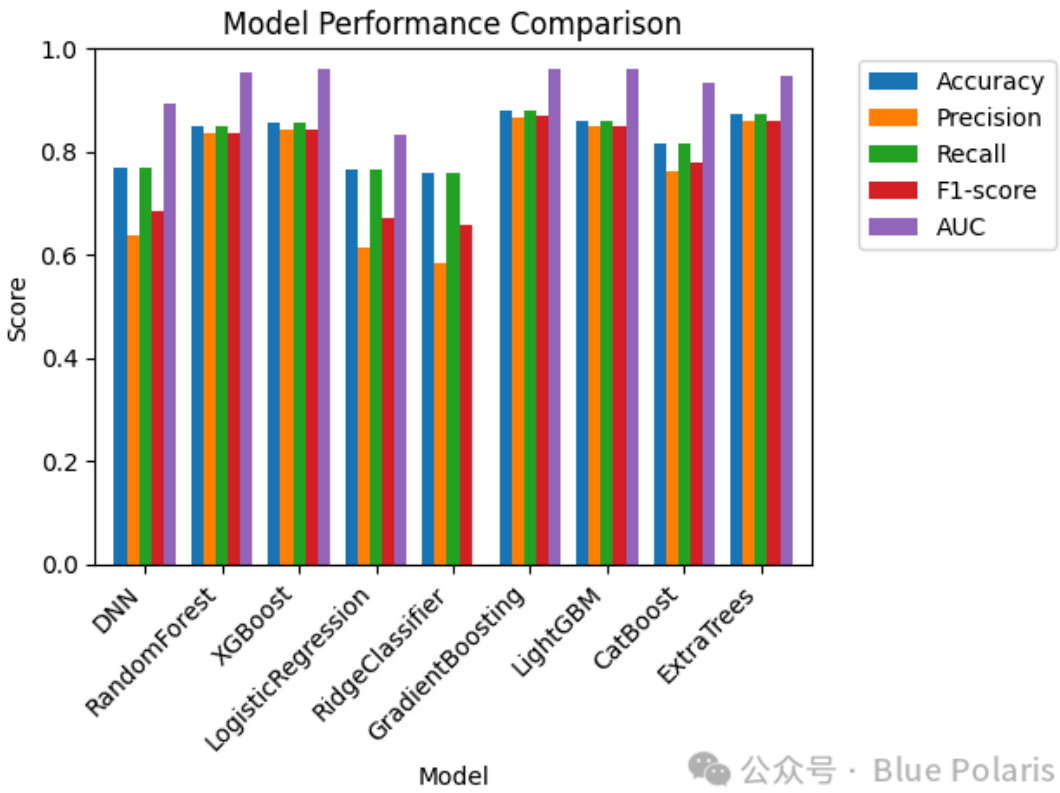

图11 机器学习模型实验

- 从整体来看，XGBoost/LightGBM模型在大多数指标上表现较为出色。
- ExtraTrees和随机森林（RandomForest）作为基于集成学习的模型，也展现出了稳定且高效的性能，特别是在准确率和精确率方面。
- 值得注意的是，逻辑回归和岭分类器虽然是相对简单的线性模型，但在某些指标上仍能保持与DNN相当的性能。
- 梯度提升模型显示出均衡和优秀的性能。
- 深度神经网络模型在此比较中表现相对较弱。

🌊05：附录

(其他未详细展示数据在此处)

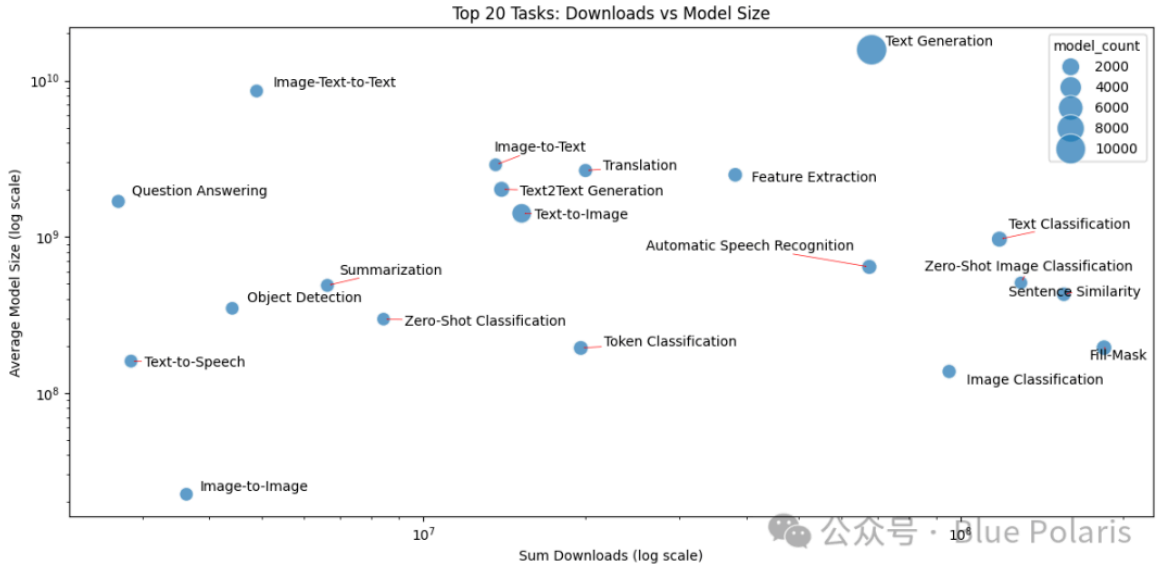

图1 Top 20 Tasks - Total downloads - Average model_size

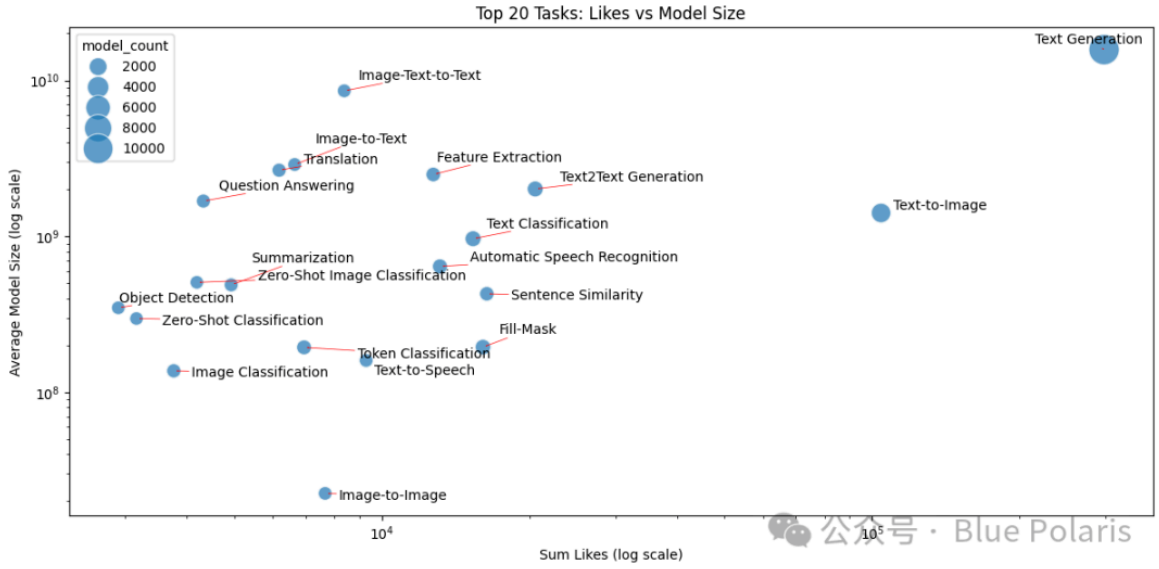

图2 Top 20 Tasks - Total likes - Average model_size

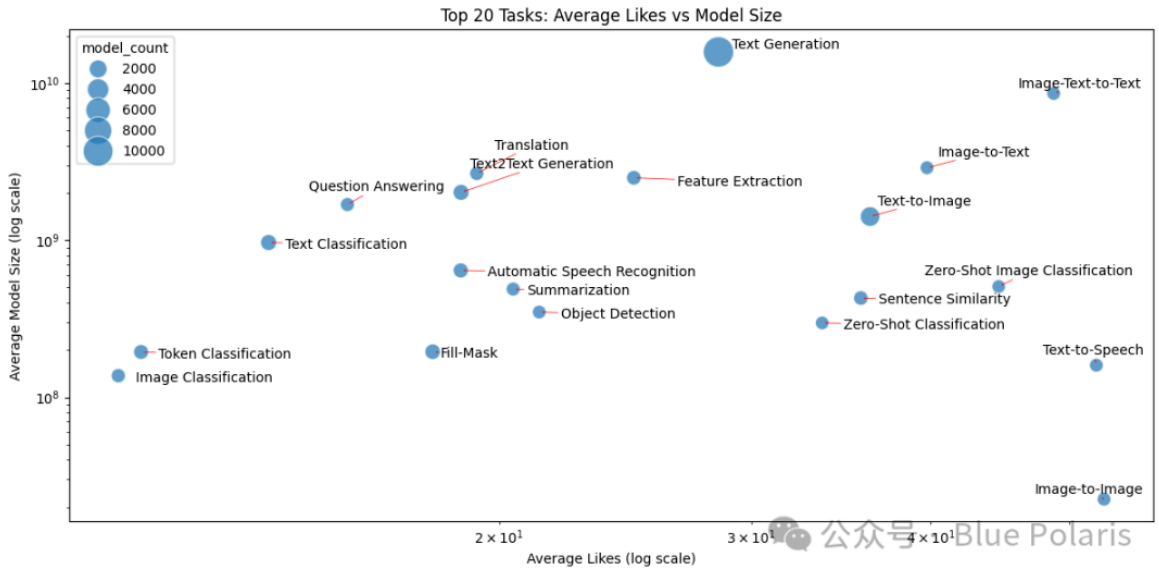

图3 Top 20 Tasks - Average likes - Average model_size

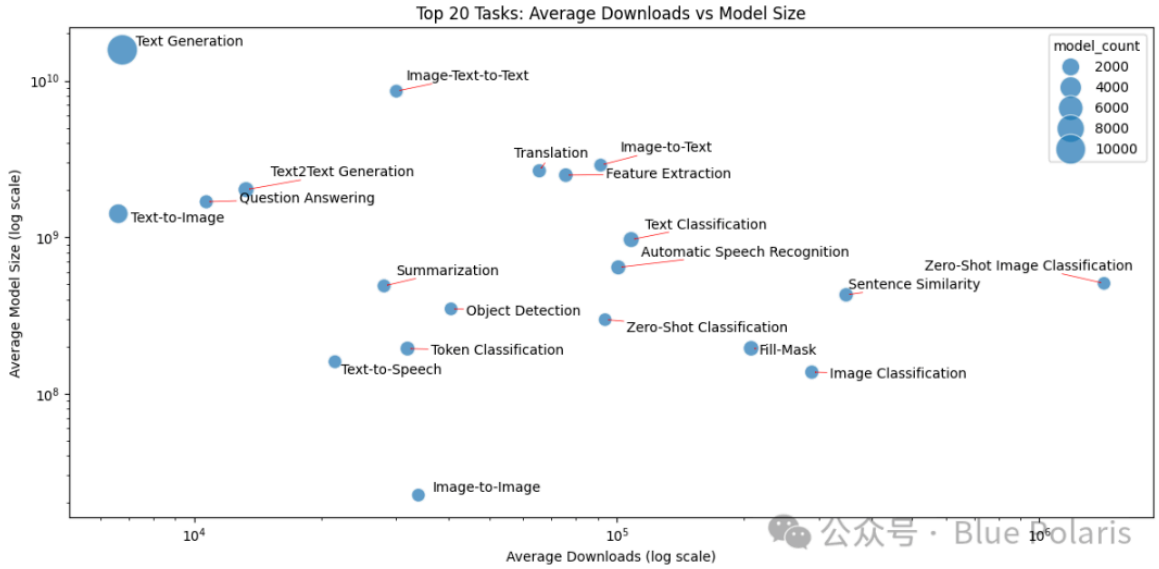

图4 Top 20 Tasks - Average downloads - Average model_size

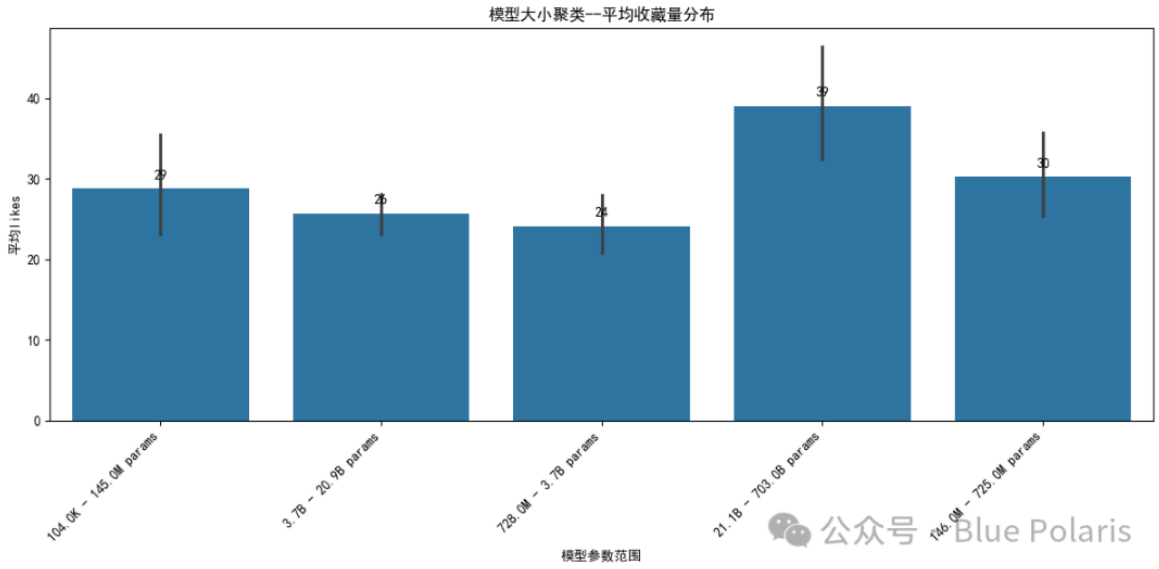

图5 model_size聚类+平均收藏量

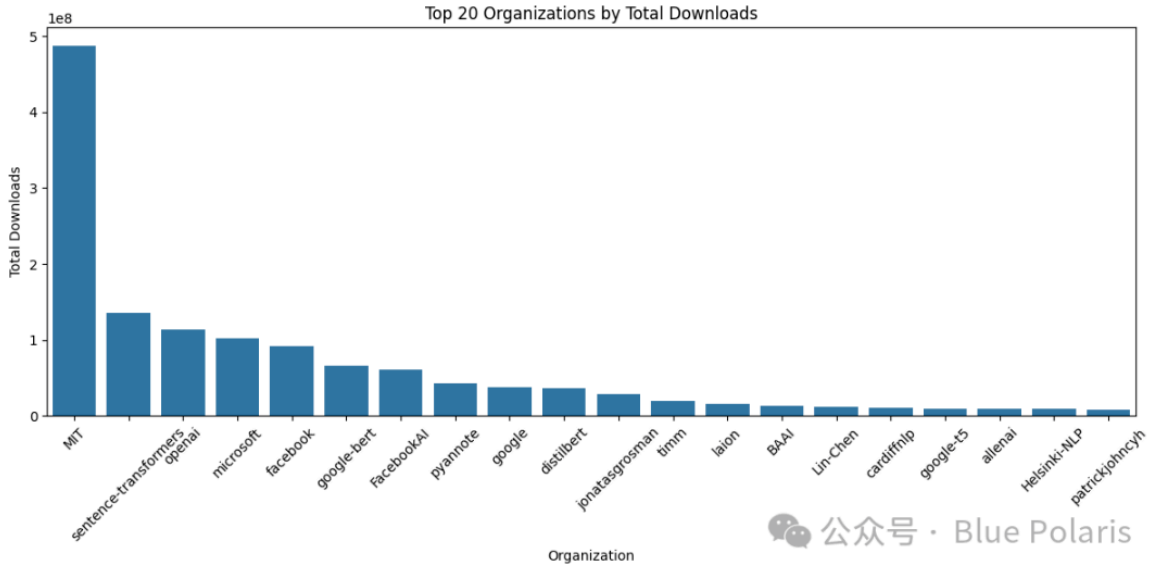
Top 20 Organizations by Total Downloads
1e8
Total Downloads
Organization
MIT
sentence-transformers
openai
microsoft
facebook
google-bert
FacebookAI
pyannote
google
distilbert
jonatasgrosman
timm
laion
BAAI
Lin-Chen
cardiffnlp
google-t5
allenai
Helsinki-NLP
patrickjohncyh
公众号 · Blue Polaris

图6

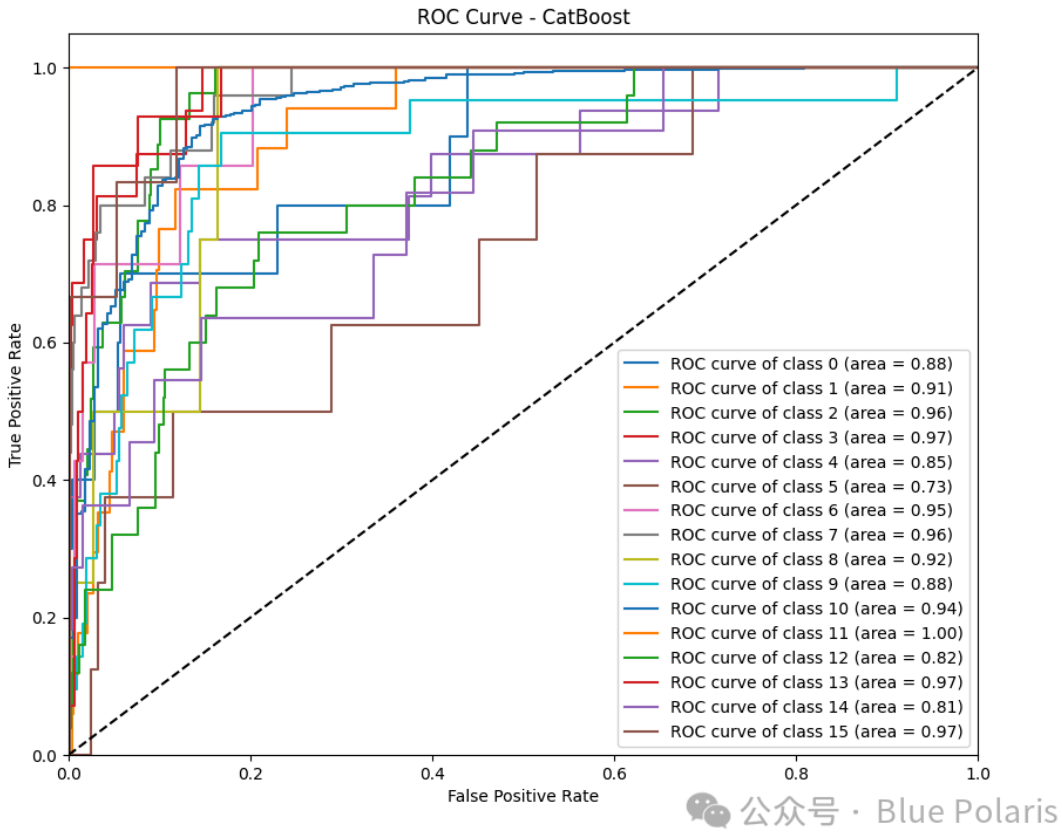
ROC Curve - CatBoost
True Positive Rate
False Positive Rate
ROC curve of class 0 (area = 0.88)
ROC curve of class 1 (area = 0.91)
ROC curve of class 2 (area = 0.96)
ROC curve of class 3 (area = 0.97)
ROC curve of class 4 (area = 0.85)
ROC curve of class 5 (area = 0.73)
ROC curve of class 6 (area = 0.95)
ROC curve of class 7 (area = 0.96)
ROC curve of class 8 (area = 0.92)
ROC curve of class 9 (area = 0.88)
ROC curve of class 10 (area = 0.94)
ROC curve of class 11 (area = 1.00)
ROC curve of class 12 (area = 0.82)
ROC curve of class 13 (area = 0.97)
ROC curve of class 14 (area = 0.81)
ROC curve of class 15 (area = 0.97)
公众号 · Blue Polaris

图7

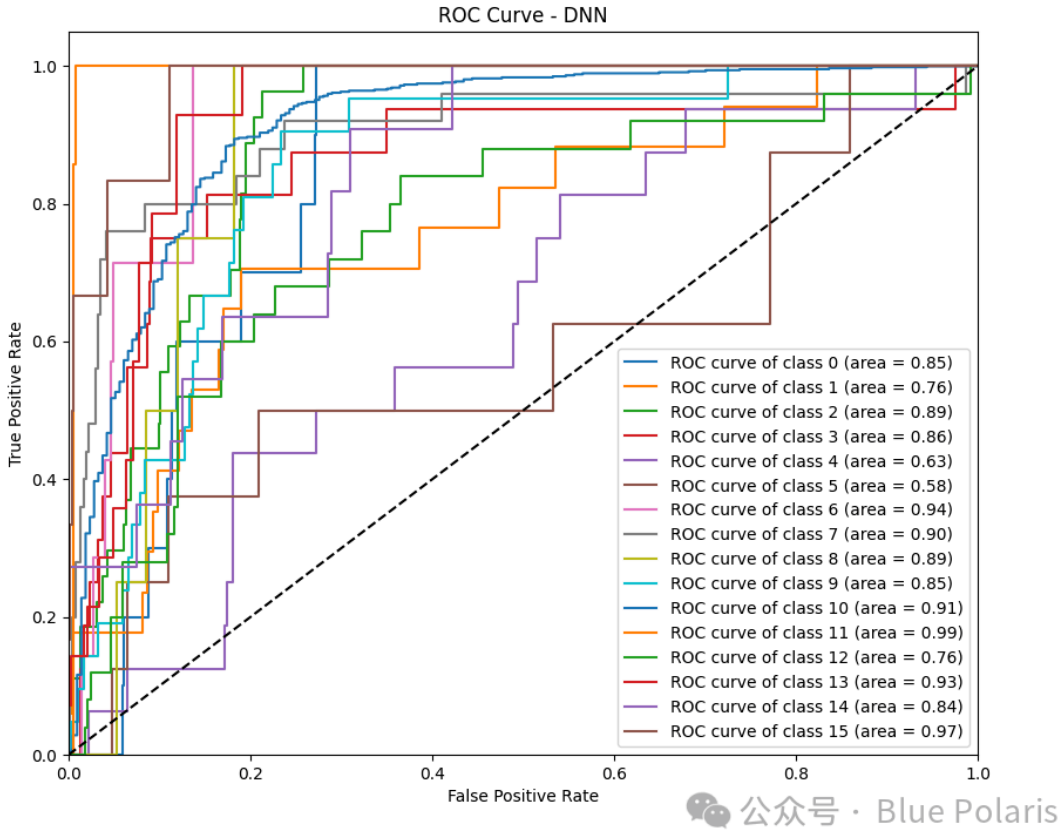

图8

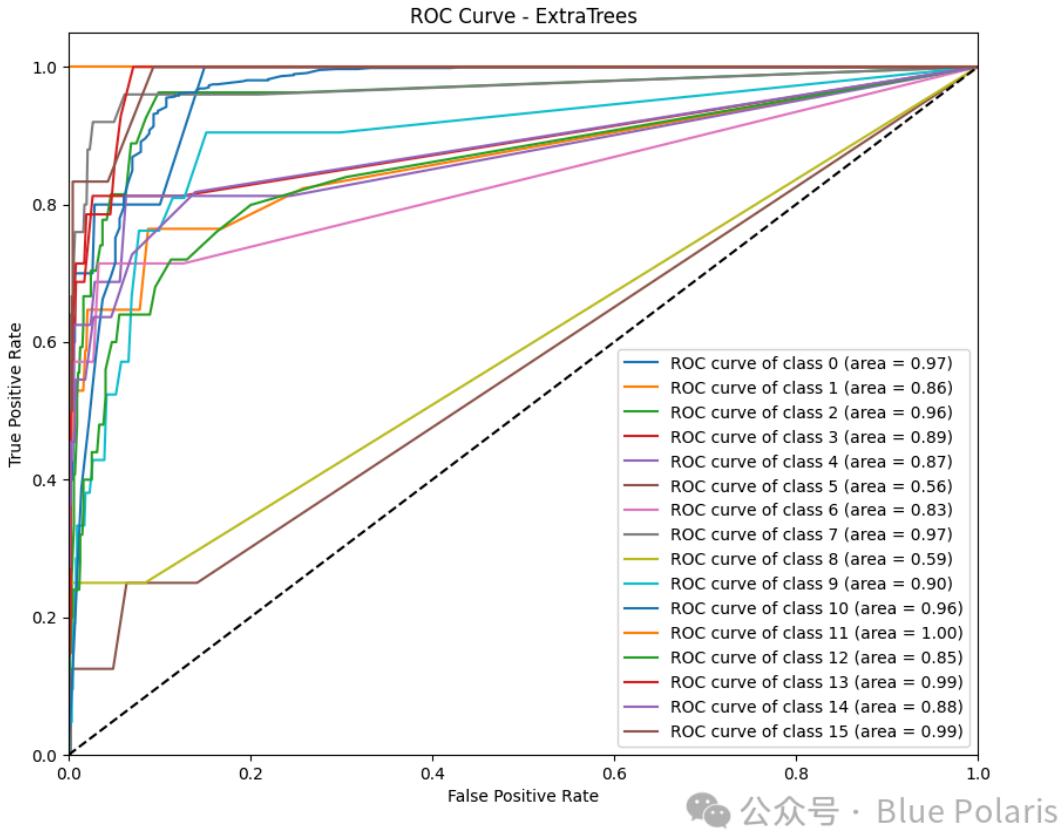

图9

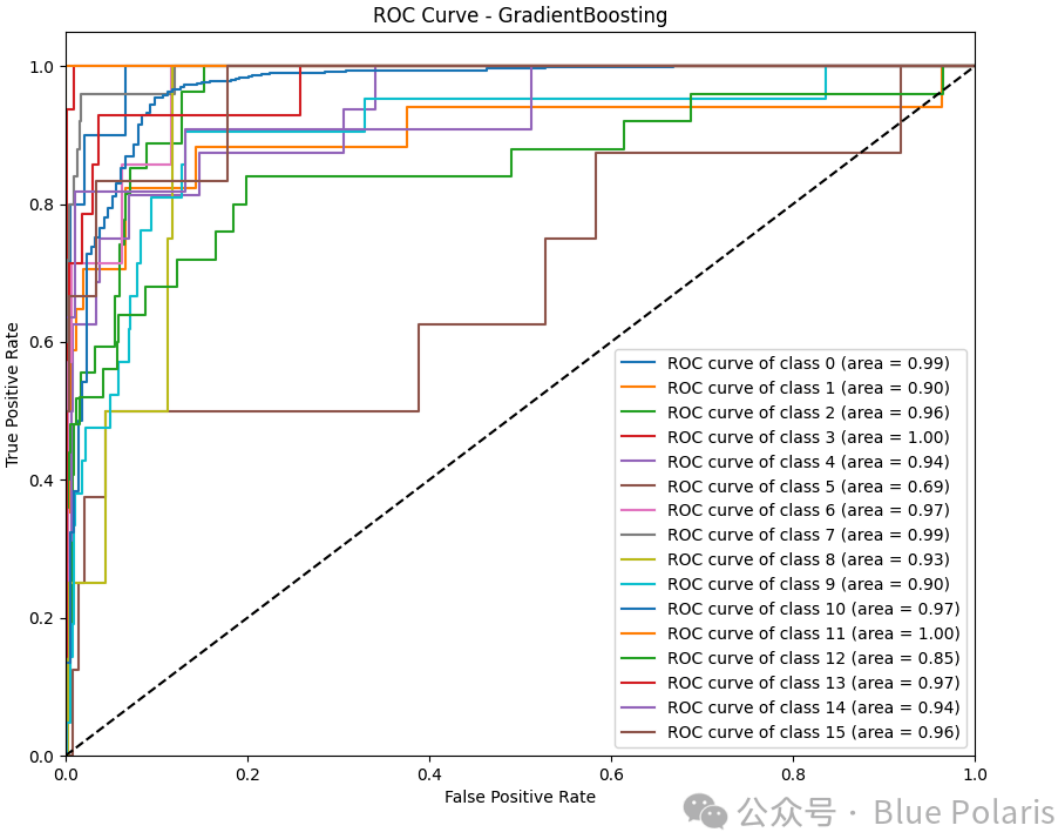

图10

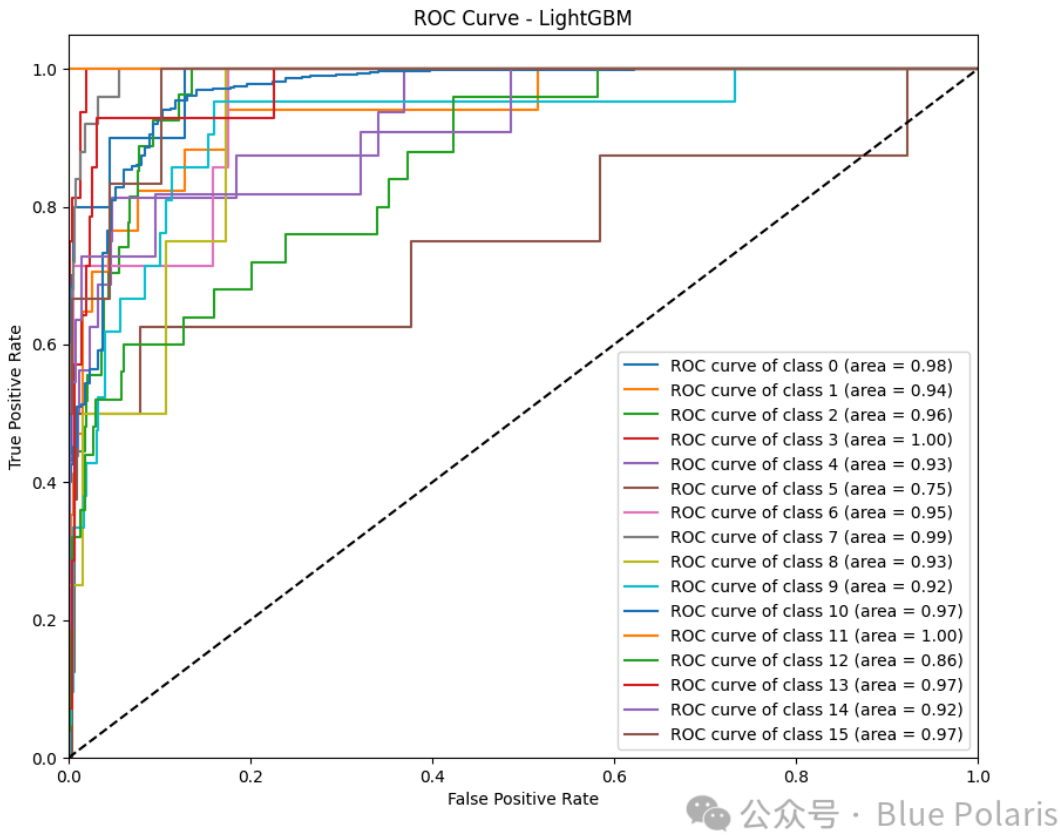

图11

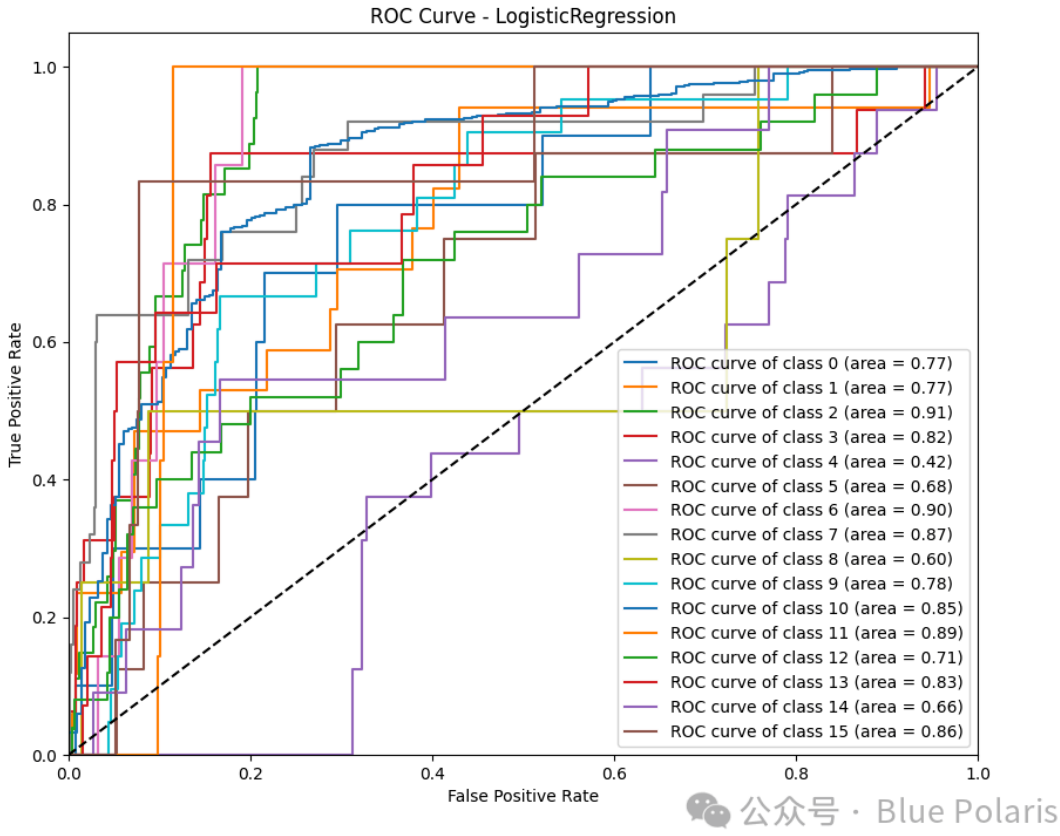

图12

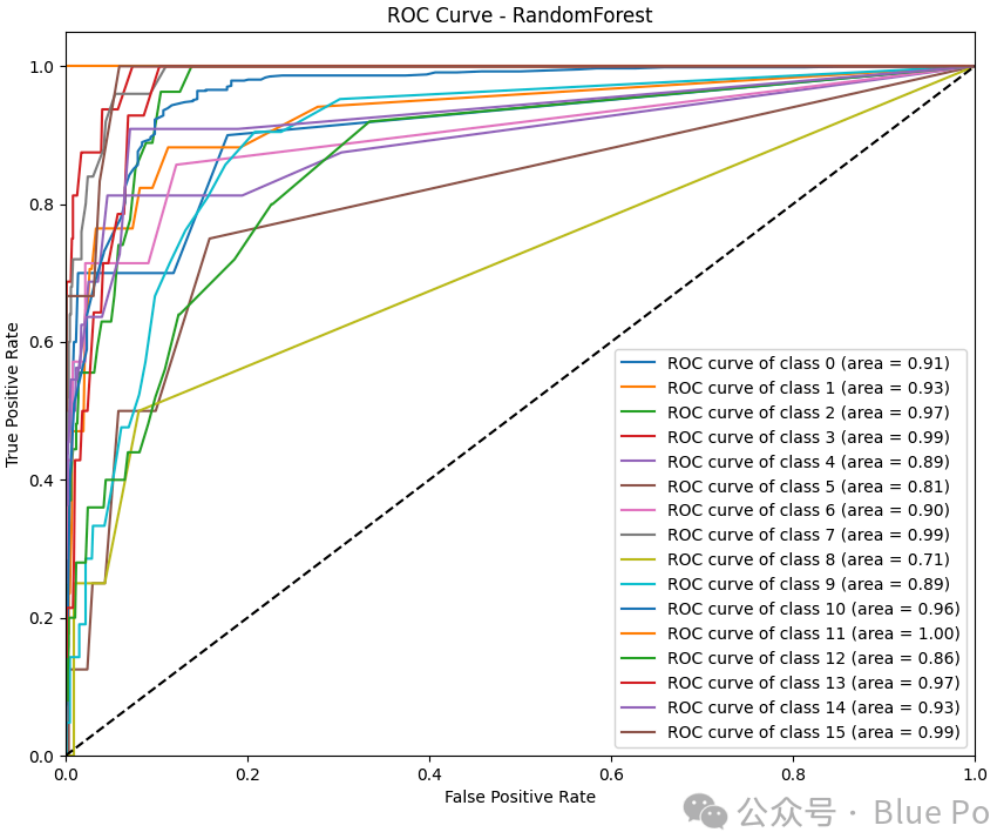

图13

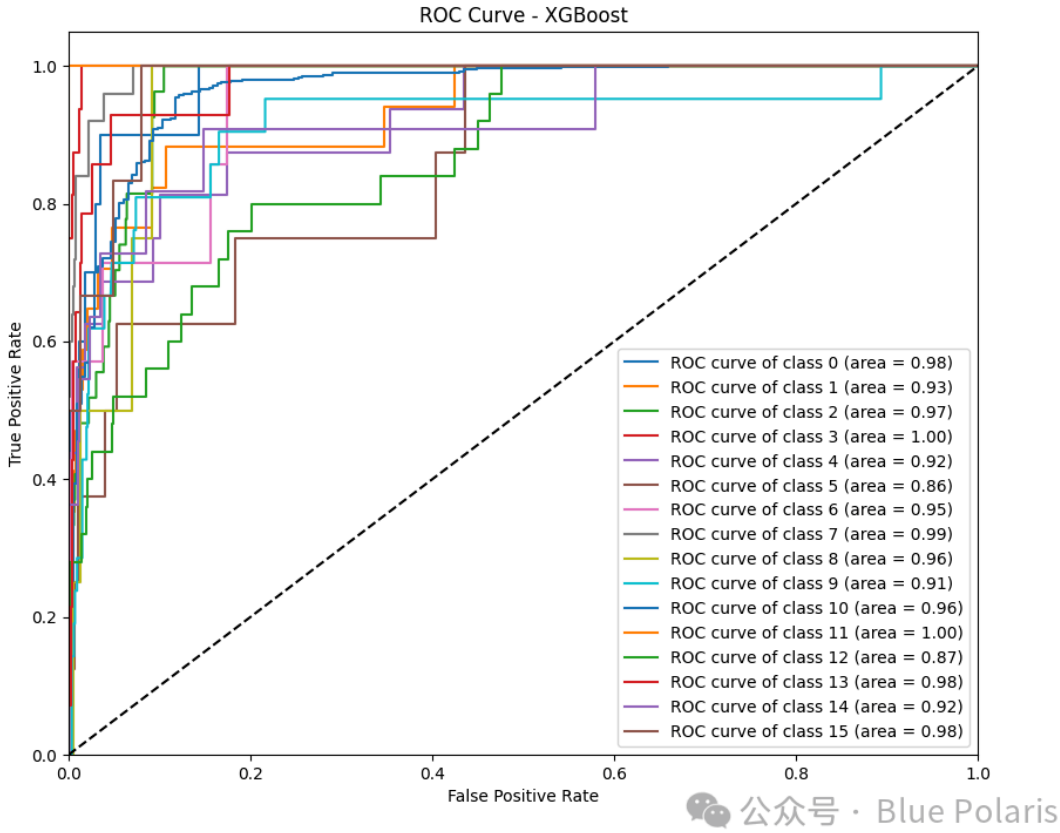

图14

内容剧情演绎，仅供娱乐

修改于2024年08月01日