各商品类别下BGM对短视频带货效果分类的探索

曹英瑞 郭礼华 廖兰宇

摘要：本文是对抖音带货短视频中BGM的研究，主要是探索不同商品类别下BGM是否可以分为带货效果好和带货效果差两类并对其加以识别，附加对不同带货类型下BGM特点的比较。我们爬取了抖音男装、女装、手机数码、食品生鲜、美妆护理、日用百货六类商品下各1400个视频，总计8400个视频，以抖音数据网站蝉妈妈提供的短视频带货预估销售额为标准，将各商品类别下视频分为带货效果较好和带货效果较差两类。我们使用了 Adobe Audition 来进行初步的音频处理，主要是将带有达人讲解声的音频进行人声消除而留下BGM，然后剔除人声消除效果差的音频，同时将音频时长统一为8s方便后续处理。初步处理音频完成后，我们使用VGGISH模型提取音频特征，得到8x128维矩阵，然后使用SVM，AdaBoost，随机森林三个模型对特征进行学习，得到分类报告，从分类报告中分析相应结论。

关键词：抖音 BGM VGGISH 机器学习

目录

第一章 绪论

1.1 研究原因

项目初期讨论研究题目时，我们认为本次研究有如下难点：

1.数据由自己获取。对于抖音带货短视频的商业化研究势必要涉及对商品销量，视频热度等的分析，而调研时我们发现这些数据在各大抖音数据网站上基本需要付费获得，且价格很高。若是人工爬取，抖音本身有反爬机制，同时这些商业数据需要大量分析才能获得，时间有限的情况下进行商品销量和热度的研究意味着要花大量时间在获取数据上，成本太高。

2.音频分析困难。老师提出的主题是与 BGM 相关，而分析 BGM 本身是一件非常困难的事，涉及到乐理和声学等各方面的知识，要对音频性质做出深入的分析，我们小组认为并不现实。

3.视频内容干扰。抖音 BGM 并不是一段独立的音频，其与视频内容的结合大多比较紧密，如海草舞短视频就是为海草舞这首 BGM 量身打造的营销视频，海草舞能够风靡抖音除开其自身魔性洗脑的旋律，与舞蹈的结合也是很大的助力。因此某些选题，如分析 BGM 在抖音的热度曲线会受到这方面(公司营销和视频内容)的干扰，而分析视频内容或者减小其干扰在技术层面上也很难实现。

在分析难点后，我们决定探索不同商品类别下 BGM 是否可以分为带货效果好和带货效果差两类并对其加以识别。选择这个研究主题的原因主要有：

1.数据获取较为简单。我们只需要爬取短视频及其销量、点赞量、转发量、评论量，相比于爬取随时间变化的曲线图要容易实现。

2.通过人工筛选，可以减少视频内容对分类的影响。在筛选视频时，我们主要挑选讲解性的视频，剔除了剧情向等类型的短视频，减少了视频形式对分类的影响。

1.2 研究意义

短视频带货日益火爆，研究如何提高短视频带货效率也有较高的实用价值。BGM 作为短视频的一个组成成分，依照生活经验我们认为合适的 BGM 能吸引观众注意力，调动观众情绪，提高短视频带货效率。我们希望通过数据分析探索不同类别下 BGM 是否可以分为带货效果好和带货效果差两类，如果可行，那么在短视频博主发布短视频时就可以参考我们的分类模型，判断自己所选 BGM 是否合适。

第二章 数据获取与处理

2.1 数据获取

本次研究的主体是抖音平台上面的带货相关的小视频，我们下载了多个类别的带货小视频并且收集了每个小视频对应的点赞量、转发量、评论量、预估销量和预估销售额作为小视频带货效果好坏的参考，数据的主要来源为蝉妈妈网站(www.chanmama.com)，视频的获取方式则通过集成工具爬取得到。

此次研究选取了男装、女装、手机数码、食品生鲜、美妆护理、日用百货六个类别的商品，这几个类别囊括了日常生活中大部分的常见物品，类别之间的消费人群也有较显著的区别，因此我们希望通过对这些类别的小视频进行分析以探究 BGM 是否对小视频的带货效果会产生显著的影响。

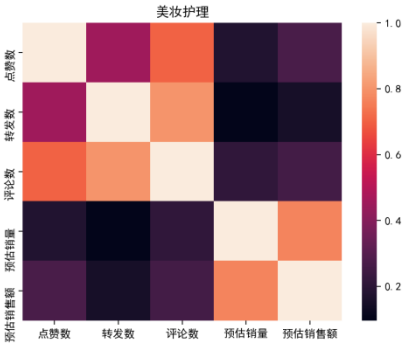
在研究的初期，我们共收集了从 2020 年 7 月 17 日至 2020 年 7 月 23 日七天内全网带货小视频的相关数据（其中女装还收集了 2020 年 7 月 16 日的数据），每个类别每天对应 500 条数据，一个类别总共收集到 3500 条数据，六个类别共收集得到 24500 条数据，数据的基本分布情况如下

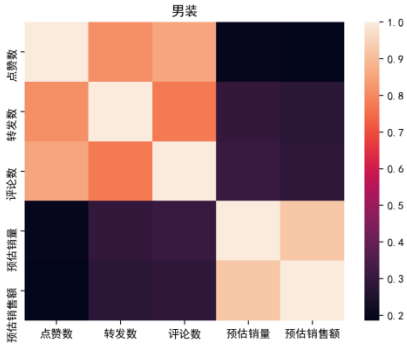
	点赞数	转发数	评论数	预估销量	预估销售额
count	24500	24500	24500	24500	24500
mean	3534.986	63.51449	141.6288	45.52102	2222.222
std	26518.71	747.9357	955.7625	435.5972	11457.08
min	0	0	0	0	0
25%	130	1	6	0	0
50%	398	4	20	5	239.6
75%	1135	18	63	19	1287
max	1752000	62000	52000	33000	659000

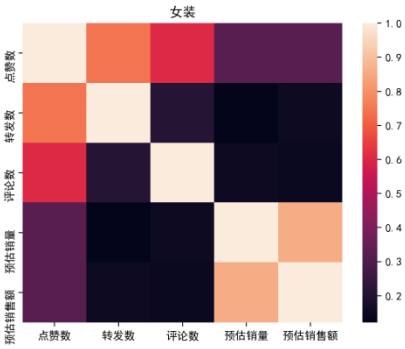
由数据分布较高的方差中我们可以看出不同小视频之间的传播效果和带货效果上确实存在较为显著的差别，接着我们对不同的统计量之间进行了相关性的分析以进一步探索他们之间的关系，由皮尔逊相关系数公式，

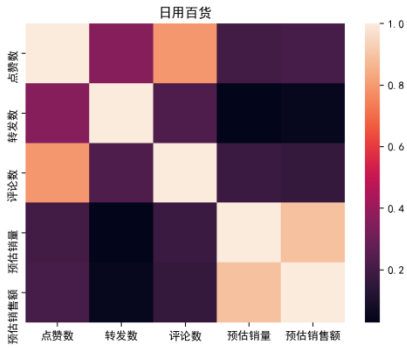
$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

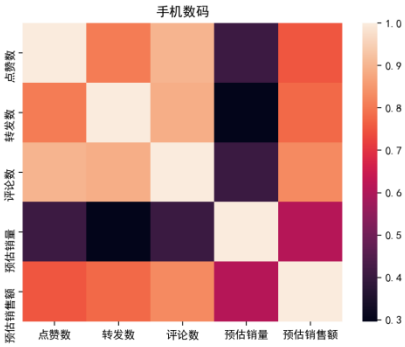
我们计算了不同统计量之间的相关性程度，并得到以下结果

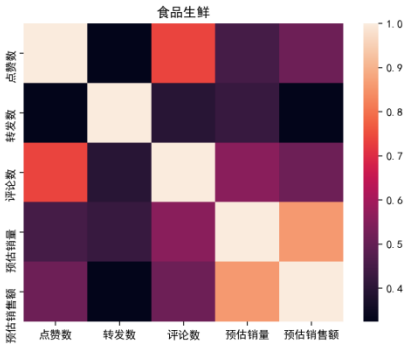

美妆护理
点赞数
转发数
评论数
预估销量
预估销售额
1.0
0.8
0.6
0.4
0.2

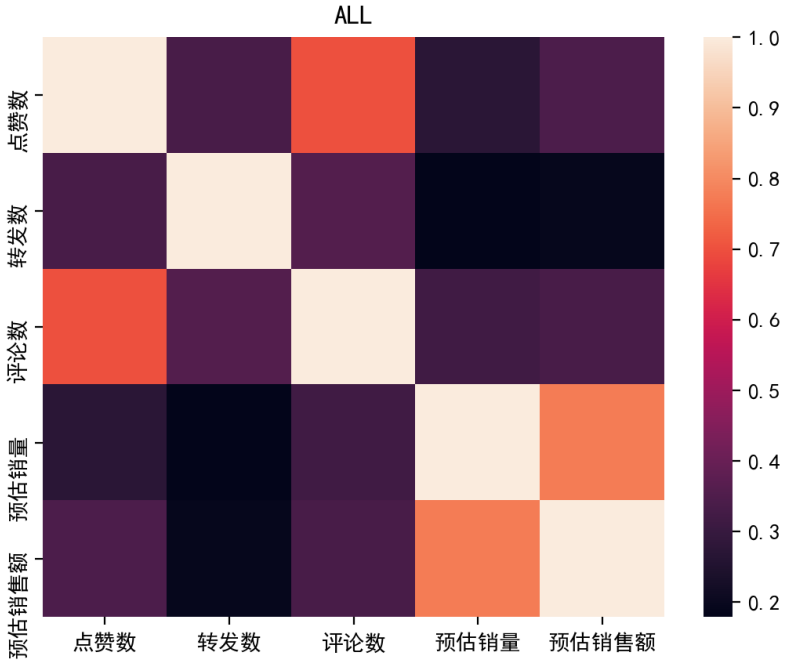

男装
点赞数
转发数
评论数
预估销量
预估销售额
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2


女装
点赞数
转发数
评论数
预估销量
预估销售额
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2


日用百货
点赞数
转发数
评论数
预估销量
预估销售额
1.0
0.8
0.6
0.4
0.2

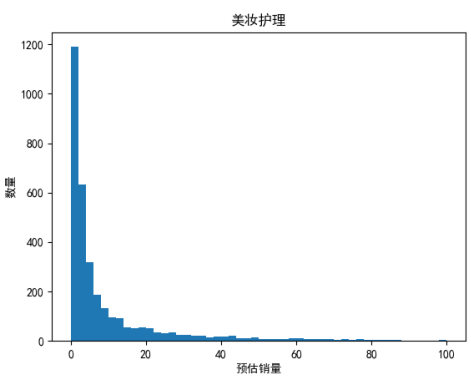

手机数码
点赞数
转发数
评论数
预估销量
预估销售额
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3

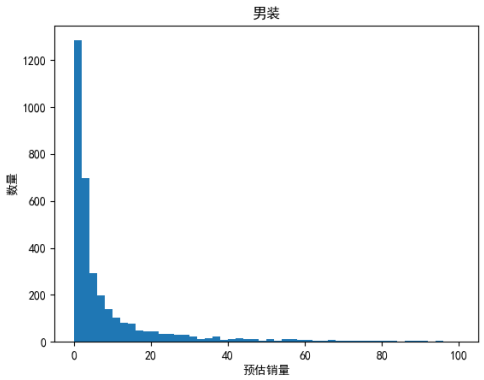

食品生鲜
点赞数
转发数
评论数
预估销量
预估销售额
1.0
0.9
0.8
0.7
0.6
0.5
0.4

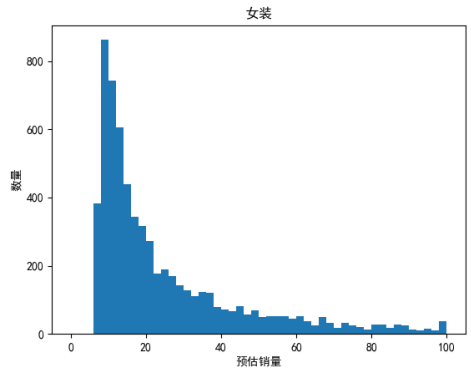

ALL
点赞数
转发数
评论数
预估销量
预估销售额
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2

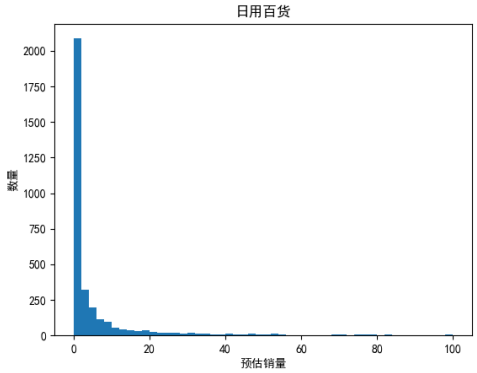
从六个类别的相关性热力图以及所有数据的热力图中我们可以初步得到，对于大多数的小视频来说，视频的点赞数、转发数及评论数之间有较为紧密的关联，而预估销量和预估销售额之间有较为紧密的关联，但这两组相互关联的变量之间的相关性却不大。因此我们认为视频的点赞量、转发数及评论数衡量的更多是小视频本身在传播过程中受观众喜爱的程度，而预估销量和预估销售额才是更加直观地衡量了小视频的带货效果的好坏，考虑到预估销售额会受到销售商品的价格的影响，所以我们决定采用数据中的预估销量作为最终衡量小视频带货效果好坏的指标。

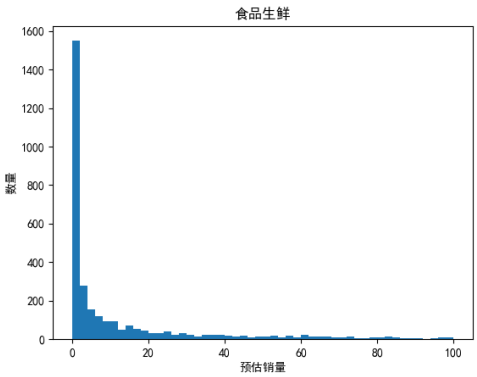
我们对预估销量的分布做了简单的统计后发现结果如下

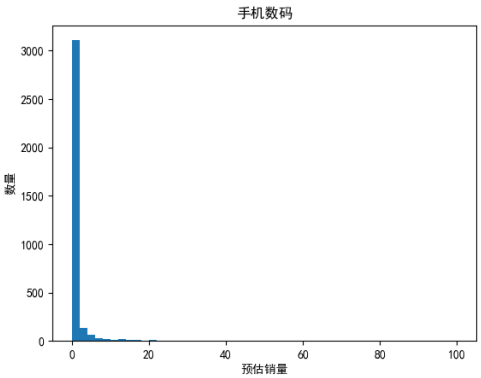


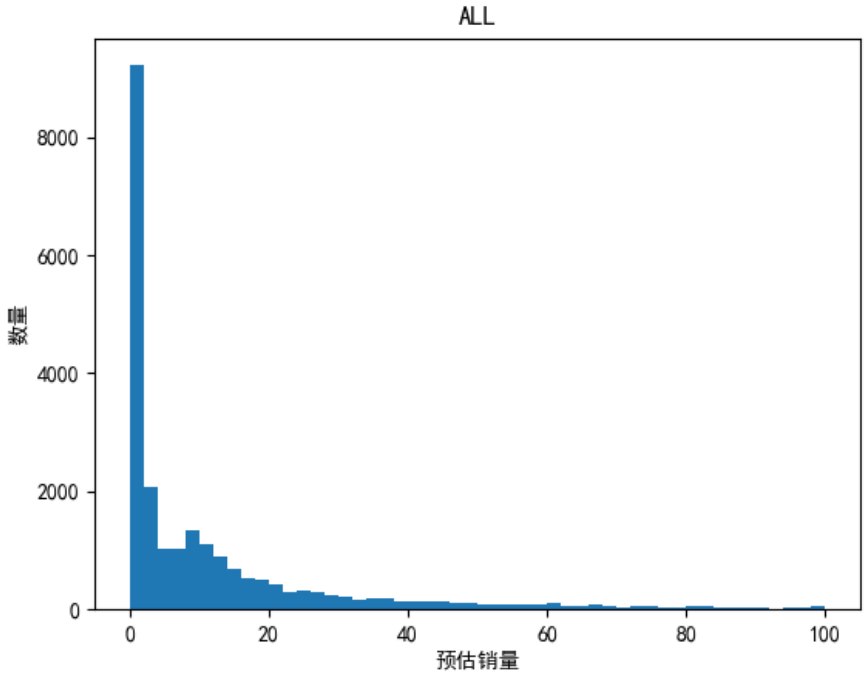












或许是因为获取的数据是来自全网带货小视频，所以许多小视频的带货效果都不是很理想，为了增大视频带货效果之间的区分度，我们决定在每一个类别里面各选取预估销量排在前 20%的小视频作为带货效果好的视频，而后 20%的小视频作为带货效果差的视频，经过抖音视频采集工具的下载之后，我们共采集到 8400 个视频，这些将作为研究下一步的数据储备。

2.2 数据处理

2.2.1 视频筛选

如何定义一条短视频带货效果好坏？在没有找到预估销售额之前，我们打算以(点赞量+转发量+评论量)/发布天数，即日曝光率来近似带货效果的指标，但是在没有转化率指标的情况下，这种近似效果并不好。在查询各大抖音数据网站后，我们找到了可以爬取的预估销售额，并且该预估销售额是基于达人粉丝数和视频曝光率综合给出的预测，于是我们打算以预估销售额作为短视频带货效果的评判指标。

我们选取了全网同类带货视频中预估销售额前 20%的短视频作为带货效果较好的样本，后 20%的短视频作为带货效果较差的样本。以预估销售额而不是预估销量作为评估指标是考虑到同类商品价格差距也会影响观众的购买欲望，进而影响销量，因此预估销售额评估效果更好。虽然粉丝基数对预估销售量的排名也有影响，但是粉丝基数较大的达人制作短视频时挑选适配 BGM 的概率更大，因此我们没有过多考虑达人粉丝基数对短视频 BGM 划分的影响。短视频的内容也是影响带货效果的因素之一，因此我们在筛选短视频时尽量挑选了纯讲解展示视频，剔除了剧情向短视频来减少视频形式的影响。

综上，我们将筛选后各类商品下预估销售额前 20%的短视频的 BGM 划分为适合该类商品的 BGM，预估销售额后 20%的短视频的 BGM 划分为不适合该类商品的 BGM。

2.2.2 音频处理

为了获取短视频的 BGM，我们使用了 Adobe Audition 对带有人声讲解的短视频进行人声消除处理，剔除人声消除后无法分辨 BGM 的音频，而纯展示无杂声的短视频则保留原音频。由于短视频时长的限制，综合考虑后我们选择统一截取 8s 音频，其中时长 20s 以内的音频取中间 8s，时长大于 20 的音频取副歌部分。

2.2.3 特征提取

本研究基于 VGGish 进行音频特征提取。VGGish 项目由 Google 的声音理解团队发布，并在 AudioSet 数据集上训练而成。生成模型包括 128 维的 embedding 特征向量，能够相对全面准确地描述音频。本研究使用 VGGish 预训练模型进行学习训练，在生成 128 维 embedding 特征向量集后进行了 PCA 白化等处理，使得数据美观度和可用性更高。由于特征向量规模受音频时长影响，本研究采用两种模式进行数据合并，即取平均值和向量整合（例如把 8*128 的向量集转为 1*1024 的向量集），并输入给下游模型进一步训练。

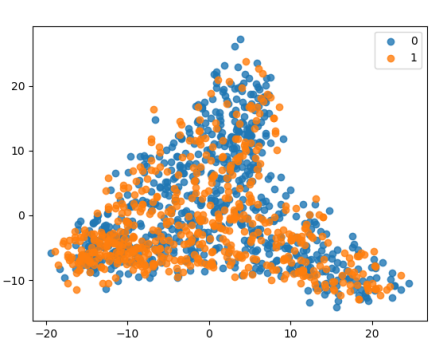
此外，音频特征分析部分主要使用 librosa 和 madmom 等的相关方法提取特征（包括频谱、节拍、和弦等），使用列表、字典等对特征数据进行整合，借助 matplotlib 等工具绘图，从而分析并得出结论。详见第四章。

第三章 机器学习训练模型

3.1 PCA 处理初步估计分类效果

由于数据量较少，为了保留更多信息，在实际训练模型时我们并未采用 PCA 处理数据，PCA 处理仅用于初步估计分类效果。

对所有类别下带货效果较好和较差的 BGM 特征数据进行 PCA 处理结果如下图所示(0 代表效果差，1 代表效果好)。可以看出两类 BGM 在 PCA 处理后较难分开，可能是因为部分在某类下效果较好的音频在另一类中被归为效果较差的音频，同时特征提取过程中以及 PCA 处理本身也损失了一部分特征信息。



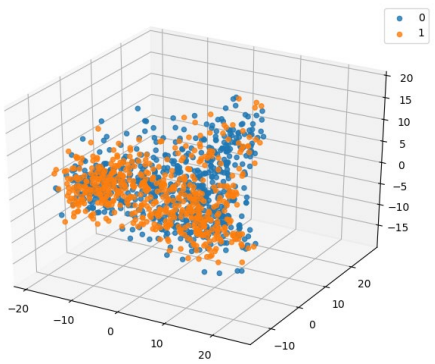


图 2-1 所有音频 PCA2 维图　　图 2-2 所有音频 PCA3 维图

在各个单独分类中效果较差和较好的两类音频的可分性比全部音频要稍好，说明在样本数据内可以尝试对音频按照带货效果好坏进行分类，具体如下图所示

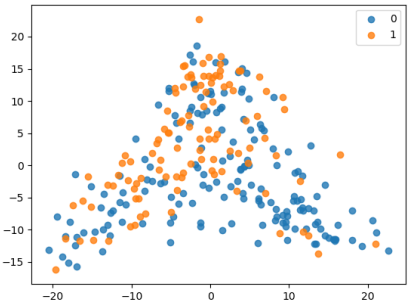


图 2-3 手机数码 BGM 分类

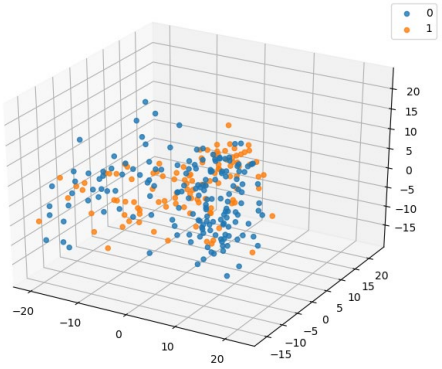


图 2-4 手机数码 BGM 分类

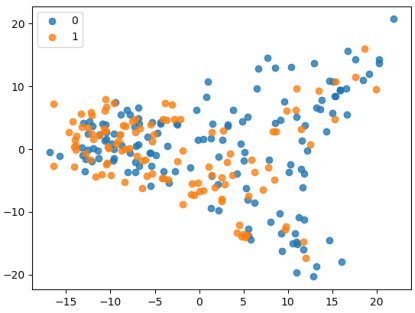


图 2-5 食品生鲜 BGM 分类

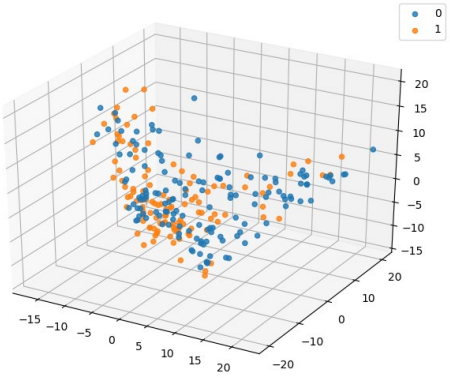


图 2-6 食品生鲜 BGM 分类

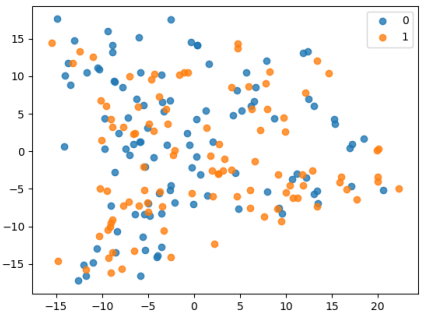


图 2-7 美妆护肤 BGM 分类

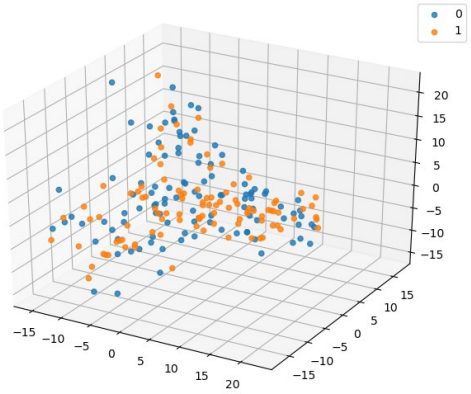


图 2-8 美妆护肤 BGM 分类

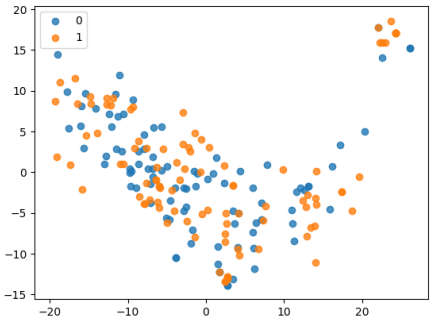


图 2-9 男装 BGM 分类

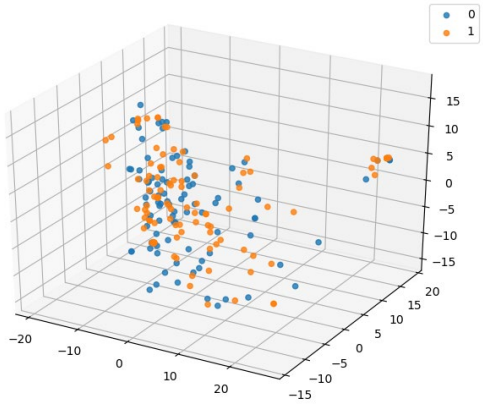


图 2-10 男装 BGM 分类

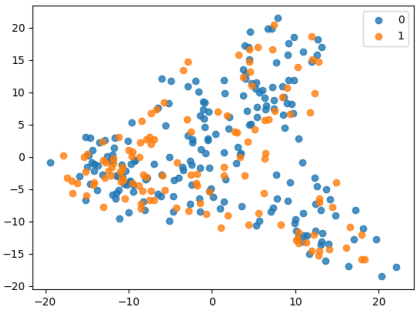


图 2-11 日用百货 BGM 分类

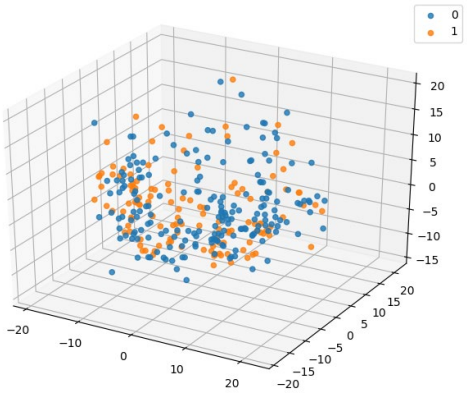


图 2-12 日用百货 BGM 分类

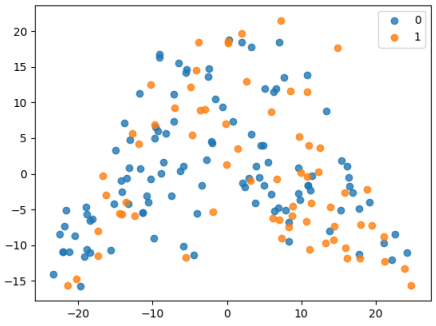


图 2-13 女装 BGM 分类

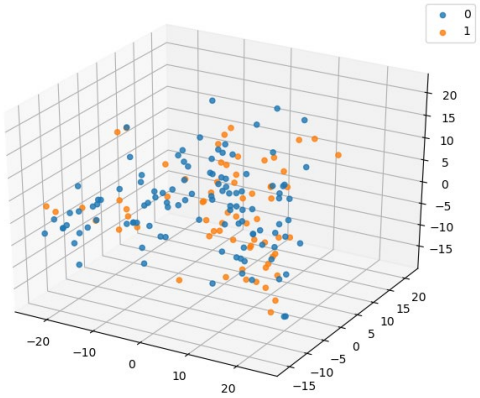


图 2-14 女装 BGM 分类

3.2 机器学习模型

3.2.1 SVM 模型处理

支持向量机（support vector machine，SVM）可以解决非线性分类问题，同时在高维数据和小样本量的情况下表现良好。这次研究我们使用了 SKlearn 下封装的 SVM 实现。

参数选择上，高斯核适用于线性不可分的情况，因此我们选择了高斯核作为核函数。在代码实现中我们以训练集：测试集=7:3 为比例随机划分训练集和测试集，在训练过程中按照 10 折交叉进行交叉验证，同时使用网格调参来寻找最佳参数组合（惩罚系数 C 和核函数参数 gamma），C 值的取值范围是$(2^{-5},2^{15})$，C 值越大对错误分类的惩罚越大，容易过拟合，而 C 值过小则容易欠拟合；gamma 的取值范围是$(2^{-9},2^{3})$，gamma 值隐含地决定了数据映射到新的特征空间后的分布，gamma 越大，支持向量越少，gamma 值越小，支持向量越多。

模型最后输出学习曲线、分类报告。

3.2.2 随机森林模型处理

随机森林能够处理高维度的数据，并且不用做特征选择，对数据集的适应能力强；由于采用了集成算法，本身精度比大多数单个算法要好，且在测试集上表现良好，由于两个随机性的引入（样本随机，特征随机），随机森林也不容易陷入过拟合，因此这次研究我们也使用了 SKlearn 封装的随机森林分类器进行横向比较。

参数选择上，我们使用了网格调参对最大弱学习器个数 n_estimators 和决策树最大深度 max_depth 进行调节。其中最大弱学习器个数 n_estimators 太小容易欠拟合，n_estimators 太大，计算量会太大，并且 n_estimators 到一定的数量后，再增大 n_estimators 获得的模型提升会很小，所以一般选择一个适中的数值，在我们的模型中设定该参数的范围是(10,300)，步长为 10;决策树最大深度 max_depth 值越大，决策树越复杂，越容易过拟合，模型中我们设定该参数的范围是(3,14)。在拟合模型过程中，我们同样打乱了数据顺序，按照训练集:测试集=7:3 的比例随机划分训练集和测试集，在训练集中按照 10 折交叉进行交叉验证拟合模型。

模型最后输出可视化学习曲线、分类报告。

3.2.3AdaBoost 模型处理

AdaBoost 是一种迭代算法，其核心思想是针对同一个训练集训练不同的分类器(弱分类器)，然后把这些弱分类器集合起来，构成一个更强的最终分类器（强分类器）。AdaBoost 很好的利用了弱分类器进行级联，可以将不同的分类算法作为弱分类器，相对于 bagging 算法和随机森林算法，AdaBoost 充分考虑每个分类器的权重。AdaBoost 大多用于分类问题，精度较高。

参数选择上，我们主要调节最大弱学习器个数 n_estimators，将范围设定为(10,300)，步长为 10。在拟合模型过程中，我们打乱了数据顺序，按照训练集:测试集=7:3 的比例随机划分训练集和测试集，在训练集中按照 10 折交叉进行交叉验证拟合模型。

模型最后输出可视化学习曲线、分类报告。

3.3 模型处理结果汇总

3.3.1 所有音频分类情况

总体分类情况如下图所示。可以看出训练集得分较高，而验证集得分较低，同时测试集的得分也不高，出现了过拟合的现象。由于我们在数据预处理时归一化了数据，同时是随机

按比例抽取训练集，因此推测导致过拟合的原因有：

1. 数据量过少，由于时间有限，我们在每类标签下只筛选出了共计 300 个左右的视频，该数据量远小于一般机器学习所需数据量大小，因此容易导致过拟合现象。

2. 由于抖音热门 BGM 数量有限，存在不同视频所用 BGM 相同的情况，部分在某类下被划分为效果好的音频在另一类中被划分为效果较差的类别，影响了总体分类效果。

3. VGGISH 提取的特征矩阵包含信息不够全面。VGGISH 虽然是近年来提取音频特征效果最好的模型之一，但它是基于谷歌的 AudioSet 训练而来，适用于各类音频，并不是专门处理音乐的模型，同时本次研究数据中大量 BGM 的歌词是中文，而 VGGISH 对音频语义的特征也有部分提取，这部分特征是针对英文提取的，因此也影响到了提取特征的准确性，导致部分样本特征区别不明显，影响训练效果。

总体来看，三个模型的分类结果并不理想，测试集上表现最好的 SVM 模型的精确度也只有 0.67，虽然模型在测试集得分不高，但在正负类中具有出现精确度在 0.65 左右的模型，我们认为带货效果好坏的音频存在部分可以区别和量化的特征。同时在数据筛选过程中，我们认为从听觉上也可以听出部分带货效果较好和带货效果较差的音频以及各类商品下筛选出的音频之间的区别。考虑到个人喜好对判断的影响，我们后续进行了音频特征分析佐证我们的想法，详细参考第四章部分。

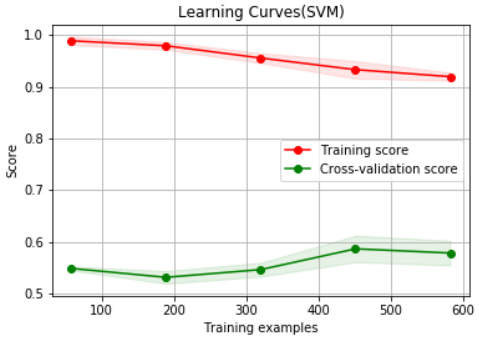


图 3-1-1 所有音频 SVM 学习曲线

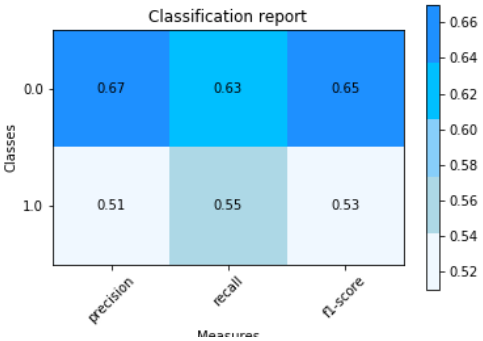


图 3-1-2 所有音频 SVM 分类报告

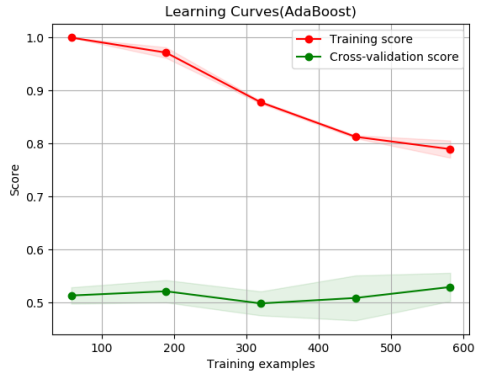


图 3-1-3 所有音频 AdaBoost 学习曲线

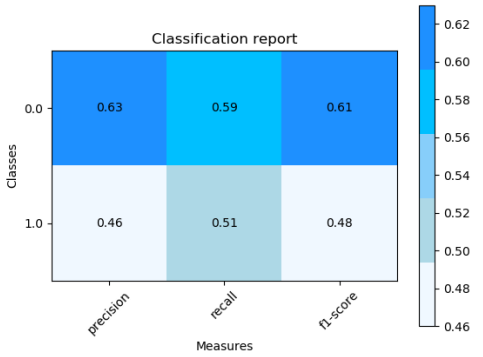


图 3-1-4 所有音频 AdaBoost 分类报告

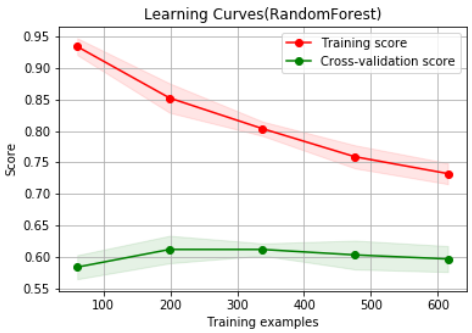


图 3-1-3 所有音频随机森林学习曲线

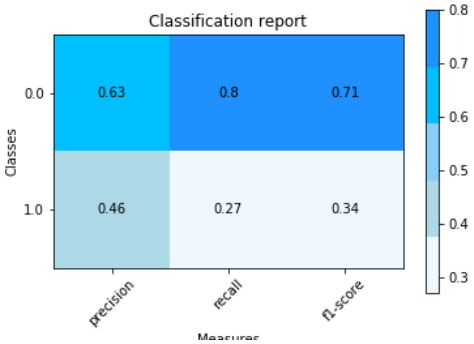


图 3-1-4 所有音频随机森林分类报告

3.3.2 食品音频分类情况

食品短视频的 BGM 效果分类图如下所示。模型平均准确率在 0.6 以上，效果最好的 AdaBoost 模型对正负样本的识别准确率达到了 0.75，且 f1 值也较高，在小数据量的情况下是一个较理想的分类效果。在人工筛选数据时，我们也发现带货效果较好的视频 BGM 直观感受上格调与“吃播”主题较搭，风格偏轻松明快，而带货效果较差的部分视频的 BGM 则较为嘈杂，影响观感。

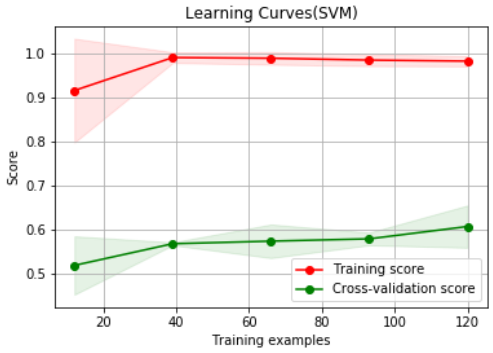


图 3-2-1 食品音频 SVM 学习曲线

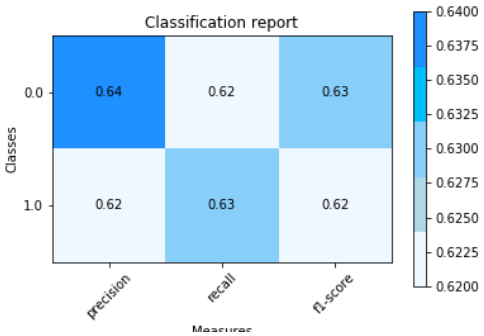


图 3-2-2 食品音频 SVM 分类报告

图 3-2-3 食品音频 AdaBoost 学习曲线

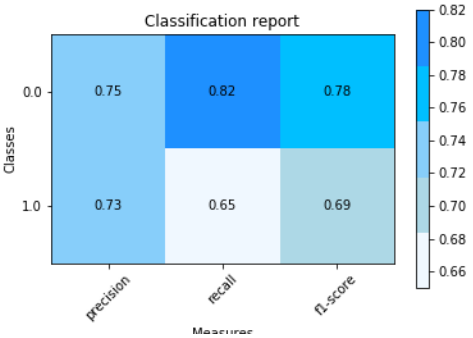


图 3-2-4 食品音频 AdaBoost 分类报告

图 3-2-5 食品音频随机森林学习曲线

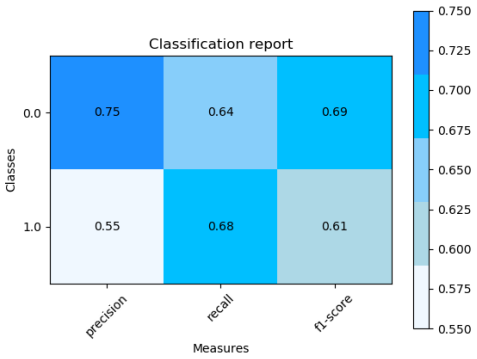


图 3-2-6 食品音频随机森林分类报告

3.3.3 手机数码音频分类情况

手机数码短视频的 BGM 效果分类图如下所示。可以看出，三个模型对正负样本的识别准确率平均值在 0.7 左右，负例的 f1 值均值在 0.75 以上，表现最好的模型是 SVM。三个模型对负例的识别效果都优于正例，可能原因是数据中正例:负例=108:164，负例样本数显著更高，所以模型学习到了更多负例的特征，更倾向于将样本识别为负例，这点在各模型分类报告的正例召回率数据中均能看出。在进一步优化分类效果时，可以尝试加大数据量，同时尽可能平衡正负样本的个数。

在人工筛选过程中，我们发现手机数码音频的正例样本中电音 BGM 较多，且打击元素较多，在样本音频分析中，音频过零率和 BPM 值的数据也佐证了我们这个观点。直观感受上，这两种元素与手机数码的带货短视频也更加匹配。

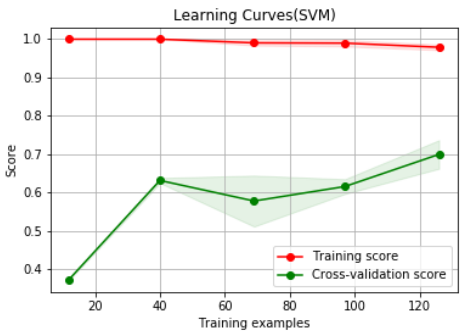


图 3-3-1 手机数码音频分类 SVM 学习曲线

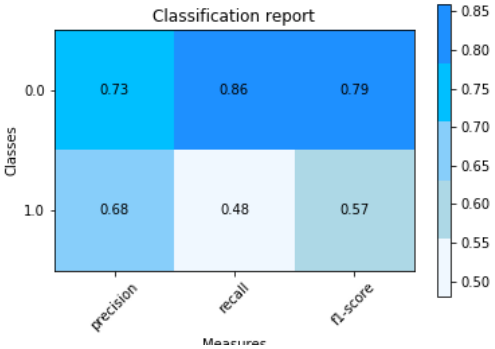


图 3-3-2 手机数码音频分类 SVM 分类报告

图 3-3-3 手机数码音频 AdaBoost 学习曲线

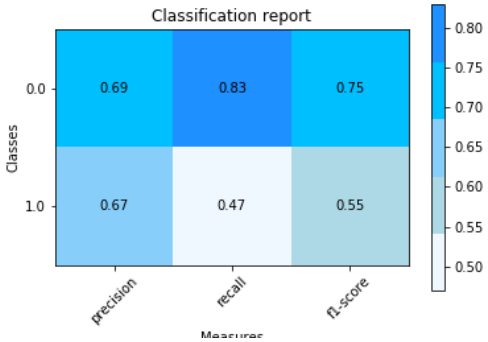


图 3-3-4 手机数码音频 AdaBoost 分类报告

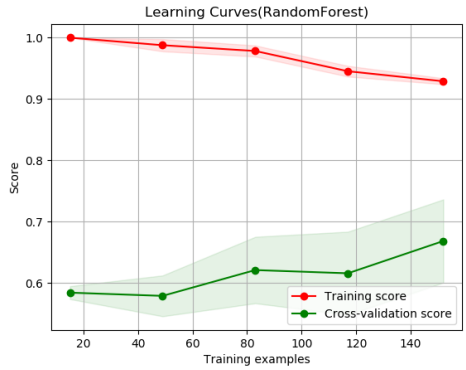


图 3-3-5 手机数码音频随机森林学习曲线

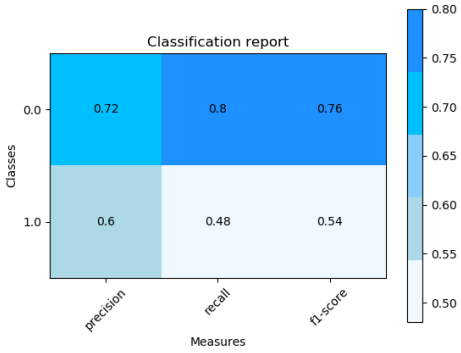
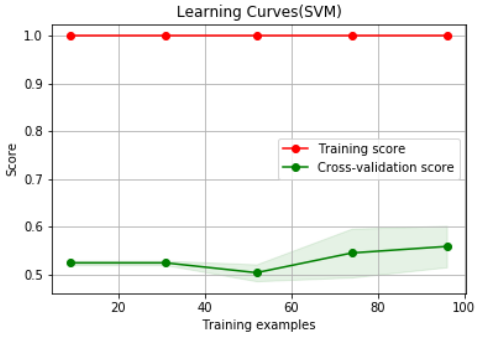


图 3-3-6 手机数码音频随机森林分类报告

3.3.4 美妆护肤音频分类情况

美妆护肤短视频的 BGM 效果分类图如下所示。可以看出，三个模型的分类效果都很差，基本无法辨别出正负样本的差别，在 PCA 初步分析中也可以看出降维损失信息之后两类样本点混杂在一起。

在人工筛选数据时，我们发现美妆护肤类视频正负样本的 BGM 大多为轻快的风格，且以女声为主。我们认为可能原因是目前美妆带货的博主以年轻女性为主，其它人群占比较小。而年轻女性对 BGM 风格搭配比较敏感，因此出现了正负样本中 BGM 风格均较为统一的现象，故模型难以区分。



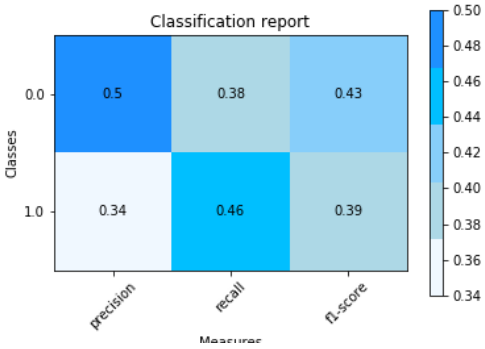


图 3-4-1 美妆护肤音频 SVM 学习曲线

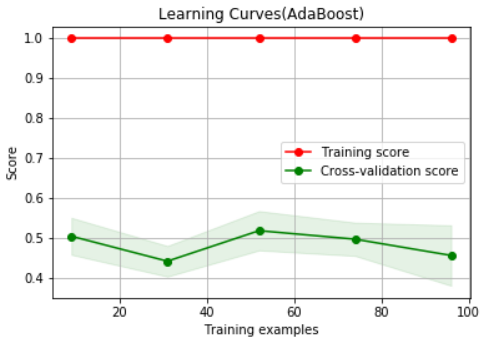


图 3-4-2 美妆护肤音频 SVM 分类报告

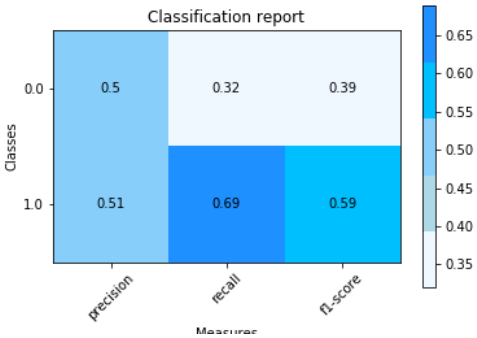


图 3-4-3 美妆护肤音频 AdaBoost 学习曲线

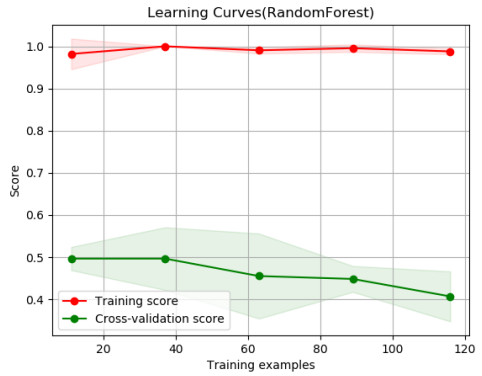


图 3-4-4 美妆护肤音频 AdaBoost 分类报告

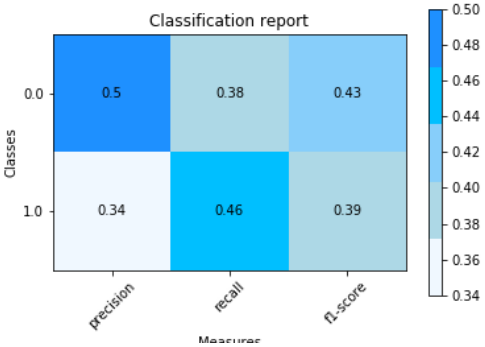


图 3-4-5 美妆护肤音频随机森林学习曲线

图 3-4-6 美妆护肤音频随机森林分类报告

3.3.5 男装音频分类情况

男装短视频的 BGM 效果分类图如下所示。SVM 模型和 AdaBoost 效果相近，其中 AdaBoost 对正例的识别效果较好，随机森林模型效果较差。

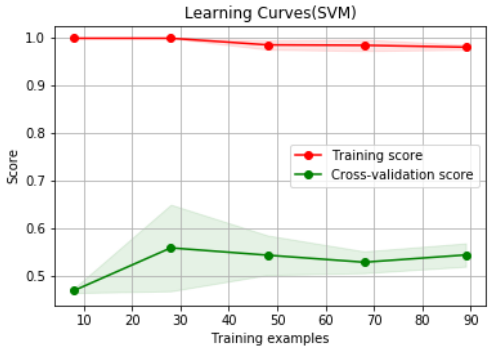


图 3-5-1 男装音频 SVM 学习曲线

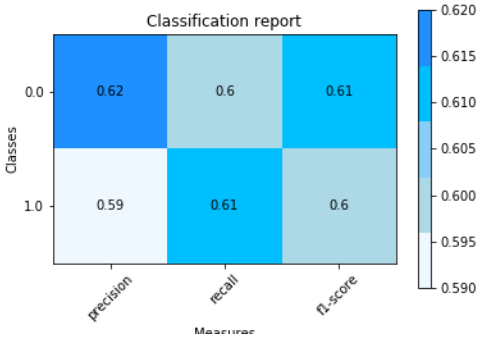


图 3-5-2 男装音频 SVM 分类报告

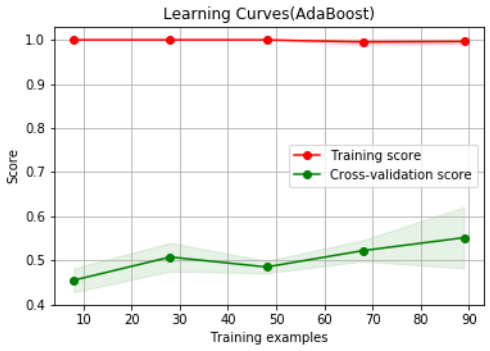


图 3-5-3 男装音频 AdaBoost 学习曲线

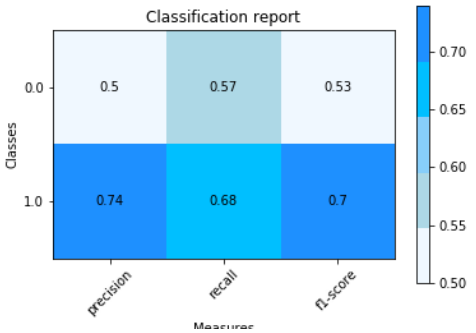


图 3-5-4 男装音频 AdaBoost 分类报告

图 3-5-5 男装音频随机森林学习曲线

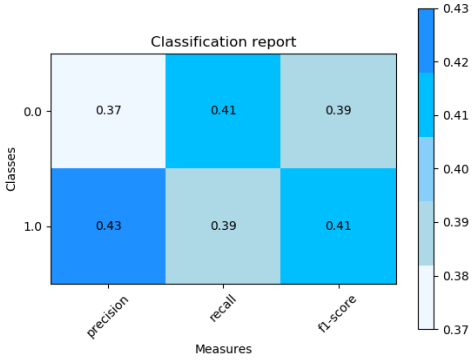


图 3-5-6 男装音频随机森林分类报告

3.3.6 女装音频分类情况

女装短视频的 BGM 效果分类图如下所示。对正例识别效果最好的是 SVM 模型，准确率达到了 0.75，但召回率很低，对负例识别率最高的是 AdaBoost 模型，准确率达到了 0.78，且召回率较高。可以发现三个模型正例的召回率均低于负例，猜测也是因为正负样本数为正：负=67:110 的原因。负例样本偏少，模型学习到负例的特征更多，更偏向于将测试集样本判断为负标签。

总体来说，女装带货带视频的 BGM 的分类效果较其他商品类的分类效果要更好，存在一定可分性。人工筛选过程中我们也发现部分销量高的中老年女装带货视频倾向于使用更有年代感，更符合中老年审美的背景音乐，而部分销量高的年轻女装带货视频则使用了更多流行歌曲；一些带货效果较差的视频在 BGM 与视频内容适配上则做的较差，使用的 BGM 虽然属于抖音热门，但风格与服装展示的视频并不搭。

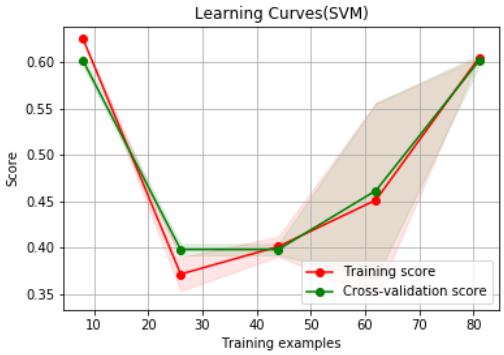


图 3-6-1 女装音频 SVM 学习曲线

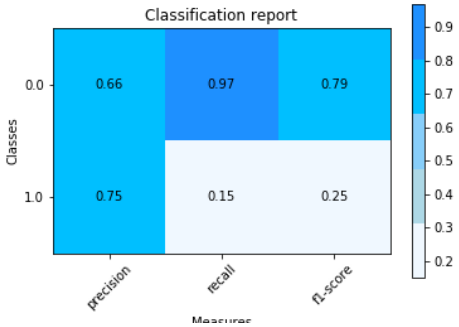


图 3-6-2 女装音频 SVM 分类报告

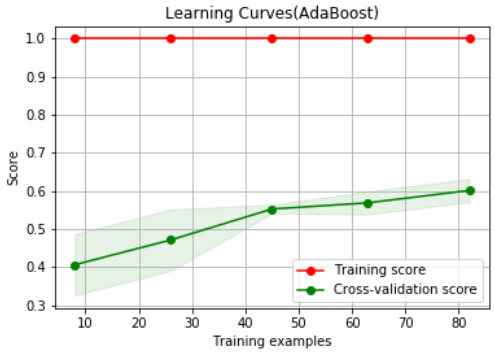


图 3-6-3 女装音频 AdaBoost 学习曲线

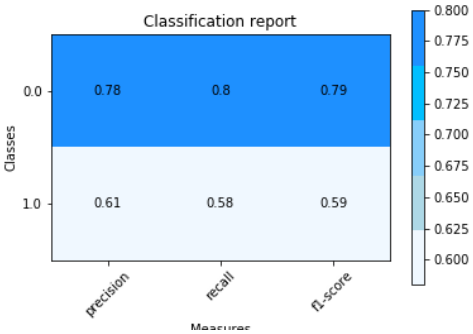


图 3-6-4 女装音频 AdaBoost 分类报告

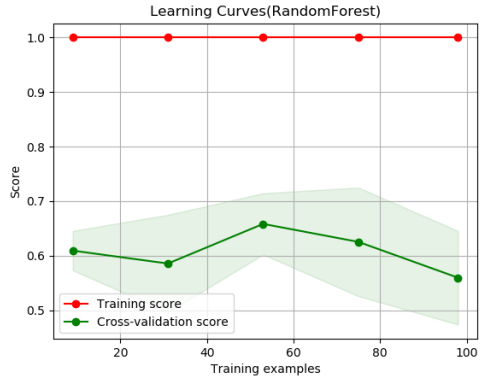


图 3-6-5 女装音频随机森林学习曲线

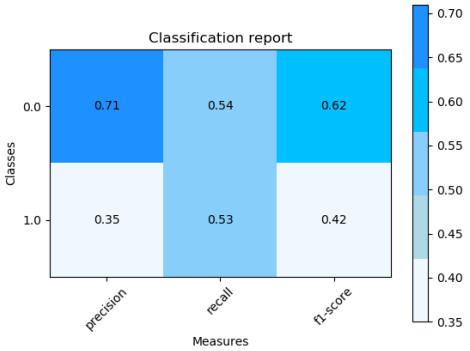


图 3-6-6 女装音频随机森林分类报告

3.3.7 日用百货音频分类情况

日用百货短视频的 BGM 效果分类图如下所示。正例:负例=126:192，模型同样出现了对负例的召回率和精确率更高的情况。

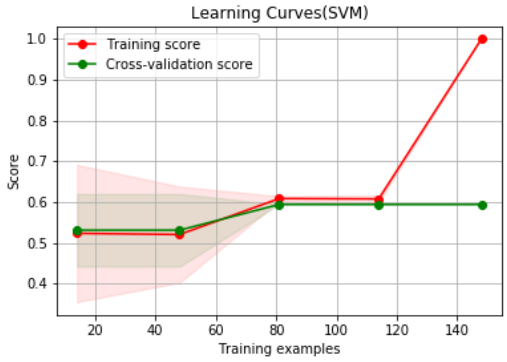


图 3-7-1 日用百货音频 SVM 学习曲线

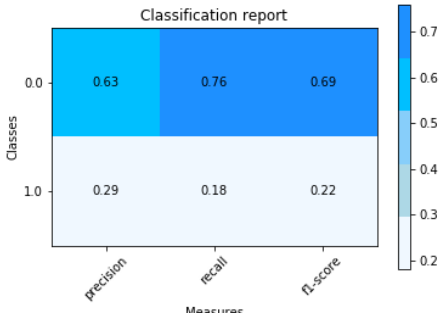


图 3-7-2 日用百货音频 SVM 分类报告

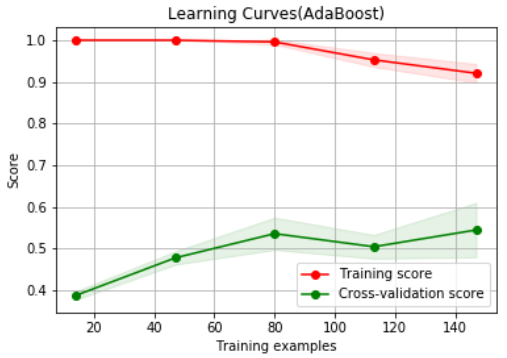


图 3-7-3 日用百货音频 AdaBoost 学习曲线

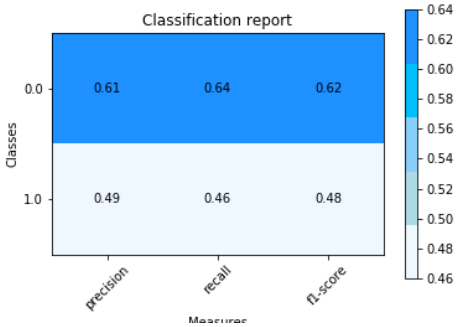


图 3-7-4 日用百货音频 AdaBoost 分类报告

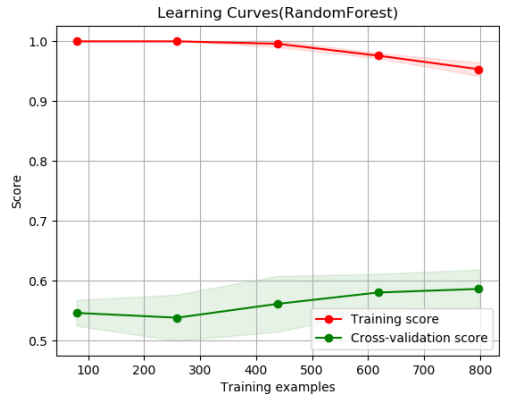


图 3-7-5 日用百货音频随机森林学习曲线

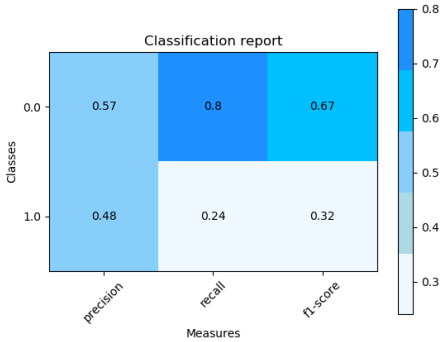


图 3-7-6 日用百货音频随机森林分类报告

第四章 音频分析

音乐是反映人类现实生活情感的一种艺术，其中背景音乐（Background Music，简称 BGM）通常可以起到调节气氛、增强情感表达等作用。艺术往往都是抽象的，音乐也不例外，故而要像数学公式那样严格定义什么是音乐几乎是不可能的；也因此，人们常通过音乐

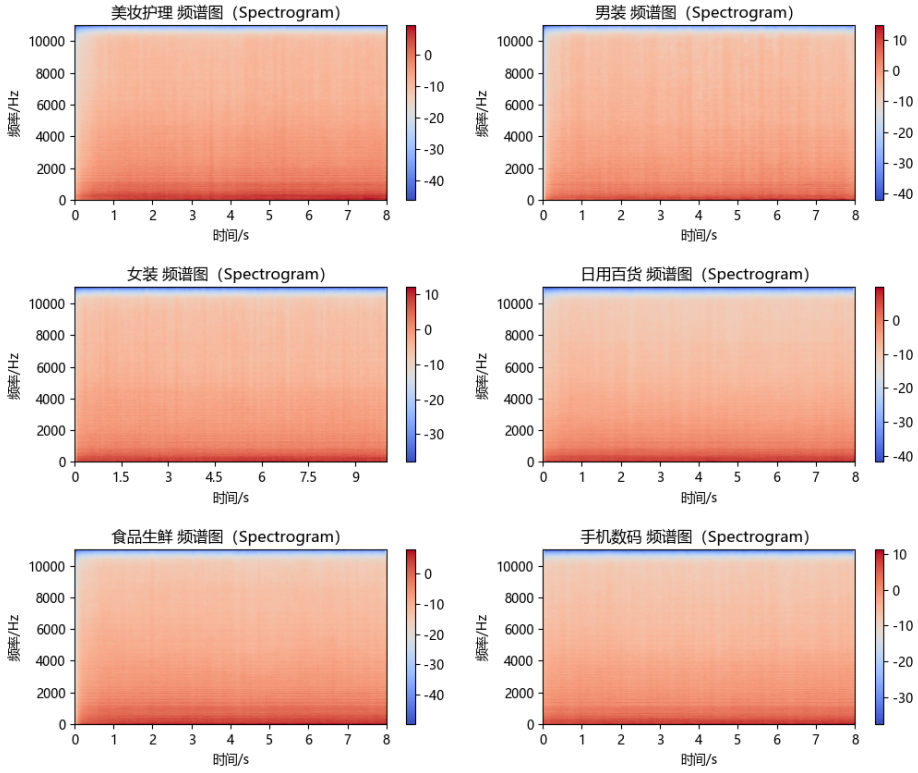
的一些特征进行描述。

音乐的要素包括旋律、节奏、力度、音色等等，在不同场合有着不同的呈现。在此次关于抖音带货视频背景音乐的研究中，我们按统一标准获取了不同带货类型的同等时长（只有女装类统一截取了 10 秒的背景音乐，其余均为 8 秒）的背景音乐，选取音乐的频率、响度、节拍、旋律等要素进行研究分析，对背景音乐和带货视频类型之间的关联有了更多的思考与见解。

以下数据源自人工筛选和截取的抖音不同带货类型视频的背景音乐，基于 python 环境下的 librosa、matplotlib 和 madmom 等工具提取和解析，以取平均值或总和的方式概括每一类的特征并进行可视化操作。

4.1 频谱分析

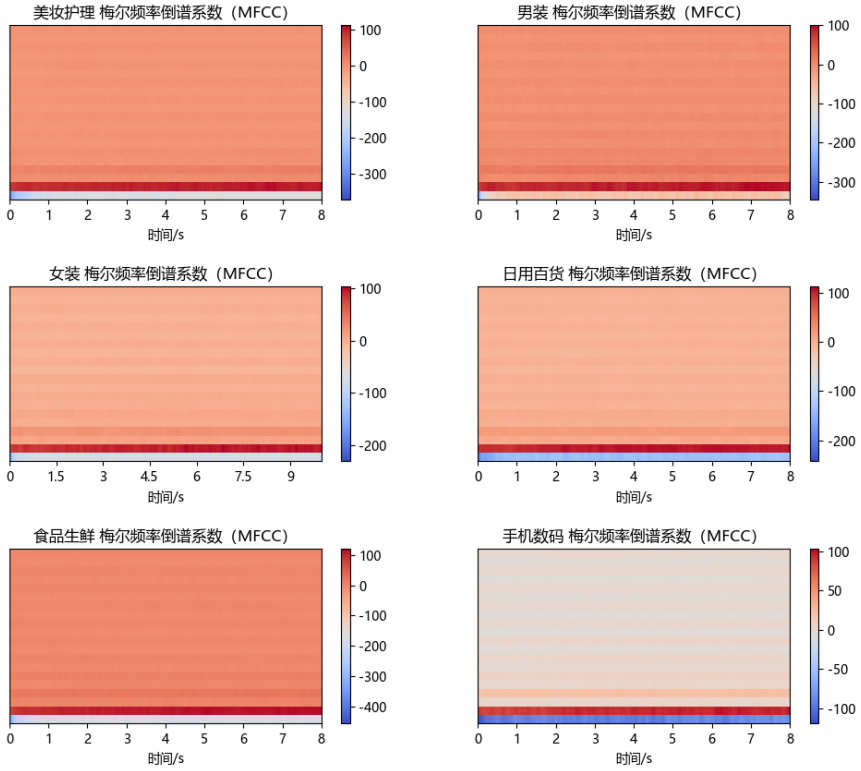
频谱是频率谱密度的简称，是描述音乐频率和能量分布的特征。下图为每个类别音频通过短时傅里叶变换后得到特征值取平均后制作的频谱图，能量转化为响度加以描述。图中横轴代表时间，纵轴代表频率，颜色代表响度（单位为分贝）。



从整体来看，每一类的频谱图基本一致，差别甚微。考虑到视频来源一致，并对音频进行过统一提取和处理，这一点容易理解。仔细观察可以发现，美妆护理与食品生鲜类音频在 0~1000Hz 频率范围内的能量相对更高更集中，且 1000Hz 附近的响度更接近于 0dB，这正是人类恰好能够听到的声音，故而相对来说，这两类的音频对于人耳更加明晰。

4.2 梅尔频率倒谱系数（MFCC）

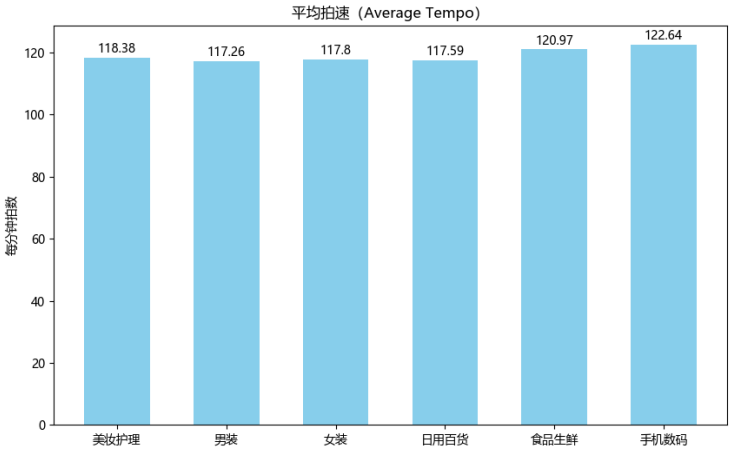
在频谱描述方面，梅尔频率倒谱系数是音频信号特征中最重要的之一，可以描述音频频率分布的包络，是语音识别等音频处理操作常用的特征，但不容易被直观理解，下图展示了梅尔频率倒谱系数提取后的可视化展示。



整体上来看美妆护理、男装和食品生鲜类特征相似，女装和日用百货类相似，手机数码类相对独特。图中参数的物理含义难以直观理解，此处暂不分析。

4.3 拍速分析

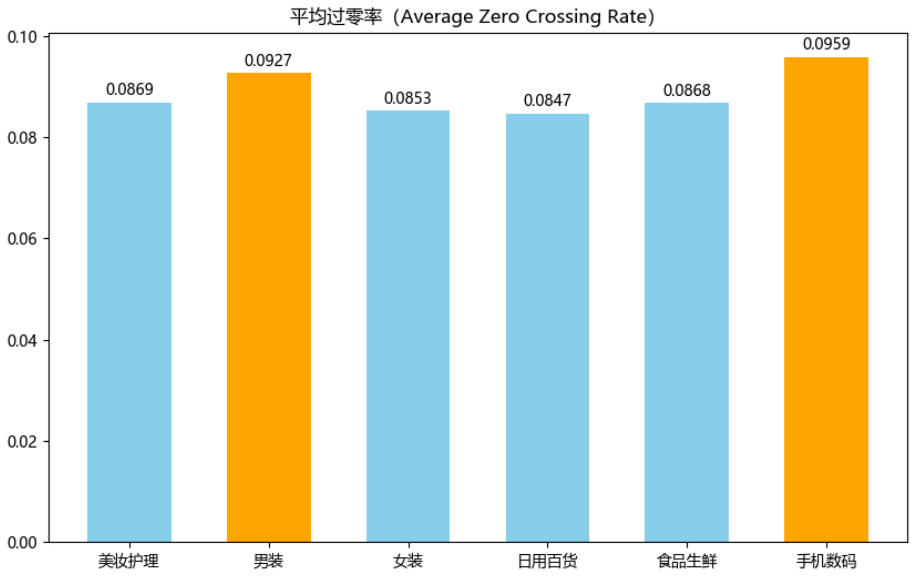
拍速是音乐每分钟演奏四分音符的个数，是描述音乐演奏速度的一个特征。下图体现了不同类型带货视频背景音乐在拍速上的差异。



可以发现，在抛开细微差距后，每一种音频的拍速均在 120 拍/分钟左右，属于略快的节奏；推测原因在于这样的节奏更契合短视频短小精炼的特点，有利于在更短的时间内吸引观众的眼球。对比发现，手机数码类带货视频的音乐节奏最快，男装的拍速最慢，这与带货种类或多或少有所关联。手机等电子产品本身更有科技感，配以快节奏的音乐可以更加凸显数码设备在性能等方面的优越表现；而男装女装等货品的价值更多在于内涵和耐用性，相对舒缓的音乐更契合这类货品的特点，赋之以时间的沉淀感。

4.4 过零率分析

过零率是指信号的符号变化的比率，是对敲击音乐、金属声音等分类以来的主要特征之一。一般过零率越高，打击感也越强。下图为各种类音频的平均过零率分布图。

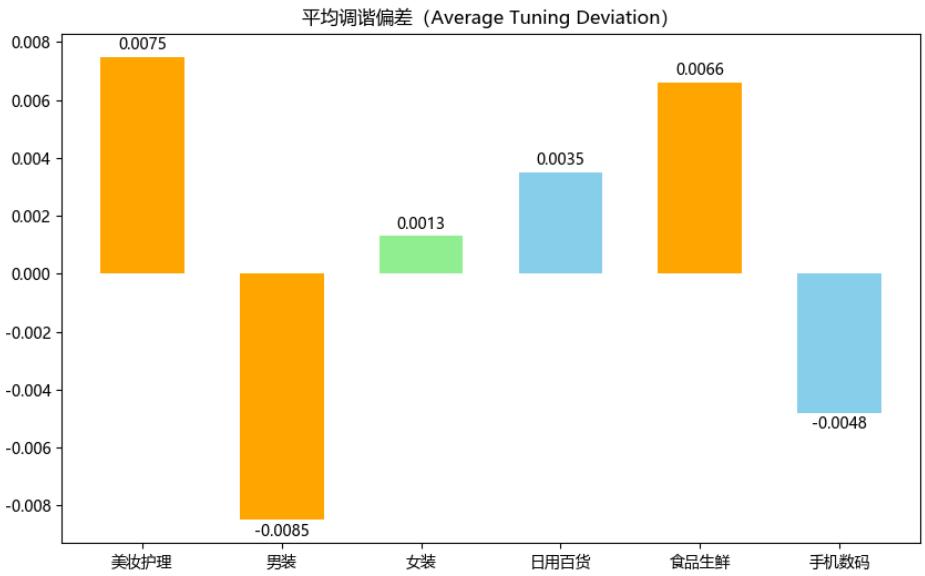


巧合的是，过零率的分布图似乎和前面拍速的分布有些相似，突出的两个类别仍为男装和手机数码，但这里两者均有相对较高的过零率。高过零率意味着更多的打击元素、金属音乐元素等，如果搭配快节奏的音乐，则更能给人在力度和音效上的冲击；如果搭配慢节奏的音乐，则会让音乐多一些质感和经典的味道。如此看来，手机数码和男装则是这两种方向搭

配的表现。

4.5 旋律分析

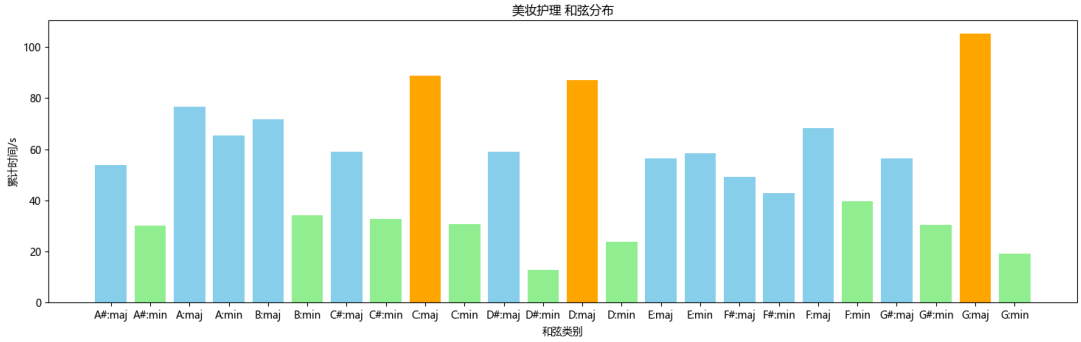
调谐是指调节频率从而与其他波产生谐振的过程，调谐偏差自然是指与谐振的偏差。在音乐中也有谐振，不仅在于声音上的共振，更在于思想上的和谐。音乐谐振给人更和谐的听觉效果，具有独特的音乐魅力。在带货视频所使用的背景音乐中，有些是纯人声，有些是人声与乐器合奏，有单乐器的纯音乐，也有多乐器的协奏，协调偏差越小，音乐更接近谐振。



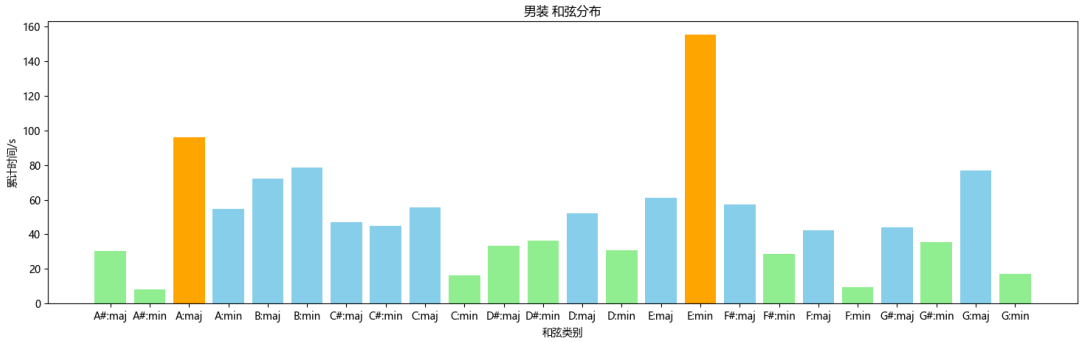
上图展示了不同带货类型视频所使用的背景音乐的平均调谐偏差，其中女装类最小，男装类最大。同样是卖服装，面对不同性别的客户时音乐选取的特点却有如此大的差别。根据此处数据，再结合视频分类时的感受和生活经验，我们认为女装类带货视频往往风格柔和，且大多面向中老年女性，这类客户也符合端庄温婉的特点；而男装类带货视频风格不一，既有面向青少年的，也有面向中老年的，既有男博主穿衣展示，也有女博主穿男装推荐货品，加上前文提到的金属质感明显，其谐振效果相对难以达到。

4.6 和弦分析

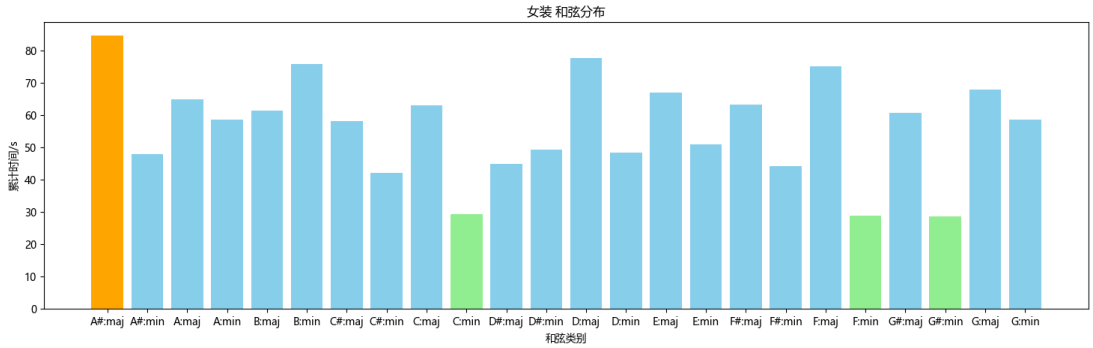
和弦是指有一定音程关系的一组声音，这一组声音同时弹奏能够表达具有一定特点的旋律，此处的特点更偏向于情感。我们对每一组声音采用的和弦进行了统计，并计算出各类和弦的总时长。考虑到每种音乐的样本数不同，此处不进行类别之间的比较，而是观察每一类在和弦使用上都有怎样的偏向。当然，由于和弦种类过多，并涉及专业的乐理知识，分析时会采取一些近似和简单处理。



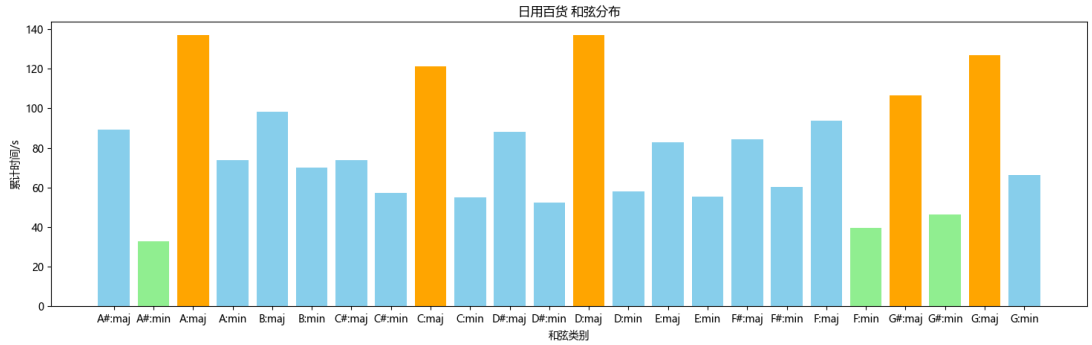
美妆类视频主要使用 C、D 和 G 和弦，且均为大和弦，旋律明亮，稳重柔和。



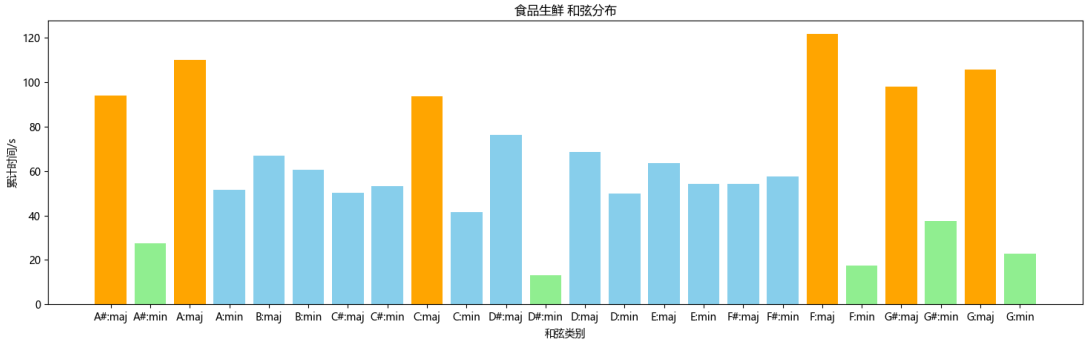
男装类视频中 E 小和弦出现比重最高，此和弦往往给人带来暗淡忧伤的情感。



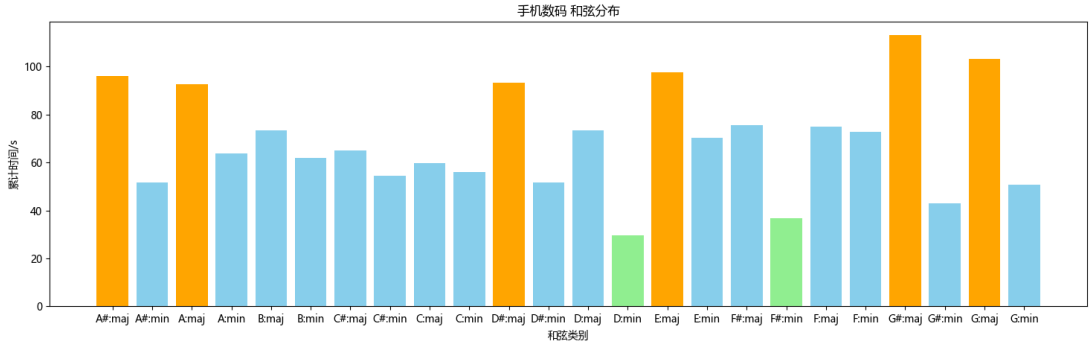
女装类视频各类和弦均有出现，大小和弦比重参差不齐，特点多样，情感丰富。



日用百货类视频以每种和弦的大和弦为主，色彩相对明快，种类也比较多样。与后两类相比大 D 和弦相对突出。



食品生鲜类视频与日用百货类和弦分布接近，属于多样化的分布，主色调明亮。与前后两类相比大 F 和弦相对突出。



手机数码类视频与前两类基本一致，情感色彩中性。与前两类相比大 E 和弦相对突出。

以上只是对和弦部分的统计和概括，事实上音乐的情感表达远不止依赖各类和弦的配合，还有演奏手法、调式选用等众多影响因素，因此这里也无法展开更多分析。

4.7 音频分析总结

综合以上所述，不同带货类型的短视频背景音乐在一些音乐特征上存在差异，这些差异与带货类型存在一定的关联，在一定程度上可人为或通过机器分类音频、推荐类型适合的音频。本文关于音频特征的分析旨在通过数据描述人类通常可感知的音频特征，对不同带货类型的抖音短视频进行简单的区分和理解；然音乐变化多样，特征不是定数，且本研究分析的数据有限、方法简陋，分析结论只留作短期内特定研究的参考，未来仍有待完善。

第五章 结论与反思

5.1 研究结论

经过机器学习对 BGM 进行分类以及音频特征的分析，我们认为带货效果好和带货效果差的音频存在可分的特征差异，选择与商品搭配的 BGM 更有利于提高短视频的带货效率。在研究的各商品类中，手机数码和女装类带货视频的 BGM 分类效果最好，原因应该是这两类商品与特定音乐类型适配性更强，样本特征更明显，分类效果更好。受样本比例影响，模

型对其余四类商品中负例的识别准确度更高，正例识别能力较差，但从准确率和召回率来看也存在可分性。

根据生活经验我们认为合适的 BGM 可以提高短视频的带货效率，并且可以通过一些特征来 BGM 判断是否“合适”，本次研究一定程度上证明了我们的观点。由于本次研究音频提取的特征是 VGGISH 提取的数字矩阵，而这些数字没有实际意义，是对音频特征的抽象刻度，对如何创作合适的带货 BGM 没有指导意义。但是在加大数据量，优化标签标记以及优化模型之后，模型分类的准确度可以继续提升，制作带货短视频的达人也可以通过我们的模型判断自己选择的 BGM 是否是“合适”的。

5.2 项目反思

本次研究是我们小组成员第一次进行数据分析的尝试，机器学习和音频分析都是在研究进程中逐步学习的，因此在学习上花费了不少时间。研究效果由老师来评判，不过我们认为本研究还可以从以下几个方面进行优化:

1.加大数据量。由于时间因素制约，我们这次只爬取了总计 8400 个视频，但是在筛选过后只剩下了 1426 个视频，分摊到各个商品类下，数据量就只是百级，这对模型拟合影响较大。在时间充足的情况下，应该可以做到将每个商品类下的音频数据提升到千级。

2.提升标签准确度。这次研究中我们以蝉妈妈网站提供的预估销售量作为分类标准，由于其他数据网站需要付费才能获取数据，我们无法横向对比该指标的准确程度。同时我们发现发布视频的达人的粉丝数范围跨度较大，虽然粉丝多的博主更有可能注意视频质量，挑选合适 BGM，正样本受影响较少，但是部分预估销量很低的视频的博主的粉丝很少，也许视频本身质量很高，而我们将这部分样本打上了负标签，在打标签的过程中还可以在这方面进行优化。

3.音频特征提取模型的优化。VGGISH 是近年来音频特征提取表现最好的模型之一，但它不是针对音乐的特征提取模型，一段时间后也许会出现更优秀的音乐特征提取模型。

4.机器学习模型还可以再优化。由于这次研究开始我们才接触机器学习，因此模型选择了 sklearn 库封装实现的三个模型，而神经网络在分类领哟也有优秀表现，未来可以尝试使用神经网络来优化分类效果。

第六章 附录

6.1 神经网络验证

6.1.1RNN 循环神经网络验证

研究最后我们使用 Keras，搭建了简单的 RNN 循环神经网络对所有音频进行分类，分类准确率并没有大的提升。

```
Epoch 1/10
125/125 [==============================] - 0s 2ms/step - loss: 0.6737 - accuracy: 0.5660
Epoch 2/10
125/125 [==============================] - 0s 2ms/step - loss: 0.6421 - accuracy: 0.5820
Epoch 3/10
125/125 [==============================] - 0s 2ms/step - loss: 0.6319 - accuracy: 0.5920
Epoch 4/10
125/125 [==============================] - 0s 2ms/step - loss: 0.6261 - accuracy: 0.6050
Epoch 5/10
125/125 [==============================] - 0s 2ms/step - loss: 0.6281 - accuracy: 0.6170
Epoch 6/10
125/125 [==============================] - 0s 2ms/step - loss: 0.6170 - accuracy: 0.6280
Epoch 7/10
125/125 [==============================] - 0s 2ms/step - loss: 0.6076 - accuracy: 0.6400
Epoch 8/10
125/125 [==============================] - 0s 2ms/step - loss: 0.6110 - accuracy: 0.6260
Epoch 9/10
125/125 [==============================] - 0s 2ms/step - loss: 0.6107 - accuracy: 0.6170
Epoch 10/10
125/125 [==============================] - 0s 2ms/step - loss: 0.5932 - accuracy: 0.6390
10/10 [==============================] - 0s 1ms/step - loss: 0.6171 - accuracy: 0.6375
loss:0.617144763469696
accuracy:0.637499988079071
```

6.1.2BP 神经网络验证

BP 神经网络验证的结果报告如下，准确率也没有明显提升

	precision	recall	f1-score	support
0	0.72	0.62	0.66	182
1	0.58	0.69	0.63	138
accuracy			0.65	320
macro avg	0.65	0.65	0.65	320
weighted avg	0.66	0.65	0.65	320

6.2 参考文献和项目地址

参考文献：
[1] 王欢.基于 VGGish 网络对音乐情感的分析[D].天津:天津商业大学,2019:15.

项目地址：https://github.com/Plutooooooo/czyBGM

6.3 小组成员信息

曹英瑞：学号 181250006 邮箱 181250006@smail.nju.edu.cn 完成题目数:200
郭礼华：学号 181250038 邮箱 181250038@smail.nju.edu.cn 完成题目数:170
廖兰宇：学号 181250082 邮箱 181250082@smail.nju.edu.cn 完成题目数:199