 DOI: https://doi.org/10.12345/cjygl.v6i2.10697

# Data Mining and Analysis of the Marketing of Lipstick Brand “MAC”

**Suping Li**

Shaoxing University of Arts and Sciences, Shaoxing, Zhejiang, 312045, China

**Abstract**

Lipstick is an important part of makeup.In recent years, the cosmetics industry has developed rapidly, and more and more cosmetics brands appear in people’s vision, and MAC has more and more competitors.In order to make merchants produce lipstick more compatible with consumers and increase their sales.Among these many factors that affect lipstick sales, By using a random forest model, the naive Bayes Algorithm analyzing which factors are crucial to lipstick sales, which predict merchant sales, And introduce the algorithm process in detail.

**Keywords**

MAC lipstick; a random forest mode; a naive Bayesian algorithm

# 关于口红品牌“MAC”营销的数据挖掘与分析

李素萍

绍兴文理学院，中国·浙江 绍兴 312045

**摘　要**

口红是彩妆的一项重要内容。近年来，化妆品行业发展迅速，越来越多的化妆品品牌出现在人们的视野里，MAC的竞争对手也越来越多。为了让商家生产出更符合消费者的口红，提高其销售量，我们在这诸多影响口红销量的影响因素中，通过用随机森林模型，朴素贝叶斯算法分析哪些因素对口红销量至关重要，根据这些因素预测商家的销售量，并把算法过程详细介绍。

**关键词**

Mac口红；随机森林模型；朴素贝叶斯算法

## 1 引言

目前化妆品行业发展迅速。根据搜集的数据发现，全球化妆行业市场规模持续增加，市场主要分布于亚太地区。而我国化妆行业目前整体规模不大，但是发展潜力巨大，但国内品牌竞争激烈，国际品牌占据国内市场大部分份额，如欧莱雅、雅诗兰黛、香奈儿等品牌。而国内品牌无论是知名度还是产品质量都不足以与国外品牌相抗衡。口红是彩妆的一项重要内容，随着技术的不断发展，MAC 口红近年来越来越被广大女性消费者所喜爱。MAC 活跃于创造时尚前沿趋势，与时尚、艺术和潮流文化的领军人才们合作。近年来，国际上化妆品行业发展迅速，越来越多的化妆品品牌出现在人们的视野里，MAC 的竞争对手也越来越多。因此 MAC 口红的销售数据对于研究数据挖掘有一定的使用价值。

论文以主要以 MAC 口红为列，用随机森林模型，朴素贝叶斯算法等分析其之前的销售情况，一方面可以为 MAC 口红如何创新自己的营销方式提供数据基础，另一方面，详细的介绍随机森林模型，朴素贝叶斯算法如何使用，把数据分析与营销相结合。发现、探索数据分析中的问题，挖掘企业的竞争力，为提高 MAC 品牌的经济利益，实现利润最大化提供数据支撑。

【作者简介】李素萍（1999–），女，中国云南昆明人，本科，从事市场营销研究。

## 2 研究背景与研究问题

### 2.1 MAC 口红概况

MAC 品牌是 1984 年在加拿大成立的，现在成为雅诗兰黛旗下的一个品牌，MAC 以其优越的质量和绚丽多姿的色彩，在彩妆届奠定了良好的基础。

MAC 口红在 1994 年被雅诗兰黛集团收购，成为雅诗兰黛旗下的子品牌，这也促进了 MAC 的快速发展，利用雅诗兰黛的销售渠道，在全球 90 多个国家销售 50 多个系列的产品。其中 MAC 口红近年来采取的营销策略，抓住大众的消费心理，利用优惠的价格，抓住网络营销的便利渠道，在经济快速发展的今天促进自身品牌的快速发展 [1]。

## 2.2 MAC 口红营销现状

据 2019 年彩妆品牌线上交易规模的数据，MAC 以 91.92% 的同比增长速度名列第一，紧跟着的是阿玛尼，同比增长为 73.09%；而后面的迪奥、圣罗兰的同比增长只有 20% 到 40% 之间 [2]。之所以其增长速度最快与以下两点有密不可分的关系。

### 2.2.1 与网红 KOL 密切合作

MAC 与韩国知名化妆师 PONY 朴惠敏联合推出了一款新系列彩妆产品。该系列彩妆产品运用了 PONY 喜爱的致爱水晶和塔罗牌插图元素，包括了眼影、唇彩、珠光 Prep + Prime Fix+ 喷雾、彩妆刷、假睫毛五个品类，其中八色眼影是主打产品。但是即便网红能依靠自身的粉丝基础在短时间内获取庞大的销量，但由于品质问题也无法长久卖下去。所以与 PONY 合作，只能是锦上添花。产品品质过硬，就能以良好的口碑长久生存下去。

### 2.2.2 与年轻人保持密切关系

MAC 在年轻化营销上，远超了很多国际大牌。从今年 1 月份推出王者荣耀口红，到 8 月份参展 China Joy，MAC 利用二次元、游戏、DIY 等 IP 与元素俘获了无数年轻消费者的心，在内容上就能做到对年轻消费者的心一击必中。

## 2.3 MAC 口红营销问题

### 2.3.1 缺乏信任感

由于现在假货泛滥，人们对对 MAC 口红的真伪，色号和上嘴效果等都持怀疑态度，缺乏信任感 [3]。

### 2.3.2 技术与安全性问题

如果是在线上进行交易，不能看到口红的真实面貌，因此人们对口红的质量和它的技术与安全性问题感到担忧。

### 2.3.3 品牌运作能力和管理能力相对于其他公司较弱

MAC 集团的管理能力还需要多向其他大型跨国公司学习，不断提高自身品牌的管理运作能力，利用技术提高自己的市场份额 [4]。

# 3 数据挖掘分析与模型构建

## 3.1 朴素贝叶斯分类

### 3.1.1 朴素贝叶斯算法原理

根据此方法，对一个未知类别的样本 x，可以先分别计算出 x 属于每一个类别 $C$ 的概率 $P(X|C_i)P(C_i)$，然后选择其中概率最大的类别作为其类别。朴素贝叶斯算法成立的前提是各属性之间互相独立。当数据集满足这种独立性假设时，分类的准确度较高，否则可能较低。另外，该算法没有分类规则输出。

### 3.1.2 计算结果

通过函数 NaiveBayes（sales_num~data6）生成判别规则，并对各类别下变量密度进行可视化，绘制了描述分和价格分的密度图，如图 1 所示。

从图 2 中可以看出，在字段描述分密度图中，销售总量类型 B 和 C 主要集中在 0.82 左右，而类型 A 主要集中在 0.73 左右；在字段价格分密度图中，类型 A 的数量比其他 2 个类型要小。

## 3.2 集成学习一

### 3.2.1 AdaBoost 算法原理

AdaBoost 算法的工作机制是首先从训练集用初始权重训练出一个弱学习器 1，根据弱学习的学习误差率表现来更新训练样本的权重，使得之前弱学习器 1 学习误差率高的训练样本点的权重变高，使得这些误差率高的点在后面的弱学习器 2 中得到更多的重视。然后基于调整权重后的训练集来训练弱学习器 2。如此重复进行，直到弱学习器数达到事先指定的数目 $T$，最终将这 $T$ 个弱学习器通过集合策略进行整合，得到最终的强学习器 [5]。

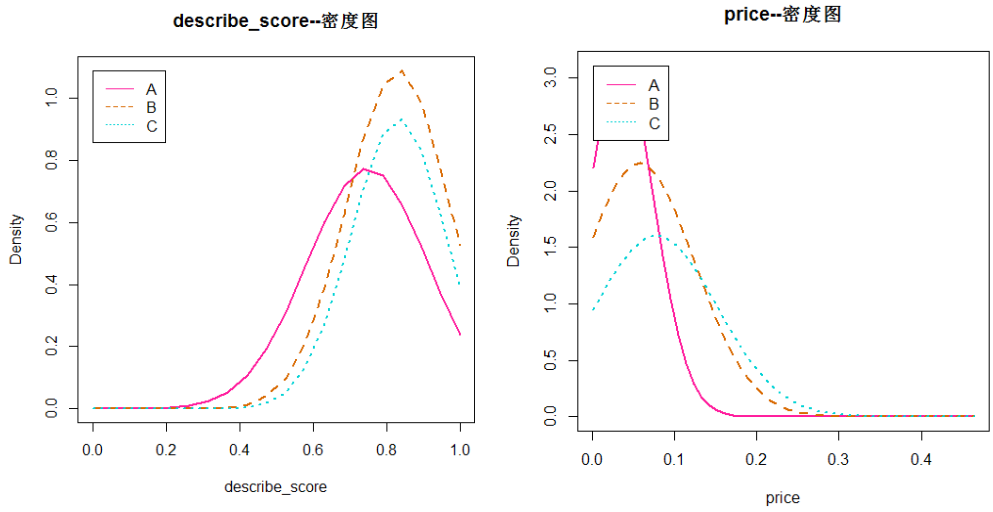


图 1 描述分和价格分的密度图

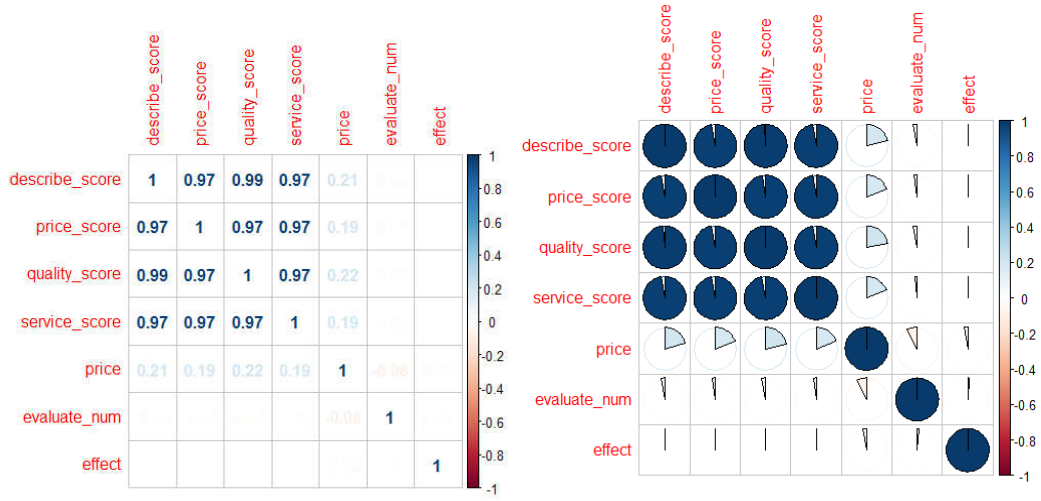


图 2 各变量相关性图

### 3.2.2 计算结果

调用函数 boosting(sales_num.~,data4,boos = TRUE, mfinal = 500) 构建模型，参数 mfinal 代表算法的迭代次数，即基分类器的个数，本实验设置其大小为 500，参数 boos 为采用各观测样本的相应权值来抽取 boostrap 样本。

通过调用函数 pred_boost=predict(boost2,data5) 进行预测，pred_boost$confusion 函数生成混淆矩阵，如表 1 所示。模型的预测错误率为 17.88%，进一步使用随机森林算法进行建模分析。

表 1 AdaBoost 预测混淆矩阵

| | A | B | C |
|---|---|---|---|
| A | 209 | 18 | 5 |
| B | 16 | 74 | 15 |
| C | 3 | 12 | 34 |

## 3.3 集成学习二

### 3.3.1 随机森林算法原理

随机森林通过自助法（bootstrap）重采样技术，从原始训练样本集 N 中有放回地重复随机抽取 k 个样本生成新的训练样本集合，然后根据自助样本集生成 k 个分类树组成随机森林，新数据的分类结果按分类树投票多少形成的分数而定 [6]。

在建立每一棵决策树的过程中，有两点需要注意采样与完全分裂。对于行采样，采用有放回的方式，也就是在采样得到的样本集合中，可能有重复的样本。这样使得在训练的时候，每一棵树的输入样本都不是全部的样本，使得相对不容易出现过拟合。之后就是对采样之后的数据使用完全分裂的方式建立出决策树，这样决策树的某一个叶子节点要么是无法继续分裂的，要么里面的所有样本的都是指向的同一个分类。

### 3.3.2 计算结果

调用函数 randomForest(sales_num ~ .-name,data=data3, ntree=500,importance=TRUE,proximity=TRUE,subset=Train_data_num) 构建随机森林的模型，其中决策树的数量为 500 棵。其 OOB 估计错误率为 29.12%，调用函数 importance(data.rf) 可以查看随机森林模型中变量的重要值，随机森林模型中两种测算方式下的自变量重要程度对比如图 3 所示。

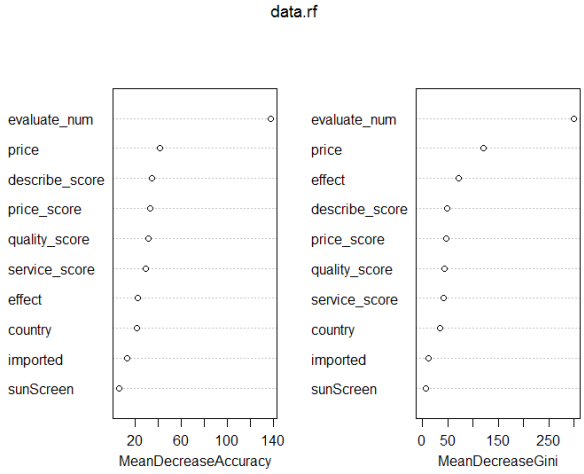


图 3 随机森林模型中两种测算方式下自变量重要对比

### 3.3.3 模型优化

影响随机森林模型的两个主要为：①决策树节点分支所选的变量个数；②随机森林模型中决策树的数量。使用 randomForest() 其两个参数都设为默认值，但在实际使用过程中，该默认参数不一定是最优的参数值，因此本实验在构建模型时进一步确定优化参数。其中对于决策树节点分支所选的变量个数的确定，采用逐一增加变量的方法进行建模，最后寻找到最优的参数 [7]。

①节点最优变量个数。

通过参数 mtry 来改变节点变量的个数，即从 1 逐渐增加到 10( 其中 10 为自变量的个数 )，如表 2 所示。从输出结果可以观察到，当决策树节点所选变量时为 5 时，模型的误判率均值是最低的，因此将参数 mtry 的值设置为 5。

**表 2 节点变量个数对应误判率**

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| 34.62% | 29.06% | 29.59% | 29.09% | 28.41% |
| 6 | 7 | 8 | 9 | 10 |
| 29.68% | 28.75% | 28.98% | 29.05% | 28.54% |

②决策树数量。

在确定了模型中决策树的节点最优变量个数之后，还需要进一步确定模型中决策树的数量。在确定该参数时，将用到模型的可视化分析。通过调用函数 plot(model,col=1:6)，绘制模型误差与决策树数量的关系图 4，可以观察到当决策树的数量大于 400 之后，模型误差趋于稳定，因此将模型中的决策树数量大致确定为 400 左右，以此来达到最优模型 [8]。

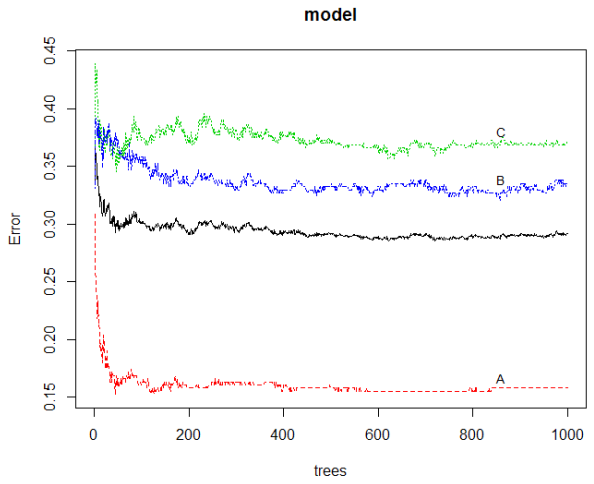


**图 4 模型误差与决策树数量关系**

在经过上述分析以后，本实验确定最优模型为决策树节点处变量个数为 5，模型中决策树数量为 400 的模型。再次进行建模，模型基于 OOB 数据的总体误判率为 29.03%；模型中决策树的节点数最少为 200 个，而节点数最多为 270 个。

### 3.4 模型评价

将构建的模型，在测试集中运行，得到如表 3 所示的错误率。可以看到，朴素贝叶斯分类错误率相差不大，已经超过 50%，AdaBoost 算法的错误率为 17.88%，而随机森林的错误率只有 6.99%，在这三种算法中，随机森林的算法最好。

**表 3 三种算法错误率**

| 算法 | 朴素贝叶斯分类 | AdaBoost | 随机森林 |
|---|---|---|---|
| 错误率 | 53.11% | 17.88% | 6.99% |

由分析可知：

①该数据集不符合朴素贝叶斯判别分析执行的前提条件——各变量条件独立，即参与建立判别分析规则的这些变量是有着显著的相关性的，很大程度上影响了预测结果的好坏 [9]；

②非数值自变量在一定程度上不能归一化处理，在预测时有一定影响；

③在随机森林中，3 个类别存在严重交叉，对模型的预测有一定影响，但模型的鲁棒性较强，预测的错误率在三个模型中最低。

## 4 结语

论文着重研究数据挖掘中朴素贝叶斯判别分析算法、AdaBoost 算法以及随机森林算法在口红销量预测中的效果。通过对常用算法的比较研究，总结了它们的长处和不足。在挖掘之前，对数据库进行了数据清理、数据转换、数据词云等数据预处理，处理了空缺数据、将连续值属性离散化，为进一步挖掘打好基础。通过论文的研究，初步实现了数据挖掘技术在商品口红的应用，但是仍然存在着一些问题需进一步研究：①在数据预处理方面还不够完善，还需要依靠其它数据库工具人工完成；②文本数据处理不强，需要进一步提高。

## 参考文献

[1] 田晓虎.基于社交网络的市场营销模式研究[J].商场现代化,2018(16).

[2] 董伟.电子商务环境下的企业营销策略分析[J].现代营销(下旬刊),2015(4).

[3] 薛蕾.电子商务环境下的企业网络营销策略探讨[J].科技经济导刊,2017(7).

[4] 张彤.浅谈电子商务环境下企业的网络营销策略[J].大众投资指南,2017(2).

[5] 张保龙,刘卫俊,黄海燕,等.基于R语言的计算机类博士学位论文主题分析研究[J].济源职业技术学院学报,2021,20(2).

[6] 朱铁锋.R软件在概率论与数理统计教学中的应用研究[J].科技经济导刊,2021,29(12):33-35.

[7] Liao Yong, Cao Wen, Zhang Kunpeng, et al. Bioinformatic and integrated analysis identifies an lncRNA-miRNA-mRNA interaction mechanism in gastric adenocarcinoma[J]. Genes & genomics,2021,43(6).

[8] 熊强.基于R语言的平面公益广告数据分析与设计优化[J].包装工程,2021,42(10):355-360.

[9] 李多多,俞兴,韩晟,等.多元线性回归分析数据可视化的R Studio软件实践[J].中国循证医学杂志,2021,21(4):482-490.