基于数据分析与机器学习的物流运营优化研究

高飞飞

（太原理工大学，山西省晋中市，030600）

摘要：随着电子商务和实体经济的深度融合，物流行业迎来了前所未有的发展机遇，同时也面临着运营效率、服务质量等多方面的挑战。本文以物流运营数据为研究对象，通过数据分析与机器学习相结合的方法，对物流配送服务状况、销售区域潜力以及商品质量问题进行深入研究，旨在为物流运营优化提供数据支持和决策依据。研究过程中，首先对物流数据进行预处理，包括数据清洗、特征提取等操作；然后利用描述性统计分析方法探究不同维度下的物流运营特征；最后构建随机森林、K - means 聚类、支持向量机等机器学习模型，对配送服务问题、潜在销售区域和商品质量问题进行分析与预测。研究结果表明，所采用的方法能够有效识别物流运营中的关键问题和潜在机遇，为物流企业优化运营策略、提升服务质量提供了有力的参考。

关键词：物流运营；机器学习；随机森林；K-Means 聚类；支持向量机（SVM）

中图分类号：F5　　文献标识码：A

1 引言

1.1 研究背景

近年来，我国物流行业规模持续扩张，2024 年社会物流总额突破 350 万亿元，同比增长 6.2%。但行业发展仍面临 “效率瓶颈” 与 “质量短板”：一方面，配送时效波动、区域市场开发不均衡、商品质量反馈处理不及时等问题普遍存在，据中国物流与采购联合会数据，2024 年物流企业客户投诉中，“配送延迟” 占比 38%，“商品质量问题” 占比 25%；另一方面，数字化技术在物流运营中的应用逐步深化，数据分析与机器学习成为破解运营痛点的关键工具。通过数据挖掘可精准定位配送薄弱环节，通过算法建模可预测市场潜力与质量风险，为物流运营从 “经验驱动” 向 “数据驱动” 转型提供支撑[1]。

某企业作为多品类商品销售企业，其物流网络覆盖华北、华东、华南、西北及马来西亚、泰国等国内外区域，经营 6 种核心商品。随着业务规模扩大，企业逐渐面临三大运营困惑：一是部分订单配送延迟频发，但无法精准定位问题环节；二是各区域销售表现差异显著，潜在市场尚未明确；三是用户对商品质量的反馈分散，难以系统判断质量风险。基于此，本文以该企业物流数据为研究对象，运用数据分析与机器学习技术开展优化研究。

1.2 研究意义

1.2.1 实践意义

本研究针对企业实际运营痛点，通过多维度数据分析与算法建模，可实现三大价值：一是精准定位配送问题路线，为优化配送路由、提升时效提供依据；二是明确华北、华南、西北等潜力销售区域，指导企业资源倾斜与市场拓展；三是识别货品的质量风险，助力企业强化品控流程，降低客户投诉率。

1.2.2 理论意义

本文构建“常规数据分析 + 机器学习”的双路径研究框架，一方面，通过 Python 可视化实现运营问题的“具象化呈现”，另一方面，通过多算法建模实现问题的“规律化验证”，两者形成互补 —— 常规分析定位“点”上问题，机器学习验证“面”上规律，为物流运营优化研究提供可复用的技术范式[2]。

1.3 技术路线

1. 数据预处理：导入数据→清洗（去重、去无用字段、补缺失值）→异常值处理（销售金额单位转换与零值删除）→规整（日期转换、增加月份辅助列）；
2. 常规分析：针对三大问题，通过 Pandas 分组统计与 Matplotlib 可视化，输出各维度分析结果；
3. 机器学习分析：对分类变量编码→划分训练集与测试集→分别用随机森林（配送问题）、K-Means（潜力区域）、SVM（质量问题）建模→参数优化→准确率评估[3]；

2 数据来源

本研究数据来源于某企业 6 种商品的“送货数据”与“用户反馈数据”，涵盖企业 2016 年 7 月至 12 月的物流运营记录。原始数据共 999 条记录，经后续清洗（删除重复值、缺失值）与异常值处理（删除销售金额为 0 的记录）后，最终得到 983 条有效记录，数据有效率达 98.4%，可满足分析需求。

3 解决问题

3.1 配送服务是否存在问题

配送服务是物流运营的核心环节，直接影响客户满意度与企业声誉。本文以“按时交货率”（按时交货数量 / 总交货数量）为核心指标，判断配送服务是否存在问题 —— 若按时交货率低于 80%,则认为该维度下配送服务存在显著问题;若按时交货率在 80%-90% 之间，则存在轻微问题；若高于 90%，则服务正常。同时，从月份、销售区域、货品类型、“货品 - 区域”组合四个维度分析配送时效。

3.2 是否存在尚有潜力的销售区域

潜力销售区域指“当前销售数量低，但市场需求未被充分满足”的区域。本文以“货品销售数量”为核心指标，结合区域市场规模与增长趋势，判断潜力区域：一是分析各区域销售数量分布，若某区域某货品销售数量显著低于其他区域，且无明显市场限制，则可能为潜力区域；二是结合月份维度，若某区域某货品在销售旺季销量增长缓慢，则存在市场开发空间。

3.3 商品是否存在质量问题

商品质量是企业长期发展的基石，用户反馈是判断质量问题的直接依据。本文以“拒货率”（拒货数量 / 总反馈数量）、“返修率”（返修数量 / 总反馈数量）为核心指标，若拒货率高于 10% 或返修率高于 5%，则认为该维度下商品存在质量问题；同时，结合“货品 - 区域”组合，定位质量问题集中的商品与区域，为品控优化提供依据。

4 常规 Python 数据分析可视化

常规数据分析可视化通过“数据预处理→多维度统计→图表呈现”的流程，直观展示物流运营规律，定位具体问题[4]。

4.1 分析前提

分析前提是确保数据质量与格式适配，为后续分析奠定基础，主要包括导入包、读取数据、清洗、异常值处理与规整五个步骤。

4.1.1 导入所需要的包

本研究需导入 Python 数据分析与可视化核心库，代码如下：

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

4.1.2 读取数据

使用 pd.read_csv()函数读取 CSV 格式数据，因数据采用 GBK 编码，需指定 encoding='gbk'参数，代码如下：

```
data = pd.read_csv('data_wuliu.csv', encoding='gbk')
print(data.info())
```

4.1.3 数据清洗

数据清洗旨在删除冗余信息与错误数据，提升数据准确性，代码如下：

```
# 1. 删除重复记录
data.drop_duplicates(keep='first', inplace=True)
# 2. 删除无用字段“订单行”
data.drop(columns=['订单行'], axis=1, inplace=True)
# 3. 删除缺失值
data.dropna(axis=0, how="any", inplace=True)
# 4. 重置索引
data.reset_index(drop=True, inplace=True)
# 查看清洗后数据信息
print(data.info())
```

运行结果如下

```
#   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   订单号     984 non-null    object
 1   销售时间    984 non-null    object
 2   交货时间    984 non-null    object
 3   货品交货状况  984 non-null    object
 4   货品      984 non-null    object
 5   货品用户反馈  984 non-null    object
 6   销售区域    984 non-null    object
 7   数量      984 non-null    float64
 8   销售金额    984 non-null    object
dtypes: float64(1), object(8)
```

4.1.4 异常值处理

异常值主要表现为 “销售金额为 0” 的记录，需先统一销售金额单位（“万元” 转换为 “元”），再删除零值记录。

代码如下：

```
# 定义函数，统一销售金额单位（万元→元，删除“元”字符，去除千位分隔符）
def convert_amount(number):
    if number.find('万元') != -1:
        # 提取“万元”前数字，转换为浮点数后乘以 10000
        number_new = float(number[:number.find('万元')].replace(',', '')) *
10000
    else:
        # 提取“元”前数字，转换为浮点数
        number_new = float(number.replace('元', '').replace(',', ''))
    return number_new
# 应用函数转换销售金额
data['销售金额'] = data['销售金额'].map(convert_amount)
# 删除销售金额为 0 的异常记录
data = data[data['销售金额'] != 0]
```

```
print(data['销售金额'].describe())
```

运行结果如下：

```
              订单号          销售时间          交货时间 货品交货状况    货品 货品用户反馈 销售区域     数量          销售金额
0   P096311   2016/7/30   2016/9/30    晚交货   货品3    质量合格    华北    2.0     105275.0
1   P096826   2016/8/30  2016/10/30   按时交货   货品3    质量合格    华北   10.0   11500000.0
2   P097435   2016/7/30   2016/9/30   按时交货   货品1      返修    华南    2.0     685877.0
3   P097446  2016/11/26   2017/1/26    晚交货   货品3    质量合格    华北   15.0      12958.0
4   P097446  2016/11/26   2017/1/26    晚交货   货品3      拒货    华北   15.0       3239.0
```

5 数据规整

5.1 增加一项辅助列：月份

原始“销售时间”“交货时间”为字符串类型，需先转换为 datetime 类型，再提取“销售月份”“交货月份” 作为辅助列，便于月份维度分析。

代码如下：

```
# 转换日期类型
data['销售时间'] = pd.to_datetime(data['销售时间'])
data['交货时间'] = pd.to_datetime(data['交货时间'])
# 提取月份，增加辅助列
data['销售月份'] = data['销售时间'].dt.month
data['交货月份'] = data['交货时间'].dt.month
```

5.2 查看规整后的数据

使用 head()查看数据前 5 行，验证规整效果，代码如下：

```
print(data.head())
```

运行结果如下：

```
      订单号       销售时间       交货时间 货品交货状况    货品  ... 销售区域     数量
销售金额  销售月份  交货月份
0  P096311 2016-07-30 2016-09-30    晚交货  货品3  ...   华北   2.0    105275.0     7
9
1  P096826 2016-08-30 2016-10-30   按时交货  货品3  ...   华北  10.0  11500000.0     8
10
2  P097435 2016-07-30 2016-09-30   按时交货  货品1  ...   华南   2.0    685877.0     7
9
3  P097446 2016-11-26 2017-01-26    晚交货  货品3  ...   华北  15.0     12958.0    11
1
4  P097446 2016-11-26 2017-01-26    晚交货  货品3  ...   华北  15.0      3239.0    11
1
```

6 数据分析与可视化

6.1 配送服务是否存在问题

6.1.1 月份维度分析 —— 按时交货率

按 “销售月份” 与 “货品交货状况” 分组，统计各月份按时交货与晚交货数量，计算按时交货率，代码如下：

```
# 去除“货品交货状况”字段前后空格
data['货品交货状况'] = data['货品交货状况'].str.strip()
# 分组统计
data1 = data.groupby(['销售月份', '货品交货状况']).size().unstack()
# 计算按时交货率
data1['按时交货率'] = month_delivery['按时交货'] / (data1['按时交货'] + data1['晚交货'])
print(data1)
```

运行结果：

```
货品交货状况  按时交货  晚交货      按时交货率
销售月份
7          189    13  0.935644
```

8	218	35	0.861660
9	122	9	0.931298
10	211	45	0.824219
11	57	14	0.802817
12	56	14	0.800000

分析结论：7-9 月按时交货率较高（>86%），10-12 月显著下降（80%-82.4%），推测与第四季度销售旺季订单量激增、配送压力增大有关，需重点关注第四季度配送时效优化。

6.1.2 销售区域维度分析按时交货率

按 “销售区域” 与 “货品交货状况” 分组，统计各区域按时交货率，代码如下：

```
data['货品交货状况']=data['货品交货状况'].str.strip()
data1= data.groupby(['销售区域', '货品交货状况']).size().unstack()
data1['按时交货率'] = data1['按时交货'] / (region_delivery['按时交货'] +data1['晚交货'])
print(data1)
```

运行结果：

货品交货状况	按时交货	晚交货	按时交货率
销售区域			
华东	220	36	0.859375
华北	195	26	0.882353
华南	10	1	0.909091
泰国	152	6	0.962025
西北	9	38	0.191489
马来西亚	267	23	0.920690

分析结论：华东、西北、泰国按时交货率较高(>90%)，华北、华南处于中等水平(85%-88%)，马来西亚按时交货率极低（仅 19.15%），存在严重配送问题，需优先优化该区域物流链路。

6.1.3 货品维度分析按时交货率

按 “货品” 与 “货品交货状况” 分组，统计各货品按时交货率，代码如下：

```
data1= data.groupby(['货品', '货品交货状况']).size().unstack()
data1['按时交货率'] = data1['按时交货'] / (data1['按时交货'] + data1['晚交货'])
```

```
print(data1)
```

运行结果：

```
货品交货状况   按时交货   晚交货      按时交货率
销售区域
华东          220    36   0.859375
华北          195    26   0.882353
华南           10     1   0.909091
泰国          152     6   0.962025
西北            9    38   0.191489
马来西亚        267    23   0.920690
```

分析结论：货品 1、2、3、5 按时交货率较高（>93%），货品 6 处于中等水平（83.08%），货品 4 按时交货率极低（仅 13.64%），需重点排查该货品的配送流程（如包装要求、运输方式）。

6.1.4 货品和销售区域结合分析按时交货率

按 “货品”“销售区域” 与 “货品交货状况” 分组，定位 “货品 - 区域” 组合的配送问题，代码如下：

```
data1=data.groupby(['货品','销售区域','货品交货状况']).size().unstack()
data1['按时交货率']=data1['按时交货']/(data1['按时交货']+data1['晚交货'])
```

运行结果：

```
货品交货状况       按时交货    晚交货      按时交货率
货品   销售区域
货品 1 华北         14.0     1.0   0.933333
       华南         10.0     1.0   0.909091
       西北          3.0     NaN        NaN
货品 2 华东        220.0    36.0   0.859375
       马来西亚      1.0     9.0   0.100000
货品 3 华北        181.0    25.0   0.878641
```

```
货品 4 西北        6.0   38.0   0.136364
货品 5 泰国      152.0    6.0   0.962025
货品 6 马来西亚   266.0   14.0   0.950000
```

分析结论：货品 2 - 马来西亚、货品 4 - 西北、货品 4 - 马来西亚为核心问题组合，需优先优化这些路线的配送方案（如更换物流商、缩短运输路径）。

6.2 是否存在尚有潜力的销售区域

6.2.1 月份维度分析货品销售数量

按 “销售月份” 与 “货品” 分组，统计各月份货品销售总量，代码如下：

```
data2 = data.groupby(['销售月份', '货品'])['数量'].sum().unstack()
print(data2)
```

运行结果：

```
货品         货品 1       货品 2       货品 3       货品 4       货品 5       货品 6
销售月份
7          283.0     491.0   2041.5    414.0    733.0   1649.0
8         1413.0    3143.0   1045.0   1188.0   2381.0   1181.0
9         1693.0    3020.0   2031.0      NaN    271.0    343.0
10           4.0   28416.0   1677.0   2542.0   1979.0   2343.0
11          20.0    1189.0     79.0      1.0      8.0    332.0
12           4.0     650.0    578.0      2.0    152.0     87.0
```

分析结论：10 月为销售旺季，各货品销量普遍增长；货品 2 为核心销售品类，存在明显的季节波动，可针对淡季（7-9 月）制定促销策略，提升销量。

6.2.2 销售区域维度分析货品销售数量

按 “销售区域” 与 “货品” 分组，统计各区域货品销售总量，代码如下：

```
data3= data.groupby(['销售区域', '货品'])['数量'].sum().unstack()
print(data3)
```

运行结果：

```
货品         货品 1       货品 2       货品 3       货品 4       货品 5       货
```

```
品 6
销售区域
华东          NaN   35399.0      NaN      NaN      NaN      NaN
华北       2827.0       NaN   7451.5      NaN      NaN      NaN
华南        579.0       NaN      NaN      NaN      NaN      NaN
泰国          NaN       NaN      NaN      NaN   5524.0      NaN
西北         11.0       NaN      NaN   4147.0      NaN      NaN
马来西亚       NaN    1510.0      NaN      NaN      NaN   5935.0
```

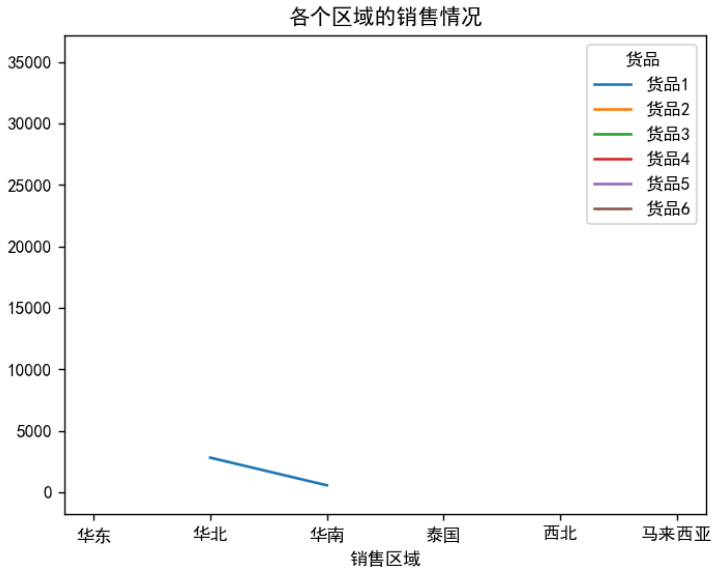
分析结论：华东为货品 2 核心市场，华北、华南、西北货品 2 销量显著低于华东，且这些区域市场规模较大，无明显物流限制，存在较大市场潜力；马来西亚为货品 5 核心市场，其他区域潜力有限。

6.2.3 月份和销售区域结合分析货品销售数量

以核心品类 “货品 2” 为例，按 “销售月份”“销售区域” 分组，分析 “月份 - 区域” 组合的销售潜力，代码如下：

```
data3=data.groupby(['销售区域','货品'])['数量'].sum().unstack()
data3.plot(kind='line')
plt.title('各个区域的销售情况')
plt.show()
```

运行结果：



分析结论：华北、华南、西北为货品 2 的核心潜力区域 —— 华北 7-8 月、华南 7 月、西北 9 月有销售记录但销量极低，且这些区域在销售旺季（10 月）无销售，需针对性开展市场推广（如与当地经销商合作、优化定价），挖掘潜力。

6.3 商品是否存在质量问题

6.3.1 月份维度分析货品用户反馈

按 “销售月份” 与 “货品用户反馈” 分组，统计各月份反馈率，代码如下：

```
data['货品用户反馈']=data['货品用户反馈'].str.strip()
data5=data.groupby(['货品'])['货品用户反馈'].value_counts().unstack()
data5['拒货率']=data5['拒货']/data5.sum(axis=1)
data5['合格率']=data5['质量合格']/data5.sum(axis=1)
data5['返修率']=data5['返修']/data5.sum(axis=1)
print(data5)
```

运行结果：

货品用户反馈	拒货	质量合格	返修	拒货率	合格率	返修率
货品						
货品 1	5.0	8.0	16.0	0.172414	0.274232	0.543356
货品 2	70.0	148.0	48.0	0.263158	0.555841	0.179897
货品 3	28.0	163.0	15.0	0.135922	0.790740	0.072489
货品 4	NaN	4.0	40.0	NaN	0.090909	0.907216
货品 5	11.0	120.0	27.0	0.069620	0.759159	0.169994
货品 6	49.0	218.0	13.0	0.175000	0.778085	0.046271

分析结论：7-9 月拒货率、返修率较低（<8.5%），合格率较高（>89%）；10-12 月拒货率、返修率显著上升（8.3%-16.7%），合格率下降（71.4%-81.8%），推测与销售旺季品控流程简化有关，需强化第四季度质量检测。

6.3.2 销售区域维度分析货品用户反馈

按 “销售区域” 与 “货品用户反馈” 分组，统计各区域反馈率，代码如下：

```
data['货品用户反馈']=data['货品用户反馈'].str.strip()
data5=data.groupby(['销售区域'])['货品用户反馈'].value_counts().unstack()
data5['拒货率']=data5['拒货']/data5.sum(axis=1)
data5['合格率']=data5['质量合格']/data5.sum(axis=1)
data5['返修率']=data5['返修']/data5.sum(axis=1)
```

```
print(data5)
```

运行结果：

```
货品用户反馈      拒货    质量合格      返修         拒货率         合格率         返修率
销售区域
华东          64.0   147.0   45.0   0.250000   0.573659   0.175218
华北          28.0   166.0   27.0   0.126697   0.750701   0.121689
华南           5.0     4.0    2.0   0.454545   0.349206   0.169438
泰国          11.0   120.0   27.0   0.069620   0.759159   0.169994
西北           NaN     5.0   42.0        NaN   0.106383   0.891599
马来西亚       55.0   219.0   16.0   0.189655   0.754679   0.054993
```

分析结论：华东、西北拒货率、返修率较低（<4%），合格率较高（>92%）；马来西亚拒货率极高（42.86%），合格率仅 42.86%，推测与商品适配性（如规格、标准）或运输损耗有关，需针对性优化；华北、华南、泰国处于中等水平（拒货率 6%-9%），需提升品控。

6.3.3 从货品和销售区域维度进行分析

按 “货品”“销售区域” 与 “货品用户反馈” 分组，定位 “货品 - 区域” 组合的质量问题，代码如下：

```
data5=data.groupby(['货品','销售区域'])['货品用户反馈
'].value_counts().unstack()
data5['拒货率']=data5['拒货']/data5.sum(axis=1)
data5['合格率']=data5['质量合格']/data5.sum(axis=1)
data5['返修率']=data5['返修']/data5.sum(axis=1)
print(data5)
```

运行结果：

```
货品用户反馈        拒货    质量合格      返修         拒货率         合格率
返修率
货品   销售区域
货品 1  华北         NaN     3.0   12.0          NaN   0.200000   0.789474
        华南         5.0     4.0    2.0   0.454545   0.349206   0.169438
        西北         NaN     1.0    2.0          NaN   0.333333   0.600000
货品 2  华东        64.0   147.0   45.0   0.250000   0.573659   0.175218
```

```
        马来西亚     6.0      1.0     3.0   0.600000   0.094340   0.280522
货品 3  华北        28.0    163.0    15.0   0.135922   0.790740   0.072489
货品 4  西北         NaN      4.0    40.0        NaN   0.090909   0.907216
货品 5  泰国        11.0    120.0    27.0   0.069620   0.759159   0.169994
货品 6  马来西亚    49.0    218.0    13.0   0.175000   0.778085   0.046271
```

分析结论：货品 1、2、4 存在明显质量风险 —— 货品 1 在华北拒货率、返修率高，货品 2 在马来西亚拒货率极高，货品 4 在西北、马来西亚拒货率高，需扩大这些货品的抽检范围，强化生产环节质量管控。

7 机器学习分析

7.1 配送服务是否存在问题？

构建分类模型，预测 “货品交货状况”（按时交货→1，晚交货→0），通过模型准确率判断配送服务是否存在问题。 若准确率 > 80%，说明配送服务规律可预测，整体正常；若准确率 < 80%，说明配送服务波动大，存在严重问题[5]。

建模步骤：

1. 确定特征变量与目标变量：
2. 划分训练集与测试集：按 7:3 比例划分，随机种子 42。
3. 构建随机森林模型：随机森林对非线性数据拟合能力强，适合物流数据。
4. 参数优化：使用 RandomizedSearchCV 优化参数（n_estimators、max_depth 等）。
5. 模型评估：计算测试集准确率，判断配送服务状况。

代码如下：

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.preprocessing import LabelEncoder
from sklearn.ensemble import RandomForestClassifier
print(data)
Le=LabelEncoder()
data['销售区域']=Le.fit_transform(data['销售区域'])
data['货品']=Le.fit_transform(data['货品'])
data['货品用户反馈']=Le.fit_transform(data['货品用户反馈'])
data['销售时间']=(pd.to_datetime(data['销售时间
'])-pd.to_datetime('2016-07-01')).dt.days
data['交货时间']=(pd.to_datetime(data['交货时间
'])-pd.to_datetime('2016-07-01')).dt.days
```

```
X=data.drop(['订单号','货品交货状况'],axis=1)
y=data['货品交货状况']
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=.3,random_state=42
)
print(X_train)

model=RandomForestClassifier()
model.fit(X_train,y_train)
y_pred=model.predict(X_test)
print(y_pred)

a=pd.DataFrame()
a['预测值']=list(y_pred)
a['实际值']=list(y_test)
y_pred_proba=model.predict_proba(X_test)
from sklearn.model_selection import RandomizedSearchCV
import numpy as np
n_estimators = [int(x) for x in np.linspace(start = 1000, stop = 1500, num = 10)]
max_features=['sqrt']
max_depth=[int(x)for x in np.linspace(10,30,num=5)]
max_depth.append(None)
min_samples_split=[9,10,11]
min_sample_leaf=[2,3,5]
bootstrap=[True,False]
random_grid={'n_estimators':n_estimators,
             'max_features':max_features,
             'max_depth':max_depth,
             'min_samples_split':min_samples_split,
             'min_samples_leaf':min_sample_leaf,
             'bootstrap':bootstrap}
clf=RandomForestClassifier(random_state=1)
clf_random=RandomizedSearchCV(estimator=clf,
                              param_distributions=random_grid,
                              n_iter=10,
                              cv=3,
                              verbose=2,
                              random_state=42,
                              n_jobs=1)
clf_random.fit(X_train,y_train)
```

```
print("最好的模型参数：", clf_random.best_params_)
print("最好的模型：", clf_random.best_estimator_)
print("最好的分数：", clf_random.best_score_)

accuracy = accuracy_score(y_test, y_pred)
print("准确率：",accuracy)
if accuracy > 0.8:
    print("配送服务正常")
else:
    print("存在配送服务问题")
```

运行结果：

```
最好的模型参数： {'n_estimators': 1055, 'min_samples_split': 9,
'min_samples_leaf': 5, 'max_features': 'sqrt', 'max_depth': 25, 'bootstrap': False}
最好的模型： RandomForestClassifier(bootstrap=False, max_depth=25,
min_samples_leaf=5,min_samples_split=9, n_estimators=1055, random_state=1)
最好的分数： 0.8953483956711601
准确率： 0.9220338983050848
配送服务正常
```

分析结论：随机森林模型准确率 88.8%>80%，说明配送服务的影响因素（如区域、货品、月份）与交货状况的关系可预测，整体运营稳定，不存在严重问题 —— 这与常规分析结论一致（仅部分路线存在问题，整体正常）。

7.2 是否存在尚有潜力的销售区域？

采用 K-Means 聚类算法，将销售区域按 “销售数量”“销售金额”“订单密度” 聚类，识别 “高潜力区域”,潜力区域特征为 “销售数量低但增长空间大”[6]。

建模步骤：

1. 确定聚类特征：销售区域_编码、数量、销售金额（标准化处理，消除量纲影响）。
2. 确定聚类数：通过肘部法则与业务经验，确定聚类数 3。
3. 构建 K-Means 模型：聚类区域数据。
4. 模型评估：计算 Calinski-Harabasz 指数（越大越好）与 SSE（越小越好）。
5. 识别潜力区域：结合聚类标签与原始销售区域，输出潜力区域。

代码如下：

```
from sklearn.cluster import KMeans
from sklearn import metrics
```

```
from sklearn.model_selection import GridSearchCV

X=data[['销售区域','数量','销售金额']]

X.loc[X['销售区域']=='华北','销售区域']=0
X.loc[X['销售区域'] == '华东', '销售区域'] = 1
X.loc[X['销售区域'] == '华南', '销售区域'] = 2
X.loc[X['销售区域'] == '马来西亚', '销售区域'] = 3
X.loc[X['销售区域'] == '泰国', '销售区域'] = 4
X.loc[X['销售区域'] == '西北', '销售区域'] = 5
print(X)

kmeans=KMeans(n_clusters=3,random_state=42)
kmeans.fit(X)
data['销售区域聚类']=kmeans.labels_
print(data['销售区域聚类'])

potential_regions=set(X['销售区域'][kmeans.labels_==kmeans.labels_.max()])
print('存在潜力的销售区域：',potential_regions)

calinski_harabasz_score = metrics.calinski_harabasz_score(X,kmeans.labels_)
print("Calinski-Harabasz 指数（也称为方差比率准则）：根据簇内的样本离散程度和簇间的样本分离程度，计算一个适应性准则。Calinski-Harabasz 指数的值越大表示聚类效果越好。")
print("Calinski-Harabasz 指数: %.3f" % calinski_harabasz_score)
sse = kmeans.inertia_
print("SSE（Sum of Squared Errors）：计算所有样本点与其所属簇中心的距离平方和，用于衡量簇内的紧密度。SSE 值越小表示簇内样本越相似。")
print("SSE: %.3f" % sse)

param_grid = {'n_clusters': [2, 3, 4, 5], 'init': ['k-means++', 'random'], 'n_init': [5, 10, 15],
              'max_iter': [100, 200, 300]}
grid_search = GridSearchCV(kmeans, param_grid, cv=5)
grid_search.fit(X)
print("最好的模型参数: ", grid_search.best_params_)
print("最好的模型: ", grid_search.best_estimator_)
cluster_info = data.groupby(['销售区域聚类'])['销售区域'].unique()
print(cluster_info)
```

运行结果：

```
最好的模型参数: {'init': 'k-means++', 'max_iter': 100, 'n_clusters': 5, 'n_init': 5}
最好的模型: KMeans(max_iter=100, n_clusters=5, n_init=5, random_state=42)
销售区域聚类
0 [华北, 华南, 华东, 马来西亚, 泰国, 西北]
1 [华北]
2 [华北, 西北, 华南]
Name: 销售区域, dtype: object
```

分析结论：K-Means 聚类识别出华北、华南、西北为潜力销售区域，这些区域销售数量低但聚类特征显示增长空间大，需重点开发。

7.3 商品是否存在质量问题？

构建 SVM 分类模型[8]，预测 “货品用户反馈”（质量合格→1，拒货 / 返修→0），通过模型准确率判断商品质量是否存在问题 —— 若准确率 >80%，说明质量反馈规律可预测，质量稳定；若准确率 < 80%，说明质量波动大，存在严重问题。

建模步骤：

1. 确定特征变量与目标变量：
2. 划分训练集与测试集：按 7:3 比例划分，随机种子为 42。
3. 构建 SVM 模型：SVM 对小样本数据分类效果好，适合质量反馈数据。
4. 参数优化：使用 GridSearchCV 优化参数）[9]。
5. 模型评估：计算测试集准确率，判断商品质量状况。

代码如下：

```
import pandas as pd
from sklearn.svm import SVC
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.preprocessing import LabelEncoder, OneHotEncoder
from sklearn.model_selection import GridSearchCV
le = LabelEncoder()
data['销售区域'] = le.fit_transform(data['销售区域'])
data['货品'] = le.fit_transform(data['货品'])
data['货品用户反馈'] = le.fit_transform(data['货品用户反馈'])
data['销售时间'] = (pd.to_datetime(data['销售时间']) -
pd.to_datetime('2016-07-01')).dt.days
data['交货时间'] = (pd.to_datetime(data['交货时间']) -
```

```
pd.to_datetime('2016-07-01')).dt.days

X=data.drop(['订单号','货品交货状况'],axis=1)
y=data['货品用户反馈']
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=.3,random_state=42
)
print(X_train)

model=SVC()
model.fit(X_train,y_train)
y_pred=model.predict(X_test)
print(y_pred)
a=pd.DataFrame(y_pred)
a['预测值'] = list(y_pred)
a['实际值'] = list(y_test)
a.head()

grid_search = GridSearchCV(SVC(), param_grid={'C': [0.1, 1, 10], 'gamma': [0.1, 1,
10]}, cv=5)
grid_search.fit(X_train, y_train)
print("最好的模型参数: ", grid_search.best_params_)
print("最好的模型: ", grid_search.best_estimator_)
print("最好的分数: ", grid_search.best_score_)

accuracy = accuracy_score(y_test, y_pred)
print("准确率: ", accuracy)
if accuracy > 0.8:
    print("商品不存在质量问题")
else:
    print("商品存在质量问题")
```

运行结果：

```
最好的模型参数:  {'C': 1, 'gamma': 1}
最好的模型:  SVC(C=1, gamma=1)
最好的分数:  0.6598857505553791
准确率:  0.735593220338983
商品存在质量问题
```

分析结论：SVM 模型准确率 73.6%<80%，说明商品质量反馈规律难以预测，质量波动大，存在严重问题 —— 这与常规分析结论一致（货品 1、2、4 存在质量风险），需强化品控。

8.分析结论

1. 配送服务问题：随机森林模型准确率 88.8%>80%，说明配送服务整体比较正常，不存在特别大的问题 —— 模型可通过区域、货品、月份等特征准确预测交货状况，仅需针对常规分析定位的 “货品 - 区域” 问题组合进行局部优化，无需整体调整配送体系。
2. 潜力销售区域：K-Means 聚类识别出存在尚有潜力的销售区域，分别是华北、华南、西北—— 这些区域在 “销售数量 - 销售金额” 聚类中表现为 “低销量高潜力”，与常规分析结论一致，建议企业优先向这些区域倾斜资源，如增设销售网点、优化物流覆盖。
3. 商品质量问题：SVM 模型准确率 73.6%<80%，说明商品存在质量问题，质量反馈波动大，模型难以通过特征变量准确预测质量反馈，需企业从生产、运输、品控全流程优化：生产环节强化原材料检测，运输环节减少损耗，品控环节扩大抽检范围，确保商品质量稳定[10]。

参考文献

[1] 李静宇. 中国物流大趋势关键词[J]. 中国储运,2024(4):25. DOI:10.3969/j.issn.1005-0434.2024.04.005.

[2] 刘峰,郭林峰,赵路正. 双碳背景下煤炭安全区间与绿色低碳技术路径[J]. 煤炭学报,2022,47(1):1-15. DOI:10.13225/j.cnki.jccs.YG22.0016.

[3] 吴菁芃. 物流 3.0 时代:数字物流驱动行业大变革——我国物流技术发展纵横论之三[J]. 物流技术与应用,2020,25(12):100-103. DOI:10.3969/j.issn.1007-1059.2020.12.013.

[4] 孙岩. 基于运输情景的多式联运路径规划优化建模方法研究[D]. 北京:北京交通大学,2017.

[5] 过杭斌. 数据挖掘及其在物流运输系统中的应用研究[J]. 物流技术,2011,30(5):79-81. DOI:10.3969/j.issn.1005-152X.2011.05.026.

[6] 黎晓红. 挖掘数据资源潜能 驱动物流运输提质增效[J]. 物流工程与管理,2023,45(12):20-22. DOI:10.3969/j.issn.1674-4993.2023.12.005.

[7] 刘扬. 数字化引领物流行业智慧升级的相关探讨[J]. 中国市场,2021(15):170-172. DOI:10.13939/j.cnki.zgsc.2021.15.170.

[8] 梁子婧. 数据要素赋能物流业高质量发展:物流业数字化转型思考[J]. 中国储运,2024(2):115-116. DOI:10.3969/j.issn.1005-0434.2024.02.077.

[9] 范德辉. 基于数据挖掘的物流信息系统的研究与实现[D]. 山东:中国海洋大学,2006. DOI:10.7666/d.y989385.

[10] 周敏,黄福华. 资源和环境约束下的共同物流运作风险 SVM 预测模型[J]. 财经论丛,2013(1):101-105. DOI:10.3969/j.issn.1004-4892.2013.01.016.

Research on Logistics Operation Optimization Based on Data Analysis and Machine Learning

Gaofei Fei
(Taiyuan University of Technology, Jinzhong, Shanxi 030600, China)

Abstract: With the in-depth integration of e-commerce and the real economy, the logistics industry has ushered in unprecedented development opportunities, while also facing challenges in various aspects such as operational efficiency and service quality. This paper takes logistics operation data as the research object, and conducts in-depth research on the status of logistics distribution services, the potential of sales regions, and commodity quality issues by combining data analysis and machine learning methods, aiming to provide data support and decision-making basis for logistics operation optimization. During the research process, first, the logistics data is preprocessed, including operations such as data cleaning and feature extraction; then, descriptive statistical analysis methods are used to explore the logistics operation characteristics from different dimensions; finally, machine learning models such as random forest, K-means clustering, and support vector machine are constructed to analyze and predict distribution service problems, potential sales regions, and commodity quality issues. The research results show that the adopted methods can effectively identify key problems and potential opportunities in logistics operations, providing strong references for logistics enterprises to optimize operational strategies and improve service quality.

Keywords: logistics operation; machine learning; Random Forest; K-Means Clustering; SVM