**某大型超市2011-2014年全球数据集数据分析报告**

严语陈

# 摘要

在零售业中，**数据分析**作为提高营销效率和预测销售趋势的关键工具，具有重要的实践意义。本文通过对某大型超市的历史数据进行**数据清洗与预处理**，并运用**探索式数据分析**（EDA）深入挖掘数据特征，旨在实现客户群体细分和未来销售趋势的预测。

首先，基于RFM模型和**KMeans聚类分析**，对客户进行有效的**细分**。在RFM模型中，依据客户的最近购买时间（Recency）、购买频率（Frequency）和购买金额（Monetary），对客户价值进行评估。随后通过KMeans算法对客户群体进行聚类，利用**数据归一化**、**模型训练与评估**以及**参数搜索**优化模型，得到合理的**细分结果**。基于这些结果，提出相应的**营销策略**，以提升客户粘性和转化率。

其次，在预测未来销售趋势方面，本文采用了**SARIMAX模型**进行时间序列分析和预测。通过对历史销售数据的**趋势分析**和**季节性时间序列分解**，揭示销售额的潜在变化规律。模型经过**参数网格搜索**和优化后，最终得出较为准确的销售预测结果，并通过**模型融合**进一步提升预测的准确性。

综上所述，本研究结合了多种数据分析技术，为超市在客户管理和销售预测方面提供了实用的解决方案，具有较强的应用价值。

**关键词：探索式数据分析、RFM模型、KMeans聚类、SARIMAX模型、客户群体细分、销售预测**

**目录**

[摘要 1](#_Toc177668849)

[一、 绪论 3](#_Toc177668850)

[（一） 行业背景与意义 3](#_Toc177668851)

[（二） 数据分析目标 3](#_Toc177668852)

[二、 数据清洗与预处理 3](#_Toc177668853)

[1. 缺失值处理 3](#_Toc177668854)

[2. 数据类型转换 4](#_Toc177668855)

[三、 探索式数据分析 4](#_Toc177668856)

[四、 目标1：客户群体细分与营销策略 6](#_Toc177668857)

[（一） RFM模型 6](#_Toc177668858)

[1. 简介 6](#_Toc177668859)

[2. 模型计算 6](#_Toc177668860)

[（二） KMeans聚类分析 7](#_Toc177668861)

[1. 数据归一化 7](#_Toc177668862)

[2. 模型训练与评估 7](#_Toc177668863)

[3. 参数搜索 7](#_Toc177668864)

[4. 结果统计 8](#_Toc177668865)

[（三） 营销策略制定 9](#_Toc177668866)

[五、 目标2：预测未来销售趋势 10](#_Toc177668867)

[（一） 数据趋势分析 10](#_Toc177668868)

[（二） 使用SARIMAX模型预测销售额 10](#_Toc177668869)

[1. 季节性时间序列分解 10](#_Toc177668870)

[2. 模型训练与评估 11](#_Toc177668871)

[3. 参数网格搜索 12](#_Toc177668872)

[4. 网格搜索优化 12](#_Toc177668873)

[5. 模型结果 12](#_Toc177668874)

[6. 模型融合 13](#_Toc177668875)

# 绪论

## 行业背景与意义

近年来，随着经济的快速发展和消费者购买力的提升，零售行业尤其是超市业态呈现出持续增长的趋势。超市作为日常生活中不可或缺的零售渠道，涵盖了从食品、日用品到家居产品等多种品类，深度嵌入消费者的日常消费场景中。根据市场调研数据显示，国内零售行业的总销售额在过去几年保持稳定增长，并在电子商务蓬勃发展的背景下，线上与线下融合的全渠道零售模式逐渐成为行业的主要趋势。

大型超市作为零售行业中的重要组成部分，面临着激烈的市场竞争与消费者偏好快速变化的挑战。随着消费者对产品质量、购物体验和个性化需求的要求不断提高，超市需要更加依赖精准的数据分析，以洞察市场动态、优化供应链、提升顾客满意度和推动销售增长。数据分析在其中的作用尤为关键，它能够帮助企业挖掘潜在的市场机会，发现运营中的问题，并通过科学决策提高整体效益。

此次数据分析的目的在于通过深入挖掘超市的销售数据，了解各类商品的销售表现和消费者偏好，从而为超市的商品布局、库存管理、促销策略等提供数据支持，最终提高超市的市场竞争力与运营效率。

## 数据分析目标

本次数据分析的主要目标是通过对超市销售数据的深入挖掘，帮助超市管理层更好地理解商品销售趋势，以获得更好的销售额和利润。具体目标如下：

1. **客户群体细分与营销策略**

客户类型多样，包括零售客户、批发客户等，不同类型客户的购买行为可能不同。制定精准的营销策略，可以吸引不同的客户群体，提升客户满意度和忠诚度。

1. **预测未来销售趋势**

超市在2011-2014年的销售数据展示了在不同国家和地区的销售情况，有可能存在季节性波动和长期趋势。 超市管理层可能对未来的销售趋势感兴趣，以便制定战略规划。

# 数据清洗与预处理

在本次数据分析过程中，为了简化分析过程和保证分析结果的准确性，我们需要对数据的缺失值，异常值等异常数据进行处理。

### 缺失值处理

通过分析得到，此数据集中Postal Code（邮政编码）字段存在大量的缺失值，并且其他数据都不存在缺失值。由于此次分析任务不涉及到邮政编码，所以我们直接将此字段去除。

### 数据类型转换

此次分析任务涉及到Order Date（下单时间）和Ship Date（发货时间）两个时间数据。通过分析得到，这些时间数据的格式不统一。我们将其进行识别后，转换成统一的datetime64时间数据类型。

# 探索式数据分析

在本次数据分析项目中，探索性数据分析（Exploratory Data Analysis，简称 EDA）作为数据处理的重要步骤，旨在通过可视化和统计分析方法，对超市的销售数据进行初步的理解与特征发掘。

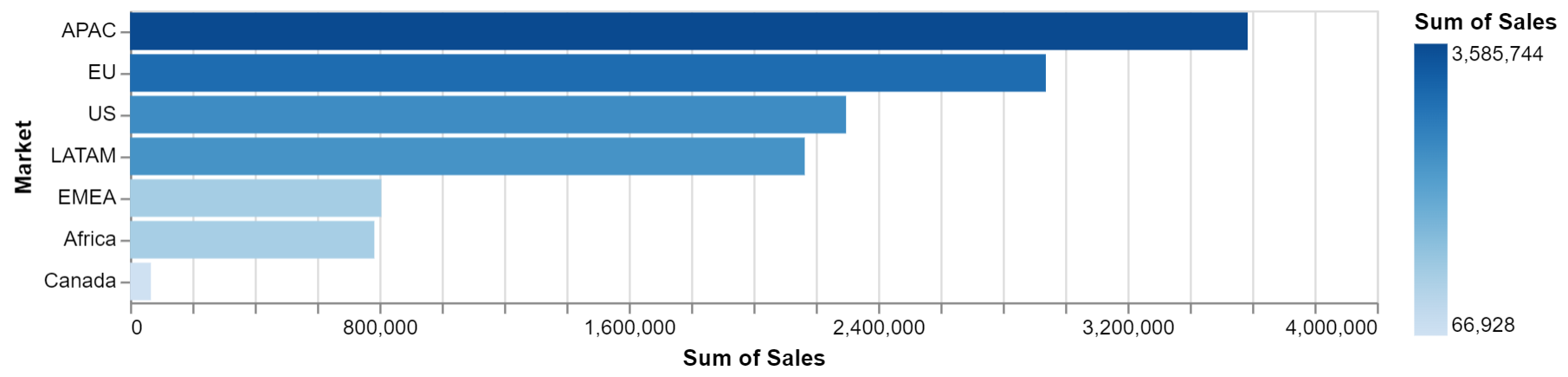


图 1 各市场销售额总和

从上图分析得出，此超市主要客户群体处于亚太地区，其次为欧洲和美国。

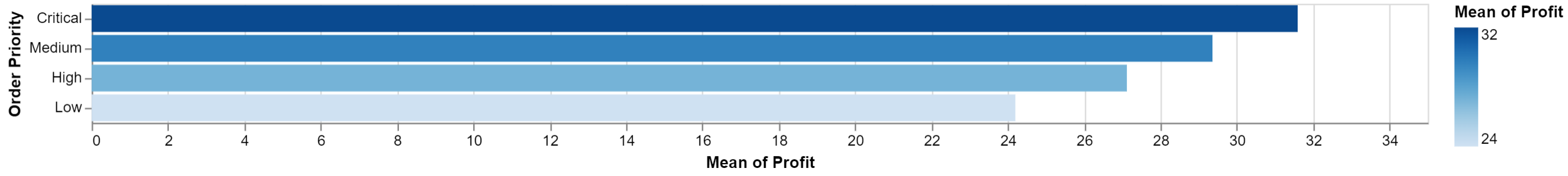


图 2 订单优先级与平均利润关系

从上图分析得出，平均利润最高的是Critical（紧急）订单，而High（高）优先级的平均利润低于Medium（中等）优先级的平均利润。

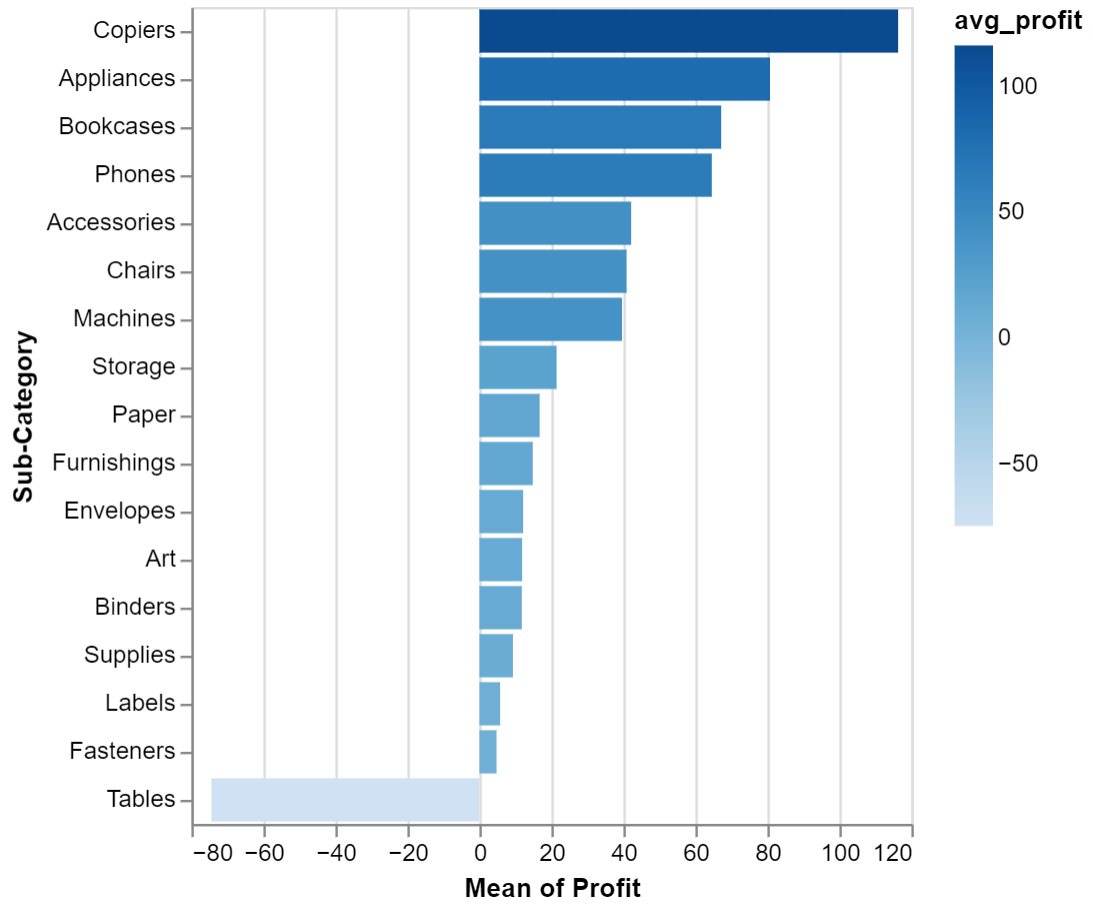


图 3 商品子类别与平均利润关系

从上图分析得出，复印机是利润最高的商品，而只有桌子的平均利润为负，我们推测这其中可能存在异常值。

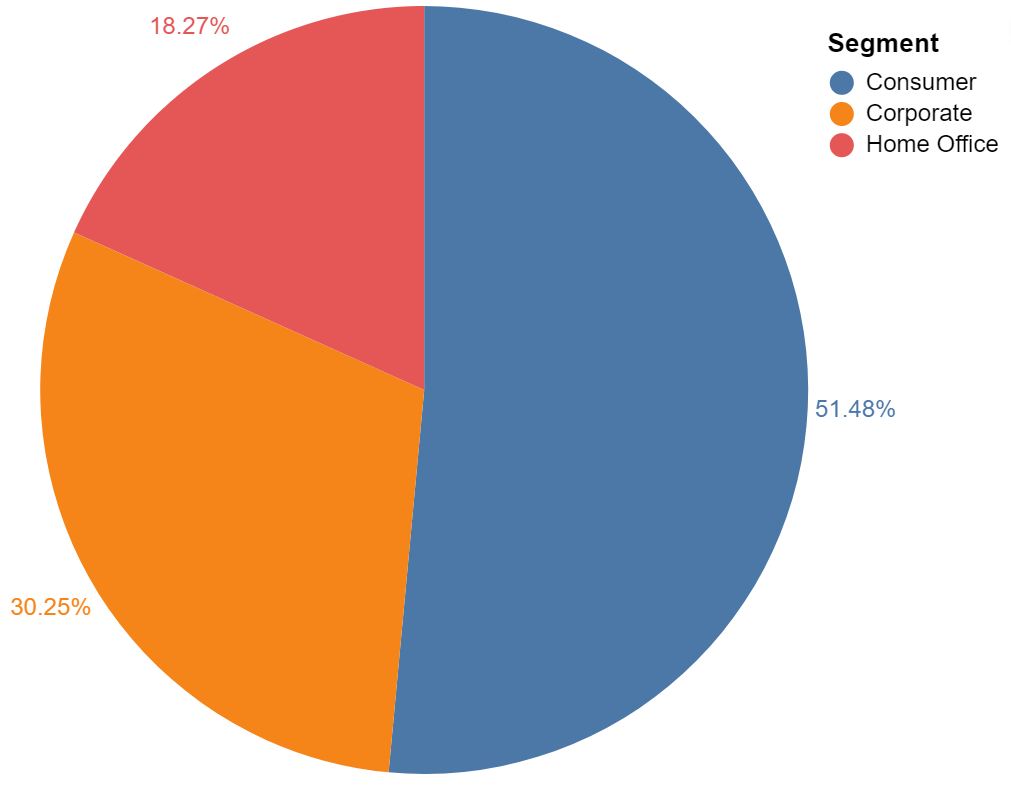


图 4 不同客户类别销售额总和

从上图分析得出，普通消费者占此市场消费额的一半，其余由公司和居家办公者组成。超市可针对普通消费者进行一定市场优化。

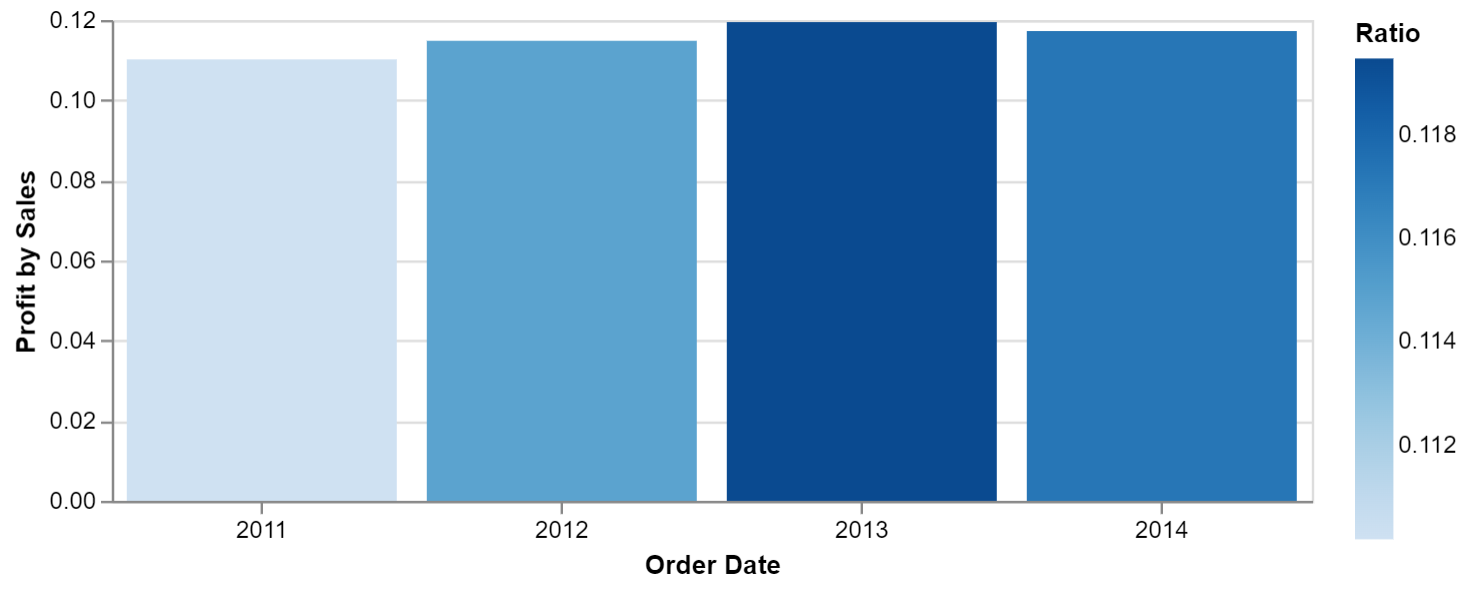


图 5 年利润比销售额走势

从上图可以看出，2011至2013年三年间利润比销售额持续上升，而2014年略有下降。

# 目标1：客户群体细分与营销策略

## **RFM模型**

### 简介

RFM模型是通过一个客户的近期购买行为、购买的总体频率以及花了多少钱3个维度来描述该客户价值状况的客户分类模型。其中R代表最近一次消费距离现在的时间 (Recency)，值越小客户价值越高。F代表某段时间内消费频率次数（Frequency），值越大客户价值越高。M代表某段时间内消费金额 (Monetary)，值越大客户价值越高。

### 模型计算

在此数据集中，时间范围为2011年1月1日至2014年12月31日，所以我们取2015年1月1日为时间的终止点来计算R值，以2011年1月1日至2014年12月31日作为时间范围计算R值和M值。R值使用客户订单总数量，M值使用客户订单总金额（Sales字段）计算。计算采用Customer ID字段作为客户的标识。

一部分计算后的数据如下：

|  |  |  |  |
| --- | --- | --- | --- |
| Customer ID | R | F | M |
| AA-10315 | 9 | 42 | 13747.41300 |
| AA-10375 | 7 | 42 | 5884.19500 |
| AA-10480 | 118 | 38 | 17695.58978 |
| AA-10645 | 27 | 73 | 15343.89070 |
| AA-315 | 3 | 8 | 2243.25600 |

表 1 一部分客户的RFM值

## KMeans聚类分析

得到RFM值后，我们就可以对其进行聚类分析以将客户细分为不同群体。

### 数据归一化

在进行聚类分析任务前，我们通常会对数据进行归一化，以消除数据特征之间的量纲影响，以提高模型的精度。

我们采用最大最小值归一化方法，具体使用sklearn库中preprocessing包中的MinMaxScaler实现。

|  |  |  |  |
| --- | --- | --- | --- |
| Customer ID | R | F | M |
| AA-10315 | 0.007156 | 0.427083 | 0.385302 |
| AA-10375 | 0.005367 | 0.427083 | 0.164803 |
| AA-10480 | 0.104651 | 0.385417 | 0.496016 |
| AA-10645 | 0.023256 | 0.750000 | 0.430070 |
| AA-315 | 0.001789 | 0.072917 | 0.062704 |

表 2 归一化后的一部分RFM值

### 模型训练与评估

我们使用sklearn库中的KMeans类进行模型的构造和训练。Sklearn提供的这个类可以接受数据，并自动进行训练，最后得出簇心，迭代次数和模型分数等数据。

对于模型评估，我们使用簇内误差平方和（within-cluster sum of squared error, SSE）作为评估模型的标准。SSE主要描述每个样本数据到簇心的距离和，SSE越小，表示数据越接近他们所属的簇中心，所以聚类效果越好。此数据可通过sklearn训练出的模型对象的interia\_属性获得。

### 参数搜索

KMeans模型需要一个参数：簇数量K。为了确定最好的K值，我们使用遍历K值对最好的模型进行搜索，并用SSE指标进行评估。

搜索结果如下图所示：

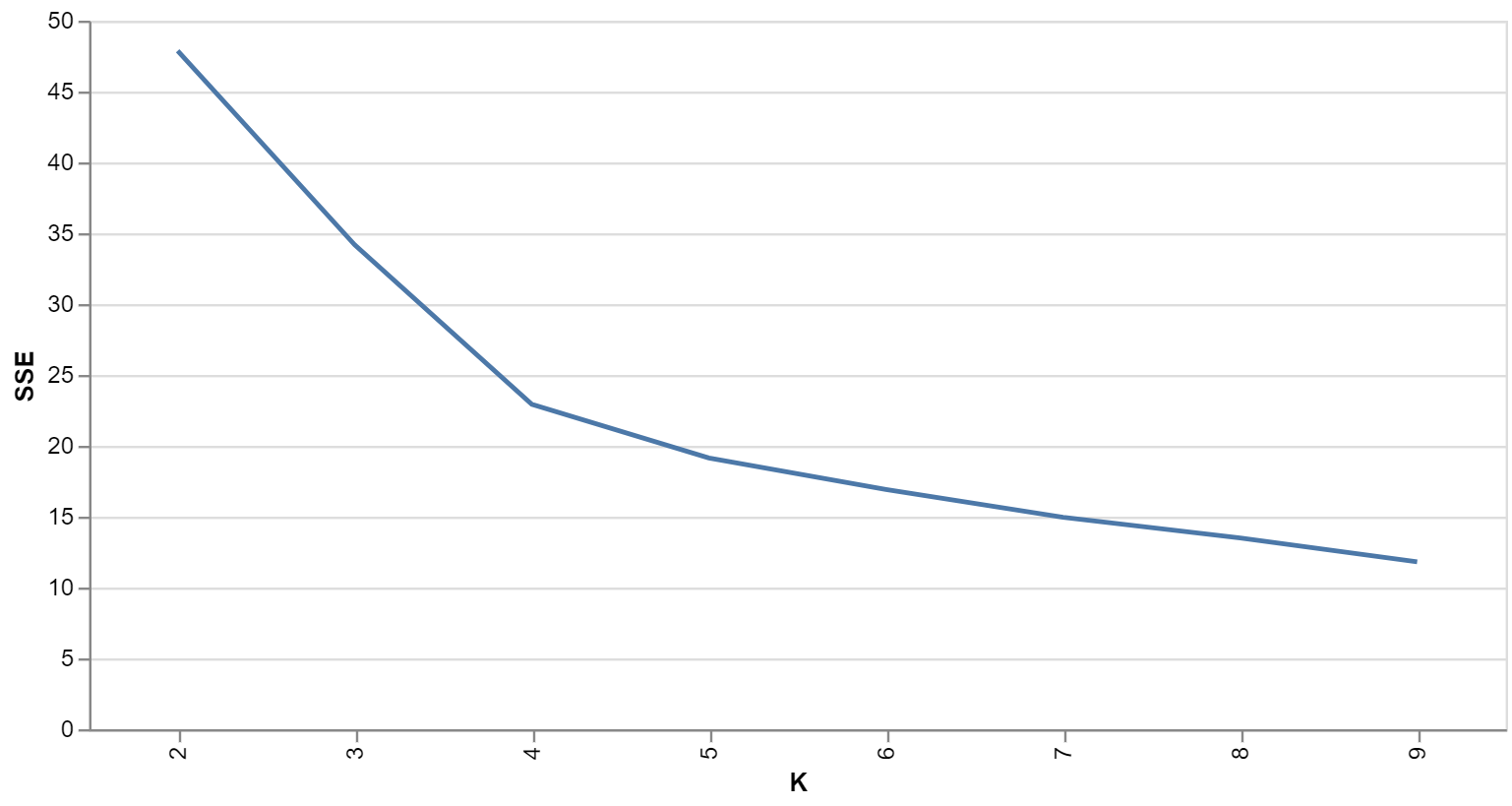


图 6 KMeans模型簇中心数量与SSE的关系

可以看到，在K=4时模型的SSE指标相对较低，且K大于4时SSE减小趋势小，所以我们选取K=4作为KMeans模型的簇中心数量。

### 结果统计

经过KMeans聚类后，我们得到了四个用户群体，接下来对这四个用户群体进行分析统计，并进行数据可视化。为了统计便利，我们按用户占比降序排列，分别标记四个用户群体为A，B，C，D。

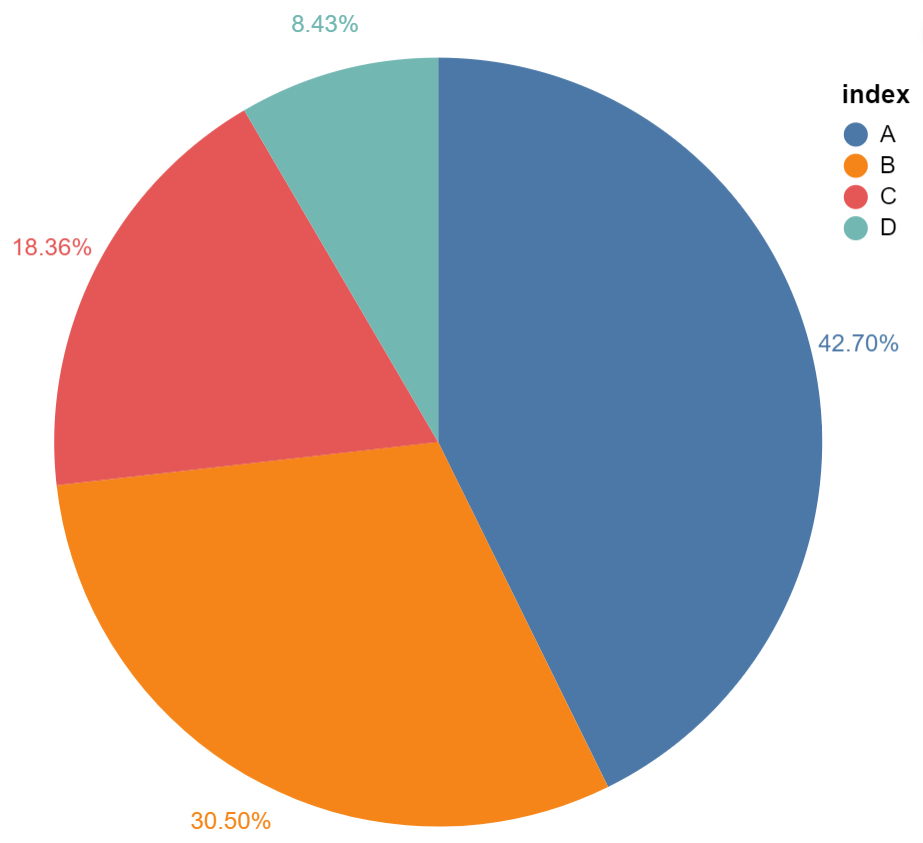


图 7 不同用户群体占比图

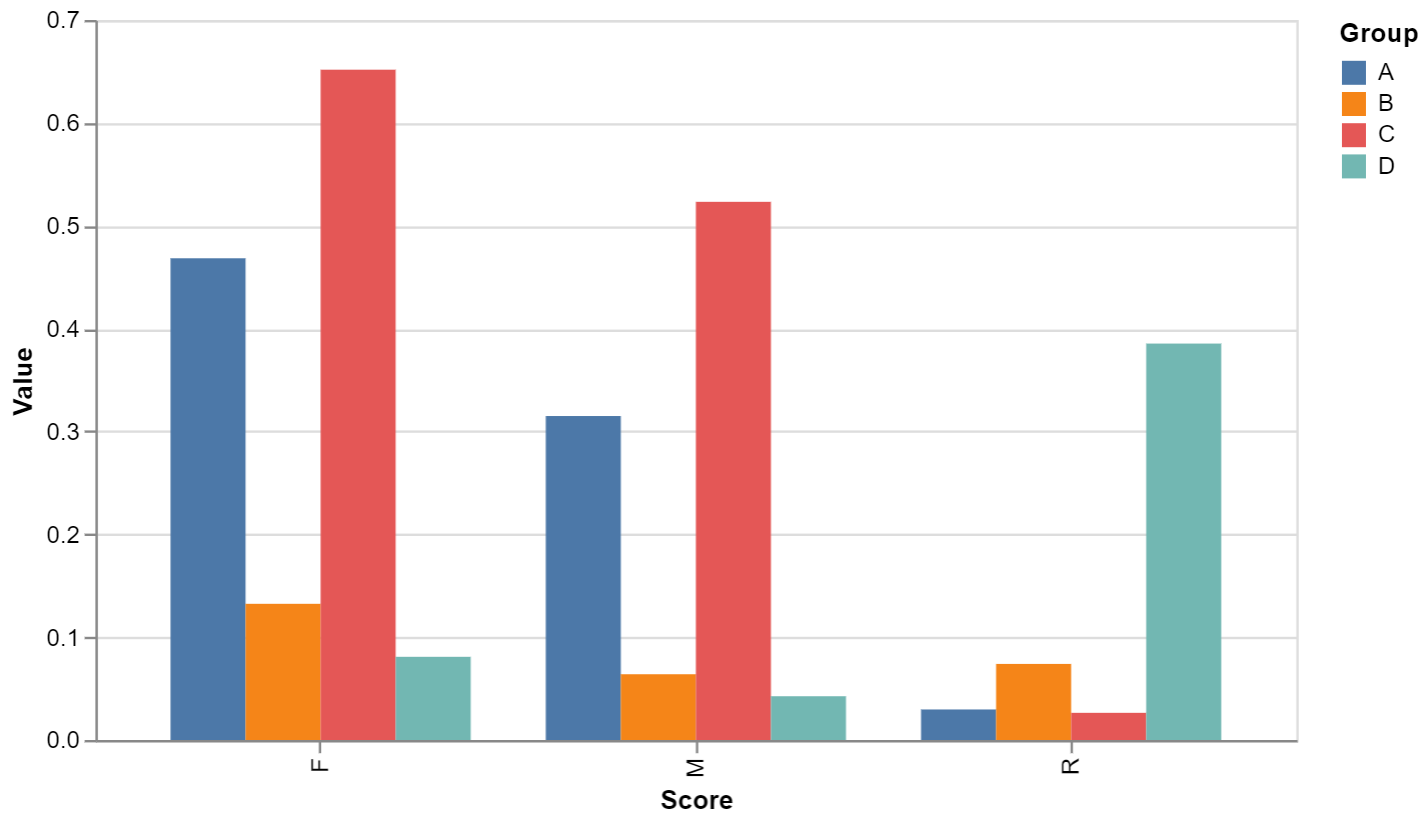


图 8 每组用户群体RFM值图

从上图可以得出：

* A组和C组F值，M值高，R值低，属于重要唤回客户，而且占比最大。
* B组F值，M值，R值都偏低，属于流失客户。
* D组F值，M值低，R值高，属于新用户。

## 营销策略制定

由统计得出，此超市的用户群体主要为重要唤回客户，其次为新客户。

针对重要唤回客户，此超市可以制定如下营销策略：

* **个性化召回优惠：** 根据他们之前的购买记录，发送针对性的优惠券或打折活动，吸引他们重新购买。例如，可以提供他们过去常购买的产品的折扣或“回归专享优惠”。、
* **忠诚度计划升级：** 通过推出更高等级的忠诚度计划、会员福利或VIP服务，给他们更多专属待遇，增加他们的归属感，促使他们再次光顾。
* **定向邮件或短信营销：** 发送个性化的温情营销信息（如“我们很想念您”），让客户感觉到企业的关注和重视。

针对新用户，此超市可以制定如下营销策略：

* **欢迎礼品或首次购买优惠：** 对新用户给予欢迎礼包或首次购买折扣，降低他们的再次购买门槛。
* **简化购买流程：** 为新用户提供简单明了的购买流程，以及便捷的退换货政策，让他们感受到无压力购物体验。

# 目标2：预测未来销售趋势

## 数据趋势分析

为了建立对数据趋势初步的认识，并挑选合适的机器学习模型，我们先将数据集中的销售额趋势进行可视化。为了保证数据一定的平稳性，我们决定以月份为单位，将销售额统计为每月的总和，并预测未来每月内的总销售额。

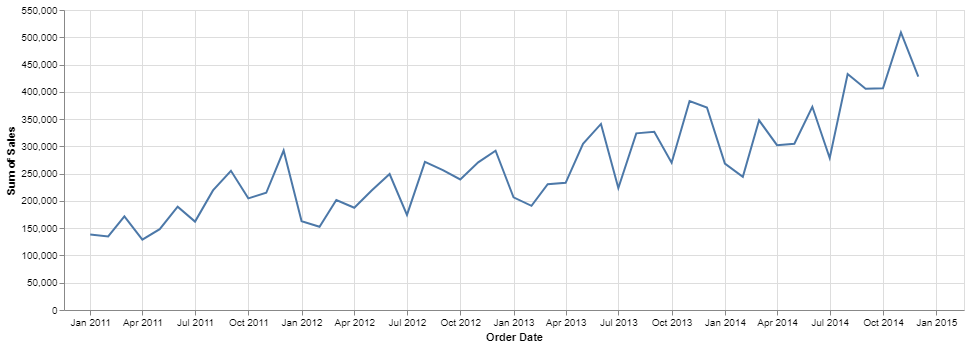


图 9 每月销售额总和趋势图

从此图可以看出，销售额总体呈现上升趋势，并且每年内1月至4月销售额普遍较低，10月至12月销售额普遍较高。由此我们可以假设销售额的趋势具有一定的季节性，可以选择季节性预测机器学习模型对其进行预测。

## 使用SARIMAX模型预测销售额

### 季节性时间序列分解

为了进一步分析数据的季节性，并判断数据的季节性模式，我们使用statsmodels库中的seasonal\_decompose函数对时间序列进行季节性分解。

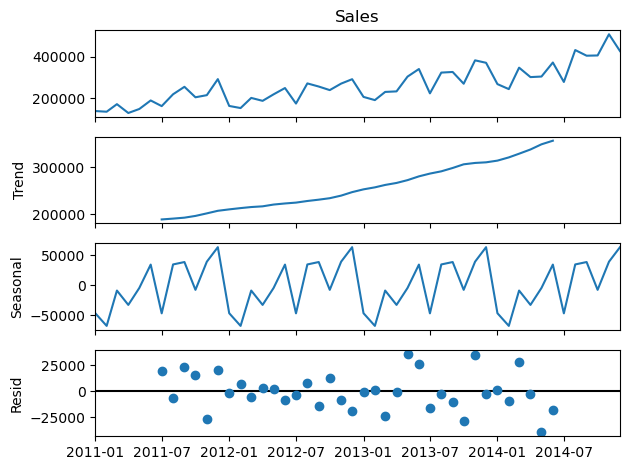


图 10 加法模型

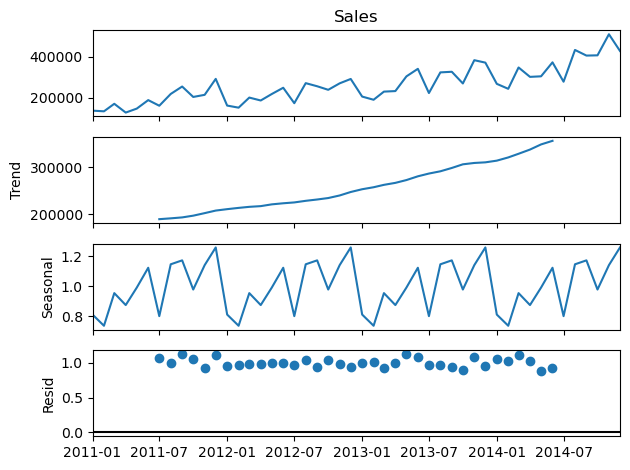


图 11 乘法模型

从得到的残差数据可以看出，加法模型的残差数据具有随机性，而乘法模型的残差数据几乎保持一致。由于残差理论上是指数据中无法进一步解释的随机噪声，所以如果残差是随机的，那么就意味着模型已经捕捉到了数据中的主要模式（趋势和季节性）。所以这里使用加法模型更加合适。

### 模型训练与评估

我们使用statsmodels库进行模型的构建和训练。因为数据集中除了日期外，没有其他与时间相关的数据，所以我们并未用到SARIMAX模型的“X”(即外部因素)。因为销售额是一维数据，所以我们使用平均绝对误差（MAE）来评估模型的质量。MAE值越小，代表预测曲线与原始数据曲线的拟合程度越高。

### 参数网格搜索

SARIMAX模型需要7个参数，非季节参数：p，d，q（非季节自回归的阶数，一步差分的次数，非季节移动平均的阶数），季节参数：P，D，Q，s（季节自回归的阶数，季节差分的次数，季节移动平均的阶数，季节长度）。其中，因为我们按月份进行预测，所以s固定为12。其他六个参数均需要使用网格搜索找出最优。经过学习发现，这六个参数通常范围都为，按此范围进行网格搜索，理论上有种组合。

在对所有组合分别训练得出模型后，我们使用MAE指标评估每个模型，并选出MAE值最小的模型。网格搜索的结果如下表所示：

|  |  |
| --- | --- |
| Parameters | MAE |
|  | 31630.618240 |
|  | 31737.731081 |
|  | 31887.605397 |
|  | 32106.862234 |
|  | 32503.936290 |

表 3 MAE升序参数组合表（前五）

从此表可以看出，为最优的一组参数组合，接下来我们会使用此参数训练最终的模型。

### 网格搜索优化

由于statsmodels库在Windows操作系统上不支持GPU计算，为了使网格搜索速度更快，我们使用了joblib库进行CPU多线程训练。这也使我们拥有进行更大规模的网格搜索的能力。

### 模型结果

我们使用上文网格搜索得到的参数组合进行训练，并预测从2011年1月1日至2018年1月1日每月的销售额总和。预测曲线和真实曲线的对比如下图所示：

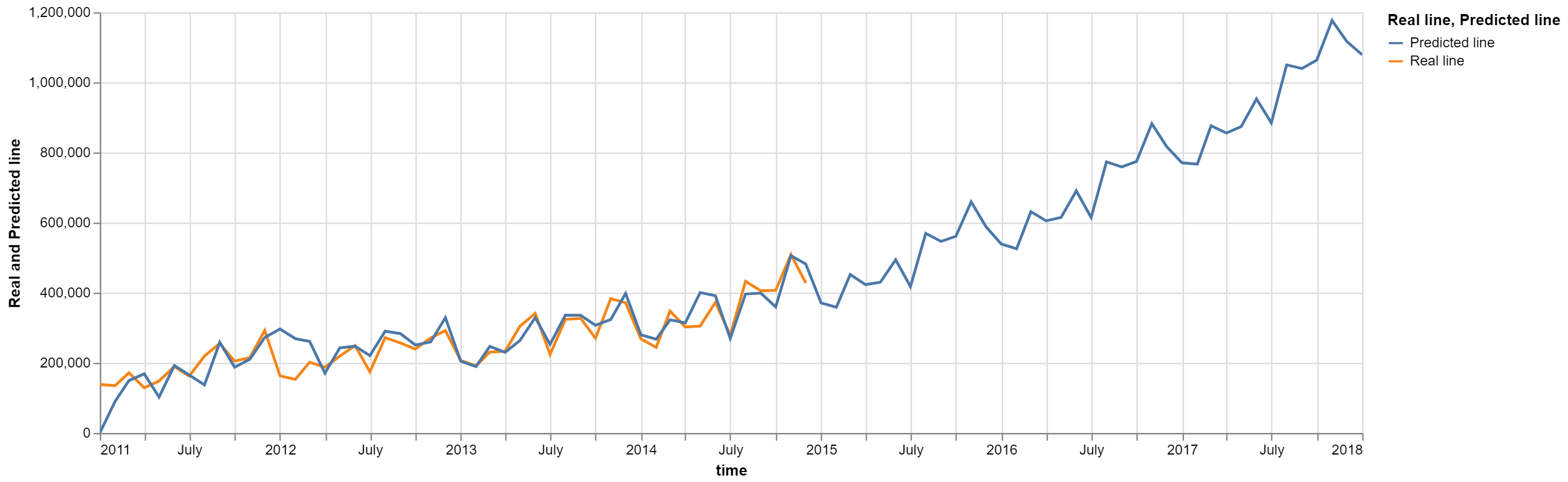
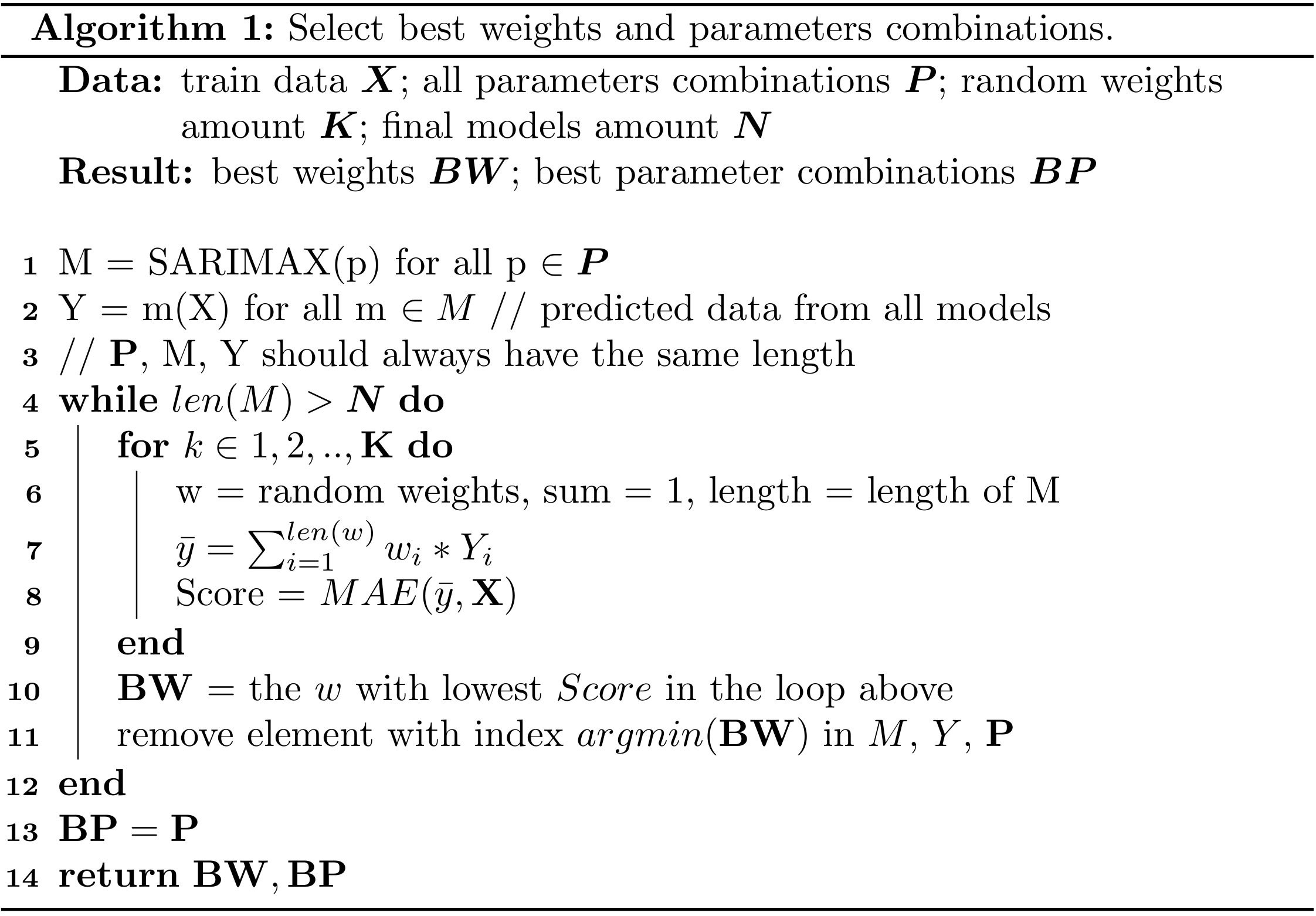


图 12每月销售额真实曲线与预测曲线图

上图中橙线为真实曲线，蓝线为预测曲线。由于模型无法预测起始时的销售量，所以0点处预测值为0，其余预测值均与真实值相似，整体曲线趋势相似。预测结果具有一定的参考价值。

### 模型融合

为了得到更好的拟合曲线，我们尝试将不同参数组合的几个模型的预测结果通过·加权平均融合。经过实践，我们先以模型MAE指标升序排序，选出前10个MAE指标最低的参数组合。接下来进行迭代，随机抽取K次权重并加权平均，选取加权平均后预测曲线MAE值最小的一组权重，找出最小的权重对应的模型并将其抛弃，直到留下N个模型为止。算法的伪代码如下：



我们使N=4, K=2000,最终生成的结果如下：

|  |  |
| --- | --- |
| Parameters | Weights |
| (4,1,3)x(0,1,0,12) | 0.28461255 |
| (4,1,3)x(0,0,2,12) | 0.18215123 |
| (0,0,1)x(0,0,3,12) | 0.34594683 |
| (0,0,4)x(1,0,2,12) | 0.18728939 |

表 4 筛选后的权重与参数组合

MAE指标数据表如下：

|  |  |
| --- | --- |
| Parameters | MAE |
| **Mixed** | **27467.20101502279** |
| (4,1,3)x(0,1,0,12) | 31737.73108055849 |
| (4,1,3)x(0,0,2,12) | 31887.60539736766 |
| (0,0,1)x(0,0,3,12) | 32106.86223401445 |
| (0,0,4)x(1,0,2,12) | 33310.67804699635 |

表 5 融合模型与单模型MAE对比

可以看到，模型融合后的MAE显著小于最好的参数组合的MAE，代表曲线的拟合程度更高。

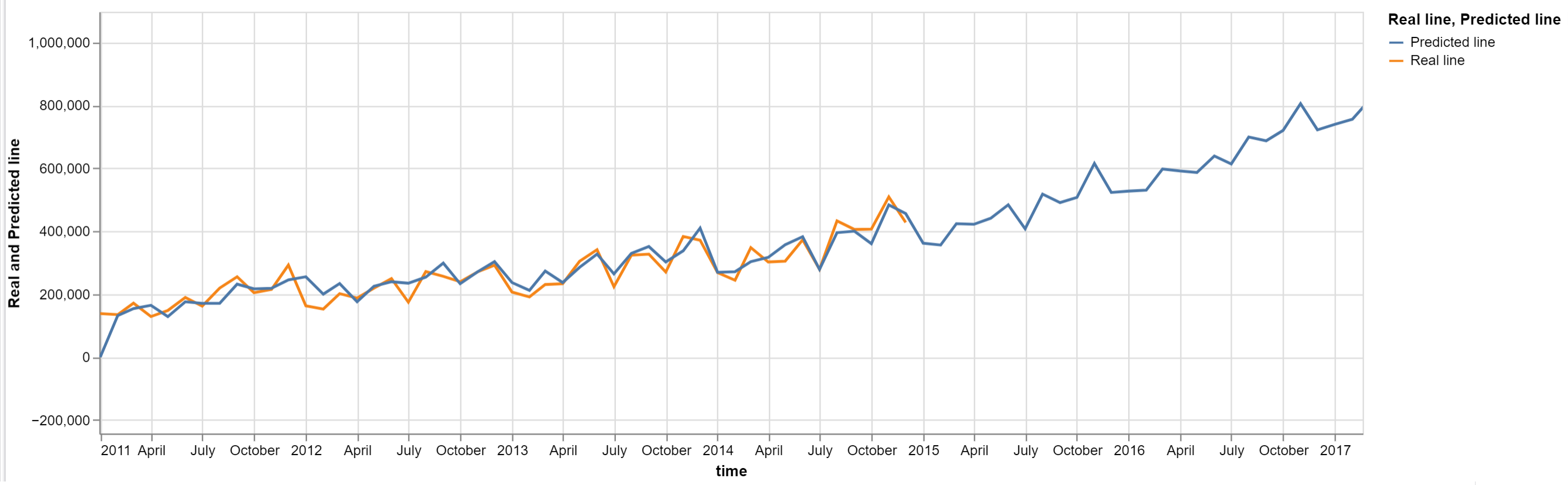


图 13 融合后模型的预测曲线