零售行业销售情况分析与利润率预测

20300750093 于可欣

随着全球经济的不断演变和科技的发展，零售行业作为经济的重要组成部分之一，逐渐成为连接生产者与消费者的重要纽带角色。在当下竞争激烈的环境中，零售商不仅需要应对日益复杂的市场需求，还需要有效地评估商品的利润与其影响因素，从而更好地适应市场需求，将利润最大化。

利润率作为零售业经济状况的关键指标之一，直接关系到商品本身的盈利能力。相比于销售额、销售量，利润率适用于不同性质企业的利润分析，可以应用于到来自不同规模企业的商品分析。同时，不同于销售额和利润数值会随着某个大订单发生剧变的特点，利润率对与已有数据是否是全面的这一事实依赖较小。无论是有多个订单记录还是仅有一个大订单记录，商品的利润率仍然是客观有效的，鲁棒性明显优于其他指标。

然而，传统的利润率分析方法往往无法全面利用已知信息，综合分析影响因素，这促使零售商寻求更为精确、高效的分析手段。在此背景下，机器学习技术成为了分析行业走势和利润率预测的不二之选。

本分析报告将基于已有商品订单记录数据集，首先通过订单基本特征的分析与可视化，深入了解零售行业的销售情况。同时，本报告将基于行业走势呈现出的特征，对商品进行有针对性的特征提取与预处理，并综合所有特征对于特定商品的利润率进行预测。

1. **零售行业的市场偏好分析**

在深入探讨机器学习方法对零售行业利润率的影响之前，本分析报告首先对数据集中未处理的特征进行简要的统计和可视化分析。通过对于数据集整体的粗略分析，本报告旨在了解市场对于零售商品的偏好，进而发掘影响商品利润率的潜在特征，为后续的分析提供基础。

数据集“超市.csv”以订单为单位记录销售数据，每条数据包含商品属性，用户属性，邮寄属性，利润属性四部分特征（具体特征如下表）。其中，商品属性中的商品类别和用户属性中的地域特征、用户类别有着较为明确的语义含义，且为市场偏好的直接反映，所以本部分主要针对这三个特征展开分析。

|  |  |  |
| --- | --- | --- |
| 商品属性 | 制造商 | 品牌名为组合字符串，没有明确语义含义 |
| 产品名 | 产品名为组合字符串，没有明确语义含义 |
| **类别、子类别** | 可反映市场对于类别的偏好，有明确语义含义 |
| 用户属性 | **国家、地区、省/自治区、城市** | 可反映不同地域市场对于商品的偏好，有明确语义含义 |
| **细分** | 可反映不同类别客户对于商品的偏好，有明确语义含义 |
| 客户名称 | 客户名为组合字符串，没有明确语义含义 |
| 邮寄属性 | 邮寄方式 | 为订单产生后的客观事实，由商家决定，无法反映偏好 |
| 订单日期 |
| 发货日期 |
| 利润属性 | 销售额、利润 | 不反映偏好，累加统计有意义 |
| 数量 | 不反映偏好，累加统计有意义 |
| 单价、折扣、利润率 | 不反映偏好，累加统计无意义 |

1. **地域差异导致的利润与成交情况分析**

首先，我们以订单为单位，对于不同地域的成交订单数建立地域分布图（图1）。从图中可以看出，成交情况呈现由西到东的递增、由北向南递增的总体趋势；大部分成交订单分布在华东区域，其中，山东省和广东省的成交数量明显高于其他省份。

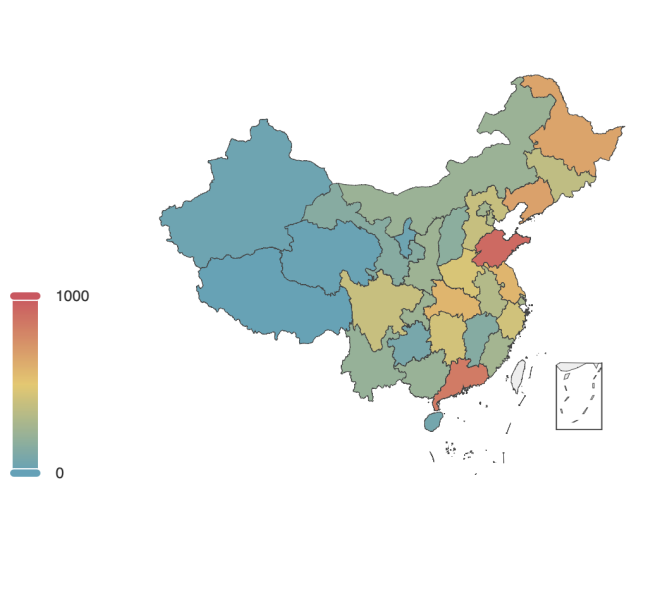


图 1 订单成交情况的地域分析

其次，我们对于具体的销售数量和利润进行分析。首先按照地域划分，对各省的销售利润进行数据统计并做可视化处理（如图2）。在这里选择利润而不是利润率的原因是：利润率的累加没有明显的意义，而利润的累加则能较好的反映地域的盈利情况。从利润图中可以看出，各地区的零售业以盈利为主，其中华东地区的利润明显高于其他省份；同时可以看到辽宁省一直为负盈利状态，且亏损严重。这说明地域对于利润有着较为明显的影响。

基于以上情况，我们结合销售数量和利润，以省为单位，查看二者随时间的变化。从图中不难看出，对于大部分的负盈利省份，其销售数量和利润都呈现局部或全局的负相关关系（以辽宁和浙江为代表），这再一次说明了我们选取利润率预测是更加客观与合理的。

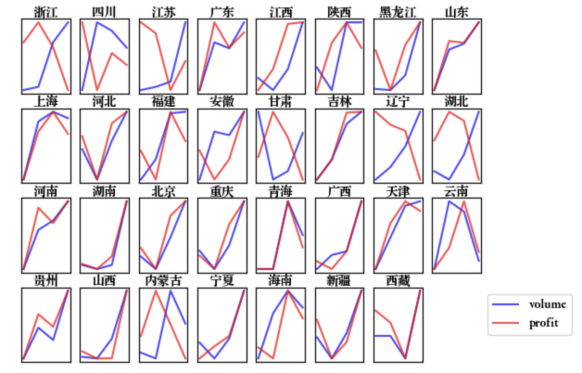
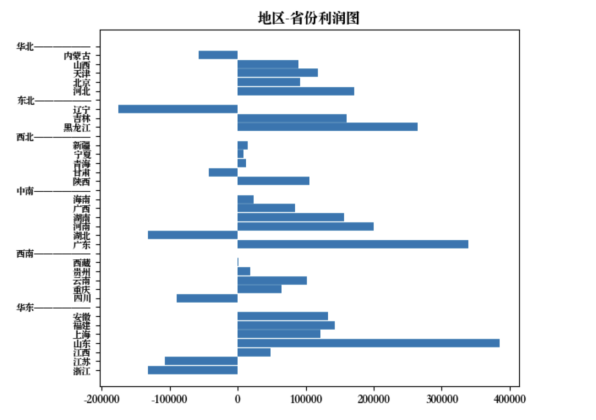


图 2 地区-省份利润图 图 3 省份利润-销量分析

1. **市场对于商品类别的偏好与盈利情况分析**

关于市场对商品类别偏好的分析，我们基于类别和子类别两个特征分别做销售量和利润的累加统计。

从子类别的销售量统计图（如图4）可以看出，市场对于装订机、椅子、收纳具有着高于其他类别的偏好，而对于设备和桌子类的商品则有着较低的需求。在类别的销售统计图（如图5）中可以看出市场对于办公用品有着明显超出其他类别的偏好。

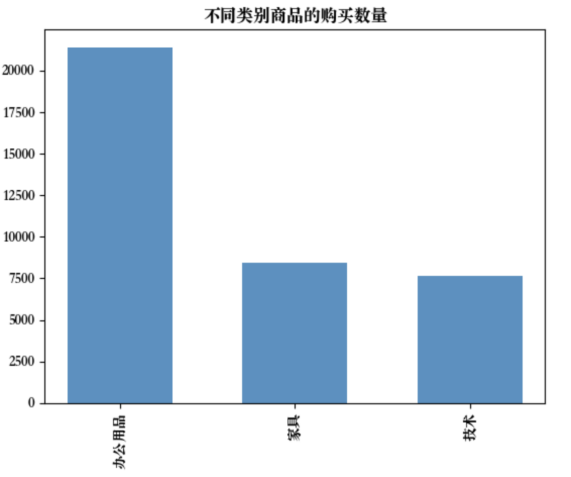
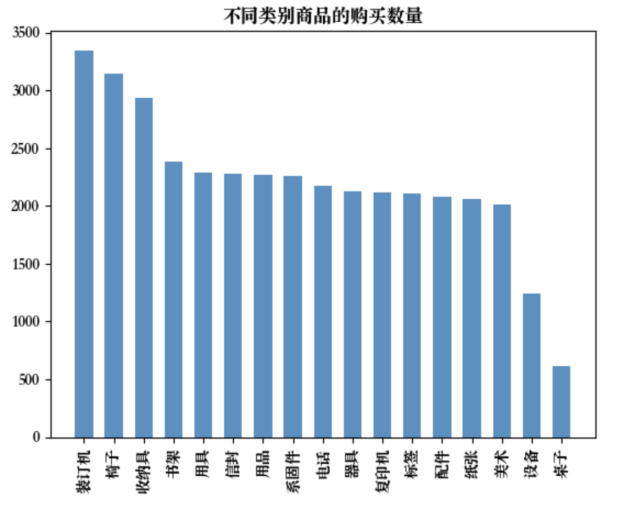


图 4 子类别-销售量分布统计图 图 5 类别-销售量分布统计图

在进行销量的分析后，我们继续分析市场偏好对盈利带来的影响。首先，各类商品的盈利是有区别的，这说明商品的利润会受到其所属类别的影响。其次，除去少部分销量少但盈利高的商品，大部分商品的盈利会受到其销量的影响，然而这不是简单的线性相关的关系：从子类别-利润图（如图6）中可以看出虽然市场对装订机的需求大，但其盈利不多；相反，设备类商品虽然需求不大，但其盈利却较高；同样的，技术类虽然有着相对低的销量，但其盈利却和办公用品类基本持平。

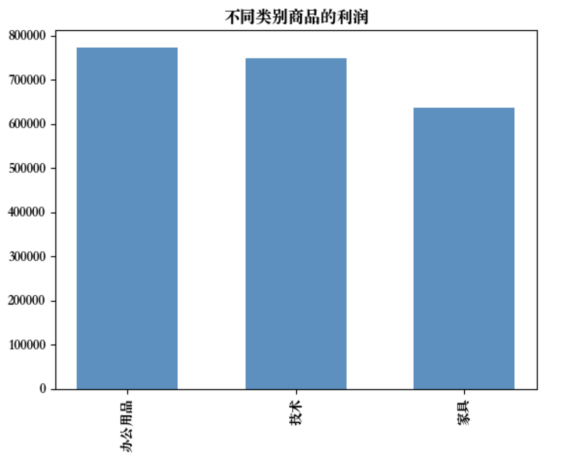
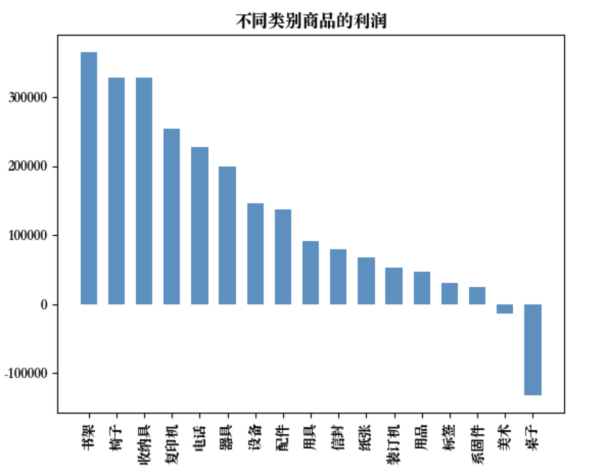


图 6 子类别-利润分布统计图 图 7 类别-利润分布统计图

1. **主要客户类型分析**

客户类型决定了市场的组成，对商品销售有着决定性作用。基于此特性，我们以订单量为统计目标，结合前两部分分析，分地域和类别对不同类别客户在市场的参与度进行统计分析。

从不同类别客户的两张分布图（图8与图9）上可以看出，无论是按地域统计，还是按类别统计，消费者和公司都是零售行业的主要售出对象。

在按地域分析（图8）中，消费者在各个省份都是市场的主要组成部分，部分省份公司的占比和消费者持平。而在按类别分析（图9）中，消费者和公司则为市场的主要组成部分，消费者和公司的占比基本持平。

以上分析说明，商品的主要面向客户类型将较大的影响商品的销售情况，进而影响到商品的利润，如果将地域性特征、商品类别、商品主要面向对象结合进行分析，可能会产生更好的利润预测结果。

1. **小结**

综上所述，我们可以对零售行业的现状进行简单总结：零售行业的主要消费群体位于华东地区，广东省和山东省为消费大省；主要客户类型为普通消费者，公司类型的客户占比较大。零售行业的商品销量和利润受商品类别、销售地区、主要面向客户影响；同时，利润会受到销售量的正向影响，但二者不是简单的线性关系。

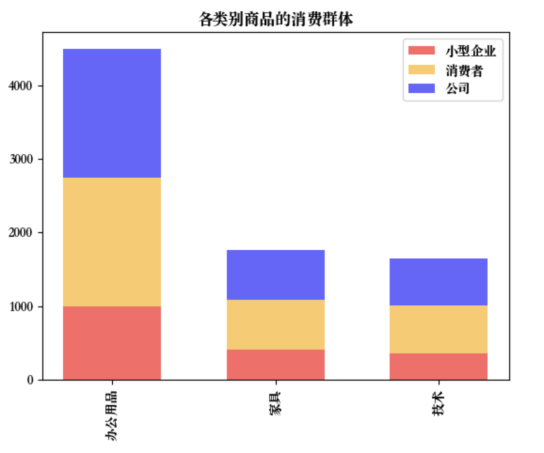
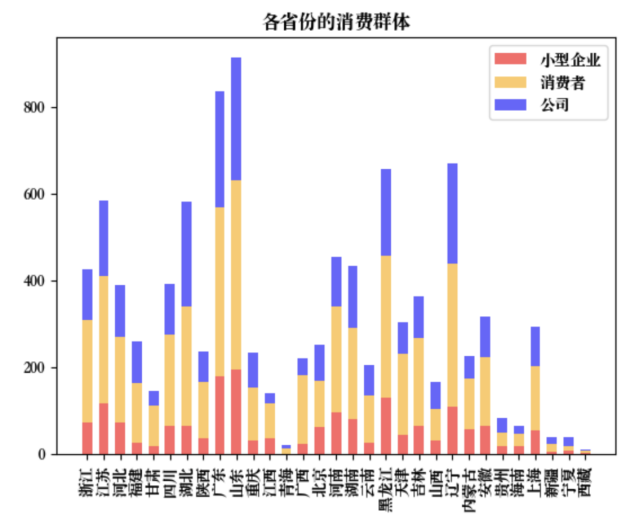


图 8 各省份的消费群体分布 图 9 各类别商品的消费群体分布

1. **特征选取与构建**

在选取特征之前，我们需要首先明确预测面向的对象。

毫无疑问地，我们的预测主要受益人将是商家。所以我们可以假设以下情景：一个不了解行业内情的商家希望通过提供商品特征，来预测他售卖的某样商品是否能获利，并根据不同特征的重要性来调整自己的销售策略，从而实现产品利润率最大化。所以，本分析报告以商品作为分析主体（以商品名区分商品），并以商品为主要面向对象进行特征提取与利润率预测。

从第一部分的分析我们可以看出，商品利润率的影响因素是多种多样的，它不仅由商品自身的属性决定，还受到零售行业主要面向的客户特征影响。同时，对于第一部分没有分析的非市场因素的特征，可能也对商品的利润有着一定的影响。综合以上特点，我们分为商品属性，用户属性，邮寄属性，利润属性四部分简要说明特征选取与构造。

对于商品属性，可选取的特征分为制造商、产品名、类别、子类别四部分。首先，产品名为区分商品的标识符，不能作为特征使用。其余三个根据第一部分的分析，都可以作为特征，但因为其值为字符串，所以对其进行编码处理（具体代码如下）。

其中，因为类别的利润在第一部分的分析中，是有排名先后的，所以对其按照排名编码。对于子类别，因为值较多，且利润排名中差距不大，所以采用顺序编码。对于制造商，直接采用顺序编码。

|  |
| --- |
| category\_dict = {'办公用品': 1, '技术':2, '家具': 3}  sub\_category\_dict = {}  item\_set = set()  for item in raw\_df['子类别']:  if item not in item\_set:  sub\_category\_dict[item] = len(item\_set) + 1  item\_set.add(item)  maker\_dict = {}  item\_set = set()  for item in raw\_df['制造商']:  if item not in item\_set:  maker\_dict[item] = len(item\_set) + 1  item\_set.add(item) |

对于用户属性，可选取的特征有用户类型，用户地理位置的三维特征（地区、城市、省/自治区），然而对于同一商品，因为数据集是以订单为单位，而每个订单的用户特征是不同的，所以我们需要对此做统计与编码处理。

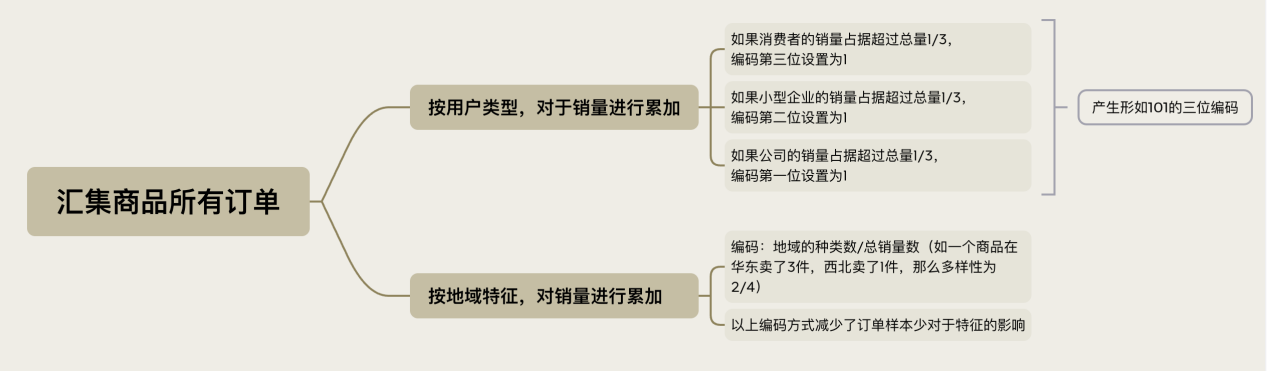
对于用户类型特征，可以通过统计不同类型消费者购买的数量来定义商品的主要客户。考虑到商品存在多类主要客户，对商品的所有订单进行如图流程处理（图 10）。然而在提取地域特征时，上述的编码方式显然不适用于地区数较多的省份和城市特征，所以相比于采用主要面向地域这一特征，我们从多样性考虑地域特征的提取，具体流程如图10所示。

图 10 用户类型和地域特征的编码处理流程图

对于邮寄属性，可选取的特征有邮寄方式、订单时间、发货时间。邮寄方式因类别较少，只需进行简单编码即可。但订单时间和发货时间的时间戳本身显然是没有意义的，我们可以对这两个特征进行相减，形成邮寄等待时间特征。

对于利润属性，利润率将作为我们预测的目标；销售额、数量、利润是商品销售后才产生的数据，然而我们的预测显然是发生在售卖之前，所以这些特征对于我们的预测没有意义，全部舍弃；单价、折扣会随着订单的变化而变化，在这里以订单销售量为权重，取加权平均作为商品这两个特征的最终值。

在进行以上处理后我们可以得到1952个商品的11个特征表（如图11）

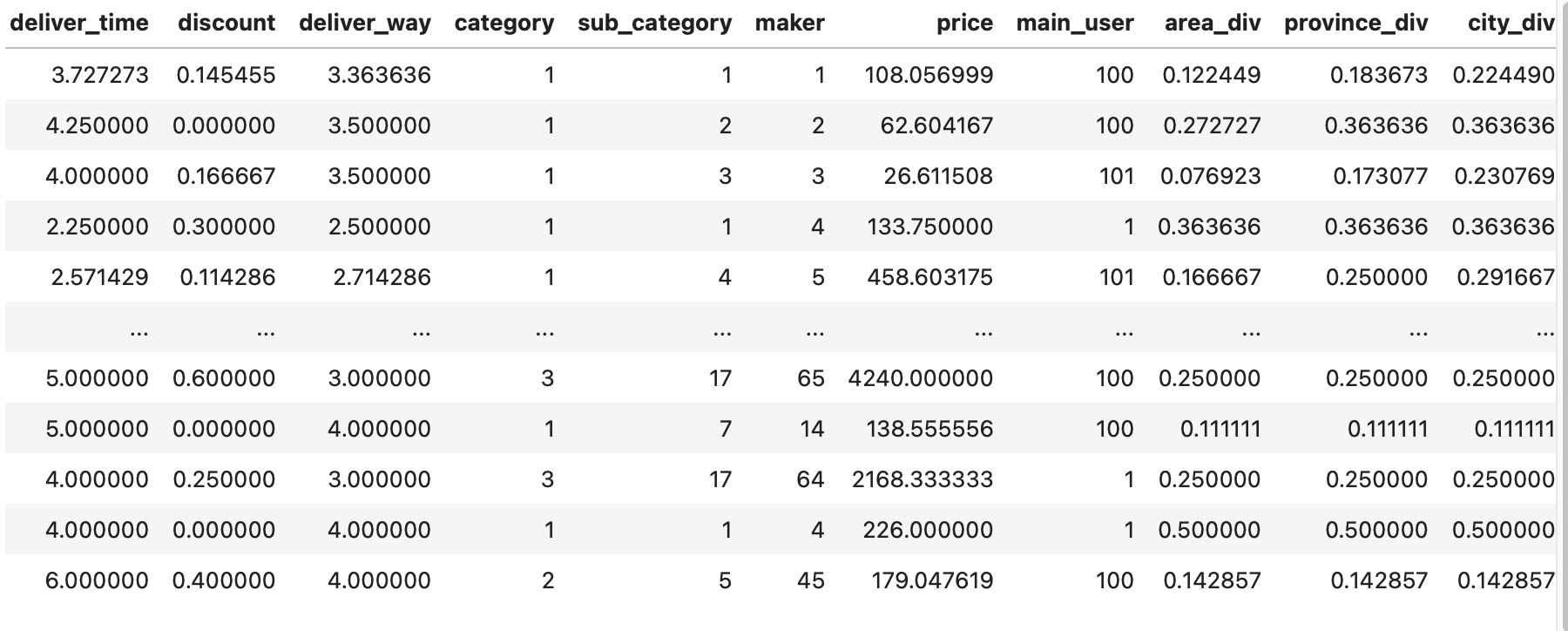


图 11 商品特征提取后的数据表

1. **特征预处理**
2. **缺失值检查**

在数据分析的初步阶段，了解数据中的缺失情况是至关重要的。这一部分处理使用 msno 库生成缺失值矩阵并将其可视化。这种可视化方式使我们能够迅速识别哪些特征存在缺失值，并可以观察到缺失值的分布模式。我们可以通过缺失值来确认之前编码方式是否适用于数据集。

|  |
| --- |
| df = pd.DataFrame(data=get\_df(raw\_df))  msno.matrix(df)  plt.show() |

如图12所示，在使用第二部分阐述的特征编码和处理方式后，特征数据仍然没有缺失的部分，说明特征数据完整，编码预处理没有问题。

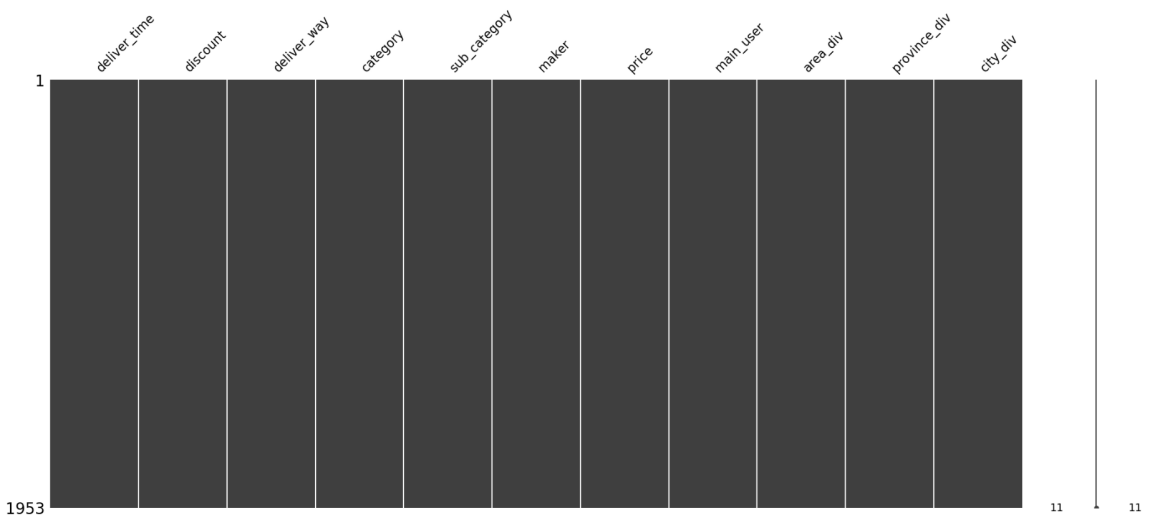


图 12 特征缺失值矩阵

1. **异常值检查**

在验证数据没有缺失后，应该对于数据进行异常值检查。针对不同特征的数值绘制箱线图与数值分布，以检验是否有异常值。

|  |
| --- |
| fig, axes = plt.subplots(nrows=3, ncols=4, figsize=(15, 10))  for i, column in enumerate(df.columns):  sns.boxplot(x=df[column], ax=axes[i // 4, i % 4])  axes[i // 4, i % 4].set\_title(column)  plt.tight\_layout()  plt.show()  fig, axes = plt.subplots(nrows=3, ncols=4, figsize=(15, 8))  for i, column in enumerate(df.columns):  row, col = divmod(i, 4)  axes[row, col].hist(df[column], bins=10, alpha=0.5)  axes[row, col].set\_title(column)  plt.tight\_layout()  plt.show() |

如图13所示，可以发现discount（折扣）、price（单价）、deliver\_way（邮寄方式）、area\_div（地区多样性）、province\_div（省多样性）、city\_div（城市多样性）有着较为长的单侧尾须，结合图14的分布可以发现原因是这些特征的分布是有偏向的，不是异常值导致的。

对于deliver\_time（运送时间），有两侧的尾部，不难发现左侧的尾部可能是因为存在较多的0值。对于运送时间，0的取值表示当日送达，这一取值貌似是不合理的。但由于数据集没有提供更加细化的时间戳，所以暂时仍然采用这样的计算方式来保持特征计算的一致性。对于右侧的尾部，结合分布来看，可能是存在厚尾正态分布情况导致的，在分布中并没有显示出明显的异常值，所以不做进一步的异常值处理。

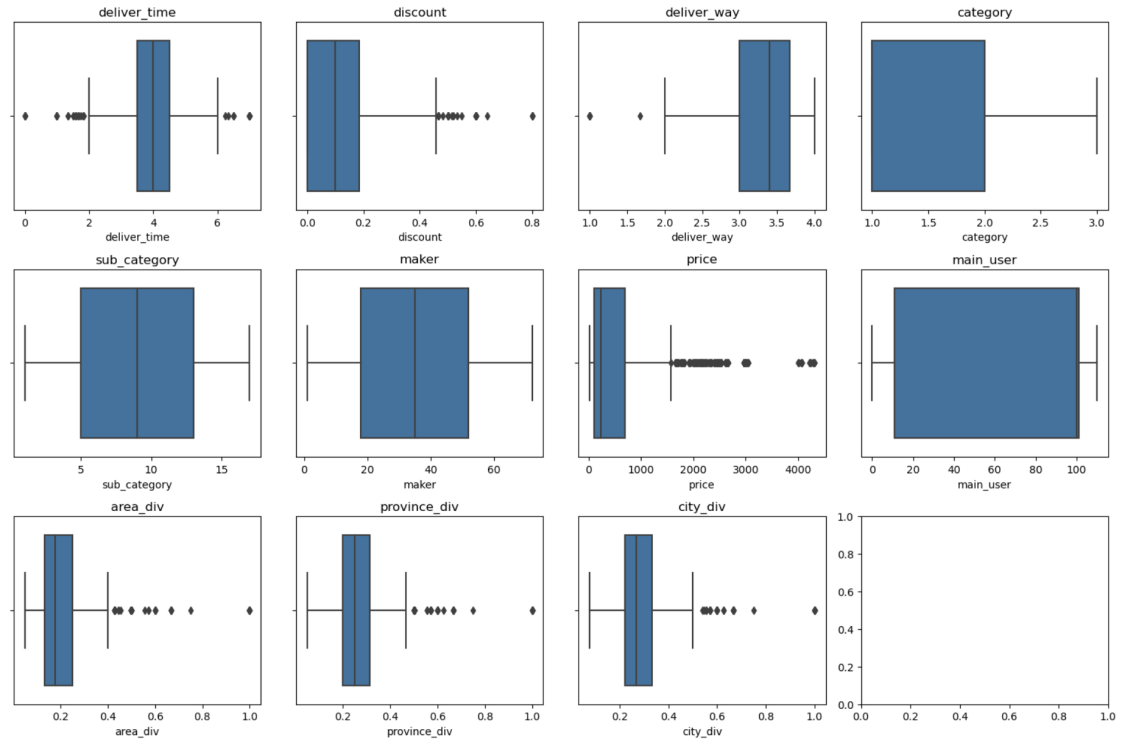


图 13 各特征的箱线图

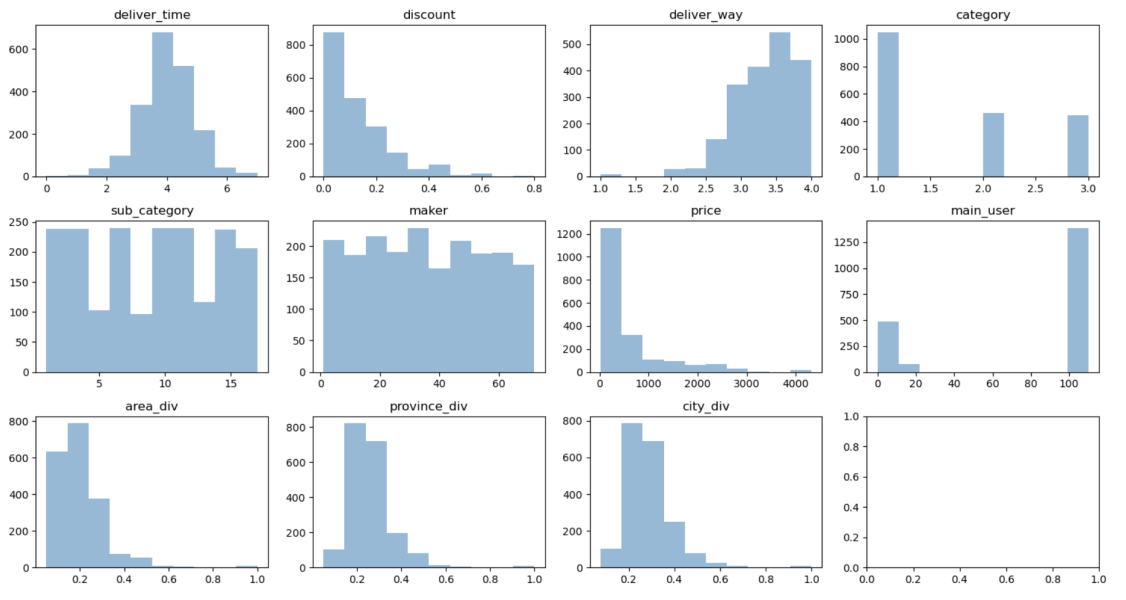


图 14 各特征的分布图

1. **特征相关性**

在确认特征提取的数值无误后，对于不同特征之间的相关性进行热度图绘制，以检查特征之间是否有强相关性，方便去除部分相似的特征，进而提升预测模型的性能。

|  |
| --- |
| correlation\_matrix = df.corr()  plt.figure(figsize=(10, 8))  sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm', fmt=".2f", linewidths=.5)  plt.title('Correlation Heatmap of Features')  plt.show() |

如图15所示，以0.7为高相关度的判定标准，可以看到deliver\_way（运送方式）与deliver\_time（运送时间）有着较强的相关性；同时，地域的三维特征之间都有着较高的相似度。对于这两组具有高相关性的特征，将根据后续的Spearman检验，选择与预测值相关性更高的一个特征作为最终特征。

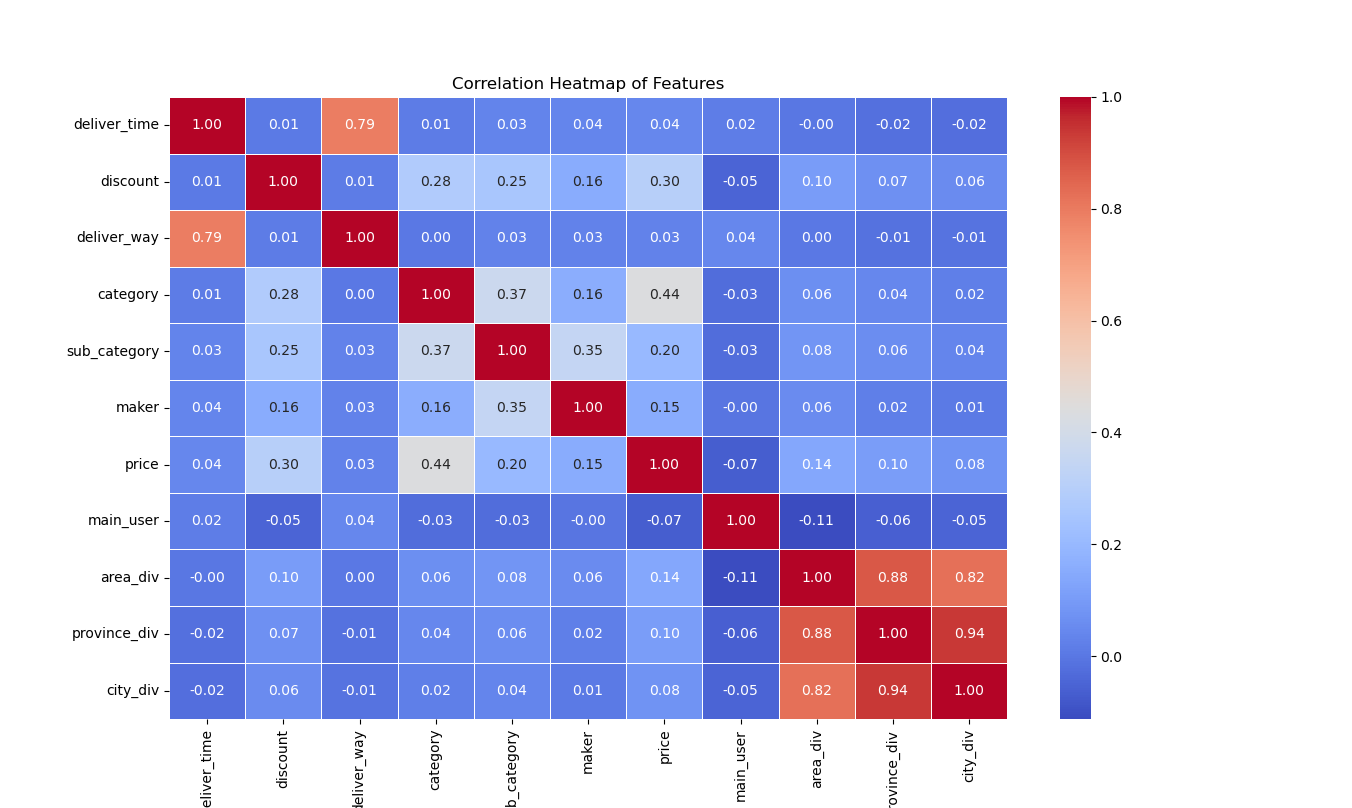


图 15 特征热度图

1. **连续特征离散化**

由于之前的特征提取存在加权平均的处理，所以部分特征的分布并不是离散的，且部分特征取值（如多样性和折扣）的变化比较细微。考虑到离散的特征有助于简化模型与提升模型的泛化能力，同时更好地适应新的数据，对于所有连续型特征进行Kmeans聚类。

考虑到不同特征的n\_cluster最优取值可能不同，所以对所有的非离散特征进行n\_cluster=3-8的聚类测试，并以silhouette\_score为标准观察聚类效果。

|  |
| --- |
| for n in range(3,9,1):  n\_clusters= n  labels.extend([i for i in range(n\_clusters)])  fig,axs=plt.subplots(nrows=2,ncols=4,figsize=(12,12))  for i in range(7):  row=i//4  col=i%4  ax=axs[row][col]  param=param\_distribute\_list[i]  temp=np.array(df[param].values.reshape(-1, 1)).reshape(-1,1)  kmeans=KMeans(n\_clusters=n\_clusters).fit(temp)  result=kmeans.predict(temp)  kmeans\_result[str(n)].append(result)  silhouette\_avg = silhouette\_score(temp, result)  score[param]['n\_cluster='+str(n)] = silhouette\_avg  counts=np.bincount(kmeans.labels\_)  centers=kmeans.cluster\_centers\_  boundary[param]=[item[0] for item in centers]  boundary[param].sort()  total=sum(counts)  row=[param,total] |

根据图16选出silhouette\_score最高的n\_cluster参数，并根据分类结果绘制各个特征聚类后的分布图。

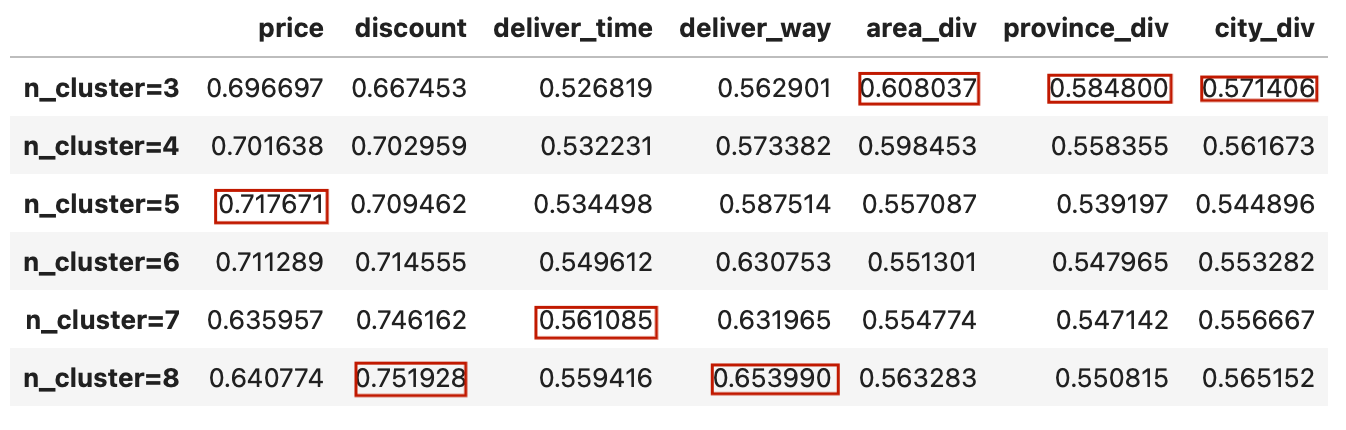


图 16 不同n\_cluster下的聚类效果

根据分布图可以看出，deliver\_time，deliver\_way，discount都有着超过1类的数据在聚类后的分布有着明显的脱离中心的离散点，且通过观察discout的silhouette\_score，可发现其有着相对明显的过拟合倾向，采用聚类可能会对特征蕴含的信息造成损失，故对于这几个特征的聚类不予以采纳。

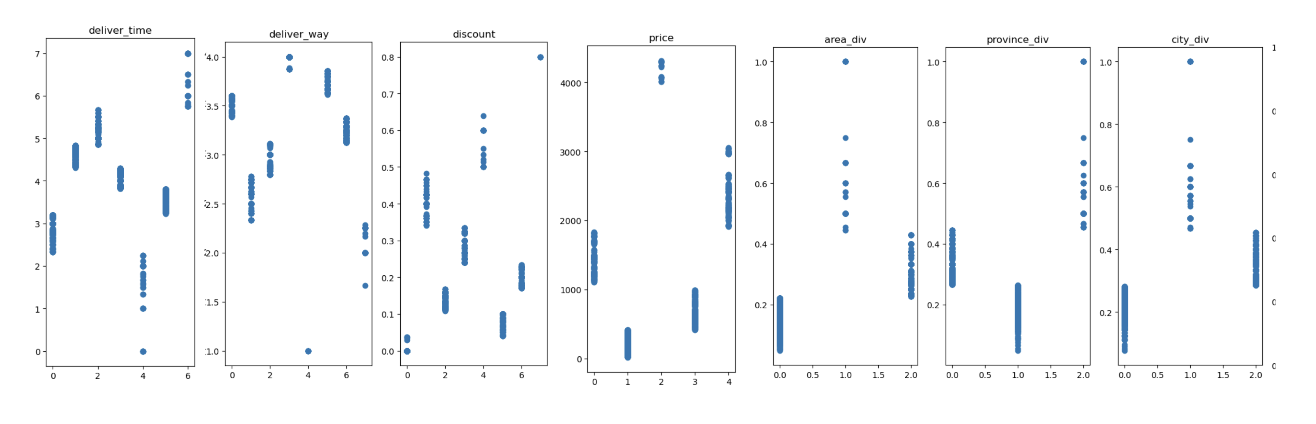


图 17 连续型特征的Kmeans聚类结果可视化

1. **特征与标签相关性检验**

根据图14的分布可以发现，大部分特征的分布是不符合正态的，故采用Spearman检验来验证各个特征和标签之间的相关关系。通过分析相关性检验，将对之前分析的具有高相似度的特征组中的特征进行最终选拔，并同时过滤部分p-value超过0.5的特征。

|  |
| --- |
| table = PrettyTable()  table.field\_names = ["Feature", "Spearman Correlation", "P-value"]  label = processed\_dict['profit\_money']  for feature\_name, feature\_values in processed\_dict.items():  if feature\_name != 'profit\_money' and feature\_name !='profit':  correlation, p\_value = spearmanr(processed\_dict[feature\_name], label)  table.add\_row([feature\_name, correlation, p\_value])  print(table) |

如图17所示，对于（deliver\_time，deliver\_way）特征组，deliver\_way的p-value明显低于deliver\_time，且其p-value低于0.5，满足筛选条件，所以选择deliver\_way作为最终特征之一。对于（area\_div, province\_div, city\_div），area\_div明显高于其余两者，所以选择area\_div作为最终特征之一。

在表中除了以上被筛选掉的特征，main\_user的p-value是唯一高于0.5的，根据标准，main\_user将不会用于最终的预测特征。

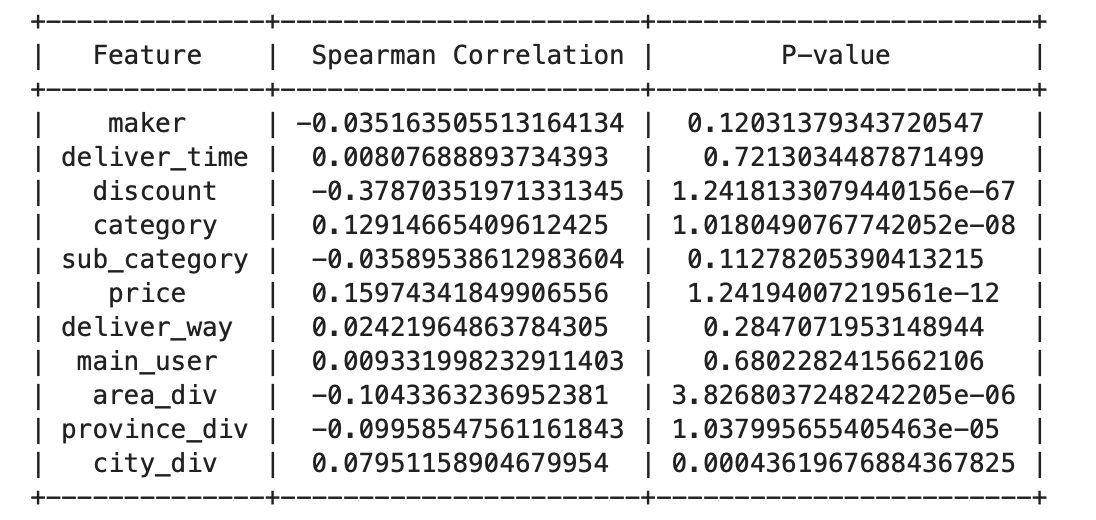


图 17 各特征的spearman相关性检验

1. **小结**

根据以上预处理，最终选择7个特征用于预测，且部分连续型特征做了合理的离散化处理，最终结果如图18所示。

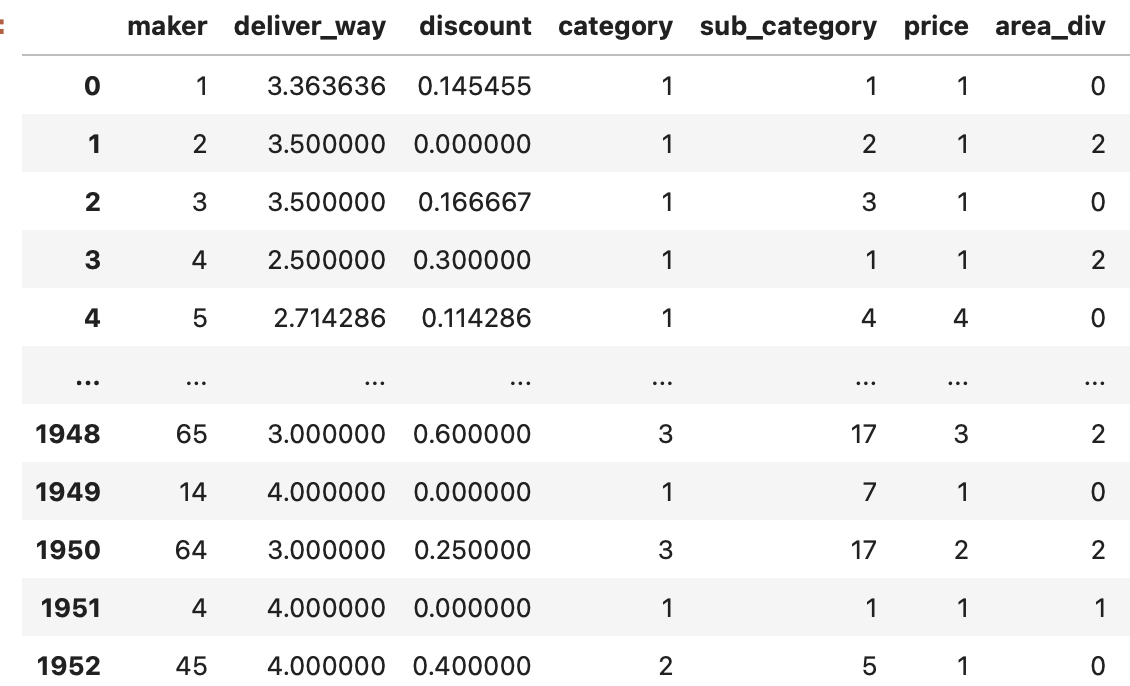


图 18 最终预处理结果图表

1. **回归预测**

因为预测目标是具有实际含义的加权利润率，为连续型变量，所以使用回归对利润率进行预测。对于回归模型的选取，本分析采用了多个常见的模型以寻求最适合数据集的模型，具体采用的模型有：线性回归, 随机森林回归, 决策树回归, 神经网络回归，梯度提升树回归, Ridge回归，Lasso回归，Elastic Net回归。

考虑到不同特征之间的数值差异较大，为了防止这一特征对于模型的影响，在预测前首先对特征进行标准化处理。在划分训练与测试集时，按照8:2的比例对数据进行划分。

|  |
| --- |
| profit\_feature, profit\_label, profit\_key\_name = get\_feature\_label(data\_df)  scaler = StandardScaler()  profit\_feature\_normalized = scaler.fit\_transform(profit\_feature)  X = np.array(profit\_feature\_normalized)  y = np.array(profit\_label)  X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42) |

回归模型评价指标有多种，例如平均绝对误差（MAE）、均方差（MSE）、均方根差（RMSE）、R2等。本次分析使用均方根差和R2作为评价指标，其中均方根差是平均拟合误差，R2反映了回归模型的拟合程度。

同时，为了确保模型的鲁棒性和准确度，同时防止过拟合问题，本案例在测试时，使用了交叉验证的方法，限于计算能力与样本数目，设置了5折交叉验证，并针对每个模型选择了不同的超参数进行网格搜索，以确保选出最佳的参数与模型。具体参数设置可看代码部分，此处挑选梯度提升树为例。

|  |
| --- |
| # 梯度提升树  gb\_model = GradientBoostingRegressor()  param\_grid\_gb = {'n\_estimators': [10, 20, 50],  'learning\_rate': [0.001, 0.01, 0.1],  'max\_depth': [3, 5, 7],  'min\_samples\_split': [2, 5, 10],  'min\_samples\_leaf': [1, 2, 4]}  grid\_search\_gb = GridSearchCV(estimator=gb\_model, param\_grid=param\_grid\_gb, scoring=scorer, cv=kf)  grid\_search\_gb.fit(X\_train, y\_train)  gb\_best\_model = grid\_search\_gb.best\_estimator\_  y\_pred = gb\_best\_model.predict(X\_test)  gb\_mse = mean\_squared\_error(y\_test, y\_pred)  gb\_r2 = r2\_score(y\_test, y\_pred) |

预测结果对比如图20与图21所示。

首先可以看到预测结果的rmse在0.2左右，最低值（梯度提升树）为0.194，r2值全部高于0.5，最高值（梯度提升树）为0.702。参考利润率的分布（如图20）可以发现，对于一个值分布主要在-1.5-0.5的预测目标来说，本次预测的rmse相对较低，模型效果较好。

在多个模型中，梯度提升树在rmse和r2的指标表现都为最佳，所以最终选择梯度提升树作为本次分析报告的预测模型。

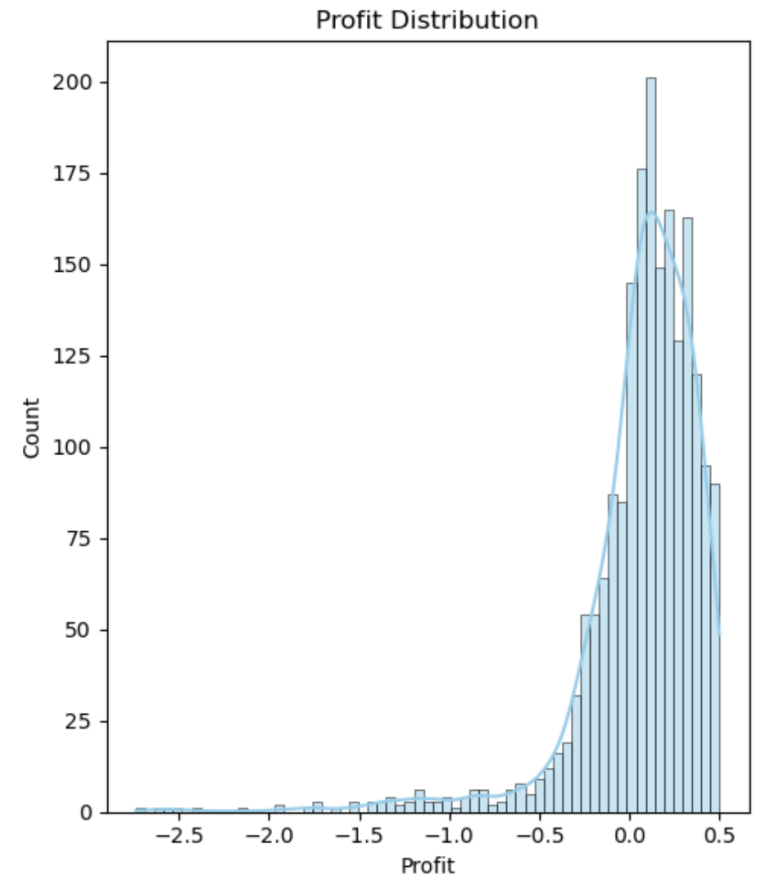


图 19 利润率分布图

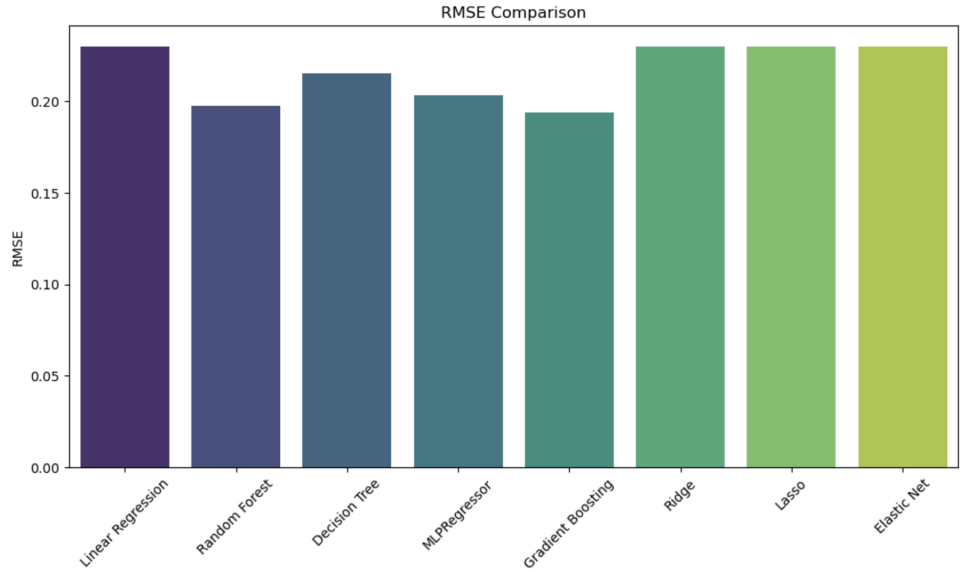


图 20 不同模型的RMSE对比

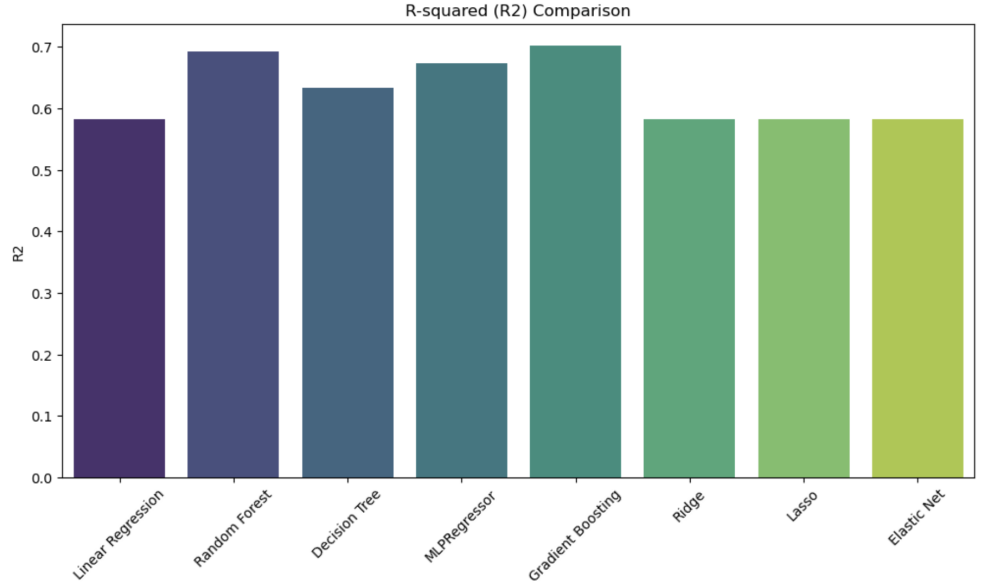


图 21 不同模型的R2对比

在选定模型后，我们查看最终选定的预测模型的特征重要性进行考察。

|  |
| --- |
| importance\_df = {'feature':profit\_key\_name, 'importance':gb\_best\_model.feature\_importances\_}  plt.figure(figsize=(10, 6))  sns.barplot(x='importance', y='feature', data=importance\_df,，orient='h',palette='viridis')  plt.title('Feature Importance') |

如图22所示，商品的折扣在模型中起着决定性作用，在特征重要性中占比接近0.8，其次商品的种类也起着较大的影响。

对于类别的影响，在第一部分已经做过可视化分析，在此不再赘述。对于折扣的影响，在第二部分的Spearman相关性（图17）分析处可以看到，折扣和利润率呈现负向关系。这说明虽然折扣在零售行业可以起到促进售出的作用，但其对商家的盈利影响大概率会呈现负向影响。如果商家想通过打折来获取更高的利润，那么商家需要对这种策略进行进一步的市场评估与考量。

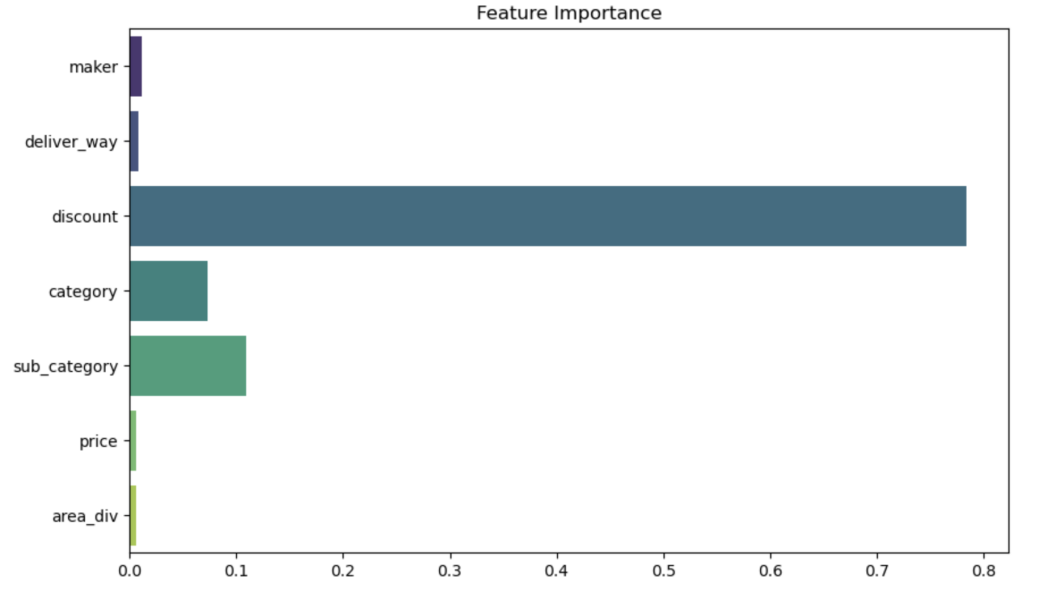


图 22 梯度提升树的特征重要性分布图

1. **预测结果分析**

本次训练出的预测模型可以根据商品的属性，在其进入市场前对于其能达到的利润率进行预测。只要商家给出其制造商、运送方式、折扣、商品类别、单价与其主要目的地域的多样性评估，预处理部分就可以对相关特征进行处理，然后梯度提升树模型就可以对这个商品做出相对准确的利润率预测。

对于已经进入市场的商品，如果商家希望得到一些关于获利的指示，那么可以参考本模型的特征重要性分析：主要来说，改变其打折策略与其主要售出商品的种类都将在一定程度上利于商家获利。具体来讲，商家需要考虑商品的打折促销是否真正能让自己获利，并考虑在一定程度上减少折扣力度以防止亏损。同时，也可以考虑售卖更受市场欢迎的商品种类，而不是死守某类特定商品。

综上所述，本次分析基于数据集的整体特征，对于零售业的整体情况进行了可视化与分析，通过分析结果揭示了行业趋势。同时，基于行业情况进行了面向商品的特征提取，并结合多种方法对特征进行预处理与筛选。在得到最终的特征之后，本次分析结合多种回归预测模型对商品的利润率进行了预测，并通过对比指标得到一个较好的预测模型。通过对预测模型的特征重要性分析，本次分析对于商家获利的策略给出了客观的建议。