基于大数据分析的社交媒体在旅游行业中的作用与趋势预测

作者：林燕勤

摘　要：

社交媒体已经成为人们生活中不可或缺的一部分。特别是在旅游行业，社交媒体不仅改变了传统的旅游信息传播方式，还深刻影响了消费者的旅游决策过程。通过大数据分析，可以揭示社交媒体在旅游行业中的具体作用，预测未来的旅游趋势，从而为企业制定更为精准的市场营销策略提供数据支持。

关键词：社交媒体， 旅游行业，旅游信息传播方式，旅游决策，大数据分析。

引言

随着社交媒体的迅速发展，越来越多的消费者在选择旅游目的地、预订旅游服务以及分享旅游体验时依赖社交媒体。通过社交媒体平台，用户可以获取他人对旅游产品和服务的评价，这种口碑效应对旅游行业的影响日益显著[1]。鉴于此，本文旨在通过大数据分析探讨社交媒体在旅游行业中的作用，并预测其未来趋势。通过研究消费者在社交媒体上的行为模式和情感倾向，可以为旅游企业提供数据驱动的决策支持，从而提升其市场竞争力。

1　研究问题

本研究主要围绕社交媒体对旅游行业的具体影响、用户在社交媒体上的行为如何影响其旅游决策、社交媒体数据是否能有效预测旅游行业的趋势以及不同类型的机器学习模型在情感分析和趋势预测中的表现展开[2]。通过分析社交媒体数据，通过研究消费者在社交媒体上的行为模式，可以为旅游企业提供数据驱动的决策支持，从而提升其市场竞争力。

2　数据集来源及数据集介绍

本研究所使用来源于 Kaggle https://www.kaggle.com/datasets/ddosad/customer-behaviour-tourism-portal 的名为“Customer_behaviour_Tourism.csv 的数据集。该数据集包含了客户在社交媒体平台上的行为数据。数据集来源于某知名旅游社交媒体平台，涵盖了大量的用户互动信息[3]。具体字段及其解释如下：

字段名称	字段解释	缺失率
UserID	用户 ID, 使用唯一标识每个用户	0
Taken_product	用户是否购买旅游产品（Yes/No）	0
Yearly_avg_view_on_travel_page	每年用户在旅游页面上的平均浏览量	0.05
preferred_device	用户首选设备（例如 iOS、Android 等）	0
total_likes_on_outstation_checkin_given	用户在外地签到时**给出的**点赞数	0.03
yearly_avg_Outstation_checkins	每年用户在外地的平均签到次数	0.01
member_in_family	用户家庭成员数量	0
preferred_location_type	用户偏好的旅游地点类型（例如Financial、Other、Medical 等）	0
Yearly_avg_comment_on_travel_page	每年用户在旅游页面上的平均评论数	0.02
total_likes_on_outofstation_checkin_received	表示用户在外地签到时**收到的**点赞数	0

week_since_last_outstation_checkin	用户自上次外地签到以来的周数	0
following_company_page	用户是否关注公司页面（Yes/No）	0.01
montly_avg_comment_on_company_pag	用户在公司页面上的月均评论数	0
working_flag	用户是否在职（Yes/No）	0
travelling_network_rating	用户的旅游网络评分（1 到 5 的评分）	0
Adult_flag	用户是否为成人（0 表示未成年，1 表示成人）	0
Daily_Avg_mins_spend_on_traveling_page	用户每日在旅游页面上花费的平均分钟数	0

3 数据集处理分析

在对数据集进行分析之前，需要进行数据清洗和预处理。原始数据内容包含 17 个字段和 11760 行。在数据预处理阶段，主要针对缺失和重复数据进行处理。首先，原始数据中存在最高 5%的缺失值，由于数量较少（未高于 20%），因此选择删除这些缺失值。经过删除后的数据量为 10455 行，且数据本身并无重复数据。针对描述性统计，对于数据的描述性统计将分为离散型和连续性 2 种方式，连续型的部分我们主要观察密度跟分布，离散型数据则是主要观察每个类别的量,作为之后模型选择的依据。

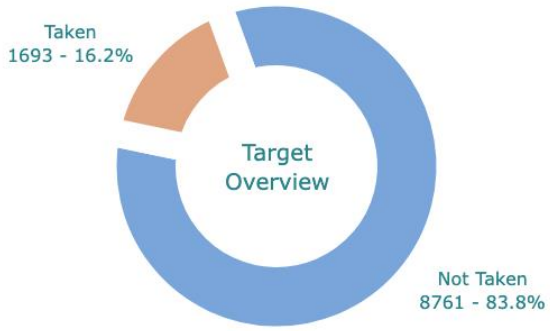
为了便于后续的分析和建模，需要对数据进行标准化和特征工程处理。由于数据之间的数量级差别较大（超过 3 个数量级），因此选择使用 MinMaxScaler 进行标准化处理。处理后的数据如下：

字段名称	标准化后示例值
Yearly_avg_view_on_travel_page	0.54825
preferred_device	1
total_likes_on_outstation_checkin_given	0.74378
yearly_avg_Outstation_checkins	0
member_in_family	0.33333
preferred_location_type	0.21429
Yearly_avg_comment_on_travel_page	0.77941
total_likes_on_outofstation_checkin_received	0.24515
week_since_last_outstation_checkin	0.88889
following_company_page	1
montly_avg_comment_on_company_page	0
working_flag	0
travelling_network_rating	0
Adult_flag	0
Daily_Avg_mins_spend_on_traveling_page	0.17857

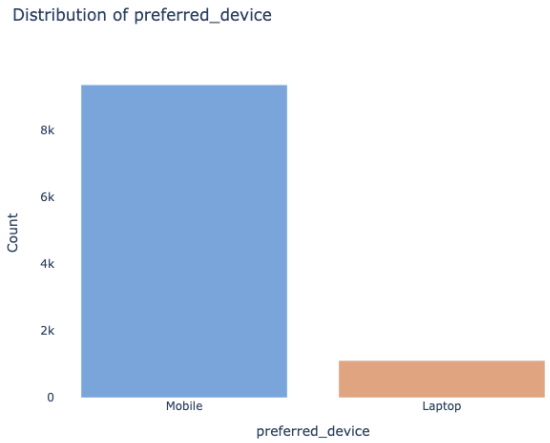
3.1 离散型数据描述

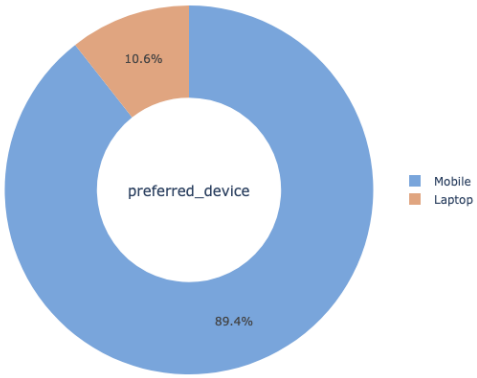
如下是关于离散型数据描述：

Feature	count	unique	top	freq
Taken_product	11760	2	No	9864
preferred_device	11760	2	Mobile	10652
preferred_location_type	11760	15	Beach	2455
following_company_page	11760	4	0	8458
working_flag	11760	2	No	9952
travelling_network_rating	11760	5	3	3672
Adult_flag	11760	4	0	5048

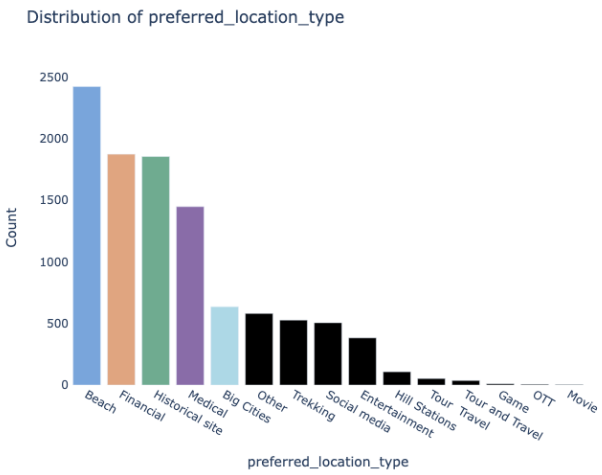


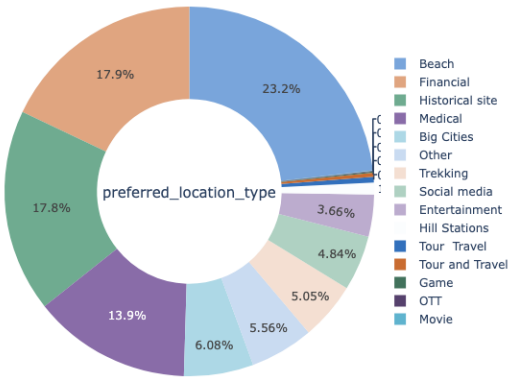
结论：用户偏好与行为(Taken_Product)显示大多数使用者没有使用过特定产品，数据存在不平衡。



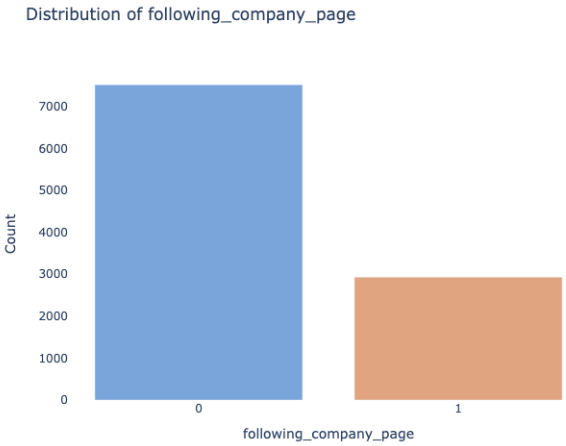


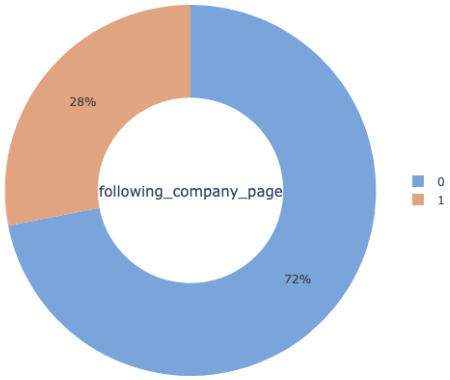
结论：关于移动电子设备参与(preferred_device)，对移动电子设备使用的高度偏好(89.4%)显示出高参与度。这也表明数据不平衡。因此优先考虑所有电子设备内容的行为优化，包括广告、网站和应用程式。



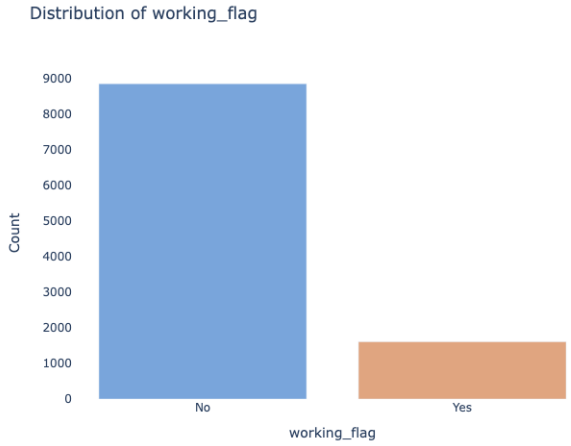


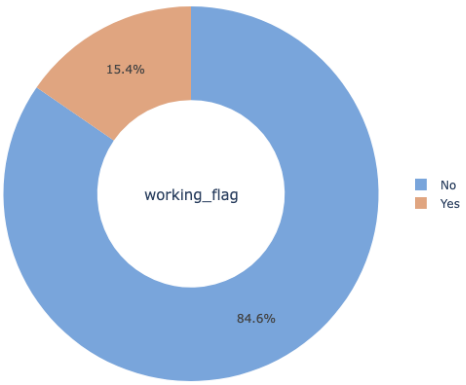
结论：关于旅游偏好(preferred_location_type)，旅行地点偏好呈现多样性，首选海滩、金融区和历史遗址。



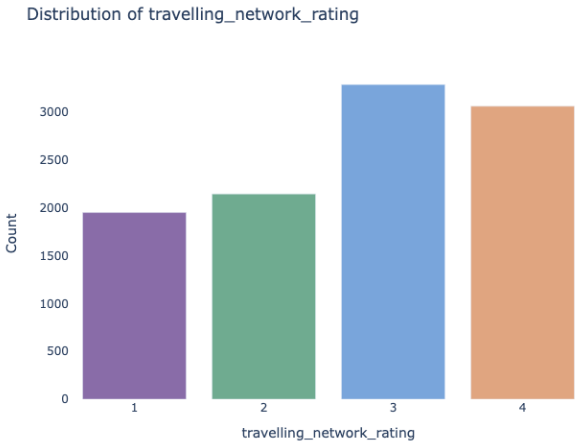


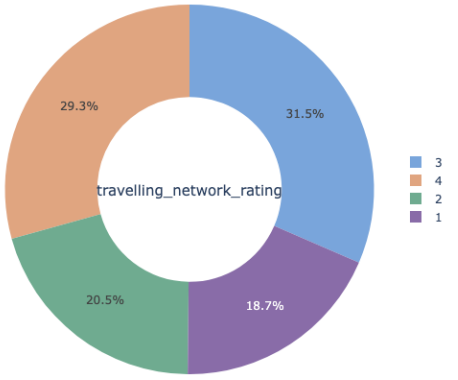
结论：关于用户与公司的互动，很大一部分人（72%）不关注公司页面(following_company_page)，这表明用户参与度存在潜在差距。需制定策略，透过提供有价值且相关的内容来增加公司页面的关注者。透过互动贴文、调查和出校活动吸引用户，以培训社群意识和忠诚度。



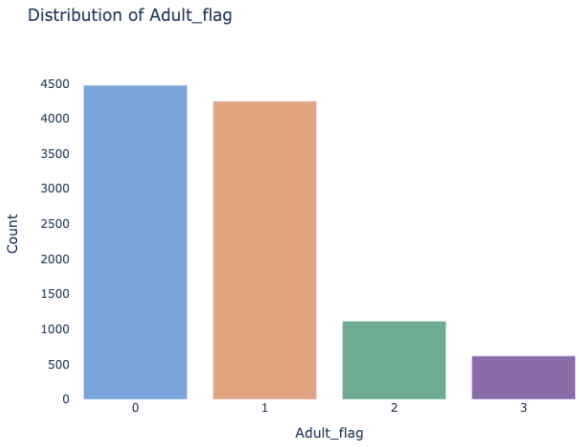


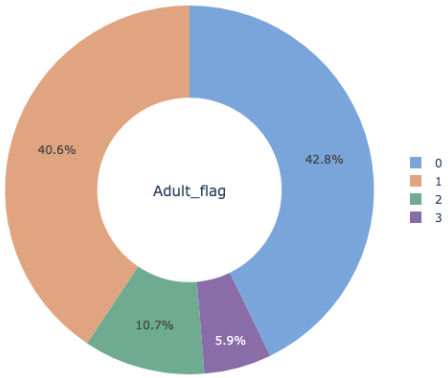
结论：关于工作影响(working_flag)，了解到 84.62%的非工作者占主导地位，这表明需要定制化营销策略。考虑为该类用户提供灵活的旅游套餐、非高峰时段折扣活特别促销。





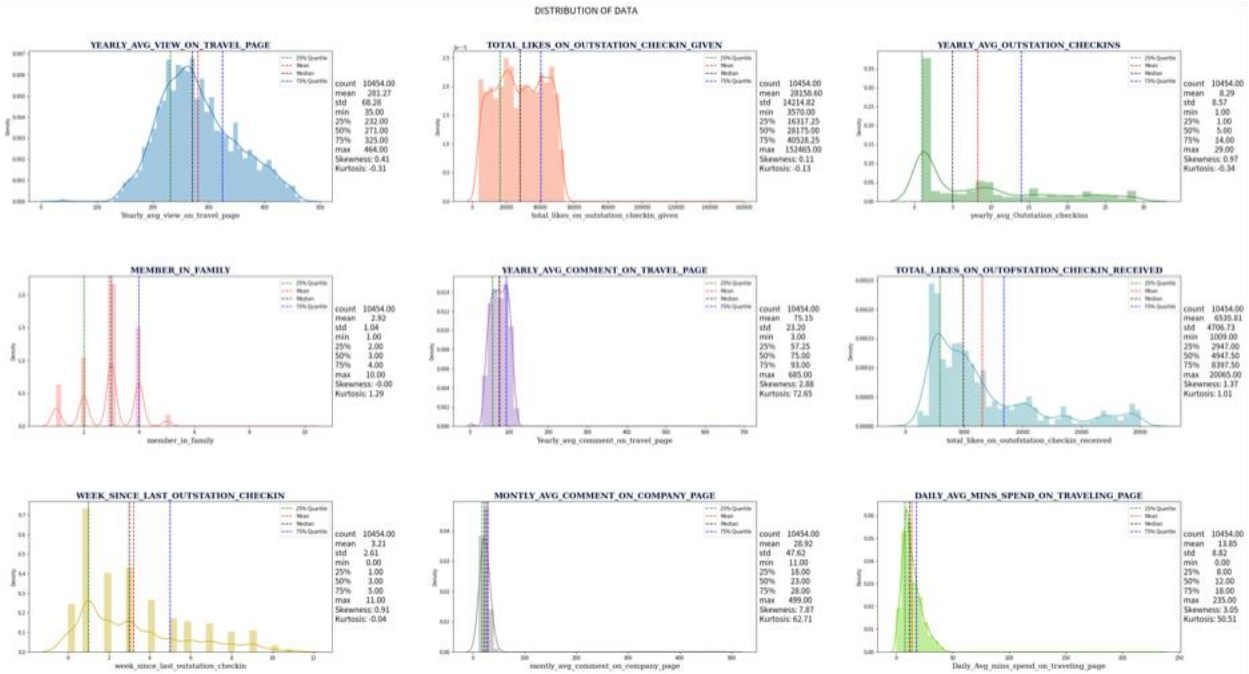
结论：解决旅游网络评级问题，评级分布，尤其是「3」31.5% 和「4」29.3%的显著百分比，凸显了解决用户回馈的重要性。实施改进以提高整体满意度并提供积极的用户体验。





结论：针对使用者细分进行人口统计定制化，考虑成人标签类别的多样化分布，其中，「0」42.8% 和「1」40.6%比例较大，根据使用者年龄层订花服务和促销可以带来更有效的效果以及由针对性的营销活动。

3.2 连续型数据描述

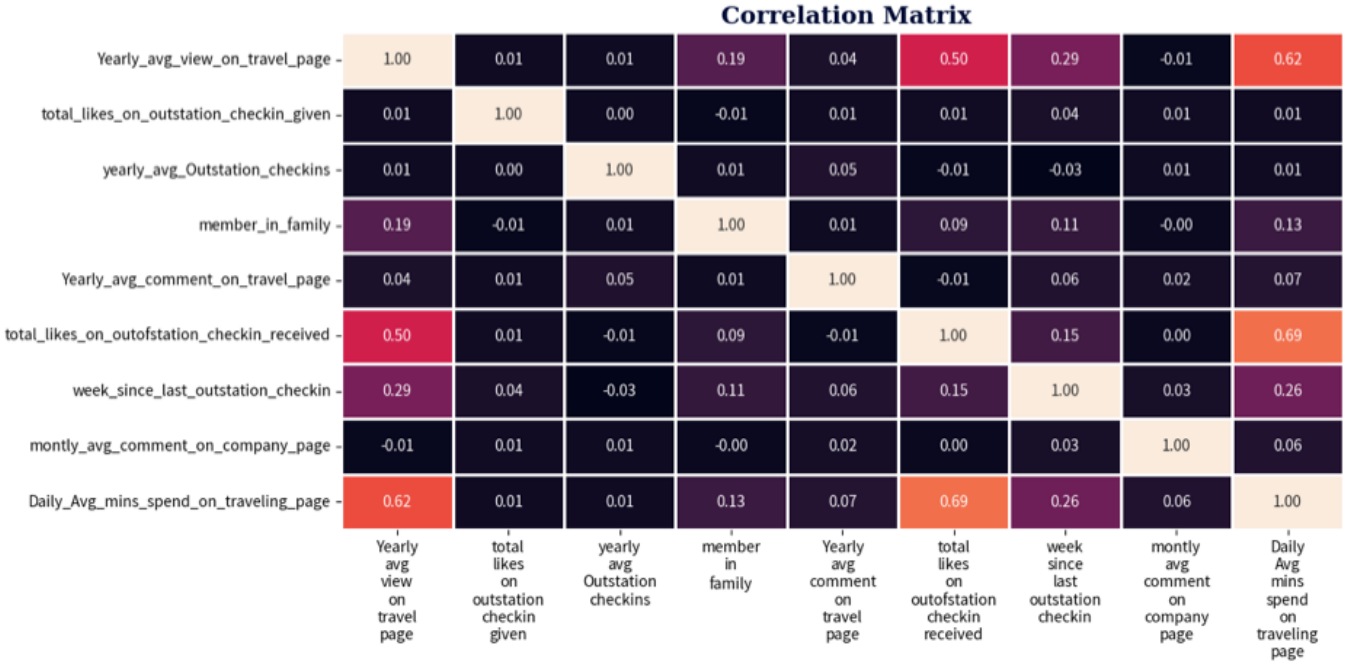


上述可视化图表结果表明：

1） Skewness：数据显示不同程度的偏度（表示分布的不对称性）和峰度（反映尾部的厚度）。例如，外站签到和收到的赞具有显著的偏度和较高的峰度，表示可能集中在某些范围内。

2） Outliers: 虽然大多数功能并未显示出极端异常值，但有些功能却表现出变异性，特别是在外站签到、收到的点赞数和家庭规模方面。

3） User Interaction Insights: 使用者通常会对旅游和公司页面发表评论，按赞数和收到的按赞数有明显差异。



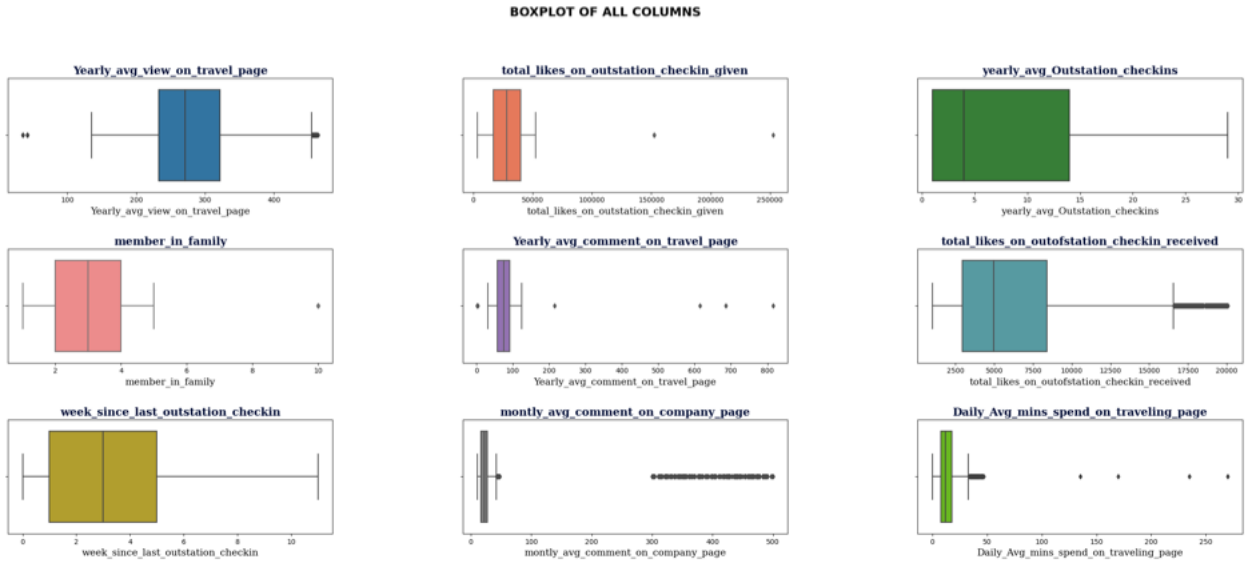
上述可视化图表结果表明：

1） 参与模式：花更多时间在平台上查看旅行内容的用户也往往会花更多时间参与旅行相关活动，旅行页面的年平均浏览量与旅行页面的日平均花费时间之间存在适度的正相关性(0.58).

2） 社交互动影响：社交互动在用户参与度中发挥着重要作用，站外签到的总点赞数与旅行页面上的日平均花费时间之间存在很强的正相关性（0.67）。这表明在签到时获得更多赞的用户可能会在平台上花费更多时间.

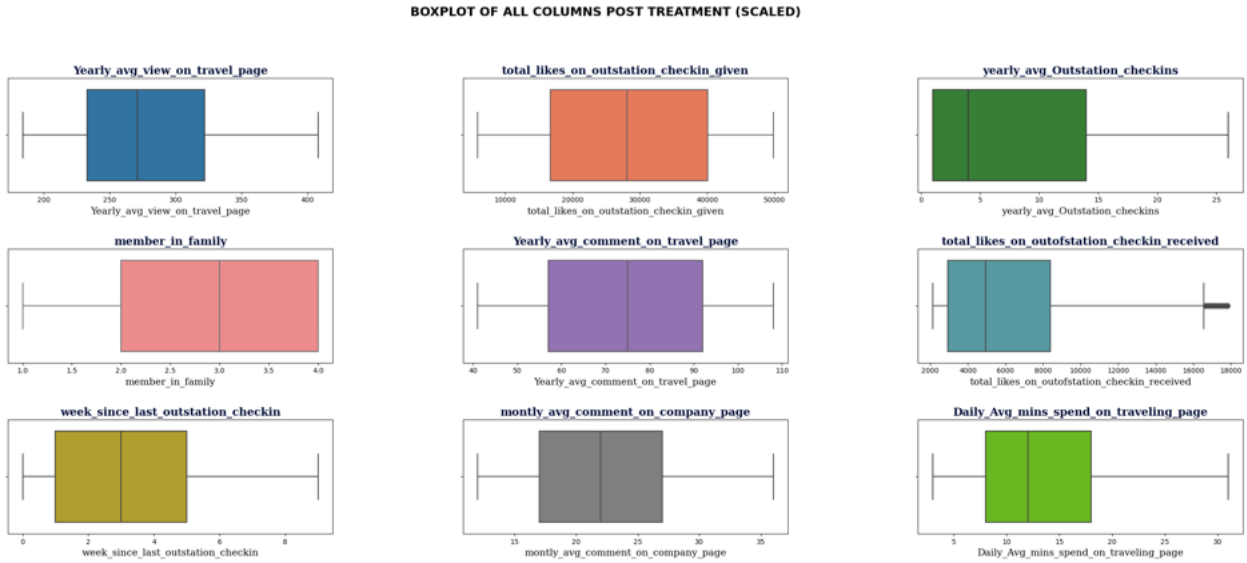
3） 内容多样性与家庭动态：旅游页面的年平均浏览量与家庭成员数量之间存在微弱的正相关性（0.19），这表明家庭动态与旅行内容的参与度之间存在轻微的关联。此外，与上次外站签到以来周数的相关性(0.28) 表明，使用者参与度可能会根据旅游相关活动的新近程度而有所不同。

3.3 异常型描述



上图为 9 个数值类型特征的分布情况, 异常值通常指的是在整体数据分布中在四分位数中超过上下限的数值。

经异常值处理后的分布如下图显示：



4 模型选择喝特征工程

在该次实验中，涉及是否购买的意愿作为预测作为分类问题。由于 `UserID` 做为用户编号, 在此不具备特征意义, 因此在此优先删除。在购买意愿方面, 采用 `Taken_product` 作为目标, 而其它 15 个字段为特征。

4.1 特征工程

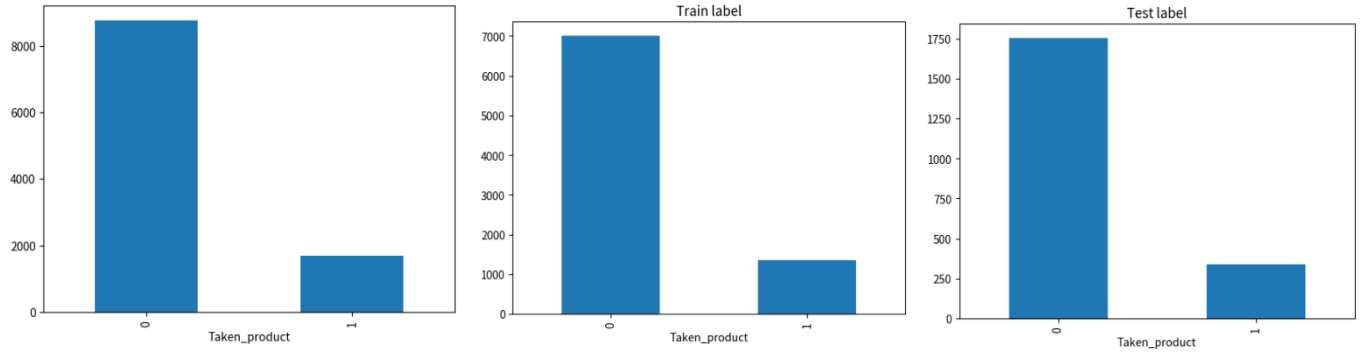
特征工程的主要目的是在建模之前确定哪些特征应该使用以及如何使用，对此我将对特征进行标签化、标准化和数据分割处理。在前面探索阶段将数值型数据进行了异常值的去除，而在离散型数据方面将对它们使用标签化(LabelEncoder)。标签化后的结果如下：

Yearly_avg_view_on_travel_page	preferred_device	total_likes_on_outstation_checkin_given	yearly_avg_Outstation_checkins	meber_in_family	prefeRred_location_type	Yearly_avg_comment_on_travel_page	total_likes_on_outofstation_checkin_received	week_since_last_outstation_checkin	following_company_page	montly_avg_comment_on_company_page	working_flag	travelling_network_rating	Adult_flag	Daily_Avg_mins_spend_on_traveling_page
1	307	1	38570	1	2	3	94	5993	8	1	12	0	1	0
0	367	1	9765	1	1	3	61	5130	1	0	23	1	4	1
1	277	1	48055	1	2	10	92	2136	6	1	15	0	2	0
0	247	1	48720	1	4	3	56	2909	1	1	12	0	3	0
0	202	1	20685	1	1	7	41	3468	9	0	12	0	4	1

由于数据之间的数量级差别较大（超过 3 个数量级），因此将会对数据进行标准化处理，在标准化方法的选择上由于数据中并不存在附属，所以选择 MinMaxScaler 的方式，标准化之后的数据版本如下：

Yearly_avg_view_on_travel_page	preferred_device	total_likes_on_outstation_checkin_given	yearly_avg_Outstation_checkins	meber_in_family	prefeRred_location_type	Yearly_avg_comment_on_travel_page	total_likes_on_outofstation_checkin_received	week_since_last_outstation_checkin	following_company_page	montly_avg_comment_on_company_page	working_flag	travelling_network_rating	Adult_flag	Daily_Avg_mins_spend_on_traveling_page
0.5483	1.0000	0.7438	0.0000	0.3333	0.2143	0.7794	0.2452	0.8889	1.0000	0.0000	0.0000	0.0000	0.000000	0.1786
0.8114	1.0000	0.0910	0.0000	0.0000	0.2143	0.2941	0.1903	0.1111	0.0000	0.4400	1.0000	1.0000	0.333333	0.2500
0.4167	1.0000	0.9587	0.0000	0.3333	0.7143	0.7500	0.0000	0.6667	1.0000	0.1200	0.0000	0.3333	0.000000	0.1429
0.2851	1.0000	0.9738	0.0000	1.0000	0.2143	0.2206	0.0491	0.1111	1.0000	0.0000	0.0000	0.6667	0.000000	0.1786
0.0877	1.0000	0.3385	0.0000	0.0000	0.5000	0.0000	0.0847	1.0000	0.0000	0.0000	0.0000	1.0000	0.333333	0.1071

在数据分割方面，我采用随机抽取 20%的数据作为测试用到的数据大小及确认。在对比目标分布的整体、训练、测试分割后的比例，以查看分割后的数据是否有着同样比例，以下可以看到分割后的数据与整体数据有着同样的差距比例。



由于在之前对于特征的相关性分析表示，特征与目标皆为弱相关，少数为中相关，如下表：

index	Taken_product
Daily_Avg_mins_spend_on_traveling_page	-0.1607
total_likes_on_outofstation_checkin_received	-0.1601
Yearly_avg_view_on_travel_page	-0.151
Adult_flag	-0.133
preferred_device	-0.0815
total_likes_on_outstation_checkin_given	-0.0527
travelling_network_rating	-0.0458
member_in_family	-0.0315
montly_avg_comment_on_company_page	-0.0116
Yearly_avg_comment_on_travel_page	-0.0084
working_flag	0.00031
preferred_location_type	0.02625
week_since_last_outstation_checkin	0.05259
yearly_avg_Outstation_checkins	0.07654
following_company_page	0.24517

因此，线性模型不是我们的首选，而是作为效果的对比，自此分类案例中首先考虑树状模型。采用树状模型除了是根据其中可能会有相对理想的效果之外，可以更进一步的说明模型当中的特征重要性。

5 模型建立和评估（及其数学公式）介绍

在训练好模型之后，对于模型效果的评估，我们将采用以下指标：准确率（Accuracy）、精确率（Precision）、召回率（Recall）和 F1 分数（F1 Score）。这些指标可以全面衡量模型的分类性能，尤其是在不平衡样本中。其中，

- 准确率（Accuracy）：正确分类的样本数占总样本数的比例。
- 精确率（Precision）：模型预测为正类的样本中真正为正类的比例。
- 召回率（Recall）：真正的正类样本中被模型正确预测为正类的比例。
- F1 分数（F1 Score）：精确率和召回率的调和平均数，综合反映模型的分类性能。

在分类模型中，由于样本类别可能不平衡，单纯考虑准确率可能会导致误导性结论。因此，我们需要结合精确率和召回率，详细分析模型在正负样本中的表现。在接下来的说明中，我们将通过交叉验证结果得出的平

均效果，以及模型在各项指标上的表现，作为最终结论的依据。以下为该次实验用到的模型及其数学公式的具体描述和实验结果。

5.1 模型介绍及数学公式

1） 逻辑回归(Logistic Regression): 是一种线性模型，用于解决二分类问题。其基本思想是通过线性回归的形式计算出事件发生的概率。逻辑回归模型的输出是一个介于 0 和 1 之间的概率值，表示某个事件发生的可能性。其模型的优点为：结构简单，结果解释性强；计算效率高，适用于大规模数据集；对线性可分数据集效果好；通过添加正则化项来防止过拟合。缺点：对非线性可分的数据集效果较差；在高维数据集上容易过拟合，需要正则化。异常值会对模型产生较大影响。核心点：使得预测值尽可能接近实际值。其用到的数学公式为：

$$P(y = 1 \mid X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\cdots+\beta_n X_n)}}$$

公式解释：

- β_0 : 截距项，表示在没有输入特征时的预测值。
- β_i : 特征 X_i 的系数，表示特征对预测结果的影响。
- X_i :输入特征，表示数据的各个维度。
- $P(y = 1 \mid X)$: 给定特征 X 时，样本属于类别 1 的概率。

2） 支持向量机(Support Vector Classifier, SVC): 是一种用于分类和回归的监督学习模型。本研究中采用分类用途（简称 SVC）支持向量机通过寻找一个最优的超平面将数据集中的不同类别分开，以最大化不同类别之间的间隔。对于分类任务，支持向量机通常被称为支持向量分类器（Support Vector Classifier, SVC）[4]。其模型的优点为：在高维空间中效果好，适用于复杂决策边界；对数据分布变化不敏感，泛化能力强；通过核方法处理非线性问题。缺点：训练时间长，特别在大规模数据集上；需要调节多个超参数；在大数据集上应用有限，训练和预测时间较长。核心点：核函数用于将数据映射到更高维空间，以便在非线性可分数据中找到线性决策边界。其用到的数学公式为：

$$\text{minimize } \frac{1}{2} \| \mathbf{w} \|^2 \quad \text{subject to } y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1,\ i = 1, \ldots, n$$

公式解释：

- $\mathbf{w}$: 权重向量，表示超平面的方向。
- b : 偏置项，决定超平面的位置。
- $\mathbf{x}_i$: 特征向量，表示第 i 个样本的特征。
- y_i : 标签，表元第 i 个样本的类别 (+1 或 -1)。
- $\frac{1}{2}||\mathbf{w}\|^2$: 目标是最小化权重向量的范数，从而最大化类间间隔。

3） 决策树分类器(Decision Tree Classifier): 是一种树形结构的分类模型。它通过一系列的决策规则将数据划分为不同的类别。决策树的每个内部节点表示一个特征上的测试，每个分支代表一个测试结果，每个叶节点表示一个类别。其模型的优点为：直观的树结构，便于理解和解释；需要对数据进行归一化活

便准话处理；能够处理缺失值。缺点：在训练数据上表现良好，但在测试数据上效果较差；对数据的微小变化敏感，容易导致模型不稳定；不能很好低捕捉特征之间的复杂关系[5]。核心点：决策树通过递归地选择最佳特征进行数据分割，直至每个分支中的数据点属于同一类活达到预设的最大深度。信息增益用于选择最佳分割特征。其用到的数学公式：

- 信息增益:

$$IG(T,a) = H(T) - \sum_{v \in \text{Values}\ (a)} \frac{|T_v|}{|T|} H(T_v)$$

- 基尼指数:

$$\text{Gini}\ (T) = 1 - \sum_{i=1}^{m} (p_i)^2$$

公式解释：

- $H(T)$:集合 T 的熵，衡量数据集的不确定性。
- T_v : 特征取值为 v 的子集。
- p_i : 样本属于第i类的概率。
- $IG(T,a)$: 通过特征 a 进行分割后信息增益的大小。
- Gini (T) :数据集 T 的基尼指数，街量节点的不纯度。

4） 随机森林分类器(Random Forest Classifier): 是一种集成学习方法，通过构建多个决策树并结合它们的预测结果来提高分类的稳定性和准确性[6]。随机森林中的每棵树都是在训练集的不同子集上独立训练的。每棵树在构建时只使用数据的一个随机子集和特征的一个随机子集。最终结果通过所有树的预测结果进行投票决定 。随机森林是 Bagging 的一个典型应用，其中基学习器是决策树。其用到的数学公式为：

- 决策树组合:

$$f(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} f_i(\mathbf{x})$$

公式解释：

- $f(\mathbf{x})$: 最终的预测结果。
- N :决策树的数量。
- $f_i(\mathbf{x})$: 第 T 棵决策树的预测吆吉果。
- Bagging (Bootstrap Aggregating) : 通过对原始数据集进行有放回的随机抽样，生成多个训练集
- 随机选择特征: 在每个节点分裂时，随机选择一部分特征来寻找最佳分割点。

5.2 模型交叉验证及模型结果展示

在建模之后进行 5 折交叉验证进一步说明模型效果，交叉验证是一种强大且灵活的模型评估方法，它通过在多个子集上训练和验证模型，提供对模型性能更准确的估计，减少了由于数据划分方式带来的偏差和方差。通过交叉验证，可以更好地评估模型的泛化能力，避免过拟合和欠拟合，同时为模型选择和参数调整提供可靠的依据[7]。在交叉验证中，数据集被划分为5个子集，每次使用其中一个子集进行验证，剩余的子集进行训练。这样重复 5 次，每次使用不同的子集进行验证，最后对 5 次的验证结果取平均值，得到模型的综合评估结果。这种方法确保了评估的可靠性和稳定性。这些优点使得交叉验证在机器学习和数据科学中被广泛应用。

在本研究中，我使用了逻辑回归、支持向量机（SVC）模型、决策分类器模型、随机森林分类器模型以及网格搜索进行参数调优。以下是对每个模型结果的详细分析：

1） 逻辑回归模型:

交叉验证分数：

Cross-Validation Scores	Mean CV Accuracy
[0.8584, 0.8546, 0.8551, 0.8589, 0.8579]	0.857

交叉验证结果表明模型在不同折叠上的表现非常稳定。**平均交叉验证准确率**：0.857，表示模型在训练和验证集上的整体表现较好。

分类报告：

类别	精确率 (precision)	召回率 (recall)	F1 分数 (f1-score)	支持 (support)
0	0.86	0.99	0.92	1753
1	0.74	0.2	0.32	338
总体	0.86	-	-	2091
宏平均 (macro avg)	0.8	0.59	0.62	2091
加权平均 (weighted avg)	0.84	0.86	0.82	2091

分类报告结果表明：模型总体准确率为 0.86，表明模型在大多数情况下能够正确分类，但在类别 1 上存在显著的识别不足。需要改进模型以提高对类别 1 的召回率，从而提高整体 F1 分数和模型性能。

2)支持向量机分类器模型：

交叉验证分数：

Cross-Validation Scores	Mean CV Accuracy
[0.88235294, 0.88761358, 0.87613582, 0.88522238, 0.87751196]	0.881

交叉验证结果表明模型在不同折叠上的表现较为稳定。平均交叉验证准确率：0.881，表示模型在训练和验证集上的整体表现较好。

分类报告：

类别	精确率 (precision)	召回率 (recall)	F1 分数 (f1-score)	支持 (support)
0	0.89	0.99	0.94	1753
1	0.93	0.37	0.53	338
总体	0.89	-	-	2091
宏平均 (macro avg)	0.91	0.68	0.74	2091
加权平均 (weighted avg)	0.9	0.89	0.87	2091

分类报告结果表明：类别 0 的精确率、召回率和 F1 分数都很高，表示模型在识别类别 0 时表现出色。类别 1 的召回率很低，仅为 0.37，表示模型未能很好地识别出类别 1 的样本。这导致了 F1 分数较低，仅为 0.53。类别 1 的召回率很低，仅为 0.37，表示模型未能很好地识别出类别 1 的样本。这导致了 F1 分数较低，仅为 0.53。

3）决策树分类器模型：

交叉验证分数：

Cross-Validation Scores	Mean CV Accuracy
[0.97513152, 0.98469632, 0.97513152, 0.98421808, 0.98133971]	0.98

交叉验证结果表明模型在不同折叠上的表现非常稳定。**平均交叉验证准确率**：0.980，表示模型在训练和验证集上的整体表现非常好。

分类报告：

类别	精确率 (precision)	召回率 (recall)	F1 分数 (f1-score)	支持 (support)
0	1	1	1	1753
1	1	1	1	338
总体	1	-	-	2091
宏平均 (macro avg)	1	1	1	2091
加权平均 (weighted avg)	1	1	1	2091

分类报告结果表明：模型在所有折叠和测试集上的表现都非常完美，没有任何误分类。所有指标（精确率、召回率和 F1 分数）均达到了 1.00，表示模型在训练集和验证集上都表现出色。可能表明模型存在过拟合现象。

4）随机森林分类器模型：

交叉验证分数：

Cross-Validation Scores	Mean CV Accuracy
[0.98708752, 0.98469632, 0.98947872, 0.9904352, 0.98564593]	0.987

交叉验证结果表明模型在不同折叠上的表现非常稳定。平均交叉验证准确率：0.987，表示模型在训练和验证集上的整体表现非常好。

分类报告：

类别	精确率 (precision)	召回率 (recall)	F1 分数 (f1-score)	支持 (support)
0	1	1	1	1753
1	1	1	1	338
总体	1	-	-	2091
宏平均 (macro avg)	1	1	1	2091
加权平均 (weighted avg)	1	1	1	2091

分类报告结果表明：模型在所有折叠和测试集上的表现都非常完美，没有任何误分类。所有指标（精确率、召回率和 F1 分数）均达到了 1.00，表示模型在训练集和验证集上都表现出色。这种完美的分类结果在实际应用中并不常见，可能表明模型存在过拟合现象。

5）由于上述结果已表明随机森林分类器模型是最优模型，所以在这里对随机森林分类器使用网格搜索（Grid Search）进行最优参数调优。网格搜索是一种用于超参数优化的穷举搜索方法。它通过遍历预定义的超参数集合，训练模型并评估其性能，选择出表现最优的超参数组合。网格搜索的主要优点是其简单直观，能够有效地找到全局最优解。在此次实验中，设置的参数搜索范围如下：

- `n_estimators`：森林中树的数量，取值范围为 [50, 100, 200]
- `max_depth`：树的最大深度，取值范围为 [None, 10, 20]
- `min_samples_split`：内部节点再划分所需的最小样本数，取值范围为 [2, 5, 10]
- `min_samples_leaf`：叶子节点所需的最小样本数，取值范围为 [1, 2, 4]

结果如下：

交叉验证分数：

Cross-Validation Scores	Mean CV Accuracy
[0.98756576, 0.98278336, 0.98708752, 0.98186992, 0.984689]	0.987

交叉验证结果表明模型在不同折叠上的表现非常稳定。平均交叉验证准确率：0.987，表示模型在训练和验证集上的整体表现非常好。

分类报告：

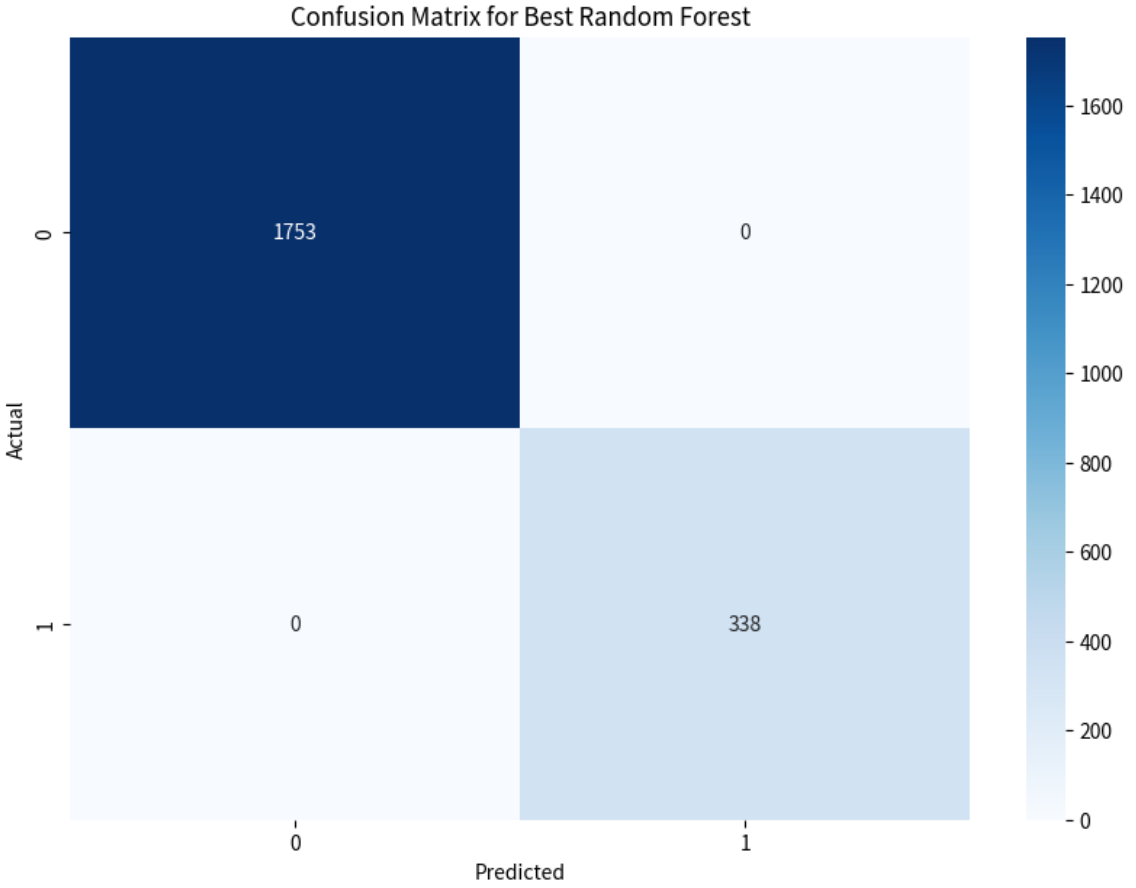
类别	精确率 (precision)	召回率 (recall)	F1 分数 (f1-score)	支持 (support)
0	1	1	1	1753
1	1	1	1	338
总体	1	-	-	2091
宏平均 (macro avg)	1	1	1	2091
加权平均 (weighted avg)	1	1	1	2091

分类报告结果表明：模型在所有折叠和测试集上的表现都非常完美，没有任何误分类。所有指标（精确率、召回率和 F1 分数）均达到了 1.00，表示模型在训练集和验证集上都表现出色。

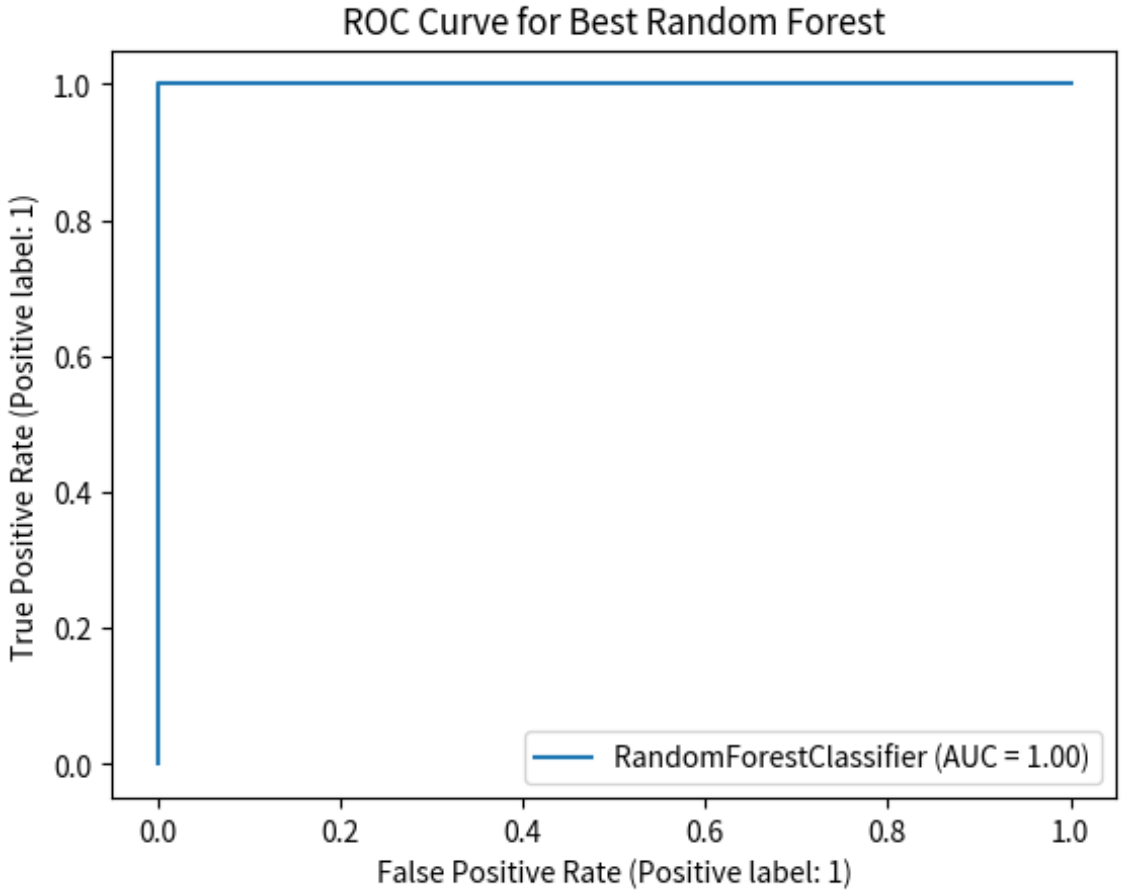
为方便以可视化图展示分析结果，如下矩阵图为辅助结果展现分析。从混淆矩阵中可以看出：

- 1753 个负类样本 (0 类) 被正确分类为负类 (True Negatives)。
- 338 个正类样本 (1 类) 被正确分类为正类 (True Positives)。
- 没有假阳性 (False Positives) 和假阴性 (False Negatives)。

这意味着模型在分类这两个类别时的表现非常出色，没有任何误分类的情况。这种完美的分类结果在实际应用中并不常见，表明模型可能存在过拟合的现象。模型在训练集上表现完美，但在实际应用中的表现可能不如在训练和验证阶段。 意味着特征中存在重要性极高的特征，这可能导致模型过度依赖这些特征，从而无法在新数据上表现出同样的效果。建议对模型进行进一步验证，例如在不同的数据集上测试，或通过交叉验证、引入噪声数据等方法评估模型的鲁棒性和泛化能力。



ROC 曲线说明



- 曲线下的面积 (AUC) 为 1.00，表示模型的分类性能极为优秀。
- 曲线几乎是一条完美的直线，这表明模型在所有可能的阈值下都能够完美地区分正类和负类。

5.3 线性回归模型在其他方面的表现

在交叉验证的方面如同预期的部分线性模型(Logistic Regression, SVC)的表现并不如树状模型(Decision Tree, Random Forest)，而在其他方面的表现如下：

Model	Class	Precision	Recall
LogitisticRegression	0	0.86	0.99
	1	0.74	0.20
SVC	0	0.89	0.99
	1	0.93	0.37
DecisionTreeClassifier	0	1.00	1.00
	1	1.00	1.00
RandomForestClassifier	0	1.00	1.00
	1	1.00	1.00

上述结果分析表明：

1） 逻辑回归在类别 0 上的表现非常好，具有高精度和高召回率，说明模型能很好地识别负类样本。但是在类别 1 上的表现较差，尤其是召回率仅为 0.20，意味着模型对正类样本的识别能力较弱。着可能是因为数据集的不平衡或者模型对类别 1 的特征学习不足。

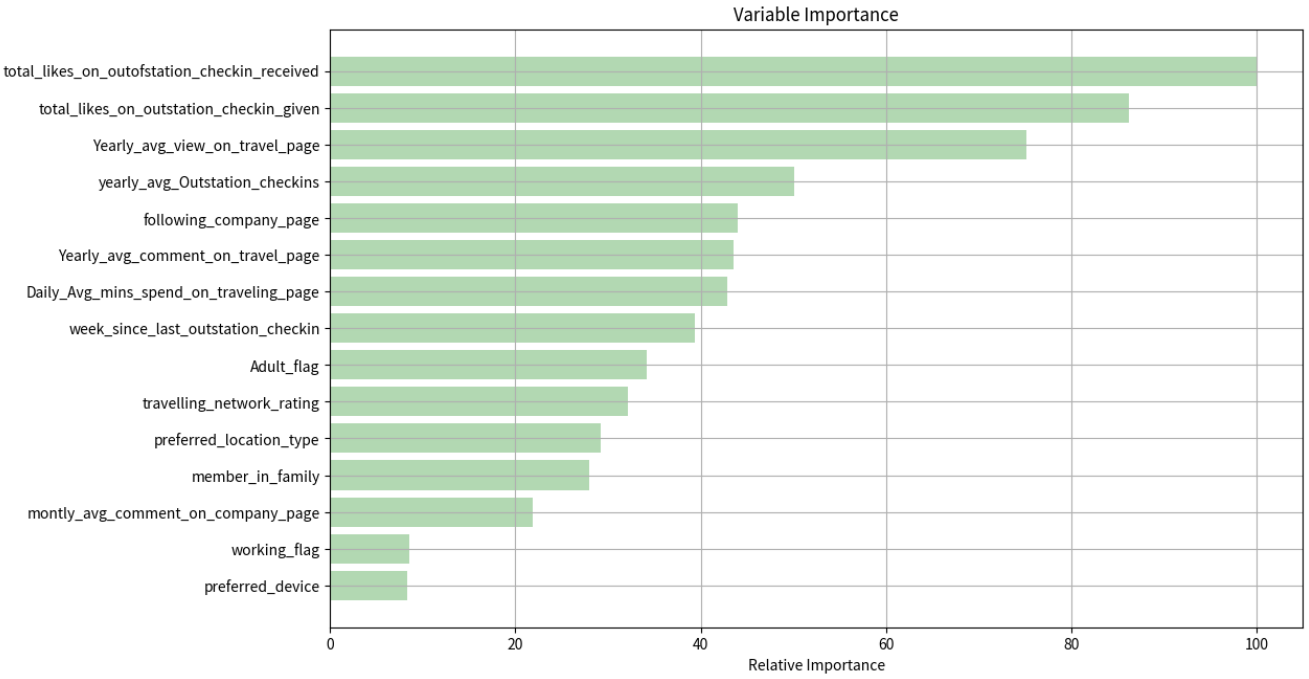
2） SVC 在类别 0 上的表现也非常好，具有高精度和高召回率。相比于逻辑回归，SVC 在类别 1 上的表现有所提升，精度达到了 0.93，召回率也提高到了 0.37，但任然存在对类别 1 的识别能力不足的问题。

3） 决策树分类器在两类上的表现都非常出色，精度和召回率都达到了 1.00。这表明模型对数据的拟合非常好，几乎可以完全正确地分类所有样本。这种现象可能是由于模型过拟合，特别是在训练集上表现优异，但在测试集或实际应用中可能会表现不佳。

4） 随机森林分类器与决策树分类器类似，也在两类上的表现非常出色，精度和召回率都达到了 1.00。随机森林通过集成多个决策树，通常能更好地防止过拟合，但在这里也表现出了极高的精度和召回率，可能同样存在过拟合的风险。

5.4 特征重要性描述

模型建立和评估之后，讨论特征的重要性有助于理解模型决策的依据。因此，在树状模型（决策树和随机森林）中，可以计算各个特征的重要性得分，以了解哪些特征对模型的决策影响最大[8]。特征重要性是通过计算每个特征对模型性能提升的贡献来确定的。 以下是随机森林模型中特征重要性的计算结果：



其中排名前三的特征明显领先其他特征（超 70%），因此我们在购买意愿的项目中可以得出初步结论，评论数以及点赞数为用户是否购买的重要特征。

6 结语

该次实验通过大数据分析和及其学习模型，能够预测用户的购买意愿和在旅游页面上的停留时间，这对旅游业有着多方面的实质性帮助，如：通过预测用户的购买意愿，旅游企业可以更有针对性地进行精准营销。预测用户在旅游页面上的停留时间，可以帮助企业提供更好的用户体验。通过预测用户行为，企业可以更好地进行资源配置。了解用户的偏好和行为，可以指导企业进行新产品的开发[9]。在分类和回归模型中，识别出了一些对用户购买意愿和停留时间有显著影响的特征。针对这些特征，企业可以采取相应的措施来强化用户行为，如：评论数和点赞数被识别为重要特征，企业可以鼓励用户留下评论和点赞，通过提供奖励（如：折扣、积分）来增加用户的互动量。了解用户偏好的旅游地点类型，可以帮助企业进行更有针对性的产品推荐。站外签到和收到的点赞数与用户在平台上的停留时间有很强的正相关性[10]。企业可以通过举办签到活动和增加社交互动功能（如：推出“每日签到赢奖励”活动），来增强用户的参与度。

附录

数据集如下：

Customer_behaviour
_Tourism.csv

"Customer_Behaviour
_Tourism_1_ipynb"的

Code URL: https://colab.research.google.com/drive/1MbgPNoqOtc6df9XSO2NU6_i56O-oo9cK?usp=sharing