数据分析作品集

Email:zs3m20@soton.ac.uk

这里记录了我的一些数据分析的项目，主要展示了对于澳洲 Sprocket Central Pty Ltd 的客户价值预测,Kaggle 电商数据分析和一些其他的学术项目。

基于机器学习技术预测新客户价值

Abstract

本次项目基于 SCPL 公司的 4000 名老客户的 20000 条消费记录，训练了决策树回归模型，模型输入新客户的性别，职业，地址等特征信息，模型输出该用户的消费价值预测。模型实验结果对于训练集的分类准确度达到了 0.65 左右，F1score 达到了 0.7 以上。

Method

首先根据消费记录，计算出每个客户的消费价值，之后去除掉无效字段，空字段，接着判断各个特征(比如性别)和价值的相关性，筛选出相关性较高的字段，之后将行业等字符性进行离散处理，接着通过 Sklearn 的预处理库对于数据进行预处理，之后在多个机器学习模型上进行训练。多次训练测试，保存训练集效果最好的模型，对新客户表的数据进行同样处理后，获取预测结果。

Results and discussions

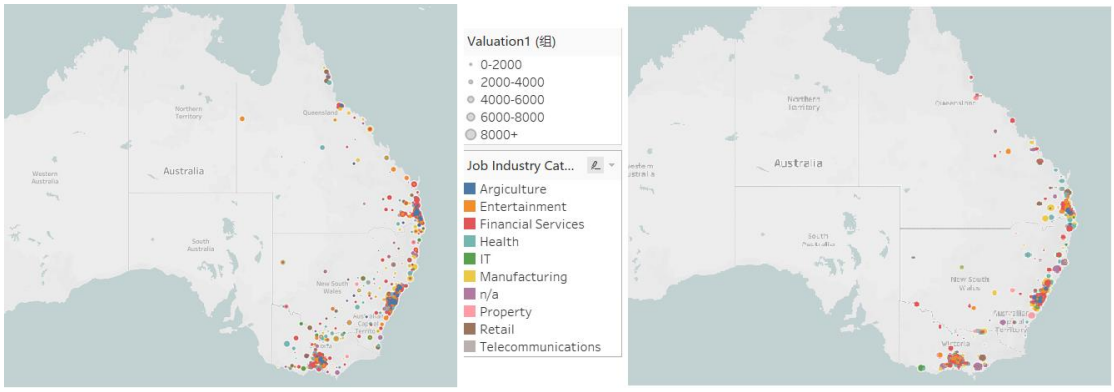


Figure 1: 左图为 4000 名老顾客分布，右图为 1000 名新顾客预测分布.

如图 1 所示，原始数据的老顾客分布情况和预测数据的新顾客预测情况大体一致，除此之外，各个行业的价值分布，以及用户的消费占比，新顾客的情况和老用户也几乎一致，因此本次决策树回归模型对于新用户的价值预测拥有一定的准确度。然而由于没有后续数据，因此很难对模型进行评估，实际应用中，模型在上线前，仍可以对于新数据进行一段时间的测试迭代，最终获得一个成熟的新顾客价值预测模型。

Data presentation

在对 1000 名新用户的价值预测完成之后，使用 Tableau 搭建了动态的可视化仪表盘，统计了 SCPL 公司年度利润，客流量分析，部分用户画像展示，产品受欢迎展示，以及新客户预测展示，仪表盘如图 2 所示。

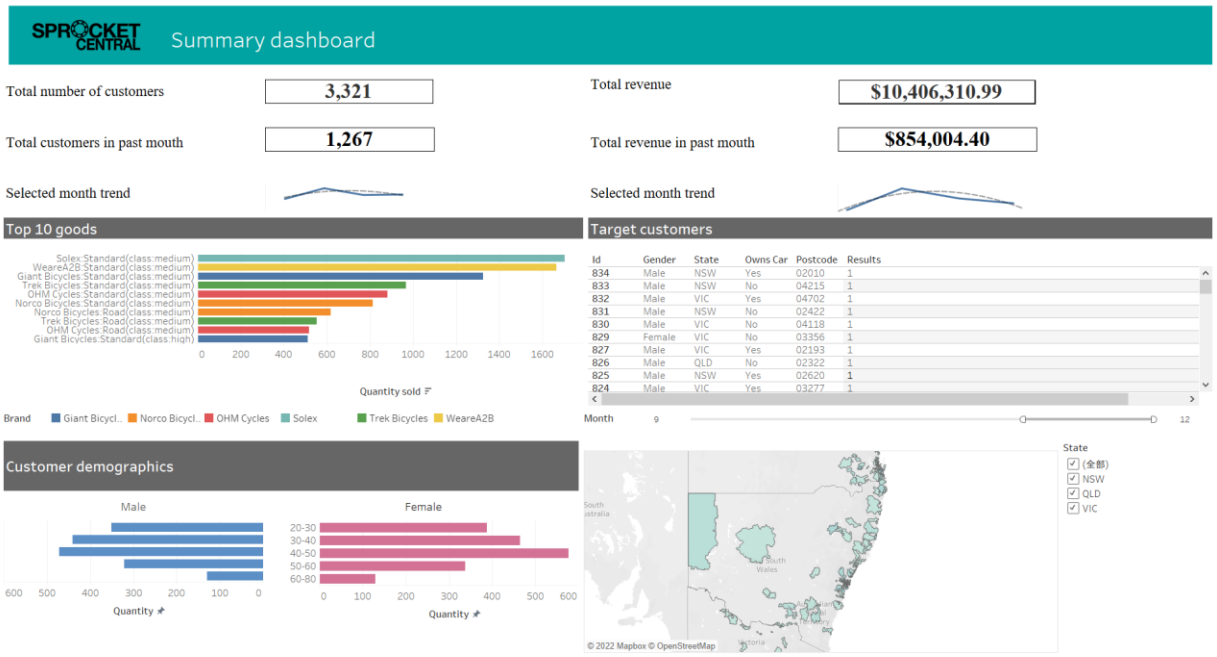


Figure 2: 动态仪表盘

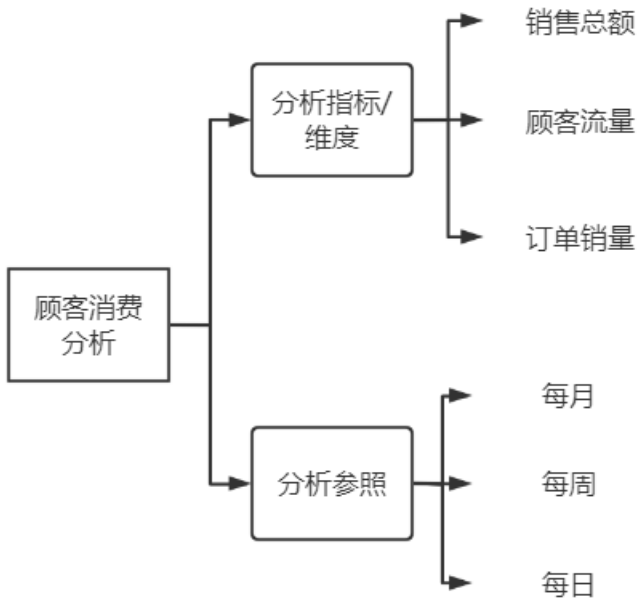
Kaggle 电商数据分析

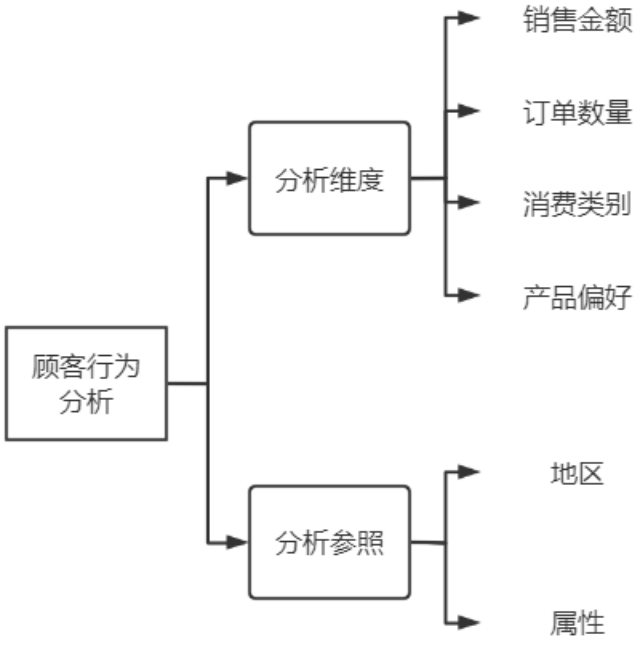
Abstract

本次数据来自 Kaggle 开源项目，数据是英国某电商企业 2010 年末到 2011 年中半年共计 20 余万条的销售数据，数据集中包含，订单 ID,订单数量，产品单价，产品描述，订单时间,订单数量，订单地区等信息。目的是对该企业当前(数据)阶段的销售情况进行数据分析。

Analyse

本次分析从三个维度展开，对企业现阶段的经济发展情况，客户行为情况，客户组成情况分别展开分析，具体分析流程如图 3 所示。


顾客消费分析
分析指标/维度
销售总额
顾客流量
订单销量
分析参照
每月
每周
每日


顾客行为分析
分析维度
销售金额
订单数量
消费类别
产品偏好
分析参照
地区
属性

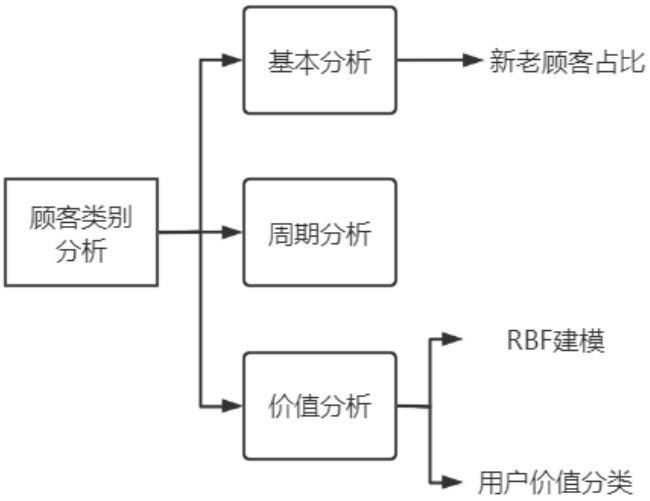


Figure 3: 分析维度

Data Cleaning

首先需要对源数据进行数据清理，经过 Excell 的数据描述，数据筛选后，发现源数据中存在，数据类型不统一，存在空值等情况。

针对时间类型不统一问题，通过 Python 提取出时间数据的年，月，日，并和消费具体时间分开，将具体时间单独提取到新的一列。面对存在空值的情况，考虑到存在大量顾客 ID 空值(约 1/4)，本次分析将在分析消费情况时，保留这些数据，分析顾客情况时筛选这些数据(两个数据源)。针对单价为零的情况，由于数据条目较少，且本身无意义，因此将数据条目为 0 的数据行进行过滤。

- **Data support**

搭建电商消费情况周报，使用 Excell 针对清洗后的数据源，搭建了一个可以根据源数据，根据选择日期动态调整的 Excell 周报，包含 GMV，周流量，日流量，日增长率，周增长率，目标完成情况，周环比等重要数据指标。

数据时间:	2011-01-24	至	2011-01-30			
	线上零售厂商动态周报					
周累计						
	累计客流量		累计GMV		地区:	All
	5845		99270.78		业务进度	
周环比					91.51%	
	客流量周环比		GMV周环比		目标:	500000
	21.82%		2.43%			
结果指标		WEEK				

日期	星期	GMV	客流量	GMV变化	客流量变化	客单价
2011-01-24	周一	17454.2	773	69.69%	-12.06%	22.58
2011-01-25	周二	23013.4	1251	31.85%	61.84%	18.396
2011-01-26	周三	17061.9	1024	-25.86%	-18.15%	16.662
2011-01-27	周四	19518.3	1372	14.40%	33.98%	14.226
2011-01-28	周五	15766.6	703	-19.22%	-48.76%	22.428
2011-01-29	周六	0	0	-100.00%	-100.00%	0
2011-01-30	周日	6456.44	722	0.00%	0.00%	8.9424

Figure 4 : 动态周报

- **Consumption Analyse**

Month

图 4 是该电商公司在 2010 年 12 月至 2011 年 6 月半年的时间内，月消费情况，顾客数量等情况的趋势图，可以发现，10 年 12 月顾客数量和订单情况并非最多，但是该月销售总值最多，因此这个月的顾客购买东西的质量应该最高，推测应为新年影响，顾客更愿意购买高质量商品送礼等。进入 11 年后销量和收益都先降低后升高，在五月收益达到一个局部峰值，推测应为 5 月节假日的原因(银行节)，流量的降低应该为圣诞节，新年后大部分用户的购物需求都得到满足，因此迎来一段时间的淡季，预计人流，和销售情况应该回逐渐升高并在 2011 年 11 月附近来到峰值。

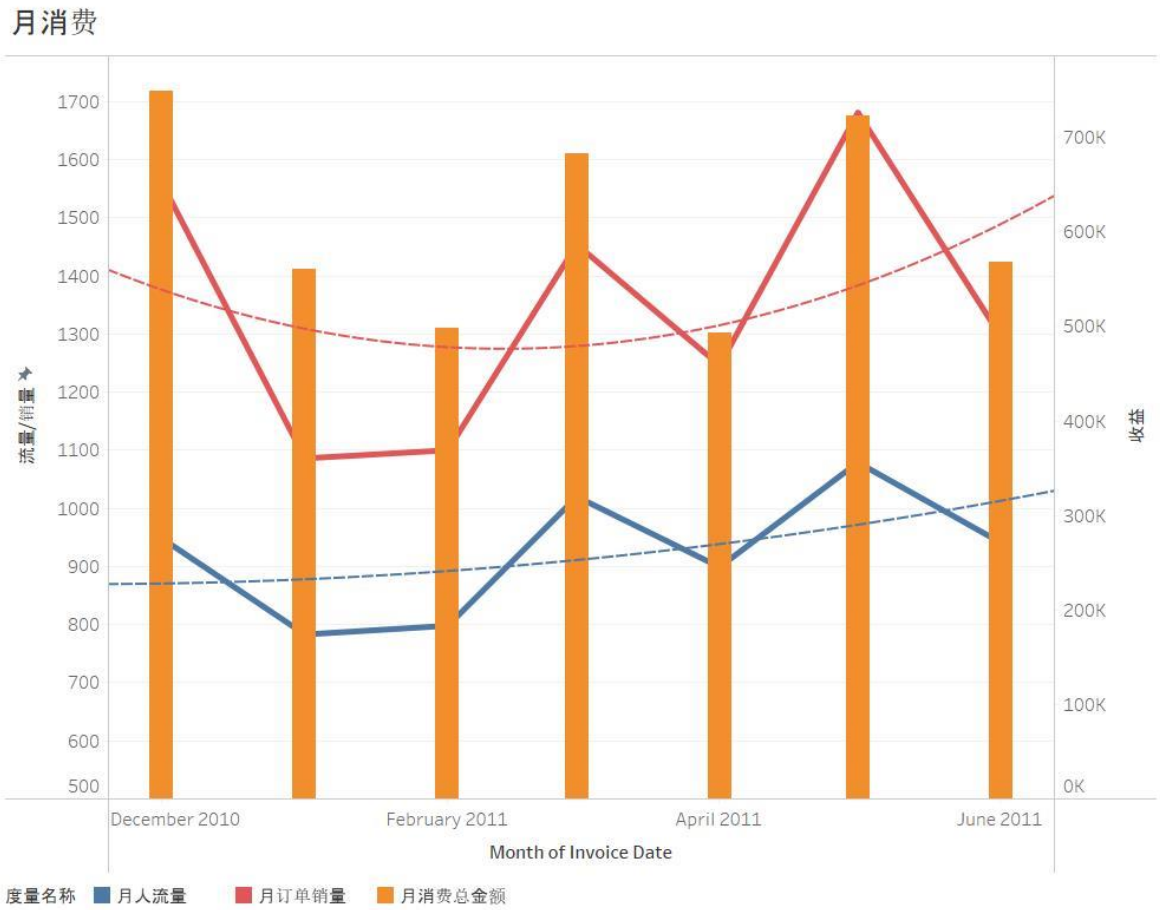


Figure 4:月消费情况

Number

图 5 中我们可以发现，随着时间的推移，好消息是，我们的老顾客消费情况稳定，坏消息是这段时间我们新顾客似乎正在下降，可能是因为企业宣传期空窗的影响。

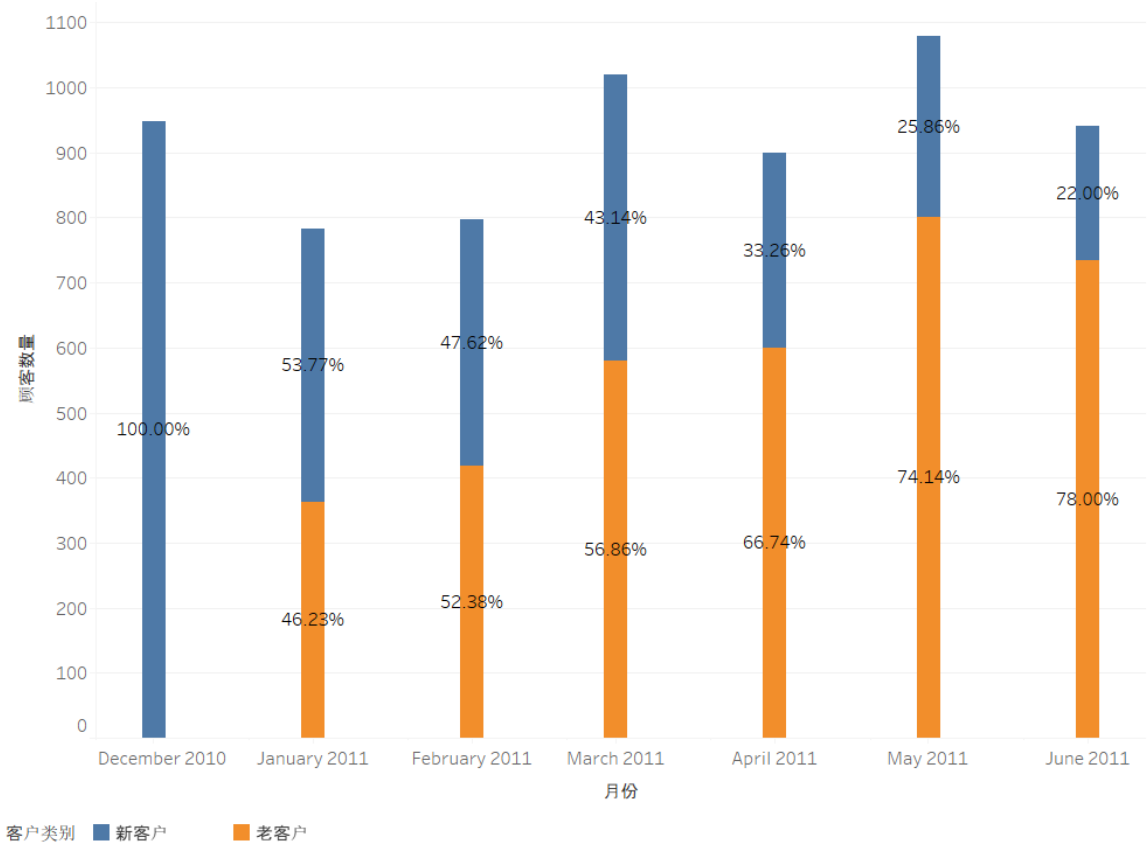


Figure 5:顾客数量/流量

Week

从图 6 中，我们发现，周中(2,3,4)的时候顾客的消费量最高，而周一和周日的消费情况最差，周六应该为企业的休息日，因此没有该日的数据。

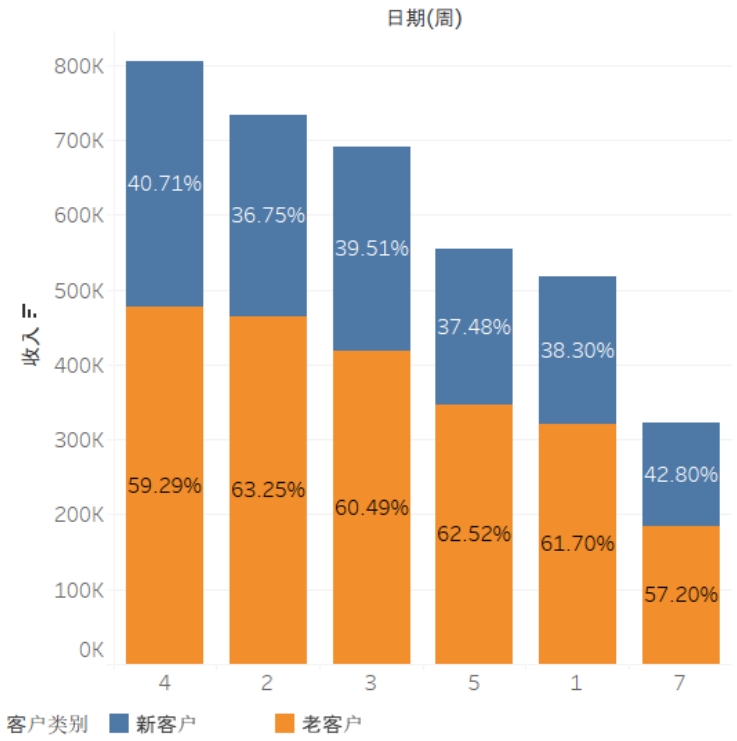


Figure6: 周消费

Day

图 7 的情况和图 6 类似，用户在开始时间和结束时间的购物行为最少，贡献消费金额的核心时间集中在 10 点到 14 点的中午时间。

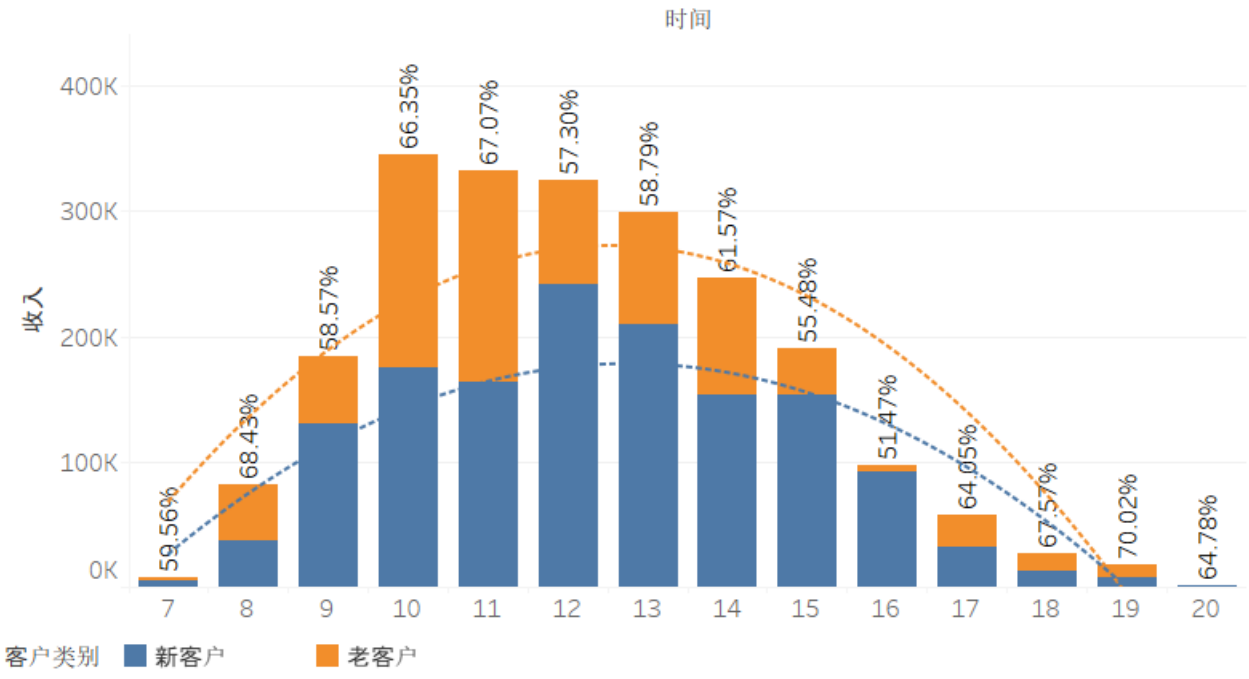


Figure 7: 日消费

- **Customer Behaviour**

图 8 中我们可以发现，现阶段大量用户的消费数量都集中在 400 以下，订单总额都集中在 1 万以下，大量以及较贵的产品售出客户都是老客户，推测应为新客户第一购买可能对企业/平台的质量不够信任。通过更深入的数据可视化，我们发现大量的客户消费总金额都在 600 以下，可见当前阶段，该电商公司缺少足够的高消费客户,核心用户集中于低消费人员。

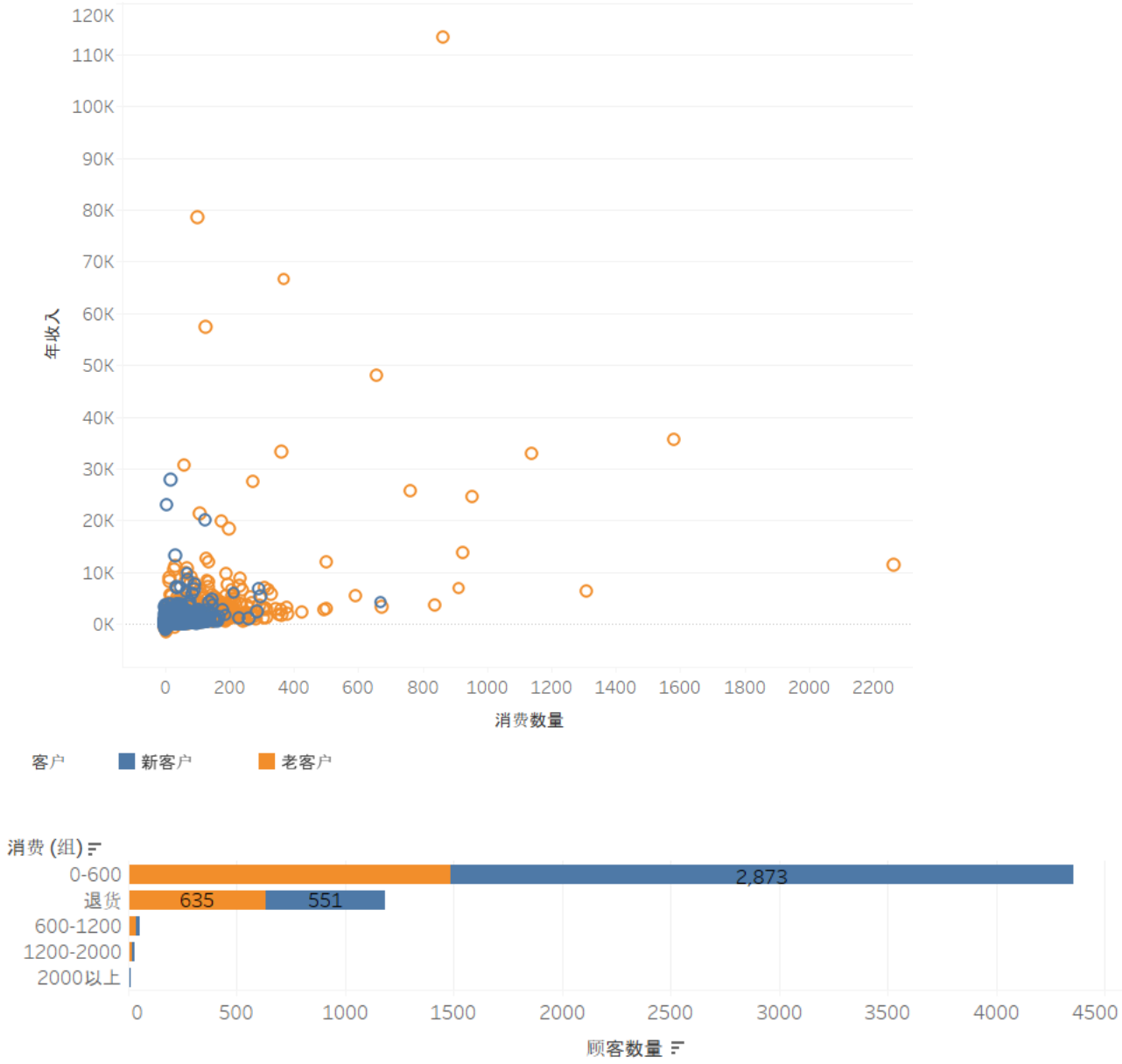


Figure 8 : 客户消费数量与收益关系

Product price

图 9 我们可以发现，当前该电商企业售卖的主流产品，单价集中在 £ 15 以下，目前高质量商品较少。

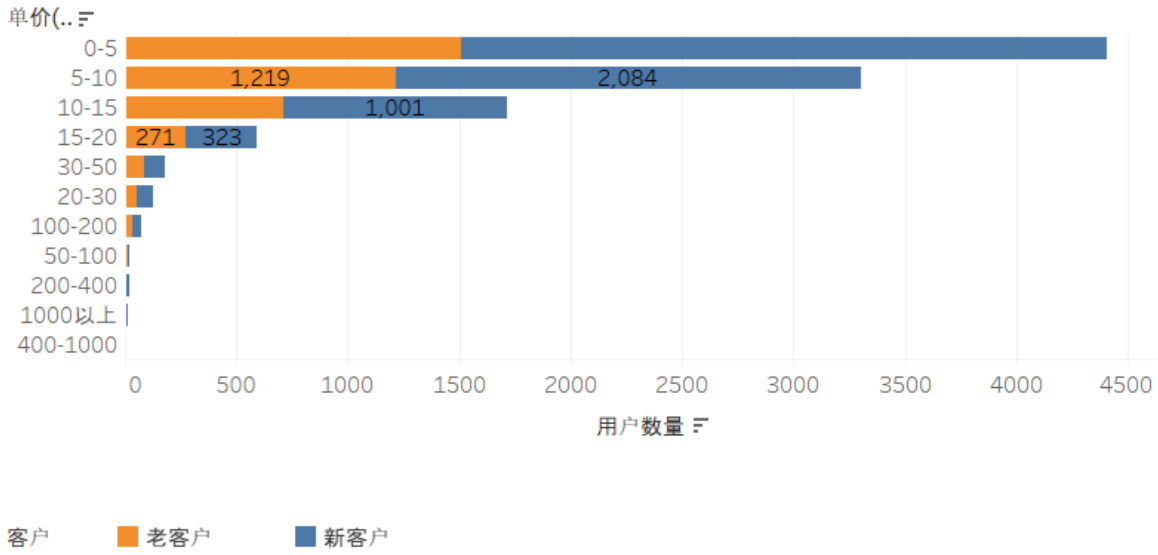


Figure 9 : 商品单价

Country

由于企业商品售出国家过多，因此本次分析只集中于消费占比 1%以上的六个国家，从图 10 中可以发现目前企业的主要客户来源仍是英国国内。

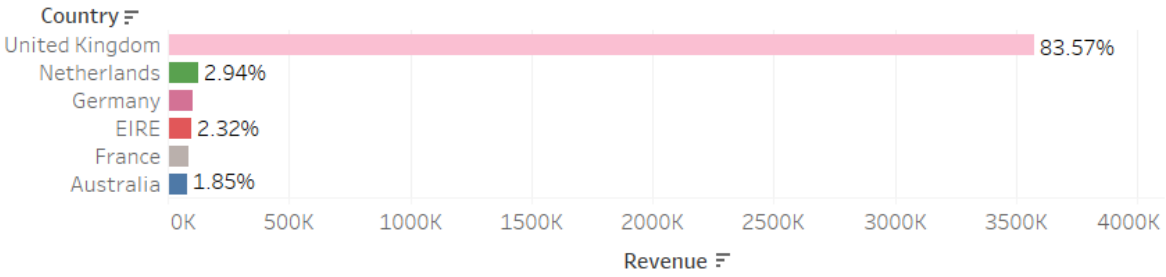


Figure10 : Top 6 Country

图 11 中为各个国家消费总量和客单价的对比图，图中我们可以发现，英国的消费总量虽然很高，但是客单价最低，澳大利亚和荷兰的顾客量小，但是客单价较高，因此这两个国家的顾客应当重点对待，

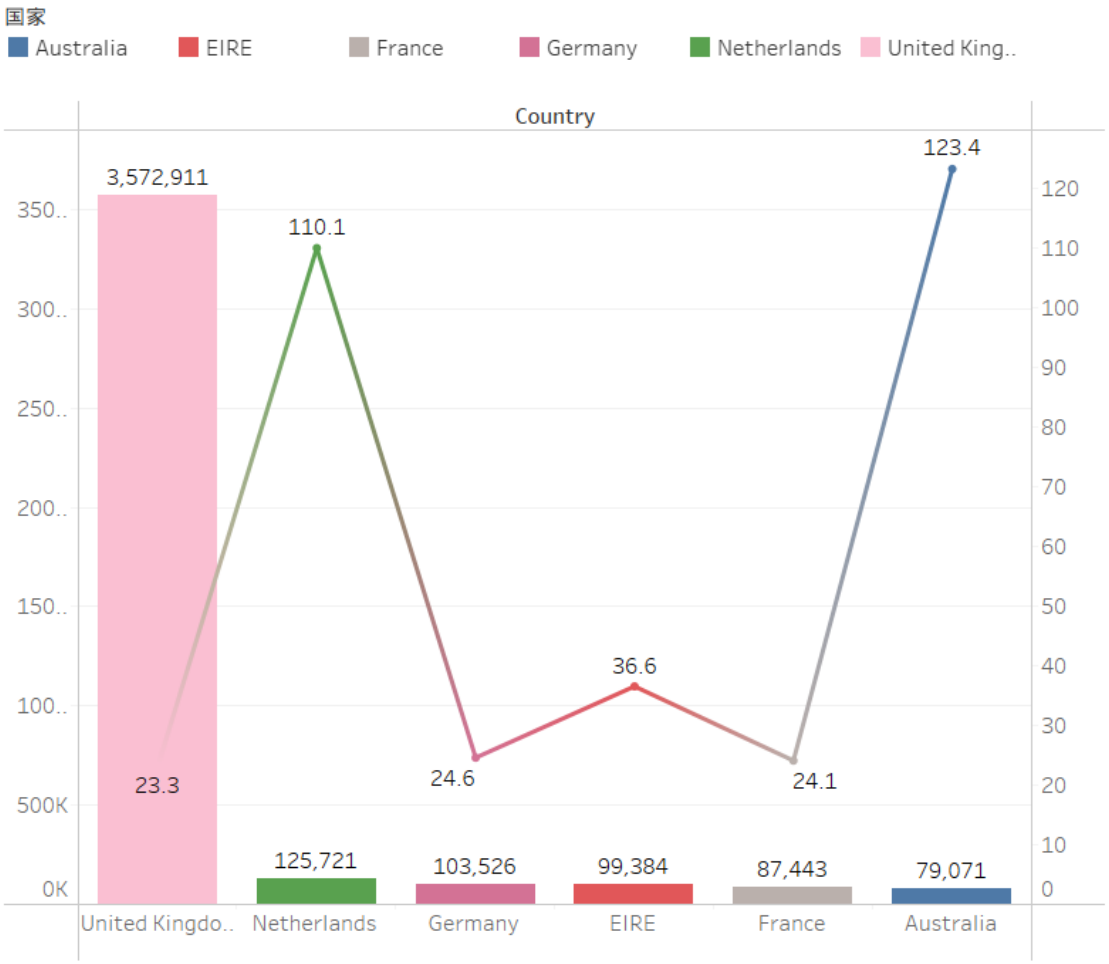


Figure 11: 客单价

- **Customer Classification**

Basic attribute

图 12 是现阶段该电商企业新老客户占比，接近于 6:4 开整体较为稳定，处于发展阶段，切新老客户占比比较健康。

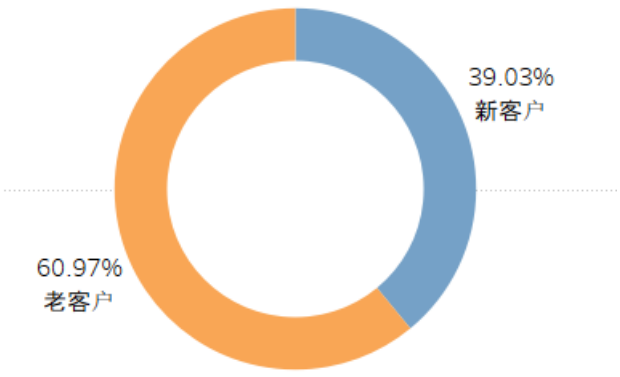


Figure12: 新老客户

Consumption cycle

根据客户消费时间周期进行分类，由图 13 可以发现，现阶段大部分顾客消费仅一次，消费周期较短，不过仍有 30%以上的客户消费周期在 100 天以上，忠实客户占比较为可观，建议现阶段企业需要给仅一次消费的客户进行深入排查，确定不消费原因，通过派发购物券等方式，提升这部分客户的消费频率。

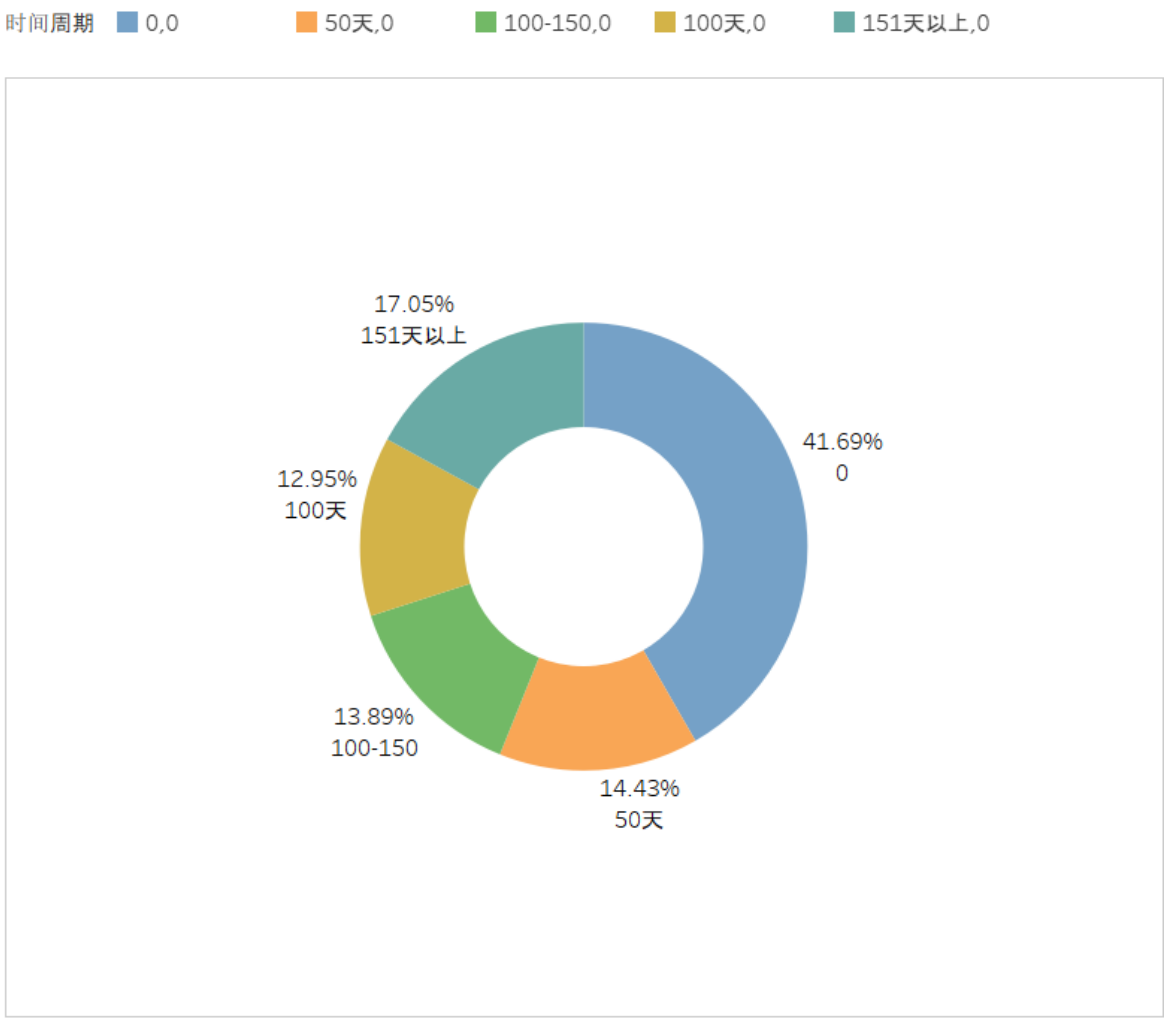


Figure 13: 客户消费周期占比

RFM

通过 Python 根据消费时间，消费金额，消费频率建立 RFM 模型对用户进行价值分类，判断用户的消费时间，金额，频率是否大于平均消费时间，若大于则为 1，小于则为 0，具体分类规则如图 14 所示。

RS分类	FS分类	MS分类	客户类型
高	高	高	高价值客户
低	高	高	重点保持客户
高	低	高	重点发展客户
低	低	高	重点挽留客户
高	高	低	一般价值客户
低	高	低	一般保持客户
高	低	低	一般发展客户
低	低	低	潜在客户

Figure 14: RFM 分类图

用户的占比如图 15 所示,可见当前阶段该电商企业拥有的高价值客户占比较为理想，但是潜在客户，以及一般发展客户仍占多数。

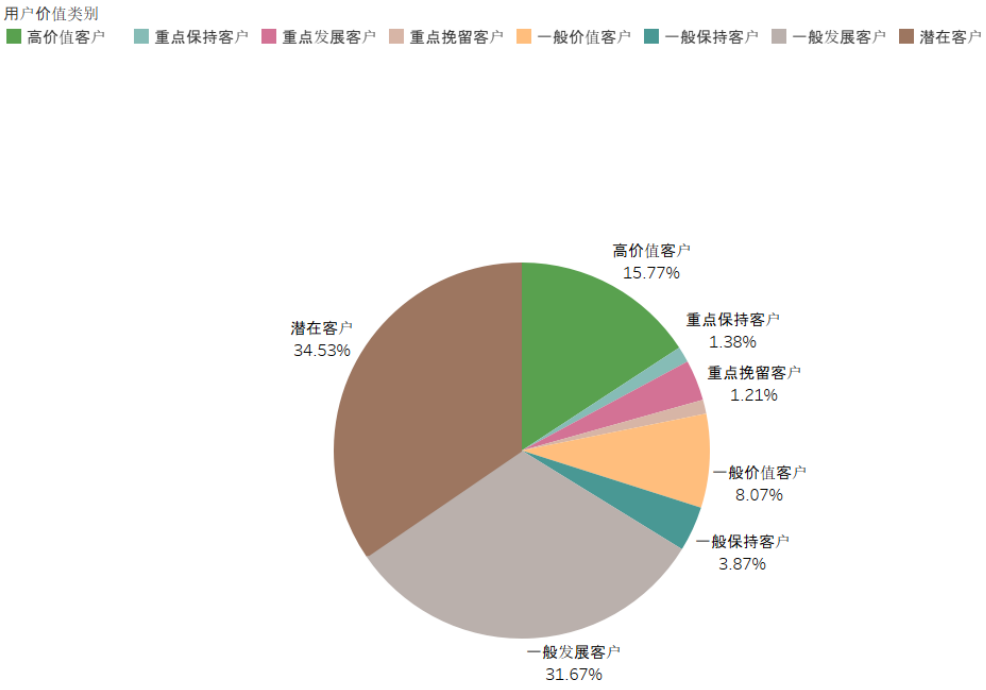


Figure 15: RFM 占比

Conclusion

1. 总的来说，该电商企业消费集中于 11，12 月与 5-6 月，消费的整体时间集中在周中和中午时间，这段时间广告的推广，服务器的维护，以及仓库的存量，企业应该提升关注度。
2. 目前企业主流商品为较为廉价商品，因此未来可以重点放在保证该类商品的质量与种类的同时，提升中高商品的宣传与商品研究，稳固现有客户的同时，开辟新的市场。
3. 该电商公司正处于一个较好的发展趋势，新老客户比较为健康，同时也有一定的忠实，高价值客户，客户主要来源在英国占总市场的 82%。需要注意的是，目前很多用户为仅消费一次，因此如何转化这部分用户的消费周期也是应该解决的问题之一。

推文真伪识别

Abstract

本次研究提出了 5 种机器学习技术基于中世纪的上万条推文进行建模训练，首先分析了数据集的数据质量，数据偏向等数据集基本特征，之后对于数据进行了去除常用词，去除停用词等预处理手段，之后采用词频提取等特征提取方法提取文本数据特征，之后在五种**机器学习模型中进行测试记录。**

Method

首先对于推文数据进行预处理，包括对于不同类型语言进行机器翻译，网址等冗余信息进行数据清理，之后采用 TF-IDF 等特征提取技术对文本数据进行特征提取，随后在多层感知机，文本 CNN，SVM，KNN 和朴素贝叶斯五种模型上进行测试，通过测试集的表现以及识别速度，成本等对五种模型进行排序。

Results and discussions

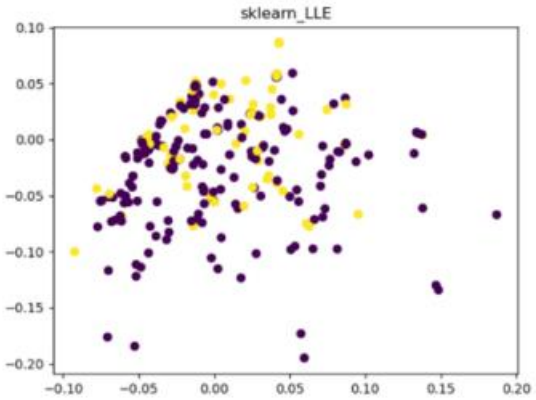


Figure16:LLE 降维

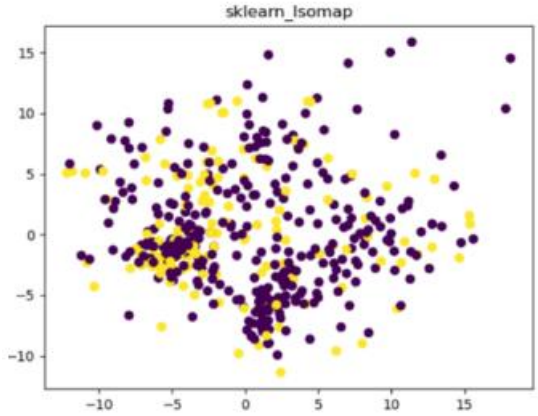


Figure17:LSOmap 降维

五种机器学习模型中，识别准确率最高的是文本 CNN，然后文本 CNN 的训练成本最高，KNN 的识别准确率最差，而且计算量很大，很次 KNN 的整体效果最差，SVM 的识别准确率略好于 KNN，同时训练过程也比较简单，五种模型中整体效果最好的是朴素贝叶斯，该模型被广泛认为适合文本分类任务，实际表现中不断训练速度快，而且准确率高。图 16，17 显示了两种特征降维技术降维后的数据情况，考虑到较多的特征数据会增加模型训练成本，因此合适的特征降维技术也应该应用到本次应用中来。