**黑色星期五销售数据深度分析及用户购买力预测、画像**

**雷云霖**

**何祎君**

**徐琳锴**

**摘要：** 黑色星期五在美国是一个购物狂潮期。在这个时期，各个商店会进行一次大规模的促销，从而引起群众大量购买物资。在购物盛典中将会产生大量的销售数据，这些数据具有非常大的商业价值，对其的挖掘可以给商家提供非常有效的建议让其提高销售额。我们尝试分析了Analytics Vidhya公司公布的2018年黑色星期五部分销售数据，用Apriori算法挖掘商品的关联规则，以此给商家商品捆绑销售提供建议。用随机森林模型用年龄、性别、职业、定居时间、居住城市、婚姻状态等属性对用户购买力进行预测，以此给商店为老顾客促销提供建议，并用线性回归模型进行比较，在十折交叉验证的框架下使用RMSE衡量模型。用随机森林、KNN、朴素贝叶斯和决策树等多种模型用购买记录预测用户性别、年龄，达到部分用户画像的功能，给精准营销提供建议，并比较这些模型的优劣，使用准确率衡量模型。我们对结果进行了合理的可视化表示，方便商户通过结果进行决策。

**关键词：** 黑色星期五，销售数据预测，用户画像，Apriori，随机森林，KNN，朴素贝叶斯，决策树

**正文**

**1．介绍**

黑色星期五（英语：Black Friday）在美国用来指每年感恩节之后的第一天。这一天通常被认为标志着圣诞购物期的正式开始，被看作是每年零售业圣诞销售业绩的晴雨表，也是一年中各个商家最看重也是最繁忙的日子之一，各个商家在这一天会进行大规模的促销，从而引起线下和线下购物热潮。2018年黑色星期五开始九小时后亚马逊平台已经售出100万件玩具、70万件时装，成交额达到6.43亿美元。最终黑色星期五全球销售总额达到7175亿美元。在如此庞大的销售量里，销售数据具有非常大是商业价值。对销售数据的分析和挖掘可以给商家销售策略的策略提供很大的帮助。

我们基于Analytics Vidhya公司公布的2018年黑色星期五部分销售数据的数据集尝试分析和挖掘2018年黑色星期五销售数据。我们的工作分成四部分，第一部分对销售数据进行探索性数据分析(Exploratory Data Analysis，EDA)，从宏观角度对数据集进行统计，并通过可视化方法对部分属性之间的关系进行探索，如年龄、婚姻状态、定居时间和性别等属性和购买力之间的关系。以此找到一些有趣、有潜在价值的知识。在第二部分工作中，我们使用Apriori算法挖掘销售商品的频繁项集。在现实生活中，商家将商品捆绑促销是一种常用并且有效的销售策略，商家可以通过挖掘商品的频繁项集来采取合理有效的捆绑策略。在第三部分工作中，我们尝试建立模型用来预测顾客的购买力，通过这个模型商家可以根据已收集到的顾客信息，预测顾客的购买力，及时向顾客推荐合适价位的商品，精准营销。这是一个回归问题，我们使用了随机森林回归模型，同时用线性回归模型进行比较。并尝试了多种建模的方案。在第四部分工作中，我们尝试对顾客画像，给已经购买过商品的顾客的性别、年龄打上标签，商家可以通过预测的标签进行个性化推荐和促销策略，提高顾客的粘性。可以知道这是一个分类问题，在这里我们尝试了多种模型，如随机森林、KNN、朴素贝叶斯和决策树等，并比较这些模型的优劣。以上这些工作我们都进行了可视化，方便结果的呈现。

本文的第二部分将介绍前人的工作，第三部分将介绍探索性数据分析，第四部分将介绍我们数据预处理的工作，第五部分将介绍我们挖掘商品的关联规则的工作，第六部分将介绍我们预测顾客购买力的工作，第七部分将介绍我们预测用户性别、年龄的工作。

**2．前人工作**

企业挖掘销售数据并以此调整自己的销售策略这是非常常见的现象，基本每个稍具规模的企业都有自己的销售分析部门。随着计算机的发展，对商业数据的挖掘变得高效简单起来。有非常多成功的商业数据分析案例，例如经典的塔吉特超市靠着分析用户所有的购物数据，然后通过相关关系分析得出高中生怀孕的案例。Zhang, Y.等人[2]使用天猫公布的数据预测B2C业务的用户行为。Jia等人[10]对电子商务用户购买力预测，对我们有很大的启示。

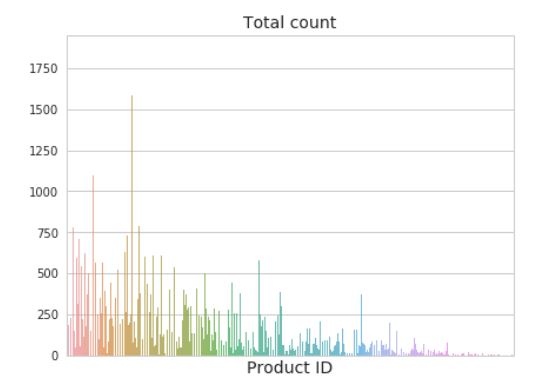
在具体算法上Apriori算法是一个非常经典、简单但十分有效的用于挖掘频繁项集的算法。早在1994年Agrawal等人[1]的工作中就提出了。并由此诞生了购物篮分析，通过对用户购买数据挖掘来找到商品的关联，Brin等人[3]的工作通过动态数据产生关联规则分析购物篮，这对我们的思路有一定启发。

随机森林由Leo Breiman[4]在2001年提出，基于决策树的思想，但性能比单棵决策树好，并能用于回归问题。A Liaw[5]等人的工作对随机森林进行了总结，给出了建模和实现方法。

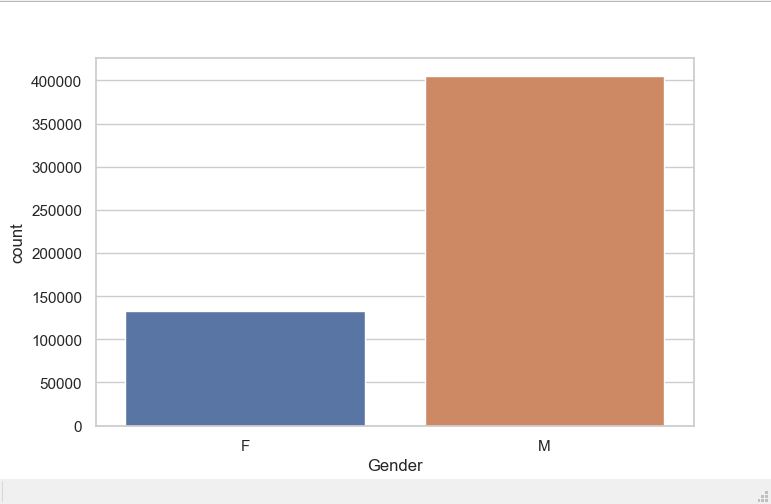
用于解决分类问题的还有Cover等人[6]提出的KNN算法。Breiman[7]提出的CART决策树，Quinlan提出的ID3决策树[8]和C4.5决策树[9]。

**3．探索性数据分析（EDA）**

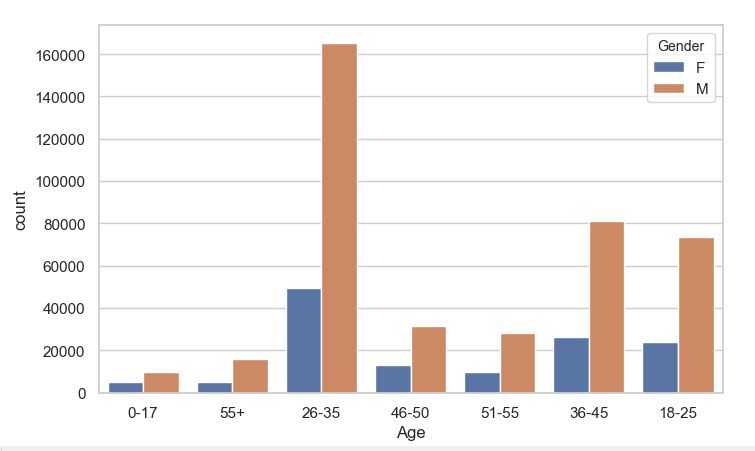
我们首先对数据集进行探索性分析，以此希望找到一些有趣的知识。



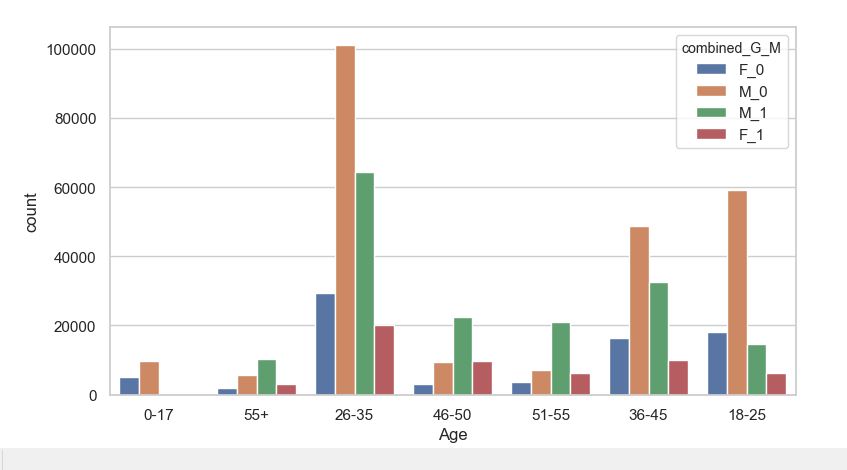
产品ID和购买者数量柱状图，不同颜色代表不同的商品类别。发现有一些类别的商品在黑色星期五销量比较好，比如家电。



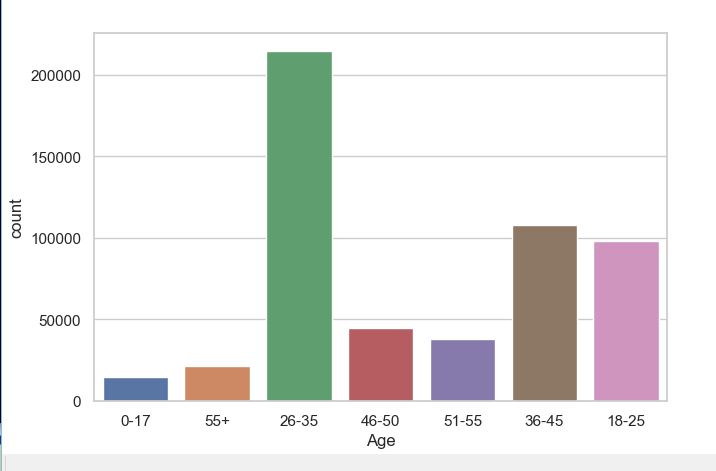
性别和人数的柱状图，发现男性远比女性多，这和中国市场情况刚好相反，我们进一步将性别按年龄拆分。



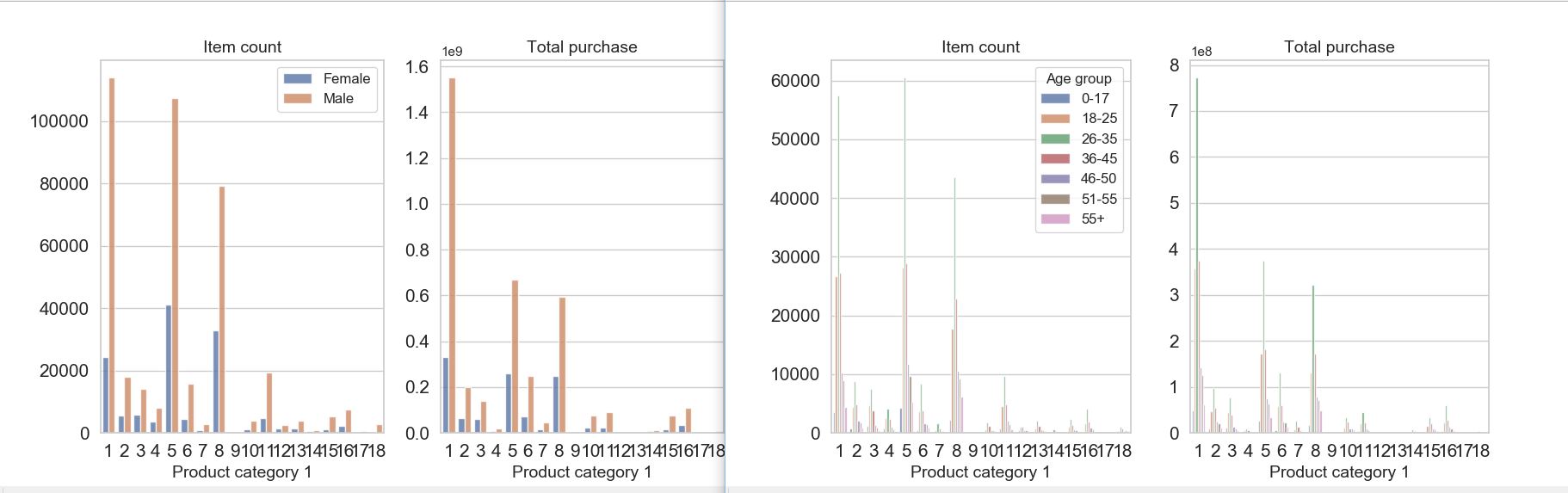
年龄、性别和购买人数的关系，和我们想象中的不一样，在黑色星期五中各年龄段的男性消费者的数量都多于女性。可能是因为部分男性消费者会为女性消费者买单，比如配偶关系。我们进一步分析婚姻状态、性别和购买者关系。



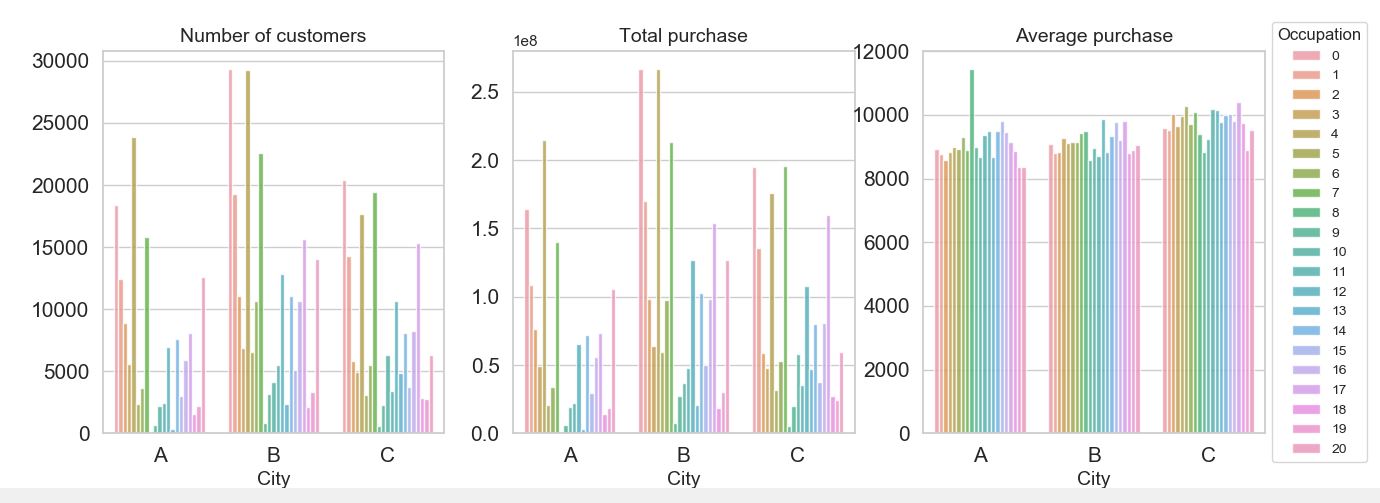
似乎单身男性消费者更多，不能证明部分男性消费者会为女性消费者买单的观点。当然也可能是为了追求女朋友。但没有更多数据支持。但还是可以看出来，不同年龄段消费力不同。



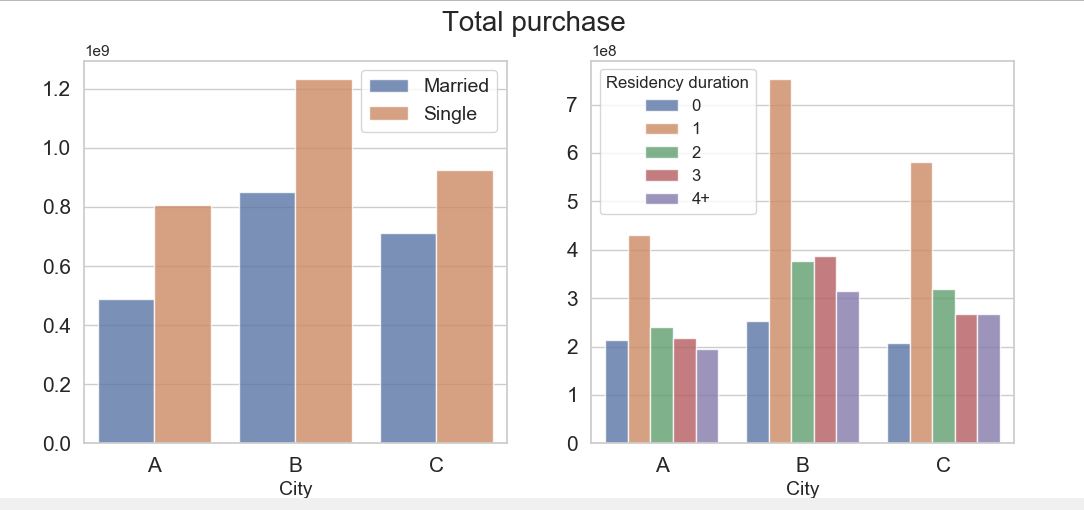
年轻人消费较高。



商品类别和年龄、性别的柱状图。单身男性购买最多。



不同城市的购买力和职业的关系，可以看出来城市B消费力最高。



居住时间、性别和城市的柱状图，可以看出居住一年的顾客消费力最高。或许是需要购置家具。

综上可以得出一些有趣的知识：

1) 男性消费力大于女性

2) 单身消费力大于已婚

3) 在城市居住一年的消费力最高

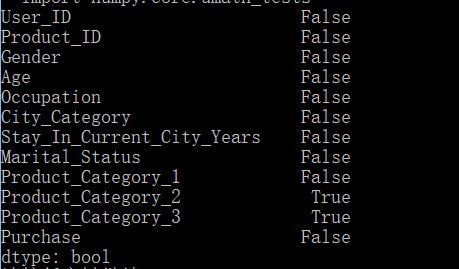
4) 商品分类1中#1，5，8销量最好

5) 产品1可以加大对单身男性的广告

**4．数据预处理**

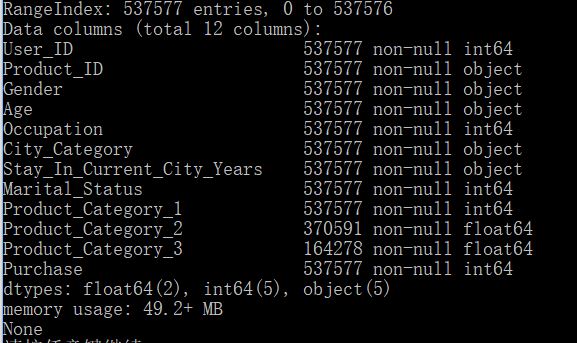
我们对数据集的数据进行初步的清理，主要包括缺失值的处理、数据转换（重新编码）。

我们首先对缺失值情况进行了分析，结果如下。



Product\_category\_2和Product\_category\_3有缺失，我们分析是有些商品没有第二级和第三级分类，所以我们可以将这些值直接用0填充。

之后我们对数据格式进行处理。我们打印出各属性的数据类型，



对于User\_ID和Product\_ID我们直接用序号即可。Genger、Marital\_Status这两个属性有二元性质，直接使用二元数值进行编码。Purchase直接用数值表示即可。对于City\_Category、Stay\_In\_Current\_City\_Years和Product\_Category这些标签属性，我们发现其标签类型有限，所以使用热编码(One-Hot)进行编码。热编码将标签属性拆分成一组二元属性，对于一个记录其对应标签处为1，其余标签处为0。这种编码的好处有便于距离计算。

最后我们按8：2的比例随机划分了训练集和测试集，用于后两个工作的模型训练和评估。

**5．挖掘商品关联规则**

挖掘商品关联规则有利于商家制定捆绑销售策略，非常具有商业价值。我们使用Apriori算法进行分析，Apriori是非常经典的频繁项集挖掘算法。根据频繁项集可以进一步挖掘关联规则。

使用Python3的mlxtend库进行频繁项集和关联规则分析，并使用Python3的nextworkx库和R语言的arulesViz库进行可视化。

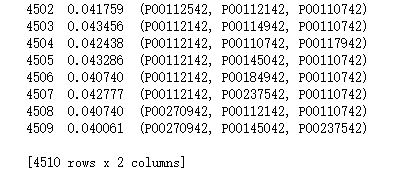
**5.1 规范格式**

统计每一种商品被每一个顾客的购买情况，统计完成后使用mlxtend库的TransactionEncoder函数进行数据格式的规范，将数据转换为mlxtend库要求的形式。

**5.1 频繁项集发现**

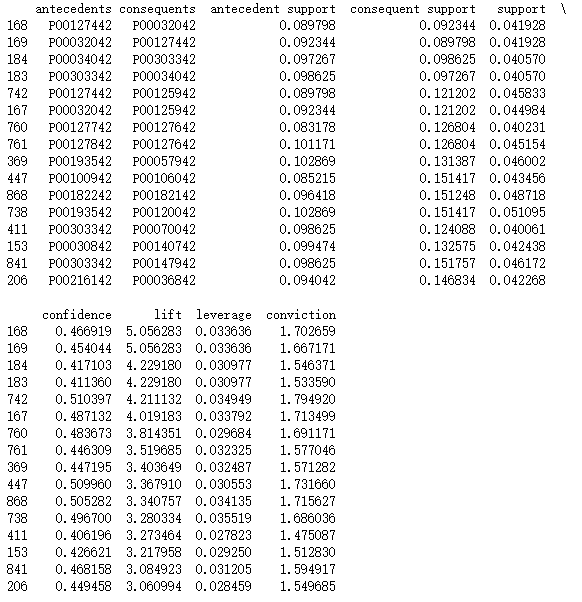
使用apriori算法统计频繁项集。将支持阈值设置为4%。

结果共发现频繁项集4510个。



**5.2 关联规则挖掘**

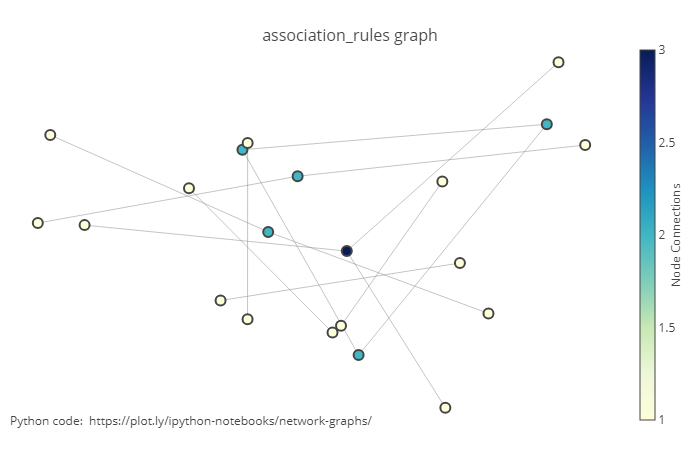
我们基于上面的频繁项集进行分析，产生关联规则。将最小置信度设为40%，最小提升度为3，共产生16条关联规则。



这些规则中没有三元规则，每条规则只涉及到两个商品，这大概可能是数据集中的商品数并不是很多，无法产生多元规则。

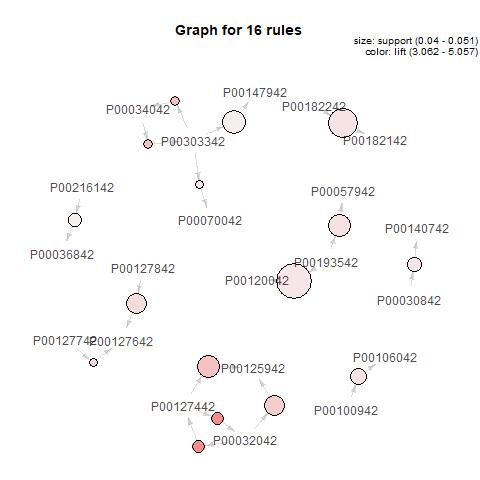
**5.3 关联规则可视化**

我们使用python3中开源的nextworkx复杂网络分析库对我们找到的关联规则进行可视化，将关联规则的项作为点，提高度作为边绘制。



实现出来的效果可以支持动态可交互式的查看，但解释性不强，无法知道哪件商品支持哪件。

于是我们又使用R语言中的arulesViz库进行可视化。



在这张图中，每一个点代表一个关联规则，其中点的大小代表规则的置信度，点颜色深浅代表支持度高低，指向点的商品代表关联规则的前项，点指向的商品代表后项。具有较好的可视化效果。

**6．预测顾客购买力**

预测顾客购买力对于商家精准营销策略有非常大的实用价值，我们基于Analytics Vidhya公司公布的2018年黑色星期五销售数据。

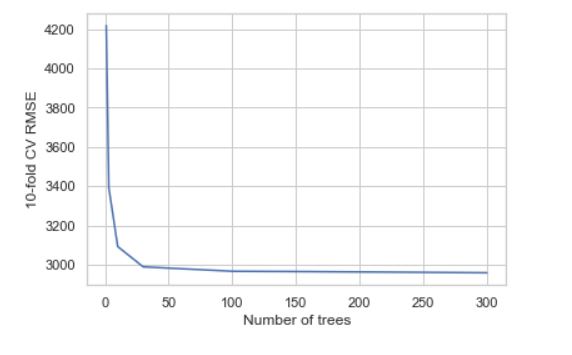
由于购买力是一个连续数值，所以对购买力的预测是一个回归问题。常用的回归问题算法有随机森林、线性回归等。一般来说随机森林的效果更好。随机森林是一个包含多个决策树的分类器，并且其输出的类别是由个别树输出的类别的众数而定，随机森林能处理大量的输入变量，具有很好的准确度和鲁棒性。于是我们尝试建立一个随机森林的模型用于通过已收集的顾客信息来预测顾客购买力。并使用线性回归来对比模型。对于模型的评价我们在十折交叉验证的框架下使用均方根误差（RMSE）来衡量。

我们使用Python3中sklearn库中封装的RandomForestRegressor来搭建模型。

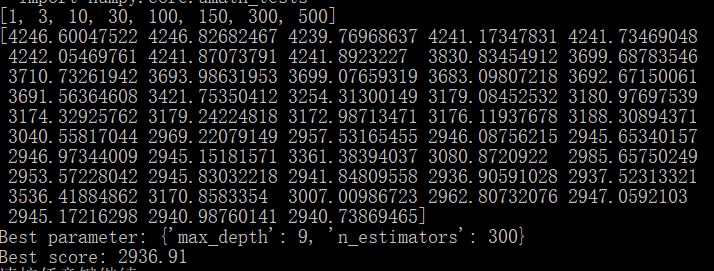
**6.1 随机森林**

**6.1.1 模型使用以及参数调整**

我们将通过除购买力之外的所有属性来预测，通过对随机森林树的数量参数进行网格搜索，得到下图

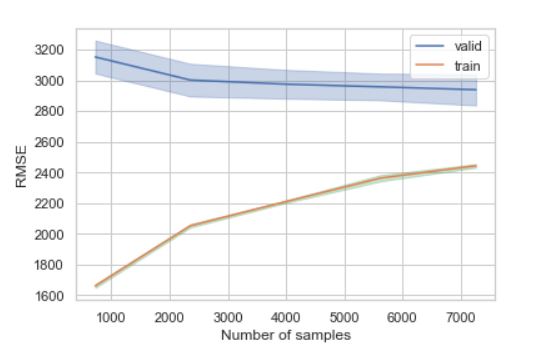


当树的数量为300左右时最优，但是在150时就已经收敛了，于是进一步网格搜索最大搜索深度



于是找到最佳参数，max\_depth:9 n\_estimatore:300。

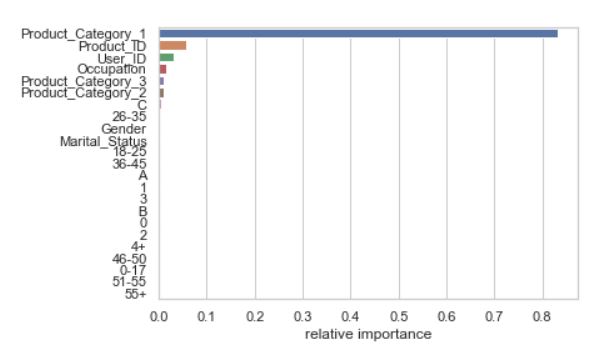
在十折交叉验证框架下，RMSE如下



在测试集中RMSE 2839.680。

**6.1.2 特征依赖情况**

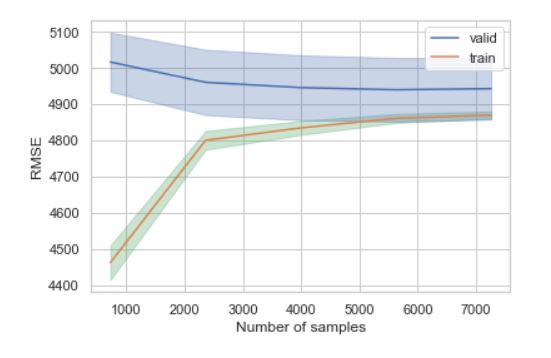
我们将模型中属性的影响力展示出来



我们发现这个模型预测对商品类别高度依赖，而在现实情况下，商家是很难知道顾客购买的商品信息，所以我们应该尝试使用低等数据，如年龄、性别、婚姻状态、职业、居住城市、定居时间等属性来预测。并且我们应该将每个顾客购买记录合并，预测总的购买力。

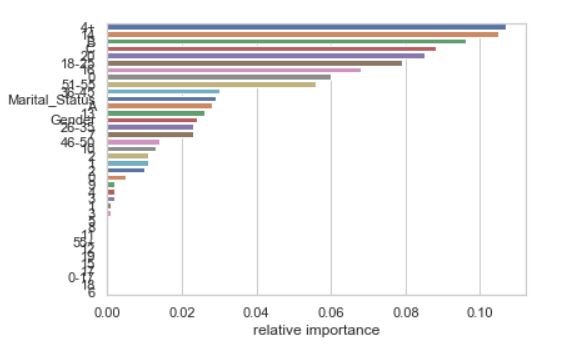
**6.1.4 模型调整**

仍然使用随机森林模型，去除商品类别、商品ID等属性，将顾客购买记录进行合并去重，按照6.1.1方法重新构建，在十折交叉验证框架下



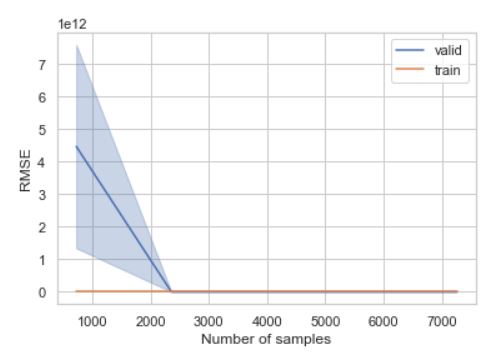
在测试集中RMSE 4877.794。

属性影响力如下图，定居时间和年龄的影响力最大。



**6.1.5 和线性回归模型对比**

我们使用线性回归模型按照同样的方法构建模型，用于对比，



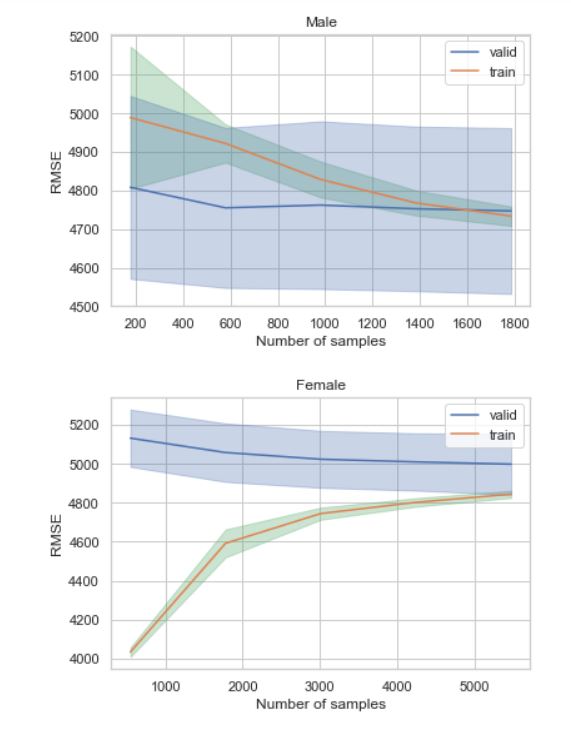
在测试集中RMSE：4885.545。

结果很令人失望，感觉随机森林的结果并没有预想中的好。

**6.1.6 进一步改进模型**

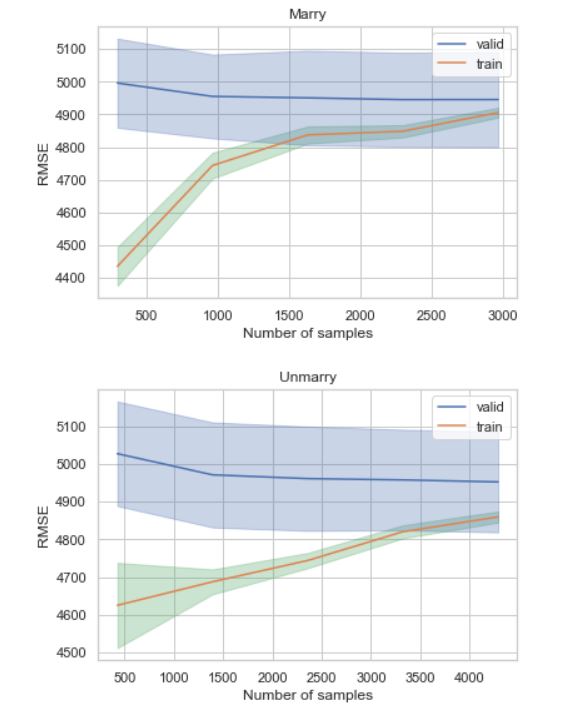
进一步分析，我们发现在随机森林模型中，或许是因为有些属性影响力不够导致了预测结果的不好，于是我们探索将数据集分别拆分为男性、女性和已婚、未婚来重新搭建模型。

我们将数据集进行拆分并按照6.1.1方法进行训练。按性别划分结果，



在测试集中男性测试集中RMSE=4706.875，女性测试集中RMSE=5132.161。

按婚姻状态划分，结果如下



在已婚测试集中RMSE=5047.171未婚测试集中RMSE=5017.173。

划分后RMSE反而变大了。

**6.1.7 结论**

在这个数据集中预测顾客购买力的效果不是很好，但我们观察到，当我们按性别划分后在男性集中模型的效果变好，而在女性集中变差，根据探索性分析中的结果我们知道男性顾客远大于女性顾客，可能是因为女性集过小导致的效果变差。所以我们认为效果的不好最有可能还是数据集太少，如果有更大的数据集按性别划分后使用随机森林模型可能会取得很好的效果。

**7．顾客画像**

**7.1用户画像概念**

用户画像又称用户角色，作为一种勾画目标用户、联系用户诉求与设计方向的有效工具，用户画像在各领域得到了广泛的应用。用户画像最初是在电商领域得到应用的，在大数据时代背景下，用户信息充斥在网络中，将用户的每个具体信息抽象成标签，利用这些标签将用户形象具体化，从而为用户提供有针对性的服务。

**7.2用户画像的优点**

用户画像可以使产品的服务对象更加聚焦，更加的专注。在行业里，我们经常看到这样一种现象：做一个产品，期望目标用户能涵盖所有人，男人女人、老人小孩、专家小白、文青......通常这样的产品会走向消亡，因为每一个产品都是为特定目标群的共同标准而服务的，当目标群的基数越大，这个标准就越低。换言之，如果这个产品是适合每一个人的，那么其实它是为最低的标准服务的，这样的产品要么毫无特色，要么过于简陋。

纵览成功的产品案例，他们服务的目标用户通常都非常清晰，特征明显，体现在产品上就是专注、极致，能解决核心问题。比如苹果的产品，一直都为有态度、追求品质、特立独行的人群服务，赢得了很好的用户口碑及市场份额。又比如豆瓣，专注文艺事业十多年，只为文艺青年服务，用户粘性非常高，文艺青年在这里能找到知音，找到归宿。所以，给特定群体提供专注的服务，远比给广泛人群提供低标准的服务更接近成功。其次，用户画像可以在一定程度上避免产品设计人员草率的代表用户。代替用户发声是在产品设计中常出现的现象，产品设计人员经常不自觉的认为用户的期望跟他们是一致的，并且还总打着“为用户服务”的旗号。这样的后果往往是：我们精心设计的服务，用户并不买账，甚至觉得很糟糕。

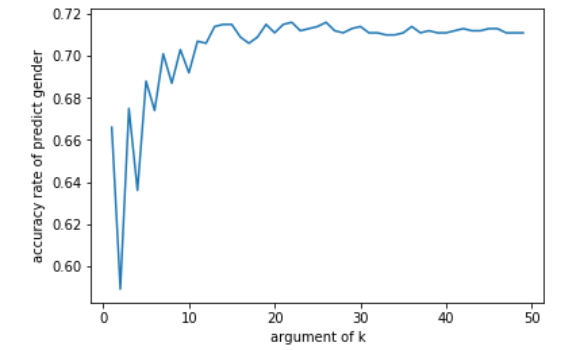
**7.3实际建模**

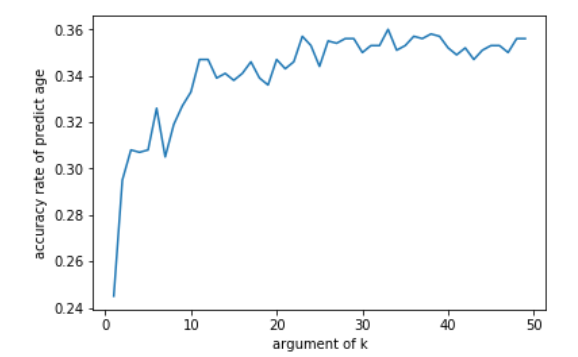
同样是基于Analytics Vidhya公司公布的2018年黑色星期五销售数据进行建模，根据数据的特点，主要分析的用户标签为年龄和性别；我们总共采用了KNN、决策树、朴素贝叶斯和随机森林四种算法模型，并分别对其结果进行分析和比较。直接使用Python3中Sklearn库进行模型的实现。

**7.3.1 K-NearestNeighbor算法**

kNN算法的核心思想是如果一个样本在特征空间中的k个最相邻的样本中的大多数属于某一个类别，则该样本也属于这个类别，并具有这个类别上样本的特性。该方法在确定分类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别。

所以该算法的核心参数为所选择的K值的大小，通过记录不同的k值所得到的对应模型的分类准确率可得到下图：





可以看出当K值在20~50范围之间时，准确率就会达到最大并趋于稳定。

其中性别预测准确率最大约为71.5%，年龄预测准确率最大约为35.5%

**7.3.2 决策树算法**

决策树(Decision Tree）是在已知各种情况发生概率的[基础](https://baike.baidu.com/item/%E5%9F%BA%E7%A1%80/32794)上，通过构成决策树来求取净现值的[期望](https://baike.baidu.com/item/%E6%9C%9F%E6%9C%9B/35704)值大于等于零的概率，评价项目风险，判断其可行性的决策分析方法，是直观运用概率分析的一种图解法。由于这种决策分支画成图形很像一棵树的枝干，故称决策树。在机器学习中，决策树是一个预测模型，他代表的是对象属性与对象值之间的一种映射关系。Entropy=系统的凌乱程度，使用于算法[ID3](https://baike.baidu.com/item/ID3), [C4.5](https://baike.baidu.com/item/C4.5)和C5.0生成树算法使用熵。这一度量是基于信息学理论中熵的概念。分类树（决策树）是一种十分常用的分类方法。他是一种监管学习，所谓监管学习就是给定一堆样本，每个样本都有一组属性和一个类别，这些类别是事先确定的，那么通过学习得到一个分类器，这个分类器能够对新出现的对象给出正确的分类。这样的机器学习就被称之为监督学习。

我们使用CART进行预测。

将训练数据载入到我们建立的的模型中得到了如下结果：

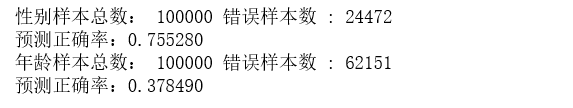


**7.3.3 朴素贝叶斯算法**

素贝叶斯法是基于贝叶斯定理与特征条件独立假设的分类方法 [1]  。最为广泛的两种分类模型是决策树模型(Decision Tree Model)和朴素贝叶斯模型（Naive Bayesian Model，NBM）。

和决策树模型相比，朴素贝叶斯分类器(Naive Bayes Classifier,或 NBC)发源于古典数学理论，有着坚实的数学基础，以及稳定的分类效率。同时，NBC模型所需估计的参数很少，对缺失数据不太敏感，算法也比较简单。理论上，NBC模型与其他分类方法相比具有最小的误差率。但是实际上并非总是如此，这是因为NBC模型假设属性之间相互独立，这个假设在实际应用中往往是不成立的，这给NBC模型的正确分类带来了一定影响。

将训练数据输入到模型中可得结果：



**7.3.4 随机森林算法**

随机森林是一个包含多个决策树的[分类器](https://baike.baidu.com/item/%E5%88%86%E7%B1%BB%E5%99%A8)， 并且其输出的类别是由个别树输出的类别的众数而定。

根据下列步骤来建造每棵树：

1、用N来表示训练用例（样本）的个数，M表示特征数目。

2、输入特征数目m，用于确定决策树上一个节点的决策结果；其中m应远小于M。

3、从N个训练用例（样本）中以有放回抽样的方式，取样N次，形成一个训练集（即bootstrap取样），并用未抽到的用例（样本）作预测，评估其误差。

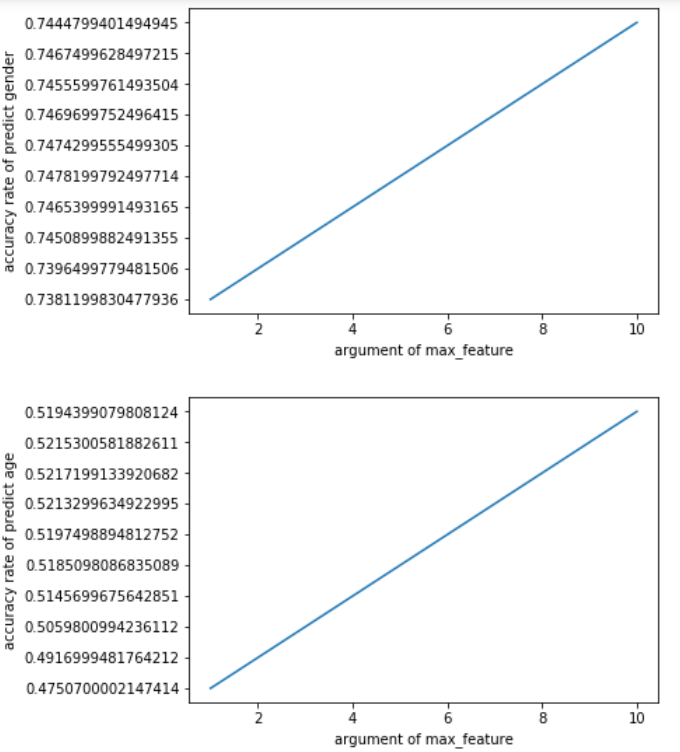
4、对于每一个节点，随机选择m个特征，决策树上每个节点的决定都是基于这些特征确定的。根据这m个特征，计算其最佳的分裂方式。

5、每棵树都会完整成长而不会剪枝，这有可能在建完一棵正常树状分类器后会被采用）。

在所建立的随机森林模型中有以下一些关键参数



我们针对max\_features(最适属性划分值的上限)进行调整得到如下视图：



**7.4 结论**

对于性别的分类朴素贝叶斯的表现最好，但与其他几种算法相差不大；对于年龄的分类随机森林算法表现最好，且显著优于其他几种算法。

**8．Reference**

[1] Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (Vol. 1215, pp. 487-499).

[2] Zhang, Y., Pang, L., Shi, L., & Wang, B. (2014). Large scale purchase prediction with historical user actions on B2C online retail platform. *arXiv preprint arXiv:1408.6515*.

[3] Brin, S., Motwani, R., Ullman, J. D., & Tsur, S. (1997). Dynamic itemset counting and implication rules for market basket data. *Acm Sigmod Record*, *26*(2), 255-264.

[4] Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

[5] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18-22.

[6] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), 21-27.

[7] Breiman, L. (2017). *Classification and regression trees*. Routledge.

[8] Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*(1), 81-106.

[9] Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.

[10] Jia, R., Li, R., Yu, M., & Wang, S. (2017, July). E-commerce purchase prediction approach by user behavior data. In *Computer, Information and Telecommunication Systems (CITS), 2017 International Conference on* (pp. 1-5). IEEE.