在线旅行社(OTA)业务数据分析

张 晗

1 介绍

酒店搜索引擎能否基于用户查询条件将匹配度最高的酒店推荐给用户将直接影响到访问者对网站的信赖度，用户可能因为短时间内（查询结果首页中）无法找到满意的酒店访问其他网站从而造成潜在的订购流失。因此订购网站希望通过用户的浏览订购记录以及查询条件预测其满意度最高的酒店，用户对于每一条查询结果的满意度可以通过是否预订或点击查看进行衡量。酒店订购网站*Expedia*提供的历史数据包括用户查询、浏览以及预订的消费记录，这份报告将通过给定数据分析与最终订购有较强关系的变量，并尝试构造能够显著影响用户反馈的特征变量。

2 数据集描述

数据集包括由40万次查询生成的990万条数据，每次查询*Expedia*平均生成25条酒店查询结果，共54个变量。每条数据包含用户对搜索结果的反馈变量：是否订购(*booking_bool*)，是否点击(*click_bool*)，订购总金额(*gross_booking_usd*)，由*Expedia*系统生成的查询结果排序(*position*)。其他变量主要可以分为用户搜索信息相关变量、酒店特征相关变量、用户信息相关变量，下表列举了其中部分变量：

prop_starrating	酒店星级
prop_review_score	用户评分
prop_brand_bool	是否连锁
price_usd	预订价格
prop_log_historical_price	历史预订均价
*prop_location_score*1	地点需求度1
*prop_location_score*2	地点需求度2
prmotion_flag	是否优惠
srch_room_count	房间预订数
srch_destination_id	酒店查询地点
position	酒店排序位置

从图一中可以看到数据集中有近一半变量存在不同程度的缺失，大部分缺失率超过60%，缺失数据中大部分为竞争网站与Expedia的价格与订购对比信息(*comp*6_*inv*, *comp*6_*rate*)、用户曾经订购酒店的平均价格和星级(*visitor_hist_starrating*)、酒店所在地需求度、酒店在查询列表中被点击概率的对数(*srch_query_affinity_score*)以及用户支付的订购总费用(*gross_bookings_usd*)。接下来考察主要变量之间的相关性，图二为变量之间相关系数热力图，颜色越深代表相关性越大，对角线代表与自身的相关系数：

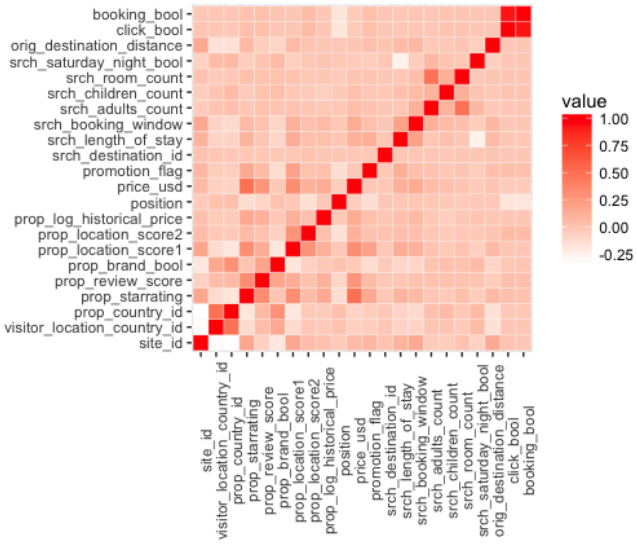

Fig 2: 变量相关系数热力图

图二中相关系数为1的变量组合为是否被点击与是否被预订，因为预订酒店代表肯定已被点击查看。其中酒店价格、酒店星级、酒店用户评分、酒店所在地需求度、酒店历史平均价格之间有较强相关性。另外酒店所在地需求度(*prop_location_score*2)与多个变量有较强相关性，*Expedia*未明确给出酒店所在地需求度计算方法，但可以合理猜测该变量值应该和其所在

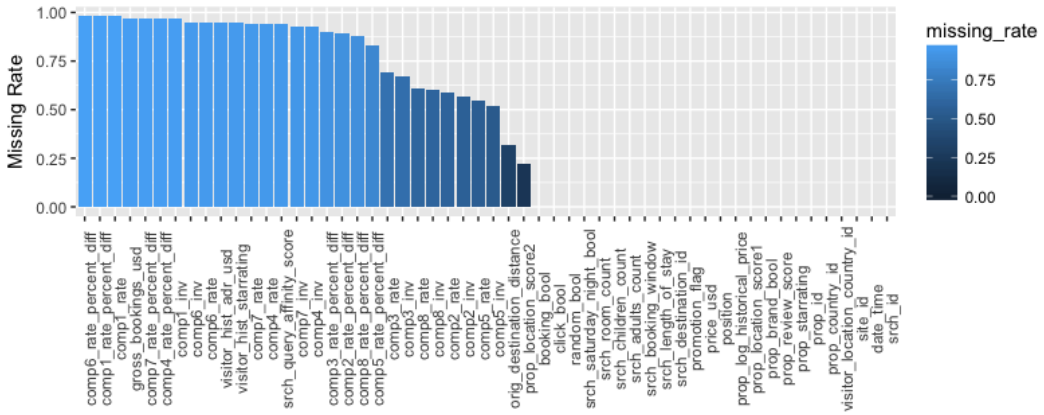

Fig 1: 数据缺失率

地与特定场所的距离以及是否处在某一特定区域等因素相关，如酒店与海滩的距离、是否在市中心等等，满足这些条件酒店的价格大概率会相对较高。另一点值得注意的是该酒店是否被预订(*booking_bool*)与酒店当前预订价格(*price_usd*)以及是否有优惠(*promotion_flag*)相关性较低，除了酒店价格可能并不直接影响预订率的直观理解之外，另外一个可能的原因可能和*Expedia*系统自身机制有关。通过对同一酒店(*prop_id*)在数次搜索数据的观察，如果当酒店在之前搜索未被预订甚至未被查看，在下一次搜索中*Expedia*有一定概率会降低该酒店价格并附加促销标签；相反如果在之前数次搜索中该酒店订购率较高，系统可能会适当增加预订价格并不会提供优惠。

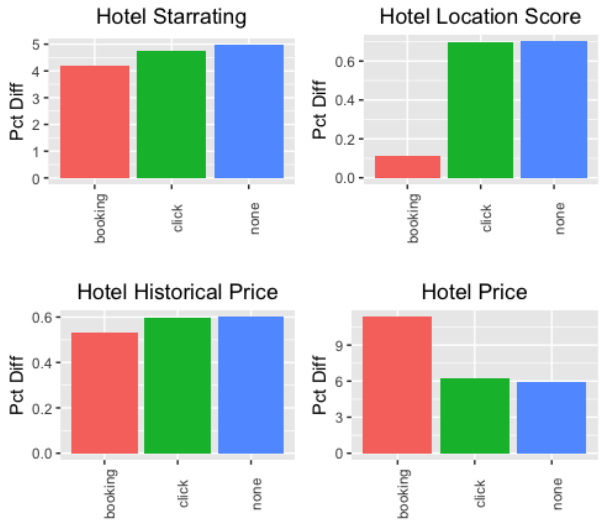

Fig 3: 订购/点击差异对比

是否被订购或被点击查看是衡量用户满意度最为重要的指标，图三显示了在被客户预订或未预定以及被点击或未被点击的情况下，酒店平均价格(*price_usd*)、用户对酒店评分(*prop_review_score*)、酒店星级(*prop_starrating*)以及酒店所在地需求度(*prop_location_score2*)的绝对差异百分比。可以看到，被预订/点击查看过的酒店与未被预订/查看酒店在价格与星级上有较为显著的差异，酒店历史平均预订价格及所在地需求度没有显著受到用户是否预订或查看的影响。此类对于不同用户反馈变量均值差异较大的变量能为推荐位置预测提供更多有效信息。

3 数据预处理

*Expedia*给定的数据中大部分变量分布有较长的尾部且含有缺失数据。下面说明对缺失数据的处理方式：

- 酒店相关的历史数据平均值：对比有缺失值酒店的所在地需求度(*prop_location_score2*)、酒店点击率对数值(*srch_query_affinity_score*)以及用户对酒店评分(*prop_review_score*)的均值都要低于曾经被订购过酒店上述属性的平均值，所以变量数据缺失酒店的这些属性也应该相对偏低，实际中将已有数据的25%分位值分配给酒店所在地需求度(*prop_location_score2*)、酒店点击率对数值(*srch_query_affinity_score*)。

酒店订购率对比		
含缺失值变量	缺失	未缺失
所在地需求度	1.48%	3.15%
酒店点击率对数值	2.73%	3.21%

酒店点击率对比		
含缺失值变量	缺失	未缺失
所在地需求度	2.62%	4.91%
酒店点击率对数值	4.32%	4.77%

- 与竞争网站的对比信息：下面的表格可以看到当Expedia与另外三家竞争网站酒店价格存在差异时各自的订购率和点击率并无显著差异，所以可以假设Expedia与竞争网站的酒店价格相等并将$comp1_rate$至$comp7_rate$的缺失值设置为0，同时也可以将价格差异百分比$comp_rate_percent_diff$设置为0。类似可以检查到$comp_inv$在取不同值时订购率及点击率无明显差异，可以将$comp_inv1$至$comp_inv7$设置为0。

酒店订购率对比				
酒店价格	相等	小于	大于	缺失
comp1_rate	2.2%	2.5%	2.7%	2.6%
comp2_rate	2.5%	2.9%	3.8%	2.4%
comp3_rate	2.9%	2.9%	3.8%	2.8%

酒店点击率对比				
酒店价格	相等	小于	大于	缺失
comp1_rate	3.6%	4.4%	4.7%	4.3%
comp2_rate	3.8%	4.2%	5.1%	4.4%
comp3_rate	3.9%	4.1%	5.2%	4.3%

- 用户曾订购酒店平均价格与星级：首先计算未缺失的用户曾经订购酒店平均星级与酒店当前星级的差，以下表格是不同星级差异区间对应的酒店订购率与点击率。由于当不同差异区间的订购率具有单调性时该特征变量对预测会更加有效，并且*Expedia*系统趋向于将与用户较为匹配酒店推荐给用户，可以将用户曾订购酒店星级平均值缺失值设置为与酒店当前星级绝对误差在1.5以内，$|starrating - visitor\_hist\_starrating| < 1.5$。

酒店订购率对比					
0 - 1	1 - 2	2 - 3	3 - 4	4 - 5	缺失
4.3%	2.8%	1.9%	1.8%	1.9%	2.8%

类似的通过计算用户曾经订购酒店价格平均值与酒店当前订购价格的差异不同区间对应的订购率与点击率，将此缺失值设置为$|log(price_usd) - visitor_hist_adr_usd| < 0.15$

酒店点击率对比					
0 - 1	1 - 2	2 - 3	3 - 4	4 - 5	缺失
4.9%	3.6%	2.6%	2.2%	2.2%	4.5%

除填补缺失数据之外，需要尽量减少极值(*outlier*)对预测的干扰。图四为部分选取变量的箱图，数据集中大部分变量存在较大比重的极值。处理方法是对该变量的最大值设定上限，选择尾部1%或5%分位变量值，根据该变量数据实际分布的情况而定。

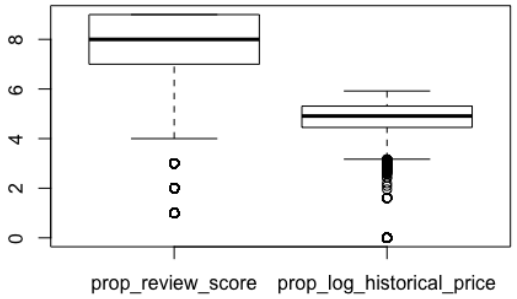

Fig 4: 部分特征变量极值分布情况

4 特征工程

*Expedia*数据集中包含54个特征变量，并非所有变量都会显著影响用户点击与订购行为，部分特征变量需要经过适当处理才能有效为预测提供信息。额外构造的特征变量主要包括：1）以各类*id*为分组计算的酒店和查询相关变量的平均值、方差；2）经过标准化的酒店数值型变量；3）基于列的排序特征；4）复合型特征。

特征变量有效性的判断依据为该特征对应订购率和点击率是否具有单调性，在下一节可以看到预测模型为基于树的排序模型，具有单调性的特征变量更有利于树模型进行空间划分。例如可以通过如下公式计算对于订购行为的酒店星级单调性：$|prop\_starrating - mean(booked\_prop\_starrating)|$，即当前酒店星级与被订购酒店平均星级的差，类似可以计算酒

店星级关于点击的单调性，下图中的酒店星级特征变量对目标变量并不具有良好的单调性。

Fig 5: 特征单调性

下面分别介绍主要的四类特征变量：

- 以id类变量进行分组统计的特征：id类变量作为随机类变量本身不会对目标变量产生显著影响，但经过对同一搜索id或酒店id下的数值型变量计算均值或方差，这些统计量可以视为某一特定搜索或酒店的独特特征，即这些统计量更完整的描述了特定酒店或查询的特点。此类特征变量另一个作用是在一定程度上减少了单条数据中可能存在的误差，某些数据明显超过了方差尺度上的合理波动范围。实际中将基于查询($srch\_id$)、酒店($prop\_id$)及查询目的地($srch\_destination\_id$)进行分组统计量的计算。

- 标准化特征变量：由于酒店部分特征在不同地区、不同时段和不同的用户查询中会有一定的差别，对酒店重要特征进行正规化可以弱化区域、时段等因素的干扰，主要通过$date_time$、$srch\_id$、$srch\_destination\_id$、$prop_id$对酒店价格和所在地需求度进行正规化。下图为通过$prop_id$标准化的酒店价格对数值对于订购率和点击率的单调性。

Fig 6: 特征单调性

- 基于列($listwise$)的特征：在同一查询内酒店特征的排名，此类特征包括酒店价格、酒店星级和用户评分在查询内的排序。例如该查询结果共五家酒店，订购价格为200、190、170、150、120，则价格190的酒店对应该变量值为2。

- 复合型特征：此类特征变量包括现有酒店订购价格与历史平均订购价格的差价、酒店星级与同一查询内酒店星级平均值的差等。

5 模型与误差评估

因为用户反馈对查询结果排序有直接影响，加入预测的用户反馈信息可以为排序提供更多有效的信息。但这样需要训练多个模型并导致计算时间过长，可能不利于网站进行实时更新与排序预测。这里使用基于决策树的$LambdaMART$[1]，该模型是为处理检索排序问题而特别设计的，下面简单介绍模型的原理。假设q_i为第i次用户查询，$H_i = \{h_{i1}, ..., h_{im}\}$为与查询$q_i$相关的酒店列表，$s_i = \{s_{i1}, ..., s_{im}\}$为这些酒店的排序得分，模型形式如下：

$$f(q_i, H_i) := \hat{s}_i = \{\hat{s}_{i1}, ..., \hat{s}_{im}\}$$

预测出的关于查询q_i的酒店排序得分$\hat{s}_i$与根据实际用户反馈得出的排序得分越接近，表示预测精确度越高。$LambdaMART$优化的损失函数如下：

$$\min_{\hat{s}_{ij},\hat{s}_{ik}} \sum_i \sum_{jk\in H_i} \left|\Delta Z_i^{jk}\right| log(1+e^{\sigma y_{ijk}(\hat{s}_{ij}-\hat{s}_{ik})})$$

$$\left|\Delta Z_i^{jk}\right| = \left|ndcg(\hat{s}-\hat{s}_i) - ndcg(\hat{s}_i^{jk}-\hat{s}_i)\right|$$

其中当s_{ij}大于、小于或等于s_{ik}时，y_{ijk}的值分别为1、0及-1。$ndcg(\hat{s}-\hat{s}_i)$代表$\hat{s}_i$的归一化折损累计增益，$ndcg(\hat{s}_i^{jk}-\hat{s}_i)$代表当元素$j$、$k$交换位置时$\hat{s}_i$的归一化折损累计增益。归一化折损累计增益(*Normalized Discounted Cumulative Gain*)为信息检索中常用的衡量排序精度的度量方式，以下为酒店查询结果排序中折损累计增益的计算方法：

$$dcg_n = \sum_{i=1}^{n} \frac{2^{rel_i}}{\log_2(i+1)}, \quad rel_i = \begin{cases} 5, & booked \\ 1, & clicked \\ 0, & none \end{cases}$$

其中i代表在一组搜索结果中的排位，排位越靠前的项获得的权重越大。归一项的计算方式与上式类似，但排序设置为能够最大化折损累计增益的顺序，即rel_i相关度高的项排位靠前。归一化折损累计增益的范围在0到1之间，相比于标准误差(*RMSE*)与指数损失函数(*Exponential Loss*)，可以更好评估查询结果排序准确度。R中gbm[2](*Generalized Boosting Regression Machine*)是对*LambdaMart*的一个实现，使用时将目标变量分布设置为*pairwise*并指定查询序号($srch\_id$)为酒店所属组别的主键。可以通过计算特征变量对损失函数的贡献程度，检验已有与新增特征变量的相对重要性，下图列举了对预测影响最大的40个特征变量及其相对重要性得分。

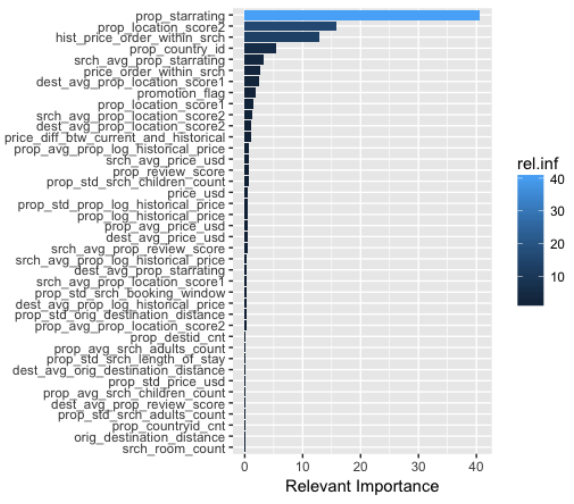

Fig 7: 排序相关变量重要性对比

图五为各特征变量对查询结果排序的相对重要性。可以看到酒店所在地需求度($prop\_location\_score2$)和酒店星级($prop\_starrating$)对排序有重要影响，其他对预测结果有较大影响的特征变量包括：同一查询中酒店当前订购价格($price_usd$)及历史平均价格排名($hist\_price\_order\_within\_price$)、同一查询中所在地需求度的均值($srch\_avg\_prop\_location\_score1$)、是否存在优惠($promotion\_flag$)、同一目的地中酒店所在地需求度平均值($dest\_avg\_prop\_location\_score1$)以及酒店历史价与现价的差。

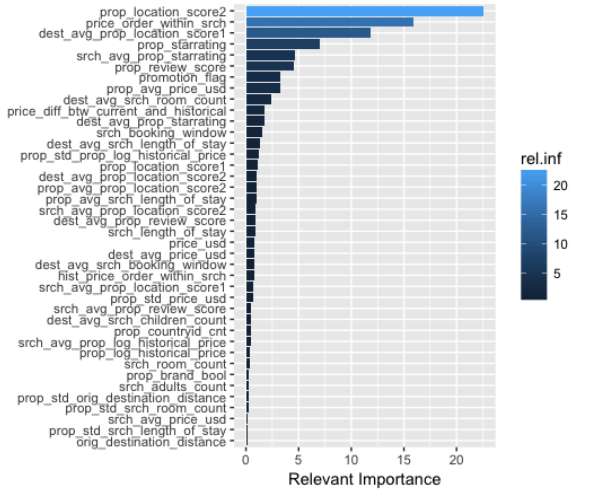

Fig 8: 订购相关变量重要性对比

在检验哪些特征变量对酒店排列顺序会产生较大影响的同时还检验了对用户反馈产生重要影响的特征变量，因为理想的酒店推荐顺序与用户是否订购该酒店直接相关，图六中可以看到两图中影响力较大的变量重叠率非常高，虽然具体相对重要性得分排序有一定差异。

由于数据量较大，实际处理时对查询序号($srch\_id$)进行不放回随机抽样，共抽取了10%查询序号及其相关数据（约100万条）。训练时使用交叉验证，抽样获得的数据集被分为五份，每次训练使用80%的数据，剩余的20%作为测试数据，共生成五组模型和预测结果。以$srch_id$为组别标识符，可以计算每一次查询结果排序的归一化折损累计增益，根据$srch_id$对所有查询排序的归一化折损累计增益求均值即可以度量预测的精度，该平均值越大代表精度越高。

实际中通过生成的五组$LambdaMART$模型计算预测结果的归一化折损累计增益平均值为0.45。

6 总结

通过使用单一模型$LambdaMART$实现了对$Expedia$酒店查询结果的排序。在模型训练之前根据缺失数据对应的订购率及点击率的单调性对其进行了估计和填充，并通过id类变量分组、特征标准化、基于列的变量排序对部分特征进行了处理与再构造。总体来说，所在地需求度、酒店星级以及订购价格会较大程度影响到推荐顺序；重要程度较高的变量中有很多是根据id类变量$prop_id$、$srch\_destination\_id$、$srch_id$分组得到的其他特征统计量（均值、方差、标准差）。酒店相关的属性（星级、订购价格、所在地得分及其衍生出的特征）比查询信息（预订房间数、成人数、订购天数等）相关特征对推荐顺序的影响更大。这一事实基本符合直觉，因为最终排序是根据以酒店为个体划分的，通过$prop_id$ 构造的特征变量可以帮助模型对酒店进行更细致的区分。

在以上模型中一个尚未考虑的问题是用户反馈行为存在位置偏差($position\ bias$)，即用户偏向于点击查询结果中位置靠前的酒店，部分原因可以归结为位置靠前的推荐与用户匹配度更高，但即使查询结果的顺序为随机生成，位置靠前酒店的点击率依然明显偏高，说明大多数用户会忽略靠后的酒店而主要查看首页中的结果。实际中可以考虑对位置靠前但未被点击的酒店施加更大的惩罚或对被点击的位置靠后酒店给予更大得分奖励，以消除位置偏差对预测的干扰。

参考文献

[1] C. Burges (2010) "From RankNet to LambdaRank to LambdaMART: An Overview," $Microsoft\ Research\ Technical\ Report\ MSR-TR-2010-82$.

[2] J.H. Friedman (2010) "Greedy Function Approximation: A Gradient Boosting Machine," $Annuals\ of\ Statistics$ 29(5) : 1189 − 1232.