**说 明 书**

**一种推荐系统中用于建模用户意愿度的模型设计**

**技术领域**

1. 本发明涉及的是推荐系统（Recommendation System）领域中，对潜在购房用户意向度进行建模和预测的设计方法。用户意向度（User’s Intention）是评估用户是否对某些商品感兴趣的重要指标，我们设计了一种新颖的建模方式和预测模型来对用户意向度进行更为准确的推断，进而帮助销售人员利用模型的预测打分更好地将精力集中在意向度较高的用户身上，从而促成更多交易，提高用户总体满意度。

**背景技术**

1. 近年来，随着数字经济的提出和大力推行，大数据（Big data）及人工智能（Artificial Intelligence）也相应地快速发展了起来，这些技术不仅引起了计算机领域的大变革，也为房地产行业的数字化转型创新带来了视角与工具。借助互联网等新型信息技术，我们能够采集用户的海量行为数据，如用户在小程序的浏览、点击与分享记录。通过数据挖掘（Data Mining）与机器学习等技术可以实现更好的用户体验与更精准的营销服务，而随着海量行为数据的积累与沉淀，信息化价值也逐渐从”业务支撑“向”推动变革“转变；信息化驱动从”业务驱动“向”数据驱动“转变。
2. 以往在房产行业中，用户信息的搜集与用户购买意向度判断完全依赖于销售人员，通过电话、线下拉人等各种方式获取用户的信息，并且由销售人员根据信息判断用户对购买房产的意向度如何，这种方式非常消耗销售人员的时间精力，容易在许多不愿意购买房产的用户身上花费过多时间。我们规划使用的数据链路会同时使用线上小程序、线下销售的数据作为用户信息，我们设计的机器学习模型能以远高于人工的效率判断用户的购买意向度。
3. 考虑到房产行业的特点和现实工业界问题，用于建模用户购买意向度的模型有三个困难：第一，从数据获取的角度上来说，相比于理论中的模型训练数据，产生于真实行业的数据会有较多的空缺值，导致模型训练时会损失很多信息，并且由于不是专业的数据标注人员做的标注，在数据标签上也有不统一不规范的问题，另外，在数字化转型的过程中，以往全人工的数据和现在半人工半自动读取的信息需要做融合统一，这两者相应的数据格式和数据分布会有一定差别。第二， 购房行为在用户的一生中的次数也极其稀少，相比于购买小商品的成交量和成交速度，购房行为的正样本数目会非常稀疏，不利于模型的训练和收敛。第三，购房行为会比购买小商品更理性，更不易受到冲动等其他因素干扰，这要求用于模型训练的数据更加精确，也对模型的预测准确度提出更高的要求。
4. 除了这三个难点之外，还有一个特点与其他的用户意向度预测问题有区别。展现用户意向度一般有多个指标，以现有的广告领域中的用户意向度预测的方法为例，用户会先点击，再激活，再转化，分别代表用户点击广告、下载app或激活商品、app内付费或下单商品，这三个指标递进式的代表用户意向度的大小，越在转化链路的后侧说明用户意向度越强。现有的广告领域中的用户意向度预测方法主要的目标是最后的转化，因为只有当发生了转化广告主才会得到实际的回报。因此前两个指标点击、激活是用来辅助最后转化的预测的。在房地产行业内，指标也有三种，到访售楼部、预签约、实际认购，从前往后用户变得稀疏，代表的意向度也变大，指标是类似的，主要的目标也类似，都是为了促成更多最终的实际认购，不过区别在于房产行业有大量的销售人员活跃在接触客户的一线，销售人员针对客户购房意愿度的建模和预判在现阶段稳稳超过大数据模型的预测，因此大数据模型不用像广告一样以最终的转化为主要目标并且想办法优化，而是可以以第一步到访售楼部作为主要目标去优化到访的预测准确度。一言以蔽之，房产行业的模型目标是为销售人员初筛客户，而不是为了预测最终的转化成功率。

**发明内容**

1. 本发明针对现有技术存在的上述不足，我们设计了一种新颖的建模方式和预测模型来对用户意向度进行更为准确的推断，进而帮助销售人员利用模型的预测打分更好地将精力集中在意向度较高的用户身上，从而促成更多交易，提高用户总体满意度。除模型设计外，我们还对原有的数据链路进行了重新规划和设计，使得数据从产生、筛选、计算、输出到反向动态调整模型及生成结果报告更连贯，适配于我们设计的模型。
2. 本发明是通过以下技术方案实现的：
3. 本发明涉及一种新颖的建模方式和新的数据链路规划，包括以下步骤：
4. 步骤1、我们首先对用户购房意愿度进行建模，将其建模为多级转化链路的序列化用户行为模型，即用户的不同行为可以反映用户不同的意愿度，而且行为之间大概率有前置条件的关系。比如在售楼场景中，用户到访售楼部能说明他有一定的购房意愿度，用户与开发商签约代表有更强的意愿度，用户最终付款认购代表成交，有最强的购房意愿度。到访、签约、认购这三者之间，一般是序列化的递进顺序，用户先到访观察情况，如果满意可能会继续签约，进一步如果做好决策准备购买就会认购，有可能用户只是到访，发现不满意就不再推进后面的流程，也有可能用户到访并且签约，但是最后因为一些原因没有付款认购，这就是多级转化链路的序列化用户行为模型。
5. 设用户初始状态为init，简写i，到访为visited，简写v，签约为contract，简写c，认购为buy，简写b，那么这个多级转化链路的最终付款概率可以表示为p(b) = p(i2v)\*p(v2c)\*p(c2b)+p(i2v)\*p(v2b)+p(i2b)。这个公式由三部分组成，第一部分概率是从初始状态到到访概率\*从到访到签约的概率\*从签约到认购的概率，代表这部分用户是按顺序走了完整的转化链路，第二部分概率是从初始状态到签约概率\*从签约到认购概率，这部分用户没有到访售楼部，直接签约并且之后认购，第三部分概率是从初始状态直接到认购的概率，这部分用户跳过了到访和签约的步骤，直接付款购买。在目前情况下，第一部分用户占绝大多数，第二部分和第三部分的用户很少。
6. 步骤2、搭建前人提出的多任务学习模型（multi-gate mixture-of-experts）。因为用户的三种转化行为：到访、签约和认购都能反应用户的购房意愿度，因此如果用三个独立的网络来建模三种行为会有一定的计算资源浪费，并且三者之间的关联不能很好的刻画，所以我们采用了多任务学习模型去同时训练、预测这三个转化行为，在该模型里，用三个独有网络分别代表用户三个转化行为，在独有网络的上游，我们使用了三个专家网络，每一个独有网络都对应上游的三个专家网络，通过门（gate）来融合三个专家网络的输出作为独有网络的输入。这样的模型既可以减少计算资源的浪费，使三个独有网络都有共同的上游专家网络层，同时共同层也让这三个任务相互影响，信息可以通过参数反向传播来实现相互流动。同时三个独立的网络也保持了三种任务的独立性，不会完全受到别的任务影响。另外一个用多任务学习模型的原因是，在售楼场景中，最终的认购是整个系统的目标，但是由于认购的正样本数非常的稀疏，单单拟合认购会导致模型误差较大，因此也引入到访和签约作为辅助指标，降低模型偏差（bias）。
7. 步骤3、在多任务学习模型MMOE的基础上，我们额外设计了聚合链路以降低正样本稀疏带来的训练困难，如图4所示，在三个专用网络分别输出ABC，将A和B的结果相乘得到的乘积作为任务B的输出，将A和B和C的乘积作为任务C的输出。之所以这么做，是因为A任务的正样本数相比于B任务会更多，而B任务的正样本数相比于C任务会更多，为了减少模型用极稀疏正样本数据做训练的难度，我们将B任务的预测值result(B)=output(A)\* output(B|A)，C任务也类似，result(C)=result(B)\* output(C|B)，为了保证这样条件概率的成立，我们将B任务的正样本调整为B或C，这样使得result(B)\* output(C|B)一定能成立，简化问题。4
8. 步骤4、我们设计了不同权重大小的损失函数。虽然系统的最终目标是认购，认购的预测准确率和召回率是最重要的指标，但是由于存在大量的线下销售人员可以跟进用户和揣摩用户购房意愿度，因此对模型来说，最重要的目标是给出一个打分，初筛意愿度较低的用户，让有限的销售人员集中精力在高分用户上，并且由于用户到访后就会有销售人员密切跟进，而到访后转化链路是可以由销售人员来进行预测和跟进的，相比于模型，销售人员能获取到更多关于用户的信息，比如表情、语气、行为，因此在后转化链路上模型不需要过于精准，输出的打分是预估到访的打分就可以，这反映在模型结构上就是将损失函数的权重设为不等值，第一个指标到访的权重会较大。
9. 步骤5、模型的输入数据在用来训练前需要经过数据清洗和数据预处理，比如空值、不确定值较多和方差极小的特征列会删除，而类别类特征比如城市会用TargetEncoder的方法转化为一维连续值。在特征列之外，数据行也需要做数据预处理，多行数据可能有相同的客户id，这代表同一个客户可能会出现在数据中多次，一部分是冗余的数据需要剔除，还有一部分是不同渠道获取的同一个客户的数据，另外还有回流客户的数据。针对后两者我们会把客户数据做正样本扩展，只要该客户有过到访或签约或认购的行为，那会把他在该时间之后的其他数据行都设为正样本，即有相同的到访或签约或认购的行为。接着会把数据归一化标准化到[-1,1]上。
10. 步骤6、我们的模型在训练时会同时用到到访、签约、认购这三维度的标签，以到访或签约或认购作为第一个专有网络的输出，以签约或认购作为第二个专有网络的输出，以认购作为第三个专有网络的输出。但是模型在预测的时候，只使用第一个专有网络的输出，将该输出通过一个分段器从0-1的概率映射到1-10的打分上去。

**技术效果**

1. 本发明与现有技术相比，其优点表现为：第一，利用后转化链路信息帮助前转化链路的指标进行训练学习，能有效提高前转化链路指标的预测稳定性和召回率，第二，充分利用购房场景与其他场景如广告、推荐的区别，即大量的销售人员，优化的目标在前转化链路，起到初筛把关的作用，第二，以三个输出指标作为训练标签，但最后输出一个指标作为总的意向度，方便训练和预测这两种对模型的使用方式。

**附图说明**

1. 图1 用户行为的序列化转化链路
2. 图2 本研究问题与其他问题的区别
3. 图3 MMOE的模型结构
4. 图4我们设计的模型结构
5. 图5 本模型与旧模型在到访指标上的比较
6. 图6 本模型与旧模型在签约指标上的比较
7. 图7 本模型与旧模型在认购指标上的比较

**具体实施方式**

1. 下面对本发明的实施例作详细说明，本实施例在以本发明技术方案为前提下进行实施，给出了详细的实施方式和具体的操作过程，但本发明的保护范围不限于下述的实施例。

**实施例1**

1. 本实施例包括以下步骤：
2. 第一步、对用户购房意愿度及其行为进行建模。
3. 对一个用户来说，他可能会收到来自开发商销售人员的电话，在电话中透露自己的一部分信息，或者有可能在小程序、app甚至直接到访售楼部来查看他感兴趣的房产，在这些线上或线下途径中都会留下自己的信息，开发商收集信息后可以用来预估该用户的购房意向度，从而对该用户增大或减小推广力度。根据数据埋点的位置，本发明中对用户的关键行为定位有到访售楼部、签约、付款认购这样三种，可以递进地表示用户购买房产意向度的大小。这条关键行为路径被称为多级转化链路。设用户初始状态为init，简写i，到访为visited，简写v，签约为contract，简写c，认购为buy，简写b，那么这个多级转化链路的最终付款概率可以表示为p(b) = p(i2v)\*p(v2c)\*p(c2b)+p(i2v)\*p(v2b)+p(i2b)。我们系统的最终目标是要尽可能提高对最终认购概率的预测准确度和召回率，从而尽可能将有意向买房的用户在早期筛选出来，提高召回率是为了让大多数意向客户能被模型筛选出来，可以接受容忍损失一定的精确率。
4. 第二步、深入理解业务场景
5. 虽然系统的最终目标是要尽可能提高对最终认购概率的预测准确度和召回率，但由于房地产行业的特点，有大量的销售人员活跃在与用户沟通交流的一线，其中有经验的销售人员是可以通过与用户沟通过程中，了解到用户对购房的意向度，对房产的喜好，从而做出针对性的推荐和跟进。这些全面多角度的预测意向度、挖掘用户偏好、个性化推荐的能力是基于大数据训练得到的模型所欠缺的。在用户到访售楼部之后就会有销售人员一对一跟进开展工作。因此在深入挖掘业务场景之后，模型侧的目标不再是提高最终认购概率的预测准确度和召回率，而是提高到访售楼部这一指标的预测准确度和召回率。
6. 第三步、设计模型结构
7. 目前的三个输出指标都可以作为反映用户意向度的指标，但也不是完全在同一个维度上反映意向度，有的用户可能不到访，直接签约，有的用户可能到访、认购，但不签约。因此本发明选择使用多任务学习模型(MMOE)来同时学习三个输出指标，这样能够在保留三个指标一定的独立性的同时，兼顾三个指标的同方向性。另外在MMOE基础上，本发明新增了三个专有网络输出的拼接，并且将到访的指标改为三个指标的或，也就是说不将第一个网络定义为到访，而是定义为到访或签约或认购这样的代表用户意向度的指标。模型结构如图4所示。假设输入数据是60维，在共享网络层使用3个专家网络，每个专家网络的结构是相同的，都是60\*75\*30单隐层的多层感知机(MLP)模型，将三个专家网络的30维输出经过门网络（gate）的线性组合后得到30维中间结果，然后输入三个独有网络中，独有网络的结构设计为30\*45\*2，最后的输出是2维结果，分别代表一个指标的正样本概率和负样本概率，两者之和为1，网络使用的损失函数是交叉熵函数，并且将三个子任务的损失函数的权重设的不等值，设为0.8，0.1，0.1.

**模拟实验结果**

1. 本实施实例的模拟实验比较了本模型和老模型在同一批数据上对到访、签约、认购这三个指标上的表现，并据此得到1-10分每个分段对某标签的正样本人数和比例。
2. 如图5所示，为到访标签上的各个分段上本模型和老模型的效果对比图。图中主要比较了各个分段上到访人数和到访的比例。可以看出，在到访人数上，本模型绝大多数到访数都集中在比较高的分数段，相比于老模型主要集中在中间分段效果会更好，即模型给出的高打分可以较好的体现用户的高购买意向度。在到访比例上与老模型差不多，在10分处会有较大的优势。
3. 如图6、7所示，为签约、认购标签上的各个分段上本模型和老模型的效果对比图。可以看出，在签约、认购的人数上，本模型可以给绝大多数达成购买的用户赋予比较高的打分，能较好的召回高质量用户，在高分段的召回比例上，本模型也比老模型有更高的转化率。
4. 通过以上模拟实验的结果，我们设计的模型在高分段的召回数和召回比例上与现有的老模型相比确实拥有巨大的优势。

**说 明 书 附 图**

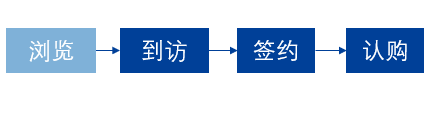


图1

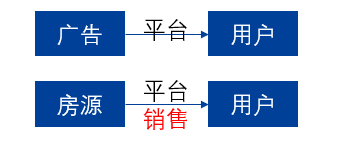


图2

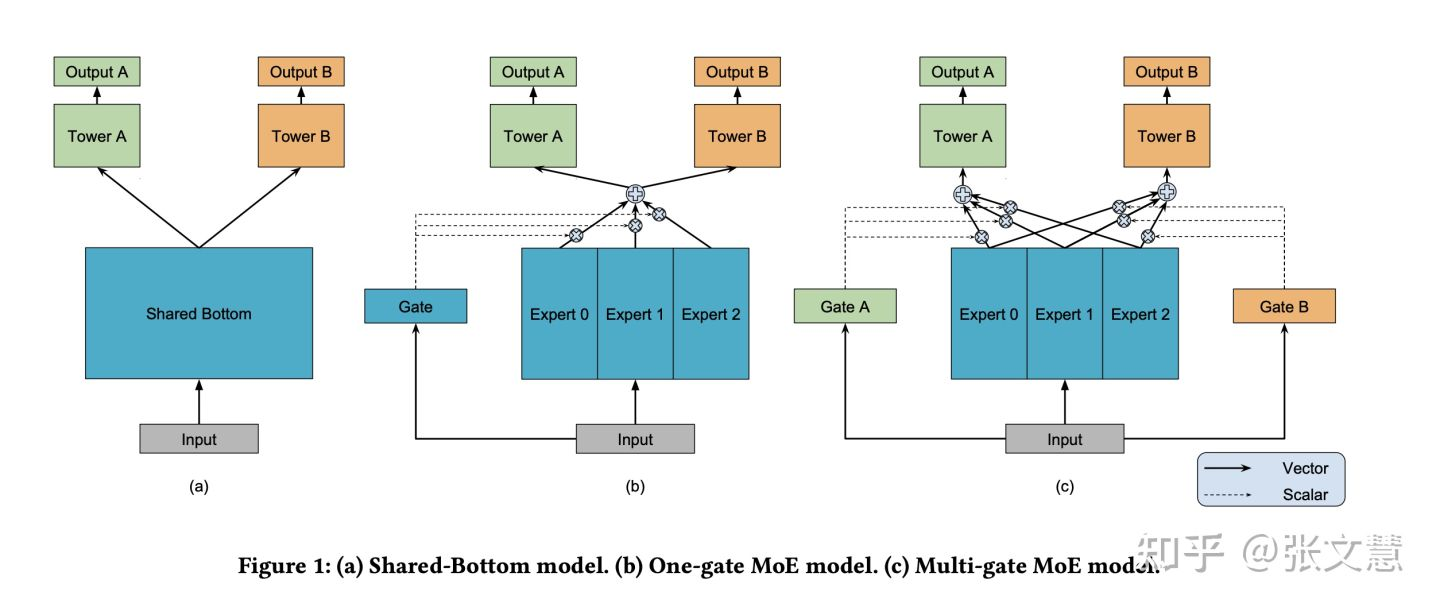


图3

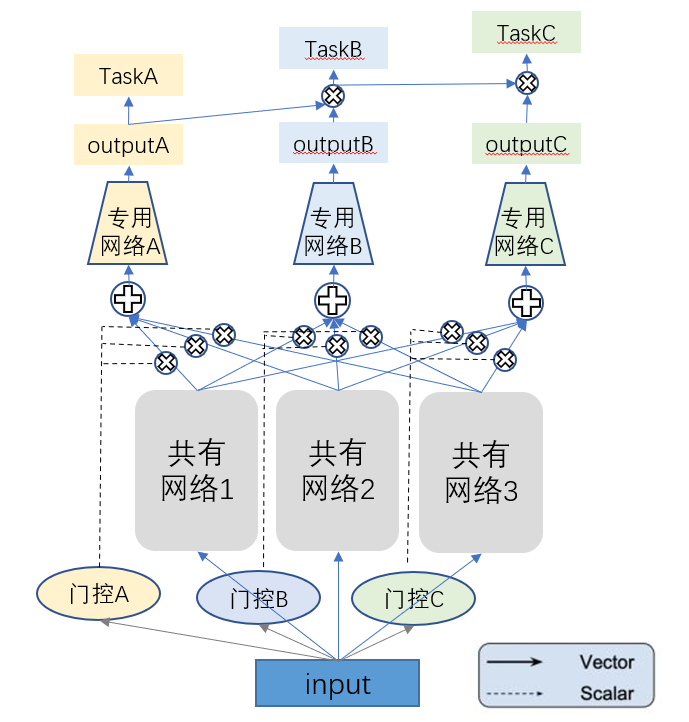


图4

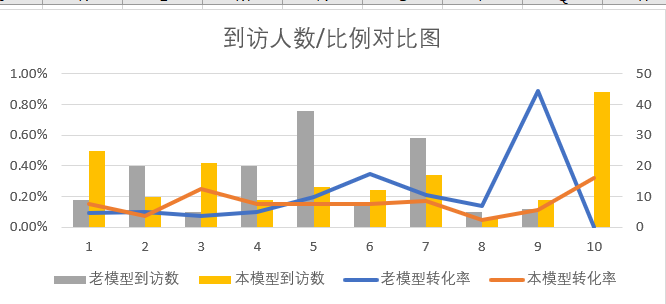


图5

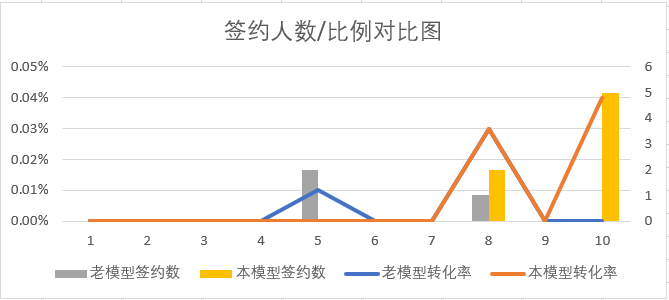


图6

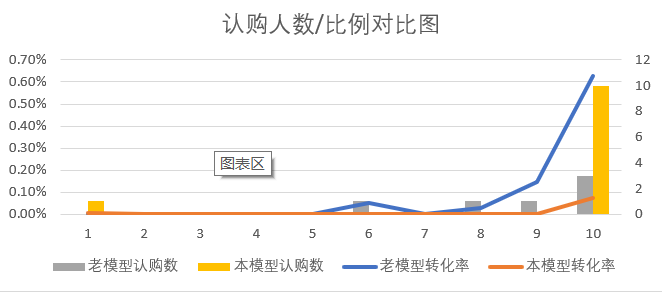


图7

**权 利 要 求 书**

一种推荐系统中用于建模用户意愿度的模型设计，包括以下步骤：

第一步、我们首先对用户购房意愿度进行建模，将其建模为多级转化链路的序列化用户行为模型，即用户的不同行为可以反映用户不同的意愿度，而且行为之间大概率有前置条件的关系。

第二步、在前人提出的多任务学习模型（multi-gate mixture-of-experts）的基础上。我们额外设计了聚合链路以降低正样本稀疏带来的训练困难，如图4所示，在三个专用网络分别输出ABC，将A和B的结果相乘得到的乘积作为任务B的输出，将A和B和C的乘积作为任务C的输出。之所以这么做，是因为A任务的正样本数相比于B任务会更多，而B任务的正样本数相比于C任务会更多，为了减少模型用极稀疏正样本数据做训练的难度，我们将B任务的预测值result(B)=output(A)\* output(B|A)，C任务也类似，result(C)=result(B)\* output(C|B)，为了保证这样条件概率的成立，我们将B任务的正样本调整为B或C，这样使得result(B)\* output(C|B)一定能成立，简化问题。

第三步、我们设计了不同权重大小的损失函数。虽然系统的最终目标是认购，认购的预测准确率和召回率是最重要的指标，但是由于存在大量的线下销售人员可以跟进用户和揣摩用户购房意愿度，因此对模型来说，最重要的目标是给出一个打分，初筛意愿度较低的用户，让有限的销售人员集中精力在高分用户上，并且由于用户到访后就会有销售人员密切跟进，而到访后转化链路是可以由销售人员来进行预测和跟进的，相比于模型，销售人员能获取到更多关于用户的信息，比如表情、语气、行为，因此在后转化链路上模型不需要过于精准，输出的打分是预估到访的打分就可以，这反映在模型结构上就是将损失函数的权重设为不等值，第一个指标到访的权重会较大。

第四步、我们的模型在训练时会同时用到到访、签约、认购这三维度的标签，以到访或签约或认购作为第一个专有网络的输出，以签约或认购作为第二个专有网络的输出，以认购作为第三个专有网络的输出。但是模型在预测的时候，只使用第一个专有网络的输出，将该输出通过一个分段器从0-1的概率映射到1-10的打分上去。

**说 明 书 摘 要**

一种推荐系统中用于建模用户意愿度的模型设计，将实际业务场景进行建模为多级转化链路的序列化用户行为模型，针对该业务场景的特点在MMOE多任务学习模型的基础上，给专有网络的终端设计了乘法拼接的输出值，使本模型可以一定程度上缓解样本稀疏性的问题，另外对三种输出标签值进行了逻辑组合，使之更适配于网络拼接的设计。本模型还分离了训练和预测输出，使模型训练时用更多的输出特征，在预测时专心输出一种特征，建立了针对该业务场景召回率更优的模型，相比于旧的模型，在各个打分段都有较高的转化率和召回率。