美团点评搜索广告点击率预估书面报告

小组：G2
成员：王道烩、胡潇、苑苑、钟凯、郑瑜

项目背景

美团点评是国内领先的智能营销平台，推广通作为其广告平台，拥有三大核心功能：更强大的生活全场景覆盖、更智能化的数字智慧和更友好的用户体验。推广通在美团平台上构建了商户与用户的智慧连接，能够助力广告主高效实现商业增长。

在移动互联网的盈利方式中，有一个经典的说法叫做“羊毛出在狗身上”，也即免费向用户提供产品获得用户的时间，通过卖广告获得利润。虽然在美团点评中外卖是利润的最大来源，搜索广告作为常见的互联网营销方式依然贡献着不俗的营收。在实际的美团场景中，当用户筛选相应的兴趣品类时，相应的商家（广告主）商品就会展示在用户看到的结果页面中。计算广告作为一个极其复杂的系统，其中有用户、广告主和平台三个利益相关方，而点击率（CTR）的预估是计算广告中重要的环节。

在本项目中，我们以美团的搜索广告为研究对象，利用美团提供的脱敏的真实广告点击数据，在给定用户、广告和上下文等信息的条件下预测广告的点击率，完成数据探索、特征工程、模型构建、效果评测以及模型分析等工作。

问题分析

定义

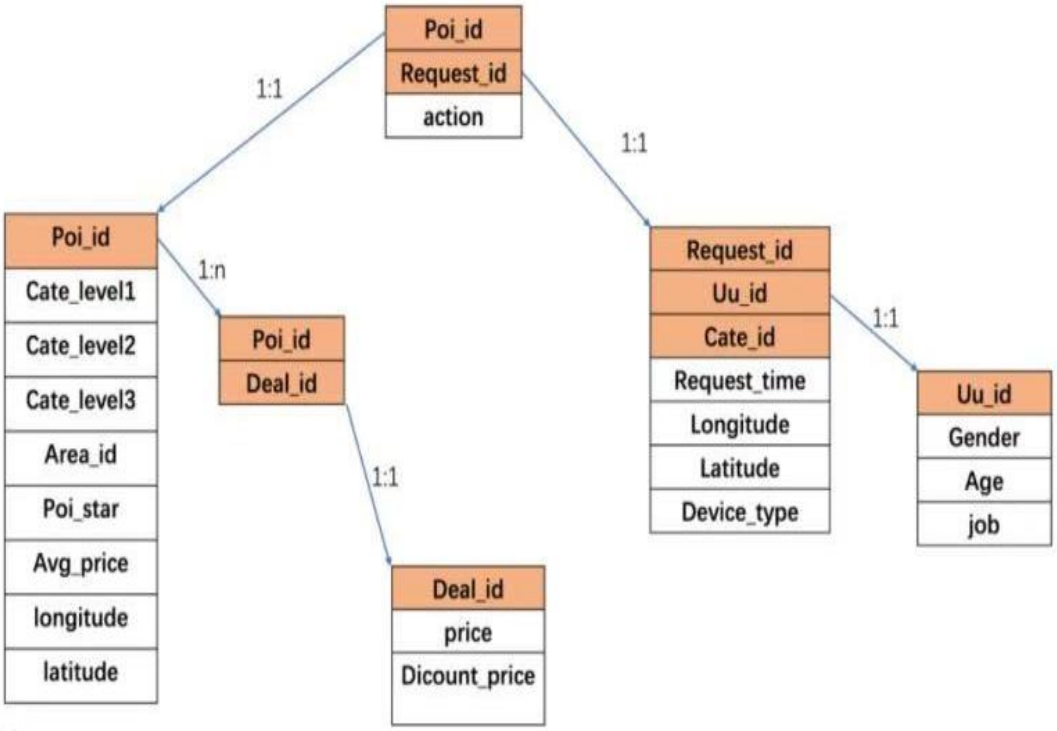
广告的点击率在给定用户、商户和上下文的条件下可以形式化的定义为：

$$pCTR = P(click=1|user,poi,context)$$

CTR 预估的目标是建立一个从特征到点击率的映射，也即学习一个函数关系 $y=f(x)$，其中 y 为点击率，x 为所有特征。

数据特点

经过对原始数据的分析，我们大概搞清了数据表的含义和关系，以关联图的形式表示如下表所示，通过不同表中的 id 对应关系，我们可以将其组合起来。



此外，通过观察分析，数据中的特征大致分为三类：用户相关特征、商家相关特征、每次交互的上下文特征。如下表所示。

分类	特征
用户特征	性别、年龄、职业……
商家特征	类别、商圈、价格、星级、经纬度……
上下文特征	时间（星期，小时）、设备、经纬度……

对于上述数据，我们将多个数据表联合起来，组成训练数据集，每一条数据由特征集合和交互结果组成，特征集合即每一次交互的上述各个特征，交互结果即用户是否点击该广告。数据集形成后，我们现做一个简单的统计，结果如下表。

总数据条数	正样本条数	负样本条数	用户数	商家广告数
4198717	92188	4106529	3961675	40472

简单分析上表数据，我们可以得到该从特征到点击率预测函数拟合问题的几个特点。首先是样本量，虽然每条样本数据不多，但样本量不算小，400 万的数量级。其次是样本不平衡性，正负样本差距明显，可以发现正样本在总样本中的比例仅为约 2.2%，这样不平衡的样本分布对于映射关系的学习非常不利。此外数据稀疏性较大，一方面是从总体可以算出平均每个用户只有 1.06 条交互数据，仔细统计也可以发现大部分用户仅有一条交互信息，因此肯定要从用户的各个子特征中学习出比较本质的关系；另一方面如商品品类等特征的总类别较多，one hot 编码也是非常稀疏的，这给模型原理的设计提出挑战。

挑战

从上述特点中，我们分析总结出该问题的几个挑战和对我们模型的相应要

求。

首先是样本条数较多，特征关系复杂，需要被较好地建模，不能仅仅拟合表面信息。这个特点要求我们采用精细的特征工程手段，调整和丰富特征信息，此外我们的模型方法要能够建模特征之间复杂的交互关系。

其次，样本条数多对我们的模型复杂度提出了要求，不能采用过于复杂的模型，要有较高的效率，提高迭代速度。

再次，因为数据的稀疏性、不平衡等特点，我们的模型还需要具有较好的泛化性，不能导致模型过拟合于训练集数据，没有建模出问题的本质特征。

方案选型

经过调研，我们发现目前解决 CTR 预估问题的模型主要有三种，作为我们的候选方案，分别是：因子模型、树模型、深度学习模型。

因子模型的特点是能够丰富特征信息，通过隐向量编码和分解机的计算机制比较容易地建模特征相互关系，开销不是很大。树模型的特点是通过在训练集上学习合适的参数，集成综合多个弱模型，计算量不大，而通过快速迭代多次调试树的结构等超参数，我们能够较好得调优模型表达能力，避免过拟合。深度模型是近年来工业界广泛使用的新方法，具有数据处理简单、表达能力强、建模特征关系更加复杂等优点。但是深度模型的缺点也很明显，一方面算法复杂度过高，对算力要求较高，另一方面在数据量不够的情况下难以拟合出泛化能力强的结果，我们的数据量、以及数据的不平衡性对于训练一个泛化能力较好的深度模型来说还是显得不够多。

综合模型复杂度和泛化能力，我们最终选择探究因子模型和树模型两种方案。

方案设计

根据之前的方案选型，我们实现了两类方案，分别是因子模型和树模型，两类方案都在在线比赛中进行了提交，现在分别介绍如下。

因子模型

基本原理

之前提到，数据具有稀疏性的特点。每个特征经过 one-hot 编码后，大部分样本数据特征是比较稀疏的。例如商品的品类，one-hot 之后有上百个数值特征，但是有且仅有一个是非零的。在真实应用场景中，解决数据稀疏性是一个不可避免的问题。在特征较为稀疏的情况下，关联某些特征，与 label 的相

关性就会提高。例如，“device_type”、“time”与“gender”的关联特征，对用户的点击会产生影响。比如在较晚的时间，可能年轻人的点击会比较多。因此，引入特征的组合是非常有意义的。通过建立多项式模型，FM 旨在解决稀疏数据下的特征组合问题。

我们以二阶多项式模型为例，简单介绍 FM 模型原理：

$$\hat{y}(\mathbf{x}) := w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \mathbf{v}_i, \mathbf{v}_j \rangle x_i x_j$$

上式的第三项即为特征组合表达。比较朴素的想法是，每两个特征之间我们可以计算一个权重系数，但是这样系数的个数为个，在 one-hot 之后，计算量过大。所以 FM 对每个特征学习一个隐向量，权重系数就等于这两个特征隐向量的内积。

数据处理

在进行特征工程之前，首先需要对数据进行预处理。

因子模型只提供每条请求记录拥有的特征，所以对于缺失值我们没有必要进行填充。

划分训练集和验证集，我们没有进行简单的随机抽样，而是按照时间戳划分，避免在训练集上引入未来信息。

原始数据正负样本比例约为 1:50，是比较不平衡的样本，由于训练时，算法对每个数据都会进行学习，多数数据样本带有的信息量比少数样本信息量大，会对分类学习过程中造成困扰。为解决正负样本不均衡的问题，我们在训练集上对负样本进行降采样，去除一些负样本，使正负样本比例约为 1:10。

特征工程

当数据预处理完成后，我们需要选择有意义的特征输入模型中进行训练，因此进行特征工程，对特征进行选择和处理十分关键。下面是对各类特征的具体处理方法。

地理位置

我们根据用户和 POI 各自的经纬度来计算距离（haversine），作为 double 类型特征。

年龄特征

年龄特征虽然是一个较为连续的值，我们对其进行分桶处理，只考虑相近年龄的特征。这里我们自定义分桶区间 0-20，20-30，30-40，40-50，50-。分桶后，我们分别使用 1，2，3，4，5 来代表各个年龄段。

时间戳

时间类特征既可以看作是连续值，也可以看作是离散值。考虑其连续特性时，我们可以计算用户点击广告的间隔时间。但是在我们的数据集上，用户的历史行为记录十分稀疏，大部分用户只拥有一次请求记录，所以我们将时间特征看作离散值，分离出一周中的天数（星期一，星期二）和小时时段作为分类特征。同时一年中的哪几个星期可能也是比较重要的特征，但是由于我们的数据集全部都在五月份，故舍弃了此特征。

double 特征

数值型特征并不符合正态分布，却呈现出长尾分布，较少的高频特征的频率占整体的较高比例。这里我们对 75%以后的极端数据做截断，以避免长尾效应。

在不同的特征指标中，其量纲或是量纲单位往往是不同的，变化区间处于不同的数量集合，如果不对数据进行处理，可能导致某些指标被忽视，影响到模型的预测效果。为了消除特征数据之间的量纲影响，我们对特征进行 Min-Max 归一化，以解决特征指标之间的可比性问题。

YouTube 推荐系统[3]提出的方法可以看作是因子模型的一种非线性处理，相比传统的线性模型和树模型，取得了较好的预测效果。受此启发，我们尝试引入非线性特征，对于 double 类型的特征，将其归一化后的结果做平方、立方、平方根、立方根处理加入特征。

虽然 double 类型属于连续特征，但是离散化处理可以使模型的学习更为准确和迅速。这里我们采用的离散方法均为分桶。

分类特征

userid，poiid，cateid 等超长整型的特征，不能将其 id 直接作为分类类别。对于分类特征，我们进行 re-index 处理。

模型效果

特征	验证集 AUC
w/o distance	0.614199
w/ distance	0.635012
w/ distance & nonlinear	0.636346

不使用距离特征进行预测，验证集 AUC 为 0.614 左右；加入距离特征后，AUC 上升至 0.635，提升了近 3%。引入非线性的距离特征后，AUC 进一步提高至 0.636。从结果我们可以发现：

1. 距离特征在美团广告 CTR 预估中非常重要
2. 在线性 FM 模型中加入非线性特征能够进一步提升性能

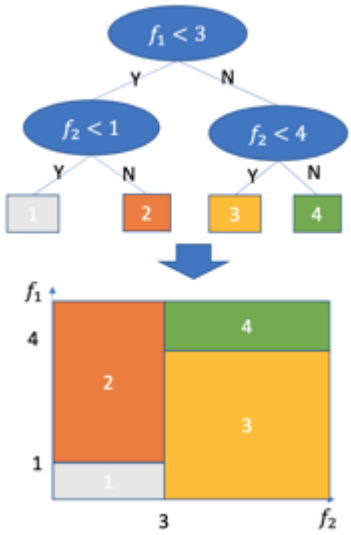
树模型

基本原理

梯度提升树的主要思想是利用决策树来进行回归，同时利用 Boosting 的方法来将多个决策树的结果进行集成得到更好的结果。

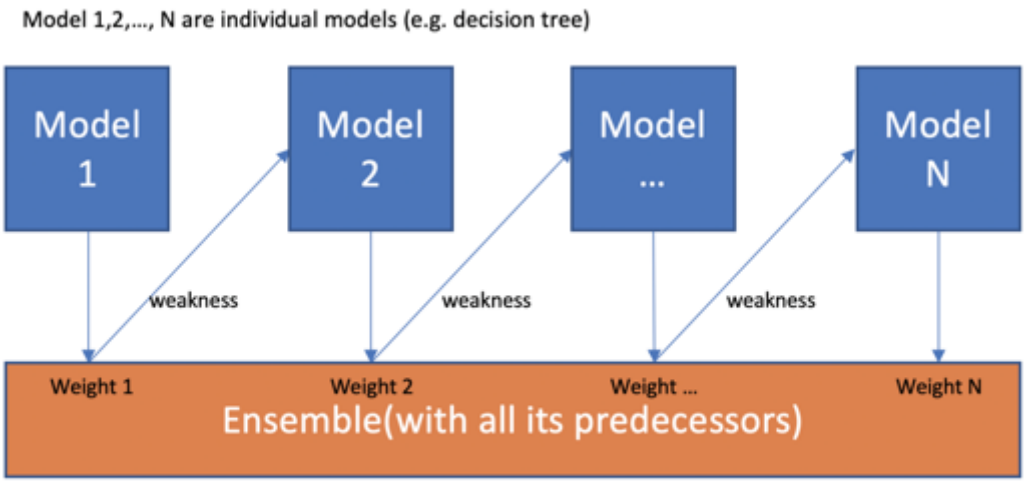
决策树的基本原理是，每一次迭代会对在所有的叶子节点中选择一个叶子节点进行分裂，由于每个样本具有很多不同的特征，每种特征又有不同的取值，所以需要遍历所有的叶子节点，然后遍历每个叶子节点的所有的特征以及这个特征的所有可能的阈值。选取阈值的方法是当前叶子节点利用当前特征以及当前阈值来进行分裂，所得到的收益是最大的。所以每一个叶子节点的生长过程中，计算的时间复杂度为 O(#num_leaf*#num_feature*#num_threshold)。决策树按照任务的类型可以分为决策分类树和决策回归树。在算法结束之后，输入一个测试样本，样本按照自己各个特征的值和模型的每个分支节点所使用的特征以及阈值，逐层向下移动，最终会落到某个叶子节点上，那么这个样本的输出的类型就是训练集数据落到该叶子节点的类型最多的数据所属的类型。对于决策回归树来说，输入测试样本，按照上述类似的形式落到某个叶子节点之后，样本的输出是训练数据落到该叶子节点的所有数据的 label 的平均值。

从本质上来说决策树相当于按照特征将样本空间进行不同的划分，然后利用每个区域中训练数据的结果来代表这个区域的 label 或者回归值。如下图所示。



决策树示例

Boosting 方法被广泛应用到机器学习领域，其基本的思想是，如果能够训练得到若干弱模型，如何将这些弱分类结合起来得到一个强模型，如图所示。该思想被广泛应用许多算法中，如 AdamBoost， 随机森林以及梯度提升树等。



One is weak, together is strong, learning from past is the best

在梯度提升树算法中，主要是利用决策树来对模型的梯度进行回归。如公式所示，在梯度提升树算法的最终结果是将多个树的结果相加。对于一个具体的训练样本，在得到了前 m-1 棵树的结果之后，希望第 m 棵树的输出和之前所有的模型加起来能够使得其输出结果和真实的 label 计算得到的损失函数能够降低。如公式所示。对损失函数进行一阶泰勒展开，可以发现当第 m 个模型的输出是沿着损失函数在前 m-1 个模型输出位置的负梯度方向，那么将这个模型累加到之前的模型之后，输出的结果的损失函数会降低。所以对于每个训练样本，第 m 棵树只需要回归损失函数在这个样本的前 m-1 棵树的结果位置的负梯度就可以了。有了这个思想，就可以不断地生成新的树模型来使得效果得到进一步提升。

- $F_m(X) = F_{m-1}(X) + f_m(X) = f_1(X) + \ ... + f_m(X)$
- $f_i(\cdot)$ is the weak learner
- We want to get $f_m(X)$ that satisfies
  - $L(F_{m-1}(X) + f_m(X), Y) < L(F_{m-1}(X), Y)$
- Calculate the negative gradients
  - $\hat{y}_i = -\partial_{F_{m-1}(x_i)} l(F_{m-1}(x_i), y_i)$

L2 loss, $\hat{y}_i = y_i - F_{m-1}(x_i)$

- Learn $f_m(X)$ to fit $\hat{Y}$ by using L2 loss
  - $f_m(X) = \arg\min_{f(X)} \sum_{i=1}^{n} (f(x_i) - \hat{y}_i)^2$
- Prove by first order Taylor Expansion
  - $l(y_i, F_{m-1}(x_i) + f_m(x_i)) = l(y_i, F_{m-1}(x_i)) + \partial_{F_{m-1}(x_i)} l(F_{m-1}(x_i), y_i) f_m(x_i)$
  - And $f_m(x_i) \approx \hat{y}_i = -\partial_{F_{m-1}(x_i)} l(F_{m-1}(x_i), y_i)$
  - Then $l(y_i, F_{m-1}(x_i) + f_m(x_i)) \approx l(y_i, F_{m-1}(x_i)) - \hat{y}_i^2 < l(y_i, F_{m-1}(x_i))$

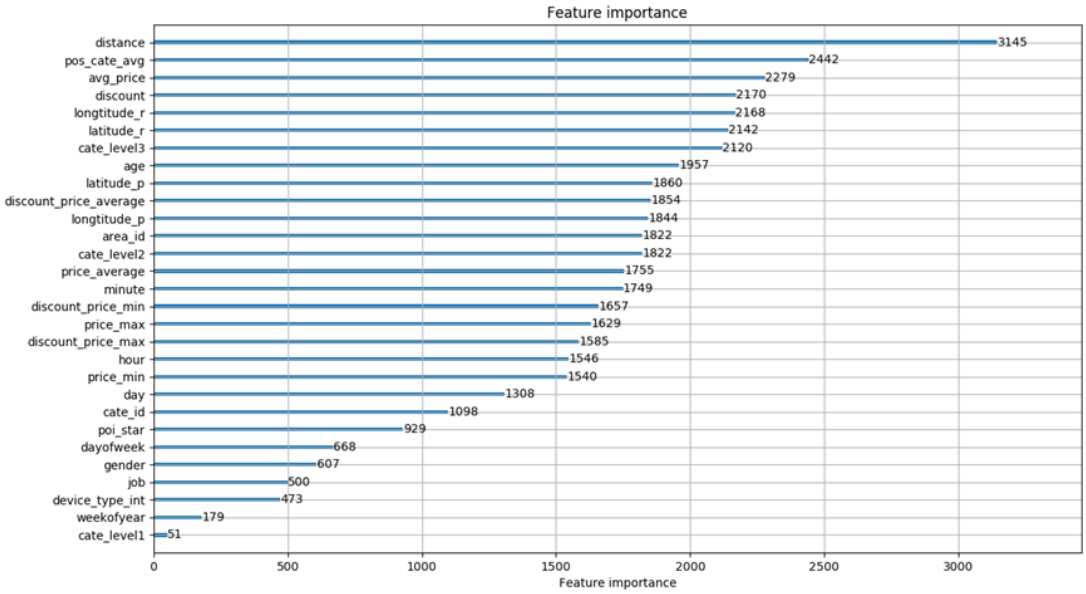
特征提取

在这部分我们除了利用了因子模型提取的距离和时间特征之外，引入了一些新的特征。对于距离特征，由于树模型的特点，我们没有进行离散化处理。同时对于其他 double 类型的特征也没有进行类似的处理。

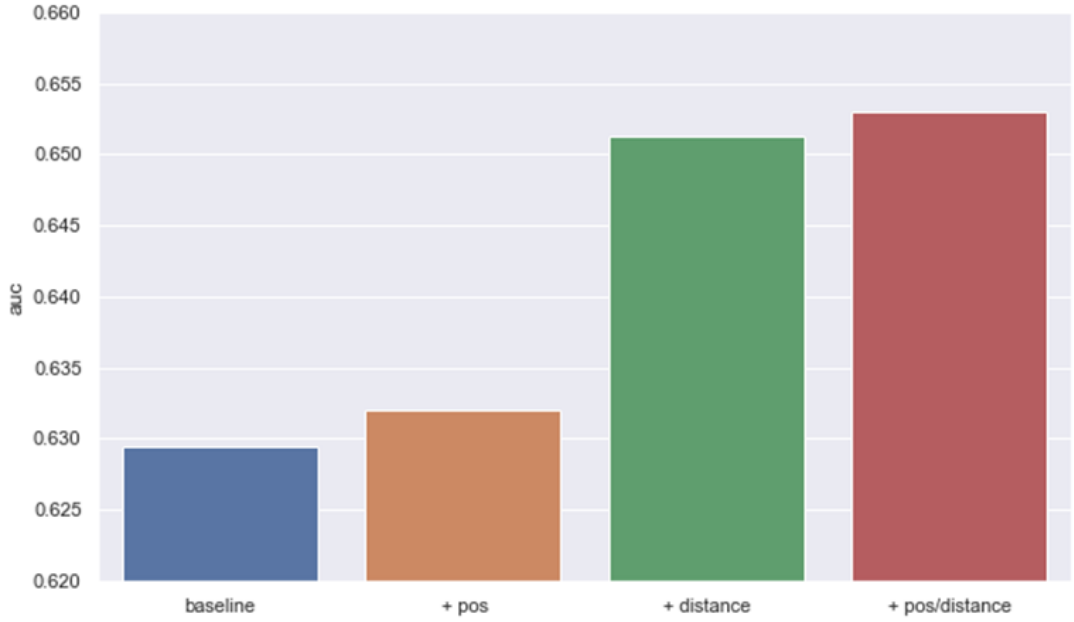
1. POS 特征。这个特征表示了在当前请求下所在的位置。由于这个信息在测试集中没有，所以我们将其处理作为商户特征。我们统计了每个商户在每种请求类型下的 POS 的平均值、最大值、最小值、中位数特征。然后再测试的时候作为商户特征加入到了数据中。这为我们的性能带来了一定的提升，具体分析见下面的部分。
2. 团单特征。对于每一个商户，其有一写团单，包括这些团单的原始价格，打折后的价格。我们对其进行了一些处理，并最终作为这个商户的特征。我们对每个商户所有的团单的打折率求平均值、最大值、最小值等统计特征。同时计算原始价格的统计特征和打折后价格的统计特征。
3. 点击率特征。我们希望统计每个商户在每种 cate 类型下的点击率，作为当前商户的特征。但是在加入这个特征之后，出现了非常严重的过拟合现象。在展示的过程汇中，指导老师也指出，可能特征的使用方法不对，做一些平滑或者去除出现较少的商户，应该对效果会有一部分提升。这也是未来值得探索的一个方向。
4. 用户历史行为建模尝试。我们希望对用户的历史行为信息进行建模，但是对数据进行统计之后，发现有超过 90%的用户在数据中只出现过一次，而且出现次数和点击率没有什么太大的关系，所以这个特征最后也没有使用。

特征重要性分析

画出各特征在树中出现的次数做为重要性参考，如下图：

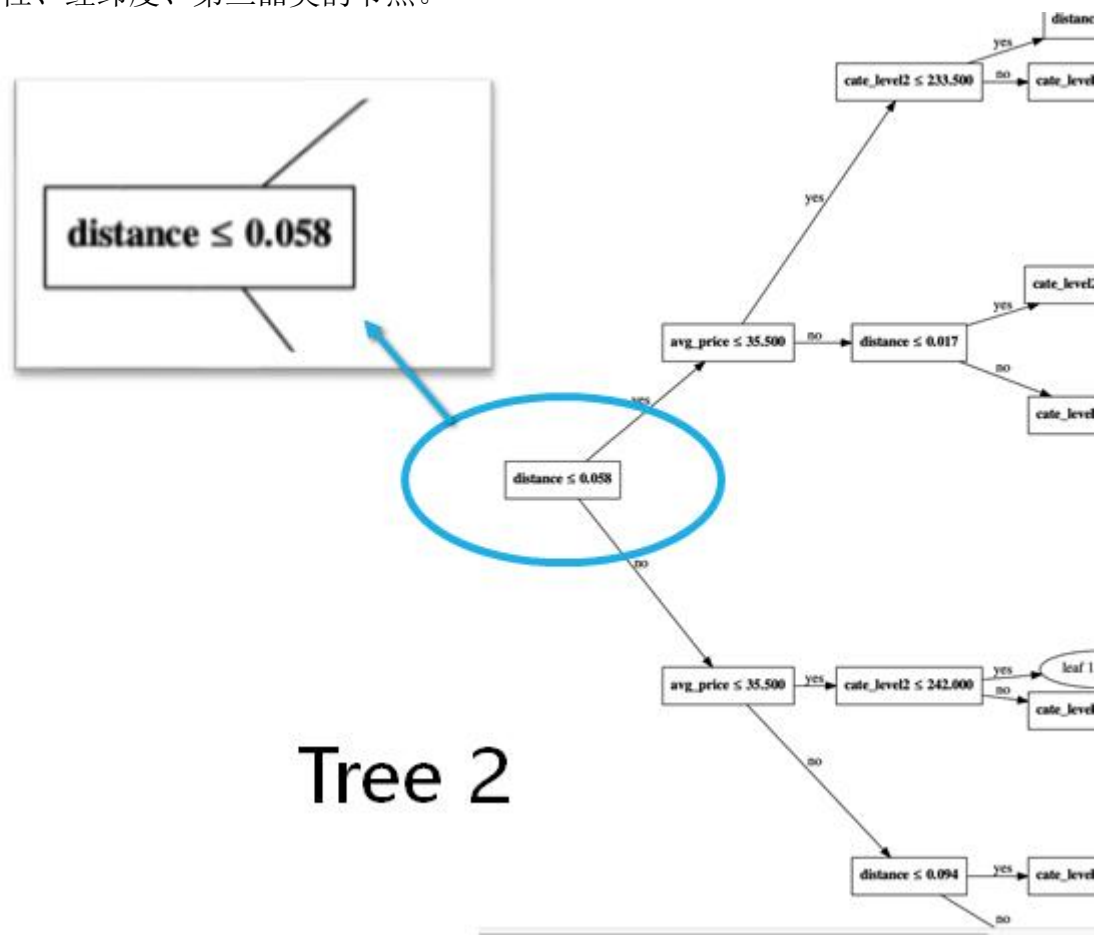


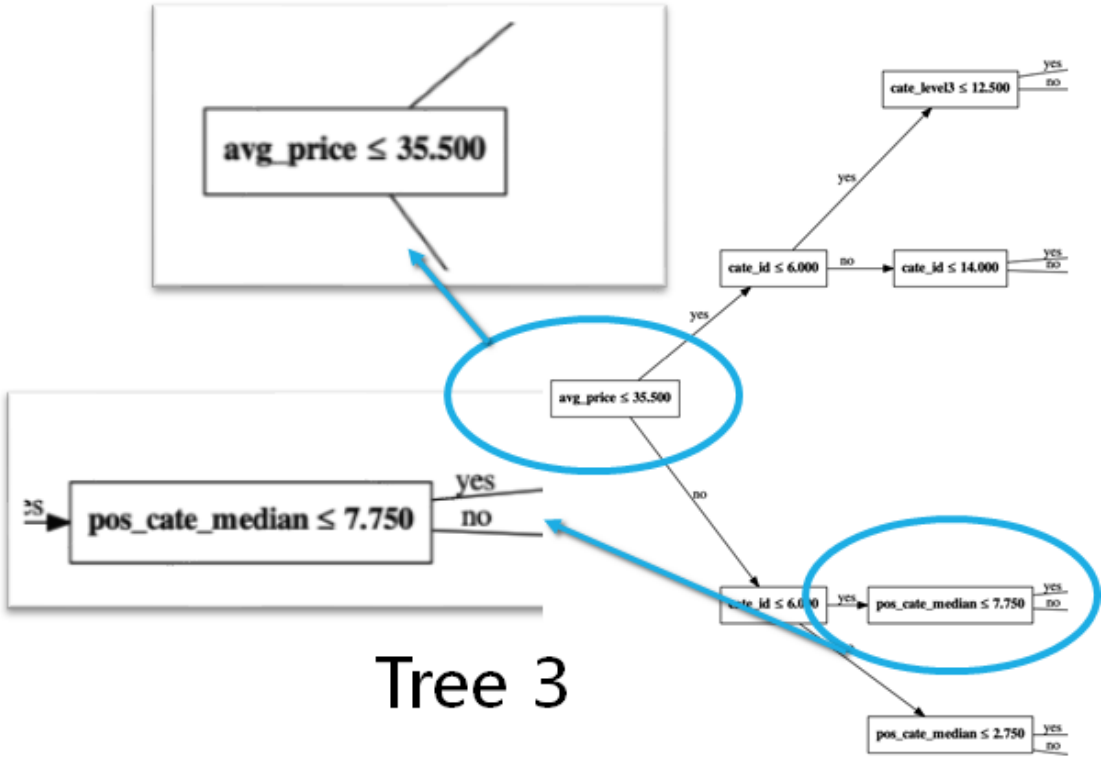
从图中可以看出距离特征在树中出现的次数遥遥领先，之后一种 pos 特征和平均价格、折扣的重要性、经纬度、第三品类的重要性也较高。
选取排行前二的距离特征和 pos 特征做消融实验，画出在测试集上的 AUC 柱状图如下：



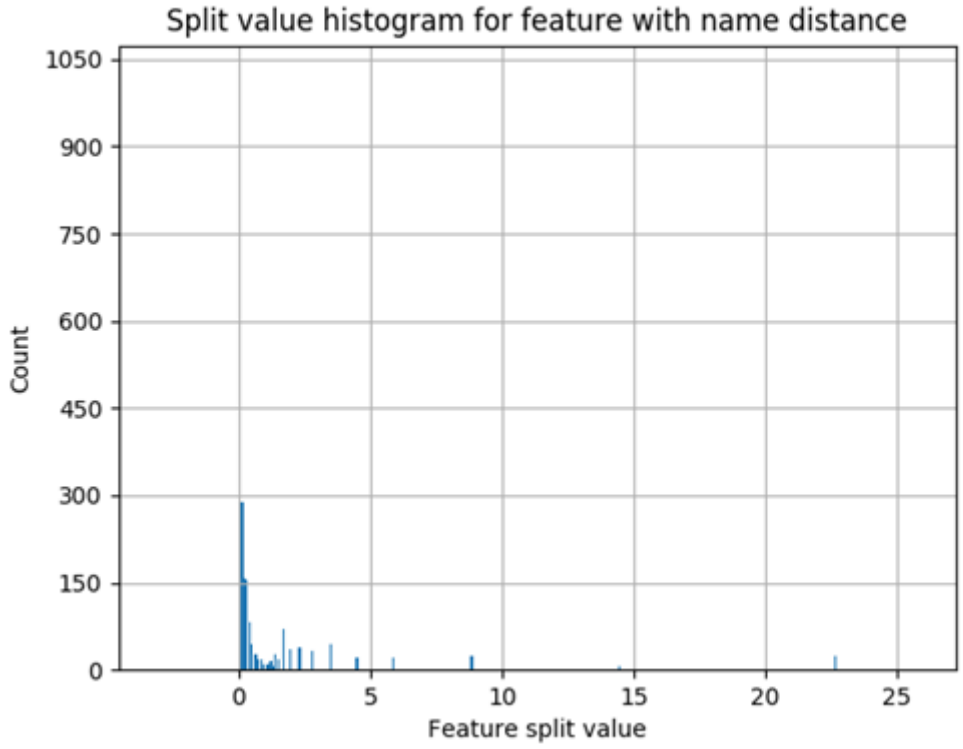
从图中可以看出距离特征的提升非常明显，pos 特征有少量提升，且两者都有的效果最好。

将最后的树模型可视化，可以看到在第二颗树、第三棵树中我们就找到很多 distance 的节点和之前提到重要性较高的 pos 特征和平均价格、折扣的重要性、经纬度、第三品类的节点。



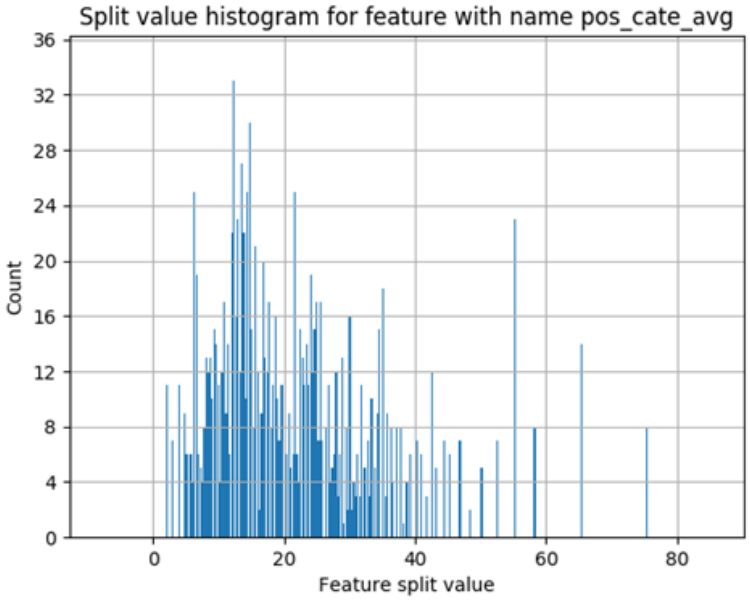


为了考察各特征在树中分离值的统计规律，可以做出分离值的柱状图，首先可观察距离特征：



从上图可见距离特征的分离值在距离小时非常地集中，这说明距离特征在

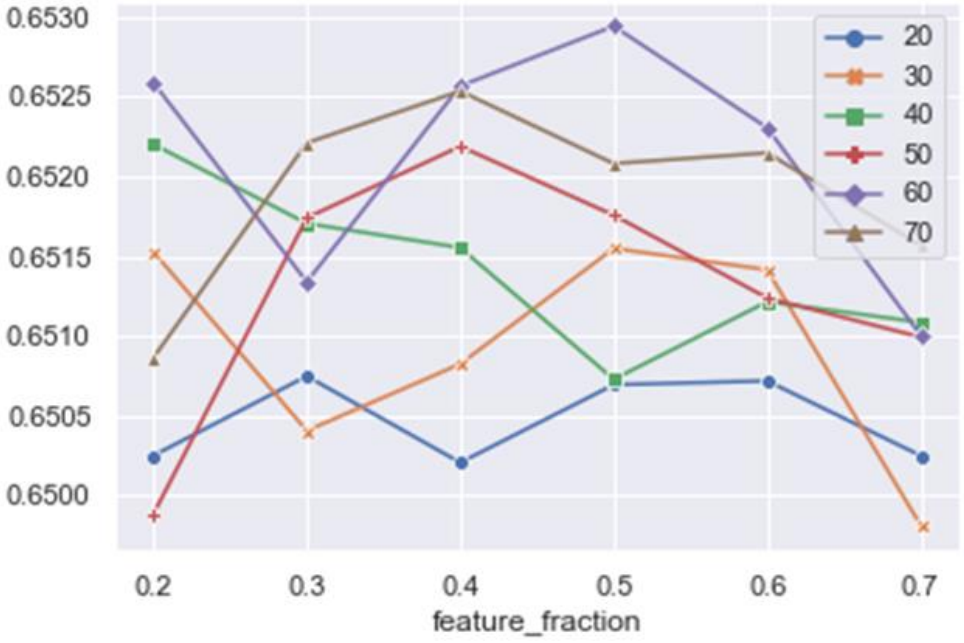
本次任务中非常地敏感，一点点距离的优势可能就会很大地影响最后的结果。



这张图是位置特征的分离值柱状图，可见非常地分散，起初我们非常疑惑，因为这也应该是 pos 越小影响越大才对，后来我们发现可能 pos 并不是越靠前越好，即使在训练集中 pos 为 3 的正样本比例也是远高于 pos 为 1 的，那么从这张图中就可以发现 pos 的值还是被分离地很好地，几乎所有值都有许多节点，说明学到的树模型能充分利用提取的 pos 特征。

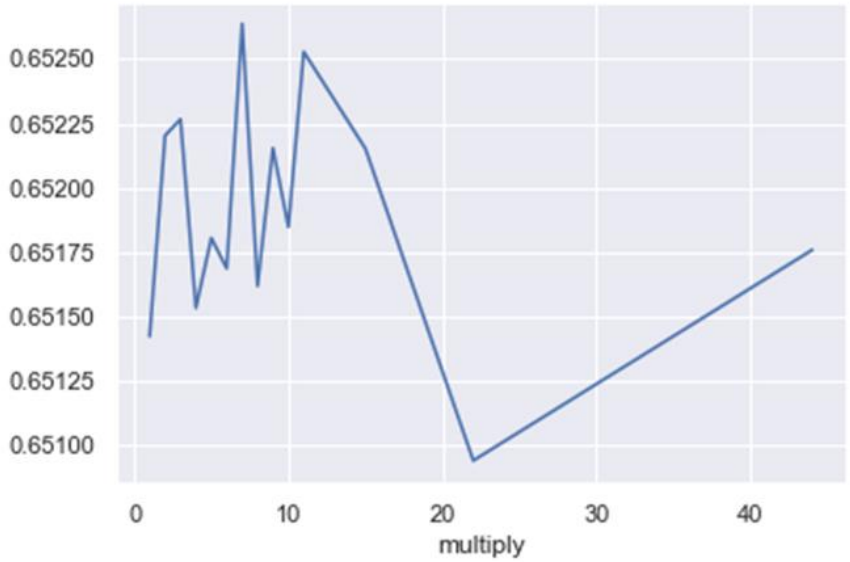
超参数分析

首先我们对叶子树和特征比例做了参数扫描：



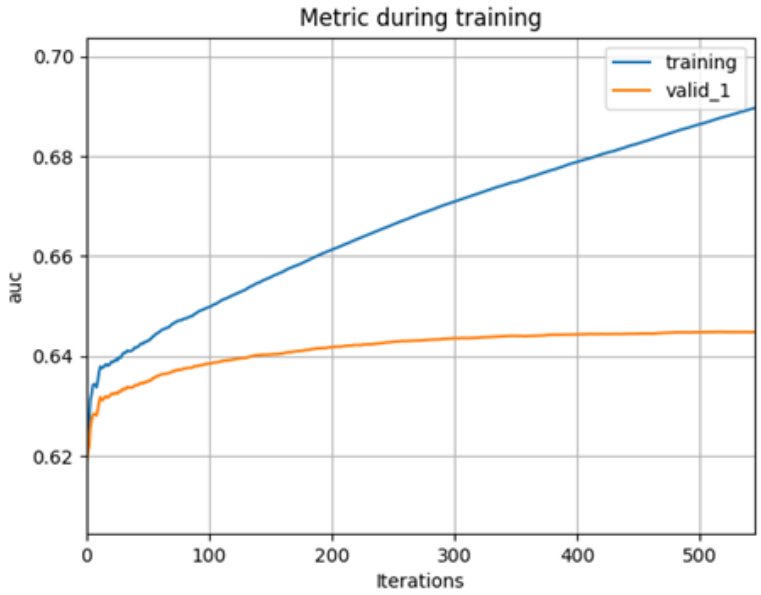
从图中可以看出随叶子节点数增多，效果是越来越好这是因为更多地节点数能表示更多地信息。但从 60 增长到 70 后就下降了，分析可能有 2 种原因：一是造成了过拟合，二是造成了冗余空间和噪音。性能随 feature_fraction 的趋势基本也是先增后减，分析原因可能是过小的比例使得模型很难选择有用的特征进行学习，而过大的比例效果不好也可能有两种原因，一是过大比列导致模型每次都倾向选择最有辨别力的特征从而忽略了某些特征，二是造成过拟合。

为了解决数据不平衡的问题又不想因为下采样导致信息损失，我们采用了简单重复正样本的上采样办法。下图是 AUC 在测试集上随正样本重复次数的变化趋势：



从图中我们可以看出没有明显的趋势刻画两者的变化，但是可以肯定的是，完全不重复的效果很差，进行过多重复正样本使得正负样本比例基本相同时的效果也不好。进行少量重复使得负样本为正样本几倍时的表现普遍较好，最后我们选择重复正样本 8 次，此时正负样本比例基本为 1:5.5。

下图是训练过程中 AUC 在测试集和训练集上的变化趋势：



从图可见到 515 轮的训练过程中，AUC 在训练集上不端提高，在测试集上缓步提高。为了防止过拟合，我们使用早停机制，当 30 轮内在测试集上 AUC 不再提高便停止训练。可见最后停止时 AUC 在测试集上变化趋势已接近水平。

模型复杂度及训练耗时

CPU	Intel(R) Xeon(R) CPU E5-2683 v3 @ 2.00GHz*30
模型大小	3.27M
轮数	~500
训练时间	196s

我们在搭载 30 个 Intel(R) Xeon(R) CPU E5-2683 v3 @ 2.00GHz 的服务器上进行实验，最后训练的轮数为 515 轮，训练时间仅 196s，模型的大小为 3.27M。可见 lightgbm 的训练还是非常快速轻量的。

结果分析

在本次大作业中，我们共尝试了两种方案，分别是因子模型和树模型，我们用多个账号提交了不同方案的结果，将两种方案的线上比赛结果总结如下表。

方法	AUC(public board)	AUC(private board)
xlearn	0.64295	0.63696
LightGBM	0.64770	0.64352

从线上比赛的结果中我们可以看出，树模型的效果显著的优于因子模型。而在美团课程的课堂展示中，我们也得到了验证，我们小组排在六个小组中的第三名，比我们成绩更好的两个组也是采用了树模型。树模型在多特征、高稀疏性的结构化数据中确实有较强的建模能力。

分工

王道烩：数据分析、数据集清洗、树模型特征工程、树模型开发与调试、报告撰写、PPT 树模型部分制作

胡潇：数据分析、树模型特征工程、树模型开发与调试、报告撰写、PPT 树模型部分制作
苑苑：数据分析、数据集清洗、因子模型特征工程、因子模型开发及调试，报告因子模型部分
钟凯：数据分析、因子模型训练及调参、报告撰写、PPT 展示
郑瑜：组织讨论及分工、因子模型特征工程、因子模型开发及调试、PPT 大纲及因子模型部分制作、报告撰写

总结

在本次大作业中，我们利用美团课程所学的内容，将机器学习的算法与技术应用在美团真实场景产生的数据中，尝试解决美团搜索广告里实际的点击率预估问题。我们小组的五名队员经过细致的讨论，分析了任务的难点和挑战，在多种可能的候选方案中采用了两种解决方案，分别是因子模型和树模型。经过分工，三位组员负责因子模型，两位组员负责树模型。我们从数据出发，分析数据集里的特征及其分布，根据数据的特点对数据集做清洗和过滤等预处理。之后我们学习探索了所选方案的原理及开源实现，了解了因子模型和树模型对于数据的要求，并根据采用的方案分别设计特征工程。在协作完成了算法模型的开发之后，我们做了多次的调参和迭代，力求找到最优的特征组合和超参数。我们在线上比赛中提交了多次，对不同模型和特征做了细致的对比。在完成比赛的过程中，我们还对模型的效果做了分析，尤其是特征的重要性以及超参数的敏感度分析。由于算力有限，我们对于模型的复杂度和迭代周期也非常关注，最终的两个解决方案的训练时间都在分钟级别，这也为我们对模型做快速迭代提供了可能。

在项目的开展过程中，我们非常注重信息共享，组员之间的协作也比较默契。虽然小组内部分成了两拨分别实现因子模型和树模型，但是对于开发过程中非常重要的环节尤其是特征工程，我们尽可能做到了信息的高效共享，在一个解决方案中非常有用的特征都会分享给另一个解决方案。在实际的项目代码开发中，我们也非常注意代码的复用性与可读性，在利用开源工具实现算法时也细致的查阅了文档说明，力求代码的简洁性以及 API 调用的规范性。在实际的协作中，我们利用腾讯文档等工具提高工作的效率，降低团队沟通的成本。

在这次大作业的完成过程中，我们也有一些遗憾。首先，我们小组的五位队员都没有打 Kaggle 比赛的经验，因此在项目进展之前本应对 Kaggle 比赛做充分的调研和了解，汲取前人的经验，也避免走弯路。事实上，在知乎和很多微信公众号中，都有 Kaggle 比赛大神做的一些分享，我们没有充分利用这些资源，在实际的特征工程和模型构建中也因此做的不够精细，这是我们最终只在六个小组中排名第三的一个原因。另外，多个模型或者同类模型不同实现方式的 bagging 和 stacking 是提升成绩的一大利器，在这一点上我们做的不好，最终提交的结果都是单模型预测的结果。看到排名第一的软院 Kaggle 大神的展示，发现他们的最优结果也只是 GBDT 不同开源实现的 stacking 时，我们也追悔莫及。

通过本次美团课程的大作业，每个组员都学到了很多内容，积累了不少经

验，相信这次大作业对于我们之后的研究甚至就业都有不小的帮助。同时，美团课程的各位讲师在本学期带来的课程讲解也让我们受益匪浅，我们了解到了实际工业界大数据的处理方式以及工作流程，对于整个大数据的技术栈有了初步的认识，从算法原理到工程实现再到实际应用都有了较为清晰的理解。最后，感谢美团课程的各位讲师，也感谢开设这门课程的老师以及助教。下面是每个组员各自的一点感想。

王道烩：

因为这是自己的研究方向是自然语言处理，第一次接触到推荐系统，所以整个流程走下来自己的收获还是非常的大的。首先是数据格式和自然语言处理具有非常大的差别。本次任务的数据结构都是表格形式的结构化数据，所以在数据处理和数据清洗这部分学习了一些工具的使用以及对于缺失值常见的填充方法。第二点感触是特征工程的重要性，有句话说的好，“模型决定性能的下限，特征决定性能的上限”。特征工程的目的是将先验知识融入数据，让模型更好地理解数据。第三点是对“no free lunch theorem”有了更深刻的认识。由于现在神经网络模型非常地火热，所有其它小组成员也尝试使用深度学习来解决这个问题。但是普遍效果不是很好。所以不存在一种模型能够直接用于所有数据，解决所有的问题，具体模型的选择是强烈依赖于数据的分布的。虽然现在的深度学习方法在很多领域都取得了非常大的突破，尤其是其自动提取特征的能力，但是传统机器学习方法任然具有其强大的生命力。引用 LightGBM 作者柯国霖的一句话“有多少人工就有多少智能，用神经网络的话，你需要结构设计；用传统模型的话，你需要特征工程。”最后就具体实现中存在的缺陷，感觉对于同一种算法，不同的实现方法对于不同的数据也有着不同的效果。所以尽可能多尝试几种框架；还有就是在模型 ensemble 方面，还有一些工作可以尝试。最后感谢各位美团金牌讲师和老师、助教的辛苦付出，感谢队友的协同合作和帮助！

胡潇：

通过本次课程大作业，第一次见识到了真实企业的数据。作为国内领先的智能营销平台，美团点评广告平台的数据可以说是非常详尽，包含了非常多的信息，可值得挖掘的潜力也非常大。虽然放出来作为比赛的数据不多，但可以想象其完整的数据规模也是海量的。从这些数据中通过人工智能技术构建预测模型预估用户对广告的点击概率，从而帮助做搜索广告进行互联网营销，也是很有实用价值和应用潜力的一件事。

在本次大作业中，第一次接触推荐相关的任务，了解了树模型、因子模型、深度模型各自的原理和优缺点，也实际地上手完整体验了从学习使用 pandas 到进行数据处理、搭建模型、构建特征、模型调试、结果分析的整个流程。还在最后的课程报告中了解、学习了其他同学的思路、方法以及他们的实验经验和总结，聆听了美团金牌讲师们的点评和指导，受益匪浅。其中最深刻的一个认识便是对机器学习方法的一个重新认识，之前接触的任务中基本深度学习的方法都比非深度学习的基本方法要更好，固我曾认为深度学习方法是“万能”的，只要模型合理，精细调参，深度模型至少能拟合传统方法在假设空间中的函数，甚至能做得更好。然而在本次大作业中体验到了由于真实数据的限制，目前已有的深度学习的方法可能效果并不好。传统方法不仅在性能上

可能会更好而且在时间复杂度、模型大小等其他方面有自己很大的优势。不过需要细致而繁琐地耗费人力进行特征工程。在提取特征这方面也是听了排名前二的小组的报告和美团讲师的点评得到了很大的启发。此外，未能使用 catboost 和更好的模型 ensemble 方法也是这次大作业的一个遗憾。

总的来说，在本次课程中，不进学到了大数据获取、存储、分析与挖掘技术的理论知识，同时还领略了大数据在美团的商业应用与实践，并通过课程 Project 体验了搭建大数据实际系统，对大数据技术及其应用有了一个全面的了解。感谢队友们对我的帮助，同时非常感谢老师和助教们为我们安排了如此一个有意义的课程，也感谢美团为我们提供的支持，感谢各位美团的金牌讲师们精彩的讲解和指导！

苑苑

本次课程作业，为我们提供了体验企业真实业务场景的机会。在这个过程中，我感受到了特征工程的重要性。最开始进行特征工程，我并没有给予足够的重视，所以第一次的结果应用于模型并没有给出理想的结果。这项繁琐又需要耐心的工作在推荐相关的工作中需要投入更多的时间和精力。将数据输入模型之前，没有进行特征的重要性及相关性分析是我们存在的不足之处，因为只有对数据特征有更多的理解和认识，才能更好的选择模型并应用于模型。在此次作业完成的过程中，我从队友身上学到了很多经验。无论是展示，还是完成报告，团队之间彼此的合作交流都促进了进步和提升。总的来说，课程和作业都让我受益匪浅。感谢老师和助教的辛勤付出，祝《大数据技术的商业应用与实践》课程越办越好！

钟凯

本次课程大作业是我第一次接触大数据相关的任务。之前总是听说大数据技术的重要性，还觉得很神秘，这次通过这门课，不但系统得学习了商业实践中大数据技术的从底层硬件系统、软件框架、处理逻辑到顶层的应用场景、算法设计等诸多技术，而且还有机会利用真实应用场景的数据进行实际任务的开发测试，收获非常多。这次的实验我们组同学都认真投入，通过几次讨论交流，我从几个能力比较强的同学那里学到了不少数据处理的理念和分析 CTR 预估这类问题的方法以及常见模型，加上自己动手处理数据、分析特征等实际操作，最终收获颇多。具体来说，在问题特点的分析和模型的选择上，由于我们对各种方法进行了比较仔细的调研和分析，结合工业界的应用经验，我们适时判断出在工业界能广泛应用的深度模型在解决这一具体的小问题上的局限性，没有过多耗费精力调试。而在因子模型和树模型的调优开发上，我们分为两组、组内独立测试、组间共享数据处理方法的具体开发计划保证了我们的效率和质量。不过这次也有一些做得不足的地方，比如对于特征更多的是从数学角度处理，没有挖掘背后具体内涵；以及不同模型的集成方法没有做特别多的尝试，在今后的学习工作中我们会总结运用本次宝贵实践的经验和教训，做得更好。最后感谢美团各位老师的指导和支持，感谢老师和助教的辛苦付出。

郑瑜

本次课程大作业，我第一次接触了真实场景下的广告数据，并且利用课程所学内容对数据做了分析和建模，运用机器学习的技术和算法解决实际问题，

和队友讨论合作。我之前一直做推荐系统相关的研究，对于题目的背景比较熟悉，因此主动承担了组长的责任，并组织了小组内的讨论和分工，自己也和另外两名队友共同负责因子模型的开发和调试。虽然在平时的研究中一直在做和深度学习技术结合的工作，但是对于数据极其稀疏的情况下我还是判断深度学习模型难以获得令人满意的效果，因此在分工时没有选择深度模型做尝试，而是将小组分成两拨分别研究因子模型和树模型。在最终的课堂展示中，其他选择这一题目的小组均在深度模型上做了失败的尝试，现在回想起来，感觉当时对于问题的全面而深刻的分析确实非常重要，提炼出任务的难点并针对难点设计解决方案使我们小组避免了在深度模型上走弯路。总的来说和各位小伙伴的讨论与合作非常愉快，队友们都非常靠谱，交流也十分流畅，有了明确的分工之后小组的协作与同步也做的不错。最后，感谢美团课程的老师和助教对我们的指导和帮助，也感谢各位美团金牌讲师的讲授。