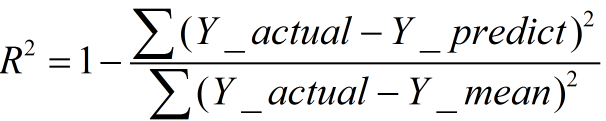
模型推荐模块：

在上文中，经过sql解析以及我们的学习索引模块，我们得到了结构化的输出数据：key以及它在相关字段上的映射position。对于来自于同一数据源的同一类数据，他们的数据分布是相似的。我们希望通过学习这些带查询数据所处的精确位置以及所属类别之间的关系，建立一个自动选择模型。当来自同一数据源的其他数据输入时，我们可以不经过复杂的训练或者去其他不相关类别的字段中寻找，而直接迅速定位到其所属类别的最可能的字段位置。换言之，就是自动的在底层进行了索引推荐。

为了实现这一目标，我们首先需要在同一数据源的不同数据类型上，应用多种类，在某些场景下，个体性能表现都较好的模型，分别进行大量的训练。这些模型包括：弹性回归，岭回归，多项式回归，决策树回归，KNN，线性回归，随机森林回归，径向基回归（rbf），GradientBoostingRegressor（梯度提升回归模型），极限树（ExtraTreeRegressor）。\*\*（可选）细说选择这些模型的原因以及应用场景。

一般而言，相同的数据在不同回归模型上的表现效果是不同的，我们记录了如下参数，用于比较模型的效果以便做出选择，他们包括：

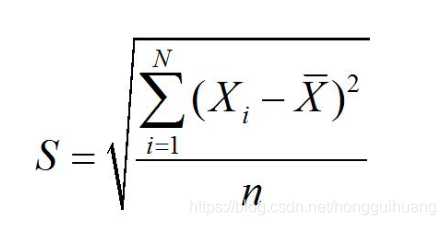
R2\_score（R方）：



Explained\_variance\_score：为了增强可解释性，引入Explained\_variance\_score。它是解释回归模型的方差得分，其值取值范围是[0,1]，越接近于1说明自变量越能解释因变量的方差变化，值越小则说明效果越差。公式如下：

https://img-blog.csdn.net/20160420140735716

RMSE(均方根误差，标准误差)：为了评价数据的变化程度引入的RMSE。RMSE的值越5244小，说明预测模型描述实验数据具有更好的精确度。公式如下



RMSE的计算公式

Mean\_absolute\_error（平均绝对误差）：为了评估拟合结果给出的位置与真实位置的偏离程度，引入Mean\_absolute\_error。公式如下：

https://img-blog.csdn.net/20160420140705013

Mean\_absolute\_error计算公式

Median\_absolute\_error(绝对中位差)：它是对单变量数值型数据的样本偏差的一种鲁棒性测量，效果比方差或者标准差更好。对于这参数的重视有助于提高模型系统的鲁棒性。

https://img-blog.csdn.net/20160420140810341

这些参数对于模型的鲁棒性，准确度，方差变化的可解释性等等都有一定程度的考量。我们收集这些参数，并且，依据他们对于模型不同方面的影响以及我们的实际需求，有侧重的设计打分模型，对于这些模型做出打分。我们注意到在面向索引的回归模型的训练中，过拟合问题往往是有利的，因为这一现象使得我们总是能查询到正确的位置。因此在设计打分函数时不用避讳这个因素。我们的想法是，设计一个线性函数，为上述参数设计合适的权重，用这个结果决定，我们的这个数据输入到底应该采用哪个回归模型进行处理。我们设计的打分函数如下所示：

公式中，，对应上述五个评估参数：均方根误差, R方, 平均绝对误差, 绝对中位差, Explained variance score。wi ， 分别表示w\_R2\_score, w\_EV\_score, w\_RMSE, w\_MAE, w\_MedAE，也就是上述五个评估参数的权重。是正则化系数。由于,并且这两个数越小，代表模型越差，所以进行了公式中的处理，使得他们与其他评估参数一样，都是在单增的情况下表明效果更好。

得到打分之后，我们利用快排对于打分进行排序，得到效果最好的模型的名字。在这之前，需要对之前提到的10个模型标注名字，方便后面对数据进行标注。我们把模型的名字记作，其中，。

然后，对于这轮数据一次标注 他们所属的最佳模型的名字，以标签的方式，附属到数据上。实际上，这种方式是比对于数据属性进行分类更加广义的标注方式，因为同一标签只能说明他们都采用这种模型可以获得最佳效果，可能具有类似的数据分布特征，然而他们的类型或者领域可能是无关的。需要注意的事，因为在来自同一数据源的同一轮数据中，他们的查询的相关字段的位置很可能是临近的。也就是说，有很大的概率，这些数据经过打分步骤之后，大部分只选择了两到三个合适的模型，而剩余的几个模型被选择的覆盖率很低。这种现象不适合我们最终的效果，因为这相当于其他的模型失去了意义。所以，每当我们训练完一轮之后，我们对数据进行打乱操作（shuffle），通过人为设定随机数种子seed, 读取到的数据的random\_state会被指定，在打乱之后才进行测试集和训练集的切割，此时数据的分布情况会被重新定义。同时，这样相当于整个数据都加入了训练，过拟合现象得到了增强，而这也是我们所欢迎的现象。

标注好的数据，最终会送到分类器中进行分类，以获知哪些key（实际上是key所在的相关字段）会选择哪个最优的模型进行回归分析。我们选择的是随机森林分类器。由于面向数据库的数据查询，通常是高维，复杂，数据种类多元化的，在短时间内提取数据特征非常困难。然而，随机森林对于处理高维的数据具有天然优势，并且不用进行特征选择，这是我们选择这一分类器的最重要原因。这个分类器在噪声过大的分类问题上可能会过拟合，但是，由于我们欢迎过拟合行为是我们的计算更加准确，因而这反而成为了我们选择他的另一个理由。此外，他对于对generlization error进行的是无偏估计，模型泛化能力强。并且，训练速度快使得这种方法容易做成并行化方法。

最后，当训练部分结束，我们面临同一数据源内的新的查询时，可以跳过开始时的回归模型训练，直接通过分类器的结果，发现这个查询的相关字段应该用哪种回归模型进行训练效果最佳，并应用它进行回归分析，获取精确结果。这样，就实现了模型的自动选择推荐任务。