**模型方案说明**

众所周知，一个学生所处的环境、班级、宿舍、以及老师都会无形中影响到该学生的发展。同时学生的行为，如考勤情况也会表明该学生这一阶段的学习状况。对于学生学习状态的把控，一直都是一个备受关注的难题。在学生存在成绩下滑的信号的时候，及时捕捉并且预警变得尤为重要。这里我们在已有的数据基础上，使用机器学习的方案，进行探索性的挖掘。并取得了一定的收效。

我们将模型要解决的问题定义为：通过学生历史成绩，以及当前学期所属的环境和行为数据，构造特征并建模来预测下一次考核时该学生的等第。最终模型达到相比于传统的规则更优秀的预测效果。下面我们介绍方案细节。

**目标变量**：**等第**

我们采用等第作为目标变量，原因是考虑掉未来考试的难度是未知的，故预测成绩会存在误差，而等第则是表明该学生在群里种的排名，一般不会受到太多考试难度影响。

**特征工程:**

|  |  |
| --- | --- |
| 历史成绩 | 上一次的等第 （直接强相关）  前三次等第均值（提高稳定性）  上一次考试等第的变化（增加趋势）  等第的方差 （表征偏科程度，以及成绩的稳定性） |
| 教师画像  班级画像 | 班级ID、教师id、该教师下平均学生等第  该班级平均等第、舍友平均成绩 |
| 学生画像 | 学生ID、性别、民族、出生日期（年）、家庭类型、政治面貌 |
| 考勤记录 | 该学生当前学期迟到次数、该学生当前学期早退次数 |
| 消费记录 | 当前学期消费次数、当前学期消费总金额、当前学期最大一笔消费金额 |

**模型：**

这里我们选择了工业界十分优秀的lightgbm 模型建模，该模型训练速度快，易部署，鲁棒性好，且具有良好的特征筛选能力。通过参数配置即可实现对于一些低质量信号及特征的过滤。

**验证：**

线下通过时序验证的思路，使用最后一个学期的考试成绩作为验证集，其他的作为训练集，模型通过学习特征和历史成绩之间的联系，从而达到预测下一个学期成绩的效果。

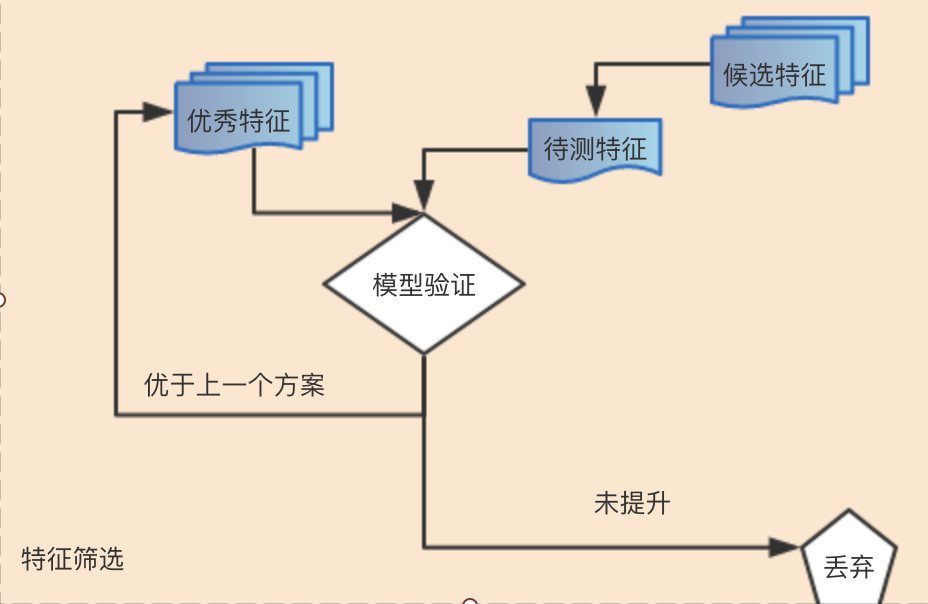
训练数据（2018-2019-1 学期之前）

测试数据（2018-2019-1）

考试记录表

**特征筛选:**

由于训练速度很快，故这里使用了包裹式特征选择方案，包裹式特征选择是指，以数据模型的实际效果为特征选择标准。在操作时，可以有前向选择，即从一个特征开始，在数据模型中每次都添加一个新的特征，然后看验证集误差是否有降低。如果误差降低了，则保留这个特征，继续添加下一个；如果误差没有降低，则删除该特征，尝试下一个。

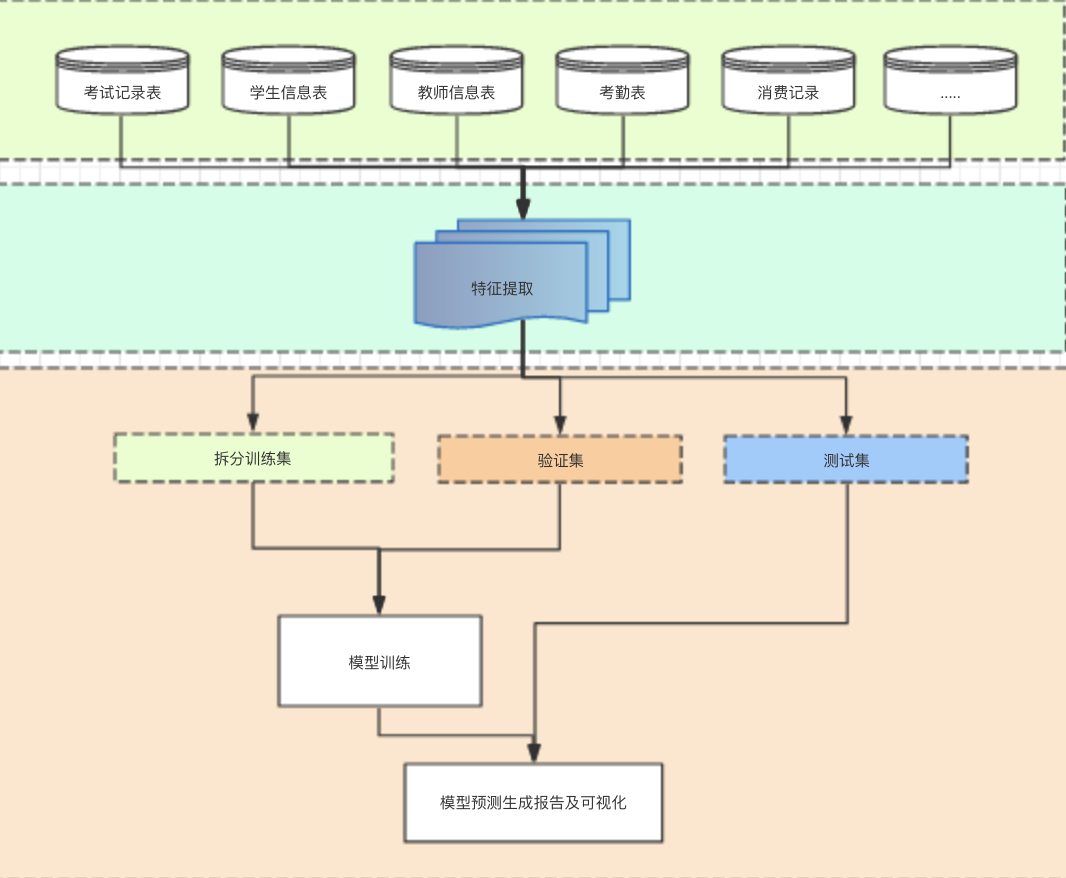


**生成报告:**

预测好的模型可以对每个学生的每个科目进行下一次考试的等第预测。最终作为一项指标落地到学生的个人页面上，同时对于影响结果的核心指标，（例如多次等第下滑、近期等第过低等）我们使用规则引擎生成预警报告，从而达到对模型结果的可解释报告呈现。



**模型架构图：**



**总结：**

通过对已有数据的建模，最终我们的模型结果的最终误差在0.2个等第左右，普遍优于传统规则预测的效果。（最近值预测、趋势预测、均值预测等），在拥有更多的数据的场景下，

模型效果可以得到更进一步的提升。

**不足之处：**

由于数据可用特征并不是特别多，所以模型并没有得到一个非常质变的提升。

我们也很期待，后续可以接触到更多的数据，进一步优化模型。达到更为精准的把握学生变化的信息。