好未来AI Lab

数据挖掘　图像处理　自然语言处理　语音

项目： 业务部门数据分析与预测
数据挖掘平台建设

优势： 数据丰富
迭代速度快

学生续报预测

1 数据分析

2 续报预测

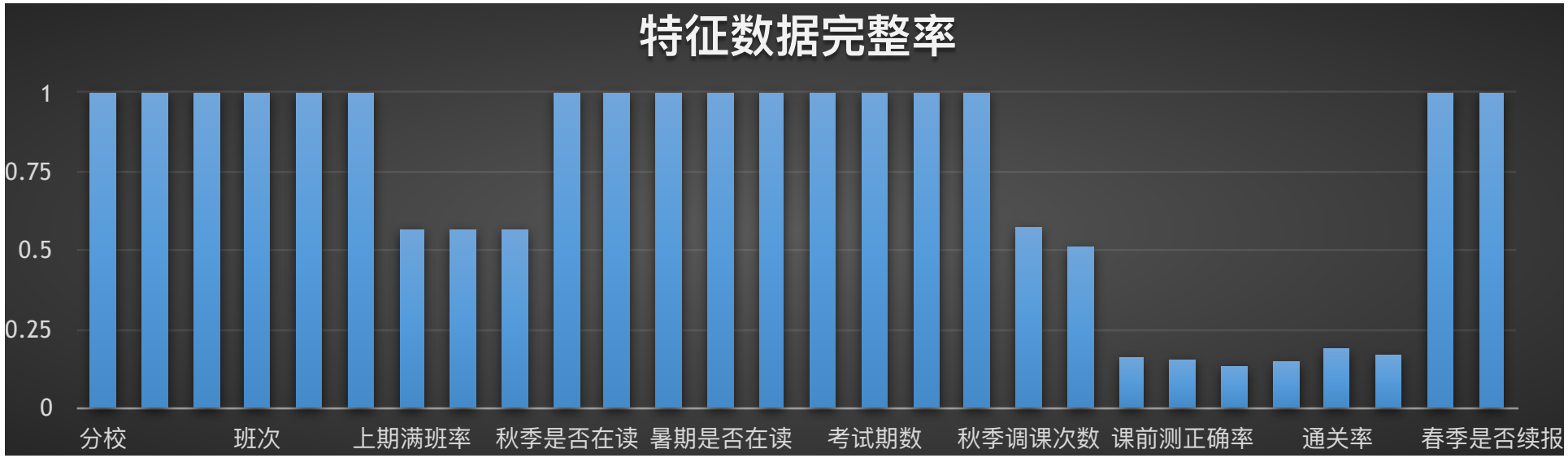
3 总结

数据分析

长期班数据： *65*万个学生样本 *(34*维特征 、*2*个待预测量：春季和寒假是否续报*)*

去除不相关特征：学员*ID*、班级*ID*、学员编码 ……

去除数据缺失较多特征：学员评分、课前测正确率 …….



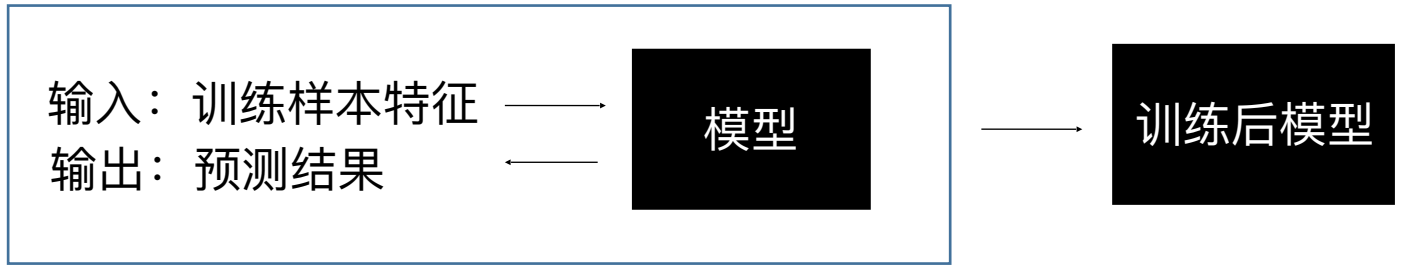
剩余20维特征：分校 年级 学科 班次 教师 是否新生 报班总课次 长期班期数….

特征预处理：异常数据消除、缺失数据填充、类别变量编码、标准化

续报预测

机器学习：

训练：



预测：

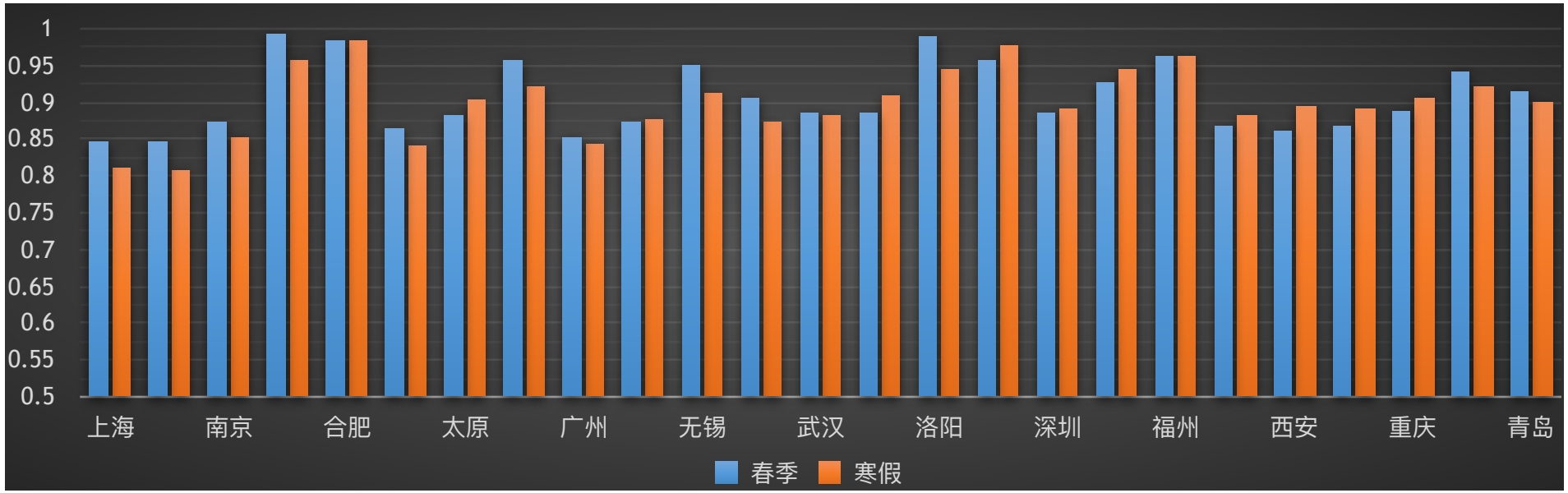
模型： 随机森林(树的数目、深度、特征数目等)

数据： 65w条学生样本（训练，验证，测试）

续报预测

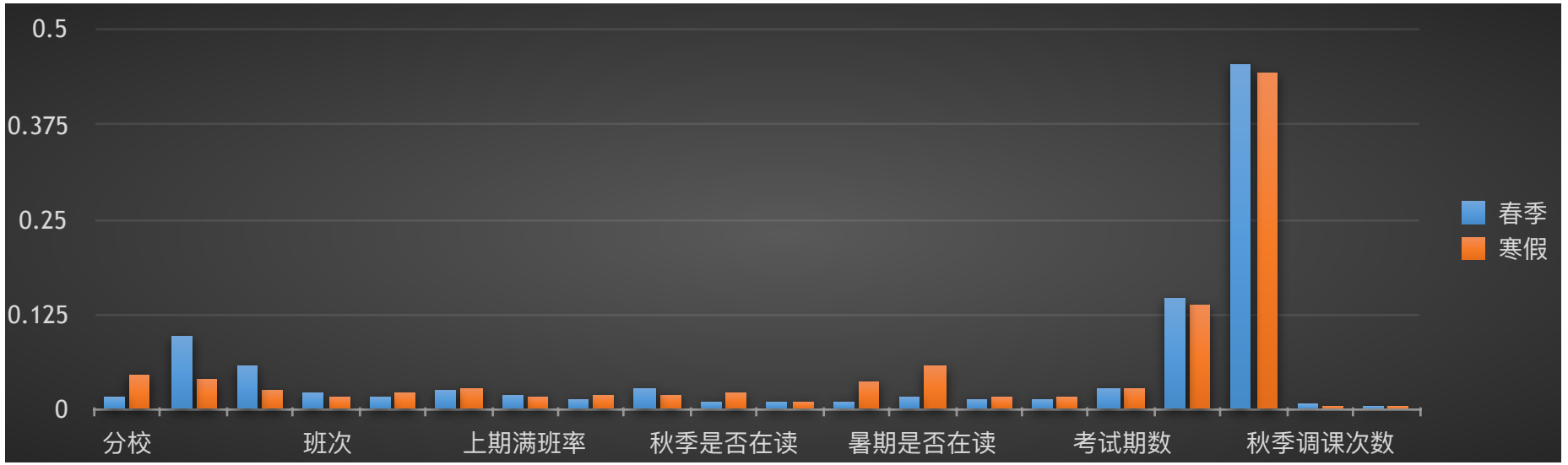
春季学生续报准确率：*87.7%*　　寒假学生续报准确率：*86.4%*

各分校学生准确度：



续报预测

特征重要性打分：



春季续报前三重要特征：长期班期数、报班总课数、年级

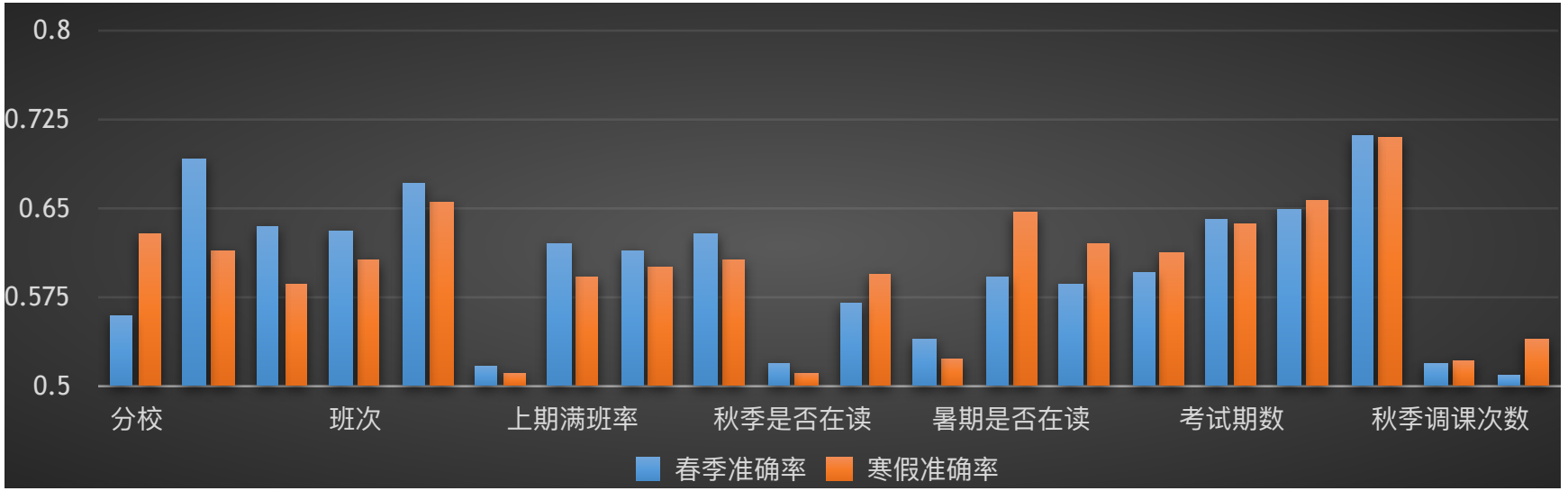
寒假续报前三重要特征：长期班期数、报班总课数、暑期是否在读

只采用前三重要特征，春季和寒假续报准确率

	春季	寒假
随机森林	*85.5%*	*85.8%*
Adaboost	*85.2%*	*85.3%*

续报预测

单个特征对应续报预测精度（每次只用一个特征）：



春季续报预测准确度前五特征：长期班期数、年级、教师、报班总课数、考试期数

寒假续报预测准确度前五特征：长期班期数、教师、暑期是否在读、报班总课数、考试期数

续报预测

每个学生续报的概率值

分校	学员ID	班级ID	学员编号	学员姓名	年份	考试周数	试听总课次	长期班期数	秋季课课次	秋季转	寒假续报预测概率	寒假是否	春季续报预测概率	春季是否续报
上海	0b1d0cbcc	2280808[illegible]	2[illegible]6402[illegible]	胡奕	2016	1.0	22.0	5.0	2.0		0.866971	是	0.935161	是
上海	1b9ba9b65	2280808[illegible]	2[illegible]2000006[illegible]	贾二乐	2016	3.0	44.0	13.0	2.0		0.[illegible]9726	是	0.908701	是
上海	17d9[illegible]27[illegible]	2280808[illegible]	2[illegible]5000060[illegible]	徐洁彦	2016	0.0	20.0	5.0			0.[illegible]8049	是	0.941	是
上海	27a434hhr	228080[illegible]	4[illegible]0000608	储晴鹤	2016	0.0	2.0	4.0	6.0	4.0	0.46[illegible]3	否	0.902431	是
上海	28be6d07[illegible]	228080[illegible]	2[illegible]4000060[illegible]	叶子睿	2016	1.0	74.0	[illegible]1.0	4.0		0.877857	是	0.890346	是
上海	2aea207h3	228080[illegible]	5[illegible]0000603	倪丞伟	2016	0.0	44.0	7.0		4.0	0.[illegible]8748	是	0.890512	是

不同续报(不续报)阈值对应精确率和召回率

	春季		寒假	
续报阈值	精确率	召回率	精确率	召回率
P > 0.7	91.5%	45.8%	92.05%	43.7%
P > 0.9	97.4%	28.3%	98.1%	30.2%

	春季		寒假	
不续报阈值	精确率	召回率	精确率	召回率
P < 0.3	92.0%	50.0%	92.8%	48.0%
P < 0.1	98.5%	31.2%	98.3%	30.8%

总结

1、利用机器学习模型，可以得到每个学生续报概率（准确度超过85%）。

2、利用机器学习模型，可以得到影响学生续报的特征的打分

谢谢