**代码：**C:\Users\28079083\OneDrive - Anheuser-Busch InBev\Desktop\2023项目\9.turnover\Best-1228 notebook

**数据：**C:\Users\28079083\OneDrive - Anheuser-Busch InBev\Desktop\2023项目\9.turnover\data

* 预测数据：截止到2023年10月底的在职员工数据
* Train\_data: 12.26, 和27 Taoye发了一波补充数据，train data是第一版模型的处理数据，后面的补充数据添加，我都是在train\_data数据上直接加的，不需要重跑前面的数据处理过程
* forecast\_fte：第一版模型的预测数据，后面也根据补充数据和新features加入做了修改，在model\_result里有预测数据的全部

**模型标签维护：**

C:\Users\28079083\OneDrive - Anheuser-Busch InBev\Desktop\2023项目\9.turnover\模型标签。Xlsx

**结果：**

**For sharon的分析+汇报：**

C:\Users\28079083\OneDrive - Anheuser-Busch InBev\Desktop\2023项目\9.turnover\for Sharon

**Sharon汇报版的模型结果：**

C:\Users\28079083\OneDrive - Anheuser-Busch InBev\Desktop\2023项目\9.turnover\model\_data\model\_result1227

**结果内容：**

变量分析，模型的混淆矩阵，预测结果，所用的train & test data都在model\_result excel里

Final\_train.xlsx是抽样前的数据，模型的训练和测试数据是根据final\_train抽样的结果

**负样本：**

Turnover Report\_Procurement\_202212 【2018-2022】

Termination Type（V列）：Termination-Involuntary的都不要

Country key只筛选中国

**正样本：**

Turnover Report\_Procurement\_202310 APAC 里面的FTE sheet看上去是2023年在职人，这是要预测的数据

目标：预测未来半年是否会离职

模型：catboost

离职数据

1. Macro Entity=BU CHINA、ZONE ASIA PACIFIC
2. LeaveType 去掉involuntary

**建模的数据操作：**

对于离职的人，把这些人之前在职的每一年的数据补充到模型数据里，数据最早到2018年

数据的流失标签：

1. Is\_fte = 1是没流失，表示在职，0表示离职，模型的标签跟is\_fte的标签有半年的gap来保证是预测未来半年是否流失
2. Duration: 也是可以用来预测是否流失，duration是这个人的预测的tenure，如果(duration-预测时间点) 距离预测时间节点超过半年,就是不会流失 （但是这个模型没建，后面可以试试）

**问题：**

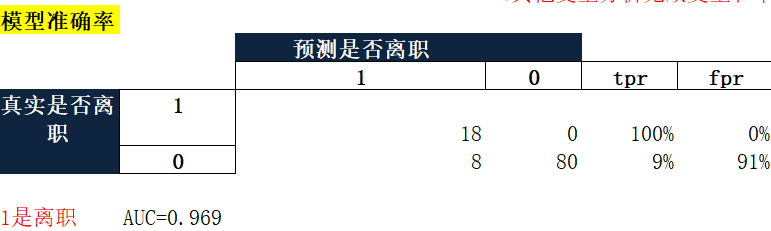
Sharon要求engagement相关的变量EE和ME的重要性很高

1. Model\_result的变量重要性里把manager Effectiveness 和employee effectiveness都强行调高了
2. EE和ME数据缺失严重，用line manager的line manager的数据和org的数据填充

**TODO:**

把代码搬到databricks上画树的图

**离职率:** distinct人-30%；不distinct 人-9%（数据里，by人by year）



**最终数据的缺失情况**

1. 模型数据的

