1. 任务基本情况

考虑到个人理论基础和动手能力较弱，故本次选题采用DC竞赛平台较简单的训练赛——员工离职预测训练赛。竞赛任务是根据31个维度的属性，1100条训练集，对员工是否已经离职其中一个维度.进行预测。上述31个维度具体说明如下：

1.Age：员工年龄

2.Attrition：员工是否已经离职，1表示已经离职，2表示未离职，这是目标预测值；

3.BusinessTravel：商务差旅频率，Non-Travel表示不出差，Travel\_Rarely表示不经常出差，Travel\_Frequently表示经常出差；

4.Department：员工所在部门，Sales表示销售部，Research & Development表示研发部，Human Resources表示人力资源部；

5.DistanceFromHome：公司跟家庭住址的距离，从1到29，1表示最近，29表示最远；

6.Education：员工的教育程度，从1到5，5表示教育程度最高；

7.EducationField：员工所学习的专业领域，Life Sciences表示生命科学，Medical表示医疗，Marketing表示市场营销，Technical Degree表示技术学位，Human Resources表示人力资源，Other表示其他；

8.EmployeeNumber：员工号码；

9.EnvironmentSatisfaction：员工对于工作环境的满意程度，从1到4，1的满意程度最低，4的满意程度最高；

10.Gender：员工性别，Male表示男性，Female表示女性；

11.JobInvolvement：员工工作投入度，从1到4，1为投入度最低，4为投入度最高；

12.JobLevel：职业级别，从1到5，1为最低级别，5为最高级别；

13.JobRole：工作角色：Sales Executive是销售主管，Research Scientist是科学研究员，Laboratory Technician实验室技术员，Manufacturing Director是制造总监，Healthcare Representative是医疗代表，Manager是经理，Sales Representative是销售代表，Research Director是研究总监，Human Resources是人力资源；

14.JobSatisfaction：工作满意度，从1到4，1代表满意程度最低，4代表满意程度最高；

15.MaritalStatus：员工婚姻状况，Single代表单身，Married代表已婚，Divorced代表离婚；

16.MonthlyIncome：员工月收入，范围在1009到19999之间；

17.NumCompaniesWorked：员工曾经工作过的公司数；

18.Over18：年龄是否超过18岁；

19.OverTime：是否加班，Yes表示加班，No表示不加班；

20.PercentSalaryHike：工资提高的百分比；

21.PerformanceRating：绩效评估；

22.RelationshipSatisfaction：关系满意度，从1到4，1表示满意度最低，4表示满意度最高；

23.StandardHours：标准工时；

24.StockOptionLevel：股票期权水平；

25.TotalWorkingYears：总工龄；

26.TrainingTimesLastYear：上一年的培训时长，从0到6，0表示没有培训，6表示培训时间最长；

27.WorkLifeBalance：工作与生活平衡程度，从1到4，1表示平衡程度最低，4表示平衡程度最高；

28.YearsAtCompany：在目前公司工作年数；

29.YearsInCurrentRole：在目前工作职责的工作年数

30.YearsSinceLastPromotion：距离上次升职时长

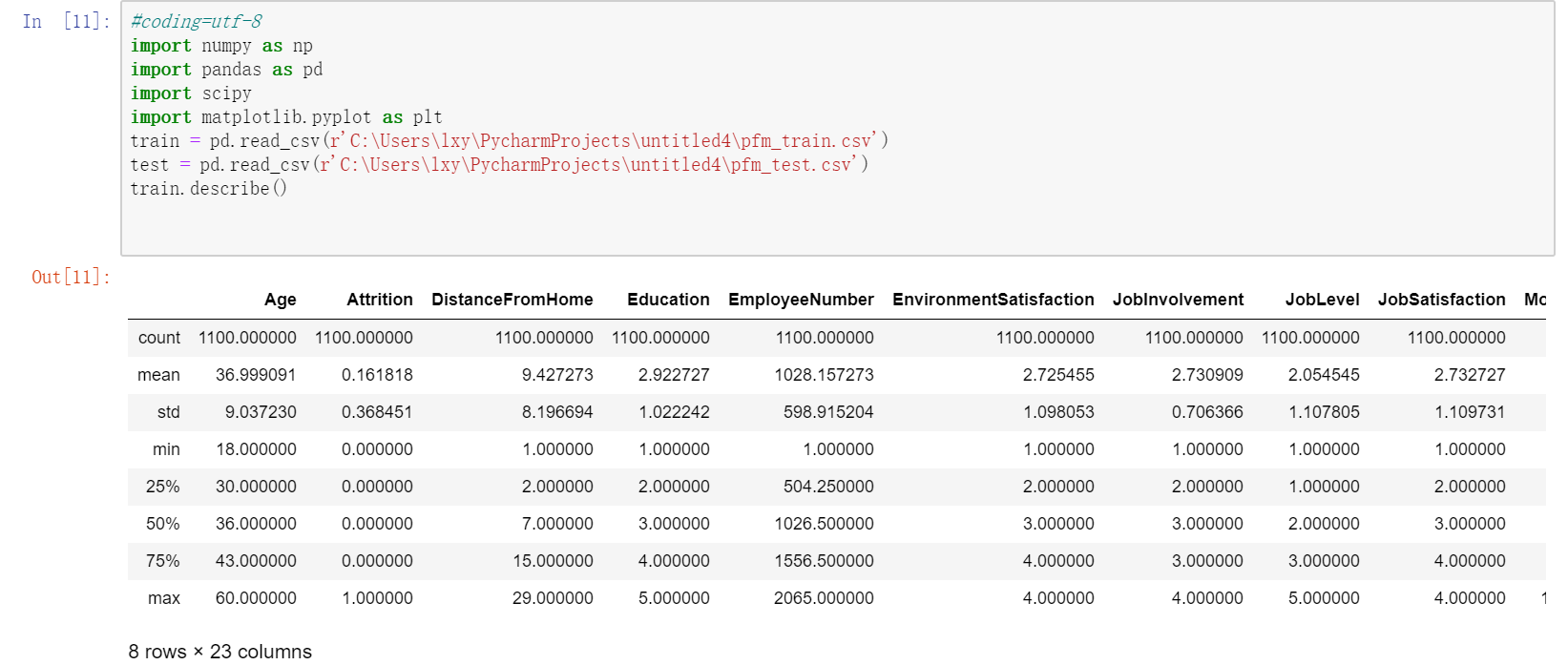
31.YearsWithCurrManager：跟目前的管理者共事年数；

二、初步构想

由于预测目标为是否离职的二元属性值，考虑采用分类算法。根据老师在课上讲授的理论知识，随机森林处理这种问题比较合适。由于随机森林方法存在一些超参数，故考虑后续采用网格搜索优化模型。另一方面，由于题目提示使用逻辑回归模型，故两种方法同时采用。由于个人计算机运算能力有限，考虑在数据预处理时进行合理的降维处理和连续数值离散化以减少运算时间。

三、数据预处理

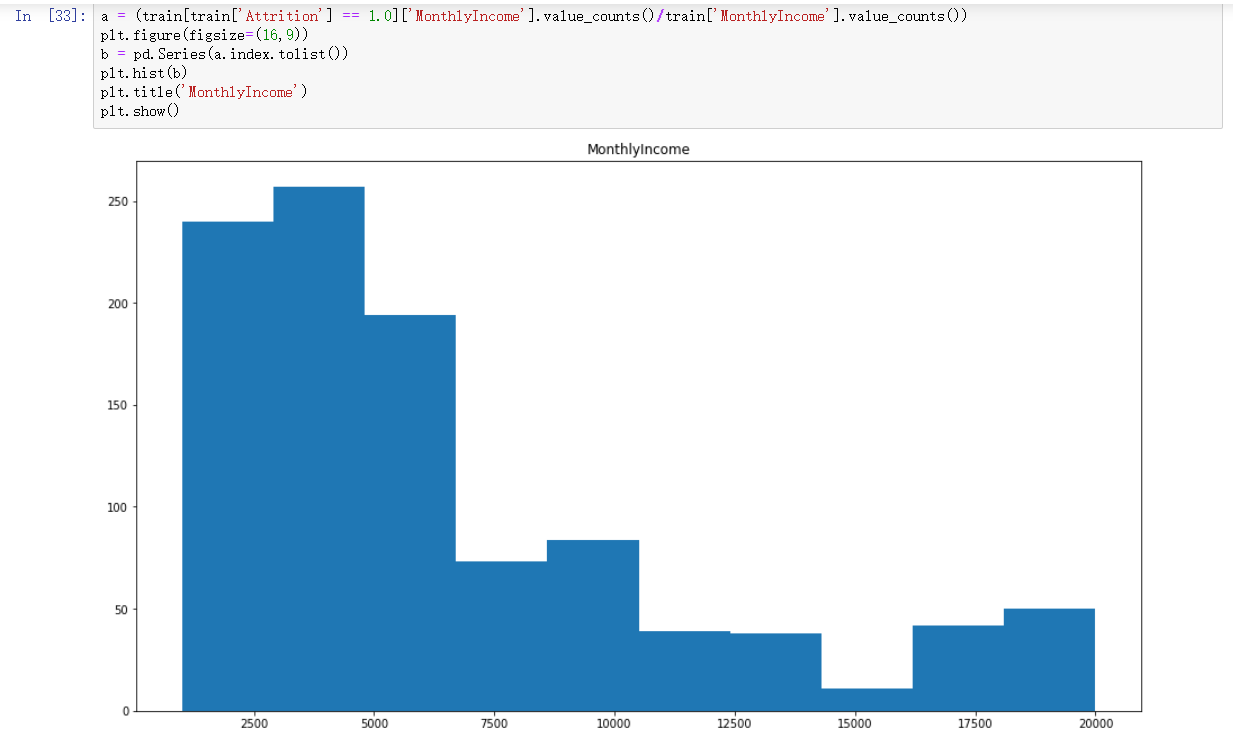
第一步，用describe观察训练集数据基本情况。观察到：1.数据无缺失，故后续无需进一步处理2.部分属性的考察，无论从常识角度出发，还是从数据本身角度出发，都几乎不对预测目标值产生影响，可以删去。如：“EmployeeNumber”，“StandardHours”，“Over18”



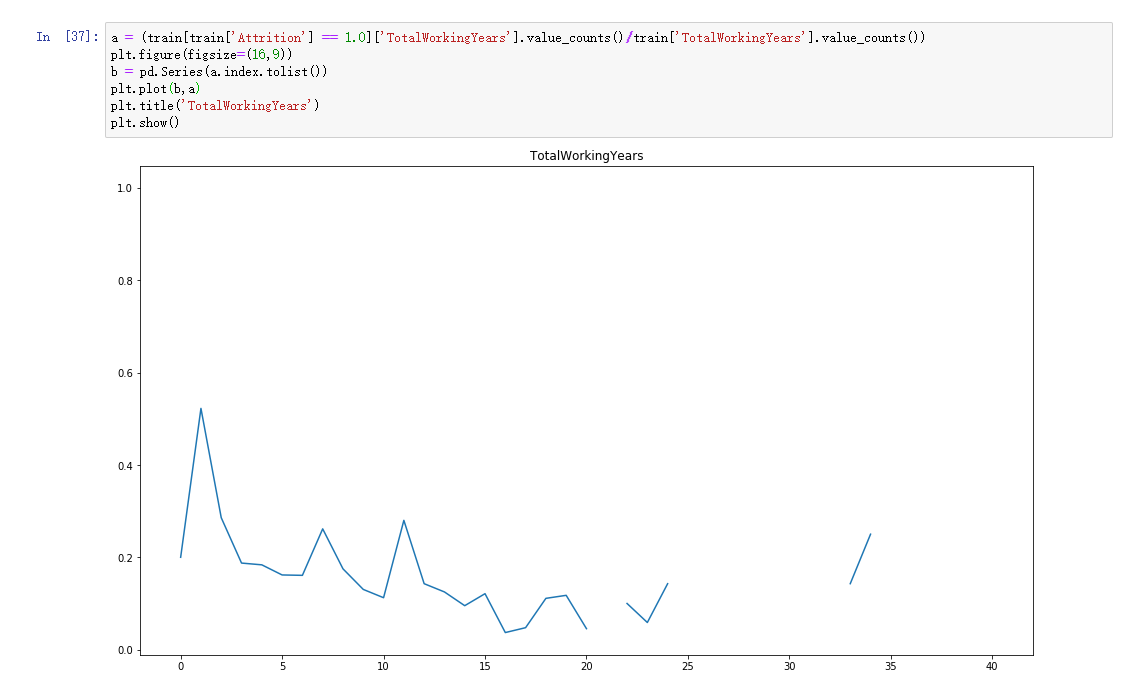
第二步，进一步考虑各个属性对目标值的影响。结果显示：1.特征值为字符串形式的各个属性，普遍具有较好的区分度（性别属性Gender除外）。2. 一部分特征值为数值形式的属性对预测目标值的影响比较容易直接得出（如：年龄小、受教育程度低、总工龄短和关系满意度低等可能导致离职率偏高），另一部分还需进一步图形化处理，才能得出准确结论。



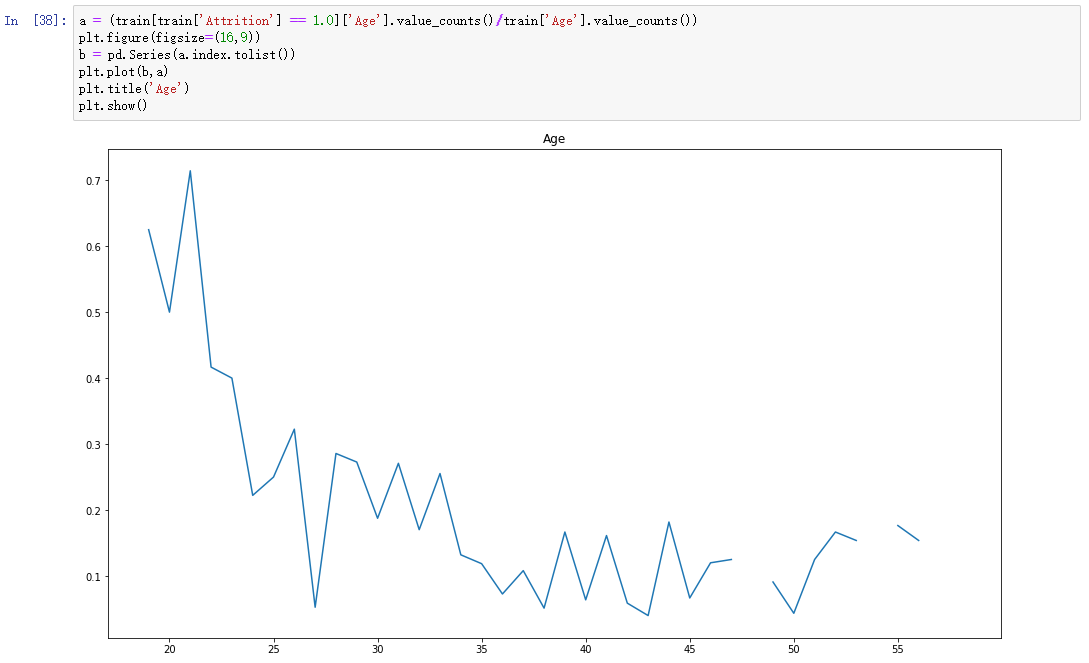
第三步，对离散值较多的属性进行图形化处理和离散化处理（如年龄、总工龄、月收入和在目前公司工作年数）。



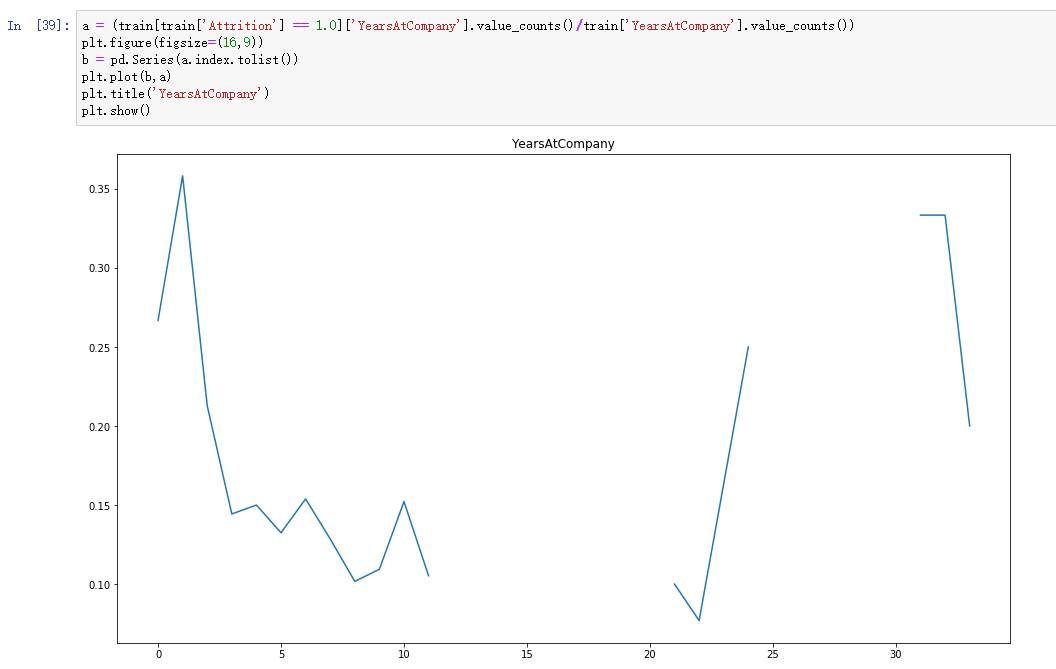
根据绘图情况大致将月收入分为三组：1000-6500.5置为0,6500.5-15000.5置为1,15000.5-20000置为2



将总工作年数为两组：0-25.5置为0，25.5-40置为1



将年龄分为两组：18-33.5置为0，33.5-60置为1

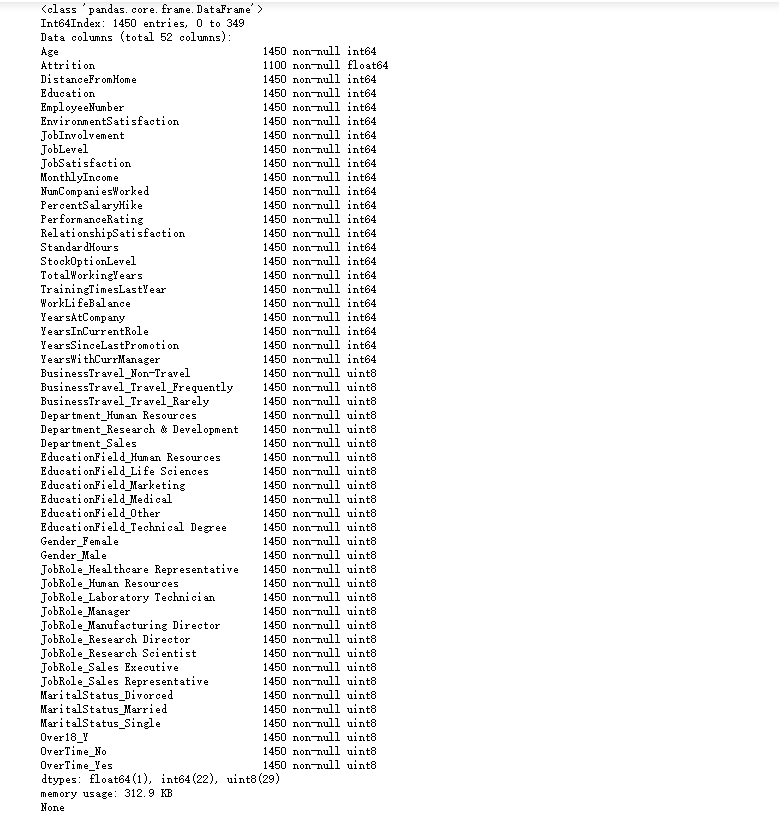


将在目前公司工作年数分为两组：0-12.5置为0，12.5-37置为1。

第四步，合并训练集和数据集，对除目标值外的属性进行相同的预处理。一是删除“EmployeeNumber”，“StandardHours”，“Over18”三个无关属性；二是用上述离散化参数完成部分属性值的离散化处理；三是对类型为“object”的属性值进行“one-hot”编码。



输出结果如下：



四、数据建模和预测



观察到两种模型的预测结果中，逻辑回归表现竟然更好，故将逻辑回归模型预测出的结果提交。





五、实验结果分析

上述操作第四步中的随机森林模型中的参数是已经利用网格搜索优化后的结果，但预测效果仍然不如逻辑回归。

