**Bank Marketing**

**分类任务实验报告**

**毛怡伟 21851168**

# 实验目的

* 熟悉数据预处理的基本方法，包括缺失值填充、数据编码。
* 熟悉pandas、scikit等数据分析库的使用。
* 熟悉机器学习算法交叉验证方法。
* 熟悉混淆矩阵、PR曲线、ROC曲线和AUC值。
* 分析正则项参数对Logistic Regression性能的影响。
* 分析最大树深度对决策树性能的影响。
* 分析随机森林随着个体数的性能收敛情况，以及个体树深度对随机森林性能的影响。
* 比较决策树、随机森林在Bank Marketing数据集上的差异，并分析产生这种从差异的原因。

# 数据预处理

## 数据描述

Bank Marketing数据集用于预测一个客户购买理财产品的成功率。

数据输入的属性包括客户基本信息，当前营销活动信息，其它信息和社会经济背景信息。

* Age 年龄（数字）
* Job 工作类型（分类：“管理员”，“蓝领”，“企业家”，“女佣”，“管理”，“退休”，“自营职业”，“服务”，“学生” “技术人员”， “失业”， “未知”）
* Marital 婚姻状况（分类：“离婚”，“已婚”，“单身”，“未知”;注：“离婚”是指离婚或丧偶）
* Education 教育（分类：“basic.4y”，“basic.6y”，“basic.9y”，“high.school”，“illiterate”，“professional.course”，“university.degree”，“unknown”）
* Default 默认信用额度？ （分类：“不”，“是”，“未知”）
* Housing 有住房贷款吗？ （分类：“不”，“是”，“未知”）
* Loan 有个人贷款吗？ （分类：“不”，“是”，“未知”）
* Contact 联系沟通类型（分类：“手机”，“有线电话”）
* Month 上一个联系月份（分类：“jan”，“feb”，“mar”，...，“nov”，“dec”）
* Day\_of\_week 最后一个联系日（分类：“mon”，“tue”，“wed”，“thu”，“fri”）
* Duration 上次联系持续时间，以秒为单位（数字）。
* Campaign 此营销活动系列期间和此客户的联系次数（数字，包括最后一次联系）
* Pdays 从上一个营销活动上次联系客户后经过的天数（数字; 999表示之前未联系过客户）
* Previous 此营销活动前联系客户的次数
* Poutcome 上一次营销活动的结果（分类：'失败'，'不存在'，'成功'）
* emp.var.rate：就业变化率 - 季度指标（数字）
* cons.price.idx：消费者价格指数 - 月度指标（数字）
* cons.conf.idx：消费者信心指数 - 月度指标（数字）
* euribor3m：euribor 3个月费率 - 每日指标（数字）
* nr.employed：员工人数 - 季度指标（数字）

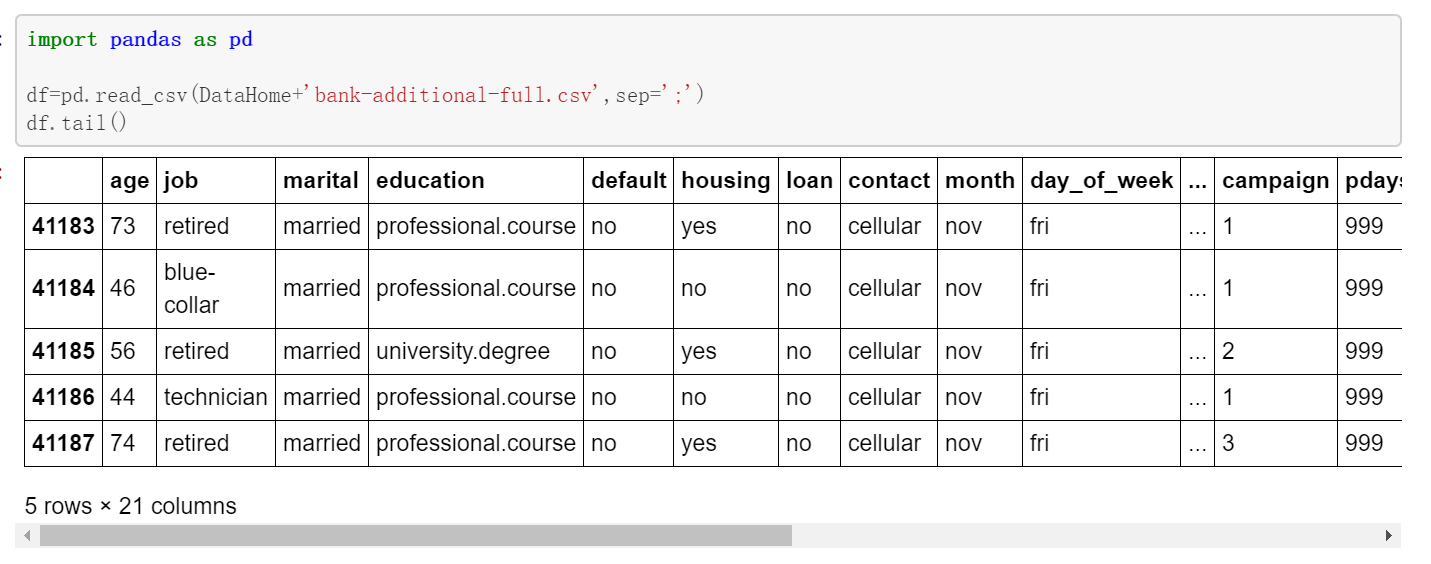
数据输出包括：

* Y 客户是否会购买产品 （分类：是、否）

另外，数据中可能存在缺失值（用unknown表示）

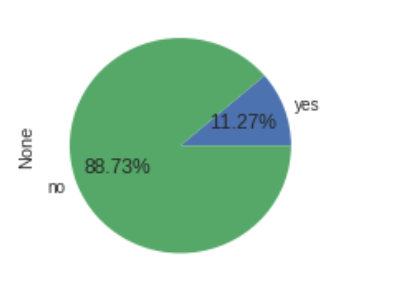
## 数据预览

实验基于pandas可以读入数据，可以通过tail方法获得样本的一些数据：



图表 1 基于pandas进行数据概览

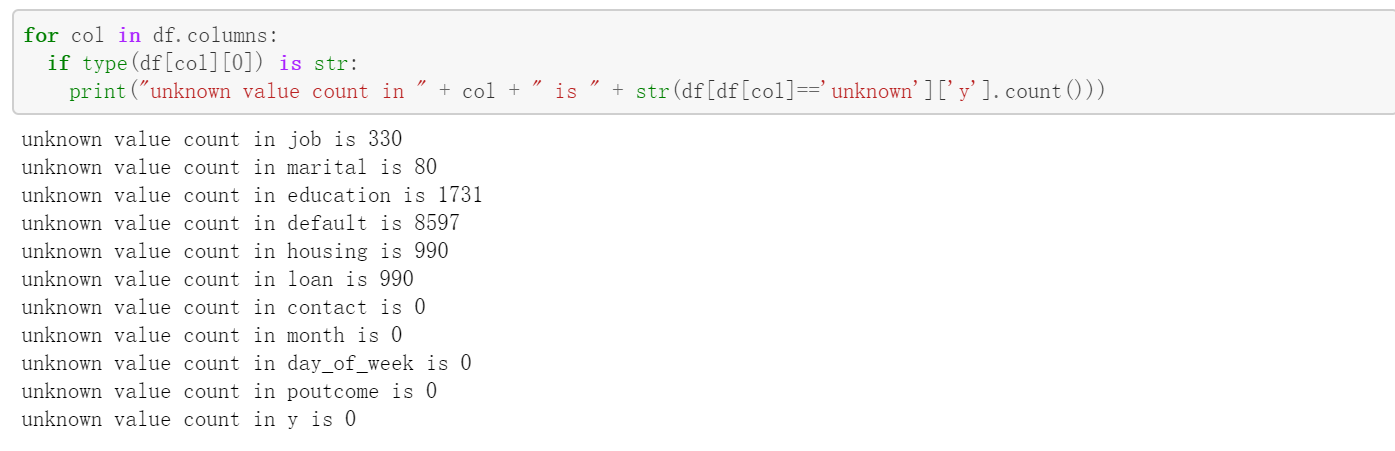
进一步分析数据中正负样本的比例，可以看到，正负样本的比例并不是很均匀，购买产品的客户占比不多。



图表 2 正负样本比例

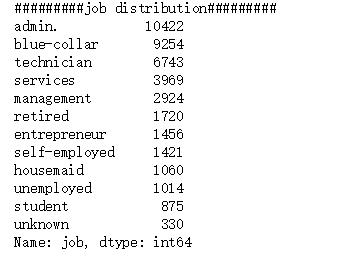
## 缺失值处理

首先分析有哪些列存在unknown数据：

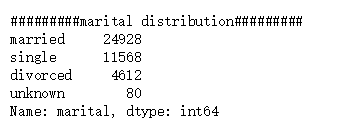


图表 3 可能存在缺失值的属性

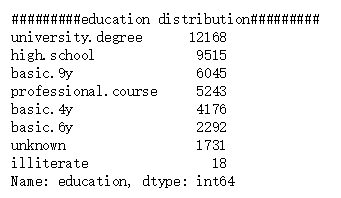
可以看到，job、marital、education、default、housing、loan这些属性都存在缺失值。我们首先通过pandas的value\_counts分析出存在缺失值的属性的各个分类的数量分布情况。



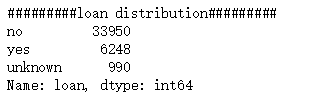
图表 4 Job属性各个分类分布情况



图表 5 marital属性各个分类分布情况

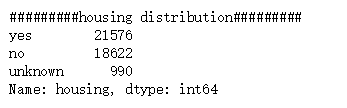


图表 6 education属性各个分类分布情况



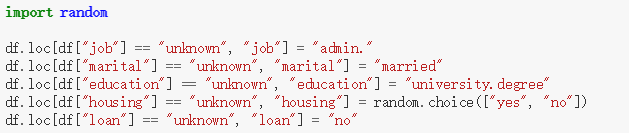
图表 7 loan属性各个分类分布情况

job、marital、education、loan这些属性，占比最高的分类的比例远远高于其它属性，因此我们直接用占比最高的分类作为缺失值的填充值。

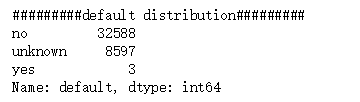


图表 8 housing属性各个分类分布情况

Housing属性yes和no的比例比较相近，因此采用随机选取的方法进行yes和no的选择。



图表 9 数据缺失值处理方法



图表 10 default属性各个分类分布情况

另外注意到default属性中yes的实例只有3个。这个属性对机器学习算法没有任何意义，因此我们直接丢弃这个属性。

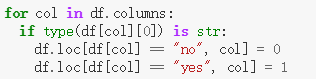


图表 11丢弃default属性

## 数据编码

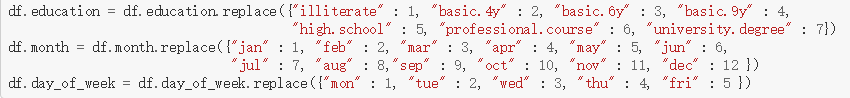
数值数据不需要进行编码，因此这里不阐述。

二分类变量包括housing和loan，直接编码为0和1即可。



图表 12 二分类变量编码方法

有序类别变量包括education, month, day\_of\_week，这种数据可以根据分类的先后顺序进行编码：



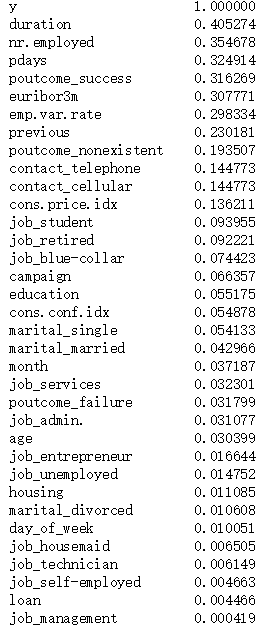
图表 13 有序类别变量编码

无序分类数据包括job, contact, poutcome, marital。无法直接和有序类别变量那样进行分类，需要转化为多维数据。这里采用pandas的get\_dummies帮助无序变量进行编码：



图表 14 为无序分类生成哑数据

经过编码后，可以分析各个属性和分类之间的相关系数：



图表 15 各个属性和y分类的相关系数

# 算法评估方法

## 交叉验证

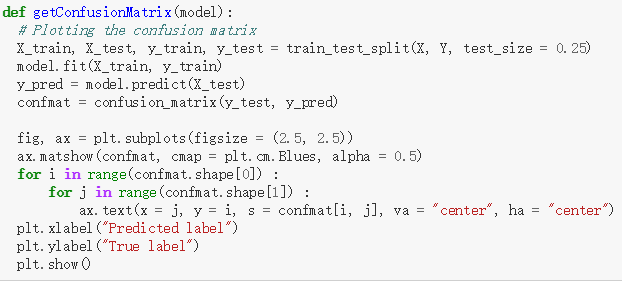
我们基于scikit的ShuffleSplit进行交叉验证方法的定义。我们定义五次五折交叉验证：



图表 16 定义交叉验证

## 混淆矩阵

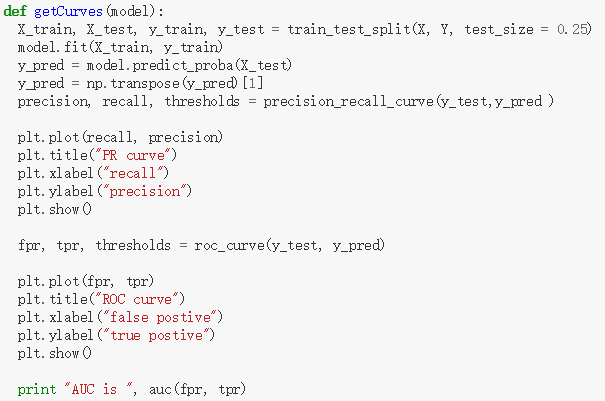
一个正确值无法反应模型的真正的性能，我们引入混淆矩阵，可以分析出购买客户和未购买客户各自的分类性能：



图表 17 混淆矩阵的绘制

## PR曲线、ROC曲线和AUC

我们基于PR曲线、ROC曲线和AUC来进一步分析算法的性能。

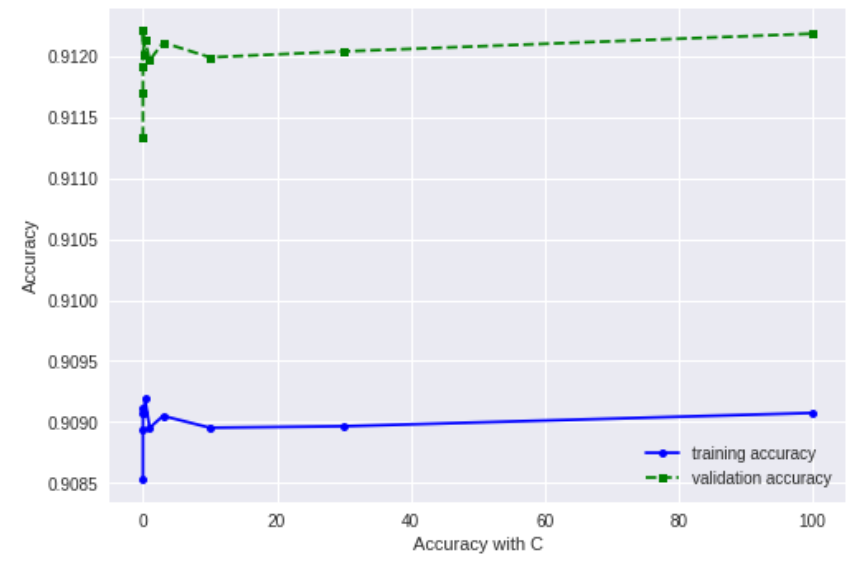


图表 18 PR曲线、ROC曲线和AUC

# Logistic Regression

## 正则化系数的调整

我们分析正则化系数对Logistic Regression的性能影响，可以看到，正则化系数对算法性能的影响基本不大。



图表 19 正则化系数和模型准确率关系

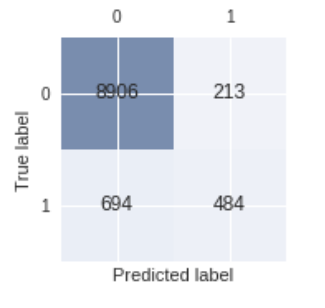
## 准确率随训练样本数的变化



图表 20 LR准确率随训练样本数的变化

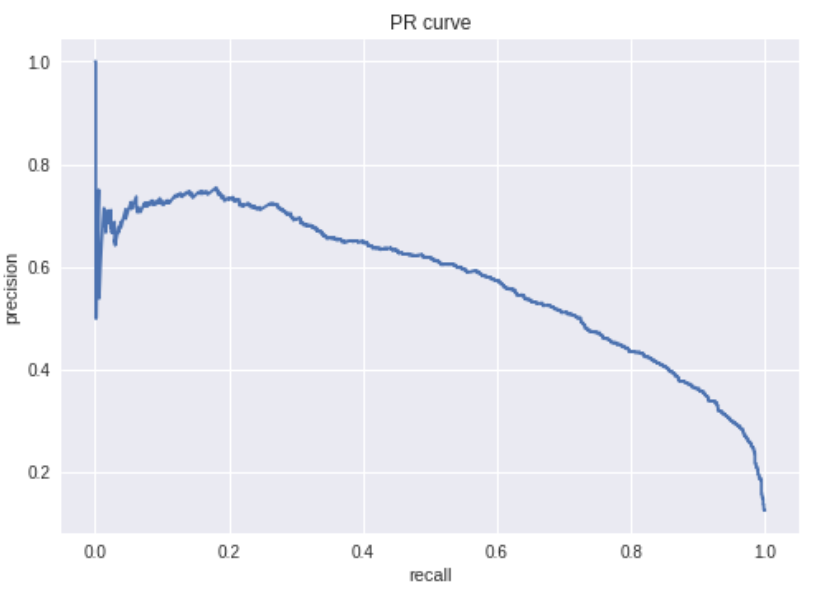
## 混淆矩阵

可以看到虽然模型整体准确率较高，但是类别的不均匀使得未订购客户的分类性能较好，而判断订购客户的分类性能表现较差。

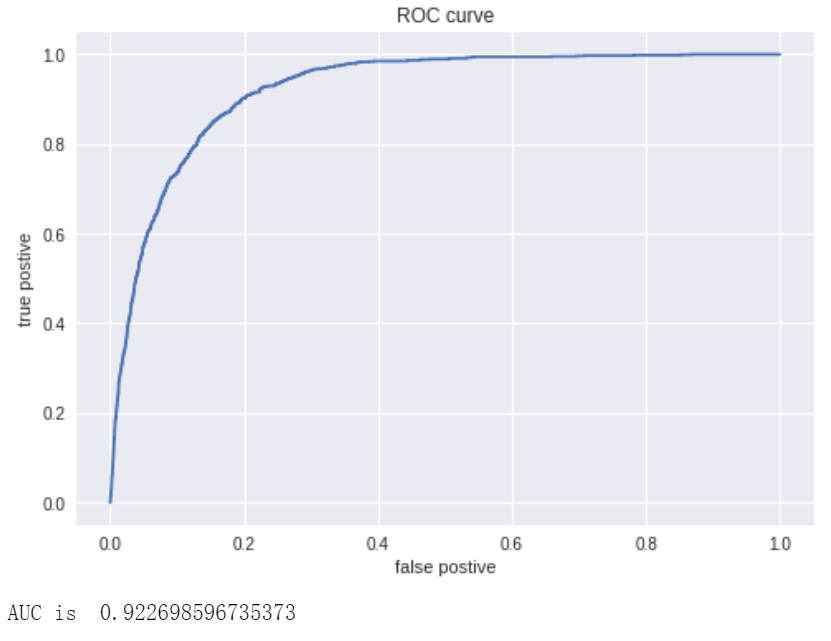


图表 21 LR混淆矩阵

## PR曲线、ROC曲线和AUC值



图表 22 LR的PR曲线

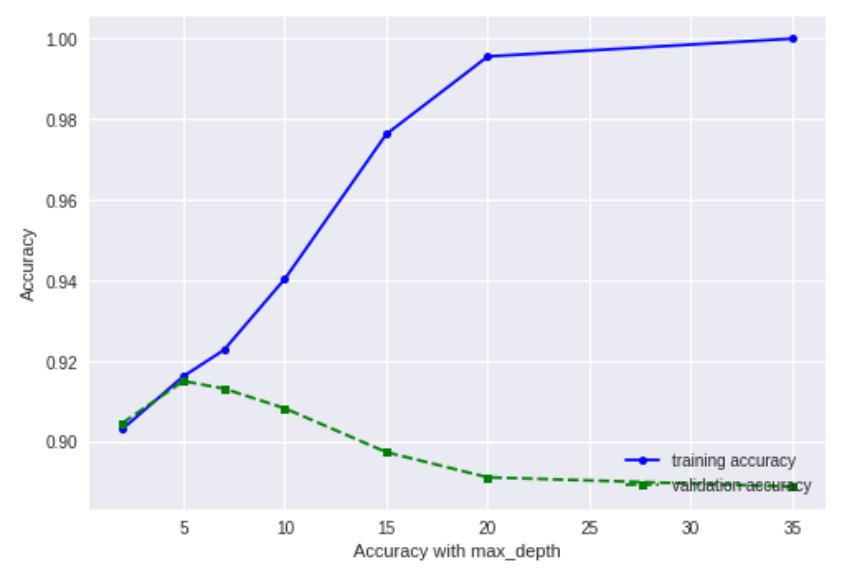


图表 23 LR的ROC曲线与AUC

# 决策树实验

## 最高树深度对性能的影响

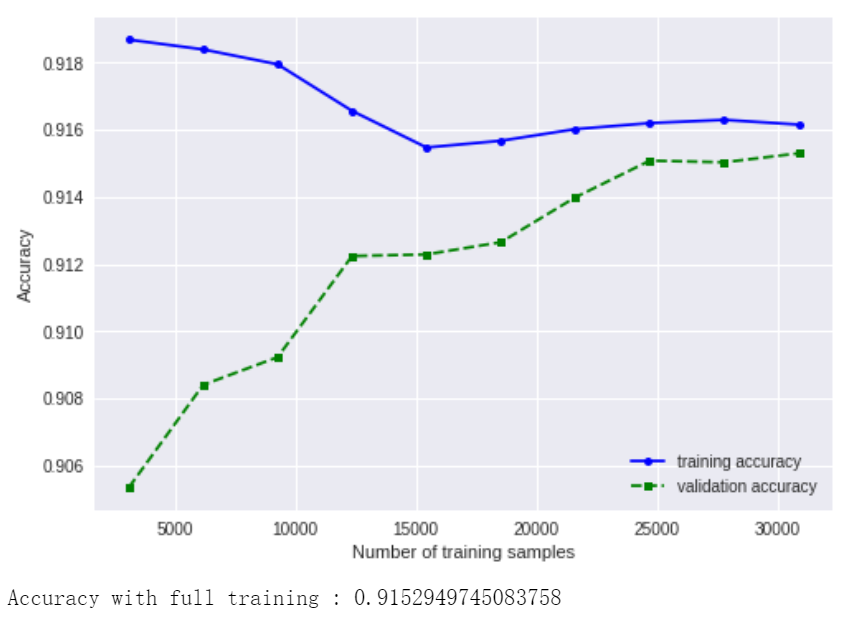
当树的高度小于5的时候，树的高度的提高可以减少泛化误差和经验误差。但是当树的高度大于5的时候，虽然树高度的提高可以减少经验误差，但是泛化误差缺开始提高，说明决策树树高度过高容易造成过拟合。



图表 24 最高树深度对决策树性能的影响

## 训练样本数对性能的影响

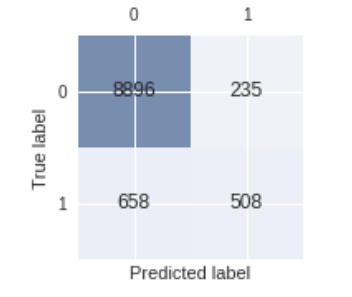
随着训练样本增加，树的泛化误差减低，经验误差有所提高。



图表 25训练样本数对决策树性能的影响

## 混淆矩阵

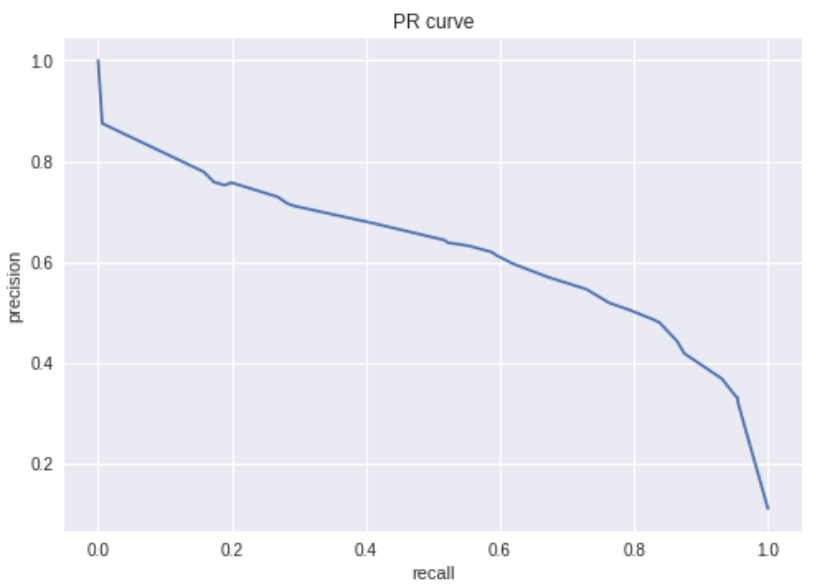
和LR一样，决策树分类中未订购客户分类的性能依旧表现不好。说明数据集不均匀对训练的影响是巨大的。



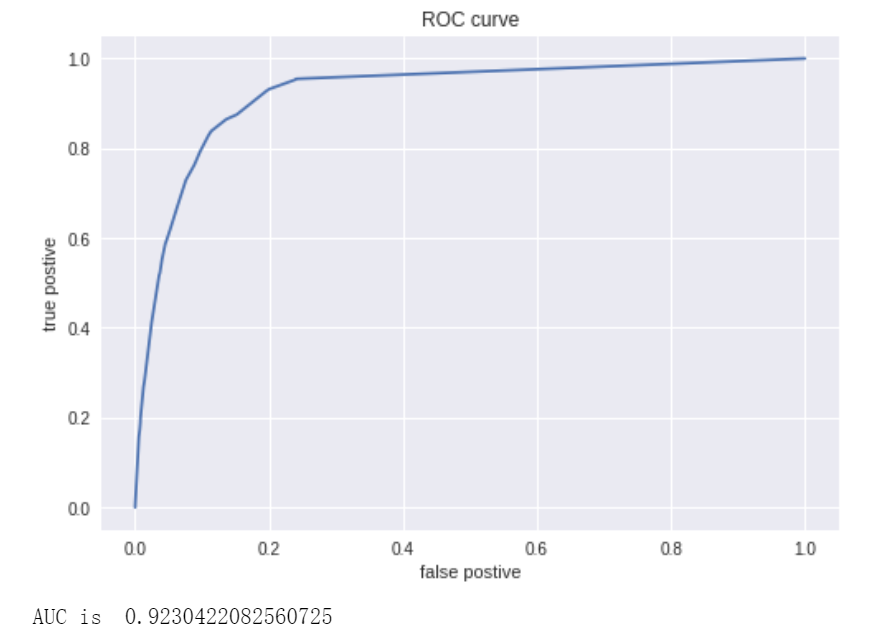
图表 26 混淆矩阵

## PR曲线、ROC曲线和AUC值

决策树的AUC和LR基本一致。



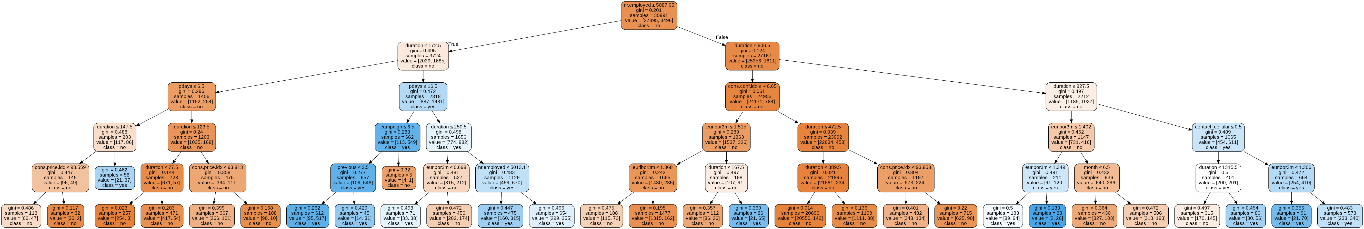
图表 27 决策树PR曲线



图表 28 决策树ROC曲线和AUC值

## 决策树可视化

从决策树可视化中可以看到，对决策影响最高的属性集中于duration、社会经济属性等。

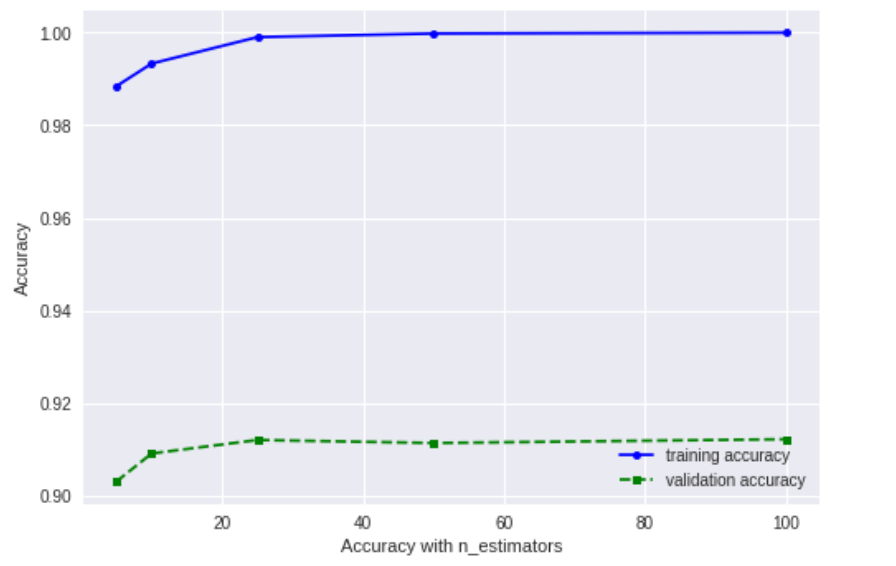


图表 29决策树可视化

# 随机森林实验

## 学习个体与随机森林性能的收敛

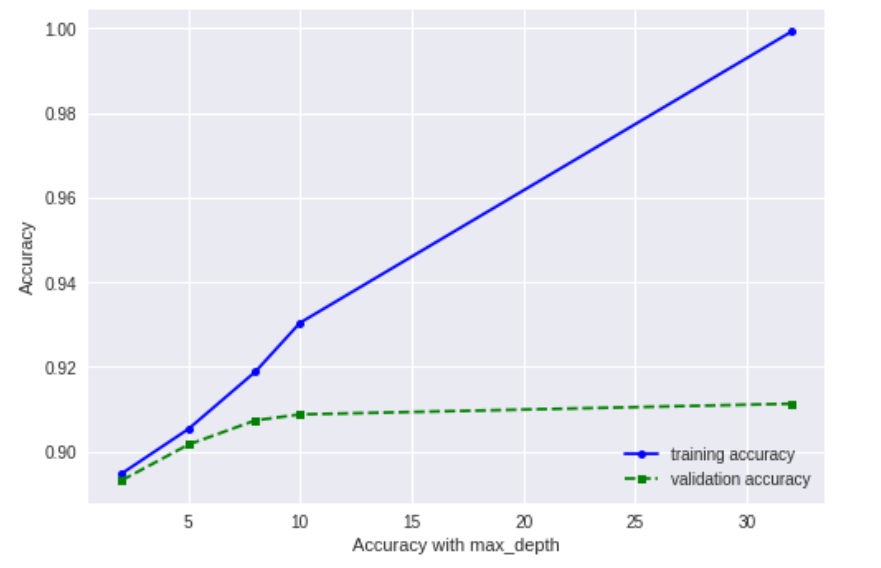
个体数达到30的时候模型性能基本实现了收敛。



图表 30学习个体与随机森林性能的收敛

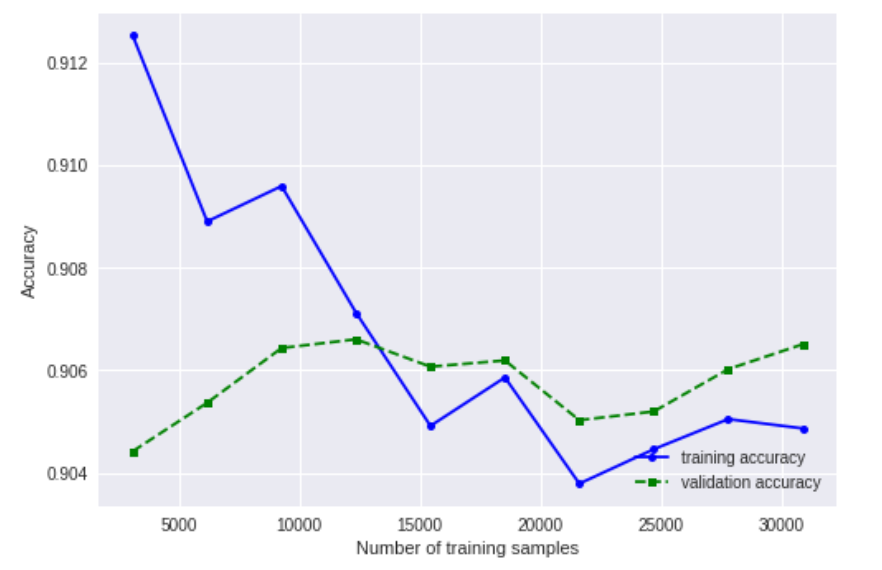
## 最大树高和模型性能关系

和决策树不同，随机森林中树高度的提高并没有造成过拟合问题，相反，树的高度越高，模型性能越好。



图表 31 最大树高和模型性能关系

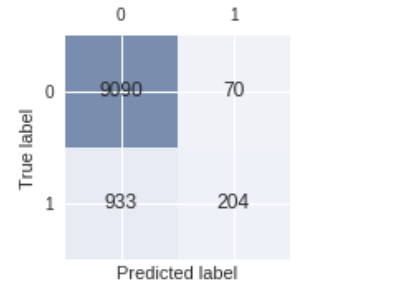
## 样本数对随机森林性能的影响



图表 32 样本数对随机森林性能的影响

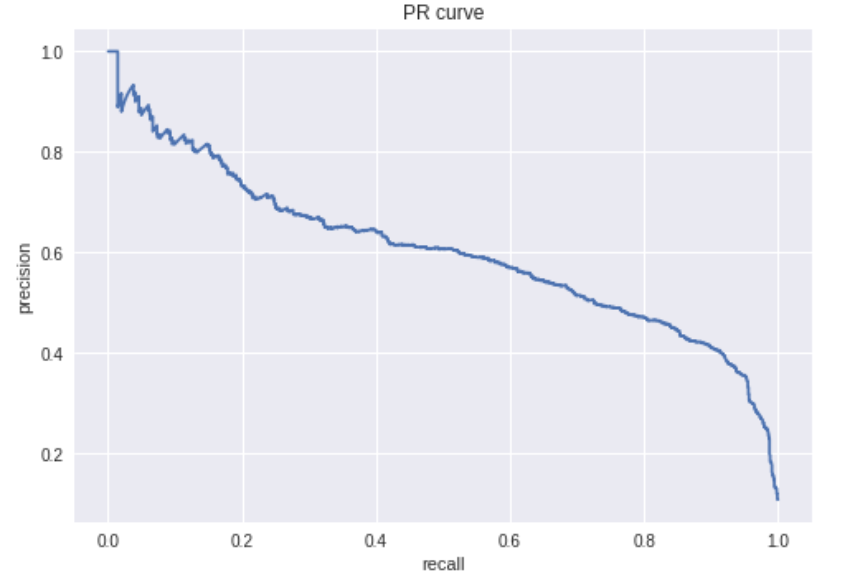
## 混淆矩阵

在混淆矩阵表现了，随机森林中更多的实例被分为了不会购买产品的客户。虽然随机森林和决策树的模型准确率基本一致，但是随机森林模型会使得银行流失更多的客户。

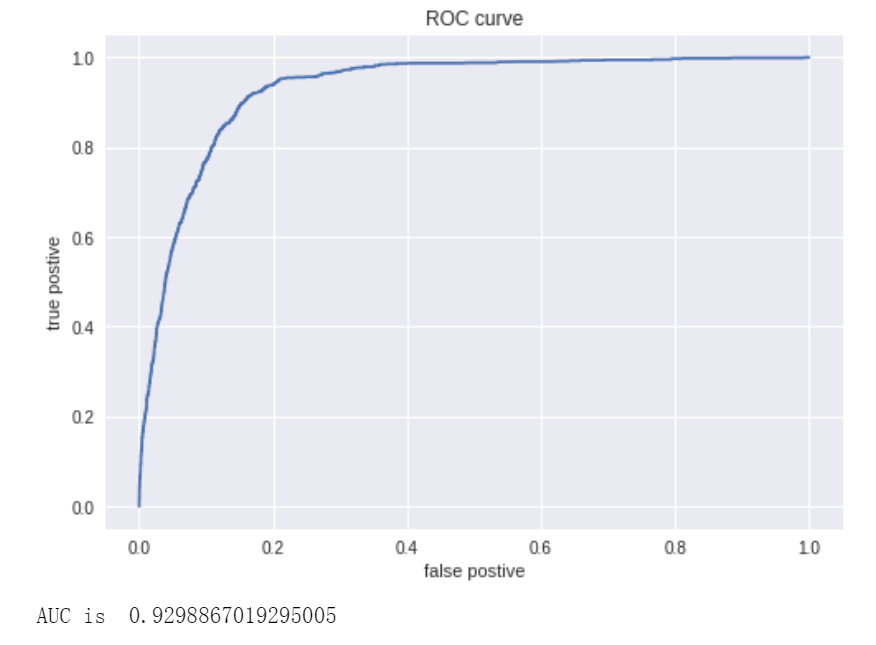


图表 33 随机森林混淆矩阵

## PR曲线、ROC曲线和AUC



图表 34 随机森林PR曲线



图表 35随机森林ROC和AUC值

## 和决策树的对比

和决策树相比，随机森林在表现上：

* 性能随着max\_depth的提升而提高，没有产生过拟合问题
* 混淆矩阵表现上多类别实例表现得更好，但是少类别实例表现更差，这可能会使得银行流失更多潜在购买客户

原因可能如下：

* 随机森林产生多个“差的学习器”，因而max\_depth的提高可以使得这些“差的学习器”性能表现更好，从而从整体上提高随机森林的性能。
* 由于随机森林选择随机选择属性构建学习个体，这导致随机森林的大部分个体忽略了决定性的属性，每个学习个体无法从过多无用属性中学习出购买产品客户的特征，因此这些学习个体会偏向预测客户未占比较高的未购买产品的客户。每个学习个体的偏向使得随机森林整体偏向将客户分为未购买产品的客户。