背景介绍：决策树是一种常见的机器学习方法，其常用于分类问题中，但若将连续变量进行离散化处理后（常用的为二分法），其也可以用于回归问题，此时称决策树为回归树。在将连续变量离散化后，假设所有的输入特征都有有限的离散域，并且有一个称为“分类”的单一目标特征。分类域的每个元素称为一个类。决策树其中每个内部(非叶节点)节点都带有一个输入特性。来自标记了输入特征的节点被标记为目标或输出特征的每个可能值，或者弧导致不同输入特征上的从属决策节点。每片叶子的树是贴上一个类或一个概率分布类,表示数据集已经被树分类到特定的类,或一个特定的概率分布(如果决策树构建良好,是偏向特定子集的类)。

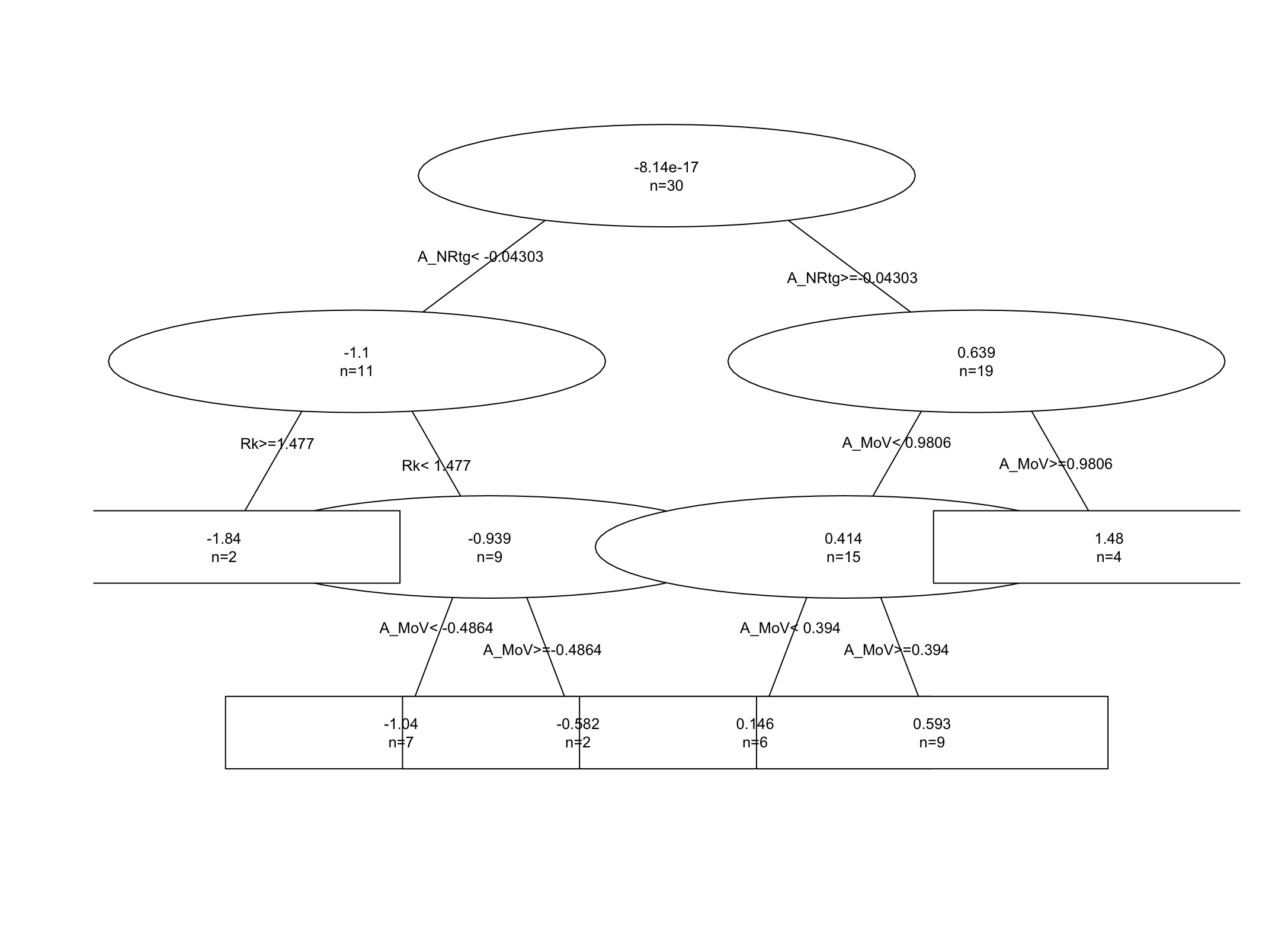
1. 将标准化好的数据投入模型f(x)中,树模型将按照信息增益最大自动生成结点。信息增益的计算方法如图：

信息增益gain越高说明数据中蕴含的信息越多，说明该自变量对因变量的贡献率越高。

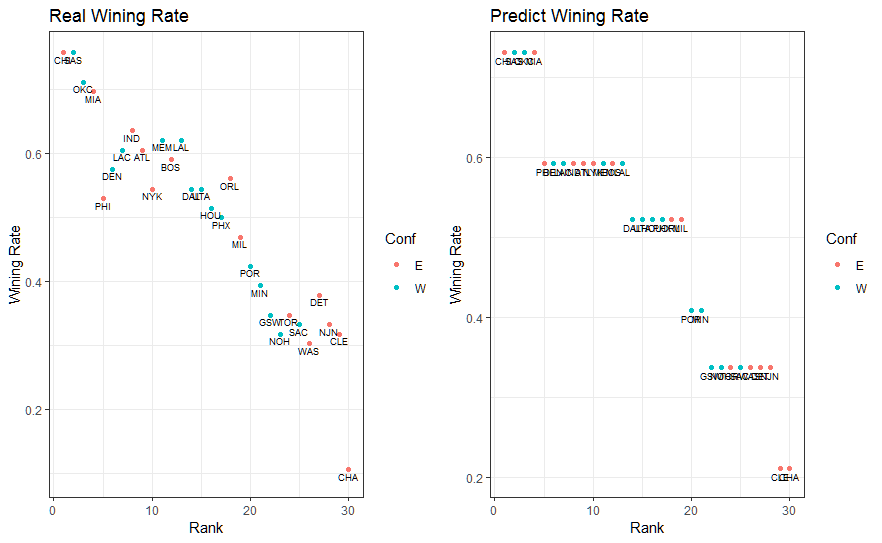
1. 如图所示：

|  |  |
| --- | --- |
| 1) root 30 29.00000000 -8.141636e-17 |  |
| 2) A\_NRtg< -0.04303368 11 2.75381800 -1.103319e+00 | |
| 4) Rk>=1.476701 2 0.92184250 -1.843950e+00 \* | |
| 5) Rk< 1.476701 9 0.49111330 -9.387344e-01 | |
| 10) A\_MoV< -0.4863959 7 0.14577990 -1.040602e+00 \* | |
| 11) A\_MoV>=-0.4863959 2 0.01845982 -5.821989e-01 \* | |
| 3) A\_NRtg>=-0.04303368 19 5.10337900 6.387637e-01 | |
| 6) A\_MoV< 0.9805847 15 1.38078400 4.139654e-01 | |
| 12) A\_MoV< 0.3940031 6 0.23847360 1.458166e-01 \* | |
| 13) A\_MoV>=0.3940031 9 0.42327230 5.927313e-01 \* | |
| 7) A\_MoV>=0.9805847 4 0.12202970 1.481757e+00 \* | |

该树的根结点处共有30个观测值，第一个节点选择了信息增益最大的Adjust\_NRtg为第一个划分属性，由于Adjust\_NRtg是连续型变量，模型计算得到最优分割点为-0.043；因此将数据集划分为大于-0.043和小于0.043的两部分。在第二步分钟共含有11个观测，以第二部分小于0.043的数据为例，在第二部分数据中继续划分数据，得到第二个节点，选择信息增益最大的Rank，将数据划分为Rank>1.48和小于1.48的两部分，第一部分得到了一个叶子节点。将第二部分继续划分下去。最终得到6个子节点。树模型如图：



1. 根据该树模型对12-13赛季的数据进行预测，返回我们的预测结果如下图：



|  |  |  |  |
| --- | --- | --- | --- |
| Team | Real | Team | Predict |
| CHA | 0.106 | CLE | 0.2119922 |
| WAS | 0.303 | CHA | 0.2119922 |
| NOH | 0.318 | WAS | 0.3373664 |
| CLE | 0.318 | TOR | 0.3373664 |
| SAC | 0.333 | SAC | 0.3373664 |
| NJN | 0.333 | NOH | 0.3373664 |
| TOR | 0.348 | NJN | 0.3373664 |
| GSW | 0.348 | GSW | 0.3373664 |
| DET | 0.379 | DET | 0.3373664 |
| MIN | 0.394 | POR | 0.4090311 |
| POR | 0.424 | MIN | 0.4090311 |
| MIL | 0.47 | UTA | 0.5226953 |
| PHX | 0.5 | PHX | 0.5226953 |
| HOU | 0.515 | ORL | 0.5226953 |
| PHI | 0.53 | MIL | 0.5226953 |
| UTA | 0.545 | HOU | 0.5226953 |
| NYK | 0.545 | DAL | 0.5226953 |
| DAL | 0.545 | PHI | 0.5924864 |
| ORL | 0.561 | NYK | 0.5924864 |
| DEN | 0.576 | MEM | 0.5924864 |
| BOS | 0.591 | LAL | 0.5924864 |
| LAC | 0.606 | LAC | 0.5924864 |
| ATL | 0.606 | IND | 0.5924864 |
| MEM | 0.621 | DEN | 0.5924864 |
| LAL | 0.621 | BOS | 0.5924864 |
| IND | 0.636 | ATL | 0.5924864 |
| MIA | 0.697 | SAS | 0.7312879 |
| OKC | 0.712 | OKC | 0.7312879 |
| SAS | 0.758 | MIA | 0.7312879 |
| CHI | 0.758 | CHI | 0.7312879 |

结论：从图中可以看出，回归树模型返回的预测值有6个，代表着6个叶子节点，每个叶子节点中都包含一些观测值，其预测结果和真实值最为相近。