Baseball4_TY

Thomas Young

2025-04-06

Load Hitters DATA

Hitters Data Set from Kaggle: https://www.kaggle.com/datasets/floser/hitters

```
hitters <- read_csv("Hitters.csv")
```

```
## Rows: 322 Columns: 20
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): League, Division, NewLeague
## dbl (17): AtBat, Hits, HmRun, Runs, RBI, Walks, Years, CAtBat, CHits, CHmRun...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

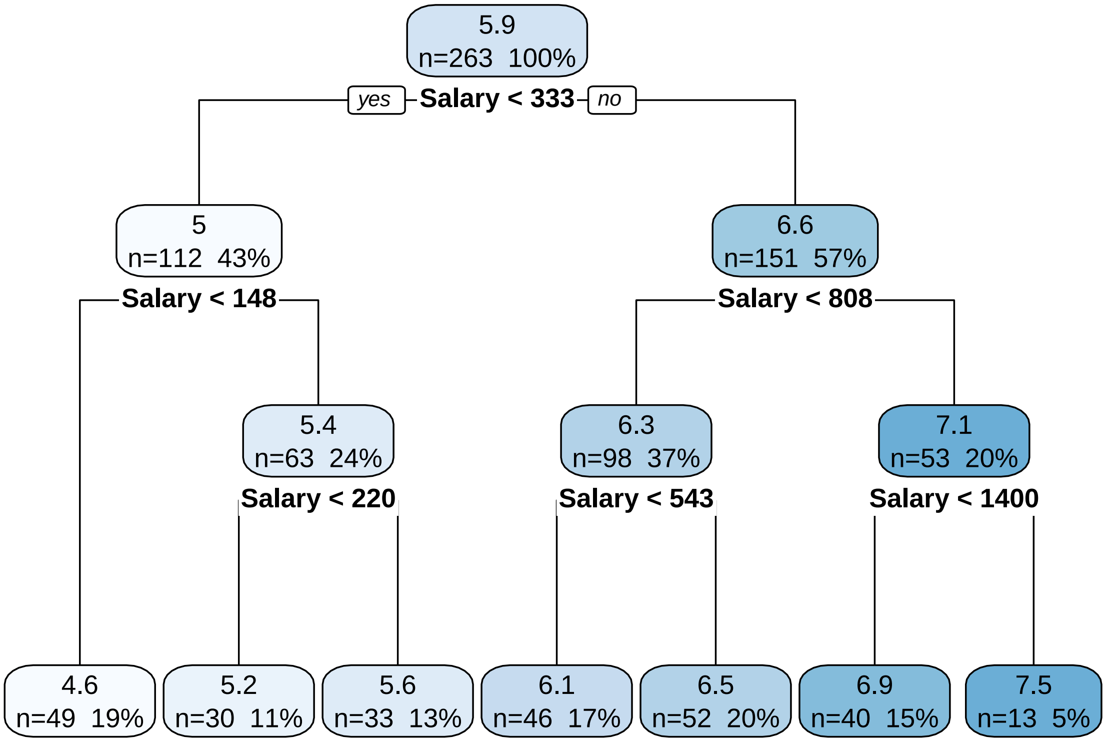
```
# clean
hitters <- hitters |>  drop_na()

# Create variable for log salary
hitters <- hitters |>  mutate(LogSalary = log(Salary))
```

Build with full data set

```
set.seed(123)
tree_model <- rpart(LogSalary ~ ., data = hitters,
                    control = rpart.control(cp = 0.01))

# Visualize the full tree
rpart.plot(tree_model, extra = 101) # extra=101 for node labels
```



```r
summary(tree_model)
```

```
## Call:
## rpart(formula = LogSalary ~ ., data = hitters, control = rpart.control(cp = 0.01))
##   n= 263
##
##           CP nsplit  rel error     xerror        xstd
## 1 0.72839437      0 1.00000000 1.00721703 0.065674949
## 2 0.09555868      1 0.27160563 0.28011572 0.019916678
## 3 0.09143698      2 0.17604695 0.20457999 0.017827296
## 4 0.02461812      3 0.08460997 0.09591707 0.008216027
## 5 0.01777830      4 0.05999185 0.06134570 0.004261465
## 6 0.01199666      5 0.04221354 0.05238338 0.003847300
## 7 0.01000000      6 0.03021689 0.04299560 0.003561517
##
## Variable importance
## Salary CAtBat  CHits  CRuns   CRBI CWalks    RBI  HmRun
##     24     15     15     15     14     14      1      1
##
## Node number 1: 263 observations,    complexity param=0.7283944
##   mean=5.927222, MSE=0.7876568
##   left son=2 (112 obs) right son=3 (151 obs)
##   Primary splits:
##       Salary < 333.3335 to the left,  improve=0.7283944, (0 missing)
##       CAtBat < 1452     to the left,  improve=0.5689379, (0 missing)
##       CHits  < 358      to the left,  improve=0.5640348, (0 missing)
##       CRuns  < 204.5    to the left,  improve=0.5568240, (0 missing)
##       CWalks < 115.5    to the left,  improve=0.5339963, (0 missing)
##   Surrogate splits:
##       CAtBat < 1452     to the left,  agree=0.897, adj=0.759, (0 split)
##       CHits  < 331      to the left,  agree=0.890, adj=0.741, (0 split)
##       CRuns  < 204.5    to the left,  agree=0.882, adj=0.723, (0 split)
##       CWalks < 120      to the left,  agree=0.871, adj=0.696, (0 split)
##       CRBI   < 121.5    to the left,  agree=0.863, adj=0.679, (0 split)
##
## Node number 2: 112 observations,    complexity param=0.09143698
##   mean=5.047731, MSE=0.2220426
##   left son=4 (49 obs) right son=5 (63 obs)
##   Primary splits:
##       Salary < 147.5    to the left,  improve=0.7616585, (0 missing)
##       CRBI   < 55.5     to the left,  improve=0.5798324, (0 missing)
##       CRuns  < 81       to the left,  improve=0.5766546, (0 missing)
##       CHits  < 132      to the left,  improve=0.5748266, (0 missing)
##       CAtBat < 624      to the left,  improve=0.5649600, (0 missing)
##   Surrogate splits:
##       CAtBat < 688      to the left,  agree=0.884, adj=0.735, (0 split)
##       CRuns  < 81       to the left,  agree=0.884, adj=0.735, (0 split)
##       CRBI   < 55.5     to the left,  agree=0.884, adj=0.735, (0 split)
##       CHits  < 171      to the left,  agree=0.875, adj=0.714, (0 split)
##       CWalks < 49.5     to the left,  agree=0.821, adj=0.592, (0 split)
##
## Node number 3: 151 observations,    complexity param=0.09555868
##   mean=6.579559, MSE=0.2079162
##   left son=6 (98 obs) right son=7 (53 obs)
##   Primary splits:
##       Salary < 807.5    to the left,  improve=0.6305182, (0 missing)
##       CRBI   < 300.5    to the left,  improve=0.2105738, (0 missing)
##       CHmRun < 78.5     to the left,  improve=0.1966182, (0 missing)
##       CHits  < 669      to the left,  improve=0.1861248, (0 missing)
##       RBI    < 80.5     to the left,  improve=0.1852984, (0 missing)
##   Surrogate splits:
##       HmRun  < 22.5     to the left,  agree=0.735, adj=0.245, (0 split)
##       RBI    < 80.5     to the left,  agree=0.735, adj=0.245, (0 split)
##       CHmRun < 91       to the left,  agree=0.715, adj=0.189, (0 split)
##       CRuns  < 757      to the left,  agree=0.715, adj=0.189, (0 split)
##       Walks  < 61       to the left,  agree=0.709, adj=0.170, (0 split)
##
## Node number 4: 49 observations
##   mean=4.581425, MSE=0.05372728
##
## Node number 5: 63 observations,    complexity param=0.01199666
##   mean=5.410413, MSE=0.0522956
##   left son=10 (30 obs) right son=11 (33 obs)
##   Primary splits:
##       Salary < 220      to the left,  improve=0.7543056, (0 missing)
##       CRBI   < 106.5    to the left,  improve=0.3471801, (0 missing)
##       Years  < 4.5      to the left,  improve=0.3376300, (0 missing)
##       CWalks < 117.5    to the left,  improve=0.3102906, (0 missing)
##       AtBat  < 405.5    to the right, improve=0.3096749, (0 missing)
##   Surrogate splits:
##       AtBat < 405.5    to the right, agree=0.810, adj=0.600, (0 split)
##       Years < 3.5      to the left,  agree=0.778, adj=0.533, (0 split)
##       Hits  < 92.5     to the right, agree=0.762, adj=0.500, (0 split)
##       Runs  < 39.5     to the right, agree=0.762, adj=0.500, (0 split)
##       RBI   < 40.5     to the right, agree=0.746, adj=0.467, (0 split)
##
## Node number 6: 98 observations,    complexity param=0.02461812
##   mean=6.313292, MSE=0.06799157
##   left son=12 (46 obs) right son=13 (52 obs)
##   Primary splits:
##       Salary < 542.5    to the left,  improve=0.7653614, (0 missing)
##       CHits  < 450      to the left,  improve=0.2434091, (0 missing)
##       CRuns  < 218.5    to the left,  improve=0.1871144, (0 missing)
##       CAtBat < 1772.5   to the left,  improve=0.1733384, (0 missing)
##       Walks  < 21       to the left,  improve=0.1713732, (0 missing)
##   Surrogate splits:
##       Hits  < 103.5    to the left,  agree=0.694, adj=0.348, (0 split)
##       Runs  < 44       to the left,  agree=0.694, adj=0.348, (0 split)
##       CHits < 459.5    to the left,  agree=0.673, adj=0.304, (0 split)
##       RBI   < 47.5     to the left,  agree=0.663, adj=0.283, (0 split)
##       AtBat < 369.5    to the left,  agree=0.653, adj=0.261, (0 split)
##
## Node number 7: 53 observations,    complexity param=0.0177783
##   mean=7.071902, MSE=0.09314787
##   left son=14 (40 obs) right son=15 (13 obs)
##   Primary splits:
##       Salary   < 1400     to the left,  improve=0.7459921, (0 missing)
##       Runs     < 93.5     to the left,  improve=0.1820475, (0 missing)
##       RBI      < 70.5     to the left,  improve=0.1682246, (0 missing)
##       CRBI     < 988      to the left,  improve=0.1555606, (0 missing)
##       Division splits as  RL,           improve=0.1469899, (0 missing)
##   Surrogate splits:
##       CHits < 2053.5   to the left,  agree=0.830, adj=0.308, (0 split)
##       Hits  < 199      to the left,  agree=0.811, adj=0.231, (0 split)
##       Walks < 74.5     to the left,  agree=0.811, adj=0.231, (0 split)
##       CRBI  < 988      to the left,  agree=0.811, adj=0.231, (0 split)
##       HmRun < 0.5      to the right, agree=0.792, adj=0.154, (0 split)
##
## Node number 10: 30 observations
##   mean=5.202107, MSE=0.01320714
##
## Node number 11: 33 observations
##   mean=5.599782, MSE=0.01252291
##
## Node number 12: 46 observations
##   mean=6.070752, MSE=0.01947143
##
## Node number 13: 52 observations
##   mean=6.527846, MSE=0.01284139
##
## Node number 14: 40 observations
##   mean=6.921624, MSE=0.02349865
##
## Node number 15: 13 observations
##   mean=7.534296, MSE=0.02415765
```

```r
# Performance
r_squared <- 1 - (sum((hitters$LogSalary - predict(tree_model, hitters))^2) /
                   sum((hitters$LogSalary - mean(hitters$LogSalary))^2))
print(r_squared)
```

```
## [1] 0.9697831
```

Interpretation

- The tree begins with 263 observations (players) and the root node (Node 1) has a predicted average LogSalary of 5.927.
- The primary split is based on Salary < 333.3335.
  - If a player's salary is less than 333.3335, they go to Node 2.
  - If their salary is greater than or equal to 333.3335, they go to Node 3.

Branch 1: Lower Salaries (Node 2, Node 4, Node 5, Node 10, Node 11)

Node 2: This node contains 112 players with lower salaries (predicted LogSalary = 5.048). - Split: The next split is based on Salary < 147.5. - If salary is less than 147.5, go to Node 4. - If salary is greater than or equal to 147.5, go to Node 5.

Node 4: This is a terminal node with 49 players and a predicted LogSalary of 4.581. These are likely players with very low salaries.

Node 5: This node contains 63 players with slightly higher salaries (predicted LogSalary = 5.410). - Split: Further split based on Salary < 220. - If salary is less than 220, go to Node 10. - If salary is greater than or equal to 220, go to Node 11.

Node 10: Terminal node with 30 players and predicted LogSalary = 5.202.

Node 11: Terminal node with 33 players and predicted LogSalary = 5.599.

Branch 2: Higher Salaries (Node 3, Node 6, Node 7, Node 12, Node 13, Node 14, Node 15)

Node 3: This node contains 151 players with higher salaries (predicted LogSalary = 6.579). - Split: Based on Salary < 807.5. - If salary is less than 807.5, go to Node 6. - If salary is greater than or equal to 807.5, go to Node 7.

Node 6: 98 players (predicted LogSalary = 6.313). - Split: Based on Salary < 542.5. - If salary is less than 542.5, go to Node 12. - If salary is greater than or equal to 542.5, go to Node 13.

Node 12: Terminal node with 46 players and predicted LogSalary = 6.071.

Node 13: Terminal node with 52 players and predicted LogSalary = 6.528.

Node 7: 53 players (predicted LogSalary = 7.072). - Split: Based on Salary < 1400. - If salary is less than 1400, go to Node 14. - If salary is greater than or equal to 1400, go to Node 15.

Node 14: Terminal node with 40 players and predicted LogSalary = 6.922.

Node 15: Terminal node with 13 players and predicted LogSalary = 7.534. These are likely players with the highest salaries.

The tree primarily uses the Salary variable to make splits, suggesting it's a strong predictor of LogSalary. As you move down the branches, the predicted LogSalary values generally increase, reflecting the salary-based splits

Overall performance of this model was 97%