NETFLIX

Gifford Tompkins
Project 04 | DSI 09
21 October 2019

Problem Statement

As netflix expands and seeks to create more content, they are looking to speed up their evaluation process for incoming show ideas, scripts and pilots.

We will attempt to build a model that will predict if a NetFlix original show will be renewed for another season.

Data Collection

We scraped our data from the *"List of original programs distributed by Netflix"* page on Wikipedia.

Title	Genre	Premiere	Seasons	Length
House of Cards	Political drama	February 1, 2013	6 seasons, 73 episodes	42–59 min.
Hemlock Grove	Horror/thriller	April 19, 2013	3 seasons, 33 episodes	45–58 min.
Orange Is the New Black	Comedy-drama	July 11, 2013	7 seasons, 91 episodes	50–92 min.
Marco Polo	Historical drama	December 12, 2014	2 seasons, 20 episodes	48–65 min.
Bloodline	Thriller	March 20, 2015	3 seasons, 33 episodes	48–68 min.

Data Cleaning:

Data Cleaning consisted of:

- Converting ‘seasons’, ‘length’ and ‘premiere’ into appropriate data columns.
- Expanding ‘genre’ into dummy columns that included each sub-genre.
- Binarizing the ‘status’ column into ‘Renewed’, or ‘Ended’
- Miscellaneous standardization of string texts.

When we finished the munging process, or data set was whittled down from 1313 to 285.

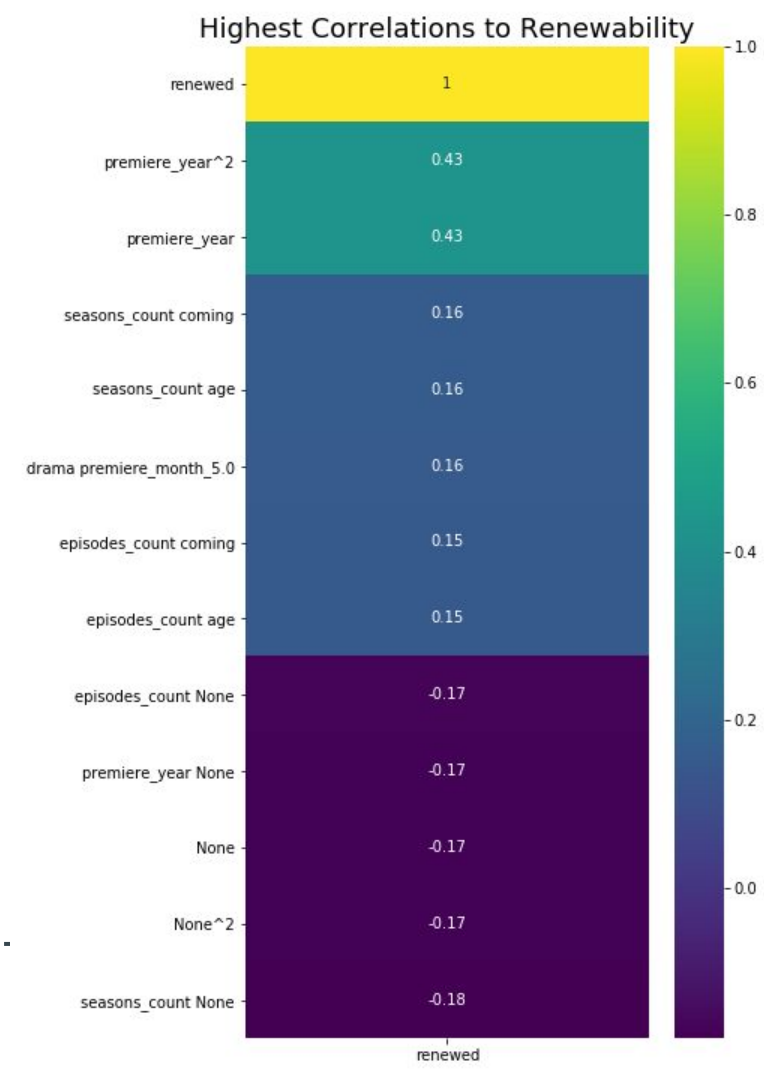
Title	Genre	Premiere	Seasons	Length
House of Cards	Political drama	February 1, 2013	6 seasons, 73 episodes	42–59 min.
Hemlock Grove	Horror/thriller	April 19, 2013	3 seasons, 33 episodes	45–58 min.
Orange Is the New Black	Comedy-drama	July 11, 2013	7 seasons, 91 episodes	50–92 min.
Marco Polo	Historical drama	December 12, 2014	2 seasons, 20 episodes	48–65 min.
Bloodline	Thriller	March 20, 2015	3 seasons, 33 episodes	48–68 min.

64.6 %

Baseline Score

After all of our data cleaning process, this is our goal to beat.

Correlative Columns

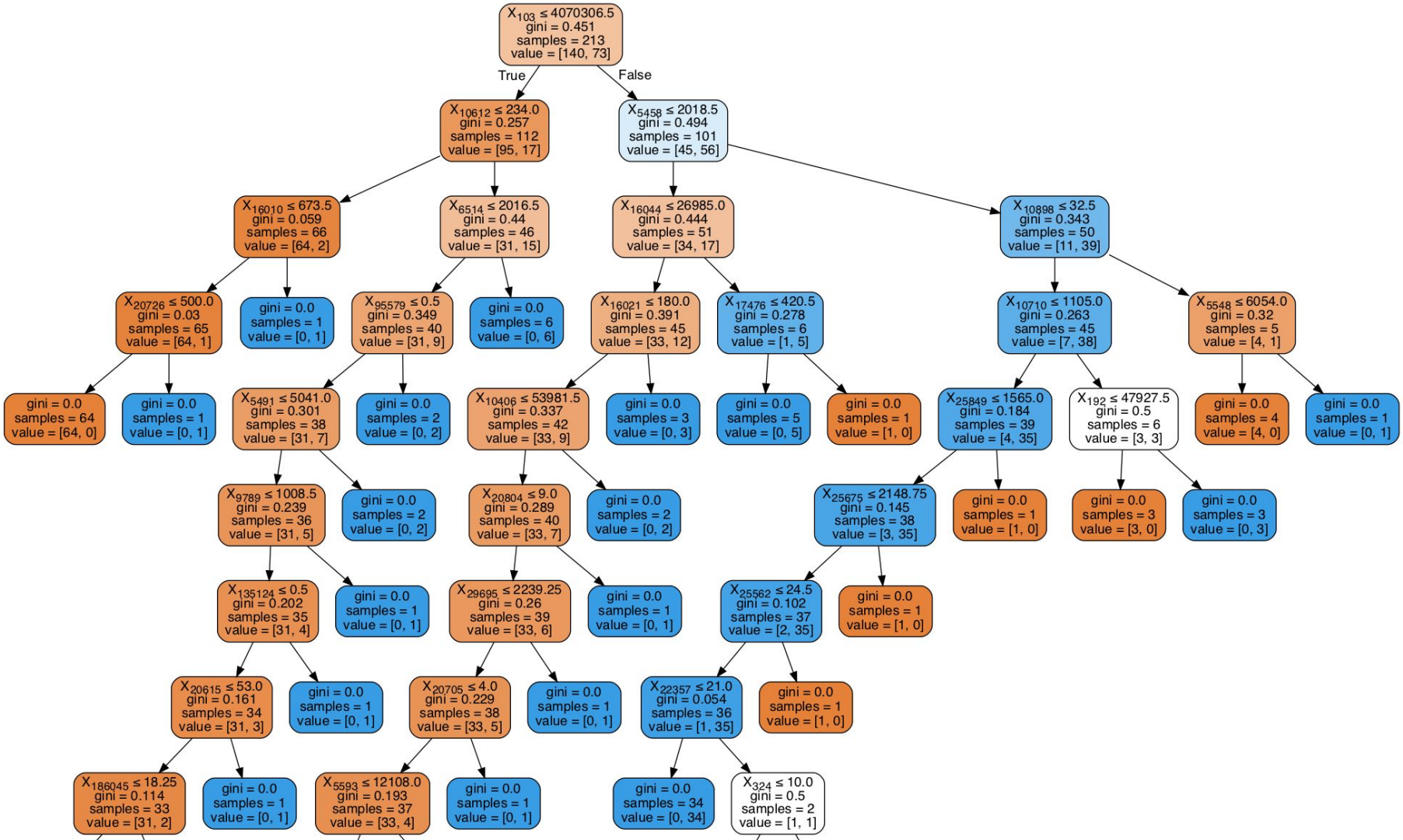


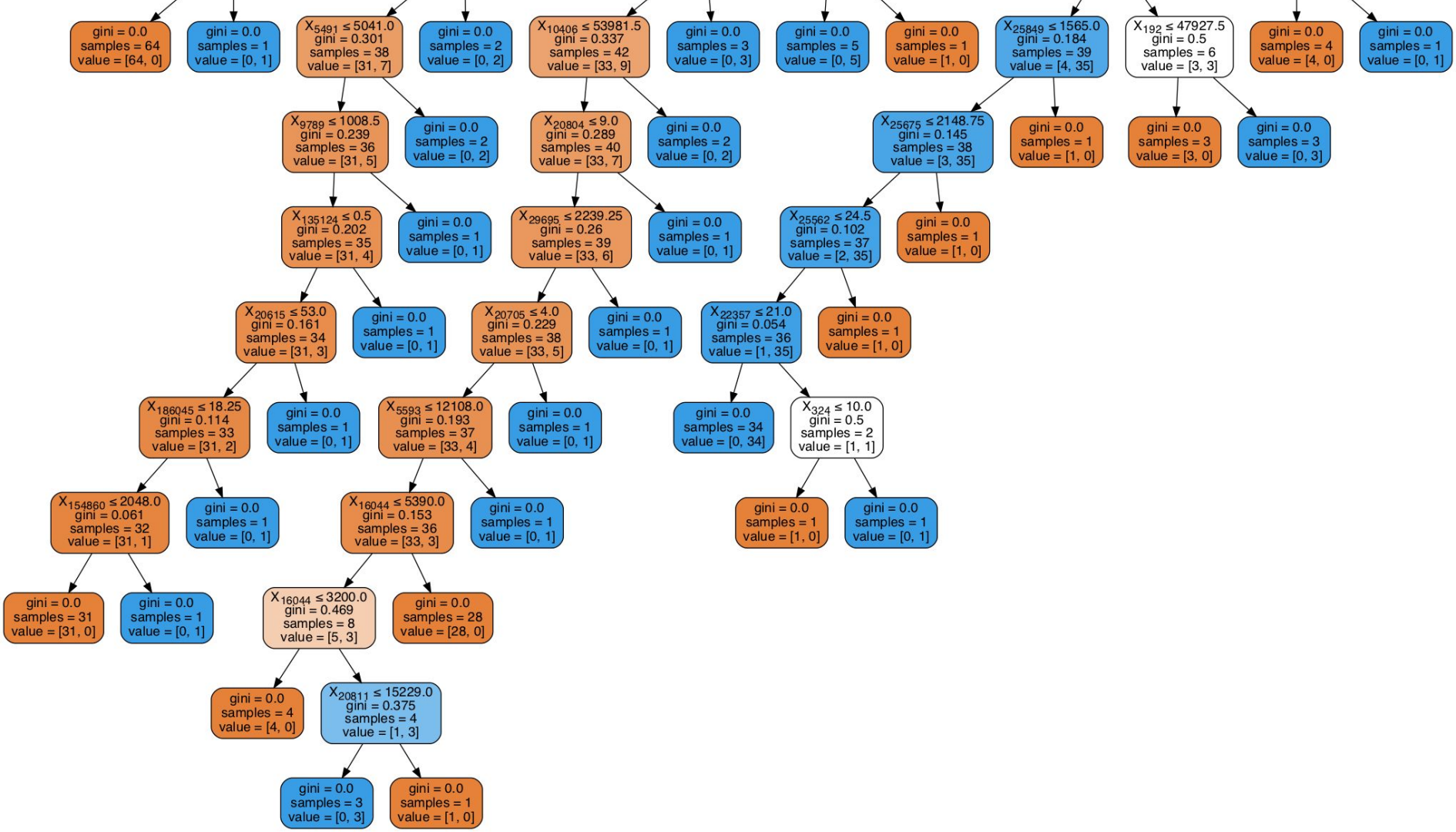
Decision Tree Classifier

Parameters:

{'max_depth': 3,
'min_samples_leaf': 9,
'min_samples_split': 2,
'random_state': 42}

Train score: 80%
Test Score: 77.8%
Improvement: 20.43% *from baseline*


X103 ≤ 4070306.5
gini = 0.451
samples = 213
value = [140, 73]
True
False
X10612 ≤ 234.0
gini = 0.257
samples = 112
value = [95, 17]
X5458 ≤ 2018.5
gini = 0.494
samples = 101
value = [45, 56]
X16010 ≤ 673.5
gini = 0.059
samples = 66
value = [64, 2]
X6514 ≤ 2016.5
gini = 0.44
samples = 46
value = [31, 15]
X16044 ≤ 26985.0
gini = 0.444
samples = 51
value = [34, 17]
X10898 ≤ 32.5
gini = 0.343
samples = 50
value = [11, 39]
X20726 ≤ 500.0
gini = 0.03
samples = 65
value = [64, 1]
gini = 0.0
samples = 1
value = [0, 1]
X95579 ≤ 0.5
gini = 0.349
samples = 40
value = [31, 9]
gini = 0.0
samples = 6
value = [0, 6]
X16021 ≤ 180.0
gini = 0.391
samples = 45
value = [33, 12]
X17476 ≤ 420.5
gini = 0.278
samples = 6
value = [1, 5]
X10710 ≤ 1105.0
gini = 0.263
samples = 45
value = [7, 38]
X5548 ≤ 6054.0
gini = 0.32
samples = 5
value = [4, 1]
gini = 0.0
samples = 64
value = [64, 0]
gini = 0.0
samples = 1
value = [0, 1]
X5491 ≤ 5041.0
gini = 0.301
samples = 38
value = [31, 7]
gini = 0.0
samples = 2
value = [0, 2]
X10406 ≤ 53981.5
gini = 0.337
samples = 42
value = [33, 9]
gini = 0.0
samples = 3
value = [0, 3]
gini = 0.0
samples = 5
value = [0, 5]
gini = 0.0
samples = 1
value = [1, 0]
X25849 ≤ 1565.0
gini = 0.184
samples = 39
value = [4, 35]
X192 ≤ 47927.5
gini = 0.5
samples = 6
value = [3, 3]
gini = 0.0
samples = 4
value = [4, 0]
gini = 0.0
samples = 1
value = [0, 1]
X9789 ≤ 1008.5
gini = 0.239
samples = 36
value = [31, 5]
gini = 0.0
samples = 2
value = [0, 2]
X20804 ≤ 9.0
gini = 0.289
samples = 40
value = [33, 7]
gini = 0.0
samples = 2
value = [0, 2]
X25675 ≤ 2148.75
gini = 0.145
samples = 38
value = [3, 35]
gini = 0.0
samples = 1
value = [1, 0]
gini = 0.0
samples = 3
value = [3, 0]
gini = 0.0
samples = 3
value = [0, 3]
X135124 ≤ 0.5
gini = 0.202
samples = 35
value = [31, 4]
gini = 0.0
samples = 1
value = [0, 1]
X29695 ≤ 2239.25
gini = 0.26
samples = 39
value = [33, 6]
gini = 0.0
samples = 1
value = [0, 1]
X25562 ≤ 24.5
gini = 0.102
samples = 37
value = [2, 35]
gini = 0.0
samples = 1
value = [1, 0]
X20615 ≤ 53.0
gini = 0.161
samples = 34
value = [31, 3]
gini = 0.0
samples = 1
value = [0, 1]
X20705 ≤ 4.0
gini = 0.229
samples = 38
value = [33, 5]
gini = 0.0
samples = 1
value = [0, 1]
X22357 ≤ 21.0
gini = 0.054
samples = 36
value = [1, 35]
gini = 0.0
samples = 1
value = [1, 0]
X186045 ≤ 18.25
gini = 0.114
samples = 33
value = [31, 2]
gini = 0.0
samples = 1
value = [0, 1]
X5593 ≤ 12108.0
gini = 0.193
samples = 37
value = [33, 4]
gini = 0.0
samples = 1
value = [0, 1]
gini = 0.0
samples = 34
value = [0, 34]
X324 ≤ 10.0
gini = 0.5
samples = 2
value = [1, 1]


gini = 0.0
samples = 64
value = [64, 0]
gini = 0.0
samples = 1
value = [0, 1]
X5491 ≤ 5041.0
gini = 0.301
samples = 38
value = [31, 7]
gini = 0.0
samples = 2
value = [0, 2]
X10406 ≤ 53981.5
gini = 0.337
samples = 42
value = [33, 9]
gini = 0.0
samples = 3
value = [0, 3]
gini = 0.0
samples = 5
value = [0, 5]
gini = 0.0
samples = 1
value = [1, 0]
X25849 ≤ 1565.0
gini = 0.184
samples = 39
value = [4, 35]
X192 ≤ 47927.5
gini = 0.5
samples = 6
value = [3, 3]
gini = 0.0
samples = 4
value = [4, 0]
gini = 0.0
samples = 1
value = [0, 1]
X9789 ≤ 1008.5
gini = 0.239
samples = 36
value = [31, 5]
gini = 0.0
samples = 2
value = [0, 2]
X20804 ≤ 9.0
gini = 0.289
samples = 40
value = [33, 7]
gini = 0.0
samples = 2
value = [0, 2]
X25675 ≤ 2148.75
gini = 0.145
samples = 38
value = [3, 35]
gini = 0.0
samples = 1
value = [1, 0]
gini = 0.0
samples = 3
value = [3, 0]
gini = 0.0
samples = 3
value = [0, 3]
X135124 ≤ 0.5
gini = 0.202
samples = 35
value = [31, 4]
gini = 0.0
samples = 1
value = [0, 1]
X29695 ≤ 2239.25
gini = 0.26
samples = 39
value = [33, 6]
gini = 0.0
samples = 1
value = [0, 1]
X25562 ≤ 24.5
gini = 0.102
samples = 37
value = [2, 35]
gini = 0.0
samples = 1
value = [1, 0]
X20615 ≤ 53.0
gini = 0.161
samples = 34
value = [31, 3]
gini = 0.0
samples = 1
value = [0, 1]
X20705 ≤ 4.0
gini = 0.229
samples = 38
value = [33, 5]
gini = 0.0
samples = 1
value = [0, 1]
X22357 ≤ 21.0
gini = 0.054
samples = 36
value = [1, 35]
gini = 0.0
samples = 1
value = [1, 0]
X186045 ≤ 18.25
gini = 0.114
samples = 33
value = [31, 2]
gini = 0.0
samples = 1
value = [0, 1]
X5593 ≤ 12108.0
gini = 0.193
samples = 37
value = [33, 4]
gini = 0.0
samples = 1
value = [0, 1]
gini = 0.0
samples = 34
value = [0, 34]
X324 ≤ 10.0
gini = 0.5
samples = 2
value = [1, 1]
X154860 ≤ 2048.0
gini = 0.061
samples = 32
value = [31, 1]
gini = 0.0
samples = 1
value = [0, 1]
X16044 ≤ 5390.0
gini = 0.153
samples = 36
value = [33, 3]
gini = 0.0
samples = 1
value = [0, 1]
gini = 0.0
samples = 1
value = [1, 0]
gini = 0.0
samples = 1
value = [0, 1]
gini = 0.0
samples = 31
value = [31, 0]
gini = 0.0
samples = 1
value = [0, 1]
X16044 ≤ 3200.0
gini = 0.469
samples = 8
value = [5, 3]
gini = 0.0
samples = 28
value = [28, 0]
gini = 0.0
samples = 4
value = [4, 0]
X20811 ≤ 15229.0
gini = 0.375
samples = 4
value = [1, 3]
gini = 0.0
samples = 3
value = [0, 3]
gini = 0.0
samples = 1
value = [1, 0]

Summary And Further Steps

There seems to be promise in being able to predict the renewability of a show in this way.

Recommendations

1. More data.
2. Play with Polynomial Features more.
3. Try a model that is a little more interpretable.