Which Show Should Go On?: Predictive Analysis and Not-for-Profit Arts Organizations in New York City

CULTURAL DATA PROJECT

Overview:

- Question(s) to be Investigated
- Underlying Economic Theory
- Introduction to the Data and Tools Utilized
- Data Preparation
- The Analysis
- Why is this Analysis relevant?
- Where to go from here?

Questions to be Investigated:

- Should Not-For-Profit Arts Organizations be subsidized in the first place?
- If so, which ones should be subsidized? Why?
- Under what circumstances should a new or existing organization receive public subsidy?

Underlying Economic Theory

- A change in economics occurred in the early 2000's
- Essentially, a man name Richard Florida proposed an alteration to Human Capital Theory: What if People are the "motor force behind regional [and economic] growth" rather than abundant amenities of geography.
- This theory was called the Creative Capital Theory.

“Cities and The Creative Class”

- High populations of these individuals:
  - Artists, entertainers, poets, novelists, nonfiction writers, editors, cultural figures, architects, actors, and designers.
- Along with certain atmospheric characteristics (such as the 3T’s: Technology, Tolerance, and Talent)
- Have been statistically correlated with regional employment growth, high-technological growth, and population growth.

The Cultural Data Project:

- The Cultural Data Project (CDP) is a non-for-profit organization which enables arts and cultural organizations to enter financial, programmatic and operational data into a standardized online form

- The type of data collected ranges includes basic organizational information, revenue, expense, marketing activity, balance sheet items, investments, loans and a wide range of non-financial information.

- The database contains over 800 variables, and over 1000 art organizations within the five boroughs for the year of 2008.

The Data:

RStudio

File Edit Code View Plots Session Build Debug Tools Help

Go to file/function | Project: (None)

FinalCDPOut

Filter

supp_gov_county_total	supp_gov_state_total	supp_gov_federal_total	attendance_total_total	web_nmbr_unique_visitors	staff_full_time_employees_total	AllPlusRev	OperatingRev	OperatingEx
0	107719	25000	3300	200000	7	181734	0	197
0	0	0	2500	NA	0	150	50301	16
0	20000	113613	0	NA	10	20144	0	128
0	66500	570400	66179	NA	51	2428413	3769190	896
0	0	0	250	57584	13	-65393	125784	214
0	23100	100000	1065	82850	10	-238373	505264	235
0	0	0	9000	525600	7	284338	497940	85
0	0	0	240	92000	8	37505	0	354
0	66645	35000	159462	NA	27	403756	2038675	770
0	0	0	1000	NA	0	0	18493	9
0	17980	28000	2000	608768	2	24	1310	17
0	5000	0	2500	NA	0	0	7000	1
0	14500	0	351000	NA	0	91	650	3
0	0	0	4000	NA	1	0	374	13
0	9300	10000	1175	NA	0	9949	10529	9
0	0	0	13000	NA	4	0	205449	19
0	20000	70000	30000	NA	3	0	208128	72
0	29000	0	5341	NA	2	6826	105662	56
0	60000	0	472	NA	5	20642	96737	54
0	40000	20000	23056	210000	14	100164	1409334	289
0	13700	20000	5167	4500	0	641	22529	9
0	467503	1648	466631	160352	32	311127	523548	646

Showing 1 to 23 of 1,015 entries

Console

R, SAS, and Tableau:

- R - Statistical Analysis Software, utilized C.5 algorithm in C50 package
- SAS - Statistical Analysis Software, cleaned, organized, and edited the data
- Tableau - Data Visualization + Analysis Software, aided in formatting and determining which art orgs. should be considered failures/successes.

C5.0 Algorithm

- Originated from its predecessors: The ID3 and C4.5 algorithm
- These two algorithms utilize entropy and information gain to determine the importance of predictors
- This is the newest update for this type of classification, and in practice has shown to use less memory, perform at a faster rate, and enhance accuracy

Data Preparation

Cleaning + Organizing the Data

- Imported the data
- Merged multiple sections
- Sorted the data
- Excised null values and replaced them with zeros
- Excised values considered to be outliers

Exporting the data

- Exported the data in CSV format
- Then imported the data into R and Tableau for further analysis

Data Preparation (Visualization)

```
LIBNAME pew 'c:\pew\';

Data a; set pew.section_01;
     zipcode = scan(zip_plus_4,1,'-');
     year = year (fy_end_date);

Data aa; set a;
where year = 2008;

Proc Sort; by org_id fy_end_date;

Data c; set pew.section_03;
year = year (fy_end_date);

Data cc; set c;
where year = 2008;

Proc sort; by org_id fy_end_date;

Data f; set pew.section_06;
year = year (fy_end_date);

data ff; set f;
where year = 2008;

Proc sort; by org_id fy_end_date;

data joe; merge aa cc ff;
by org_id year;

proc sort; by org_id year;
```

The Analysis

The data was split into two sections for classification analysis and divided up by borough:

- Community Outreach
  - Total Attendance, Total Unique Website Views
- Economic Outreach
  - Amount of Jobs Supplied

The Analysis (Continued)

- I then utilized the median values of each category as the initial benchmark, and generated a SubsidizeMore/SubsidizeLess column based off of the Farebox Recovery Ratio

```
data joeboi3; set joeboi2;

if county = 'New York' then
        if  fareboxrecovery < .2320 then SubsidizeorNah = "Subsidize";
        else SubsidizeorNah = "Nah";

if county = 'Kings' then
        if  fareboxrecovery < .2698 then SubsidizeorNah = "Subsidize";
        else SubsidizeorNah = "Nah";

if county = 'Richmond' then
        if  fareboxrecovery < .1642 then SubsidizeorNah = "Subsidize";
        else SubsidizeorNah = "Nah";

if county = 'Queens' then
        if  fareboxrecovery < .1238 then SubsidizeorNah = "Subsidize";
        else SubsidizeorNah = "Nah";

if county = 'Bronx' then
        if  fareboxrecovery < .1878 then SubsidizeorNah = "Subsidize";
        else SubsidizeorNah = "Nah";
```

The Analysis (Continued)

FinalCDPOut × | CDP Scripts.R* ×

Source on Save | Run | Source

```
2  CDPDataset <- read.csv("~/Desktop/CDPDataset.csv")
3  View(CDPDataset)
4
5  #Shuffling the dataset
6  set.seed(9850)
7  g <- runif(nrow(CDPDataset))
8  CDP1 <- CDPDataset[order(g),]
9  levels(CDP1$SubsidizeorNah)[1] = "missing"
10
11 # load the package
12 library(C50)
13
14 #Have to get rid of Operating revenue, along with org_name
15 CDP1$OperatingRev = NULL
16 CDP1$org_name = NULL
17
18 #create the tree using 75% of results
19 #-14 = thing we are trying to predict, first arg. = predictors, sec arg. = target
20 Tree1 <- C5.0(CDP1[1:760, -14], CDP1[1:760, 14])
21
22 #Outputs a very long tree
23 summary(Tree1)
24
25 #Guage the accuracy of your model utilizing test data
26 predictions <- predict(Tree5, CDP1[761:1015, ])
27 table(CDP1[761:1015,14], predictions)
28
29 # boost model (aka enhance performance)  by running the
30 # classification system through a series of models
31 Tree6 <- C5.0(CDP1[1:760, -14], CDP1[1:760, 14], trials = 10)
32
33 #determine the rules that occur at various steps within the tree
34 rulesTree6 <- Tree6 <- C5.0(CDP1[1:760, -14], CDP1[1:760, 14], rules = 10)
35 summary(rulesTree6)
```

19:82 (Top Level) R Script

- The dataset was imported into R and I utilized the C.5 classification algorithm to generate a classification tree, and investigated the precision of the created model.

Examining The Tree:

```
C5.0 [Release 2.07 GPL Edition]         Fri Dec 18 02:20:18 2015
-------------------------------

Class specified by attribute `outcome'

Read 760 cases (14 attributes) from undefined.data

Decision tree:

TypeofInstitute in {Individual Entities,Performing Group,Schools/University/}:
:...attendance_total_total > 7200:
:   :...TotalSupp <= 210391: Nah (27)
:   :   TotalSupp > 210391:
:   :   :...OperatingEx > 695925: Nah (62/7)
:   :       OperatingEx <= 695925:
:   :       :...supp_gov_state_total > 22500: Subsidize (6)
:   :           supp_gov_state_total <= 22500:
:   :           :...TotalPublicSupp <= 25640: Subsidize (4/1)
:   :               TotalPublicSupp > 25640: Nah (5)
:   attendance_total_total <= 7200:
:   :...TotalSupp <= 57846:
:       :...attendance_total_total > 1600: Nah (38/4)
:       :   attendance_total_total <= 1600:
:       :   :...TotalPublicSupp <= 11300: Nah (51/18)
:       :       TotalPublicSupp > 11300: Subsidize (9/1)
:       TotalSupp > 57846:
:       :...OperatingEx <= 75392: Subsidize (14)
:           OperatingEx > 75392:
:           :...TypeofInstitute = Individual Entities: Nah (3)
:               TypeofInstitute = Schools/University/:
:               :...staff_full_time_employees_total <= 4: Subsidize (2)
:               :   staff_full_time_employees_total > 4: Nah (4)
:               TypeofInstitute = Performing Group:
:               :...TotalSupp <= 351470:
```

Summary Data about the Tree:

RStudio

File Edit Code View Plots Session Build Debug Tools Help

Go to file/function

Source

Console ~/

```
Evaluation on training data (760 cases):

            Decision Tree
          ----------------
          Size      Errors

            40  174(22.9%)   <<


           (a)   (b)    <-classified as
          ----  ----
           266   109    (a): class Nah
            65   320    (b): class Subsidize


        Attribute usage:

        100.00% TypeofInstitute
         61.97% attendance_total_total
         42.37% TotalSupp
         33.55% fndgrp
         31.32% staff_full_time_employees_total
         25.92% OperatingEx
         17.37% TotalPublicSupp
         16.45% TotalPrivateSupp
         11.97% AllPlusRev
          7.63% supp_gov_state_total
          5.79% supp_gov_federal_total
          4.34% web_nmbr_unique_visitors


Time: 0.0 secs
```

Evaluating the Rules:

Source

Console ~/

```
Rules:

Rule 1: (49/4, lift 1.8)
        attendance_total_total > 1600
        TotalSupp <= 57846
        TypeofInstitute in {Individual Entities, Performing Group,
                            Schools/University/}
        ->  class Nah  [0.902]

Rule 2: (6, lift 1.8)
        TotalPrivateSupp <= 6490
        TypeofInstitute = None of the Above
        ->  class Nah  [0.875]

Rule 3: (6, lift 1.8)
        staff_full_time_employees_total > 4
        TypeofInstitute = Schools/University/
        ->  class Nah  [0.875]

Rule 4: (5, lift 1.7)
        web_nmbr_unique_visitors <= 175866
        TotalPrivateSupp <= 10100
        TypeofInstitute = Service Orgs.
        ->  class Nah  [0.857]

Rule 5: (4, lift 1.7)
        fndgrp = CIG
        TypeofInstitute = None of the Above
        ->  class Nah  [0.833]

Rule 6: (15/3, lift 1.5)
        staff_full_time_employees_total > 28
        fndgrp = OTH
        TypeofInstitute = Arts Center/Council
        ->  class Nah  [0.765]
```

Summary of the Results:

```
> predictions <- predict(Tree5, CDP1[761:1015, ])
> table( CDP1[760:1015,14], predictions)
Error in table(CDP1[760:1015, 14], predictions) :
  all arguments must have the same length
> table( CDP1[761:1015,14], predictions)
            predictions
             Nah Subsidize
  Nah         64        67
  Subsidize   38        86
```

Summary of the Optimized Results:

```
  no non-missing arguments to min; returning Inf
> Tree6 <- C5.0(CDP1[1:760, -14], CDP1[1:760, 14], trials = 10)
> Tree6

Call:
C5.0.default(x = CDP1[1:760, -14], y = CDP1[1:760, 14], trials = 10)

Classification Tree
Number of samples: 760
Number of predictors: 13

Number of boosting iterations: 10 requested;  6 used due to early stopping
Average tree size: 14.5

Non-standard options: attempt to group attributes

> predictions1 <- predict(Tree6, CDP1[761:1015, ])
> table(CDP1[761:1015,14], predictions1)
           predictions1
            Nah Subsidize
  Nah        79        52
  Subsidize  56        68
> 
```

The Overall Analysis:

- Essentially, the model (whether boosted or not) predicted approximately 42% of the data values accurately

- The percentage could be enhanced by either altering the way the algorithm traverses throughout the tree, or by utilizing different algorithms. Also GraphViz can be incorporated to show what the entire tree looks like.

Where to go from Here:

- In the future, the author would like to utilize multiple machine learning algorithms, and conduct the same procedure, and then compare which one possesses the best trade off between accuracy and efficiency.

- Additionally, the author would like to introduce a clustering aspect into the model to hopefully offer other potential predictors. The author hopes that there will be more research conducted on the value of predictive analytics within the arts industry. More research on such a topic could create additional financial opportunities for arts organizations, and even create unforeseen economic and social benefits for those within the arts, and also for the surrounding community.

I Would Like to Thank:

- Jonathan Peters, Ph.D., Professor of Finance at The College of Staten Island
- Nora Santiago Urban Analyst at the College of Staten Island
- Dr. Kress Ph.D., Vice President of Economic Development and Civic Prosperity
- And our professor Dr. Chun