Nora Keenan
April 14, 2019

Springboard Capstone Project 1

Final

Pew Research Center - Libraries 2015

Description of Dataset: This dataset includes 2004 responses to a nationwide telephone survey conducted by Princeton Survey Research Associates for the Pew Research Center in March and April 2015. The purpose of the survey was to assess library utilization across the United States. This survey is one of a series of periodic assessments on the question of library usage conducted for the Pew Research Center.

Surveyors asked, among other things, whether the respondent had visited a library or bookmobile in the previous twelve months, what they did at the library, their opinions on fundamental library services, their internet access/usage, as well as demographic information on age, education, ethnicity, dwelling type, zip code and income.

The quality of the dataset is high. Few data values are missing and the survey was designed to minimize response bias. Respondents are spread across geographic, income, age and education classifications.

Problem: Analysis of the questionnaire results by the Pew Research Center found that libraries do a good job of appealing to patrons with higher education and income levels. The intent of my project is to determine what aspects of libraries appeal to those with a high school or lower level of education. By identifying these features libraries can build and expand these programs to improve the education levels and, by extension, incomes of those in their communities. Libraries could also use the results to prioritize funding or approach schools, businesses or other education partners with specific projects.

To determine these features I will begin with an analysis of basic statistics (mean, median, mode, standard deviation) and feature correlation. I will use the data in a predictive model that seeks to accurately categorize the respondents by whether or not they visited the library.

Possible Project Extension Ideas: An interesting additional option for this project would be to overlay United States census data with the model results. This analysis could pinpoint the counties and states that would most benefit from specialized library outreach and education programs. I could also try find survey results from other years and determine if the results could be combined. This would provide a larger dataset on which to build the model.

Data Collection and Wrangling

Description of Dataset: The dataset consists of 2004 responses to a nationwide telephone survey regarding library usage. The data were made available in an Excel spreadsheet consisting of 1 row of column headers, 2004 rows of survey responses and 142 columns. Each column contains the response to an individual question within the survey.

Data Cleaning Steps: Upon deeper review of the data I discovered that the copy of the survey provided with the data did not match the responses. The data are the results from the 2015 survey, but the survey itself is from 2016. Most individual questions cannot be matched between the data set and the survey. Consequently the data set will be reduced to those features, mostly demographic, that can be matched with confidence. The final dataset was reduced to 21 columns.

Missing values in categorical variables: Some questions were only asked if the respondent answered a previous question in a certain way. The questions where a response is not expected will be updated to show “NA” for “Not Applicable.”

Number of Children if Respondent is a Parent: Responses 0-5 represent the actual number of children under 18. Eight (8) represents “Don’t Know” and 9, “Refused”. Eight (8) and 9 responses will be replaced with “Unknown.”

Education Level: The survey breaks down education level into seven categories. I will create an ordinal representation of the values.

Income Level: The survey breaks down income level into nine categories. I will create an ordinal representation of the values.

Outliers: Due to the nature of the dataset and its few variable data points, there are few outliers. Most responses consist of one of the following four values: 1) Yes, 2) No, 8) Don't Know, or 9) Refused.

Columns in the dataset (description in parentheses):

sex
age
marital (Marital Status)
is_parent (Yes/No)
education_level
emplnw (Employment Status)
disabled (Yes/No)
party (Political Party Affiliation)
ideology (conservative, moderate or liberal)
race
income
hh1 (# of people in the household)
reg_voter (Registered Voter)
email_use (Uses email yes/no)
mobile_phone (Has a mobile phone)
home_int (Has internet at home)
broadband (Has broadband at home)
smartphone (Has a smartphone)
library_onsite (Has visited a physical library location)
library_website (Has visited the library website)
visit_freq (Visit frequency)

Additional Data Cleaning Steps:

1. Addition of a flag column for whether or not the respondent had ever visited the library: has_visited.
2. Added a column to create an ordinal value for income and education levels: inc_ordinal, educ2_ordinal.
3. Renamed columns to be more clearly understood and representative.
4. Updated data values to reflect the survey response rather than the number representation. E.g., changed 1 to "Male" and 2 to "Female".
5. Added columns for categorical variables using the Pandas get_dummies function.

Exploratory Data Analysis and Inferential Statistics

Initial Findings

After cleaning the dataset and creating ordinal columns for those categoricals with a logical order (education and income), I chose to analyze the correlations between library visitation (has_visited) and the other columns.

Strongest Positive Correlations:

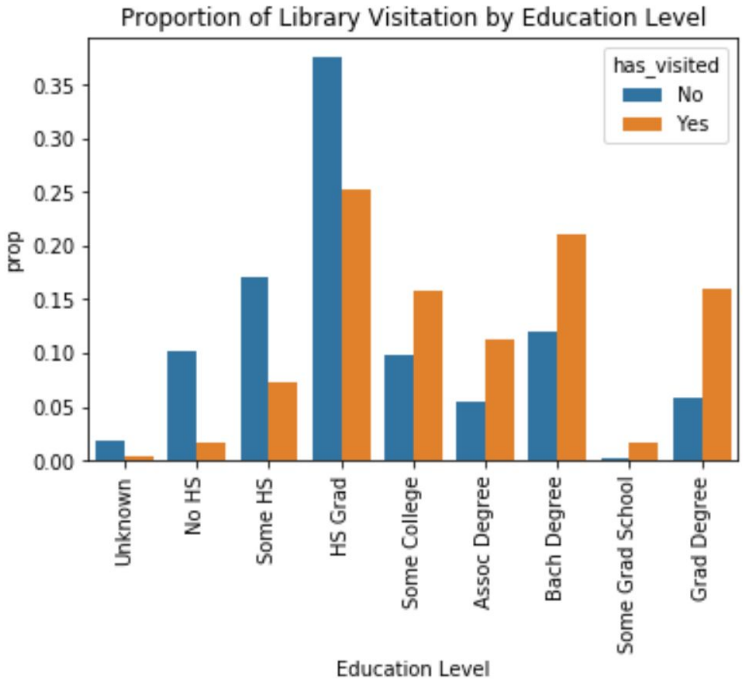
High Speed Broadband at Home	0.266
Has Internet at Home	0.265
Education Level	0.255
Uses Email	0.252
Income Level	0.174

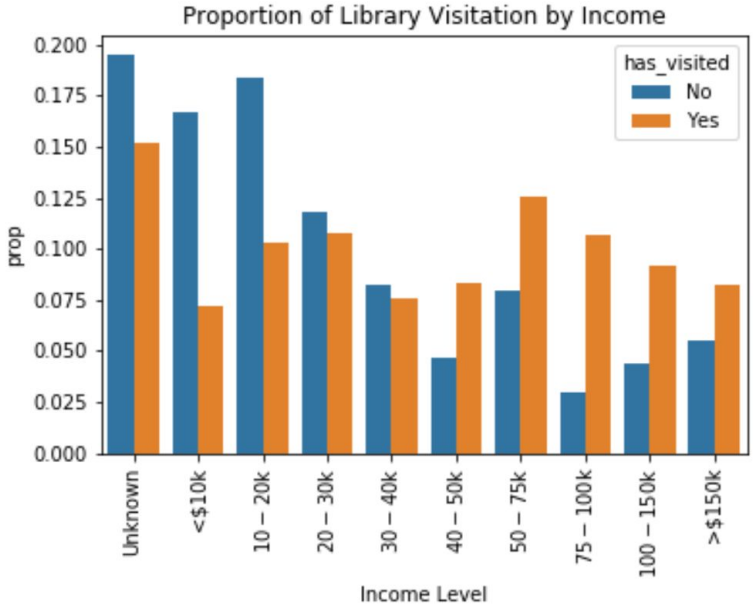
Strongest Negative Correlations:

Unknown Broadband at Home	-0.258
Doesn't Use Email	-0.251
Unknown Home Internet	-0.209
No High School	-0.184
Probably a Registered Voter	-0.146

Visualizations:

Two themes in the correlation analysis are education level and income. In the bar graphs below, it is apparent that as income and education levels increase, so does the likelihood that someone will have visited the library.





Inferential Statistics

In order to determine if the correlation and visual evidence portrays a statistically significant feature, I completed a one sided z test on the difference in proportions for visitation by high school graduates. In the bar graph is appears that respondents with no more than a high school are less likely to have visited the library than not have visited. A proportional comparison of the two options was determined to be the appropriate statistic.

1 sided z test with a significance level of 0.01
Is education level a predictor for library visitation?

Values:
Population 1: High school diploma and hasn't visited
Proportion= .375342
n = 365

Population 2: High school diploma and has visited
Proportion -= .0.251373
n = 1639

The proportion of visitors with a high school diploma is equal to or greater than those with a high school diploma that don't visit. $H_0 = P_1 <= P_2$
The proportion of visitors with a high school diploma is lower that nonvisitors. $H_{alt} = P_1 > P_2$

$$p = \frac{(p1*n1+p2*n2)}{(n1+n2)}$$

$$p = \frac{(.375342*365+.251373*1639)}{365+1639}$$

p=0.274

Standard Error

$$SE = \sqrt{p(1-p) * (\frac{1}{n1} + \frac{1}{n2})}$$

$$SE = \sqrt{(0.274(1-0.274) * (\frac{1}{365} + \frac{1}{1639}))}$$

SE = 0.000666

z score

$$z = \frac{p1-p2}{SE}$$

$$z = \frac{.375342-.251373}{.000666}$$

z = 186.14

Probability of a z value of 186.14 = 0.0
0.0 is less than 0.01. The null hypothesis is **rejected**.
High school graduation status is a significant predictor of library visitation.

Statistical, correlation and visual analysis above provides areas of interest for the design of a subsequent machine learning model. Income, education level and internet access are features likely to factor into a successful predictive model for library visitation.

Machine Learning Analysis

I approached this machine learning problem by recognizing the almost total categorical nature of the data. The size of the dataset (2004 samples) is large enough to warrant analysis by machine learning algorithms that work best with categorical data.

These are the steps I followed in my machine learning process:

1. Transformed data to be most amenable to machine learning analysis.
    a. Assigned get_dummies transformation to categorical features
    b. Dropped features that were surrogates for the label, has_visited.
2. Created X and y variables consisting of a dataframe of features (X) and the label (y).
3. Divided the X and y variables into train and test sets using train_test_split function in sklearn. Seventy percent (70%) of the data was used as the training set and 30% was used for the test set.
4. Due to the highly categorical nature of the data, a classifier model was determined to be the most appropriate.
5. Applied the following models to the training and test data:
    a. Random Forest Classifier
    b. AdaBoost Classifier
    c. Gradient Boost Classifier
    d. KNeighbors Classifier
6. Calculated Train and Test set accuracy against each model.
7. Calculated AUC_ROC values for each prediction set.
8. Generated a Confusion Matrix for each model.

Best Fit Results:

Algorithm Name	Train Accuracy	Test Accuracy	AUC_ROC Value
Random Forest	0.907	0.821	0.565
AdaBoost	0.837	0.822	0.577
Gradient Boost	0.887	0.822	0.562
KNeighbors	0.823	0.818	0.520

Algorithm Name	Confusion Matrix
Random Forest	Predicted No Yes All True No 18 89 107 Yes 19 476 495 All 37 565 602
AdaBoost	Predicted No Yes All True No 21 86 107 Yes 21 474 495 All 42 560 602
Gradient Boost	Predicted No Yes All True No 17 90 107 Yes 17 478 495 All 34 568 602

KNeighbors	

Predicted	No	Yes	All
True			
No	6	101	107
Yes	8	487	495
All	14	588	602

Best Machine Learning Model - AdaBoost

Of the algorithms tested, the most successful was the AdaBoost as demonstrated by the following:

1. Tied with Gradient Boosted Classifier for Test Accuracy
2. Higher ROC_AUC value (A: 0.577 vs GB: 0.562)
3. Does a better job of predicting “no” visitors: (A: 21 vs GB: 17)

AdaBoost combines multiple classifiers to increase the accuracy of classifiers. AdaBoost is an iterative ensemble method that combines multiple poorly performing classifiers to build a high accuracy strong classifier. AdaBoost assigns feature weights by training on each data sample. The model coefficients are then combined to create a general model of the dataset. AdaBoost works best on data that is not highly variable, as outliers can skew the overall model coefficient calculations, creating a model that isn’t very accurate.

AdaBoost Equation

Given: $(x_1, y_1), \ldots, (x_m, y_m)$ where $x_i \in \mathcal{X}$, $y_i \in \{-1, +1\}$.
Initialize: $D_1(i) = 1/m$ for $i = 1, \ldots, m$.
For $t = 1, \ldots, T$:

- Train weak learner using distribution D_t.
- Get weak hypothesis $h_t : \mathcal{X} \to \{-1, +1\}$.
- Aim: select h_t with low weighted error:

$$\varepsilon_t = \Pr_{i \sim D_t}[h_t(x_i) \neq y_i].$$

- Choose $\alpha_t = \frac{1}{2} \ln\left(\frac{1 - \varepsilon_t}{\varepsilon_t}\right)$.
- Update, for $i = 1, \ldots, m$:

$$D_{t+1}(i) = \frac{D_t(i) \exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

where Z_t is a normalization factor (chosen so that D_{t+1} will be a distribution).

Output the final hypothesis:

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right).$$

Fig. 1 The boosting algorithm AdaBoost.

Conclusion

None of the machine learning models did a great job of predicting whether or not a respondent had visited the library. I believe that the classifier type of algorithm was the correct one, but that the features had too many response options. With each response, the algorithm had fewer and fewer samples to work with.

If I were to continue working with this dataset, I would look for results from additional surveys and combine them to make a larger dataset. I would also combine some of the demographic responses into groups of similar types.