Predicting Patron Patterns for the San Francisco Public Library System

Hanif Lumsden

University of Maryland Global Campus

DBST 667 Data Mining

Dr. Majid Shaalan

8/4/2022

Abstract

Data mining algorithms are used to predict the behaviors of a population. The library dataset is analyzed to gain insight into patron behavior since library usage entails collecting data regarding patrons and information. The analysis involves multiple linear regression modeling of renewal and checkout data and decision tree modeling for age range and supervisor district. Overall, said algorithms provide a general picture of how these models forecast patron patterns. The best analysis for the library data set is the decision tree. Optimally, neural networks will give the best analysis.

Keywords: linear regression, decision tree, library circulation system, patron patterns, data mining

Introduction

San Francisco’s integrated library system (SFILS) contains bibliographic records of items, patrons, and circulation data (DataSF, 2016). Essentially, the data represents usage of the SFILS and is composed of 420,000 records. Using data mining classification techniques, frequency association, and classifiers will predict expected library activity and patron type, specifically, renewal and checkout behavior along with behavior based on age and location.

Background Information

The public library analyzed is an open, public domain for all visitors of the city and county (San Francisco Public Library, n.d.; Wittmann et al., 2019). SFILS aims to provide a safe and pleasurable atmosphere for patrons to study, work, learn, and use library services by providing a patron code of conduct that governs library use (San Francisco Public Library, n.d.). A historical data set provides complete anonymity that can give insight into patron behavior. Libraries can better suit the patrons if they understand them.

Method

Dataset Selection

This library data is provided by the San Francisco city and county, more specifically, the publishing department of the public library, representing the use of inventoried items from the years 2003 to 2016 (DataSF, 2019). Patron records are provided under the Public Domain Dedication License (DataSF, 2016). Columns consist of a patron type code where patrons are assigned a categorical number for records-sake; patron type definition that describes the patron type number; total checkouts and total renewal columns representing a total number of checkouts and renewals made by the patron; age range is a string column of the from 0 to 9 years, 10 to 19 years, 20 to 24, 25 to 34, 35 to 44, 45 to 54, 55 to 59, 60 to 64, 65 to 74, and 75 and over; home library code which is an alphanumeric identifier; home library definition defines the alphanumeric identifier or where the library the patron signed up at; circulation active month and year columns are when the patron was last active; notice preference and definition is the code and preferred contact information; provided email address is a boolean true and false attribute; year patron registered; outside of the county is another boolean true, and false attribute; and supervisor district is a number that corresponds with the district number (DataSF, 2019).

Method Selection

Multiple linear model regression will be used for registration and checkout by year and a decision tree for predicting patron behavior by age and likely patron type and behavior based on location (Ansari et al., 2020).

Data Preparation and Preprocessing

Missing data values from the Supervisor District column are replaced with 0. Any row containing a null value is deleted. Using the is.na() function, the data frame is checked for not applicable values. Values are deleted using the na.omit() function. For multiple linear model regression and decision tree analysis, the data attributes where the data type is a string or non-numeric are converted to a numeric factor. This includes patron type definition, age range, library code and definition, circulation month and year, preference code and definition, email address, and outside of the county. For the decision tree, the data is shuffled using the sample function.

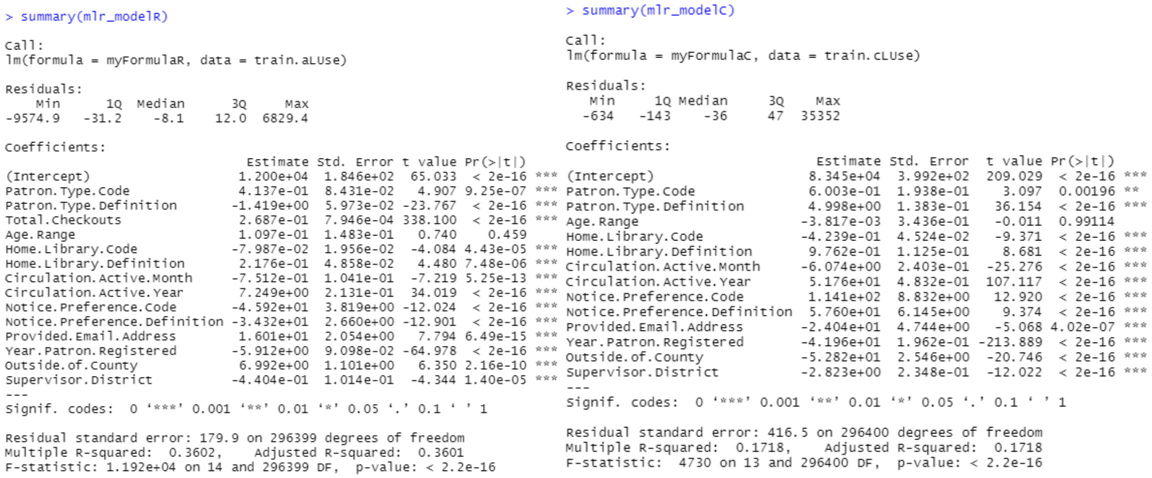
Discussion

Method Analysis

Dependent data variables total checkout and renewals are identified for multiple linear models. The model shows all attributes are dependent variables except age range, such that all dependent variable attributes have a p-value less than 0.05. The dataset has over 400000 attributes, so the model factoring in renewals has an accuracy of 36.02 percent, and the modeling factoring in checkouts has an accuracy of 17.18 percent.

Figure 1

Summary of multiple linear regression model for renewal (left) and checkouts (right)



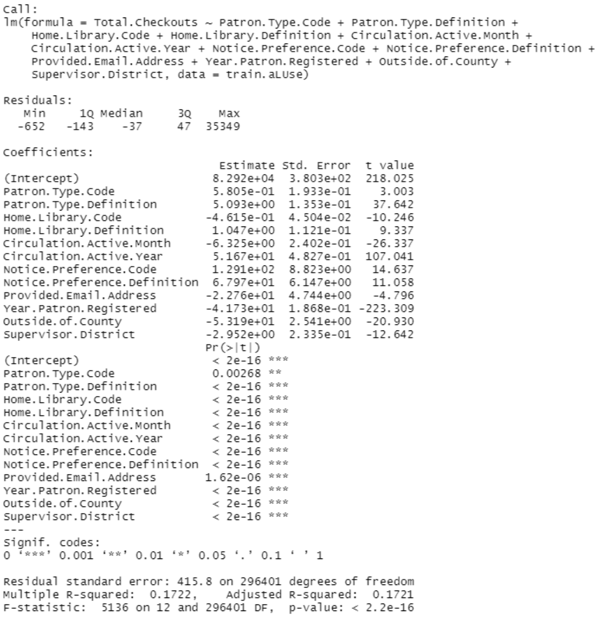
```
> summary(mlr_modelR)

Call:
lm(formula = myFormulaR, data = train.aLUse)

Residuals:
    Min      1Q  Median      3Q     Max 
-9574.9   -31.2    -8.1    12.0  6829.4 

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  1.200e+04  1.846e+02  65.033  < 2e-16 ***
Patron.Type.Code             4.137e-01  8.431e-02   4.907 9.25e-07 ***
Patron.Type.Definition      -1.419e+00  5.973e-02 -23.767  < 2e-16 ***
Total.Checkouts              2.687e-01  7.946e-04 338.100  < 2e-16 ***
Age.Range                    1.097e-01  1.483e-01   0.740    0.459    
Home.Library.Code           -7.987e-02  1.956e-02  -4.084 4.43e-05 ***
Home.Library.Definition      2.176e-01  4.858e-02   4.480 7.48e-06 ***
Circulation.Active.Month    -7.512e-01  1.041e-01  -7.219 5.25e-13 ***
Circulation.Active.Year      7.249e+00  2.131e-01  34.019  < 2e-16 ***
Notice.Preference.Code      -4.592e+01  3.819e+00 -12.024  < 2e-16 ***
Notice.Preference.Definition -3.432e+01  2.660e+00 -12.901  < 2e-16 ***
Provided.Email.Address       1.601e+01  2.054e+00   7.794 6.49e-15 ***
Year.Patron.Registered      -5.912e+00  9.098e-02 -64.978  < 2e-16 ***
Outside.of.County            6.992e+00  1.101e+00   6.350 2.16e-10 ***
Supervisor.District         -4.404e-01  1.014e-01  -4.344 1.40e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 179.9 on 296399 degrees of freedom
Multiple R-squared:  0.3602,	Adjusted R-squared:  0.3601 
F-statistic: 1.192e+04 on 14 and 296399 DF,  p-value: < 2.2e-16
```

```
> summary(mlr_modelC)

Call:
lm(formula = myFormulaC, data = train.cLUse)

Residuals:
   Min     1Q Median     3Q    Max 
  -634   -143    -36     47  35352 

Coefficients:
                              Estimate Std. Error  t value Pr(>|t|)    
(Intercept)                  8.345e+04  3.992e+02  209.029  < 2e-16 ***
Patron.Type.Code             6.003e-01  1.938e-01    3.097  0.00196 ** 
Patron.Type.Definition       4.998e+00  1.383e-01   36.154  < 2e-16 ***
Age.Range                   -3.817e-03  3.436e-01   -0.011  0.99114    
Home.Library.Code           -4.239e-01  4.524e-02   -9.371  < 2e-16 ***
Home.Library.Definition      9.762e-01  1.125e-01    8.681  < 2e-16 ***
Circulation.Active.Month    -6.074e+00  2.403e-01  -25.276  < 2e-16 ***
Circulation.Active.Year      5.176e+01  4.832e-01  107.117  < 2e-16 ***
Notice.Preference.Code       1.141e+02  8.832e+00   12.920  < 2e-16 ***
Notice.Preference.Definition  5.760e+01  6.145e+00    9.374  < 2e-16 ***
Provided.Email.Address      -2.404e+01  4.744e+00   -5.068 4.02e-07 ***
Year.Patron.Registered      -4.196e+01  1.962e-01 -213.889  < 2e-16 ***
Outside.of.County           -5.282e+01  2.546e+00  -20.746  < 2e-16 ***
Supervisor.District         -2.823e+00  2.348e-01  -12.022  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 416.5 on 296400 degrees of freedom
Multiple R-squared:  0.1718,	Adjusted R-squared:  0.1718 
F-statistic:  4730 on 13 and 296400 DF,  p-value: < 2.2e-16
```

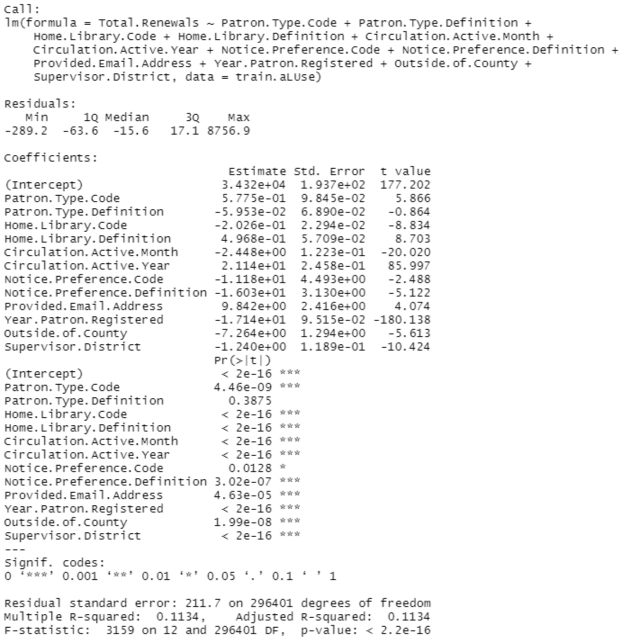
Minimum adequate models are generated to simplify the model. Most attributes are dependent variables, so accuracy does not increase; it decreases. There is a decrease in accuracy due to the contradiction in the purpose. Based on practice, p-values less than 0.05 are considered for the minimum adequate model. The only difference between the minimum adequate model and the initial multiple linear regression model is not factoring in the age range this test run. The accuracy of the model amounts to 17.22 percent considering the renewal attribute and 11.34 percent considering checkouts:

Figure 2

Summary of the minimum adequate model for renewal (left) and checkouts (right)



```
Call:
lm(formula = Total.Checkouts ~ Patron.Type.Code + Patron.Type.Definition +
    Home.Library.Code + Home.Library.Definition + Circulation.Active.Month +
    Circulation.Active.Year + Notice.Preference.Code + Notice.Preference.Definition +
    Provided.Email.Address + Year.Patron.Registered + Outside.of.County +
    Supervisor.District, data = train.aLUse)

Residuals:
   Min     1Q Median     3Q    Max
  -652   -143    -37     47  35349

Coefficients:
                              Estimate Std. Error  t value
(Intercept)                  8.292e+04  3.803e+02  218.025
Patron.Type.Code             5.805e-01  1.933e-01    3.003
Patron.Type.Definition       5.093e+00  1.353e-01   37.642
Home.Library.Code           -4.615e-01  4.504e-02  -10.246
Home.Library.Definition      1.047e+00  1.121e-01    9.337
Circulation.Active.Month    -6.325e+00  2.402e-01  -26.337
Circulation.Active.Year      5.167e+01  4.827e-01  107.041
Notice.Preference.Code       1.291e+02  8.823e+00   14.637
Notice.Preference.Definition 6.797e+01  6.147e+00   11.058
Provided.Email.Address      -2.276e+01  4.744e+00   -4.796
Year.Patron.Registered      -4.173e+01  1.868e-01 -223.309
Outside.of.County           -5.319e+01  2.541e+00  -20.930
Supervisor.District         -2.952e+00  2.335e-01  -12.642
                            Pr(>|t|)
(Intercept)                  < 2e-16 ***
Patron.Type.Code             0.00268 **
Patron.Type.Definition       < 2e-16 ***
Home.Library.Code            < 2e-16 ***
Home.Library.Definition      < 2e-16 ***
Circulation.Active.Month     < 2e-16 ***
Circulation.Active.Year      < 2e-16 ***
Notice.Preference.Code       < 2e-16 ***
Notice.Preference.Definition < 2e-16 ***
Provided.Email.Address      1.62e-06 ***
Year.Patron.Registered       < 2e-16 ***
Outside.of.County            < 2e-16 ***
Supervisor.District          < 2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 415.8 on 296401 degrees of freedom
Multiple R-squared:  0.1722,    Adjusted R-squared:  0.1721
F-statistic:  5136 on 12 and 296401 DF,  p-value: < 2.2e-16
```

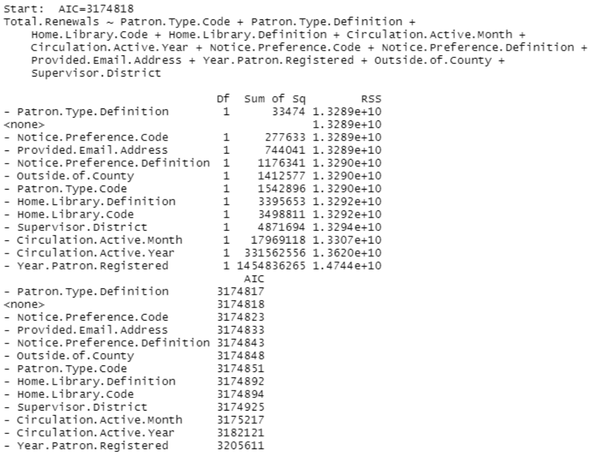


```
Call:
lm(formula = Total.Renewals ~ Patron.Type.Code + Patron.Type.Definition +
    Home.Library.Code + Home.Library.Definition + Circulation.Active.Month +
    Circulation.Active.Year + Notice.Preference.Code + Notice.Preference.Definition +
    Provided.Email.Address + Year.Patron.Registered + Outside.of.County +
    Supervisor.District, data = train.aLUse)

Residuals:
   Min     1Q Median     3Q    Max
-289.2  -63.6  -15.6   17.1 8756.9

Coefficients:
                              Estimate Std. Error  t value
(Intercept)                  3.432e+04  1.937e+02  177.202
Patron.Type.Code             5.775e-01  9.845e-02    5.866
Patron.Type.Definition      -5.953e-02  6.890e-02   -0.864
Home.Library.Code           -2.026e-01  2.294e-02   -8.834
Home.Library.Definition      4.968e-01  5.709e-02    8.703
Circulation.Active.Month    -2.448e+00  1.223e-01  -20.020
Circulation.Active.Year      2.114e+01  2.458e-01   85.997
Notice.Preference.Code      -1.118e+01  4.493e+00   -2.488
Notice.Preference.Definition -1.603e+01  3.130e+00   -5.122
Provided.Email.Address       9.842e+00  2.416e+00    4.074
Year.Patron.Registered      -1.714e+01  9.515e-02 -180.138
Outside.of.County           -7.264e+00  1.294e+00   -5.613
Supervisor.District         -1.240e+00  1.189e-01  -10.424
                            Pr(>|t|)
(Intercept)                  < 2e-16 ***
Patron.Type.Code            4.46e-09 ***
Patron.Type.Definition        0.3875
Home.Library.Code            < 2e-16 ***
Home.Library.Definition      < 2e-16 ***
Circulation.Active.Month     < 2e-16 ***
Circulation.Active.Year      < 2e-16 ***
Notice.Preference.Code        0.0128 *
Notice.Preference.Definition 3.02e-07 ***
Provided.Email.Address      4.63e-05 ***
Year.Patron.Registered       < 2e-16 ***
Outside.of.County           1.99e-08 ***
Supervisor.District          < 2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 211.7 on 296401 degrees of freedom
Multiple R-squared:  0.1134,    Adjusted R-squared:  0.1134
F-statistic:  3159 on 12 and 296401 DF,  p-value: < 2.2e-16
```

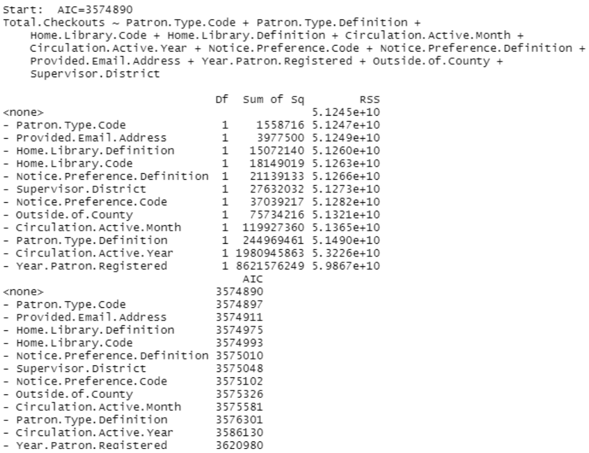
Based on the residuals, both models have an extensive range with the renewal and checkout trading off in terms of range. The Akaike Information Criterion (AIC) is fetched to determine which model fits better. Both AIC values for the models are relatively the same, so this distinction, or lack of it, is not considered due to lack of variety.

Figure 3

Akaike Information Criterion for both renewal (left) and checkout (right) consideration



```
Start:  AIC=3174818
Total.Renewals ~ Patron.Type.Code + Patron.Type.Definition +
    Home.Library.Code + Home.Library.Definition + Circulation.Active.Month +
    Circulation.Active.Year + Notice.Preference.Code + Notice.Preference.Definition +
    Provided.Email.Address + Year.Patron.Registered + Outside.of.County +
    Supervisor.District

                               Df  Sum of Sq        RSS
- Patron.Type.Definition        1      33474 1.3289e+10
<none>                                       1.3289e+10
- Notice.Preference.Code        1     277633 1.3289e+10
- Provided.Email.Address        1     744041 1.3289e+10
- Notice.Preference.Definition  1    1176341 1.3290e+10
- Outside.of.County             1    1412577 1.3290e+10
- Patron.Type.Code              1    1542896 1.3290e+10
- Home.Library.Definition       1    3395653 1.3292e+10
- Home.Library.Code             1    3498811 1.3292e+10
- Supervisor.District           1    4871694 1.3294e+10
- Circulation.Active.Month      1   17969118 1.3307e+10
- Circulation.Active.Year       1  331562556 1.3620e+10
- Year.Patron.Registered        1 1454836265 1.4744e+10
                                   AIC
- Patron.Type.Definition       3174817
<none>                         3174818
- Notice.Preference.Code       3174823
- Provided.Email.Address       3174833
- Notice.Preference.Definition 3174843
- Outside.of.County            3174848
- Patron.Type.Code             3174851
- Home.Library.Definition      3174892
- Home.Library.Code            3174894
- Supervisor.District          3174925
- Circulation.Active.Month     3175217
- Circulation.Active.Year      3182121
- Year.Patron.Registered       3205611
```

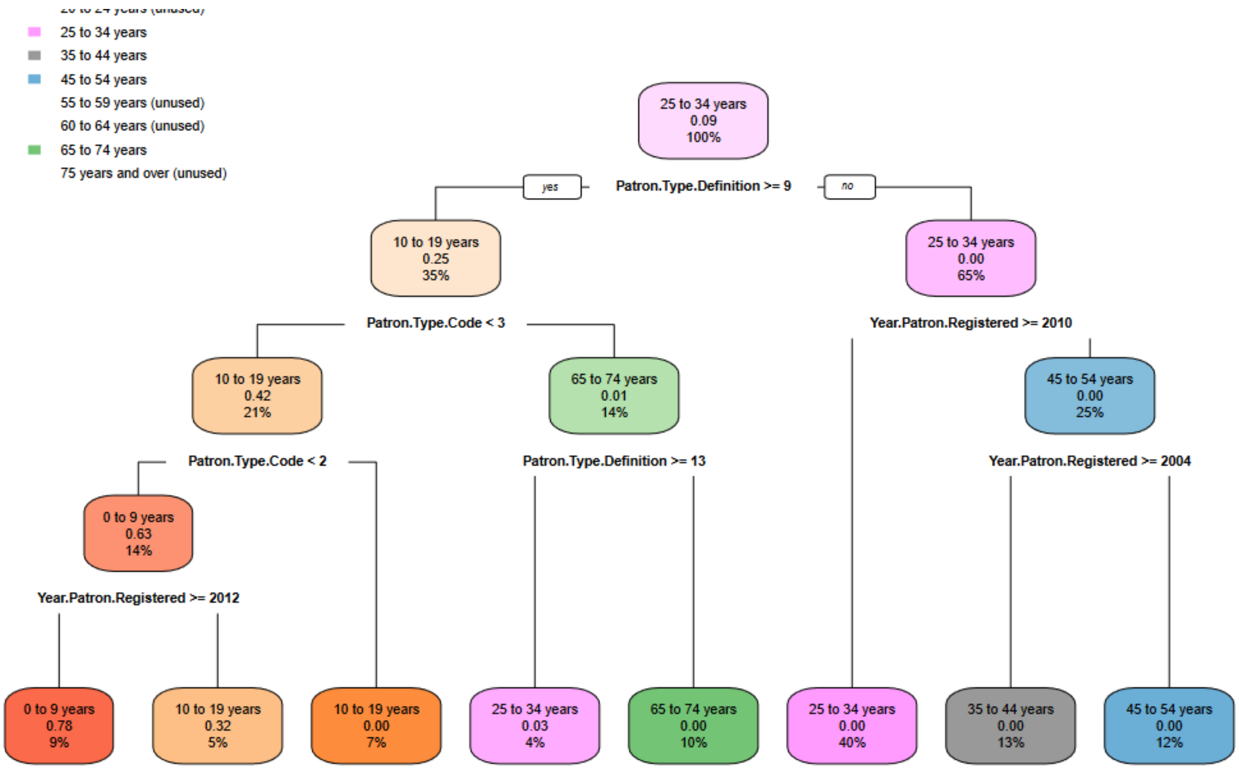


```
Start:  AIC=3574890
Total.Checkouts ~ Patron.Type.Code + Patron.Type.Definition +
    Home.Library.Code + Home.Library.Definition + Circulation.Active.Month +
    Circulation.Active.Year + Notice.Preference.Code + Notice.Preference.Definition +
    Provided.Email.Address + Year.Patron.Registered + Outside.of.County +
    Supervisor.District

                               Df  Sum of Sq        RSS
<none>                                       5.1245e+10
- Patron.Type.Code              1    1558716 5.1247e+10
- Provided.Email.Address        1    3977500 5.1249e+10
- Home.Library.Definition       1   15072140 5.1260e+10
- Home.Library.Code             1   18149019 5.1263e+10
- Notice.Preference.Definition  1   21139133 5.1266e+10
- Supervisor.District           1   27632032 5.1273e+10
- Notice.Preference.Code        1   37039217 5.1282e+10
- Outside.of.County             1   75734216 5.1321e+10
- Circulation.Active.Month      1  119927360 5.1365e+10
- Patron.Type.Definition        1  244969461 5.1490e+10
- Circulation.Active.Year       1 1980945863 5.3226e+10
- Year.Patron.Registered        1 8621576249 5.9867e+10
                                   AIC
<none>                         3574890
- Patron.Type.Code             3574897
- Provided.Email.Address       3574911
- Home.Library.Definition      3574975
- Home.Library.Code            3574993
- Notice.Preference.Definition 3575010
- Supervisor.District          3575048
- Notice.Preference.Code       3575102
- Outside.of.County            3575326
- Circulation.Active.Month     3575581
- Patron.Type.Definition       3576301
- Circulation.Active.Year      3586130
- Year.Patron.Registered       3620980
```

The decision tree is modeled to determine better fit for the library data:

Figure 4

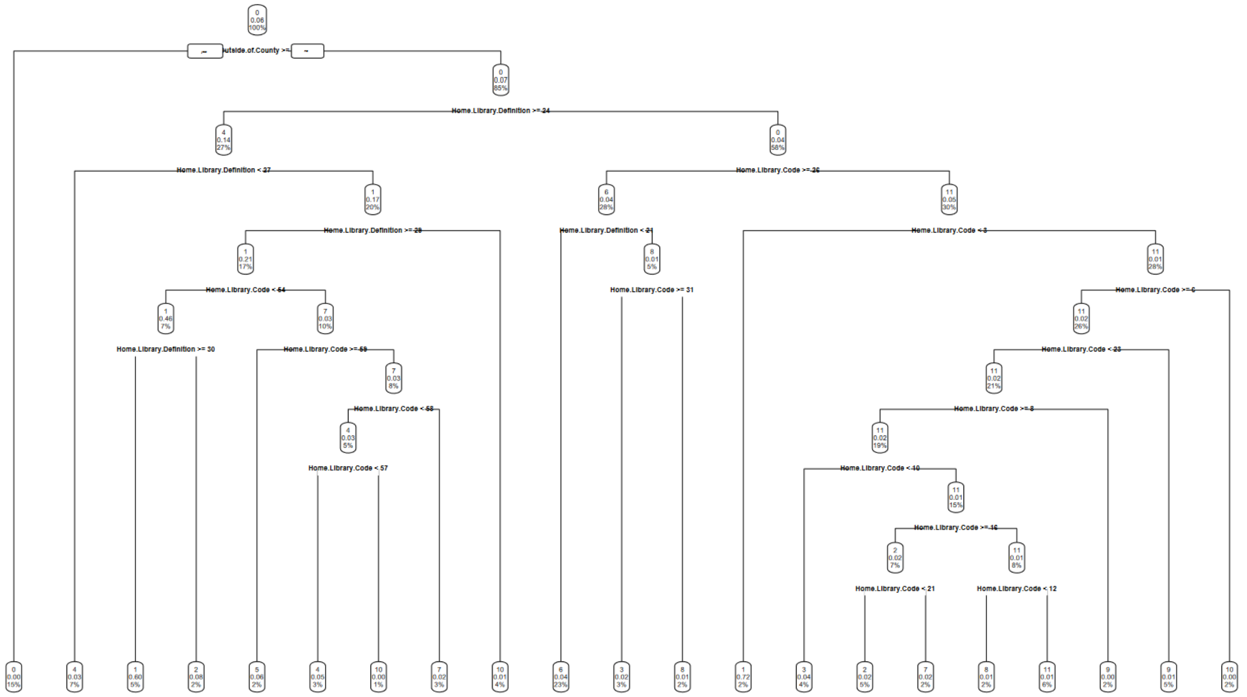
Decision tree considering Age Range



At the start of the root node, the probability that the patron is 25 to 34 years is 9 percent. If the patron is not labeled as a staff, senior, welcome or young adult, juvenile, or patron definition type less than 9, the node goes to the correct node, where guest and adults are between the ranges 25 to 34, which in turn predicts year patron registered and amounts to a very low value. When said yes to patron definition type over 9, 35 percent of patrons with a type definition over 9 are 25 percent likely between the ages of 10 to 19. The type code is predicted. It is determined that the majority of patrons that were kids that registered past 2012 are between the ages of 0 to 9, and before 2012 were between the ages of 10 to 19. The patron type code being more significant than 3 means that the patron is an adult. It is determined that patron type definition greater than 13 or comprising of staff and young adults are between the ages of 25 to 34 on a scale of 3 percent.

Figure 5

Decision tree considering supervisor district



Note: Full resolution screenshot is viewed here.

This is a much more detailed decision tree considering the supervisor district or patrons' location. There are many insights, including the prediction that 6 percent of patrons will remain outside the county. From then on, home library definition and code are emphasized or the description where the patron is registered originally registered in home library definition, then home library code which can change depending on when the patron wants to adjust their destination library. Essentially, a predictor based on location is displayed with many potential outcomes. The decision tree is a better fit for the library circulation data based on the accuracy of the models expounded in **Method Performance**.

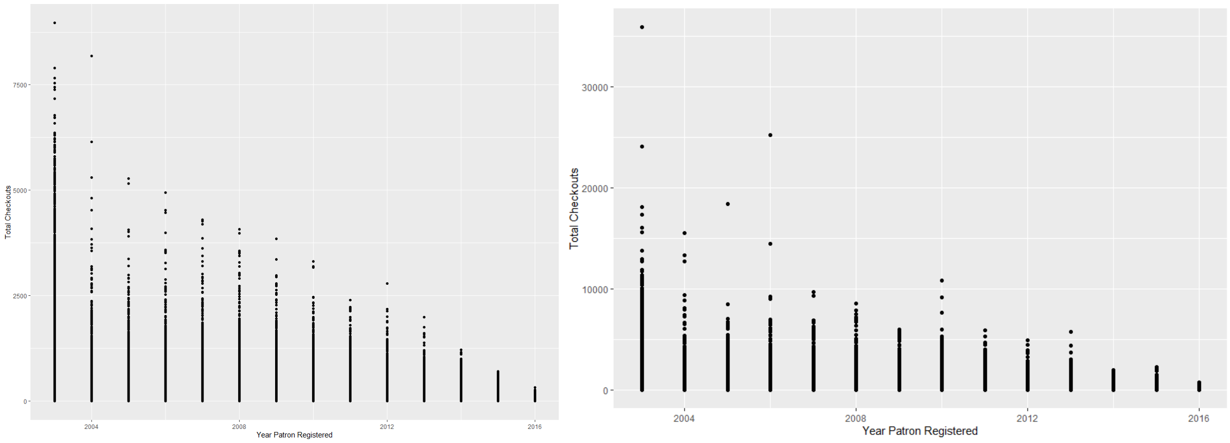
Results

Method Performance

The distribution of checkout and renewal activity as a function of year is modeled on a regular plot using the ggplot2 library. Below is the plot:

Figure 6

Total checkouts (left) and renewals (right) plot v year

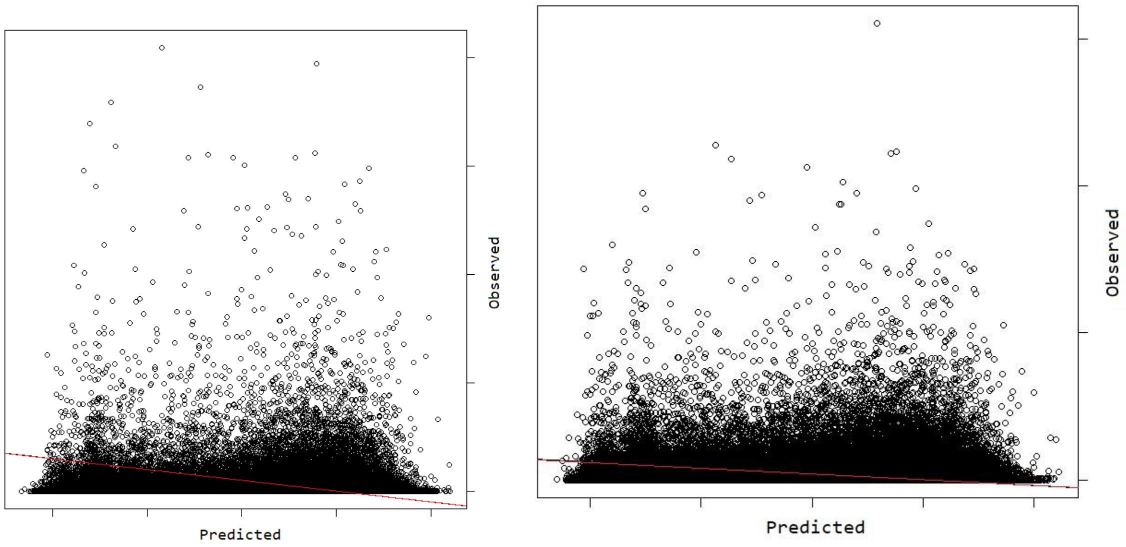


Clear from these two graphs, the majority of the activity happened during the year 2003 when this particular patron dataset was first gathered. Activity decreases over time. It is assumed that technology is particular a factor in the decrease.

It must be emphasized that the distance between two points. According to the predictive model, the majority of values are clustered along the fit, and both summaries share the same general regression behavior. There are many outliers, however, and this is very indicative of the accuracy model.

Figure 7

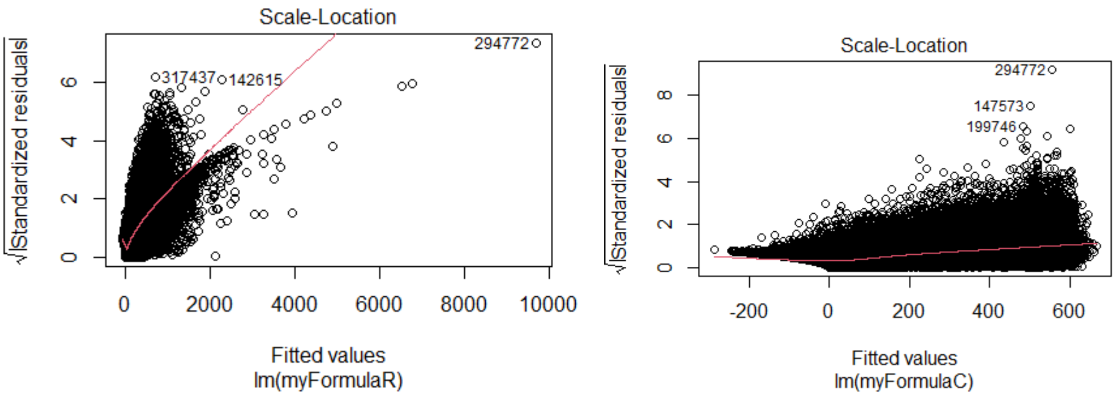
Multiple linear regression of renewal (left) and checkout (right)



A more indicative fit is presented, showing the lack of distribution uniformity along the regression line:

Figure 8

Scale location of multiple linear regression model of renewal (left) and checkout (right)



The decision tree model fits the data better. Consider a dependent and independent variable in supervisor district and age range, respectively. The accuracy amounted to 48.04 percent for the age range decision tree and 51.8 percent for the supervisor district decision tree as opposed to accuracy being below 20 percent for the multiple linear regression model.

The two methods used while analyzing the entire data set amounted to very slow performance on an 8 GB ram computer. When handling this data set, the majority of data should be deleted. Doing this will lead to better performance and potentially better accuracy.

Literature Comparison

Ansari et al. (2020) predicted behavioral models using many data mining techniques ranging from neural networks to decision trees. Decision trees are inquired about since the best results from this report came from decision tree classification. Ansari et al. (2020) used the decision tree algorithm to predict loan period, referrals, and user delays, ranging from 59.3 percent to 90 percent. The neural network algorithm produced the greatest accuracy rates, up to 99 percent, close to the actual library behavioral patterns (Ansari et al., 2020).

Han et al. (2022) utilize an algorithm entitled 'fast update' or FUP. The FUP will scan the database once and produce a computation to increase library operation efficiency (Han et al., 2022). Hans uses association rule mining and, with high accuracy, determines borrowing behavior. The decision tree model done for the San Francisco library amount to around 50 percent, which is around 10 percent less than the least accurate run for Ansari et al. (2020) though we were modeling different attributes and data. The range of accuracy for Ansari et al. (2020) shows that while a decision tree can be an effective model for library circulation data, neural network algorithms that assess the same data produce a more accurate model and depiction of actual user behavior. The FUP algorithm will likely better fit the borrowing behavior for the San Francisco library circulation data over multiple linear regression analysis.

Conclusion

The San Francisco public library circulation data from 2003 to 2016 is analyzed using decision tree classification and multiple linear regression modeling. Tests are run while maintaining most of the data. More accurate models will likely be produced if most data is first randomized and deleted. This will have a significant effect on performance as well. Regardless, the decision tree model is a better fit when compared with the multiple linear regression model based on the accuracy produced. It is determined through a literature review that neural networks are a better fit for library circulation data. Analysts should consider this paper and neural network algorithms when dealing with library circulation data.

References

Ansari, N., Vakilimofrad, H., Mansoorizadeh, M., & Amiri, M. R. (2020). Using data mining techniques to predict user's behavior and create recommender systems in the libraries and information centers. *Global Knowledge, Memory and Communication*, *70*(6/7), 538-557. https://doi.org/10.1108/gkmc-04-2020-0058

DataSF. (2016). *San Francisco library usage*. Kaggle. https://www.kaggle.com/datasets/datasf/sf-library-usage-data

DataSF. (2019). *Library usage*. San Francisco Open Data. https://data.sfgov.org/Culture-and-Recreation/Library-Usage/qzz6-2jup

Han, C., Yu, W., Li, X., Lin, H., & Zhao, H. (2022). A new fast algorithm for library circulation data mining based on FUP. *Scientific Programming*, *2022*, 1-8. https://doi.org/10.1155/2022/1683099

San Francisco Public Library. (n.d.). *Guidelines for library use*. https://sfpl.org/about-us/guidelines-library-use

Wittmann, R., Neatrour, A., Cummings, R., & Myntti, J. (2019). From Digital Library to Open Datasets: Embracing a "Collections as Data" Framework. *Information Technology and Libraries*, *38*(4), 49-61. https://doi.org/10.6017/ital.v38i4.11101