Final Report for CKME 136 Capstone Project

Submitted by: Josephine Choi

Submitted for : CKME 136

Submitted on: Apr 5, 2020

Submitted to: Dr. Uzair Ahmad

## Introduction

Using a dataset with patron records from San Francisco Public Library, this paper would explore an question that is essential to administrators of a public library: that is what are the demographics of their frequent users or infrequent users? Through reviewing literatures written about this topic, this study wish to find out what are the known about these users, and then to see if these presumptions shall be confirmed by the findings from this dataset. An exploratory data analysis (EDA) would also be conducted to fully understand the dataset before applying predictive data analysis.

This dataset was obtained through San Francisco Open Data project and could be found in the Kaggle website. It contained 42000+ patron record, with fifteen variables, including age range, home library code, year patron registered, total renewals and total checkouts in the period from 2003 to 2016. This study planned to use total checkouts – and maybe total renewals as well - as independent variables, turning this into a classification problem. Based on the number of annual total checkout (calculated by dividing the total checkouts and/or renewals with 2016 minus year patron registered), patrons would be divided into groups labelled as frequent users or infrequent users. Then, this dataset would be used to apply binary classification algorithms, including random forest, logistic regression, and KNN . Treating this as an two-class imbalanced learning problem, I will address the potential skewness towards the majority class by using Imlearn, a Python library for applying synthetic minority oversampling techniques (SMOTE), to oversample the minority class. Finally, I will use 10-folds cross validation process to evaluate the effectiveness using the mean of roc\_auc.

## Background: location and the institution

From information found in San Francisco, California (CA) profile, we know that San Francisco is an urban region with population size 884 K in 2017. The growth in the city was quite rapid in the period we will study, as demonstrated in the 13.9 % population change since 2000 (San Francisco, California (CA) profile).There is no indication that the city-county is serving an aging population, as the median resident age was 38.9 years. California median age was 36.5 years (Ibid.).

The estimated household income in 2017 was $110 K, much higher comparing to that of California ($71 K). The cost of living, however, was also notably high. In 2017, the estimated median house or condo value was over 1 million. That was a huge jump from its value in 2000 ($422 K). The March 2019 of cost living index in San Francisco is 173.6. That is very high, as the U.S. average was 100. The percentage of the population living in poverty was 10 % in 2017 (Ibid.).

In public librarianship, there is a school of thought that emphasizes on the professional and social responsibility that public librarians have towards the underserved marginalized populations. San Francisco Public Library (SFPL) is one of the libraries that have taken commitment to social justice seriously. Lilienthal (2011) reported that SFPL has Health and Safety Associates (HaSA), hired as part-time interns for the library. This group, comprised of formerly homeless men and women who were clients of the city's Homeless Outreach Team (HOT), provide peer counselling and offer information about food and shelter to patrons experiencing homelessness (Lilienthal 2011). Comito (2015) noted that SFPL hired full-time social workers to help address homelessness in the community, and this was shown in their partnerships and programming. The library, partnering with Lava Mae, a non-profit organization that retrofitted decommissioned city buses with shower facilities, offers free showers to the public. Comito quoted Leah Filler, a global community engagement coordinator of Lava Mae, who pointed out,

"The library is unique (as a partner) because they are not a homeless service provider, they are a public library. What’s great about them is that they are one of the few institutions in San Francisco that have adamantly kept their doors open to...all members of the public, including people who are homeless. In doing so they preserve public restroom access, which is extremely limited right now in [the city]. So they are known as one of the only places people can go to use the restrooms, get on the computer, have a quiet space to read for the day."(ibid.)

As a library professional, for me SFPL's commitment to social justice is a good reminder that, while public libraries administrators are often positioned to validate its value with quantitative measures, part of its institutional value would always be incommunicable through numbers. In the case of San Francisco, a program serving the affluent population may generate higher statistics, given that such population is likely to have the leisure time to use the service. However, it does not feel right, for me as a library professional as well as a human being, to say that we should overlook the 10 per cent population that is in poverty.

Thus, as a learner of quantitative methods and predictive analysis, I am acutely aware of the limitation of this dataset, when used to evaluate the value of library. To go a step further, while I am interested in using this dataset to explore the demographic characteristics of frequent users/ infrequent users of the library, I note that I should not attach judgment to these users based on such labels. A user can borrow ten books a year, but read each in depth with great pleasures and erudite insight; a user can borrow a thousand books a year, but read only few and flip through most. The categories "frequent user" and "infrequent user" simply could denote the usage pattern. Neither does it characterize how the materials are being used, nor would it denote the frequency of visits. Furthermore, reading one book can make all the difference in a person's life; therefore, this in itself can already be an act worth celebrating.

## Literature regarding circulation activities

A search in Library-specific abstract and index (Library, Information Science & Technology Abstracts (LISTA)) finds no results related to this particular dataset. This shows that, while there are exploratory data analysis (EDA) being done with this dataset in Kaggle, such attempts are likely to be conducted by data science enthusiasts, not library professionals.

Ideally, we can find literature about studies done in the area of public libraries. If there is none, a broader scope (which covers all kind of libraries) will be used. In LISTA (a library specific database), the most useful subject term is "LIBRARY circulation analysis". At the first glance, it does look like it is more being done in academic libraries. We use subject terms LIBRARY circulation analysis" and "public libraries" to conduct the search.

Library circulation statistics are important to public libraries. In fact, librarians have gone so far to fake library usage as attempts to save low-usage but valuable items in their collections (Brunvand Mar-Apr. 2017). It is however not common for public library professional to use patron data to understand the patrons. American public libraries use an index called "Index of American Public Libraries Circulation" to gauge whether their circulation statistics is proportional to the expenditures (Cree and Yoon. 2005). This study, trying to tap into the power of data science to advance library science, is an novel idea.

Cree and Yoon reported that in 2004, that is, in the earlier part of the period our data is covering, the Index of American Public Libraries Circulation showed a decrease in both adult and juvenile circulations. This indicate, "a change trend in public library use" (ibid. Cree and Yoon).

When seeing a surge or a drop in circulation statistics, public library professionals have explained the changes by attributing to various factors: mild weather (“Library Circulation Booming in Lake County, Indiana"), a popular addition - such as graphic novels (Raiteri, 2006) , career-related books for a population with high unemployment rate (“Library Circulation Booming in Lake County, Indiana") to the collection.

## Key Insights from Exploratory Data Analysis (EDA)

The EDA informs us that some initial cleanups and modifications are required before an algorithm can be applied.

First, while at the first glance, this dataset includes 15 variables, some of the variables are duplications. For instance, Patron Type Definition, Notice Preference Definition, and Home Library Definition interpret the Patron Type Code, Notice Preference Definition, and Home Library Code and turn the codes into categories, without adding information values. These categories would have to be removed.

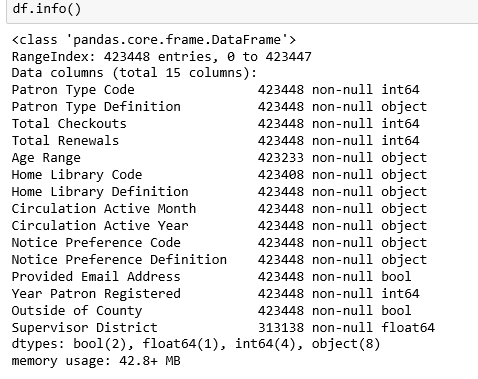


Figure 1 df.info (data description of the raw dataset)

Clearly, three variables - Supervisor District, Age Range, and Home Library Code - have missing values. Supervisor District is missing in approx. 25% of the dataset, so these records definitely could not be imputated. It is worth looking into why such the number is so big. Age Range, Home Library County is small enough, so it can be considered to be imputated. We will do it if we are using these for prediction. "Supervisor District" is an automatically populated fields and will be left blank for users who are outside of country. That will explain high volume of null values in this particular field.

From literature review, I found out that San Francisco Public Library (SFPL) considers equity and social justice as their service priority. They welcome patrons without fixed address - individuals who are homeless - to use their facilities. If the patron record without supervisor district indeed signifies that these are records from vulnerable population, the null value is meaningful and could add power to the model. For that reason, I have filled the null fields with value “12”.

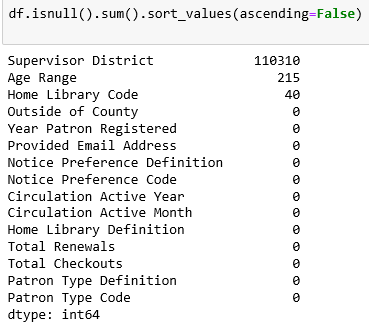


Figure 2 Variable with Null values

Plots have been created for various variables, such as Year Patron Registered, Provided Email Address (True/False), Notice Preference, and Patron Type. The field Year Patron Registered shows continuous increase in the period studied - the year "2003" is likely the year records got imported from old system to new system, so it is not surprised that this category would have a high count; 2016 data is likely to be incomplete, and hence does not reflect the trend. For the other variables, there are not much insight, except that the patrons of SFPL appear

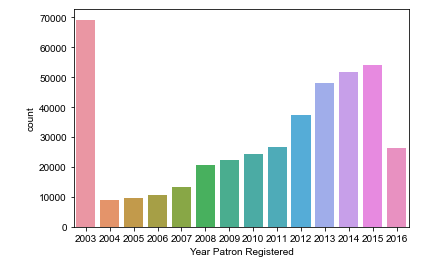


Figure 3 Histogram of Patron Year Registered

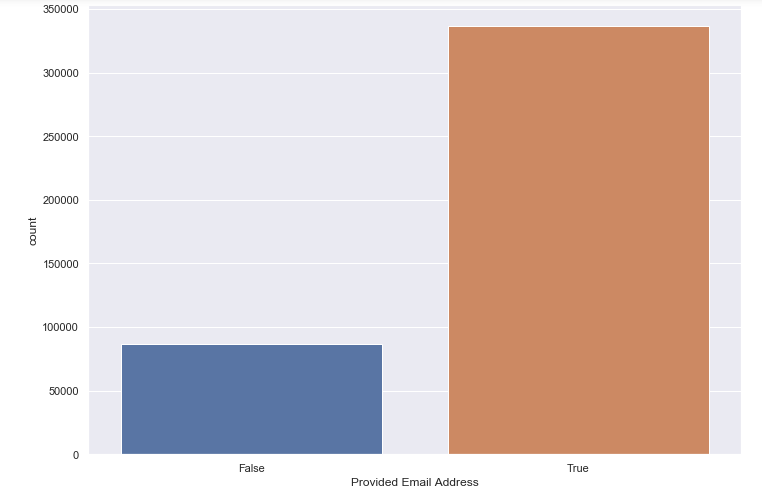


Figure 4 Count Plot of Provided Email Address

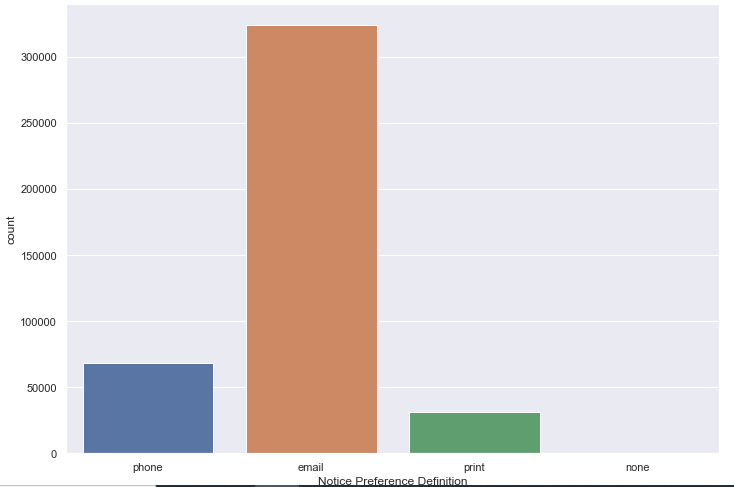


Figure 5 Notice Preference Definition

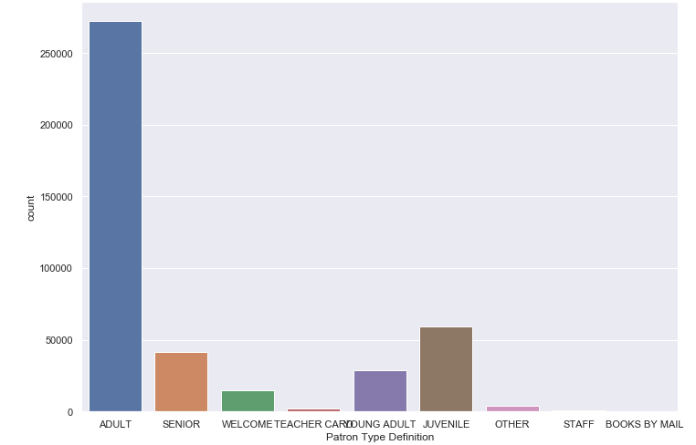


Figure 6 Count plot of Patron Type Definition

The variable “Age range” provides valuable demographic information of the patron. At the first glance, it looks like the data is not normally distributed. However, note that the intervals are not the same: the width varies from 5 years to 10 years. Because of this, this is not a histogram that could be used to determine whether there is any skewness in the data. If we redistribute "25 to 34 years","35 to 44 year", "45 to 54 years" so that each of these category would have an interval of ten years, it looks like the data would peak in the newly category of "30 to 39 years". This presumption is based on the likelihood that both "25 to 34 years" and "35 to 44 years" could be equally split into two parts and redistributed. Based on our literature review, we know that the median age of San Francisco is 38.9 year. This seems to be aligned with our findings with the library data.

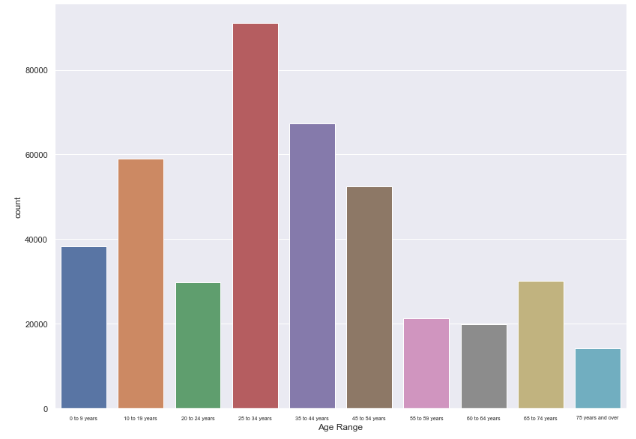
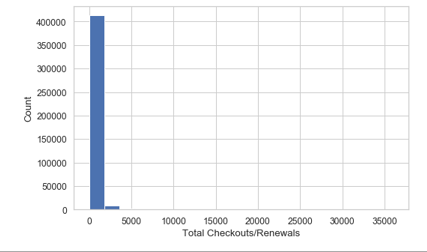


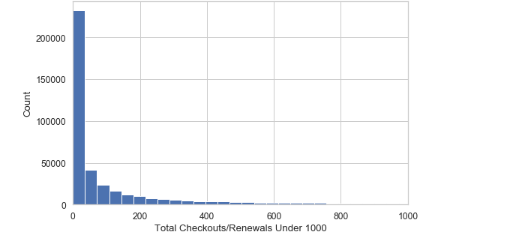
Figure 8 Count plot of Age Range

## Determining the threshold for labelling

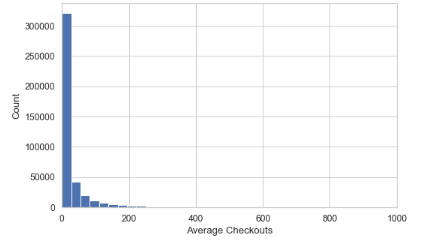
In this dataset, the numerical variables “total checkouts” and “total renewals” are used to identify the frequent and infrequent users. However, because these two variables are cumulative numbers for each user, “avg\_cko” is also calculated to avoid penalizing users who just receive their library cards only in recent years.





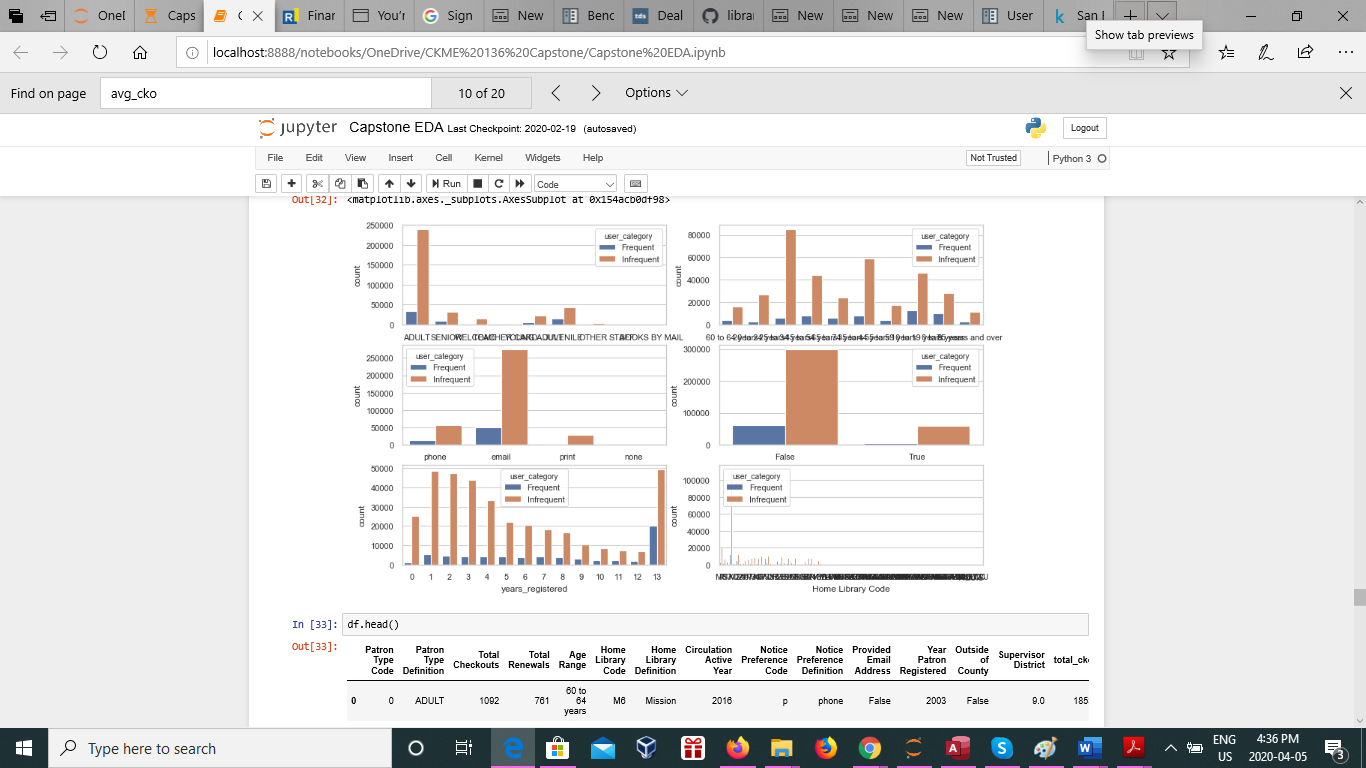


In graphs plotting both Total Checkouts + Renewals and Average Checkouts, the plots show that our data is heavily skewed to the right. This is crucial information to note: not only would this determine the threshold we use to classify the users into ‘frequent’ and ‘infrequent’ users, it also informs us that our data is imbalanced.

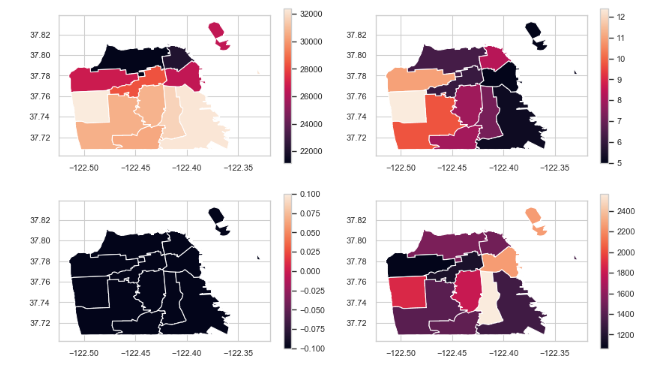


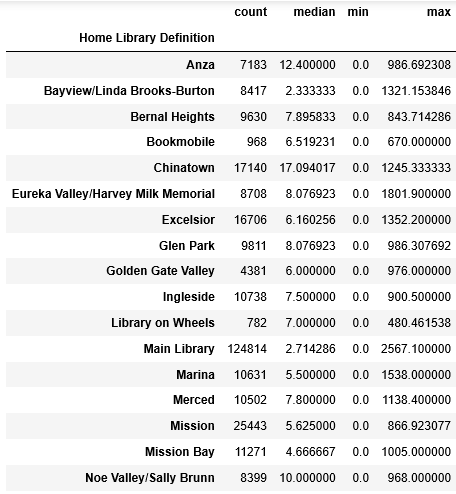
From the above histogram of the average checkouts, we note that most of the users have an average of less than fifty checkouts. Therefore, we will label users with less than fifty checkouts as “infrequent”, and the ones with more than fifty checkouts as “frequent”.

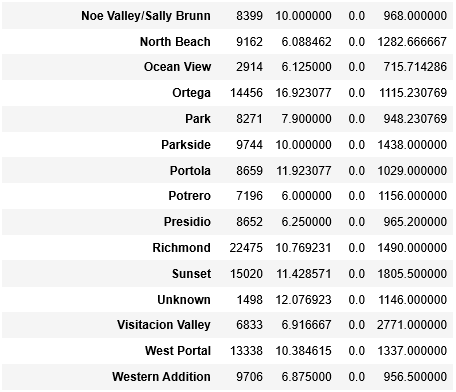
Other graphs were plotted quickly to see if we can find any additional insights based on this additional label. None was found at the stage of EDA.



In addition, because this dataset comes with a shapefile for the supervisor district, we have plotted the aggregated values of the average checkouts ( count, mean, min, and max) based on the supervisor district using Geopanda. The second plot ( top right graph below) shows that the districts in the east side have higher average checkouts, followed by the region in the north. The main branch of SFPL, which has high counts and high average, is in the east end. The northern districts also have higher count (probably higher number of registered users). This information could be meaningful for the library administrators.





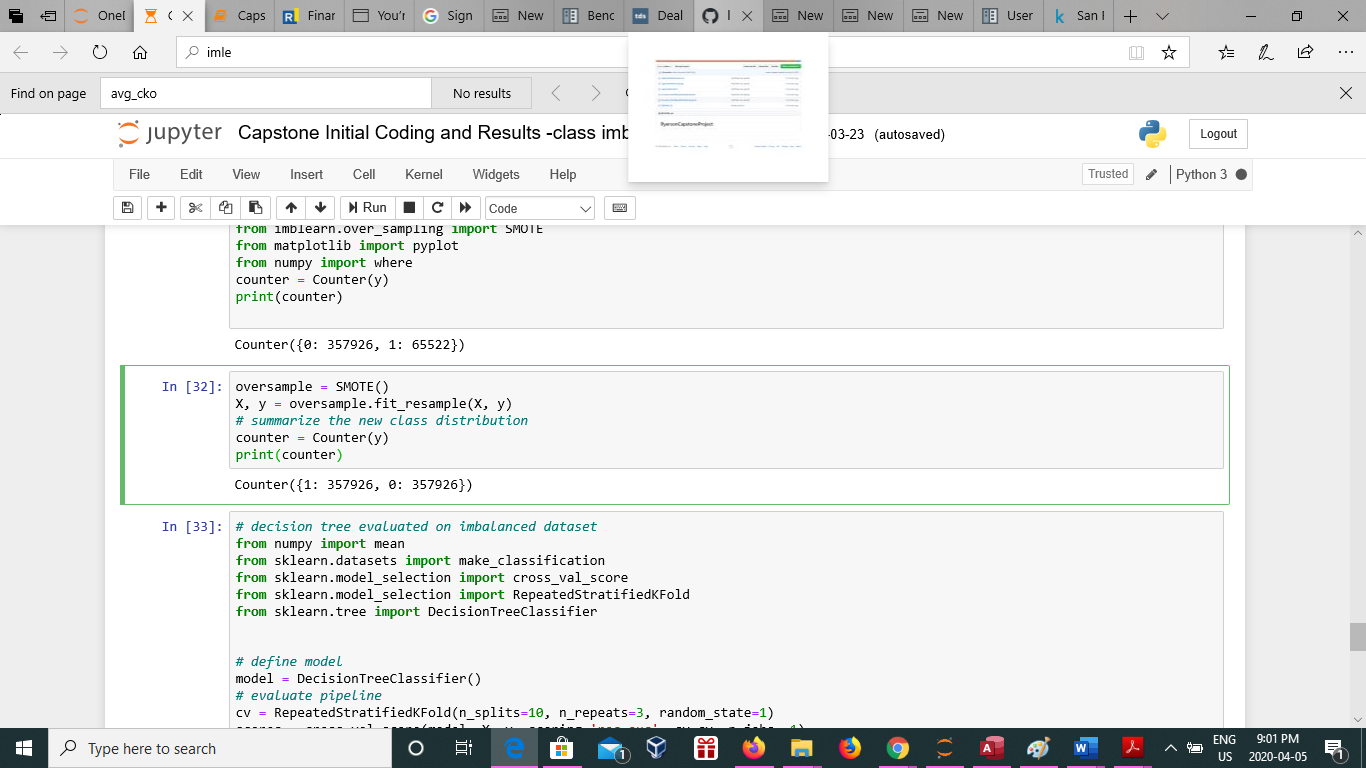


## Imbalanced Data

Counter({0: 357926, 1: 65522})

While neither could I find literature specifically dealing with this dataset nor about applying algorithms to study library circulation, I did review some articles about dealing with imbalanced data. Techniques to handle imbalanced data, as noted by Krawczyk (2016), has wide real-life applications, including cancer malignancy grading, credit card fraud detection, software defect prediction, sentiment analysis, and text mining (Krawczyk 223). My application here is under the category of activity recognition – the detection of rare or less-frequent activities. The minority class is usually the label a research is more likely interested in; yet, due to its smaller size, it can be ignored by algorithms if this issue is not addressed. For this particular dataset, using average annual checkouts larger than 50 as the criteria to identify the frequent users, I found that approximately 18 % of the dataset belongs to the minority class. While the dataset is not severely imbalanced – it is noted that it is not uncommon for class imbalances to be on the order of 100:1, 1,000:1, and 10,000:1 (He and Garcia 2009: 1264) – it does require actions, such as using sampling methods and changing evaluation methods.

To address the imbalance class, I apply Synthetic Minority Oversampling Technique (SMOTE) using a Python library known as imlearn to oversample the minority class. He and Garcia provides a comprehensive review of the development in the area of imbalanced learning, and observes that both oversampling the minority class and undersampling the majority class are two common techniques used to correct the skewness in the data. However, each approaches have its own potential drawbacks – undersampling can lead to overfitting, while oversampling may “cause the classifier to miss important concepts pertaining to the majority class’ (He and Garcia 1267). SMOTE provides option to apply both oversampling and undersampling. It also has options to apply the procedure with advanced variations such as Borderline-SMOTE and Adaptive Synthetic Sampling (ADA-SYN) algorithms. However, for simplicity’s sake, I opt for simply doing oversampling the minority class. The goal is to increase the minority class to the same number as the majority class, resulting a partially synthetic data space with sample size of 715,852. This was done prior to applying algorithms. The image below shows that the sample size has changed from the original 423,448 to 715,852, with 50 % data labelled as 0 (‘infrequent users’) and the other half as 1 (‘infrequent users’).



## Methodology

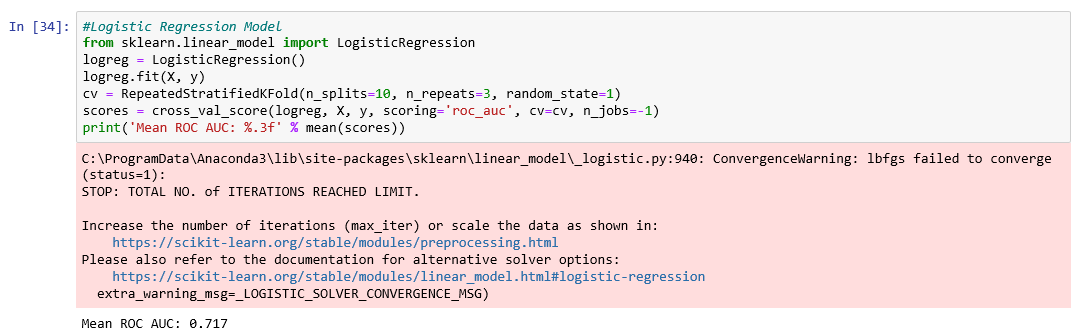
In brief, based on the findings in the EDA, the following procedures have been done prior to apply the three algorithms (random forest, logistic regression, and KNN).

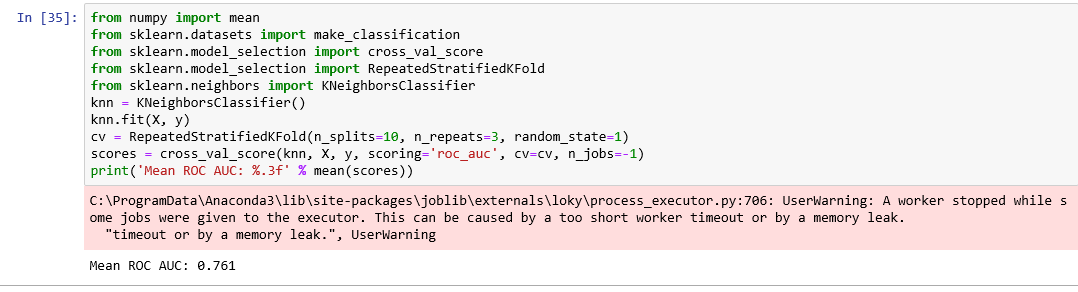
* Remove duplicated variables ( Patron Type Definition, Notice Preference Definition, and Home Library Definition) and a variable with low information value (Circulation Active Month)
* Create new variables, for instance, calculate the number of years the patrons have a library card ( 2016 minus “Year Patron Registered” = “years\_registered”), average annual checkouts (‘avg\_cko’)
* Turn categorical variables, such as Age Range, into numerical variable using median. Turn some categorical variables into Boolean (e.g. Provide email address, outside county)
* Fill missing values in the variable “supervisory district” with numerical value ‘12’. The records with null values could be because these users have no fixed address, and hence it can be valuable information
* Based on average annual checkouts, identify the users with more than 50 checkouts as ‘frequent’ (1). The rest is ‘infrequent’ (0).
* Create dummy data using data\_dum. Discard the first column.
* Apply SMOTE by using Imlearn in Python to oversample the minority class.
* Apply the three algorithms, random forest, logistic regression, and KNN.
* Evaluate the result with ten-fold cross validation using mean ROC AUC (mean of the area under the ROC curve.

## Results

The three algorithms, including random forest, logistic regression and K-neighbor classifier (KNN) has been used to classify this data set.







Among the three algorithms, the random forest seems to perform best classifying the data, achieving 82 % mean ROC AUC. The runner up is the KNN, which has 76.1 % mean ROC AUC. The last one is Logistic Regression, with about 71.7 % mean ROC AUC. I should note that both logistic regression and KNN took significantly more time to run, resulting error message due to timeout. I also attempted to run a fourth algorithm, Support Vector Machine (SVM) for this project. However, it simply took too long to run and thus was abandoned eventually.

## Conclusion

Using this open dataset from San Francisco Public Library, I attempt to answer a key question that intrigue library professionals: who are the frequent users and what characteristics do they share? Studying the circulation statistics is useful to validate a library’s value to the society, and by using techniques in data science, combined with domain knowledge in library science, I hope to gain insights through both exploratory data analysis and predictive analysis.

My exploration is limited by the granularity of this dataset. For instance, because this dataset only provides the total checkouts over multiple years for each users, it is not possible to explore this data as a time series data. As a result, the number of features (variables) included in this dataset is relatively small. Thus, there is no much need for dimension reduction.

However, through working on this dataset, I did learn invaluable techniques in dealing with imbalanced date. I researched about SMOTE and learned about how to the reason behind oversampling and undersampling. Because imbalanced learning is so prominent, I have no doubt I would be able to apply what I learnt here on other datasets.

In addition, further works could be done to improve the result. I would be interested in learning more about kernel modification methods and use SVM to achieve results. Undersampling the majority class, combining SMOTE with ADAboost, as well as using data cleaning techniques such as Tomek links are some of the other techniques and could be applied by using Imlearn.

Works Cited

Brunvand, Amy. "Fake patrons as civil disobedience." American Libraries, Mar.-Apr. 2017, p. 9. Gale OneFile: CPI.Q, <https://link-gale-com.ezproxy.lib.ryerson.ca/apps/doc/A488605139/CPI?u=rpu_main&sid=CPI&xid=aab27c28>. Accessed 3 Feb. 2020.

Comito, Lauren. “DPL, SFPL Offer Innovative Homeless Services.” Library Journal, vol. 140, no. 17, Oct. 2015, pp. 13–16. EBSCOhost, search.ebscohost.com/login.aspx?direct=true&db=lxh&AN=110530174&site=ehost-live.

Cree, Charity L., and Mijung Yoon. “Public Library Circulation Decreases, Expenditures Rise.” American Libraries, vol. 36, no. 10, Nov. 2005, pp. 57–58. EBSCOhost, search.ebscohost.com/login.aspx?direct=true&db=lxh&AN=18738597&site=ehost-live.

He, Haibo, and E. A. Garcia. "Learning from Imbalanced Data." IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, 2009, pp. 1263-1284.

Krawczyk, Bartosz. "Learning from Imbalanced Data: Open Challenges and Future Directions." Progress in Artificial Intelligence, vol. 5, no. 4, 2016, pp. 221-232.

“Library Circulation Booming in Lake County, Indiana.” Library Journal, vol. 108, no. 9, May 1983, p. 869. EBSCOhost, search.ebscohost.com/login.aspx?direct=true&db=lxh&AN=7572831&site=ehost-live.

Lilienthal, Stephen M. “The Problem Is Not the Homeless.” Library Journal, vol. 136, no. 11, June 2011, pp. 30–34. EBSCOhost, search.ebscohost.com/login.aspx?direct=true&db=lxh&AN=61428299&site=ehost-live.

Luzius, Jeff. “A Look at Circulation Statistics.” Journal of Access Services, vol. 2, no. 4, Jan. 2005, pp. 15–22. EBSCOhost, doi:10.1300/J204v02n04\_03.

Raiteri, Steve. “Graphic Novels.” Library Journal, vol. 131, no. 5, Mar. 2006, p. 58. EBSCOhost, search.ebscohost.com/login.aspx?direct=true&db=lxh&AN=20052151&site=ehost-live.

San Francisco, California (CA) profile: population, maps, real .... Retrieved February 15, 2020, from <http://www.city-data.com/city/San-Francisco-California.html>