# **Popular library in state of North Carolina**

*Robert Nadar*

*03/30/2020*

**Introduction**

1.1 Background. There is a huge effort taken by each states to set aside some budget for all the public libraries. Each year libraries include various activities like “Do it yourself” training, kids activities, infant activities like dance and story book classes. More and more users are getting attracted to public libraries due to fact of more services provided by them. I took this topic because it fascinated me to identify what are the popular libraries in a particular state and what are the factors that determine the popularity of a library

1.2 Problem Data : the availability of public dataset for all the users in North Carolina state , their usages, the book they read, how often they checkout a book, what genre of book they like to borrow, how quickly they can finish the book. We need to get the public library dataset along with the list of all libraries in the state. We also need the Best sellers list, popular books, New York best sellers to build the catalog of books

**Data Section**

Once the problem data is identified it is necessary to identify all the dataset available from various sources. We will start with use of different steps in data science methodology. Now that the business requirement is already laid, it is time to find the approach we are going to take to collect, understand, analyze and prepare the data. The first step includes collecting the data making use of publicly available library dataset, foursquare api to get the location of the library, publicly available popular books dataset like New york best sellers list. We are going to scrap away the PII data and only keep the publicly available data during the stage of data preprocessing. Now we analyze the data make sure what are the individual fields/columns are relevance and remove the unwanted fields. we will convert the text column to int by using transpose method so that we can fit model on the data. During the whole process of analyzing the data, we will make use of different plots to understand the data to its depth

## **Dataset**

1. Libraries in North Carolina
2. Popular books
3. Foursquare API

**Methodology and Analysis**

In this project we will direct our efforts on data collection. data preparation, generating a model, evaluating the model and finally plotting the data in the map to identify the popular library.

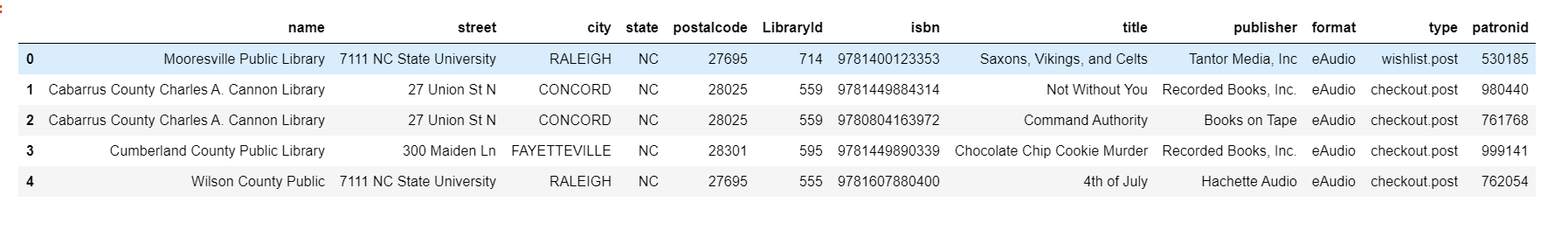
**First step** - we are going to identify the unique postal code of all the libraries and then use the "geolocator" libraries to get the location information like latitude and longitude. this is necessary further to plot the location of the libraries in the map. We are going to do it in a separate data frame to avoid multiple hits to geolocator api thereby avoiding time-out issue. Next, we will merge the location data into the original data frame. We will only consider the relevant columns like name, postal code, longitude, latitude information and ignore all other columns

**Second step** - Once the initial dataset is completed, we will have rough information of popular titles based on the no of checkouts information on each library. we will use that information to find the most popular library with help of dataframe.describe method. Next we will use Foursquare api to find all the venues close to the popular library to identify what are the venues is helpful to decide whether a library is popular or not

**Third step** - We will be creating a model to identify the popular library. Since we need to find out whether a library is popular or not, it falls under category of binary classification(yes/no), so we would be using logistic regression to build a model and evaluate the dataset. We will use different evaluation technique like logloss function, confusion matric etc. to verify the results

In the **fourth** and the **final** step, we will use the map to plot all the popular libraries in the state of North Carolina and use plot markers to mark the libraries with no of checkouts

* **Snapshot of libraries users in State of North Carolina**
* This is the public dataset which we are using for generating the model. This contains all the information of library location and books used by the different users as part of their digital catalog



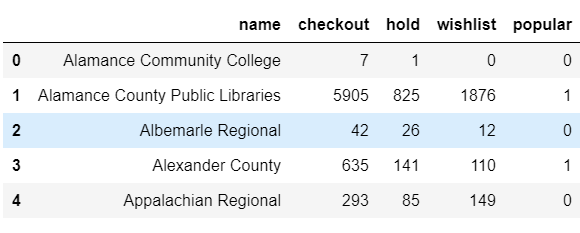
**Result:**

**Final Dataset**

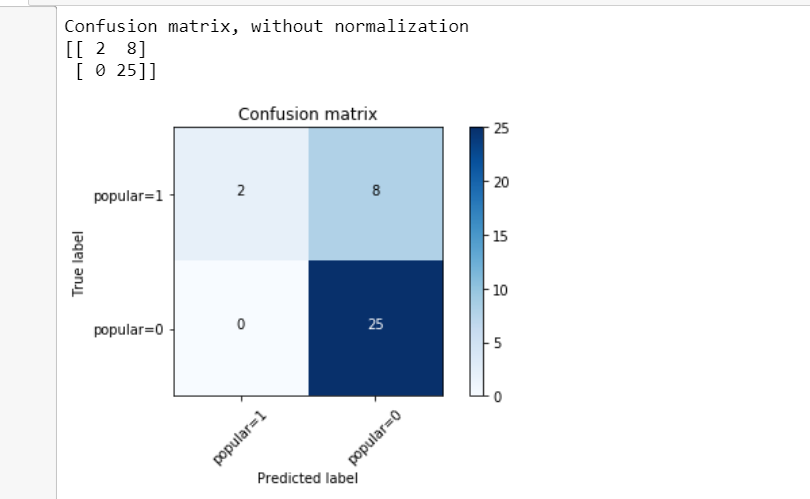
This is the final dataset used for the model evaluation. We will be using StandardScaler preprocessing to bring all the row values close to each other thereby avoiding overfit of the model

we have used logistic regression to build a model and evaluate the dataset. We used the following evaluation techniques

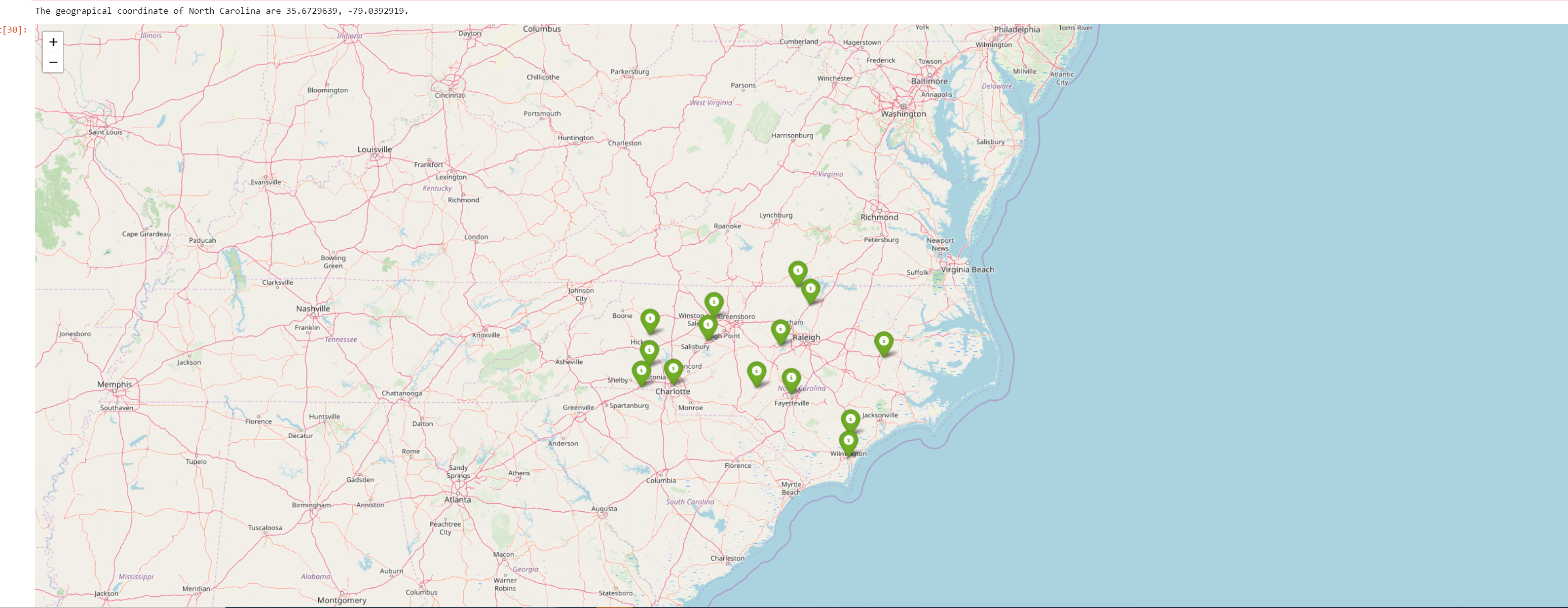
1. Jaccard Index
2. Confusion Matrix
3. Log Loss



**Confusion Matrix**



**Map Plot (Popular Libraries)**

****

**Conclusion**

* Purpose of this project was to identiy the popular libraries in the state of North Carolina. We saw from the dataset that based on no of checkouts information we could analyse and get all the popular libraries , however that was not enough to be an optimal solutions, so we made use of data science and built a logistic regression model to evaluate the popular libraries and we identified there are many factors like nearby venues that determine the popularity of the library. this result can be used by different stakeholders to build a new library in the state of North Carolina.