Logo

Description automatically generated

**Data 2206 – Capstone**

Professor: Thomas Francescutti

What impacts funding

Ontario Public Libraries?

Strong Libraries – Strong Communities

Analysis Report

Student Names

Jaclyn Perron 100456664

Priyanshi Jain 100851063

Chirag Bhanushali 100853897

Zehra Nur Ozer 100761241

November 2022

Contents

[1 Problem Description 2](#_Toc120012927)

[1.1 Business Goal 2](#_Toc120012928)

[1.2 Data Analysis Goal 3](#_Toc120012929)

[2 Data Description 3](#_Toc120012930)

[3 Data Preparation Details 4](#_Toc120012931)

[4 Data Analysis Solution 4](#_Toc120012932)

[5 References 5](#_Toc120012933)

# Executive Summary

focus for this project was to understand the relationship between the funding and other factors collected from various sources and investigate which have the most substantial impact on funding.

Using this information, the team generated a model to forecast the upcoming year’s funding expectations and a dashboard which would display the features we determined had the greatest impact on funding.

First data was gathered from multiple sources included from the libraries themselves, the Ontario data website, and the stats Canada website. We then preformed correlation analysis to determine which factors were the most significant. There were over 70 features with correlation higher than 0.8.

Using these highly correlated features a predictive model was generated which had an underfitting problem. This was addressed by reducing the features further. Then there was an overfitting problem which also required a reduction in features. Our final model was an extra tree regression model with an accuracy of 87%.

The features selected to be displayed in the dashboard were the current funding amounts, the predicted funding amounts, expenses (due to it making up 4 of the top 5 most correlated features), population (due to being the most significance feature to the predictive model) and the ratio of spending on print vs electronic material (due to the main goal of Ontario public libraries being to provide materials).

If implemented this predictive model and dashboard could allow the government to estimate future funding needs for libraries, make business decision such as deciding where to build new libraries based on their predicted funding requirements, and help the government more accurately budget for funding of Ontario libraries.

# Problem Description

## Business Goal

Ontario Government funds the Ontario Public Libraries every year. As the data analysis team, our focus for this project was to understand the relationship between the funding and other factors collected from various sources and investigate which have the most substantial impact on funding. With this information, the team generated a model to forecast the upcoming year’s funding expectations.

As a starting point, in order to get a better understanding of the funding needs of the libraries in Ontario, this project looked to discover the factors which have the most significant impact on the funding requirement of Ontario’s libraries. As the business goal, the team was responsible for answering the questions to the stakeholders as below:

* Public
* Ministry of Heritage, Sport, Tourism and Culture Industries
* Regions, Towns and Municipalities
* Federation of Ontario Public Libraries (Supports Ontario public libraries for research and development, marketing as well as representation to government)
* American Library Association (Provides resources to the public library community)
* Canadian Federation of Library Associations
* The Sound Practices in Library Services to Aboriginal Peoples (Provides information on First Nation public library services and public library services for Indigenous peoples.)

## Data Analysis Goal

This project aimed to explore publicly available information to discover the factors related to funding. We intended to explore the libraries' internal and external factors. Some examples from the data that were investigated are:

* Hours of operation of the libraries,
* Number of cardholders of the libraries,
* Unemployment rate of the city where the library is located,
* Age distribution of the population in the surrounding area and many more.

The purpose of this project is to be able to determine what factors affect the libraries' funding needs. Suppose these factors are determined to be things within the libraries’ control, such as their hours of operation. In that case, we recommend that libraries review their hours of operation to decrease funding requirements. Suppose the factors are determined to be outside of library control, such as the age distribution of the served population. In that case, it can be recommended that new libraries be built in preferred target areas.

Overall, the data analysis goal can be summarized as forecasting future funds for Ontario libraries by analyzing and understanding the data better. To achieve a broader approach, the team decided to add other public datasets, such as census data, crime data, and income data.

# Data Description

The team analyzed the primary dataset from Ontario Government’s public source (Ontario Public Library Statistics) ”2020\_ontario\_public\_library\_statistics\_open\_data,” as can be seen in the appendix in figure 1. This is a 420 KB dataset that is an annual collection of self-reported data from 380 Ontario public libraries with 327 variables. The supplementary dataset used in the analysis were a census profile data, crime rate data, and income data for the same year. This is the official Census Program data as published on the official government website. The Census is distributed every ten years and contains information such as population density by city, age distribution, marital status, etc. The team cleaned all datasets with OpenRefine which a data cleaning software which can help shape our understanding the data.

# Data Preparation Details

Once we had the desired data, we prepared it for the solution by first cleaning the data. The data sets had different naming conventions for the city column which was the column we intended to merge on. This was fixed by separating the columns based on delimiters until the city names were separate and consistence across all the data sets, this was done using OpenRefine.

Another problem with merging on city was that some of the datasets did not have the same number of cities. This would result in a loss of those rows if we were to merge all the data sets into one. To avoid the loss of data we first merged each dataset with the survey data to determine correlation with funding so that any datasets which were not relevant could be excluded from the final data. We ended up with the following data sets:

* A dataset which contained only the survey data
* A dataset which contained the survey data merged with the census data
* A dataset which contained the survey data merged with the income data
* A dataset which contained the survey data merged with the crime data

Then correlation analysis was performed on these 4 datasets in python. All variables which were more than 0.8 correlated with funding were selected and printed into a new dataset which could be further analyzed. This new data set of highly correlated features contained 73 columns from the survey data, 3 columns from the census data, 1 column form the crime data and no columns from the income data.

The data set being mostly made up of features from the survey data was expected because many of the columns in the survey dataset are features which would be clearly correlated, such as the number of staff and the number of librarians, or the number of card holders and the number of materials circulated. For the income it was assumed that lower income would result in more people using free services like libraries so seeing that it was not highly correlated was unexpected, in fact the correlation was very low with a correlation coefficient of only 0.13. From the crime data the column which was correlated was the actual incident column. From the census data the three correlated columns were population, total private dwellings, and total occupied private dwellings.

# Data Analysis Solution

The intention of this analysis was to use the correlated features found to generate a dashboard which would show those features which impact funding and to generate a model which would predict funding needs.

**4.1 The model**

Starting with the model, we began by generating a multivariate regression model and a tree regression model which were both not viable. This was to the models being extremely underfit as evidenced by a negative value being returned for R2. To address this problem, we determined the cause which was the many zero values of many of the features. After removing these features were left with an extremely overfit model. To address this problem, we used extra tree regression which was designed to deal with overfitting. Despite using extra tree, the R2 was still high enough to suggest overfitting so we investigated and reduced the number of features a second time. To lower the R2 enough to get the R2 to drop below 0.99 we had to reduce the number of features to one, population.

This has some advantages and disadvantages. The advantages are that we can rely on population forecasting already being done. There are population forecasting models which already investigate what each cities population might be next year, 10 years from now and so on. If we could lean on forecasting models already in use, it would make our model much more versatile. Instead of guessing or doing simple forecasting to determine what the population of the desired city at the desired time would be, we could take the value for population from existing models. This would result in a more accurate prediction. Another advantage is that with less features the model is less computationally intense which means time and money saved. The disadvantage is that with only one feature the model could become unstable. In the end we decided that these advantages out weighted the disadvantages.

In order to use this model, any value of population can be input, and the model will return a funding value which has an accuracy of 87% and from the MAE interval we can be confident that the actual funding required by a population of that size would be within approximately $17,000 of the value given by the model.

**4.2 The dashboard**

The dashboard consists of 6 main visuals which display important insights about funding and the things which impact funding and can be seen in the appendix in figure 2.

The first visual is yellow text which displays the total funding distributed in the top right of the dashboard. This value updates when cities are selected to show how much funding was given to the specific city in the year of 2021.

The second visual is the bar graph on the far left of the dashboard listing every city which receives library funding in the province of Ontario and the amount of funding they received in 2021 represented by a bar. Clicking on a city or its corresponding bar updates the yellow text at the top left to show the specific value for funding which that city received.

There are two scatter plots in the bottom middle. These two scatter plots show those features which impact funding that we determined were the most important. These were expense and population. Of the top 5 most correlated features found while preforming the correlations, 4 of them were expense related. Population was also very heavily correlated with funding having a correlation coefficient of 0.97. Additionally, population was the feature selected for the final model.

In the bottom right there is a stacked bar chart which displays how much money was spent on electronic materials vs print materials as a way of gaining some insight on where the funding was going. Facilitating equity of access to information and contributing to education literacy and lifelong learning (ontario.ca) is the goal of Ontario Public Libraries, seeing how much is spent on materials which support this goal helps the dashboard display the narrative of the impact of funding on Ontarian communities.

The last visualization in this dashboard is the line graph which displays funding forecasts generated by the extra tree regression model for various populations from 0 to 3 million. This is to aid in understanding the relationship between population and funding and can also be used to estimate funding. This line graph can be used to estimate funding by hovering the mouse over the population you wish to estimate and looking at the value shown for funding for that population.

**4.3 Final thoughts**

This analysis had two purposes the first was to create a dashboard which would display things which impact funding. This was done by finding features correlated with funding and displaying them in visualizations. This was accomplished by creating a dashboard showing two features (expense and population) which are correlated with funding as well as displaying funding information for the previous year as well as the funding forecasts given by the model.

The second purpose was to create a model which would forecast funding. This was done by generating a extra tree regression model. This model can be used to determine funding of Ontario libraries based on the population of the community the library serves. In this way the model can be used to determine future funding needs of libraries by looking at population forecasts for Ontario cities and to determine what the funding needs of new libraries being built would be by looking at the population of the location where the new library would be built.

# Data Analysis Solution

From this project we can conclude that there are many factors which are correlated with library funding. Of those factors we determined population to be the most impactful. By determining population, the government of Ontario can estimate fundings with 87% accuracy. Implementing this dashboard and predictive model would have the following benefits:

* It would allow the government to estimate funding for future libraries
* It would help the government to be aware of the budget amount to timely and accurately fund the libraries
* It would aid the government in making future business decisions for the libraries such as if new libraries should be built in areas where population is growing, if upgrades should be made to existing libraries’ infrastructure and assets
* It would allow the government to track funding, expenses, and population in a single page dashboard.

Implementing this dashboard and predictive model would have the following drawbacks:

* The data is updated yearly which means that the data will fall behind as the year goes on
* The data is in different formats each year making it difficult to use APIs or other tools which would automatically keep the data up to date

Considering these drawbacks, we recommend the implementation of both more frequent data gathering and consistent formatting across the data. In conclusion despite the drawbacks this project would be an asset to the government of Ontario as it would allow for better data driven decisions to be made regarding the libraries which serve Ontarian communities.

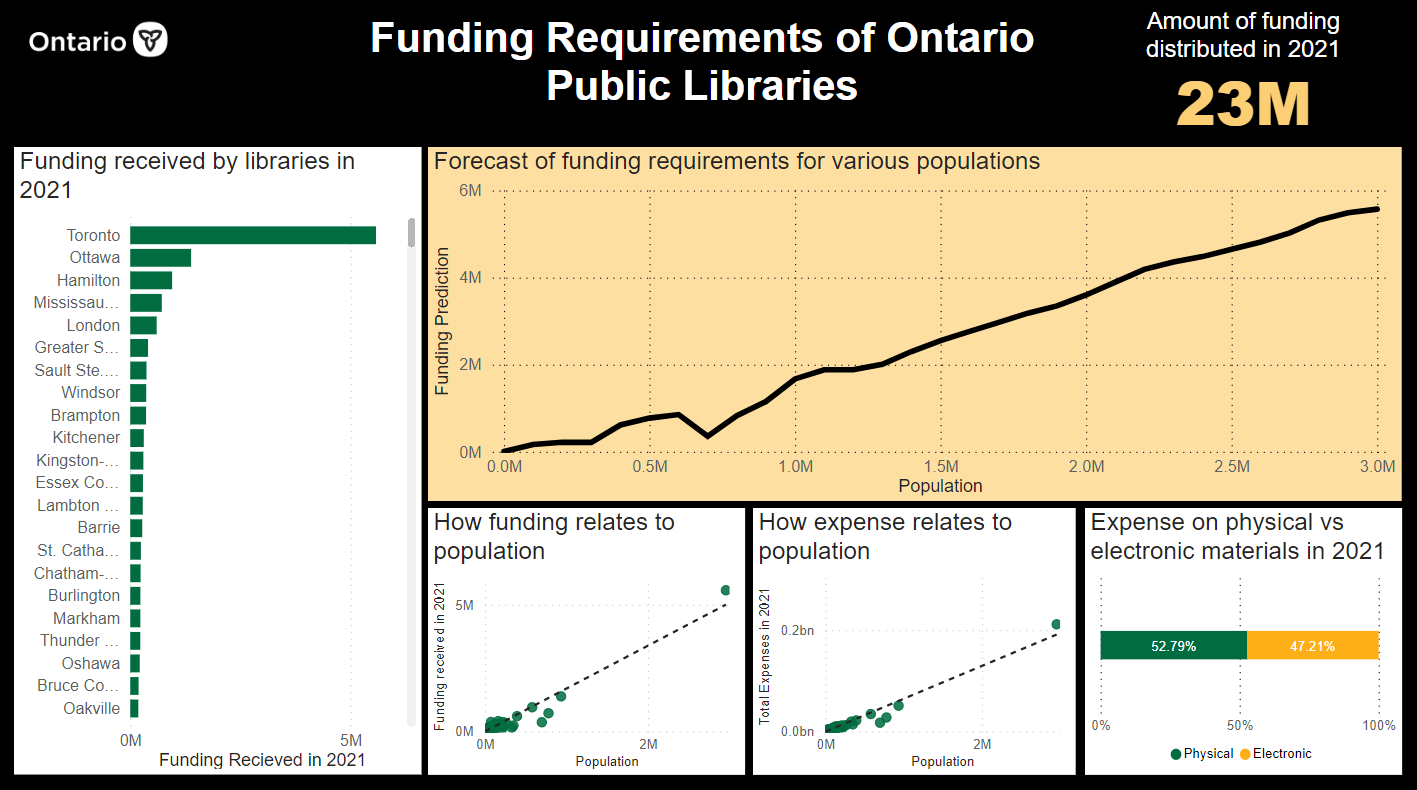
# Appendix

Figure 1 – Survey dataset

Table

Description automatically generated

Figure 2 - Dashboard



# References

(2022, 08 22). Retrieved from Census of Population: https://www12.statcan.gc.ca/census-recensement/pc-eng.cfm

(2022). Retrieved from Ontario public library statistics: https://data.ontario.ca/en/dataset/ontario-public-library-statistics

*ontario.ca*. (2022, 3 10). Retrieved from Ontario public libraries: https://www.ontario.ca/page/ontario-public-libraries