AN ANALYSIS OF DEMAND FOR PUBLIC LIBRARY BOOKS REQUESTED FOR DELIVERY ONLINE BASED UPON GENRE

Technical Report

May 16th 2021

National College of Ireland

Student Name: M. Gerard Smyth, Number: x16473414,

Email address:lynaussa@gmail.com Course: BSc (Hons) in Computing

Specialization: Data Analytics

Declaration Cover Sheet for Project Submission

SECTION 1 Student Details

Name: M. Gerard Smyth
Student ID: x16473414
Supervisor: Noel Cosgrave

SECTION 2 Confirmation of Authorship

The acceptance of your work is subject to your signature on the following declaration:

I confirm that I have read the College statement on plagiarism (summarised overleaf and printed in full in the Student Handbook) and that the work I have submitted for assessment is entirely my own work.

Signature: M. Gerard Smyth Date: 16/05/2021

Table of Contents

1. Executive Summary

This report documents the approach the researcher's project will take in order to answer the question posed. This project aims to address demand issues that may exist within the infrastructure of the public library system in Ireland, starting with localised analysis of the books that are requested for borrowing online from local libraries in order to better see what genres of books have the highest demand for books based upon genre. The dataset is made up of a local library's online lending service, which lists each book and how many times it has been borrowed online. The document intends to assist the reader in understanding what took place in the project and how the approach used helped to complete its various milestones.

December 2019 marked the mid-point of the project, a stage at which many of the techniques that were needed in order to dive further into the datasets, would only be taught in semester two starting February 2020.

Due to various circumstances including the COVID-19 pandemic, May 2021 was the date marking the completion of the project; the proceeding document summarising all of the steps that were taken throughout the duration of this project and the manner of its completion.

Using a dataset provided by the Rush Public library, this project was able to analyse and determine which genre groups of books were the most successful from the genres currently made available by public libraries, and predict which genres would be popular in the coming years.

1 Introduction

1.1 Background

The basis for this project stemmed from anecdotal experience of interacting with the public library system. In many circumstances I found that books that were in particularly high demand and that often the Library didn't seem to respond to the popularity of any but the most popular books in terms of looking to the future of meeting demands within their local readership. This often resulted in particular books that I wanted to read being out of stock quite often, as they were titles of middling commercial success that were nonetheless in high demand from a particular subset of readers in the area.

In many instances the only real way to remedy a lack of titles was to directly place a request for those specific titles that I personally wished to see better stocked or represented within the library itself. Naturally this wasn't a remedy that was ideal and indeed the majority of the time the demand of a single reader was not enough to warrant entirely new titles being added to the shelves of a public library that often did not have the greatest budget to work from. This area has long been one that has been vexing for me personally, having often spent evenings in the local library near my secondary school that as such left me pining for some way of improving the ways in which libraries could meet the demand for their readership on the limited budgets available to the public works.

For a long time I was unaware of how to even begin to approach this solution due to my education lying in computing, but in time my thoughts turned to a different means than "This issue would be more easily solved if the public works' had better funding" to instead ruminating on the possibility of providing information to public libraries about what the demands they were not meeting are, especially as my specialisation in data analysis took root in the fourth year of college. What if I could collect a dataset of loans from local libraries and analyse them in ways that libraries often do not have the means to implement themselves?

Data mining can often be somewhat difficult to implement within the context of a physical medium, especially a locally run public resource. As a result, those local resources being provided a simple and clear means with which to explore what titles or genres are in demand from a readership could majorly improve their ability to predict which titles will be popular in future or what titles are in demand at the moment and may need extra copies added to inventory. This in turn would allow local libraries to not only recognize the demand for a title before it is released, but to provide an argument for better funding should the titles in question be too expensive for the library to acquire.

Hopefully answering these questions will provide insights into what sorts of genres in local libraries may need attention which in turn will allow the local libraries any datasets are taken from a means to tackle the demands of any future titles in the popular if perhaps a little more obscure genres than the massive pop-hits that span less granular descriptors such as “Fantasy” or “Teen-Lit” which are both ever-popular genres that may not provide as much insight as subgenres that may slip between the cracks. Being able to specifically target which of these genres is more popular than they may first appear due to being somewhat more disparate will provide a good baseline with which libraries can better meet the needs of their communities without necessarily having to secure major funding increases - spending wisely instead of spending more.

As the year went on the unfortunate reality of the Covid pandemic made the act of publicly entering a library an impossibility for many people for large swathes of the year, meaning that online lending became more popular than ever. To this end, investigating the trends in online lending became more important than ever - as while the pandemic was an outlier, being able to pinpoint any genres that were being underserved could provide an essential component for providing a physical, free reading experience for people who are stuck at home for any number of reasons. This has certainly provided a unique view of the issue in terms of how to assist libraries, but it is one that will be just as valuable as what titles a library will have in stock for those who walk in following the end of the lockdown.

1.2 Aims

Below is a list of the project’s aims and objectives:

Aim 1: The first aim is to find a dataset of a public library’s list of inventory and lending history for at least a year. This will be limited to a single public library’s inventory rather than several as this is a study to find if this information can provide any insights regarding local libraries to begin with. In order to find the datasets that will be required for this research, the most important body to contact is a number of local libraries that could potentially provide their records, if possible. These records will not contain the individual profiles of those who borrowed each book, but they should contain the statistics of how many times a book has been lended and the genres it contains.

Aim 2: The second aim is to make use of open-source technologies that don't require any kind of buy-in from a user in order for the insights to be found, thus making such programs less prohibitive to end-users in local libraries.

Aim 3: The third aim will be to transform the dataset to pull out information needed for the project e.g. Turnover Rate, Borrows, Genre (where it's possible to do so).

Aim 4: The fourth aim is to dive deep into the data and compare information across each genre and see if there is a particular genre or group of genres that tend to be popular and thus may indicate future popularity.

Aim 5: The fifth aim is to use machine learning algorithms and based on results, see if there is a trend in the genres that are popular and may continue to yield popular titles in the future.

Aim 6: To complete documentation with views on this study, including ideas on how to help improve the ability of local libraries to recognize information about what books may be in a higher demand upon release and use the data found here or in future to make an argument for such instances taking place.

1.3 Technologies

R Studios

The project will make extensive use of the R Studios program to construct this project, which is an open-source integrated development environment for the R Programming Language which itself is a language for statistical computing and graphical visualisation of those statistics. This will be the language that the project is written in. R studios can usually be used in conjunction with the dataset provided by libraries to take in the information contained within and convert it to a format with which the data analysis can take place.

R Language

R language will be used to build this project and display information regarding the various aspects of the dataset, as well as the language used in the transformation step on the data. The libraries being used for this project are Caret, rvest, data_table, lattice and ggplot2.

HTML

HTML is a markup language designed to be displayed on a web browser. Due to the file type of the dataset provided by the library being of a HTML format, this will be the initial type of technology taken in and transformed by R.

1.4 Research

At this point in the project, research was required in order to check and ensure that no similar study had been completed on too similar a topic. After searching through Google Scholar and dataset sites such as Kaggle and UCI depository, I have found that there are no existing studies of this nature to be found.

There are a number of studies that have been conducted regarding IT integration into the Irish public libraries system and the improvement of interlending, but nothing leaning upon the nature of the popularity of genre as a means to predict release popularity and other such predictive elements and far more studies were concerned with academic libraries, rather than public ones. In general the amount of information regarding the level with which studies have been carried out upon the lending trends of libraries is little to nonexistent, which leads to a need to look at studies carried out on a similar types of topics regarding public libraries that overlap with the research being conducted in this study.

Focusing upon the technological element of libraries and how public libraries are meeting the needs of readership, two papers surfaced as the primary reference points. The first of these articles was found on the website academia.edu from as recently as 2020 titled “An intersectional quantitative content analysis of the LGBTQ+ catalogue in Irish public libraries" and provides information regarding the representation of LGBTQ+ content within books stocked by public libraries. This study found that in Ireland usage of libraries by the LGBQT+ were dominated by gay males moreso than any other proportion of that community. It was found that “50% of the sample indicated that their library only sometimes catered to their sexual orientation and a further 29% said that it did not cater to them on this basis at all(Hicks, P. & Kerrigan, P., 2020)”. In addition it also acknowledged that this particular lack of focus was in part due to a lack of published material from these underrepresented groups until recently but were trending upwards. One of the possible solutions to this lack of representation was that “Public libraries can also draw upon community resources already in place, to help with the creation of LGBTQ+ reading lists.(Hicks, P. & Kerrigan, P., 2020)”.

This article makes it abundantly clear that some elements of Public library readership are neglected, and that there were steps required to remedy it. This information was particularly useful in highlighting the factors of how an endeavor to highlight popular genres already existing in a library could help to highlight the ways in which those genres were ignored due to a broader notion of LGBTQ+ being a monolith.

“What effect has technology integration in Irish libraries had on the public librarian’s skillset?”(Kenny, N., 2020) is the name of the second study which can be found on the website esource.dbs.ie. This paper is strongly focused upon the element of integration of information technology within the public library system. This paper references the idea of the integration of information technology being a layered process that required a strong basis of communication among staff. It also generally observed that integration of this

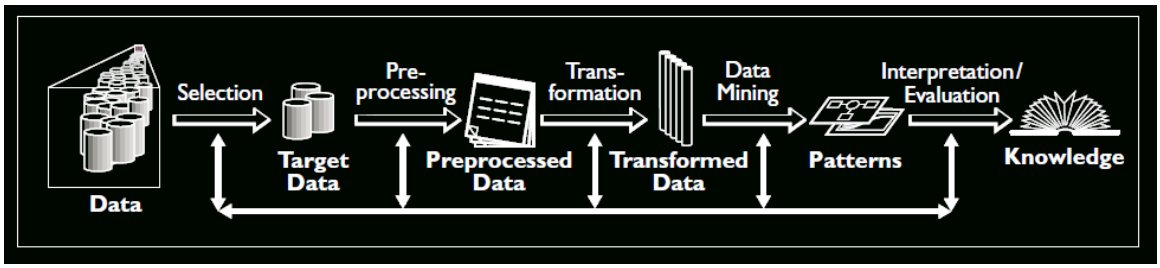
technology was nor inherently linked to a necessity for a younger staff due to several interviews with various individuals in the library network(Kenny, N., 2020). A common thread to be found regarding IT integration was that " a lack of technical expertise does not necessarily diminish a librarian's contribution to the overall service. Other participants largely shared this view, with all of them citing the collaborative nature of operating a public library branch, with different skill sets working in tandem to deliver a strong overall service"(Kenny, N., 2020) and that "participants agreed that given the rate of technological integration in libraries, library staff need to regularly upskill to be adequately familiar with new technologies."(Kenny, N., 2020) with consistent observed factors of a digital divide being down to the lack of up to date hardware and internet access.

The second paper offers a separate insight into one of the challenges that have to be overcome when approaching public libraries: the technological literacy of potential users who are to make use of this study. While RStudio and R are themselves completely free to use - understanding the basics of how it is used and how best to interpret the analysis provided by any functions offered in this study will rely upon the ability of at least a few members of staff in a public library being able to interface with the somewhat layman unfriendly code and run through the steps necessary to make the same findings that this paper does. While this paper does not directly address these concerns, none the less they are important factors to consider when creating programs that require training in this form of technology for the best results. It also presents that these same users are far from incapable of such with the correct technology to support such efforts being in their hands, presenting the opportunity for a strong process of integration of predictive algorithms as a means of improving the supply for an unmet demand.

The preceding research alternatively presents that there is a noted gap in some areas for what libraries are providing and their ability to recognize what needs it is not meeting while also showing that many libraries can make space for the integration of new technologies so long as the means for that integration are properly provided in the form of up to date hardware and internet, as well as upskilling that properly introduces the technology. This will allow for this project to make inferences in the conclusion as to how to better meet needs and how to integrate this free technology with its user base. While acknowledging that due to the nature of the studies that they do not provide strong evidence that the study of genre will allow for demands to be met more successfully, it allows us to at least make a strong case for the notion that needs are not being met and that more proactive methods need to be taken to meet those needs. This will be discussed further in "Conclusions" when the project has been completed, and the results can be properly compared against the findings of the research discussed here or discussed further with regards to integration.

1.5 Structure

The methodology approach I will be using for my project is the knowledge discovery and data mining(KDD). See steps in diagram below:



Shawndra.pbworks.com. (2019).

Selection: the data for my project will come from the public libraries online lending service, provided to the public upon request. Specifically in this project's case, the dataset has been provided by Rush Library(County Dublin).

Pre-processing: during this stage I will be accessing the data from the sources and focusing on cleaning the data to prepare it for the transformation phase and to obtain consistent data. An example of this is that the HTML formatting of the information taken from the online library lending dataset will need to be formatted into contiguous columns and rows to be used properly by the project.

Transformation: This process of the KDD involves pulling the clean and basic dataset and generating better data so that the data frame can properly be used in the next step. The data mining stage methods here include a number of steps such as dimension reduction,feature selection, and extraction, and record sampling, and attribute transformation such as discretization of numerical attributes and functional transformation.

Data Mining: this step the data will be analyzed in Rstudio . Data mining is used to translate problems into effective results. The data mining process will extract useful patterns from my dataset by making use of the Negative Binomial Generalized Linear Model.

Evaluation: This process of the KDD is concluded by taking the mined data and interpreting the results allowing the end user to produce visualizations of the knowledge gained.

1.6 Definitions, Acronyms and Abbreviations

In this section I have placed a list of Definitions, Acronyms and Abbreviations to help the reader better understand what certain parts of the document are referring to.

KDD: Knowledge discovery in databases is the process of discovering useful knowledge from a collection of data. This highly popular data mining technique is a process that involves a series of steps known as data preparation and selection, data transformation/cleansing, incorporation of prior knowledge on data sets, data mining and interpreting accurate solutions from the observed results.

Programming application: A Programming application is a User interface that allows a user to use a set of functions or protocols to do such things as create software or manipulate datasets.

Visualization: Visualisation functions allow a user to visually display and present numerical data, particularly a graphical one that simplifies abstract elements of code for visual consumption. This might include anything from a simple XY graph of one dependent variable against one independent variable to a virtual reality which allows you to fly around the data.

Storage: A state of being kept in a place when not being used: the state of being stored somewhere. Examples of this in the context of computing are a hard drive on a computer or on a cloud storage system which allows you to remotely store your data.

GitHub: A web-based Git repository hosting service. It offers all of the distributed version control and source code management (SCM) functionality of Git as well as adding its own features.

Google Drive: Google Drive is a personal cloud storage service from Google that lets users store and synchronize digital content across computers, laptops and mobile devices, including Androidpowered tablet and smartphone devices

Dropbox: Dropbox is a free cloud storage service for sharing and storing files including codebases, photos and videos. Files can be shared with others by providing them with a link to the dropbox folder containing the relevant files.

Admin: Shorthand for administrator - someone who controls or has elevated access to files and system functions, as is most commonly the case for computing.

UCI: The UCI Machine Learning Repository is a group of databases, domain theories, and data generators that are used extensively by the machine learning community for the empirical analysis of machine learning algorithms and other forms of digital data analysis.

GUI: Graphical User Interface this is the visual aspect of the project, and in my case a dashboard like tableau will be used to provide the visual output.

NBGLM(Negative Binomial Generalized Linear Model): A linear regression model used for the purpose of modeling count variables, usually for overdispersed count outcome variables.

Caret:

2 System

2.1 Requirements

The following section documents all the Requirements that will be necessary for this project to progress, as the project progresses this may grow and changes may happen; if this is to occur an explanation as to why will be provided to explore these changes.

2.1.1 User Requirements Definition

The client requires a study and analysis of books sorted by genre to find correlations between the genre of titles and their population so as to predict if future titles from this genre will also be popular. The objectives will be to acquire datasets from these local libraries and proceed to clean the datasets so that the data left will be relevant information such as title, genre, copies, lending and turnover rates for information to be extrapolated from the following data mining. Datasets should be obtained for free on request from local libraries upon request.

2.1.2 Data requirements

The main data requirement that will be utilised in order to complete this project will be a dataset pertaining to a library's online lending service:

2.1.2.1 The primary dataset is the **siteWideReportByTitle** dataset that has been graciously provided by the Rush local library. This shows a list of all books by title and includes information such as the name of the titles, the number of those titles held by the library and the turnover rate for the books based on the number of copies there are and the number of books that have been borrowed. from The following is a list of attributes provided with the dataset and what they mean:

- Title - The name of the book contained within the library.
- Author - The name of the Author(s) who wrote the book.
- ISBN - The International Standard Book Number(a numeric commercial book identifier which is intended to be unique) the title is associated with.
- Category - A composite of all genres that the book is a part of. This category can contain one or several books depending on how many titles share the same composite genres they are labelled with.

- Fiction/Non-Fiction - Stipulates whether the book is a Fiction or nonFiction(true to life events, people and history) book.
- Genre - A list of genres that the book is ascribed.
- Publisher - The name of the company or individual responsible for publishing the book.
- Released - The date that the book was first released to the public.
- Initially Added On - The date that one or several books of this title was/were first added to the library's current collection.
- Last Added On - The last date upon which one or several books of this title was/were added to the library's current collection.
- Copies - The number of copies that the library currently has.
- Loans - The number of times that the library has loaned a book.
- Reserves - The number of reservations that have been made for a book was currently unavailable due to all copies being out.
- Turnover Rate - The ratio of titles being loaned in comparison to the number of copies available. eg if there are two copies and one copy is borrowed, the turnover rate is 0.5.

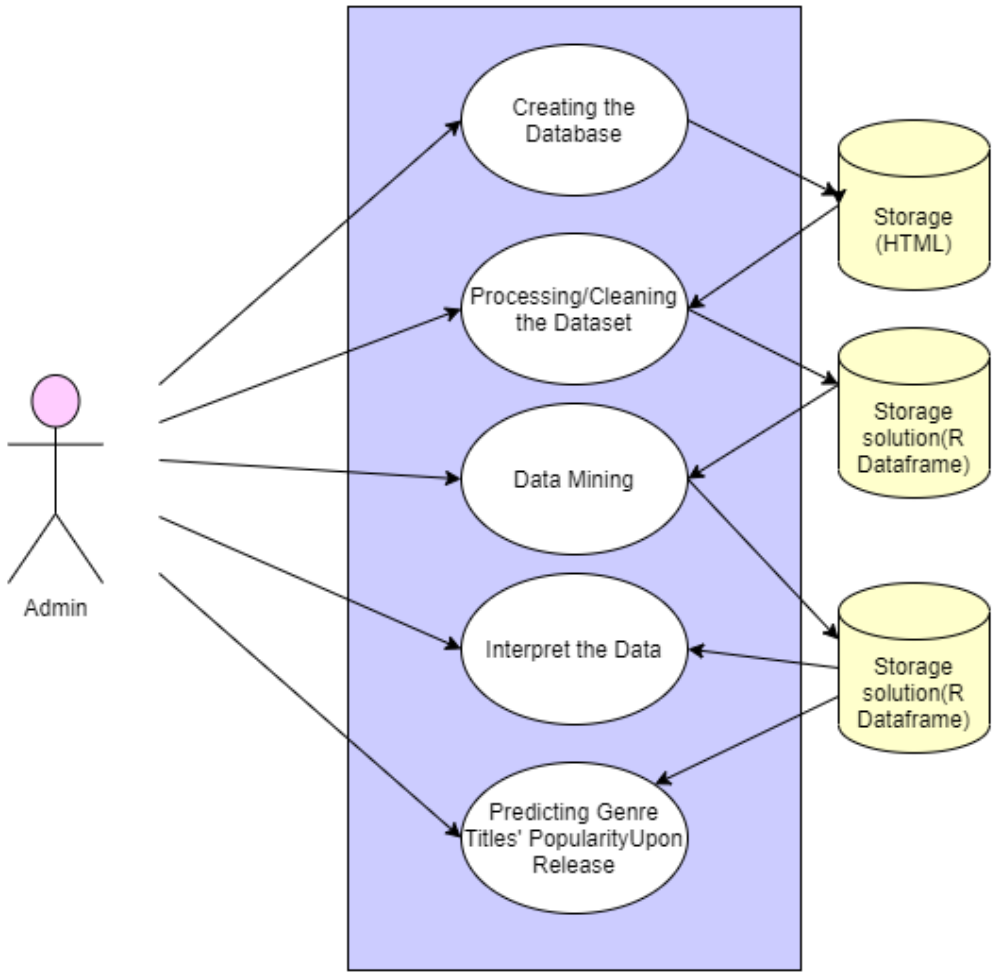
2.1.3 Functional requirements

The functional requirements listed are the ways in which a user will interact with the system so that the study can be concluded. Being the primary user, my functional requirements will describe how I as the user will interact with the various systems in order to conclude this study ass appropriate e.g. The user accesses the programming application, pulls in a dataset from Storage, transforms the dataset to the appropriate dimensions for use in an environment using R, cleans the dataset by dropping attributes that will be not needed and exits the application.

The below use cases are the most important functional requirements for accessing the data that is needed to make my project viable. I will provide the techniques and methods used to achieve these use cases in detail below. The functional Requirements will be listed in order of with which these requirements will be needed to be observed during the project's development.

This section has been significantly altered since the midpoint report, as the functional requirements were effectively split from a single functional requirement into several that were in need of their own use cases and explanations, due to their individual complexities.

Use Case Diagram - Functional Requirements



2.1.3.1 Requirement 1 - Creating a Database

Description & Priority:

This requirement would be considered a level 1 priority and be the first thing that will happen, creating a database is the most important requirement in the study without it there is no place to import the datasets to and from meaning no analysis can take place.

The user

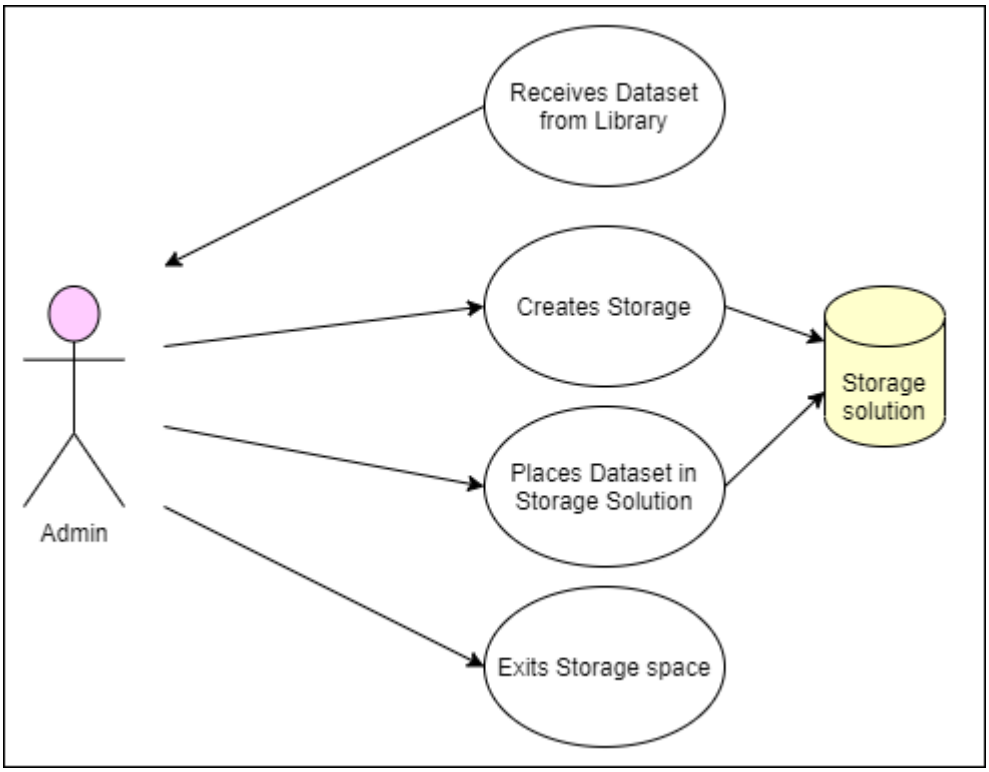
***Use Case**:*

The administrator accesses the application in order to create a Storage which will house all relevant datasets for the duration of this project.

Scope:

The scope of this use case is to create a Storage using a database application; in order to have a location from which I can access all datasets acquired.

Use Case Diagram - Creating a Database



Flow Description:

Precondition:

Dataset must be accessible via the Local Library's correspondence if it is to be placed in the storage created by the admin.

Activation:

This use case starts when the administrator accesses an email or web application and creates a folder within which to store the data.

Main flow:

1. The Admin opens a web application.

2. The Admin creates/selects an existing Storage space.

3. The Admin imports dataset from local directory to Storage using the web/email application.

4. The Admin exits storage location.

5. The exits programming application

Exceptional flow:

<Algorithm error>

1. The system states an error has occurred and that the dataset is corrupt in some fashion upon an attempt to download.

2. The admin checks the dataset saved in the local directory and either finds and corrects the problem or contacts the local library to check if a different version is available.

3. The use case continues from step 3 of the main flow.

Termination:

The Storage is created properly with the dataset imported from the web/email application; the use case is terminated.

Post condition:

The Storage is set properly with the stored data sets imported and waiting to be used.

2.1.3.2 Requirement 2 - Processing the Data

Description & Priority

This requirement would be considered a level 2 priority, as transforming the data so that it can be properly read by the programming language and cleaning the data is essential in order to get most accurate figures and statistics.

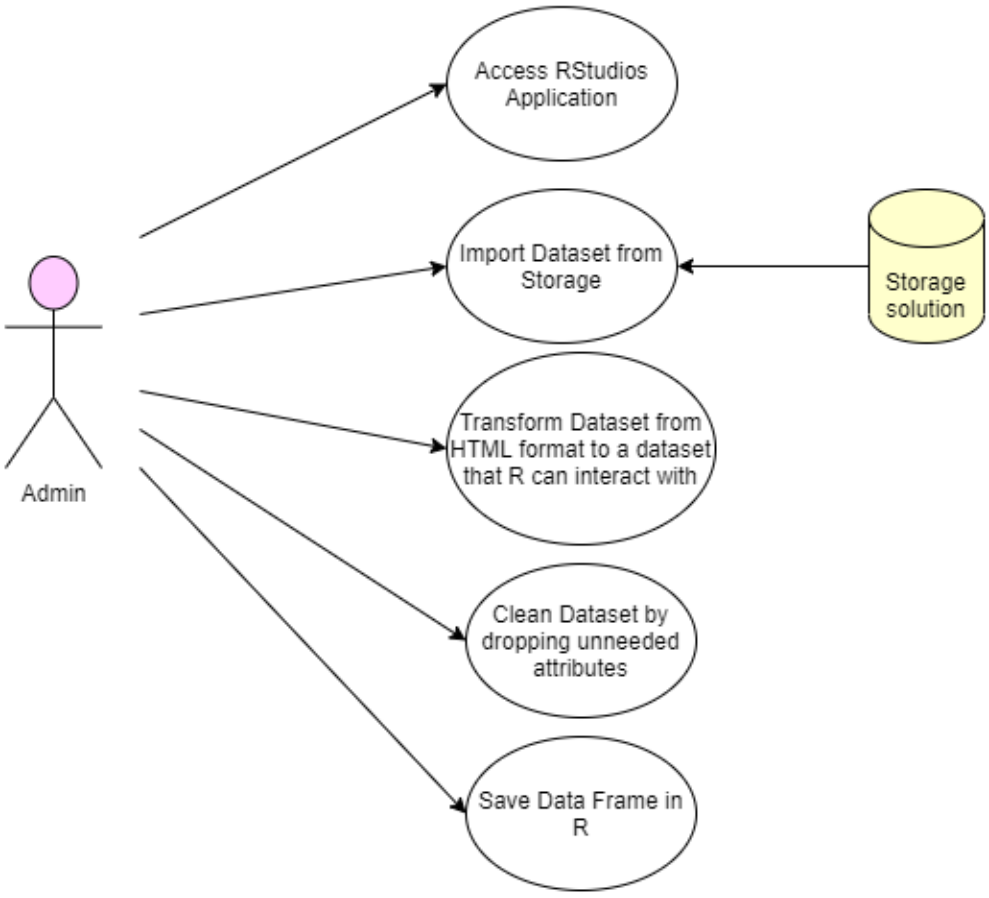
Use Case

The administrator accesses the stored data and transforms the data frame so that it is formatted correctly, before beginning to strip away all of the irrelevant information.

Scope

The scope of this use case is to clean the dataset in order to remove all irrelevant or redundant information, columns or attributes within the dataset that won't relate to this project.

Use Case Diagram - Processing the Data



Flow Description

Precondition:

The dataset is in a wait state prior to being retrieved.

Activation:

This use case begins at the point in which the Admin accesses the Programming Application.

Main flow:

6. The Admin accesses the Programming Application(in this case RStudio).

7. The Admin imports the dataset from the storage space from its HTML format.

8. The Admin transforms the dataset from its prior formatting the information into a list of contiguous columns that can be properly recognized by the R language.

9. The admin cleans the data by removing all unneeded columns, thereby creating a new data set with only attributes that are needed.

10. The admin saves this dataset as a dataframe in R and **optionally** exports it to a CSV file.

11. The admin may now close RStudio.

Exceptional flow:

<Algorithm error>

4. The system is unable to retrieve the dataset.

5. The admin checks all file paths are correct in order to ensure that the application can access the dataset.

6. The use case continues at position 7 of the main flow.

2.1.3.3 Requirement 3 Data Mining

Description & Priority

This requirement has a high priority as Mining the data is required for the whole project to succeed and is essential in finding patterns that exist in the dataset that has been gathered and transformed/cleaned thus far.

Use Case

The administrator calls in the dataset and uses aggregation and model to find correlating patterns within the dataset and documents results.

Scope

The scope of this use case is to mine the data in order to find patterns within the dataset that could provide insights into the data's contents.

Use Case Diagram - Data Mining

Flow Description

Precondition:

The dataset is in a waiting state within the storage.

Activation:

This use case starts when an accesses the Programming application in order to call in the dataset.

Main flow:

12. The Admin accesses the Programming Application

13. The calls an existing transformed and cleaned data frame into the .

14. The uses a regression model on the dataset in order to find patterns

15. The Admin documents all of the results of the data mining.

16. The Admin saves the findings as part of a dataframe.

17. The Admin Exits the Programming Application.

Exceptional flow:

<Algorithm error>

7. The programming application can't process the data to create models, display graphs or other information extrapolated from the dataset.

8. The Admin closes the programming application.

9. The use case continues at the activation stage before the main flow.

Termination:

The dataset mining has been completed and results have been documented fully; the use case is now complete.

Post condition:

The data frame is back in a restful state ready to be accessed as it is needed.

2.1.3.4 Requirement 4: Interpreting the Data and Predicting Genre Titles Upon Release

Description & Priority

This requirement would be considered a lower priority than previous functional requirements, interpreting the data comes at the end and is the explanation given on the results found. All the same, interpretation of the data is vital in order to present the findings in an easy to understand method for consumption by both the Admin and those extant to the study.

Use Case

The admin accesses the stored data frame of information that was mined from the data set and begins to interpret the data visually.

Scope

The scope of this use case is to interpret and to find the most popular titles by Genre so as to predict what future titles of that genre may also be popular using the data mining information previously received.

Use Case Diagram - Interpreting the Data

Flow Description

Precondition

The data frame containing the data mining's output is awaiting retrieval from storage.

Activation

This use case starts when an admin accesses the programming application and calls in the data frame containing the data mining's output from storage.

Main flow.

18. The Admin calls in the data frame created by the output of the data mining.

19. The admin Interprets patterns found in the data mining.

20. The admin documents the interpretations found by using the programming application.

21. The admin interprets the data by comparing the output against the alpha.

22. The admin then sends the data frame back to storage.

Exceptional flow

<Algorithm error>

10. The application is unable to export data.

11. The admin checks data created by the data mining along with any export destinations and fixes the problem that has occurred.

12. The use case continues at position 3 of the main flow.

Termination

The data frame's mined data visualizations have been successfully exported and demonstrated the use case is complete.

Post condition

The data frame is in a restful state in the storage.

2.1.3.5 Non-Functional Requirements

The below section details all of the non-functional requirements that will be part of this study. As with the previous section, multiple changes were made to this section to accommodate for any shifts in focus and to expand upon priorly underdeveloped topics.

2.1.3.6 Performance/Response time requirement

While high performance and response times are seen as a high priority in most system architectures and would benefit this study to have access to with this large dataset it is not fully required for this project. The scope of this project allows the user to analyse the data provided in their own time without any particular time constraints.

2.1.3.7 Recover requirement

Recoverability is a top priority in this project. Making sure that all data is fully recoverable in event of hardware failure or server errors is hugely important for the success of the project. Cloud storage such as Dropbox, GitHub and Google drive will be utilized in order to back up all data relating to this project.

2.1.3.8 Robustness requirement

Robustness is not a requirement for this project. (See2.1.3.7 Recover Requirement.)

2.1.3.9 Security requirement

The raw data retrieved for this project comes from local libraries upon request and therefore does not have a high security protection. The project which will be developed can only be fully accessed from a personal desktop that is fully password protected both on the machine itself and for all of the information that would be accessed via web applications.

2.1.3.10 Reliability requirement

The data set is composed by government entities in the form of aggregated data taken from the lending history of libraries over a number of years. Upon new data being made available or provided new tests will be carried out to find more accurate results.

2.1.3.11 Maintainability requirement

The system is a once-in-once-out design and needs no maintainability once created.

2.1.3.12 Availability requirement

Data will be available to the system at all segments of the project scope.

2.1.3.13 Resource utilization requirement

Hardware such as a desktop/laptop or other computing device will have to be provided along with internet access and backup storage devices, programming and visualization programs will also be needed in order for this study to be carried out.

2.1.3.14 Extendibility requirement

The project could be very easily extended in the future depending on the question that is to be answered regarding the content of the library's inventory but as of now there are no plans for extension.

2.2 Design and Architecture

Describe the design, system architecture and components used. Describe the main algorithms used in the project. (Note use standard mathematical notations if applicable).

An architecture diagram may be useful. In case of a distributed system, it may be useful to describe functions and/or data structures in each component separately. The following diagram shows the Architecture at a high-level view that has been utilized in this project; which is made up of programming applications which will allow for the manipulation of data stacked-on top of a database used to hold and retrieve data sets.

Figure 2 Project Architecture

2.3 Implementation

The main algorithms used in this project have been three forms of penalized regression - which were needed due to the non-integer nature of the dependent variable of the Turnover Rate of books from the library's records, which made use of a number of formulae to find the correlating factors involved in the name of an author. These models were created using the cleaned data.

```
342 samples
134 predictors

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 308, 307, 307, 308, 308, 307, ...
Resampling results across tuning parameters:

  alpha  lambda        RMSE       Rsquared   MAE
  0.1    0.0001955347  0.8125873  0.5754409  0.5654430
  0.1    0.0004517106  0.8120558  0.5756925  0.5651602
  0.1    0.0010435101  0.8050174  0.5792456  0.5613573
  0.1    0.0024106438  0.7915919  0.5857479  0.5546414
  0.1    0.0055688999  0.7723201  0.5935420  0.5449946
  0.1    0.0128648815  0.7581188  0.5951843  0.5430486
  0.1    0.0297195455  0.7774250  0.5722316  0.5582896
  0.1    0.0686560065  0.8715163  0.5019526  0.6130394
  0.1    0.1586042836  1.0187731  0.3994910  0.6991016
  0.1    0.3663964752  1.1537107  0.3358522  0.7821757
  0.2    0.0001955347  0.8105733  0.5774559  0.5640674
  0.2    0.0004517106  0.8092639  0.5782330  0.5632186
  0.2    0.0010435101  0.8001193  0.5839220  0.5579976
  0.2    0.0024106438  0.7829203  0.5935958  0.5483913
  0.2    0.0055688999  0.7562998  0.6097451  0.5335728
  0.2    0.0128648815  0.7349415  0.6197738  0.5271102
  0.2    0.0297195455  0.7485465  0.6110407  0.5345076
  0.2    0.0686560065  0.8529411  0.5514539  0.5911703
  0.2    0.1586042836  1.0015635  0.4709587  0.6914098
  0.2    0.3663964752  1.1597783  0.4235805  0.7882866
```

```
> coef(model$finalModel, model$bestTune$lambda)
135 x 1 sparse Matrix of class "dgCMatrix"
                                                  1
(Intercept)                             9.277881e-01
Category                                .
Fiction                                -2.107619e-01
Released                                .
First_Added                            -2.467675e-05
Last_Added                              .
Copies                                 -3.638524e-01
Loans                                   4.483684e-01
Reserved                               -1.988010e-03
`17th_Century`TRUE                     -7.986347e-02
`18th_Century`TRUE                      3.290637e-01
`19th_Century`TRUE                      8.987527e-02
`20th_Century`TRUE                      .
`21st_Century`TRUE                      .
AncientTRUE                             .
EuropeanTRUE                            .
`General_World History`TRUE             8.486837e-02
Middle_AgesTRUE                        -3.607800e-01
Social_ScienceTRUE                      2.582525e-03
MilitaryTRUE                           -2.690796e-01
`Politics_and Current Affairs`TRUE      .
`World_War I`TRUE                       .
`World_War II`TRUE                      .
HistoricalTRUE                          .
PoliticalTRUE                          -2.133141e-01
Australian_FictionTRUE                 -1.406893e-01
BritishTRUE                             .
Historical_FictionTRUE                  2.931056e-03
War_FictionTRUE                        -1.210378e-01
Earth_ScienceTRUE                       .
ArchitectureTRUE                       -1.380718e-01
EngineeringTRUE                         2.618729e-04
North_AmericanTRUE                     -2.472126e-01
AsianTRUE                               5.410064e-01
BusinessTRUE                            .
AustralianTRUE                          2.581888e-01
CelebrityTRUE                          -1.833290e-01
ClassicalTRUE                          -2.237418e-01
MusicTRUE                              -1.360461e-05
MusicianTRUE                           -3.234933e-03
Performing_ArtsTRUE                     .
International_RelationsTRUE             .
ReligiousTRUE                           1.703072e-01
```

```
> model$bestTune
   alpha     lambda
96     1 0.01286488
```

Figure 2.3.1 - Penalized Regression(ElasticNet)

Additionally a major factor of the project was transforming the largely R unfriendly html based list of books and categories into a format that could be parsed and read correctly by the R language. This takes up a particularly large chunk of the file used to clean the database. This piece of code allows for all of the cleaning of the model and the proceeding modeling to take place.

```
# Define the dimensions of the HTML page
categories_marker <- '<td colspan="12" style=" vertical-align: middle;"><p style="overflow: hidden; text-indent: 0px; ">'
categories <- ''
left_col_css = 'td[style=" vertical-align: middle;background-color: #F0F0F0; "]'
left_col_css_alternate = 'td[style="overflow: hidden; text-indent: 0px; "]'
right_col_css = 'td[style=" vertical-align: middle;background-color: #F0F0F0; text-align: right;"]'
rows <- xml2::read_html(filename) %>% html_nodes('tr')
row_counter <- 0
all_categories <- c()
cat("Creating category list ")
# Proceed to take in the information from the HTML page and add them to each row as appropriate until all of the columns have been processed
for (row in rows) {
  first_cell = row %>% html_node('td:nth-child(2)') %>% html_text %>% as.character()
  if (!is.na(first_cell) & first_cell != 'Title' & first_cell != "") {
    cells <- row %>% html_nodes('td')
    for (cell in cells) {
      if (startsWith(as.character(cell),categories_marker)) {
        categories <-  unlist(
          strsplit(
            html_text(
              cell %>% html_nodes(css = 'p[style="overflow: hidden; text-indent: 0px; "]') %>%
              html_nodes(css='span[style="font-family: Arial, Times New Roman; color: #000000; font-size: 13px; line-height: 1.1499023; font-weight: bold;"]')
            ), ","
          )
        )
        categories <- unlist(lapply(categories,FUN=trimws))
        all_categories <- unique(c(all_categories,categories))
      }
    }
  }
}
```

Figure 2.3.2 - First half of Transformation of HTML file into a readable dataframe

```
for (category in all_categories) {
  books_df[sub(' ','_',sub('&','and',category))] <- as.logical(c())
}
for (row in rows) {
  row_counter <- row_counter + 1
  cat(paste("Processing table row:",row_counter,"of",length(rows),"\n"))
  first_cell = row %>% html_node('td:nth-child(2)') %>% html_text %>% as.character()
  if (row_counter > 11 & !is.na(first_cell) & first_cell != 'Title' & first_cell != "") {
    cells <- row %>% html_nodes('td')
    for (cell in cells) {
      if (startsWith(as.character(cell),categories_marker)) {
        categories <-  unlist(
          strsplit(html_text(
          cell %>% html_nodes(css = 'p[style="overflow: hidden; text-indent: 0px; "]') %>%
            html_nodes(css='span[style="font-family: Arial, Times New Roman; color: #000000; font-size: 13px; line-height: 1.1499023; font-weight: bold;"]')
        ),", "))
      }
    }
    left_cells <- row %>% html_nodes(css=left_col_css)
    left_cells_alternate <- row %>% html_nodes(css=left_col_css_alternate)
    if (length(left_cells) == 10) {
      left_cols <- html_text(left_cells)
      right_cols <- html_text(row %>% html_nodes(css=right_col_css))
      new_row <- data.frame(t(c(left_cols,right_cols)))
      for (category in all_categories) {
        new_row[sub('&','and',category)] <- sub('&','and',category) %in% categories
      }
      colnames(new_row) <- colnames(books_df)
      df_list <- list(books_df,new_row)
      books_df <- rbindlist(df_list)
    } else {
      if (length(left_cells_alternate) == 10) {
        left_cols <- html_text(left_cells_alternate)
        right_cols <- html_text(row %>% html_nodes(css=right_col_css))
        new_row <- data.frame(t(c(left_cols,right_cols)))
        for (category in all_categories) {
          new_row[sub('&','and',category)] <- sub('&','and',category) %in% categories
        }
        colnames(new_row) <- colnames(books_df)
        df_list <- list(books_df,new_row)
        books_df <- rbindlist(df_list)
      }
    }
  }
}
```

Figure 2.3.3 - Second half of Transformation of HTML file into a readable dataframe

2.4 Graphical User Interface (GUI) Layout

No GUI will be introduced to the project at this time. This is a significant change from the midpoint report for this project in which the GUI was planned to be outputted using Tableau, however as it currently stands this project will be a descriptive analysis producing outputs on datasets that have been acquired without a need for an explicit graphical interface for the user.

2.5 Testing

The test carried out in this section is an indicator of how good or valid the data is. As this data analysis project relies heavily on the data being valid, each test case checks if the data behind the function is correct, allowing the end user to know that the data here is of good quality.

The testing tools used in this project were largely speaking the same RStudio tools that were used in designing the project. As RStudio includes a debugging tool, it will inform the user of any errors made. This in turn allowed me to find issues line by line in the process of developing this system. In turn this

2.5.1 **Predictive Models**

Predictive models are used to forecast or predict values whether they be future values of a given variable or what class a certain object falls into e.g. What will the weather be like tomorrow?

2.5.1.1 Test Case 1: Ridge Regression

Success: In order for this to be a success the model must run from start to finish and return a forecasted value within a 95% confidence.

Failure: This test is a failure if the model fails to return a forecasted value. Check Model and parameters of data and test again until success.

Expected Result: A predicted genre that will have a high turnover rate within the 95% confidence interval.

Observed Result: This test returned a predicted genre that will have a high turnover rate within the 95% confidence interval.

2.5.1.2 Test Case 2: Lasso Regression

Success: In order for this to be a success the model must run from start to finish and return a forecasted value within a 95% confidence.

Failure: This test is a failure if the model fails to return a forecasted value. Check Model and parameters of data and test again until success.

Expected Result: A predicted genre that will have a high turnover rate within the 95% confidence interval.

Observed Result: This test returned a predicted genre that will have a high turnover rate within the 95% confidence interval.

2.5.1.3 Rest Case 3: ElasticNet Regression

Success: In order for this to be a success the model must run from start to finish and return a forecasted value within a 95% confidence.

Failure: This test is a failure if the model fails to return a forecasted value. Check Model and parameters of data and test again until success.

Expected Result: A predicted genre that will have a high turnover rate within the 95% confidence interval.

Observed Result: This test returned a predicted genre that will have a high turnover rate within the 95% confidence interval.

2.6 Evaluation

In this section, the project will be reviewed or more so the method applied to the project will be reviewed. The KDD (knowledge discovery in databases) allows a user to select a data source and easily navigate through a structured path in order to come to a conclusion about the data.

While there exist alternative methods that can be used to produce a data analytics project such as SEMMA or CRISP-DM and both offer attractive means to substantially garner

information and results from databases, the KDD approach was chosen for its linear path and ability to clearly outline each step along the way - something which has been highly useful in guiding the process of this study.

The KDD methodology worked very efficiently with the dataset being used throughout this project. It provided an excellent structure that allowed for a definitive guide on where to begin and where to end with the project's trajectory. This left very little room for the project to skew off track during any one of its steps. Each part of the KDD method allowed for goals or milestones to be implemented meaning that a very clear progression could be seen through the project from choosing a data source and cleaning it to mining and displaying the knowledge gained. Overall, the KDD methodology highly suited a linear project of this type which is clearly evidenced by the results it has produced.

3 Results

The results of the three models used for this project will be given subsections of the same experiment as they were all similar, but require elaboration enough as to why each was less useful or provided poorer results in some regards than the other models. In this way the reader can parse each of these models from one another and the results they yielded along with the suitability of the model for this database and study.

3.1 Experiment 1: Penalized Regression

Using three separate Penalized Regression models, this project was able to find the appropriate regression model that best shows the genres of books that will have a high turnover rate in the future with a 95% confidence rate. Unfortunately due to the complexity of these regression models no vizualisation was possible, but the output of each model was interpreted all the same.

3.1.1 Ridge Regression

Whilst producing a prediction within the confidence bounds of 95%, Ridge Regression also produced a far higher RMSE value than Lasso Regression and ElasticNet Regression, which was somewhat indicated by its radically higher positive coefficient output values than those two models and far fewer negative correlations present within the output. In addition it also produced very different predictions to those two models, with very few of the same genres appearing there.

Figure 3.1.1.1 - Ridge RMSE

The figure above shows the high RMSE for the model when training the data via a split, which is very high in comparison to the other two models.

Meaningful Ridge:

EuropeanTRUE	1.364784e-02
Literary_FictionTRUE	6.585372e-03
PhotographyTRUE	3.740704e-02
Visual_ArtsTRUE	4.524884e-02
BiologyTRUE	2.842440e-03
RomanceTRUE	2.375768e-02

Figure 3.1.1.2 - Meaningful Ridge values

The figure above shows the list of gathered results that were outputted by the coefficient that produced a prediction within the confidence bounds, predicting that these genres would be the most likely to have a high turnover rate in the future and thus the genres that should be paid attention to when determining what sections could see an increase in titles, with particular attention paid to Biology. However due to the high RMSE and the radically different outputs, the confidence in this model is somewhat lower than the others.

3.1.2 Lasso Regression

The Lasso regression model provided more consistent results than the Ridge model, with a significantly lower RMSE and what out of the three models could be considered to be a middle point, albeit only by virtue of being the second best fit model for this testing. With a confidence level of 95% the output of positive coefficient values was somewhat more evenly spread and the values associated with those outputs were far closer to the confidence level.

RMSE	Rsquare
0.8037492	0.4253989

Figure 3.1.2.1 - Lasso RMSE

The figure above shows the lower RMSE for the model when training the data via a split, which is much lower than the previous RMSE. This RMSE along with the evenness of the output inspires more confidence in the fit of the model for this problem, albeit slightly less than the final model - though it does provide support in its output and similar RMSE for the final model's own output.

Meaningful Lasso:

Social_ScienceTRUE	6.896515e-03
Historical_FictionTRUE	1.868479e-02
EngineeringTRUE	2.158541e-02
Humorous_FictionTRUE	2.562020e-02
MysteryTRUE	2.996154e-02
Futuristic_AdventureTRUE	8.324364e-13

Figure 3.1.2.2 - Meaningful Lasso values

With this said, the data still put out 6 genres that while shared largely with ElasticNet, are somewhat different. The figure above shows the list of gathered results that were outputted by the coefficient that produced a positive prediction within the confidence bounds of 95%, predicting that these genres would be the most likely to have a high turnover rate in the future. Here Futuristic_Adventure has a particularly interesting value that ElasticNet bears similarities to in its output, with a strong prediction that the genre would provide a high turnover rate for future books added to the library within the genre.

3.1.3 ElasticNet Regression

There is no question that of the three Penalized Regression models used, ElasticNet was the best fit. Not only bearing the lower RMSE, it also bore a number of genre predictions consistent with the Lasso Regression's output again with smoother results. With a 95% confidence level for this test in much the same manner as the other model testing that took place, ElasticNet produced the most meaningful results so far. This in combination with the good RMSE provides a strong indication that these predictions can be taken with the most confidence.

RMSE	Rsquare
0.7996445	0.424696

Figure 3.1.3.1

The figure above shows the low RMSE for the model when training the data via a split, which is the lowest compared to the other two models, but quite a bit closer to Lasso than to the Ridge Regression model. This indicates that it is the best fit model in this instance to an extent, as finding the Penalized Regression model's best fit can be somewhat unreliable and the methods of doing so are advanced and a fairly new topic.

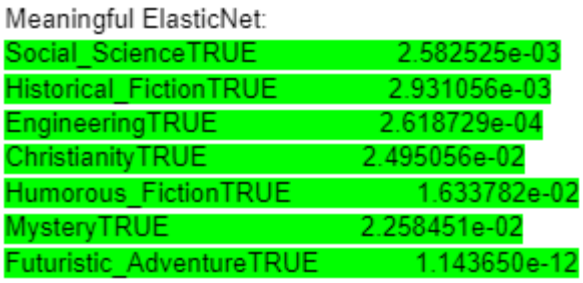


Figure 3.1.3.2 - Meaningful ElasticNet output

The figure above shows the list of gathered results that were outputted by the coefficient that produced positively predictions within the confidence bounds, predicting that these genres would be the most likely to have a high turnover rate in the future and thus the genres that should be paid attention to when determining what sections could see an increase in titles, with particular attention paid to Futuristic_Adventure in much the same way as Lasso, with a notable increase in the value associated with that genre. With that said, the number of genres in common with Lasso can be somewhat helpful in determining that these genres are ones that are particularly notable and worth looking into in terms of their popularity by way of their Turnover Rates.

4 Conclusions

In conclusion, based on the dataset provided by the local Public, this study has found that there is a means to predict the future turnover rate of books in a public library based upon genre. When observing the results of the best fit model's output, it can be observed that the titles that belong to these genres are likely to have a high turnover rate in the future based upon their genre.

While many of the genres of these books are somewhat to be expected with the likes of Social Science, Historical Fiction, Engineering and Mystery all being expected, the popularity of humorous literature and Futuristic Adventure while not unprecedented, may be a surprise as far libraries are concerned, and more attention being paid to these genres would provide some level of relief for groups seeking out these books but may not be having their demand be met.

In particular as the research in the research section showed, social science books are in high demand, but only a small amount of these actually reflect a number of broader LGBTQ+ community members, meaning that in turn this demand may not be getting met in the best way possible by the library due to a simple lack of knowledge that the demand even exists.

Some results from this test may be a little too broad, as futuristic adventure can cover genres from hard science-fiction to urban science-fiction and may encompass a larger group of books than even the others might.

Additionally some of these predictions are not as helpful in showing the prediction of future genre turnover rates for books but can equally provide an insight as to why books of that type should be consistently updated. Biology in particular is a matter always seeing progress and as such a library being made aware that these kinds of books will continue to have a high turnover rate is proof that there is a need to pay close attention to the matter and ensure that out of date books on biology are removed or replaced by up to date volumes.

Whilst interesting results have been found in the current research, it must be said that this project was limited by the level of access that was given to library records. Whilst the access allowed provided an answer as to what the most popular genres were within an online borrowing setting, unrestricted access to the entire local library's database including

physical records would have allowed for a much more intensive amount of research to be conducted about the turnover rate for walk-in physical borrowings.

5 Further development or research

This application works as a strong bedrock that can be expanded upon in a few ways.The potential evolution of this tool would be to expand the data that would be observed. Patterns of borrowing from certain users could provide insights into what kinds of books that would make for good recommendations for them or if perhaps their needs as a reader aren't being met. Furthermore attempted borrowings could be recorded, which would allow for a demand for books that are currently out of stock to be more accurately tallied.

At the current time, the limited availability of properly stratified library data is the biggest limiting factor upon the development of more sophisticated processes for finding patterns in borrowing patterns of users. Gaining access to the number of books being physically accessed in local libraries rather than ones accessed via an online tool could provide a much more sophisticated web of information across larger spaces. Additionally this could lead to a larger web of information shared between libraries via information gleaned from this research and allow for better intersectionality between libraries that could lead to a need for even buying as many books as before, but rather meeting demand with intersectionality between libraries in a proactive manner.

6 References

6.1 R:The R Project for Statistical ComputingR-project.org. (2019). R: The R Project for Statistical Computing. [online] Available at: https://www.r-project.org/ [Accessed 26 Nov. 2019].

6.2 Shawndra.pbworks.com. (2019). [online] Available at: http://shawndra.pbworks.com/f/The%20KDD%20process%20for%20extracting%20useful%20knowledge%20from%20volumes%20of%20data.pdf [Accessed 5 Dec. 2019].

6.3 MySQL :: MySQL 5.7 Reference Manual :: 1.3.1 What is MySQL? Dev.mysql.com. (2019). MySQL :: MySQL 5.7 Reference Manual :: 1.3.1 What is MySQL?. [online] Available at: https://dev.mysql.com/doc/refman/5.7/en/what-is-mysql.html [Accessed 6 Dec. 2019].

6.4 Kenny, N. What effect has technology integration in Irish libraries had on the public librarian's skillset?.(2020) [online] Available at: https://esource.dbs.ie/handle/10788/4181 [Accessed January 2020].

6.5 Hicks, P. & Kerrigan, P. An intersectional quantitative content analysis of the LGBTQ+ catalogue in Irish public libraries.(2020) [online] Available at: https://www.academia.edu/41749178/An_Intersectional_Quantitative_Content_Analysis_of_the_LGBTQ_Catalogue_in_Irish_Public_Libraries [Accessed January 2020].

6.6 Kuhn, M. The Caret Package.(2019) [online] Available at: https://topepo.github.io/caret/ [Accessed March 2021].

6.7

7 Appendix

7.1 Project Proposal

7.1.1 Objectives

The goal of this project is to create an application that can provide astute recommendations of books that are underrepresented in the library based on the categories of book title, author and section. This application that takes datasets of books borrowed from the library and measures interest based on the time they spent borrowed, over the course of a number of years. Using data processing this measure of interest can then be studied to see if the demand of borrowers is being met and to what degree it is. The user interface in turn should be not only be accessible (with minimal training required bar the input of the dataset itself) but sport an attractive visual output as well.

To summarise, the objectives are:

- To create a simple output of information based on borrowing habits.
- To provide a form of automated feedback for libraries.
- To create a system that is easy to use and requires few file inputs.
- To provide a program that requires relatively little technical training to use.
- To provide a clean and simple interface that is appealing.

7.1.2 Background

The core of the idea for this project came from a discussion surrounding book availability in local libraries. Often libraries will juggle some books between them whenever supply is short, a fall-back to cut spending if the demand of the readership is high. While this can sometimes be down to trends, often times there may not be enough copies of a book or other provided resource even between these libraries. From this there arose the idea of a program that can provide concrete insights into borrowing trends that will enable libraries to build evidence that a book, author or section is in high demand that is not being properly met.

7.1.3 Technical Approach

Development

For the development of my project I will be using R & python, which I will use for transformative purposes. From there I can organise that data into a data frame in R and run a data analysis on it with my data model. I may use MySQL for a database and I will be using Tableau for data visualisation.

Implementation

For implementation of the project I will need to adapt the dataset I currently have from the selected library so as it accurately reflects the times and number of volumes that are

being borrowed at any given time from the library. In addition, a full listing of books in the library will need to be acquired in order to compare one set of data against the other to determine just how many of a book or author the library has.

1. ***Project Management***

The approach I'm most comfortable with when it comes to project management would be a scrum methodology with weekly/twice weekly tasks that must be completed by the end of the week. Review what went well, what didn't and implement that back into the planning process.

7.1.4 Special Resources Required

2. ***Software***

The software I will need to use is R with Rstudio, Python, MySQL and Tableau, for which I have acquired a year-long student license.

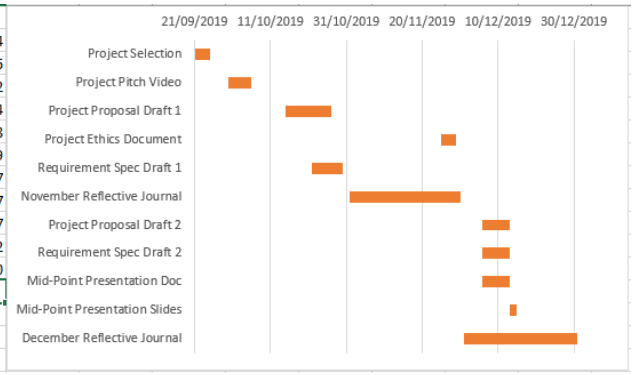
3. ***Hardware***

No specific hardware needed for this project, and as such the current desktop I have available will be sufficient enough to process any solutions for this project.

7.1.5 Project Plan

Below can be seen a Gantt chart laying out the activities that have been carried out and have yet to be carried out as of the 13/12/2019. Most of these activities were successful, though the drafts were somewhat undercooked in the early stages.

Task ID	Task	Start	End	Days
1	Project Selection	21/09/2019	25/09/2019	4
2	Project Pitch Video	30/09/2019	06/10/2019	6
3	Project Proposal Draft 1	15/10/2019	27/10/2019	12
4	Project Ethics Document	25/11/2019	29/11/2019	4
5	Requirement Spec Draft 1	22/10/2019	30/10/2019	8
6	November Reflective Journal	01/11/2019	30/11/2019	29
7	Project Proposal Draft 2	06/12/2019	13/12/2019	7
8	Requirement Spec Draft 2	06/12/2019	13/12/2019	7
9	Mid-Point Presentation Doc	06/12/2019	13/12/2019	7
10	Mid-Point Presentation Slides	13/12/2019	15/12/2019	2
11	December Reflective Journal	01/12/2019	31/12/2019	30



This project plan has been a useful tool for tracking progress(and the periods with a lack of it) and as such I will be using another chart over the Christmas break and come the second semester.

7.1.6 Technical Details

Proposed Technologies:

- R Language & Rstudio
- Python

- MySQL
- Tableau

7.1.7 Evaluation

There will be system tests conducted using the first dataset and then a second dataset that will be plugged into the application to make sure that the implementation of the application is working correctly.

For the purposes of testing with an end-user, I am hoping that I can provide the application to the NCI library and request they test it using some information they compile based on the format I recommend to them. For this, I am hoping that this feedback can be anonymously performed in-house in the library, thus ensuring feedback is given by someone knowledgeable enough to be able to utilise the program.

7.2 Monthly Journals / Meeting Minutes

7.2.1 January 2020

Student name: Michael Smyth - x16473414

Programme (e.g., BSc in Computing) Month: January

2. My Achievements:

I was guided in looking at alternative solutions for developing my application with the use of R and Python to be used. This is due to the relative ease in which a member of library staff can be instructed in the use of the application in addition to the open source nature of the technology which will allow for full accessibility for libraries without spending on expensive licenses such as those for Tableau - something which could provide a heavy obstacle for strictly budgeted libraries.

3. My Reflection:

This choice of software has helped steer the project in a new direction that will be time consuming to work on, but ultimately more beneficial in the long run. However necessary it may have been, dropping the focus on Tableau has been a blow to the time spent learning to use the program, with a month or so of study having been lost to this change in direction.

The changeover came at a not so convenient time with the lack of guidance over the winter break and the lateness in the month from which classes resumed. For this reason a great deal of work will have to be done to make up the difference in February where a strong bedrock of work on this new direction should be laid down.

4. Intended Changes:

The changes this month were not entirely extensive due to the work to be done on exams, largely speaking the changes to the project were in direction. More extensive changes are expected next month to coincide with the changeover from tableau to a fully R based application getting off the ground. The intended changes for next month are that there will be a skeleton for the new application created and discussed in supervisor meetings.

5. Supervisor Meetings:

N/A. Due to the lack of classes in the month there was not much time to organise meetings.

7.2.2 August 2020

Student name: Michael Smyth - x16473414

Programme (e.g., BSc in Computing) Month: August

6. My Achievements:

As per the previous reflective journal entry I had been directed to acquiring borrowing records from local libraries, as college library borrowing records were not indicative of the borrowings of the general public. Due to the lockdown that took place through a large portion of the Summer I was unfortunately unable to continue work on the project - largely due to the lack of resources that could be acquired from the library. This changed in late July, when I was able to once again contact the library and speak to them about acquiring records for the project. After some time exchanging information I acquired the records pertaining to online borrowing - as unfortunately I could not be given physical records for Rush, Malahide or Skerries libraries.

7. My Reflection:

While the acquisition of the data is a good step, the data itself is unfortunately in a HTML format and as such is not usable for an R program to take information for investigation of its components. With this in mind some way to parse this information without the data having to be inputted by hand will need to be developed. This, along with the fact that these records are of online borrowings and not in-library borrowings was an unforeseen issue that will take time to address, but ultimately these are roadblocks that can be accounted for and worked around.

8. Intended Changes:

Changes this month were limited due to there needing to be a way to parse the information within the HTML record and reformat it into an excel file that can then be properly looked into, though having taken the records and having placed them in an excel file by hand, I intend to look into further ways to develop the program using the data I already have.

7.2.3 January 2021

Student name: Michael Smyth - x16473414

Programme (e.g., BSc in Computing) Month: January 2021

9. My Achievements:

With the library data having been acquired in full a number of issues arose regarding the type of data that it was. The formatting was not of the correct format for this data to be processed for data mining and as such was in need of transformation. This month work was done on looking into solutions for such a transformation.

The path to this point has been rather rocky due to the challenges of full-time work which has lead to difficulties in networking with my supervisor, and in turn the online resources for working with various aspects of R and the classifiers I want to explore the usage of will be are somewhat limited and bogged down in jargon that can be hard to decipher or immerse oneself in.

10. My Reflection:

Due to the limited scope of the data provided by the library, there will be a number of issues in exploring the dataset properly. This is due to a complication:

As this data provided only accounts for physical copies borrowed via an online service rather than the usual physical walk-in, walk-out system that many libraries would make the majority of their lending through, this data is not complete for the purposes of studying the lending trends of local libraries. As such, this project will have to pivot to provide a more specialist view of borrowing habits online, which may provide a different view.

Intended Changes:

The change in direction this month was largely to the target of the project itself, once again pivoting from a local brick and mortar library to instead the Irish online lending services provided online from the past two years.

7.2.4 February 2021

Student name: Michael Smyth - x16473414

Programme (e.g., BSc in Computing) Month: February 2021

11. My Achievements:

An R program for transforming the data has been fixed up and used so as to make using data mining techniques on this dataset viable within the framework of an R program.

12. My Reflection:

Continued difficulties with R on my personal system and the lack of viability to use college machines for such an endeavor meant the progress was slow for properly testing the program that converted the data frame to a usable format for data mining. Largely speaking R has been a difficult language to interface with in some regards, seemingly more ideal for physical feedback in a controlled environment rather than the digital space student-supervisor interaction has been relegated to due to the outbreak.

13. Intended Changes:

Continue to convert the database into a more model-friendly layout, including dummy-encoding any variables that are currently stored as characters rather than integers.

14. Supervisor Meetings:

Due to the transient nature of work weeks and digital workspaces, as usual a number of smaller interactions took place over the month of March, in which pointers and feedback was given on various elements of the project including what classifiers were best to use along with help in creating a functional system that allowed the information within the data frame to be parsed.

7.2.5 March 2021

Student name: Michael Smyth - x16473414

Programme (e.g., BSc in Computing) Month: March 2021

15. **My Achievements:**

Having been guided to take a look at using classifiers or regression models for the purposes of the data mining needed for my project I have proceeded to begin investigating which types of models will best fit my data.

16. My Reflection:

The testing of different classifiers and regression models based on the data frame available and the predictions I hope to make for the project has been one that was a long time coming but was impeded by the difficulty of transforming the data.

17. Intended Changes:

Testing Classifiers and regression - I hope to test Naive Bayes, Multiple Linear Regression and Negative binomial generalized linear model.

18. Supervisor Meetings:

Due to the transient nature of work weeks and digital workspaces, as usual a number of smaller interactions took place over the month of March, in which pointers and feedback was given on various elements of the project including what classifiers were best to use along with help in creating a functional system that allowed the information within the data frame to be parsed.

7.2.6 April 2021

Student name: Michael Smyth - x16473414

Programme (e.g., BSc in Computing) Month: April 2021

19. **My Achievements:**

A regression model rather than a classification has been settled upon as the primary modelling to be undertaken on the data to be used in this project, which will allow for exploration of the data(limited as it may be) in order to find which elements are the most revealing.

20. My Reflection:

With the data further cleaned and reduced to numerical factors within the dataset, regression from hereon will be very simple to test and allow for headway to be made with testing each one on the datasets. Already a Negative binomial generalized linear model has been ruled out, as the data for the dependent variable(Turnover Rate) contains non-Integers.

21. Intended Changes:

The changes this month included working on getting an appropriate model for the study that has been transformed for viable use within the project.The findings from this classifier will move the project close to its completion and provide some form of conclusion on the viability of exploring library records with data mining.

22. Supervisor Meetings:

Due to the transient nature of work weeks and digital workspaces, as usual a number of smaller interactions took place over the month of March, in which pointers and feedback was given on various elements of the project including what classifiers were best to use along with help in creating a functional system that allowed the information within the data frame to be parsed.

7.3 Other Material Used

Any other reference material used in the project for example evaluation surveys etc.