US
University of Sussex

Data Analytics & Predictive Modelling for Library Occupancy

Author: **Andreas Kyratzis (196980)**
Technical Supervisor: **Enrico Scalas**

Master of Science in Data Science

School of Mathematical and Physical Sciences

University of Sussex

August 2019

Abstract

This research project focuses on understanding patterns of occupancy of the Library building of the University of Sussex by using Data Science Research methods. The first objective of this project is to uncover data insights through the use of Exploratory Data Analysis techniques using a dataset containing the individual logs of people entering and leaving the library building. Through data cleaning and pre-processing and then through the use of plots, the patterns of occupancy are shown, and hypotheses are tested.

The second objective of this research is on using Predictive Modelling approaches to forecast the timeline of the Library occupancy. For this task, ARIMA, SARIMA and Random Forest modelling procedures have been used and evaluated to select one model as the best fit model for this dataset.

Acknowledgments

Firstly, I would like to express my gratitude and appreciation to my supervisor, Dr Enrico Scalas, for the critique he has provided during the several stages of this project and feedback that has allowed me to improve various parts of the project.

I would also like to thank my friends and family for being my side through this whole journey by showing their love, their support and trust through my educational path and through this project.

Lastly, I would like to thank Antony Groves for providing me with the dataset that was used for the project and for his notes on using and understanding the data of the Library.

Table of Contents

Table of Figures

1. Introduction

Universities have used libraries through the years to enable students as well as faculty to receive information that they would find applicable to be used for their educational needs as well as for inspiration to future writings and discoveries using a material that would otherwise be difficult to find.

Due to the age of the internet, universities have created new perspectives of a library; through online databases of the material that would normally be based on a physical library. While online libraries have attracted most of the original audience of students and faculty of universities, physical libraries remain a big part of university life and people tend to attend libraries for various purposes. Some of the reasons that people still use libraries are the physical books and publications available to them for either in house reading or to take home renting which enables people to take the material outside of university for a small amount of time. Other purposes of visiting the library building include the use of the silent reading areas where students and faculty/staff can read or work in silence. Furthermore, the library provides computers that could be used for any kind of work needed to be done.

To enter and leave the library building, an access card is required which has to be scanned for access to and out of the building to be given. For any internal students or faculty, the uni card is the one used. By using this pass mechanism, the library IT department can collect data of all In and Out logs of the building. Since the introduction of this data keeping method, the library has attracted over 19 million columns of data in their dataset. Having this big amount of data available, the insights that could be uncovered if the data was analysed could enable the university to make adjustments to the library operations. This kind of adjustments would be towards providing customer satisfaction for students and faculty as well as any other kind of library attendees.

To get this kind of insights, the IT staff of the library has provided the dataset of the entry/exit stats that will be used for Exploratory Data Analysis as well as Predictive Modelling for the forecasting of future occupancy of the library. The data analysis as well as the predictive modelling techniques used for this project answer questions that can help the library staff understand how the past, present and the future of the library is shaped and how attendees attend and what their backgrounds are.

Some of the basic questions answered through this research are: “What kind of students tend to visit the library the most?”, “What School spends the most time in the library?”, “Is it necessary to have the library open 24/7?”. Getting the overall time spend from each particular group of people is a highlight of this research as it can show various patterns that could be overlooked by a naked eye. Hypothesis based on the findings of plotting the various results shows the correlation between the school of study and time spend in the library and how different each school visits and spends time in the library.

The predictive modelling techniques used can also provide the forecasted occupancy of the library using the training data available. To provide a forecasted timeline of the occupancy of the library, the use of time series forecasting was considered since the occupancy is correlated with the date and time which meant that each day and time may have a different occupancy rate than any other day or time of the day.

To effectively predict the timeline of the occupancy of the library, three models were considered to be used. Having strong connections with time-series predictions as we are going to examine later on based on literature, ARIMA modelling, SARIMA modelling and Random Forest modelling were all used and will be evaluated in this research on how well they can forecast the future occupancy through evaluation methods.

Taking on all of the aforesaid results of this research, the Library staff can work on improving their services based firstly on the customer’s needs and secondly on the Library’s own needs.

2. Literature Review

When looking at this research title, the greater cause is to get insights through data of the library building. At the grand scheme of Data Analytics usages in library domains, numerous data analytics methods have been put in use to get insights of library data through the years and a research that puts emphasis on understanding how libraries operate and invest in various resources or scenarios through data analysis and manipulation is ‘Library assessment and data analytics in the big data era: Practice and policies’. This research has helped understand and put emphasis on the ethical implications of working with data of such high volume as well as how libraries require certain data assessments to be able to use their resources to their best capabilities. As the research suggests there are challenges to using big data and there are still areas where the use of big data at scale for analysis is very hard as the results of the analysis may provide wrongful insights. The study focuses on integrating behavioural user studies in understanding library usage which has helped our research project in understanding patterns that do not fall right into the visualisations of the analytical steps taken in our Exploratory Data Analysis [1,2].

As the aim of the Library of the University of Sussex is to improve its services and to provide customer satisfaction to a greater degree, this research project had to examine ways of using the data-driven resources that the library has been using, to improve their services [3]. Applying data-driven tools for capturing information of behavioural nature is a challenging process. The analytical steps taken to produce insights that are worth looking at need to have certain characteristics such as ease of understanding or numerical values that can be used to easily capture insights. A study of the online use of a library in Singapore and the procedures that were taken to understand insights has helped our research produce trustful visualisations that are easily understood and that provide insights that can help the library understand its occupancy rates from the different groups of people interacting with it. This study has also shared how Predictive Modelling can be used in this data domain to produce forecasts [4].

Predictive modelling on occupancy rates is an area of work that has been evolving over the years through various researches. As predicting occupancy rates at any kind of venue or building can provide solutions to problems such as overcrowding or security breaches or panic situations is very important and very beneficial to the certain causes, many pieces of research have emphasized providing correct predictions to produce forecasts of occupancy rates for their particular venues of interest. There are two parts of literature that focus heavily on understanding patterns of occupancy and predicting occupancy rates in the hotel industry and as it is based on physical occupancy rather than online occupancy rates, it has provided additional expertise on understanding and using Machine Learning approaches to predict accurate forecasts of time series data as the data examined is for hotel room occupancy levels and is of similar characteristics to the library data used in our research. [5,6].

As this particular research project has already been examined by another student, Marhioudaki (2018), the focus is in understanding what has already been done correctly or not and improve further on understanding the patterns that we are interested in seeing through the use of the library data sets. As seen in the previously worked on project, the emphasis was on using exponential smoothing techniques to predict the occupancy forecast but as seen in the project the results were not as good as it was needed to effectively predict the occupancy rates. The use of ARIMA model has shown a good understanding of the dataset but as the data used in the project was not pre-processed to a good extent, the results lacked accuracy. This research project has overall enabled our project to use extensions of the basic ARIMA model to create better forecasts that accurately represent all of the characteristics of our data set. [7]. As exponential smoothing approaches in time series forecasting have been used for a while in both economics as well as finance sectors, there were several examples on their use and their model predictions to understand whether or not the use of a model such as Holt-Winters was a good fit for this research area. As seen through literature, ARIMA has more parameters which enable a greater degree of accuracy to its predictions while the Holt-Winters approach only has three parameters. Most researches on the use of the two have shown that the predictions of the two models share good accuracy overall. A great example of the use of the Holt-Winters method in library data shows how the algorithm can predict the occupancy levels of another library in order to use the correct amount of staff for several time periods in the library [8]. However, taking a look at the results of Mathioudaki's use of Holt-Winters and ARIMA, it was decided to enhance the ARIMA model that she has used to use seasonal characteristics which in return would provide lower prediction errors. [7,9]. 'Some Aspect of modelling and forecasting multivariate time series' is a resource that has been used to appropriately see how each different component of a time series plays a role on the performance of the model used [10]. Understanding the practicality of time series modelling has also helped in choosing the correct models to test for their predictions on our data [11].

Finding certain correlations between the students and their library usage has been a priority for this project. By correlating students' backgrounds with the amount of time spent in the library we can understand whether a school or course requires additional reading of material on top of the lecture and tutorial material that they provide and whether or not a school's courses are theoretical or practical in form. While data of student performance was not available to our project, there is an interesting research piece on finding correlations between library usage and student performance [12], which helped shape correlations that could be tested in our research as the correlation between library visits and school of study. The performance of students of universities in also a heavily researched areas and researchers have tried linking school of study with student performance in the past [13]. Another paper seeks to link library usage with students' grades while examining data but as this area of the university data cannot be researched on, the research project will only seek to use the literature to understand the patterns of university students visits to the library [14].

Through the usage of several different Python libraries, the research project was able to observe patterns of occupancy as well as predict the occupancy rates of the library. The documentation of the libraries used has helped with understanding the parameters as well as the output of the methods that would have been used [15,16,17,18,19] .

3. Methodology

As the main focus of this research is on getting insights of the library usage of different audiences through analysing the dataset that has been provided, a Data Science Research Methods process was undertaken to fully get an understanding of the whole dataset. A Data Science Research Method process requires Data Pre-processing as an initial plan on getting the best out of the data available by the use of cleaning and handling techniques. Continuing with the process, Exploratory Data Analysis which helps on getting plotted results is the second part of this project. During this phase, the questions that arise from Library occupancy rates are going to be answered through statistical analysis and visualisation of the results of the analysis. The visualisations of the representations of the data show how certain hypothesis arise through the exploratory data analysis. While Exploratory Data Analysis provides good enough information to see patterns of occupancy through the data, a Predictive Modelling process will enable to forecast the future of the occupancy rates. In this section, three models will be looked at, tested and evaluated to showcase the most fit model for this certain dataset.

3.1 Programming Language and Libraries in Use

As this is a Data Science Research project, the most widely used programming language for this particular cause is Python. Python is a high-level programming language which is used in various domains of work. As it is dynamic and can be used for both Object-Oriented programming as well as structured type programming it fits most projects as the language that can be used. Adding to that, Python has an enormous number of libraries available and free to use that cater to Data Science and Machine Learning. The programming language chosen to work with during this research project as highlighting the reasons for use above is Python. The open-source libraries that have been used in the project are matplotlib, Pandas, NumPy, StatsModels and scikit-learn. The libraries used enabled the use of predictive modelling approaches as well as visualisation of the results of the data analysis.

3.2 University of Sussex Data

The University of Sussex Library, the entity under research, provides various services to its audience both online and in their facility at the University of Sussex. As the focus of this research is to uncover patterns of occupancy of the library building, the online services of the library will not be investigated. As of the physical library services; one of the most common services that the library provides is the study areas where one could study in various scenarios such as quiet and silent or group study areas. A few more common services provided by the library are the Computing clusters which can be used by the people in the library for various workflows as well as the Printing and Photocopying facilities which enable one to print any kind of material needed. The Library also provides study rooms can be rented for group study in private. Another not so popular service of the University of Sussex library is the café and restaurant that they house which mostly caters to already-in-library people and not people entering the library just to eat or drink. Access to the building is by the university card for any kind of student as said before but for any kind of student without the card on sight, the library provides reference only cards that can be used to enter or exit the library for 24 hours only. Visitors can also request visitor passes to the library from the reception desk and any alumni of the university can also get a membership card. Another group of people that the library of the University of Sussex caters to are students of other universities through SCONUL which is a service that enables students who are at different universities campuses to use the libraries of those external to their universities. The Library building is open 24-hours a day most of the year which means that people access the library almost at all times. The only part of the year that the library building is not open 24-hours a day is during the Summer Vacation which is from July to September. During this period the library is open from 8:30 am to 11:00 pm on weekdays and noon to 7:00 pm on weekends.

The University of Sussex library provides an online service which enables users to see how busy the building is; a service which is in the work by getting the metrics of the gate stats and computing the number of people in the library. This service enables ease of use to the people that are going to attend the library building by letting anyone that would visit the library that it is overcrowded thus providing customer satisfaction. While this service works well for anyone that wants to attend the library at a certain point, forecasting the occupancy of the library will enable the staff of the Library to find other ways to work around such events such as if the forecasted occupancy is higher than estimated during a certain period they could extend the building or provide more sitting areas for the bigger number of people that would attend.

3.3 Data Collection

Data collection is the process of collecting information from one or more areas of operations to fit a particular cause; in our case a research project. Data collection for this project was done by the University of Sussex Library IT Staff which collects and manages data from entries and exits of each person to the Library building. The process involves a person using his card to enter or exit the building. The card in use is either the university ID card for any students, faculty and staff and reference only cards for any kind of visitors. As the card holds information for each person such as level of study and school of study for all students and relevant information for staff or faculty such as the department of work, the Library IT staff get this information for each person entering or exiting the building. The collection of this data has been happening for a large enough period to have an enormous amount of data available for analysis and predictive modelling. The data available to the project though derives from the year 2013 up until before the start of this project, in April of 2019.

3.4 Dataset Description

The library IT Staff made available two datasets that could be used for this research project. The two datasets are “entry-stats-2013-onwards” which is the log file of the entries and exits to the university library from 2013 up until April 2019 and “room-bookings-2013-onwards” which holds data of each library room booking such as the Booking ID, room description, the time of booking, the start time of the booking and the end time of the booking as well as the type of user which describes the level of study or if the user is part of staff or visitor and the secondary category of user which describes school of study or additional details for staff members or visitors. As the Room Bookings dataset doesn’t provide much helpful information on understanding the occupancy levels of the overall Library, it was decided that it would not be used for this research project. The rest of this report will solely be using data that is found in the “entry-stats-2013-onwards” to better understand occupancy patterns of the Library building.

3.4.1 Entry-Exit Stats Dataset

The aforesaid dataset has four columns that will be all be used for the fulfilment of this research project. The four columns are Timestamp which denotes the time and date of each visit and is of Date and Time type, the Direction which defines the direction of user passage at the gate; either “I” which means the user enters the library or “O” which defines that the user exits the library. Furthermore, the other two columns of the dataset are Category 1 which defines the type of user. For students, this is the level of study as covered above, for example, Undergraduate or Postgraduate or their different variations and for any other types of users, their backgrounds such as staff or a visitor or their occupation title. The last column of the dataset, Category 3, further describes each user; for students, this shows the school of study, for staff and faculty it shows the department of work and for visitors, it adds additional information to their background. While the dataset holds an enormous amount of data that can be used, there is a large need for data cleaning and pre-processing to get the best out of the data available. The following section shows how the above problem was handled.

3.5 Data Cleaning and Pre-processing

Data cleaning is the process of working on the data available to make it more readable and easier to work with by either deleting some variables or rows of the dataset or changing and correcting inaccurate information from the dataset after careful look and investigation of the information that is collected in the dataset. Such procedures are executed before any analysis of the dataset as the changes done to it could shape results differently. In the case of this research project, there was no fabrication of results through data cleaning procedures as the objective of cleaning the data was to only better understand the information that is held.

The cleaning procedure of the dataset was initiated by deleting any data before the year 2017. This workload was undertaken due to categories 1 and 3 not being readable as they had numbers as values which numbers were very difficult to be interpreted as there were not explained as to what each number meant as far as categories stand. Even though some of the numbers were given certain categories as advised by the IT staff, the rest of the numbers didn’t have any meaning. This data-wrangling procedure was arranged and executed due to data keeping at the time of the entry of those logs or rows of the dataset was not organized to the degree that the Library IT staff kept and inputted data into the dataset from 2017 onwards.

Also, several categories values in both Category 1 and Category 3 had duplicates with similar names. As this would prove to show messy data in the plotted results later on the research, the decision taken was to come up with a single word value for each category that meant to say the same thing but with different letters; therefore, values were re-entered through string manipulation programming techniques. To conclude with the data cleaning and pre-processing process, after a careful investigation on the dataset, several fields had the value default. As any field with default value wouldn't help to provide any insights to our cause of research, the rows that had fields with the value default were removed completely.

After carefully further investigating the dataset, there was no more any cleaning or pre-processing that was required. The procedure done was ethical as no rows of the dataset were subconsciously removed to fabricate the results of this research. Having a clean dataset, the research can move on with uncovering the insights that lay within the data in the following section which is Exploratory Data Analysis.

3.6 Exploratory Data Analysis

Exploratory Data Analysis is in Data Science terms the process of investigating the data available through analytical programming techniques to discover various patterns hidden within the data or spot various abnormalities in the data. Additionally, in Exploratory Data Analysis, certain hypotheses are tested to see whether assumptions taken at the early stages of the research are valid or not. The tested hypotheses are tested and visualised, so the results are easily understood through the various visualisation techniques.

3.6.1 Overall Library Interactions

As our main objective was to get an understanding of occupancy rates of the library during various points, the goal was to plot data that would help us understand patterns and create a hypothesis.

As the dataset contained direction for each log which identifies entries and exits the data that was plotted had to be seen as both entries and exits together at this stage as the goal of identifying patterns of certain categories entering or leaving the library didn't require the exclusivity of use of either entries of exits.

The first plot that was decided to be processed and visualised was the number of logs per Category 1, identifying the level of study of the people interacting with the library and the staff or visitors counts that interact with the library. This plot will show if the level of study correlates with the overall interaction with the library building.

The second plot used was showing the number of logs per Category 3, identifying the schools of study overall interaction with the library. Therefore, this plot will show the correlation between the school of study and library building interactions.

Other areas that the patterns of occupancy may appear are timely frames that would show an increase or decrease of occupancy. Consequently, the next plot decided to be visualised was the daily rate of interactions to the library without specifying school of study or level of study. This plot will show the correlation of day to the overall interaction of people with the library building. Continuing with visualizing time data, the next pattern that we are interested to look for is whether certain months attract more people to the library; this was done by plotting the count of logs per month. Lastly, when looking just for correlations with the time of interaction with the library it was decided to see how each year of the three years of data available had the most amount of interactions of people with the library. As the dataset ends in April of 2019, the percentages of can't represent the future interactions of people with the library for the rest of 2019; suggesting that this plot will only care about showing how 2017 and 2018 have been for the University of Sussex Library as far as the number of people entering and exiting the library building.

Other aspects of library interactions that this research wanted to find insights on were how the most attentive schools tend to spend their time in the library. It was best to show the top three schools in terms of interactions to the library building as the other schools had a huge difference in terms of interactions. By looking at daily, monthly and yearly graphs we would see which school visits the most and when and that would highlight hypotheses that we would see if there were correlations between certain factors.

As the focus was not only on the school of study but on the level of study as well, the same analytical spends were taken for the discovery of the top 3 categories of users or else the level of study for both daily and monthly plots were we could confirm if the level of study correlates with the number of interactions to the library.

3.6.2 Total Time spend per Category 1

The total time by each Category 1 or level of study was something that would provide insights and create hypotheses that could prove to be of importance to the research. As the total time spent was not something given by the dataset, there had to be calculations that could provide us with the numbers of total time spend inside the library of each Category 1. To compute this result, the direction column of the dataset was used to identify entries and exits to and from the library building. After figuring out all different entries and exits to the library for each specific category 1 using flags, the focus was on getting the absolute time difference of entries to exits. This calculation was done by the use of NumPy's timedelta64 data type which enabled the subtraction and addition of date and time data types so that we would have the overall total time spend in the library for each Category 1. To put the computation into simpler abbreviations, the 'I' overall time of each category was subtracted from the 'O' overall time of the said category, repeating for all different categories. Then by plotting the results of this computation, we would have the visual representations of the Total time spend of each category to the library building we can see if there are any patterns in the occupancy based on the level of study.

3.6.3 Total Time spend per Category 3

The same procedure that was taken for the discovery of the overall total time of each Category 1 was also taken to compute the overall total time of each Category 3 so we could see if the patterns also correspond with the school of study or if there are any other insights that we could unveil. The plots of the visual representations will be shown in the next chapter which is Findings.

3.7 Predictive Modelling

3.7.1 Time-Series forecasting

Looking at the data available for analysis and predictive modelling, the first notable discovery we can unveil is that the data rows have a timestamp variable. This timestamp variable can be used in predictive modelling approaches to predict a future timeline that is based not only on the features that are used in normal Machine Learning models but also on the time component that would treat observations differently to each other based on their time component. The library of the University of Sussex depends on the timeline as several timeline factors matter to their projected occupancy rates. Some of the factors may be weekdays and weekends or before noon times and past noon times or Summer holidays or the most important factor for a library of a University, assessment periods. As these factors shape the occupancy rates differently, caring about time in the predictions that are to the produced is a requirement.

To forecast the timeline of the library occupancy, three main Machine Learning models were considered to be used as the literature suggests their use and their proposed results as being very accurate. To evaluate the certain models that are used to select one as the one that would predict the most accurate predictions, evaluation techniques are analysed in this section to select the best evaluation method for this specific research project.

3.7.2 Models

3.7.2.1 ARIMA modelling

ARIMA stands for Auto Regressive Integrated Moving Average. This particular model is the model that is most notably used in Time Series forecasting. The individual words of the name ARIMA stand for:

AR - Auto Regressive: The term auto regressive implies a stochastic process where the output value of a said variable is depended upon the value that it had prior to it. With that said, the ARIMA model is differentiation function of stochastic equations.

I - Integrated: With this model being integrated, the time series of events has to differenced to be made into a stationary time series. With a model being integrated, a non-stationary time series will be made stationary by differencing of the data points.

MA - Moving Average: This terminology implies the use of the lag between the observation and the prediction error as a dependency in a moving average model which is applied to the lagged observation. The lagged observations are key components of the model.

The notation for ARIMA is ARIMA (p, d, q). The p stands for the number of autoregressive terms, d is the number of differencing needed, and q is the number of lagged forecast errors in the prediction equation. The mathematical formulae of the ARIMA (p, d, q) model is seen below.

$$\phi_\rho(B)(1-B)\,dY_t = \theta_q(B)a_t \cdot$$

ϕ_ρ are the auto regressive parameters that have to be estimated and θ_q are the moving average parameters that have to be estimated while a_t are the prediction errors that need to be estimated as well. If there is no differencing value, then the Arima model can be denoted to an ARMA model. In the same manner leaving any of other three values as zero can denote the model to their predecessors AR, I, MA, ARMA.

3.7.2.2 Seasonal ARIMA modelling

SARIMA or else Seasonal ARIMA is a model that is based on the normal ARIMA model but cares about seasonality in time series data or in other terms repeated cycles. It uses 4 extra elements that need to be configured on top of the p, d and q of ARIMA models. In general, a SARIMA model is a non-stationary ARMA model with seasonal characteristics. Seasonal non-stationarity can be seen is periodic processes in seasons where each season varies slightly from the other seasons, meaning that the seasonal component value varies from season to season.

The four elements added are P, D, Q, m where the P, D and Q elements are for the seasonal equivalents of the trend elements p, d and q of ARIMA. m stands for the number of time steps for a single seasonal period. The notation therefore for a Seasonal ARIMA model is the following:

$$ARIMA\ (p, d, q) \times (P, D, Q)\, m\ \cdot$$

3.7.2.3 Random Forest

A random forest is an implementation of multiple decision trees. To get a Random Forest model out of decision trees, the mean prediction from the individual decision trees. This means that the focus will be on using Random Forest as a Regressor rather than a classification model, but the model can be used for both regression and classification tasks.

As decision trees are prone to overfitting, Random Forest regressor tend to fix the problem by the use of bagging. Bagging is the practice of training each decision tree on a different sample of the dataset which in return helps with preventing overfitting. Random Forest, therefore, reduces correlation issues that individual decision trees have by using a subsample at each split of dataset thus making the decision trees de-correlated.

While there are benefits to using Random Forest model, by taking a subsample, trends in time series data are not able to be observed as each different decision tree takes a sample which will not show the trends and seasonality of the whole dataset. This means that the prediction errors of using Random Forest model may be higher than from the prediction errors of SARIMA model which takes into account trends as well as seasonality in data and ARIMA model which takes into account trends in data.

3.7.3 Evaluation of Models

The evaluation of the Machine Learning models that have been used is a critical process to the research project as providing actual scores of how well a model predicts the future occupancy rates of the library shows proof that the predictions will be accurate. The evaluation metric that is used on this project is Root Mean Square Error (RMSE) while also looking at its differences with Mean Absolute Percentage Error (MAPE) which is a metric that is equally used in Machine Learning projects.

3.7.3.1 RMSE

Root Mean Square Error (RMSE) is the standard deviation of the prediction errors acquired from using Machine Learning models on data. The prediction errors or else residuals are a metric of the distance of the regression line data points are, meaning that the Root Mean Square Error is showing the how spread out the prediction errors are. The reason for using this evaluation method is to understand if their prediction errors are small enough in number to use the certain model evaluated. This evaluation method is commonly used for the evaluation of forecasting and regression procedures in various Machine Learning experiments.

The decision to use Root Mean Square Error metric for evaluation is based on the fact that it is more useful to evaluations when large errors are particularly undesirable instead of Mean Absolute Percentage Error (MAPE) which punishes large errors the same way as small errors. By using RMSE, the focus was on punishing large errors. By punishing large errors, seasonality of data will be taken into account as the data that is to be modelled has seasonal characteristics. The seasonal characteristics are as examined in the SARIMA model, the various assessment periods, the summer school breaks and holidays.

4.Findings

This following chapter will showcase the findings of the exploratory data analysis as well as predictive modelling predictions on the sample data. The results of the finding will be showcased through visual plots and will be examined in-depth to understand the patterns shown and to come to certain realisations.

4.1 Exploratory Data Analysis Findings

4.1.1 Entries and Exits Combined for pattern recognition

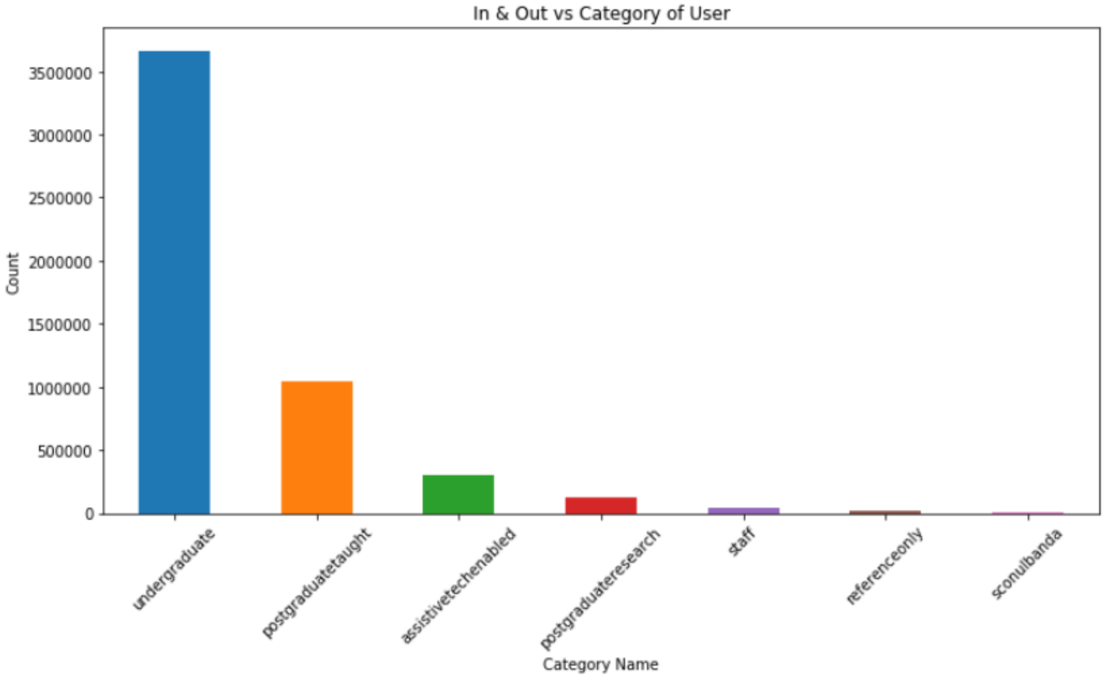

Figure 1 – In & Out vs Category of User Plot

As we can in Figure 1, Undergraduate students are the first category in the number of visits/exits. Second most entries and visits or put simply interactions with the library building has the Postgraduate Taught group which has less than 50% interactions than the Undergraduate group. As stated in [20], only a quarter of the overall students registered count is Postgraduate students therefore we can understand why Undergraduate students have the most amount of interactions; they are the majority out of all the different types of students. As the plot shows, Assistive Tech Enabled is the third category with the most interactions, this group is of people with certain disabilities which can use special or assistive computer equipment to help with their learning inside the library. The categories following in number of interactions are Postgraduate Research students, staff, reference only card holders which are mostly for people that have lost their library access card or have daily passes to the library and SCONUL Band A, which as stated in 3.2 University of Sussex Data, is a service that enables students of other universities to use the library of the University of Sussex.

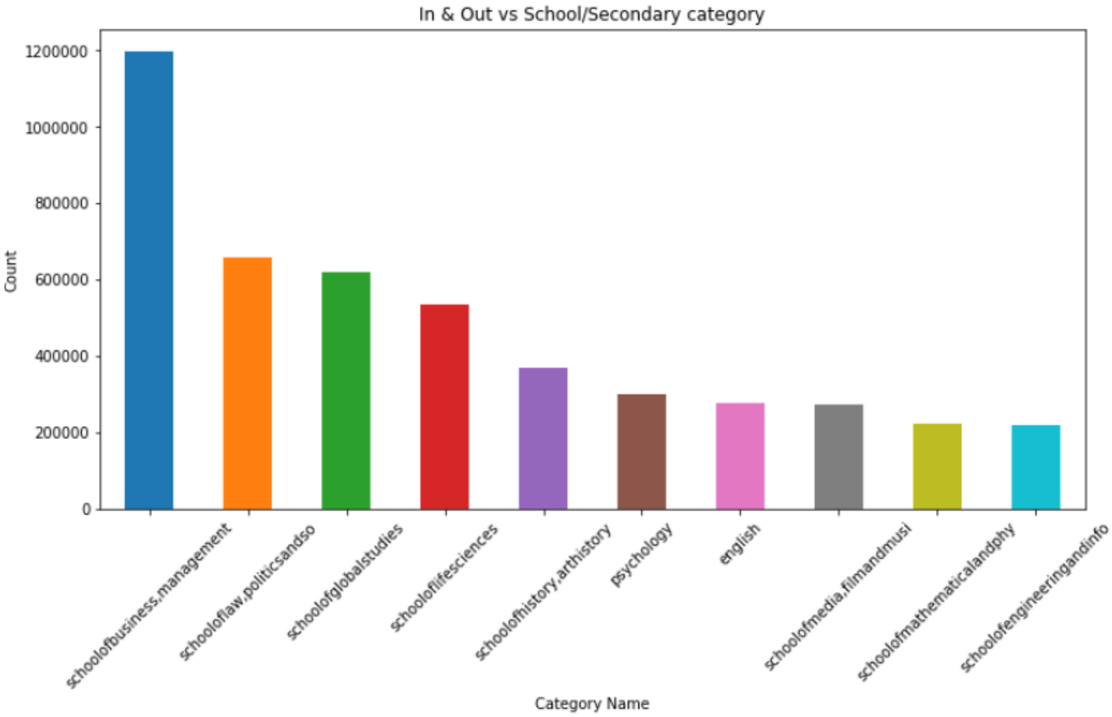

Figure 2 – In & Out vs School/Secondary Category

Looking at Figure 2, we can observe the different school of studies and how their interaction with the library is shaped into counts. As it can be seen the School of Business and Management has the most amount of entries and exits to the library while the School of Law, Politics and Sociology has the second most interactions and the School of Global Studies has the third most interactions. The schools that follow in interactions are the School of Life Sciences, the School of History, Art History and Philosophy, the School of Psychology, the School of English, the School of Media Film and Music, the School of Mathematical and Physical Sciences and the School of Engineering and Informatics. Having a first look at the schools that have the most interactions, these schools offer theoretical courses and courses that require the use of books for the majority of their learning outcomes. Thus, a hypothesis at this stage is that library material loans correlate with library interactions heavily. This hypothesis will be tested at a later stage to confirm or deny the above statement.

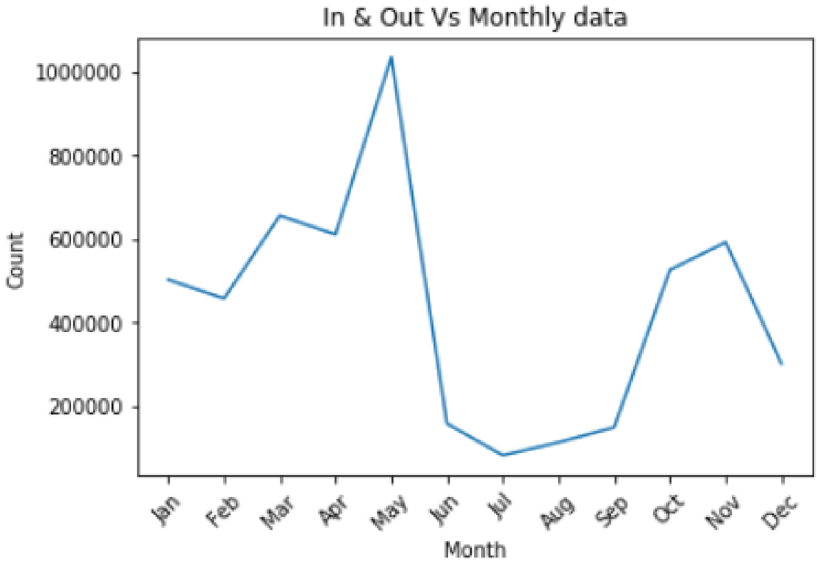

Figure 3 – In & Out Vs Monthly Data

In Figure 3, we can observe the overall daily interactions to the library. The library interactions peak in May and we can also see that March, as well as November, are months that attract people to the library more than the rest of the months. May for the University of Sussex is the End of Year assessment and examination period which can prove that students tend to visit the library the most during their exam and assessment periods. Secondly, March, as well as November, are time periods that schools tend to put out various assessment deadlines, therefore, students visit the library during those periods to either loan library material for out of campus study or for in-library reading and studying [20].

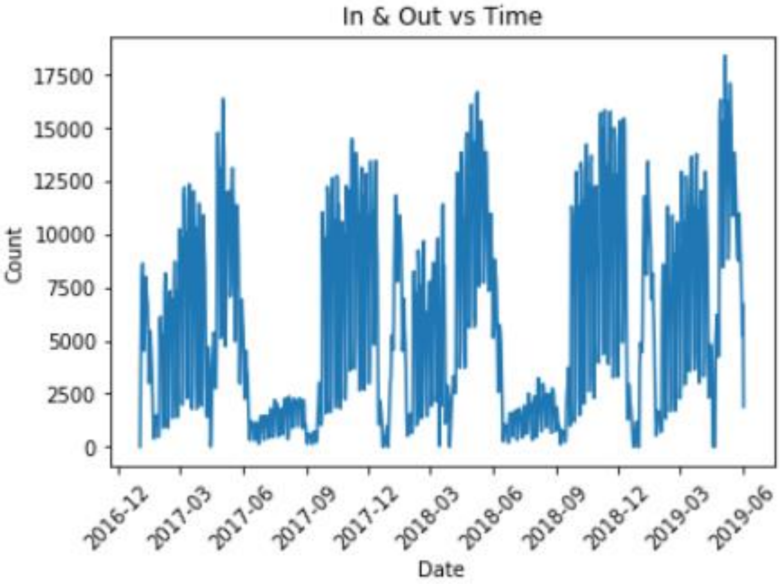

Figure 4 – In & Out Vs Time

Figure 4 shows the number of monthly interactions for the whole timeline of the dataset. By plotting this visualisation, we can understand in more depth how the certain seasonal characteristics of our data have an impact on the number of interactions with the library. We can see how the monthly patterns observed in Figure 3 continue through the different years and are of a yearly routine.

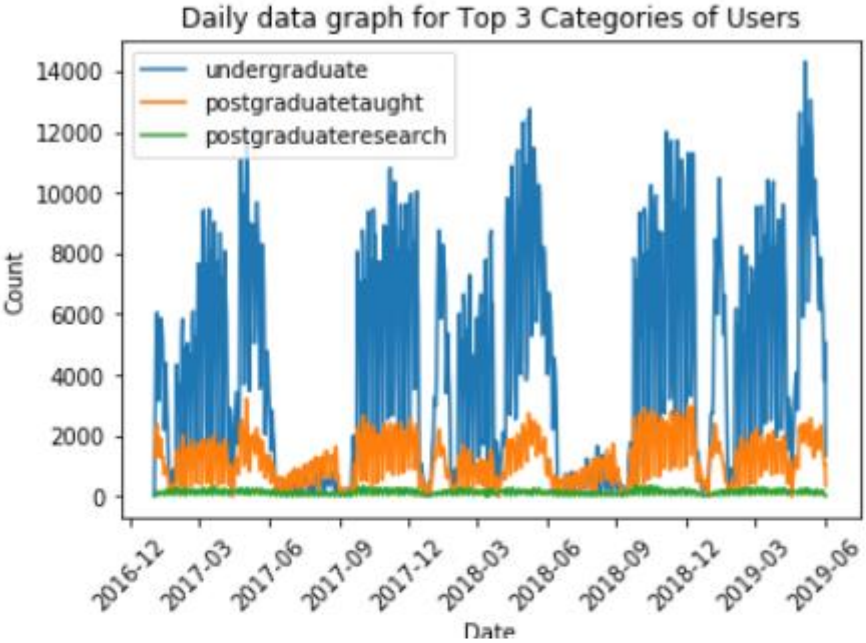

Figure 5 – Daily Data Graph for Top 3 Categories of User

Figure 5 shows a plot that specifically cares about providing insights for the top three categories of users as of daily interactions. We can see that Undergraduate, Postgraduate Taught and Postgraduate Research students are the top three categories of user as for daily interactions. As seen in the previous figures, interactions of students with the library peak in the months: May March and November. While Undergraduate students tend to visit the library the most as they are the most of any different types of students, Postgraduate students tend to visit the library in a very systematic way and visit the library throughout the year. This has to do with Postgraduate courses being twelve months long which means that as they do not have summer breaks, their visits to the library continue being at the same level throughout the year. On the other hand, as Undergraduate courses end in June, the summer break from June to September does not attract many Undergraduate students to the library which is understandable as there is no need for studying except for various late submissions or re-sits taken.

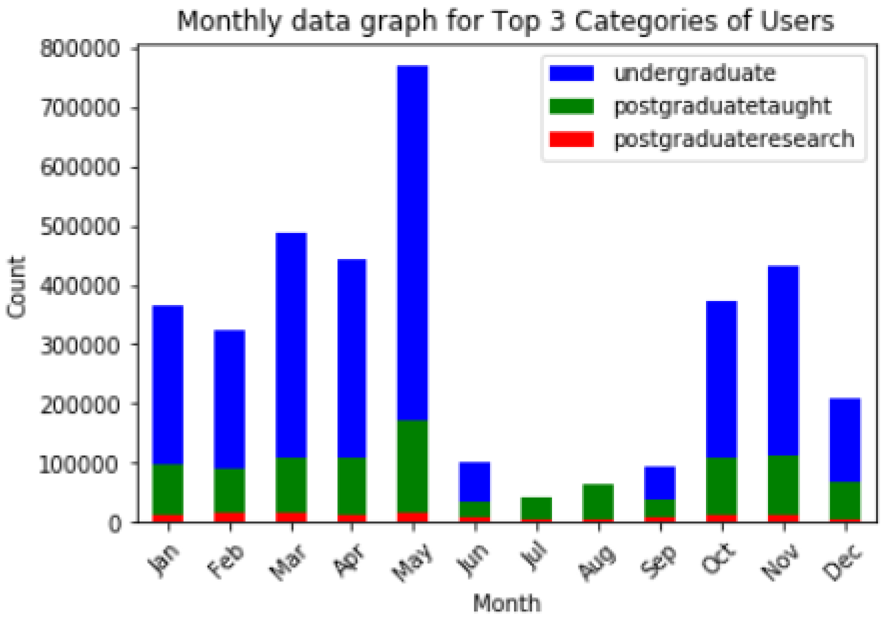

Figure 6 – Monthly Data Graph for Top 3 Categories of Users

We can further examine the patterns of the different types of students that we got insights to from Figure 5 in Figure 6 which shows the top three categories of users in monthly interactions. Same patterns arise when looking at daily interactions as well as monthly interactions.

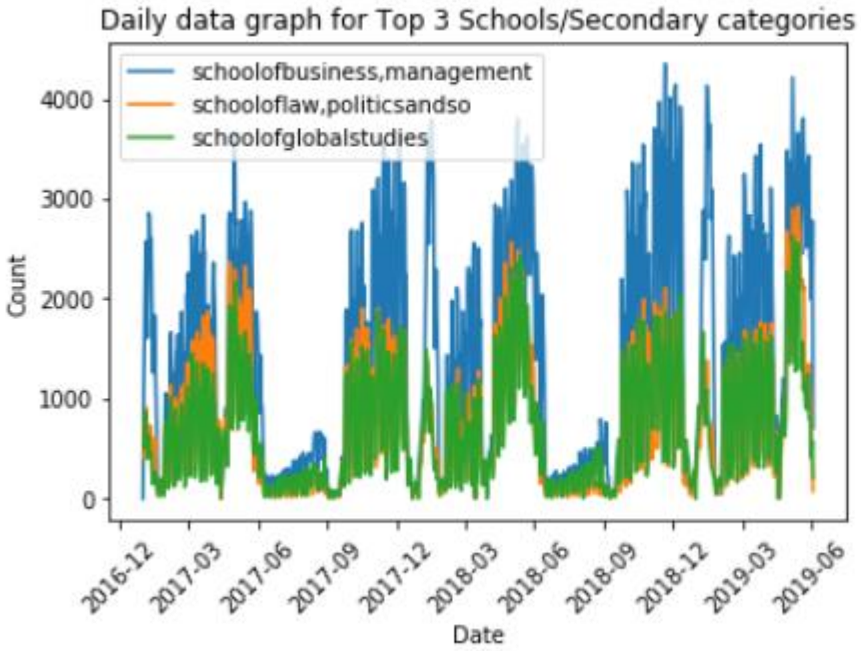

Figure 7 – Daily Data Graph for Top 3 Schools/Secondary Categories

While understanding the patterns of the several types of students is critical to understanding the overall occupancy of the library; understanding patterns of occupancy of the different schools is equally beneficial to the overall research. In Figure

7 we can observe the top three schools in term of daily interactions to the library. As previously examined the School of Business and Management as well the School of Law Politics and Sociology and the School of Global Studies are the schools that tend to visit the library the most. Their daily distribution tends to stay at the same pattern throughout the days examined. As the School of Business and Management is the school seems to have a large difference in numbers with the other two schools in interactions, the number of students registered to the School has to be the reason why it is ahead of the other two schools while also follow the same daily distribution of visits to the building. A hypothesis that was earlier in the research suggested that can be seen again in Figure 7 is that the three schools seen are mostly theoretical study-based schools with emphasis on hard copy material reading.

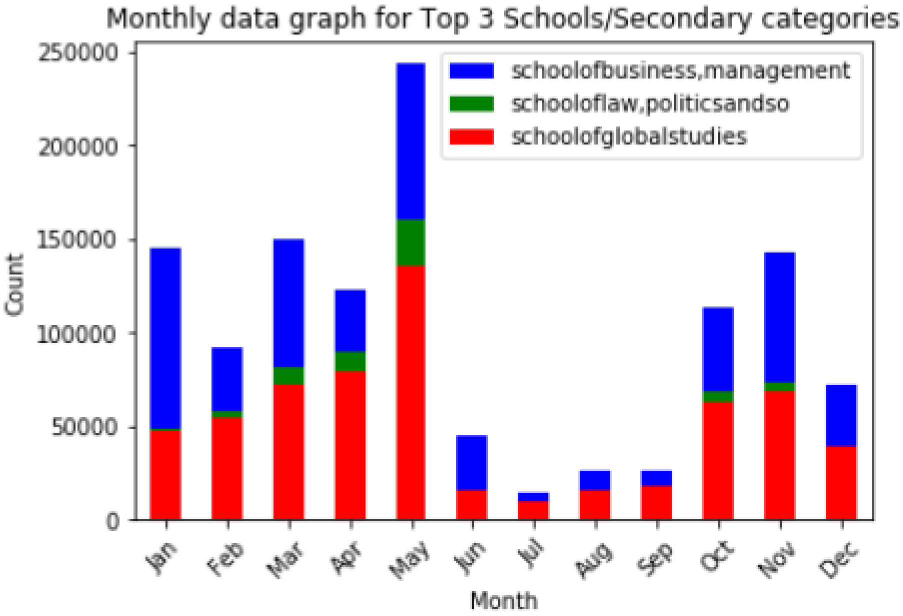

Figure 8 – Monthly Data Graph for Top 3 Schools/Secondary Categories

In Figure 8, a representation of the three top schools in number of monthly interactions, we can see that the School of Business and Management visits the library the most and their attendance peaks in May which is reasonable as it is the End of Year assessment and examination period. The correlation of school of study and number of interactions can be seen again in figure 8.

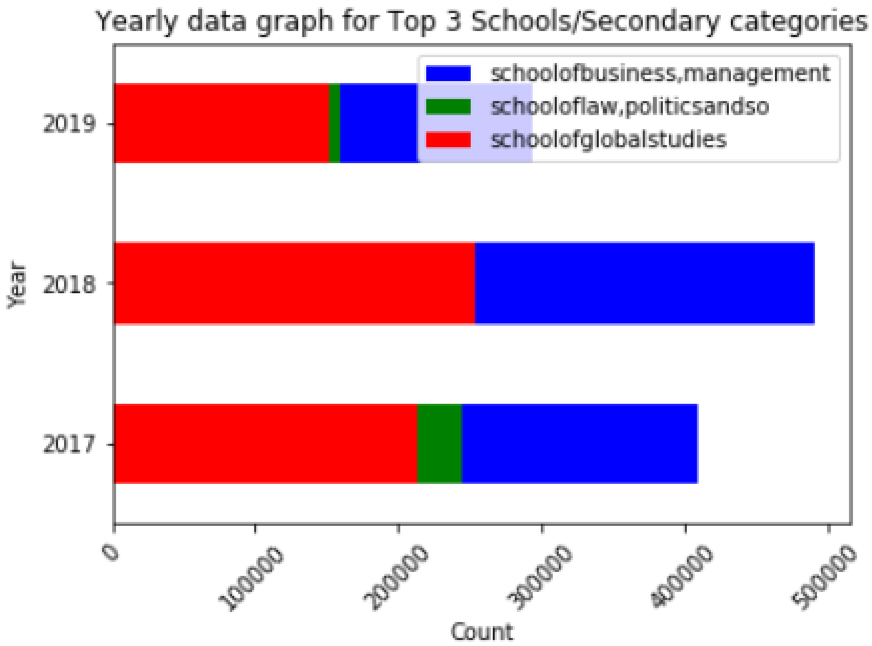

Figure 9 – Yearly Data Graph for Top 3 Schools/Secondary Categories

Figure 9 shows how the yearly interactions are shaped. As the dataset does not have the full values of the interaction of the year 2019 but only up until April, we can speculate how the interaction are being distributed. As seen, there is an increase of interactions to the library from 2017 to 2018 overall but we can also observe that the School of Law Politics and Sociology did not visit the library in 2018 as much as it did in 2017 and the first third of 2019.

4.1.2 Total Time Spend for pattern recognition

Understanding the total time spend by each type of students as well as by each different school can validate certain correlations that we have seen through the previous plots or even deny the claims of those correlations. The following Figures 10 and 11 will help show how the type of students and school of study plays a role on how much time is spend in the library building.

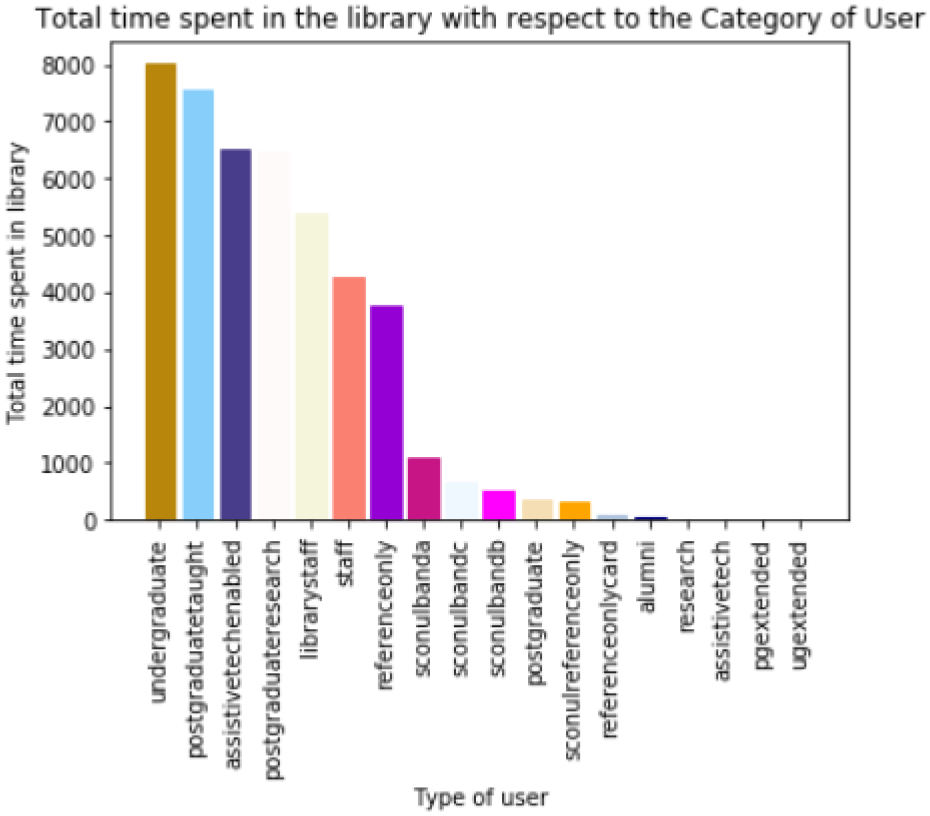

Figure 10 – Total Time spend in the library with respect to the Category of User

Computing the total time spend in the library for each different category of user, we can see in Figure 10 that Undergraduate students tend to spend the most time in the library while Postgraduate student are the second most and Assistive Tech Enabled card holders are the third category in most time spend in library; something that we have seen through Figure 1 as well. Even though the category Undergraduate has the most time spend in the library, postgraduate students have several different categories under their name as an entity; Postgraduate Taught, Postgraduate Research, Postgraduate and Postgraduate Extended which their overall total comes up being way more than the total number of hours Undergraduate students spend in the library.

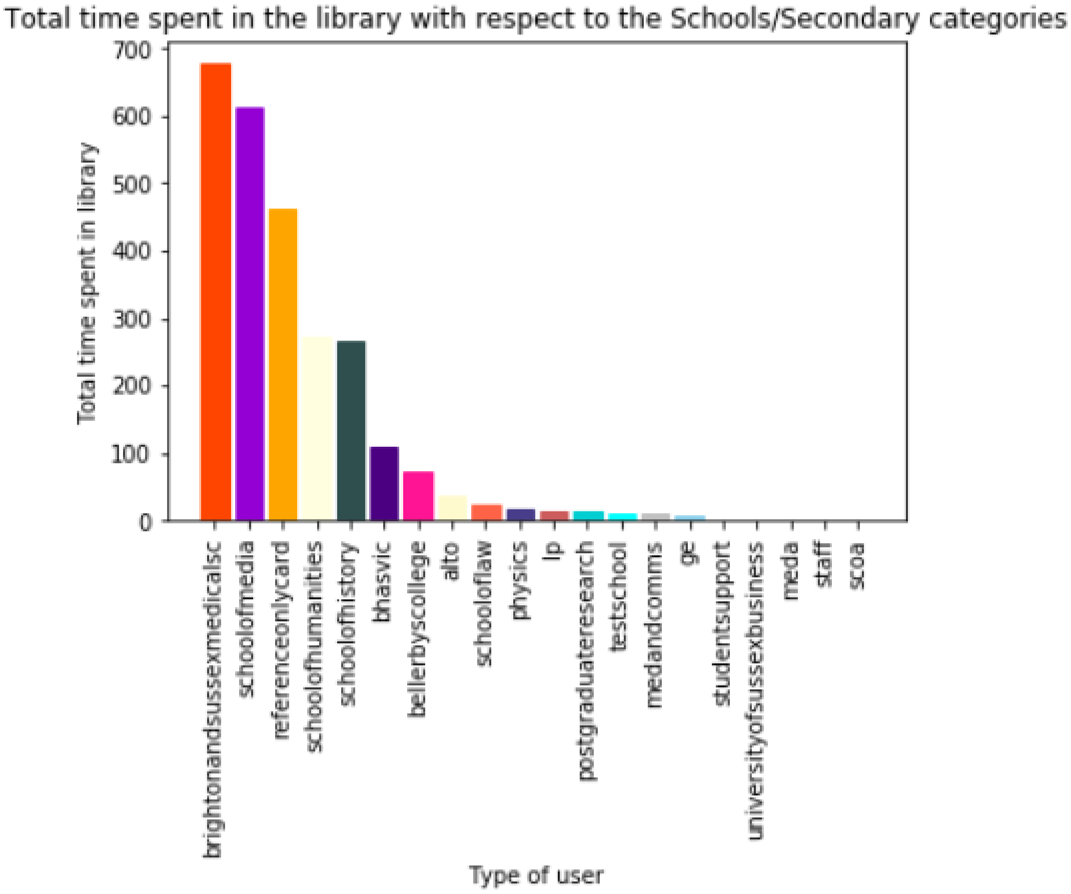

Figure 11 – Total time spend in the library with respect to the Schools/Secondary Categories

In Figure 11 we can see the total time spend by the various schools or secondary categories. The school which spends the most total time in the library is the Brighton and Sussex Medical School with School of Media having the second most time spend in total. Therefore, based on the results of the plot on the previous slide, the Business and Management school that had the most entries and exits spends a very little amount of time, therefore, it is best to assume that the activity they partake is book loans and returns. On the other hand, the Brighton and Sussex Medical School seem to be using the library to read material hence the amount of time spent. Other schools that visit the library to read material or study are School of Humanities and School of History as they spend a lot of time in the library while not being one of the top schools in terms of the number of interactions with the library building.

4.2 Predictive Modelling Findings

In this section, the plots of the predictive modelling approaches used in the training data show how effectively each different model predicts the occupancy of the library. As the use of the whole dataset would require very heavy computationally, it was decided to use the data of the top types of users or the Category 1. By looking at the plots we can understand if the said model is predicting the occupancy correctly by having a signal of the actual data alongside it to confirm the results. Each different model plot will be looked at individually and then the Root Mean Square Error will be evaluated to see how each model performs.

4.2.1 ARIMA modelling

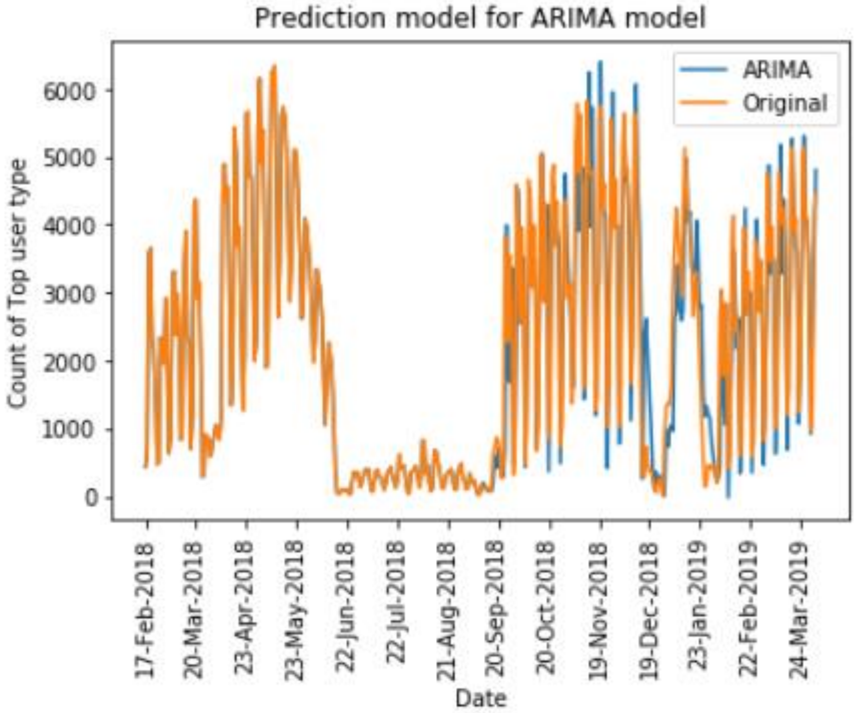

Figure 12 – Prediction Model for ARIMA model

ARIMA modelling cares about the trend characteristics of our dataset therefore it should perform well in the short term. This statement can be seen in practice in Figure 12, as the two signals orange for the original data and blue for the prediction are almost identical for the first few months of the predictions. We can also observe that in the long term, the outliers in our data are not predicted very well which is due to the fact that seasonality of data is not taken in the formula of the ARIMA model. Overall, the ARIMA model provides good predictions for what it can predict without using seasonality as a characteristic of the model.

4.2.2 SARIMA modelling

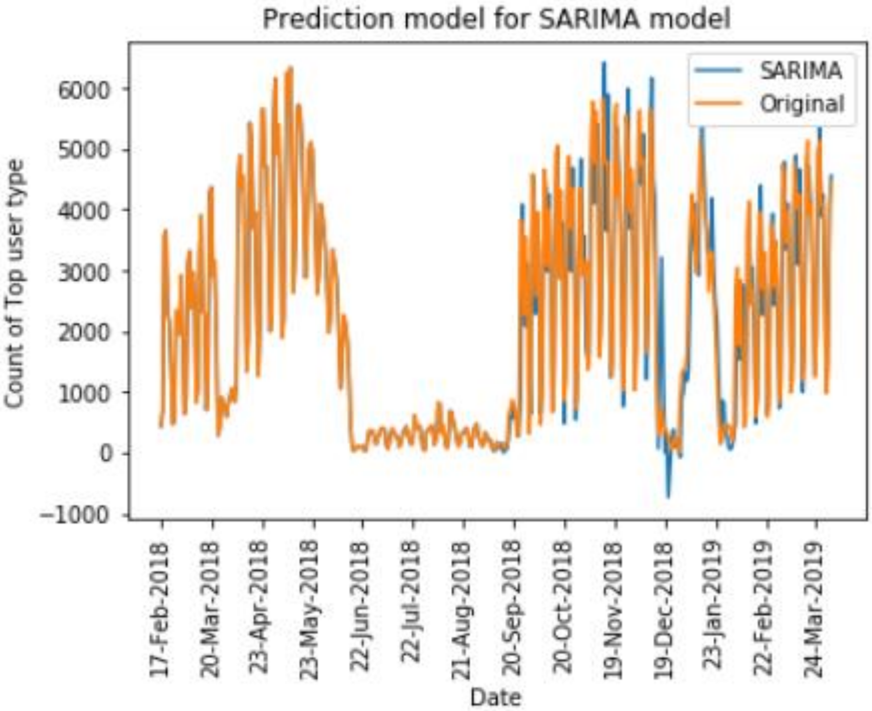

Figure 13 – Prediction model for SARIMA model

As seen above, in Figure 13, the SARIMA model shows great predictions both in the short-term future as well as the long-term future. The prediction given from using the training data with the SARIMA model provides us with results that surpass the prediction of the ARIMA model and the reason for that is its use of the four seasonal components it has on top of the three of ARIMA. As we can detect, the last seventy days of the two hundred days predicted, the prediction is almost the same of the original data which shows that with more training data, the predictions would have been even better with more precision. Seasonal characteristics seem to be providing greater detail to the predictions.

4.2.3 Random Forest

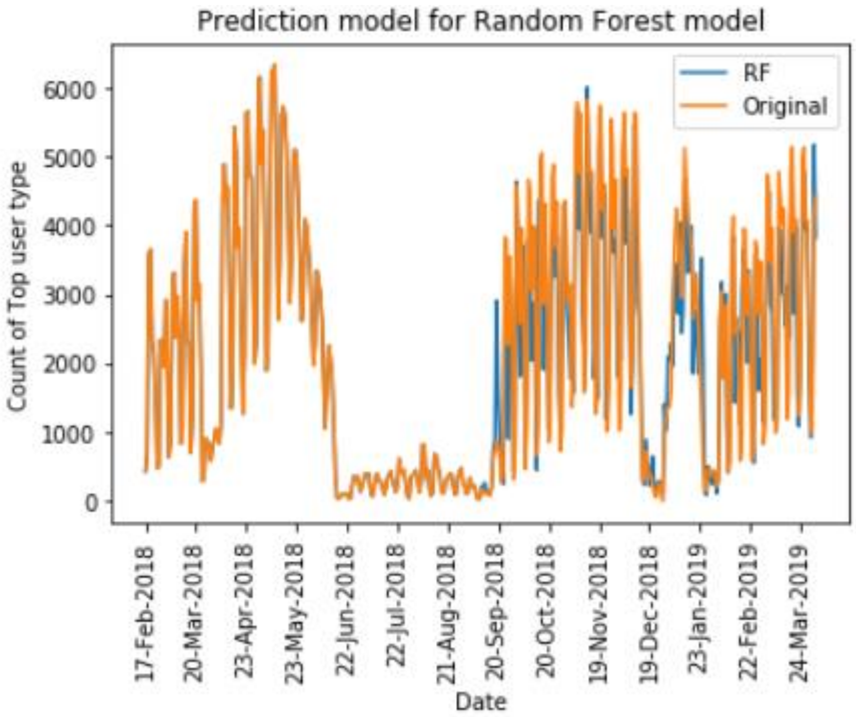

Figure 14 – Prediction Model for Random Forest Model

In Figure 14 we can see how the Random Forest regressor predicts the occupancy rates. As seen in the plot the first half of the prediction is almost identical to the original data. Moving on though we can see that the predictions are not extremely good in the long term. As we can see in the area that represents the data from 19-Decemeber-2018 to 23-January-2019, the Random Forest Regressor can not pick up the seasonal characteristic and therefore can not predict the huge difference in occupancy from December to January. We can see that this model works well overall but as this is a Random Forest, the trend and seasonal characteristics can not be picked up. Furthermore, the random sample for each decision tree means that the predictions may have been way worse in other scenarios. Overall this model would work well for short term forecasting of the library occupancy but would fail to predict well in the long-term.

4.2.4 Root Mean Square Error Evaluation

By using the Root Mean Square Error metric for evaluation on our models, the three values we got were for ARIMA modelling 864.782, for SARIMA modelling 698.468 and for our Random Forest Regressor model 749.216. As seen through the plots of the prediction tests in the previous section, SARIMA predicts the forecast with higher accuracy and with fewer prediction errors than the other two models. As SARIMA covers both the trend in the data as well as the seasonality of the data, it creates a better-forecasted timeline of the occupancy.

While looking at Random Forest Regressor which does not cover trend or seasonality through its model, we can see that it gets a lower amount of prediction errors through our RMSE rather than ARIMA which cares about the trend in data. While this is something that would mean that Random Forest predicts the occupancy time series more accurately; the occurrence of each data point is not taken into account. Therefore, since this project focuses heavily on understanding how each different month or period of the year shapes the occupancy, the Random Forest RMSE is dismissed as a reason to consider this model to be used instead of ARIMA which covers trends in data.

While also looking at the RMSE score of ARIMA, we can observe that as it is higher than its’ successor SARIMA, thus seasonality in data plays a big role when predicting such data sets’ time series. Even though that trend in data is covered in ARIMA, seasonality in conjunction with the trend has to be considered and computed for the best model predictions of this certain data set’s timeline and its characteristics.

5. Conclusion

This research project had as objectives to figure out patterns of occupancy rates of the library building of the University of Sussex by using Exploratory Data Analysis techniques in conjunction with Predictive Modelling to forecast the occupancy rates for the future.

For the first objective, understanding the patterns of the historical data that the library had kept through the years, the focus was on seeing if the level of study correlates with library visits and if the school of study correlates to library visits. As seen through 4.1 Exploratory Analysis Data Analysis Findings, the insights that we have uncovered will be presented in this chapter.

There is a correlation between the school of study and the number of interactions to the library as well as the time spent in the library. We have seen that schools that are based on theoretical material for their courses tend to visit the library a lot but we have also seen that those specific schools do not spend a lot of time in the library which means that their visits are mostly for library material loans and returns which are services that do not require a lot of time spent. We have also seen that the schools with the most time spend in the library do not visit the library as much which means that they spend a lot of time on each visit, either for silent studying or for group studying in private rooms. W e have observed the same pattern for Undergraduate and Postgraduate visits and time spend in the library as Undergraduate students tend to visit the library the most as they have the most amount of interactions with the gate of the library but Postgraduate students who are also way less in count tend to spend a lot more time in the library. We have also seen that the assessment periods are the times that the library gets its most visits as visits to its facilities are ways of preparing for any assessment.

For our second objective, forecasting the occupancy of the library by using Predictive Modelling approaches, the focus was on testing various models that work well with time series forecasting and evaluating the said models with Root Mean Square Error to come to a realization of which one of the models can predict the occupancy rates to a greater extent by caring about the prediction errors.

As seen in 4.2 Predictive Modelling Findings, Seasonal ARIMA can forecast the time-series with less amount of prediction errors than ARIMA and Random Forest Regressor. This insight has been evaluated with the use of Root Mean Square Error which ensures that out of these three models, SARIMA can be used for the forecasting of the library occupancy rates in the future as it provides accurate predictions. When caring about short-term and long-term predictions as we could understand from the plots generated, ARIMA and Random Forest both can predict well on short-term forecasts but fail to produce an equally good prediction on long-term as seasonal characteristics are not used in these models.

When it comes to further research, as this has been the first attempt of using SARIMA modelling on predicting the occupancy rates of the library building, a general extension to this project could have been some kind of secondary evaluation procedure taken to cover any kind of information that was left out or has not been covered to its full extent. Such an example could be Cross-Validation of the model predictions to future proof of its use in practice.

Another concern when it comes to the predictions acquired is that the data used is only two and a half years long meaning that we can not see exactly how each different year shapes the occupancy of the library. If the data set given was of higher quality, the prediction may have been more accurate and could put more emphasis on the differences of each different year when it comes to SARIMA modelling since the other two models used would not predict more accurately when given more historical data.

The data collection process at the University of Sussex Library, as this research has shown could be improved which would help understand additional patterns of occupancy that this research was not able to unveil. A suggestion would have been to have a column of services used by the person in each different log but that could also be a completely different data set which enables to see each different log of the library services system thus covering book loan procedures and use of workstation computers and book renting. These kinds of advancements to the data collection methods used by the University of Sussex Library IT Staff could provide in return better insights into the occupancy of the library.

Bibliography

[1] Chen, H., Doty, P., Mollman, C., Niu, X., Yu, J. and Zhang, T. (2015). Library assessment and data analytics in the big data era: Practice and policies. Proceedings of the Association for Information Science and Technology, [online] 52(1). Available at: https://onlinelibrary.wiley.com/doi/full/10.1002/pra2.2015.14505201002 [Accessed 9 Apr. 2019].

[2] Burrell, Q. and Cane, V. (1982). The Analysis of Library Data. Journal of the Royal Statistical Society. Series A, [online] 145(4). Available at: https://www.jstor.org/stable/2982096?seq=1#page_scan_tab_contents [Accessed 11 Apr. 2019].

[3] Shill, H. and Tonner, S. (2003). Creating a Better Place: Physical Improvements in Academic Libraries, 1995–2002. College & Research Libraries, [online] 64(6). Available at: http://ht- tps://crl.acrl.org/index.php/crl/article/viewFile/15626/17072 [Accessed 12 Apr. 2019].

[4] Lu, N., Song, R., Heng, D., Gottipati, S., Tay, C. and Tay, A. (2017). `Using Data Analytics for Discovering Library Resource Insights { Case from Singapore Management University. [online] Available at: http://ink:library:smu:edu:sg=sisresearch=3835= [Accessed 9 Apr. 2019].

[5] Law, R. (1998). Room occupancy rate forecasting: a neural network approach. International Journal of Contemporary Hospitality Management, [online] 10(6), pp.234-239. Available at: https://www.emeraldinsight.com/doi/abs/10.1108/09596119810232301 [Accessed 10 Apr. 2019].

[6] Tang, C., King, B. and Pratt, S. (2016). Predicting hotel occupancies with public data. Tourism Economics, [online] 23(5), pp.1096-1113. Available at: https://journals.sagepub.com/doi/abs/10.1177/1354816616666670 [Accessed 10 Apr. 2019].

[7] Mathioudaki, M. (2018). Library: What's the use?. Postgraduate. University of Sussex.

[8] Ahmadi, M., Dileepan, P., Murgai, S. and Roth, W. (2008). An exponential smoothing model for predicting traffic in the library and at the reference desk. The Bottom Line, 21(2).

[9] Brownlee, J. (2017). How to Create an ARIMA Model for Time Series Forecasting in Python. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/ [Accessed 15 Apr. 2019].

[10] Jenkins, G. and Alavi, A. (1981). SOME ASPECTS OF MODELLING AND FORECASTING MULTIVARIATE TIME SERIES. Journal of Time Series Analysis, 2(1), pp.1-47.

[11] Tavgen, A. (2018). Time series modelling. [Blog] Medium. Available at: https://medium.com/@ATavgen/time-series-modelling-a9bf4f467687 [Accessed 3 Apr. 2019].

[12] Open.ac.uk. (2018). Open University Library Data Project | Analysing the relationship between library use and student retention and student performance. [online] Available at: http://www.open.ac.uk/blogs/librarydata/ [Accessed 11 Apr. 2019].

[13] Badr, G., Algobail, A., Almutairi, H. and Almutery, M. (2016). Predicting Students' Performance in University Courses: A Case Study and Tool in KSU Mathematics Department. Procedia Computer Science, [online] 82. Available at: https://www.sciencedirect.com/science/article/pii/S1877050916300266 [Accessed 8 Apr. 2019].

[14] Renaud, J., Britton, S., Wang, D. and Ogihara, M. (2015). Mining library and university data to understand library use patterns. The Electronic Library, [online] 33(3). Available at: https://www.emeraldinsight.com/doi/abs/10.1108/EL-07-2013-0136 [Accessed 9 Apr. 2019].

[15] Matplotlib.org. (2019). Matplotlib: Python plotting — Matplotlib 3.1.1 documentation. [online] Available at: https://matplotlib.org/ [Accessed 7 Jul. 2019].

[16] Numpy.org. (2019). NumPy — NumPy. [online] Available at: https://www.numpy.org/ [Accessed 6 Jul. 2019].

[17] Pandas.pydata.org. (2019). Python Data Analysis Library — pandas: Python Data Analysis Library. [online] Available at: https://pandas.pydata.org/ [Accessed 5 Jul. 2019].

[18] Scikit-learn.org. (2019). scikit-learn: machine learning in Python — scikit-learn 0.21.3 documentation. [online] Available at: https://scikit-learn.org/stable/ [Accessed 13 Jul. 2019].

[19] Statsmodels.org. (2019). StatsModels: Statistics in Python — statsmodels v0.10.1 documentation. [online] Available at: https://www.statsmodels.org/stable/index.html [Accessed 10 Jul. 2019].

[20] Sussex.ac.uk. (2019). Facts and figures : Rankings and figures : About us : University of Sussex. [online] Available at: https://www.sussex.ac.uk/about/facts/facts-figures [Accessed 15 Jul. 2019].