Advances in Computing 2023, 13(1): 1-8
DOI: 10.5923/j.ac.20231301.01

Predicting Electrical Energy Consumption for Commercial Buildings Using Data Science: A Case Study of the University of Zambia

Michelo Mtolo*, Grayson Himunzowa

Department of Electrical and Electronics Engineering, University of Zambia, Lusaka, Zambia

Abstract This study investigates and analyses the prevailing energy consumption challenge and practices at the University of Zambia. For this reason, Data Science was proposed in this study. An orange toolkit which is a data Visualisation, Machine learning and data mining toolkit from Anaconda open-source distribution software and Microsoft power BI were used to achieve the objective and answer the research questions of this study. The process involved data collection of energy consumption from a specified target area which was the University of Zambia through the office of the Residence Engineer, raw energy consumption datasets were collected for 10 years from 2009 to 2019 in form of Microsoft Excel and stored using Microsoft OneDrive software, the data was thoroughly cleaned, understood and analysed, then based on that, a model was developed in Orange toolkit using Linear regression as a machine learning tool to give us the predictions, with the main data split into test and training data, the model was able to analyse the data and do the prediction of energy consumption for the year 2020, later data visualisation was achieved through Microsoft power BI software, hence completing the Data Science cycle and meeting the research objectives. Results show total monthly energy consumption for 2020 based on the historical data and other changing factors, with an accuracy of 92.4% test on training data and 77.4% test on test data against the actual 2020 data. The research has identified expected energy consumption through predictive analysis at the University of Zambia, proper energy management is a matter of concern at the University of Zambia.

Keywords Data Science, Energy Consumption, Machine Learning, Linear Regression, Training Data, Test Data

1. Introduction

As to date smart meters are being deployed in millions of buildings, providing bidirectional communication between the utility companies and energy consumers, which has given rise to the generation of extensive volumes of data with high velocity and veracity attributes. Such data have a time-series notion, typically consisting of energy usage measurements of component appliances over a time interval [1]. This data stored in computers can be used to predict energy consumption for consumer companies to optimise energy use and forecasting demand. This is possible nowadays due to the advent of data science, capable of taking these large volumes of time series data and facilitates decision through the transformation of meaningful information. This has revolutionised utility providers for learning customers energy consumption behaviour, making it possible to predict electrical energy consumption for the future. Utility companies are constantly working towards determining the best methods to improve profit and reduce costs by introducing programs, such as demand response and demand-side management, that best align with the consumer's energy consumption behaviour. Despite having a marginal success in achieving the goals of such programs, sustainable results are yet to be accomplished [2].

It is challenging to understand consumers behaviour individually and tailor strategies that suggest energy-saving plans. Furthermore, the relationship between the parameters affecting energy consumption patterns and human behaviour is non-static [2]. Consumer behaviour is dependent on weather and seasons which has a variable impact on energy consumption decisions. Thereby, actively engaging consumers in personalised energy management by facilitating well-timed feedback on energy consumption and the related cost is key to steering suitable energy saving schemes or programs [3]. Therefore, designing models that are capable of analysing energy time series from smart meters which are capable of intelligently forecasting energy usage is very critical.

Data science builds systems and algorithms to discover knowledge, detect patterns, and generate useful insights and predictions from large-scale data [4].

The purpose of this study is to develop a model for

* Corresponding author:
michelomtolo@gmail.com (Michelo Mtolo)

Received: Jan. 2, 2023; Accepted: Jan. 13, 2023; Published: Jan. 16, 2023
Published online at http://journal.sapub.org/ac

predicting electrical energy consumption using data science. The University of Zambia Great East Road Campus was used as a case study where historical electrical energy data was collected. The study carried field evaluation of the energy consumption by taking historical data sets for 10 years (from 2009 to 2019). The data were collected through the residence engineer's office in form of excel documents.

Historical data sets gathered were analysed and enumerated. An orange toolkit which data science toolkit from Anaconda open-source distribution software was used to achieve the objectives and answer the research questions of this study.

2. Literature Review

2.1. Data Science

Data science is an inter-disciplinary field that uses algorithms, processes, scientific methods, and systems to extract knowledge and insights from many structural and unstructured data. This allows efficiency, in managing costs, identifying new business opportunities, and results in boosting market advantage as compared to the old traditional method which used excel documents to manage different types of data [5].

To define data science and improve data science project management, begins with its life cycle. The first stage in the data science pipeline involves business understanding where you ask relevant questions and define objectives for the problem to be tackled. The next stage is data mining where you gather the data necessary for the project. Followed by data cleaning, a stage constituting fixing the inconsistencies within the data. Next is an equally critical stage which is mainly the exploration of data. Here data scientists form a hypothesis from analysing the observed data. Next is feature engineering where the data scientist selects more meaningful and constructive data from the raw data. Predictive modelling is the next stage where data. scientist train machine learning models, evaluate their performance and use them to make predictions. The last stage is data visualisation where the data scientist communicates the findings with key stakeholders using plots and interactive visualisations [5].

2.2. Application of Data Science in Commercial Buildings Electrical Energy Management

Data science techniques have been frequently used to support and improve the basic aspects of energy efficiency and management. Accordingly, this section focuses on applications of Data Science that can do the following: predicting the electrical energy demand required for the efficient operation of a building, analysing the economic and commercial impact of user electrical energy consumption, detecting, and preventing electrical energy fraud.

3. Methodology

3.1. Research Approach

This research was conducted in a highly technical space and drew participation from the University of Zambia. The study used quantitative methods and took the form of a case study of the University of Zambia.

3.2. Research Tools

To test provide responses to the questions, it is necessary that we observe the behaviour of data when exposed to scientifically proven systems. This will help to obtain empirical evidence that can be relied upon and plausible. There are many simulations software tools available to formulate predictive model. For the purposes of this study, three (3) tools were identified suitable for this study. The identified are Anaconda open-source distribution software (Edition 2020), Orange Data Science Toolkit (Edition 2020), and Microsoft Power BI (Edition 2020).

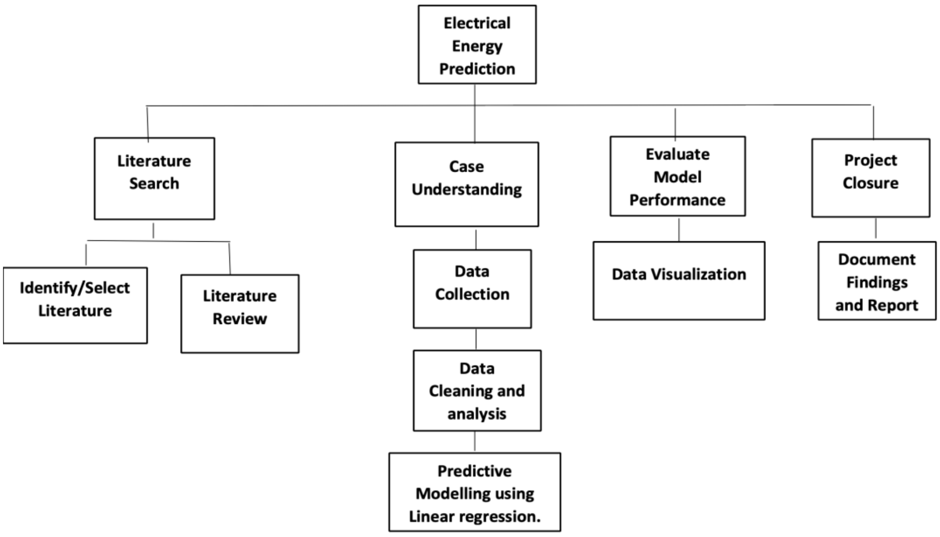

Electrical Energy Prediction
Literature Search
Identify/Select Literature
Literature Review
Case Understanding
Data Collection
Data Cleaning and analysis
Predictive Modelling using Linear regression.
Evaluate Model Performance
Data Visualization
Project Closure
Document Findings and Report

Figure 1. Research approach

3.3. Data Collection Methods and Instruments

Data collection of this research involved collecting the university's historical electrical energy consumption data for a period of ten years from 2009 to 2019. The historical data were Zambia Electrical Supply Corporation monthly statements which amounted to 120 months was collected from the resident engineer's office. Historical university status from 2009 to 2019, historical exam status from 2009 to 2019, historical season status from 2009 to 2019, historical school activities from 2009 to 2019 and lastly student's population from 2009 to 2019 were collected using Microsoft excel and stored using Microsoft OneDrive software. The primary data was historical electrical energy consumption data. The literature in chapter two will be used as well to help deduce the gathered facts and conclude the study.

3.4. Data Cleaning and Analysis

The data collected was cleaned and analysed in this way, the historical university status, was categorised in two scenarios: university closed or university open. The historical exam status was also taken into consideration and categorised in three scenarios, namely: no exam, midterm exam and end term exam. The historical season status was also categorised in four scenarios, namely: Autumn, winter, summer, and rain season. The historical school activity categorised in five scenarios, namely: arrival of students, classes, study break, term break and vacation. Lastly the student population was categorised as full-time students and part time students. This data is critical to provide accurate total electrical consumption predictions. This was done in Microsoft excel document.

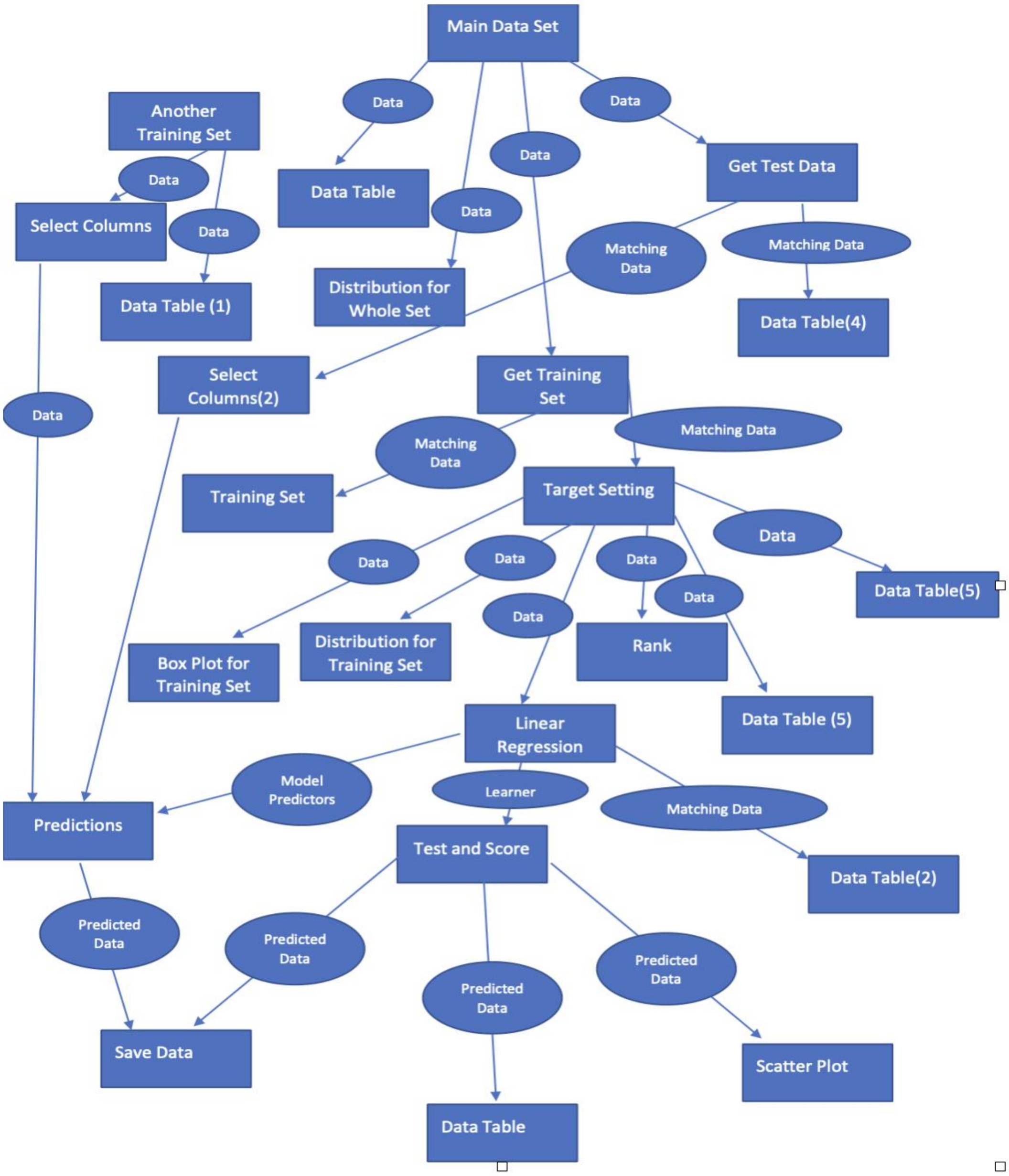


Figure 2. Model Design and Experimentation

3.5. Model Design and Experimentation

Linear regression was the machine learning algorithm used in this model because of its efficiency in building machine learning projects, scientifically proven and reliably predict the future and it being a long-established statistical procedure hence well understood and can be trained very quickly [6]. The model has six sections that are critical in its operation, these are: main data (where you feed processed and analysed energy dataset), another training set(where you feed energy dates to be predicted, 2020 energy consumption), get test data (2018-2019 energy data sets, to validate the model), get training data(Energy data from 2010-2017, this is to build the model), target settings (Which variables are critical in the data to be predicted) and rank section (To check the accuracy of the Model).

3.6. Equations

The linear regression is used in the model consisting of a predictor variable and a dependent variable related linearly to each other. In case the data involves more than one Independent variable, then linear regression is called multiple linear regression models. The below given equation is used in this research.

Multiple linear regression: Y = a + b1X1 + b2X2 + b3X3 + ... + btXt + u

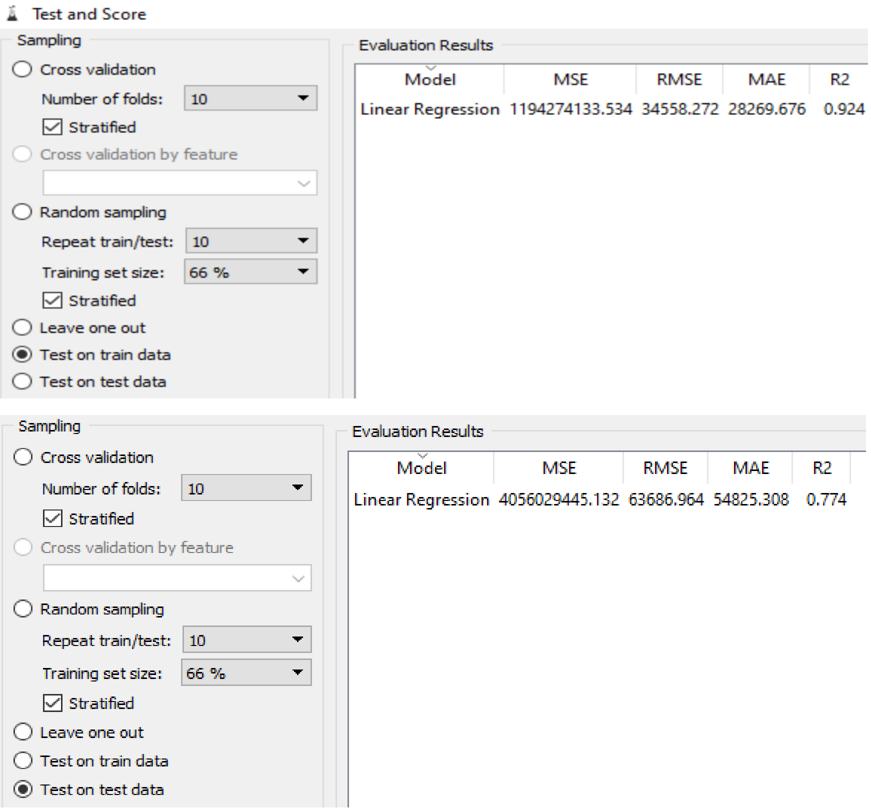
Where: Y = the variable that you are trying to predict (dependent variable). X = the variable that you are using to predict Y (independent variable). a = the intercept. b = the slope. u = the regression residual [7].

4. Findings

The results analysis presents the results generated from the model. There are nine results that are discussed in this part, these are results accuracy, model Predicted electrical consumption data, Microsoft Power BI Visualised: predicted electrical consumption by season status, predicted electrical consumption by exam status, predicted electrical consumption by school activity, predicted electrical consumption by student population, predicted electrical consumption by month and lastly predicted total electrical consumption for the year 2020.

4.1. Results Accuracy

The results accuracy was evaluated in Test and Score widget. Test and Score widget can be used to view your desired learning algorithms accuracy [8]. Mean Squared Error (MSE) of an estimator measures the average of the squares of the errors that is, the average squared difference between the estimated values and the actual value [9]. Root Mean Square Error (RMSE) is a frequently used measure of the differences between values (sample or population values) predicted by a model, or an estimator and the values observed [10]. Mean Absolute Error (MAE) is a measure of errors between paired observations expressing the same phenomenon [11]. Coefficient of determination score varies between 0 and 100% or 0 and 1. It is the proportion of the variance in the dependent variable that is predictable from the independent variable [12].



Test and Score

Sampling

○ Cross validation
Number of folds: 10
☑ Stratified
○ Cross validation by feature
○ Random sampling
Repeat train/test: 10
Training set size: 66 %
☑ Stratified
○ Leave one out
● Test on train data
○ Test on test data

Evaluation Results

Model	MSE	RMSE	MAE	R2
Linear Regression	1194274133.534	34558.272	28269.676	0.924

Sampling

○ Cross validation
Number of folds: 10
☑ Stratified
○ Cross validation by feature
○ Random sampling
Repeat train/test: 10
Training set size: 66 %
☑ Stratified
○ Leave one out
○ Test on train data
● Test on test data

Evaluation Results

Model	MSE	RMSE	MAE	R2
Linear Regression	4056029445.132	63686.964	54825.308	0.774

Figure 3. Test and Score Report

Figure 3 above shows the accuracy of the model, representing test on train data and test on test data, respectively. Table 1 summarises the accuracy results which is 92.4% and 77.4% on train data and test data, respectively. An accuracy of 0-40% is poor, while 40-65% is average, 65-85% is good and 85-100% id excellent [13].

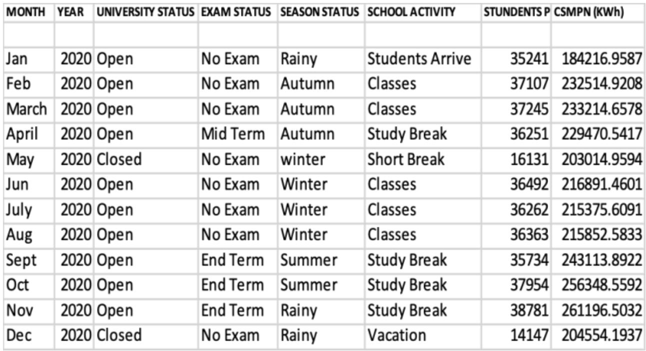
Table 1. R-Squared Report

DATA SET	R-SQUARED	(%)
Train Data	0.924	92.4%
Test Data	0.774	77.4%

4.2. Predicted Electrical Consumption Data from model

In table 2 below, the 2020 predicted electrical consumption data is presented which is based on exam status, university status, season status, season activity and student population.

Table 2. Predicted Electrical Consumption Data from Model



MONTH	YEAR	UNIVERSITY STATUS	EXAM STATUS	SEASON STATUS	SCHOOL ACTIVITY	STUNDENTS P	CSMPN (KWh)
Jan	2020	Open	No Exam	Rainy	Students Arrive	35241	184216.9587
Feb	2020	Open	No Exam	Autumn	Classes	37107	232514.9208
March	2020	Open	No Exam	Autumn	Classes	37245	233214.6578
April	2020	Open	Mid Term	Autumn	Study Break	36251	229470.5417
May	2020	Closed	No Exam	winter	Short Break	16131	203014.9594
Jun	2020	Open	No Exam	Winter	Classes	36492	216891.4601
July	2020	Open	No Exam	Winter	Classes	36262	215375.6091
Aug	2020	Open	No Exam	Winter	Classes	36363	215852.5833
Sept	2020	Open	End Term	Summer	Study Break	35734	243113.8922
Oct	2020	Open	End Term	Summer	Study Break	37954	256348.5592
Nov	2020	Open	End Term	Rainy	Study Break	38781	261196.5032
Dec	2020	Closed	No Exam	Rainy	Vacation	14147	204554.1937

5. Discussions of Results

5.1. Predicted Electrical Consumption Data by Season Status

The visualised data above indicates that cumulatively winter season in 2020 has the highest consumption followed by autumn, rainy and lastly summer. Consumption of. electricity rises in the winter from autumn because the days are shorter hence more lighting usage and some rooms heat with electricity usage, either for their primary heating equipment, such as electric furnaces or heat pumps, or with secondary heating equipment, such as space heaters or electric blankets [14]. Electrical consumption is generally high in the summer months when demand peaks in the afternoon as buildings and businesses are using air conditioning appliances on hot days [15].

5.2. Predicted Electrical Consumption Data by Exam Status

The visualised data above indicates that cumulatively during no exams in 2020 has the highest consumption followed by end of term examination period and lastly midterm examination period. Consumption of electricity are high when they are no exams because it is mostly. winter and autumn period, students spend more time in rooms this period and use heating appliances and more lighting [16]. During end term and midterm which is mostly summer and rainy season student spend less time in their rooms and don't use any heating appliances, moreover they spend more time in outdoor study areas and library to study during exams.

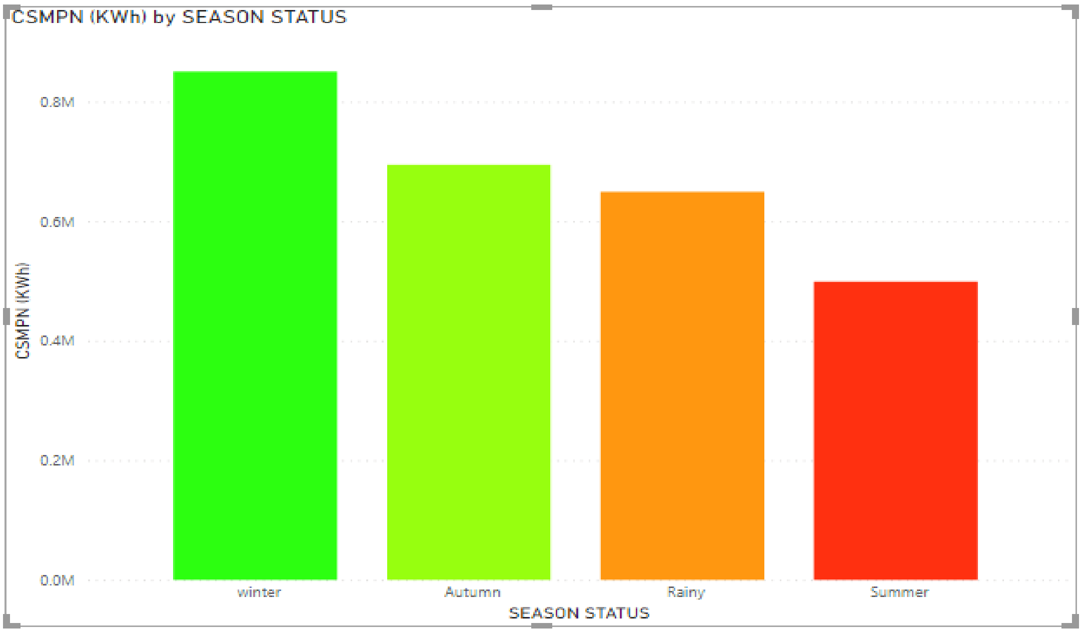


Figure 4. Data by Season Status

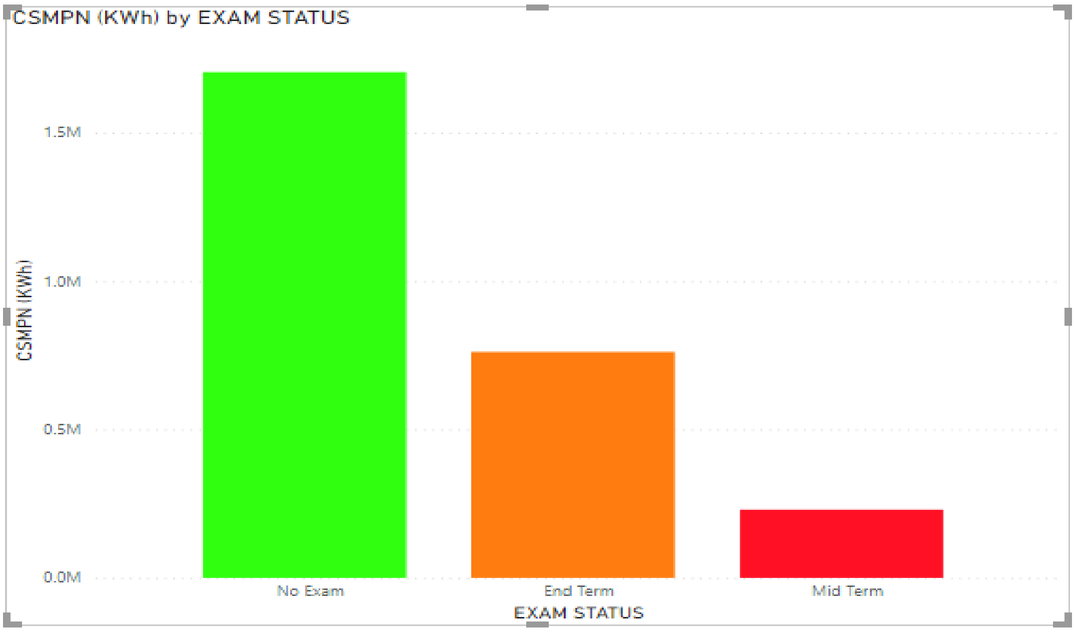


Figure 5. Data by Season Exam status

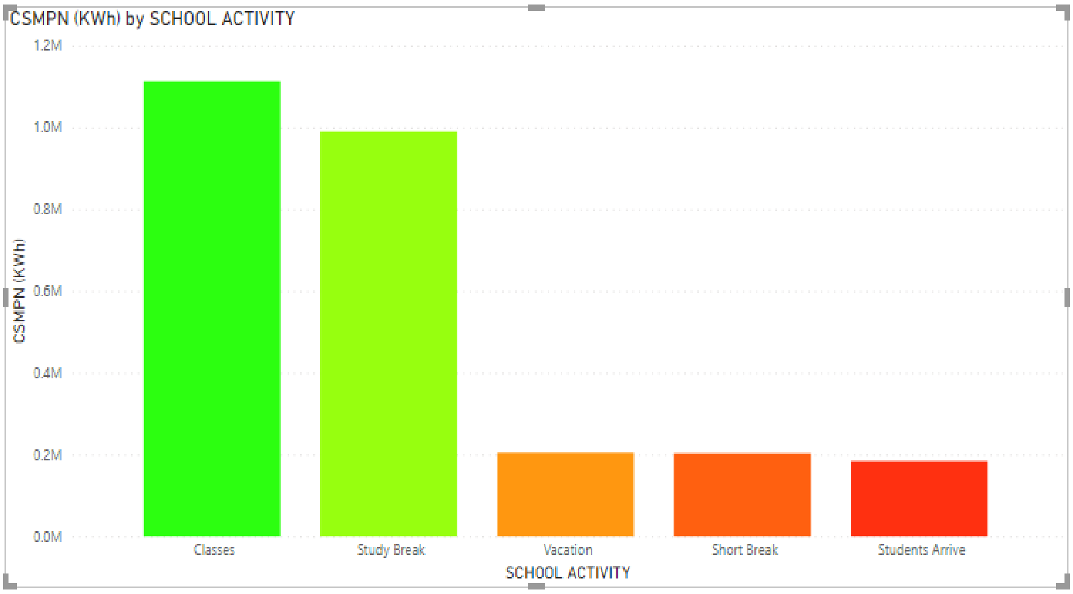


Figure 6. Data by Season School Activity

5.3. Predicted Electrical Consumption Data by School Activity

The visualised data above indicates that cumulatively during classes in 2020 has the highest consumption followed by study break, vacation, short break and lastly arrival of students. Consumption of electricity are high when they are no classes and on study break because it is mostly winter and autumn period as well, students spend more time in classes and dormitories during this period [17]. During vacation, short break and arrival of students which is mostly summer and rainy season student spend less time in their rooms and the student population reduces as students go on vacation.

5.4. Predicted Electrical Consumption Data by Student Population

The visualised data above indicates that when schools are open, and all students are on campus the electrical consumption is at its highest on comparison when schools are closed when the university as less students. Based on the collected data the university is closed twice in an academic year, that is during vacation and short breaks, during winter and autumn when we only recorded short breaks and no vacation but mostly during this period the university is open and has the highest number of students as a result electrical consumption is at its highest.

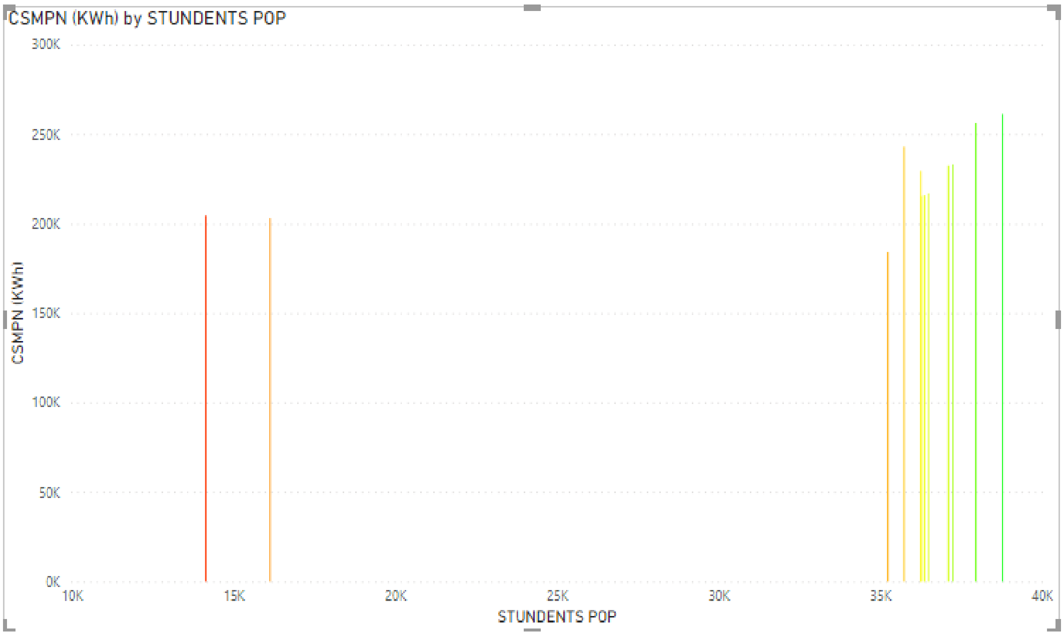


Figure 7. Data by Season Student Population

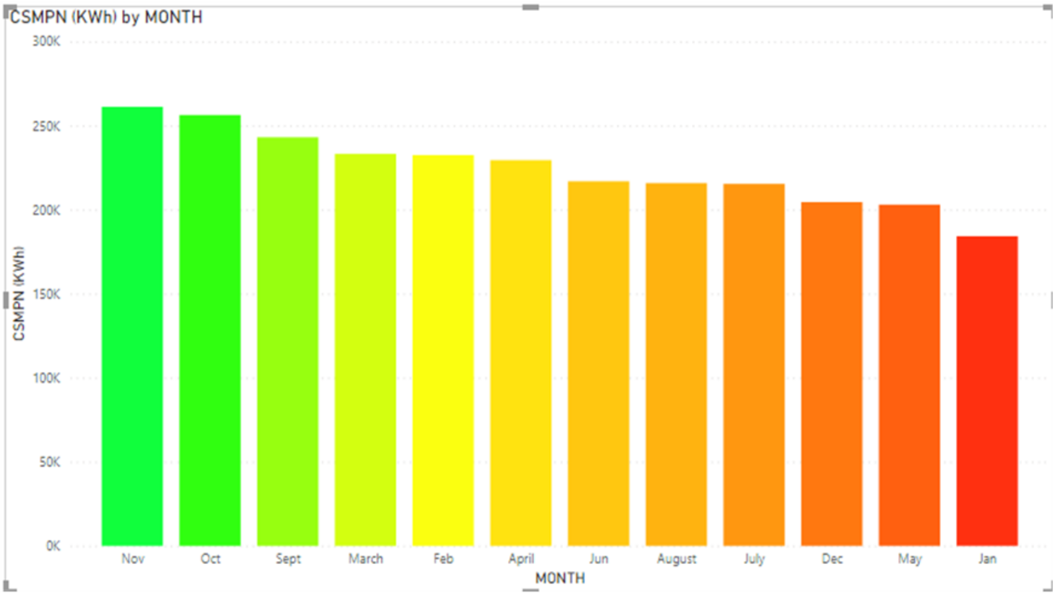


Figure 8. Data by Month

5.5. Predicted Electrical Consumption Data by Month

The visualised data above indicates that in the months of November, October and September have the highest consumption, these months are in summer, other months included during this season is January when students arrive hence less population and lesser consumption level, and December when students go on vacation but only a smaller number of distance students remain on campus. Total hourly electricity consumption is generally highest in the summer months when demand peaks in the afternoon as rooms are using air conditioning on hot days [18]. Based on the collected data March, February, April is Autumn and their only classes with student population high at during months. June, August, July, and May are categorised in winter period, schools are open during this period, classes are on hence student population is high and students spend more time in rooms During the winter months, hourly electricity load is less variable but peaks in both the morning and the evening [19].

6. Conclusions and Recommendations

This study has predicted electrical energy consumption using data science at the University of Zambia. Linear regression the suitable machine learning algorithm used in the formulated model achieved a higher accuracy in the prediction of 92.4% on train data and 77.4% on test data, this concentrated on the following eight (8) predicted areas.

Based on predicted data for 2020 obtained and the observations of this study and others, this research recommends the use of Data Science to predict electrical energy consumption in higher learning institutions. Microsoft Power BI provides clearer, and data driven visualisations and shown in chapter 4 of this study. In summary this study makes the following recommendations. The machine learning algorithm for the Data Science prediction model must be carefully selected to meet the objective based on the raw data. This study recommends that the software selected such as Microsoft Excel spread sheet, One Drive and Power BI work well in the Data Science life cycle. To achieve a higher accuracy with Data Science predictions more raw data is required. It is also recommended Anaconda open-source distribution software and orange tool kit latest versions should be installed on a computer with minimum 8GB RAM and core I5.

This work focused on total electrical energy consumption prediction at the university of Zambia. In future this work could be extended in other sectors such as the mines, agriculture health and finance sector where electrical energy consumption is high. Furthermore, future studies could look at better machine learning algorithms in detail and consider them in predicting electrical energy consumption.

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to God for giving me special grace to accomplish this work. This whole study was not without challenges yet at no time did I doubt his presence in the storm.

Secondly, I would like to thank my Research Supervisor, Dr. Grayson Himunzowa for his unfading support and guidance throughout my study and research period. I would like to thank my father, Mr. Raban Mtolo and mother, Mrs Pimpa Mtolo for unwaveringly supporting my study.

REFERENCES

[1] Mdpi," Energy Usage Measurement for usage usage," https://www.mdpi.com/1996-1073/11/2/452/pdf.

[2] Research gate, "Building Occupant Behaviour on Energy Efficiency and Methods,"https://www.researchgate.net/publication/324589706.

[3] Mdpi, "Data Mining Applications in Understanding Consumer behaviour," https://www.mdpi.com/2018-1073/12/22/4287/htm.

[4] Mdpi, "Data Science and the Energy Sector," https://www.Datascienceintheenergyandutilities.com.

[5] Data Driven Approaches, "Steam Load Prediction in Buildings," https://www. A data-driven approach for steam load prediction in buildings.com.

[6] Kumar Rohit, "linear regression modeling and assumptions" https://towardsdatascience.com -dcd7a201502a- -2018.

[7] Brian Beers, "Regression Models in Machine Learning" https://www.investopedia.com/terms/r/regression.asp.

[8] Ng Wai Foong "Test & Score widget can be used to test your desired learning algorithms on the dataset." https://www.linkedin.com/in/wai-foong-ng-694619185/,2019.

[9] Wikipedia "Mean Squared Error" https://en.wikipedia.org/wiki/-wikipedia-2018.

[10] Wikipedia "Root Mean Square Deviation"https://en.wikipedia.org/wiki/-wikipedia-2018.

[11] Wikipedia "Mean Absolute Error" https://en.wikipedia.org/wiki/-wikipedia-2018.

[12] Walker. R "Mean Squared Error and variance in regression analysis" https://www.bmc.com/blogs/mean-squared-error-r2-and-variance-in-regression-analysis/-2019.

[13] Kok Wei Khong "predictive accuracy and precision threshold." https://www.predictiveaccuracy.li.org.

[14] Tyler Hodge "Hourly electricity consumption varies throughout the day and across seasons," https://www.eia.gov/todayinenergy/detail.php?id=42915-2020.

[15] Nilay Manzagol" Winter residential electricity consumption expected to increase from last winter," https://www.eia.gov/todayinenergy/detail.php?id=29112-2019.

[16] Tyler Hodge "Hourly electricity consumption varies throughout months," https://www.eia.gov/todayinenergy/detail.php?id=42915-2020.

[17] Tyler Hodge "Hourly electricity consumption varies by population https://www.eia.gov/todayinenergy/detail.php?id=42915-2020.

[18] Nilay Manzagol" Winter residential electricity consumption expected to increase by season and month," https://www.eia.gov/todayinenergy/detail.php?id=29112-2019.

[19] Nilay Manzagol" Winter residential electricity consumption expected to increase by day," https://www.eia.gov/todayinenergy/detail.php?id=29112-2019.