Electricity consumption prediction

Group Four

Group Members

Lean Nyaboke

Derrick Kuria

Paul Mwaura

Melisa Michuki

Business overview

Electrical energy is a form of energy resulting from the flow of electric charge. Electrical energy may be either potential energy or kinetic energy, but it's usually encountered as potential energy, which is energy stored due to the relative positions of charged particles or electric fields.

People use electricity for lighting, heating, cooling, and refrigeration and operating appliances, computers, electronics, machinery, and public transportation systems. The industrial sector generates and uses nearly all of the direct use of electricity.

This leads to us demystifying electric energy consumption. Electric energy consumption is the actual energy demand made on the existing electricity supply. Due to demand, electrical energy producers sell this product per Kilowatt depending on its usage.

From the data collected, we have observations of customers ranging from household users who use very little electrical energy to industrial users who use a lot of it.

However, there are many sources of energy that are also used but biomass from plants is a very common source since it is cheap and easily accessible to the common mwananchi.

Business objective

Electricity suppliers would stand to save millions of shillings if they would decrease their peak demand charge. This can only be possible based on the ability to predict electricity consumption on a daily basis. Our project is aimed at solving this problem by coming up with a solution that can predict electricity consumption on an hourly basis. We will rely on natural factors to build the model. These factors include; temperature, pressure and wind speed.

Other objectives include:

- To highlight current electricity consumption by customers
- To predict the required electrical energy to be produced
-
- To identify opportunities based on the customers' behaviour to optimize electrical energy consumption
- To identify patterns in electrical energy consumption

Definition of Terms

Some of the common technical terms that you will come across include;

- **Energy** is measured in J and kJ. **Power** is the rate of using energy and is measured in W and kW. Fuel bills show energy used in kWh, and the cost of this can be calculated if the cost per kWh is known.
- **Current** is the rate of flow of electric charge. A potential difference (**voltage**) across an electrical component is needed to make a current flow through it. An electric current flows when **electrons** move through a conductor, such as a metal wire. The moving electrons can collide with the **ions** in the metal. This makes it more difficult for the current to flow and causes **resistance**.

The Temperature Effect

Temperature affects how electricity flows through an electrical circuit by changing the speed at which the electrons travel. This is due to an increase in resistance of the circuit that results from an increase in temperature. Likewise, resistance is decreased with decreasing temperatures. Therefore, a high resistance leads to a low current while a low resistance leads to high current.

Power(P) = Voltage(V) * Current(I)

So what does this mean for us? Having a higher temperature would mean that the resistance will be high hence, the current will be low. Thus, an increase in temperature would, in turn, lead to lower electricity consumption and a decrease in temperature leads to an increase in electricity consumption.

Voltage(V) = Current(I) * Resistance(R)

From the above two formulas, we can deduce the following formula.

Power(P) = Current(I) * Current(I) * Resistance(R)

Each of these quantities is measured using different units:

- Voltage is measured in volts (V)
- Current is measured in amps (A)
- Resistance is measured in ohms (Ω)
- Power is measured in watts (W)

The Pressure Effect

Resistivity is manifested by scattering of electrons and phonons; a temperature increase results in an escalation of the number of electron-phonon and electron-electron scattering events leading to an increase in electrical resistivity.

However, pressure has the opposite effect; It reduces the inter-atomic spacing and the atomic vibration amplitude resulting in decreasing electrical resistivity with increasing pressure. Therefore, this means that an increase in pressure leads to an increase in electricity consumption.

NB/

Pressure effects on electric current might be quite low to notice but it does affect the resistance to the flow of current in a conductor.

Importance of solving this problem

The power and lighting company would like to understand the different environmental factors that would cause electricity consumption to fluctuate when other factors are constant (ceteris paribus).

Other than the environmental impact on energy production and propagation, there could be other factors that cause electricity consumption to vary.

These causes could be:

1. The client Factor - People could use various dubious and illegal methods to modify and amend energy consumption by avoiding electricity meters and Tokens to consume power directly.
2. The government factor- Governments may offer regulations that will affect supply and demand.

Impact of this project

This project is done to provide solutions for the lighting to grasp the energy spectrum in terms of external factors such as temperature and pressure. Through this project, the Lighting Company can understand the following:

1. The trend of Energy and electricity consumption between the year 2013 and 2017.
2. The impact of each of the environmental factors on electricity consumption.
3. How to leverage the project results to improve energy production and propagation to the clients and end-users.
4. An accurate prediction also allows us to make better decisions in terms of cost and energy efficiency

Research questions

This involves formulating questions that define the business goals that the data science techniques can target. Based on our research, we came up with the following questions;

1. How does temperature affect electricity Consumption?
2. How does pressure affect electricity Consumption?
3. How does wind speed affect electricity Consumption?
4. How does var 1 affect electricity Consumption?
5. How does var 2 affect electricity Consumption?
6. What is the trend and the stationarity in electricity consumption between 2013 to 2017?
7. Which model would be appropriate for making predictions from the data provided?

Project management

Resources

- Dataset
    - We obtained our data from an electricity generation company supplies electricity to a Ugandan city. This can be found in the following link [Electricity Consumption Predictor]

- Software
    - Google collabs- Data analysis, visualisations and modelling
    - Github - The main repository for our work
    - Google Docs - Project documentation
    - Trello, Google sheets - Project management

Assumptions

To carry out our analysis, we made a few assumptions on our data:

- The data provided by the company was accurate and up to date.
- The data were consistent
- Data collection was comprehensive and reliable data collection methods were used.

Constraints

Some of the data was concealed due to data privacy and this proved challenging to make sense of how it impacted our analysis.

Data understanding

The dataset provided had 8 variables and 26496 observations. Data was mainly collected from 2013-07-01 to 2017-06-23. The dataset has 8 variables: ID, temperature, pressure, windspeed, var1, var2, datetime, and electricity consumption (in MWh).

The data will be split into two with the training set having the first 23 days of every month and the test set having the 24th day to the end of the month.

The table below gives a description of the variables given;

Column Name	**Description**
ID	Unique ID
datetime	Date and time info on record
Temperature	Temperature at that hour
var1	Anonymized feature variable 1
var2	Anonymized feature variable 2
pressure	Pressure at that hour
windspeed	Wind speed at that hour
electricity_consumption	Electricity consumption in MWh

Data preparation

This includes cleaning the data and selecting the variables that would be used to conduct our univariate, bivariate and multivariate analysis.

Data cleaning was the first step taken in the preparation of our data. During the cleaning process, we found out that there were no duplicates and null values. We decided to split the datetime column to obtain the exact date and time in separate columns and dropped the original column. The reason as to why we are splitting this column is because we need to analyze how electricity consumption varies on an hourly basis.

Moreover,the date and time column was further formatted to fit the datetime data type and we drew out the specific date to ensure that we could also identify common holidays and dropped since it was assumed that more electricity was consumed during these days hence would affect our analysis.

The data set was split into numerical and categorical data. Using the numerical data, we were able to list and plot out the outliers and found that the electricity consumption columns had the highest number of outliers, making up to 1.8% of the data.Most of the outliers were dropped to prevent them from affecting the central tendencies.

Data Analysis

After carrying out data cleaning and making the data fit the format we needed it to be in,we carried out our analysis.The analysis is broken into three stages as shown below.

Univariate Analysis

1.) Central tendencies

From our analysis of the central tendencies of each variable,we were able to find out that during the seventh month of the year we have a spike in electricity consumption.

Moreover,during the month of January we see a slightly lower record of electricity consumption.

We also looked into the mean consumption on a daily basis where we found out that there is an upward trend toward the end of the week in electricity consumption. Electricity consumption increases from Friday up to Monday where consumption reduces. Therefore Wednesday has the least mean slightly above 290MW, while Monday has the highest consumption, slightly above 306MW.

In addition, when analysed the mean consumption on an hourly basis,we found out that electricity was consumed most from around 8pm to around 2 am dropping off during the day with the time when it is least consumed being 3 pm.

2.) Frequency distributions

We carried out analysis of some of the categorical data.When plotting out the Var2 ,we found out A was the most frequent value in the column.

We also looked into the consumption distribution of electricity where we found out that the electricity consumption frequency plot is skewed to the left, with the majority of the consumption between 200MW and 400MW.This was found to be the same even after looking into the quartiles since all the plots were skewed to the left.

Bivariate Analysis

1.) Pearson Correlation Coefficient

Pearson's Correlation Coefficient helps to find out the relationship between two quantities.When carrying out our analysis,we found out that var1 was positively correlated to temperature while pressure and temperature were negatively correlated

2.) R Squared Correlation

R^2,like correlation, tells you how related two things are. However, we tend to use R^2 because it's easier to interpret. R^2 is the percentage of variation (i.e. varies from 0 to 1) explained by the relationship between two variables.

Below is a table containing the R^2 scores of different values;

:

	Features	r**2
4	windspeed	5.706505e-02
2	var1	1.793283e-02
1	temperature	1.374850e-02
12	Dayofmonth	9.596925e-03
10	Q	2.558151e-03
16	Peak	2.414006e-03
15	Work	2.414006e-03
11	Dayofyear	2.166563e-03
13	Weekofyear	1.891069e-03
8	Month	1.661453e-03
14	Holiday	4.858659e-04
6	Hour	4.581254e-04
9	Year	3.947586e-04

17	Weekend	3.611049e-04
0	index	7.919861e-05
5	var2	2.242213e-05
3	pressure	8.219949e-07
7	Day	1.294018e-07

3.) Average electricity consumption by day of the month

We looked into the average energy consumption by day of the month and found that there was a sharp drop in its consumption on the 10th day and the peak days were from the 19th to the 20th day.

4.) Average electricity consumption by month

We also looked into the average energy consumption across the year where we discovered the months between July & October have peak electricity consumption rates with August taking the lead.

5.) Average electricity consumption by day of the week

In addition to the above analysis,we found that Wednesday was the day with the least electricity consumption followed closely by Friday .The weekends and Mondays got higher rates as compared to the rest of the week.

6.) Average electricity consumption by hour

We found that during the night ,from around 9pm to around 1am ,people tended to consume more electricity followed by a sharp drop towards 3 pm.

7.) Effect of temperature on electricity consumption

We plotted the average temperature and electricity consumption and based on the data we had,the region of focus experienced winter from June to September.From our analysis, we were able to derive that an increase in temperature would, in turn, lead to lower electricity consumption and a decrease in temperature leads to an increase in electricity consumption.

Modelling

Our project was mainly a time series problem where we were to work on making a model which would make electricity consumption predictions on an hourly basis.
We also came across other models that could be used and found that we could work with LSTM(Long short-term memory) and XGBoost because they could work with multiple variables.

Seasonality and Stationarity

Before we carried out modelling, we had to check for autocorrelation and seasonality since they would have affected our model.

When checking for seasonality,we found that the trend in electricity consumption is almost stationary as it neither increases or decreases.We also looked into the autocorrelation and partial autocorrelation and found that even after 40 lages, the line does not get inside the Confidence Interval meaning the data does not have seasonality.

Univariate models

- ARIMA

Autoregressive integrated moving average, or ARIMA, is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends.This model is a univariate model as it only takes one variable/feature at a time.It performed poorly because the predictions are way off from the actual values.

- SARIMA

It is a better model as compared to ARIMA as it combines the ARIMA model with the ability to perform the same autoregression, differencing, and moving average modeling at the seasonal level.

For this model,it performed relatively better than ARIMA.However,since it is a univariate time series model, hence it can only allow for one feature variable

The benefit of using this model is because it might be very useful and might provide you with more insights concerning your dependent variable. in our case(Electricity Consumption).

Since the above models did not satisfy the problem statement,we decided to explore some multivariate models.

Multivariate models

- VARMAX

  The VARMAX class in statsmodels allows estimation of VAR, VMA, and VARMA models (through the order argument), optionally with a constant term (via the trend argument).

- Vector Autoregression(VAR)

  The Vector Autoregression (VAR) method models the next step in each time series using an AR model. It is the generalization of AR to multiple parallel time series, e.g. multivariate time series.

- Vector Moving Average(VMA)

  We leave out the exogenous regressor but now include the constant term.

- Vector Autoregression Moving Average(VARMA)

  This method models the next step in each time series using an ARMA model. It is the generalization of ARMA to multiple parallel time series, e.g. multivariate time series.

However,for the above models,they work best when dealing with fewer feature variables, therefore, it also takes a long time for the model and the more the number of features,the longer it takes to train the model.

- XG BOOST

  XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework.

  For this model in particular,we had to fine-tune it by trying to use reasonable values based on the MSE.This model performed relatively well.The accuracy is at **82.44%** and the rmse is **82.44.**

Feature Importance

We plotted out the feature importances and from the chart, we could see the importance of the feature for our prediction in descending order.

With this plot we were able to identify some of the features that we needed to drop.These included ; var2,month,Q,peak and weekend.

After dropping the above features,we were able to build a much better model.

This model performed better.The accuracy is at **80.57%** and the rmse is **82.44.**

Below is a table with the predictions of the first five rows of the datasets

	Date_time	Actual	Prediction
0	2016-10-01 01:00:00	264.0	232.904205
1	2016-10-01 02:00:00	252.0	275.941956
2	2016-10-01 03:00:00	231.0	260.814941
3	2016-10-01 04:00:00	246.0	261.100494
4	2016-10-01 05:00:00	252.0	228.843979

- LSTM

Sequential class model is a linear stack of Layers. You can create a Sequential model and define all of the layers in the constructor.

We have used;

1.LSTM Layer

2.Dense Layer

Why are we using LSTM Layer?

Typical RNN uses information from the previous step to predict the output.But if only the previous step is not enough,that is long term dependency.If we use RNN using all previous steps ,the explosion/vanishing gradient problem is encountered.

LSTM can solve this problem, because it uses gates to control the memorizing process.

We defined the LSTM with 128 neurons in the first hidden layer and 1 neuron in the output layer for predicting electricity consumption.

This model had an accuracy of **74.62%** and rmse **126.69.**This was generally a good score.

The above work can be found through the following link: [Electricity consumption predictor model]

Conclusion

XGB model performed quite well with an

- RMSE score of 82.44%
- MAPE of 19.4%

LSTM Performance:

- RMSE score of 126.69
- MAPE of 25.39%

From the above summary of the models used,we came to a conclusion that XG Boost would be a better model since XGB had a lower MAPE compared to LSTM.

That said, we trust the XGBoost model performance more than the LSTM model.

However,more Hyperparameter tuning and also cross validation can be done in a bid to reduce the Mean Percentage error of the model.

Recommendations

From our analysis,here are the recommendations we came up with;

- Advise the company when to subsidize and to add the cost of electricity bills using the peak and off peak months for profitability.

- Incase of energy inadequacies , the company knows the best time for power rationing during the day which is between 11 AM - 6 PM when the electricity is relatively consumed.

- The company now comprehends the average consumption and can allocate resources to areas that may have low consumption may be due to high costs.

Future Plans

Some of the things we would like to carry out in the future with this project are ;

- Implement the project for a lighting company.
- Find a dataset that looks into other aspects that affect electricity consumption other than environmental factors eg government regulations and global markets.